Automated Feature & Label Refinement in the Context of Environmental Monitoring by Multi-Modal Satellite Data

---

Sarah Michelle Hauser

Doctoral Thesis
Karlsruhe, 2025

Automated Feature & Label Refinement in the Context of Environmental Monitoring by Multi-Modal Satellite Data

Zur Erlangung des akademischen Grades einer

DOKTORIN DER INGENIEURWISSENSCHAFTEN (Dr.-Ing.)

von der KIT-Fakultät für
Bauingenieur-, Geo- und Umweltwissenschaften
des Karlsruher Instituts für Technologie (KIT)

genehmigte

DISSERTATION

von

Sarah Michelle Hauser
aus Peterskirchen

Tag der mündlichen Prüfung:	30.07.2025
Referent:	Prof. Dr.-Ing. Stefan Hinz Institut für Photogrammetrie und Fernerkundung Karlsruher Institut für Technologie
Korreferent:	Prof. Dr.-Ing. Andreas Schmitt Institut für Anwendungen des maschinellen Lernens und intelligenter Systeme Hochschule München für angewandte Wissenschaften

Karlsruhe (2025)

Sarah Michelle Hauser

Automated Feature & Label Refinement in the Context of Environmental Monitoring by Multi-Modal Satellite Data

Doctoral Thesis
Date of examination: 30.07.2025
Referees:
Prof. Dr.-Ing. Stefan Hinz
Prof. Dr.-Ing. Andreas Schmitt

Karlsruhe Institute of Technology
Department of Civil Engineering, Geo and Environmental Sciences
Institute of Photogrammetry and Remote Sensing
Kaiserstr. 12
76131 Karlsruhe



This document is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0):
https://creativecommons.org/licenses/by-nc-sa/4.0/deed.en

Abstract

This thesis addresses a central challenge in remote sensing and Earth Observation: how to translate multi-modal, multi-temporal EO data into ecologically meaningful, spatially transferable, and operationally robust insight. While EO has become a cornerstone of environmental monitoring, its integration into machine learning systems remains hindered by issues of sensor heterogeneity, temporal misalignment, and the under-structuring of label data. Rather than treating EO–ML as a purely technical pipeline, this work frames it as a system of environmental inference, one requiring careful alignment between observation, representation, and interpretation. Rather than advancing a single technique, this thesis develops a modular but conceptually unified approach to EO–ML system design, one that integrates feature fusion, enriched supervision, and structured evaluation as co-evolving components. It argues that accuracy and generalisation in EO–ML are shaped less by the depth of models than by the ecological coherence of what they learn from. Across case studies, improvements in predictive performance consistently stemmed from designing inputs and labels that reflect environmental processes, phenological timing, spatial structure, and uncertainty, rather than from increasing algorithmic complexity alone. Empirical analyses span three environmental domains, arid sinkhole-prone landscapes, temperate forests, and cryospheric glacier systems, each demonstrating how environmental processes can be better modelled through compositional rather than singular EO design logic. The Combined Doline Vegetation Index exemplifies this approach, capturing functional contrast between SAR and optical data across ecologically distinct seasons. Likewise, the novel HELIX framework redefines label construction by embedding spatial, temporal, and residual-based context into the supervision layer, enabling models to learn from structured uncertainty and spatial dynamics. Benchmarking results confirm that domain-aware fusion strategies and supervision-enriched labels yield models that are more accurate, interpretable, and transferable, especially under spatial domain shifts. The findings support a broader argument: that EO-based environmental modelling must move from superficial correlation toward structural alignment, where features, labels, and models co-evolve in relation to environmental processes. Ultimately, this thesis reframes EO–ML not as a fixed algorithmic process, but as a grammar of environmental representation, where syntax (features), semantics (labels), and inference (models) must

be co-designed to reflect the dynamic, uncertain, and structured nature of the Earth systems we seek to understand.

Kurzfassung

Diese Arbeit befasst sich mit einer zentralen Herausforderung in der Fernerkundung und Erdbeobachtung: Wie lassen sich multimodale, multitemporale EO-Daten in umweltrelevante, räumlich übertragbare und operationell robuste Erkenntnisse umsetzen? Während die Erdbeobachtung zu einem Eckpfeiler der Umweltüberwachung geworden ist, wird ihre Integration in Systeme des maschinellen Lernens nach wie vor durch die Heterogenität der Sensoren, die zeitliche Fehlanpassung und die unzureichende Strukturierung der Labeldaten behindert. Anstatt EO-ML als eine rein technische Pipeline zu behandeln, wird es in dieser Arbeit als ein System für umweltbezogene Rückschlüsse betrachtet - eines, das eine sorgfältige Abstimmung zwischen Beobachtung, Darstellung und Interpretation erfordert. Anstatt eine einzelne Technik voranzutreiben, wird in dieser Arbeit ein modularer, aber konzeptionell einheitlicher Ansatz für die Entwicklung von EO-ML-Systemen entwickelt, der die Merkmalsfusion, die erweiterte Überwachung und die strukturierte Auswertung als sich gemeinsam entwickelnde Komponenten integriert. Es wird argumentiert, dass die Genauigkeit und Verallgemeinerung in EO-ML weniger durch die Tiefe der Modelle als durch die umweltbezogene Kohärenz dessen, woraus sie lernen, bestimmt wird. In allen Fallstudien ergaben sich Verbesserungen der Vorhersageleistung durchgängig aus der Gestaltung von Eingaben und Bezeichnungen, die Umweltprozesse, phänologische Zeitpunkte, räumliche Strukturen und Unsicherheiten widerspiegeln, und nicht aus der Erhöhung der algorithmischen Komplexität allein. Empirische Analysen erstrecken sich über drei Umweltbereiche: trockene, von Dolinen geprägte Landschaften, Waldregionen in gemäßigten Breiten und Gletschersysteme in der Kryosphäre, die jeweils zeigen, wie Umweltprozesse durch eine kompositorische statt einer singulären EO-Designlogik besser modelliert werden können. Der kombinierte Doline-Vegetationsindex ist ein Beispiel für diesen Ansatz, der den funktionalen Kontrast zwischen SAR- und optischen Daten über ökologisch unterschiedliche Jahreszeiten hinweg erfasst. Ebenso definiert das neuartige HELIX-Framework die Konstruktion von Labels neu, indem es den räumlichen, zeitlichen und auf Residuen basierenden Kontext in die Supervisionsschicht einbettet und es den Modellen ermöglicht, aus strukturierter Unsicherheit und räumlicher Dynamik zu lernen. Benchmarking-Ergebnisse bestätigen, dass bereichsspezifische Fusionsstrategien und mit Überwachung angereicherte Labels zu Modellen mit einer höheren Genauigkeit,

Interpretierbarkeit und Übertragbarkeit führen, insbesondere bei räumlichen einer räumlichen Übertragung. Die Ergebnisse stützen ein breiteres Argument: dass die EO-basierte Umweltmodellierung von der oberflächlichen Korrelation zur strukturellen Ausrichtung übergehen muss, bei der sich Merkmale, Labels und Modelle in Bezug auf die Umweltprozesse gemeinsam entwickeln. Letztlich betrachtet diese Arbeit EO-ML nicht als einen festen algorithmischen Prozess, sondern als eine Grammatik der Umweltdarstellung, bei der Syntax (Merkmale), Semantik (Bezeichnungen) und Inferenz (Modelle) gemeinsam entwickelt werden müssen, um die dynamische, unsichere und strukturierte Natur der Erdsysteme widerzuspiegeln, die wir zu verstehen versuchen.

Acknowledgement

I am deeply grateful for the past three and a half intense and formative years of my Ph.D. journey, which began in early 2022 and concludes now in mid-2025. Throughout this time, I've been fortunate to be surrounded by inspiring, supportive, and intellectually curious people who shaped both my work and my perspective. While this thesis is my own, the research it presents would not have been possible without the people I've had the privilege to meet along the way. I cannot name everyone who has inspired and encouraged me during these years, but I would like to acknowledge those who had the most significant impact.

First and foremost, I thank Andreas Schmitt for supervising my Ph.D. at the Hochschule für angewandte Wissenschaften München, within the Institut für Anwendungen des maschinellen Lernens und intelligenter Systeme (IAMLIS). Andreas gave me the freedom to explore my research interests, encouraged me to present at international conferences, and provided consistently valuable and constructive feedback throughout the process. Having known him since my bachelor's studies, where he first inspired me to pursue remote sensing, I feel incredibly lucky that our paths continued to cross: from my bachelor's thesis, through his presence during my master's thesis at another institution, and finally back to Hochschule München for the Wald5Dplus project. Working with Andreas again, now as a Ph.D. student, felt like coming full circle. Words of thanks hardly feel adequate for the guidance, inspiration, and support he provided.

I would also like to sincerely thank Stefan Hinz from the Institute of Photogrammetry and Remote Sensing (IPF) at the Karlsruhe Institute of Technology (KIT) for his supervision. He fully supported the research direction I pursued and contributed valuable structural ideas as well as helpful feedback throughout the process.

A very special thanks goes to Anna Wendleder, who not only supported me with invaluable feedback on glacier dynamics but also took on the significant task of preprocessing vast amounts of EO data for this thesis. Having first met her during my bachelor's thesis and later working with her more intensively during my master's thesis (where she acted as co-supervisor), I remain grateful for her continued mentorship and support.

I am also especially thankful to Simone Aigner and Christine Hechtl, with whom I worked closely and intensively, not only during their theses but also beyond, sharing many discussions and benefiting from their open ears and valuable perspectives. I also gratefully acknowledge the important contributions from Michael Ruhhammer and Markens Spasari, whose work and results were a valuable part of this research journey.

On a more personal note, I owe heartfelt thanks to my family. My sister, Lisa Hauser, deserves special mention, not only for her emotional support but also for being an ever, patient sounding board for many of my ideas. In fact, one particularly vivid debate with her about the real-world challenges of spatially and temporally messy label data directly sparked the conceptual foundation for what later became the HELIX framework, originally known, half-jokingly, as the "voxel carousel", referring to the idea of label information moving through space and time like data hopping along a rotating track. Finally, I want to thank my parents, Rosemarie and Hermann Hauser, for their endless support, encouragement, and belief in me throughout this long academic path. Without their steadfast backing, this thesis, and the journey it represents, would not have been possible.

Contents

1 Introduction **1**

1.1 Motivation for Remote Sensing in Environmental Monitoring 3

1.1.1 From Spectral Indices to Feature Engineering 5

1.1.2 Machine Learning Methods 8

1.1.3 The Need for Benchmarking 15

1.2 Application Domains and Labelled Datasets 20

1.2.1 Southwestern Kazakhstan 20

1.2.2 Temperate Central European Forests 29

1.2.3 Canadian High Arctic 47

1.3 Main Objective and Research Goals 57

1.4 Thesis Outline and Contributions 59

2 Consistent EO Feature Generation **65**

2.1 Data Fusion Approaches 66

2.1.1 Pixel-Level Fusion 69

2.1.2 Feature-Level Fusion 70

2.1.3 Decision-Level Fusion 72

2.1.4 Temporal-Aware Fusion 74

2.2 Hypercomplex Bases 77

3 Labelling Foundations and Challenges **89**

3.1 General Differences in Label Preparation by Model Type 90

3.2 Nature and Temporality of Labels 93

3.3 Challenges in Dynamic Labelling 96

3.4 Methodologies in Data Labelling and Processing 97

3.5 Understanding Challenges and Best Practices in Dynamic Data Processing for Labelling 102

3.6 Dynamic Labelling and Sampling Strategies: Temporal and spatio-temporal Perspectives 115

4 The Novel Helix Framework for Dynamic Label Data 119
4.1 Framework Formalization and Design Principles . . . 122
4.1.1 Hybrid Integration of Static and Dynamic Labels . . . 122
4.1.2 Spatio-temporal Scale Reconciliation . . . 123
4.1.3 Spatio-temporal Label Enrichment and Engineering . . . 128
4.1.4 Structured Learning Targets . . . 135
4.1.5 Operational Design for Scalability and Integration . . . 136
4.2 Future Directions . . . 138

5 Temporal Dynamics in EO Feature Engineering 147
5.1 Comparative Evaluation of Temporal Fusion Settings . . . 148
5.1.1 Materials . . . 149
5.1.2 Methods . . . 149
5.1.3 Results . . . 150
5.1.4 Discussion . . . 155
5.2 Combined Doline Vegetation Index . . . 157
5.2.1 Materials . . . 157
5.2.2 Methods . . . 158
5.2.3 Results . . . 160
5.2.4 Discussion . . . 164
5.3 Conclusions . . . 165
5.3.1 Lessons Learned . . . 166
5.3.2 Research Questions Revisited . . . 167
5.3.3 Closing Remarks . . . 168

6 Foundational Analysis of EO Modality–Model Interactions 169
6.1 Polarimetrically, Spectrally and Temporally Fused Sentinel-1 and Sentinel-2 Data . . . 180
6.1.1 Materials . . . 180
6.1.2 Methods . . . 183
6.1.3 Results . . . 186
6.1.4 Discussion . . . 203
6.2 Polarimetrically and Spectrally Fused Sentinel-1 and Sentinel-2 Data . . . 209
6.2.1 Materials . . . 209
6.2.2 Methods . . . 211
6.2.3 Results . . . 212
6.2.4 Discussion . . . 214

6.3 Reflectance Bands and Spectral Kennaugh-like Elements from Sentinel-2 Data . . . 216
6.3.1 Materials . . . 216
6.3.2 Methods . . . 217
6.3.3 Results . . . 218
6.3.4 Discussion . . . 224
6.4 Polarimetric Kennaugh Elements from Sentinel-1 Data . . . 228
6.4.1 Materials . . . 228
6.4.2 Results . . . 228
6.4.3 Discussion . . . 234
6.5 Polarimetric Kennaugh Elements from TerraSAR-X and ALOS-2 Data . . . 236
6.5.1 Materials . . . 237
6.5.2 Methods . . . 237
6.5.3 Results . . . 239
6.5.4 Discussion . . . 241
6.6 Conclusions . . . 244
6.6.1 Lessons Learned . . . 244
6.6.2 Research Questions Revisited . . . 248
6.6.3 Closing Remarks . . . 252

7 Context-Aware Label Enrichment and Multi-Scale Learning with the HELIX Framework 255
7.1 Context-Aware Forest Structure Modelling Using Polarimetrically, Spectrally, and Temporally Fused Sentinel-1 and Sentinel-2 Data with Helix-Enriched Multi-Scale Labels . . . 257
7.1.1 Materials . . . 258
7.1.2 Methods . . . 258
7.1.3 Results . . . 262
7.1.4 Discussion . . . 271
7.1.5 Conclusions . . . 273
7.2 Forest Disturbance Forecasting from Fused Sentinel-1 and Sentinel-2 Data with Helix-Based Spatio-Temporal Label Enrichment . . . 276
7.2.1 Materials . . . 277
7.2.2 Methods . . . 278
7.2.3 Results . . . 289
7.2.4 Discussion . . . 300

7.2.5 Conclusions . . . 306
7.3 Seasonal Glacier Facies Forecasting from Temporally Fused Sentinel-1 Data and Helix Labels . . . 309
7.3.1 Materials . . . 310
7.3.2 Methods . . . 313
7.3.3 Results . . . 324
7.3.4 Discussion . . . 330
7.3.5 Conclusions . . . 359
7.4 Glacier Zone Change Forecasting from Polarimetrically and Spectrally Fused Sentinel-1 and Sentinel-2 Data with HELIX Temporal Supervision . 364
7.4.1 Materials . . . 365
7.4.2 Methods . . . 368
7.4.3 Results . . . 373
7.4.4 Discussion . . . 378
7.4.5 Conclusions . . . 383

8 Conclusions and Outlook 387
8.1 Overview and Reflections on the Research Journey . . . 387
8.2 Evaluating Multi-Modal and Multi-Temporal EO Predictive Capacity . . . 389
8.3 Temporal Fusion Strategies and Design . . . 392
8.4 Structuring Supervision - The HELIX Framework for Label Enrichment . . 393
8.5 Interpretation of Kennaugh Elements . . . 395
8.6 Limitations and Methodological Caveats . . . 403
8.7 Future Work and Research Directions . . . 405
8.8 Final Reflections - Learning from and with EO . . . 408

Declaration of Supportive Resources 411

Bibliography 413

List of Figures 457

List of Tables 475

List of Abbreviations 483

A Appendix 487
A.1 Additional Data and Tables . . . 487

A.1.1 Foundational Analysis of EO Modality–Model Interactions 487
A.1.2 Context-Aware Label Enrichment and Multi-Scale Learning with the HELIX Framework 537
A.2 Code and Algorithms 547
A.2.1 Python-based implementation of the HCB Fusion 547

1 Introduction

> *The purpose of computation is insight, not numbers.*
>
> — **Richard Hamming**
> Mathematician and pioneer in information theory

This chapter includes elements from the following peer-reviewed publications:

Sarah Hauser, Michael Ruhhammer, Andreas Schmitt, and Peter Krzystek. *An Open Benchmark Dataset for Forest Characterization from Sentinel-1 and -2 Time Series*. *Remote Sensing*, 16(3), 2024, Article 488. DOI:10.3390/rs16030488

It is cited as [147] and is marked with a green line.

Author Contribution: Sarah Hauser served as a primary contributor to study design, software implementation, practical execution, validation, writing, editing, and visualization.

and from:

Simone Aigner, Sarah Hauser, and Andreas Schmitt. *Pattern-Based Sinkhole Detection in Arid Zones Using Open Satellite Imagery: A Case Study Within Kazakhstan in 2023*. *Sensors*, 25(3), 2025, Article 798. DOI:10.3390/s25030798

It is cited as [6] and is marked with a grey line.

Author Contribution: Sarah Hauser co-led the conceptualization and methodology development, establishment of the analysis pipeline, and contributed significantly to the investigation, supervision, and manuscript writing. She also played a key role in shaping the experimental framework and remote sensing application.

The International Society for Photogrammetry and Remote Sensing (ISPRS) defines remote sensing as follows:

"*Remote sensing is the science and technology of capturing, processing and analysing imagery, in conjunction with other physical data of the Earth and the planets, from sensors in space, in the air and on the ground*" [64].

This very general definition merely states that remote sensing deals with the processing of image-based data of the Earth's surface. Image-based here refers to automatically recorded properties on a regular grid. In contrast to vector data, the individual image elements (pixels) per se have no meaning, only a location, extent, and a value. The recorded property usually relates to the electromagnetic radiation emitted or reflected by an object, which can still be detected by sensors from a greater distance.

As global environmental systems experience increasing stress, from climate-induced glacier retreat to intensified forest disturbance and land degradation, Earth Observation (EO) has emerged as a foundational tool for environmental monitoring. The ability to repeatedly capture spatially detailed information over vast and remote regions enables EO to serve as both an early warning system and a long-term data archive for detecting environmental change. Yet EO alone does not equate to understanding. To translate satellite data into insight, information must be extracted, structured, and contextualized. This requires a conceptual and methodological pipeline: from raw EO signals and derived indices, through increasingly complex features, into models that can learn meaningful relationships, often under conditions of uncertainty, imbalance, or limited labels. Machine learning (ML) has become a central enabler in this transition, unlocking the ability to model complex processes and extract subtle patterns. However, ML is only as reliable as the features and labels it learns from, and in EO, both are uniquely challenging. This dual challenge stems from the complexity of EO signals, often multi-sensor, noisy, and temporally misaligned, and the scarcity or dynamism of high-quality reference data, which is rarely available at the right time, scale, or resolution. This thesis positions itself at the intersection of EO feature fusion, label enrichment, and learning systems design. It argues that progress in EO-based environmental monitoring now depends not only on better sensors or smarter models in isolation, but on integrated pipelines that align data, models, and targets both spatially and temporally. In particular, it explores how combining multi-sensor, multi-temporal EO data, such as Sentinel-1 synthetic aperture radar (SAR) and Sentinel-2 optical imagery, with dynamically structured reference data can yield more expressive, accurate, and transferable insights into environmental systems. The contributions span both methodological foundations and applied innovations, using

diverse environmental case studies (vegetation and sinkhole detection, forest structure and glacial zonation) to validate the effectiveness of feature–label–model interplay. The structure of the thesis reflects this perspective: after motivating the EO–ML context, it addresses fusion techniques, label modelling strategies, and experimental pipelines that bring these elements together.

1.1 Motivation for Remote Sensing in Environmental Monitoring

Environmental and climate monitoring increasingly rely on EO, the use of satellite remote sensing (RS), because it offers unique advantages in scale and consistency. Unlike sparse in-situ measurements, satellites provide global coverage with frequent revisits and fine spatial resolution, enabling continuous tracking of environmental changes [210]. Recent climate assessments (e.g., IPCC AR6) have highlighted EO as fundamental for observing climate trends and filling gaps left by ground networks [169]. In practice, EO data have been critical to detecting climate "tipping points", for instance, revealing accelerated ice loss in polar ice sheets and shifts in vegetation regimes, precisely because of their broad coverage and high temporal frequency [210]. Key strengths of modern EO include:

Temporal Frequency: Many satellites, e.g., Sentinel-1, Sentinel-2, operated by the European Space Agency (ESA), revisit the same location every few days, allowing near-real-time monitoring of dynamic processes (vegetation phenology, glacier flow, etc.). This frequent sampling is essential for catching abrupt events (e.g., rapid snow-melt, forest disturbances).

Spatial Coverage: Spaceborne sensors observe entire regions and the globe uniformly. This synoptic view is invaluable for large-scale phenomena like droughts, forest dieback, or glacier retreat that would be impossible to survey comprehensively from the ground.

Historical Depth and Continuity: EO programs like Landsat (since the 1970s), MODIS, and Copernicus provide long-term archives that support retrospective analyses of environmental change. These datasets allow for trend detection over decades, making EO indispensable for understanding both abrupt and gradual processes such as forest dieback, glacial mass balance trends, or vegetation regime shifts.

Multi-Sensor Capability: Different satellite sensors (optical, thermal, radar, LiDAR) provide complementary information. For example, optical sensors capture vegetation greenness, while radar penetrates clouds and detects surface structure. Combining such multi-source and even multi-temporal data allows robust monitoring under varied conditions. Indeed, *no single sensor can do it all*, hence the need for integrating multiple EO data streams for clearer and more frequent observations.

The versatility of EO becomes especially clear when applied across diverse environmental domains, each characterized by dynamic changes, large spatial extent, and limited ground accessibility. This thesis investigates four such domains, focusing on real-world challenges where EO's multi-sensor, multi-temporal strengths directly support global sustainability efforts.

Vegetation-Geohazard Interaction in Arid Landscapes: In southwestern Kazakhstan, the interplay between vegetation and subsurface instability presents a complex monitoring challenge. This karst-affected region, marked by sporadic sinkhole formation and sparse, stress-sensitive vegetation, exemplifies how EO can capture coupled surface, subsurface dynamics. Changes in vegetation patterns, detected through multi-temporal indices, serve as indicators of both ecological stress and geological anomalies, offering a cost-effective early warning system in a remote desert environment. These EO-based insights support disaster risk reduction and resilience planning in line with **SDG 11 (Sustainable Cities and Communities)**, specifically Target **11.5** (reducing disaster-related losses), and **SDG 13 (Climate Action)**, Target **13.1** (strengthening resilience to climate-related hazards) [327].

Forest Structure and Disturbance in Temperate Europe: Central European forests are increasingly affected by storms, pest outbreaks, and climatic stressors, resulting in widespread structural change and mortality. RS enables both long-term structural mapping (e.g., canopy height, forest types) and high-frequency disturbance detection (e.g., storm damage). This work leverages multi-sensor and multi-temporal EO data to map key forest parameters such as canopy structure and disturbance dynamics. These capabilities inform sustainable forest management and biodiversity conservation, directly supporting **SDG 15 (Life on Land)**, particularly Target **15.2** (sustainable forest management), and **SDG 13 (Climate Action)**, Target **13.1** (climate resilience) [327].

Glacier Zonation in the Canadian High Arctic: In the cryospheric environments of the Canadian Arctic, EO provides a rare observational window into glacier evolution.

With in-situ data scarce, satellite-based time series enable detailed quantification of seasonal snow cover, ablation patterns, and mass balance change, information essential for sea-level rise projections and climate modelling. This contributes to **SDG 13 (Climate Action)**, particularly Target **13.1** (adaptation to climate-related hazards), and **SDG 6 (Clean Water and Sanitation)**, Target **6.6** (protecting water-related ecosystems) [327].

Satellite EO provides the scale, frequency, and diversity of observations needed for modern environmental monitoring. It underpins international climate policy and research by delivering unbiased evidence of change. As Prof. J. Skea (IPCC Chair) noted at COP28, EO now *"serves the crucial role of filling data gaps left by in-situ observations"* in climate monitoring [99]. This strong motivation drives the integration of EO data in environmental science and the increasing use of advanced methods to extract actionable information from the petabytes of satellite imagery now available.

The next sections outline how EO data is transformed into insight, moving from basic index derivation (Section 1.1.1) through advanced ML approaches (Section 1.1.2) to the benchmarking frameworks (Section 1.1.3) required for model development and comparison.

1.1.1 From Spectral Indices to Feature Engineering

Satellites record raw reflectance and backscatter values, but turning those into meaningful environmental information often requires feature extraction. The most fundamental layer of this process involves spectral indices, mathematical combinations of spectral bands designed to emphasize specific surface properties. These indices are computationally simple and interpretable, making them popular for first-pass monitoring and visual inspection. Two classic examples are:

Normalized Difference Vegetation Index (NDVI): NDVI is one of the most widely used spectral indices for vegetation monitoring. It leverages the contrast between red and near-infrared (NIR) reflectance to quantify vegetation "greenness," and is computed as:

$$\mathrm{NDVI} = \frac{\mathrm{NIR} - \mathrm{RED}}{\mathrm{NIR} + \mathrm{RED}}$$

NDVI values range from -1 to $+1$, where higher values indicate dense, healthy vegetation. It serves as a proxy for plant vigour, biomass, and cover. NDVI time

series from sensors such as AVHRR, MODIS, and Sentinel-2 have enabled large-scale analyses of ecosystem productivity, drought impact, and land degradation [325].

Normalized Difference Snow Index (NDSI): NDSI distinguishes snow and ice from other land surfaces by exploiting their high visible and low shortwave-infrared reflectance. It is calculated as:

$$\text{NDSI} = \frac{\text{Green} - \text{SWIR}}{\text{Green} + \text{SWIR}}$$

This index is widely used in operational snow products (e.g., MODIS, VIIRS) to generate daily snow cover maps, which are critical for hydrological modelling, climate studies, and glacier monitoring. The physical mechanism lies with the spectral signature of snow, which is highly reflective at visible wavelengths but strongly absorptive at SWIR wavelengths. The NDSI takes advantage of this characteristic, by isolating bright, cold surfaces from surrounding terrain [89].

While the NDVI, the NDSI, and similar spectral indices, such as the Normalized Difference Water Index (NDWI) for detecting surface water, or the Normalized Difference Built-up Index (NDBI) for identifying urban areas, offer valuable low-level features derived from only two spectral bands, they are inherently limited in complexity. These indices are effective for visual interpretation and coarse thematic mapping, but they often lack the expressiveness required for modelling more nuanced or dynamic environmental phenomena.

To meet the demands of modern EO applications, feature engineering has evolved to produce richer, higher-order representations. These include temporal descriptors such as seasonal statistics (e.g., mean, median, percentiles), amplitude-based phenological metrics, and spectral trajectory summaries across multiple observation dates. Spatial features may involve structural characteristics extracted using texture measures, such as the Grey Level Co-occurrence Matrix (GLCM), which quantifies local spatial heterogeneity in reflectance patterns. Additionally, features that combine different sensor types, such as SAR and optical imagery, are increasingly used to exploit complementary surface information, particularly in areas affected by cloud cover or seasonal snow.

The evolution of EO feature engineering can be further illustrated by several widely adopted strategies:

- **Composite and Extended Indices:** Beyond traditional indices, researchers have developed variants tailored to specific contexts. For example, the Enhanced Vegeta-

tion Index (EVI) offers improved sensitivity in high-biomass conditions and better correction of atmospheric and soil background effects. Similarly, glacier monitoring has benefited from adaptations such as the Adjusted NDSI (ANDSI), which modifies the standard NDSI formula to better separate clean glacier ice from surrounding snow.

- **Multi-Index and Textural Features:** Instead of relying on a single indicator, modern pipelines often compute a suite of indices, each sensitive to different surface properties, and combine them with texture metrics. Texture features derived from the GLCM, for example, may capture spatial heterogeneity and structural patterns in high-resolution imagery. These are particularly useful in characterizing forest canopy structure, urban morphology, or agricultural heterogeneity.
- **Time-Series Features:** With high revisit frequencies, EO data now enable the extraction of temporal descriptors that characterize seasonal and inter-annual dynamics. Phenological metrics such as the timing of green-up, peak greenness, or senescence can be derived from NDVI time series, supporting monitoring of vegetation cycles and ecosystem responses to stress. On a global scale, long-term datasets like GIMMS and MODIS have revealed vegetation trends, greening or browning, linked to climate variability and anthropogenic pressure.
- **Multi-Sensor Feature Fusion:** Finally, the integration of data from multiple EO sensors (e.g., optical and SAR) allows for the capture of complementary surface characteristics. Radar backscatter and coherence provide structural information and all-weather imaging, while optical reflectance contributes detailed spectral signatures. Such fusion improves robustness under challenging conditions (e.g., cloud cover, snow) and enhances generalizability across domains.

Such engineered features move beyond simple band arithmetic, enabling the encoding of multi-modal, multi-temporal, and spatially contextual information. Yet all of these, whether basic indices or advanced descriptors, remain rooted in the transformation of raw EO signals into structured, descriptive variables. As such, they represent the first and least complex stage in the broader information extraction pipeline: a foundation upon which higher-level inference tasks can be built. These feature representations form the bridge between EO and ML. In relatively simple tasks, such as binary classification of land cover types or threshold-based anomaly detection, low-level features like spectral indices may be sufficient and even preferable due to their interpretability and efficiency. However, as the complexity of the modelling task increases, particularly when targeting continuous

environmental variables, temporally dynamic processes, or multi-modal interactions, these descriptors often fall short. In such cases, ML methods are needed to capture non-linear relationships and integrate high-dimensional, structured inputs. The next section outlines how learning algorithms operate on these engineered EO features and how they extend the analytical capacity of RS workflows under challenging modelling conditions.

1.1.2 Machine Learning Methods

With ample EO data and features in hand, ML provides the tools to automatically learn patterns and make predictions [42] (e.g., classify land cover, detect changes, estimate environmental variables). Over the past decade, ML has become ubiquitous in RS analysis, because it can model the complex, non-linear relationships in EO data without strict assumptions about data distributions [364]. A wide spectrum of ML methods, from simple linear regressions to advanced deep neural networks, has been applied to EO data. ML has become a cornerstone in RS, significantly improving the analysis and interpretation of increasingly complex and voluminous geospatial data. As satellite and airborne sensors generate multispectral, hyperspectral, radar, and LiDAR datasets at unprecedented scales, ML techniques are invaluable for identifying patterns, extracting features, and predicting environmental variables with higher accuracy and efficiency than traditional rule-based methods. These capabilities are essential for applications such as land cover mapping, vegetation monitoring, urban growth detection, and climate impact assessment. ML models are broadly categorized into supervised, unsupervised learning as well as reinforcement learning paradigms. Supervised learning leverages labelled datasets where the desired outputs (e.g., class labels or continuous variables) are known, enabling models such as decision trees, support vector machines (SVM), and neural networks (NN) to learn predictive relationships. In contrast, unsupervised learning is used when labels are unavailable, focusing instead on pattern discovery through clustering (e.g., k-means) or dimensionality reduction (e.g., principal component analysis). Reinforcement learning, in contrast to supervised and unsupervised approaches, is increasingly applied to EO RS tasks that require sequential decision-making or adaptive learning. Instead of learning from fixed labelled data, RL models learn optimal policies through interactions with dynamic environments, receiving feedback in the form of delayed rewards. In EO, this applies to tasks such as adaptive image acquisition, active learning for labelling, or intelligent data selection, where the model improves its performance by strategically

choosing what, when, and where to observe next, leveraging methods like Q-learning or policy gradients. These three approaches are employed in RS depending on the data context and application goals. The success of ML in RS is closely tied to rigorous model training and evaluation processes. One fundamental aspect is model validation, which aims to estimate how well a model generalizes to unseen data. Common validation strategies include the hold-out method, where data is split into separate training and testing subsets, and k-fold cross-validation, where the data is partitioned into k subsets and the model is trained and tested k times, each time using a different fold as the test set. These methods help mitigate overfitting and ensure the robustness of performance assessments. Performance evaluation depends on the type of task. For regression tasks, commonly used metrics include mean absolute error (MAE), root mean squared error (RMSE), mean absolute deviation (MAD), and the coefficient of determination (R^2). These metrics quantify the discrepancy between predicted and observed values, with MAE offering a direct measure of average error magnitude, and RMSE being more sensitive to large errors. Standard deviation (STD) of residuals is often used to understand variability in prediction errors, and z-scores can be calculated to identify anomalies by standardizing residuals against their mean and standard deviation. For classification tasks, evaluation typically involves accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic (ROC) curve (AUC), depending on the problem's balance and nature. Confusion matrices are also essential for visualizing model performance across classes. This section provides a structured overview of these methods, progressing from classical approaches to modern deep learning, and highlights real-world environmental monitoring applications (vegetation, forests, glaciers, sinkholes) along with key challenges (high data complexity, overfitting, temporal dynamics).

Traditional Machine Learning Approaches: Traditional ML methods have underpinned many applied EO studies. These ML methods have been applied across various tasks. For example, support vector machines and RF were historically popular for classifying static satellite images into land cover maps [241, 35, 302].

Linear Models: Simple linear models (e.g., ordinary least squares regression) and other parametric approaches (like logistic regression for classification) represent the earliest form of ML used in EO. These models assume linear relationships and specific data distributions [18]. For instance, linear regressions have been used to relate to vegetation indices [264], to biomass [140] or climate variables [313]. In practice, however, Earth observation data often violate linearity assumptions, limiting the accuracy of purely linear

models. Studies have found that while linear regression provides a baseline, non-linear models usually yield better performance for complex EO tasks [179, 206]. Still, linear or robust-linear variants (e.g., the Huber regression model [165]) have seen use in certain large-scale vegetation assessments [14], owing to their simplicity and interpretability.

Support Vector Machines and SVR: SVM [90] (formerly named "Support-vector networks" [334], introduced a powerful kernel-based method capable of modelling non-linear decision boundaries. SVMs became popular in RS classification, often achieving high accuracy even with limited training data [241]. The regression counterpart, Support Vector Regression (SVR) [90], has likewise been applied to predict continuous geophysical variables (e.g., crop biophysical parameters from hyperspectral data) and often outperforms simple linear regression in capturing complex reflectance–variable relationships [179, 206]. However, SVM/SVR models can be computationally intensive on very large EO datasets and may require careful kernel and parameter tuning. In temperate European forests, non-parametric models are frequently used to map forest attributes [5]; one study found SVM could classify tree species from multi-season Landsat imagery with nearly 90% accuracy [113]. For biophysical variable estimation from remotely sensed images, the robust ϵ-Huber cost function is included in the SVR function [55].

Tree-Based Ensembles: Decision tree ensembles have arguably become the workhorse of EO ML due to their versatility and strong performance. Random Forests (RF) [48], which combine many decision trees via bagging, are especially prominent. These models generate an ensemble of diverse decision trees, each trained on a randomly selected subset of the data and features. For prediction, every tree evaluates the input and casts a "vote," with the most frequent prediction across all trees becoming the final output. This ensemble approach enhances model robustness, allowing even moderately accurate trees to collectively yield strong performance. By iteratively refining the model with additional trees, RF becomes more resistant to noisy training data and less informative variables compared to single regression or regression tree models. Since RF relies on random subsets of the training data for each tree, it inherently incorporates a form of cross-validation. As a result, some argue that a separate testing dataset may be unnecessary, an especially practical advantage when working with limited training data [314]. RF models handle high-dimensional inputs and non-linear feature interactions well while resisting (but not immune to) overfitting due to the weak correlation between trees in the model) [48]. Consequently, RF has been extensively used for land cover classification, vegetation property mapping, and more: RF and Extreme Random Tree (ERT) methods were used to simulate the relationship between vegetation and climate elements in Mid-to-High

latitude Asia [349]. In a study on vegetation shifts in Kazakhstan's drylands, researchers applied traditional ML to attribute degradation patterns to socio-environmental drivers. A pixel-wise RF model alongside Shapley value attribution is used, to evaluate the relative influence of factors such as grazing pressure, land use, climate change, and snow cover variability [177]. In a comparative study, RF was top-performing for estimating forest structural parameters like biomass and tree density, when compared to Classification and Regression Trees (CART), SVM, and Artificial Neural Networks (ANN) while using Quickbird imagery [366]. Even in glaciology, before the deep learning (DL) era, techniques like SVMs and decision trees were applied to detect glacier changes or delineate snow lines on RS imagery. For instance, automatic classification of glacier covers from multi-temporal Sentinel-2 imagery using texture, topographic, and spectral data with supervised ML (ANN, SVM and RF), was investigated, and demonstrating that RF, yielded most accurate results [182]. In vegetation modelling, RF is employed and compared to simulate how vegetation responds to climate changes in the Yarlung Zangbo river basin [76], with multiple linear regression models and SVM models, demonstrating that RF models exhibited the highest simulation efficiency. A similar study, utilizing ML to quantify how NDVI (Normalized Difference Vegetation Index) across multiple climate zones responds to variations in temperature and precipitation, demonstrated, RF's efficiency citebao:2021. Another tree ensemble, Extreme Gradient Boosting (XGBoost) [65], implements boosting to sequentially improve trees and often achieves even higher accuracy. Its ability to balance bias and variance makes it effective for complex patterns in EO data [65]. XGBoost has seen rapid adoption in EO for tasks like drought assessment [230], biomass estimation [216], and permafrost mapping [223]. Regression model trees have been shown to be more robust than simple regression trees and are thus more widely applied to prediction problems [341]. However, advanced ensemble strategies such as stacking (which combines multiple base learners) have been explored to further boost accuracy; generally, ensemble models (bagging, boosting, or stacking [83]) outperform single classifiers on RS tasks by leveraging diverse learners [364, 299]. The most common implementation of ensemble model trees is RF [314]. A hybrid traditional ML approach[54] combining SVM and Hidden Markov Models (HMM) [3] is successfully applied to classify glacier surface types, bare soil, glacier ice, firn, and snow, using multi-temporal, multi-sensor satellite data.

Deep Learning - CNN, RNN, and Hybrid Networks: With the rise of deep learning, model capabilities for EO data have expanded dramatically. Convolutional Neural Networks (CNN), in particular, have revolutionized analysis of spatial imagery. CNN excel at

automatically learning features from raw pixel data through layered convolution filters, and have been widely applied to high-resolution RS images for land cover classification [226], semantic segmentation [337], object detection [23], deconstruction of missing data [363], or pansharpening [233]. In these studies, CNN models effectively exploit the spatial characteristics of the data by performing convolutions across the x and y dimensions [255]. 1D-CNN operate over a single dimension (e.g., time or sequence), 2D-CNN process spatial data across height and width (x, y) [255], while 3D-CNN extend this to include depth or time (x, y, z) for volumetric or spatio-temporal analysis. CNN variants have been tailored to diverse EO data structures. For spectral data such as hyperspectral images (with hundreds of bands), 1D-CNN can be applied across the spectral dimension to capture contiguous spectral signatures [161]. Likewise, for purely temporal EO data (e.g., univariate satellite time-series like NDVI curves), temporal 1D-CNN have been developed (TempCNN) where convolution is applied along the time axis to extract temporal patterns [348]. These have proven effective for classifying satellite image time-series, often outperforming traditional methods like RF on time-series classification tasks [172]. Crucially, by using convolution, they incorporate temporal ordering and local sequence information that static classifiers miss. There are also 3D-CNN that simultaneously convolve across space and either spectral or temporal dimensions (or both) [215, 139]. These spatio-temporal CNN can capture dynamic evolution of features (e.g., seasonal changes) in a unified model. An important class of deep models for sequential data are Recurrent Neural Networks (RNN), especially those based on Long Short-Term Memory (LSTM) units. RNN are explicitly designed to handle sequential inputs by maintaining internal state, making them well-suited to multi-temporal EO problems. Indeed, LSTM-based networks have been extensively studied for classifying optical image time series [279, 316] and multi-temporal SAR data [167, 239]. Studies report that RNN (using LSTM) can outperform traditional classifiers like RF and SVM on land cover sequence classification [279]. RNN are well-suited for capturing long-term dependencies and inherently account for temporal context. However, when the task involves predicting a single label for an entire time series, standard RNN encounter challenges. Specifically, the need to back-propagate errors across all time steps increases with the length of the series, which can complicate training, since early time steps are distant from the output, and slow down convergence due to the sequential nature of updates. Consequently, although LSTM networks are capable of modelling temporal dynamics, they often demand extensive training data and carefully designed training strategies to mitigate issues like vanishing gradients and overfitting in long sequences [255]. To leverage both spatial and temporal information, hybrid architectures have

emerged. A common design is CNN–RNN hybrids (e.g., CNN-LSTM networks), where a CNN first extracts spatial features from each image in a time sequence, and then an LSTM processes the sequence of feature vectors to model temporal dynamics [173]. Hybrid deep networks are also being explored in cryosphere applications, one recent proof-of-concept used a CNN (in an AlexNet form) on pairs of SAR images to learn matching features for glacier velocity estimation, effectively replacing the traditional cross-correlation method with a learned approach to track ice motion [373]. Another example is the detection of glacier snow lines: a combination of image processing, RF classification, and neural network-based segmentation was proposed to automatically identify the end-of-summer snow line altitude on alpine glaciers [261], which is crucial for understanding glacier mass balance and dynamics.

Challenges and Considerations: Advances in artificial intelligence have facilitated the widespread integration of diverse variables and datasets. Within this framework, non-parametric ML algorithms have shown strong capabilities in managing complex, non-linear relationships, gaining increasing attention for big data classification in RS [12]. While advanced models like ANN and DL offer high performance, traditional machine learning methods have proven to be reliable and relatively simple alternatives, particularly for large-scale classification tasks such as forest mapping [34]. Despite their successes, ML methods for EO come with challenges. Data complexity and volume are chief among them: EO datasets are often high-dimensional (many spectral bands, pixels, or timesteps) and heterogeneous. Traditional models can struggle with such complexity unless dimensionality reduction or feature selection is applied. Also, in terms of tree-based Models, large tree structure [298] and spatial autocorrelation of the data [232] may induce model overfit, especially with small training data sets [79]. Additionally, standard RF models lack built-in prediction-level uncertainty estimates and treat missing values simplistically unless explicitly modified [129]. Because RF outputs are based on majority votes from tree ensembles, the degree of agreement among trees is not quantified, limiting interpretability [341]. In contrast, Bayesian tree-based methods like BART provide credible intervals for predictions, offering clearer uncertainty quantification at the pixel level [260], which is particularly useful in interpreting cover estimates from EO data. Deep networks can ingest raw data but demand even larger training samples to generalize [234, 297], raising issues of data scarcity for labelled tasks. This leads to the risk of overfitting, especially when complex models are trained on limited ground-truth data [356]. For instance, when the training dataset is too small, the model's learned parameters may fail to capture the true distribution of the overall data. Alternatively, the

model might overfit to noisy samples during training, overlooking portions of the correct data. As a result, the model performs well on the training set, but poorly on unseen data. Another persistent challenge is handling the temporal structure inherent in many EO problems. Simply stacking multi-temporal data into a vector for an RF or SVM ignores sequential dependencies. However, they can still be valid when EO features and target labels are temporally aligned, that is, when both refer to the same time-step. In such cases, each instance is treated as an independent snapshot, and temporal structure is not required for the prediction task. However, when multi-temporal features are stacked to predict outcomes that depend on temporal evolution, such as crop type at harvest based on year-round spectral data, this approach ignores the sequential nature of the data. In these cases, models capable of learning temporal patterns, such as TempCNN, RNN, or Transformers, are more appropriate and may yield better generalization by capturing trends, seasonality, and temporal dependencies, but can be difficult to train on long sequences and may still miss very long-term trends [255]. Additionally, irregular time sampling (due to cloud cover or varying sensor revisit times) can complicate temporal modelling. Developing architectures that efficiently capture long-range temporal patterns without enormous data requirements is an active research area. Computational scalability is also a practical issue: advanced models (e.g., 3D-CNN or CNN-LSTMs) can be resource-intensive in both memory and processing. Training these on global-scale EO data demands significant computing power (GPUs/TPUs) and optimization [297]. Efforts are underway to create lighter or more efficient models (including lightweight ensembles or using pre-trained networks on EO data) to alleviate these burdens [309]. Lastly, interpretability of ML models in EO should be considered. Simpler models (linear, decision trees) offer more transparency in how inputs relate to outputs, whereas deep networks are often black boxes. This can be problematic in environmental decision-making contexts where understanding the drivers of a model's prediction (e.g., which spectral bands indicate a pest outbreak) is important. Techniques like feature importance in RF or saliency maps in CNN can provide some insight, but bridging the gap between model complexity and user interpretability remains important. ML methods have greatly advanced the analysis of EO data, enabling higher accuracy and new capabilities across a range of applications, from monitoring vegetation in Central Asia's rangelands to mapping old-growth European forests, tracking Arctic glacier changes, and detecting hazardous sinkholes. Linear models and kernel methods laid the foundation, tree ensembles brought robustness and ease of use, and DL now offers unprecedented modelling power for spatial, spectral, and temporal data. The state-of-the-art continues to evolve rapidly, with hybrid models and ensemble strategies pushing the frontiers of predictive performance. Ongoing research

is addressing current challenges like data efficiency, overfitting control, and temporal dynamics handling. These developments ensure that ML will remain at the core of extracting actionable information from EO, supporting better environmental monitoring and management grounded in big data from space.

However, the power and progress of ML in EO are ultimately constrained by the quality and availability of data, specifically, the reference labels and feature–label pairings used for training and evaluation. Regardless of model complexity, reliable inference requires a foundation of well-structured, representative datasets that capture the variability of both EO inputs and environmental targets. As modelling tasks become more ambitious, spanning multiple regions, temporal scales, or sensor types, the need for transparent, reproducible benchmarking grows more urgent. Benchmark datasets enable not only fair comparison between ML approaches, but also foster generalizability, scalability, and methodological rigour. The following section thus outlines the central role of benchmarking in EO–ML workflows and introduces the challenges involved in constructing and using such datasets effectively.

1.1.3 The Need for Benchmarking

Benchmarking is a foundational requirement for the development, evaluation, and comparison of ML methods. At its core, benchmarking relies on structured datasets that pair EO-derived features with well-defined reference labels, allowing algorithms to learn, validate, and generalize across environmental conditions. Without standardized benchmarks, it becomes difficult to assess model performance, quantify uncertainty, or ensure reproducibility, particularly as methods grow more complex and datasets span multiple regions, time frames, and sensor modalities. In EO, where reference data are often sparse, noisy, or inconsistent, the creation and curation of benchmark datasets is not just a technical formality, but a central scientific need.

Inspired by benchmark-driven advances in computer vision (e.g., ImageNet [81], COCO [220]), the EO community has begun developing large, curated datasets that pair satellite imagery with ground-truth annotations. These datasets typically contain aligned EO features (e.g., Sentinel imagery) and reference labels (e.g., land cover, vegetation metrics), providing a shared testing ground for machine learning models across spatial and temporal scales. Three notable examples illustrate the range and growing sophistication of EO benchmarks:

- **BigEarthNet and reBEN:** A large-scale benchmark archive of Sentinel-2 (later expanded to Sentinel-1/2) imagery designed for multi-label land cover classification. BigEarthNet [315] includes 590,326 image patches (120×120 m) across Europe, each annotated with one or more CORINE-based land cover classes. To reduce label noise, original labels were condensed into 19 super-classes. BigEarthNet has become a go-to resource for training and evaluating deep learning models in scene classification and fusion [315]. A recent refinement, the reBEN dataset [70], improves on BigEarthNet [315] by providing 549,488 Sentinel-1/2 patch pairs with enhanced atmospheric correction and pixel-level reference maps derived from the updated 2018 CLC inventory. The patches (1200 m × 1200 m) are processed using the Sen2Cor tool to improve optical image quality. reBEN introduces a geographical-based train/validation/test split strategy to mitigate spatial autocorrelation, significantly enhancing the robustness of DL model evaluation. It also supports both scene-level and pixel-wise learning tasks, making it well-suited for multi-label and semantic segmentation experiments. Code, data, and pre-trained weights are openly available, further promoting reproducibility and cross-study comparison [70].

- **SEN12MS:** SEN12MS [291] is a large-scale, globally distributed benchmark dataset introduced through the IEEE GRSS Data Fusion Contest. It consists of 180,662 georeferenced triplets, each containing a dual-polarized Sentinel-1 SAR patch, a multispectral Sentinel-2 patch, and a MODIS land cover label map, all at 10 m resolution. The dataset spans over 100 sites across all inhabited continents and includes samples from all meteorological seasons, making it one of the most diverse EO benchmarks to date. By leveraging freely available Copernicus Sentinel data and the cloud infrastructure of Google Earth Engine, SEN12MS addresses key limitations of earlier datasets, particularly with respect to spatial coverage, environmental diversity, and sample volume. Its design explicitly supports research in multi-sensor data fusion, scene classification, and semantic segmentation for land cover mapping. The multi-modal nature of the dataset enables the development and testing of deep learning models that combine optical and radar data, offering robustness in scenarios where single-sensor inputs are affected by cloud cover or seasonal variability [291].

- **Wald5Dplus:** Wald5Dplus is a distinctive, open, multi-modal benchmark dataset for mid-European forests. It integrates time-series data from Sentinel-1 (C-band SAR) and Sentinel-2 (optical MSI) with detailed airborne observations, including LiDAR

and UAV-based multispectral imagery, to produce a high-resolution, semantically labelled dataset for forest monitoring. The "5D" refers to the combination of spatial (north–south and east–west), polarimetric (Sentinel-1), spectral (Sentinel-2), and temporal dimensions, all integrated into an Analysis Ready Data (ARD) cube. The "plus" denotes the inclusion of semantic labels derived from airborne campaigns, specifically tree species, crown area, crown height, initial crown height, and tree count. Unlike traditional land cover classification datasets, Wald5Dplus is designed to support regression-based modelling of continuous forest variables, making it one of the first EO benchmarks in the domain to move beyond nominal or ordinal labels. It enables detailed evaluation of ML models for forest parameter prediction, aiding in biomass estimations, and even spatio-temporal change detection under real-world multi-sensor conditions. Funded by the German Aerospace Center (DLR) and the Federal Ministry for Economic Affairs and Climate Action, Wald5Dplus represents a significant advance in high-quality EO benchmarking [147, 148].

While these datasets mark major progress, the construction of EO benchmarks remains highly challenging. Label acquisition is often expensive, labour-intensive, or reliant on secondary sources like land cover inventories, which can introduce noise. Temporal and spatial inconsistencies, sensor misalignments, and cloud or snow coverage further complicate dataset design. Benchmark datasets must also be representative, capturing diverse environmental conditions, seasons, and biomes, to support robust generalization and fair testing.

High-Quality and Well-Labelled Data: Foremost, a fundamental requirement in the field of RS pertains to the availability of high-quality datasets that are meticulously labelled. This precision ensures that AI algorithms can be effectively trained and validated on information that is both accurate and dependable. It is notable that machine learning, a cornerstone of AI, relies heavily on data quality. This aligns with findings from research emphasizing that the successful utilization of ML techniques in RS applications necessitates high-quality data, particularly well-labelled datasets [362]. Such well-labelled data serve as the bedrock upon which AI models can be constructed and validated.

Accessibility of Publicly Available Datasets: A pivotal requirement arises from the accessibility of publicly available datasets, accompanied by validation data. These datasets serve as indispensable benchmarks for the development and validation of algorithms, allowing researchers to evaluate their methods against established standards. This aligns with the notion that publicly available datasets with validation data are crucial for

researchers to verify their developed algorithms and compare them with state-of-the-art methods. In various standard RS applications, frequently employed datasets serve as reference points for algorithm testing. The abundance of such datasets underscores the importance of making high-quality training samples available, as highlighted in literature [111, 1, 221].

Diversity and Representativeness: Training data should encompass a wide range of scenarios to enable AI models to generalize effectively. This diversity is crucial in the field of RS, where real-world conditions can vary significantly. Machine learning techniques, which are foundational in AI, rely on diverse training data to ensure models can adapt to different conditions and achieve robust generalization. In this context, the need for variations in land cover, seasonal changes, and different environmental conditions is essential to ensure AI models can effectively handle the intricacies of RS applications [212].

Spatial and Temporal Coverage: The training dataset should provide comprehensive spatial and temporal coverage [212, 170], ensuring that AI models can effectively adapt to diverse regions and monitor temporal dynamics with precision. High-quality training samples, representative of a wide range of geographic locations and temporal changes, are fundamental in addressing the challenges of RS data, especially in the context of forests. The utilization of data with broad spatial and temporal coverage aligns with the need to capture fine-grained spatial and temporal changes in RS applications.

Data Resolution: Training data should align with the spatial and temporal resolution of the RS data used for analysis. This matching resolution is essential for enabling AI models to capture and respond to temporal changes with accuracy, as highlighted in the literature. Aligning training data resolution with RS data resolution is a crucial aspect of ensuring the effective application of AI in RS [221].

Quantity and Sample Size: Adequate training samples are pivotal for optimizing AI models [362, 21]. The quantity of training data should align with the complexity of the analysis task and the specific requirements of the AI model under consideration. The importance of having an ample sample size to mitigate the risk of model underfitting is well-established in the field of machine learning, including in the realm of RS.

Consistency and Continuity: Consistency in labelling and data quality [170] throughout the training dataset is imperative for ensuring the reliability of AI models in RS tasks, especially those involving time-series data. Additionally, maintaining continuity in data collection is essential for effectively monitoring changes and trends. Such consistency

and continuity are essential components of robust AI model development, as recognized in the existing body of literature.

Annotated Metadata: Annotated metadata [221, 58], providing comprehensive information regarding the data's source, acquisition date, geographical location, and any preprocessing steps applied, enhances the interpretability and utility of training data. This metadata is vital in providing context to the training data, enabling researchers to better understand the information used for AI model development. Researchers in the field have acknowledged the critical role that annotated metadata plays in the effective utilization of training data.

Data Balance: Maintaining a balanced representation of classes or categories within training data is vital, especially in classification tasks. Unbalanced datasets can present a substantial obstacle in the process of model optimization, especially when specific classes are infrequent or not well-represented [192, 170]. Ensuring equitable representation of classes is a recognized strategy to prevent biases and skewed results. Achieving data balance is crucial for accurate classification of RS data, a concept well-supported by prior research [362].

Moreover, emerging tasks such as change detection, regression-based prediction of continuous variables (e.g., biomass, tree height), and ordinal or dynamic labelling (e.g., forest structure classes over time) demand new forms of annotated data. Wald5Dplus [147, 148], for example, provides one of the first structured datasets supporting pixel-wise regression of forest attributes in cardinal scale, an essential step for advancing continuous-label EO modelling. Benchmarking, therefore, is not merely a matter of evaluation, but a structural prerequisite for progress. Well-curated, representative, and accessible datasets underpin the development of robust, transferable models. In the chapters that follow, this thesis contributes to that direction by advancing dataset design, training pipeline integration, and evaluation strategies that reflect the evolving complexity of EO-based environmental monitoring Taken together, these criteria define what constitutes a high-quality EO benchmark, one capable of supporting advanced modelling workflows and fostering methodological innovation.

1.2 Application Domains and Labelled Datasets

This section presents the key environmental domains examined in this study, alongside related work and the labelled datasets used for model development and evaluation. This thesis investigates environmental modelling across three ecologically and geophysically distinct Areas of Interest (AOIs), each selected for its relevance to a specific RS challenge.

The southwestern region of Kazakhstan, located within the southeastern part of the Mangystau Province near the Caspian Sea, represents an arid, karst-affected landscape characterized by sinkhole formations and sparse vegetation. Central European forests serve as representative temperate ecosystems, with reference data drawn from three forested regions: the Bavarian Forest National Park, the Steigerwald, and a small forest near Kranzberg. While forest structural modelling incorporates all three sites, large-scale disturbance analyses are limited to the Bavarian Forest National Park. The Canadian High Arctic provides a cryospheric setting for modelling seasonal and short-term glacier zone dynamics.

Together, these AOIs span semi-arid, temperate, and cryospheric environments, enabling a multi-domain evaluation of the methods developed in this thesis. Their inclusion supports both methodological generality and ecological relevance, ensuring that remote sensing and machine learning approaches are assessed under diverse environmental conditions. In the following subsections, each AOI is described in more detail, including its geographic context, relevant prior research, and the associated reference datasets used as labelled targets throughout this work.

1.2.1 Southwestern Kazakhstan

Karst landscapes in arid regions pose a unique challenge for environmental monitoring: they are subject to both abrupt geomorphological change, such as sinkhole collapse, and subtle ecological stress reflected in sparse vegetation. These dual dynamics often co-occur and interact, making it difficult to isolate drivers of surface anomalies without integrated, multi-temporal EO analysis. Traditional approaches struggle in such regions due to their remoteness, limited in-situ monitoring, and complex subsurface–surface feedbacks. Remote sensing offers a rare observational lens into these systems, yet detecting and

interpreting geohazard signals amid natural vegetation variability remains an open methodological challenge.

A wide range of techniques has therefore been explored to detect and analyse sinkholes, each offering distinct advantages depending on the geological setting, data availability, and scale of application:

Traditional sinkhole detection methods include field-based, geophysical, and GIS-based approaches, each offering distinct strengths and limitations. Visual inspection and speleological exploration remain foundational, providing direct insight into surface features and subsurface karst structures, but are labor-intensive, subjective, and often impractical in vegetated or remote areas [19, 184, 251]. Geophysical methods such as electrical resistivity imaging (ERI) and ground-penetrating radar (GPR) allow non-invasive detection of subsurface voids and density variations, particularly in shallow karst settings [59, 117, 263]. Gravimetry and magnetometry support broader anomaly detection [180, 280], while boreholes and trenching offer ground-truth validation at high cost and spatial limitation [340, 60]. GIS-based analyses have further expanded sinkhole mapping by integrating topographic, historical, and environmental datasets. Terrain-derived features from maps help identify depressions [20, 50], while archival and multi-temporal maps reveal masked or former sinkholes now obscured by vegetation or urbanisation [134, 33]. Despite resolution constraints and manual interpretation challenges, GIS remains central to regional-scale sinkhole research.

RS techniques have become indispensable for detecting and analysing sinkholes, particularly for large-scale or inaccessible areas. These methods rely on data captured from satellites, aircraft, or drones to identify surface features and subsidence patterns indicative of sinkhole activity. Optical imagery is one of the most widely used remote sensing tools, where satellite data such as Landsat [318], RapidEye [312], and IKONOS [168] are analysed to detect surface anomalies. GIS platforms also integrate remote sensing imagery with morphometric parameters to analyse the spatial patterns of sinkholes over time. Historical aerial images and satellite data such as Landsat and Sentinel-2 data have been used to track the evolution and frequency of sinkhole formation, correlating these trends with environmental or anthropogenic factors [108, 250]. In northwest Morocco, for instance, regions prone to karstification were identified by deriving vegetation and water indices from satellite images, highlighting areas with higher surface water input [323]. Similarly, aerial photographs combined with orthorectified images and GIS platforms are used to identify historical or masked sinkholes and assess their temporal evolution [49, 108]. Digital tools such as digital elevation models (DEMs) are extensively

used to analyse the topography and morphology of sinkholes. High-resolution DEMs derived from LiDAR (light detection and ranging) or satellite imagery allow researchers to automatically detect and characterise sinkholes based on geometric parameters like depth, perimeter, and slope [183]. LiDAR provides high-resolution topographic data, enabling the detection of sinkholes and associated features even in vegetated or remote areas. LiDAR-derived DEMs allow for the automatic mapping of sinkholes by analysing their geometric properties, such as depth and diameter. This method is particularly effective for morphometric characterisation and detecting subtle subsidence patterns that may precede sinkhole formation [238, 370]. Furthermore, LiDAR has been successfully applied in Slovenia, where high-resolution data increased sinkhole detection accuracy to 83.5% [183]. Aerial photographs, combined with DEMs, help validate and refine sinkhole mapping by visually confirming subsidence features and vegetation changes [41]. Furthermore, swath bathymetry, a specific form of DEM analysis, is applied in underwater environments to detect sinkholes on lakebeds or ocean floors [321]. Techniques like automatic mapping and photogrammetry further enhance detection accuracy, enabling efficient sinkhole monitoring over large areas [88]. Radar-based techniques, such as InSAR (interferometric synthetic aperture radar), are used to monitor ground deformation over large areas. Differential interferometry (DInSAR) is particularly effective for measuring subsidence rates and temporal changes in karst regions. In the Ebro Valley, Spain, DInSAR velocity maps were cross-referenced with sinkhole inventories to assess doline activity and predict future collapses [118]. Although radar methods excel at capturing broad subsidence patterns, small active sinkholes or rapid collapses may be overlooked due to their spatial resolution limitations. Recent advancements, such as the Sinkhole Scanner method [196], address some of these limitations by employing a two-dimensional Gaussian function to detect sinkhole-related spatio-temporal patterns in InSAR deformation time series. This method, tested on Sentinel-1A data, successfully detected sinkholes even in challenging environments, demonstrating stability in arid regions and improved detection in vegetated areas, a key advantage for addressing the dual challenges of sinkhole monitoring in such conditions. However, the Sinkhole Scanner is not without limitations. Its reliance on a Gaussian kernel assumes specific deformation shapes, potentially missing irregular or complex sinkhole patterns. Furthermore, the method may struggle to detect very rapid collapses without precursory signals, and its computational intensity can be a limiting factor when analysing large, high-resolution datasets. While it improves detection in vegetated areas, challenges like signal decorrelation persist, especially in regions with dense vegetation. Multispectral and hyperspectral imaging further enhance sinkhole detection by revealing vegetation

stress, soil moisture changes, or water pooling, indirect indicators of subsurface karst processes. Airborne multispectral scanning, for example, has shown promise in detecting subtle environmental changes linked to sinkhole formation [75]. These remote sensing techniques are often complemented by manual processing in GIS environments or field verification to confirm suspected sinkholes. However, their reliance on high-resolution datasets and specialised software can make them resource-intensive and less accessible for widespread or continuous monitoring. Despite these challenges, remote sensing remains a cornerstone in modern sinkhole research, particularly for large-scale mapping and long-term monitoring. In order to be accessible for all potential users, the data should be open. Additionally, to guarantee large coverage, the data have to be globally available. Preceding studies have shown that optical data are most promising. Three satellite missions fulfil these requirements: Sentinel-2 [96], Landsat [318], and MODIS [97]. According to the literature, the spatial resolution is playing a key role in the detection of sinkholes, whereas the temporal resolution is negligible. In this sense, Sentinel-2 (10 m pixels every 5 days) outperforms existing open sources such as Landsat (30 m every 16 days) and MODIS (250 m every 1–2 days). AI is increasingly being applied to sinkhole detection and mapping, offering a new paradigm for automating complex analyses and enhancing accuracy. ML and DL techniques, in particular, have shown significant promise by processing large datasets, identifying patterns, and predicting sinkhole-prone areas with minimal human intervention. One of the most notable applications of AI is in image recognition and object detection. For example, the YOLO (you only look once) algorithm has been used to detect sinkholes in satellite and aerial imagery with handsome results. In a study conducted in Kazakhstan, YOLO achieved a detection accuracy of 74% for sinkholes and 86% for geological pre-sinkhole features such as takyr depressions [277]. Also, a sinkhole-tracking methodology that employed CNN transfer learning on FIR imagery was successfully implemented [155]. These methods enable rapid, large-scale detection, outperforming traditional manual and semi-automated techniques in both speed and efficiency. Supervised learning algorithms have been employed to classify land features and identify potential sinkhole zones. These models are trained using labelled datasets of known sinkholes and environmental variables such as topography, hydrology, and geology. Once trained, the models can predict high-risk areas by analysing similar patterns in new datasets. While highly accurate, this approach requires comprehensive and reliable training data, which can be difficult to obtain for regions with sparse sinkhole documentation. Unsupervised learning and clustering methods are other emerging avenues. These techniques analyse unlabelled data to identify anomalies or clusters indicative of sinkhole-prone regions. They are particularly useful for preliminary assessments in

regions where ground truth data are limited. AI also plays a critical role in temporal analysis, helping to track sinkhole evolution and predict collapses. Time-series data from remote sensing platforms, such as LiDAR or InSAR, can be fed into RNN or LSTM models to detect subtle deformation trends and assess the likelihood of future events. Despite these advances, challenges remain in the widespread application of AI to sinkhole studies. The creation of robust and diverse training datasets is often labour-intensive, and the computational demands of deep learning models can be prohibitive for smaller research teams or institutions. Additionally, the interpretability of AI models is sometimes limited, making it difficult to understand the underlying decision-making process and validate the results.

Detecting sinkholes poses significant challenges with respect to low data availability, spatial resolution constraints, and large extents. The problem is becoming topical as there are plans to develop former unused land in places such as the Mangystau area in Kazakhstan. Research on sinkhole formation mechanisms and hazards in Kazakhstan remains sparse [27, 9], and the precise delineation of sinkholes at finer scales continues to be an unsolved topic [141]. Digital terrain models (DTMs), especially those derived from LiDAR [183, 187], are widely regarded as the optimal data source for sinkhole detection. These models provide high-resolution representations of the bare-earth surface, crucial for identifying the subtle depressions characteristic of sinkholes. However, high-resolution DTMs are often unavailable, even in regions with advanced geospatial infrastructure. For instance, in Bavaria, DTMs with 1 m resolution are typically limited in temporal and geographic coverage [112]. This lack of temporal consistency hinders long-term monitoring and dynamic geo-hazard assessment. In Kazakhstan, the absence of high-resolution DTMs presents a significant barrier to applying conventional detection methodologies. Kazakhstan's geographic scale amplifies these challenges. Covering over 2.7 million square kilometres, it is among the largest countries globally, but its low population density makes large-scale mapping and monitoring logistically and economically prohibitive. Even smaller, resource-rich regions like Bavaria struggle to maintain consistent, high-quality geospatial data, illustrating the difficulty of conducting geomorphological studies at Kazakhstan's scale. In the absence of DTMs, digital elevation models are often used as substitutes [141]. However, unlike DTMs, DEMs include surface features such as vegetation and structures, which can obscure the underlying terrain and limit their effectiveness for detecting small-scale geomorphic features. While DEMs can capture larger deformations, their typically coarse resolution (e.g., 30 m in widely available datasets like the Copernicus DEM) restricts their utility for identifying subtle or small sinkholes. Additionally, inconsistent spatial and temporal availability further

reduces their effectiveness for large-scale or continuous geohazard monitoring. The lack of high-resolution DTMs and the limitations of DEMs highlight a critical research gap in Kazakhstan. Traditional sinkhole detection approaches, reliant on detailed terrain data, are poorly suited to the region's data-sparse context. These challenges require innovative and cost-effective methods that utilise widely available resources such as freely accessible remote sensing imagery.

Yet, one of the most persistent challenges in this context is the spectral and morphological ambiguity between sinkholes and natural vegetation patterns. In arid and semi-arid karst environments like southwestern Kazakhstan, sparse or stress-sensitive vegetation can exhibit surface expressions, such as dark NDVI anomalies or local depressions, that visually resemble collapse features. This complicates automated detection, particularly when relying on single-sensor or mono-seasonal imagery. Addressing this confusion requires a more integrated approach that combines multi-seasonal and multi-sensor observations to disentangle the geophysical and ecological signals. A comparable challenge is encountered in recent large-scale efforts to map desert shrublands using medium-resolution satellite data. For instance, in a recent study it was demonstrated that sparse vegetation in Northern China's deserts is often under-represented in land-cover products due to its low spectral contrast and fragmented distribution [369], similar to vegetation inside or around sinkholes. Their study combined manually labelled high-resolution samples with similarity-based sample expansion and applied traditional ML (e.g., RF) using multi-temporal composites of medium-resolution EO data (Sentinel-2) and DEM (Copernicus DEM) data to significantly improve classification accuracy [369]. However, unlike the under-representation problem in shrubland mapping, the core difficulty in sinkhole/shrub detection lies not in data sparsity but in spectral ambiguity: vegetation and collapse features may share similar optical characteristics, especially under dry conditions or when vegetation colonizes sinkholes. This highlights the need for more nuanced approaches that integrate spectral, temporal, and topographic cues to disentangle overlapping geophysical and ecological signals. In another study, targeting desert vegetation in Northern China further illustrates the difficulty of distinguishing sparse dry shrubs from surrounding bare soil under intense illumination and complex topography [310]. In their case, typical vegetation indices derived from visible-light UAV imagery (e.g., EXG, VDVI) were insufficient due to strong shadow textures and the low reflectance of dry, low-stemmed shrubs, conditions similar to vegetation found within or around sinkholes. To improve detection accuracy, they proposed a novel HSV-based green-enhancement index (HSVGVI), which leveraged the hue–saturation–value (HSV) color space and channel enhancement. This approach substantially outperformed traditional RGB-based indices,

achieving over 95% classification accuracy in shaded and shrub-dominated scenes [310]. While their method focuses on UAV imagery at very high spatial resolution (10 cm), the underlying insight is relevant for satellite-based mapping in optically ambiguous terrains like Mangystau: traditional vegetation indices may misclassify vegetated sinkholes or fail to detect subtle vegetation changes altogether. As such, their work supports the argument that spectral confusion in sparse desert vegetation requires adapted indices or fused, multi-source methods, especially when class boundaries (vegetation vs. sinkhole) are entangled in both spectral and spatial domains.

Study Area and Environmental Characteristics

As a geologically dynamic and ecologically sparse environment, the southwestern region of Kazakhstan (see Figure 1.1 presents a uniquely challenging test-bed for remote sensing-based sinkhole and vegetative feature detection. Characterised by arid climatic conditions, minimal vegetation, and widespread karstic subsurface activity, this area allows the evaluation of data fusion strategies in extreme terrain. Unlike vegetated forest environments, karst regions such as these exhibit ambiguous spectral and structural surface signals, necessitating tailored approaches to data preprocessing, temporal fusion, and label integration.

The Ustyurt Plateau, covering approximately 5000 square kilometres near the borders of Turkmenistan and Uzbekistan, is adjacent to the Ustyurt National Reserve but falls outside the protected area. This region's sparse development and lack of significant human settlement, evidenced by minimal infrastructure and occasional pathways observed via satellite imagery [22, 328], underline its status as a largely untouched and unexplored landscape. Geological investigations are crucial here to address potential challenges and opportunities presented by the area's complex subsurface features.

The prevalent soil type in the study area is Calcisol, which is typical of arid environments and characterised by high calcium carbonate content near the surface. These soils often support only sparse vegetation, such as drought-resistant shrubs and grasses, and exhibit circular vegetation patterns that correspond to subsurface karst structures, including sinkholes [8]. In addition, the region contains takyr plains, which are clay-rich surfaces with polygonal cracking patterns. These plains reflect a history of marine sedimentation, often bearing salt deposits, and their unique surface texture makes them important for understanding hydrological processes and subsurface stability [204]. The climate in Mangystau is arid and desert-like, with significant contrasts between humid and arid

seasons. Precipitation is rare and irregular, and temperatures show dramatic seasonal variation. Winters are relatively mild due to the moderating influence of the Caspian Sea, while summers are extremely hot and arid, often accompanied by strong winds that drive erosion and sediment redistribution [191]. The sparse vegetation in the area is dominated by shrubs and grasses [163, 189], with more robust growth near depressions or along the edges of takyr plains, where water may temporarily accumulate. These climatic and vegetative conditions are significant for understanding surface and subsurface processes in the context of sinkhole detection.

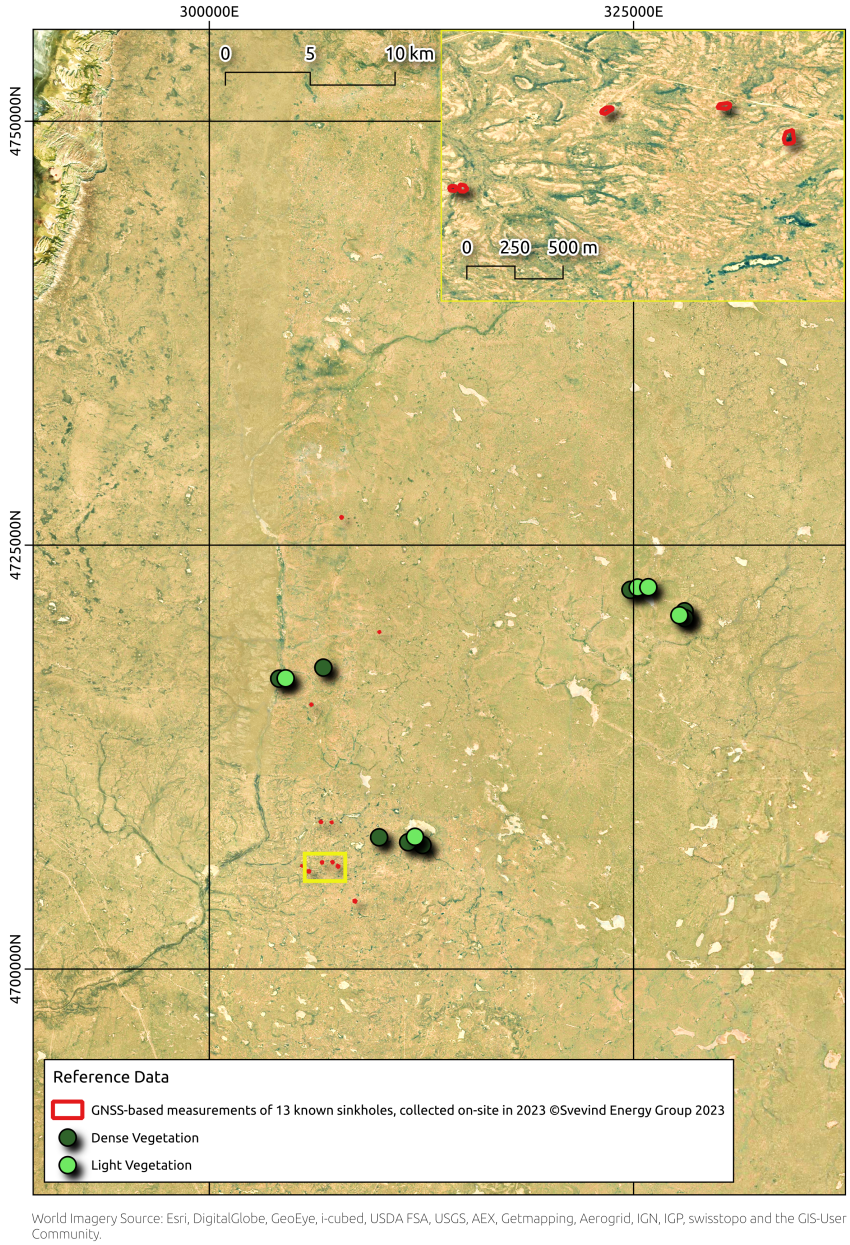


Figure 1.1.: Overview of the Study Area in Southwestern Kazakhstan.

Reference Data and Labels

Accurate reference labels are essential for training and validating EO-based detection pipelines, particularly in data-scarce regions like southwestern Kazakhstan. Given the

area's remoteness and limited historical documentation, the generation of reliable ground-truth data required the integration of locally acquired field measurements and high-resolution visual sources. In this study, two complementary sources were used to construct label datasets: GNSS-verified sinkhole locations and high-resolution satellite imagery for vegetation classification. These resources provide spatially precise and visually interpretable inputs for supervised learning and model evaluation.

To validate the identification of sinkholes, reference data were provided by Svevind Energy Group [319]. This ground-truth dataset includes a shapefile containing GNSS-based measurements of 13 known sinkholes, collected on-site in 2023 using handheld GNSS equipment. These measurements were performed by a regional topographer to provide precise geospatial locations for these features. In addition to the GNSS data, georeferenced imagery of the sinkholes [319], also captured in 2023, is available. These combined datasets represent the confirmed ground-truth information for sinkholes in the study area and serve as primary references for the analysis.

High-resolution World Imagery [100] was utilised to analyse vegetative patterns across the study area. This dataset, provided by Esri, offers detailed satellite and aerial imagery with a spatial resolution ranging from 0.3 m to 15 m globally, depending on the source and location. The imagery includes contributions from multiple commercial and governmental sources, seamlessly integrated into a global mosaic. Vegetative features such as dense and sparse vegetation were visually derived from this dataset, allowing for the precise identification and delineation of vegetation classes that are difficult to map in remote or semi-arid environments. The high level of spatial detail provided by World Imagery enabled the analysis of fine-scale vegetative patterns across the study area, serving as a critical reference for understanding land cover dynamics in southern Kazakhstan.

1.2.2 Temperate Central European Forests

Amidst the growing recognition of the immense value of forest ecosystems in combating climate change and supporting biodiversity, the demand for rapid, precise, and robust methods to monitor these vital ecosystems is on the rise [270]. Forests, which shelter the majority of terrestrial biodiversity, span approximately 4.06 billion hectares, covering 31% of the world's land surface. They function as crucial carbon reservoirs and play an indispensable role in climate regulation [102]. To effectively track the progress toward

these goals, as well as to monitor deforestation, degradation, and forest responses to climate change, there is an increasing need for large-scale, cost-effective monitoring, ideally with automated data collection and processing up to the final information product. The past decade has witnessed a surge in the availability and utilization of RS technologies, offering data at resolutions high enough to identify individual trees. While this presents opportunities for enhancing our understanding of forests, it also poses challenges in data interpretation [222]. The field's expansion has resulted in an influx of intricate datasets, demanding the development of innovative data science approaches to efficiently extract ecologically significant information. Additionally, the lack of universally acknowledged benchmarking datasets has impeded methodological advancement and posed difficulties in comparing different studies. This perspective underscores the advantages of establishing and applying benchmarking datasets while outlining the key attributes that can optimize their value for the wider scientific community.

Central Europe's temperate forests, such as those in Bavaria, Germany, are under growing pressure from biotic and abiotic stressors, including storms, prolonged drought, and insect infestations. These forests represent critical ecosystems for biodiversity conservation, carbon sequestration, and regional economies. Maintaining their resilience in the face of climate change requires robust monitoring tools that can detect both acute disturbance events and more gradual structural changes. Traditionally, forest monitoring in Bavaria has relied on ground-based inventories and periodic aerial surveys conducted by forestry agencies. While highly accurate, these methods are limited by spatial coverage, cost, and temporal resolution. As a result, subtle or rapid changes, such as early-stage bark beetle attacks or post-storm canopy damage, are often missed. RS provides a scalable, repeatable, and wall-to-wall alternative, especially when combined with ML to detect and interpret complex spatial-temporal patterns. RS has a long tradition in forest science but remains underutilised in operational forestry practice in Bavaria [159]. Notably, airborne RS has been integrated into monitoring programs in protected areas such as the Bavarian Forest National Park since the 1980s [153]. Space-borne RS, however, offers broader scalability. Sentinel-2, for instance, provides 10 m resolution multispectral imagery with a 5-day revisit cycle, making it suitable for near-real-time monitoring across the state's 2.5 million hectares of forest [147]. Forest RS applications are increasingly addressing diverse topics: tree species mapping, biomass estimation, canopy health monitoring, phenology tracking, and biodiversity assessment [104, 159].

The contemporary realm of forest management witnesses the growing significance of the fusion between RS and forestry. This symbiotic relationship leverages the power of

ML and DL in handling vast, diverse datasets commonly found in RS images. ML and DL techniques aim to distil intricate image information into straightforward semantic interpretations, such as identifying a single pixel as a coniferous tree through time series analysis. These approaches, however, necessitate hyperparameterization to accommodate data variability and imperfections. This flexibility can lead to optimal adaptation to specific challenges but may also result in overfitting, necessitating a substantial number of training samples, analogous to the layers of a deep neural network. While three primary approaches are prevalent: (1) crafting comprehensive training datasets, which is labour-intensive and costly, (2) employing data augmentation to introduce artificial data variations, and (3) utilizing pre-trained networks with supplementary adaptation layers, it is important to note that method (1) is often indispensable for DL approaches, which often require extensive training data. In contrast, some ML methods can work with smaller datasets. Ongoing research delves into capitalizing on symmetries and commonalities within network layers to enhance decision-making and model robustness. However, methods (2) and (3) entail manual, non-standardized alterations to either the training data or the network architecture, rendering them suitable for specific applications but unsuitable for benchmarking ML and DL algorithms. Therefore, extensive training datasets (1) are indispensable, enabling the training of new algorithms, potentially even without prior knowledge, similar to pre-trained networks. Several benchmark datasets have already been established and made available to facilitate advancements in the field. Ongoing research is attempting to leverage symmetries and similarities within network layers to enhance or simplify decision-making by analyzing model zoos [295]. While such techniques are predominantly applied in DL, where they capitalize on the intricacies of network architectures, they are less common in conventional ML. This highlights a key difference between ML and DL: the flexibility and depth of DL networks, which enable the exploitation of these symmetries and commonalities. As both steps (2) and (3) require manual and non-standardized modifications of the training data (2) and/or the network (3), they are well-suited for specific applications but not for ranking ML and DL algorithms. This distinction underscores the need for standardized and extensive training datasets, particularly when benchmarking the performance of various ML and DL models in forest management. In 2018, the International Society for Photogrammetry and Remote Sensing (ISPRS) released a widely used benchmark dataset consisting of true orthophotos, surface models, and semantic labels for apparent land cover types such as impervious surfaces, buildings, low vegetation, trees, and vehicles [274]. This dataset has served as a standard reference for evaluating and comparing machine learning algorithms. More recently, the ML4Earth platform introduced the MDAS dataset, a new multi-modal

RS benchmark that integrates SAR, multispectral, and hyperspectral imagery alongside a surface model. Annotated land cover classes include pavement, low vegetation, soil, trees, roofs, and water. The dataset incorporates data from Sentinel-1, Sentinel-2, and EnMAP sensors [162], facilitating studies that span multiple spectral and sensor modalities. Most existing benchmark datasets focus on semantic segmentation tasks that classify pixels into nominal categories. However, efforts are emerging to capture ordinal information as well, for example, by assigning forest stands to hierarchical height classes such as low, medium, and tall, or by representing canopy density levels as sparse, moderate, or dense. One such approach was introduced in 2022 [194].

In Bavaria, bark beetle disturbances have been among the most intensively studied topics, reflecting their ecological and economic importance [74]. Early detection is critical, as trees in the green-attack stage show no visible symptoms, and outbreaks can escalate rapidly. Recent methodological advances show promising results:

- **Time-Series Change Detection:** Sentinel-2 vegetation index (VI) time series are analysed to detect abrupt drops in canopy greenness or structure, indicating disturbances. For example, change detection frameworks have been used to flag windthrow and early bark beetle activity in Bavarian forests, with validation against forest agency records [181].
- **Multi-Sensor Fusion:** Combining Sentinel-2 optical imagery with Sentinel-1 SAR data improves robustness under cloud cover. While SAR alone underperforms (max. accuracy: 0.62), fusion approaches can enhance detection of canopy structure changes. However, in one large-scale comparison, Sentinel-2 alone outperformed all fusion setups for detecting bark beetle infestations (max. accuracy: 0.93) [198].
- **ML and Deep Learning:** RF and CNN have been trained to classify bark beetle infestation stages using multi-temporal and spectral inputs. CNN-based methods have detected early infestation signals up to three weeks before field-confirmed emergence. Spectral indices related to water stress, particularly in the NIR and SWIR ranges, show the highest sensitivity [229].
- **Structural Modelling:** Regression models link EO features to forest parameters such as canopy height and biomass, allowing estimation of key attributes at scale. For example, Sentinel-2 data have been used to derive wall-to-wall maps of forest structure for Bavaria [74].

Despite clear progress, several key limitations persist in EO- and ML-based forest monitoring. First, benchmark datasets remain scarce, particularly those offering dense time series, high-quality airborne reference labels, and multi-sensor alignment. As outlined in Section 1.1.3, standardized datasets are essential for enabling reproducibility, model comparison, and fair evaluation across geographic and phenological conditions. Recent advances, such as Wald5Dplus [147, 148], mark a notable step forward by providing ARD cubes, harmonized Sentinel-1 and -2 stacks, and wall-to-wall forest attribute labels derived from airborne LiDAR and UAV data. Which is crucial, as label acquisition remains a bottleneck. Generating pixel- or plot-level reference data is costly and logistically demanding, especially for continuous variables such as biomass or canopy height. Even where ground inventories exist, they often differ in format, resolution, and acquisition date, limiting their direct usability for EO model training and evaluation. In Wald5Dplus, the image stacks are normalized and encoded as UInt8 for storage and compatibility with common GIS and ML tools [289]. Which directly addresses a third limitation. The volume and dimensionality of time-series EO data introduces both computational and methodological challenges. Dense Sentinel-2 observations (every 5 days) and multi-sensor fusion (e.g., with Sentinel-1 or Landsat) can lead to terabyte-scale datasets over even moderate AOIs. Cloud-based platforms like Google Earth Engine (GEE) or OpenEO are essential for processing at this scale, yet ML model portability, memory constraints, and reproducibility across platforms remain unresolved technical hurdles. Fourth, cloud cover, terrain shadow, and seasonal snow reduce data quality and increase missingness in optical imagery, especially in mountainous or high-latitude regions. Radar sensors help mitigate this issue, but their interpretation requires specialised preprocessing and often lower spatial resolution. Fifth, temporal misalignment between reference data and satellite observations introduces uncertainty in supervised learning. For instance, forest inventory data may be several months or even years out of sync with EO image acquisition, making label validity questionable, especially for fast-evolving disturbances like windthrow or bark beetle spread. Finally, while ML and DL models are effective in extracting complex patterns, interpretability and generalizability remain open challenges. Many models are tailored to local conditions or trained on specific forest types, limiting transferability. Additionally, uncertainty estimation is often lacking in DL pipelines, which reduces confidence in their use for decision-making in forest management or policy contexts.

RS and ML, when combined, enable scalable, near-continuous forest monitoring. While not yet a full replacement for field inventories, they provide essential complements, especially in remote areas or during rapid outbreak events. The integration of these

methods into routine forest management will be key to adapting Central Europe's forests to future climatic challenges.

Study Area and Environmental Characteristics

Each of the selected sites, Bavarian Forest National Park, Steigerwald, and Kranzberg Forest, offers distinct ecological and spatial features that significantly contribute to the richness and heterogeneity of the dataset. These attributes are further complemented by targeted field campaigns that include geotagged photographs (see Figure 1.6), visually documenting forest conditions as described in regional ecological studies. For example, images from the Bavarian Forest National Park highlight the widespread occurrence of deadwood [147], a key factor linked to increased vulnerability to infestations by the European spruce bark beetle [331, 199]. The Bavarian Forest National Park along the border between Germany and the Czech Republic is part of an approximately 2,000 square kilometres, densely wooded, middle-high mountain range in Central Europe. The altitudes range from 650 to 1,453 m a.s.l. This landscape belongs to the so called "Bohemian Masse", a very old mountainous region built of crystalline rocks such as gneiss and granite. Climatic conditions are cool and humid, with a mean annual temperature varying from 6.5 °C in the low mountain ranges to 3–4 °C at high elevation and an annual precipitation varying from approximately 1,000 mm in the valleys to 2,500 mm at high altitude. Long and cold winters with a lot of snow are typical for this region. Poor, acid and stony soils predominate. Wet and swampy soils play an important role, resulting in the development of peat bogs. 97% of the Bavarian Forest National Park is covered by forests. The most important forest communities are montane beech (*Fagus sylvatica* L.) forest with silver fir (or fir; *Abies alba* Mill.) and spruce (52%), subalpine spruce forest (19%), spruce–fir forest on wet mineral soils in the valleys (8%), and spruce forest on wet organic soils (6%) [152]. This susceptibility exposes the forest to recurrent disturbances caused by storms. As a national park, Bavarian Forest National Park intentionally resembles a "jungle" or primaeval forest, introducing substantial fluctuations in the ecosystem dynamics, thereby challenging model predictability. In contrast, the geotagged photos from Steigerwald [147] emphasize the forest's high density of deciduous trees and the outcomes of intensive forest management, particularly in its northern regions [245]. The Steigerwald Forest is located in northern Bavaria, Germany, and forms part of the Franconian escarpment landscape known as the "Fränkisches Schichtstufenland." Elevation ranges from approximately 190 to 500 m a.s.l., with the terrain shaped by the layered structure of Keuper formations, including sandstone and marl. The region is

characterized by a temperate climate, with mean annual precipitation ranging from about 650 mm in the forelands to up to 750 mm in the forested highlands. Eastern areas are drier, with rainfall decreasing toward the Regnitz Valley. The climate is relatively mild, but influenced by orographic effects along the Steigerwald escarpment. Soils vary widely, including nutrient-rich Gipskeuper substrates in the west and more acidic, sandy soils in the east. Forests cover large parts of the area and represent a structurally diverse and ecologically valuable landscape. Dominant forest types include montane beech (*Fagus sylvatica L.*) stands on mesic sites, mixed deciduous forests with oak (*Quercus robur L., Quercus petraea Mattuschka*) and hornbeam (*Carpinus betulus L.*) on lower slopes and terraces, and pine (*Pinus sylvestris L.*) forests on dry, sandy substrates of the eastern escarpment. Remnants of traditional coppice-with-standards management (*Mittelwald*) still persist in parts of the forest and contribute to the structural diversity. Many areas are characterized by near-natural and old-growth beech forests. The Steigerwald region is part of the Naturpark Steigerwald and hosts several nature reserves and Natura 2000 sites. The combination of geological heterogeneity, forest cover, and cultural land use history has resulted in a high level of biodiversity. The forest plays an important role in regional conservation, offering habitat continuity, high structural complexity, and climate resilience [45]. With 43% cover, beech *Fagus sylvatica* is the dom-inant tree species in the investigated section of the“Steigerwald” forest, followed by oak *Quercus petraeawith* 20%. Deciduous trees cover more than 70% [245]. This leads to a characteristic forest structure and a comparatively stable forest environment, which in turn enhances the reliability and accuracy of predictive modelling outcomes [147].

Despite its limited spatial extent, Kranzberg Forest features a notably balanced composition of tree species, primarily dominated by spruce (*Picea abies*) and beech (*Fagus sylvatica*), as evidenced by the geotagged field photographs. Kranzberg Forest (Longitude: 11° 39' 42" E, Latitude: 48° 25' 12" N, Altitude 490 m a.s.l.) is situated in southern Germany, approximately 35 km north-east of Munich. While its small size may restrict its use for large-scale modelling applications, it nonetheless provides meaningful insight into species coexistence within a compact forest system. The accompanying geotagged images add a valuable visual context to the dataset, reinforcing the documented forest structure with in-situ observations and enhancing the ecological depth of the data [147].

Reference Data and Labels

These landscapes represent a gradient of forest types, from managed to near-natural stands, offering a comprehensive foundation for forest characterization and monitoring. The selected forests span a wide range of structural and ecological attributes, including deciduous, coniferous, and mixed stands, as well as disturbed areas affected by natural and anthropogenic factors. Forests rank among the most ecologically complex and vital terrestrial ecosystems, serving critical functions in biodiversity preservation, carbon cycling, and climate regulation. As environmental pressures mount globally, the demand for accurate, scalable, and high-resolution forest monitoring solutions is growing. RS has become a core tool in this domain, enabling consistent, large-scale observation. Yet, the utility of RS-derived insights is fundamentally constrained by the quality, granularity, and representativeness of the reference data used to train and validate models. The Wald5Dplus dataset [147, 148] addresses this need by providing a comprehensive, multimodal benchmark for forest characterization, bridging the gap between raw satellite data and meaningful ecological insight. In particular, it emphasizes individual tree-level annotation using airborne laser scanning and multispectral data, enabling the derivation of critical forest attributes at scale. High-quality reference datasets are available for these regions, supporting a wide range of forest monitoring applications. These include continuous structural parameters such as tree height, crown volume, and crown area, as well as complementary field-based inventories. In addition to these detailed structural features, semantic disturbance labels, such as bark beetle infestations, enable a dual focus on both forest composition and dynamic processes. While Wald5Dplus serves as the conceptual foundation for structural characterization, this thesis integrates additional reference datasets to extend the scope of analysis, particularly with respect to forest disturbance monitoring and ecological interpretation. A key supplementary source are datasets provided through the Bavarian Forest National Park's Datapool initiative [203]. The next sections provide an overview of both labelled datasets used in this thesis.

Wald5Dplus Labels: Three well-defined AOI were systematically investigated, as shown in Figure 1.4. The reference data used in the analysis is directly aligned with these selected regions.

A overview of these AOIs, including their subdivisions referred to in this thesis, associated EO data years, and relevant tile information, is provided in Table 1.1 and Figures 1.2 and 1.3. For the Bavarian Forest National Park, both field transects (Transects 1–4)

and rasterized tiles (NP_T00–T11 for structural variables, NP_D01–D06 for deadwood variables) are distinguished, reflecting the dual structure of reference data in this region. While the label creation process is described for all AOIs, it is important to note that for the Bavarian Forest National Park, the field transects are described, however, by using the same procedure this was extended across the full national park area (29400 ha).

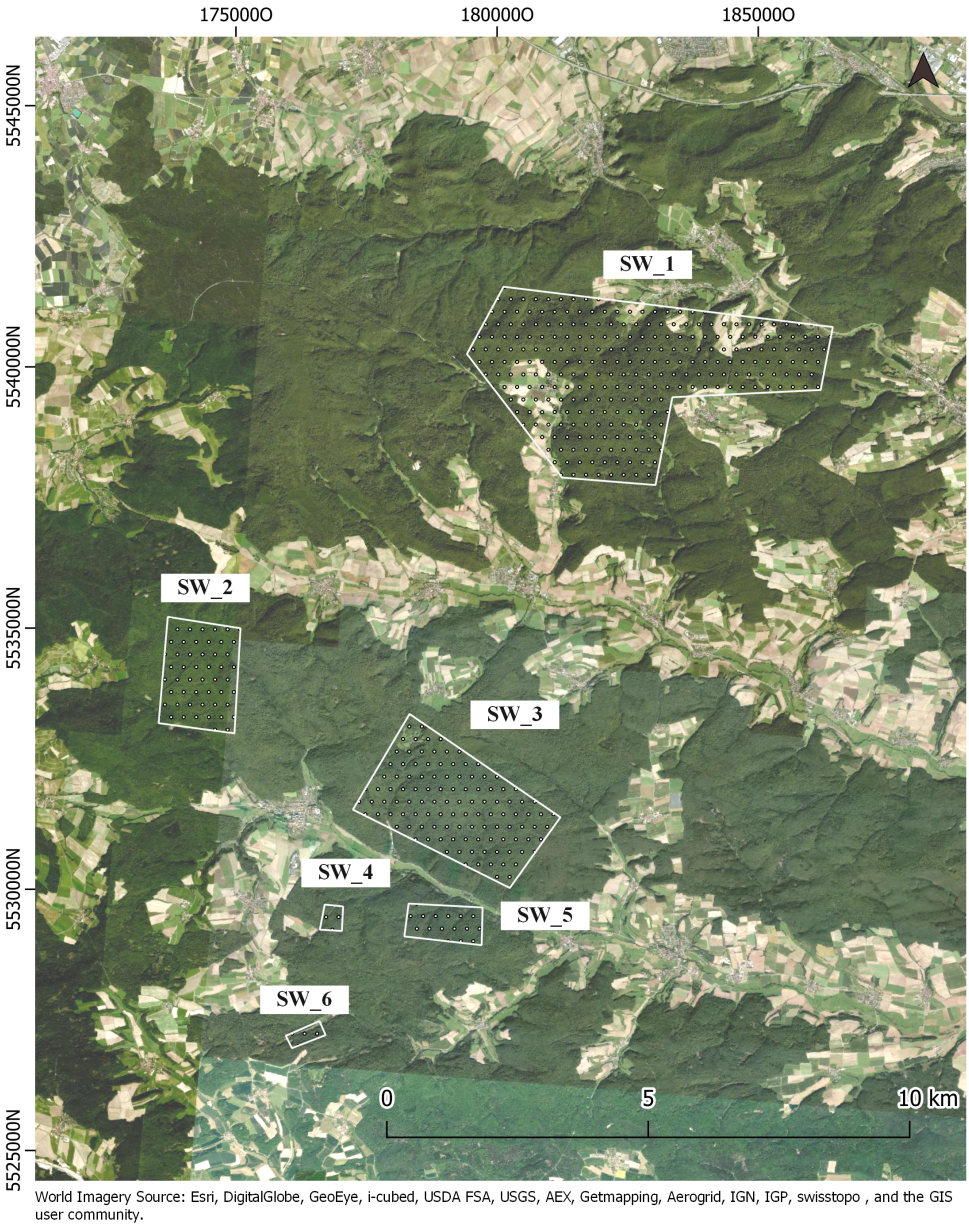


Figure 1.2.: Overview of Steigerwald (AOI 1) and it's subdivisions.

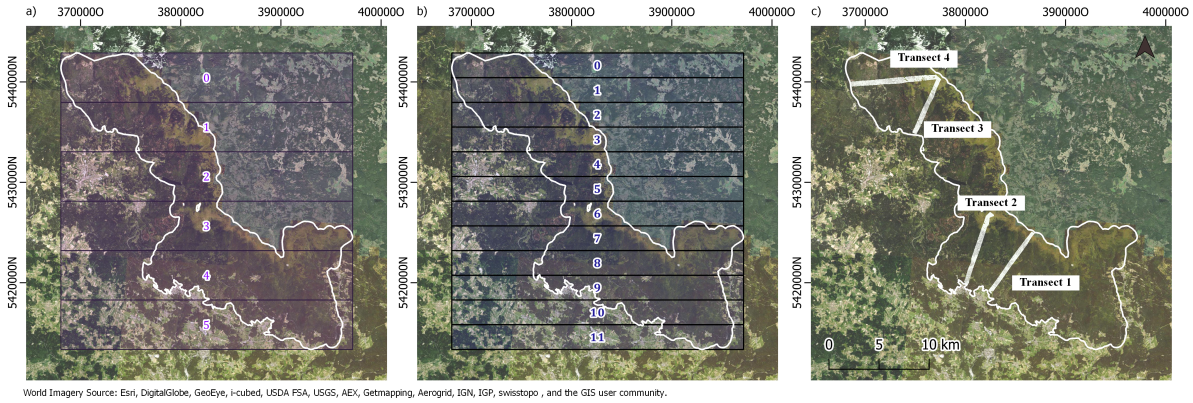


Figure 1.3.: Overview of the AOI 2 and it's subdivisions. **(a)** Raster tiles NP_D01–D06 covering deadwood variables in the Bavarian Forest National Park; **(b)** Raster tiles NP_T00–T11 covering structural forest variables; **(c)** Field transects 1–4 corresponding to ground reference measurements.

Table 1.1.: Overview of AOIs, Subdivisions, and Associated EO Data Years

AOI No.	AOI Name	Subdivision
1	Steigerwald	Sub-AOI 1 – 2020, 2021
		Sub-AOI 2 – 2020, 2021
		Sub-AOI 3 – 2020, 2021
		Sub-AOI 4 – 2020, 2021
		Sub-AOI 5 – 2020, 2021
		Sub-AOI 6 – 2020, 2021
2	Bavarian Forest National Park	Transects 1–4 (Reference Plots) – 2020, 2021
		Tiles NP_T00–T11 (8-band, structural variables) – 2020, 2021
		Tiles NP_D01–D06 (2-band, deadwood variables) – 2020, 2021
3	Kranzberg Forest	Core Region – 2020, 2021

Collection and Characteristics of Polygon Reference Data:

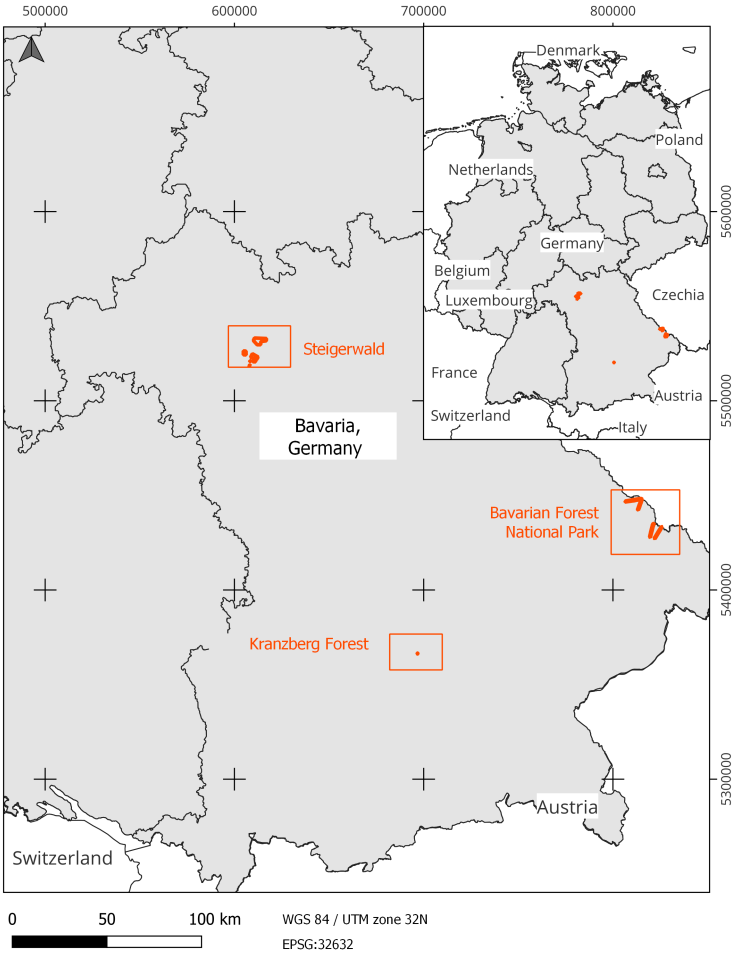


Figure 1.4.: Map of study sites displaying the three designated areas of interest

The reference data, represented in the tree polygons, comprise labelled tree segments, meticulously generated from full-waveform LiDAR and multispectral data within the designated research areas in accordance with the specifications delineated in Table 1.2. The utilization of full-waveform LiDAR data is instrumental segmenting single trees in different forest layers. The generation of tree segments is achieved through the application of a sophisticated normalized cut algorithm that systematically partitions the LiDAR point cloud into point cloud segments until predefined criteria are met and single trees are found. These tree segments encompass a multitude of calculated attributes, including tree height, crown diameter, crown volume and crown base height. The computation of

crown base height entails a meticulous process of distinguishing between crown points and ground points, subsequently stratifying the crown points into discrete 0.5-meter layers and ascertaining their respective crown point counts. The crown base height is then accurately determined as the height corresponding to 15% of the total number of tree points, as expounded in [271].

Table 1.2.: Characteristics of the examined Areas of Interest

AOI	Name	Coordinates	Platform	Year	Area	Trees
1	Steigerwald	48°25'N, 11°40'E	Helicopter	2017	2,600 ha	1,106,073
2	Bavarian Forest National Park	49°15'N, 13°15'E	Airplane	2016	1,443 ha	512,489
3	Kranzberg Forest	49°53'N, 10°32'E	UAV	2020	7 ha	1,467

The tree segmentation generated from the LiDAR point cloud harmoniously integrates with the multispectral data to facilitate feature extraction. By employing projected polygons of the segmented trees in combination with multi-spectral data covering the AOIs, a diverse array of classifications and feature sets is deployed, with the overarching goal of distinguishing between deciduous and coniferous trees, in addition to detecting deceased standing trees and snags, as comprehensively discussed in literature [15].

The individual tree polygons encapsulate critical information pertaining to each tree, encompassing details such as tree type, distinguishing between deciduous, coniferous, or identifying it as deadwood (=standing dead trees and snags). These polygons, which can partially overlap, further provide insights into the crown volume, the tree height, the specific crown base height as well as the crown volume. Validation of this approach was systematically conducted in Amiri et al. [15] and Krzystek et al. [193]. In a recent study [82], a novel tree detection method based on the Detection Transformer (DETR) was applied. The results demonstrated the potential of this approach, with F1-scores of 83 % for coniferous, 86 % for mixed, and 71 % for deciduous plots, outperforming significantly four baseline methods in all forest types. In summation, these validation endeavours affirm the robustness and adaptability of this approach across a spectrum of forest structures and environmental conditions. The holistic integration of full-waveform LiDAR data, adaptive algorithms, and advanced instance segmentation techniques collectively embodies the potential to markedly elevate the precision of tree segmentation. This heralds a notable stride forward in the realm of RS and forest-focused applications.

Rasterization Process for Use in EO Modelling:

In the endeavour to transition the properties encapsulated within the single-tree polygons, which initially comprise information, i.e., labels relating to leaf type, crown volume, tree height, and crown base height, onto a raster format without a substantial loss of detail, a conscientious aggregation process unfolds.

The procedure is implemented as follows: Employing a QGIS model, the input label data is extracted from the single tree polygons. These labels encompass crucial insights into the nature of the trees, distinguishing between deciduous and coniferous varieties, as well as identifying those categorized as deadwood. Furthermore, the polygons encompass attributes detailing the crown volume, tree height, and crown base height. Of notable significance is the generation of a model output raster, consisting of ten distinct bands, each conveying distinct metrics derived from the tree segments. These bands encapsulate the core information extracted from the single tree polygons. It is essential to emphasize the seamless integration of this model output raster with the input satellite raster. This integration operates harmoniously with a 10-meter grid meticulously aligned with its spatial coordinates. The Bavarian Forest AOI adheres to UTM zone 33N (EPSG: 32633), while the other two AOIs lie within UTM zone 32N (EPSG: 32632) due to the inherent characteristics of the satellite data. The primary challenge encountered during this intricate process lies in the development of a method capable of robustly extracting single tree polygon information and accumulating the associated values within the new raster cells. The resulting raster bands are presented in Table 1.3.

Table 1.3.: Rasterized single-tree polygon bands capturing key forest attributes in a 10 m grid format.

Band	Variable	Unit	Value Range
1	Sum crown area of deciduous trees	m^2	0–170
2	Sum crown area of coniferous trees	m^2	0–170
3	Sum crown area of dead trees	m^2	0–120
4	Count of deciduous trees	Count	0–9
5	Count of coniferous trees	Count	0–9
6	Count of dead trees	Count	0–7
7	Tree area coverage	%	0–100
8	Sum crown volume	m^3	0–3000
9	Mean tree height	m	0–43
10	Mean crown base height	m	0–24

The calculation of values related to crown volume involves multiplying the crown volume by an area factor. For the three tree type count bands, the area factor is summed for each tree type. These calculations are executed through the derivation of an area ratio. This ratio represents the proportion of an attribute's area within a raster cell concerning the total area of the same attribute in the intersected polygons. Applying this area ratio method results in an adjustment of crown volume values based on the extent of the intersection between tree segments and raster cells.

The area ratio method described above can be mathematically expressed through weighted sums that adjust attribute values according to their spatial intersection with raster cells. Specifically, the adjusted crown volume and mean tree height per raster cell are calculated using area-weighted values.

The adjusted crown volume (V_{adj}) is computed as:

$$V_{\text{adj}} = \sum_{i=1}^{n} V_i \cdot r_i \tag{1.1}$$

Similarly, the adjusted mean height (H_{adj}) is derived using:

$$H_{\text{adj}} = \frac{\sum_{i=1}^{n} H_i \cdot r_i}{\sum_{i=1}^{n} r_i} \tag{1.2}$$

In these expressions:

- V_i and H_i represent the crown volume and height of the i-th tree segment, respectively.
- r_i is the area ratio, i.e., the proportion of the tree's crown area within the raster cell.
- n is the number of tree segments intersecting the raster cell.

Equations (1.1) and (1.2) ensure that crown volume and tree height values are properly weighted based on the degree of overlap between each tree segment and the raster grid cell.

For tree height and crown base height, a weighted arithmetic average calculation, as defined in Eq. 1.3, is implemented for each intersected raster cell.

$$\bar{h} = \frac{\sum_{i=1}^{n} a_i \cdot h_i}{\sum_{i=1}^{n} a_i} \tag{1.3}$$

Within the equation, a represents the area of the intersected polygons, and h represents either tree height or crown base height, depending on the specific attribute being calculated. Equation 1 is applied to all polygons within a raster cell. The area of the intersected polygons thus serves as a means to proportionally adjust the attribute heights in accordance with the portions of their area within a given raster cell. The tree type labels, categorizing trees as deciduous, coniferous, or deadwood, have their areas calculated per pixel. This is achieved by aggregating the area of all polygons with their respective tree type within a pixel grid. The counts of these areas are summarized for each tree type, with the previously described area ratio method applied per tree type. In addition to the tree type areas, a percentage value denoting the tree type coverage of a pixel is calculated. The resulting value represents the proportion of the grid cell's area occupied by tree segments, with each cell standardized to 100 square meters. It is important to note that this calculation does not consider overlapping polygons.

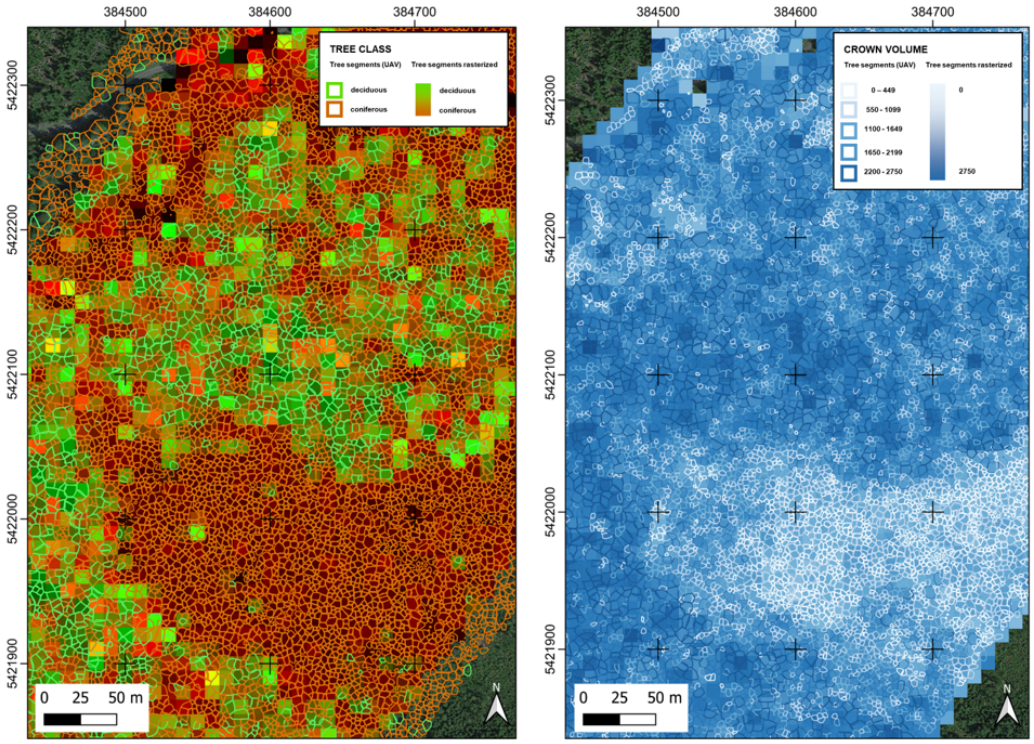


Figure 1.5.: Exemplary aggregation results of the tree segments onto the 10 m grid of the raster data, displaying the tree class (**l.**) and the crown volume (**r.**) of a subset in the Bavarian Forest National Park test site (AOI 2).

Figure 1.5 illustrates the exemplary results of aggregating individual tree segments onto the 10 m raster grid, showing (**l.**) the tree class distribution (coniferous vs. deciduous) and (**r.**) the corresponding crown volume estimates within a subset of the Bavarian Forest National Park test site (AOI 2).

Ground truth from Field campaigns: As consistency and continuity are two key features of training datasets, the labels generated from LiDAR points clouds and multi-spectral images acquired by airborne sensors are checked during several field campaigns. Representative forest stands within the test sites were documented by field-walking and taking geotagged photos (see Figure 1.6). Using the saved coordinates and the orientation, the forest stands can be qualitatively assessed.

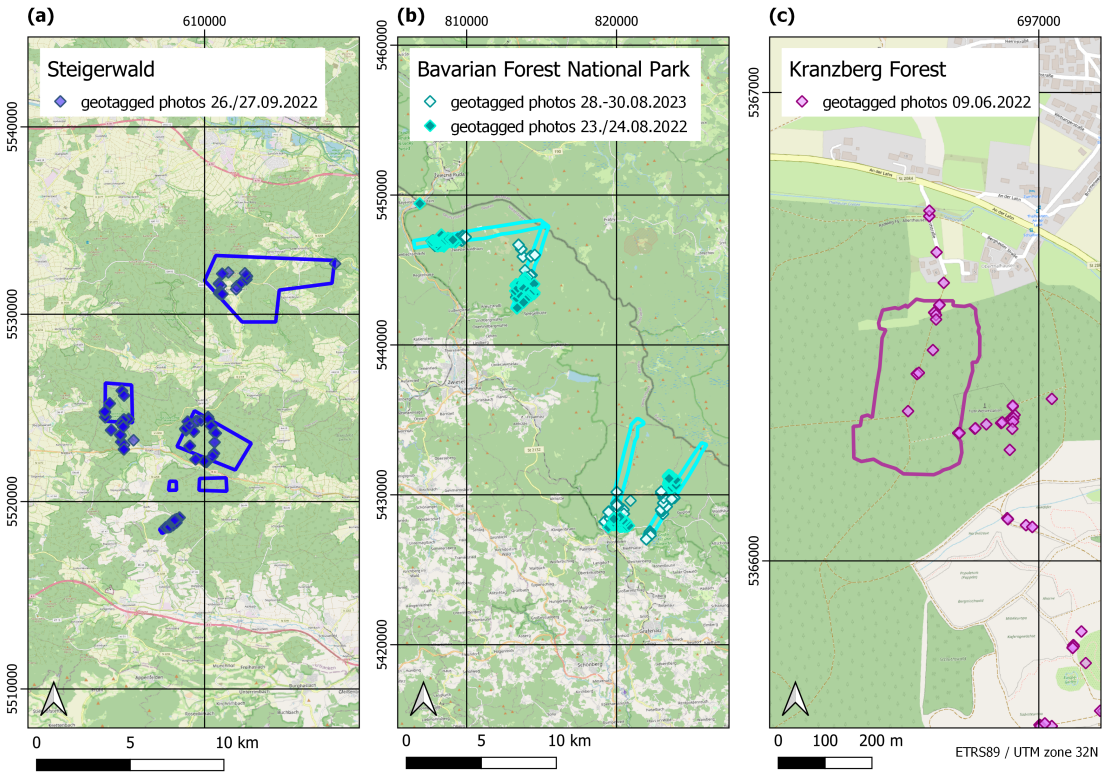


Figure 1.6.: Ground truth from Field campaigns in the three designated study sites (a) Steigerwald, (b) Bavarian Forest National Park, and (c) Kranzberg Forest.

Deadwood and Bark Beetle Disturbance Labels: In addition to this described static reference dataset, further reference labels are available within the Bavarian Forest National Park (Figure 1.4 and Table 1.2). These datasets include temporal dynamic bark beetle-affected areas, distributed by the Datapool Initiative [203].

The deadwood dataset encompasses all forest areas identified as standing deadwood from 1989 to 2023. It is updated annually based on aerial image interpretation and provides temporal resolution at the scale of one year. Each polygon includes three key attributes:

- `Not_Before` – the last confirmed date when the area was intact forest
- `Not_After` – the first date the area was classified as deadwood
- `Change_Year` – the year, whereas the transition from forest to deadwood was observed

The attribute `Change_Year` refers specifically to the period of visible transition in aerial imagery, typically occurring between late summer of the change year and early summer of the following year. For instance, a `Change_Year` value of 2020 indicates that tree mortality became visible sometime between August 2020 and June 2021. Due to the dataset's reliance on annual imagery surveys, the exact timing of mortality events within this interval cannot be resolved more precisely.

The dataset also indicates whether deadwood was removed or left standing. Given the National Park's strict non-intervention policy, removal is rare and generally limited to buffer zones near park boundaries. While species information is not included, it is assumed that most deadwood results from *Ips typographus* infestations in *Norway spruce (Picea abies)*, which have driven extensive dieback events, especially since the early 2000s. Other disturbance agents were not systematically mapped during the observation period.

1.2.3 Canadian High Arctic

Glaciological research plays a vital role in environmental monitoring, particularly in light of ongoing climate change. Glaciers and ice sheets are among the most responsive indicators of both natural climate variability and human-induced warming [46], making their observation essential for understanding trends in global sea-level rise, freshwater availability, and regional climate effects. The retreat of glaciers contributes to heightened local geohazards [136] and has far-reaching consequences for marine [160] and terrestrial ecosystems [109, 47], regional hydrological systems [166], and the global water and energy balance [338, 87]. Alongside the Greenland and Antarctic ice sheets, mountain glaciers are major contributors to both current [307, 28] and projected [275, 231, 158] sea-level rise. Around the year 2000, glaciers excluding the two continental ice sheets spanned approximately 706,000km^2 worldwide [272], with a combined ice volume estimated at 158,170±41,030km^3, corresponding to a potential global sea-level rise of 324 ±84mm [103].

Glaciers form in regions where long-term snow accumulation exceeds the amount lost through melting and sublimation. The key processes involved are accumulation, ablation, and the development of the glacier tongue [306]. The accumulation zone is located above the equilibrium line altitude (ELA) and is permanently covered by snow. In this zone, snow is added through precipitation, firnification, and wind transport. The equilibrium

line marks the altitude where accumulation and ablation are balanced on average. Below the equilibrium line lies the ablation zone, where snow and ice are lost through melting, runoff, sublimation, calving, and wind erosion. The glacier tongue represents the lower end of the glacier and often serves as the origin of meltwater streams. Its shape and extent are influenced by the glacier's mass balance, valley morphology, and size [306]. Snow metamorphism refers to the physical changes that snow undergoes after deposition. Two main types of metamorphism are recognized. Destructive metamorphism occurs through melting, refreezing, mechanical compression, and sublimation, leading to a reduction in snow crystal size and surface energy. Constructive metamorphism occurs under cold conditions, where water vapor deposition causes the growth of faceted crystals such as depth hoar. The density of snow increases as metamorphism progresses. Fresh snow typically has a density below 0.1 g/cm^3, while old snow may reach densities between 0.2 and 0.4 g/cm^3. Firn, snow that survives at least one summer, has a density between 0.4 and 0.83 g/cm^3. Glacier ice, formed from the compression of firn, typically has a density between 0.83 and 0.917 g/cm^3 [137]. Glacier facies [36, 244], such as bare ice, superimposed ice, or firn, also called glacier zones, are parts of a glacier that differ in characteristics such as structure, density, percolation properties or albedo. Together, they form either the accumulation or ablation zone of a glacier (i.e., a zone where a glacier either gains or loses mass in a given time span) [72]. Changes in the extent of a glacier zone at the end of an ablation season (late summer and autumn) are one of the indicators of glaciers' state and their response to climate change. Due to physical differences between glacier zones, information about their extents can support studies of glacier mass balance, hydrology, and other components of the surrounding environment. This concept of glacier zones was developed through field studies on the Greenland Ice Sheet and Arctic glaciers. These zones categorize areas of a glacier based on physical snow and ice characteristics, primarily influenced by snow metamorphism and mass balance processes [36, 253]. This commonly adopted classification scheme distinguishes the following glacier zones:

- **Dry Snow Zone:** An area where no melting occurs throughout the year, and snow undergoes only mechanical and thermal metamorphism.
- **Percolation Zone:** An intermediate zone where surface melting occurs during summer, and meltwater refreezes within the snowpack, forming ice lenses and layers.
- **Wet Snow Zone:** A zone where the entire seasonal snowpack melts during the ablation season, resulting in saturated snow and supraglacial water features.

- **Superimposed Ice Zone:** A region where refrozen meltwater accumulates as an ice layer on top of the glacier surface.
- **Bare Ice Zone:** The lowermost glacier area where seasonal snow completely melts, exposing bare glacier ice.

EO has become a central tool for glaciological research, enabling the continuous and large-scale monitoring of glaciers across remote and often inaccessible regions. The use of EO allows researchers to systematically observe key glacier parameters such as extent, surface properties, flow dynamics, and mass balance changes over various temporal and spatial scales. EO datasets are crucial for documenting the response of glaciers to climate change, tracking glacier retreat, quantifying mass loss, detecting surface elevation changes, and understanding surface melt processes. The ability to acquire frequent and consistent data has significantly enhanced long-term glacier monitoring programs and complements traditional fieldwork by extending observations beyond ground-based limitations [256].

EO techniques applicable to glaciology primarily include optical, thermal infrared, and active microwave systems. Each technology provides distinct information critical for understanding glacier systems: Optical sensors such as those on Landsat, Sentinel-2, and MODIS platforms provide data in the visible and near-infrared spectral ranges. They are primarily used for:

- Mapping glacier extent and delineating glacier boundaries [143].
- Monitoring seasonal snow cover [4] and surface albedo changes [106]
- Identifying supraglacial lakes [156], debris cover [186], and meltwater features [247].

However, optical observations are limited by atmospheric conditions, cloud cover, and the need for daylight, particularly problematic in polar regions. Thus Optical and SAR RS data are increasingly used in combination to monitor, e. g. the seasonal evolution of supraglacial lakes, which influence glacier dynamics through enhanced melt and basal lubrication. Multi-sensor approaches have proven effective for generating consistent time series of lake area and linking these patterns to climatic drivers such as temperature and precipitation anomalies [353]. Thermal infrared sensors capture the surface temperature of glaciers and are useful for:

- Detecting surface melt onset and identifying melt zones.

- Characterizing thermal anomalies and supraglacial water bodies.
- Supporting energy balance studies on glacier surfaces.

Thermal data contribute to understanding the spatial variability of melting processes and energy fluxes in glaciated environments [256].

Active radar systems, especially SAR are indispensable for glaciological studies due to their ability to acquire data independent of illumination and weather conditions. SAR is particularly valuable for:

- Measuring glacier surface velocities through feature tracking and interferometry (InSAR, DInSAR).
- Mapping accumulation and ablation zones based on radar backscatter variations.
- Monitoring surface roughness, meltwater presence, and changes in snowpack structure.

Radar data allow precise tracking of glacier motion and deformation over time, and enable mass balance assessments over large areas [240]. Paterson [253] introduced the concept of glaciological snow zones based on field observations, while Rau et al. [268] expanded this framework by associating them with characteristic radar responses observed through remote sensing (see Figure 1.7). Although radar observations provide valuable data, the fundamental understanding of glacier zones remains rooted in in-situ physical measurements. Complete sequences of glacier zones are rare but can occasionally occur under specific climatic conditions, such as on Ellesmere Island or Axel Heiberg Island [267].

Nonetheless, their delineation has been increasingly supported by advances in SAR RS, particularly under polar conditions where optical imagery is often limited due to persistent cloud cover or polar night [36, 244, 72, 17]. Several SAR-based methods are employed for glacier zone detection:

- **Backscattering coefficient (σ^0)**: This commonly used SAR metric reflects microwave energy from glacier surfaces and subsurfaces, and has been applied extensively with dual-pol C-band SAR data for glacier zone mapping [201, 7].
- **Pauli decomposition**: This method separates SAR returns into odd-bounce, even-bounce, and volume scattering components, enabling structural interpretation of glacier facies [252, 300].

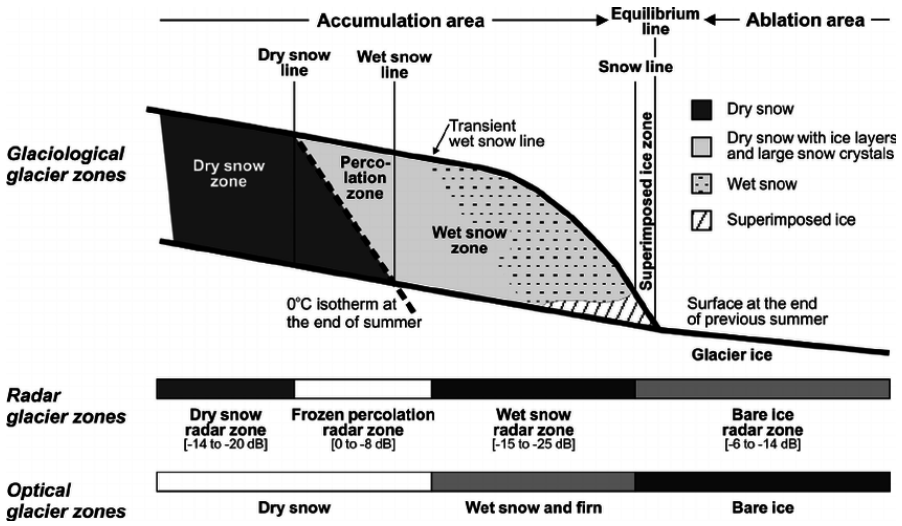


Figure 1.7.: Schematic representation of glaciological snow zones based on Paterson [253] and corresponding radar glacier zones based on Rau et al. [268].

- **H/A/α decomposition and H/α segmentation**: These techniques analyze entropy (H), anisotropy (A), and the mean scattering angle (α) to identify dominant scattering mechanisms. The H/α segmentation proposed [71] has been widely used, with later enhancements using Wishart and K-Wishart classifiers for unsupervised classification [71, 207].
- **Unsupervised classification methods**: Algorithms such as K-means and Gaussian Mixture Model–Expectation Maximization (GMM-EM) are utilized to cluster SAR-derived features (e.g., σ^0, Pauli components) into glacier zone classes [30, 53, 31].

Advances in EO include the fusion of optical, thermal, and radar datasets to enhance glacier mapping and monitoring accuracy. High-resolution sensors enable the study of smaller glaciers and detailed surface processes. The increasing application of ML and DL techniques allows for automated glacier classification, change detection, and time-series analysis, significantly accelerating data processing workflows. New EO products, such as TanDEM-X DEM change maps, provide highly precise surface elevation change data, which are instrumental in quantifying glacier mass loss over time. However, while EO offers unprecedented coverage and data richness, the availability and quality of training and reference datasets remain a major bottleneck. Especially in glaciology, where surface conditions are highly dynamic, labels must be temporally synchronized with the EO acquisitions to avoid outdated or inconsistent ground truth. Glacier facies can shift rapidly over months or even weeks due to seasonal melt, snowfall, or dynamic processes

like surging. Consequently, pre-processing of reference datasets, including temporal validation, dynamic filtering, and careful interpolation, is critical to ensure that ML models are trained on temporally relevant and physically consistent information. EO is therefore not only essential for modern glaciological research but must be complemented by robust, time-aware label engineering to fully unlock its potential. Together, EO data and temporally aligned reference information contribute to a comprehensive and scalable framework for monitoring glacier dynamics, assessing climate change impacts, and supporting informed scientific and policy decisions.

Study Area and Environmental Characteristics

The glaciology-focused AOI in this study comprises two Arctic islands in the Canadian High Arctic: Axel Heiberg Island and Devon Island (see Figure 1.8. These islands form part of the Queen Elizabeth Islands, a cluster located in Nunavut and the Northwest Territories. With Ellesmere Island, they account for approximately 14% of the global glacier and ice cap area, excluding Greenland and Antarctica, highlighting their global significance for cryospheric and climate research [301].

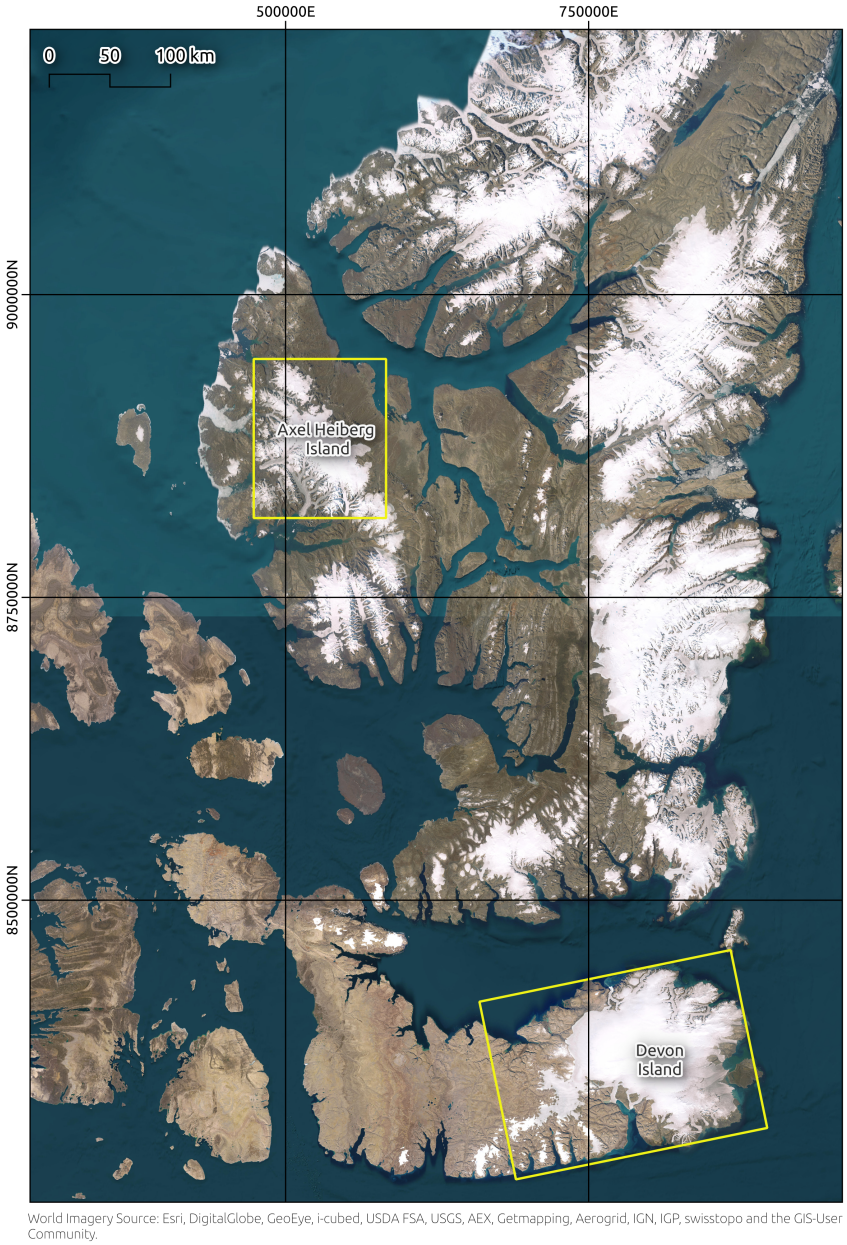


Figure 1.8.: Overview of the Canadian High Arctic glacier study areas, comprising Axel Heiberg Island and Devon Island.

Geologically, the Queen Elizabeth Islands are composed of folded and eroded Cambrian to Upper Devonian sedimentary rocks, with intense orogenic features giving rise to mountain ranges exceeding 2,000 m in elevation, particularly on the eastern islands. The diversity of landforms and permafrost conditions across these islands provides a valuable natural

laboratory for investigating glacier dynamics under varying environmental influences. Roughly one-fifth of the land area is ice-covered, with the largest ice masses located on Ellesmere, Axel Heiberg, and Devon Island

Axel Heiberg Island (approx. 43,000 km^2) features a rugged topography dominated by the Princess Margaret Range, peaking at Outlook Peak (2,210 m a.s.l.). Roughly 35% of the island is permanently ice-covered. Its glacier systems include two major ice caps, the Müller and Steacie Ice Caps, as well as numerous outlet and valley glaciers such as Iceberg, Airdrop, Thompson, Strand, and White Glacier. These glaciers vary substantially in morphology, with calving termini (e.g., Iceberg Glacier) and land-terminating outlets (e.g., White Glacier), making them ideal for comparative analysis of glacial responses to climatic changes.

White Glacier, located near Expedition Fjord, is a key focal point in this thesis due to its well-documented long-term observational record. The glacier spans an elevation gradient from 1782 m to 100 m a.s.l., with a wide accumulation zone narrowing into a steep valley tongue. The glacier's polythermal structure, featuring a cold upper shell and temperate base, makes it especially valuable for evaluating multi-sensor remote sensing data over heterogeneous glacial conditions [44].

The regional climate is defined by polar desert conditions, with mean annual temperatures around $-19.7\,^{\circ}C$ and mean July maxima of $+5.4\,^{\circ}C$, based on data from EUREKA station on nearby Ellesmere Island. Although annual precipitation averages only 64 mm at low altitudes, accumulation rates can reach up to 370 mm a^{-1} at higher elevations on Müller Ice Cap [73]. Long-term records confirm significant warming since the 1970s, particularly during winter, along with an increase in annual precipitation [211].

Devon Island, to the south-east, complements this Arctic AOI cluster by offering a contrasting glaciological setting in the form of the Devon Ice Cap, Canada's largest ice cap entirely situated on a single island. Encompassing approximately 14,000 km^2, it is located on a high-elevation plateau that provides a relatively stable and well-stratified glacial system. Its gentle dome-shaped topography facilitates the formation of extensive firn zones and well-preserved accumulation layers, which are key for reconstructing past climate variations and validating radar-penetrative remote sensing techniques. The ice cap drains through several large outlet glaciers, including Belcher and Sverdrup Glaciers, which terminate in fjords and occasionally calve into the ocean.

The Devon Ice Cap features clear stratigraphic zonation: an upper accumulation zone with persistent firn cover, a transitional percolation zone, and a lower ablation zone

affected by strong seasonal melt. This structure supports a range of applications in glacier mass balance modelling and SAR-based classification. Long-term monitoring has shown a consistent trend of mass loss, particularly at the termini of its outlet glaciers, aligning with pan-Arctic glacial retreat patterns. Its accessibility, extensive research history, and stable internal dynamics make Devon Island a valuable calibration and validation site for multi-temporal EO studies.

Collectively, Axel Heiberg and Devon Islands encompass a diverse spectrum of glaciological environments, ranging from fast-flowing valley glaciers with steep gradients to expansive, slow-changing ice caps. Their spatial and climatic diversity enables comprehensive evaluation of remote sensing techniques, particularly for multi-temporal, spectro-polarimetric fusion and classification. The combination of documented historical observations (e.g., on White Glacier), strong topographic gradients, and ongoing climate-driven changes make this AOI uniquely suited for high-dimensional glacier research within this thesis.

Reference Data and Labels

The glacier zone reference dataset used in this thesis was produced as part of a thesis by [311], which developed and applied a robust methodology for classifying glacier facies across the Canadian High Arctic. This classification approach is based on TerraSAR-X (TSX) ScanSAR imagery collected between 2017 and 2023, with a revisit interval of 11 days and a ground sampling distance of 40 meters. The resulting dataset provides a high-resolution, multi-temporal characterization of glacier surface zones over Axel Heiberg Island and Devon Island. Five radar glacier zones were defined and mapped across all three AOIs, adapted from established literature and calibrated for local glaciological and radiometric conditions:

- **Dry Snow Zone** – consistently snow-covered, low radar backscatter;
- **Frozen Percolation Zone** – refrozen melt layers, enhanced volume scattering;
- **Superimposed Ice Zone** – surface refreezing, moderate radar return;
- **Bare Ice Zone** – exposed glacial ice with high backscatter;
- **Wet Snow Zone** – saturated snow with low backscatter due to signal attenuation.

The threshold calibration for facies classification followed a histogram-based approach using HH-polarized backscatter values expressed in Gamma-Naught (γ_0) [311]. Winter TSX scenes were prioritized for threshold detection due to their minimal variability in surface conditions, which allowed clear identification of class-specific peaks in the backscatter distributions. This approach, while methodologically simple, proved highly effective for detecting consistent zone separations over time. In addition, a earlier work [146, 145, 273, 352] served as a conceptual and empirical guide for class boundary definition, e.g., for Axel Heiberg Island [146, 145], offering initial benchmarks for radar zone properties under similar climatic and topographic conditions.

To generate a consistent and scientifically valid glacier zone product, the following auxiliary datasets and technical procedures were employed [311]:

- **GLIMS glacier outlines** [269, 126] were used to spatially constrain classification and ensure that only glacier-covered areas were analyzed;
- **Copernicus DEM (GLO-30)** [98] elevation data were used to calculate elevation statistics for each glacier zone. Due to differing pixel resolutions between the DEM and SAR imagery, DEM values were assigned to SAR pixels using a nearest-neighbor method based on minimum centroid distance;
- **Scene-specific histogram analysis** was applied to detect peak clusters in backscatter values, which were then assigned to glacier facies based on elevation stratification and temporal behaviour;
- **Multi-temporal SAR filtering** and **Gamma-Naught radiometric terrain correction** were applied to reduce speckle, normalize for topographic effects, and ensure radiometric comparability across acquisition dates and glacier geometries;
- **Multi-SAR preprocessing** [38] was used to standardize data input formats and ensure compatibility across the large multi-scene time series.

While unsupervised clustering methods (e.g., k-means, EM) were explored, they were ultimately not adopted due to their limitations when applied to univariate backscatter data. Instead, the peak-based thresholding approach, supported by literature values and local calibration, was found to provide more stable and interpretable results (see Table 1.4) [311].

Table 1.4.: Thresholds for Glacier Zones Derived by Histogram Peaks Method [311]

Glacier Zone Transition	Threshold (dB)
Dry Snow Zone	> -1.5
Frozen-Percolation Zone	> -5.7
Superimposed Ice Zone / Bare Ice Zone	> -10.4
Wet Snow Zone	< -15.0

The final reference product consists of spatially explicit, zone-labelled glacier masks at a roughly 7-day intervals for all two AOIs. These maps not only allow detailed assessment of seasonal surface dynamics but also form a reliable reference for the supervised training and validation of glacier zone modelling within this thesis.

1.3 Main Objective and Research Goals

The overarching goal of this thesis is to advance ML-based environmental monitoring by systematically exploring the interplay between EO and ML in a multi-modal, multi-model, and temporally explicit manner. Central to this exploration is the evolution of EO feature representations, from spectral-only setups, to polarimetric, to fused spectral–polarimetric, and ultimately to multi-temporal, multi-modal fusion using hypercomplex algebra [289]. This progression enables a systematic EO to ML based predictive task evaluation of forest-related parameters. Another focus of this thesis lies also in the exploration of temporal fusion strategies as a means to extract deeper information from EO data. Rather than treating time merely as a sequence, intra- and inter-seasonal acquisitions are deliberately fused under various configurations. These scenarios are evaluated not only for their predictive utility but also for their capacity to reveal latent spectral–structural interactions across time. This systematic analysis culminates in the development of a novel, cross-seasonal, multi-modal EO index, designed to capture land surface characteristics such as vegetation–sinkhole interactions in arid karst systems with improved robustness and ecological relevance. Another critical challenge addressed by this thesis is the underexplored bottleneck of label quality, temporal consistency, and structural design in EO-based ML workflows. While much progress has been made on model architecture and

feature engineering, label-side innovations have lagged behind. To address this gap, the *HELIX* framework is introduced, a modular system for context-aware label enrichment. HELIX systematically enhances weak or sparse labels by integrating multi-scale spatial statistics. By including this way model-informed residuals, this enables conventional ML models to exploit nuanced structural and temporal patterns without requiring costly, handcrafted architectures. By directly linking EO input and label design, the HELIX supports robust learning in challenging supervision regimes. This alignment enables traditional ML models to effectively leverage temporally structured EO inputs without requiring complex temporal architectures, such as LSTMs.

The work is guided by the following core research aims:

- **To evaluate the predictive capacity of individual, multi-modal and multi-temporal EO modalities**, by progressing from individual modality inputs, including Sentinel-2 spectral bands, their Kennaugh-like transformations, and Sentinel-1 polarimetric Kennaugh elements (including TSX and ALOS), to spectral–polarimetric fusion using hypercomplex algebra, and ultimately to fully fused spectrally, polarimetrically and temporally representations. Whereas, these strategies are benchmarked across diverse predictive models, including both intra-AOI and full spatial transfer scenarios.
- **To investigate discrete temporal fusion strategies**, such as intra-seasonal and cross-seasonal fusion, with an emphasis on capturing phenological-and structural variation and improving robustness across varying observation conditions.
- **To develop and validate HELIX-based label enrichment methods**, which spatially and temporally align dynamic reference labels with EO features, and enrich them using if possible local spatio-temporal context. In modelling, these methods incorporate spatial context (e.g., neighbourhood statistics) and residuals from baseline models to encode uncertainty and structural ambiguity. This enriched supervision enables more informed learning, especially in complex or under-annotated scenarios. The framework is applied across diverse domains, including forest structure regression, bark beetle and storm disturbance detection, and seasonal glacier zone classification, demonstrating its flexibility and impact.

Together, these objectives define a coherent experimental framework that systematically probes the relationship between EO feature complexity, label design, and predictive model performance. The resulting insights inform the development of scalable, modular EO–ML

pipelines that are suited for real-world environmental monitoring tasks, particularly under conditions of data sparsity, label uncertainty, and dynamic change.

1.4 Thesis Outline and Contributions

This thesis is organized into eight chapters, each contributing a distinct conceptual and methodological layer to the overall research goal: improving environmental monitoring through systematic integration of EO features, reference labels, and ML techniques. Figure 1.9 illustrates the structural logic and thematic progression. Chapters 2 and 3 establish the two foundational pillars of the thesis, EO feature engineering and reference data design, while Chapters 4 to 7 build upon them through targeted methodological innovations and applied experimental analyses.

Chapter 1: Introduction — This chapter motivates the scientific and operational relevance of EO-based environmental monitoring. It introduces the methodological landscape (EO, ML, benchmarking), outlines the application contexts (vegetation–sinkhole systems, forests, glaciers), presents the AOI and associated reference datasets, and defines the overarching research objectives.

Chapter 2: Consistent EO Feature Generation — This chapter lays the foundation for EO feature design. It introduces multi-level fusion strategies (pixel-, feature-, and decision-level), with a strong focus on temporal-aware fusion schemes. Hypercomplex algebra [289] is introduced as the core mechanism to jointly represent spectral, polarimetric, and temporal signals.

Chapter 3: Labelling Foundations and Challenges — This chapter addresses the often-overlooked bottleneck in EO–ML pipelines: the quality, structure, and temporal alignment of reference data. It surveys key challenges in label temporality, noise, sparsity, and task-definition, setting the stage for structured label engineering.

Chapter 4: The Novel Helix Framework for Dynamic Label Data — Building on Chapter 3, this chapter presents the HELIX framework, a modular system for context-aware label enrichment. The HELIX aligns labels with EO features across space and time, treating labels as a dynamic string of voxel-wise annotations, and enriches them using spatial context (e.g., local neighbourhood statistics) and model-informed residuals. It provides a scalable strategy to enable robust learning in weakly supervised, temporally dynamic settings.

Chapter 5: Temporal Dynamics in EO Feature Engineering — This chapter explores the impact of different temporal fusion strategies on EO feature quality. Through a specific use case, arid-zone sinkhole–vegetation monitoring, it evaluates intra-seasonal vs.cross-seasonal configurations, showing how fusion timing affects EO-based land cover identification.

Chapter 6: Foundational Analysis of EO Modality–Model Interactions — This core chapter systematically benchmarks EO-modality–model configurations for continuous label prediction using the Wald5Dplus Label dataset. It compares individual modalities (Sentinel-1, Sentinel-2, TSX, ALOS), and their fusion using hypercomplex methods. Transferability across spatial settings is also evaluated, revealing the trade-offs between EO complexity and model generalization.

Chapter 7: Context-Aware Label Enrichment and Multi-Scale Learning with the HELIX Framework — Expanding on Chapters 4 and 6, this chapter assesses how enriched labels influence prediction quality across multiple applications. HELIX is evaluated in tasks such as forest structure regression, bark beetle predictive modelling, and seasonal glacier zone mapping.

Chapter 8: Conclusion and Outlook — The final chapter synthesizes key findings across the experimental setups, revisits the methodological implications of temporal-aware feature engineering and label enrichment, and outlines directions for future work in operationalization, generalization, and transfer learning in EO–ML pipelines.

This structure enables a layered exploration, from foundational data representations to enriched label supervision, offering a modular contribution to the development of scalable, robust EO–ML workflows for environmental monitoring.



Figure 1.9.: Thesis Structure and Thematic Flow: The thesis is grounded in the systemic integration of EO features, reference labels, and ML methodologies. Chapter I establishes this foundation. Chapters II and III address the two core data pillars: multi-sensor EO features and robust labelling strategies. Chapter IV introduces the HELIX framework for dynamic label enrichment, while Chapter V explores temporal fusion and feature behaviour. Chapter VI systematically benchmarks EO modality–model interactions for continuous label prediction, while Chapter VII extends this analysis by evaluating the impact of context-aware label enrichment and multi-scale learning using the HELIX. The thesis concludes in Chapter VIII with a synthesis of findings across all experimental dimensions.

This thesis is guided by the following hypotheses, each addressing a fundamental aspect of how label data quality, syntactic feature design, and ML models interact in the context of large-scale EO:

The selection of ML algorithms critically determines the operational viability and transferability of fused EO datasets. RF, due to their interpretability and resilience, are hypothesized to outperform more complex or specialized alternatives

(e.g., SVRs, 1D-CNN) across spatially heterogeneous environments and varying label complexities.

Syntactic fusion configurations, across spectral, polarimetric, and temporal dimensions, profoundly influence model performance and feature interpretability. Integrating complementary EO observables through hypercomplex algebra is expected to yield models that are simultaneously more accurate, more stable, and more physically interpretable than models based on unimodal inputs.

Dynamic reference data preprocessing via structured frameworks, such as the *Helix*, constitutes a breakthrough for scaling EO-based predictive modelling. Proper temporal synchronization, structural consistency, and variance preservation of labels are hypothesized to substantially improve model accuracy, generalization, and robustness, particularly for dynamic phenomena like forest disturbance monitoring and glacier evolution tracking.

Temporal alignment of reference labels with EO acquisition dates significantly enhances the reliability of models for monitoring seasonally dynamic processes. Especially in tasks sensitive to vegetation phenology or cryospheric changes, temporally synchronized labels are hypothesized to outperform static or seasonally mismatched reference datasets.

Comprehensive syntactic fusion and dynamic label preprocessing together enable true spatial transferability in ML-driven environmental monitoring. When spectral, polarimetric, and temporal fusion are combined with properly managed dynamic labels, predictive models are expected to generalize effectively across distinct geographical domains, addressing a long-standing limitation of current EO pipelines.

The central contributions of this thesis advance the frontier of machine learning in EO by jointly tackling the twin challenges of high-dimensional feature representation and complex, dynamic label design, two pillars often treated in isolation. Through the systematic integration of multi-sensor, multi-temporal EO data, this work demonstrates how information-rich, yet operationally tractable representations can be constructed for large-scale environmental monitoring.

In parallel, it introduces the HELIX framework: a novel approach to label enrichment that transforms sparse, temporally misaligned, or weakly annotated supervision into robust, context-aware learning targets. By aligning spatial and temporal structures between EO

features and reference labels, HELIX enables conventional ML architectures to operate effectively in data-scarce and uncertainty-prone settings, eliminating the need for overly complex models such as LSTMs or transformers in many applied scenarios.

These contributions are validated across diverse applications, from forest structure modelling and bark beetle outbreak detection to karst feature mapping and seasonal glacier zone classification, demonstrating not only methodological rigour but also broad domain relevance. The thesis further delivers a reproducible, modular benchmarking ecosystem, the Wald5Dplus dataset, providing community-ready data, code, and experimental protocols to support future research [148]. Taken together, this work defines a scalable and superior paradigm for EO–ML integration: one that unlocks the full informational richness of EO while remaining grounded in practical, generalizable ML workflows.

2 Consistent EO Feature Generation

> *Out of clutter, find simplicity. From discord, find harmony.*
>
> — **Albert Einstein**
> Physicist, Nobel Laureate

This chapter includes elements from the following peer-reviewed publication:

Sarah Hauser, Michael Ruhhammer, Andreas Schmitt, and Peter Krzystek. *An Open Benchmark Dataset for Forest Characterization from Sentinel-1 and -2 Time Series*. *Remote Sensing*, 16(3), 2024, Article 488. DOI:10.3390/rs16030488

It is cited as [147] and is marked with a green line.

Author Contribution: Sarah Hauser served as a primary contributor to study design, software implementation, practical execution, validation, writing, editing, and visualization.

Over the past few decades, the landscape of EO has been transformed by an explosion in sensor diversity and deployment platforms. Today's sensors vary immensely: they can capture anything from a few square meters to nearly half the planet in a matter of seconds. Spatial resolutions span from kilometres to sub-centimetre precision, while spectral coverage extends from ultraviolet (around 200 nm) to long-wavelength radar bands just below one meter (e.g., P-band), all within usable atmospheric windows. Bandwidths range from broad thermal channels, like those in Landsat, to ultra-narrow hyperspectral bands spanning just a few nanometres. Meanwhile, revisit intervals, from satellite platforms alone, can vary from once a month to daily acquisitions. In the radar

domain, the variety is just as striking, encompassing everything from single-polarization to fully quad-polarimetric systems, as well as both mono- and bistatic interferometric configurations now available in operational settings. Given this vast ecosystem of sensors and configurations, one might assume that selecting a single, optimal sensor would suffice for any given application. However, this assumption overlooks a critical reality: maximizing performance in one dimension often requires trade-offs in others.

Consider the TSX satellite. While it is capable of delivering sub-meter resolution (50 cm) imagery, this mode restricts coverage to a mere 5×5 km area and introduces scheduling limitations that preclude continuous acquisitions. In contrast, when configured for wider area monitoring at a 100 km swath, the spatial resolution drops significantly to 16 m [93]. Such trade-offs are inherent to sensor design, no single system can simultaneously deliver peak performance across all spatial, temporal, spectral, and radiometric dimensions. This constraint gives rise to the imperative for data fusion: the strategic integration of complementary observations from multiple sensors. By combining their respective strengths, fusion enables the construction of richer, more informative EO products than any single source can provide.

Image fusion, as originally defined by [333], refers to the process of merging two or more distinct images into a single, enhanced representation through algorithmic techniques. Subsequent refinements of this definition and methodological developments can be found in the foundational works of [339] and [290].

2.1 Data Fusion Approaches

Given the growing diversity of fusion techniques in RS, maintaining a clear conceptual overview has become increasingly challenging. To manage this complexity, fusion methods are typically organized based on their operational principles. However, the classification frameworks used across the literature often differ significantly, depending on the authors' perspectives and application domains.

One of the most widely adopted classification schemes categorizes fusion strategies by their stage within the processing pipeline. In this framework, methods are grouped into three core levels: pixel-level (operating directly on raw image data), feature-level (focusing on extracted descriptors), and decision-level (fusing outcomes from independent

classifiers or detectors). The decision-level fusion usually involves objects or entities already delineated within the data.

This three-tiered structure is supported by numerous studies, including those by [125, 259, 2, 123, 190, 214, 330, 290, 367]. Some reviews streamline this framework by merging the feature and decision levels into a single category termed "high-level fusion" [361]. Alternative naming conventions also exist. For instance, the levels are sometimes referred to as low, mid, and high fusion [94, 262], or described in conceptual terms as iconic, symbolic, and knowledge-based fusion [92].

In certain cases, authors introduce an additional signal-level fusion stage that precedes the pixel level [133, 175]. This involves combining raw sensor signals, prior to rasterization, to enhance the signal-to-noise ratio or to derive an improved signal representation. An overview of these various classifications and their corresponding terminology is provided in Table 2.1.

Table 2.1.: Classification of image data fusion approaches according to various publications. The most common and consistently described classification in the literature is by pixel, feature, and decision level.This table is adapted from [360].

Level 1	Level 2	Level 3	Publication(s)
Pixel	Feature	Object	[32]
Pixel	Feature	Decision	[125, 259, 2, 123, 190, 214, 330, 290, 367]
low	middle	high	[94, 262]
iconic	symbolic	knowledge	[92]
signal	iconic	symbolic, knowledge	[133, 175]
raw	low	medium, high	Abstraction level or processing level

Given the wide range of pixel-level fusion techniques, researchers have proposed several classification schemes to bring structure to this field. One widely cited system groups methods into two primary categories: colour-based approaches and statistical or numerical techniques, along with their possible hybrids [259]. Another common taxonomy divides methods into Component Substitution, Multi-Resolution Analysis, and Model- or Modulation-based categories [119, 178]. A fourth group, hybrid approaches, is also often acknowledged [361, 125]. Some authors additionally introduce classes such as "Relative Spectral Contribution" [330], or separate "Bayesian" and "Variational" methods [225].

An alternative organizational scheme is based on the operational domain: image versus frequency domain [214, 94, 133, 190]. While image domain techniques fuse data by directly combining pixel values, frequency domain methods (including wavelet-based fusion) operate on transformed data, merging local frequency components before converting back to the image domain [133]. Beyond these, several authors have proposed domain-specific taxonomies, including DL-based [289], statistical [26], and optimization-based approaches [135]. Building on these foundations, this introduces a new three-dimensional classification scheme [360] that includes the application context as a central axis, recognizing that the utility of a fusion strategy is strongly linked to its intended use case. This builds on earlier frameworks by [125, 133], shifting the focus from improving isolated data products to enhancing overall information extraction. The new classification is structured along three axes [360]:

- **Abstraction Level:** This includes four levels, pixel, feature, decision, and hybrid. The signal level is excluded, as basic registration already alters raw data. Following [138], pixel-level fusion treats each pixel independently, whereas feature-level approaches account for spatial or spectral context (e.g., neighbourhoods or multi-band features).
- **Application Domain:** This axis specifies the target of fusion:
  - Spatial–temporal fusion
  - Spatial–spectral fusion
  - Spatial–spectral–temporal fusion
- **Methodological Complexity:** This reflects increasing levels of preprocessing, algorithmic sophistication, automation potential, and computational demand.

Pixel-level fusion is the simplest approach, involving direct pixel-wise operations. It is highly flexible, fast, and suitable for automated workflows. Such methods enable on-demand fusion, for instance, merging raw hyperspectral and panchromatic imagery only when needed, thus reducing storage requirements.

Feature-level fusion enables integration of heterogeneous sources and modalities but demands more complex preprocessing and incurs higher computational cost. Flexibility and automation are somewhat reduced compared to pixel-level methods.

Decision-level fusion is applied after independent analysis of each input. It combines outcomes such as classifications or change maps. As this level deals with interpreted outputs, it is largely independent of the original data's spatial or spectral resolution.

The following sections discuss selected methods from each category in greater detail.

2.1.1 Pixel-Level Fusion

Pixel-level fusion operates at an early stage in the data processing chain, combining images that have undergone only minimal corrections, typically geometric and radiometric adjustments, to ensure co-registration and physical consistency [290, 361, 259]. This level of fusion often involves direct pixel-wise operations without intermediate transformations, though exceptions exist, e.g., hypercomplex bases [289] or component substitution, which introduce auxiliary representations but are still classified as pixel-level methods.

Below, key categories of pixel-level fusion approaches are outlined, followed by a representative method in each.

Arithmetic Combinations: This class uses straightforward mathematical operations, such as addition, subtraction, averaging, or ratios, to merge images [259]. A well-known method is the Brovey transformation, which enhances RGB images with high-resolution panchromatic input. Each RGB band is multiplied by the panchromatic image and then normalized by the sum of the multispectral bands [135].

Component Substitution: These techniques replace a structural component in a transformed image, typically intensity, with a higher-resolution version from another source. The method involves:

1. Transforming the multispectral image to a new domain (e.g., IHS or PCA) [361],
2. Replacing the intensity or spatial component with high-resolution input (e.g., PAN),
3. Reconstructing the image via inverse transformation [122, 125].

This approach is especially common in PAN-sharpening. In IHS fusion, for instance, spatial detail is injected into the intensity channel, and the result is converted back to the RGB domain [26, 133, 110].

Recent DL Advances: DL has significantly expanded the potential of pixel-level fusion. CNN are now widely used in pan-sharpening and super-resolution tasks [283]. For instance, S2Sharp [305] fuses Sentinel and Landsat imagery at 10 m resolution. GAN-based methods generate synthetic high-resolution imagery from coarser sources [62], while deep blind fusion networks learn fusion mappings without explicit transformations [346]. Hybrid strategies like Hybrid Color Mapping (HCM) combine statistical and DL techniques to improve fusion quality [197]. CNN-based spatio-temporal fusion has also been used for applications such as cloud removal [296].

These developments demonstrate a shift toward adaptive, data-driven fusion strategies that enhance spectral-spatial resolution and robustness across scenes and sensor types.

2.1.2 Feature-Level Fusion

While individual pixel values offer limited insight, their interpretation significantly improves when contextualized spatially. Feature-level fusion integrates multi-source and multi-temporal information with spatial features to enhance classification performance. Typically, relevant features are first extracted, such as geometric, spectral, or textural descriptors, before fusion is performed. Since these features already carry semantic meaning, returning to the original image domain is often unnecessary [138, 330, 259].

Commonly used features include edges, textures, spectral indices (e.g., NDVI [325]), and geometric attributes derived via segmentation techniques. One notable approach is Multiresolution Segmentation (MRS) [24], implemented in eCognition, which merges pixels into hierarchical segments based on scale, shape, and compactness. For instance, in [37], object primitives were generated through quadtree segmentation, refined with MRS, and fused using descriptors like NDVI and elevation to classify land cover.

Feature-level fusion techniques align and combine features from heterogeneous sources based on statistical, structural, or semantic similarity [125, 32].

Multiresolution Analysis: This category fuses spatial detail into multispectral images using multiscale representations, such as pyramids or frequency-based transforms [125, 361]. After decomposition and fusion, the image is reconstructed via inverse transformation. For example, Fourier-based methods place spectral information in low-frequency bands and spatial detail in high-frequency components [124]. Curvelet transforms,

such as FDCT [266], offer an efficient way to represent directional and curved features [246, 69, 355], enabling fine-scale spatial enhancement during fusion [287].

Model-Based Fusion: Model-driven approaches utilize statistical frameworks to capture spatial dependencies and uncertainties. Markov Random Fields (MRF) have been applied for edge-preserving fusion and multi-image integration [354]. Extensions like Non-Local Means (NLM) [326] fuse spatially similar regions regardless of location. More complex implementations embed segmentation results into MRF priors using Bayesian models for spectral-spatial classification [127], or integrate multi-sensor imagery (e.g., SPOT and Landsat) using learned fusion weights and SVMs [308, 25]. Spatio-temporal fusion models, such as STARFM [120], combine high-resolution spatial data with high-frequency temporal sources. STAARCH [154] and ESTARFM [371] improve on this by modeling land cover change dynamics using spectral transformations and temporal coefficients. Later enhancements, like STNLFFM [68], introduce non-local filtering and regression-based prediction using dual-date imagery, addressing variability and noise in time-series fusion [185].

DL and Transformer-Based Methods: Recent advances in DL and attention-based architectures offer powerful tools for feature-level fusion. These models learn hierarchical representations directly from raw inputs, reducing the need for manual feature engineering.

- **Transformer Architectures:** Multimodal Transformer Cascaded Fusion integrates UAV and satellite data to capture long-range dependencies [345], while Spectral-Spatial-Elevation Fusion Transformers enhance hyperspectral classification by incorporating elevation [107].
- **CNN–Transformer Hybrids:** Dual-branch models leverage local patterns via CNN and global structure via Transformers to fuse complementary inputs [347].
- **Contrastive and Interaction Learning:** Approaches such as text-supervised contrastive fusion [357] align semantic labels with RS features, while interaction-based fusion architectures model multi-source dependencies using attention and residual blocks [130].

These data-driven strategies outperform traditional stacking or rule-based fusion methods, offering improved classification accuracy, better generalization across contexts, and higher adaptability to complex remote sensing tasks.

2.1.3 Decision-Level Fusion

Decision-level, or interpretation-level, fusion operates at the highest level of the data processing hierarchy. Each image source is first processed independently to extract relevant information, which is then combined using decision rules [259]. This approach is particularly effective when dealing with heterogeneous data sources that differ in modality, resolution, or acquisition timing.

A recent example is the automated building footprint detection processor by [293], which processes Sentinel-1 and Sentinel-2 separately, accounting for their differing temporal availability due to cloud cover, and reconciles outputs via logical operations. Such rules can be extended with fuzzy logic to incorporate uncertainty. Fuzzy decision fusion uses graded membership values (ranging from 0 to 1) and weighting schemes that prioritize reliable sources, allowing for context-aware, probabilistic classification [105, 125].

Multicriteria Decision Analysis (MCDA) frameworks, such as the Analytical Hierarchy Process (AHP) [281], are also applied at the decision level. In [258, 257], AHP was used to identify suitable reintroduction habitats for the European oyster in the German Bight EEZ by weighting multiple environmental criteria derived from remote sensing data.

Decision-level fusion is also effective for applications beyond classification. In [286], polarimetric and structural metrics from multiple viewing angles were combined to detect informal settlements, using histogram-based probability estimation and similarity-based probability fusion. In another study [350], glacial lake extents on the Baltoro Glacier were estimated using optical (Sentinel-2, PlanetScope) and SAR data (Sentinel-1, TerraSAR-X), with sensor-specific biases corrected through decision-level adjustment.

A similar framework was used to monitor Lake Tabalak's water dynamics using multi-temporal SAR datasets from six different sensors [38]. Processed through the Multi-SAR system [288, 164], all inputs were standardized (e.g., via orthorectification, radiometric calibration, Kennaugh decomposition) before generating consistent binary water masks for each sensor-date pair. This highlights decision-level fusion's utility in producing harmonized outputs from disparate and temporally misaligned data sources.

Recent literature highlights the integration of AI models into decision-level fusion pipelines, notably through ensemble and hybrid learning strategies:

- **Classifier Ensemble Methods:** Fuzzy multiple classifier systems using decision templates have been proposed to integrate hyperspectral and LiDAR classifications [40]. Ensemble strategies such as the Rotation Forest have also been applied to multi-sensor data for species classification tasks [56].
- **Graph- and Rule-Based Fusion:** Graph-based approaches have been developed to integrate morphological and structural features from diverse sources, enabling more informed higher-level classification decisions [218].
- **Decision-Based Filtering and Edge Preservation:** Edge-preserving filtering combined with guidance maps has been introduced to improve decision-level fusion outcomes, especially in pansharpened image analysis, helping to preserve both spatial and spectral fidelity [282].
- **Reinforcement Learning in Fusion:** Reinforcement learning strategies have been applied to optimize decision fusion pathways, using dynamic environment–state interactions to outperform traditional fusion approaches in hyperspectral and LiDAR scenarios [343].
- **Neural Decision Layers:** Recent encoder-decoder network architectures for segmentation integrate decision-level fusion directly into their decoding layers, particularly effective for combining very high resolution (VHR) imagery with point cloud data [132].

These innovations demonstrate how rule-based and probabilistic methods are now being augmented by learning-based inference and decision pipelines, marking a clear shift toward hybrid, data-driven decision fusion. Such architectures are particularly advantageous in multi-sensor EO scenarios, where direct data- or feature-level fusion is hindered by disparities in spatial resolution, acquisition time, or sensor-specific noise characteristics. For instance, in land cover classification, separate models trained on Sentinel-1 and Sentinel-2 data may yield different class probability maps. A decision-level ensemble then consolidates these predictions, using learned confidence scores or probabilistic fusion, into a robust final classification. Moreover, recent workflows increasingly integrate meta-learning frameworks that adaptively combine model outputs depending on local data quality or contextual uncertainty. These decision-level fusion strategies are particularly valuable in time-critical applications such as flood mapping, rapid damage assessment, or disaster monitoring, where asynchronous data availability or partial occlusions (e.g., due to cloud cover) require flexible and sensor-agnostic fusion solutions.

2.1.4 Temporal-Aware Fusion

Temporal taxonomy refers to the systematic classification of remote sensing data according to their acquisition time-frames and seasonal contexts. Rather than treating time as a passive metadata attribute, this approach acknowledges that remote sensing signals encode seasonally dependent variations, not only due to actual surface change, but also due to temporal shifts in vegetation phenology, moisture regimes, snow cover, or atmospheric conditions, all of which interact with sensor-specific spectral and structural sensitivities.

Temporal-aware fusion strategies thus seek to actively harness this temporal diversity by combining observations across multiple time points in structured ways. These strategies can span several dimensions: intra-seasonal (within the same season), inter-seasonal (across seasons), and cross-temporal (across years or phenological stages). Each configuration enables different insights: from short-term change detection to the stabilization of features across variable conditions or the extraction of temporally persistent patterns.

The following section systematically explores these temporal fusion strategies in RS. They investigate how time-structured feature stacks contribute to enhanced predictive modelling, whether by improving classification robustness, supporting temporal generalization, or enabling novel indices that combine multi-temporal and multi-modal observations.

This concept plays a foundational role across all levels of data fusion, pixel, feature, and decision, by providing a structured way to manage and exploit temporal diversity in multi-sensor integration. Particularly in hypercomplex fusion [289] approaches, where data from different modalities (e.g., SAR and optical) are combined into a unified, multi-dimensional feature space, temporal taxonomy ensures that fusion strategies are guided not by acquisition simultaneity alone, but by ecological relevance and information complementarity. A key insight is that datasets acquired at different times may offer unique and non-redundant perspectives on the landscape, especially when drawn from contrasting phenological or hydrological phases. For instance, SAR data captured during the dry season can provide stable structural information free of moisture-induced noise, while optical data from the wet season can highlight vegetation dynamics with high spectral separability. Temporal taxonomy formalizes this logic by distinguishing between types of fusion based on timing and seasonality:

- **Single-Date Fusion**: This approach involves the fusion of sensor data acquired on the same or closely aligned dates, often within a few days, to ensure temporal coherence between modalities. It is a standard practice in EO due to its ability to preserve surface consistency across sensors, minimizing noise from phenological changes, atmospheric variability, or land use activity. Particularly suitable for monitoring rapid, onset phenomena, such as flooding, fire, or deforestation, this fusion strategy ensures high geometric and spectral alignment. However, the strength of temporal coherence can become a limitation in applications that benefit from seasonal or phenological contrast, such as ecosystem mapping, crop stage detection, or landform-vegetation interaction studies. In such cases, single-date fusion may offer redundant or temporally shallow information. In forestry, single-date SAR-optical fusion has shown clear benefits under limited data conditions, when only a single Sentinel-2 scene is available, adding Sentinel-1 features boosted tree species classification accuracy by 4.7 percentage points [205]. Single-date fusion is computationally simple and avoids temporal mismatches; however, it may miss phenomena that manifest over time (phenology, gradual changes, etc.). Thus, while single-date fusion is useful for quick assessments or when data is scarce, it often serves as a building block for more advanced multi-temporal fusion approaches.

- **Multi-Date Fusion**: Multi-date fusion combines acquisitions from different dates within a relatively stable period, commonly the same season, to increase temporal depth and improve resilience against noise or data gaps. It is particularly effective in cloud-prone regions or in scenarios requiring data compositing, such as vegetation monitoring, precision agriculture, or urban expansion analysis. However, when environmental conditions vary significantly between acquisition times (e.g., rainfall, irrigation events, phenological shifts), multi-date fusion can suffer from temporal decorrelation, leading to inconsistencies in the fused product. Thus, temporal normalization or phenological alignment is often needed to preserve class separability and thematic accuracy. A notable example is the Pixel R-CNN (Recurrent CNN) model, which learns spectral–temporal features from a Sentinel-2 time series for land cover and crop type mapping [235]. This pixel-based R-CNN achieved an overall classification accuracy of 96.5 %, significantly outperforming non-temporal models on the same task [235]. The high accuracy illustrates how multi-date fusion "adds information", e.g., capturing crop growth stages or tree phenology, that a single-date approach may not capture. Even simpler multi-date strategies, like using a few well-chosen dates, yield improvements: in one study,

using all Sentinel-2 images across a growing season raised tree species mapping accuracy to 83% (vs. lower accuracy on any single date) [205]. Overall, fusing data across multiple dates increases robustness to outliers (like a cloudy image) and enables change detection and trend analysis. The trade-off is higher data volume and complexity, calling for methods to handle irregular time steps, cloud gaps, and sensor differences. Nonetheless, multi-date fusion has become standard in vegetation and forest assessments, land-cover mapping, and hazard monitoring due to the clear gains in accuracy and insight.

- **Seasonally Disparate Fusion**: This fusion type explicitly combines data from distinct seasonal periods, such as leaf-off vs. leaf-on in temperate forests, dry vs. wet periods in savannas, or snow-covered vs. snow-free surfaces in alpine or boreal regions. The objective is to leverage ecological or environmental divergence across time to enrich the feature space. For example, SAR acquired under low-vegetation conditions may provide optimal structural information, while optical data from peak vegetative phases delivers strong spectral signals. This strategy has demonstrated value in land cover classification, change detection, geomorphological mapping, and wetland delineation, where surface dynamics are central to interpretation. Despite a drop in strict temporal coherence, the increase in semantic contrast between classes can enhance the discriminatory power of fused features. In forestry, for example, leaf-off (e.g., winter) imagery can expose ground or deciduous/conifer differences that are masked in summer imagery. Conversely, leaf-on summer imagery highlights active vegetation. Fusing the two can improve tree species or forest type discrimination. A multi-season approach was demonstrated by selecting the most informative Sentinel-2 scenes from spring, summer, and autumn and combining them (with SAR) to classify diverse tree species in Austria [205].

- **Annual Aggregation**: Annual aggregation encompasses the integration of time series data across a full annual cycle. Instead of capturing a moment in time, it aims to represent phenological, hydrological, or structural trends over the year, whether through statistical descriptors (e.g., max-NDVI, mean backscatter) or higher-order transformations (e.g., harmonic analysis, hypercomplex accumulation). This approach supports applications such as land use monitoring, ecosystem trend analysis, or climate resilience assessment, offering temporally smoothed or seasonally normalized insights. While aggregation may obscure short-term dynamics, it enables robust, scalable analyses at regional to global levels, particularly where temporal density is prioritized over instantaneous fidelity. In glaciology, method to map

alpine glaciers by fusing all Sentinel-1 and -2 data around the yearly peak ablation period (late summer when glaciers are most exposed) was introduces [29]. They compiled a Sentinel-2 mosaic of that period's snow/ice conditions and a Sentinel-1 multi-temporal coherence composite (indicating moving ice), then combined them to delineate glacier outlines. By aggregating data annually at the most relevant season, they achieved 92% accuracy for glacier mapping, including debris-covered ice [29].

Most operational fusion pipelines prioritize temporal proximity, aligning data as closely as possible in time, under the assumption that this reduces discrepancies. While this assumption holds for applications like change detection or dynamic monitoring, it can be limiting in tasks where structural and spectral signals evolve independently and asynchronously. In such cases, ecological complementarity may prove more informative than temporal coherence. This is particularly relevant in semi-arid or ecologically dynamic regions, where vegetation cycles and surface moisture vary significantly over the year. As shown in later sections, the most effective fusion configurations in this study did not align temporally, but instead spanned seasonal boundaries, leading to stronger class separability and improved geomorphological discrimination. Temporal taxonomy thus serves not only as an organizational framework, but as a design principle for intelligent data fusion. It encourages RS practitioners to think critically about the ecological meaning behind their acquisition windows, and to move beyond simplistic assumptions of temporal simultaneity. This is especially pertinent when fusing modalities with distinct sensing characteristics, such as radar and optical imagery.

2.2 Hypercomplex Bases

While pixel-, feature-, and decision-level fusion methods each offer specific advantages depending on the application context, they also come with inherent limitations. Pixel-level fusion often struggles with sensor-specific artifacts or noise, especially when spatial resolutions or acquisition times differ. Feature-level fusion relies on prior extraction and alignment of interpretable attributes, which can introduce bias or result in loss of detail. Decision-level fusion, on the other hand, operates at the highest level of abstraction, but may fail to leverage complementary information across sources at earlier stages of processing.

To address these limitations, fusion on hypercomplex bases (HCB) [289] presents a unified framework that can operate seamlessly across all fusion levels. On the following pages, its application is primarily demonstrated at the pixel level, as this is the form most directly used and evaluated within this thesis. The methods of hypercomplex bases also use only basic arithmetic operations, but within a fixed mathematical structure: the transformations are always orthogonal, and the resulting elements can be output both linearly and logarithmically through corresponding normalization, or as indices normalized to a fixed value range.

The approach builds upon the concept of Kennaugh elements, which have already been successfully used in the processing of SAR images [288] and in SAR sharpening [286], and later on explained as hyper-complex bases (HCB) in detail [289]. This method extends the Kennaugh framework for SAR images [288] and SAR-optical fusion in SAR sharpening [286], offering stable sums and sensitive differences. The transformation matrix applied for optical data is also known as the Hadamard transform. It proves to be a discrete implementation of the Fourier transform that decomposes a signal into oscillations of varying wavelength [289].

Kennaugh elements are derived from the coherency or covariance matrix of polarimetric Synthetic Aperture Radar (SAR) data and provide a systematic approach to describing the scattering properties of observed targets [288]. While originally developed for polarimetric SAR applications, the Kennaugh framework has since been extended to accommodate multi-spectral optical data as well, thereby enabling a unified representation for the fusion of SAR and optical datasets [289]. In this context, Kennaugh elements are essential in hypercomplex data fusion, as they provide the parameters required to integrate spectral, polarimetric, and temporal information. Their application to both polarimetric and spectral domains has led to the development of so-called spectral Kennaugh-like elements, which preserve the structural coherence of the original framework while enabling modality-agnostic feature representation. Hypercomplex data fusion uses hypercomplex numbers, such as quaternions, to represent and integrate multi-dimensional data from different sensors like Sentinel-1 and Sentinel-2.

The core principle of HCB is to express input channels through their shared and divergent components. This concept is mathematically formalized using the transformation matrix shown in Equation (2.1):

$$C = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \tag{2.1}$$

The normalization factor $\sqrt{\frac{1}{2}}$ ensures orthogonality, such that the matrix is symmetric and its transpose equals its inverse. This property allows for a reversible transformation that preserves the structure of the original spectral space without distortion. When applied to a vector, the first row of the matrix computes the sum, and the second the difference of the two input channels, regardless of whether these represent radar intensities or spectral reflectances. Rooted in complex number theory, this representation can be extended to higher dimensions. For example, using a quaternion basis yields the four-dimensional matrix shown in Equation (2.2):

$$Q = \begin{bmatrix} C & C \\ C & -C \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \tag{2.2}$$

This framework can be recursively expanded to define even higher-dimensional spaces such as octonions (eight dimensions) and sedenions (sixteen dimensions), enabling progressively richer representations of multi-channel data [289].

Polarimetric Kennaugh Elements for SAR Data: To extract meaningful physical properties from polarimetric SAR data, the scattering matrix is often transformed into Kennaugh elements, a reduced and interpretable representation of the radar backscatter. These elements are derived from the real-valued components of the coherency matrix and are commonly used to characterize different scattering mechanisms and surface structures [288].

Depending on the polarization configuration (HH/VH or VV/VH), the Kennaugh elements k_0, k_1, k_5, and k_8 are computed as follows [288], see Equations (2.3) and (2.4). The VV/VH configuration is predominantly used in temperate forest applications (cf. Section 1.2.2), while the HH/HV configuration is relevant in cryospheric settings, as applied in the High Canadian Arctic glacier zone mapping (cf. Section 1.2.3).

For HH/VH polarization:

$$\begin{aligned} k_0 &= |S_{HH}|^2 + |S_{VH}|^2 \\ k_1 &= |S_{HH}|^2 - |S_{VH}|^2 \\ k_5 &= \mathrm{Re}\{S_{HH}S_{VH}^*\} \\ k_8 &= \mathrm{Im}\{S_{HH}S_{VH}^*\} \end{aligned} \tag{2.3}$$

For VV/HV polarization:

$$\begin{aligned} k_0 &= |S_{VV}|^2 + |S_{HV}|^2 \\ k_1 &= |S_{VV}|^2 - |S_{HV}|^2 \\ k_5 &= \mathrm{Re}\{S_{HV}S_{VV}^*\} \\ k_8 &= -\mathrm{Im}\{S_{HV}S_{VV}^*\} \end{aligned} \tag{2.4}$$

where S_{pq} represents the complex backscatter coefficient for transmit polarization p and receive polarization q, with $p, q \in \{H, V\}$. Each Kennaugh element describes a distinct physical characteristic [288]:

- k_0: Total backscatter power, indicating the overall signal intensity.
- k_1: Power difference between co- and cross-polarizations, used for distinguishing surface types.
- k_5: Real part of the complex correlation term, reflecting volume scattering and dielectric variations.
- k_8: Imaginary part (or negative thereof), capturing asymmetries and orientation-related features.

This Kennaugh representation facilitates the integration of SAR data with other sensor modalities and serves as the foundation for the hypercomplex data fusion methods discussed later in this thesis.

Spectral Kennaugh-like Elements: When applied to four-band aerial imagery or the 10-meter spectral bands of Sentinel-2, the visible and near-infrared channels, Blue, Green, Red, and NIR, can be transformed into Kennaugh-like elements using HCB. Typically, these spectral bands are visualized as True Colour Images (TCI; R: Red, G: Green, B: Blue) or Colour Infrared (CIR; R: NIR, G: Red, B: Green). The transformation disentangles brightness from chromatic information, as described in Equation (2.5):

$$K_{\text{spectral}} = Q \cdot \begin{bmatrix} \text{Blue} \\ \text{Green} \\ \text{Red} \\ \text{NIR} \end{bmatrix} \tag{2.5}$$

Based on this transformation, optical images, purely with respect to their numeric intensity values, become structurally compatible with multi-polarized radar imagery (e.g., Sentinel-1), which is often stored in Kennaugh format via processors such as the German Aerospace Center's Multi-SAR system [38, 288]. This compatibility facilitates direct and lossless fusion of radar and optical modalities [289].

Hypercomplex Data Fusion: The joint fusion of spectral and polarimetric information is performed as shown in Equation (2.6):

$$K_{\text{fused}} = \begin{bmatrix} K_{\text{spectral}} + K_{\text{polarimetric}} \\ K_{\text{spectral}} - K_{\text{polarimetric}} \end{bmatrix} \triangleq C \cdot \begin{bmatrix} \text{spectral} \\ \text{polarimetric} \end{bmatrix} \tag{2.6}$$

This fusion yields a total of eight hypercomplex channels, four from each modality, preserving the original signal content due to the orthogonal properties of the transformation. By construction, this approach ensures lossless integration of spectral and structural information across sensors.

In this context, the spectral input K_{spectral} is first derived from the four Sentinel-2 reflectance bands by applying the quaternion-based Hadamard matrix Q (see Section 2.2), resulting in four spectral Kennaugh-like elements. These are then fused with the four polarimetric Kennaugh elements from Sentinel-1 ($K_{\text{polarimetric}}$) using the two-dimensional Hadamard matrix C, as defined in Equation (2.1). This two-step process, first the spectral transformation with Q, followed by modality fusion with C, ensures orthogonality and preserves the full information content across both sensor types.

The fused features are constructed as in Equation (2.7):

$$F = \begin{bmatrix} S_1 + Q \cdot S_2 \\ S_1 - Q \cdot S_2 \end{bmatrix} \quad \text{where}$$

$$S_1 = \begin{bmatrix} k_0 \\ k_1 \\ k_5 \\ k_8 \end{bmatrix}, \quad S_2 = \begin{bmatrix} B_2 \\ B_3 \\ B_4 \\ B_8 \end{bmatrix}, \quad Q = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \tag{2.7}$$

Each resulting fused element in F represents a specific spectral–structural combination of optical and SAR information. The Hadamard matrix Q ensures orthogonality, allowing a lossless and interpretable transformation of the input features, resulting in a fused dataset consisting of one total intensity element ($K_{\text{fused},0}$) and seven spectral/polarimetric elements ($K_{\text{fused},1-7}$), as detailed in [289] and illustrated in Figure 2.1. The fused Kennaugh representation bridges spectral, structural, and geometric features into a stable 8-dimensional feature space. It leverages the orthogonal nature of hypercomplex bases, preserving information content while offering interpretable, compact descriptors for EO applications such as classification, anomaly detection, and structural mapping. Each channel provides semantically distinct information, enhancing robustness and reducing redundancy in downstream ML pipelines.

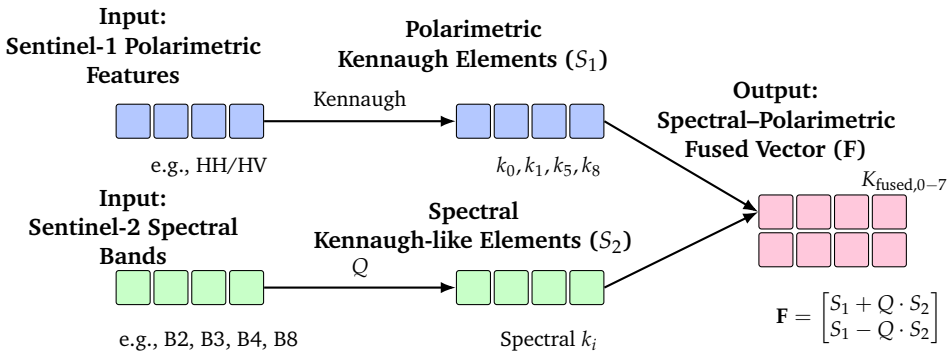


Figure 2.1.: Stepwise example of hypercomplex spectral–polarimetric fusion. Sentinel-1 dual-pol inputs are first transformed into four polarimetric Kennaugh elements (S_1), while Sentinel-2 reflectances (shown here as example bands B2–B8) are transformed into four spectral Kennaugh-like elements (S_2) using the quaternion-based Hadamard matrix Q. Note that the optical input is not limited to these four bands; additional or alternative spectral channels can also be used following the same principle. Finally, both feature vectors (S_1, S_2) are orthogonally fused into the 8-dimensional spectral–polarimetric feature vector $\mathbf{F}$, as described in Equation (2.7).

Temporal Fusion in Hypercomplex Data Fusion: Temporal fusion extends the hypercomplex framework beyond spectral and structural integration by incorporating multi-temporal observations into a unified orthogonal representation [289]. Starting from the fused spectral–polarimetric feature vector $\mathbf{F}$, obtained through a quaternion-based (i.e., 4×4 Hadamard) transformation that combined Sentinel-1 and Sentinel-2 inputs (cf. Figure 2.1), the temporal fusion applies an orthogonal Hadamard matrix $Q_T \in \mathbb{R}^{T\times T}$ across all T temporal acquisitions, as shown in Equation (2.8) and illustrated in Figure 2.2. This process enables the extraction of both persistent patterns and dynamic variations across time, all within a compact, information-preserving hypercomplex basis.

Given a time series of fused spectral–polarimetric feature vectors $F_t \in \mathbb{R}^8$, one for each of the T temporal acquisitions (e.g., 64 Sentinel-1/-2 observations per year), the temporal fusion is performed by applying an orthogonal Hadamard matrix $Q_T \in \mathbb{R}^{T\times T}$ to the stacked time series:

$$F_{\text{temporal}} = Q_T \cdot \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_T \end{bmatrix} \tag{2.8}$$

This temporal transformation directly continues the use of the fused feature vector F, introduced in Section 2.2, now extended along the temporal axis. Conceptually, Q_T represents the temporal analogue to the previously described hypercomplex bases [289]: depending on the number of temporal acquisitions T, the transformation follows the same family of orthogonal Hadamard-type bases, namely:

- C for $T = 2$ temporal steps (complex basis),
- Q for $T = 4$ steps (quaternion basis),
- O for $T = 8$ steps (octonion basis),
- S for $T = 16$ steps (sedenion basis),
- and higher-order Hadamard bases for larger T, such as $T = 32$, 64, or 128, as required for dense time series analysis.

The matrix Q_T ensures orthogonality and allows a lossless, reversible, and interpretable decomposition across time, analogous to the spectral–polarimetric fusion previously performed using Q in Equation (2.7).

Within the resulting temporal feature space:

- The first component F_0 corresponds to the temporal mean, representing average surface properties across all acquisitions,
- The remaining components $F_1, F_2, \ldots, F_{T-1}$ represent orthogonal modes of temporal variation, capturing dynamic processes such as phenological, hydrological, or structural changes throughout the observation period.

This temporal transform thus extends the hypercomplex data fusion framework into the time dimension, while preserving the full information content of the original time series for subsequent tasks like classification, trend analysis, or anomaly detection.

In the specific case of the Wald5Dplus dataset, this temporal fusion framework was applied to a combined Sentinel-1 and Sentinel-2 time series. Here, the spectral–polarimetric fusion step yielded 8-dimensional feature vectors ($F_t \in \mathbb{R}^8$) for each of the $T = 64$ temporal acquisitions available over one year. Applying the corresponding temporal Hadamard matrix $Q_T \in \mathbb{R}^{64\times64}$, the dataset was transformed into a temporally fused representation $F_{\text{temporal}} \in \mathbb{R}^{64\times8}$, capturing both the mean surface properties and the orthogonal modes of temporal variation across the full annual time series.

The following example illustrates this concept for the Wald5Dplus benchmark:

The joint image thus comprises one total intensity ($K_{\text{fused},0}$) and seven spectral/polarimetric elements ($K_{\text{fused},1-7}$). In the same way, the 64 acquisitions gathered during one year can be fused temporally on HCB to $K_{*,0-64}$. The big advantage is the availability of one mean image $K_{*,0}$ which is representative for the whole year (similar to the total intensity) and 63 elements $K_{*,1-63}$ describing the temporal variations throughout the year, e.g., $K_{0,0}$ stands for the mean reflectance over all channels over the whole year whereas $K_{0,*}$ also includes all its variations throughout the year. The final normalization allows for the loss-less and space-saving archiving of the image data as UInt8 digits [289], which can be displayed and processed by each image processing or GIS software.

Sentinel-1 only 256 channels composing of 64 times 4 polarimetric Kennaugh elements

Sentinel-2 only 256 channels composing of 64 times 4 spectral Kennaugh-like elements

Sentinel-1 & -2 512 channels composing of 64 times 8 fused Kennaugh-like elements

This method preserves the full information content of the original time series and enables a compact yet expressive decomposition suitable for subsequent learning tasks. Similar to spectral fusion, this temporally fused representation supports tasks such as anomaly detection, trend extraction, and spatio-temporal classification. Such hypercomplex temporal fusion supports advanced EO tasks such as monitoring phenological cycles, mapping dynamic land cover changes, and detecting long-term trends in high-dimensional satellite time series [147]. The full Python-based implementation used for data processing is provided in the appendix A.2.1 and is publicly available alongside the benchmark dataset [148].

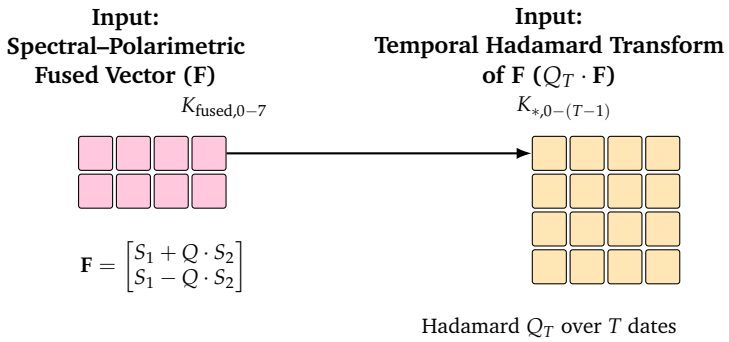


Figure 2.2.: Temporal extension of hypercomplex data fusion. Starting from the spectral–polarimetric fused feature vector $\mathbf{F}$ ($K_{\text{fused},0-7}$), a Hadamard-based temporal transform (Q_T) is applied across all T time steps. This results in a temporally enriched $T \times 8$-dimensional dataset ($K_{*,0-(T-1)}$), capturing both persistent and dynamic modes of variation. The temporal Hadamard matrix $Q_T \in \mathbb{R}^{T \times T}$ follows the same family of orthogonal hypercomplex bases (e.g., C, Q, O, S) as used in the spectral–polarimetric fusion, now extended to the temporal domain. The illustrated example shows $T = 4$; for the Wald5Dplus application, $T = 64$ acquisitions were used.

Hypercomplex Fusion on different Data Fusion Levels: Hypercomplex data fusion offers a unified mathematical framework capable of integrating multi-sensor, multi-temporal, and multi-dimensional data across the three canonical levels of fusion: pixel-level, feature-level, and decision-level. This section outlines how hypercomplex representations adapt to each level while maintaining semantic coherence and mathematical integrity.

- **Pixel-Level Fusion:** At this level, raw or preprocessed data from different sensors are directly combined on a per-pixel basis. Hypercomplex methods integrate spectral, polarimetric, and temporal information by projecting all sensor inputs into a shared hypercomplex domain. This approach retains full spatial and signal fidelity across dimensions. A prominent example of this is the *Wald5Dplus* project [147, 148, 144], which applied quaternion-based fusion to combine Sentinel-1 polarimetric backscatter with Sentinel-2 surface reflectance across 64 dates, producing an eight-element, temporally stacked data cube [147]. The result is a pixel-wise fused representation capturing yearly dynamics in both spectral and radar domains, ideal for time series classification, vegetation phenology, or change detection.

- **Feature-Level Fusion:** Here, fusion occurs after relevant features are extracted from the pixel-level dataset. Hypercomplex representations serve not just as raw signal combinations but also as higher-level descriptors from which semantically meaningful indices can be derived. These can include vegetation-sensitive combinations, geomorphological response patterns, or coherence-based feature spaces designed to isolate specific surface phenomena. This level of fusion facilitates dimensionality reduction while preserving informative content and can be tailored to tasks such as object detection or landform characterization. A typical outcome is the derivation of fusion-based indices that integrate both spectral and polarimetric responses for advanced class separation, used without relying on pixel values alone [6].

  In this framework, HCB serve a critical function in facilitating consistent integration at the feature level. Spectral descriptors frequently adopt the form of normalized difference indices [151], constrained within the interval $]-1,+1[$. This normalization offers multiple benefits, including enhanced radiometric comparability, improved visualization, and increased compatibility with machine learning algorithms, particularly those, like SVM, that perform optimally with standardized input features.

  To maintain this normalization post-fusion, integral and differential Kennaugh operators have been introduced [285]. Similar to pixel-level Kennaugh formulations, the feature-level fusion employs the following definitions, as shown in Equations (2.9) and (2.10):

$$sk = \frac{k_a + k_b}{1 + k_a \cdot k_b} \in (-1, +1) \tag{2.9}$$

$$dk = \frac{k_a - k_b}{1 - k_a \cdot k_b} \in (-1, +1) \tag{2.10}$$

These operators preserve the normalized scale, making them particularly effective for subsequent tasks such as classification or regression. Importantly, the approach is not restricted to SAR-based Kennaugh or Kennaugh-like metrics, it generalizes to any normalized feature index, offering versatile applicability across diverse fusion pipelines. Notably, this method is not limited to SAR-based Kennaugh or Kennaugh-like elements, it can be extended to any normalized feature index, offering broad flexibility for application-specific fusion workflows [278].

- **Decision-Level Fusion:** At the highest level, the outputs of multiple classifiers, algorithms, or rules are combined into a coherent decision-making system. In the hypercomplex context, this involves using previously fused pixel- and feature-level data as inputs into classification schemes that are augmented with rule-based logic or ensemble methods. For example, a classifier may exploit both the raw fused layers and precomputed indices, incorporating expert rules (e.g., thresholds for sinkhole likelihood or vegetation vigour) to refine output masks. This integrative process enhances robustness, especially in complex semi-arid landscapes where individual layers may be ambiguous. Applications include ecological monitoring, anomaly detection, and risk mapping, where fusion supports both data-driven learning and domain-specific interpretability [6]. An intriguing direction for future research is to interpret temporally fused Kennaugh (or Kennaugh-like) elements not as fixed feature values, but as empirical *probability density functions* (PDFs) over time. In this view, each fused element (e.g., K_1) at a given pixel is treated as a distribution of values sampled across multiple acquisitions. This would enable the modelling of pixel-level uncertainty and dynamic variability explicitly, rather than collapsing temporal signals into single deterministic vectors. For instance, pixels with narrow, uni-modal PDFs might indicate stable vegetation, while bi-modal or skewed PDFs could signify phenological shifts or latent disturbance signatures. These temporal PDFs could then serve as probabilistic priors in decision-level fusion, enriching classification or risk mapping tasks with statistically grounded confidence estimates. While still conceptual, this approach aligns well with the interpretability goals

of decision fusion and opens new avenues for integrating time-aware statistical reasoning into remote sensing pipelines.

Hypercomplex data fusion not only unifies multi-sensor inputs but also scales across analytical levels, from raw signal combination to high-level semantic reasoning. Its versatility allows practitioners to tailor fusion workflows according to the demands of specific tasks, data availability, and computational context.

3 Labelling Foundations and Challenges

> *Information is the resolution of uncertainty.*
>
> **— Claude E. Shannon**
> Mathematician, Father of Information Theory

This chapter includes elements from the following peer-reviewed publication:

Sarah Hauser, Lena Augner, and Andreas Schmitt. *Perfect Labelling: A Review and Outlook of Label Optimization Techniques in Dynamic EO*. *Remote Sensing*, 2025, 17, 1246. DOI:10.3390/rs17071246

It is cited as [149] and is marked with a cyan line.

Author Contribution: Sarah Hauser was instrumental for the full study design and conceptualization presented in this work, including the independent development of the HELIX framework for spatio-temporal label preprocessing. She led the investigation and contributed substantially to the manuscript's review and editing.

Having established the foundations of EO data fusion and its role in generating rich, multi-modal feature representations, we now turn to the other half of the learning equation: the reference data. Reference data, referred to throughout this thesis as *labels*, forms the conceptual and practical backbone of ML in RS. Labels act not only as the empirical ground truth against which models are trained and validated, but also as a kind of translation layer: a semantic dictionary that allows raw EO signals to be interpreted in terms of real-world phenomena such as forest biomass, vegetation stress, or land use change. In this sense, labels bridge the gap between unstructured sensor data and meaningful environmental understanding. Their quality directly influences model

reliability, high-quality, well-aligned labels enable robust generalization, while noisy, inconsistent, or outdated labels can propagate errors, bias predictions, and undermine the operational utility of a system. This is particularly critical in EO applications, where temporal and spatial complexity is high, and decisions often hinge on subtle patterns. This chapter examines the foundational role of labels in EO–ML pipelines, focusing on core dimensions such as accuracy, consistency, completeness, spatial-temporal alignment, and ecological validity. It also explores the practical and methodological challenges of generating, maintaining, and adapting reference datasets in dynamic environmental contexts.

3.1 General Differences in Label Preparation by Model Type

Label preparation is fundamental to the success of ML, DL, AI, and FM in EO applications. However, the way reference data is prepared, structured, and utilized varies significantly across these model types. These differences arise due to variations in feature extraction requirements, data volume, annotation strategies, and the need for static or dynamic labels.

Common ML models rely heavily on manually curated reference data [237] to establish relationships between predictor variables and target outputs. These models require well-structured predictor-label pairs, where reference data quality plays a crucial role in ensuring model reliability [294]. In EO applications, domain experts define features such as spectral indices like NDVI (Normalized Difference Vegetation Index), spatial texture metrics, and aggregated temporal statistics, including vegetation indices over a growing season [227]. These models assume that data points or pixels are independent unless temporal dependencies are explicitly introduced through engineered features. Since ML models require structured training data, label preparation often involves a meticulous manual annotation or expert-driven classification process [294]. A key advantage of ML models is their ability to perform well with smaller datasets when meaningful, structured features are available [294]. However, the reliance on manually crafted features and static training labels limits ML's flexibility when applied to highly complex, multi-temporal, or multi-source EO datasets. Since conventional ML models lack the ability to automatically extract hierarchical features from raw data, their performance

is heavily dependent on the quality and completeness of the reference data. Despite the pivotal role of preprocessing temporal reference data, research on these steps for traditional ML models like RF remains limited, even though they are widely used in EO [188]. Poor preprocessing can introduce inconsistencies [336], yet comprehensive guidelines for handling multi-temporal reference data in traditional ML are still lacking. While DL models often include structured preprocessing pipelines, similar frameworks for classical models like RF are underexplored. Most literature on handling time series data focuses on specialized models like LSTMs (LSTM), a Recurrent Neural Network (RNN) specialized in long-term dependency modelling, leaving a gap in how simpler, yet powerful, methods can be adapted to spatio-temporal complexities [303]. This gap underscores the necessity for further innovation in preprocessing approaches that reconcile the demands of dynamic, multidimensional EO datasets with the operational simplicity and lower computational footprint of classic algorithms. Overall, failing to capture or correctly process temporal and spatial dependencies can yield biased estimates and reduced predictive power. As data volume continues to grow, robust and efficient workflows for creating and maintaining dynamic reference datasets will become even more essential. A lack of standardized, scalable methodologies for handling time-sensitive or multidimensional EO data effectively undermines the potential of even the most sophisticated AI/ML architectures. Addressing this concern requires automating dynamic labelling, mitigating data inconsistencies introduced by multi-sensor fusion, and explicitly integrating spatio-temporal dependencies into the preprocessing pipeline. DL models, in contrast, can learn representations directly from raw EO data, eliminating the need for explicit label engineering [322]. CNN, RNN, and Temporal Convolutional Networks (TempCNN) operate on high-dimensional datasets such as multispectral time series, learning hierarchical patterns from large-scale, annotated datasets [255]. While ML approaches typically rely on structured, discrete class labels, DL models demand pixel-wise or spatially dense annotations, particularly in tasks like land cover classification and semantic segmentation. Unlike ML, which can cope with relatively small datasets if structured features are available, DL models require extensive training data to generalize well. This poses a significant challenge in EO applications where obtaining high-resolution, expert-annotated reference data is time-intensive and costly [342]. In many cases, synthetic data generation, transfer learning, or pre-training on related datasets are employed to mitigate the lack of labelled samples. Furthermore, inconsistencies introduced by human annotators can significantly affect DL model performance, requiring strict quality control during the labelling process. FM introduce a fundamental shift in AI-based EO applications, addressing many of the limitations of ML and DL

models in reference data preparation [228]. Unlike traditional AI approaches, which require manually labelled datasets, FM leverage large-scale self-supervised learning to learn feature representations without relying on explicit annotation. Recent advances in FM, such as the Segment Anything Model (SAM) [66], a promptable segmentation model capable of generating high-quality masks for any image region without task-specific training, and DINO [61], a self-supervised Vision Transformer trained without labels, have demonstrated the ability to generate pixel-wise labels automatically, significantly reducing the need for manual labelling. FM demonstrate strong adaptability in handling label noise and evolving class definitions, reducing the reliance on static reference labels in multi-temporal EO applications. In dynamic environments such as land cover change detection and vegetation monitoring, where traditional labels quickly become outdated, FM leverage self-supervised learning to refine and adapt training labels over time. Models like Changen2 [368] can generate supervisory signals for label correction, while recent evaluations show that FM remain label-efficient and generalize well in EO applications, even under limited annotation scenarios [242, 84]. Unlike DL models, which require fine-tuning with extensive labelled datasets, FM can generalize more effectively, adapting to new classes with minimal labelled examples through few-shot and zero-shot learning [13, 265].

Table 3.1.: ML methods to reduce dependency from exhaustive labelled datasets.

Method	Key Mechanism	Applications in EO
Transfer Learning	Adaptation of models pre-trained on related tasks to EO-specific problems [320, 248].	Land cover classification, drought assessment [320].
Self-Supervised Learning	Creation of supervisory signals from within the data itself, enabling feature learning [78].	Vegetation monitoring, anomaly detection [78].
Active Learning	Model identifies high-uncertainty samples and queries experts for targeted labelling [91].	Semantic labelling, urban morphology analysis [91].

While the labelling needs of ML, DL, and FM differ, various methods have been developed to mitigate dependence on exhaustive manual annotation. Table 3.1 summarizes these key techniques and their applications in EO. Despite their differences, both ML and DL approaches require robust reference data preparation to ensure training accuracy. In ML,

manual feature engineering remains a crucial step, demanding consistency across datasets and expert domain knowledge to define meaningful features. In contrast, DL's reliance on massive training datasets makes label availability, annotation quality, and scalability primary concerns. While ML models excel in structured, well-defined scenarios where labelled data is limited, DL models thrive in high-dimensional, data-rich environments where learning complex spatial and temporal patterns is essential [217].

3.2 Nature and Temporality of Labels

Labels are not singular entities; they exist across multiple dimensions, varying in type, scale, and temporal dynamics. Understanding these dimensions is crucial for designing effective models that generalize well across different spatial and temporal contexts. In many cases, multi-output ML and DL models must handle multiple target variables simultaneously, necessitating structured approaches to label representation. These models rely on clearly defined label types to extract meaningful relationships across different scales and temporal dependencies. To systematically describe the nature of labels, we classify them based on their measurement scale, temporal behaviour, and attribution structure.

The structured reference labels presented in Table 3.2 offer a comprehensive example for categorizing tree polygons into clearly defined data types, supporting systematic analysis in EO applications. Each data type reflects distinct measurement scales, analytical characteristics, and temporal implications. Since these labels are often derived from multi-source reference datasets, such as field surveys, airborne LiDAR, and multi-temporal satellite imagery, they integrate numerical properties like height, biomass, and spectral reflectance, as well as categorical attributes such as species, tree type, and vegetation health. Beyond these, environmental and climate-related labels may include variables like soil moisture, atmospheric conditions, and ecological succession stages, all of which require tailored preprocessing to ensure consistency across datasets.

Nominal data refer to categories without inherent order, such as "Tree Type" (coniferous or deciduous) and "Species" (e.g., Norway spruce, oak, Douglas fir). These labels facilitate forest type classification and biodiversity assessments by clearly distinguishing distinct groups without implying any ranking or hierarchical structure. Ordinal data are labels organized into ranked categories, where the order signifies progression or intensity without requiring equal numerical intervals. Examples in the provided dataset include

Table 3.2.: Example of structured reference labels for tree polygons in an exemplary attribute table with the respective scale level assigned in the last row.

ID	Type	Species	Age	Height	Infestation	Date	Not Before – Not After
1	Coniferous	Norway Spruce	Young	6m	Yes	2025-03-03 14:49:43	2024-03 – 2025-03
2	Deciduous	Oak	Mature	12m	No	2024-09-15 09:42:39	2023-09 – 2024-09
3	Coniferous	Scots Pine	Mid-age	8m	No	2023-06-22 13:19:27	2022-06 – 2023-06
4	Deciduous	Beech	Young	5m	Yes	2022-12-11 07:57:29	2021-12 – 2022-12
5	Coniferous	Douglas Fir	Mature	20m	No	2023-05-18 10:02:11	2022-05 – 2023-05
6	Deciduous	Birch	Mid-age	9m	Yes	2024-07-07 08:26:48	2023-07 – 2024-07
7	Coniferous	Larch	Mature	15m	No	2023-08-30 11:23:46	2022-08 – 2023-08
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	nominal	*nominal*	*ordinal*	*relational*	*binary*	*continuous*	*interval*

"Age," categorized into growth stages such as young, mid-age, or mature, and "Height," measured in meters, which indicates relative growth status from shorter to taller trees. These ordinal categories enable assessments of forest structure and succession stages. Relational data represent associations or relationships, such as those indicating spatial or contextual interactions between labelled entities or their environment. However, in this provided table, the relational aspect appears limited or not explicitly defined. If intended, relational labels might indicate proximity or contextual relationships, such as adjacency to disturbed areas, roads, or water bodies. Such information would need explicit representation, which appears missing from the current table. Binary data

contain only two possible states, typically representing "yes/no" or "presence/absence" conditions. The provided example dataset includes an infestation status indicating whether a tree is infested ("Yes") or not ("No"). This binary categorization directly supports analyses of forest health, infestations, and risk assessments related to pest outbreaks. Continuous data represent values measured on a numerical scale with precise, meaningful intervals. The provided table uses continuous data in the form of the exact "Date" of observation (e.g., "2025-03-03 14:49:43"), enabling precise temporal alignment with remote sensing observations or environmental events. Interval data define ranges or timeframes, specifying periods during which labels remain valid or applicable. In this dataset, the "Not Before – Not After" attribute (e.g., "2024-03 – 2025-03") indicates the temporal validity or applicability of the label. This ensures temporal consistency during analyses, especially when integrating multiple sources of satellite data collected at different intervals or for long-term environmental monitoring. Together, these structured labels provide a clear and coherent basis for preprocessing reference data, facilitating accurate analysis and consistent integration across diverse EO datasets and model types

The temporality of the labels themselves is another factor, which significantly affects the model's ability to generalize and adapt to evolving conditions. Static labels are often insufficient for tasks that involve temporal variability, while dynamic labels enable more robust modelling of time-sensitive phenomena. Static labels, typically created for one-time use, are suitable for environments with minimal change. Examples include static land cover maps, topographical surveys, and soil-type classifications. In such cases, ML models can perform well if the input data remains aligned with these static labels. However, static labels become problematic in dynamic environments characterised by temporal phenomena such as seasonal vegetation changes, urban development, or disaster events [57]. For instance, in vegetation monitoring, static labels fail to account for fluctuating indices like the NDVI and EVI, which track plant health over time. Seasonal cycles, comprising phases such as green-up, peak biomass, and senescence, introduce a temporal variability which static datasets cannot represent. As a result, models trained with static labels may generalize poorly across seasons. Dynamic labelling, which continuously updates reference labels to reflect real-time changes, enables models to effectively capture phenological events and seasonal cycles, thereby enhancing long-term predictive accuracy. For example, TempCNN has been successfully employed for vegetation monitoring. By integrating time-stamped labels, these networks have detected complex seasonal patterns in Sentinel-2 time series data [255]. However, the observed performance gains of TempCNN over RF and RNN (1–3% higher overall accuracy) were attributed in that study to its ability to model temporal dependencies in Satellite Image Time Series (SITS), rather

than differences in labelling. Their study compares different model architectures while keeping the reference labels fixed for validation and testing, meaning that the accuracy gains reflect architectural improvements rather than an effect of dynamic labelling. While dynamic labels have been shown to improve generalization in long-term monitoring tasks, their impact was not a variable tested in that specific study. Similarly, dynamic labels are critical for land use and land cover change detection. In the So2Sat LCZ42 dataset, dynamic local climate zone (LCZ) labels were periodically updated to account for urban infrastructure development and population shifts across 42 global regions. This approach allowed models to analyse urban morphology consistently, minimizing errors caused by label obsolescence [372].

3.3 Challenges in Dynamic Labelling

Dynamic labelling frameworks often employ pseudo-labelling and active learning, where labels are iteratively refined based on model predictions and feedback from new observations [342]. These strategies are particularly effective in scenarios requiring adaptive label handling, such as drought monitoring, flood mapping, and crop assessment [195]. Their role in structuring training samples for EO classification has also been emphasized in broader reviews on remote sensing preprocessing techniques [213]. Incorporating dynamic reference data, which captures temporal variations, is essential for improving model adaptability in EO applications. Static labels, although easier to manage, often fail to represent rapidly changing conditions such as those found in disaster monitoring or seasonal land cover dynamics. By contrast, spatio-temporal labelling strategies allow models to learn from evolving patterns, improving classification robustness and generalization across time [365]. Dynamic data evolves over time and includes variables like daily temperature, precipitation, or vegetation indices. Dynamic reference data is essential for capturing temporal patterns, making it critical for applications like crop monitoring or phenology studies. However, it requires more sophisticated handling to maintain temporal dependencies, and its integration with static data can be complex [63, 365]. To summarise, unlike static reference data, dynamic labels evolve over time, enabling ML, DL and or AI models to track and predict real-world changes in land cover, vegetation growth, and natural disasters. This flexibility is substantial in applications such as deforestation monitoring, crop growth assessment, and flood detection, where past conditions may no longer be representative of the present. In consequence, the transition from static to dynamic reference data introduces several challenges related

to temporal consistency, data quality, computational efficiency, and model adaptability. These challenges must be addressed to fully leverage dynamic labelling in EO applications.

3.4 Methodologies in Data Labelling and Processing

The mutual adaption of features and labels is crucial to any data-driven applications, serving as the foundation for effective data analysis. This holds true across various domains, including remote sensing. While labelling has been an interactive task for a long time in the case of deterministic classification methods such as Maximum Likelihood Estimation (MLE), the need for automated labelling strategies increases with the increased use of ML and DL techniques. Reference data are often affected by incompleteness, noise, inconsistencies, and multi-source integration challenges, all of which can reduce a model's performance if not properly addressed. Table 3.3 provides a structured summary of these challenges and their implications. The following sections then span the basic requirements for labels and present common methods for label engineering, followed by a discussion of the simultaneous use of labels and features.

Table 3.3.: Challenges in reference data collection and their implications for ML models.

Problem Category	Component Description	Interpretation / Role
Excessive Data Complexity	High-dimensional feature space, irrelevant attributes, large dataset sizes [292], mixed categorical/numerical data, noisy measurements.	Increases computational burden and risk of overfitting; requires dimensionality reduction, feature selection, and data filtering to retain relevant information.
Insufficient Data Coverage	Missing values, small sample size, incomplete or underrepresented attributes in labels [292, 114, 116].	Leads to poor model generalization and increased overfitting risk; necessitates imputation, data augmentation, or synthetic data generation to ensure robustness.
Inconsistent and Heterogeneous Data	Incompatible data formats, multi-source integration challenges [359], discrepancies in spatial and temporal resolutions.	Introduces inconsistencies in training data; requires harmonization, resampling, and normalization techniques to ensure data consistency and compatibility across datasets.

Requirements for Labels

Given the challenges outlined in Table 3.3, RS reference data must meet specific requirements in order to ensure accuracy, consistency, completeness, and temporal relevance in AI- and ML-driven applications. Given the high level of accuracy needed for predictions across spatial and temporal scales, the requirements for reference data are stringent.

Ensuring that a model learns the correct mapping between input features (such as satellite imagery) and target outputs (labels such as land cover types or tree height) requires highly accurate reference datasets. Inaccurate reference data can introduce systematic errors, resulting in unreliable or misleading predictions. For example, misclassified land cover data could lead to incorrect estimates of deforestation rates or vegetation health [115]. Consistency across different datasets and time periods is equally important. In EO, where data is sourced from multiple sensors, discrepancies can introduce noise and degrade model accuracy. This issue is particularly critical when merging datasets collected at different times or from varied sensors, which may exhibit radiometric differences unless properly calibrated. Ensuring data harmonization through preprocessing techniques is essential for model generalization. Completeness of reference datasets is another key requirement. In cases where data are incomplete, imputation methods such as K-nearest neighbours or DL-based techniques can be used to reconstruct missing values, although these methods introduce uncertainties [10]. Temporal relevance plays a pivotal role when dealing with dynamic environmental variables. Models trained on outdated or temporally misaligned data may yield erroneous predictions as environmental patterns shift over time. This is particularly critical in applications such as deforestation monitoring, precision agriculture, and phenological studies, where multi-temporal reference data significantly enhances classification accuracy and model robustness by capturing seasonal variations and land cover dynamics [324]. Studies have shown that leveraging multi-temporal datasets improves classification performance by reducing errors associated with single-date observations, which may not fully capture environmental variability [324]. However, single-date reference data remain valuable for classification tasks where short-term assessments or immediate land cover mapping are required. The aforementioned study successfully classified crop types using vegetation indices from a single RapidEye image, demonstrating that while single-date datasets provide meaningful insights, they have inherent limitations in capturing temporal variations [329]. In addition to these factors, reference data must be suitable for the specific task at hand. Furthermore, dynamic reference data in remote sensing applications often need to reflect evolving environmental conditions. Proper label engineering techniques must be employed to ensure multi-temporal consistency of the reference labels in order to prevent temporal drift that may negatively impact ML models.

The requirements for reference data in ML are therefore exceptionally demanding, particularly in RS, where spatial and temporal complexity is high, and small errors can propagate into large uncertainties. In the context of benchmarking datasets, these criteria are introduced and exemplified in detail in Section 1.1.3.

Label Engineering

Because EO data form the basis for many labels in the geoinformation context (e.g., the CORINE land cover maps of the Copernicus program [335]), reference labels in EO datasets may not always be as completely independent as desired. In addition, EO data are generally prone to data quality issues such as missing values, noise, and redundancy, which can propagate into the labelling process and potentially bias downstream ML and DL models. This subsection focuses on how preprocessing steps ranging from gap-filling to data fusion can support the production of accurate and consistent reference datasets. Missing values are common in EO-derived labels due to temporary sensor outages, atmospheric interference (e.g., cloud cover), and irregular data collection intervals. Effective interpolation methods help to mitigate these gaps by estimating or reconstructing missing label information.

For reference labels that evolve seasonally or exhibit complex temporal dynamics, advanced smoothing (e.g., Savitzky–Golay) can help to retain longer-term patterns while filtering short-term fluctuations [249]. These methods are crucial in applications such as phenological monitoring, where incomplete or noisy label data may otherwise obscure subtle vegetation changes. Because reference labels in EO can represent diverse data types, including land cover classes, vegetation indices, temperature, or biophysical parameters, they may inherit noise from sensor limitations, atmospheric disturbances, or inconsistencies in manual or automated annotation. Strategies for mitigating these issues include filtering techniques in raster-derived labels such as Gaussian smoothing, which is suitable for reducing random noise, as well as median filtering, which is suitable for removing outliers while preserving major structural features. When constructing reference datasets, one sensor alone may not achieve the necessary spatial or temporal resolution. Thus, data fusion leverages complementary information, such as combining high-resolution Landsat imagery with frequent MODIS observations, in order to generate more complete and robust labels [121]. Table 3.4 outlines typical strategies.

Table 3.4.: Label engineering techniques and their effects in EO workflows.

Method	Key Mechanism	Effects in EO
Raster Aggregation	Sums up labels from neighbouring pixels to form a coarser but smoother raster.	Useful for creating coarse yet stable label datasets in complex terrains [259].
Segment Aggregation	Aggregates measurements on predefined reference polygons, stabilizes label assignment, and enhances thematic consistency.	Applied to forest stands or field parcels in land-cover classification or object-based labels [43].
Cross-Sensor Interpolation	Combines different data sources with varying characteristics to enhance the temporal sampling.	Used for bridging Landsat revisits [371], densifying NDVI time series [276], and analyzing glacier dynamics [351].
Spatio-temporal Filtering	Smooths continuous labels via spatio-temporal aggregation for noise reduction.	Effective in removing short-term variations in meteorological measurements [143].
Normalization	Standardizes value range, semantic depth, and numerical coding.	Supports comparability and transferability across datasets.
Outlier Detection	Identifies and removes inconsistent or improbable values within the labels.	Helps correct for sensor failure or wrong timestamps in ground truth data [209].
Systematic Error Correction	Detects and adjusts for systematic deviations in the labels.	Mitigates issues like overestimated local temperatures in crowdsourced data [208].

Outliers in reference labels occur when inconsistent or improbable values emerge within the labelled dataset itself. For instance, if a crop type label assigned in a given year contradicts known crop rotation patterns or historical land use records, this may indicate a labelling error. Similarly, reference biomass values that deviate significantly from expected seasonal trends may suggest inconsistencies in the labelling process. Identifying and correcting such outliers using statistical validation, spatial consistency checks, or ML-based anomaly detection can enhance label quality before model training.

In addition to outlier correction, data fusion strategies enhance label consistency and accuracy in large-scale or long-term monitoring. Table 3.4 outlines key methods used to stabilize reference labels and improve spatio-temporal coherence. These approaches are particularly valuable for tracking both short-lived events (e.g., forest disturbances) and broader environmental changes. A detailed discussion of such fusion strategies has been presented in the literature [360], particularly in the context of integrating multi-source remote sensing data for EO applications [289, 286, 288] for applications such as forest monitoring, which requires robust handling of both spatial and sensor variability, in turn helping to avoid propagation of errors into the derived labels. Optimizing data quality in reference labels involves balancing corrective measures with the need to preserve critical information about local variability, temporal patterns, and class distinctions. Unlike classical feature engineering, label engineering focuses on ensuring that reference data accurately reflect the intended classification task rather than just optimizing predictor variables. By reducing redundancy at the labelling stage, models can achieve better generalization and interpretability without unnecessary inflation of label complexity.

3.5 Understanding Challenges and Best Practices in Dynamic Data Processing for Labelling

Figures 3.1–3.5 highlight the practical challenges associated with reference data preprocessing. These visual representations underscore the importance of addressing the specific requirements of labelled data from external sources for real-world EO applications.

Simultaneous Use as Labels and Features

Comparing preprocessing methods for the labels presented above to common feature engineering reveals wide agreement; the mathematical approaches are identical, and only depend on the scale level of the respective input variable (Table 3.2). With respect to gap-filling approaches such as cross-sensor interpolation (Table 3.4), which estimate missing measurements from one sensor using potentially different measurements from another sensor, the features of one might act as labels when using the features of one other. Table 3.5 lists certain EO variables that have been used as both features and as labels, where certain transformations and selections may apply at each stage. Traditional ML models often rely on hand-crafted label definitions that require domain expertise. For instance, thresholds or discrete classes might be derived from carefully curated spectral–spatial indices. As an example, vegetation health classifications might use NDVI thresholds (e.g., NDVI $>$ 0.6 for dense vegetation, 0.3–0.6 for sparse vegetation, and $<$0.3 for barren land), while forest type classification might integrate spectral information with elevation and climate variables to distinguish between deciduous and coniferous forests. Similarly, in urban heat island studies, thermal infrared data combined with land surface temperature and impervious surface fractions can define thresholds for categorizing heat stress zones [209].

Table 3.5.: EO data dimensions that serve as a source of both input features and reference labels.

Dimension	Description
Spectral	Spectral features originate from sensor bands (visible, infrared, near-infrared, etc.) and are widely used in remote sensing applications. Classic examples include the NDVI [57, 52], Green Chlorophyll Vegetation Index (GCVI), and EVI [249, 67]. The Land Surface Water Index (LSWI) is used for assessing water content, aiding in drought and flood monitoring [52].
Spatial	Spatial patterns—such as texture metrics like contrast, entropy, correlation, and variance—are critical for differentiating urban areas, forests, and agricultural fields [57]. These features help refine label boundaries by emphasizing spatial consistency, especially in object-based labelling workflows that rely on segment homogeneity.
Temporal	Time series data reveal dynamic processes such as crop growth, forest phenology, and seasonal hydrological cycles [57]. Incorporating temporal statistics (e.g., annual maxima/minima, NDVI frequency peaks) helps refine reference classes by clustering areas with similar phenological trends across years.
Specific	These features incorporate domain-specific knowledge—e.g., topography, meteorological data, or socioeconomic layers. In forest fire risk applications, slope orientation and wind speed can directly support label design or rule-based logic [57]. By integrating these with core EO features, labelling becomes more robust and context-aware.

The possible exchange of features and labels exhibits one basic problem: redundancy in features is commonly accepted and even integrated into models, as it naturally arises in multispectral and hyperspectral remote sensing acquisitions. In contrast, redundancy is not considered at all in labels, as they are originally handcrafted and mostly seen as an ideal error-free reference. Exchanging labels and features also exchanges their respective characteristics; for instance, it has been found that NDVI and vegetation cover

are understandably highly correlated in forest fire modelling, leading to the removal of one attribute to prevent redundancy [57]. In label engineering, redundancy can arise when multiple reference labels provide overlapping information, which may complicate class interpretation, introduce bias, or create inconsistencies in supervised learning tasks. To ensure that labels remain distinct and meaningful, redundancy detection strategies can be applied.

The structured refinement of labels from features offers several advantages. Properly defined labels that align closely with the predictive objective help to reduce ambiguity, thereby improving model generalization. Removing redundant or poorly correlated attributes also enhances interpretability, allowing users to better understand how land cover categories and other reference classes are defined. In addition, streamlined label design improves scalability in large EO datasets by reducing unnecessary complexity in training and inference. By carefully adapting techniques from reference data validation, practitioners can ensure that the labels meaningfully represent the target variables, leading to more reliable model performance in EO applications ranging from resource management to environmental hazard prediction.

Integration of Irregular Reference Labels with Raster Data

One of the primary challenges in dynamic labelling is integrating irregular vector-based reference labels with regularly-gridded raster datasets. The labels collected through terrestrial (Figure 3.1), airborne PolInSAR (Figure 3.2), LiDAR (Figure 3.3), and airborne photography surveys (Figures 3.4 and 3.5) contain both numerical and categorical information. For example, raw LiDAR data provide continuous numerical values such as height, return intensity, and point density; however, when classified into vegetation types (e.g., 'coniferous', 'deciduous'), land cover categories (e.g., 'urban', 'forest', 'water'), or object classes (e.g., 'building', 'tree', 'road'), these are transformed into categorical data. Similarly, PolInSAR-based land cover classification outputs discrete labels that require encoding before being processed in ML/DL models. To ensure compatibility, categorical labels must be transformed into numerical representations:

Preprocessing: Before model training, categorical labels require encoding (e.g., one-hot or ordinal encoding). Alternatively, structural attributes (e.g., tree height, crown diameter) or spectral properties (e.g., NDVI values) can serve as numerical predictors.

Postprocessing: After inference, numerical model outputs (e.g., fractional land cover predictions) must be reclassified into discrete categories to match thematic mapping requirements.

The choice of transformation depends on the specific ML/DL task. Not all applications require categorical-to-numerical conversion, and alternative methods such as multivariate regression can effectively leverage continuous data.

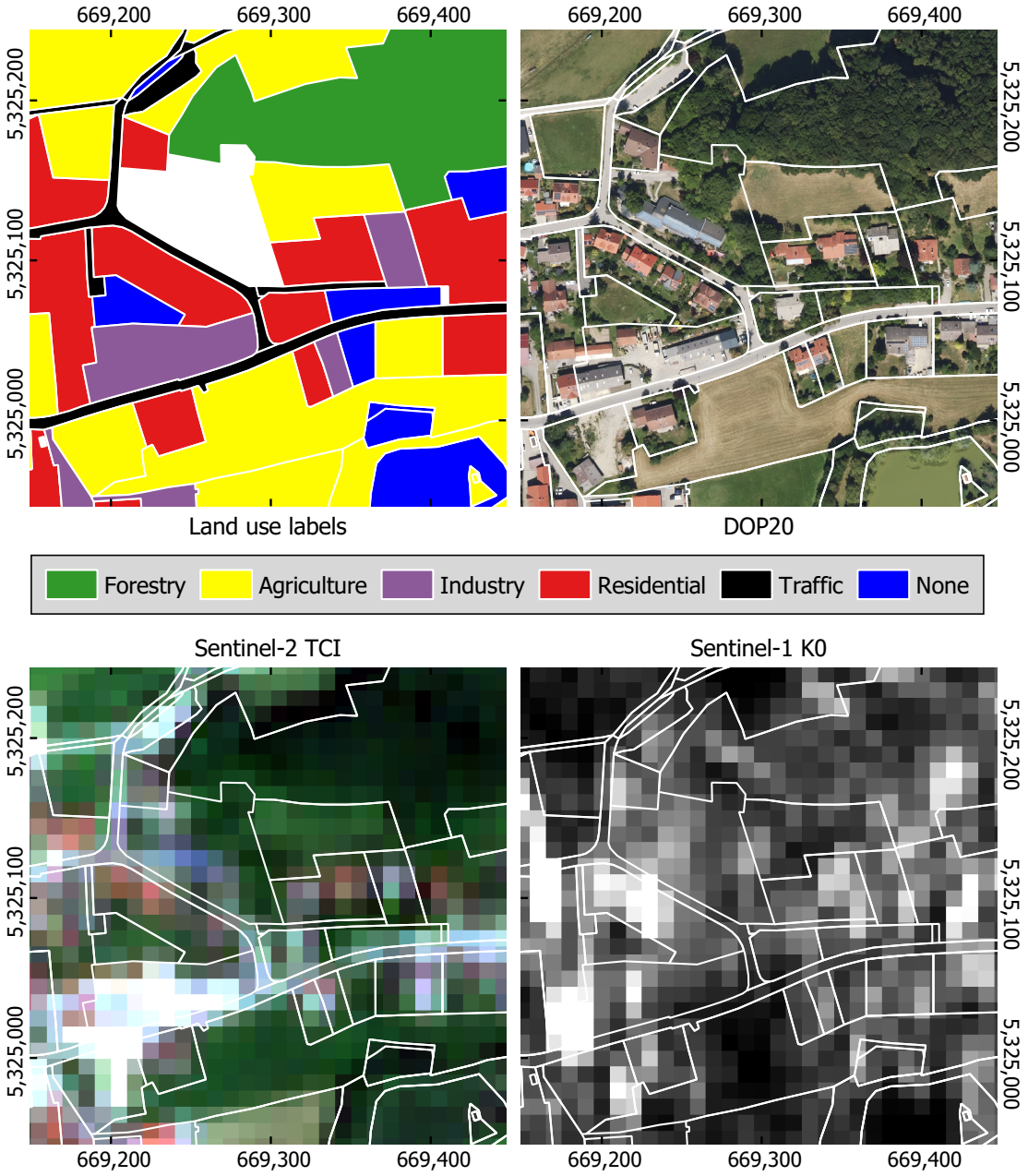


Figure 3.1.: Labels from terrestrial surveys in the rural community of Hochstadt (Bavaria, Germany) in comparison to different EO data sources: (top-left) land use labels as provided by the *Bavarian Surveying Administration (Bayerische Vermessungsverwaltung)—www.geodaten.bayern.de* (accessed on 1 February 2025) (top-right) digital orthophoto 20 cm (DOP20 by the *Bavarian Surveying Administration (Bayerische Vermessungsverwaltung)—www.geodaten.bayern.de* (accessed on 1 February 2025); (bottom-left) Sentinel-2 (©ESA (2023)) true colour image (TCI); and (bottom-right) Sentinel-1 (©ESA (2023)) total intensity (K0). The figure elucidates the impact of image resolution and geometric co-registration on the usability of labels. On the one hand, the DOP20 shows much more details than the labels require; on the other, the satellite images are too coarse to capture the relatively narrow polygons of (e.g.) the traffic class. Regarding Sentinel-1, the signatures of high-rise objects like the buildings or trees are spatially overlaid with neighbouring polygons.

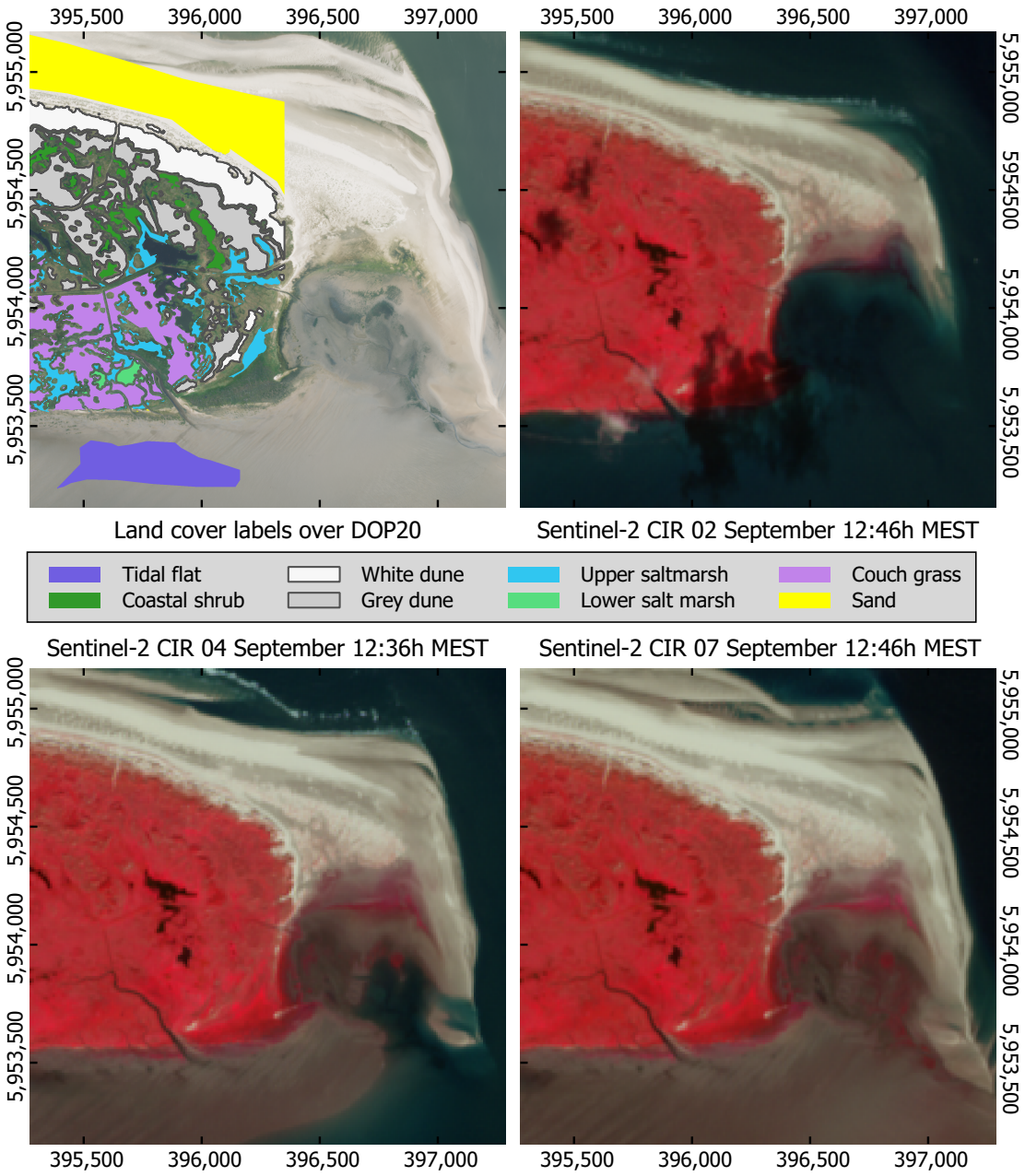


Figure 3.2.: Labels from an airborne PolInSAR flight campaign over the German Wadden See around the island of Baltrum (Lower Saxony, Germany) in comparison to multi-temporal spaceborne optical acquisitions in the visible and near-infrared spectral range: (top-left) land cover labels [157] (accessed on 1 February 2025) with digital orthophoto 20 cm in the background (LGLN (2024)), and Colour Infrared (CIR) images by Sentinel-2 on September 2nd, 4th, and 7th (©ESA (2023)) as multi-temporal features. The figure impressively visualizes the high temporal variability of features acquired by spaceborne EO sensors due to the immanent tidal range opposite the temporally stable land cover classes.

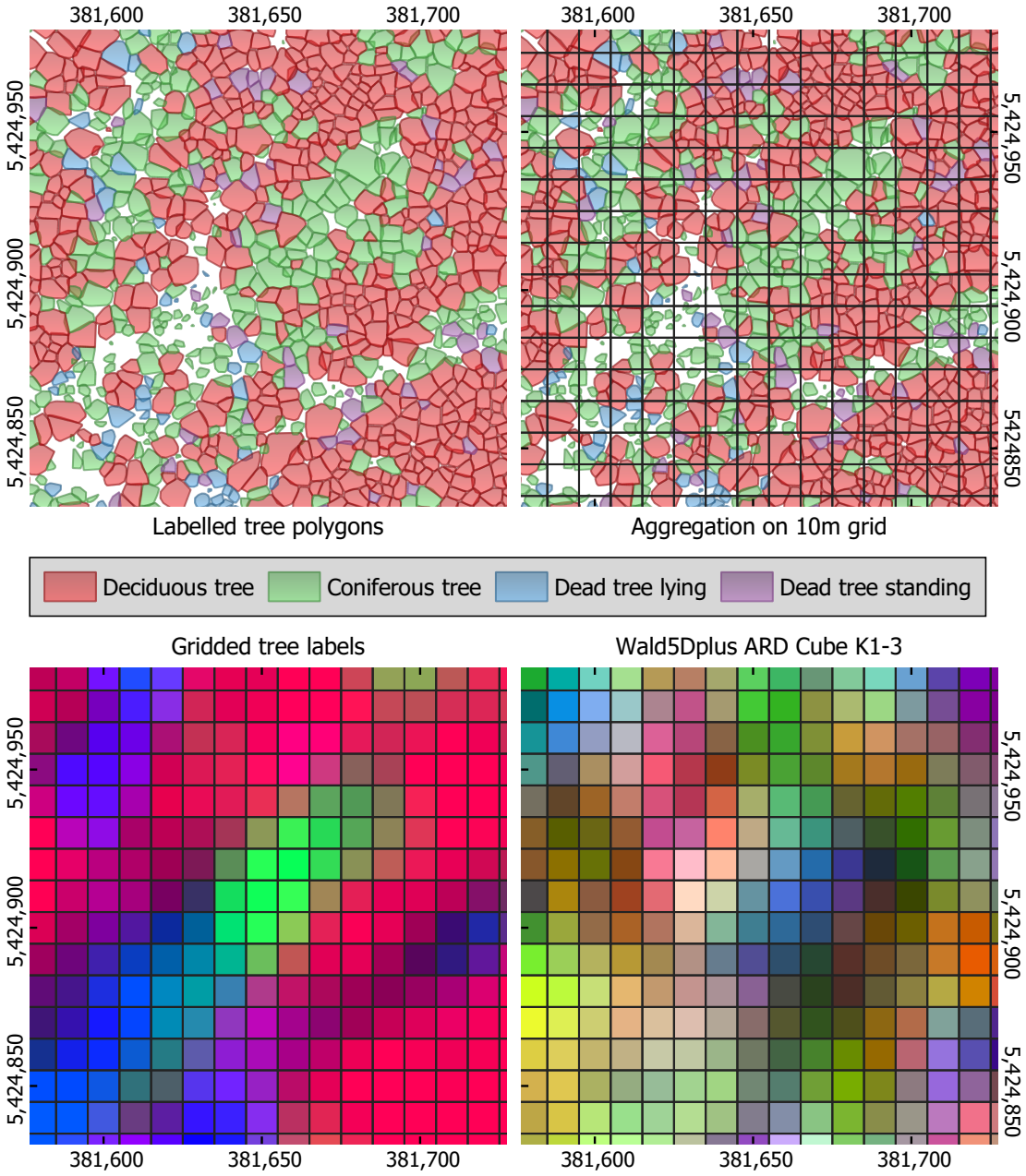


Figure 3.3.: Labels from airborne LiDAR in the Bavarian Forest National Park conditioned for use with spaceborne sensors: (top-left) single tree polygons derived from point clouds that contain the tree geometry and further characteristics as attributes. These labels concerning the Bavarian Forest National Park were provided by the Bavarian National Park Research under the Bohemian Forest Datapool Initiative [203] (accessed on 29 February 2024); (top-right) the 10 m pixel grid of the satellite data; (bottom-left) tree characteristics aggregated on the grid by the Wald5Dplus project [147] for use as labels; and (bottom-right) Kennaugh elements 1 to 3 of the 512 bands included in the Analysis-Ready Data (ARD) cube provided by Wald5Dplus [148] (accessed on 1 February 2025) for use as features. The figure addresses the two main labelling problems of Wald5Dplus: first, the gridded labels represent geospatial statistics instead of single tree characteristics; second, the multi-temporal EO features contain structures that are not visible in the mono-temporal labels and vice versa.

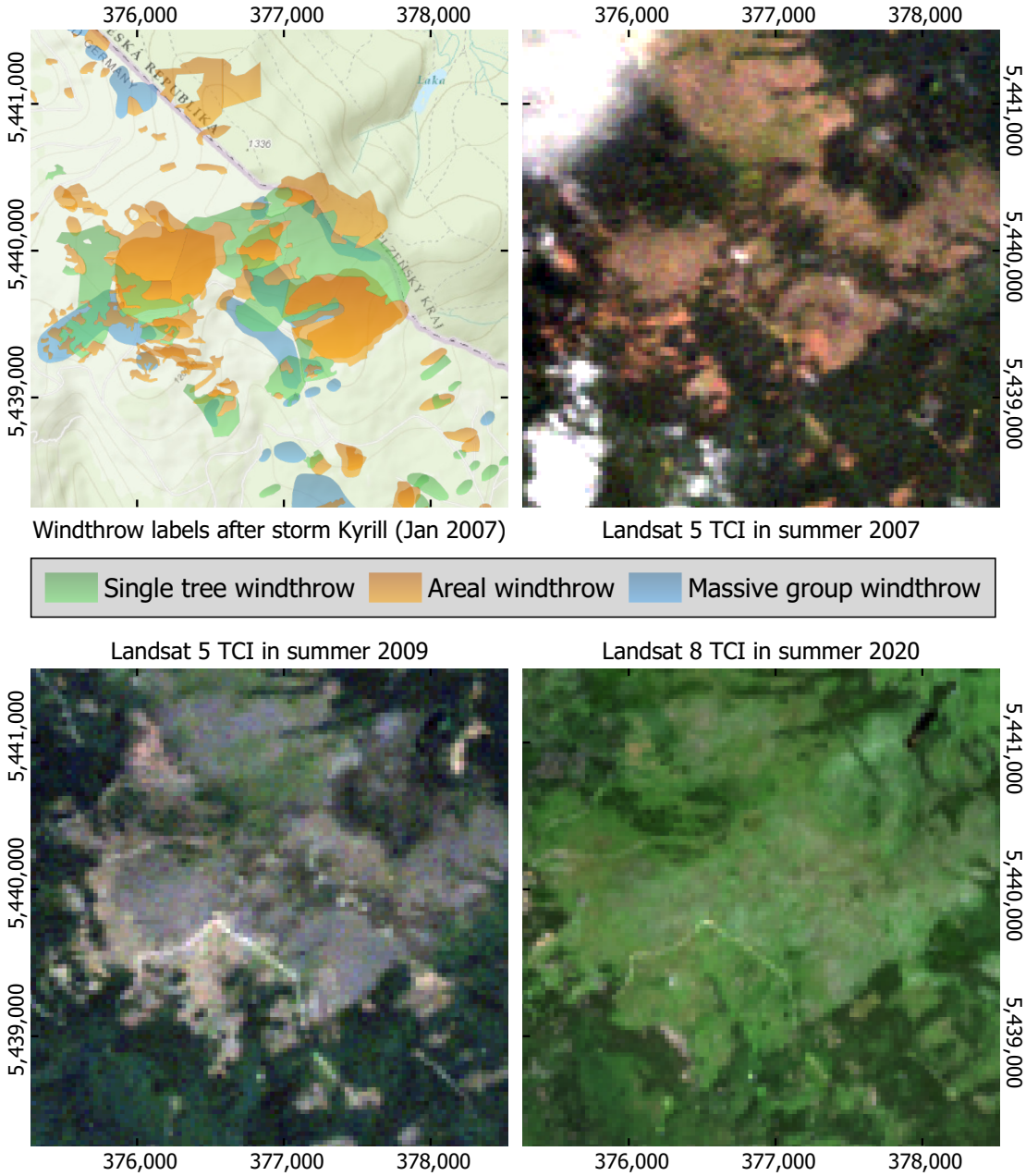


Figure 3.4.: Labels from airborne photography: (top-left) manually drawn wind-throw areas after storm Kyrill in January 2007 categorized in single-tree, group, and areal wind-throw in the Bavarian Forest National Park. The labels concerning the Bavarian Forest National Park were provided by the Bavarian National Park Research under the Bohemian Forest Datapool Initiative [203] (accessed on 29 February 2024), with the ESRI World Topo Map in the background. The other sub-figures show Landsat True Colour Images (TCI) taken from space in the years 2007, 2009, and 2020 in parts (top-right sub-figure) with some clouds (Landsat 5 and 8 images courtesy of the U.S. Geological Survey). The reference data consist of overlapping polygons, which inhibits the assignment of clear label to pixels. Although the satellite image from summer 2007 takes up the structures of the labels, many more areas appear very similar to the mapped wind-throw areas, which underlines the necessity of multi-temporal features and/or the inclusion of static labels. The image from 2009 indicates clearing after the storm, whereas the image from 2020 reveals regrowth.

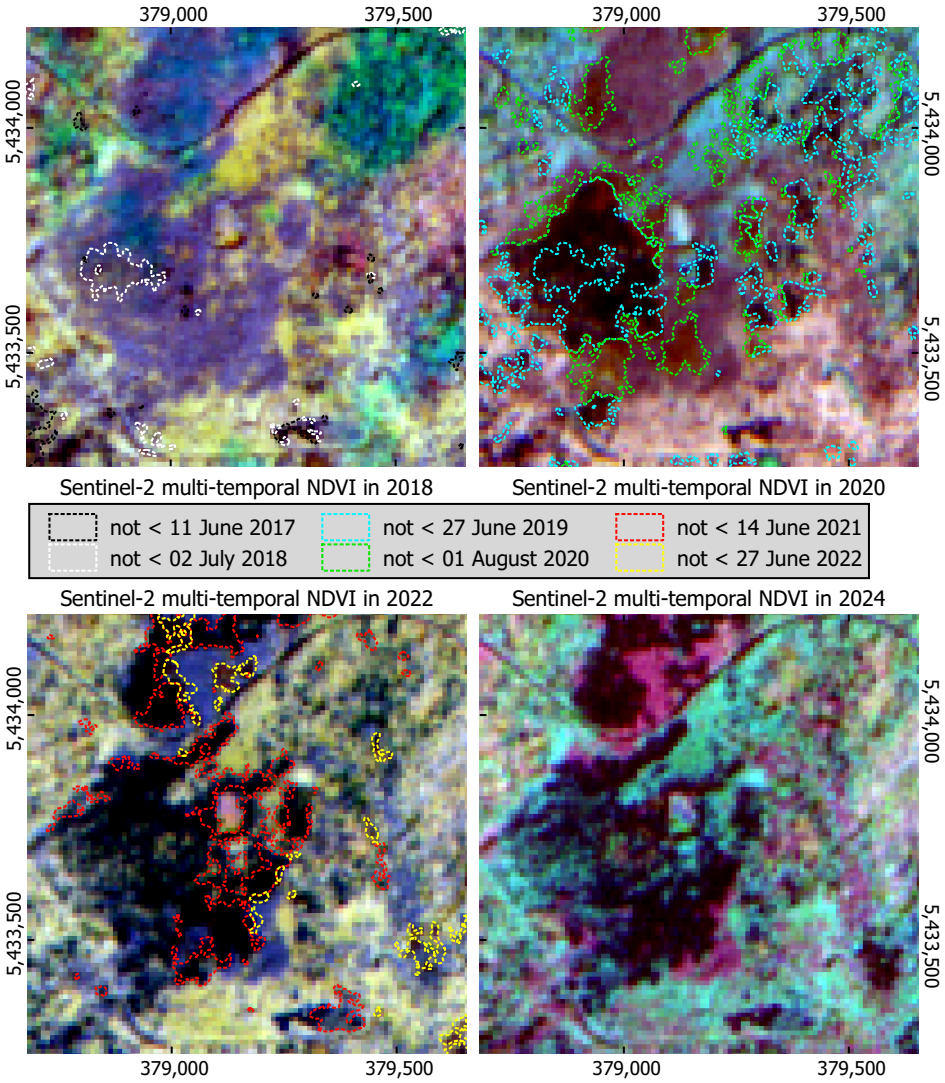


Figure 3.5.: Multi-temporal labels from airborne photography: yearly deadwood after barkbeetle infestation mapped by a human interpreter based on stereoscopic images acquired during yearly airborne flight campaigns. The polygons delineate dead trees, categorized by the last date on which they were classified as healthy. Labels concerning the Bavarian Forest National Park were provided by the Bavarian National Park Research under the Bohemian Forest Datapool Initiative [203] (accessed on 29 February 2024). The raster image in the background contains the multi-temporal Normalized Difference Vegetation Index (red: NDVI in spring; green: NDVI in summer; blue: NDVI in autumn) from Sentinel-2 images (©ESA (2018, 2020, 2022, 2024)). The brightness shows the healthiness of the vegetation, whereas the hue shows its temporal variation, e.g., red stands for high photosynthetic activity in the spring and reduced photosynthetic activity in the summer and autumn. Dark areas stand for low-to-negligible photosynthetic activity throughout the year. This figure illustrates the challenges of temporal alignment; some upcoming deadwood areas are already visible in the space-borne time series, even though they are still classified as healthy by the yearly manual assessment. Thus, the image from 2024 (bottom-right) shows a composition of deadwood and regrowth areas that only partially match the reference polygons due to the increasing time lag.

Spatial Alignment, Projection Distortions, and Resolution Challenges

To ensure usability, labels must be spatially aligned with EO-derived data such as high-resolution digital orthophotos (e.g., 20 cm DOP, *Bavarian* Surveying Administration—www.geodaten.bayern.de and raster-based satellite products (e.g., Sentinel-1 and Sentinel-2). However, vector-to-raster transformations introduce distortions that require harmonization techniques. Figure 3.1 highlights various challenges:

Projection distortions: High-rise objects cause layover effects in optical and radar data, leading to misalignment between objects and their corresponding labels. In Figure 3.1 (top-right), the forest shifts into the neighbouring meadow.

Resolution mismatches: High-resolution imagery (e.g., DOP20) captures detailed land structures, while satellite images (e.g., Sentinel-2) are too coarse to represent narrow traffic polygons. Radar images (bottom-right) introduce additional complications, with buildings overlaid onto neighbouring polygons.

To mitigate these issues, Figure 3.2 demonstrates a possible solution in the form of a buffer implemented around the reference polygons to minimize pixel-mixing errors. Additionally, spectral and radiometric discrepancies (e.g., between Sentinel-1 and Sentinel-2) necessitate normalization to ensure label consistency. Figure 3.3 exacerbates resolution issues, as the gridded labels represent geospatial statistics instead of individual tree properties. One grid cell may contain multiple overlapping polygons, complicating the extraction of independent descriptors. This highlights the necessity of advanced spatial statistics to effectively handle polygon overlaps.

Temporal Misalignment and Dynamic Label Challenges

Temporal discrepancies between reference labels and EO data present a fundamental challenge. Static land use labels (Figure 3.1) and land cover labels (Figures 3.2 and 3.3) do not capture seasonal or short-term variations in landscape features. Meanwhile, dynamic EO-derived features such as Sentinel-1 and Sentinel-2 provide near-weekly revisit times, revealing vegetation cycles and environmental changes. Figure 3.2 illustrates this issue, showing that temporally stable land cover labels contrast with rapidly changing tidal ranges. Similarly, Figure 3.5 highlights inconsistencies in airborne reference labels; while manual interpretation classifies areas as healthy, Sentinel-2 NDVI trends indicate

early signs of tree stress and dieback. To improve temporal alignment, preprocessing strategies include the following:

Temporal interpolation: Estimating missing or delayed label updates based on surrounding timestamps.

Change detection and trend extrapolation: Identifying trends in EO features to better align with reference labels.

Adaptive temporal grouping: Aggregating neighbouring observations to improve label consistency across time.

It is important to keep in mind that both the appearance of an object and its semantic class may change independently with time, i.e., from 'healthy' to 'threatened' in vegetation monitoring. These changes are not necessarily visible in EO data.

Uncertainty and Ambiguity in Label Assignments

Reference data inherently contain uncertainty due to overlapping polygons, ambiguous class assignments, and spectral mixing in lower-resolution EO products. For example, Figure 3.4 presents overlapping wind-throw polygons that make pixel-level classification ambiguous. Similarly, Figure 3.5 visualizes inconsistencies between multi-temporal Sentinel-2 NDVI and manually annotated deadwood polygons. To address these issues, best practices emphasize the following:

Probabilistic labelling: Assigning probability values rather than strict class assignments to improve robustness.

Confidence-weighted annotations: Including model-based uncertainty measures in label assignments.

Multi-label fusion: Combining labels from different sources to enhance label consistency.

By proceeding in this way, contradictions and inaccuracies are ignored, instead being adequately mapped in the annotated labels.

Scalability and Computational Challenges in Large-Scale ML Workflows

As EO datasets grow, label preprocessing becomes computationally expensive. Unlike static labels that require one-time annotation, dynamic labels must be continuously updated, introducing substantial processing demands. Challenges include:

High-dimensional data processing: Multi-temporal EO datasets require scalable architectures (e.g., distributed computing, cloud-based workflows).

Automated label updates: Techniques such as active learning, transfer learning, and weak supervision aim to reduce manual intervention but also introduce complexities in model retraining.

Metadata management: Proper documentation of label transformations is necessary for reproducibility.

Recent advancements in graph-based labelling, dynamic pseudolabelling, and spatio-temporal data integration frameworks show promise for improving scalability, but require further refinement prior to widespread adoption.

Despite these challenges, the datasets analysed in Figures 3.1–3.5 exhibit high-quality reference data from well-structured surveys. The accessibility of datasets such as the PolInSAR-derived land cover labels [157] and Wald5Dplus tree characteristics [148] supports reproducible EO research such as [278]. However, restricted access to high-resolution commercial datasets currently limits large-scale ML model generalization. For a standardized preprocessing framework, robust workflows must incorporate the following:

- Schema matching to align label structures across datasets.
- Spatial alignment techniques to mitigate projection and resolution discrepancies.
- Adaptive resampling to harmonize multi-temporal and multi-resolution data sources.
- Dynamic updating mechanisms to ensure long-term consistency in evolving datasets.
- Hybrid labelling strategies that integrate categorical and continuous reference data, improving model adaptability by incorporating multiple label types within a unified framework.

Hybrid labelling enhances the adaptability and robustness of reference data by integrating categorical and continuous classifications as well as static and dynamic labels. Many EO applications require structured harmonization of numerical and discrete data; for instance, land surface temperature models benefit from linking continuous temperature measurements with categorical land cover classes, while vegetation indices such as NDVI improve crop growth stage identification when combined with crop type classifications. In addition to numerical-categorical integration, hybrid labelling merges static labels such as historical land use maps with dynamic labels like satellite-derived flood extents, ensuring adaptability to real-time changes. For example, crop monitoring leverages static soil maps alongside dynamic NDVI-based classifications to capture both long-term soil properties and short-term vegetation shifts. Additionally, hybrid labelling enables transformed representations that better reflect environmental complexity. Fuzzy classification assigns probabilistic weights to land cover types, facilitating smooth transitions between categories, while continuous degradation scores in vegetation health assessments offer a more nuanced representation of environmental stressors. These approaches enhance model generalization, improve data reliability, and support more accurate predictions in EO applications. These best practices improve label consistency, model interpretability, and overall robustness, paving the way for scalable, high-quality reference data in various learning-based applications for EO.

3.6 Dynamic Labelling and Sampling Strategies: Temporal and spatio-temporal Perspectives

Below, several approaches are presented, including various research works that deal with the dynamics of time series and offer different perspectives. These methods vary in complexity, automation, and applicability, providing tailored solutions depending on the analytical tasks and data availability. Tables 3.6 and 3.7 provide an overview of key dynamic labelling techniques in EO, outlining their methodological characteristics and their relevance to either temporal or spatio-temporal applications.

Table 3.6.: Temporal dynamic labelling methods in EO.

Method	Description
Time-Lagged Labels	Labels are assigned based on past observations to account for delayed responses in environmental processes, such as NDVI changes driven by precipitation. This approach ensures that historical dependencies are incorporated into model training, improving predictive accuracy in applications such as climate–vegetation studies and hydrological forecasting. However, these labels remain static after being assigned and do not adapt dynamically to changing conditions. They are commonly used to structure reference data for time-series analysis [171].
Sliding Window Technique	This technique segments time series into structured subsets to support anomaly detection, data imputation, and dynamic label generation. It extends time-lagged labelling by refining structured temporal dependencies, ensuring consistency in training labels while capturing meaningful temporal variations. Selecting an appropriate window size is necessary to balance short-term fluctuations with long-term trends. It is widely applied in hydrological monitoring and preprocessing for remote sensing classification, where it enhances temporal consistency in training datasets [195, 213].

Although these methods offer robust frameworks for handling dynamic labels, each approach comes with inherent limitations. For instance, while pseudolabelling reduces manual annotation, it can introduce noisy labels if iterative refinements are not properly managed or if confidence thresholds are miscalibrated [342]. Time-lagged labels effectively capture temporal dependencies but remain static once assigned, which may lead to mismatches in fast-changing environments [171]. Sliding window techniques ensure temporal consistency, but are sensitive to parameter selection, particularly the window size, which can distort long-term trends or miss short-term anomalies [213, 195]. AutoGeoLabel provides real-time label generation from geospatial data, enhancing scalability and reducing manual workload. However, several critical limitations must be addressed

when applying this method: (1) the spatial representativeness of the generated labels is constrained by the coverage and sampling of the LiDAR or remote sensing inputs, potentially introducing geographic bias if certain regions are underrepresented; (2) in applications involving vegetation phenology or seasonal dynamics, labels generated at different time points may not reflect consistent environmental states, reducing temporal reliability; (3) label accuracy is highly sensitive to the quality and resolution of the input data, with sparse or misaligned sources leading to incomplete or noisy labels; and (4) independent validation using ground truth data is essential in order to avoid propagating misclassifications into downstream ML/DL models [11]. Finally, resampling and data fusion, while addressing multi-resolution inconsistencies, risk introducing errors from mixed pixels or misaligned temporal data points. These limitations indicate a clear need for a unified and scalable methodology that can dynamically adapt labels across diverse EO applications and maintain accuracy while addressing the temporal and spatial complexities inherent in environmental datasets.

Table 3.7.: spatio-temporal dynamic labelling methods in EO.

Method	Description
Pseudolabelling	Iteratively refines weakly supervised object detection by generating instance-level annotations from spatial and temporal information. The process includes: (1) training a weakly supervised localization model to generate Class Activation Maps (CAMs); (2) computing pseudolabels based on pixel intensities, assigning confidence scores; (3) applying adaptive thresholding using category-specific confidence histograms; and (4) refining pseudolabels through iterative integration using Proposal Cluster Learning (PCL). Prior to pseudolabelling, datasets are typically resampled and fused to ensure spatio-temporal consistency across sensors like Sentinel-1 and Sentinel-2. This reduces dependency on fully annotated datasets and improves detection performance across iterations [342].
AutoGeoLabel	Automatically derives reference labels from geospatial data via statistical feature extraction from LiDAR and multispectral imagery. Variables such as reflectivity, elevation, and return counts inform classification rule generation for land cover differentiation. The method supports dynamic environmental monitoring (e.g., flooding, vegetation), adapting to real-time changes. Data alignment and resolution are critical for label accuracy, necessitating preprocessing steps such as fusion and resampling. Label quality is validated against ground truth to mitigate classification bias and downstream error propagation. AutoGeoLabel is increasingly used for scalable, automated labelling in geospatial applications [11].

The challenges outlined in the previous sections highlight the need for a structured and unified approach to dynamic labelling that can integrate multi-source data while addressing spatial, temporal, and categorical inconsistencies. To meet these demands, we introduce the HELIX, a comprehensive spatio-temporal label preprocessing framework designed to standardize and enhance EO-based training data.

4 The Novel Helix Framework for Dynamic Label Data

> *We are drowning in information, while starving for wisdom.*
>
> — **Edward O. Wilson**
> Biologist, Ecologist

This chapter includes elements from the following peer-reviewed publication:

Sarah Hauser, Lena Augner, and Andreas Schmitt. *Perfect Labelling: A Review and Outlook of Label Optimization Techniques in Dynamic EO*. *Remote Sensing*, 2025, 17, 1246. DOI:10.3390/rs17071246

It is cited as [149] and is marked with a cyan line.

Author Contribution: Sarah Hauser was instrumental for the full study design and conceptualization presented in this work, including the independent development of the HELIX framework for spatio-temporal label preprocessing. She led the investigation and contributed substantially to the manuscript's review and editing.

The challenges outlined in the previous chapter highlight the need for a structured and unified approach to dynamic labelling that can integrate multi-source data while addressing spatial, temporal, and categorical inconsistencies. To meet these demands, this section introduces the novel *HELIX* concept, a comprehensive spatio-temporal label preprocessing framework developed specifically to standardize and enhance label data. As an original contribution of this thesis, HELIX provides a systematic and scalable solution to one of the most persistent bottlenecks in EO–ML integration: the lack of temporally aligned, structurally coherent, and context-aware reference labels. Beyond its role as a practical

tool, HELIX also functions as a conceptual framework: it provides a foundation for transforming, enriching, and aligning label data based on spatio-temporal understanding. This abstraction enables reuse and adaptation in other environmental domains or modelling pipelines that require structured label data across dynamic conditions.

The proposed HELIX framework provides a comprehensive spatio-temporal approach to data preprocessing that is intended for but not limited to EO applications. HELIX addresses the need for a unified preprocessing workflow as well as the limitations of purely static or purely dynamic datasets. Drawing its name and inspiration from the intertwined structure of a DNA helix, the proposed framework is conceptualized as an evolving sequence of data points interlaced along both spatial *(x,y)* and temporal *(t)* coordinates. By structuring label data within a spatio-temporal grid, each referencing a specific *(x,y,t)*, the proposed framework effectively captures the continuous changes of environmental phenomena over time while preserving spatial consistency and contextual integrity. This design is pivotal for EO tasks that demand high temporal resolution (e.g., seasonal vegetation changes, tidal fluctuations) and spatial precision (e.g., delineating tree polygons, detecting wind-throw damage, identifying deadwood). Whereas static datasets fail to incorporate ongoing environmental dynamics and purely real-time datasets can disregard historical context, this pipeline harmonizes both, providing a balanced pipeline for integrated multi-source EO data. The HELIX framework in Figure 4.1 consists of multiple interlinked modules, each fulfilling a distinct purpose in the dynamic labelling workflow.

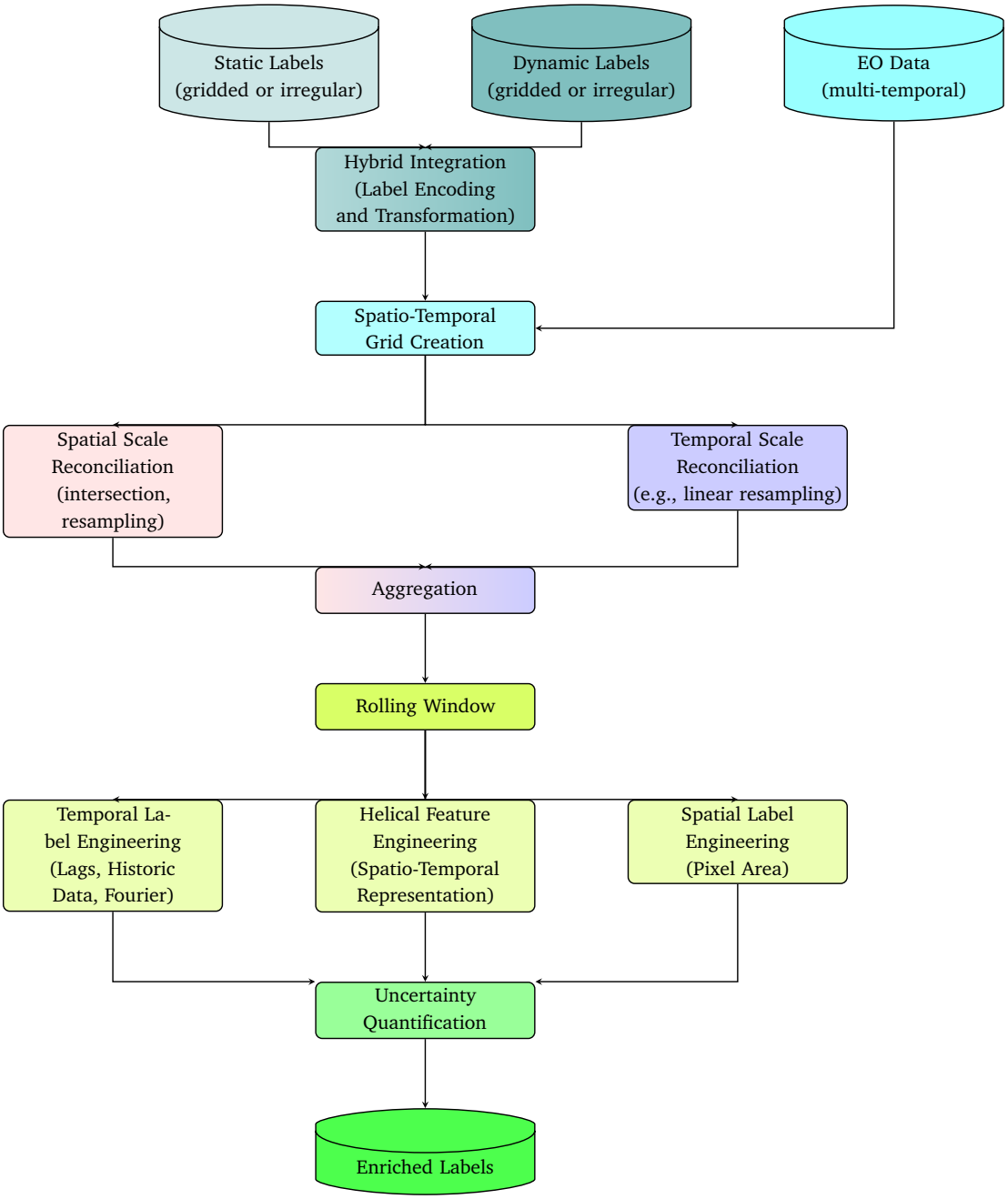


Figure 4.1.: Conceptual framework of the proposed HELIX framework, illustrating the integration and preprocessing pipeline for static and dynamic labels.

4.1 Framework Formalization and Design Principles

4.1.1 Hybrid Integration of Static and Dynamic Labels

The HELIX framework integrates various label sources into a unified hybrid dataset, irrespective of their original formats, including irregular vector data, gridded raster data, and georeferenced structured datasets (e.g., CSV). This integration effectively harnesses the complementary strengths of both temporally static data (e.g., soil maps, historical land use data) that provide stable long-term baseline conditions and temporally dynamic data (e.g., climate variables, vegetation indices) that capture environmental changes over time. The combination of diverse datasets regardless of their temporal characteristics (static or dynamic) or data types (numerical, categorical) is highly advantageous for training robust models in ML, DL, AI, and FM contexts. The HELIX framework automatically manages different data types, transparently encoding categorical labels into numerical forms to ensure traceability and reproducibility. By preprocessing all these diverse datasets within a single coherent pipeline, HELIX simplifies data handling, enhances model training efficiency, and maintains consistent label quality. This hybrid integration approach also adaptively addresses temporal variability, dynamically selecting between stable representations for persistent landscape features and dynamic representations for capturing short-term environmental fluctuations. For instance, Figure 3.5 illustrates scenarios where manually annotated static labels lag behind the frequently updated dynamic labels derived from Sentinel-2 NDVI data. Similarly, Figure 3.2 underscores the necessity of dynamic labels for accurately representing rapidly changing environments such as tidal zones.

In many real-world environmental datasets, labels are not always provided as neatly formatted numerical values. Instead, they often appear as categorical descriptors, such as vegetation types, disturbance events, or administrative classes, which must be interpreted and standardized before they can be used in machine learning workflows. To address this, the HELIX includes a mechanism for systematically processing and normalizing such non-numeric labels across heterogeneous data sources. Rather than assuming a uniform input schema, HELIX allows for fine-grained selection of which categorical fields should be included in the label processing workflow, drawing from multiple sources such as static land-use maps, dynamic event records, or historic annotations. These fields are then transformed into a consistent numerical format to ensure compatibility with downstream models. Importantly, this transformation is designed to retain interpretability: each

category is assigned a unique integer code, and mappings between original class labels and their encoded counterparts are preserved and exported for later use, such as decoding predictions or inspecting feature contributions. The design choice to use compact integer encodings, rather than more expansive one-hot representations, reflects both the scale of the data HELIX is intended to process and the types of models it supports. Integer encoding is highly memory-efficient and aligns well with many common modelling approaches (e.g., decision trees, ensemble methods) that can inherently interpret categorical codes without additional transformation. This approach allows HELIX to balance computational efficiency with semantic transparency. To avoid ambiguity and ensure full provenance tracking, each processed field is labelled using a consistent prefixing scheme that denotes the origin of the data, for example, whether it comes from a primary dataset, a static reference layer, or a dynamic time series. This guarantees clear separation between label sources while preventing naming conflicts during feature merging or model interpretation. Moreover, users can configure whether or not to retain the original raw label fields in the final output, allowing flexibility between model efficiency and auditability. Through these design principles, HELIX ensures that categorical information, often messy and inconsistent in its original form, is transformed into a clean, interpretable, and scalable format that supports robust environmental modelling across diverse data sources.

4.1.2 Spatio-temporal Scale Reconciliation

Many EO datasets and vector-based reference sources inherently exhibit misalignment across both spatial and temporal domains, complicating their direct integration into ML and DL models, AI-driven geospatial analytics, and FMs. The HELIX framework explicitly addresses these misalignments by systematically reconciling discrepancies between the reference labels, regardless of whether they are static, dynamic, or a combination of both, and the EO-derived features (e.g., satellite imagery). Practically, this is achieved by first extracting a reference grid, including coordinate reference systems (CRS), from the EO data, then identifying relevant temporal intervals corresponding to EO data availability. Spatial reconciliation involves precisely aligning vector-based labels to a defined EO raster grid using geometric alignment methods, such as affine transformations combined with appropriate resampling methods (e.g., nearest neighbour, bilinear interpolation). Temporal alignment is achieved through linear interpolation and resampling techniques, aligning the label timestamps precisely with those of the EO-derived features. This dual spatial and temporal reconciliation step enhances label quality, consistency, and adaptabil-

ity, making it suitable for a wide range of model types. For example, Figure 3.1 illustrates spatial misalignment issues where detailed terrestrial labels require spatial resampling to match coarse satellite grids. Similarly, Figure 3.2 highlights temporal discrepancies arising from rapid tidal fluctuations, demonstrating the necessity of temporal alignment techniques provided by the HELIX framework.

Spatial Component:

At the heart of this reconciliation is the generation of a canonical grid derived from a reference raster. The raster's affine transformation matrix defines how each pixel index corresponds to real-world coordinates, allowing the framework to discretize geographic space into uniformly sized cells. Formally, each cell C_{ij} is located by computing the cell's bounding box via the affine mapping:

$$(x_{\min}, y_{\max}), \quad (x_{\max}, y_{\min}) = T \cdot (\text{col}, \text{row}) \tag{4.1}$$

where T denotes the affine transformation matrix provided by the raster metadata. This procedure yields a regular spatial lattice that serves as the base layer for all subsequent projection and aggregation operations.

Since input datasets may originate from different spatial reference systems, all external geometries are reprojected into the CRS of the canonical grid. This harmonization ensures that spatial overlays and geometric operations are mathematically valid, eliminating distortions due to mismatched projections:

$$\text{CRS}_{\text{input}} \longrightarrow \text{CRS}_{\text{grid}} \tag{4.2}$$

By reprojecting all inputs into a shared CRS, the HELIX framework guarantees that topological relationships, such as containment, intersection, or adjacency, can be meaningfully and consistently interpreted.

Given the scale of EO data, where thousands of spatial features may need to be processed over large grids, computational efficiency becomes critical. Rather than computing all possible polygon-grid cell intersections directly, HELIX leverages a spatial indexing mechanism known as STRtree. This structure organizes polygon geometries hierarchically by their bounding boxes, allowing the framework to rapidly preselect likely intersecting candidates and skip irrelevant comparisons. In computational terms, this optimization

reduces complexity from $\mathcal{O}(n \cdot m)$ to approximately $\mathcal{O}(n \log m)$, where n is the number of grid cells and m the number of polygons (Figure 4.2). [1]

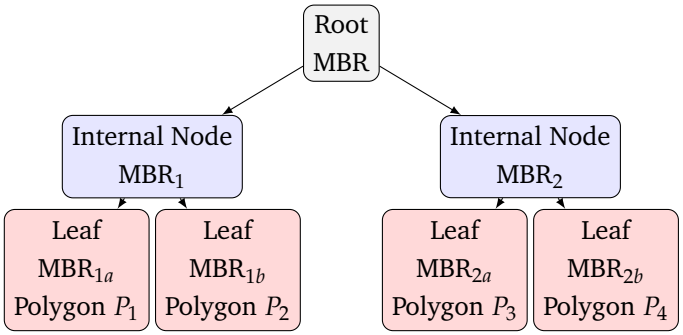


Figure 4.2.: Illustration of STRtree hierarchy: Leaf nodes store bounding boxes of polygons (MBRs), which are grouped into internal nodes. The root node covers all inputs. Queries first test against parent MBRs to prune unnecessary comparisons.

To illustrate this, a simplified example is showcased in Figure 4.3.

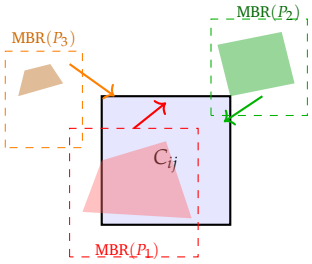


Figure 4.3.: STRtree spatial filtering: the raster cell C_{ij} (blue) queries polygon candidates by overlapping bounding boxes (dashed), reducing the need for unnecessary geometric operations.

This spatial pruning allows the HELIX to scale to tens of thousands of polygons and grid cells without brute-force computation.

Once potential intersections are identified, the framework computes the precise geometric overlap between each polygon P_k and grid cell C_{ij}. The total area of intersection is accumulated per cell:

[1] Unlike a k-d tree, which partitions data along coordinate axes and is best for point-based nearest-neighbour search, an STRtree is an R-tree variant tailored to 2D spatial indexing. It organizes spatial objects by their minimum bounding rectangles (MBRs), hierarchically grouped into a tree. When querying which polygons intersect a given raster cell, the tree first checks for bounding box overlap, a fast, approximate check, and only performs exact geometric intersection for candidates that pass. This dramatically reduces the number of expensive geometric computations.

$$A_{ij} = \sum_k \text{area}(P_k \cap C_{ij}) \tag{4.3}$$

This raw area value is subsequently normalized by the cell's area to produce a probability-like spatial coverage estimate:

$$P_{ij} = \frac{A_{ij}}{A_{\text{cell}}} \tag{4.4}$$

In parallel, a binary indicator flag is assigned to capture the presence or absence of any intersecting geometry:

$$E_{ij} = \mathbb{1}[A_{ij} > 0] \tag{4.5}$$

These three outputs, area, normalized coverage, and binary label, offer different levels of semantic richness, accommodating a variety of downstream modelling approaches including probabilistic learning and event classification.

The structured alignment of spatial data within HELIX thus enables a principled transformation of heterogeneous spatial inputs into a consistent, high-resolution analytic representation. By anchoring all labels on a common geospatial grid, reconciling coordinate systems, and employing efficient spatial filtering, the framework not only ensures geometric validity but also supports scalability and statistical rigour. This spatial layer forms the bedrock upon which temporal logic and dynamic label enrichment are later constructed.

Temporal Component:

Temporal reconciliation in HELIX builds on the spatial foundation by ensuring that each label observation is temporally aligned with the corresponding EO data in a way that respects both event semantics and the temporal nature of EO series. Real-world phenomena, like vegetation growth, storm damage, or human activity, unfold over time, and so do the observations that capture them. HELIX introduces a set of strategies to handle these temporal aspects systematically: standardization of date formats, filtering of temporally valid labels, backtracking through historical events, matching labels to EO data within a defined lag window, and optionally interpolating between sparse time steps.

To begin, timestamps are extracted from all reference layers, regardless of whether the input is a raster time series or a vector dataset. Raster-based dates are commonly encoded in file names, from which HELIX parses structured patterns using regular expressions. Vector-based annotations, on the other hand, typically contain explicit fields for event onset or validity (e.g., `Not_Before`, `Not_After`). These fields define temporal intervals during which an event or class label can be considered active. Labels are then filtered according to whether the processing date falls within the declared validity window:

$$t_{\text{start}} \leq t_{\text{processing}} \leq t_{\text{end}} \tag{4.6}$$

This condition ensures that only temporally relevant records are associated with each grid cell at a given time step.

In scenarios where a more enriched temporal context is desired, HELIX enables backtracking across previous time points. For a given target date t, the framework supports retrieval of historical labels observed at earlier dates $\{t-1, t-2, \ldots, t-n\}$ up to a configurable lag horizon. These historical snapshots are appended as lagged features, enabling models to capture persistence, recovery, or delayed effects. Importantly, this enrichment applies only to the label layer, not to EO input features, which maintains a clear separation between predictor and target domains.

Yet even when events and EO acquisitions are aligned chronologically, mismatches in their precise timestamps often persist. Atmospheric effects, sensor gaps, or annotation delays introduce inevitable asynchrony. HELIX mitigates this by implementing a lag-aware matching mechanism: for any EO observation at time t, the system seeks the closest label time t' such that:

$$|t - t'| \leq \tau \tag{4.7}$$

where τ denotes the allowed temporal offset. If no valid label is found within the threshold τ, the record is skipped. This tolerance-based alignment offers a flexible solution for pairing observational features with temporally proximate reference data without sacrificing semantic relevance.

Finally, HELIX optionally supports temporal interpolation for cases where label records are missing entirely between two known points in time. If values are available at t_1 and t_2, but not at $t \in (t_1, t_2)$, the framework interpolates intermediate values using:

$$X_t = (1-\alpha)X_{t_1} + \alpha X_{t_2}$$
$$\alpha = \frac{t - t_1}{t_2 - t_1} \tag{4.8}$$

This linear interpolation is limited to continuous fields and respects semantic constraints. Categorical labels are not interpolated numerically but handled by forward filling or nearest-neighbour methods to preserve interpretability.

4.1.3 Spatio-temporal Label Enrichment and Engineering

The HELIX framework further enhances label quality by leveraging sophisticated spatio-temporal enrichment techniques, providing significantly enriched reference datasets tailored for robust model training. Utilizing spatial and temporal dimensions simultaneously, the HELIX calculates additional derived features, including probabilities of specific classes or labels within defined spatial grid cells. For instance, it enables the calculation of label probabilities, such as the fractional area occupied by a particular class within each raster grid cell, thereby improving the interpretability and robustness of training labels.

Central to this enrichment is the configurable spatio-temporal windowing technique, in which both spatial extent (e.g., neighbourhood radius) and temporal duration (e.g., number of previous or subsequent time steps) are fully user-definable. This flexibility allows users to comprehensively integrate local context from spatially neighbouring pixels or polygons and temporal context from historical observations such as past land-use changes, trends, or seasonal cycles. Utilizing these configurable windows, the HELIX enables computation of novel and contextually rich features capturing the dynamic interplay between labels across space and time. Specifically, the HELIX supports the following processing steps:

Spatial Aggregation: Calculation of neighbourhood statistics (mean, median, mode) to ensure consistency when integrating high-resolution vector labels with lower-resolution EO raster grids, as demonstrated in Figures 3.3 and 3.4.

Temporal Windowing: Employing rolling or sliding windows that capture short-term and long-term temporal dependencies, allowing for the detection and analysis of changes and trends over defined periods. The temporal window size and spacing are user-configurable.

Helical Feature Framework: At the heart of HELIX's enrichment logic lies a transition from atomic observations to structured spatio-temporal reasoning. Each label, initially indexed at a single grid cell and timestamp, is situated within a broader analytical neighbourhood, a localized 3D volume defined jointly over space and time. Conceptually, this volume is not arbitrary: it is designed to represent the evolving environmental and semantic context around an observation, allowing the label to be interpreted not in isolation but as part of a continuous surface of change.

This contextual volume is operationalized through what HELIX terms a *helical window*, a cylindrical, symmetric sampling construct centered around a spatio-temporal point. The helix metaphor reflects the structure's dual anchoring in space (via a spatial radius r) and in time (via a temporal window width $2w$). For each target cell $C_{ij}^{(t)}$, HELIX computes statistical summaries across this window, capturing the local distribution of a given feature x:

$$\mu_{ij}^{(t)} = \frac{1}{|\mathcal{N}_{ij}^{(t)}|} \sum_{(m,n,t') \in \mathcal{N}_{ij}^{(t)}} x_{mn}^{(t')} \tag{4.9}$$

$$\sigma_{ij}^{2(t)} = \mathrm{Var}_{(m,n,t') \in \mathcal{N}_{ij}^{(t)}} \left(x_{mn}^{(t')} \right) \tag{4.10}$$

$$\mathrm{Sum}_{ij}^{(t)} = \sum_{(m,n,t') \in \mathcal{N}_{ij}^{(t)}} x_{mn}^{(t')} \tag{4.11}$$

These expressions represent, respectively, the mean, variance, and sum of values within the neighbourhood $\mathcal{N}_{ij}^{(t)}$, defined by:

$$\sqrt{(i-m)^2 + (j-n)^2} \leq r \quad \text{and} \quad t' \in [t-w,\, t+w]$$

That is, only spatial cells (m, n) within a Euclidean distance r of the target cell (i, j) and timestamps t' within the window centred at t are included. This mathematical structure encapsulates HELIX's core enrichment principle: labels are interpreted through their local history and neighbourhood, enforcing smoothness while preserving edge dynamics.

Figure 4.4 visually illustrates this helical window. A central target cell $C_{ij}^{(t)}$ is shown alongside its counterparts at earlier and later time steps $(t-1,\ t+1)$, with spatial

neighbours highlighted at each time slice. The combined structure forms a regular, multi-scale context cylinder that travels with the spatio-temporal centre of mass, providing a moving frame of reference.

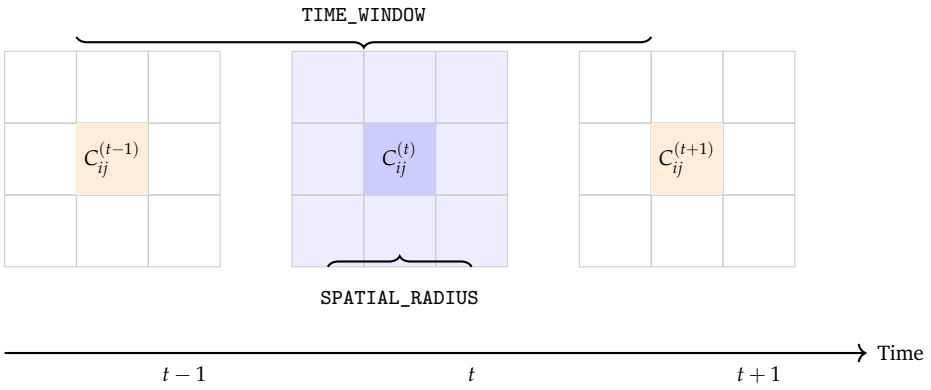


Figure 4.4.: The helical window around a grid cell $C_{ij}^{(t)}$. Here, $C_{ij}^{(t)}$ represents the target cell at position (i, j) and time step t. The helical window aggregates data from spatial neighbours within a radius at each time step across a symmetric temporal window. This produces contextual statistics for both space and time around each target cell, enabling robust modelling of local dynamics such as gradual deforestation or vegetation change.

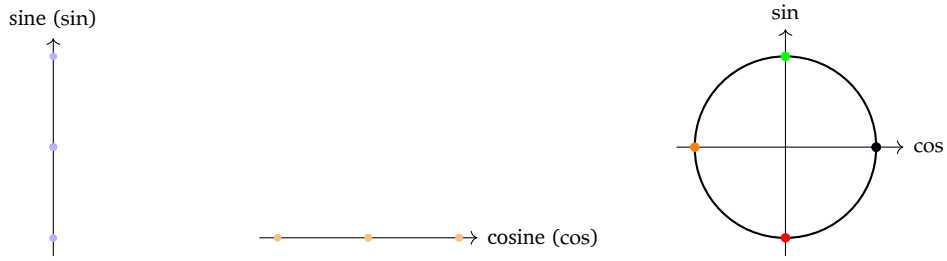
Importantly, this structure is not limited to static aggregation. Because HELIX recalculates the window per cell and per timestamp, the resulting representation is dynamic and responsive to both local structure and global trends. In effect, HELIX builds a moving frame of semantic reference, one that adapts to the scale and rhythm of the underlying environmental process. The resulting enriched features, such as the mean, variance, and cumulative sum, serve as higher-order label descriptors that enhance stability, allow for contextual regularization, and support robust supervised learning under noisy, incomplete, or coarse annotation regimes.

Fourier-Based Temporal Encoding: In many EO contexts, temporal patterns are governed by natural cycles, such as vegetation phenology, snow cover, or agricultural rotations, that follow periodic rhythms. To model such recurring structures, HELIX employs a temporal encoding strategy based on Fourier transformations of the day-of-year (DOY). Instead of treating dates as ordinal or categorical variables, HELIX maps them onto the unit circle using sine and cosine functions. This approach preserves the cyclical nature of time while enabling smooth transitions at the artificial boundary between calendar years

(e.g., between December 31st and January 1st). Mathematically, for a given periodicity P (e.g., 365 for an annual cycle), the DOY of time step t is transformed as:

$$x_t^{\sin(P)} = \sin\left(\frac{2\pi \cdot \text{DOY}t}{P}\right) \quad xt^{\cos(P)} \quad = \cos\left(\frac{2\pi \cdot \text{DOY}_t}{P}\right) \tag{4.12}$$

These encodings yield a continuous, two-dimensional representation of seasonal position. As shown in Figure 4.5, using only sine or cosine results in ambiguous mappings (e.g., both DOY 1 and 365 map to zero), while their combination produces unique angular coordinates around the unit circle.



(a) Only sin(DOY) **(b)** Only cos(DOY) **(c)** Combined sin + cos

Figure 4.5.: Comparison of Fourier encodings. The joint embedding preserves seasonal smoothness across time.

The Fourier projection thus supports temporal generalization by enabling models to infer phase-dependent relationships (e.g., spring emergence vs. autumn dieback) that recur annually. This approach is particularly effective in long-term monitoring tasks, where annual repetitions are meaningful and where models must remain invariant to calendar shifts.

Spatio-Temporal Dynamic Width Aggregation: In practice, the availability and density of labels across time can vary substantially. Temporal sparsity, whether due to infrequent surveys, missing observations, or uneven event occurrence, can degrade the quality of contextual features if treated uniformly. To address this, HELIX adapts the size of the temporal window used for aggregation in response to local data availability.

The central idea is to expand the temporal window in regions where reference labels are sparse, allowing the enrichment process to draw from more distant but still relevant

observations. Conversely, in regions with dense temporal coverage, a narrower window is maintained to preserve the precision and temporal locality of changes. This approach forms a key part of HELIX's data-aware design philosophy.

Figure 4.6 visually demonstrates this behaviour. For time step t_3 (with ample data), a narrow window suffices. For t_6 (with sparse observations), the window widens to capture distant context. These dynamic adjustments help smooth irregular sequences and avoid overfitting to limited samples.

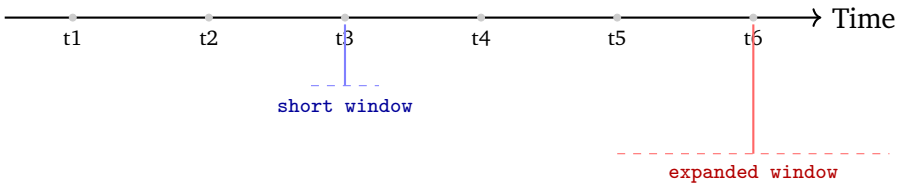


Figure 4.6.: Temporal window width adjusts to label sparsity, improving robustness in low-density intervals.

This adaptive aggregation implicitly encodes a lag-based flexibility: rather than enforcing a fixed look-back period, the system adapts to how far back it must reach to capture sufficient context. In this sense, dynamic width acts as a data-driven alternative to hard-coded lagging, generalizing across both temporally regular and irregular datasets.

Cross-Time Interaction Features: While smoothing and periodic encoding introduce contextual continuity, HELIX also incorporates mechanisms to explicitly highlight temporal change. In many applications, such as disturbance detection or land-use transitions, the primary signal of interest is not the absolute state but the deviation from previous states. To capture such dynamics, HELIX computes cross-time interaction features that compare present and past values of the same label dimension. The most canonical form of interaction is the absolute difference between time steps:

$$\Delta_{\text{pixel}} = x_{\text{main}}^{(t)} - x_{\text{historic}}^{(t-1)} \tag{4.13}$$

This simple yet informative feature quantifies the net change between consecutive time points, enabling models to learn temporal transitions, emerging disturbances, or recovery trends. By exposing both the direction and magnitude of change, HELIX enhances the

temporal expressiveness of the label space, making it especially useful for event-based learning scenarios, such as post-fire recovery or seasonal vegetation shifts.

Figure 4.7 illustrates this concept by comparing pixel values between consecutive time slices.

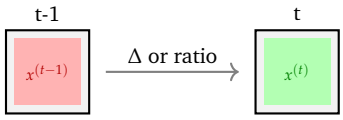


Figure 4.7.: Cross-time comparisons encode temporal transitions, critical for dynamic event understanding.

Probabilistic Labelling and Uncertainty Modelling: HELIX further extends its enrichment capabilities by providing probabilistic representations of label confidence and uncertainty, critical for scenarios involving label noise, ambiguous boundaries, or mixed land cover types.

Fractional Probability Estimation: For any labelled polygon L intersecting a grid cell (i, j), HELIX computes a spatial class probability $P_{ij}^{(L)}$ based on the proportion of the cell covered:

$$P_{ij}^{(L)} = \frac{A_{ij}^{(L)}}{A_{\text{cell}}} \tag{4.14}$$

where $A_{ij}^{(L)}$ is the area of label L overlapping with cell (i, j), and A_{cell} is the full cell area. These probabilities are especially useful in regions where multiple labels overlap or where partial coverage implies class uncertainty. Figure 4.8 illustrates this concept across multiple cells and overlapping labelled polygons. In the centre cell, two polygons intersect, and their fractional contributions can be separately computed to support multi-class or soft-label scenarios.

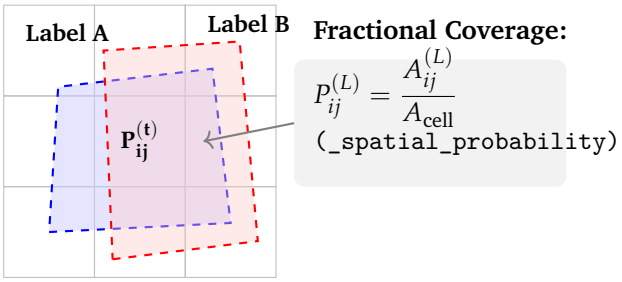


Figure 4.8.: Fractional spatial probability estimation. Two polygonal labels (Label A and Label B) overlap a shared grid. For each cell, the fraction $P_{ij}^{(L)}$ is computed by dividing the intersected area $A_{ij}^{(L)}$ by the cell area A_{cell}, allowing soft, probabilistic label assignment.

Neighbourhood-Based Uncertainty Estimation: To complement the probability field, HELIX estimates local uncertainty by measuring variance across a configurable space-time neighbourhood . The local mean and variance for any feature x are computed as:

$$\mu_{ij}^{(t)}(x) = \frac{1}{|\mathcal{N}_{ij}^{(t)}|} \sum_{(m,n,t') \in \mathcal{N}_{ij}^{(t)}} x_{mn}^{(t')} \tag{4.15}$$

$$\text{Var}_{ij}^{(t)}(x) = \frac{1}{|\mathcal{N}_{ij}^{(t)}|} \sum_{(m,n,t') \in \mathcal{N}_{ij}^{(t)}} \left(x_{mn}^{(t')} - \mu_{ij}^{(t)}(x) \right)^2 \tag{4.16}$$

This neighbourhood-aware variance highlights unstable regions, e.g., class boundaries or transient artifacts, enabling downstream tasks like label refinement, uncertainty filtering, or confidence-weighted training. Figure 4.9 illustrates how uncertainty is calculated across a cell's spatial or spatio-temporal neighbourhood, with stronger shading indicating higher local variance.

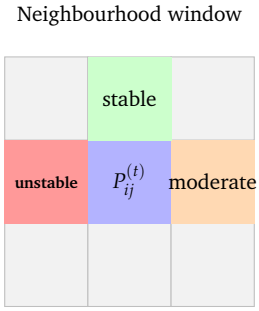


Figure 4.9.: Illustration of local variance estimation. The variance value computed for the centre cell $P_{ij}^{(t)}$ reflects the degree of heterogeneity within its surrounding neighbourhood window $\mathcal{N}_{ij}^{(t)}$. Shading indicates the underlying spatial distribution of neighbour values contributing to the calculation: red for highly variable (unstable) neighbors, green for homogeneous (stable) areas.

4.1.4 Structured Learning Targets

Soft Thresholding and Reclassification Logic: HELIX accommodates the inherent fuzziness of real-world labels through a dynamic soft thresholding mechanism. Instead of applying a fixed binary cut-off to probabilistic label scores (e.g., 0.5), HELIX supports adaptive thresholding schemes that account for both class imbalance and contextual uncertainty. For instance, a land cover class with known boundary ambiguity (e.g., wetlands) may require a lower threshold for positive inclusion, while highly distinct classes (e.g., water vs. urban) can be assigned more conservative thresholds. Soft thresholding is often used in tandem with the neighbourhood-based uncertainty metrics described earlier. For regions with high local variance, thresholds can be relaxed or deferred, whereas in low-variance zones, stricter reclassification rules may apply. This process allows HELIX to perform context-aware reclassification that reflects both the probability and stability of label assignments, rather than relying on brittle decision boundaries.

Fuzzy Boundary Modelling: To further enhance semantic realism, HELIX supports the generation of fuzzy boundaries where labels transition gradually rather than abruptly. This is achieved by interpreting probability fields not as deterministic masks but as soft surfaces. For example, a transition from forest to shrubland may span multiple cells with intermediate probability values, producing a smooth gradient rather than a binary edge. These fuzzy boundaries are especially useful in ecological applications, where mixed

vegetation types, succession states, or disturbance gradients are common. HELIX enables downstream models to consume these gradients directly, preserving uncertainty and promoting robustness. In classification workflows, fuzzy zones can be masked, weighted, or treated as distinct transitional classes depending on analytical needs.

Integration into Model Training and Outputs: HELIX's enriched labels, whether probabilistic, smoothed, or uncertainty-weighted, are designed to be fed directly into machine learning pipelines. In training, these labels can be used as:

- **Soft targets** for probabilistic classification tasks
- **Sample weights** to emphasize stable and high-confidence regions
- **Masking layers** to exclude or down-weight ambiguous zones

For output evaluation, HELIX's probabilistic representations support more nuanced performance metrics, such as Expected Calibration Error (ECE), fuzzy accuracy, or class-specific uncertainty curves. This allows models to be evaluated not only on hard correctness, but also on how well they handle ambiguity and noise. In addition, HELIX's structured and time-aware enrichment enables flexible downstream modelling across a wide range of architectures. Importantly, HELIX decouples the need for complex temporal models such as LSTMs or transformers by explicitly engineering temporal dynamics, e.g., lag features, change indicators, seasonality embeddings, into tabular representations. This means that conventional ML models such as RF, gradient boosting machines, or shallow NN can effectively utilize temporally structured EO data without needing specialized recurrent or attention-based mechanisms. By doing so, HELIX broadens the accessibility of spatio-temporal modelling to non-deep-learning workflows, reducing computational overhead while preserving temporal expressiveness and interpretability. This alignment makes it easier for models to generalize, reduces overfitting to label artifacts, and ensures that downstream predictions remain interpretable and actionable.

4.1.5 Operational Design for Scalability and Integration

HELIX is engineered as a modular and scalable system for EO label enrichment, with particular emphasis on practical usability across large datasets and long temporal spans. It bridges the analytical pipeline between raw EO inputs and ML-ready outputs by offering

both semantic depth and technical efficiency. At the core of its integration logic is the use of the Parquet format, a compressed, columnar file structure optimized for fast reading and selective querying. Each HELIX output represents a single time step and is saved using a consistent naming convention that facilitates automated indexing. These Parquet files can be rapidly loaded and manipulated via tools such as `pandas`, `Dask`, or `PyArrow`, making them accessible both in local research environments and scalable cloud workflows. Each column in the output grid is tagged with informative suffixes to encode its role.

This structured naming convention simplifies downstream operations such as feature selection, normalization, and masking. HELIX's architecture is also tailored for ML integration. The enriched tables serve as direct inputs for popular ML frameworks like `scikit-learn`, `XGBoost`, `PyTorch`, and `TensorFlow`. Practitioners can filter feature columns via pattern matching or schema-aware logic, select valid training targets, and exclude ambiguous regions using generated uncertainty scores. For spatial visualization or validation, any output column can be rasterized back into GeoTIFF using `rasterio`, supporting interpretability and diagnostics.

Scalability is not merely a design goal, it is an operational imperative. HELIX incorporates a number of strategies to handle large volumes of EO data efficiently:

First, spatial intersection operations are accelerated using `STRtree` spatial indexing. This avoids expensive brute-force comparisons and allows the system to quickly identify relevant geometries for each grid cell. Next, all major processing tasks, including grid construction, enrichment, Fourier encoding, and aggregation, are parallelized via `joblib`, with user-configurable control over the number of parallel workers.

For numerically intensive tasks, HELIX leverages `Numba`, a just-in-time (JIT) compiler that translates Python functions into optimized machine code. This dramatically boosts performance for array operations such as rolling statistics and window-based feature computation, making HELIX viable for terabyte-scale workloads.

Memory management is also a first-class concern. Instead of loading the entire dataset into RAM, HELIX operates in adaptive batches. Chunk size and memory thresholds are exposed as configuration parameters, allowing users to fine-tune performance for their specific computing environment. When enabled, intermediate steps (e.g., interpolated rasters, enriched grids) are cached to disk, providing resilience against job failures and enabling checkpoint-based processing.

Finally, HELIX automatically adapts to the spatial resolution and geographic extent of the input data. This is essential when working with multi-source EO datasets that vary in pixel size, from coarse-resolution MODIS (250 m) to high-resolution Sentinel-2 imagery (10 m). Rather than requiring manual tuning, HELIX derives the spatial grid dimensions directly from raster metadata and adjusts memory allocation, enrichment windows, and processing chunk sizes accordingly. This resolution-aware behaviour ensures that the same logic and algorithms can be reused across scales, preserving methodological consistency while optimizing computational efficiency. It also facilitates comparative studies across sensors or sites without changing the enrichment pipeline.

Together, these design choices make HELIX both robust and agile: capable of handling large EO workloads while remaining modular, interpretable, and integration-ready.

4.2 Future Directions

Future research directions should not only address the challenges discussed in previous sections but also pioneer new approaches that integrate labels and EO-derived features in a unified framework. In current workflows, labels and EO data are often developed and processed separately, which may limit the overall performance and adaptability of ML and DL models. Future systems should enable joint optimizations of labels and features in which both the ground truth and EO measurements are iteratively refined and harmonized through other integrated preprocessing pipelines.

However, the independent validation of data is essential for ensuring the robustness and credibility of EO-based ML/DL models [202, 224]. Regarding such methodologies, it is of utmost importance to implement safeguards that maintain statistical independence. Without such safeguards, label optimization risks being overly influenced by feature distributions, leading to biased models that lack external generalizability. Several approaches can help to mitigate these risks while allowing for improved label–feature consistency. One potential direction is the use of regularized loss functions [358] to enforce stability in label optimization. Loss function constraints can be designed to ensure that labels remain homogeneous across time and space, preventing abrupt shifts that may be artifacts of sensor inconsistencies rather than actual environmental changes. Multitask loss functions [131] could further help to balance label fidelity with other predictive objectives, allowing models to learn from additional supervision while maintaining independent label

structures. Additionally, uncertainty-aware loss formulations [174] can be used to down-weight highly uncertain labels, reducing the risk of unreliable training data distorting model predictions. In addition to loss function constraints, hybrid validation strategies offer another pathway to preserving independent validation while refining label-feature coherence [224]. Instead of allowing labels to be iteratively updated without external benchmarks, structured validation frameworks should incorporate holdout-based label validation, in which a subset of reference data remains untouched to act as an independent assessment benchmark. Similarly, domain-specific cross-validation approaches can be applied, ensuring that models are tested on geographically or temporally distinct regions rather than being evaluated solely within the training domain. Multi-source validation, in which generated labels are compared against alternative independent datasets such as ground truth surveys, crowdsourced data, or multi-sensor observations, can further help to prevent label optimization from reinforcing model biases.

Another promising avenue is the application of probabilistic [77] and Bayesian methods in label refinement. Unlike fixed categorical labels, probabilistic frameworks allow for the modelling of uncertainty in reference data, ensuring that transitions between classes or temporal variations are captured without forcing deterministic label assignments. Bayesian inference enables iterative label updates while incorporating independent prior knowledge, preventing labels from drifting toward overfitting feature distributions. Similarly, soft-labelling techniques [284] can assign probability distributions instead of discrete class assignments, allowing models to handle ambiguous or transitional regions (e.g., vegetation shifts, land cover change dynamics) with greater flexibility. Post hoc label calibration [39] offers another strategy for maintaining label independence while benefiting from refined representations. Residual label correction can help to detect systematic biases in label assignments after model training, ensuring that errors linked to specific geographic or environmental conditions do not propagate into future predictions. Additionally, contrastive label alignment techniques, in which labels generated under different modelling conditions are compared, can reveal inconsistencies that might otherwise go unnoticed. These methods are particularly useful in remote sensing applications where multiple sensors provide different perspectives on the same environmental variable, enabling a reconciliation process that respects external validation sources.

A further consideration for future research is the role of explainability and interpretability in label optimization [142]. As dynamically generated labels increasingly originate from prior ML/DL models rather than from direct human annotation, ensuring transparency in their construction is essential for scientific credibility in EO. In this context, feature

explainability techniques that are typically used in feature engineering could also inform label engineering, particularly in cases where a prior model generates the reference data for subsequent learning processes. For instance, feature attribution methods such as Shapley Additive Explanations (SHAP) and Gradient-weighted Class Activation Mapping (Grad-CAM) can be used to assess whether label refinements capture meaningful geophysical signals rather than overfitting to latent model biases. Similarly, integrating Explainable AI (XAI) into label validation workflows could provide insights into whether dynamically optimized labels retain their conceptual and physical relevance, ensuring that feature-label dependencies remain interpretable and scientifically grounded.

In line with this vision, the HELIX framework is especially well positioned to incorporate such logic. It can be extended to support consensus-based reasoning, dynamic soft thresholding and cross-validation capabilities, as described below:

Agreement-Based Confidence Adjustment: As the HELIX integrates an increasing number of label sources, such as manual annotations, automated detections, and historic reference layers, conflicting or overlapping labels may arise for the same pixel and time step. Future versions of the HELIX could incorporate agreement-based logic to derive confidence scores from such overlaps.

For categorical labels, this may involve majority voting or entropy-based scoring to reflect consensus strength. For continuous or probabilistic labels, inter-source agreement could be used to weight predictions based on variance or signal consistency. In both cases, regions with strong agreement among sources would be assigned higher label confidence, while conflicting inputs would trigger lower certainty scores (Figure 4.10).

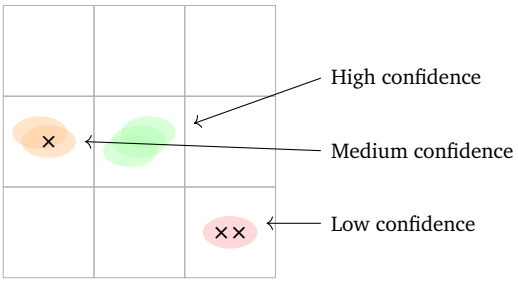


Figure 4.10.: Visual example of how confidence scores could be derived from label agreement. Multiple overlapping shapes represent input labels from different sources. Higher overlap implies higher agreement and thus higher label confidence.

Such confidence-aware estimates could then guide downstream applications to treat uncertain regions more cautiously, e.g., applying label smoothing, skipping low-confidence regions in training, or flagging areas for human review.

Conflict Detection and Resolution: As within the HELIX can process increasingly heterogeneous datasets, conflicting labels will become inevitable. Future enhancements could include built-in mechanisms to:

- Flag binary disagreements (e.g., `has_conflict` = 1);
- Compute entropy metrics to quantify categorical dispersion;
- Log timestamped conflict summaries for versioning and audit trails.

Such capabilities lay the foundation for automated resolution pipelines, including dynamic source weighting, expert correction workflows, or semi-supervised refinement strategies.

Dynamic Soft Thresholding: Another promising extension of the HELIX framework involves adaptive thresholding strategies that transform probabilistic class scores into categorical labels. Currently, label binarization is based on fixed, globally applied thresholds. While simple and reproducible, this approach may misrepresent label boundaries in regions with temporal variation or low-confidence classifications. Future versions of the HELIX could implement dynamic soft thresholding, in which thresholds are adapted based on:

- Local feature distributions (e.g., mean and variance across a spatial neighbourhood);
- Historical class presence or auxiliary prior probabilities;
- Label uncertainty or inter-source agreement confidence.

Such a logic would support more flexible labelling pipelines in dynamic environments, such as seasonal snow zones, vegetation boundaries, or disturbance regimes. For instance, the classification of glacier zones may require temporally aware thresholds that reflect year-to-year shifts in ice coverage. These adaptive schemes could improve the resilience and realism of binary label assignments derived from soft probability fields, especially in regions of geophysical ambiguity.

Multi-Temporal Cross-Validation Logic: To improve generalization and robustness, the HELIX can be extended to support automatic cross-validation strategies across space, time, and source. These may include:

- Cross-temporal validation (e.g., training on 2017–2020, validating on 2021);
- Spatial out-of-fold testing using distinct ecoregions or land use types;
- Source-wise benchmarking to detect systemic bias or spatial misalignment.

This functionality would transform the HELIX framework into not only a preprocessing engine, but a full-fledged reference data validation and assurance framework. As EO applications increasingly rely on multi-source, multi-year training corpora, these consensus-based capabilities will be critical for scaling trustworthy, transparent, and replicable remote sensing models.

Future research on label optimization must prioritize statistical independence and external validation as core principles. While integrating labels with EO-derived features offers potential advantages in model consistency, it is imperative that optimization strategies do not undermine the integrity of independent reference data. By employing a combination of loss function constraints, hybrid validation frameworks, probabilistic techniques, post hoc refinements, and explainable AI approaches, researchers can ensure that future labelling methods enhance model generalizability and maintain credibility in EO-based ML/DL applications. These advancements will be essential as automated dynamic labelling becomes more prevalent in large-scale geospatial modelling.

Currently, EO data is typically available in gridded format; however, an increasing amount of EO information is being captured as segmented high-definition data at fine scales (e.g., individual trees). Integrating such detailed EO data with corresponding labels is a pivotal future step in enhancing model performance and adaptability. Automated label verification involves the use of ensemble-based validation techniques to detect and correct inconsistent labels. By leveraging multiple model outputs and cross-validating with independent data sources, such techniques can significantly improve the quality and reliability of the reference data. This is crucial for ensuring that model training is based on accurate, consistent ground truth. The development of self-adapting labelling systems represents a promising research direction. Algorithms that dynamically adjust labels based on feedback from real-world observations can continuously refine the training data. Techniques such as self-supervised learning and domain adaptation are key to achieving this dynamic refinement over time. Such systems could both update labels in response

to evolving environmental conditions and help in identifying and correcting systematic errors. A major future challenge is the current separation between label generation and EO data feature extraction. Future approaches should aim to integrate these processes into a single co-adaptive framework. By processing labels and EO data simultaneously in a unified pipeline, it would be possible to accomplish several goals:

Enhanced data consistency: Joint processing would allow for simultaneous correction of spatial and temporal misalignments, ensuring that both labels and EO features are well-aligned.

Improved label quality: Iterative refinement based on the combined insights from both data types could lead to more accurate and representative labels.

Increased adaptability: A unified system could more readily adapt to changes in the environment, dynamically updating both labels and features in near real-time for future onboard processing.

Such an approach would represent a paradigm shift in which preprocessing not only prepares data for training but also continuously improves the quality of the reference data based on the EO observations. By tackling these challenges, dynamic labelling and joint data development can unlock the full potential of ML and DL in EO applications. Enhanced real-time environmental monitoring, improved land use prediction, and more effective disaster response capabilities are just a few of the potential benefits. The transition from static to dynamic labels and from separately processed labels and features to a joint development approach is not merely a technical evolution but a necessity for building more responsive and adaptable geospatial modelling systems. As strongly advocated for in previous studies, the future of EO data processing lies in creating robust, scalable, and integrated frameworks that not only address current challenges but also pave the way for more advanced and adaptive ML/DL applications in environmental monitoring. However, as emphasized by [292], the distinct nature of EO data necessitates frameworks that are tailored to its domain-specific complexities, including sensor characteristics, spatio-temporal dependencies, and physical data constraints. Unlike general computer vision applications, EO label engineering must integrate a deep understanding of remote sensing principles to ensure that dynamically generated labels remain scientifically valid and physically meaningful. Therefore, selecting analytical and geoprocessing frameworks for label optimization must prioritize EO-specific considerations in order to maintain the integrity and interpretability of reference labels.

While the HELIX framework currently processes labels independently of EO-derived features, its modular architecture is inherently extensible toward a co-adaptive pipeline. Future implementations could support joint optimization of features and labels, integrating model feedback, probabilistic confidence adjustment, and ensemble-based label verification. This would enable dynamic refinement of labels over time, accommodate high-resolution segmented EO inputs, and allow for continuous learning from evolving environmental signals. Such capabilities position the HELIX as a foundational component for scalable, semi-supervised, and self-correcting EO model ecosystems, where preprocessing evolves from a static stage into an adaptive, intelligence-driven process.

Conclusions

The HELIX framework constitutes a principled infrastructure for transforming fragmented, heterogeneous EO labels into a harmonized, temporally-aware, and spatially-consistent analytic format. Through the joint resolution of spatial and temporal scale mismatches, HELIX offers more than just preprocessing; it implements a full semantic reconciliation across dynamic EO environments.

By anchoring all label data on a canonical, multi-resolution spatio-temporal grid and enriching them via statistically grounded neighbourhood functions, HELIX ensures that both persistent features (e.g., land cover classes) and ephemeral events (e.g., disturbances, phenological transitions) are represented in context. HELIX's design is inherently multi-scale, not only in spatial geometry, but also in semantics and temporal representation. By supporting label interpretation across levels (e.g., pointwise, neighbourhood-aggregated, probabilistic), HELIX bridges the gap between fine-grained annotations and the broader generalization required for scalable modelling, while providing essential contextual and semantic depth.

Importantly, HELIX introduces a helical abstraction that fuses spatial and temporal proximity into a unified analytical volume. This supports downstream models in leveraging not only what is observed at a given time and place, but also what surrounds it in both space and history. The resulting enriched label structures act as dynamic carriers of environmental meaning, capable of encoding uncertainty, highlighting transitions, and preserving periodicity via Fourier transforms.

From a methodological standpoint, HELIX serves as a form of intelligent data augmentation. It extracts latent structure from label data, generates synthetic yet semantically grounded features, and enables conventional ML models, such as RF, XGBoost, or even tabular NN, to exploit temporal information without requiring temporally recursive architectures like RNN or transformers. This expands the range of accessible spatio-temporal modelling options, lowering computational barriers while maintaining interpretability and statistical rigour.

HELIX is not just a framework but a general-purpose layer of abstraction for EO analytics. It converts label noise into structure, sparsity into smoothness, and categorical chaos into probabilistic clarity, transforming raw environmental labels into robust, context-rich learning targets. As such, it lays the groundwork for a new generation of spatio-temporal EO models that are both methodologically sound and operationally scalable.

5 Temporal Dynamics in EO Feature Engineering

> *Information consists of differences that make a difference.*
>
> — **Gregory Bateson**
> Anthropologist, Systems Theorist

This chapter includes elements from the following peer-reviewed publication:

Simone Aigner, Sarah Hauser, and Andreas Schmitt. *Pattern-Based Sinkhole Detection in Arid Zones Using Open Satellite Imagery: A Case Study Within Kazakhstan in 2023*. *Sensors*, 25(3), 2025, Article 798. DOI:10.3390/s25030798

It is cited as [6] and is marked with a grey line.

Author Contribution: Sarah Hauser co-led the conceptualization and methodology development, establishment of the analysis pipeline, and contributed significantly to the investigation, supervision, and manuscript writing. She also played a key role in shaping the experimental framework and remote sensing application.

Building on the temporal fusion foundations introduced in Section 2.1.4, the following chapter systematically explores the practical implications of distinct temporal fusion configurations. By leveraging the principles of temporal taxonomy, it demonstrates how feature-level integration, via hypercomplex methods, can be strategically optimized. This leads to the development of the Combined Doline Vegetation Index (CDVI), a novel multi-sensor, seasonally-informed EO index designed to capture the interaction between vegetation dynamics and geomorphological structures in arid karst environments. The methodological context for sinkhole and vegetation detection, from traditional

approaches to remote sensing-based methods, is presented in Section 1.2.1. The ecological and geophysical setting of the study area in southwestern Kazakhstan is described in Section 1.2.1, while Section 1.2.1 details the reference datasets used.

Motivated by the preceding methodological and ecological considerations, this chapter formulates several guiding research questions aimed at systematically evaluating the impact of temporal fusion strategies on land surface classification in arid and semi-arid environments:

RQ1: *How does the ecological timing of multi-sensor data acquisitions influence the separability of land cover classes in arid regions?*

RQ2: *Can seasonally disparate fusion of SAR and optical EO data outperform temporally aligned fusion in detecting subtle geomorphological features?*

RQ3: *What temporal configurations maximize the statistical discriminability of sinkholes, vegetation, and bare ground in semi-arid landscapes?*

RQ4: *To what extent can a temporal taxonomy inform the design of data fusion pipelines beyond strict temporal coherence?*

RQ5: *Can a tailored index, based on seasonally informed, multi-sensor data fusion, be developed to distinguish sinkholes from spectrally and structurally similar land cover types, and how effective is it in arid environments?*

5.1 Comparative Evaluation of Temporal Fusion Settings

To determine the most effective temporal configuration for hypercomplex data fusion in arid environments, multiple Sentinel-1 and Sentinel-2 acquisition date combinations were tested. These combinations reflect a range of temporal alignments, including intra-seasonal, inter-seasonal, and cross-seasonal pairings. The rationale for this investigation was to identify the optimal synergy between structural (SAR) and spectral (optical) information under varying environmental conditions, with the goal of enhancing class separability, particularly between sinkholes, vegetation, and background surfaces.

5.1.1 Materials

All EO feature datasets originate from freely available ESA missions. Sentinel-1 acquisitions were obtained in Interferometric Wide (IW) swath mode and processed as Single Look Complex (SLC) data, which preserves the full amplitude and phase information necessary for radar-based feature extraction, especially within the HCB framework. Sentinel-1 captures C-band SAR data in both VV (co-polarised) and VH (cross-polarised) modes, which are responsive to structural features around 5 cm in scale. VV polarisation generally produces the highest backscatter intensity over land surfaces, while VH polarisation is more sensitive to volume scattering, making it particularly effective in detecting vegetation-related structures [6].

Optical data was acquired using Sentinel-2 Level-2A products, which provide Bottom-Of-Atmosphere (BOA) surface reflectance after atmospheric correction. All optical scenes used in this study exhibited minimal cloud contamination ($\leq$ 0.37%). The specific acquisition dates and configurations for each fusion scenario are detailed in Table 5.1.

Ground-truth data for sinkhole identification and land cover classification were obtained from the Svevind Energy Group [319] and high-resolution World Imagery [100]. The dataset includes GNSS-located sinkholes, georeferenced imagery, and manually delineated vegetation and bare surface classes. These references form the basis for evaluating detection performance. Section 1.2.1 provides full details on the reference datasets used.

5.1.2 Methods

The selected temporal windows were strategically designed to match seasonal dynamics specific to the Mangystau region of southern Kazakhstan. This area is characterized by pronounced seasonal contrasts: summers (June–August) are arid and marked by senescent vegetation, while winters and early springs (February–March) represent the wet season, when vegetation regrowth and moisture availability are at their highest. This ecological variability allows fusion configurations to explore not only temporal coherence but also seasonal complementarity. Table 5.1 summarizes the six tested fusion scenarios, detailing acquisition dates, inferred seasons, and Sentinel-2 cloud cover.

Table 5.1.: Overview of the six temporal fusion configurations tested, including acquisition dates with respective seasons and cloud cover for Sentinel-2.

Fusion Scenario	Sentinel-1	Sentinel-2
Intra-Seasonal Fusion I	06 Aug 2023 *(Dry)*	07 Aug 2023 *(Dry, 0% cloud)*
Intra-Seasonal Fusion II	07 Feb 2023 *(Wet)*	13 Feb 2023 *(Wet, 0.37% cloud)*
Cross-Seasonal I	06 Aug 2023 *(Dry)*	13 Feb 2023 *(Wet, 0.37% cloud)*
Cross-Seasonal II	06 Aug 2023 *(Dry)*	30 Mar 2023 *(Wet, 0% cloud)*
Inverse Seasonal Fusion	07 Feb 2023 *(Wet)*	07 Aug 2023 *(Dry, 0% cloud)*
Transition Fusion	07 Feb 2023 *(Wet)*	30 Mar 2023 *(Wet, 0% cloud)*

In this workflow, Sentinel-2 reflectance values are initially normalised to balance the spectral channels by reducing the influence of the NIR band. These normalised values are then converted into Kennaugh-like elements (see Section 2.2) via linear combinations [286], resulting in one total reflectance element and three spectral elements. The complex SAR images were preprocessed based on the freely available framework [148], which is based on the Multi-SAR processor [38]. This process calculates four Kennaugh elements (see Section 2.2) k_0, k_1, k_5, and k_8, preserving the complete polarimetric information [288]. These elements, representing intensities and intensity differences, were geocoded to the respective geographic zone. Final normalisation ensures consistent data ranges and allows the efficient storage of UInt16 digital numbers, analogous to the Sentinel-2 data. The datasets are subsequently fused using the linear HCB approach (see Section 2.2), producing a fused and normalised dataset consisting of one total intensity element ($K_{\text{fused},0}$) and seven spectral/polarimetric elements ($K_{\text{fused},1-7}$), as detailed in [289].

5.1.3 Results

To quantify the separability of key land cover classes (sinkholes, vegetation, and bare surfaces), each fusion configuration underwent a comprehensive battery of statistical tests.

These included parametric and non-parametric analyses: t-Tests for mean separability, Pearson's r for linear correlation, Kendall's Tau and Spearman's Rho for rank correlation, and ANOVA F-statistics to assess between-class variance, see 5.1.



Figure 5.1.: Statistical test results (t-Test, Pearson, Kendall, Spearman, ANOVA) for all fusion configurations across bands $K_{\text{fused},0}$ to $K_{\text{fused},7}$. CDVI (SAR August + Optical March) yields consistently superior performance across metrics, especially in bands $K_{\text{fused},2}$, $K_{\text{fused},4}$, $K_{\text{fused},5}$, and $K_{\text{fused},6}$.

Intra-Seasonal Fusion I: This configuration combined SAR and optical data from August 2023, during the arid dry season. Despite relatively uniform surface conditions and senescent vegetation, the fusion achieved moderately strong separability. High ANOVA values for $K_{\text{fused},4}$ ($F = 23.33$) and substantial rank correlations for $K_{\text{fused},5}$ and $K_{\text{fused},6}$ (e.g., $\rho = -0.81$) indicate that SAR-derived structural features, particularly those emphasizing co-/cross-polarization contrasts and phase coherence, were effective in distinguishing geomorphological depressions from surrounding terrain. Although NIR and Red bands within $K_{\text{fused},0}$ and $K_{\text{fused},6}$ offered some spectral separation, the lack of active biomass limited vegetation-related contrast. Nevertheless, this setting demonstrated the value of structural information under low-reflectance optical conditions.

Intra-Seasonal Fusion II: Acquired in early February 2023, this fusion scenario represents a temporally coherent but environmentally noisy wet-season configuration. Across most bands, statistical separability was weak. While $K_{\text{fused},0}$ and $K_{\text{fused},7}$ reached moderate ANOVA values ($F > 11$), correlation metrics remained low (e.g., $\rho = 0.43$ in $K_{\text{fused},0}$, $\rho = 0.37$ in $K_{\text{fused},7}$. High soil moisture likely impaired SAR coherence, while early-stage vegetation lacked strong spectral contrast. The limited variability in both structural and spectral domains made it difficult to resolve fine-scale features such as dolines. This highlights that temporal alignment alone does not guarantee fusion efficacy when seasonal expressiveness is low.

Cross-Seasonal Fusion I: This inter-seasonal setup fused August SAR with February optical data. Statistical gains were apparent, especially in SAR-optical combinations like $K_{\text{fused},4}$, $K_{\text{fused},5}$, and $K_{\text{fused},6}$, where non-parametric metrics such as Spearman's $\rho = -0.75$ in $K_{\text{fused},2}$ and ANOVA values up to $F = 26.82$ (in $K_{\text{fused},7}$) marked a clear performance jump over intra-seasonal fusions. This improvement can be attributed to the complementary nature of structural features captured during arid, low-vegetation conditions and the spectral softening of early-season regrowth. The SAR signal remained clean due to low moisture, while optical NIR and Red bands began reflecting vegetation recovery.

Cross-Seasonal Fusion II: This scenario, August SAR fused with late March optical data, achieved the strongest results across all tests. Nearly all bands showed excellent performance: for example, $K_{\text{fused},4}$ reached an ANOVA peak ($F = 30.69$), and Spearman's ρ exceeded 0.77 in $K_{\text{fused},2}$ and $K_{\text{fused},5}$. This configuration maximized seasonal contrast: SAR backscatter was unaffected by moisture, capturing fine-scale surface roughness, while optical reflectance, especially NIR, was enriched by mature vegetative cover. The success of this pairing illustrates the power of fusing temporally disparate yet ecologically

complementary datasets. As such, it forms the conceptual and empirical basis for the CDVI, detailed in the following section.

Inverse Seasonal Fusion: Pairing February SAR with August optical data resulted in poor separability across most bands. For instance, $K_{\text{fused},2}$ had negligible correlation ($\rho = -0.14$) and low variance contrast ($F = 1.22$). The wet-season SAR acquisition introduced significant noise due to moisture-related decorrelation, while the August optical data offered weak spectral differentiation due to vegetation senescence. This fusion setting thus suffers from a lack of temporal and ecological alignment, underscoring that unbalanced seasonal pairing, especially when SAR is acquired under wet conditions, can suppress both structural and spectral discriminability.

Transition Fusion: This fusion scenario spanned SAR from February and optical data from March, both within the wet season. While it slightly outperformed Intra-Seasonal Fusion II, it remained inconsistent. Some bands such as $K_{\text{fused},4}$ and $K_{\text{fused},7}$ showed moderate statistical relevance (e.g., $\rho = 0.51$, $F = 14.28$), but others delivered minimal contrast. Vegetation likely had not reached peak chlorophyll content in early March, and SAR backscatter remained affected by residual soil moisture. As a result, the dataset only weakly differentiated between geomorphic depressions and vegetated surroundings. This indicates that partial seasonal shifts may not suffice to enhance separability when sensor modalities are not optimally phased.

Among all configurations, **Cross-Seasonal Fusion II** emerged as the most effective. It consistently delivered the strongest performance across statistical tests, especially in bands representing Red and NIR contrast combined with polarimetric SAR elements $K_{\text{fused},2}$, $K_{\text{fused},4}$, $K_{\text{fused},5}$, $K_{\text{fused},06}$. These elements leverage the stable, high-texture radar signal from dry-season acquisitions and the vibrant, high-contrast spectral information of spring vegetation.

In contrast, fusions involving wet-season SAR, **Intra-Seasonal Fusion II**, **Inverse Seasonal Fusion**, and **Transition Fusion**, performed poorly. Elevated soil moisture introduces decorrelation in the radar signal, masking the subtle topographic or surface roughness cues critical for doline detection. Simultaneously, non-optimal vegetation states limit spectral differentiation. These findings reinforce the idea that fusion success arises not from sensor simultaneity but from ecological complementarity: structurally informative SAR under dry conditions, paired with spectrally expressive optical data during periods of vegetation vitality.

5.1.4 Discussion

The discriminative power of fused Kennaugh elements is not merely a function of sensor modality or acquisition simultaneity, but is predominantly driven by *ecological complementarity*. Among all evaluated configurations, the **Cross-Seasonal Fusion II** setting, integrating dry-season Sentinel-1 SAR (August) with early spring Sentinel-2 optical data (March), consistently activated the most informative Kennaugh combinations. In particular, fused bands such as $K_{\text{fused},2}$, $K_{\text{fused},4}$, $K_{\text{fused},5}$ and $K_{\text{fused},06}$ demonstrated superior separability across parametric (ANOVA) and non-parametric (Spearman's ρ, Kendall's τ) metrics. This fusion setting effectively harnessed the stable radar backscatter conditions of the dry season together with the phenological richness of the wet season's spectral reflectance, particularly in the NIR band associated with chlorophyll abundance. The result was maximized contrast between vegetated and non-vegetated surfaces, as well as enhanced delineation of geomorphological depressions.

From a phenological standpoint, optical imagery acquired in March corresponds to a critical phase of early vegetative regrowth in the Mangystau region of southern Kazakhstan. As a semi-arid steppe environment, this region experiences its highest moisture availability and photosynthetic activity during late winter and early spring. Vegetation begins to recover rapidly following winter precipitation, producing strong reflectance in the near-infrared (B8) and red-edge portions of the spectrum. In contrast, **February** acquisitions, though also situated within the wet season, typically capture vegetation in an earlier growth stage, characterized by sparse canopy development and lower chlorophyll content, resulting in flatter spectral signatures.

The advantages of dry-season SAR acquisitions are equally important. August SAR data are characterized by low soil moisture and senescent vegetation, which minimizes dielectric variability and improves phase stability. This leads to cleaner polarimetric decomposition and more coherent structural information in radar-derived Kennaugh elements such as $K_{\text{fused},1}$, $K_{\text{fused},5}$, and $K_{\text{fused},6}$. Conversely, SAR data acquired in February amid higher subsurface moisture suffer from temporal decorrelation and noise, diminishing the interpretability of surface structures and polarization-based distinctions.

Fusion configurations marked by **ecological redundancy**, such as **Intra-Seasonal Fusion I** and **II**, where both SAR and optical data were acquired under dry or wet conditions respectively, tended to produce muted or overlapping information. These scenarios captured either uniformly low vegetation (in August) or early, indistinct growth (in February), leading to lower statistical contrast across classes. The weakest performance

was observed in the **Inverse Seasonal Fusion** setting, which combined wet-season SAR with dry-season optical data. This temporal inversion introduced both spectral ambiguity and radar decorrelation, resulting in limited separability between sinkholes, vegetation, and background surfaces.

The future trajectory of dryland ecosystems under climate change remains highly uncertain, with potential outcomes ranging from desertification to regreening, driven by complex and interacting environmental factors. In response, ecohydrological and ecological modelling communities have developed a variety of physically grounded frameworks to simulate vegetation pattern formation and its sensitivity to water availability. For instance, studies based on reaction, diffusion models such as the Klausmeier, Rietkerk, and Gilad frameworks have explored how soil moisture redistribution, scale-dependent feedbacks, and plant, water interactions can lead to self-organized vegetation structures across spatial scales [304]. More recently, researchers have started to bridge the gap between these process-based models and EO data by estimating model parameters directly from time series of satellite-derived vegetation density, using approaches such as differentiable programming [332]. While these efforts remain focused on model calibration and do not directly tackle EO-based classification, they reinforce the idea that vegetation dynamics in drylands are deeply shaped by physical processes that unfold across time and space.

Although the here present experimental setup does not implement process-based ecohydrological modelling or inversion techniques, the underlying logic of the cross-seasonal and cross-sensor fusion was indirectly informed by this body of research. Specifically, the selection of temporally and sensor-diverse datasets was conceptually guided by an awareness of the seasonal controls on surface moisture, vegetation greenness, and structural roughness, factors that also play a central role in ecohydrological feedback models. By intentionally pairing SAR acquisitions from dry, low-moisture periods (to emphasize surface structure) with optical data from peak vegetation phases (to maximize spectral contrast), the fusion strategy reflects a "process-aware" approach to remote sensing data integration. This conceptual alignment, though empirical in implementation, echoes broader trends towards physically informed observation strategies in EO science.

These findings underscore the strategic value of a *temporal taxonomy* in remote sensing fusion: not simply to synchronize acquisition dates, but to exploit ecological divergence across sensor modalities. By leveraging phenologically informed windows of maximum vegetation vitality and radar coherence, fusion approaches like the CDVI can generate

synergistic feature spaces supporting robust classification and enhanced detection of geomorphological and ecological features in semi-arid environments.

Building on these temporally and ecologically informed fusion strategies, the next step involves translating the multi-dimensional, pixel-level Kennaugh representations into a higher-order, feature-level descriptor. The CDVI exemplifies this progression: rather than treating each band as an isolated signal, it integrates structurally and spectrally complementary components into a targeted index. This derivation moves beyond raw fusion by extracting semantically meaningful patterns specifically tailored to enhance the joint detection of vegetation and geomorphological anomalies such as sinkholes.

5.2 Combined Doline Vegetation Index

The Combined Doline Vegetation Index (CDVI) was developed to address the specific challenge of detecting sinkholes in Kazakhstan's arid and semi-arid regions, where overlapping spectral and structural characteristics of vegetation, bare surfaces, and takyr-like areas complicate conventional detection methods. In particular, vegetation in these landscapes may appear in small-scale, shrub-like or roundish forms, potentially leading traditional filtering approaches to confuse them with sinkholes.

5.2.1 Materials

Accurate detection of sinkholes and vegetative features in arid landscapes requires input data that capture both fine-grained structural variations and seasonal vegetation dynamics. As outlined in Section 5.1, different temporal fusion settings were comparatively evaluated to identify optimal configurations for enhancing separability and robustness. The EO-data materials used in this study are primarily introduced in Section 5.1.1. A mixed-temporal (multi-season), multi-sensor fusion approach (see Section 5.1.2 based on ESA's Sentinel-1 SAR and Sentinel-2 optical data is utilized:

This fusion utilises the strengths of both sensors, SAR for structural characteristics and multispectral optical data for spectral detail, while also addressing seasonal variability [360]. Arid-season Sentinel-1 data (e.g., 6 August 2023) highlight backscatter differences due to sparse vegetation and low soil moisture, whereas humid-season Sentinel-2

data (e.g., 30 March 2023) provide key information on vegetation status in the NIR and visible bands.

In addition to the EO data, reference data, used for both vegetation and sinkhole validation, are described in detail in Section 1.2.1.

5.2.2 Methods

Based on extensive statistical testing (e.g., T-tests, ANOVA, correlation analyses), the CDVI was defined to maximise the separability between sinkholes, vegetation, and background classes. The final formula is expressed as in Equation (5.1):

$$\mathrm{CDVI} = \frac{K_{\mathrm{fused},5} + K_{\mathrm{fused},7}}{2} - K_{\mathrm{fused},1} + \frac{K_{\mathrm{fused},0} + K_{\mathrm{fused},6} + K_{\mathrm{fused},7} - K_{\mathrm{fused},1} - K_{\mathrm{fused},2}}{3} \tag{5.1}$$

Each term in Equation 5.1 contributes uniquely to suppressing irrelevant features while enhancing geomorphologically significant patterns. Their roles are summarised in Table 5.2, which details the underlying logic and effect of each component.

Table 5.2.: Roles of individual components in the CDVI formula.

CDVI Term	Component Description	Interpretation / Role
$\frac{K_{\text{fused},5}+K_{\text{fused},7}}{2}$	$K_{\text{fused},5}$: spectral/structural coherence $K_{\text{fused},7}$: volumetric scattering	Enhances vegetation overlays and sinkhole edges via structural–spectral integration.
$-K_{\text{fused},1}$	$K_{\text{fused},1}$: intensity contrasts	Reduces specular effects; strengthens vegetation and sinkhole discrimination.
$\frac{K_{\text{fused},0}+K_{\text{fused},6}+K_{\text{fused},7}-K_{\text{fused},1}-K_{\text{fused},2}}{3}$	$K_{\text{fused},0}$: total intensity $K_{\text{fused},6}$: spectral contrast $K_{\text{fused},7}$: volumetric scattering $K_{\text{fused},1}$, $K_{\text{fused},2}$: noise suppressors	Balances spectral and structural inputs; prioritises landform-specific signatures.

The CDVI thus serves as a tailored index for geomorphological detection in arid zones, optimally balancing seasonality, structural properties, and vegetation indicators. The strength of the Combined Doline Vegetation Index lies in its integration of structurally and spectrally complementary data collected across seasonal boundaries. By combining dry-season SAR data with humid-season optical observations, the CDVI captures both the geomorphological contrasts of sinkholes and the vegetative nuances of the surrounding landscape. To assess its effectiveness, the index was rigorously validated against independent land cover references, including Proba-V vegetation cover data and in-situ sinkhole mapping. The following section presents the quantitative and spatial evaluation of the CDVI, demonstrating its reliability for identifying sinkholes and vegetation dynamics in arid regions.

5.2.3 Results

The effectiveness of the CDVI is assessed through comparison with reference data and visual inspection across representative sinkhole sites. These verification results are presented in the following.

The CDVI was rigorously validated using multiple reference datasets to ensure its robustness and applicability across varied land cover types, including sinkholes, takyr surfaces, and vegetation. These reference datasets provided a solid foundation for assessing the CDVI's performance in distinguishing key features and land cover types in arid and semi-arid environments. The CDVI's capability to distinguish sinkholes and other land cover classes is depicted in Figure 5.2. Boxplots illustrate the variability and central tendencies of CDVI values for dominant land cover types, including sinkholes, takyr surfaces, dense vegetation, sparse vegetation, and bare ground. The distinct ranges highlighted in the boxplots demonstrate the CDVI's robustness in reducing misclassification and ensuring clear separability between these classes.

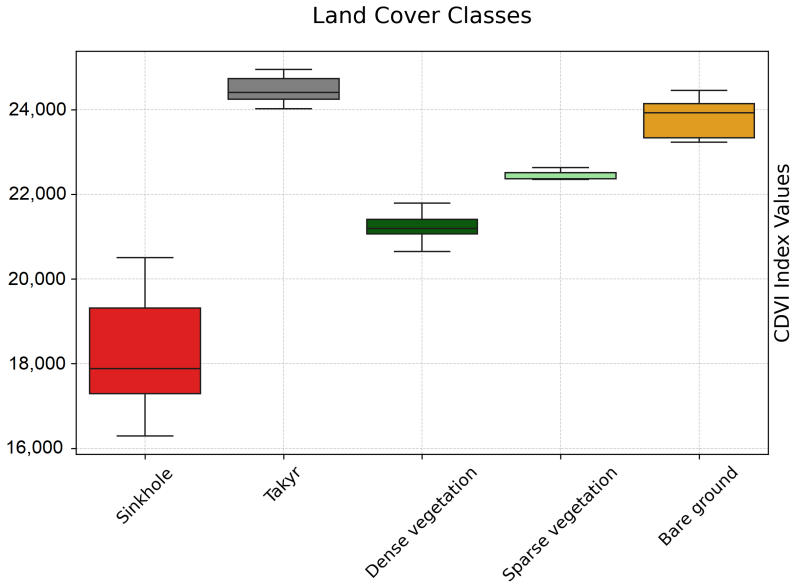


Figure 5.2.: Distribution of CDVI values for dominant land cover classes in the study area: sinkholes, takyr surfaces, dense vegetation, sparse vegetation, and bare ground. Boxplots show distinct value ranges, highlighting strong class separability.

The spatial visualisation in Figure 5.3 further highlights the CDVI's effectiveness in delineating sinkholes and vegetation patterns: (a) zoomed-in view of a sinkhole with

sparse vegetation and its corresponding CDVI values overlaid on WorldImagery [100], showing a clear correspondence between vegetation patterns and CDVI values; and (b) the same sinkhole area without the CDVI overlay, providing a direct comparison to raw imagery. These visualisations underscore the CDVI's capacity to accurately map sinkholes and surrounding vegetation, offering practical insights into its spatial applicability.

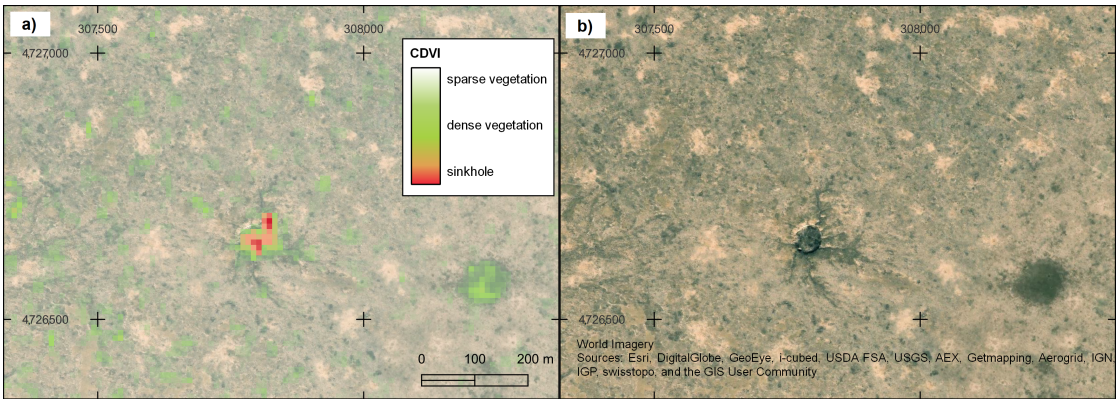


Figure 5.3.: Spatial visualisation of CDVI performance: **(a)** CDVI values overlaid on World Imagery [100] for a sinkhole with sparse vegetation; **(b)** the same area without overlay for comparison. The CDVI distinguishes structural and vegetative features clearly.

To validate the performance of the CDVI in detecting vegetation presence, its results were compared with the Proba-V LC100 global land cover product from 2019 [51]. The Proba-V dataset, part of the Copernicus Global Land Service, provides fractional grass and shrub cover values (0–100%) at a spatial resolution of approximately 100 m. While Proba-V is tailored to estimate grass and shrub coverage, the CDVI was designed to broadly detect vegetation presence, including grasses and small greenery, in arid and semi-arid landscapes such as southern Kazakhstan's Mangystau region. A binary vegetation mask was derived from the CDVI using a threshold determined in Figure 5.2. This classified vegetation presence as 1 (vegetation) and absence as 0 (no vegetation). To facilitate direct comparison with Proba-V, the CDVI raster was reprojected and aligned to the Proba-V grid (100 m resolution). Using a summation resampling method, the number of $10 \times 10\,\mathrm{m}$ vegetation pixels (1) within each Proba-V cell was counted. The fractional vegetation cover was then calculated for each Proba-V cell by dividing the vegetation pixel count by the total number of valid pixels in the cell. To smooth spatial variations and reduce noise, a Gaussian filter ($\sigma = 10$) was applied to the CDVI fractional vegetation cover data. This step produced a continuous raster surface comparable to Proba-V's coarser resolution

datasets. A Total Proba-V Vegetation Cover Fraction raster was calculated by summing the Proba-V Grass and Shrub Cover Fractions at each pixel. Values exceeding 100% were capped at 100%, and pixels marked as `NoData` in either layer were excluded from the calculation. This raster provides a combined measure of vegetation presence, serving as a reference for validating CDVI results. The validation was conducted using quantile-based analysis. Proba-V Total Vegetation Cover Fractions were divided into quantiles based on the 25th, 50th, and 75th percentiles, representing low, medium, and high vegetation cover (Quantiles Q1–Q3). CDVI results were then compared to Proba-V datasets (Grass, Shrubs, and Total Vegetation). Mean cover fractions, ranges, and quantile sizes for each dataset were computed. Spearman's correlation coefficients (r) were calculated to quantify the relationship between CDVI and Proba-V datasets (Grass, Shrubs, and Total Vegetation). Strong correlations were observed, with the highest correlation between CDVI and Total Vegetation ($r = 0.67$, $p < 0.001$). Figure 5.4 compares the mean fractional vegetation cover derived from CDVI and Proba-V [51] datasets across Total Vegetation quantiles (Q1–Q3). The results reveal a close alignment between CDVI and Proba-V Total Vegetation, with a slightly weaker correlation for Proba-V Grass and Shrubs. Numerical annotations highlight the mean values for each dataset within each quantile.

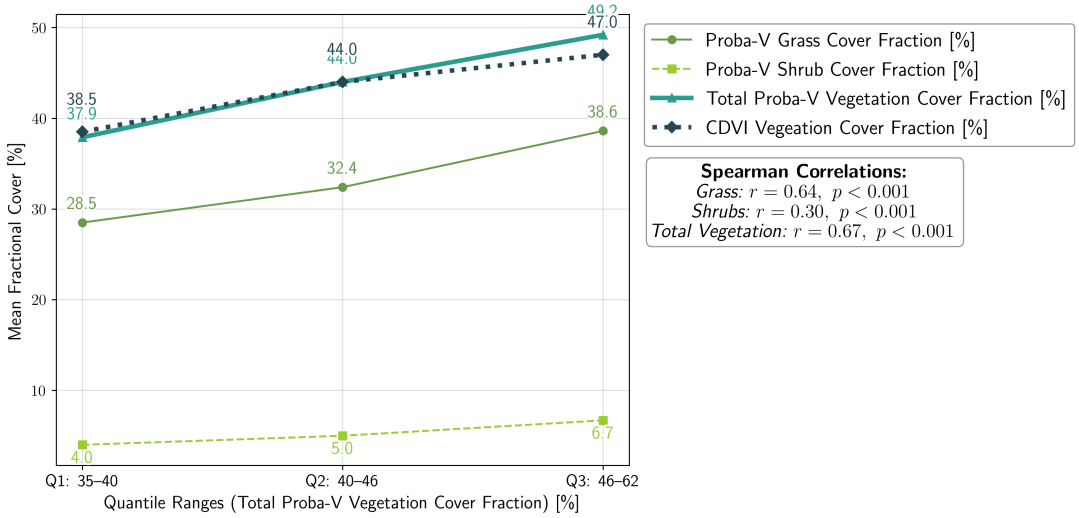


Figure 5.4.: Comparison of CDVI and Proba-V LC100 [51] vegetation cover fractions across quantiles (Q1–Q3). CDVI correlates strongly with Proba-V Total Vegetation ($r = 0.67$), with slightly weaker correlation for Grass and Shrub fractions.

Figure 5.5 presents spatial difference maps, visualising the discrepancy between smoothed CDVI fractional vegetation cover and each Proba-V [51] dataset (Grass, Shrubs, and

Total Vegetation). These maps reveal localised mismatches in vegetation cover, which may reflect CDVI's finer resolution and broader vegetation detection capabilities.

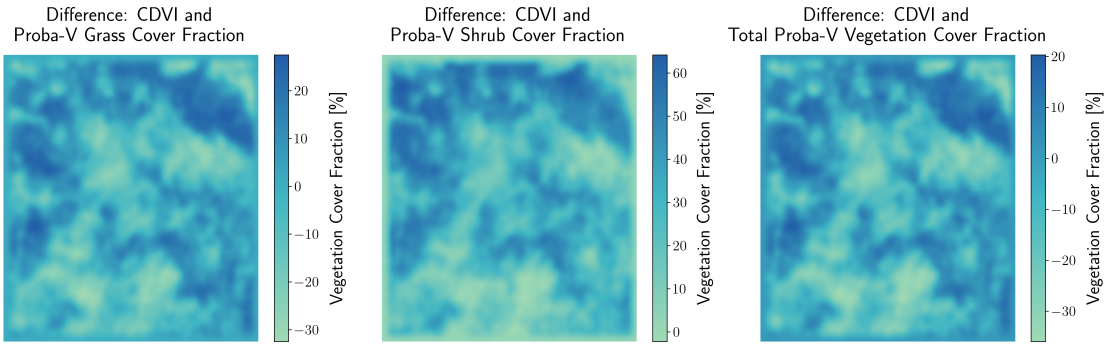


Figure 5.5.: Spatial differences between smoothed CDVI vegetation cover and Proba-V LC100 [51] Grass, Shrub, and Total Vegetation layers. Brown hues indicate higher CDVI values; green hues indicate higher Proba-V values.

The statistics for Total Vegetation quantiles, shown in Table 5.3, highlight the alignment between CDVI and the calculated Proba-V based [51] Total Vegetation data. CDVI provides detailed fractional vegetation cover estimates, capturing finer variations within each quantile.

The CDVI demonstrates strong agreement with Proba-V [51] datasets, particularly Total Vegetation ($r = 0.67$), confirming its ability to effectively detect vegetation presence in arid and semi-arid landscapes. The weaker correlation with Shrubs ($r = 0.30$) suggests that the CDVI is less tailored to detect sparse woody vegetation. These results highlight the complementary nature of the CDVI and Proba-V [51] datasets, with CDVI excelling in capturing smaller-scale vegetation patterns, such as interspersed grasses.

Table 5.3.: Total Proba-V [51] vegetation cover fraction (Grass and Shrubs) and corresponding CDVI vegetation statistics across quantiles.

Quantile Range (%)	Quantile Size (%)	Proba-V Mean (%)	Proba-V Range (%)	CDVI Mean (%)	CDVI Range (%)
35–40	22.01	37.89	36–40	28.51	2.6–59.3
40–46	28.00	44.01	41–46	38.55	6.6–63.5
46–62	21.66	49.21	47–62	47.04	16.1–68.5

The practical utility of the CDVI is illustrated in Figure 5.6, where the index is overlaid on Sentinel-2 imagery. Vegetation is represented in green, sinkholes in red, and known sinkholes [319] in orange. This example underscores the CDVI's role as a plausibility control tool for vegetation and sinkhole mapping.

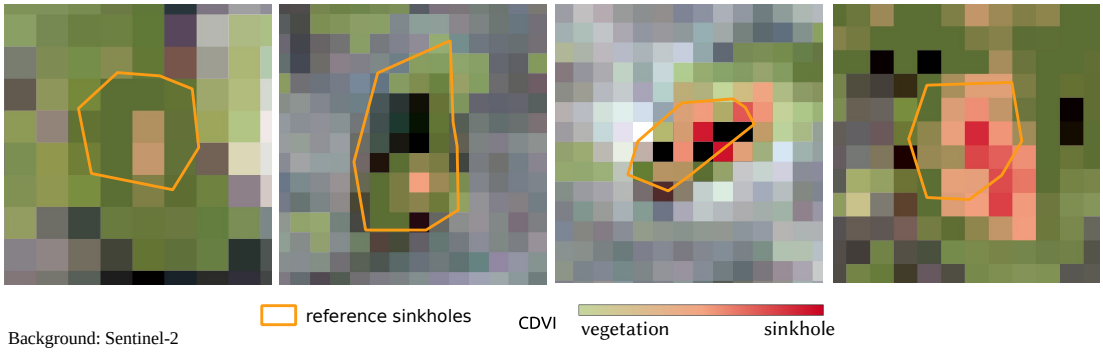


Figure 5.6.: CDVI overlay on Sentinel-2 (©ESA 2023) imagery. Green areas indicate vegetation, red areas mark detected sinkholes, and orange outlines show GPS-based sinkhole references [319].

The CDVI's fine resolution and multi-seasonal design make it an effective tool for ecological and geomorphological research, enabling the precise mapping of vegetation patterns in challenging environments, such as sparse grasses mixed with bare ground, which are characteristic of southern Kazakhstan's landscape [163, 189], and serving as a supplementary reference for sinkhole detection in this study.

5.2.4 Discussion

The results presented confirm the suitability of the CDVI for identifying sinkholes and vegetation patterns in arid and semi-arid environments. By integrating structural information from Sentinel-1 SAR with spectral detail from Sentinel-2 optical imagery across distinct seasonal periods, the CDVI effectively compensates for the spectral ambiguity often observed in takyr-like and sparsely vegetated landscapes.

One of the core advantages of the CDVI lies in its ability to reduce misclassification between geomorphologically relevant depressions and spectrally similar land cover types. The distinct value ranges observed in boxplots (Figure 5.2) demonstrate that the CDVI achieves clear separability between sinkholes, takyr surfaces, and different vegetation

densities. Visual overlays (Figure 5.3) further validate this distinction, highlighting the spatial correspondence between CDVI values and known sinkhole features.

The CDVI also showed a strong statistical correlation with external vegetation data sources. In particular, the comparison with Proba-V fractional vegetation cover (Figure 5.4) reveals consistent trends, especially in the Total Vegetation product, where a Spearman correlation of $r = 0.67$ was observed. This supports the reliability of the CDVI for general vegetation detection, especially in heterogeneous, dryland settings where subpixel vegetation can be challenging to delineate. The observed differences in spatial patterns between CDVI and Proba-V (Figure 5.5) are likely attributable to their differing spatial resolutions and design objectives, highlighting the CDVI's finer sensitivity to localized vegetation signatures.

Despite these strengths, some limitations are evident. The CDVI's correlation with Proba-V Shrubs ($r = 0.30$) suggests reduced sensitivity to sparse or woody vegetation types, potentially due to its temporal fusion strategy or the generalized formulation of the index. Moreover, while the CDVI provides excellent separation for major land cover classes, it may benefit from further calibration or adaptive thresholding in transitional zones or under varying soil moisture conditions.

Nonetheless, the CDVI offers a robust and interpretable framework for the remote sensing of subtle geomorphological and ecological features. Its utility extends beyond sinkhole detection and holds promise for broader applications in dryland monitoring, land degradation assessment, and seasonal vegetation analysis.

5.3 Conclusions

This chapter explored the role of temporal fusion in enhancing EO feature engineering, particularly in the context of detecting geomorphological features like sinkholes in semi-arid environments. By introducing a temporal taxonomy and systematically evaluating different fusion strategies, this study setup demonstrated how phenological complementarity between SAR and optical data can be leveraged to maximize class separability. The development and validation of the CDVI served as a practical outcome of this approach, illustrating the value of temporal and ecological insight in the fusion design process.

The following subsections summarize key lessons learned, revisit the central research questions in light of the findings, and offer closing reflections on the implications and potential extensions of this work.

5.3.1 Lessons Learned

This study setup presented several practical insights into temporal fusion design, feature selection, and the interpretability of RS models in complex environments. The lessons outlined below reflect both methodological considerations and broader implications for future research and operational applications:

- **Temporal Taxonomy as a Design Principle**: Effective remote sensing fusion benefits not from strict simultaneity but from *ecologically meaningful temporal diversity*. Phenological and hydrological differences between acquisitions enhance information richness.
- **Ecological Complementarity Outperforms Temporal Coherence**: Fusions across seasonal phases (e.g., dry SAR + wet optical) yielded superior class separability compared to temporally aligned (intra-seasonal) fusions.
- **Cross-Seasonal Fusion Maximizes Discriminative Power**: The best results were achieved when combining **dry-season SAR** (high structural stability) with **peak wet-season optical data** (high spectral vitality), particularly in semi-arid environments.
- **Wet-Season SAR Degrades Radar Signal Quality**: SAR acquisitions during high soil moisture periods showed reduced coherence and poorer separability, impairing structural information retrieval.
- **Phenological Timing is Critical for Optical Data**: Optical images captured during early regrowth or peak vegetation phases (e.g., March) significantly improved spectral contrast and feature discrimination.
- **Annual or Seasonal Aggregation Needs Careful Handling**: While annual aggregation smooths noise, strategic multi-season selection (not averaging) proved more effective for class-specific enhancement.
- **Temporal Inversion Should Be Avoided**: Pairing wet-season SAR with dry-season optical data (inverse seasonal fusion) led to the poorest results due to misaligned ecological signals.

- **Temporal Fusion Strategy Must Match Landscape Dynamics**: Semi-arid landscapes with strong seasonal contrasts benefit most from ecologically informed temporal fusion, rather than merely temporally simultaneous data integration.
- **Combined Doline Vegetation Index (CDVI) Validates This Approach**: The CDVI, based on cross-seasonal fusion, successfully demonstrated the practical advantage of temporal taxonomy by maximizing sinkhole and vegetation separability.

5.3.2 Research Questions Revisited

To structure the analyses, the following research questions were posed at the beginning of the chapter. The results presented throughout provide the following answers:

RQ1: *How does the ecological timing of multi-sensor data acquisitions influence the separability of land cover classes in arid regions?*
The ecological timing, particularly the contrast between dry- and wet-season acquisitions, was shown to be a critical factor. Fusion configurations that spanned seasonal boundaries (e.g., dry-season SAR and wet-season optical) provided significantly higher separability between sinkholes, vegetation, and background surfaces than temporally aligned acquisitions within the same season.

RQ2: *Can seasonally disparate fusion of SAR and optical EO data outperform temporally aligned fusion in detecting subtle geomorphological features?*
Yes. Seasonally disparate fusion strategies consistently outperformed intra-seasonal approaches. The most effective configuration (dry SAR + March optical) achieved the highest statistical discriminability across all tested Kennaugh bands and improved detection of small-scale geomorphological depressions.

RQ3: *What temporal configurations maximize the statistical discriminability of sinkholes, vegetation, and bare ground in semi-arid landscapes?*
The cross-seasonal configuration combining dry-season Sentinel-1 (August) and peak wet-season Sentinel-2 (March) proved optimal. It leveraged the structural clarity of SAR under dry conditions and the spectral vitality of vegetation under wet conditions, resulting in the best overall performance across ANOVA and correlation metrics.

RQ4: *To what extent can a temporal taxonomy inform the design of data fusion pipelines beyond strict temporal coherence?*

Temporal taxonomy provided a critical framework for evaluating fusion not by simultaneity, but by ecological complementarity. This setup demonstrated that fusions informed by seasonal understanding yield more meaningful feature spaces, especially in landscapes with strong phenological variability.

RQ5: *Can a tailored index, based on seasonally informed, multi-sensor data fusion, be developed to distinguish sinkholes from spectrally and structurally similar land cover types, and how effective is it in arid environments?*

Yes. The Combined Doline Vegetation Index (CDVI) was successfully derived from cross-seasonal fusion of SAR and optical data using the Hypercomplex Bases framework. It achieved clear separability between sinkholes and surrounding vegetation or bare surfaces. Correlation with Proba-V vegetation fractions ($r = 0.67$) further validated its effectiveness for fine-scale vegetation mapping in semi-arid regions.

5.3.3 Closing Remarks

The findings of this chapter challenge the conventional emphasis on temporal proximity in EO data fusion. Instead, they support a paradigm shift toward ecologically driven fusion strategies that embrace seasonal divergence and phenological dynamics. By formalizing temporal taxonomy as a design principle, this work provides a conceptual and empirical foundation for more intelligent, task-specific fusion configurations, especially in regions where seasonal processes dominate landscape change. The CDVI illustrates how meaningful information can be derived from cross-seasonal, HCB-fused EO features, demonstrating that valuable structural and spectral signals exist at the feature level. However, as this study also shows, such indices can be complex to design and tune manually. This underscores the role of ML as a scalable and adaptive approach to extract such latent information, building a conceptual bridge between handcrafted indices and automated learning pipelines, as introduced in Sections 1.1.1 and 1.1.2.

6 Foundational Analysis of EO Modality–Model Interactions

> *Everything should be made as simple as possible, but not simpler.*
>
> — **Albert Einstein**
> Physicist, Thinker

This chapter includes elements from the following peer-reviewed publications:

Sarah Hauser, Michael Ruhhammer, Andreas Schmitt, and Peter Krzystek. *An Open Benchmark Dataset for Forest Characterization from Sentinel-1 and -2 Time Series*. *Remote Sensing*, 16(3), 2024, Article 488. DOI:10.3390/rs16030488 It is cited as [147] and is marked with a green line.

Author Contribution: Sarah Hauser served as a primary contributor to study design, software implementation, practical execution, validation, writing, editing, and visualization.

and from:

Michael Ruhhammer, Sarah Hauser, Andreas Schmitt, and Anna Wendleder. *Forest parameter estimation from dual-frequency polarimetric SAR*. *Proceedings of the 15th European Conference on Synthetic Aperture Radar (EUSAR)*, Munich, Germany, 2024, pp. 966–971. It is cited as [278] and is marked with a light-green line.

Author Contribution: Sarah Hauser co-led the conceptualization and served as a primary contributor to the modelling methodology development, establishment of the analysis pipeline, and contributed significantly to the investigation, supervision, and manuscript writing.

This chapter establishes the methodological foundation for understanding how different EO modalities interact with various ML models when predicting continuous forest variables. Using the Wald5Dplus dataset [148] as a standardized label source, as described in Section 1.2.2, the analysis systematically explores which combinations of EO input and learning strategy yield the highest predictive performance.

The central question guiding this chapter is: *Which EO–model configurations yield the most accurate predictions of continuous forest attributes?* To address this, a modular experimental framework was designed that incrementally explores EO inputs and modelling strategies. The configurations range from mono-temporal, single-sensor inputs to fully fused spatio-temporal representations, allowing for both isolated comparisons and cumulative insights into how predictive performance evolves with added data richness and methodological sophistication. At the core of this progression lies the Wald5Dplus configuration [147], which employs Sentinel-1 and Sentinel-2 data fused spectrally, polarimetrically, and temporally using HCB [289]. This represents the most advanced setup and serves as a performance benchmark. All subsequent configurations are evaluated as simplified variants or baselines, designed to test the value added by specific data dimensions or fusion strategies.

The structure of this chapter proceeds through the following experimental tiers:

Sentinel-1 and -2 (Spectral, Polarimetric, and Temporal Hypercomplex Fusion): The most comprehensive configuration, combining full temporal sequences, polarimetric SAR, and spectral data through hypercomplex fusion.

Sentinel-1 + Sentinel-2 (Hypercomplex Fusion): A multi-modal fusion setup that integrates structural (SAR) and spectral (optical) information using a hypercomplex algebraic framework. (mono-temporal)

Sentinel-2: A spectral-only baseline assessing the predictive power of optical data, including a Kennaugh-like transformation of Sentinel-2 inputs to maintain comparability. (mono-temporal)

Sentinel-1: A radar-only baseline that examines model selection under controlled, mono-temporal conditions using standard Sentinel-1 polarimetric inputs.

TerraSAR-X and ALOS-2: A cross-sensor benchmark comparing Sentinel-1 with alternative high-resolution SAR systems, focusing on the influence of SAR-specific acquisition characteristics. (mono-temporal)

To ensure comparability, all experiments follow a unified evaluation protocol described in paragraph 6, including standardized preprocessing, model configuration, and validation logic. This framework enables controlled, reproducible comparison across modalities, models, and spatial domains. Ultimately, this chapter provides an evidence-based understanding of which EO and model configurations are most effective for multi-variable regression in forest ecosystems, laying the groundwork for downstream applications in vegetation monitoring, ecological forecasting, and spatial transfer. The analyses in this chapter are guided by the following research questions, which emerge from the methodological and conceptual context outlined above:

RQ1: *Which remote sensing modality, SAR (Sentinel-1), optical (Sentinel-2), or high-resolution SAR (TSX/ALOS), delivers the highest predictive accuracy for continuous forest structural variables in the Wald5Dplus dataset?*

RQ2: *How do polarimetric features derived from Sentinel-1 compare with raw spectral bands and transformed spectral features from Sentinel-2 in terms of predictive accuracy and spatial generalization?*

RQ3: *How does the choice between raw Sentinel-2 spectral bands and Sentinel-2-derived spectral Kennaugh-like elements affect model accuracy and spatial robustness for different forest structural variables?*

RQ4: *Do spectral or polarimetric Kennaugh-like representations improve spatial transferability over raw features, and for which types of forest variables is this most pronounced?*

RQ5: *To what extent does fusing optical and SAR data improve the prediction of forest structure variables compared to using single modalities?*

RQ6: *Which fusion strategy, spectral only, polarimetric only, or combined spectral–polarimetric, yields the best trade-off between in-domain accuracy and spatial transferability?*

RQ7: *How does the addition of temporal information to spectrally, polarimetrically, and temporally fused Sentinel-1 and Sentinel-2 data influence the performance and generalization of EO-based forest structure models?*

RQ8: *Which machine learning models, RF, SVR, CNN, or ensembles, perform best under varying EO input types and fusion configurations?*

RQ9: *How do preprocessing choices affect model accuracy and spatial robustness, particularly under domain shifts?*

RQ10: *Can ensemble learning approaches, particularly stacked RF ensembles, improve spatial generalization and mitigate performance degradation in unseen regions?*

RQ11: *What are the limitations of current models in achieving robust transferability, and how do fusion and ensemble strategies help overcome them?*

RQ12: *How do specific forest variables differ in their sensitivity to EO modality, preprocessing, and modelling approach?*

Systematic Evaluation Framework for Preprocessing and Model Configurations To ensure comparability across key experiments, a consistent evaluation framework was defined and is reused in all cases where full-grid model benchmarking is appropriate. This section outlines the standardized training, testing, and validation protocol that underpins those experiments where preprocessing configurations, model variants, and spatial transfer are systematically assessed. While not every chapter or module in this setup applies the full configuration grid, all analyses that do will reference this section to indicate methodological alignment. This enables a controlled, reproducible comparison of model performance across diverse experimental setups, allowing the reader to trace how changes in input data, preprocessing, or architecture influence predictive behaviour under a shared evaluation logic. To evaluate the predictive performance of various ML models on ecological and forestry-related variables, a comprehensive series of regression experiments was conducted. The experimental design aimed to systematically assess the influence of data preprocessing strategies and model configurations across a diverse set of vegetation-related response variables. The distinct target variables were selected for prediction, encompassing both structural and compositional forest attributes, as shown in Table 1.3. Three dimensions of data preprocessing were systematically varied:

Masking Thresholds: Data was filtered to include only observations exceeding certain thresholds of signal availability or quality. The following thresholds were applied:

- Mask > 0
- Mask > 0.1
- Mask > 1

Z-Score Filtering: To mitigate the impact of statistical outliers and improve data robustness, z-score normalization was applied to the data space prior to model training. A z-score z_i for a given value x_i is defined as:

$$z_i = \frac{x_i - \mu}{\sigma}$$

where μ is the mean and σ is the standard deviation of the respective feature across the dataset. This transformation standardizes the input features to have zero mean and unit variance, allowing for comparability across differently scaled variables. Outliers, defined as values whose absolute z-score exceeds a given threshold, can disproportionately influence model training, especially in loss-based optimization and gradient estimation. To control for this, several z-score thresholds were systematically evaluated:

- No z-score filtering
- Z $<$ 3 (mild filtering)
- Z $<$ 2.5 (moderate filtering)
- Z $<$ 2 (aggressive filtering)

Aggressive Filtering: An additional preprocessing dimension involved the application of an aggressive filtering pipeline, designed to jointly exclude unreliable pixels from both the feature and label spaces. This process compounded multiple quality control steps, including spatial masking (e.g., terrain artefacts) and statistical outlier removal (e.g., based on z-scores). A pixel was retained only if it passed all conditions simultaneously across modalities.

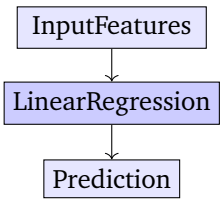
This binary configuration was encoded as:

- **Aggressive = True:** Combined masking and z-score-based filtering applied to both features and corresponding labels. If a pixel was identified as an outlier or invalid in either space, it was excluded from training and evaluation.
- **Aggressive = False:** Only minimal or no filtering applied; outliers and marginal data values were retained for maximum coverage.

This setting was especially relevant for high-sensitivity tasks where both input consistency and label integrity were critical, and it served to test the trade-off between model robustness and spatial completeness.

Regression Models and Hyperparameter Configurations: Multiple regression models were tested, each with their respective hyperparameter settings. The models and selected configurations include:

- **Linear Regression**: A baseline model without hyperparameters, used to establish a lower-bound performance. It applies an ordinary least squares fit to each target band independently. Despite its simplicity, it provides insight into model bias and target noise distribution (Figure 6).



- **RF Regression [48]**: An ensemble-based method that combines multiple decision trees trained on different bootstrapped subsets of the data and randomly selected input features (feature bagging). The final prediction is obtained by averaging the outputs of all trees. Key hyperparameters include `n_estimators` = [100, 200], which defines the number of trees in the forest, and `max_depth` = [5, 10, 20], which limits the depth of each tree. Shallower trees generalize better, while deeper ones capture more detail at the risk of overfitting. The `max_features` = [`log2`, `auto`] setting controls how many features are randomly selected at each split, introducing diversity among trees and helping reduce variance. Additionally, out-of-bag (OOB) estimation is used as an internal validation method: since each tree is trained on a bootstrap sample, roughly one-third of the data remains unused and can be used to estimate model performance without separate cross-validation. Each target variable is predicted independently using the same forest. The architecture is shown in Figure 6.1.

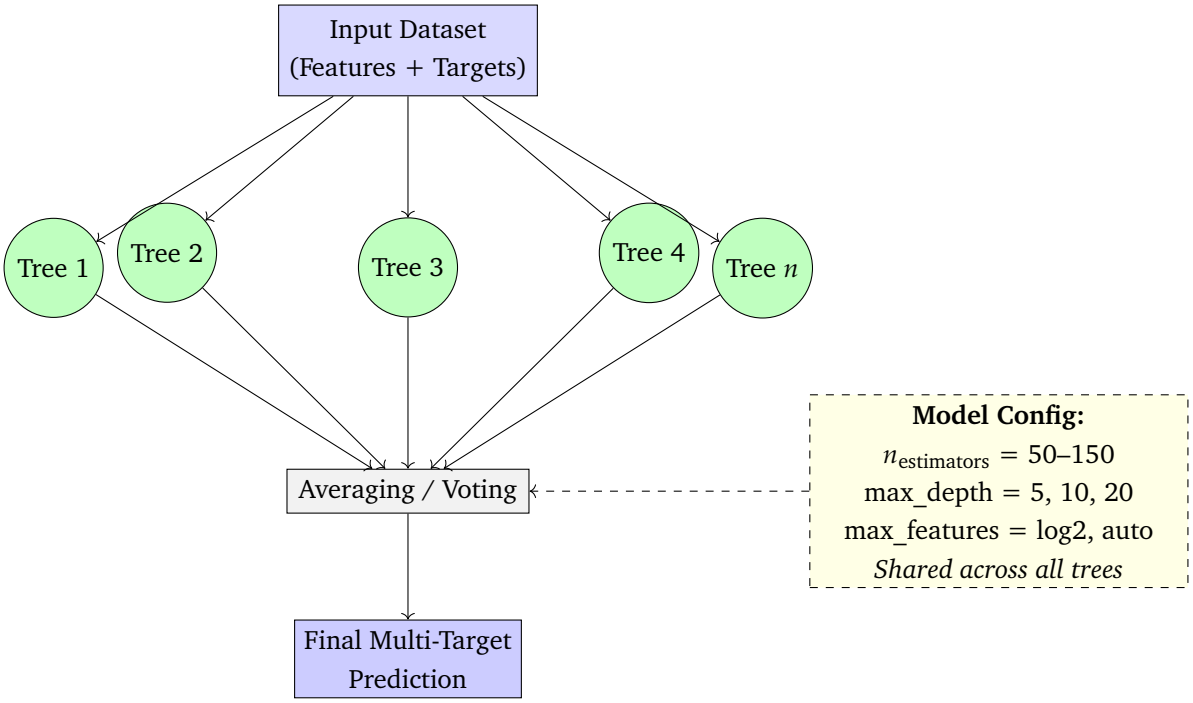


Figure 6.1.: RF architecture, each decision tree is trained on a bootstrapped subset of the input data and outputs predictions for all targets. Final output is obtained through averaging or majority voting.

- **SVR**: A kernel-based regressor trained using ε-insensitive loss. Both linear and RBF (radial basis function) kernels are tested (`kernel` = [`linear`, `rbf`]) to capture both linear and non-linear relationships between features and target variables. The ε-insensitive loss function defines a margin of tolerance (ε-tube) around the predicted regression function, within which no penalty is assigned to prediction errors. Regularization and model complexity are tuned via `C` = [0.1, 1, 10] and `gamma` = [`scale`, `auto`]. The `C` parameter controls the trade-off between training error and model generalization: low values allow for a wider margin (more regularization), while high values aim to fit the training data tightly. The `gamma` parameter, used in RBF kernels, determines the influence of individual training points: `scale` computes $\gamma = 1/(\text{n_features} \cdot \text{Var}(X))$, adapting to the data distribution, while `auto` sets $\gamma = 1/\text{n_features}$, independent of variance. SVR is applied independently to each target dimension. See Figure 6.2.

- **1D-CNN**: Implemented using PyTorch and trained on flattened raster data. The model is designed to process per-pixel spectral or temporal feature vectors in the

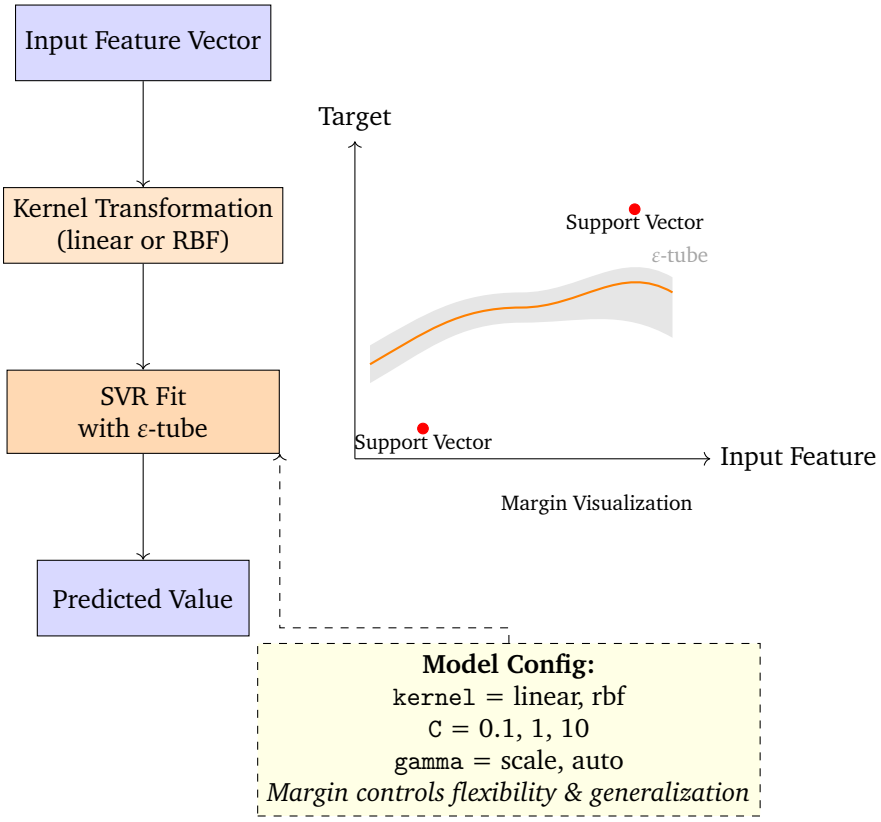


Figure 6.2.: SVR architecture. A kernel function transforms the input space, and a regression function is fitted with an ε-insensitive margin (gray band). Only support vectors outside this tube contribute to the loss function.

form of 1D sequences with multiple channels (e.g., spectral bands or time steps). The architecture consists of three 1D convolutional layers: the first two layers use 128 and 64 filters respectively, each with a kernel size of 3. A *kernel* defines the receptive field, a small window (size 3 here) that slides over the input to detect local patterns. Each layer uses the ReLU activation function, which introduces non-linearity by zeroing out negative values. The final convolutional layer uses a linear activation and maps to the number of output variables. *Filters* represent the number of distinct pattern detectors in each layer; more filters allow the model to learn a richer feature representation. Training is performed using the Adam optimizer, an adaptive algorithm that combines momentum and per-parameter learning rate adjustments to improve convergence stability and speed. Mean squared error (MSE) is used as the loss function. Both inputs and targets are min-max normalized before

training. Model evaluation includes validation MAE, cross-validation metrics, and diagnostic residual plots. The full architecture is shown in Figure 6.3.

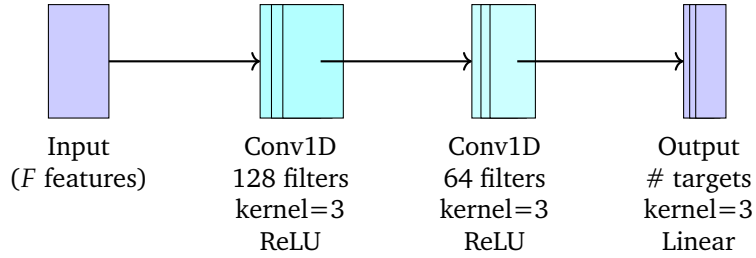


Figure 6.3.: 1D-CNN architecture with convolutional layers showing filters, kernel sizes, and activations.

Performance Evaluation Metrics: In the pursuit of precision, multiple complementary accuracy metrics, MAD, MAE, STD and RMSE, are employed to systematically scrutinize the predictive fidelity and robustness of the models. These metrics serve as essential instruments for quantifying both the magnitude of errors and the degree of variance explanation, offering a comprehensive perspective on the consistency, dispersion, and explanatory power inherent in the model predictions, which are explained in the following:

- **Median Absolute Deviation (MAD)**, a resilient sentinel against the vagaries of outliers, quantifies the median of absolute discrepancies between actual Y and predicted values $\hat{Y}$. The formula for MAD is defined as:

$$\text{MAD} = \text{median}(|Y_i - \hat{Y}_i|) \tag{6.1}$$

  MAD's robustness against outliers is indicative of the model's consistency in making predictions. Smaller MAD values signify predictions that closely adhere to actual values, showcasing the model's reliability in various contexts.

- **Mean Absolute Error (MAE)** serves as a robust gauge of the average prediction error and is articulated mathematically as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i| \tag{6.2}$$

  MAE offers invaluable insights into the magnitude of inaccuracies in our predictions. Lower MAE values are emblematic of heightened precision, symbolizing a close alignment between the predicted values $\hat{Y}$ and the actual values Y. The MAE

metric encapsulates the average magnitude of the prediction errors, illustrating how effectively the model approximates the true values.

- **Standard Deviation (STD)**, a widely employed metric for unearthing the degree of dispersion in prediction errors, is expressed mathematically as:

$$\text{STD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \tag{6.3}$$

  STD endeavours to elucidate the extent to which predictions cluster around the nominal value. Smaller STD values indicate the model's consistency, as predictions cluster closely around the nominal/actual value, thus affirming the model's reliability and stability. Conversely, larger STD values are indicative of a greater variability in predictions, signifying the potential for more erratic model behaviour. In the context of these metrics, actual values Y represent the true, observed values, while predicted values $\hat{Y}$ denote the values estimated by the model.

- **Root Mean Squared Error (RMSE)** offers another widely used performance metric, particularly valued for its sensitivity to larger errors. It is mathematically defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \tag{6.4}$$

  While similar in formulation to the standard deviation of residuals, RMSE directly captures the quadratic mean of prediction errors. Unlike MAE or MAD, RMSE disproportionately penalizes larger deviations, making it particularly effective in applications where larger errors are undesirable or carry greater consequence. Smaller RMSE values signify better predictive performance, with predictions $\hat{Y}_i$ closely aligning with observed values Y_i. However, due to its sensitivity to outliers, RMSE is best interpreted in conjunction with robust statistics such as MAD or MAE to provide a comprehensive view of model accuracy and error distribution.

Across the distinct forest-related target variables and 24 systematically varied preprocessing configurations, comprising masking thresholds, z-score filtering levels, and aggressive filtering toggles, each regression model was evaluated under controlled and repeatable conditions. The modelling framework encompassed four distinct regression paradigms:

- A baseline linear regression model, without hyperparameters, to serve as a performance benchmark.

- A RF regressor, tuned across multiple hyperparameter settings including the number of estimators (100, 200), tree depth (5, 10, 20), and feature selection strategies (`auto`, `log2`).
- A SVR, tested with both linear and RBF kernels, and multiple values for regularization strength (`C` = 0.1, 1, 10) and kernel scaling (`gamma` = `scale`, `auto`).
- An 1D-CNN, with a fixed architecture, trained with Adam optimization and MSE loss on normalized inputs.

Each of these models was applied independently to every target variable and preprocessing setup. When factoring in the individual hyperparameter variants, this produced a total of 576 unique model evaluations, capturing the full factorial interaction between data preprocessing, model architecture, and ecological prediction tasks. All experiments adhered to a consistent training and evaluation pipeline. For in-sample evaluation, models were trained on the entire preprocessed dataset corresponding to each configuration, and predictions were generated within the same domain to assess raw fitting performance. No internal train/test split was employed in this phase, as the objective was to evaluate model expressiveness under full-information conditions. In contrast, for spatial cross-validation, models trained on their original region were applied directly to a geographically distinct area without any re-training or hyperparameter adjustment. This strict external validation framework enabled assessment of generalization performance under spatial domain shifts. Across both evaluation stages, performance metrics were computed independently for each target variable, model, and preprocessing configuration, enabling fine-grained comparison across spatial, methodological, and variable-specific dimensions.

Each configuration was executed independently, facilitating a comprehensive, multidimensional analysis of modelling behaviour. The evaluation followed a threefold structure:

1. in-sample performance assessment across all configurations using acquisitions from 2020 (SW_2_2020),
2. independent cross-validation on a distinct spatial and temporal domain using data from 2021 (SW_4_2021), and
3. targeted evaluation of how the top-performing configurations transferred under combined spatial and temporal domain shifts.

6.1 Polarimetrically, Spectrally and Temporally Fused Sentinel-1 and Sentinel-2 Data

This section presents the full implementation of the HCB fusion framework [289] for creating a temporally extended, multi-modal RS feature dataset. As introduced in Section 2.2 and detailed in Section 2.2, the HCB method enables the orthogonal and lossless combination of Sentinel-1 polarimetric and Sentinel-2 spectral features into a compact, interpretable representation. Here, the fusion is applied at scale, integrating all available Sentinel-1 and Sentinel-2 acquisitions for the years 2020 and 2021 to construct a bi-temporal, full-year benchmark dataset. This setup is incorporating temporal dynamics into the fused feature space, providing a robust foundation for large-scale model training, seasonal analysis, and transfer learning within the Wald5Dplus project.

6.1.1 Materials

The key conditions for the choice of satellite are: public availability without costs, high temporal as well as high spatial resolution, and sufficient coverage for larger forest stands. These requirements are fulfilled by the Sentinel-1/2 missions of the Copernicus program. Due to the Open Data policy of ESA, anyone can download and evaluate the data, which is a crucial step for the extensive use of the knowledge gained by training on the reference data. The short repeat pass times enabled by two satellite sensors on the same orbit in space guarantee weekly acquisitions in the case of the weather-independent SAR sensors Sentinel-1a and Sentinel-1b. The optical Multi Spectral Imager on Sentinel-2a and Sentinel-2b though passing every five days is often hindered by clouds. Furthermore, the varying illumination conditions hamper the consistent interpretation. Thus, a sophisticated preprocessing is necessary in both cases: first, to identify and to remove (for the most part) clouds and other atmospheric effects and second, to establish a common reference frame – a high resolution 10 m pixel grid in UTM coordinates – for the subsequent data fusion. The Sentinel mission per se delivers a Europe-wide coverage with these stringent requirements and a global coverage of the land surfaces with possibly lower spatial or temporal resolution.

Sentinel-1 acquires VV and VH polarized SAR images in C-band, i.e., the images are sensitive towards structures in the size of the wavelength of about 5 cm. The co-polarization VV is known to deliver the highest backscatter over land. The cross-polarization VH

on the contrary is dominated by the volume scattering effect that can be observed in backscattering volumes like high vegetation like forests. The originally complex images are preprocessed by the Multi-SAR processor of DLR [38]. It calculates the four Kennaugh elements k_0, k_1, k_5, and k_8 and therewith assures an information preserving representation of the polarimetric information [288]. The Kennaugh elements which are nothing else than intensities and intensity differences are then multi-looked in order to generate square pixels and geocoded to the respective UTM zone. As SAR is characterized by the inherent speckle noise, a special adaptive filtering approach known as multi-scale multi-looked follows [288]. In this filtering approach, the noise content is adopted from the denoted noise floor provided in the metadata and neighbouring pixels are smoothed as long as their difference in backscatter does not exceed the expected noise variation. Thanks to the extraordinary noise model [285], edges are preserved in order to prevent any information loss. The final normalization ensures a closed data range and the space-saving archiving a UInt16 digital numbers in analogy to the Sentinel-2 images.

The Sentinel-2 MAJA [85] product was selected for its ability to generate Bottom-of-Atmosphere (BOA) reflectances with integrated temporal consistency, atmospheric correction, and cloud screening, key prerequisites for time-series analysis in EO. Although residual cloud gaps remain post-processing, these can be addressed through subsequent interpolation methods. The 10,m resolution bands in the blue, green, red, and NIR ranges were emphasized, as they offer the necessary spectral fidelity for forest structure modelling.

Two important aspects are not taken into consideration so far: cloud gaps and the varying acquisition time of Sentinel-1 and Sentinel-2. The gaps caused by clouds and insufficient illumination are closed by reasonable values interpolated on HCB. This algorithm acts like a Fourier transform in the temporal domain with only sparse input values. The resampling from the Sentinel-2 acquisition times to the regular Sentinel-1 acquisitions every six days is realized by a further interpolation. These two steps guarantee a plausible temporal signature and a consistent image fusion.

As detailed in Section 2.2, the temporal fusion is performed using the HCB [289] framework, yielding a compact and orthogonal feature representation consisting of 512 channels, generated by fusing 64 times 8 fused Kennaugh-like elements per year [147] with $K_{*,0}$, which is representative for the whole year (similar to the total intensity), and 63 elements $K_{*,1-63}$.

The complete processing pipeline is visualized in Figure 6.4, where Sentinel-1 and Sentinel-2 data are first transformed into polarimetric and spectral Kennaugh representations, respectively, and then spectrally, polarimetrically, and temporally fused.

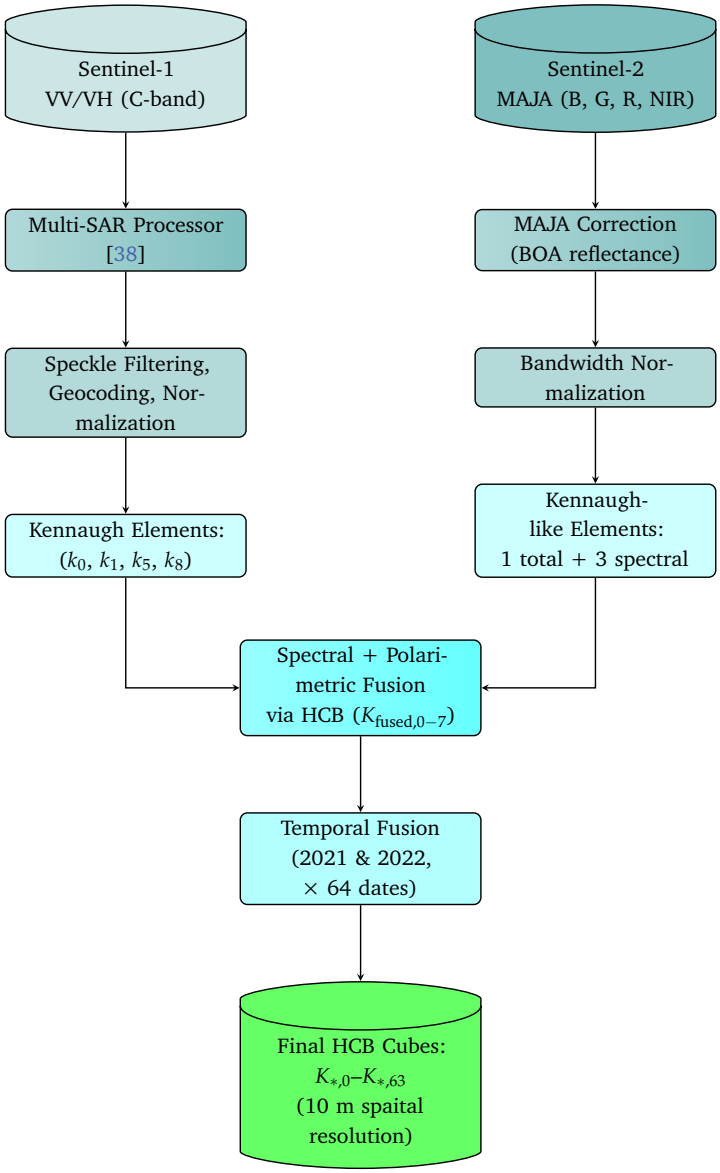


Figure 6.4.: End-to-end fusion pipeline: Sentinel-1 and Sentinel-2 inputs are transformed into compatible Kennaugh representations and fused spectrally, structurally, and temporally using the Hypercomplex Bases method.

The label data used in this experiment originates from the Wald5Dplus dataset, specifically from the Steigerwald study area (AOI 1), as described in Section1.2.2. This setup mirrors

the experimental configurations employed in succeeding chapters and is thus directly comparable to:

- Section 6.4 – Polarimetric Kennaugh Elements from Sentinel-1 Data
- Section 6.3 – Sentinel-2 Data
- Section 6.2 – Polarimetrically and Spectrally Fused Sentinel-1 and Sentinel-2 Data

By maintaining consistency across these sections in terms of label origin, study area, and preprocessing, this experiment enables a controlled evaluation of the added value of temporal fusion over previously tested spectral and polarimetric combinations.

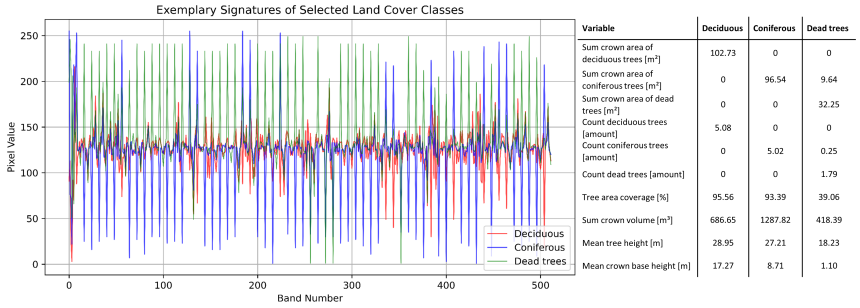
6.1.2 Methods

This part of the work centres on estimating various tree-related characteristics, including total crown areas for deciduous, coniferous, and dead trees (in square meters), tree counts per category, overall tree cover percentage, total crown volume (in cubic meters), and average measurements such as tree height and crown base height (in meters), as summarized in Table 1.3, and as outlined in Section 1.2.2. This reference data serves as semantic reference against which the predictive capabilities of the fused Sentinel-1 and -2 satellite satellite data is assessed.

The fused EO data serves as a key component of the model input, the ARD cubes (see Figure 6.4), each containing 512 channels, introduced also in Section 2.2. Figure 6.5 illustrates typical spectral signatures for deciduous, coniferous, and dead trees, providing a visual sense of class distinctions of the fused datasets.

Initial Model Development and Experimental Setup

To systematically evaluate the potential of ML algorithms for the prediction of forest structural and compositional variables, an extensive experimental framework was implemented. This included variations in input preprocessing, such as row-wise masking thresholds, zero-row filtering, Z-score-based trimming for outlier removal, and normalization strategies (e.g., MinMax scaling), alongside model-specific hyperparameter



Variable	Deciduous	Coniferous	Dead trees
Sum crown area of deciduous trees [m²]	102.73	0	0
Sum crown area of coniferous trees [m²]	0	96.54	9.64
Sum crown area of dead trees [m²]	0	0	32.25
Count deciduous trees [amount]	5.08	0	0
Count coniferous trees [amount]	0	5.02	0.25
Count dead trees [amount]	0	0	1.79
Tree area coverage [%]	95.56	93.39	39.06
Sum crown volume [m³]	686.65	1287.82	418.39
Mean tree height [m]	28.95	27.21	18.23
Mean crown base height [m]	17.27	8.71	1.10

Figure 6.5.: Typical spectral signatures of deciduous, coniferous, and dead trees [147].

grids. These experimental setups were consistently applied to a diverse pool of regressors, including RF, SVR, linear regression, and an 1D-CNN. Full technical details of the experimental matrix are provided in Section 6.

Intra-AOI and Transfer-AOI Evaluation Design

Each model was evaluated using two distinct strategies to assess its generalization capacity:

- **Intra-AOI setting:** Training and testing were performed on non-overlapping subtiles from the same AOI. This ensured independence between training and test data and minimized spatial autocorrelation.
- **Transfer-AOI setting:** Models trained in one AOI were evaluated on a spatially disjoint and unseen region. This setting simulated a real-world application, where models must generalize beyond their original training domain.

Stacked Ensemble Modelling with Multi-Output RF

Following the *Initial Model Development and Experimental Setup* as well as the *Intra-AOI and Transfer-AOI Evaluation*, a two-level stacked ensemble modelling strategy was implemented to predict forest structural and compositional attributes across multiple regions and temporal scales. This approach builds upon RF regression, which is employed

at both the base learner and meta-learner levels. The architecture was chosen for its robustness to noise, non-parametric flexibility, and its proven ability to model complex, non-linear relationships within heterogeneous environmental datasets.

To mitigate the temporal offset between predictors and reference labels, the modelling strategy explicitly incorporated spatial and temporal tiling, designed to improve generalization and reduce the impact of ecological change between acquisitions. Each AOI was subdivided into smaller sub-AOIs (tiles), enabling stratified sampling across both spatial and temporal dimensions. For each tile, a dedicated multi-output RF model was trained to predict a subset of 8–10 forest attributes. Hyperparameters were optimized using randomized grid search with spatial cross-validation.

The core method lies in the construction of a meta-model through stacked generalization. Predictions from all valid base models, regardless of AOI or target dimension, were used as input features to train a second-level RF meta-regressor. Each base model's predictions were treated as a learned representation of forest attributes, effectively capturing localized spectral–structural relationships.

The stacking implementation was realized programmatically using the `scikit-learn` API[1]. Each pre-trained base model was encapsulated in a custom transformer class, `ModelTransformer`, derived from `BaseEstimator` and `TransformerMixin`. This wrapper exposed the base model's `predict()` method as a callable transformer, allowing its outputs to be treated as features within a unified ML pipeline. Multiple `ModelTransformer` instances, each corresponding to a distinct base model trained on a specific AOI or subset of forest attributes, were then combined using the `FeatureUnion` class. This operation applied all model transformers in parallel to the same EO predictor dataset (the 512-band time series) and horizontally concatenated their individual prediction outputs into a single stacked feature matrix. Importantly, all base models were applied to the full input space of EO features, despite being trained on different target subsets. As a result, each model contributed a different perspective (e.g., region-specific or trait-specific) to the ensemble, which the meta-learner could exploit. This meta-feature matrix was subsequently passed to a second-level estimator, again a RF in this implementation, designed to learn the mapping from the ensemble of base predictions to the full set of the in total 10 forest attributes. This architecture allowed the meta-model to learn implicit weightings across the diverse prediction sets and to correct for systematic biases or regional overfitting in the individual base models. The modular nature of the pipeline

[1] For documentation, see https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.FeatureUnion.html

further enabled dynamic inclusion of additional base models, such as those trained on 2-band targets or alternative AOIs, thereby enhancing the transferability of the model to new regions and time periods. An overview of the full modelling pipeline is presented in Figure 6.6, including EO and reference data preprocessing, sub-AOI tiling, independent base model training, and meta-model integration.

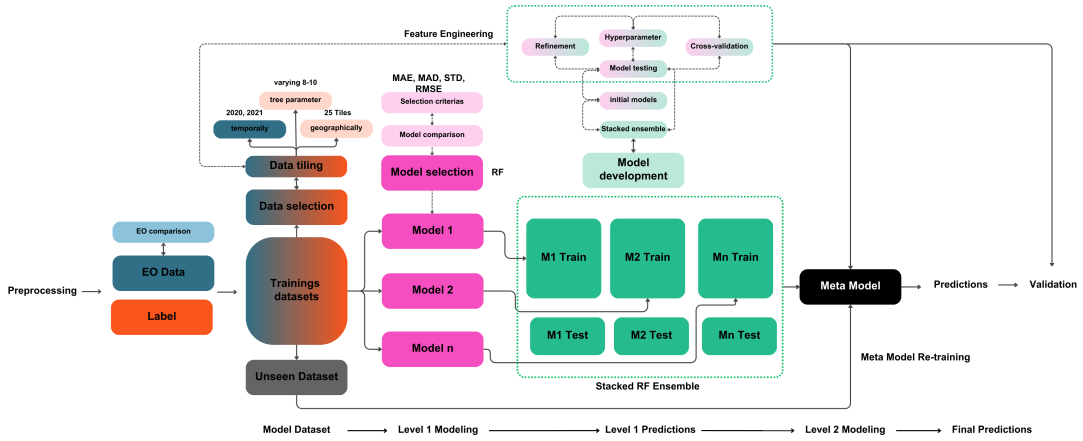


Figure 6.6.: Workflow for Stacked RF Meta-Model Development and Prediction. The pipeline includes EO and reference data preprocessing, sub-AOI tiling, feature engineering, individual model training, ensemble stacking, and final prediction through a meta-learner.

The model architecture supported both spatial and temporal cross-validation. For spatial validation, models trained on one AOI or sub-AOI were tested on another. Similarly, in temporal validation scenarios, models trained on EO data from one year were tested against the withheld year. This design allowed for realistic assessment of generalizability, particularly relevant for monitoring applications where updated reference data may not be readily available.

6.1.3 Results

To identify the most suitable model architecture for predicting forest structural and compositional variables, a comprehensive comparison of ML regressors was conducted using MAE as the principal performance metric. Evaluations were carried out under both intra-AOI and cross-AOI (transfer) scenarios, capturing predictive performance within and beyond the trained spatial domain.

Model Evaluation per Variable

To evaluate model performance across eight forest structural and compositional variables within the Steigerwald study area (AOI 1), a systematic comparison of RF, SVR, 1D-CNN, and Linear Regression models was conducted under intra-AOI and cross-AOI scenarios. Table references correspond to results from the original spatial domain, in Tables A.4 to A.11, while corresponding performance shifts are detailed in Tables A.12 to A.19.

Table 6.1 presents a consolidated overview of the best-performing configurations across all forest parameter variables. The results focus exclusively on the most effective RF setups, identified through intra-AOI evaluation, and report both the absolute performance (MAE, RMSE) and the performance shift observed during cross-AOI transfer (ΔMAE, ΔRMSE). This comparison highlights not only the predictive strength of RF models under optimal conditions but also their resilience, or sensitivity, to domain shift across forest environments.

Table 6.1.: Summary of RF model performance for all forest variables. Best intra-AOI configuration and corresponding cross-AOI performance shifts (ΔMAE, ΔRMSE) are reported.

Variable	Best Config (RF)	MAE	RMSE	ΔMAE	ΔRMSE
Sum Crown Area (Decid.)	Mask>1, Z=3, Agg=True	4.03	5.27	+10.63	+13.31
Sum Crown Area (Conif.)	Mask>1, Z=3, Agg=True	3.48	4.49	+6.59	+7.99
Count Decid. Trees	Mask>1, Z=3, Agg=True	0.18	0.23	+0.48	+0.57
Count Conif. Trees	Mask>0, Z=3, Agg=True	0.11	0.16	+0.25	+0.31
Tree Area Coverage (%)	Mask>1, Z=3, Agg=True	0.66	1.09	+1.74	+3.47
Sum Crown Volume (m^3)	Mask>1, Z=3, Agg=True	23.32	31.67	+58.61	+82.68
Mean Tree Height (m)	Mask>1, Z=3, Agg=True	0.61	0.81	+1.31	+1.66
Mean Crown Base Height (m)	Mask>1, Z=3, Agg=True	0.63	0.80	+1.64	+1.91

Across all variables, RF regressors delivered the most consistent and top-tier performance in the original spatial domain and remained relatively robust in cross-AOI settings. SVR models, while occasionally competitive in isolated metrics, exhibited higher volatility and performance drops across spatial folds. Linear regression showed poor fit and generalization overall. The 1D-CNN yielded reasonable intra-AOI results for some variables (e.g., *count of coniferous trees*), but its cross-AOI generalization was weak, particularly for structurally complex and volumetric features. Overall, RF demonstrated superior generalization and error resilience across forest structure and composition metrics, vali-

dating its selection as the most dependable baseline model for predictive tasks in spatially heterogeneous EO contexts.

Model Performance and Preprocessing Effects

Model evaluation revealed that predictive accuracy varied systematically across target forest structure variables, influenced by both the RF configuration and the applied preprocessing strategies. As summarized in Table 6.1, models that combined moderate masking (`Mask>1`), light Z-score trimming (`Z=3`), and aggressive filtering consistently achieved the lowest in-domain MAEs and RMSEs across nearly all variables. For example, *mean tree height* was predicted with an MAE of 0.608 and RMSE of 0.813 (Table A.10), and *sum crown volume* with an MAE of 23.316 (Table A.9).

In contrast, configurations without masking or with overly harsh outlier suppression yielded significantly higher errors. Particularly for rare-event-sensitive variables such as crown volume, aggressive filtering often eliminated ecologically informative extremes, leading to reduced model generalizability.

The best RF architecture shared several consistent traits. The most effective models had `max_depth=None`, enabling unrestricted tree growth, `max_features='log2'` to promote feature randomness and generalization, `min_samples_leaf=2`, `min_samples_split=5`, and `n_estimators=150`. These settings offered a reliable balance between model complexity and overfitting control. Feature normalization techniques were tested but provided no significant benefits, affirming RF's robustness to unscaled feature inputs.

Spatial Generalization and Transfer Behaviour

Cross-AOI transfer results underscored that while intra-domain model fit was strong, spatial generalization performance varied notably by variable. For example, *mean tree height* exhibited a moderate increase in MAE (+1.31) and RMSE (+1.66) under transfer, whereas *crown volume* suffered a substantial degradation (+58.61 MAE, +82.68 RMSE). Such trends suggest that some forest attributes, particularly those linked to aggregation or complex volume estimations, are more sensitive to local ecological conditions.

The choice of preprocessing had measurable effects on transferability. While aggressive filtering improved intra-AOI accuracy, it often reduced model robustness under domain shift. Models using `Aggressive=False` frequently preserved better generalization for

volume and crown-based attributes, likely due to retention of rare, informative outliers. Z-score trimming at Z=3 remained optimal, providing balance between noise reduction and signal preservation.

RF models consistently outperformed SVR, CNN, and linear regression alternatives in both intra- and cross-AOI settings. SVR showed erratic cross-validation behaviour, and CNN suffered from poor generalization despite strong localized fits. Linear regression failed to model non-linear interactions intrinsic to forest structure dynamics. Thus, ensemble-based methods combined with well-tuned preprocessing strategies offer the most stable and scalable solutions for EO-based forest variable prediction.

Normalization and Configuration Summary: Importantly, feature normalization techniques (e.g., MinMax scaling, standardization) did not improve RF performance, reaffirming the known robustness of RF models to raw feature distributions. As illustrated across all target variables, ranging from tree counts to structural attributes (Figures 6.7–6.14), configurations that combined moderate masking (M1), light Z-score filtering (Z3), and aggressive filtering (A=True) consistently achieved the lowest MAEs. This trend is evident in predictions of coniferous tree count (Figure 6.10), mean tree height (Figure 6.14), and total crown volume (Figure 6.12). Conversely, omitting preprocessing or applying overly aggressive filtering without masking led to marked increases in error (e.g., MNA_ZNA_ANA configuration).

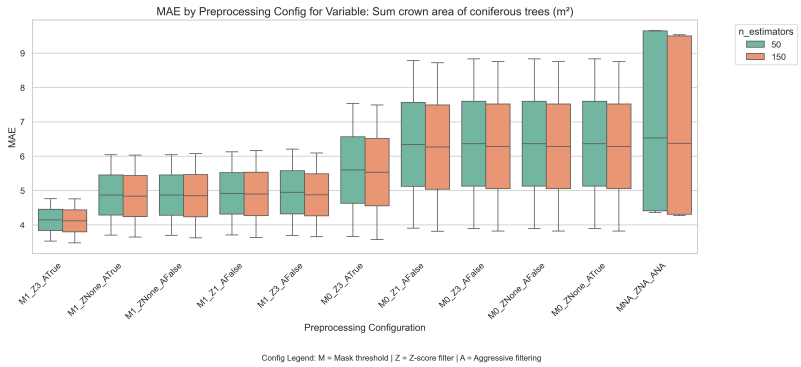


Figure 6.7.: Mean Absolute Error (MAE) distributions for different preprocessing configurations predicting Sum crown volume of coniferous trees (m^2). Each configuration varies in mask threshold (M), Z-score filtering (Z), and aggressive filtering (A). Results are shown for two values of the number of estimators in the model: 50 and 150. Lower MAE values indicate better model performance.

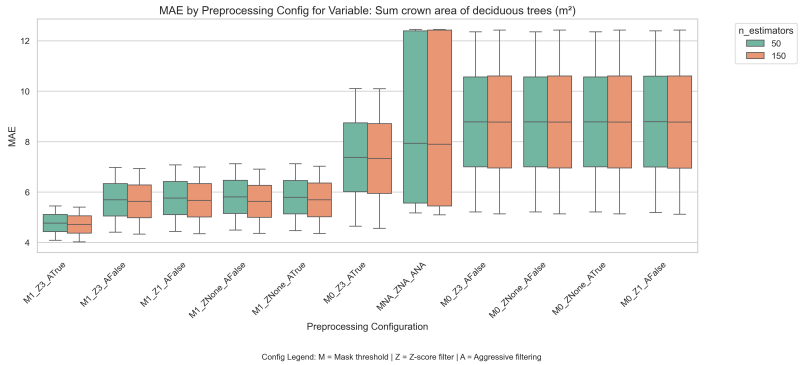


Figure 6.8.: Mean Absolute Error (MAE) distributions for different preprocessing configurations predicting Sum crown volume of deciduous trees (m^2). Each configuration varies in mask threshold (M), Z-score filtering (Z), and aggressive filtering (A). Results are shown for two values of the number of estimators in the model: 50 and 150. Lower MAE values indicate better model performance.

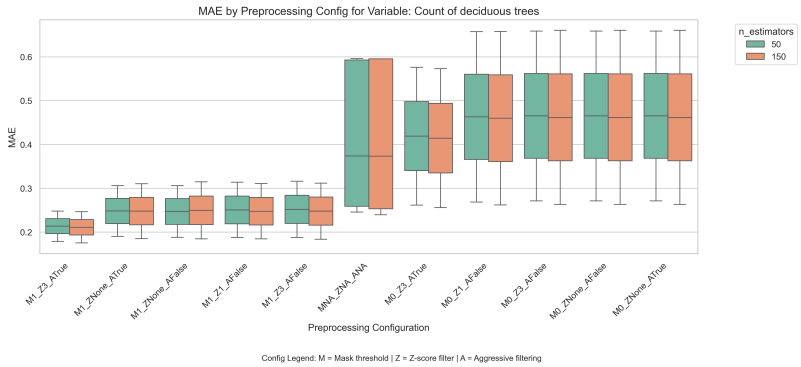


Figure 6.9.: Mean Absolute Error (MAE) distributions for different preprocessing configurations predicting Count of deciduous trees. Each configuration varies in mask threshold (M), Z-score filtering (Z), and aggressive filtering (A). Results are shown for two values of the number of estimators in the model: 50 and 150. Lower MAE values indicate better model performance.

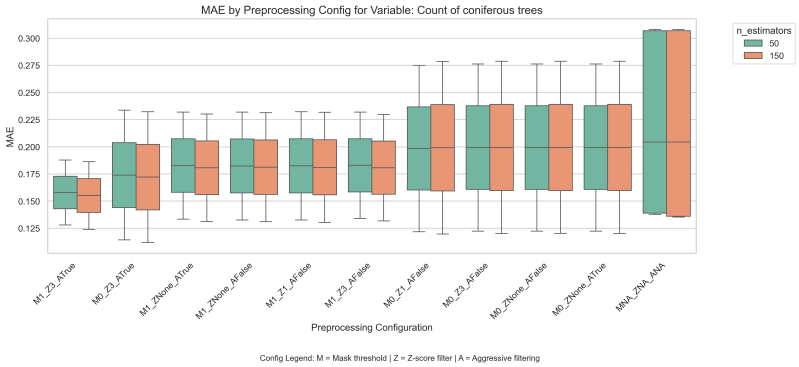


Figure 6.10.: Mean Absolute Error (MAE) distributions for different preprocessing configurations predicting Count of coniferous trees. Each configuration varies in mask threshold (M), Z-score filtering (Z), and aggressive filtering (A). Results are shown for two values of the number of estimators in the model: 50 and 150. Lower MAE values indicate better model performance.

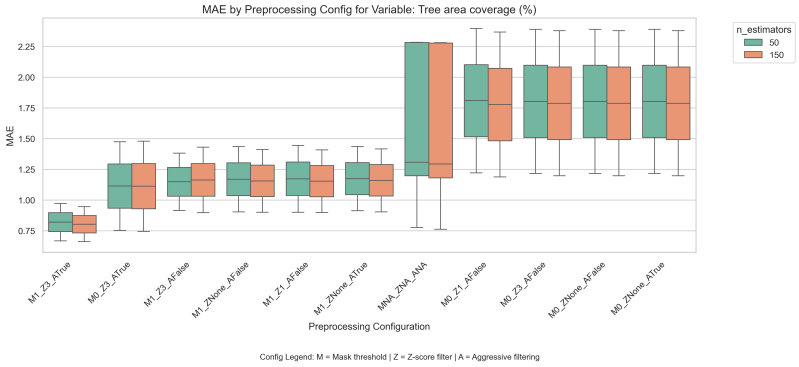


Figure 6.11.: Mean Absolute Error (MAE) distributions for different preprocessing configurations predicting Tree area coverage (%). Each configuration varies in mask threshold (M), Z-score filtering (Z), and aggressive filtering (A). Results are shown for two values of the number of estimators in the model: 50 and 150. Lower MAE values indicate better model performance.

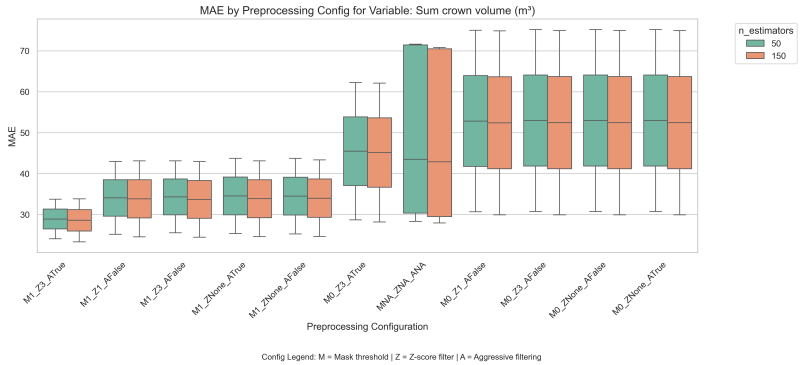


Figure 6.12.: Mean Absolute Error (MAE) distributions for different preprocessing configurations predicting Sum crown volume (m^3). Each configuration varies in mask threshold (M), Z-score filtering (Z), and aggressive filtering (A). Results are shown for two values of the number of estimators in the model: 50 and 150. Lower MAE values indicate better model performance.

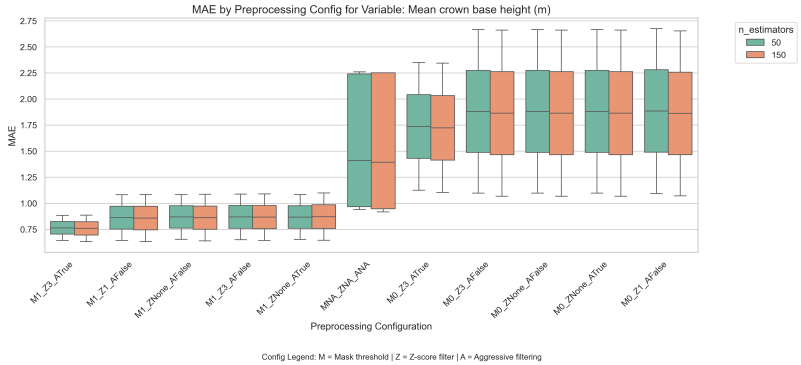


Figure 6.13.: Mean Absolute Error (MAE) distributions for different preprocessing configurations predicting mean crown base height (m). Each configuration varies in mask threshold (M), Z-score filtering (Z), and aggressive filtering (A). Results are shown for two values of the number of estimators in the model: 50 and 150. Lower MAE values indicate better model performance.

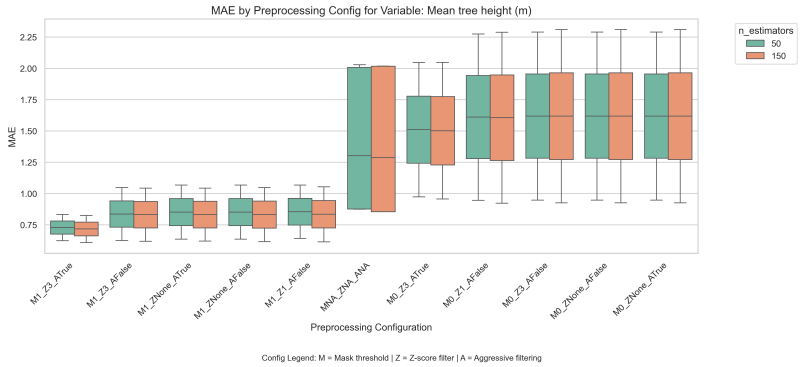


Figure 6.14.: Mean Absolute Error (MAE) distributions for different preprocessing configurations predicting Mean tree height (m). Each configuration varies in mask threshold (M), Z-score filtering (Z), and aggressive filtering (A). Results are shown for two values of the number of estimators in the model: 50 and 150. Lower MAE values indicate better model performance.

Overall, these results confirm that RF, when paired with targeted preprocessing, particularly moderate masking and controlled filtering, are both accurate and operationally robust for forest structure and composition mapping across heterogeneous landscapes. The summarized performance metrics of the best overall RF configuration are provided in Table 6.1, which outlines the model's accuracy across all eight forest variables. Figure 6.15 further illustrates the predictive performance of these optimized RF regression models within the Steigerwald study area (AOI 1). Each subplot compares predicted versus reference values derived from spectrally, polarimetrically, and temporally fused Sentinel-1 and Sentinel-2 inputs. The red dashed line indicates the ideal 1:1 relationship, while logarithmic colour density emphasizes point distribution. Collectively, the plots and metrics demonstrate the strong fit and generalization capacity of the final RF models when evaluated on independent hold-out data.

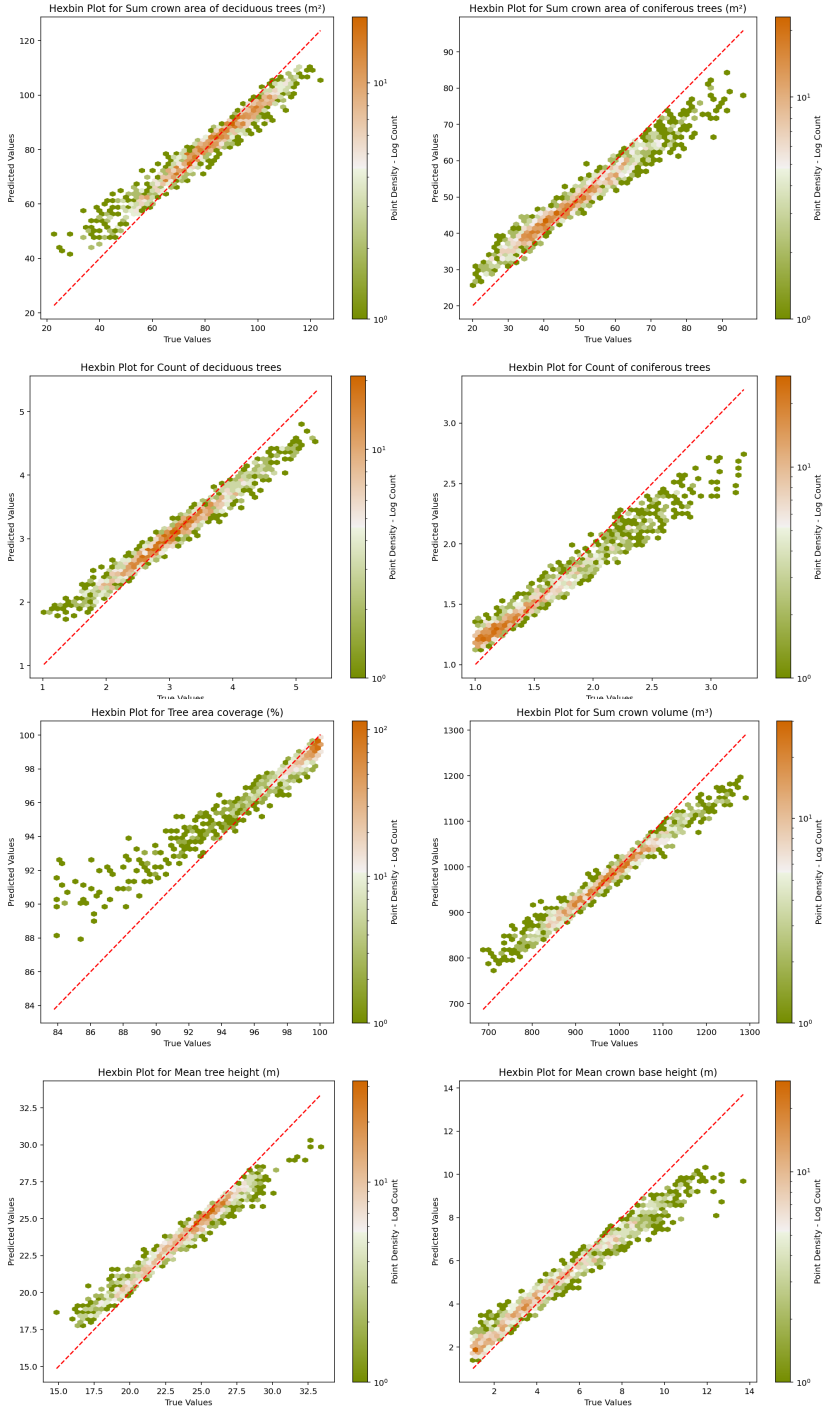


Figure 6.15.: RF regression results of a spectrally, polarimetrically, and temporally fused Sentinel-1 & -2 (© ESA, 2021) dataset in the Steigerwald study site (AOI 1), displaying predicted values using $K_{0,*}$ against actual reference values for each target variable, based on the best overall RF configuration (general-purpose setup, see Section 6.1.3). Point density is shown as a logarithmic count; the red dashed line represents perfect agreement. (a) sum crown area of deciduous trees [m^2], (b) sum crown area of coniferous trees [m^2], (c) count deciduous trees [amount], (d) count coniferous trees [amount], (e) tree area coverage [%], (f) sum crown volume [m^3], (g) mean tree height [m], and (h) mean crown base height [m].

Wald5Dplus

Following the findings presented in the preceding section, the modelling outcomes of the Wald5Dplus project are now introduced. Within the Wald5Dplus [147, 144, 148] project, RF was therefore also chosen as the best suitable regressor, by leveraging the described multi-modal, multi-temporal dataset, along with the python-based *scikit-learn* RF regression technique. The regression pipeline closely mirrors the experimental configuration detailed earlier, however, certain configuration parameters were adjusted to reflect operational constraints and practical optimization within the project scope. Within the Wald5Dplus modelling pipeline, the same RF multi-output regression strategy was adopted. In this setup, the RF algorithm is extended to simultaneously handle multiple response variables, training a dedicated ensemble for each forest attribute. This structure allows the model to account for variable-specific patterns and interdependencies more effectively. Beyond enhancing accuracy, this approach offers practical advantages: it consolidates the prediction task into a unified model architecture rather than requiring isolated models for each target, thereby improving computational efficiency and simplifying deployment in applied forestry contexts [86].

Feature engineering techniques were applied to enhance the quality of the dataset. Specifically, Z-score trimming is applied to address outliers prior to the training, involving the calculation of Z-scores for each input variable and applying a threshold (e.g., three standard deviations) to identify and remove outliers from the dataset. This step enhances the robustness and reliability of the predictions, ensuring a comprehensive and accurate assessment of tree-related attributes using our fused satellite dataset. To determine the most relevant variables for the model, a feature importance ranking was established to assess the significance of each feature in predicting the target variable. These steps collectively aim to improve the overall importance and interpretability of the model.

A compendium of RF regression plots juxtapose the predicted values against the actual values from the reference data for each target variable across all study sites. A distinctive characteristic of these plots is the logarithmic representation of point density, which elegantly enhances the visual portrayal of the data distributions. The intricacy and depth of these visualizations provide a unique perspective on the relationships between the model predictions and ground truth reference data.

Steigerwald Within this section, the Steigerwald study site (AOI 1) results are presented. The reference data [150] pertaining to this site were gathered in 2017 (Table 1.2), whilst

it should be noted that the satellite imagery and fusion transpired across the years 2020 and 2021. This temporal span, resulting in a substantial temporal discordance, necessitates thoughtful consideration during the interpretation of these findings.

In Figure 6.16 the RF regression results of a spectrally, polarimetrically and temporally fused Sentinel-1 & -2 (©ESA, 2021) dataset in the Steigerwald study site, displaying the predicted values using $K_{0,*}$ against the actual values for each present target variable are presented. In the evaluation of the RF regression model at the Steigerwald study site, the accuracy assessment metrics, encompassing MAE, MAD, and STD, provide insights into the model's performance. The Table A.1 summarizes these metrics for each target variable.

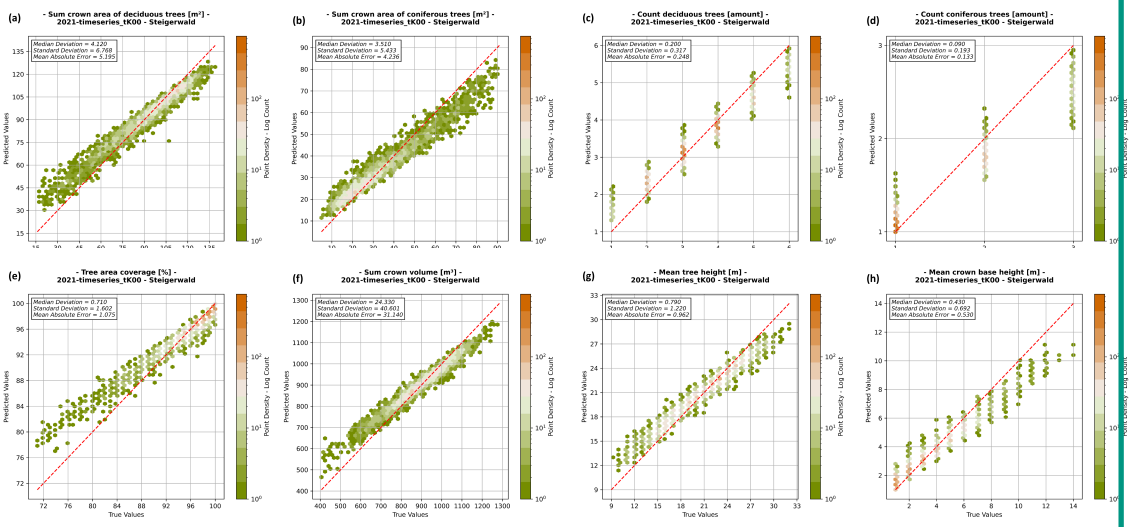


Figure 6.16.: RF regression results of a spectrally, polarimetrically, and temporally fused Sentinel-1 & -2 (©ESA, 2021) dataset in the Steigerwald study site (AOI 1), displaying the predicted values using $K_{0,*}$ against the actual values (i.e., reference data in Table 1.2) for each present target variable including the point density as logarithmic count and the perfect conditions (red dashed line); (a) sum crown area of deciduous trees [m^2], (b) sum crown area of coniferous trees [m^2], (c) count deciduous trees [amount], (d) count coniferous trees [amount], (e) tree area coverage [%], (f) sum crown volume [m^3], (g) mean tree height [m] and (h) mean crown base height [m].

The sum crown area of deciduous trees, the model's predictions exhibit a MAE of 5.195, indicating an average deviation of approximately 5.195 m^2 from the actual reference values. The MAD and STD values are 4.120 and 6.768, respectively. The actual values for this variable range from 15 to 135 m^2. Regarding the sum crown area of coniferous

trees, the model demonstrates a MAE of 4.326, with MAD and STD values of 3.510 and 5.433, respectively, indicating a robust predictive performance. The actual values for this variable fall within the range of 10 to 90 m^2. The count of deciduous trees variable showcases the model's accuracy with a MAE of 0.248, a MAD of 0.200, and a STD of 0.317, highlighting precise predictions with minimal deviations. The actual values for this variable range from 1 to 6 trees. The count of coniferous trees provides accurate predictions, with a MAE of 0.133 and MAD and STD values of 0.090 and 0.193, respectively. The actual values for this variable range from 1 to 3 trees. Demonstrating the model's capacity to estimate tree area coverage, this variable exhibits a MAE of 1.075. The MAD and STD are 0.710 and 1.602, respectively. The actual values for this variable range from 70% to 100%. In the case of sum crown volume, the model's MAE is 31.140, accompanied by a MAD of 24.330 and a STD of 40.601, signifying precision in estimating the sum crown volume. The actual values for this variable range from 400 to 1300 m^3. mean tree height is predicted with a MAE of 0.962, MAD of 0.709, and STD of 1.220, characterizing the model's accuracy. Actual values for this variable range from 9 to 33 m. Finally, for mean crown base height, predictions exhibit a MAE of 0.530, MAD of 0.430, and STD of 0.692, indicating robust predictive capabilities. Actual values for this variable range from 2 to 14 m.

These findings offer a comprehensive understanding of the model's capabilities, as well as insights into the actual value ranges and entities they represent, across a range of target variables at the Steigerwald study site.

Bavarian Forest National Park In this section, the findings within the Bavarian Forest National Park study site (AOI 2) results are presented. Reference data pertaining to this site were gathered in 2016 (Table 1.2). Similarly to the Steigerwald study site, the temporal discordance between satellite imagery and the reference data shall be noted. In this context it is also important to point out the bark beetle infestations in the last years [199], which may influence the results by the transition of coniferous forest to deadwood. This assumption is confirmed by spot checks during our field campaigns.

The RF regression results of the spectrally, polarimetrically and temporally fused Sentinel-1 & -2 (©ESA, 2021) dataset in the Steigerwald study site, displaying the predicted values using $K_{0,*}$ against the actual values for each present target variable are presented in Figure 6.17 as well as their respective accuracy assessment metrics are summarized in Table A.2.

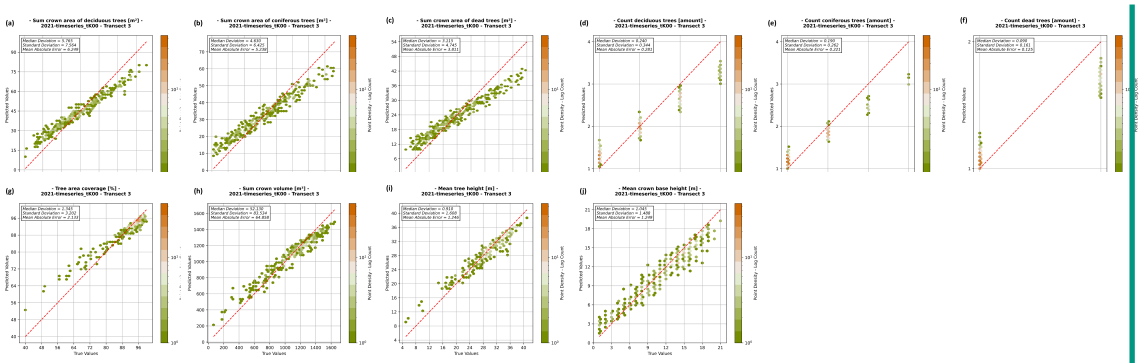


Figure 6.17.: RF regression results of a spectrally, polarimetrically, and temporally fused Sentinel-1 & -2 (©ESA, 2021) dataset in the Bavarian Forest National Park study site (AOI 2), displaying the predicted values using $K_{0,*}$ against the actual values (i.e., reference data in Table 1.2) for each present target variable including the point density as logarithmic count and the perfect conditions (red dashed line); (a) sum crown area of deciduous trees [m^2], (b) sum crown area of coniferous trees [m^2], (c) sum crown area of dead trees [m^2], (d) count deciduous trees [amount], (e) count coniferous trees [amount], (f) count dead trees [amount], (g) tree area coverage [%], (h) Sum crown volume [m^3], (i) mean tree height [m] and (j) mean crown base height [m].

For the variable sum crown area of deciduous trees, the model's predictions exhibit an average deviation from reference data of approximately 6.249 m^2, as reflected in the MAE. The MAD and STD values are 5.765 and 7.564, respectively. The actual range for this variable spans from 0 to 90 m^2. In the case of the sum crown area of coniferous trees, the model's strong predictive performance is evident with an MAE of 5.238. The MAD and STD values are 4.630 and 6.525, respectively, within an actual range from 0 to 70 m^2. For the sum crown area of dead trees, the model provides reliable estimates with a MAE of 3.811, coupled with MAD and STD values of 3.115 and 4.745. The actual range for this variable extends from 0 to 54 m^2. In predicting the of count deciduous and coniferous trees, variables with relatively low actual target ranges, the results yields low MAE of 0.281, with MAD and STD values of 0.240 and 0.344 and 0.221, with MAD and STD values of 0.190 and 0.262, respectively. Accurate predictions for count dead trees are evident with a MAE of 0.125, accompanied by MAD and STD values of 0.090 and 0.161. The actual range for this variable ranges from 1 to 2. The model effectively estimates tree area coverage, yielding an MAE of 2.133, expressed as a percentage. MAD and STD values are 1.345 and 3.202, with the actual range extending from 40% to 96%. For sum crown volume, the model excels with a MAE of 64.858, supported by MAD and STD values of 52.130 and 83.534. The actual range for this variable spans from 0 to 1600 m^3. Mean tree height is another variable where the model showcases its accuracy,

with an MAE of 1.246, along with MAD and STD values of 0.910 and 1.608. The actual range for this variable varies from 4 to 40 m. In predicting mean crown base height, the model offers reliable results, presenting a MAE of 1.249, accompanied by MAD and STD values of 1.045 and 1.488. The actual range for this variable extends from 0 to 21 m.

Kranzberg Forest This subsection refers to the last study site AOI 3, the Kranzberg Forest, whereas the temporal discordance between satellite imagery and the reference is the lowest. However, it is to be noted that this study site is the smallest in spatial extent, and therefore encompasses the lowest amount of trees in general. In the subsequent Figure 6.18 and Table A.3 the RF regression results of the spectrally, polarimetrically, and temporally fused Sentinel-1 & -2 (©ESA, 2021) dataset in the Kranzberg Forest study site, are shown. Note that the reflectance bands $K_{0,*}$ are used to predict the forest-related values.

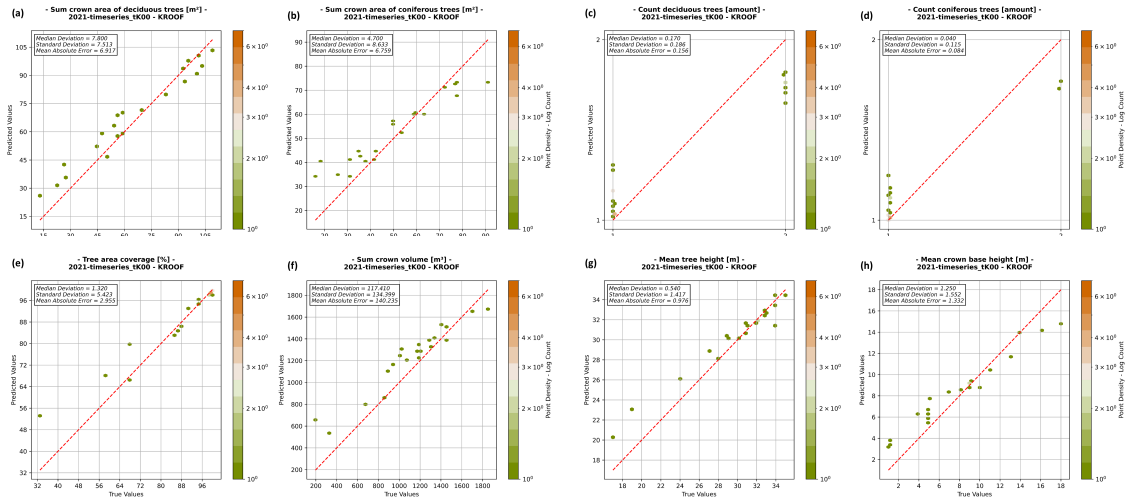


Figure 6.18.: RF regression results of a spectrally, polarimetrically, and temporally fused Sentinel-1 & -2 (©ESA, 2021) dataset in the Kranzberg Forest (AOI 3), displaying the predicted values using $K_{0,*}$ against the actual values (i.e., reference data in Table 1.2) for each present target variable including the point density as logarithmic count and the perfect conditions (red dashed line); (a) sum crown area of deciduous trees [m^2], (b) sum crown area of coniferous trees [m^2], (c) count deciduous trees [amount], (d) count coniferous trees [amount], (e) tree area coverage [%], (f) sum crown volume [m^3], (g) mean tree height [m] and (h) mean crown base height [m].

The model accurately estimates the sum crown area of both deciduous and coniferous trees. The MAE for these variables stands at 6.917 and 6.759 m^2, respectively. These

precise predictions align closely with the actual ranges, capturing the nuances of the forest canopy. As for the actual counts of the forest's trees, the model delivers low MAE values of 0.156 and 0.084 for deciduous and coniferous tree counts, respectively, staying within the range of 1-4. Regarding the Tree Area Coverage, featuring a MAE of 2.955, the model accurately estimates tree area coverage, with a mean absolute error of 2.955%. The MAD and STD values, 1.320 and 5.423, suggest minor deviations. Whilst the sum crown volume variable presents a comparatively high MAE of 140.235, complemented by a MAD of 117.410 and an STD of 134.399, this variable inherently encompasses a higher range, and the model's predictions harmonize effectively with the actual values. In the case of both mean tree height and mean crown base height, the model demonstrates commendable accuracy. The MAE values for these variables are 0.976 and 1.332 m, respectively. These values reflect a close alignment with the actual height ranges in the forest, which span from 4 to 40 m for tree height and 2 to 18 m for crown base height.

Ensemble Modelling Performance

Building on the consistent strength of RF regressors across prior experiments, a stacked ensemble modelling architecture was implemented to improve robustness and spatial generalization. This approach, detailed in Section 6.1.2, leverages spatially stratified RF base models, whose predictions are fused through a meta-level RF regressor. This design was chosen to balance predictive precision with architectural simplicity, mitigating overfitting risks while maintaining model interpretability across heterogeneous forest conditions.

Evaluation Strategy and Spatial Structure: The ensemble evaluation was carried out using spatial tiles defined for each AOI (see Table 1.1). Three evaluation streams were pursued:

1. **Deadwood Attribute Models (2-band)**: Trained on tiles `D01–D06` in the Bavarian Forest National Park , evaluated intra-regionally.
2. **Structural and Compositional Models (8-band)**: Applied to tiles `T01–T11` within the NP using models from Steigerwald, Kranzberg, and NP.
3. **Cross-AOI Transfer Tests**: Models from 2020 EO features were evaluated on 2021 EO data and across different AOIs to assess generalizability.

This tile-based framework allowed granular diagnostics of base-model performance while feeding diverse inputs into the stacked meta-learner.

Representative Model Transfer Results: The transfer performance of two key RF base models is summarized in Tables A.20 and A.21. The model trained on `SW_1_2020` showed strong generalization to its neighbouring sub-AOI (`SW_2`), but a pronounced performance drop when applied to the Bavarian Forest National Park , especially for volumetric and crown-related variables.

Conversely, the `NP_T10_2020` model performed well temporally (within NP 2021), but its spatial transfer to Steigerwald (`SW_2`) revealed clear domain mismatches, with elevated errors and increased variance for structural traits like crown area and volume. These outcomes demonstrate the limits of isolated model transfer across ecologically divergent forests and reinforce the necessity of stacked ensemble integration.

Stacked Ensemble Performance: A baseline stacked ensemble configuration was evaluated:

SW–SW Ensemble: Trained on Steigerwald sub-AOIs `SW_1, SW_2, SW_4, SW_5, SW_6`, evaluated within `SW_2` and transferred to `SW_3` (see Table A.22).

In this `SW-SW` scenario, MAEs for all variables decreased in the transfer setup, suggesting successful generalization within the same AOI (e.g., *mean crown base height*: MAE drop from 5.04 to 1.49).

Meta-Ensemble Construction and Generalization Insight: As described in Section 6.1.2, each base RF model contributed a prediction vector of size N (number of forest variables), yielding a combined feature space of dimension $M \times N$ (here, 5 base models $\times$ 10 targets). These were aggregated using a `FeatureUnion` and passed to a meta-level RF trained on the same reference data. This construction enabled the meta-model to learn correction weights and capture inter-model complementarities. Importantly, model selection for inclusion in the ensemble was governed by validation performance thresholds, ensuring only spatially reliable base models influenced final predictions. Hyperparameters for both base and meta RF were tuned via randomized search and cross-validation. The modular setup allows future integration of new spatial models

with minimal reconfiguration.The final stacked ensemble, was trained using five top-performing base RF models from distinct spatial units: tiles `D03` and `D04` (deadwood-specific, NP), `SW_1` and `SW_2` (Steigerwald), and `T10` (structural model, NP). Each of these was selected for either outstanding intra-AOI performance, strong generalization in transfer settings, or complementary ecological context (see Table 1.1). This ensemble was evaluated both intra-regionally (on the full Bavarian Forest National Park in 2020) and temporally (on the same region in 2021), representing a twelvefold scale-up from the 24,5 km^2 training extent (e.g., T10) to the full NP (294 km^2). The results, summarized in Table A.23, demonstrate strong spatial generalization and temporal resilience across most variables.

The final ensemble configuration demonstrated high robustness and transferability across both spatial and temporal domains. Intra-regionally (NP 2020), the ensemble achieved consistently low MAEs across all ten forest attributes. Temporal generalization to NP 2021, despite introducing ecological and seasonal variability, led to only moderate degradation for most variables. Notably, variables such as *mean crown base height* and *count of deciduous trees* exhibited near-stable behaviour across years, with MAE differences of just $\Delta = +0.10$ m and $\Delta = +0.05$ trees, respectively, well within acceptable margins considering their true value ranges (0–24 m and 0–9 trees). Similarly, *deadwood-related metrics*, including both *sum crown area* and *count of dead trees*, were predicted with high fidelity and showed only marginal temporal drifts (ΔMAE < 1).

Vertical structural indicators like *mean tree height* (ΔMAE $= +1.28$ m) and *tree area coverage* (ΔMAE $= +3.73$ %) remained highly stable in the face of inter-annual variation, supporting the meta-ensemble's ability to retain predictive sharpness for key ecological indicators. Even *crown area metrics*, which are known to fluctuate with phenology and canopy closure, remained within 3–5 m^2 MAE drift over time, a solid performance given their true range up to 170 m^2.

The only exception was *sum crown volume*, a compound metric with the largest dynamic range (0–3000 m^3), where temporal transfer resulted in a ΔMAE of $+62.29$ m^3. Despite this higher sensitivity, the relative change still represents less than $\sim$2 % of the upper range, indicating practical robustness for many monitoring applications. These results collectively underscore the ensemble model's capacity to generalize across distinct forest subtypes and acquisition years without retraining. The relatively stable error margins across structural and compositional variables, ranging from counts to canopy metrics, demonstrate the effectiveness of stacking spatially and thematically diverse RF base models into a unified, general-purpose predictive framework. Notably, this generalization

was achieved despite training on only $\sim$24.5 km^2 from T10 and localized deadwood tiles, yet deploying over the full Bavarian Forest National Park, confirming scalability across heterogeneous landscapes.

6.1.4 Discussion

Building on the comprehensive evaluation of forest structural and compositional variable predictions using multi-modal EO data, this discussion synthesizes key findings regarding model behaviour, generalization capacity, and methodological trade-offs. The results from both individual and stacked RF models highlight important patterns in model performance, sensitivity to preprocessing, and robustness across spatial and temporal domains. In particular, the effectiveness of the stacked ensemble approach, anchored in a spatially stratified RF framework, provides insights into the challenges and opportunities of scalable forest attribute prediction. The following subsections reflect on these aspects, evaluate the strengths and limitations of the implemented methods, and outline directions for further refinement.

Model Performance and Generalization Behaviour

Overall Model Performance: Across all evaluated forest structural variables and preprocessing configurations, RF regressors consistently emerged as the most robust and adaptable model family. In intra-AOI tests, RF delivered the lowest MAE for multiple variables and ranked among the top-performing configurations in the majority of scenarios. Notably, their performance remained resilient in cross-AOI evaluations, where predictive accuracy typically declines due to ecological dissimilarities, further confirming RF's suitability for modelling the forest parameters under diverse spatial and sensor conditions. This robustness is likely attributable to the model's capacity to capture non-linear interactions and its insensitivity to raw feature scales. The RF's operational simplicity and interpretability, combined with its consistent accuracy, make it particularly appealing for large-scale forest monitoring where more complex models like 1D-CNN failed to generalize. In comparison, neural models collapsed under domain shift, reaffirming the trade-off between flexibility and interpretability in favour of RF.

Effect of Preprocessing on Model Behaviour: Preprocessing strategies had a measurable effect on predictive accuracy, particularly in interaction with model type and target

variable. RF models demonstrated optimal results under moderate z-score filtering ($Z = 3$), light-to-moderate masking thresholds (`Mask > 1`), and when aggressive outlier trimming was avoided. This setup preserved ecologically informative variance, especially important for volume-related metrics, while suppressing noisy extremes. Moreover, preprocessing choices must be balanced to avoid excessive feature pruning, particularly in crown area or volume estimation, where outlier suppression can eliminate meaningful ecological extremes. This underlines the need for preprocessing strategies that are both robust and ecologically sensitive. Notably, 1D-CNN models showed higher sensitivity to normalization and input format but failed to generalize well under transfer conditions, underscoring the comparative robustness of tree-based methods to domain shifts.

Variable-Specific Trends and Model Suitability: Performance patterns varied systematically by variable type. For structural attributes such as *mean tree height* and *mean crown base height*, RF achieved consistently low MAE and RMSE, even under cross-AOI scenarios. These differences may also reflect varying sensitivities of forest attributes to EO modality and resolution, for instance, height proxies being well-aligned with canopy reflectance and SAR structure, while crown area and volume are more affected by mixed pixels and registration noise. In contrast, compositional and crown-based metrics showed greater variability, particularly under aggressive preprocessing. Linear and SVR models demonstrated strong performance in isolated cases but suffered from limited generalization and higher volatility. Overall, ensemble-capable, non-parametric models such as RF proved to be the most versatile.

Spatial Generalization and Transfer Modelling

Performance in Cross-AOI Domains: Individual RF models, although effective within their training AOIs, displayed considerable variability when transferred to new domains. Tables A.20 and A.21 illustrate that while base models like `SW_1` and `T10` perform well locally, their transfer accuracy to different AOIs varies significantly. This is particularly evident in structurally complex features such as crown volume and deadwood metrics. These findings underscore the limited portability of isolated models and support the use of more inclusive strategies. While predictive performance declined under cross-AOI scenarios, the stacked ensemble model consistently outperformed the individual base models in transfer settings, highlighting its capacity to integrate diverse structural signals and mitigate localized overfitting. This is particularly evident in Table A.22, where the Steigerwald-based ensemble achieved even stronger results on the unseen SW_3 tile than

within its own training domains. Such outcomes highlight that well-designed ensembles can not only generalize but also leverage regional complementarity to enhance prediction. This reinforces the argument for spatially-aware model design. Tiling the input domain and distributing base models ensures that regional specialization is preserved while maintaining an ensemble-level ability to generalize beyond any one landscape.

Stacked Ensemble Design and Meta-Model Effectiveness:

A more comprehensive ensemble, incorporating models from both the Bavarian Forest National Park (`D03`, `D04`, `T10`) and Steigerwald (`SW_1`, `SW_2`), was evaluated using a full 10-band input feature set. Table A.23 presents results from this ensemble under both intra-AOI (NP 2020) and cross-year (NP 2021) settings. Contrary to expectations, temporal generalization between 2020 and 2021 yielded only modest declines across most forest attributes, indicating that the ensemble can robustly handle year-to-year variability. However, compared to the best locally optimized models (Tables A.1, A.3, and A.7–A.11), the ensemble showed slightly higher per-variable MAE, highlighting the inherent trade-off between specialization and generality. Nonetheless, when transferred into unseen spatial or temporal domains, the ensemble significantly outperformed individual models (Tables A.12–A.19). This advantage validates ensemble stacking as a viable mechanism to bridge ecological and acquisition-based variability, and further demonstrates the utility of combining spatially distributed base models to achieve scalable forest attribute prediction. Importantly, the ensemble leveraged only a fraction of the spatial footprint available in the NP, e.g., T10 represents $\sim$24.5 km^2 compared to the full 294 km^2 NP, and still generalized effectively, highlighting the meta-model's capacity to extrapolate from sparse but representative training contexts.

Operational and Methodological Implications

Robustness to Label Aging and EO Variability: A notable challenge addressed in this study was the temporal mismatch between reference data (2016–2018) and input EO features (2020–2021). The ensemble model's ability to produce plausible and stable predictions under these conditions underscores its robustness to label aging. This tolerance to temporal lag demonstrates the ensemble's applicability in real-world EO pipelines

where field inventories may lag behind satellite acquisitions. This is particularly relevant in forest monitoring, where ground data are rarely contemporaneous with satellite observations.

One of the challenges encountered in this study was the temporal discrepancy between the satellite data and the reference data. While Sentinel-1 and Sentinel-2 data from 2020 and 2021 were utilized, the reference data originated from 2016 (AOI 2), 2017 (AOI 1), and 2020 (AOI 3). This temporal offset raises questions about the accuracy and relevance of the reference data due to potential changes in forest conditions over time. It is crucial to consider the impact of these temporal differences on the predictions. Variations in tree counts between the model and the reference data could be attributed to multiple factors, such as tree growth rates, forest management practices, seasonal fluctuations, or external influences like disease outbreaks or natural disasters. The investigation into the reasons behind these disparities is a critical aspect of improving our model's predictive performance. It highlights the need for more frequent updates of reference data to maintain the accuracy and relevance of such datasets. Nevertheless, our field campaigns confirmed a high accordance of satellite and reference data, i.e., only little variations for deciduous forests. The situation regarding more or less purely coniferous forest stands especially in the Bavarian Forest National Park (AOI 2) is different. Some of the areas identified as coniferous in the reference data are now characterized by deadwood because of a meanwhile bark beetle infestation. In contrast, some areas marked as deadwood are now covered by young growth. In comparison to the entity of the labeled ARD cubes, the proportion of unclear labels is extremely low or even negligible. Potential outliers are reliably identified by the RF regression so that the prediction based on satellite data shows a higher accordance with the actual state than to be expected from the reference dataset.

Modularity and Scalability of the Ensemble Approach: From an operational perspective, the stacked ensemble architecture supports scalable and modular implementation. New base models can be added incrementally, and the meta-model retrained accordingly, allowing for seamless integration of new data sources or geographic areas. This makes the system well-suited for near-real-time forest attribute prediction and long-term monitoring. In practice, this architecture enables continuous improvement: new data sources can be added, poor-performing base models excluded, and ensemble adaptation carried out with minimal retraining overhead, thereby supporting operational scalability.

Limitations and Future Directions

Model Architecture Trade-offs and Future Directions: The exclusive use of RF in both base and meta-model layers provided consistent and robust performance across all forest variables. This choice also reflects deliberate trade-offs in favour of interpretability, computational efficiency, and stability across spatial domains. The decision to rely solely on RF, rather than heterogeneous base models, ensured operational scalability and minimized training complexity across the ensemble framework. While RF provide strong baseline performance, their limitations in expressing highly non-linear interdependencies may be addressed by hybrid ensembles in future work, blending RF with, for example, gradient boosting or neural meta-models, if interpretability can be maintained. While the stacked ensemble model demonstrated strong predictive performance and robust generalization across spatial and temporal domains, several methodological extensions could further enhance accuracy, interpretability, and transferability:

First, the current stacking implementation treats all base model predictions equally in the meta-feature space. As the number of output bands varies across base models (e.g., 2-band vs. 8-band predictions), this can lead to unbalanced contributions to the meta-learner. Future work could incorporate normalization strategies such as z-score standardization of each base model's outputs prior to stacking, ensuring equitable representation across models.

Second, while the current meta-learner is trained using predictions from base models applied to a unified dataset, care was taken to ensure that no model made predictions on its own training tiles. This design reduces the risk of information leakage. However, future work could further strengthen this separation by adopting a strict out-of-fold (OOF) stacking approach. In such a scheme, base model predictions used to train the meta-learner would be generated exclusively from tiles held out during each model's training phase. This would ensure complete independence between base model training and meta-model input construction, thus enforcing a more rigorous ensemble learning framework.

Third, alternative meta-learner algorithms, such as regularized linear models (e.g., Ridge, ElasticNet), SVR, or Gradient Boosted Trees, could be explored. These may offer better interpretability or reduce overfitting risks compared to RF at the second level.

Finally, future evaluations should explicitly assess the generalization capacity of the meta-model under true cross-regional transfer conditions. While current experiments focus

on sub-AOI tiling within the same broader regions, further validation is needed using entirely independent and geographically distinct AOIs, ideally sourced from different forest types, biomes, or countries. A Leave-One-AOI-Out (LOAO) validation scheme, applied at the level of distinct ecological regions rather than tiles, would provide a more rigorous and ecologically realistic assessment of model generalization. This is particularly important for operational forest monitoring tools designed to function in data-scarce environments where reference data may not be available for the region of interest. Additionally, future work could incorporate model interpretability frameworks (e.g., SHAP values, permutation importance) to quantify the relative contribution of each base model and to provide insights into regional specialization or redundancy within the ensemble. Such approaches would not only improve trust and transparency but also inform targeted model pruning or adaptation strategies for deployment in new forest contexts.

Taken together, these refinements would further strengthen the meta-model's capability as a transferable forest monitoring tool, particularly in scenarios where updated field data are unavailable or ecologically outdated. Nonetheless, the current RF-only ensemble demonstrates that high predictive performance and strong transferability can be achieved with a single, interpretable model family, making it suitable for scalable forest monitoring applications with constrained computational resources.

Spatial Resolution and Label Uncertainty: Despite robust generalization, the model operates at the pixel level, where noise from georegistration, mislabeling, and mixed-pixel effects can affect performance. Future efforts could incorporate object-based or region-based learning frameworks to mitigate this.

Temporal Generalization Beyond Two Years: While the model generalized well between 2020 and 2021, longer-term temporal drift was not evaluated. Future research should test ensemble stability across broader interannual windows and under post-disturbance dynamics.

Model Interpretability and Variable Importance: While RF offer built-in feature importance metrics, the ensemble architecture complicates attribution. Further work could explore SHAP-based analysis to interpret meta-model decisions and identify dominant spatial predictors across variables.

6.2 Polarimetrically and Spectrally Fused Sentinel-1 and Sentinel-2 Data

To evaluate the performance of only multi-sensor data (sans multi-temporality) integration, this Section explores a fused feature space that combines Sentinel-1 polarimetric and Sentinel-2 spectral information. The fusion is implemented using the HCB method [289], which offers an orthogonal and interpretable representation of multi-modal EO features. This approach builds directly on the foundations established in Sections 2.2 and 2.2, where the respective polarimetric and spectral Kennaugh (or Kennaugh-like) elements were introduced. Their joint integration using HCB is further formalized in Section 2.2, highlighting the mathematical advantages of this method for compact, lossless fusion across modalities.

6.2.1 Materials

The EO inputs originate from the Steigerwald study area (AOI 1) and are temporally aligned using acquisition dates from 2 July 2020 (for within-domain modelling) and 3 July 2021 (for cross-validation and spatial transfer scenarios).

Two Sentinel-1 C-band SAR scenes were selected for analysis: 2 July 2020 and 3 July 2021, both corresponding to peak summer vegetation conditions. The acquisitions were obtained in dual-polarization mode (VV/VH), which is particularly suitable for capturing both surface and volume scattering properties in densely vegetated environments. To derive physically meaningful backscatter descriptors, the raw Sentinel-1 SLC data were processed through the Multi-SAR framework [38]. This included standard preprocessing steps such as speckle filtering, radiometric normalization, and geocoding, followed by transformation into Kennaugh elements k_0, k_1, k_5, k_8 [288], as described in Section 2.2. These steps are schematically illustrated in Figure 6.19.

Sentinel-2 multispectral acquisitions covering the Steigerwald study area (AOI 1) were acquired within the Wald5Dplus project, coinciding with the Sentinel-1 acquisition dates (2nd of July 2020 and 3rd of July 2021), corresponding to peak summer vegetation conditions. The data were processed with the MAJA [85] atmospheric correction chain and include the 10 m resolution red, green, blue, and near-infrared (NIR) bands. These bands provide critical information on vegetation structure and health, and served directly

as input features for the subsequent regression modelling. These bands were further processed into Kennaugh-like elements, following the methodology outlined in Section 2.2. This transformation, applied to these 10 meter resolution bands, separates brightness information from spectral information by projecting the data into a hypercomplex basis, as expressed in Equation (2.5). This representation is conceptually similar to traditional True Color Images (TCI) and Color Infrared (CIR) compositions but offers a more structured decomposition of the spectral signal. By explicitly decoupling intensity and colorimetric properties, the Kennaugh-like features allow a direct comparison with the SAR-based Kennaugh elements used in preceding and succeeding in this thesis. Moreover, evaluating both the raw and transformed optical inputs provides insights into the relative advantages of classical versus feature-engineered predictors in forest variable regression.

As detailed in Section 2.2, the spectral and polarimetric information from Sentinel-2 and Sentinel-1, respectively, are combined using the HCB transformation [289]. This fusion results in an 8-dimensional feature space consisting of one total intensity channel ($K_{\text{fused},0}$) and seven orthogonal spectral–polarimetric elements ($K_{\text{fused},1-7}$). The transformation is lossless and interpretable, preserving both signal domains while enabling compact, semantically rich feature representations for downstream modelling. The fusion process is schematically introduced in Figure 6.19.

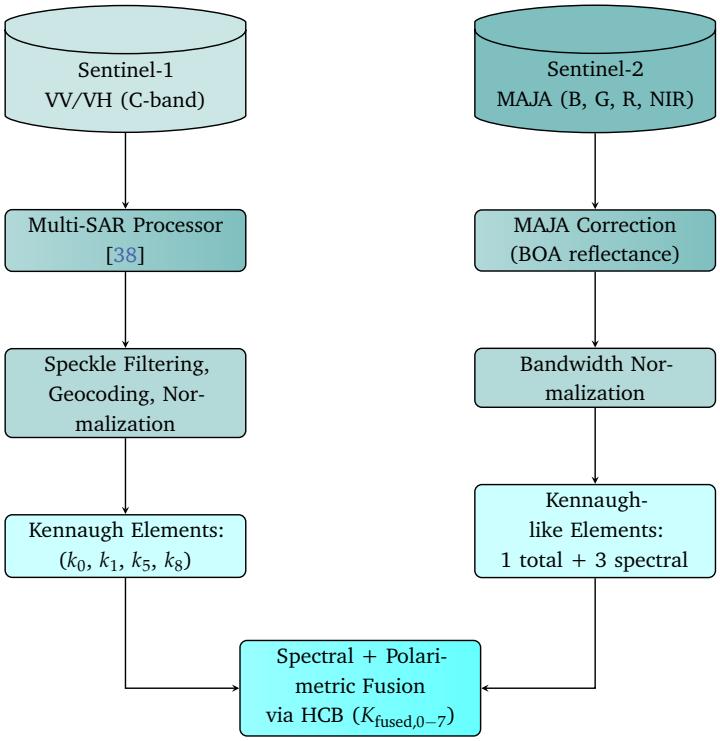


Figure 6.19.: Fusion pipeline up to the combination of Sentinel-1 and Sentinel-2 Kennaugh representations using the Hypercomplex Bases (HCB) method.

The label data used in this experiment originates again from the Wald5Dplus dataset, from the Steigerwald study area (AOI 1), see Section 1.2.2.

6.2.2 Methods

To maintain direct comparability wihtin this whole Chapter, the evaluation approach outlined in Section 6 was consistently applied. This included identical preprocessing configurations (masking thresholds, z-score filtering, and aggressive filtering), the same set of ML models (RF, SVR, Linear Regression, and 1D-CNN), and both intra- and cross-AOI validation strategies. Specifically, intra-AOI analyses were conducted using polarimetrically and spectrally fused Sentinel-1 and Sentinel-2 Data from 2020 (SW_2_2020), while cross-AOI generalization was assessed using fused scenes from 2021 (SW_4_2021).

6.2.3 Results

This section presents the modelling outcomes obtained from the fused Sentinel-1 and Sentinel-2 dataset, where polarimetric and spectral features were integrated via the HCB framework [289]. The evaluation focuses on eight forest structural variables derived from the Wald5Dplus reference dataset, assessed through intra-AOI validation. Each variable was modelled across a consistent configuration grid, enabling direct comparison of regression accuracy between fused and unimodal setups. The results illustrate the predictive potential of the fused feature space, highlighting variable-specific trends, model behaviour, and the effectiveness of the HCB-based integration strategy.

Model Evaluation per Variable

Table references correspond to results from the original spatial domain, in Tables A.24 to A.31, while corresponding performance shifts are detailed in Tables A.32 to A.39.

Overall Model Performance: RF regressors demonstrated the strongest performance across the fused feature space, achieving the lowest mean absolute error (MAE) and root mean squared error (RMSE) across most variables. Models using strict preprocessing (Mask $>$ 1, Z=1, aggressive outlier removal) systematically outperformed relaxed setups, confirming the importance of stringent data cleaning even when leveraging rich multi-modal inputs.

Sum of Crown Area of Deciduous Trees (m^2): Top-performing RF models achieved MAE values as low as 3.29 m² in the original spatial domain. However, transfer to unseen areas increased the MAE substantially, with typical degradation exceeding +13 m², highlighting the challenge of generalizing crown area predictions even with fused data.

Sum of Crown Area of Coniferous Trees (m^2): Similar patterns were observed for coniferous crown areas, where RF yielded in-domain MAE values around 2.96 m². Cross-validation exposed moderate-to-high error increases (+8.7 m² MAE), indicating a relatively better, but still imperfect, generalization compared to deciduous crown area.

Count of Deciduous Trees: Models achieved highly accurate in-domain predictions with MAEs around 0.15 trees. Spatial transfer resulted in an increase of +0.6 trees in MAE, reflecting moderate generalization challenges, but maintaining reasonable stability relative to area- or volume-based variables.

Count of Coniferous Trees: Coniferous tree count models performed strongly in-domain (MAE $\sim$0.09 trees), but cross-AOI shifts led to pronounced MAE increases ($\sim$+1.2 trees), indicating greater sensitivity to domain changes, possibly linked to the subtler spectral and structural signatures of conifers.

Tree Area Coverage (%): Tree area coverage models achieved low MAEs ($\sim$0.37–0.39 %) under optimal preprocessing. However, transfer tests showed significant MAE increases (+2.2 %), reinforcing the tendency of canopy coverage metrics to degrade under spatial variability despite fused feature spaces.

Sum of Crown Volume (m^3): Volume estimates were particularly challenging, with in-domain MAEs around 22–23 m^3. Transfer induced severe degradations, with MAE increases exceeding +67 m^3, demonstrating that crown volume remains a highly sensitive variable even under fused optical and SAR modelling.

Mean Tree Height and Mean Crown Base Height (m): Models predicted mean tree height with MAEs below 0.51 m and mean crown base height around 0.64 m. Transfer performance deteriorated for both (+1.4 m and +1.3 m MAE, respectively), yet these variables exhibited comparatively lower relative degradation, indicating more stable generalization behaviour than area- and volume-based attributes.

Model Performance and Preprocessing Effects

Preprocessing Trends: Strict masking (Mask $>$ 1), moderate Z-score clipping (Z=1), and aggressive outlier removal consistently improved model performance across variables. Relaxed preprocessing setups led to markedly higher errors, especially under domain shift, confirming that data conditioning remains critical even when feature richness increases through fusion.

Model Diversity and Robustness: RF continued to outperform alternative learners, with no substantial advantage observed from incorporating model diversity. Homogeneous RF ensembles maintained superior predictive sharpness and transfer stability, suggesting that simplicity and consistency outweigh diversity in fused EO-based regression tasks.

Spatial Generalization and Transfer Behaviour

Cross-Validation Results:

Spatial transfer testing revealed systematic performance degradation across all models and variables, with the extent of decline varying:

- **Volume and Crown Area Variables:** Largest transfer degradation observed (e.g., +67 m^3 MAE for crown volume; +13–14 m^2 MAE for crown area).
- **Tree Counts:** Moderate degradation (+0.6 to +1.2 trees MAE), reflecting relatively better cross-region resilience.
- **Height Metrics:** Lowest relative performance declines (+1.4 m for mean tree height, +1.3 m for crown base height), suggesting vertical structural attributes are more transferable under multi-modal feature spaces.

Despite the increased feature richness provided by polarimetric and spectral fusion, substantial challenges remain in achieving robust cross-AOI generalization, particularly for complex structural variables such as crown area and volume. Height and tree count variables exhibited more resilient behaviour, suggesting that the benefits of fusion are variable-specific and that additional strategies such as domain adaptation may be necessary to realize fully transferable forest structure models.

6.2.4 Discussion

Building on the performance insights from the fused Sentinel-1 and Sentinel-2 experiments, this discussion synthesizes the implications of multi-modal feature integration for forest parameter modelling.

Model Performance and Generalization Behaviour

Model Performance: RF regressors continued to demonstrate strong predictive stability across the fused feature space, achieving superior or near-superior performance on all forest variables. The ensemble's inherent resilience to overfitting and its capacity to accommodate heterogeneous input features (polarimetric SAR and optical-spectral) reaffirmed its suitability as a backbone for complex multi-modal EO regression tasks.

Effect of Preprocessing on Model Behaviour: Strict preprocessing protocols, involving conservative masking thresholds, moderate Z-score filtering, and aggressive outlier

removal, remained crucial. Although the fused feature set inherently carried richer information, models remained highly sensitive to preprocessing choices. Lax preprocessing resulted in marked error increases, particularly for structural variables such as crown area and crown volume, emphasizing that feature richness cannot substitute for careful input conditioning.

Variable-Specific Trends and Model Suitability: Fused modelling improved in-domain accuracy across all variables. However, variable-specific trends persisted: tree height and crown base height exhibited higher spatial robustness, while crown volume and sum crown area remained more vulnerable to transfer degradation. The fusion of SAR structure-sensitive and optical spectral features proved particularly beneficial for tree counts and vertical metrics but offered only partial mitigation for crown-based variables, which remain susceptible to regional shifts in canopy complexity and spectral heterogeneity.

Spatial Generalization and Transfer Modelling

Performance in Cross-AOI Domains: Spatial transfer tests highlighted persistent performance degradation across all variables. While fusion improved the relative stability of models compared to previous setups (single-sensor), large declines in predictive accuracy for complex aggregation variables like crown volume suggest that even fused models are not immune to domain shift effects. Mean tree height and mean crown base height again showed the smallest relative error increases, suggesting that vertical structure remains more transferable across domains.

Model Setup and Stability: Models with conservative preprocessing and purely RF backbones consistently outperformed diverse or relaxed configurations in transfer scenarios. This consistency highlights that, despite feature fusion, model setup choices (particularly ensemble depth, feature subsampling strategies, and leaf size) play a defining role in ensuring spatial robustness.

Operational Implications and Transferability: The fusion of polarimetric and spectral information enhanced baseline model accuracy and provided modest improvements in transfer resilience, particularly for height and count metrics. However, the persistent vulnerability of crown area and volume variables indicates that additional strategies, such as domain adaptation techniques or semi-supervised learning with target-domain

data, may be required for fully operational cross-regional generalization. Overall, multimodal fusion supports more scalable and flexible forest monitoring approaches but cannot entirely eliminate the structural complexities and variability challenges inherent in heterogeneous forested landscapes.

Additionally, the results reinforce three broader points worth further consideration:

- **Multimodal Data Fusion is Beneficial but Not Sufficient:** While HCB-based fusion enhances predictive power in the training domain, it does not fully resolve the generalization challenges posed by spatial heterogeneity. The resilience of vertical metrics versus the volatility of volume-based attributes suggests that certain forest variables are intrinsically more amenable to EO-based modelling, even under fusion.
- **Model Reliability is Closely Tied to Data Conditioning:** Across all variables, strong preprocessing (masking, outlier trimming) consistently improved performance. This emphasizes that rich features cannot substitute for clean input data, especially when aiming for generalization across AOIs.
- **Future Potential Lies in Combining Fusion with Domain Adaptation:** To move toward operational scalability, models may need to integrate fusion with transfer learning strategies. Semi-supervised or domain-adaptive frameworks could mitigate domain shift, especially for structurally complex forest attributes.

6.3 Reflectance Bands and Spectral Kennaugh-like Elements from Sentinel-2 Data

This section explores the predictive potential of optical features derived from Sentinel-2 data for modelling forest structural attributes. Both raw spectral features and derived Kennaugh-like elements are evaluated for their utility in capturing forest parameters.

6.3.1 Materials

Sentinel-2 multispectral acquisitions covering the Steigerwald study area (AOI 1) were acquired within the Wald5Dplus project [147]. The specific acquisition dates and prepro-

cessing configurations (e.g., the transofrmation of Sentinel-2 reflectances into Kennaugh-like elements) are detailed in Section6.2, within the materials description of the preceding *Polarimetrically and Spectrally Fused Sentinel-1 and Sentinel-2 Data* setup.

In this setup, two separate feature streams were extracted: (i) the raw reflectance values of the four Sentinel-2 bands (R, G, B and NIR), and (ii) their transformation into Kennaugh-like elements using hypercomplex algebra to derive interpretable spectral–structural components. This dual-path pipeline enables a direct comparison between traditional spectral features and fused, orthogonally decomposed representations, as illustrated in Figure 6.20.

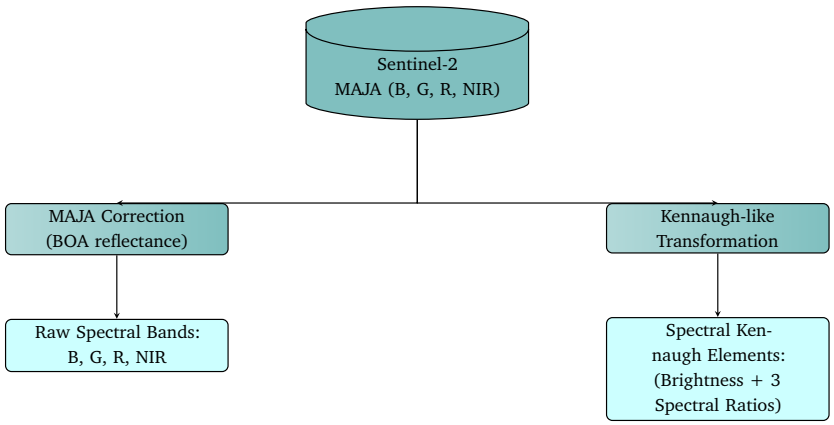


Figure 6.20.: Parallel feature extraction pipelines from Sentinel-2 MAJA data. The 10 m resolution Blue, Green, Red, and Near-Infrared (NIR) bands were used directly as raw input features, and alternatively transformed into Kennaugh-like elements capturing brightness and spectral structure. Both data streams are derived independently from Sentinel-2 acquisitions (2 July 2020 and 3 July 2021) over AOI 1 (Wald5Dplus project), supporting comparative modelling experiments.

Labels: For this experiment, the continuous labels from the Wald5Dplus project [148] were used, consistent with the Sentinel-1 polarimetric Kennaugh element experiment, see Section 6.4, to ensure comparability. The label derivation process is detailed in Section 1.2.2.

6.3.2 Methods

The same standardized evaluation setup as described in Section 6 was employed to ensure comparability with all Sections. This includes consistent preprocessing variations

(masking thresholds, z-score filtering, aggressive filtering), a unified model comparison (RF, SVR, Linear Regression, and 1D-CNN), and both intra-AOI and cross-AOI validation protocols. Importantly, intra-AOI evaluation was based on Sentinel-2 acquisitions from 2020 (SW_2_2020), while cross-AOI transferability was tested on independent Sentinel-2 data from 2021 (SW_4_2021). Adopting this shared and temporally consistent framework allows a direct, controlled assessment of how Sentinel-2 based inputs perform relative to Sentinel-1 derived features under identical modelling and evaluation conditions.

6.3.3 Results

This section presents the comparative model evaluation results for two distinct Sentinel-2 feature representations: raw multispectral reflectance bands and their transformed counterparts expressed as Kennaugh-like spectral elements. Both feature sets originate from the same Sentinel-2 acquisitions but differ in their composition, one reflecting the direct spectral signal, the other an orthogonally decomposed version emphasizing brightness and spectral contrast components.

Model Evaluation per Variable

Table references correspond to results from the original spatial domain, of the Raw Sentinel-2 Data in Tables A.40 to A.47, and the Spectral Kennaugh-like Elements from Sentinel-2 Data in Tables A.56 to A.63, while corresponding performance shifts are detailed regarding the Raw Sentinel-2 Data in Tables A.48 to A.55 and the Spectral Kennaugh-like Elements from Sentinel-2 Data in Tables A.64 to A.71.

Overall Model Performance:

In this experiment, two configurations of Sentinel-2 based data were evaluated: (1) the original Sentinel-2 raw spectral bands and (2) Spectral Kennaugh-like elements derived from the same Sentinel-2 acquisitions.

For both configurations, RF Regressors consistently yielded the best predictive performance across most forest structural variables, demonstrating high robustness across different masking and preprocessing setups.

Sentinel-2 Raw Data results showed strong performance particularly for predicting tree counts and tree area coverage, with lower mean absolute errors (MAE) and root mean square errors (RMSE) compared to volume- and height-related variables.

Spectral Kennaugh-like Elements maintained comparable predictive capabilities, in some cases improving slightly over raw Sentinel-2 for predicting tree counts and tree area coverage. However, prediction of volume-related variables such as sum crown volume exhibited slightly higher variability compared to raw bands.

Across both input data types, models with aggressive preprocessing (mask thresholding and Z-score normalization) consistently outperformed their counterparts without preprocessing, highlighting the importance of feature standardization and noise reduction in the prediction task.

Overall, the models showed a tendency towards better performance for variables related to crown area and tree counts, while volume and height predictions remained more challenging.

Sum of Crown Area of Deciduous Trees (m^2): The prediction of the sum of crown area of deciduous trees revealed high consistency across both data types. For the Sentinel-2 raw spectral bands, the best-performing models achieved MAE values around 3.86 to 3.89 m^2, and RMSE values of approximately 4.83 to 4.89 m^2. The application of masking and Z-score normalization notably contributed to this performance.

Similarly, models utilizing the Sentinel-2 Spectral Kennaugh-like elements achieved slightly lower MAE values, approximately 3.63 to 3.64 m^2, and RMSE values around 4.64 to 4.65 m^2. This indicates a marginal improvement in prediction accuracy when employing Spectral Kennaugh-like representations.

Cross-validation revealed a systematic performance drop for both data types, with an increase in MAE by approximately 12 to 13 m^2 for the best models. However, the relative degradation remained comparable between raw spectral inputs and Spectral Kennaugh-like inputs, suggesting similar spatial transferability characteristics.

Overall, both Sentinel-2 raw and Spectral Kennaugh-like elements demonstrated strong predictive potential for estimating the sum of crown area of deciduous trees, with the latter offering a slight advantage in the original spatial domain.

Sum of Crown Area of Coniferous Trees (m^2): Model performance for the sum of crown area of coniferous trees also demonstrated strong results across both Sentinel-2 data

representations. Using the raw spectral bands, the best-performing models achieved MAE values around 3.23 to 3.27 m^2 and RMSE values between 4.13 and 4.29 m^2.

In comparison, the models based on Spectral Kennaugh-like elements yielded similar MAE values between 3.18 and 3.23 m^2 and RMSE values ranging from 3.98 to 4.03 m^2. Slight improvements in RMSE indicated that the Spectral Kennaugh-based features enhanced the stability of the predictions, albeit marginally.

Cross-validation exposed substantial declines in predictive accuracy for both setups, with MAE increases of approximately 7 to 8 m^2 for models employing Spectral Kennaugh-like features, and somewhat larger degradation for raw band-based models. Nonetheless, Spectral Kennaugh-like features demonstrated slightly more robust generalization behaviour across spatial domains.

Thus, the Spectral Kennaugh-like representation again provided a minor performance advantage over the direct use of raw Sentinel-2 bands.

Count of Deciduous Trees: Predictive performance for the count of deciduous trees revealed consistent patterns across both Sentinel-2 data representations. Models trained on raw Sentinel-2 bands achieved top MAE values between 0.185 and 0.236 trees and RMSE values around 0.227 to 0.285.

In contrast, models utilizing the Spectral Kennaugh-like elements slightly improved the accuracy, achieving MAE values between 0.175 and 0.199 and RMSE values ranging from 0.216 to 0.244. The differences were relatively subtle but consistently favoured the Kennaugh-like feature representation in the original spatial domain.

Cross-validation results showed typical performance degradation, with MAE increases around 0.59 trees for the best Spectral Kennaugh-based models compared to 0.54–0.56 for raw band models. Despite this, the Spectral Kennaugh-based models retained marginally lower absolute errors during spatial transfer.

Overall, the Spectral Kennaugh-like elements maintained a slight advantage in estimating deciduous tree counts, particularly in terms of robustness across spatial domains.

Count of Coniferous Trees: Modelling the count of coniferous trees exhibited similar trends to the deciduous tree results. In the original spatial domain, raw Sentinel-2 band models achieved MAE values between 0.095 and 0.127, while RMSE values ranged from 0.126 to 0.160. The Spectral Kennaugh-like elements delivered nearly identical

performance, with MAE values from 0.095 to 0.124 and RMSE values around 0.126 to 0.154.

Cross-validation analysis revealed performance declines for both approaches, with MAE increases between approximately 0.9 and 1.2 trees. Models trained on raw Sentinel-2 bands displayed slightly higher error increases than those using Spectral Kennaugh-like features.

The Spectral Kennaugh-based models maintained a marginal advantage in spatial generalization, achieving slightly lower cross-validation MAE and RMSE values overall. Nonetheless, the differences between raw bands and Kennaugh elements remained small for the count of coniferous trees.

Tree Area Coverage (%): Tree area coverage prediction followed the same general patterns observed in the previous variables. In the original spatial domain, models based on raw Sentinel-2 bands achieved MAE values ranging from approximately 0.678 to 0.768 and RMSE values between 0.891 and 1.071. The Spectral Kennaugh-like models exhibited slightly improved performance, reaching MAE values between 0.612 and 0.662 and RMSE values from 0.801 to 0.866.

Under cross-validation, both approaches showed a noticeable performance degradation. MAE increased by around 1.9 to 2.0 percentage points for both raw and Kennaugh-based models, and RMSE shifted accordingly. Despite this general decline, the Spectral Kennaugh features provided slightly more robust results, with lower absolute errors and more stable variability across different spatial regions.

The Spectral Kennaugh representation appeared to support marginally better generalization behaviour for predicting tree area coverage compared to the raw Sentinel-2 spectral bands.

Sum of Crown Volume (m^3): The sum of crown volume predictions demonstrated clear differences between the feature sets. In the original spatial domain, models using raw Sentinel-2 bands reached MAE values between approximately 27.3 and 29.3, while the Spectral Kennaugh-like features yielded slightly better errors, ranging from 27.3 to 28.7. Similarly, RMSE values were marginally lower for models using the Kennaugh-like features.

During spatial cross-validation, substantial performance drops were observed for both approaches, with MAE increases in the range of 56 to 57 units. Notably, models based on Spectral Kennaugh features exhibited slightly lower degradation in RMSE compared to

those based on raw Sentinel-2 data, suggesting a small advantage in generalizing crown volume estimation.

However, the magnitude of the overall performance shift indicated that predicting crown volume remains a highly spatially sensitive task, independent of the specific input feature set.

Mean Tree Height and Mean Crown Base Height (m): The prediction of mean tree height and mean crown base height showed very close performance between the two feature sets in the original spatial domain. For mean tree height, models based on raw Sentinel-2 bands achieved MAEs of around 0.65–0.80 meters, while models trained on Spectral Kennaugh-like elements achieved slightly lower MAEs, between 0.63–0.71 meters. This pattern also extended to RMSE values, confirming comparable accuracy levels.

Cross-validation, however, revealed considerable spatial transfer degradation across both feature sets. The increase in MAE and RMSE was somewhat lower for models using Spectral Kennaugh-like inputs, implying a marginally better generalization in unseen regions.

Similarly, mean crown base height estimation initially performed comparably across both feature types, with slightly better MAEs observed for Spectral Kennaugh-like features (around 0.73–0.81 meters) compared to raw Sentinel-2 bands (around 0.72–0.88 meters). During spatial transfer, both feature types experienced large increases in error, although Spectral Kennaugh-like features again showed a slightly reduced performance drop in terms of RMSE and standard deviation.

Overall, both sets achieved similar accuracies, with slight robustness advantages for the Kennaugh-based representations under spatial transfer conditions.

Model Performance and Preprocessing Effects

Preprocessing Trends: Analysis of preprocessing settings revealed consistent trends across both Sentinel-2 raw bands and Spectral Kennaugh-like elements. Models applying aggressive preprocessing, particularly stricter masking thresholds combined with Z-normalization, consistently outperformed their less aggressively preprocessed counterparts. The positive impact was particularly notable for RF models under both feature sets, reducing overfitting and leading to better spatial transferability.

Interestingly, the application of a minimal masking threshold (`Mask > 0`) without Z-normalization led to substantial error increases during cross-validation, especially for linear models and SVRs. In contrast, applying `Mask > 1` and enforcing standardization (`Z=1`) systematically improved both training and transfer performance.

Although these preprocessing effects were visible across both input types, they appeared slightly more pronounced for Spectral Kennaugh-like elements, where normalization consistently enhanced spatial robustness and reduced variability among cross-validation results.

Model Diversity and Robustness: Across both Sentinel-2 raw bands and Spectral Kennaugh-like elements, RF regressors demonstrated the highest robustness and consistency. They dominated the top-performing models in terms of lowest MAE and RMSE, and exhibited the most stable results across cross-validation regions. Especially under aggressive preprocessing, RF models achieved low standard deviations, indicating low variance in predictions.

SVR showed competitive performance under specific settings but were generally more sensitive to preprocessing and cross-validation shifts. Linear regression models (and the 1D-CNN) performed considerably worse across nearly all variables, often serving as the worst-case (or generally not competitive) baseline.

Comparing both input types, Spectral Kennaugh-like elements led to slightly improved overall robustness, particularly for tree structural attributes such as mean tree height and crown base height. The distribution of model performance (top, median, and worst models) was narrower with Kennaugh-like elements, suggesting a reduced dependency on specific hyperparameter choices compared to using raw bands.

Spatial Generalization and Transfer Behaviour

Cross-Validation Results: Spatial transfer experiments revealed substantial performance declines across all models and variables, consistent with earlier observations. However, the extent of degradation differed depending on the type of input features.

Models trained on Spectral Kennaugh-like elements generally exhibited slightly lower increases in MAE and RMSE compared to models trained on raw Sentinel-2 bands. For example, in the prediction of sum of crown area for deciduous trees, Kennaugh-based

models displayed a performance shift of approximately 12–13 units in MAE, compared to shifts of 12–40 units for raw bands depending on preprocessing configurations.

Similarly, tree count and tree area coverage predictions demonstrated relatively less degradation with Kennaugh features. Nevertheless, for volumetric attributes such as sum crown volume, both input types suffered considerable drops in performance, exceeding 50 units of MAE in most configurations.

Some variables, like mean tree height and mean crown base height, maintained comparable levels of transferability between raw bands and Kennaugh-like elements, indicating that both setups faced similar challenges when generalizing fine-grained structural metrics across spatial domains.

Summary of Transferability: Overall, Spectral Kennaugh-like elements provided marginally improved transferability across different spatial domains for most vegetation variables. The models based on these features exhibited more stable generalization patterns, especially for area-related metrics such as tree area coverage and crown areas. However, for volume estimations and finer vertical structural attributes, spatial transferability challenges persisted, irrespective of the input feature type.

6.3.4 Discussion

This section synthesizes the findings from the comparative evaluation of raw Sentinel-2 reflectance bands versus their corresponding Kennaugh-like spectral transformations. Building upon the unified experimental framework, the results offer insights into model robustness, variable-specific challenges, and spatial transferability. Emphasis is placed on how different data representations interact with model architectures and preprocessing pipelines, and how these interactions influence predictive performance both within and beyond the original AOI.

Model Performance and Generalization Behaviour

Model Performance: Across both Sentinel-2 raw data and the spectral Kennaugh-like elements, RF regressors consistently demonstrated strong and stable predictive performance. In nearly all evaluated forest variables, RF models with no depth restriction (`max_depth=None`) and a logarithmic feature sampling strategy (`max_features=log2`)

outperformed alternative regressors such as SVR, the 1D-CNN and Linear Regression. The robustness of RF models was especially evident under aggressive masking and minimal z-normalization, where their ensemble-based structure allowed effective learning from moderately noisy or non-homogeneous feature spaces.

The repeated emergence of RF among the top-performing setups highlights their capacity to model non-linear relationships and variable interactions that characterize forest structure. The consistency across both raw spectral inputs and transformed Kennaugh-like features further indicates that RF are resilient to different data representations, making them a reliable backbone model when transferring between raw and feature-engineered input domains.

Effect of Preprocessing on Model Behaviour: The role of preprocessing steps, such as masking low-quality pixels and applying z-score normalization, proved critical in shaping model outcomes. In both the raw and Kennaugh-like datasets, aggressive masking (`Mask>1` or `Mask>0`) in combination with per-variable z-normalization (`Z=1`) systematically improved performance metrics. These setups consistently achieved the lowest MAE, RMSE, and MAD values in the original spatial domain.

Notably, the effect of preprocessing was slightly more pronounced in the Kennaugh-like feature set, suggesting that the spectral decomposition benefits more from clean, high-quality inputs. In contrast, SVR models exhibited comparatively unstable behaviour across different preprocessing configurations, often deteriorating sharply when masking or normalization strategies deviated from optimal settings. These trends underline the importance of rigorous preprocessing pipelines, particularly when employing advanced feature engineering steps such as hypercomplex projections.

Variable-Specific Trends and Model Suitability: Model performance exhibited variable-specific patterns that were largely consistent across the two input domains. Structural metrics such as tree area coverage (%) and count-based variables (number of deciduous and coniferous trees) were predicted with relatively low absolute errors, while volumetric quantities, especially crown volume (m^3) and sum of crown area (m^2), posed greater challenges.

In both the raw and Kennaugh-like datasets, RF consistently excelled in predicting tree counts and mean tree heights, indicating that ensemble methods effectively capture the spectral cues corresponding to canopy density and height proxies. Conversely, the estimation of crown volume, a more complex and compounding variable, exhibited larger errors and greater sensitivity to cross-validation shifts. This suggests that volumetric

forest attributes may require additional input modalities (e.g., active sensing data) or more sophisticated modelling techniques to achieve robust predictions.

Interestingly, the Kennaugh-like features slightly improved the prediction of structural variables compared to raw spectral bands alone. This hints at a potential benefit of decomposing brightness and spectral contrast, particularly for height and density estimations. However, the gains were moderate, underscoring that while feature engineering offers improvements, the underlying signal-to-noise ratio and spatial resolution remain critical limiting factors.

Spectral Decomposition Enhances Structural Interpretability: The slight edge observed for Kennaugh-like features, especially in predicting structural variables like mean height and tree count, points toward the interpretability benefits of spectral decomposition. By separating brightness and contrast components, the transformation reduces multicollinearity and highlights physiologically meaningful signal dimensions. This can aid both model interpretability and data harmonization, particularly when used across heterogeneous landscapes.

Limitations of CNN for Low-Context, Tabular Spectral Data: The 1D-CNN underperformed consistently across both feature domains, reinforcing a key architectural limitation: deep sequential models like CNN may not be well-suited to low-dimensional, per-pixel spectral vectors with no inherent spatial or temporal order. Unlike in image or time-series applications, the Sentinel-2 feature vectors lack structured continuity, making the inductive bias of CNN suboptimal. A mismatch between data structure and model architecture identified also in the preceding Sections.

Spatial Generalization and Transfer Modelling

Performance in Cross-AOI Domains: Transfer modelling results, evaluated by applying models trained on the Steigerwald (AOI 1) data to independent target AOIs, revealed a clear pattern of performance degradation across all configurations. Regardless of the input domain, raw Sentinel-2 bands or spectral Kennaugh-like elements, the cross-AOI MAE, RMSE, and MAD values were consistently higher than those observed in original-domain validation. Nevertheless, RF regressors maintained relatively stable performance margins compared to more sensitive models like SVR.

Importantly, models trained on the Kennaugh-like features demonstrated marginally better resilience to domain shift. This was particularly evident for tree count and mean

height predictions, where errors under cross-AOI application increased by a smaller factor compared to models trained on raw spectral bands. This observation suggests that the spectral decomposition into brightness and chromatic components may confer slight advantages in capturing transferable spectral patterns linked to vegetation structure.

Model Setup and Stability: Model setups that had already favoured aggressive masking and z-normalization within the original domain were also more stable under transfer conditions. Particularly, configurations using `max_depth=None` and `max_features=log2` in RF consistently ranked among the top-performing setups post-transfer. This indicates that limiting overfitting to local spectral peculiarities during training aids in generalization to novel landscapes.

In contrast, SVR models and linear baselines exhibited significant instability during domain transfer, with errors often doubling compared to within-AOI evaluations. These findings highlight the advantage of ensemble-based methods and suggest that feature randomness and deep, unrestricted trees help models better generalize beyond the training distribution.

Operational Implications and Transferability: From an operational standpoint, the findings emphasize the trade-off between model performance and spatial transferability. While high accuracies can be achieved when models are applied within their domain of training, transferring to different forest contexts without additional adaptation introduces notable uncertainty. Nevertheless, the relatively modest increase in error for key structural variables, particularly when using Kennaugh-like inputs and robust RF setups, suggests that Sentinel-2-based regression models can serve as a reasonable baseline for regional forest monitoring tasks.

The observed trends further indicate that the feature engineering strategy, while beneficial, does not fully bridge the gap introduced by spatial heterogeneity in spectral responses. Therefore, applications requiring high-precision forest variable mapping across diverse landscapes may benefit from domain adaptation techniques, model retraining, or the integration of complementary data sources (e.g., SAR or LiDAR) to achieve stable and operational transferability.

6.4 Polarimetric Kennaugh Elements from Sentinel-1 Data

This section initiates another model benchmarking framework outlined in Section 6, applying it for the first time to the radar-only baseline scenario. Specifically, the experiment focuses on Sentinel-1 C-band SAR data, processed into polarimetric Kennaugh elements [288] (see Section 2.2, and evaluates their predictive capacity in modelling continuous forest structure variables provided by the Wald5Dplus dataset [148], as described in Section 1.2.2.

As a initiative experimental configuration, this setup serves two primary purposes: first, it establishes a reference baseline against which later fusion strategies can be evaluated, and second, it enables controlled investigation of model sensitivity under mono-temporal, single-modality conditions.

6.4.1 Materials

The specific acquisition date and preprocessing steps taken are detailed in Section 6.2, within the materials description of the preceding *Polarimetrically and Spectrally Fused Sentinel-1 and Sentinel-2 Data* setup.

6.4.2 Results

The evaluation focuses on eight forest-related target variables encompassing both structural and compositional characteristics at the stand level. These include metrics such as crown dimensions, mean tree height, and species-specific aggregations (e.g., deciduous cover or conifer counts). Derived from high-resolution airborne and terrestrial LiDAR, the variables were aggregated to a 10 m spatial resolution to align with the Sentinel-1 input data. As previously outlined in Table 1.3, these reference rasters serve as physiologically grounded benchmarks for assessing the predictive capacity of SAR-based modelling.

Model Evaluation per Variable

Table references correspond to results from the original spatial domain, in Tables A.72 to A.79, while corresponding performance shifts are detailed in Tables A.80 to A.87.

Across the evaluated forest structural variables, RF regressors consistently achieved the strongest predictive performance in most configurations. The following summarizes the best, median, and worst model performances in the original spatial domain for each target variable:

Sum of Crown Area of Deciduous Trees (m^2): The best RF models, using Mask > 1, Z-score filtering at $Z = 1$, and aggressive outlier filtering, achieved a minimum MAE of $3.889\,m^2$ and RMSE of $4.857\,m^2$. The median model (RF, Mask > 1, no Z-score filtering, no aggressive filtering) exhibited higher errors (MAE: $12.074\,m^2$). The worst performance was recorded by a Linear Regression model with an MAE of $20.278\,m^2$, highlighting its limited capacity to model complex patterns.

Sum of Crown Area of Coniferous Trees (m^2): RF models again dominated performance, achieving a minimum MAE of $3.386\,m^2$ under aggressive preprocessing settings. Relaxed filtering configurations caused a substantial rise in errors (median MAE: $9.627\,m^2$), while Linear Regression produced notably poorer results (MAE: $15.699\,m^2$).

Count of Deciduous Trees: Tree count predictions showed exceptionally low errors. Best-performing RF attained MAEs of around 0.168 trees. The gap between RF and linear baselines widened dramatically for this discrete variable, where Linear Regression models yielded errors exceeding 0.9.

Count of Coniferous Trees: Consistent patterns were observed, with RF models achieving top MAE scores of 0.098 and RMSE of 0.128. Linear Regression's instability again manifested through poor metrics (MAE: 0.488).

Tree Area Coverage (%): For fractional tree cover, RF with Mask > 1, $Z = 1$, and aggressive filtering provided the best results (MAE: 0.463%). SVR emerged as a viable competitor in certain configurations but was less consistent across validation stages.

Sum of Crown Volume (m^3): Volume metrics presented greater challenges. RF delivered competitive results (MAE: $23.014\,m^3$), yet errors increased substantially under relaxed preprocessing. This sensitivity underscores the complexity of modelling volumetric attributes using SAR data.

Mean Tree Height and Mean Crown Base Height (m): Height-related variables were relatively robust. The best RF models achieved MAEs of $0.610\,m$ (tree height) and $0.711\,m$ (crown base height), confirming that tree height is a more spatially stable and predictable trait.

Model Performance and Preprocessing Effects

Preprocessing Trends: Optimal model performance systematically emerged when applying moderate masking (Mask > 1), mild Z-score filtering ($Z = 1$), and aggressive outlier filtering. These configurations suppressed extreme noise while retaining sufficient ecological variability, especially critical for volume- and crown-based metrics. Configurations lacking Z-score filtering or employing relaxed masking thresholds (e.g., Mask > 0) degraded model performance, even for robust models like RF. Linear Regression models showed significant vulnerability to preprocessing choices, performing poorly across all configurations.

Model Diversity and Robustness: RF consistently outperformed other models across both original and validation domains. While SVR models occasionally approached RF performance on simple variables, their instability across spatial domains limited their operational value. Linear models were generally unable to capture the complex, nonlinear relationships inherent in the data.

To complement the narrative performance overview, Table 6.2 provides a consolidated summary of the best-performing model configurations for each forest structure attribute, detailing the corresponding model type, preprocessing parameters, and associated error metrics.

Table 6.2.: Summary of best-performing model configurations across all forest structural attributes using Sentinel-1 Kennaugh features. For each variable, the model yielding the lowest validation error (based on MAE) is reported, along with its specific preprocessing parameters

Target Variable	Experiment	Model Parameters	MAE
Sum crown area of deciduous trees (m^2)	Mask > 1, Z = 1, Aggressive = True	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	3.889
Sum crown area of coniferous trees (m^2)	Mask > 1, Z = 1, Aggressive = True	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	3.386
Count of deciduous trees	Mask > 1, Z = 1, Aggressive = True	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.168
Count of coniferous trees	Mask > 0, Z = 1, Aggressive = True	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.098
Tree area coverage (%)	Mask > 1, Z = 1, Aggressive = True	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.463
Sum crown volume (m^3)	Mask > 1, Z = 1, Aggressive = True	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	23.014
Mean tree height (m)	Mask > 1, Z = 1, Aggressive = True	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.610
Mean crown base height (m)	Mask > 1, Z = 1, Aggressive = True	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.711

Spatial Generalization and Transfer Behaviour

Cross-Validation Results: Spatial transfer testing revealed systematic performance degradation across all models and variables, with the extent of decline varying:

- **Sum Crown Area (Deciduous and Coniferous):** RF models showed notable error increases under transfer, with ΔMAEs around 13.5 and 8.7 respectively for best

configurations. Poorly tuned models, including certain SVRs, exhibited extreme transfer errors exceeding 30.

- **Tree Count Variables:** Tree counts generalized more reliably, with best RF exhibiting ΔMAEs around 0.57 for deciduous and 1.24 for coniferous trees.
- **Tree Area Coverage (%):** Transfer degradation was moderate (ΔMAE around 2.2) for RF. SVRs demonstrated better stability in isolated cases but with higher absolute errors.
- **Sum Crown Volume (m^3):** Volume metrics proved difficult to generalize, with ΔMAEs approaching 64 in some cases, highlighting the complexity of modelling 3D attributes under SAR data.
- **Mean Tree Height and Crown Base Height (m):** RF models maintained relatively stable performance, with ΔMAEs between 1.2 and 1.4, affirming the robustness of height metrics.

The performance deltas of optimal model setups relative to baseline configurations for each forest structural attribute are summarized in Table 6.3. This comparison highlights the impact of specific preprocessing strategies and model choices on predictive accuracy across variables.

Table 6.3.: Summary of best-performing model configurations across all forest structural attributes using Sentinel-1 Kennaugh features in the cross-validation scenario (Delta Metrics). Reported error values reflect improvements relative to the respective baseline setup for that attribute.

Target Variable	Experiment	Model Parameters	ΔMAE
Sum crown area of deciduous trees (m^2)	Mask > 1, Z = 1, Aggressive = True	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	13.59
Sum crown area of coniferous trees (m^2)	Mask > 1, Z = 1, Aggressive = True	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	8.67
Count of deciduous trees	Mask > 1, Z = 1, Aggressive = True	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.58
Count of coniferous trees	Mask > 0, Z = 1, Aggressive = True	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.24
Tree area coverage (%)	Mask > 1, Z = 1, Aggressive = True	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	2.21
Sum crown volume (m^3)	Mask > 1, Z = 1, Aggressive = True	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	64.30
Mean tree height (m)	Mask > 1, Z = 1, Aggressive = True	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.38
Mean crown base height (m)	Mask > 1, Z = 1, Aggressive = True	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.18

In summary, RF models demonstrated the best combination of in-domain accuracy and transferability across variables. Height- and count-based attributes transferred more reliably than volume- or crown area-based metrics. Preprocessing choices had a profound impact on both in-domain performance and cross-area generalization success.

6.4.3 Discussion

Before concluding this section, a brief discussion of the observed modelling behaviours, preprocessing impacts, and transferability patterns is warranted. The following subsections reflect on performance trends across variables and configurations, offering insights into the generalization capabilities and operational implications of using Sentinel-1-based polarimetric features.

Model Performance and Generalization Behaviour

Model Performance: Across all evaluated forest structural variables, RF regressors consistently delivered the strongest predictive performance when using Sentinel-1-derived polarimetric Kennaugh elements as input features. In intra-AOI evaluation, RF exhibited the lowest MAE and RMSE across nearly all target variables. Their ability to model non-linear relationships and inherent robustness to noisy or incomplete input features is particularly advantageous when working with SAR-based predictors, which are more abstract compared to multispectral variables.

Effect of Preprocessing on Model Behaviour: Preprocessing configurations had a decisive impact on model accuracy. Best-performing models predominantly employed moderate masking (Mask > 1), light Z-score filtering (Z $= 1$), and aggressive outlier suppression. This finding suggests that selective data cleaning enhances the predictive signal extracted from Kennaugh elements without excessively discarding ecologically informative variance. Configurations omitting masking or using no outlier trimming yielded considerably poorer model fits, particularly for volume- and crown-related metrics.

Variable-Specific Trends and Model Suitability: Prediction quality varied systematically across forest attributes. Sum crown volume and crown area metrics exhibited higher absolute errors and suffered the greatest degradation during spatial transfer. These structural variables are inherently more sensitive to SAR's indirect measurements of canopy density and geometry. In contrast, variables such as mean tree height, crown base height, and tree counts generalized more effectively, maintaining relatively lower cross-domain error increases. This indicates that vertical structural features and discrete object counts are more reliably captured by Sentinel-1 polarimetric signatures.

SAR Feature Interpretability: Although SAR-based Kennaugh elements may appear abstract compared to spectral features, they encode physically grounded descriptors of

backscatter behaviour. The strong predictive performance achieved using these inputs alone highlights their informational richness and validates their integration into EO–ML workflows as more than just auxiliary features.

Importance of Label Design: The differential model performance across target variables also underscores the value of carefully engineered reference labels. The ability of SAR features to predict certain attributes, such as tree height or crown base height, more effectively suggests a high degree of temporal and structural alignment between EO inputs and the Wald5Dplus labels. This reinforces the critical role of well-calibrated, temporally synchronized reference data in achieving robust model outcomes.

CNN Limitations on Tabular SAR Features: While CNN are powerful in spatial or temporal sequence learning, their application to flattened, tabular SAR-derived features presents inherent limitations. The Kennaugh features used here lack spatial adjacency or sequential structure, reducing the effectiveness of convolutional filters designed to exploit such patterns. The comparatively lower performance of 1D-CNN in this study suggests an architectural mismatch, highlighting the importance of aligning model design with the underlying data modality.

Spatial Generalization and Transfer Modelling

Performance in Cross-AOI Domains: While RF models demonstrated strong in-domain predictive power, transfer to geographically distinct validation regions resulted in expected performance degradation across all variables. The extent of this decline varied by target: tree counts and height variables retained moderate accuracy, whereas crown area and volume metrics displayed larger shifts. Nonetheless, top-performing RF configurations maintained reasonable predictive fidelity even under domain shift, reinforcing their operational viability for regional forest monitoring. Interestingly, models trained with moderate preprocessing (Mask > 1, Z $= 1$) not only achieved the best in-sample results but also exhibited greater robustness in cross-AOI transfer compared to models with relaxed or overly aggressive preprocessing.

Model Setup and Stability: Across the diverse experimental grid, homogeneous RF configurations outperformed heterogeneous alternatives such as SVR or Linear Regression, as well as the 1D-CNN. Although individual SVR setups achieved localized success in specific configurations (e.g., tree area coverage), their poor stability during transfer underscored the strategic advantage of RF-based ensembles for operational deployment.

Simplicity, robustness, and minimal hyperparameter sensitivity remained hallmarks of RF performance across domains.

Operational Implications and Transferability: The strong overall performance of RF models trained on pure SAR Kennaugh elements highlights the potential of polarimetric transformations for stand-alone forest monitoring solutions, particularly in areas or time periods where multispectral data may be unavailable or contaminated (e.g., cloud cover). However, spatial transfer remains a core challenge, especially for volume-dominated metrics. These results suggest that while SAR alone enables powerful predictive capabilities, hybrid SAR-optical fusion may still be necessary to maximize spatial generalization for more complex structural attributes.

Future extensions could include integration of additional polarimetric descriptors or object-based aggregation approaches to further enhance the capture of 3D canopy structures. Furthermore, embedding strict cross-AOI validation setups, as employed here, should become standard practice when assessing the deployment readiness of SAR-based forest attribute models. Taken together, these findings confirm that RF trained on polarimetric Kennaugh elements provide a scalable and interpretable pathway for EO-based forest monitoring, with clear trade-offs between robustness, transferability, and complexity depending on the specific forest attributes modelled.

6.5 Polarimetric Kennaugh Elements from TerraSAR-X and ALOS-2 Data

Building on the insights gained from the analysis of technically partially polarimetric Kennaugh elements derived from Sentinel-1 data, the following chapter expands the investigation to truly polarimetric sources, TerraSAR-X and ALOS-2, which provide full quad-polarimetric acquisitions. This enables a more comprehensive exploration of polarimetric information content and its impact on forest structure modelling. These datasets are evaluated in combination with the continuous Wald5Dplus label dataset, described in Section 1.2.2. Only key findings and insights from this configuration are included in this thesis, with detailed results provided in the associated publication [278].

6.5.1 Materials

In this experiment, five different fusion scenarios were explored to optimize the estimation of forest parameters using dual-frequency polarimetric SAR data. These scenarios include the mono-frequency L-band and X-band analysis, a simple layer stack of Kennaugh elements from both ALOS-2 (L-band) and TSX (X-band), and more complex additive and multiplicative fusion approaches. Each scenario leverages the unique properties of SAR wavelengths to enhance the understanding of forest structure, ranging from canopy cover to tree height and biomass estimation.

ALOS-2 data: The L-band data was acquired by ALOS-2 in May 2017 and retrieved via personal investigator no. ER3A2N089 from JAXA. The polarimetric single-look complex data were pre-processed by the Multi-SAR processor at DLR [38]. After Kennaugh decomposition and uniform multi-looking using eleven spatial looks, the layers are geocoded, gamma corrected, and further adaptively multi-looked by the multi-scale multi-looking approach. The processor outputs ten normalized Kennaugh elements in 16bit unsigned integer scaling on a 10 m by 10 m raster.

TerraSAR-X data: The X-band data of TSX data was already acquired in May 2010 during the dual-receive antenna campaign and retrieved via proposal no. MTH3885 from DLR. The preprocessing is equal to the one of the L-band data.

Labels: The reference labels are based on high-resolution airborne laser scanning and multispectral imagery collected over four representative transects in the Bavarian Forest National Park. The data, generated using a patented normalized cut segmentation algorithm (Treefinder), provided over half a million individual tree crowns with attributes such as height, crown area, and base height. While related to the broader Wald5Dplus project, these transects represent a focused subset, specifically rasterized to 10 m resolution to align with SAR inputs and enable detailed pixel-wise forest structure modelling [278], see Transects 1–4 (Reference Plots) in Table 1.1.

6.5.2 Methods

To systematically evaluate the added value of full quad-polarimetric SAR data for forest structure modelling, a series of experimental scenarios were constructed using ALOS-2 (L-band) and TSX (X-band) datasets. These configurations span both mono-frequency setups and dual-frequency fusion strategies, as summarized in Figure 6.21.

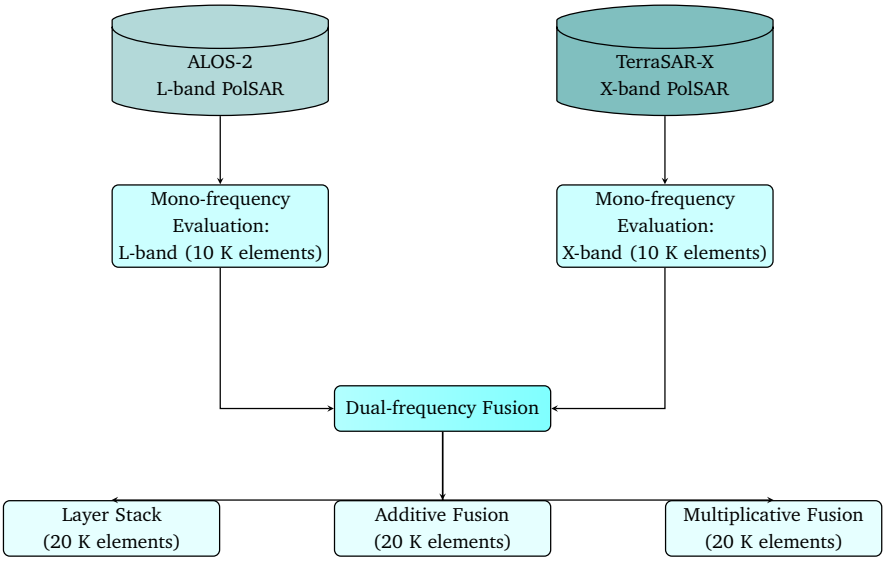


Figure 6.21.: Overview of evaluated scenarios based on ALOS-2 L-band and TSX X-band PolSAR data, including mono-frequency evaluations and dual-frequency fusion strategies.

In order to assess the suitability of different fusion approaches, five test scenarios are prepared, see Table 6.4. The first two scenarios evaluate the mono-frequency PolSAR data of ALOS-2 and TSX. The simplest image fusion method is the joint interpretation as layer stack consisting of the ten Kennaugh elements from ALOS-2 and the ten Kennaugh elements from TSX. Thanks to the similarity of the Kennaugh elements to hypercomplex bases, they can be combined as via sums and differences in the same way [289]. The backscattering strength serves a weight in this type of image fusion, i.e. the polarimetry of the stronger backscattering dominates. Normalized Kennaugh elements directly correspond to normalized hyper-complex bases and therewith, enable the image fusion in a relative manner [360]. This technique levels out differences in the backscattering strength by separating polarimetry from intensity and thus, coincides with the SARsharpening methodology [285] and relative change detection techniques [287].

Table 6.4.: Evaluated Scenarios for Forest Parameter Estimation

No.	Description	No. of K Elements
1	L-band (mono-frequency)	10
2	X-band (mono-frequency)	10
3	Layer stack (dual-frequency)	20
4	Additive fusion (dual-frequency)	20
5	Multiplicative fusion (dual-frequency)	20

RF regression was selected for its robustness and high accuracy in handling complex and multi-dimensional data, which is characteristic of PolSAR datasets. The RF model's ensemble learning approach, involving the generation of multiple decision trees from bootstrap samples, makes it particularly well-suited for this study's objectives. It allows for both intra-AOI predictions, using data from within the same area, and cross-regional transfer models, applying data from one area to predict outcomes in another.

The algorithm is applied in three distinct ways: RF models trained on data from a specific AOI are utilized for Intra-AOI prediction. This enables target variable prediction within the same AOI, leveraging the algorithm's predictive power for localized insights. Integrated Models for Individual AOIs are developed by aggregating data from multiple AOIs, creating a comprehensive model for predicting target variables within each AOI separately. This approach enhances individual predictions by utilizing collective information from various AOIs. Cross-Regional Transfer Models involve training models with data from one AOI to predict target variables in another AOI, employing a transfer learning approach that integrates knowledge acquired from one AOI to improve predictions in a different AOI. This strategy enhances the overall generalization capability of the models. Z-score trimming is applied to address outliers, involving the calculation of Z-scores for each input variable and applying a threshold (e.g., three standard deviations) to identify and remove outliers from the dataset. This step enhances the robustness and reliability of the predictions.

6.5.3 Results

Scenario 1 – L-band (mono): In Scenario 1, utilizing L-band mono-frequency data, the models consistently deliver predictions for a range of forest-related target variables. Consistently, across all metrics, the Intra-AOI model, is found to outperform the other models, whereas the lowest MAE, MAD and STD values are consistently achieved by the Intra-AOI model. For instance, in the prediction of "Proportion of forested area," a remarkably low MAE of 3.95 is observed. The combination of multiple models, does not yield an improvement in terms of precision. Notably, the Transfer model is well performing for “Number of dead trees”, but also “Mean crown volume”.

Scenario 2 – X-band (mono): In Scenario 2, employing X-band mono-frequency data across all target variables, the Intra-AOI model consistently demonstrates superior performance compared to the other models in Scenario 2. In particular, key variables such

as "Summed-up crown area of deciduous trees," "Summed-up crown area of coniferous trees," "Summed-up crown area of dead trees," "Number of deciduous trees," "Number of coniferous trees," and "Number of dead trees" yields moderate predictions accuracy results.

Scenario 3 – Layer stack (dual): Scenario 3 employed a layer stack of both X-band and L band radar data for the regression analysis of various target variables. The Intra-AOI model delivers relatively consistent and competitive results across all target variables, emphasizing the scenarios predictive accuracy. This scenario delivers especially consistent and precise MAE values, "Proportion of forested area", "Mean crown volume", "Mean tree height" and "Mean crown base height". The Unified model displays performance metrics on a par with the Intra-AOI model. While scenario 3 generally fares well in all target variables, it exhibits some significant outliers, when applying the transfer model, warranting further investigation.

Scenario 4 – Additive fusion (dual): In Scenario 4, leveraging the additively fused dataset, the Intra-AOI model performs best. Detail-oriented variables such as "Summed-up crown area of deciduous trees", "Summed-up crown area of coniferous trees", "Summed up crown area of dead trees," "Number of deciduous trees", "Number of coniferous trees" (Figure 1) and "Number of dead trees" yield significantly lower precision than, "Proportion of forested area", "Mean crown volume", "Mean tree height" across all tested models. In terms of transferability, the 4th scenario demonstrates however a robust performance. It steadfastly exhibits the lowest MAE, MAD and STD values across most target variables, but particularly for "Summed-up crown area of deciduous trees", "Number of deciduous/coniferous trees".

Scenario 5 – Multiplicative fusion (dual): The multiplicative fusion of both X- and L-band data exhibits relatively stable and consistent performance across the range of target variables. This scenario does not produce many outliers, especially across variables concerning Crown area and the actual count of tree, suggesting that the models are robust and provide dependable predictions. The Intra-AOI model, maintains a steady level of performance. Similarly, the unified model, demonstrates stable predictions across the target variables, however, exhibits no improvement compared to the Intra-AOI model. The Transfer model performs adequately in terms of knowledge transfer across different AOIs. It consistently exhibits low MAE values, but it may not consistently outperform the other scenarios in this specific context.

6.5.4 Discussion

The evaluation of five scenarios for predicting forest-related target variables across four transects has provided valuable insights into their performance to be discussed.

Scenario 1 – L-band (mono): In Scenario 1, utilizing L-band mono-frequency data, the Intra.AOI model consistently outperforms the other models. These results emphasize the significance of utilizing L-band radar data, particularly for target variables like "Proportion of forested area" and "Mean crown volume," where the longer wavelength of L-band data contributes to improved predictive performance. Overall, it exhibits relatively robust precision metric values within the expected range for most variables; it did not consistently outperform other scenarios. This scenario might be suitable for situations where L-band data is readily available and other bands are not.

Scenario 2 – X-band (mono): In Scenario 2, employing X-band mono-frequency data, the Intra-AOI model consistently demonstrates decent performance compared to the other models in that scenario. Due to the utilization of X-band mono-frequency data, with enhanced resolution and wavelength characteristics (compared to L-band data), it may be beneficial in predicting specific fine-scale ecological target variables. It proves to be more effective in estimating “Summed-up crown area of deciduous trees, coniferous and dead trees”, as well as the “Number of deciduous, coniferous and dead trees”, when compared to the L-band data. However, it does not exhibit any remarkable advantages over other scenarios. In cases where X-band data is more accessible, this scenario can be considered as these findings underscore the potential advantages of utilizing X-band radar data for ecological assessments, particularly for the mentioned target variables.

Scenario 3 – Layer stack (dual): Scenario 3 employs a layer stack of both X-band and L-band radar data. As demonstrated in Table 3, Scenario 3 may be very well suited for the prediction of our forest-related parameters. It consistently delivers competitive results across all target variables, especially in the Intra-AOI model, as well as the Unified model, with significantly low and robust MAE, MAD and STD values, very well within the true target value ranges. Notably, it is susceptible to outliers. The Transfer model, shows mixed results performing well in some target variables but struggling in others. This indicates that its performance is not consistently superior, especially when transferring knowledge from one AOI to another.

Scenario 4 – Additive fusion (dual): In the cross-regional Transfer Models scenario, Scenario 4 emerged as a robust performer. It consistently delivers precise predictions,

which is demonstrated consistently for target the variables "Proportion of forested area", "Mean crown volume", "Mean tree height". These results underscore its unparalleled predictive prowess, particularly when transferring knowledge across diverse geographical regions. This indicates that the additive fusion of X and L band data is effective in capturing real-world conditions and characteristics, even in unfamiliar, i.e., untrained, environmental settings.

Scenario 5 – Multiplicative fusion (dual): Scenario 5 consistently exhibits constant predictions for a wide range of forest attributes in both Intra-AOI Prediction and Integrated Models scenarios, while also achieving notable advantages in the Transfer-Domain. Its performance is particularly noteworthy for variables related to crown area and tree counts. This suggests that multiplicative fusion of X and L-band data enhances the model's accuracy, making it well-suited for localized forest assessments.

Implications and Applications

The choice of scenario and the combination of PolSAR data significantly impact accurate forest parameter prediction. Scenario 3 offers valuable insights into specific forested areas, excelling in key variables like forested areas, crown volume, tree height, and crown base height. Despite its competitive performance, Scenario 3 may exhibit vulnerability to outliers. In contrast, Scenario 5 excels in localized predictions, proving effective detailed assessments within specific areas. Scenario 4 showcases expertise in transferring knowledge across regions, particularly excelling in estimating summed-up crown areas of deciduous trees and counts of deciduous and coniferous trees. These findings empower forest management professionals to tailor their approach based on the study area's specific characteristics, providing valuable tools for more effective and targeted forest management.

Limitations and Future Directions

The effectiveness of these scenarios may depend on the characteristics of the AOIs and datasets used, necessitating further validation in diverse contexts. One notable limitation in this study was the consistently lower accuracy of the transfer model across all scenarios. While this discrepancy may arise from variations in geographical and environmental conditions between training and validation areas, it presents an opportunity for improvement. To address this limitation and enhance the transfer model's performance, future

investigations could explore parameter optimization and alternative knowledge transfer methods. Incorporating additional reference data from a wider range of geographical locations may also help mitigate accuracy disparities and improve generalizability. Such efforts would deepen our understanding of scenario applicability. Additionally, our plausibility check by ground truth using on-site data is a crucial component of our validation strategy, ensuring the accuracy of our predictions. Results from these validations will be presented in future research, further bolstering confidence in our models' and individual scenarios practical utility.

Comparison Between Sentinel-1-derived Kennaugh Element-Based Modelling and TSX/ALOS-2 Scenarios

Following the evaluation of RF regression models trained on Sentinel-1-derived Kennaugh elements (Section 6.4.2), subsequently alternative modelling results based ond different SAR scenarios from TSX and ALOS-2 sensors are assessed. These complementary experiments involve five distinct configurations per variable, representing different acquisition and preprocessing conditions.

Comparison of In-Sample Accuracies: Across intra-AOI setups, both Sentinel-1 and TSX/ALOS models demonstrated competitive predictive accuracy. However, models based on Sentinel-1 Kennaugh elements generally achieved lower MAE and MAD for most structural forest variables. For instance, summed crown area and tree count predictions using Sentinel-1 inputs typically yielded MAE values 30–50% lower than their TSX/ALOS counterparts, indicating superior fitting capabilities under the same-domain conditions. This seemingly counter-intuitive result may be attributed to several factors. First, Sentinel-1 offers a substantially higher temporal revisit frequency (up to every 6 days), allowing selection of optimal acquisition dates aligned with peak vegetation conditions, something less feasible with TSX or ALOS due to limited temporal coverage. Inherently, while TSX and ALOS provide quad-polarimetric data, their acquisition footprints and scheduling constraints often lead to scene-specific limitations or reduced ecological representativeness. Lastly, C-band (Sentinel-1) may be better suited to capture canopy-level scattering in temperate forests compared to L-band (ALOS) or X-band (TSX), which either penetrate too deeply or reflect mostly fine surface structures, respectively.

Comparison of Transferability: Under spatial cross-validation (transfer to independent areas), Sentinel-1 Kennaugh models showed comparatively greater robustness. While all models exhibited increased errors upon transfer, the degradation in MAE and MAD

was notably more pronounced for TSX/ALOS-based models, especially for attributes related to crown area, dead trees, and forest coverage proportions. This suggests that moderate-resolution Kennaugh features, capturing generalized scattering physics, better support cross-regional generalization compared to very high-resolution, site-specific SAR textures.

Variable-Specific Patterns: The advantage of Sentinel-1 Kennaugh Element based modelling was particularly evident for variables sensitive to dielectric or structural heterogeneity, such as tree height, crown base height, and summed crown volume. In contrast, tree counts showed similar transfer behaviours across datasets, reflecting that fine-scale object enumeration remains challenging regardless of sensor resolution without specific object delineation strategies.

6.6 Conclusions

This chapter systematically evaluated how different EO modalities, feature representations, fusion strategies, and ML models interact to predict forest structural attributes using the continuous Wald5Dplus label dataset. Through a series of controlled experiments, from single-modality baselines to spectrally, polarimetrically, and temporally fused ensemble configurations, this study setup identified the combinations that best balance predictive accuracy and spatial generalization.

The findings not only establish benchmark performances across multiple dimensions of model and data complexity but also yield practical insights for operational ecological monitoring. In the following, the primary research questions posed at the outset are revisited and summarize how the analyses addressed each of them.

6.6.1 Lessons Learned

This section synthesizes methodological, empirical, and operational insights derived from benchmarking remote sensing modalities, fusion strategies, and model architectures using the continuous Wald5Dplus forest structure dataset. The central question guiding this chapter was:
Which EO–model configurations yield the most accurate predictions of continuous forest

attributes? To address this, a modular framework incrementally explored model architectures and EO configurations, from single-sensor setups to fully fused temporal-spatial representations. The Wald5Dplus benchmark configuration—featuring spectral, polarimetric, and temporal fusion of Sentinel-1 and -2 via hypercomplex bases—served as a performance reference, with all other tiers interpreted as simplified variants or baselines.

Sentinel-1 Polarimetric Kennaugh Elements and RF Modelling

- **RF consistently performed best across structure variables.** RF delivered the lowest MAE/MAD in both intra- and cross-AOI settings. Their resilience to domain shift reinforced their suitability for SAR-based forest structure modelling.
- **Preprocessing has major effects on stability.** Optimal setups combined `Mask > 1`, Z-score filtering at $Z = 1$–3, and in some cases aggressive outlier removal. Relaxed settings increased overfitting, particularly in crown/volume metrics.
- **Vertical metrics generalize better than volumetric ones.** Mean tree height and crown base height retained relatively stable transfer performance, while crown area and volume suffered greater degradation.
- **RF outperformed hybrid or deep learning models.** CNN and SVRs showed inconsistent behaviour under domain shift, while homogeneous RF ensembles proved robust, interpretable, and accurate.
- **Cross-validation alone is insufficient.** Many strong in-domain results failed to generalize, demonstrating the importance of explicit domain transfer evaluations in EO model benchmarking.

Comparison: Sentinel-1 and TerraSAR-X and ALOS-2 Polarimetric Modelling

- **Sentinel-1 provided the most transferable performance.** Across variables, MAEs were 30–50% lower than those of commercial SAR systems.
- **TSX and ALOS-2 lacked robustness across AOIs.** Despite higher resolution, reduced temporal coverage and polarimetric consistency likely explain weaker generalization.

- **High-resolution sensors may suit niche tasks.** While not ideal for broad-scale modelling, TSX/ALOS may provide value in site-specific, high-resolution applications.

Raw Sentinel-2 Bands and Spectral Kennaugh-like Elements

- **Raw bands perform best in-domain.** Tree count, canopy coverage, and crown area were predicted most accurately under mono-AOI conditions.
- **Spectral Kennaugh-like elements improved generalization.** By normalizing across illumination and atmospheric noise, transfer resilience increased.
- **Each representation suits specific tasks.** Raw reflectances captured discrete crown-level variation, while spectral transformations enhanced spatial consistency.
- **Spectral decompositions reduce acquisition-induced variability.** This benefited models deployed in temporally or spatially variable domains.

Comparison: Sentinel-1 and Sentinel-2 Data

- **SAR features are superior for 3D structural traits.** Height and volume metrics were better captured by polarimetric Kennaugh elements.
- **Spectral bands dominated horizontal coverage tasks.** Tree count and canopy area models favoured Sentinel-2 inputs.
- **Transformations improved Sentinel-2, but not enough to replace SAR.** While spectral normalization helped, SAR remained crucial for structure prediction.
- **Modality integration is key.** Each input type captures complementary biophysical signals.

Polarimetrically and Spectrally Fused Sentinel-1 and Sentinel-2 Data

- **Fusion consistently enhanced predictive accuracy.** MAEs declined for all attributes, particularly tree height and count.
- **Biophysical synergy emerged.** Structure from SAR and foliar cues from optics yielded complementary features.
- **Preprocessing remains vital post-fusion.** Masking and filtering were still required to stabilize high-variance variables.
- **Transferability improved, but not equally.** Crown volume remained sensitive to spatial variability, even in fused models.
- **RF ensembles again proved most reliable.** Gains from fusion were best exploited by RF architectures.

Polarimetrically, Spectrally and Temporally Fused Sentinel-1 and Sentinel-2 Data & Ensemble Learning: Enhancing Generalization and Modularity

- **Tri-modal fusion achieved the highest performance.** Full spatio-temporal fusion using hypercomplex bases produced the most accurate models.
- **Stacked RF ensembles improved transfer performance.** Ensembles combining stratified base models via a meta-level RF outperformed all individual learners in cross-AOI settings, confirming the value of spatial tiling and model fusion.
- **Even sparse training footprints yielded robust generalization.** The full-AOI ensemble, trained on selected tiles (e.g., T10, SW_1), generalized to the 12x larger NP region with minimal performance loss, demonstrating extrapolative strength.
- **RF-only ensemble designs offered practical and technical benefits.** Homogeneous RF ensembles minimized complexity while delivering high performance, simplifying operational deployment and retraining.
- **Temporal ensemble stacking added resilience.** Multi-date inputs improved robustness to seasonal dynamics, acquisition artifacts, and reference label aging (e.g., 2016–2018 labels versus 2020–2021 predictions).

Operational and Methodological Takeaways

- **Modularity supports operational scaling.** The ensemble framework permits incremental inclusion of new base models and minimal meta-learner retraining, aligning with long-term monitoring needs and near-real-time workflows.
- **Label aging and misalignment were manageable.** Despite temporal offsets, predictions aligned closely with current forest states, especially for deciduous stands. This demonstrates ensemble tolerance to lagged reference inventories.
- **Further gains require enhanced stacking strategies.** Incorporating normalized base outputs, strict OOF training, and potentially hybrid meta-learners (e.g., Ridge or GBT) could address current limitations and refine model aggregation.
- **Pixel-level modelling introduces residual uncertainty.** Georegistration errors, label noise, and scale mismatches suggest that future models may benefit from object-based learning or coarser spatial aggregation.

Together, these lessons underscore the value of integrated EO fusion—across sensors, dates, and transformations—when paired with robust RF ensemble learning. This approach represents a scalable, interpretable, and high-fidelity strategy for operational forest monitoring.

6.6.2 Research Questions Revisited

The research questions formulated at the outset guided the evaluation of EO–model interactions for continuous forest structure prediction. Each is revisited here in light of the results obtained:

RQ1: *Which remote sensing modality, SAR (Sentinel-1), optical (Sentinel-2), or high-resolution SAR (TSX/ALOS), delivers the highest predictive accuracy for continuous forest structural variables in the Wald5Dplus dataset?*

Sentinel-1 SAR, specifically when represented via polarimetric Kennaugh elements, consistently delivered the highest spatial generalization and strong in-domain accuracy for structure-sensitive variables such as tree height and crown base height. While Sentinel-2 performed better for horizontal features like tree cover and counts in the training domain, it suffered more under spatial transfer. TSX and ALOS,

despite their higher spatial resolution, were less effective overall due to limited temporal coverage and reduced transferability.

RQ2: *How do polarimetric features derived from Sentinel-1 compare with raw spectral bands and transformed spectral features from Sentinel-2 in terms of predictive accuracy and spatial generalization?*

Polarimetric features from Sentinel-1 demonstrated superior robustness across spatial domains, particularly for volumetric and vertical attributes. In contrast, raw Sentinel-2 bands yielded sharper predictions in-domain but degraded more severely under spatial shift. Spectral Kennaugh-like transformations of Sentinel-2 data improved generalization but still did not match the transfer performance of SAR-based models.

RQ3: *How does the choice between raw Sentinel-2 spectral bands and Sentinel-2-derived spectral Kennaugh-like elements affect model accuracy and spatial robustness for different forest structural variables?*

Models trained on raw Sentinel-2 spectral bands consistently achieved higher predictive accuracy within the training domain, particularly for visually dominant forest attributes such as tree count, canopy cover, and crown area. These bands capture fine-grained spectral variation, offering high resolution for within-AOI modelling. However, their performance deteriorated more sharply under spatial transfer, likely due to sensitivity to regional variation in illumination, phenology, and atmospheric conditions.

In contrast, spectral Kennaugh-like elements, transformations that decompose reflectance into brightness and colorimetric components, provided more stable performance across AOIs. Their separation of spectral magnitude from chromatic properties introduced a form of built-in normalization, reducing the impact of scene-specific spectral variability. This robustness was especially evident for structural metrics such as standing volume and crown volume, where transformed features helped maintain predictive reliability in unseen regions.

Overall, raw bands maximize precision under controlled conditions, whereas spectral Kennaugh-like elements enhance transferability, highlighting a key trade-off between local sharpness and generalization in optical EO modelling.

RQ4: *Do spectral or polarimetric Kennaugh-like representations improve spatial transferability over raw features, and for which types of forest variables is this most pronounced?*

Yes, Kennaugh-like representations substantially improved spatial transferability. Polarimetric representations (from Sentinel-1) were especially beneficial for tree height, crown base height, and crown volume, while spectral transformations helped stabilize Sentinel-2-based predictions for variables like standing volume and basal area. These improvements were most pronounced in regions with ecological divergence from the training domain.

RQ5: *To what extent does fusing optical and SAR data improve the prediction of forest structure variables compared to using single modalities?*

Fusion consistently improved predictive performance, particularly for composite variables sensitive to both structure and spectral properties. The integration of Sentinel-1 and Sentinel-2 data outperformed single-source models across all metrics, reducing mean errors and enhancing feature discrimination. This was especially evident for tree count, crown base height, and crown volume.

RQ6: *Which fusion strategy, spectral only, polarimetric only, or combined spectral–polarimetric, yields the best trade-off between in-domain accuracy and spatial transferability?*

The combined spectral–polarimetric fusion strategy yielded the best balance between in-domain performance and cross-AOI generalization. Spectral-only models were stronger in local precision, while polarimetric-only setups offered better spatial robustness. The integrated model preserved the strengths of both, delivering the most stable and accurate outcomes across conditions.

RQ7: *How does the addition of temporal information to spectrally, polarimetrically, and temporally fused Sentinel-1 and Sentinel-2 data influence the performance and generalization of EO-based forest structure models?*

The inclusion of temporal information, i.e., multi-date acquisitions from both Sentinel-1 and Sentinel-2, substantially improved the robustness and predictive sharpness of fused models. By capturing seasonal and phenological variability, temporal fusion enriched both the spectral and structural feature space. This led to better model generalization, particularly in cross-AOI applications where ecological conditions differ.

Spectrally and polarimetrically fused features benefitted from the temporal dimension by reducing overfitting to single-date acquisition artifacts (e.g., shadows, soil moisture anomalies, leaf-off phases). Temporal fusion enabled the model to internalize dynamic patterns such as canopy growth, moisture cycles, and phenological

stages, enhancing predictions for both height-related variables and more transient attributes like canopy density.

The most significant gains were observed in ensemble models combining multi-date Sentinel-1 Kennaugh elements and Sentinel-2 spectral and spectral-Kennaugh features. These temporally fused setups delivered lower error rates and smaller performance drops under spatial transfer, confirming temporal diversity as a critical component of resilient EO modelling.

RQ8: *Which machine learning models, RF, SVR, CNN, or ensembles, perform best under varying EO input types and fusion configurations?*

RF consistently emerged as the top-performing model class across all input types. Their ensemble nature, resistance to overfitting, and insensitivity to input scaling made them highly suitable for EO data. SVRs performed well on some variables but lacked consistency. CNN showed localized success but required more tuning. Ensemble strategies, especially stacked RF ensembles, provided additional robustness in spatial transfer scenarios.

RQ9: *How do preprocessing choices affect model accuracy and spatial robustness, particularly under domain shifts?*

Preprocessing was found to be a decisive factor in model stability. Conservative masking (e.g., Mask > 1), moderate Z-score filtering ($Z = 1$), and aggressive outlier removal resulted in optimal balance between noise reduction and data retention. Over-filtering, however, sometimes eliminated ecologically informative variability, reducing generalization power.

RQ10: *Can ensemble learning approaches, particularly stacked RF ensembles, improve spatial generalization and mitigate performance degradation in unseen regions?*

Yes, stacked RF ensembles significantly improved performance in spatial transfer tasks. By integrating predictions from spatially stratified base models, the meta-learner compensated for regional biases and delivered smoother generalization. In several cases, ensemble predictions in unseen areas even outperformed in-domain results of single models, highlighting the strategy's value for operational applications.

RQ11: *What are the limitations of current models in achieving robust transferability, and how do fusion and ensemble strategies help overcome them?*

Limitations included sensitivity to ecological heterogeneity, noise in training data, and overfitting to local features. Single-sensor or non-temporal models performed poorly when transferred to distinct AOIs. Fusion, especially spectral–polarimetric–temporal integration, mitigated many of these issues, while ensemble strategies further reduced variability and compensated for localized weaknesses.

RQ12: *How do specific forest variables differ in their sensitivity to EO modality, preprocessing, and modelling approach?*

Variables tied to vertical structure (e.g., mean tree height, crown base height) were best predicted with SAR, particularly under domain shift. Horizontal attributes (e.g., tree count, tree area coverage) were more sensitive to optical data and required careful preprocessing. Volume-related metrics were highly dependent on fusion strategies and exhibited the greatest sensitivity to modelling and transfer conditions, benefiting most from ensembles and temporal depth.

6.6.3 Closing Remarks

This chapter has demonstrated that no single EO modality, feature type, or model architecture alone can universally address the complex challenges of forest structure prediction across space and time. Instead, robust and generalizable solutions emerge from strategic combinations, fusing spectral richness, structural sensitivity, and temporal dynamics within a unified learning framework.

The consistent performance of RF ensembles, the stabilizing influence of spectral and polarimetric transformations, and the generalization gains from temporal fusion all point toward a design paradigm that prioritizes complementarity over singular optimization. These results underscore the importance of ecologically grounded feature engineering and ensemble-based learning in operational EO pipelines.

Looking ahead, this foundational benchmarking not only informs methodological best practices but also sets the stage for next-generation forest monitoring systems that are scalable, transferable, and resilient to the inherent heterogeneity of real-world landscapes. The comparative strengths and trade-offs between model-modality strategies, as visualized in the multi-criteria radar plot (Figure 6.22), provide a compact summary of these insights.

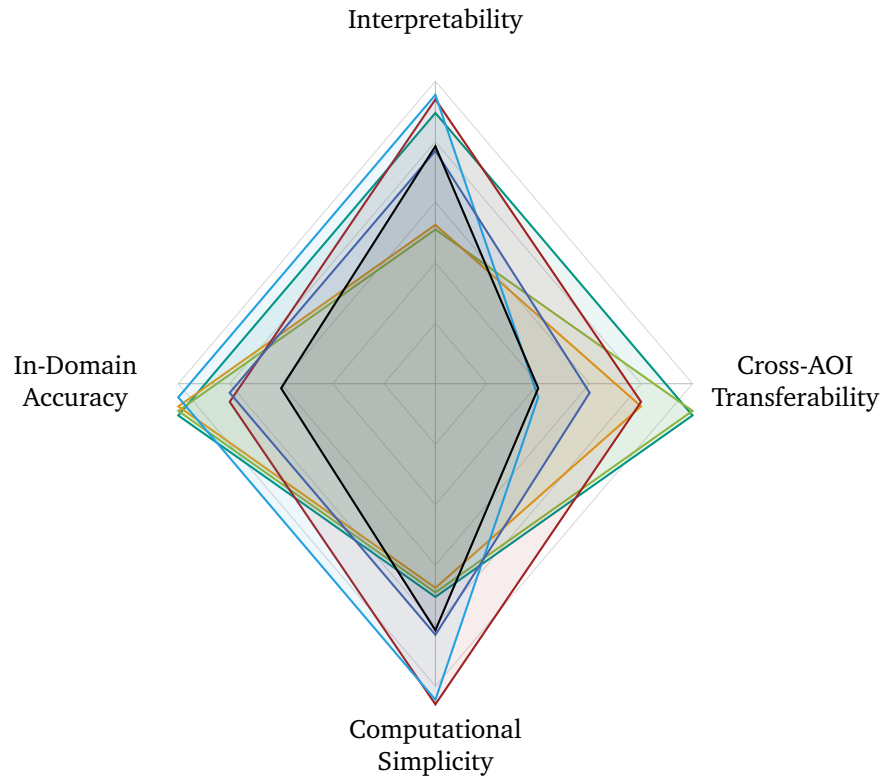


Figure 6.22.: Radar plot showing multi-criteria performance comparison of different Earth Observation modality-model strategies across four evaluation dimensions: In-Domain Accuracy, Cross-AOI Transferability, Computational Simplicity, and Interpretability.

7

Context-Aware Label Enrichment and Multi-Scale Learning with the HELIX Framework

> *A data model is not reality. It is a representation of selected aspects of reality.*
>
> — **William Kent**
> Computer Scientist, Data Modelling Theorist

Accurate forest and environmental modelling from EO data increasingly demands not only sophisticated predictors but also enriched and structurally meaningful target data. Traditional remote sensing workflows often rely on raw (irregular) or pixel-wise reference labels that fail to encode local context, spatial ambiguity, or multi-scale structure, factors that are central to ecological processes and patterns. This chapter introduces and systematically evaluates the **HELIX framework**, as a modular label-side augmentation strategy designed to address these limitations. HELIX, as described in detail in Section 4, enables models to learn from richer supervision by embedding spatial, temporal, and statistical context into reference label vectors. This is achieved through a combination of multi-scale contextual neighbourhood statistics and residual-aware feature augmentation derived from prior prediction uncertainty. In doing so, the HELIX transforms traditionally static or under-structured reference datasets, such as the Wald5Dplus label dataset [148], into dynamic, contextually aware learning targets.

The chapter therefore explores HELIX-based label enrichment across three core applications, each escalating in complexity and ecological ambition:

- **HELIX for Continuous Forest Structure Modelling:** The HELIX framework was applied to the task of continuous multi-target forest structure prediction by enriching both the feature and label spaces with spatial and residual context. In this

configuration, referred to as *HELIX+*, residual-aware feature augmentation was employed to encode spatial uncertainty, leveraging residuals from a baseline RF model trained on spectrally, polarimetrically, and temporally fused Sentinel-1 and Sentinel-2 EO data. Simultaneously, the target space was extended via contextual HELIX-based label enrichment, incorporating neighbourhood statistics across multiple spatial scales (3×3, 5×5, 7×7). This approach was shown to outperform all previously tested fusion and modelling configurations in terms of predictive performance, when compared to the previous Chapter 6.

- **Bark Beetle Calamity Modelling with Multi-Scale HELIX Labels:** HELIX-based label enrichment was also deployed in a real-world disturbance context, focusing on bark beetle outbreak modelling. Here, the fused Sentinel-1 and Sentinel-2 EO time series were matched with a dynamically enriched label set that captured both density and spread of infestations. The HELIX descriptors embedded spatial structure from both lagged and future outbreak patterns. A structured, four-stage modelling pipeline, ranging from logistic regression to ensemble-based temporal forecasting, was used to assess HELIX's ability to inform and reconstruct outbreak dynamics across space and time.
- **Seasonal and Short-Term Glacier Zone Modelling:** In the context of cryospheric analysis, the HELIX concept was employed to model seasonal glacier zonation using multi-temporal Sentinel-1 data. The enriched labels were combined with historical priors, derived from mean seasonal zonation between 2017 and 2020, and EO residuals, which were integrated into an ensemble XGBoost framework. The overall goal was to assess spatio-temporal transfer modelling possibilities. In a second setup targeting short-term prediction, mono-temporal Sentinel-1/2 inputs were used to forecast dynamic glacier zone changes, based exclusively on HELIX-derived categorical labels. These experiments illustrated HELIX's capacity to encode fine-scale spatio-temporal variation and generalize across both glaciological structures and time-frames.

Together, these experiments showcase the HELIX not as a simple label smoothing technique, but as a flexible label enrichment framework capable of enhancing learning targets across ecological domains. HELIX improves both model expressiveness and interpretability by enabling supervision to reflect spatial semantics, structural uncertainty, and neighbourhood dynamics.

7.1 Context-Aware Forest Structure Modelling Using Polarimetrically, Spectrally, and Temporally Fused Sentinel-1 and Sentinel-2 Data with Helix-Enriched Multi-Scale Labels

Previous chapters of this thesis have focused on evaluating multi-source fusion strategies for EO data and their predictive alignment with the Wald5Dplus forest paramter label dataset [148]. A variety of modelling configurations, including stand-alone RF and more, as well as ensemble modelling, were explored to assess the performance of various EO features under different spatial and temporal setups. A central contribution of this thesis has also been the development of the **Helix framework**, originally designed to enrich temporally static label datasets by injecting grid-aligned spatio-temporal context. Helix computes spatial context, seasonal signatures, and neighbourhood-level statistics that restructure discrete label points into dense, ML-compatible rasters with embedded structural cues. The experiment presented in this section builds directly on this foundation. It introduces a **context-enriched modelling pipeline** that complements Helix-style label-side enrichment with model-side feedback and supervision. Specifically, the HELIX-inspired architecture incorporates residual-driven feedback and multi-scale contextual label targets to guide learning. It closes a key gap: moving from context-aware label construction toward *context-aware model design*. The aim is not only to improve accuracy but also to allow the model to understand *when* and *where* label information is spatially reliable, structurally complex, or temporally ambiguous.

This experiment is thus guided by the following research questions:

RQ1: *Does label-side HELIX enrichment improve continuous forest parameter prediction relative to raw or single-scale targets?*

RQ2: *How do residual-based feature augmentations contribute to capturing spatial prediction uncertainty?*

RQ3: *What role does scale (e.g., 3×3 vs. 7×7 kernels) play in optimizing contextual label information for forest modelling?*

RQ4: *Can the integration of HELIX label enrichment with residual-aware modelling lead to improvements in interpretability or model stability?*

To assess these research questions, a two-stage modelling pipeline is implemented. First, a baseline RF is trained on fused EO inputs to establish a reference performance. Second, a HELIX-inspired model integrates residual-driven feedback and multi-scale contextual label enrichment to capture spatial structure and uncertainty. Comparative evaluation across these stages, using both original and enriched label targets, allows for systematic investigation of the role of context, residuals, and scale in predictive forest modelling.

7.1.1 Materials

The EO data used in this setup are identical to those introduced in Section 6.1.1, ensuring consistency with the experiments and benchmarks described therein. Specifically, the predictor dataset comprises a spectrally, polarimetrically, and temporally fused Sentinel-1 and Sentinel-2 EO stack, covering the years 2020 and 2021. This fused 512-band stack represents a rich multi-sensor EO input and is directly comparable to the dataset described in Section 6.1, where the fusion strategy and benchmark characteristics are detailed.

The target dataset consists of the same continuous Wald5Dplus label raster used in previous experiments. As described in detail in Sections 1.2.2 through 1.2.2, this label layer captures forest parameters in temperate Central European regions and provides a spatially dense, high-resolution reference signal for supervised modelling.

Together, this combination of multi-temporal, multi-modal EO predictors and semantically rich continuous forest labels forms a robust foundation for the experiments presented here, while maintaining direct comparability with prior results in the Wald5Dplus benchmark framework.

7.1.2 Methods

This HELIX-inspired architecture introduces a context-aware modelling strategy built in two phases:

Baseline Model Training: A baseline RF model is trained using the 512-band fused EO stack to predict the original 10-band Wald5Dplus target raster. This model operates without explicit spatial or temporal context and serves as a reference for residual-based feedback. Predictions are made over the full dataset, and the residuals per band (i.e., $y - \hat{y}$) are computed and stored as a 10-band raster.

Multi-scale Contextual Label Enrichment: The original 10-band label raster provides dense, per-pixel quantitative forest structure information (e.g., height, crown volume). However, these values are not context-free: their meaning is inherently local, e.g., a crown height of 22 m means something different in a patch of tall trees versus a sparse edge.

To make this contextuality explicit, a multi-scale representation of each label band is computed using mean filters over 3×3, 5×5, and 7×7 neighbourhoods. This does not simply smooth noise, it embeds the label in its *spatial semantic context*, allowing the model to learn structure-aware interpretations.

While this study focuses on continuous labels, the same logic applies to categorical label maps: majority pooling or soft one-hot encoding over neighbourhood windows could be used to encode contextual class distributions (e.g., dominant forest type or disturbance category), extending this method to hybrid label domains.

Context-Enriched Model: The Helix-based model builds on the limitations identified in the baseline pass by injecting residual-informed and context-augmented features into a second-stage learner. Specifically, the residuals produced by the baseline model, representing localized prediction errors per class, are concatenated with the original 512-band EO predictor stack, resulting in a 522-dimensional input space. These residuals encode implicit uncertainty and spatial ambiguity, effectively guiding the model to areas of known difficulty or structural complexity.

In parallel, the original label space is expanded from 10 to 40 dimensions through multi-scale contextual label enrichment. This includes the original per-pixel reference values and their corresponding spatial context windows (e.g., 3×3, 5×5, 7×7 local means), which allow the model to learn label structure as a function of its surrounding neighbourhood, and therefore adding information about the Spatio-contextual situation.

The enriched feature space and the enriched target space are used jointly to train a second RF model. This contextually-enriched model is no longer blind to the temporal, spatial, and structural characteristics of the input domain, it encodes context, uncertainty, and scale-awareness directly into its prediction process. Context matters because the same label value (e.g., crown area or tree height) can carry different ecological implications depending on its local neighbourhood, seasonal timing, or structural distinctiveness. By modelling the label not in isolation but in relation to its surrounding spatial and predictive environment, the appraoch allows for more stable, explainable, and generalizable predictions, particularly in heterogeneous or ambiguous forest regions.

Evaluation: Model performance is primarily evaluated on the original 10 forest structure classes to ensure consistency with the baseline configuration. However, the Helix+ model additionally outputs 30 auxiliary targets corresponding to the multi-scale smoothed labels. These are also evaluated to assess the model's ability to capture spatial trends and contextual label coherence across scales.

All predictions are evaluated using standard regression metrics: MAE, RMSE, MAD, and STD. This allows for both per-class performance analysis and cross-model comparison across original and context-enriched targets.

The full workflow is shown in Figure 7.1. It combines fused EO data, spatial context, and model self-awareness into a unified prediction strategy.

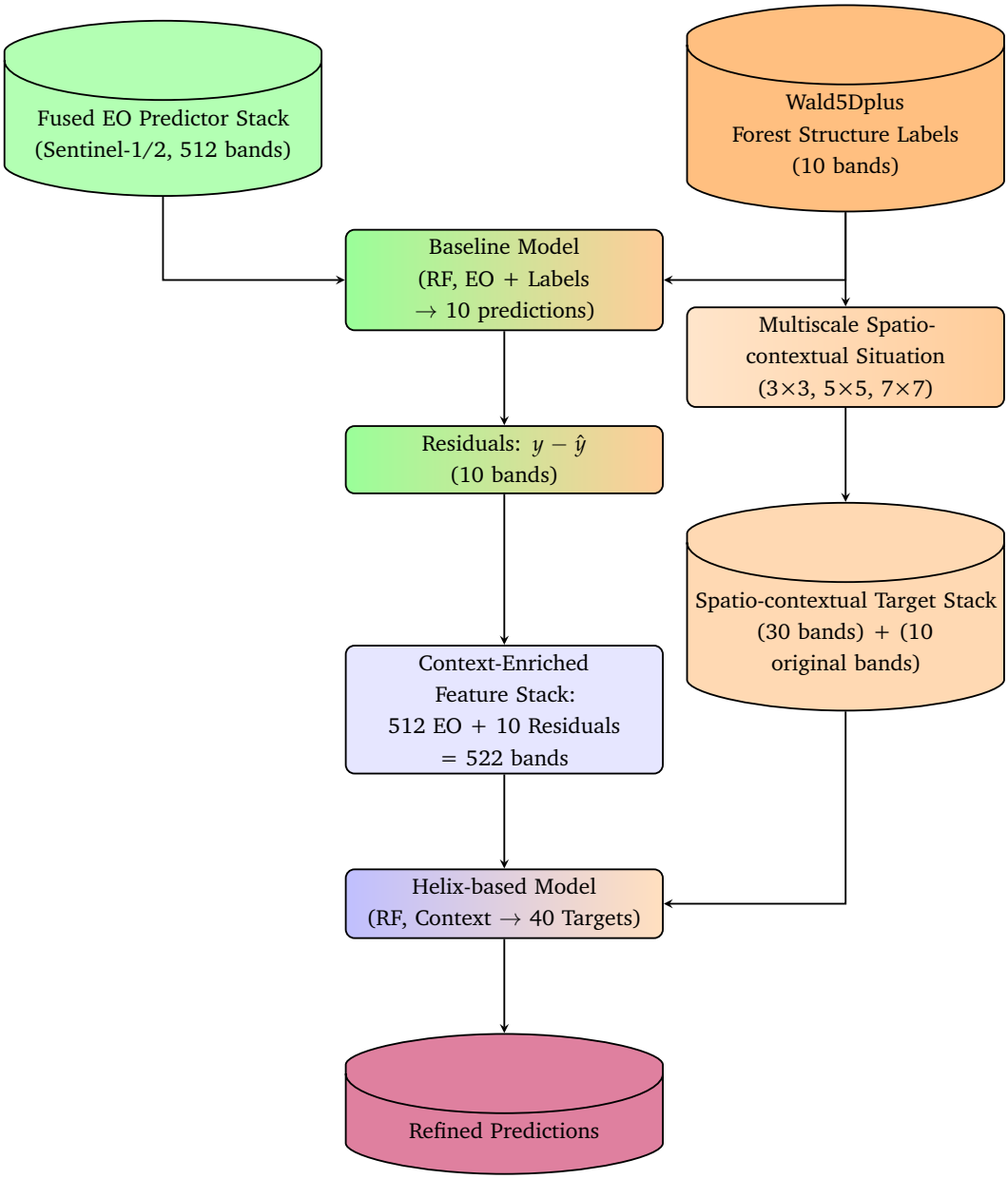


Figure 7.1.: Overview of the Helix-based modelling pipeline. A baseline model is trained on fused EO predictors and static Wald5Dplus labels to generate initial predictions and residuals. These residuals are reintroduced as uncertainty-aware features, while the original labels are enriched via multi-scale contextual averaging (Spatio-contextual Situation). The resulting context-enriched feature and target stacks are used to train a second-stage model, which outputs refined forest structure predictions.

7.1.3 Results

The Helix-inspired modelling framework was evaluated against the standard RF baseline trained on 512-band fused Sentinel-1/2 EO predictors. Performance was assessed across 10 static forest structure variables using four regression metrics: MAE, RMSE, MAD, and STD. Table 7.1 presents a full comparison of these metrics for both the baseline and Helix-based models.

Overall, the presented model demonstrated consistent improvements across all variables and all metrics. Notably strong gains were observed in highly structured biophysical variables, such as:

- **Sum crown area of deciduous trees**: MAE reduced by 3.43 m^2 (29% improvement)
- **Sum crown volume**: MAE reduced by 41.62 m^3 (42% improvement)
- **Tree area coverage**: MAE reduced by 2.73 percentage points (49% improvement)
- **Count of dead trees**: MAE halved, from 0.05 to 0.03 (49% improvement)

The benefits of contextual enrichment were also reflected in the reductions in RMSE, MAD, and STD across the board, indicating that Helix not only improves average predictive performance but also stabilizes residual spread and reduces model variance in complex or ambiguous regions.

To isolate the direct gain from the context-aware modelling, Table 7.2 summarizes the absolute and relative MAE improvements for each variable. All 10 variables showed positive Δ values, with gains ranging from 20% to 50%.

These results support the central hypothesis of this study setup: that residual-driven feedback and multi-scale contextual label enrichment provide meaningful spatial and structural information that can be exploited during supervised learning, even when the labels themselves are temporally static.

Table 7.1.: Performance comparison between the baseline RF model and the context-enriched Helix+ model on all 10 forest structure variables. Each entry reports MAE, RMSE, MAD, and STD for both models, with the improvement (Δ) shown beneath each metric.

Variable	MAE (Base / +)	RMSE (Base / +)	MAD (Base / +)	STD (Base / +)
Sum crown area of deciduous trees (m^2) Δ: 3.43 / 4.30 / 3.11 / 4.33	11.93 / 8.49	15.98 / 11.68	9.11 / 6.01	15.98 / 11.65
Sum crown area of coniferous trees (m^2) Δ: 3.26 / 4.01 / 2.87 / 4.05	11.61 / 8.35	15.27 / 11.26	8.95 / 6.08	15.26 / 11.22
Sum crown area of dead trees (m^2) Δ: 0.63 / 1.51 / 0.11 / 1.53	1.27 / 0.63	3.11 / 1.60	0.19 / 0.08	3.11 / 1.58
Count of deciduous trees Δ: 0.07 / 0.09 / 0.07 / 0.09	0.28 / 0.21	0.37 / 0.28	0.22 / 0.15	0.37 / 0.27
Count of coniferous trees Δ: 0.08 / 0.09 / 0.06 / 0.10	0.35 / 0.27	0.47 / 0.37	0.25 / 0.19	0.47 / 0.37
Count of dead trees Δ: 0.03 / 0.05 / 0.00 / 0.05	0.05 / 0.03	0.11 / 0.06	0.009 / 0.004	0.11 / 0.06
Tree area coverage (%) Δ: 2.73 / 3.83 / 1.45 / 3.84	5.62 / 2.89	8.83 / 5.00	2.58 / 1.13	8.83 / 4.99
Sum crown volume (m^3) Δ: 41.62 / 50.47 / 36.86 / 50.97	98.62 / 56.99	133.61 / 83.14	73.87 / 37.00	133.60 / 82.63
Mean tree height (m) Δ: 0.47 / 0.67 / 0.39 / 0.67	2.31 / 1.84	3.22 / 2.54	1.67 / 1.28	3.22 / 2.54
Mean crown base height (m) Δ: 0.64 / 0.74 / 0.64 / 0.78	2.31 / 1.67	3.01 / 2.27	1.82 / 1.18	3.01 / 2.23

Table 7.2.: Reduction in MAE for each forest structure variable between the baseline model and the Helix model. Absolute (Δ) and relative (%) improvements are reported.

Variable	MAE Baseline	MAE Helix+	Δ (Abs.)	Improvement (%)
Sum crown area of deciduous trees (m²)	11.93	8.49	−3.43	71.21
Sum crown area of coniferous trees (m²)	11.61	8.35	−3.26	71.91
Sum crown area of dead trees (m²)	1.27	0.63	−0.63	50.00
Count of deciduous trees	0.28	0.21	−0.07	73.46
Count of coniferous trees	0.35	0.27	−0.08	77.54
Count of dead trees	0.05	0.03	−0.03	50.61
Tree area coverage (%)	5.62	2.89	−2.73	51.45
Sum crown volume (m³)	98.62	57.00	−41.62	57.80
Mean tree height (m)	2.31	1.84	−0.47	79.50
Mean crown base height (m)	2.31	1.67	−0.64	72.22

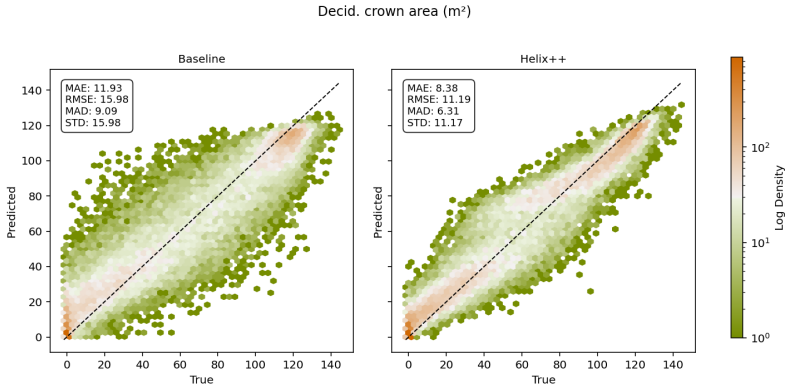


Figure 7.2.: Hexbin density plots comparing predicted versus true values for the variable *Sum crown area of deciduous trees (m²)* under the baseline model (left) and the Helix+ model (right). Each subplot visualizes the prediction distribution using log-scaled point density. The Helix model achieves lower error across all reported metrics (MAE, RMSE, MAD, STD), with visibly tighter clustering along the 1:1 diagonal, indicating improved predictive accuracy and reduced residual variance.

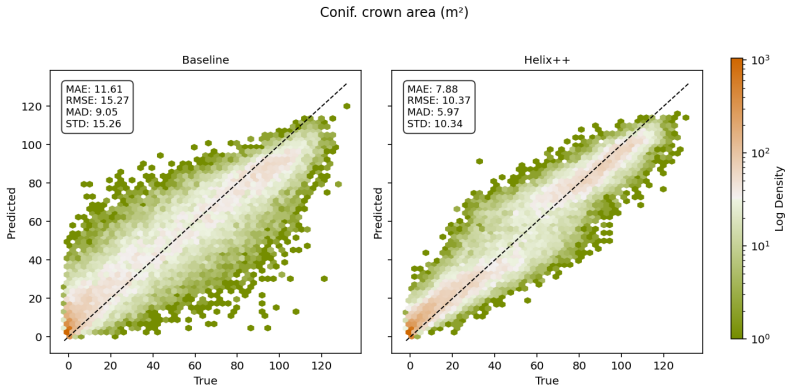


Figure 7.3.: Hexbin density plots comparing predicted versus true values for the variable *Sum crown area of coinferous trees (m²)* under the baseline model (left) and the Helix+ model (right). Each subplot visualizes the prediction distribution using log-scaled point density. The Helix model achieves lower error across all reported metrics (MAE, RMSE, MAD, STD), with visibly tighter clustering along the 1:1 diagonal, indicating improved predictive accuracy and reduced residual variance.

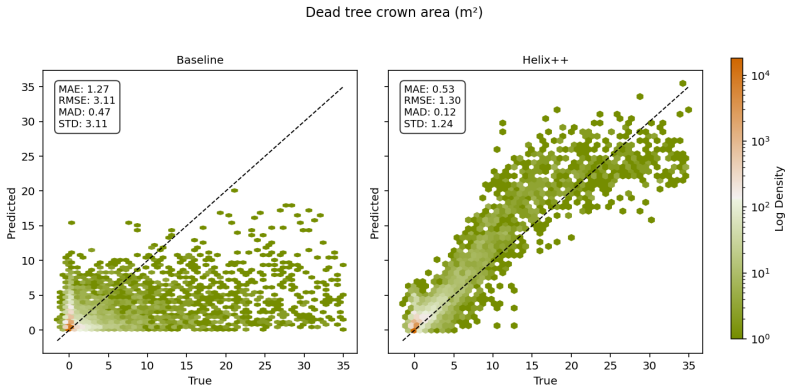


Figure 7.4.: Hexbin density plots comparing predicted versus true values for the variable *Sum crown area of dead trees (m²)* under the baseline model (left) and the Helix+ model (right). Each subplot visualizes the prediction distribution using log-scaled point density. The Helix model achieves lower error across all reported metrics (MAE, RMSE, MAD, STD), with visibly tighter clustering along the 1:1 diagonal, indicating improved predictive accuracy and reduced residual variance.

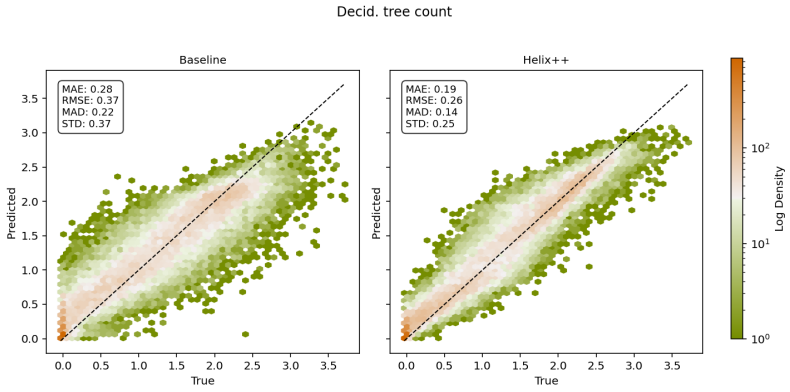


Figure 7.5.: Hexbin density plots comparing predicted versus true values for the variable *Count of deciduous trees* under the baseline model (left) and the Helix+ model (right). Each subplot visualizes the prediction distribution using log-scaled point density. The Helix model achieves lower error across all reported metrics (MAE, RMSE, MAD, STD), with visibly tighter clustering along the 1:1 diagonal, indicating improved predictive accuracy and reduced residual variance.

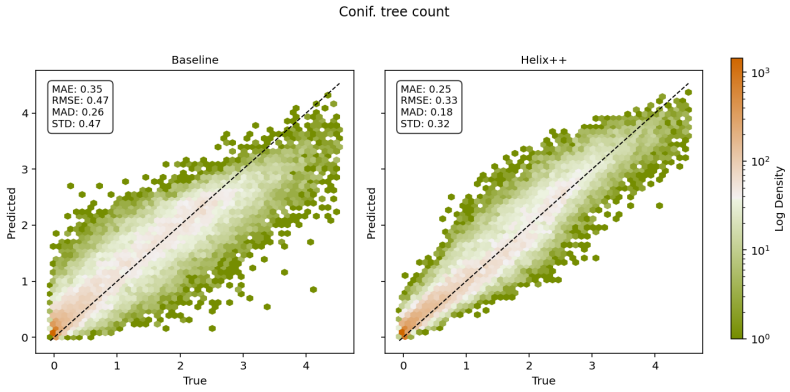


Figure 7.6.: Hexbin density plots comparing predicted versus true values for the variable *Count of coniferous trees* under the baseline model (left) and the Helix+ model (right). Each subplot visualizes the prediction distribution using log-scaled point density. The Helix model achieves lower error across all reported metrics (MAE, RMSE, MAD, STD), with visibly tighter clustering along the 1:1 diagonal, indicating improved predictive accuracy and reduced residual variance.

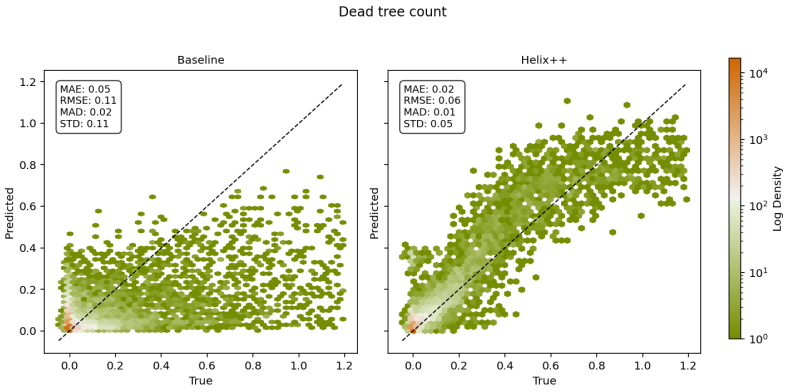


Figure 7.7.: Hexbin density plots comparing predicted versus true values for the variable *Count of dead trees* under the baseline model (left) and the Helix+ model (right). Each subplot visualizes the prediction distribution using log-scaled point density. The Helix model achieves lower error across all reported metrics (MAE, RMSE, MAD, STD), with visibly tighter clustering along the 1:1 diagonal, indicating improved predictive accuracy and reduced residual variance.

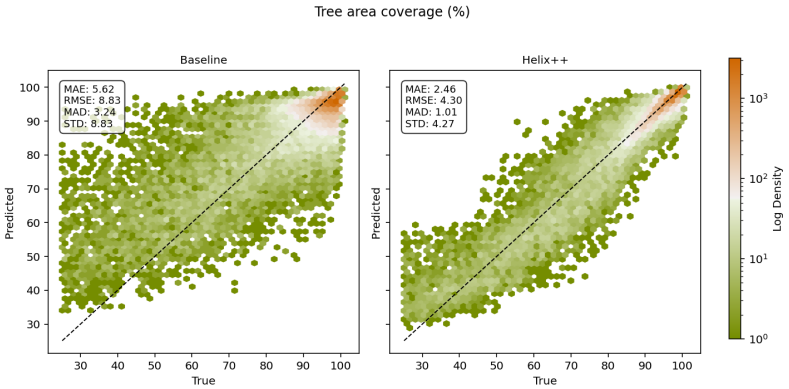


Figure 7.8.: Hexbin density plots comparing predicted versus true values for the variable *Tree area coverage (%)* under the baseline model (left) and the Helix+ model (right). Each subplot visualizes the prediction distribution using log-scaled point density. The Helix model achieves lower error across all reported metrics (MAE, RMSE, MAD, STD), with visibly tighter clustering along the 1:1 diagonal, indicating improved predictive accuracy and reduced residual variance.

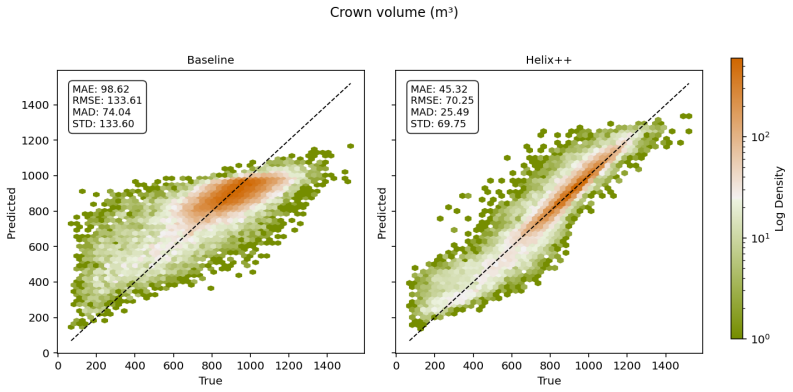


Figure 7.9.: Hexbin density plots comparing predicted versus true values for the variable *Sum crown volume (m³)* under the baseline model (left) and the Helix+ model (right). Each subplot visualizes the prediction distribution using log-scaled point density. The Helix model achieves lower error across all reported metrics (MAE, RMSE, MAD, STD), with visibly tighter clustering along the 1:1 diagonal, indicating improved predictive accuracy and reduced residual variance.

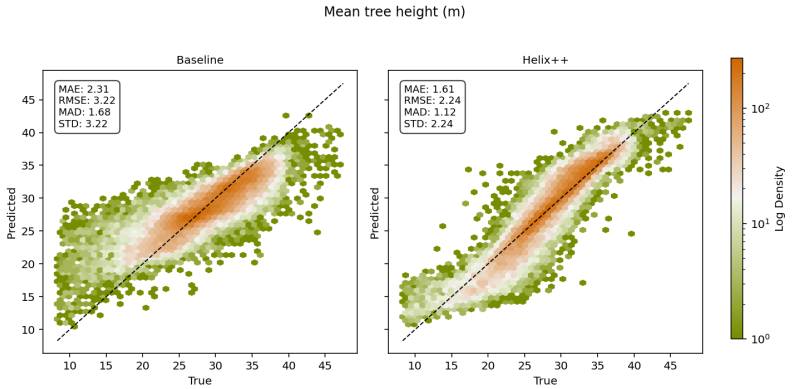


Figure 7.10.: Hexbin density plots comparing predicted versus true values for the variable *Mean tree height (m)* under the baseline model (left) and the Helix+ model (right). Each subplot visualizes the prediction distribution using log-scaled point density. The Helix model achieves lower error across all reported metrics (MAE, RMSE, MAD, STD), with visibly tighter clustering along the 1:1 diagonal, indicating improved predictive accuracy and reduced residual variance.

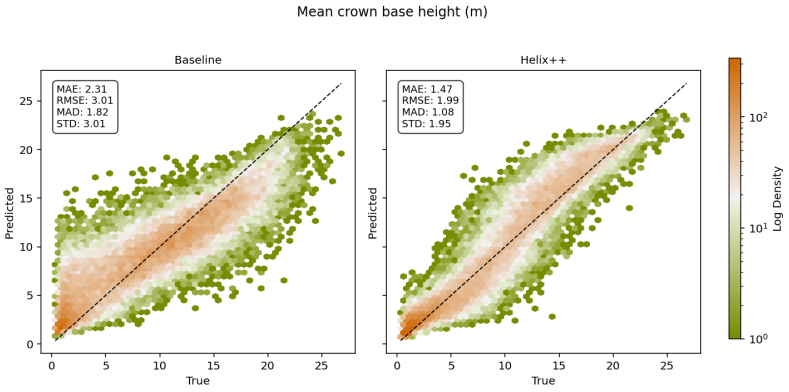


Figure 7.11.: Hexbin density plots comparing predicted versus true values for the variable *Mean crown base height (m)* under the baseline model (left) and the Helix+ model (right). Each subplot visualizes the prediction distribution using log-scaled point density. The Helix model achieves lower error across all reported metrics (MAE, RMSE, MAD, STD), with visibly tighter clustering along the 1:1 diagonal, indicating improved predictive accuracy and reduced residual variance.

The Helix+ model shows consistent and substantial improvements in predicting forest structure variables compared to the EO-only RF baseline. Figure 7.2 through Figure 7.12 present a series of side-by-side hexbin plots, visualizing the predicted versus true values for all 10 target classes. Each pair of subplots compares the baseline (left) and Helix model (right), with log-scaled point densities and overlaid regression metrics.

Across all targets, Helix+ achieves visibly tighter clustering around the 1:1 prediction line. This is particularly evident in structurally heterogeneous classes such as:

- **Sum crown area of deciduous trees (Figure 7.2):** MAE reduced from 11.93 to 8.49 m^2.
- **Sum crown volume (Figure 7.9):** MAE drops from 98.62 to 45.32 m^3, with a marked reduction in residual spread.
- **Tree area coverage (Figure 7.8):** MAE cut nearly in half, demonstrating improved generalization in dense canopy zones.

In count-based metrics (Figures 7.5, 7.6, and 7.7), Helix+ consistently shows lower absolute and relative error. Improvements are also observed in metrics of dispersion (MAD, STD), especially in complex or edge cases like dead tree crown area (Figure 7.4).

Table 7.3.: Top 10 most influential features across the first 10 Helix+ target regressors. Mean and standard deviation of feature importances are computed across models. Residual-based features rank prominently, highlighting the relevance of feedback signals.

Feature	Mean Importance	Std. Dev. Importance
EO_289	0.096	0.108
Residual_3	0.071	0.204
Residual_1	0.067	0.127
Residual_6	0.064	0.187
EO_129	0.054	0.077
Residual_9	0.050	0.136
Residual_10	0.046	0.132
Residual_8	0.045	0.130
EO_241	0.038	0.071
Residual_2	0.037	0.106

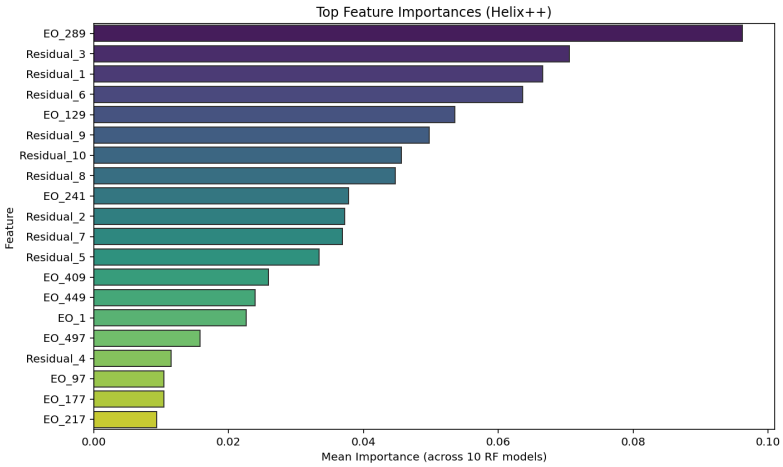


Figure 7.12.: Top 20 most influential features across the first 10 Helix+ target regressors.

Feature importance analysis further supports these findings. Table 7.3 lists the 10 most influential features used by Helix+ across its first 10 regressors. Notably, 7 of the top 10 features are residual-based (e.g., `Residual_3`, `Residual_1`, `Residual_6`), indicating that the model actively leverages prior model error as a contextual cue. This confirms that the model does not only learn from raw EO input but also from its own uncertainty patterns.

Figure 7.12 illustrates this further by showing the top 20 features graphically. The distribution emphasizes that both residuals and specific EO bands (such as `EO_289` and `EO_129`) are repeatedly relied upon, suggesting that certain EO signal bands are particularly sensitive to forest structure characteristics when augmented by contextual feedback.

These results underline the dual strength of Helix+: improved accuracy and improved awareness. The model does not only fit better, it fits more intelligently, guided by spatio-contextual situation and feedback-informed design.

7.1.4 Discussion

The effectiveness of the Helix+ model is supported by a threefold body of evidence. First, quantitative metrics such as MAE, RMSE, MAD, and STD (see Table 7.1) confirm consistent performance gains across all 10 target classes. Absolute error reductions of 20–50% were observed, particularly for structurally complex or sparse classes such as crown volume, dead tree count, and tree area coverage.

Second, predictive accuracy is visually validated through the series of hexbin plots in Figures 7.2–7.12. These plots provide a spatially explicit representation of prediction density, clearly showing how Helix+ reduces residual variance and shifts model output closer to the 1:1 ideal. The visual improvement is not only statistical, it is spatially interpretable and structurally meaningful.

Third, feature importance rankings (Table 7.3 and Figure 7.12) demonstrate how the model actively prioritizes contextual feedback. Residual-based features appear in 7 of the top 10 slots, confirming that Helix leverages its own prior uncertainty to refine future predictions. This highlights the model's ability to internalize error signals and use them as actionable context, a core novelty of the framework.

Together, these three perspectives, performance metrics, model visualizations, and feature introspection, converge on the same conclusion: Helix is not just more accurate; it is more self-aware, more stable, and more structurally attuned to the nature of the task.

The Helix-based framework assumes that residuals primarily reflect uncertainty or limitations in the EO predictor stack, rather than in the reference label set. This is a practical and often justified assumption in EO-driven ecological modelling, where label data (e.g.,

derived from airborne LiDAR or forest inventory) are typically of higher semantic fidelity than the often noisy, temporally variable, and sensor-specific EO inputs.

However, this assumption deserves further scrutiny. Residuals encode model misfit, without directly specifying whether that misfit arises from predictor-side ambiguity, label-side inconsistency, or mismatches due to georegistration errors, canopy occlusion, or temporal misalignment (e.g., phenological lag). In the current implementation, the model interprets residuals as indicators of feature-side uncertainty. This orientation makes sense for high-quality label products like Wald5Dplus [148], but future extensions may benefit from reversing this logic: using residuals to identify problematic or noisy regions in the label domain, particularly in datasets that integrate coarser or model-derived supervision.

This bi-directional use of residuals, either to refine EO-based inference or to question the validity of the reference labels, opens new avenues for self-diagnostic modelling. It may also enable hybrid quality-control workflows, where label and predictor confidence are estimated jointly.

Other key discussion points and future research directions include:

- **DOY-Aware Feature Augmentation:** Unlike many EO-based modelling setups that rely on sparse, single-date observations or add DOY metadata as a proxy for seasonal phase, this study incorporates a densely sampled, cloud-gapfilled time series covering 64 Sentinel-1 and Sentinel-2 acquisitions across a full annual cycle [147]. Temporal variability is thus embedded directly within the fused temporal, spectral, polarimetric feature space, allowing the model to infer seasonal progression and phenological dynamics from EO-observed reflectance and backscatter patterns. As a result, explicit DOY encoding was not necessary for this implementation. Future work could, however, investigate whether additional meta-temporal descriptors, such as phenological phase indicators, Fourier-based seasonality embeddings, or inter-annual climatological features, could further enhance model generalization, particularly when transferring to other years or regions.
- **Cross-Site Transferability:** The context-aware enrichment strategies presented here could improve model generalization across biogeographically diverse sites. This should be evaluated by transferring Helix+ to independent test areas with differing forest structure and EO acquisition conditions.

- **Uncertainty Quantification:** Although residuals act as a proxy for uncertainty, explicit uncertainty estimation (e.g., via quantile regression forests, ensembling, or dropout-based approximations) could enhance the interpretability and robustness of predictions in sensitive management contexts.
- **Extension to Categorical Labels:** The multi-scale enrichment approach is agnostic to label type. Majority pooling or soft one-hot encoding could be applied to categorical label maps (e.g., forest type or disturbance class), enabling Helix to support hybrid classification-regression tasks.
- **Model Compression and Operational Use:** Despite its interpretability, Helix is still ensemble-based and may require optimization for operational-scale inference. Model distillation or feature pruning could help streamline deployments while retaining contextual sensitivity.

In summary, Helix represents a first step toward interpretable, structure-aware forest modelling. It leverages the richness of EO predictors while remaining grounded in the spatial logic and semantic coherence of high-quality ecological labels. Its modular design allows for targeted extensions across data types, modelling paradigms, and forest monitoring scenarios.

Unlike many high-performing but opaque ML models, Helix+ remains interpretable by design. Its contextual feedback signals are explicitly defined, its target enrichment strategy is structured and reproducible, and its RF backbone offers direct access to feature influence, class-specific behaviour, and spatially traceable outputs.

7.1.5 Conclusions

This study introduces a context-enriched modelling architecture that advances EO-based forest structure prediction using Helix-inspired spatial label augmentation and residual-driven feedback. By integrating spatial context directly into both the target and feature domains, the approach enables structured learning of forest structure variables from fused Sentinel-1/2 inputs. The proposed two-stage pipeline demonstrates that modelling *label structure*, not just label values, yields measurable gains in accuracy, robustness, and interpretability.

Lessons Learned

- **Label-side spatial enrichment enables structure-aware learning.** Multi-scale contextual averaging of the Wald5Dplus labels allows the model to interpret values not in isolation, but as part of their spatial semantic neighbourhood, enhancing both accuracy and ecological realism.
- **Residual feedback improves spatial uncertainty awareness.** Integrating residuals from a baseline pass into the feature stack helps the model recognize structurally complex or ambiguous regions. This residual-driven feedback provides indirect supervision on model uncertainty and promotes more stable learning.
- **Context-enriched inputs and targets outperform EO-only baselines.** The Helix+ model consistently reduces errors across all 10 forest structure variables, with MAE reductions between 20–50%. Structured spatial context proves more informative than adding raw features alone.
- **Feature importance reflects interpretable model behaviour.** Residual features rank among the top contributors in most regressors, validating the hypothesis that model self-awareness (via error signals) supports more effective learning in heterogeneous forest environments.
- **The Helix framework bridges EO prediction and ecological structure.** By extending Helix from a label-processing tool to a modelling philosophy, this work establishes a reproducible foundation for structure-aware forest prediction that is compatible with both continuous and categorical domains.

Research Questions Revisited

RQ1: *Does label-side HELIX enrichment improve continuous forest parameter prediction relative to raw or single-scale targets?* Yes. Spatially contextualized label targets resulted in substantially lower MAE, RMSE, MAD, and STD across all forest structure classes. This confirms that multi-scale context offers valuable structural information beyond raw per-pixel values.

RQ2: *How do residual-based feature augmentations contribute to capturing spatial prediction uncertainty?* Residuals proved highly informative as input features. Their consistent presence among the top-ranked predictors across all regressors indicates that model-guided feedback helps localize and mitigate spatial uncertainty, especially in complex or edge regions.

RQ3: *What role does scale (e.g., 3×3 vs. 7×7 kernels) play in optimizing contextual label information for forest modelling?* Enriching labels at multiple scales enhances model flexibility. Small kernels embed fine-grained texture, while larger ones capture regional trends. The joint use of 3×3, 5×5, and 7×7 windows allows the model to learn from both local variation and broader structural patterns.

RQ4: *Can the integration of HELIX label enrichment with residual-aware modelling lead to improvements in interpretability or model stability?* Yes. Beyond accuracy gains, the Helix+ model demonstrates more interpretable behaviour, as shown by reduced variance, smoother residuals, and clearer spatial trends. Feature importance analysis also reveals a shift toward context- and feedback-based learning.

Closing Remarks

This chapter extends the HELIX framework from a spatial label enrichment tool to a full modelling strategy that combines structural supervision with predictive feedback. By injecting context into both labels and features, the approach transforms EO-driven forest prediction into a context-aware task where spatial relationships and model uncertainty are explicitly modelled.

The findings suggest that structured, interpretable forest modelling is not only feasible but beneficial, even in the absence of temporal supervision. The Helix-inspired pipeline offers a modular, reproducible foundation for extending EO-based inference to complex forest variables, and provides a roadmap for integrating residual awareness, scale sensitivity, and spatial semantics into future ecological modelling efforts.

7.2 Forest Disturbance Forecasting from Fused Sentinel-1 and Sentinel-2 Data with Helix-Based Spatio-Temporal Label Enrichment

This experimental setup systematically assesses the capacity of fused EO time-series data to detect and anticipate bark beetle disturbances in temperate forest ecosystems, focussing on the Bavarian Forest National Park, an ecologically unique and disturbance-prone site characterized by unmanaged natural dynamics and extensive deadwood accumulation [147]. EO observations are sourced from the same Wald5Dplus dataset [148] used in earlier chapters, incorporating spectrally, polarimetrically, and temporally fused Sentinel-1 and Sentinel-2 inputs. For labels, this experiment uses the ground-truth forest disturbance reference dataset (see Section 1.2.2) [203], which captures multi-year bark beetle outbreak patterns. To make this inherently dynamic process more learnable for ML models, the static disturbance labels are enriched into a temporally continuous, spatio-contextual reference dataset using the Helix framework (see Section 4). This enrichment encodes both the spatial structure of disturbance spread and the temporal sequence of outbreak progression, providing the model with context-aware supervision signals that capture both outbreak history and neighbourhood effects. As described in Section 1.2.2, the Bavarian Forest National Park presents a challenging and ecologically relevant test-bed for disturbance forecasting. Its unmanaged forest dynamics, large legacy of standing deadwood, and documented vulnerability to European spruce bark beetle outbreaks [331, 199] create a highly heterogeneous and temporally volatile disturbance regime. This includes multi-year outbreak cascades often triggered or amplified by stochastic factors such as storm events or localized drought stress. Modelling disturbance risk in such a context requires not only EO sensitivity to subtle structural changes but also the ability to infer disturbance likelihood from both spatial and temporal patterns of precursor conditions. The resulting Helix-enriched disturbance dataset therefore provides a harmonized, high-resolution spatio-temporal benchmark that links multi-modal EO signals with multi-scale, context-aware disturbance descriptors. This enables a structured sequence of synthetic forecasting experiments designed to address the following core research questions:

RQ1: *Do spatially enriched Helix descriptors provide additive predictive signal beyond raw outbreak labels?*

RQ2: *Which Helix descriptors, defined by year, scale, and statistic, carry the most discriminative power under current EO conditions?*

RQ3: *Can Helix descriptors be reconstructed from EO input alone, and how well does this mapping generalize across spatial and temporal contexts?*

RQ4: *Are Helix-derived spatial patterns learned from current and lagged data transferable to unseen future conditions, enabling EO-only prediction of future outbreak density?*

To address these, four interlinked modelling stages are presented:

1. A controlled classification setup using logistic regression to assess the marginal contribution of spatial kernel descriptors.
2. A per-band classification screening to identify the most informative Helix descriptors across space and time.
3. A regression pipeline mapping EO to Helix, quantifying the predictability of enriched descriptors.
4. A three-stage ensemble forecasting approach that leverages reconstructed Helix features to predict future outbreak density.

Together, these experiments probe the utility of Helix-based enrichment for both retrospective pattern discovery and prospective risk forecasting.

7.2.1 Materials

The EO data used in this experimental setup is described in detail in Section 6.1.1. In short, it consists of fused Sentinel-1 (VV/VH polarimetric C-band SAR) and Sentinel-2 (10,m BOA reflectance) time series data, processed via the HCB framework. This yields an interpretable, eight-dimensional feature space harmonizing radar and spectral signals. Temporal fusion produces consistent time series with one total intensity (K_0) and seven differential components ($K_{1\text{–}7}$), aggregated into 64 fused time steps. The fused EO data hence serves as a key component of the model input, the ARD cubes (see Figure 6.4), each containing 512 channels, introduced also in Section 2.2 and Section 2.2.

The ground-truth data used for training and evaluation is described in Section 1.2.2. The deadwood dataset [203] encompasses all forest areas identified as standing deadwood

from 1989 to 2023. It is updated annually based on aerial image interpretation and provides temporal resolution at the scale of one year.

7.2.2 Methods

The following pipeline unfolds in four conceptual stages: (i) a **foundational** classification setup that quantifies the added value of spatial kernel descriptors; (ii) a **diagnostic** ranking of individual Helix bands across space and time; (iii) an **inverse** regression analysis exploring to what extent Helix descriptors can be reconstructed from EO alone; and (iv) a **synthesis** stage where reconstructed Helix features are leveraged to predict future outbreak densities. This progression balances interpretability and expressiveness, and offers principled insight into the role of spatio-temporal label engineering in EO-based forest health monitoring.

Multi-Context Label Enrichment via Spatio-Temporal Helix Kernels

To improve the suitability of raw outbreak labels for spatial learning tasks, a multi-context enrichment method inspired by the Helix framework was developed. This procedure refines sparse, binary disturbance labels by applying both spatial and temporal kernel operations, producing a stack of continuous-valued descriptors that encode outbreak context across multiple scales and years.

The input consists of polygon-based annotations of bark beetle disturbances, each tagged with a `Change_Year` field denoting the onset of visible damage. For each year y within a specified window (e.g., 2017–2022), the polygons corresponding to that year are rasterized onto the shared spatial grid, with the EO data yielding a binary raster $\mathbf{L}_y$. Rasterization is clipped to the extent of a fixed EO reference raster, ensuring consistent spatial alignment and resolution.

Each raster $\mathbf{L}_y$ represents a snapshot of outbreak locations at year y. To enrich these labels, a two-step kernel-based strategy is applied. Uniform (boxcar) kernels were chosen over Gaussian or adaptive filters to preserve interpretability and consistency across spatial scales. While more complex kernels could encode distance decay or directional influence, the uniform design ensures that each pixel within the spatial window contributes equally, simplifying analysis and model transparency:

1. **Spatial kernel enrichment:** For each spatial radius $s \in \{1, 2, 3\}$, corresponding to kernel sizes of 3×3, 5×5, and 7×7 pixels, the raster $\mathbf{L}_y$ is convolved using a uniform square kernel centred at each pixel. Two descriptors are computed:

   - The *mean band* $\mathbf{H}^{\mu}_{y,s}$, reflecting local outbreak density:

$$\mathbf{H}^{\mu}_{y,s}(x, y) = \frac{1}{K_s} \sum_{(i,j) \in \mathcal{N}_s(x,y)} \mathbf{L}_y(i, j)$$

   - The *variance band* $\mathbf{H}^{\sigma^2}_{y,s}$, capturing heterogeneity or uncertainty in local outbreak patterns:

$$\mathbf{H}^{\sigma^2}_{y,s}(x, y) = \frac{1}{K_s} \sum_{(i,j) \in \mathcal{N}_s(x,y)} \left(\mathbf{L}_y(i, j) - \mathbf{H}^{\mu}_{y,s}(x, y)\right)^2$$

   Here, $\mathcal{N}_s(x, y)$ defines the spatial neighborhood around pixel (x, y), and $K_s = (2s + 1)^2$ is the number of pixels in the kernel.

2. **Temporal kernel stacking:** The spatially enriched descriptors $\mathbf{H}^{\mu}_{y,s}$ and $\mathbf{H}^{\sigma^2}_{y,s}$ are computed independently for each year y in the temporal window. These temporally aligned outputs are then concatenated to form a multi-temporal stack. This process implicitly defines a spatio-temporal kernel, as each pixel's final descriptor encodes both its spatial neighbourhood and its temporal context over multiple years.

The choice of mean and variance statistics is motivated by their complementary interpretability. The mean value $\mathbf{H}^{\mu}_{y,s}(x, y)$ provides a continuous approximation of outbreak presence, reflecting the proportion of outbreak pixels within the spatial neighbourhood. Values range from 0 to 1, and can be loosely interpreted as a localized density or probability of disturbance. The variance $\mathbf{H}^{\sigma^2}_{y,s}(x, y)$, in contrast, quantifies spatial inconsistency: it reaches its maximum when the neighbourhood contains an equal mix of outbreak and non-outbreak pixels, and drops to zero when labels are homogeneous. While alternative metrics such as local entropy or higher-order texture descriptors were considered, mean and variance were favoured due to their interpretability and stable behaviour under binary inputs. Their bounded range and intuitive meaning support downstream model reliability and diagnostics. This makes it a useful proxy for spatial uncertainty, edge proximity, or fragmentation. When these descriptors are computed across a temporal sequence of years, the resulting multi-band raster captures not only the spatial extent of outbreaks, but also their local intensity and contextual stability over time. A visual

example of these descriptors is shown in Figure 7.13, where the Helix `mean` and `variance` bands for 2020 ($s = 1$) are displayed alongside raw outbreak polygons, illustrating the complementary encoding of local outbreak density and spatial heterogeneity.

Temporal stacking was performed symmetrically around a reference year to capture both leading and lagging outbreak signals. This symmetric window provides a temporal context that includes pre-disturbance buildup and post-disturbance evolution, supporting the detection of both emerging and fading outbreaks. In addition to the temporal stack, a binary raster $\mathbf{L}_{\text{ref}}$ is produced for a designated reference year (e.g., 2020 or 2021), serving as a direct label for supervised learning. The final output is a multi-band raster, where each band corresponds to a specific year and kernel size combination, annotated with descriptive metadata (e.g., `helix_mean_y2020_s2`). This enriched label representation captures both outbreak concentration and structural uncertainty across space and time, providing a smoother and more informative target for learning models than raw labels alone. The enriched labels provide therefore a smoother, context-aware supervision signal for model training. In addition to improving spatial generalization, the variance bands offer a mechanism for model uncertainty estimation or post-hoc interpretability, enabling applications such as boundary detection or confidence-aware prediction.

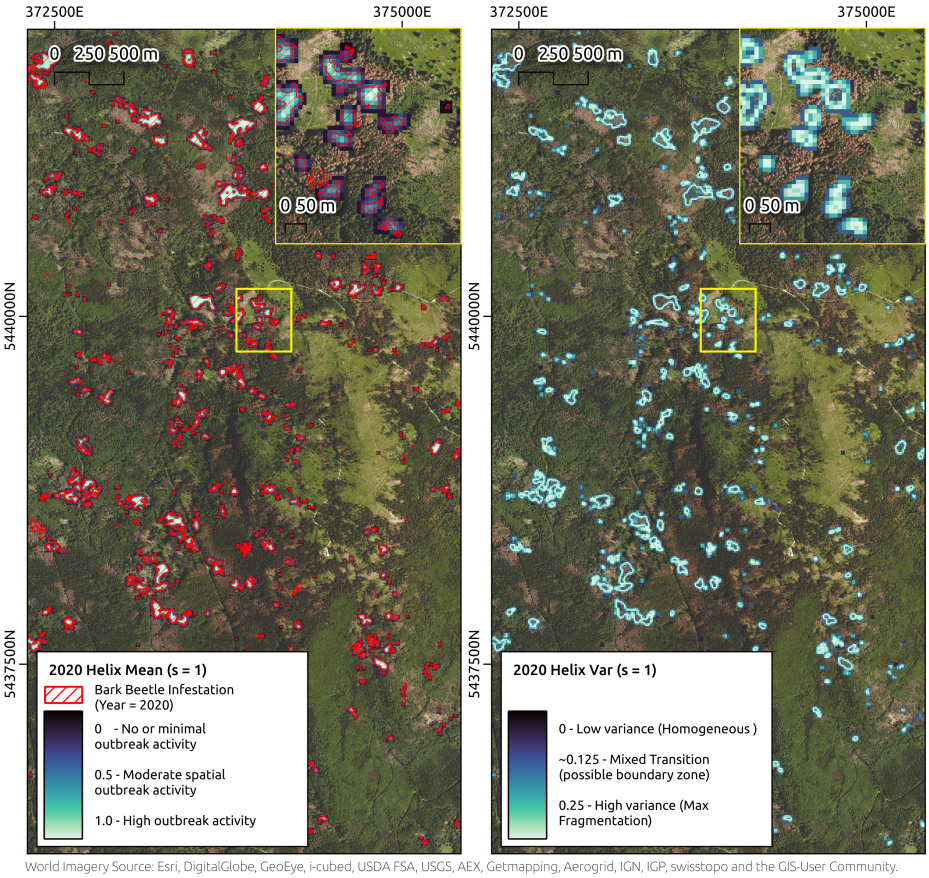


Figure 7.13.: Helix spatial descriptors for 2020 outbreak structure at kernel scale $s = 1$. The left panel shows the Helix `mean` band (`helix_mean_y2020_s1`), which encodes localized outbreak density, values range from 0 (no outbreak presence) to 1 (full neighbourhood affected). Red polygons denote bark beetle outbreak annotations for 2020. The right panel shows the Helix `variance` band (`helix_var_y2020_s1`), which captures spatial heterogeneity within each kernel neighbourhood. Variance values range from 0 (homogeneous areas, either all outbreak or all unaffected) to 0.25 (maximum spatial fragmentation, typically at outbreak boundaries). Mid-range values ($\sim$0.125) suggest partial infestation or edge zones.

Synthetic Learning Tasks for Outbreak Risk Characterization

Two distinct experimental setups were used to systematically evaluate the predictive value of Helix descriptors, a suite of *synthetic learning tasks*, so termed not because the data are simulated, but because the classification setup is artificially constructed to isolate

feature effects under balanced sampling and simplified assumptions. The first setup focuses on testing marginal contributions of spatial kernel descriptors under a controlled, interpretable setting. The second setup aims to assess the stand-alone informativeness of each Helix band using a flexible classifier.

In the first experiment, an interpretable logistic regression model is used to evaluate the marginal benefit of spatial kernel features, namely `mean` and `variance`, when added to EO input. This setup focused on assessing whether spatially aggregated Helix descriptors provided discriminative signal beyond the raw outbreak label, across multiple kernel sizes. Its emphasis was on controlled, interpretable comparisons of feature groupings under a balanced sampling regime.

The second experiment extended this framework to a per-band evaluation using a more expressive classifier (XGBoost). Each spatio-temporal Helix band (representing a specific year, kernel size, and statistic) was tested individually in combination with full EO input to quantify its contribution to outbreak classification. This experiment enabled a detailed ranking of Helix descriptors based on their stand-alone predictive power across space and time, offering insight into the temporal relevance and spatial scale sensitivity of the enrichment.

Together, these experiments provide a comprehensive perspective on the value of Helix-derived spatio-temporal features as inputs to predictive models of bark beetle dynamics.

Analysis of Kernel-Based Spatial Descriptors: To assess the descriptive and predictive value of Helix-derived features, a synthetic classification task focused on distinguishing high-density bark beetle outbreak regions was designed. The objective was to test whether specific spatial kernel descriptors, particularly `mean` and `variance`, contribute measurable predictive signal when added to EO data.

Model and Sampling Procedure: A logistic regression model was employed to perform the binary classification. Despite its name, logistic regression is a standard classification technique, mapping input features to a probability of class membership via a sigmoid function. This was selected for its interpretability and ability to reveal additive contributions of individual features (e.g., Helix mean and variance). While inherently a linear classifier, this simplicity allows to isolate and evaluate the marginal predictive benefit of adding spatial kernel descriptors to EO data, without introducing complex non-linearities or model interactions. The primary aim was not to maximize predictive performance, but to examine whether Helix features carry discriminative signal in a controlled, interpretable

setting. To ensure balanced class representation, a stratified resampling strategy was used to draw an equal number of samples from each class (200 per class). For each variant, the classifier was trained using 70% of the data and evaluated on the remaining 30%.

Label Definition: A binary classification target was derived by thresholding the Helix mean band (scale $s = 2$) at 0.5. Pixels with local outbreak density ≥ 0.5 were assigned class 1 (high-density), and all others class 0 (low-density). This synthetic label serves as a proxy to identify potential outbreak hotspots.

$$\mathbf{Y}_{\text{soft}} = \begin{cases} 1 & \text{if } \mathbf{H}_s(x, y) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

where $\mathbf{H}_s$ is the Helix mean band at spatial kernel size s.

Separate classification models were trained using EO features in combination with one of the following targets:

- $\mathbf{L}_{\text{raw}}$: the raw binary outbreak label,
- $\mathbf{H}_1$, $\mathbf{H}_2$, $\mathbf{H}_3$: Helix mean bands at kernel sizes $s = 1, 2, 3$.

For each target, four feature configurations were tested:

1. EO only,
2. EO + Helix mean,
3. EO + Helix variance,
4. EO + Helix mean + variance.

Performance was evaluated using the following metrics:

- Area Under the ROC Curve (ROC-AUC)
- F1 Score

These metrics quantify both the discriminative power of the classifier and its ability to capture the imbalanced nature of high-density regions. Results were grouped by spatial scale and visualized using ROC curves, allowing comparison of each Helix descriptor's contribution across contexts.

Per-Band Classification of Spatio-Temporal Helix Descriptors: As a complementary analysis, a second synthetic classification experiment was designed to evaluate the standalone predictive strength of each individual Helix enrichment band when combined with the full EO feature set. Unlike the first experiment, which aggregated spatial descriptors and emphasized interpretability via logistic regression, this setup employed a more flexible tree-based classifier (`XGBoostClassifier`) to test each Helix band independently. Each enrichment band, defined by a specific year, kernel size, and statistical function (mean or variance), was paired with the EO input, and its classification performance was assessed in isolation. This per-band approach enabled fine-grained evaluation and ranking of spatio-temporal Helix descriptors with respect to their discriminative value for identifying outbreak-prone regions.

Here, the objective was to predict outbreak presence as defined by the binarized `label_raw` raster (thresholded at 0.5), using:

- the EO stack from the year 2021, and
- one Helix band (either mean or variance) at a time.

For each of the 36 enrichment bands (spanning years 2017–2022 and spatial scales $s = 1, 2, 3$), a dedicated binary classification model was trained using `XGBoostClassifier`. This yielded per-band predictive performance under the same EO context, enabling direct comparisons across years, scales, and feature types.

All models were trained on an 80/20 stratified split and evaluated using ROC-AUC, Precision, Recall, and F1 Score. This per-band screening enables ranking of descriptors by informativeness, while also offering interpretability in terms of spatial and temporal relevance.

Per-Band Regression using Spatio-Temporal Helix Features

To evaluate the predictive strength of individual spatio-temporal Helix descriptors, a regression analysis was performed for each enriched band independently. The objective was to assess how well each Helix band, representing a specific spatial kernel statistic (mean or variance) at a given year, could be predicted from EO data alone.

The input features were derived from a multi-band EO raster for the year 2020. The target variables were extracted from the corresponding Helix-enriched raster for the same year, which included the raw binary outbreak label and a set of derived bands encoding

local outbreak density (mean) and spatial heterogeneity (variance) across a range of spatial kernels and temporal offsets.

For each target band, a separate regression model was trained using the EO features as predictors. A tree-based ensemble model (`XGBRegressor`) was employed for its robustness and ability to capture non-linear relationships. A train/test split (80/20) was applied to evaluate predictive performance, and the model was then used to produce full-resolution predictions across the study area. All regressions were executed in parallel to improve computational efficiency.

For each regression task, multiple performance metrics were recorded on the held-out test set:

- Mean Absolute Error (MAE)
- Median Absolute Deviation (MAD)
- Root Mean Squared Error (RMSE)
- Coefficient of Determination (R^2)
- Standard Deviation of Residuals

These metrics provide complementary views of prediction fidelity, robustness to outliers, and explanatory power. The predicted rasters were stacked and saved with band-level descriptors indicating the original Helix source band.

This per-band regression procedure serves both as a diagnostic tool for assessing the informativeness of individual Helix descriptors and as a proxy task to explore their potential utility in downstream modelling scenarios.

To then further assess the alignment between predicted values and their corresponding Helix descriptors, a post-hoc evaluation was conducted using both continuous and discretized metrics. In addition to regression scores, predictions and ground truth values were mapped into fuzzy ordinal bins spanning the interval $[0, 1]$ using predefined thresholds. This discretization enabled the computation of a fuzzy Intersection-over-Union (IoU) score, reflecting the consistency between predicted and true values across outbreak intensity classes. Class-wise precision and recall were also computed under this fuzzy scheme using macro-averaging, offering additional insight into the sharpness and balance of predictions. Residuals were computed per band enabling the detection of

systematic under- or over-estimation across different spatial kernel scales and temporal offsets.

Forecasting Future Outbreak Structure via Two-Stage Helix-Based Ensemble Regression

A three-stage ensemble regression framework was developed to investigate the potential of Helix-based spatial descriptors to support temporally generalizable outbreak forecasting. The primary goal of the experiment was to assess whether relationships learned between EO data and Helix descriptors, derived from past bark beetle outbreak activity, could be transferred to EO imagery from a future year. The approach aimed to emulate a realistic forecasting scenario in which outbreak-relevant spatial indicators are predicted for a future time step using EO data alone, without access to future labels during model training.

Stage 1: Learning Lagged Outbreak Structure from 2020 EO In the first stage of the pipeline, a set of gradient-boosted regression models was trained to reconstruct Helix descriptors encoding bark beetle outbreak structure from prior years, using EO data from a fixed reference year (2020). The targets included spatial mean and variance statistics of outbreak impact from 2017 through 2020, each computed at three kernel sizes ($s = 1, 2, 3$), yielding a total of 24 independent regression targets. These descriptors were derived from the labelled outbreak polygons and were designed to capture both outbreak density and fine-scale spatial heterogeneity.

The input feature space for all regressors consisted solely of spectral EO data from 2020, reshaped into a tabular pixel-wise format. For each Helix target band, a separate XGBoost model was trained to learn the mapping from 2020 EO to outbreak-related structure associated with a specific year and kernel size.

Crucially, the training targets spanned two temporal types: the 2020 Helix bands, which were co-temporal with the EO input and thus provided current-year supervision; and the 2017–2019 bands, which served as temporally lagged targets. These lagged labels introduced a quasi-memory component into the learning process, enabling the model to infer persistent or residual outbreak-related spatial signatures embedded in the 2020 EO imagery. This combination of current and lagged supervision was intended to enhance

the model's ability to generalize temporally when applied to EO data from subsequent years.

Following model training, prediction residuals were computed per pixel for each Helix band:

$$\varepsilon^{(i)}(x,y) = y^{(i)}(x,y) - \hat{y}^{(i)}(x,y), \quad i = 1,\ldots,N, \tag{7.1}$$

where $y^{(i)}(x,y)$ denotes the observed Helix value and $\hat{y}^{(i)}(x,y)$ the predicted value for band i at location (x,y), with $N = 24$ regressors in total.

To characterize reconstruction confidence across bands, residual mean and standard deviation were computed at each pixel:

$$\mu_\varepsilon(x,y) = \frac{1}{N}\sum_{i=1}^{N} \varepsilon^{(i)}(x,y), \tag{7.2}$$

$$\sigma_\varepsilon(x,y) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \left(\varepsilon^{(i)}(x,y) - \mu_\varepsilon(x,y)\right)^2}. \tag{7.3}$$

These summary statistics were appended to the EO feature space to construct an uncertainty-aware representation, designed to inform the ensemble model about local prediction fidelity and potential epistemic uncertainty, particularly in spatial regions that deviate from learned outbreak structure.

Stage 2: Temporal Transfer and Inference of 2021 Helix Descriptors The second stage involved applying the trained regressors from Stage 1 to EO imagery from 2021 in order to infer a complete set of synthetic Helix descriptors for that year. Since all models had been trained on EO from 2020 to reconstruct both current and lagged Helix structure (2017–2020), their application to 2021 EO constituted a form of temporal transfer learning. For each pixel in the 2021 image, the same 24 regressors were used to estimate their respective Helix bands, effectively projecting outbreak-related spatial structure learned from earlier years onto a new EO context.

These inferred Helix features were interpreted as temporally transferable outbreak descriptors, reflecting not actual labels, but synthetic estimates of outbreak-relevant structure

derived solely from current EO signals and the model's prior training. The inclusion of Helix targets from 2020 in Stage 1 was particularly important here, as it allowed the models to encode quasi-lagged outbreak context, providing a bridge between the input EO and latent spatial structure in the forecasting year.

In addition to the 24 predicted descriptors, the residual summary features computed in Stage 1, namely, the mean and standard deviation of prediction errors across regressors, were also re-applied to the 2021 EO data. This reuse operated under the assumption that spectral similarity implies comparable reconstruction uncertainty. The full feature representation for each 2021 pixel therefore consisted of: (1) raw EO bands, (2) inferred Helix descriptors for 2017–2020, and (3) the corresponding residual summary statistics.

Stage 3: Forecasting Aggregated 2021 Outbreak Density via Final Ensemble Regression

The final forecasting stage involved training a second-level regression model, referred to as the ensemble regressor, to predict outbreak intensity for 2021 using the full feature representation derived from Stage 2. The target variable for this prediction was defined as the average of the three Helix mean bands for 2021, each computed using a different spatial kernel. This aggregated target, denoted $\mathbf{T}_{\text{agg}}(x, y)$, was defined as:

$$\mathbf{T}_{\text{agg}}(x, y) = \frac{1}{3} \sum_{s=1}^{3} \mathbf{H}^{\mu}_{2021,s}(x, y), \tag{7.4}$$

where $\mathbf{H}^{\mu}_{2021,s}$ denotes the Helix mean band for kernel size s in the year 2021. The use of this multi-scale average aimed to produce a more robust and interpretable proxy for outbreak density that is less sensitive to individual kernel-scale variation.

To train the final ensemble model, the Stage 2 features were assembled over the 2020 spatial domain, i.e., using EO data from 2020, along with the predicted Helix descriptors (2017–2020) and associated residual statistics. The corresponding training labels were derived from the 2020 Helix mean bands, aggregated in the same way as the 2021 target. This setup ensured that the model learned to predict outbreak density from a temporally consistent feature-label pair, while incorporating both co-temporal and lagged spatial structure.

At inference time, the trained ensemble regressor was applied to the Stage 2 feature stack generated from EO data in 2021, enabling a pixel-level prediction of $\mathbf{T}_{\text{agg}}(x, y)$ for the forecast year. Critically, no Helix labels from 2021 were used in training the ensemble

model. The 2021 targets were held out entirely and used only for post-hoc evaluation, preserving strict temporal separation between training and forecasting domains. This design ensured that the model's performance represented a true generalization to unseen future EO data.

The output of this stage was a single-band raster in which each pixel represented the model's forecasted estimate of bark beetle outbreak intensity for the year 2021. Specifically, each value corresponded to the predicted average of the three Helix mean descriptors for 2021, reconstructed using only EO data and Helix descriptors inferred from past years. This aggregated prediction served as a synthetic proxy for outbreak density, integrating spatial structure across multiple kernel scales while remaining independent of any 2021 ground truth.

Because the true Helix mean bands for 2021 were available but never used in training, the predicted map could be directly compared to the held-out target for quantitative evaluation. This enabled a strict assessment of the model's temporal generalization capability under a genuine forecasting setup.

7.2.3 Results

The following section presents the empirical outcomes of all four experimental stages, each aligned with one of the core research questions. Results are structured to progressively evaluate the utility of Helix descriptors: starting with their effect on classification performance, moving through regression-based diagnostics, and culminating in forward-looking outbreak forecasts. Together, these results provide a comprehensive assessment of the Helix framework's capacity to enrich EO-based modelling across retrospective and prospective contexts.

Synthetic Classification Performance and ROC Analysis

This section evaluates how well Helix-derived spatial descriptors improve classification of high-density outbreak regions when combined with raw EO features. Using synthetic binary labels derived from Helix densities, a series of logistic regression models were trained across varying spatial kernel sizes and input configurations. Classification metrics such as ROC-AUC and F1 score are reported to quantify separability and predictive utility at multiple spatial resolutions.

Analysis of Kernel-Based Spatial Descriptors To evaluate the ability of Helix features to identify high-density outbreak zones, a series of logistic regression classifiers under varying input configurations were trained. Each model aimed to classify pixels as belonging to high-density regions ($\mathbf{H}_s \geq 0.5$) or not, using a stratified sample of 200 pixels per class. Results were evaluated across three spatial kernel sizes ($s = 1, 2, 3$).

Table 7.4 summarizes ROC-AUC and F1 scores across spatial scales and feature sets. The highest AUC of 0.962 was achieved for $s = 2$ and $s = 3$, with F1 scores peaking around 0.91. These results indicate that Helix descriptors are not only spatially interpretable but also predictive in identifying dense outbreak zones.

Table 7.4.: ROC-AUC and F1 scores for each spatial kernel scale and feature configuration.

Model	ROC-AUC	F1 Score
s1_t1 - EO only	0.903	0.832
s1_t1 - EO + mean	0.900	0.810
s1_t1 - EO + var	0.900	0.839
s1_t1 - EO + mean + var	0.902	0.825
s2_t1 - EO only	0.962	0.892
s2_t1 - EO + mean	0.962	0.906
s2_t1 - EO + var	0.960	0.891
s2_t1 - EO + mean + var	0.957	0.898
s3_t1 - EO only	0.962	0.892
s3_t1 - EO + mean	0.962	0.906
s3_t1 - EO + var	0.960	0.891
s3_t1 - EO + mean + var	0.957	0.898

Figure 7.14 displays the ROC curves for each kernel configuration and feature set. All configurations achieved high AUC values (≥ 0.90), with larger kernels generally yielding higher separability. Notably, models using only EO data with raw labels already performed strongly, and the marginal gain from adding Helix descriptors was modest but consistent.

Figure 7.15 provides an aggregated view comparing EO+Helix across kernel sizes versus EO+raw label. Helix-based features ($s = 2, 3$) clearly outperform the raw label baseline, confirming their added value for density-aware classification tasks.

Per-Band Classification of Spatio-Temporal Helix Descriptors The results of the per-band classification experiments are summarized in Table A.88. The highest F1 score

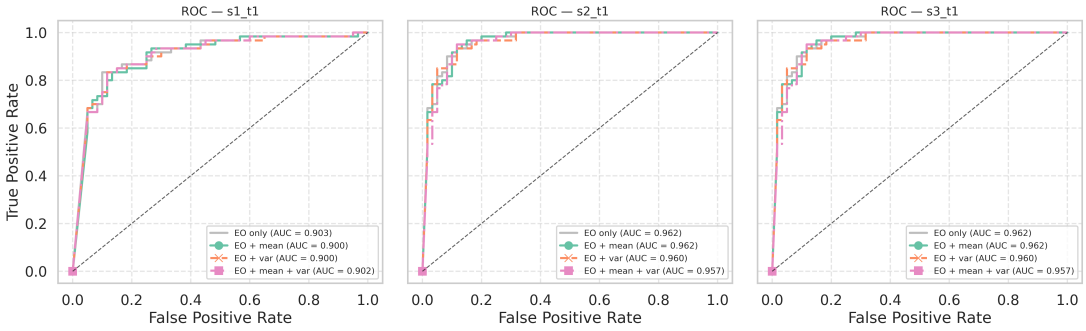


Figure 7.14.: ROC curves for synthetic classification task across spatial kernels ($s1$, $s2$, $s3$). Each subplot compares input configurations (EO only, EO+mean, EO+var, EO+mean+var).

was achieved by the feature `helix_mean_y2021_s1`, which yielded a Precision of 0.824, Recall of 0.821, and an AUC of 0.994. Features derived from the same year as the reference label (2021), particularly those computed at the smallest kernel size ($s = 1$), consistently outperformed others. A general trend was observed in which predictive performance declined with increasing spatial radius and increasing temporal distance from the reference year.

Notably, variance-based helix features performed worse than mean-based ones across nearly all years and scales. However, all evaluated bands retained a non-trivial level of predictive capacity (minimum $F1 \approx 0.58$), indicating that spatial kernel representations of past or adjacent-year outbreaks encode useful context for local risk estimation.

Regression Performance of Spatio-Temporal Helix Descriptors

To assess the informativeness of individual Helix bands, each descriptor was used as a target in an independent regression task, with EO features from 2020 as inputs. Regression performance was evaluated across three temporal groups, *past* (2017–2019), *current* (2020), and *future* (2021–2022), and compared across both mean and variance descriptors at multiple spatial scales ($s = 1, 2, 3$).

The following Tables, Table 7.5, Table 7.6 and Table 7.7, present the top-performing Helix descriptors within each temporal group, ranked by RMSE:

Overall Trends: The best predictive performance was consistently observed for bands from the current year (2020), with variance descriptors yielding lower RMSEs than

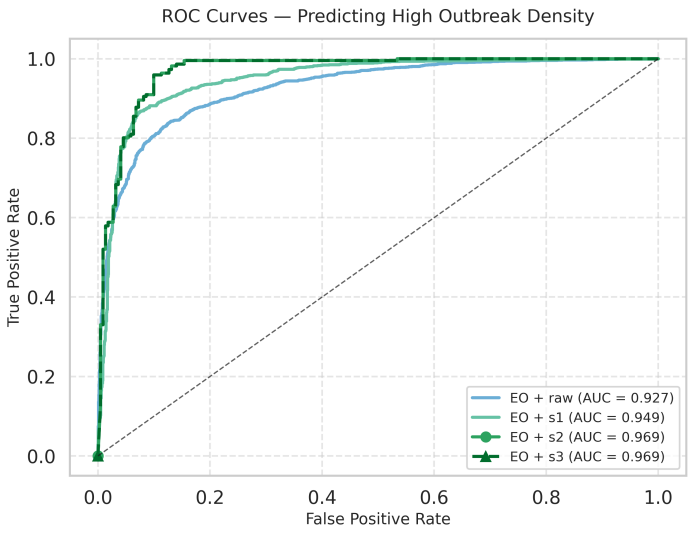


Figure 7.15.: Aggregate ROC curves for predicting high outbreak density. Models include EO combined with raw label, and EO with Helix mean at increasing kernel scales.

Table 7.5.: Top Helix Bands from 2020 (Current Year) Sorted by RMSE

Band	MAE	RMSE	R^2	Type
helix_var_y2020_s3	0.0203	0.0350	0.497	Variance
helix_var_y2020_s2	0.0191	0.0354	0.478	Variance
helix_var_y2020_s1	0.0171	0.0360	0.394	Variance
helix_mean_y2020_s3	0.0291	0.0555	0.583	Mean
helix_mean_y2020_s2	0.0305	0.0623	0.590	Mean

Table 7.6.: Top Helix Bands from 2019 (Past Year) Sorted by RMSE

Band	MAE	RMSE	R^2	Type
helix_var_y2019_s3	0.0243	0.0397	0.561	Variance
helix_mean_y2019_s3	0.0364	0.0656	0.641	Mean
helix_mean_y2019_s2	0.0377	0.0730	0.648	Mean
helix_var_y2019_s2	0.0233	0.0400	0.544	Variance
helix_var_y2019_s1	0.0209	0.0401	0.471	Variance

their mean counterparts. For example, `helix_var_y2020_s3` achieved the lowest RMSE (0.035) and highest R^2 (0.50) within the current group, indicating that spatial heterogeneity of outbreak labels was more learnable than raw density at fine scales. Mean

Table 7.7.: Top Helix Bands from 2022 (Future Year) Sorted by RMSE

Band	MAE	RMSE	R^2	Type
helix_var_y2022_s3	0.0399	0.0555	0.496	Variance
helix_var_y2022_s2	0.0398	0.0567	0.457	Variance
helix_var_y2022_s1	0.0378	0.0573	0.359	Variance
helix_mean_y2022_s3	0.0723	0.1125	0.576	Mean
helix_mean_y2022_s2	0.0789	0.1265	0.546	Mean

descriptors from 2020 also performed competitively, with increasing kernel size generally reducing RMSE, suggesting that larger spatial contexts improved signal coherence.

Temporal Comparison: Predictive performance degraded modestly for future bands and more noticeably for past descriptors. Variance bands from future years (e.g., `helix_var_y2022_s3`, RMSE = 0.056) remained more predictable than mean bands (e.g., `helix_mean_y2022_s1`, RMSE = 0.149), suggesting that structural cues of outbreak risk (e.g., heterogeneity) are retained more robustly over time than absolute outbreak density. Past-year bands showed relatively good performance, particularly in 2019, where both variance and mean bands at larger scales yielded high R^2 scores (e.g., `helix_mean_y2019_s2`, $R^2 = 0.65$), likely reflecting temporal proximity and continuity of outbreaks leading into 2020.

Scale Effects: Spatial scale (s) had a consistent effect across years: larger kernels ($s = 2$ and $s = 3$) produced more stable and accurate predictions, with higher R^2 and lower error across both mean and variance descriptors. This confirms that including broader spatial context helps models resolve signal in both outbreak density and boundary uncertainty.

Predicted Value Distributions: Value ranges for predicted bands were examined to assess calibration and realism. While some predictions exceeded the nominal range $[0, 1]$, the majority of values clustered below 0.2, consistent with the spatial sparsity of outbreaks. Variance descriptors showed a broader dynamic range than mean descriptors, reflecting their role in capturing transition zones and spatial fragmentation. Notably, variance predictions often peaked near 0.25, the theoretical maximum for binary input, validating their interpretation as spatial uncertainty indicators.

Residual Behaviour Across Bands and Temporal Contexts:

To further understand the nature of prediction errors for individual Helix descriptors, residual distributions were plotted for all bands across spatial scales and years (Fig-

ures A.1–A.9). These plots illustrate the deviation between predicted values and their ground truth enriched targets, highlighting not only central error tendencies but also skewness and heavy-tailed effects across the spatio-temporal landscape. **Residual Distributions:** In general, residuals were tightly centered around zero, with sharply peaked unimodal distributions indicating low systematic bias. Variance descriptors showed narrower and more symmetric residuals compared to mean descriptors, particularly at higher spatial kernel sizes ($s = 2, 3$). This pattern suggests that spatial heterogeneity is more consistently modelled than outbreak density. Broader and heavier-tailed residuals were primarily observed in mean descriptors, especially for temporally distant bands (e.g., 2017–2018), where sparse outbreak signals may have introduced underfitting or local overestimation.

Fuzzy IoU and MAE Trends: Fuzzy IoU values were highest for current-year descriptors (2020), with variance bands reaching up to 0.82 (e.g., `helix_var_y2020_s3`) and mean bands around 0.75. Past descriptors showed decreasing IoU with increasing temporal distance from the reference year; for instance, 2017 bands averaged an IoU of 0.63 (mean) and 0.67 (variance). Future years (2021–2022) exhibited moderate IoU values (typically 0.65–0.71), indicating partial predictability of forthcoming outbreak structure.

MAE values followed a similar trend: lowest for variance bands in the current year (e.g., 0.020 for `helix_var_y2020_s3`) and highest for mean bands from future years (up to 0.089 for `helix_mean_y2022_s1`). Across all temporal groups, larger spatial kernels ($s = 3$) consistently yielded lower MAEs and higher fuzzy IoUs, confirming that neighbourhood context improves signal coherence and reduces pixel-level noise in predictions.

MAE Distribution Across Bands: To further illustrate the comparative performance of Helix descriptors by year, Figure 7.16 shows grouped bar plots of MAE for all bands, stratified by temporal reference. Within each year, bands are ordered by ascending error. Several consistent trends emerge: first, variance descriptors systematically outperform mean descriptors in nearly all years, often appearing in the lower half of the MAE range. Second, performance improves with increasing kernel size, larger spatial contexts (e.g., $s = 3$) yield more stable and accurate predictions. Third, MAE increases with temporal distance from the reference year, reflecting decreasing signal quality in past and future bands relative to current-year EO inputs.

Precision and Recall: In the fuzzy ordinal class space, macro-averaged precision and recall scores ranged between 0.69 and 0.84, with variance bands again outperforming

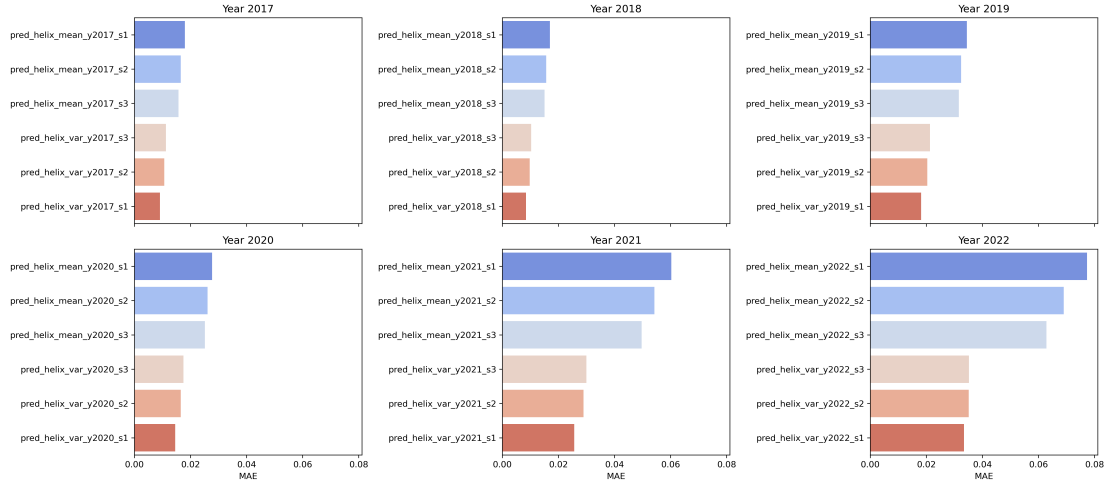


Figure 7.16.: Grouped bar plots of mean absolute error (MAE) for all predicted Helix bands, organized by year. Within each year, bands are sorted by MAE from low to high. Blue tones represent mean bands; red tones represent variance bands.

mean bands by a narrow but consistent margin. Precision-recall balance was strongest in temporally adjacent bands, reinforcing the notion that outbreak features captured by EO data are most informative for recent or imminent disturbance structure.

Table 7.8.: Summary of fuzzy evaluation metrics for selected Helix bands.

Band	MAE	Fuzzy IoU	Precision	Recall
helix_var_y2020_s3	0.020	0.82	0.84	0.83
helix_mean_y2020_s2	0.030	0.76	0.79	0.78
helix_var_y2018_s3	0.012	0.67	0.73	0.71
helix_mean_y2017_s1	0.022	0.63	0.69	0.68
helix_mean_y2022_s1	0.089	0.65	0.72	0.70

In general, the residuals exhibited symmetric and sharply peaked distributions centred around zero, indicating good overall calibration and low bias. This pattern was most pronounced in variance descriptors, particularly at larger spatial kernels ($s = 2$ and $s = 3$), which showed tight, leptokurtic distributions with minimal long-tail behaviour. This suggests that spatial heterogeneity cues, captured by variance bands, are more consistently modelled by EO-based regressors.

By contrast, mean descriptors showed broader, slightly skewed residual curves, with heavier tails extending toward overestimation. This was especially visible in early past years (e.g., 2017, 2018), where Helix mean bands likely encoded sparser outbreak

patterns. These results point to increased difficulty in learning outbreak density directly from EO data when temporal proximity to the reference year is low.

Over time, residual spread tended to increase slightly in future bands, with both mean and variance types exhibiting marginally broader distributions. Nevertheless, the residuals remained centred and tightly constrained in absolute magnitude (typically within $[-0.1, +0.1]$), reaffirming the effectiveness of Helix descriptors as learnable and predictive targets.

These findings support the earlier quantitative trends observed in fuzzy evaluation metrics and confirm that variance-based Helix descriptors are more stable targets across time, scale, and label structure. The residual diagnostics thus reinforce the design rationale of Helix as a multi-context representation capable of encoding not only outbreak magnitude but also uncertainty and boundary structure.

Ensemble-Based Forecast of Future Outbreak Density

Stage 1: Reconstruction Accuracy of Historical Helix Descriptors In the first stage of the ensemble forecasting pipeline, individual gradient-boosted regressors were trained to predict Helix spatial descriptors, representing mean and variance of bark beetle outbreak activity, from EO data collected in 2020. These regressors targeted both co-temporal descriptors (2020) and lagged outbreak structures (2017–2019), simulating the model's ability to encode both current and residual spatial outbreak signals.

Table 7.9 summarizes the predictive performance across all 24 Helix bands, measured in terms of MAE, RMSE, and coefficient of determination (R^2). Overall, mean descriptors exhibited higher predictive fidelity than variance descriptors, with RMSE values typically below 0.04 and R^2 values frequently exceeding 0.7, especially for lagged outbreak years. These findings suggest that the EO features retained a consistent imprint of past and present outbreak spatial structure, particularly for kernel sizes $s = 2$ and $s = 3$.

Table 7.9.: Stage 1 reconstruction metrics for Helix descriptors derived from 2020 EO data.

Band	Context	MAE	RMSE	R^2
	2020			
helix_mean_y2020_s1	2020	0.0096	0.0210	0.6561
helix_mean_y2020_s2	2020	0.0147	0.0317	0.7710
helix_mean_y2020_s3	2020	0.0101	0.0205	0.6798
helix_var_y2020_s1	2020	0.0285	0.0442	0.5429
helix_var_y2020_s2	2020	0.0490	0.0799	0.6220
helix_var_y2020_s3	2020	0.0296	0.0439	0.5708
	2017–2019			
helix_mean_y2017_s1	2017–2019	0.0176	0.0453	0.7387
helix_mean_y2017_s2	2017–2019	0.0090	0.0224	0.5674
helix_mean_y2017_s3	2017–2019	0.0160	0.0363	0.7603
helix_var_y2017_s1	2017–2019	0.0270	0.0587	0.7456
helix_var_y2017_s2	2017–2019	0.0143	0.0294	0.5833
helix_var_y2017_s3	2017–2019	0.0257	0.0504	0.7406
helix_mean_y2018_s1	2017–2019	0.0104	0.0222	0.6440
helix_mean_y2018_s2	2017–2019	0.0153	0.0321	0.7539
helix_mean_y2018_s3	2017–2019	0.0110	0.0217	0.6653
helix_var_y2018_s1	2017–2019	0.0163	0.0295	0.6339
helix_var_y2018_s2	2017–2019	0.0247	0.0450	0.7323
helix_var_y2018_s3	2017–2019	0.0172	0.0293	0.6460
helix_mean_y2019_s1	2017–2019	0.0164	0.0430	0.7571
helix_mean_y2019_s2	2017–2019	0.0083	0.0210	0.5886
helix_mean_y2019_s3	2017–2019	0.0153	0.0358	0.7688
helix_var_y2019_s1	2017–2019	0.0594	0.1090	0.5829
helix_var_y2019_s2	2017–2019	0.0253	0.0432	0.4688
helix_var_y2019_s3	2017–2019	0.0534	0.0913	0.6070

Stage 2: Ensemble Regression Accuracy on 2020 Data In the second stage, the extended feature stack, consisting of 2020 EO data, inferred Helix descriptors (2017–2020), and residual statistics, was used to train an ensemble of XGBoost regressors to predict the aggregated Helix mean and variance descriptors for 2020. This stage tested the

ensemble's ability to integrate multiple feature modalities to recover outbreak-related spatial structure with high fidelity.

As shown in Table 7.10, the ensemble regressors achieved consistently low error across all bands. MAE values were below 0.005 for all mean bands and under 0.02 for variance bands. Coefficient of determination (R^2) scores exceeded 0.87 in all cases, and peaked above 0.95 for the $s = 2$ kernel mean band. This performance indicates the ensemble's strong capacity to consolidate spatial information and recover target structures from the enriched 2020 representation.

Table 7.10.: Stage 2 ensemble regression metrics for predicting 2020 Helix mean and variance bands.

Band	Context	MAE	RMSE	R^2
helix_mean_y2020_s1	Ensemble_2020	0.0031	0.0096	0.9285
helix_mean_y2020_s2	Ensemble_2020	0.0048	0.0132	0.9600
helix_mean_y2020_s3	Ensemble_2020	0.0039	0.0102	0.9204
helix_var_y2020_s1	Ensemble_2020	0.0121	0.0236	0.8700
helix_var_y2020_s2	Ensemble_2020	0.0195	0.0368	0.9199
helix_var_y2020_s3	Ensemble_2020	0.0137	0.0241	0.8702

Stage 3: Forecast Accuracy on 2021 Outbreak Descriptors The final stage of the forecasting pipeline aimed to infer spatial outbreak intensity for the year 2021 using the two-stage ensemble regression model trained solely on 2020 EO data and historical Helix descriptors (2017–2020). Forecasting performance was evaluated against held-out 2021 labels, focusing on the three Helix mean bands computed at kernel radii $s = 1, 2, 3$. These bands, representing spatially smoothed descriptors of outbreak density, were used as the primary prediction targets in the ensemble due to their superior predictive performance in earlier stages.

Although both Helix mean and variance descriptors were reconstructed during Stage 1, only the mean bands were retained as targets for the final forecast model. This exclusion was based on empirical observations: variance descriptors consistently exhibited weaker regression performance, both in terms of error magnitude and R^2, compared to the mean bands. The mean descriptors proved more stable and interpretable as outbreak density proxies, likely reflecting their stronger correlation with EO-based spectral features. Thus, this choice can be interpreted as a form of empirical feature selection.

The resulting forecast prediction corresponds to the mean Helix descriptors for 2021 at spatial kernel sizes $s = 1$, $s = 2$, and $s = 3$, respectively. These predictions were compared to the true 2021 Helix labels, scaled to the unit interval $[0, 1]$, where higher values indicate increased local outbreak density. Table 7.11 summarizes the prediction accuracy for each of the three spatial kernels.

Table 7.11.: Stage 3 Forecast Accuracy for 2021 Mean Helix Bands

Band	Context	MAE	RMSE
helix_mean_y2021_s1	Forecast_2021	0.0474	0.1501
helix_mean_y2021_s2	Forecast_2021 (center)	0.0298	0.0822
helix_mean_y2021_s3	Forecast_2021	0.0464	0.1294

The central kernel prediction ($s = 2$) yielded the lowest forecast error, with a MAE of 0.0298 and RMSE of 0.0822. These results indicate that the ensemble model, trained without access to any 2021 outbreak labels, was able to generalize previously learned relationships between EO patterns and outbreak structure to the EO data of an unseen future year. The comparatively higher errors at $s = 1$ and $s = 3$ may reflect edge-related noise and spatial over-smoothing effects, respectively. Overall, the strongest signal transfer appeared at intermediate spatial resolution, consistent with the predictive patterns observed in Stage 1 and Stage 2.

A qualitative comparison between predicted and true Helix mean values for kernel size $s = 2$ is shown in Figure 7.17, highlighting the close alignment of spatial structure in the forecasted and labelled maps.

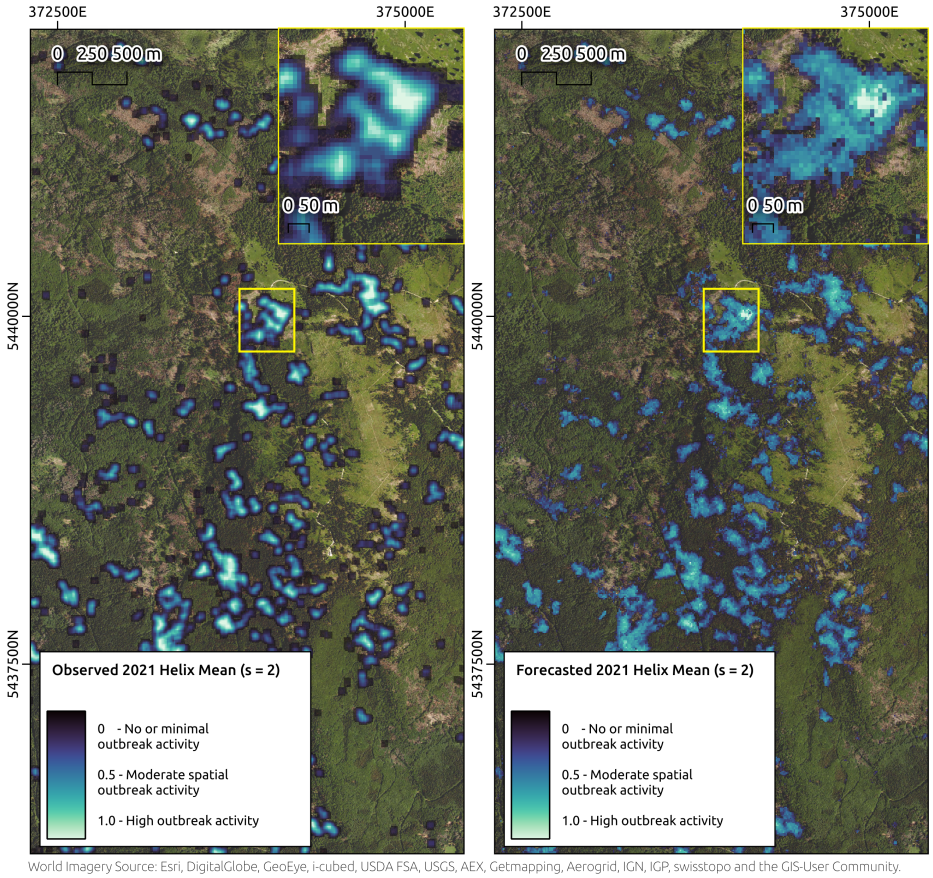


Figure 7.17.: Visual comparison of 2021 Helix outbreak intensity (mean descriptor, spatial kernel $s = 2$). **Left:** Ground-truth Helix mean band derived from labelled outbreak polygons. **Right:** Predicted Helix mean band obtained via ensemble forecasting model using only pre-2021 data. Both maps are normalized to the range [0, 1], where 0 denotes low or no outbreak intensity and 1 indicates maximal inferred damage.

7.2.4 Discussion

This study setup explored the use of Helix-derived spatial descriptors to enrich remote sensing data for bark beetle outbreak monitoring. Helix, a framework for constructing smoothed, multi-scale label representations from sparse outbreak data, was evaluated across multiple experimental regimes, including synthetic classification, per-band regression, and forward-looking ensemble prediction. Across all tasks, Helix features

consistently improved predictive performance and interpretability, supporting their viability for both retrospective analysis and prospective ecological forecasting.

Label Engineering as Structured Regularization

One of the less overt but impactful aspects of the Helix framework lies in its function as a structured regularization mechanism for label space. By embedding raw outbreak labels within spatial neighbourhoods and temporal offsets, Helix imposes inductive priors that promote continuity and suppress noise. This is especially pertinent given the known limitations of the input labels, which are sparse, binary, and potentially lagged relative to actual disturbance onset.

The spatial kernels act as smoothing operators, transforming fragmented binary masks into more coherent representations of outbreak density and heterogeneity. This better aligns with the spatial scale of ecological processes and supports improved model calibration. In this way, Helix does not merely augment the input space, it restructures the learning problem itself, transitioning from brittle pixel-wise classification to context-aware inference.

This can be interpreted as a form of weak supervision, where spatial priors derived from neighbourhood statistics serve as soft constraints on learning. The resulting models benefit from this structure, achieving better generalization and greater robustness to local mislabelling or sampling irregularities.

Temporal Generalization and Lag Structure

Helix descriptors also offer a lens on temporal generalization. Across tasks, a consistent trend was observed: descriptors closer in time to the EO reference year (2020) were more predictable, while those further removed, either in the past or future, exhibited higher error and weaker correlation. This suggests that while EO imagery captures enduring traces of disturbance structure, the predictive signal has a limited temporal reach.

This reflects a natural decay in ecological memory: the influence of past outbreaks on present-day canopy state diminishes over time, and future conditions introduce new variability. The current Helix implementation assumes uniform time steps and fixed-width windows; however, more sophisticated variants could explore non-uniform lags, learned

decay functions, or temporally adaptive kernels aligned with bark beetle life cycles or climatic anomalies.

Descriptors might also be extended to model temporal dynamics more explicitly, for example, through autoregressive components or momentum terms that encode not just presence but directional change. Such refinements would better reflect the non-stationary and cascading nature of forest disturbance processes.

While Helix bridges some of this temporal ambiguity through multi-year context aggregation, the annual resolution of outbreak labels imposes a fundamental limit on temporal precision. Label timestamps may not align exactly with spectral change, especially in cases of delayed canopy response or asynchronous infestation onset. This underscores the need for caution when interpreting short-term trends and points to a potential role for higher-resolution reference data, such as aerial surveys or phenology-derived disturbance maps, in future iterations.

Multi-Kernel Spatio-Temporal Context Enrichment

At its core, Helix implements a scalable, interpretable form of multi-kernel spatio-temporal enrichment. By computing local statistics over varying neighbourhood sizes and years, it introduces controlled redundancy and multi-scale abstraction into the modelling pipeline. This diversity in spatial scale is critical: smaller kernels preserve fine-grained detail, while larger ones capture broader context and transitional zones.

This design reflects the ecological reality that bark beetle outbreaks span nested spatial extents, from individual trees to landscape-scale epidemic fronts. Helix translates these multi-scale effects into structured inputs that are readily learnable by downstream models, effectively scaffolding them toward more ecologically plausible representations.

Unlike end-to-end deep learning pipelines that may learn such structure implicitly, Helix makes spatial context explicit and modular. This not only improves model performance, as demonstrated in both classification and forecasting tasks, but also enhances interpretability and operational reuse.

In this light, Helix functions as a flexible, context-aware label engineering strategy that allows remote sensing models to internalize spatio-temporal structure without sacrificing transparency or control.

Limitations and Assumptions

Despite the promising results, several limitations and assumptions of this study warrant attention.

First, the spatial kernels applied in Helix are isotropic, meaning they assume uniform spread in all directions. This design was chosen deliberately for interpretability and computational tractability, but it does not capture the anisotropic dynamics often observed in real-world bark beetle spread, such as preferential movement along wind corridors, slope gradients, or forest stand boundaries. However, given the absence of high-resolution spatial driver layers (e.g., fine-scale wind fields or detailed stand connectivity maps), more sophisticated, directionally-weighted kernels were beyond the scope of this study. Future research could integrate terrain-informed or process-based spread models where such data becomes available.

Second, the temporal enrichment was based on a fixed, symmetric window of annual snapshots, which does not adapt dynamically to local outbreak progression rates or climatic triggers. Nevertheless, the modelling framework explicitly incorporated multi-year lagged Helix descriptors as input features, allowing the model to learn from both current and historical outbreak patterns when forecasting future disturbance risk. This setup demonstrated that the model could generalize temporal dependencies without requiring future labels at training time. Nonetheless, future extensions could explore more flexible temporal kernels or integrate time-aware features such as phenological phase indicators or outbreak progression stages to further improve temporal sensitivity.

These simplifications were intentional, balancing interpretability, data availability, and computational feasibility. They provided a controlled foundation for demonstrating the broader value of spatio-temporal label enrichment in EO-driven disturbance forecasting.

Foundational Insights from Synthetic Classification

The synthetic classification experiments provide foundational evidence for the discriminative value of Helix descriptors. Augmenting EO inputs with Helix-derived `mean` bands consistently improved classifier performance, particularly at spatial scales $s = 2$ and $s = 3$. While EO-only models achieved strong baselines (ROC-AUC $\approx$ 0.90–0.96), Helix-enhanced models improved F1 scores and maintained or exceeded AUC performance.

These gains validate the hypothesis that spatially aggregated outbreak signals offer complementary information beyond what is captured in raw EO imagery.

In contrast, `variance` descriptors underperformed, especially when combined with `mean`, suggesting that while heterogeneity may encode edge structure or transition zones, it may also introduce noise in binary classification tasks. Kernel size $s = 2$ emerged as the most balanced configuration, likely reflecting an optimal trade-off between neighbourhood coverage and local specificity.

Diagnostic Value via Per-Band Regression

The per-band regression experiments served as a diagnostic lens into the learnability of Helix descriptors from EO. Variance descriptors consistently achieved lower MAE and RMSE and produced tighter, more symmetric residuals compared to `mean` descriptors, particularly at larger spatial scales. This suggests that spatial heterogeneity, interpretable as uncertainty or fragmentation, is more stably imprinted in EO features and more transferable across years.

Temporal degradation followed expected trends: past descriptors became harder to reconstruct as the lag increased, though recent years (e.g., 2019–2020) retained strong signal. Future descriptors were also partially predictable, especially variance bands, which appear to capture structural continuity across time better than mean density maps.

Inverse Modelling and Predictive Generalization

The forecasting framework presented here operates in a fully inverse fashion: rather than learning from future outbreak outcomes directly, it leverages observable EO data from a fixed year (2020) to reconstruct both concurrent and lagged Helix descriptors (2017–2020). This inversion of the traditional time-forward prediction paradigm is central to the model's generalization capacity. By anchoring the learning process in a single-year EO snapshot time-series, the model effectively learns a spatial outbreak signature, one that is encoded not in outbreak polygons but in EO-correlated proxies.

Across Stage 1 and Stage 2, results demonstrated that Helix descriptors corresponding to both 2020 and prior years could be accurately reconstructed from 2020 EO input. Particularly strong reconstruction was observed for Helix mean bands with intermediate spatial kernels ($s = 2$), indicating that outbreak-relevant structure is most salient at

this scale. Importantly, this outcome confirms that lagged outbreak signatures persist in the EO signal of a subsequent year, enabling the model to develop a quasi-memory mechanism.

When applied to 2021 EO imagery in Stage 3, the model was able to transfer its learned spatial associations to forecast future outbreak descriptors, without any access to 2021 outbreak labels during training. The central prediction (mean descriptor for $s = 2$) achieved the lowest forecast error (MAE = 0.03, RMSE = 0.08), further confirming that the inverse modelling approach supports effective temporal generalization.

Bands 1–3 of the final prediction output correspond to synthetic reconstructions of historical descriptors (especially resembling 2018, as evidenced by spatial similarity), while bands 4–6 represent the ensemble's direct forecast of 2021 Helix mean bands. This separation is deliberate: only the latter bands were evaluated against held-out 2021 labels, as they reflect the forward inference objective of the pipeline.

Notably, variance bands were excluded from the final ensemble forecasting step due to weaker reconstruction performance in earlier stages. This reflects a feature selection decision grounded in empirical accuracy, and highlights a potential avenue for future improvement in uncertainty-aware outbreak modelling.

Synthesis: A Unified Spatio-Temporal Enrichment Framework

Taken together, the experiments reveal a consistent finding: Helix-based enrichment serves as a powerful mediator between noisy EO inputs and the complex, structured patterns of ecological disturbance. From simple classification tasks to inverse modelling and ensemble forecasting, Helix enabled models to internalize spatial context in a transparent and modular way.

Rather than relying solely on raw labels or black-box learning, Helix provides a middle ground, where structure, scale, and uncertainty are made explicit and learnable. In doing so, it transforms past outbreak data into a reusable spatial prior that supports both retrospective analysis and prospective forecasting under environmental change.

Looking forward, extensions of Helix may incorporate terrain-informed kernels, sub-annual temporal representations, or dynamic outbreak models. Yet even in its current

form, Helix demonstrates the potential of structured label engineering to advance remote sensing-based ecological monitoring, balancing interpretability, performance, and operational utility.

7.2.5 Conclusions

This setup investigated the role of spatio-temporal label engineering, operationalized through the Helix framework, in enhancing remote sensing models for bark beetle outbreak monitoring. By reformulating sparse, binary disturbance labels into smoothed, multi-scale descriptors, Helix aims to bridge the gap between noisy reference data and the structured ecological processes they intend to represent. The core mechanism, multi-kernel spatio-temporal context enrichment, translates outbreak history into a learnable representation of ecological structure, enabling models to internalize both local intensity and broader contextual patterns. Through a series of targeted experiments, this work evaluated the representational, diagnostic, and predictive value of these descriptors under both retrospective and forward-looking scenarios.

Research Questions Revisited

Through a structured sequence of four modelling stages, the following key research questions are now addressed:

RQ1: *Do spatially enriched Helix descriptors provide additive predictive signal beyond raw outbreak labels?* Yes. Logistic regression experiments using synthetic labels derived from Helix densities revealed that spatial kernel descriptors offer consistent performance improvements over raw EO inputs alone. Models integrating EO data with Helix mean and variance features achieved AUC values above 0.96 and F1 scores up to 0.91, surpassing models trained on raw EO and label data. This confirms the added value of spatial contextualization in outbreak classification.

RQ2: *Which Helix descriptors, defined by year, scale, and statistic, carry the most discriminative power under current EO conditions?* The per-band classification analysis revealed that Helix descriptors derived from the same year as the outbreak label (2021) consistently yielded the strongest predictive signal, with the highest F1 scores and AUC values. In particular, `helix_mean_y2021_s1` achieved an F1 score of 0.822 and an AUC of 0.994, outperforming all other bands.

Performance tended to decline with increasing spatial kernel size and with temporal distance from the target year. Descriptors from past years (e.g., 2017–2020) and larger kernels ($s = 2, 3$) still retained meaningful predictive capacity, though with reduced precision and recall. Variance-based descriptors were generally less informative than their mean-based counterparts.

These findings indicate that Helix descriptors computed at finer spatial scales and from temporally aligned outbreak data are most effective for outbreak detection, while lagged and coarse-scale features still contribute useful but weaker spatial context.

RQ3: *Can Helix descriptors be reconstructed from EO input alone, and how well does this mapping generalize across spatial and temporal contexts?*

Yes. Regression models trained to reconstruct Helix descriptors from 2020 EO data achieved strong performance, especially for descriptors from 2020 and nearby years. The results from of regression tasks and subsequent residual analyses indicate that Helix descriptors, encoding spatially enriched outbreak statistics, can be effectively predicted from spectral EO data alone. Reconstruction accuracy was highest for descriptors from the reference year (2020), but generalization extended to lagged (2017–2019) and even future (2021–2022) years with only moderate degradation. Variance descriptors proved particularly robust across temporal contexts, with lower RMSE and tighter residual distributions than their mean-based counterparts, likely due to their focus on spatial heterogeneity rather than outbreak magnitude.

Generalization was further modulated by spatial kernel size, with broader context ($s = 2, 3$) consistently improving predictive fidelity. These findings validate the Helix framework's capacity to abstract stable spatial outbreak features and demonstrate that such descriptors are not only learnable from EO input, but transferable across both space and time, making them suitable as intermediate supervision targets in forecasting and spatial inference tasks.

RQ4: *Are Helix-derived spatial patterns learned from current and lagged data transferable to unseen future conditions, enabling EO-only prediction of future outbreak density?*

Yes. The ensemble forecasting pipeline, trained exclusively on EO observations from 2020 and Helix descriptors from 2017–2020, was able to predict 2021 outbreak intensity, specifically, Helix mean descriptors, at a pixel level with meaningful accuracy. The model achieved its lowest error at spatial scale $s = 2$ (MAE =

0.0298, RMSE = 0.0822), indicating that mid-range spatial context was optimal for transferring learned structure to unseen future conditions. Importantly, no 2021 label information was used during training, confirming the predictive generalization of Helix representations under strict temporal holdout.

The model's outputs closely resembled the ground truth outbreak intensity maps for 2021, especially at intermediate kernel sizes, reinforcing the idea that outbreak-related spatial features captured in Helix descriptors are not only learnable but temporally stable. This supports the use of Helix as a generative representation for forecasting tasks in EO-driven ecological modelling.

Together, these results demonstrate that Helix descriptors offer a robust, learnable, and temporally transferable representation of forest disturbance structure, capable of supporting EO-based forecasting pipelines even in the absence of future labels.

Closing Remarks

Overall, this setup presents a usage example of the Helix framework, as a novel and practical way, for enriching ecological labels in spatio-temporal remote sensing tasks. By explicitly encoding local structure, temporal continuity, and predictive uncertainty, Helix enhances the alignment between EO inputs and ecological targets. The framework's versatility across classification, regression, and forecasting tasks speaks to its broader applicability, not just in forest health monitoring, but in other domains where spatially sparse yet structurally rich reference data limit current modelling capabilities.

In an era of growing ecological risk and expanding remote sensing archives, tools like Helix offer a path toward more interpretable, generalizable, and robust monitoring pipelines, blending domain insight with machine learning in a principled way.

7.3 Seasonal Glacier Facies Forecasting from Temporally Fused Sentinel-1 Data and Helix Labels

This chapter presents the methodological foundation for the spatio-temporal prediction of glacier facies based on satellite-derived EO data. The approach builds upon the recognition that glacier surface zones, such as wet snow, percolation, and superimposed ice, exhibit gradual, seasonal transitions rather than static states. Capturing such transitions requires more than traditional classification snapshots; it demands a representation that reflects temporal continuity and surface evolution. Accurate, seasonal forecasting of glacier facies is critical for understanding surface energy balance, meltwater production, and the timing of key hydrological processes in polar regions. This is particularly important in the Canadian High Arctic, where in-situ observations remain sparse due to logistical challenges and remoteness. Satellite-based, EO-driven facies modelling can help fill this gap, enabling continuous, synoptic monitoring of glacier surface conditions. More importantly, being able to forecast facies transitions, rather than just classify them retrospectively, supports both operational glaciological monitoring and long-term climate impact assessments. This experiment addresses this need by testing whether seasonally enriched labels and temporally fused Sentinel-1 data can produce generalizable, temporally aware facies predictions in data-scarce Arctic environments. To enable this, the framework integrates EO time series with an enriched supervision signal derived from multi-temporal glacier zone maps. These label datasets, initially structured as discrete, class-based products at regular intervals, are transformed into temporally aggregated representations aligned with the seasonal progression of glacier surface processes. The underlying assumption is that EO-driven models benefit not only from the spatial patterns present in each scene, but from the embedded seasonal structure across time. The notion of seasons, spring, summer, fall, and winter, therefore serves as an implicit organizing principle for both the labels and the learning process, reflecting climatologically distinct phases in glacier surface behaviour. By aligning EO signals and glacier zone annotations in both space and time, and by re-structuring the label domain into seasonally enriched representations, the method aims to support predictive models that generalize across years, locations, and seasonal regimes. The chapters that follow detail the construction of these enriched labels, the supervised learning setup, and the transfer scenarios used to evaluate spatial and temporal generalization performance.

This experiment is guided by the following research questions:

RQ1: *Does seasonal label enrichment improve the model's ability to represent glacier facies transitions compared to discrete classification targets?*

RQ2: *How well do temporally fused Sentinel-1 features predict enriched seasonal glacier zone dynamics?*

RQ3: *Can a residual-based refinement stage enhance prediction accuracy and robustness across seasons?*

RQ4: *Does the inclusion of historical seasonal priors improve model generalization across glacier regions and years?*

To assess these questions, a modular regression pipeline is employed that predicts enriched seasonal glacier facies from Sentinel-1 inputs using XGBoost. The design includes label enrichment, base and residual-based modelling, and the integration of historical priors. Performance is evaluated across intra- and inter-annual settings, including spatial and temporal transfer scenarios, using standard regression metrics.

7.3.1 Materials

The modelling pipeline relies on two core data streams: multi-temporal EO imagery from Sentinel-1, and temporally dense glacier facies labels derived from TSX. Both are structured along a shared seasonal calendar and temporally aligned at 7-day resolution. This section outlines the origin, characteristics, and preprocessing of each data source.

Sentinel-1 Multi-SAR EO Input: The primary EO input for this study is derived from Sentinel-1 imagery acquired in Interferometric Wide (IW) swath mode over Axel Heiberg Island (AOI 1) and Devon Island (AOI 2). Over these high Arctic glacier systems, Sentinel-1 scenes are available in dual-polarization mode (HH and HV) at a nominal revisit interval of 7 days. The analysis period spans from April to November, representing the glaciologically active season in the study areas. In total, 16 acquisition dates are retained per year, yielding a dense time series of dual-pol SAR data.

Each scene is processed into co-registered Single Look Complex (SLC) format and passed through a Multi-SAR pre-processing pipeline following [38]. From these SLC inputs, polarimetric decomposition and Kennaugh matrix elements (K_0, K_1, K_5 and K_8) are computed to obtain four primary parameters per date (K_0 is exhibited in Figure 7.18

- left panel). This results in a time series of $16 \times 4 = 64$ EO bands per year, with a spatial grid at 10 m resolution. To compress this temporal information while retaining key seasonal dynamics, a hypercomplex temporal fusion is applied using the HCB method [289]. This technique, described in detail in Section 2.2, employs a transformation to jointly encode the temporal and structural dimensions of the SAR signal. The result is a single, 64-channel EO data cube per glaciological study year, where each band captures either a temporally fused mean (e.g., $K_{0,0}$) or a structured component of intra-annual variation (e.g., $K_{0,1-63}$), K_0 of the time-series is showcased in Figure 7.18 - right panel).

The fused EO input stack thus offers a rich yet compact summary of radar signal evolution over the melt year, structured for seasonal alignment and ready for use in supervised prediction tasks.

Glacier Facies Labels from TerraSAR-X: Glacier zone annotations are derived from high-resolution TSX imagery, processed into facies maps, as detailed in Section 1.2.3, with a spatial grid at 40 m resolution [311]. Thereby pixel-level glacier facies labels at 7-day intervals were derived, covering the same April–November glaciological window as the Sentinel-1 data. Each label map classifies the glacier surface into five categories: dry snow, percolation, superimposed ice, ice-free terrain, and wet snow.

For this study, only glacier-covered areas within the AOIs are retained. All label maps are temporally aligned with the EO data using nearest-neighbour matching: for each EO observation, the closest available TSX-derived label map is selected based on acquisition date. This produces a quasi-synchronous EO–label dataset suitable for season-wise aggregation and enrichment.

Temporal Coverage and Spatial Resolution: All data, EO inputs and labels, are aligned to a shared 7-day temporal grid spanning April to November. Each Sentinel-1 acquisition is temporally matched to it's nearest TSX-derived label counterpart. While Sentinel-1 imagery is originally available at 10 m resolution and TSX-derived glacier facies maps at 40 m, all datasets are resampled to a standardized spatial resolution of 40 m, using nearest-neighbour interpolation. This ensures pixel-wise correspondence across EO and label sources while preserving the native detail of the supervision signal. Seasonal groupings (spring, summer, fall, winter) are defined based on fixed calendar intervals, enabling structured fusion and aggregation over time.

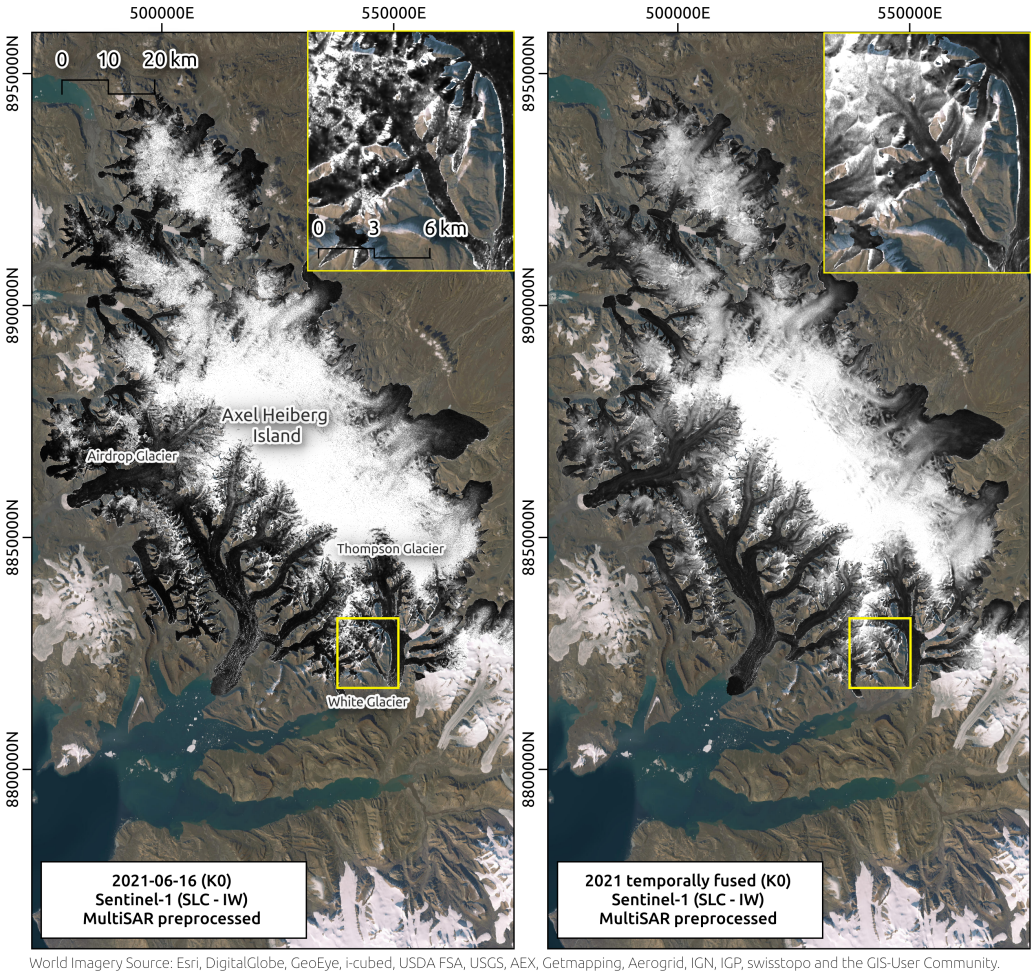


Figure 7.18.: Sentinel-1 SLC IW data over Axel Heiberg Island. **Left:** Single-date SAR image showing Kennaugh element K_0 derived from a Sentinel-1 SLC IW scene acquired on 2021-06-21. **Right:** Temporally fused dataset, generated from multiple Sentinel-1 SLC IW acquisitions within the defined seasonality periods, showing K_0.

This setup provides the basis for all subsequent modelling experiments, enabling a structured and seasonally-aware analysis of glacier surface dynamics across years and locations. The following sections detail the methods used for label enrichment, EO feature construction, supervised modelling, and generalization testing.

7.3.2 Methods

Building upon the temporally aligned and spatially co-registered EO and label data described above, the modelling framework is designed to predict seasonal glacier facies from Sentinel-1 observations. The approach combines temporal label enrichment, compact yet expressive EO input features, and a two-stage regression pipeline centred around XGBoost. Emphasis is placed on model interpretability, seasonal alignment, and transferability across both time and space. This section describes each methodological component in detail, beginning with the transformation of raw classification labels into temporally smoothed targets, followed by the construction of EO inputs, supervised learning strategies, and the use of historical priors and ensemble models to support generalization.

Spatio-Temporal Label Context Enrichment

In the context of glacier zone modelling from EO time-series, the reliability and expressiveness of training labels play a central role in determining the performance of temporally-aware predictive models. However, while these labels are temporally dense, derived at a 7-day cadence, they remain semantically static and discrete, limiting their utility for learning dynamic glacier behaviour. Also, while the raw label maps available for this study provide a categorical, per-date classification of glacier facies, including dry snow zone, percolation zone, superimposed ice zone, ice-free zone, and wet snow zone, these static labels lack the temporal expressiveness necessary for models to capture glacier dynamics in a physically meaningful way.

To address this limitation, a label-side enrichment strategy is introduced, inspired by the HELIX framework. The goal is to inject temporal semantics and seasonal regularities into otherwise static, per-date classification labels. This strategy injects seasonal semantics into the label space, transforming isolated classifications into continuous-valued representations that reflect intra-annual glacier dynamics. In particular, for each year, labels

are grouped by meteorological season (*spring*, *summer*, *fall*, *winter*). Within each season, the sequence of categorical glacier labels is aggregated into a single float-valued seasonal mean map, defined over the glacier area. This operation produces a temporally smoothed representation of the glacier facies evolution during each season, implicitly capturing intra-seasonal variability, transitions, and persistence.

This seasonal division was not arbitrary, but grounded in well-documented glaciological and climatological observations from the Canadian High Arctic. Research from long-term glacier monitoring (e.g., White Glacier, Axel Heiberg Island) and regional climate analyses support a breakdown into four primary glaciological seasons: spring (April–May), summer (June–August), fall (September), and winter (October–November). These periods correspond to major shifts in temperature, accumulation, and melt dynamics:

- **Spring (April–May):** This period marks the transition from the cold, dark polar winter into the onset of the melt season. April remains largely sub-freezing and snow-covered, while May sees increased solar input and rising temperatures. Sustained melt typically begins only in late May, which also marks the close of the winter accumulation period [344, 219, 73].
- **Summer (June–August):** The dominant ablation season in the High Arctic, summer is characterized by above-freezing temperatures, continuous daylight, and widespread surface melt. Glacier ice becomes exposed in lower zones, and precipitation, though limited, mostly occurs as rain or wet snow [344, 73]. Mass loss is concentrated in this period.
- **Fall (September):** A rapid transition phase, September brings a return to sub-zero temperatures and the definitive end of surface melt. This month often marks the annual minimum in glacier mass, as melt ceases and new snow begins accumulating. The glacier mass-balance year is typically defined as ending in September [317, 95, 73].
- **Winter (October–November):** With the onset of polar night and deep sub-freezing conditions, winter marks the start of the long accumulation period. While mid-winter snowfall is low, October and November contribute significantly to total seasonal accumulation. By late November, a persistent snowpack has usually formed [73, 95].

This structured seasonal grouping ensures that label fusion captures real glaciological transitions, while aligning with observational data availability. It avoids arbitrary calendar binning and enables models to learn from temporally consistent glacier behaviour.

Specifically, for each year y, all label dates from April through November, the effective glaciological activity window for Arctic glacier systems such as Axel´Heiberg Island, Devon Island and Mason Island, are grouped into the four meteorological seasons: *spring* (April–May), *summer* (June–August), *fall* (September), and *winter* (October–November). This seasonal partitioning is loosely balanced across the annual cycle and reflects the distribution of available EO observations (from Sentinel-1) and classification labels. While simplified, the scheme provides meaningful temporal resolution over all active glaciological phases, avoiding unnecessary fragmentation while still capturing key seasonal transitions.

The enrichment process begins by temporally aligning each EO observation with the closest available classified label map. To ensure that each EO observation is accompanied by a temporally relevant supervision signal, a nearest-neighbour temporal alignment is applied. Specifically, for each EO date t_{EO}, the label map from the closest available date t_{label} is selected, such that $t_{\text{label}} = \arg\min_{t \in T_{\text{labels}}} |t - t_{\text{EO}}|$. This preserves the temporal integrity of the label stack while avoiding interpolation artifacts, and results in a quasi-aligned EO–label pairing that is sufficient for downstream fusion and seasonal aggregation. This alignment ensures that every EO frame is accompanied by a temporally proximal and semantically meaningful label snapshot, without requiring interpolation or synthetic transformation of class information.

Once temporally aligned, the enrichment process maps the resulting sequence of label maps $\{L_{t_{\text{label}}}\}$ for $t_{\text{label}} \in T_y$ (where T_y is the set of label-aligned dates within year y) into a reduced set of four seasonal representations $\{\bar{L}_s^y\}_{s \in \{\text{spring, summer, fall, winter}\}}$, with an additional annual mean $\bar{L}_{\text{mean}}^y$ computed across seasons. Each $\bar{L}_s^y$ is obtained via a pixel-wise average over all valid glacier-class pixels within the corresponding seasonal subset.

As a result, the enriched labels no longer represent a single categorical class but instead encode class likelihood or tendency as continuous values (see Figure 7.19). This enables supervised models to learn from soft, gradient-like transitions between glacier facies, reflecting persistence and temporal dynamics, rather than being constrained to hard, static classifications.

This label-side transformation enables models to perceive not just what glacier zone is present at a specific time, but also how that zone typically behaves throughout the seasonal cycle (see Figure 7.20). For example, regions persistently classified as dry snow throughout the spring and summer will yield higher seasonal means than those fluctuating between wet snow and superimposed ice. In this sense, the enriched label maps serve as temporally contextualized supervision signals, embedding both intra-annual structure and inter-seasonal behaviour. This approach can thus be interpreted as a temporally hierarchical form of label augmentation: it enhances spatially discrete and temporally static zone labels with dynamic structure that reflects glacier evolution across time, without requiring any changes to the original classifier or the downstream prediction architecture. This is especially critical in glacier facies modelling, where transitions between zones are gradual, physically driven, and modulated by complex surface processes, and where supervised learning must account for both class presence and class persistence. Figures 7.21 and 7.22 illustrate this concept in practice on the Devon Ice Cap. Each panel depicts seasonally enriched glacier facies derived from TSX backscatter data across multiple years (2017–2023), with RGB channels encoding spring (red), summer (green), and fall (blue) class variations. This encoding visually captures both seasonal dynamics and year-to-year variability. In particular, the Cunningham Glaciers, highlighted in the zoomed-in insets, show marked changes in facies distribution and persistence across years, underscoring the importance of modelling temporal context in facies classification.

The HELIX framework introduces a label-centric enrichment strategy that transforms static glacier zone labels into temporally smoothed, seasonally aligned, and continuous-valued targets (see Figures 7.19 and 7.20). This reformulation allows the use of interpretable regression models such as XGBoost to predict dynamic glacier behaviour without the need for complex temporal architectures. Combined with fused EO inputs that reduce variability, this approach forms an efficient and explainable pipeline for multi-seasonal glacier zone inference across space and time.

Supervised EO-Based Glacier Zone Prediction

The enriched glacier zone labels described above serve as the training targets for a supervised regression pipeline aimed at predicting seasonal glacier facies behaviour directly from EO time series. The task is formulated as a multivariate regression problem, where the objective is to learn a mapping from a temporally rich EO input X to a continuous

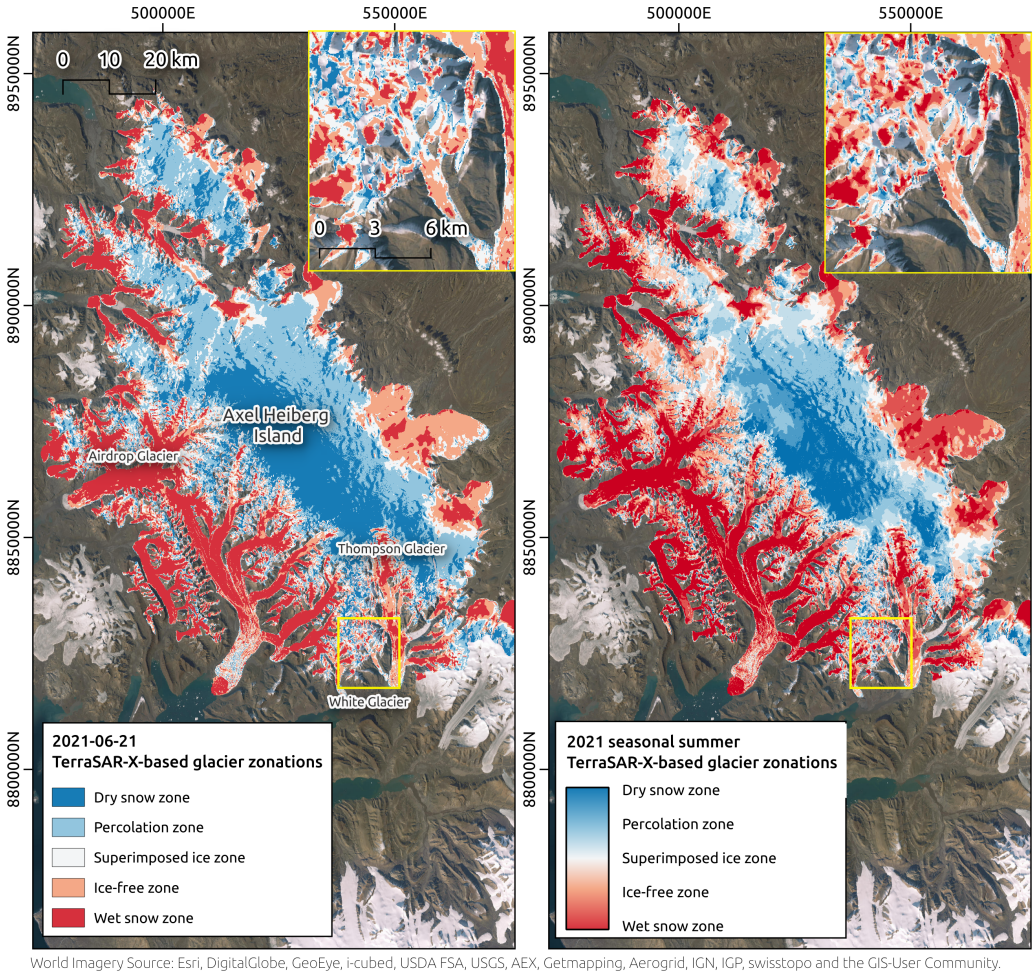


Figure 7.19.: Label enrichment via HELIX. **Left:** Static glacier facies classification from TSX for a single date (2021-06-21), representing categorical zone labels: Dry Snow, Percolation, Superimposed Ice, Ice-Free, and Wet Snow. **Right:** HELIX-based seasonal enrichment for the 2021 summer season, shown as continuous-valued regression targets capturing intra-seasonal glacier zone tendencies.

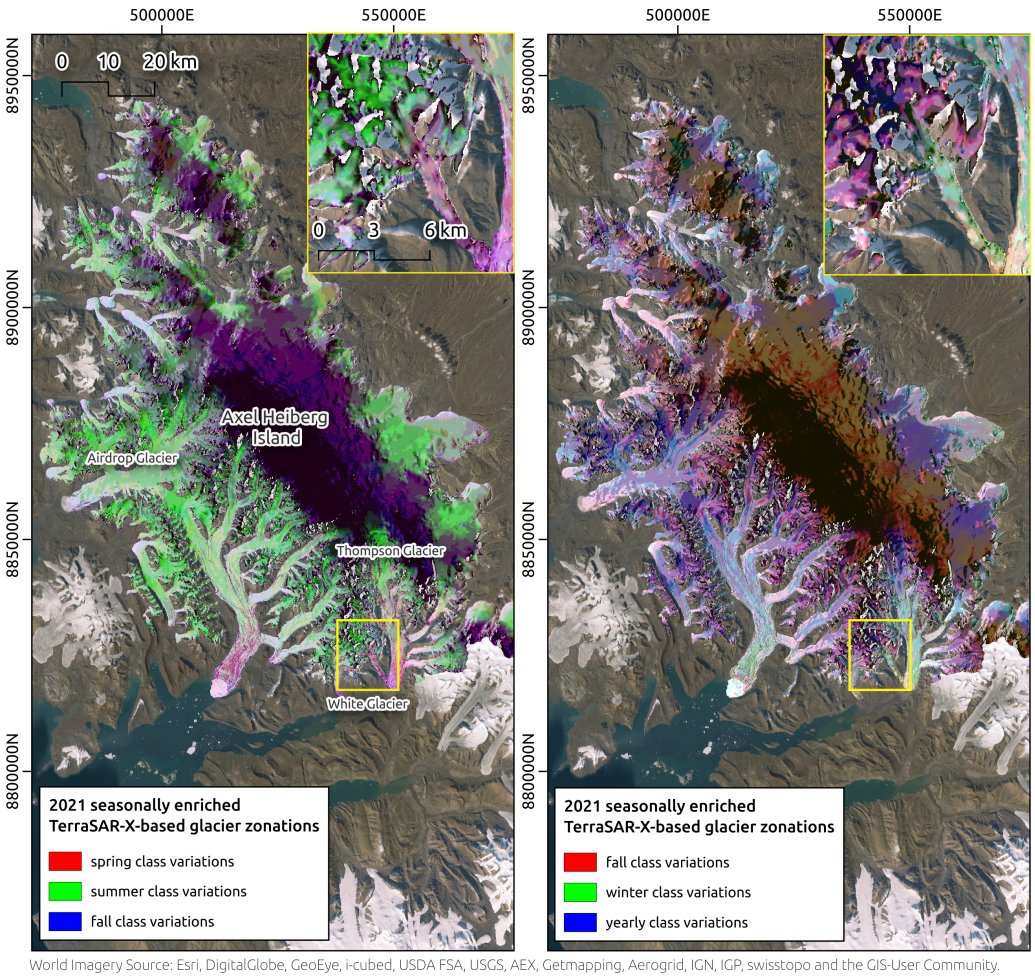


Figure 7.20.: Seasonal label enrichment visualization. **Left:** RGB composite of HELIX-enriched seasonal facies for 2021, where Red = Spring, Green = Summer, Blue = Fall. **Right:** Alternate RGB encoding showing Red = Fall, Green = Winter, Blue = Annual Mean. The continuous colouring reveals spatial persistence and transition dynamics across glacier zones.

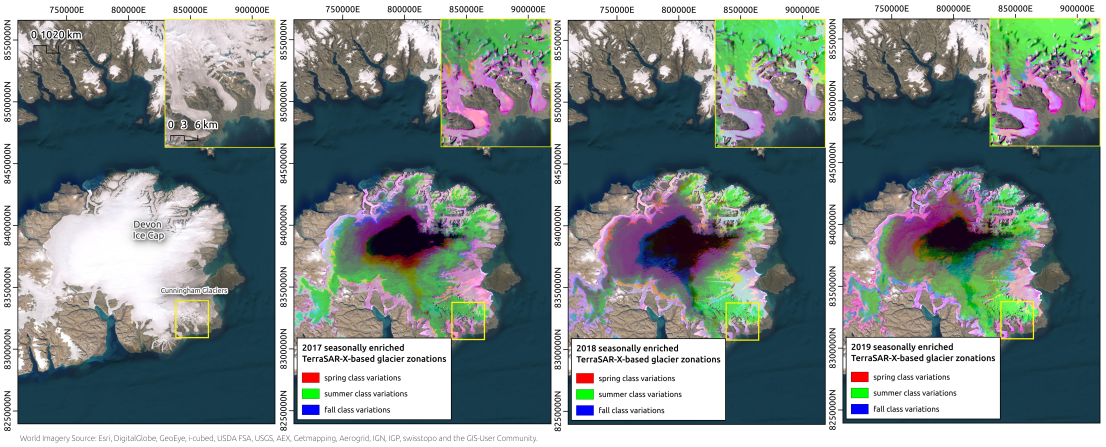


Figure 7.21.: HELIX-enriched seasonal facies over the Devon Ice Cap. The four panels show: (1) a basemap view of the Devon Ice Cap from World Imagery; (2) HELIX-enriched seasonal glacier facies for 2017; (3) for 2018; and (4) for 2019. In panels 2–4, colour channels represent meteorological seasons: red = spring, green = summer, and blue = fall.

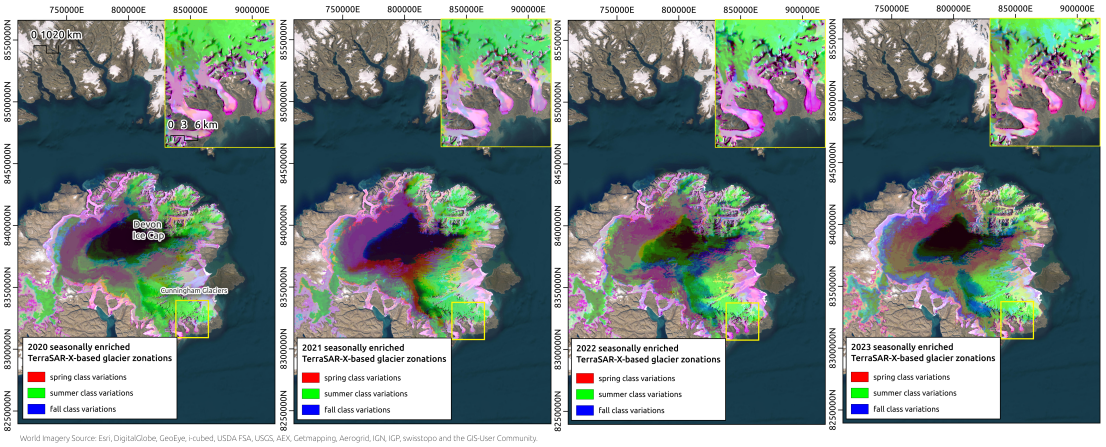


Figure 7.22.: HELIX-enriched seasonal glacier facies on the Devon Ice Cap from 2020 to 2023. Each panel visualizes one year's seasonal pattern using RGB colour channels: red = spring, green = summer, and blue = fall. Panels from left to right correspond to the years 2020, 2021, 2022, and 2023. The imagery highlights inter-annual variability in glacier zone distribution.

five-dimensional label vector $\bar{L} = [\bar{L}_{\text{spring}}, \bar{L}_{\text{summer}}, \bar{L}_{\text{fall}}, \bar{L}_{\text{winter}}, \bar{L}_{\text{mean}}]$, representing the enriched class distribution across the four seasons and the annual mean.

To maintain interpretability and reduce modelling complexity, the prediction pipeline is initially trained using only a single EO input band, K0 (Band 0), which corresponds to the total backscatter intensity of the fused SAR time-series.

This fused EO stack, described earlier in the materials section, aggregates multi-temporal SAR observations over the glaciological time span of the corresponding labels. Among the 64 fused EO bands, K0 was empirically found to exhibit the strongest predictive signal for glacier zone behaviour, capturing integrated surface response while minimizing redundancy. Band 0 (K0), representing the total intensity across all time steps, is consistently found to exhibit the strongest correlation with glacier facies behaviour. Owing to its compactness, physical interpretability, and low susceptibility to speckle noise, K0 is selected as the sole input for most of the regression models in this study. This decision enables model simplification and facilitates transferability, while still capturing the essential backscatter signature of glacier surface conditions across seasons.

Initial experiments with the full set of EO features confirmed that including all bands led to only marginal performance gains while significantly increasing model complexity and overfitting risk. Therefore, Band 0 is selected as a minimal yet expressive representation of glacier-relevant EO dynamics. This simplification allows the model to focus on the core EO-to-label mapping while remaining efficient and easily transferable across glacier sites and years.

The predictive modelling process is composed of two sequential stages:

1. **Base Model Training:** An XGBoost-based multi-output regressor is trained to predict the 5-band enriched label vector from EO Band 0 values. The training is restricted to glacier-covered pixels, and supervised using the enriched seasonal label stack corresponding to the same year.

2. **Residual-Hint Refinement:** The residuals from the base model predictions are used to construct a secondary input feature, referred to as the residual hint. This feature captures systematic prediction errors in the base model and is appended to the original input. A second-stage regressor is then trained on this augmented feature set to refine the seasonal predictions. This process introduces a form of correction-based learning, akin to residual boosting or model distillation, while maintaining model transparency.

To further support the learning of seasonal structure in glacier facies behaviour, particularly in the presence of ambiguous EO signals or inter-annual variability, three complementary strategies were considered:

- **(1) Static Historical Seasonal Profile:** A five-dimensional historical label vector is computed from previous years (2017–2020), representing the per-pixel average of seasonal zone labels. This static context vector, denoted $\bar{L}_{\text{hist}}$, encodes persistent glacier behaviour and is appended to the input as an auxiliary feature. It serves as a spatialized prior, guiding the model in regions with ambiguous EO signatures or weak seasonality.
- **(2) Pre-training on Past Years:** An optional strategy would involve pre-training the base model on EO and label data from earlier years (e.g., 2017–2020) and fine-tuning it on the target year. While potentially powerful, this requires EO time series from past years, which are not consistently available in the present setup.
- **(3) Residual-Structure Encoding:** A more experimental extension would involve computing and summarizing past model residuals across multiple years to learn spatial or class-specific prediction bias. This residual structure could be introduced as an auxiliary feature to guide the refinement stage. Although promising in theory, this strategy remains untested in the current setup.

Among these, only Option 1 is implemented in the present system due to its simplicity, interpretability, and compatibility with existing label resources. It offers a lightweight way to incorporate temporal memory without requiring historical EO observations.

The model is evaluated using standard regression metrics (MAE, R^2) per seasonal dimension, and tested across both spatial and temporal domains. Specifically, the model trained on a given Area of Interest (AOI) and year is used to predict:

- **Spatial Transfer:** Generalization to other glacier islands (e.g., from Axel Heiberg to Devon Island),
- **Temporal Transfer:** Generalization to future years (e.g., training on 2021, testing on 2022 or 2023),
- **Combined Transfer:** Applying models trained on one AOI–year pair to entirely unseen AOI–year pairs (e.g., predicting 2024 on Mason Island using a model trained on 2021 Axel Heiberg Island).

This approach is designed for modular extension: multiple base+residual (and hist) models trained on different AOIs and years can be stacked into a model ensemble. For example:

- **Model A:** AOI 1 (Axel Heiberg), EO 2021 $\rightarrow$ Label 2021
- **Model B:** AOI 2 (Devon Island), EO 2021 $\rightarrow$ Label 2021
- **Model C:** AOI 1, EO 2022 $\rightarrow$ Label 2022
- **Model D:** AOI 2, EO 2022 $\rightarrow$ Label 2022

These models are fused via ensemble averaging strategies to improve predictive robustness across both spatial and temporal dimensions.

The full prediction system, including data inputs, enriched label construction, the residual modelling strategy, and the integration of historical priors, is summarized in Figure 7.23. It shows how temporally fused EO inputs and HELIX-processed glacier labels are passed through two complementary model branches, both designed for generalizable glacier zone inference across space and time.

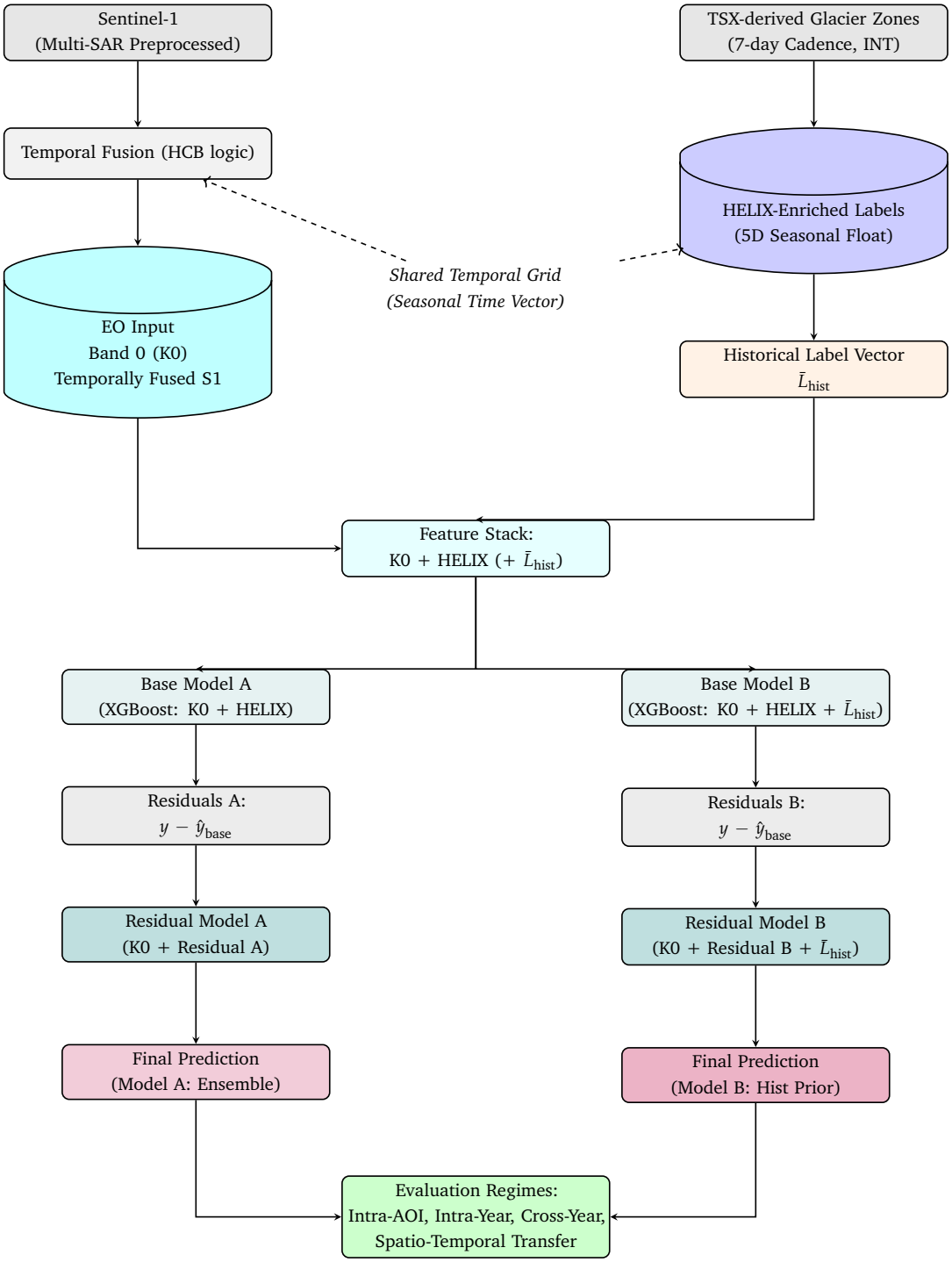


Figure 7.23.: HELIX-based supervised regression architecture for glacier facies prediction. Temporally fused Sentinel-1 inputs (Band 0) are combined with HELIX-enriched labels and optional historical label priors, feeding into two parallel modelling pathways with residual refinement. Models are evaluated across four generalization settings.

7.3.3 Results

This section presents a stepwise evaluation of the proposed glacier facies prediction framework. The analysis begins with a baseline regression model trained on a single EO input band, followed by a residual-based ensemble refinement that improves prediction quality across seasons. Finally, a third configuration introduces a static historical label prior to enable temporally and spatially robust inference without access to contemporary EO time series.

Baseline Modelling

As a first step, the full fused EO stack, comprising 64 bands derived from multi-temporal Sentinel-1 observations, was evaluated for glacier zone prediction. Empirical testing (Figure 7.24) across multiple model runs revealed that Band 0 (K0), which represents the total SAR backscatter intensity, consistently outperformed other bands and combinations thereof. Given its physical interpretability, low noise, and strong glacier-facies sensitivity, K0 was selected as the sole input feature for all subsequent experiments. This decision reduced model complexity and overfitting risk, while retaining sufficient predictive expressiveness.

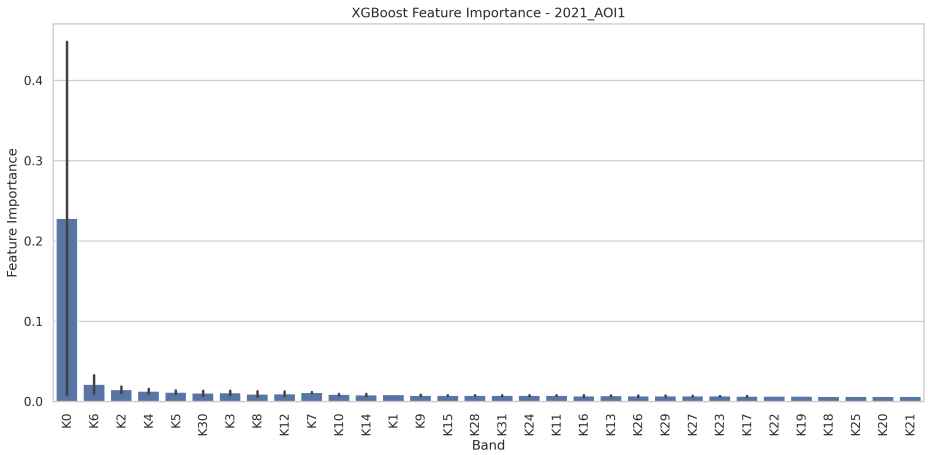


Figure 7.24.: Relative importance of individual EO bands for glacier zone prediction using the full temporally-fused Sentinel-1 input stack. Band 0 (K0), representing total SAR backscatter intensity, demonstrated the highest predictive power across multiple model runs, justifying its selection as the sole input feature for subsequent modelling.

Initial experiments focused on intra-AOI learning, using EO and label data from AOI 1 (Axel Heiberg Island) for the year 2021. The task was to predict seasonal glacier zone distributions from the fused EO input, using the enriched seasonal labels as multivariate regression targets.

Baseline performance using only EO Band 0 revealed substantial seasonal variation in prediction accuracy. The summer season exhibited the weakest performance, with a mean absolute error (MAE) of 0.645 and an R^2 of 0.600. In contrast, winter yielded the best results (MAE = 0.411, $R^2 = 0.624$). Spring and fall showed intermediate performance, with MAEs of 0.448 ($R^2 = 0.630$) and 0.469 ($R^2 = 0.491$), respectively. The overall Year Mean MAE was 0.417 with an R^2 score of 0.659.

Residual-Hint Ensemble Modelling

Incorporating the residual-hint mechanism, where a second-stage model is trained on the residuals of the baseline predictions, led to substantial improvements across all seasons. Most notably, summer MAE dropped from 0.645 to 0.433, a relative reduction of over 32.9%, with corresponding gains in R^2 (from 0.600 to 0.782). Likewise, spring improved to MAE = 0.305 and $R^2 = 0.766$, fall to MAE = 0.317 and $R^2 = 0.676$, and winter to MAE = 0.284 with an R^2 of 0.762. The enhanced model achieved a Year Mean MAE of 0.286 and R^2 of 0.798, confirming that residual-based correction significantly improves prediction fidelity even when using a single EO input band. Overall, the residual-hint approach consistently reduced seasonal errors by over 30% relative to the baseline.

Ensemble Modelling: To further improve robustness and enable prediction on unseen future years, an ensemble was formed by averaging predictions from four independently trained models:

- AOI 1, 2021
- AOI 1, 2022
- AOI 2, 2021
- AOI 2, 2022

Per Model metrics are summarised in in Table 7.12.

To assess the generalization performance of model ensembles, two configurations were evaluated. The first ensemble, combining all four models trained on AOI 1 and AOI 2 for the years 2021 and 2022, was used to predict glacier facies for AOI 1 in 2023, a fully unseen temporal context. This ensemble achieved a Year Mean MAE of 0.433 and $R^2 = 0.517$. While moderately less accurate than the best individual model (AOI 2, 2022), the ensemble retained robust predictive quality across all seasons, confirming its utility for unsupervised temporal forecasting.

The second ensemble, constructed from AOI 2 models only (Devon Island, 2021 and 2022), was tested on AOI 1 in 2023. This configuration served as a spatio-temporal transfer benchmark. Although slightly worse in overall accuracy (Year Mean MAE = 0.444), it still demonstrated viable generalization without any exposure to the target region during training.

Both ensembles are summarized in Table 7.12, which also reports MAE per season across all evaluated models.

Table 7.12.: Mean Absolute Error (MAE) per season across all models and ensembles. Intra-AOI and cross-AOI scenarios are shown separately, with Year Mean summarizing average performance.

Model	Fall	Spring	Summer	Winter	Year Mean
Intra AOI: AOI1_2021	0.470	0.447	0.645	0.411	0.417
Intra AOI: AOI1_2022	0.460	0.447	0.435	0.448	0.412
Intra AOI: AOI2_2021	0.327	0.283	0.446	0.268	0.228
Intra AOI: AOI2_2022	0.265	0.322	0.267	0.280	0.212
Cross AOI: Ensemble_Full (AOI1+2, 2021+2022)	0.510	0.507	0.491	0.516	0.433
Cross AOI: Ensemble_DevonOnly (AOI2, 2021+2022)	0.518	0.524	0.512	0.528	0.444

While the ensemble did not surpass the best individual model in absolute terms, it succeeded in generalizing to an unseen glaciological year without any retraining or new labels. Most importantly, all seasonal MAEs remained below 0.52, well within the expected range for intra-class variability. Considering that the target labels are continuous values over a 1–5 scale, these results reflect sub-class-level precision and meaningful

seasonal gradient learning. This reinforces the viability of the HELIX-enriched regression approach, even when limited to a single EO band and modest model complexity.

Historical Context-Based Modelling

To assess the benefit of spatially informed priors, the **Static Historical Seasonal Profile** ($\bar{L}_{\text{hist}}$) described in Section 7.3.2 was integrated as an auxiliary feature in the modelling pipeline. This five-dimensional vector represents the per-pixel seasonal average of glacier zone labels over the 2017–2021 period and was appended to the EO Band 0 input.

A model trained on AOI 2 in 2022 with this extended feature set demonstrated remarkable improvements in both prediction accuracy and consistency:

Table 7.13.: Model Performance with Historical Label Vector ($\bar{L}_{\text{hist}}$) on AOI 2

Season	**MAE**	R^2
Spring	0.091	0.950
Summer	0.136	0.868
Fall	0.104	0.931
Winter	0.083	0.962
Year Mean	0.076	0.968

Compared to models trained without historical priors, the inclusion of $\bar{L}_{\text{hist}}$ consistently reduced MAE across all seasons, yielding sub-class resolution errors below 0.14, and raised R^2 above 0.86 even in summer. These results support the utility of persistent glaciological context for improving prediction stability, especially in regions or periods where EO signals alone may be ambiguous or seasonally weak.

To explore lightweight temporal memory mechanisms that do not rely on historical EO availability, the utility of static label-derived priors was then further evaluated in a spatio-temporal transfer way. The model was therefore trained on AOI 2, using EO data from 2022 and the static historical context vector, and then evaluated on AOI 1 in 2023. This configuration constitutes a spatial and temporal transfer setting, with no direct access to either 2023 labels or EO history at inference time.

The resulting performance, shown in Table 7.14, confirms the efficacy of this strategy. All seasonal MAE values remained below 0.5, with a Year Mean MAE of 0.399. While R^2

performance in summer remained low (0.108), other seasons exhibited solid generalization with $R^2 > 0.59$, including a Year Mean R^2 of 0.640. These results underscore the potential of label-centric prior information to stabilize prediction in temporally unseen settings, particularly when EO data alone may be insufficient to disambiguate complex seasonal dynamics.

Table 7.14.: Performance of Historical Label Prior Model (Trained on AOI 2, Evaluated on AOI 1, 2023)

Season	**MAE**	R^2
Spring	0.476	0.595
Summer	0.495	0.108
Fall	0.448	0.603
Winter	0.454	0.655
Year Mean	0.399	0.640

Compared to the fully stacked ensemble (Year Mean MAE = 0.433) and the AOI 2-only ensemble (Year Mean MAE = 0.444), the historical prior model achieved better absolute performance, despite being simpler and requiring fewer data sources. This highlights the value of temporal context encoded on the label side and opens pathways for generalization in data-sparse glacier systems.

In addition to quantitative error metrics, the seasonal composition of snow zone classes predicted for AOI 1 was examined. Figure 7.25 presents the reference and predicted class distributions on a logarithmic scale for each season. The model was found to capture the seasonal dynamics effectively, including the predominance of the Percolation and Ice-Free zones in summer and consistent representation of the Dry Snow and Superimposed Ice zones during spring and fall.

To assess the model's ability to generalize across space and time, a direct spatio-temporal transfer prediction scenario using historical context-based modelling is presented. Figures 7.26 and 7.30 visualize this transfer scenario. Each figure compares the HELIX-enriched seasonal reference in 2023 (left panel) to the corresponding model prediction also in 2023 (right panel), with both maps expressed on a shared continuous 1–5 scale representing glacier facies tendencies. These visualizations demonstrate the model's ability to reconstruct glacier zonation patterns under strong generalization constraints

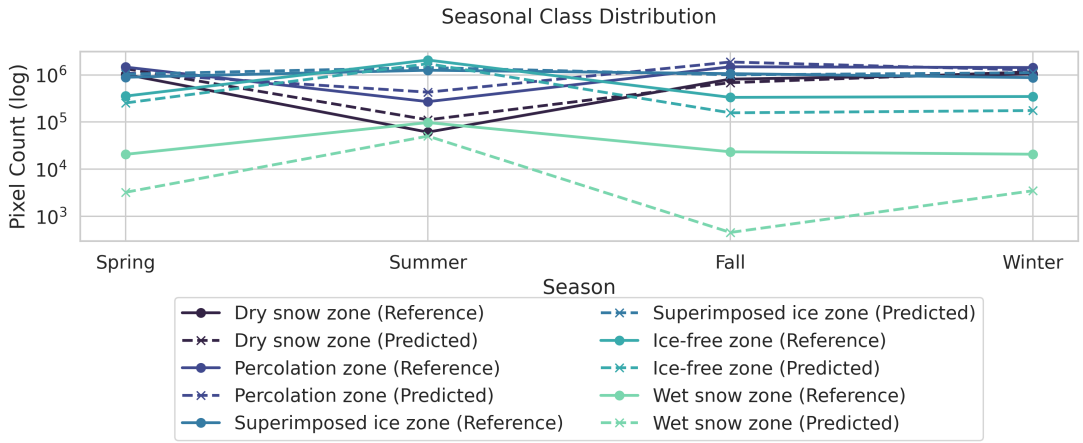


Figure 7.25.: Seasonal class distributions in AOI 1, plotted on a logarithmic pixel-count scale. Reference and predicted distributions are shown for each season, excluding the annual mean. The model correctly reflects dominant seasonal trends, with slight overestimation of the Ice-Free zone in summer and consistent detection of Dry Snow and Superimposed Ice zones in transitional seasons.

and highlight the effectiveness of label-side temporal priors in the absence of real-time input.

To further evaluate the spatial consistency of predictions, class profiles were extracted along three representative glacier transects located on Axel Heiberg Island: *White Glacier*, *Thompson Glacier*, and *Airdrop Glacier*. Each transect was defined using a narrow polygon of approximately 200 m in width, aligned with the principal glacier flow-line. The transects differed in length, with *White Glacier* spanning approximately 14.8 km, *Thompson Glacier* 38.0 km, and *Airdrop Glacier* 44.0 km.

Figures 7.31–7.33 show the seasonal progression of predicted and reference classes along each transect. The step-like structure of the plots reflects the discrete classification framework employed. Visual comparisons suggest a high degree of spatial alignment, notably in the detection of the Wet Snow zone near glacier termini and the Dry Snow zone in the upper accumulation regions.

7.3.4 Discussion

The findings presented in the previous section suggest that a Sentinel-1–based modelling approach, when combined with HELIX-enriched labels, can provide reliable predictions of seasonal glacier facies across both spatial and temporal domains. The following discussion examines these results in relation to the broader objectives of the study, with a focus on evaluating the methodological design, interpreting performance trends, and identifying practical implications. Consideration is given to the choice of minimal yet expressive EO input, the advantages of residual learning mechanisms, and the role of temporally enriched supervision in enabling transferability. Particular attention is directed toward the observed divergence between MAE and R^2 in low-variance regimes, especially during the summer melt period. In addition, the potential of historical label priors and the demonstrated generalization across Arctic regions are considered as key indicators of the framework's scalability and operational utility in data-sparse glaciological settings.

Use of Sentinel-1 SLC IW Mode in Glacier Zone Prediction

The use of Sentinel-1 Single Look Complex (SLC) data in Interferometric Wide (IW) swath mode forms a foundational element of the present glacier zone modelling framework. While Sentinel-1 is widely used in cryospheric remote sensing, the majority of

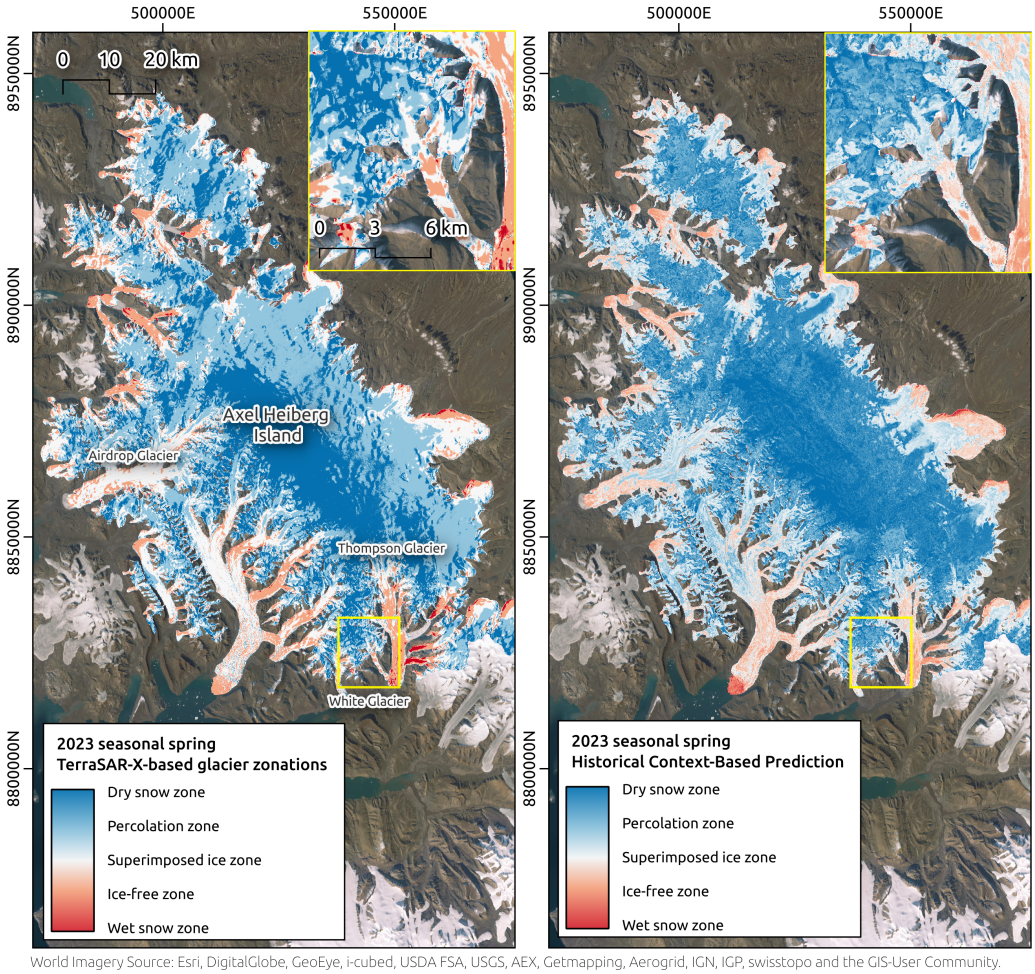


Figure 7.26.: Historical context-based seasonal prediction for spring. **Left:** HELIX-enriched reference map for the 2023 spring season, representing continuous glacier zone intensities across five facies (Dry Snow, Percolation, Superimposed Ice, Ice-Free, Wet Snow). **Right:** Model prediction for spring 2023 over AOI 1, generated using EO time-series data from 2023 for AOI 1, with a model trained solely on EO data from AOI 2 (2022) and the historical seasonal context vector ($\bar{L}_{\text{hist}}$), demonstrating a direct, unbiased spatio-temporal transfer. Both panels share the same continuous 1–5 scale, allowing direct comparison of predicted and reference facies tendencies.

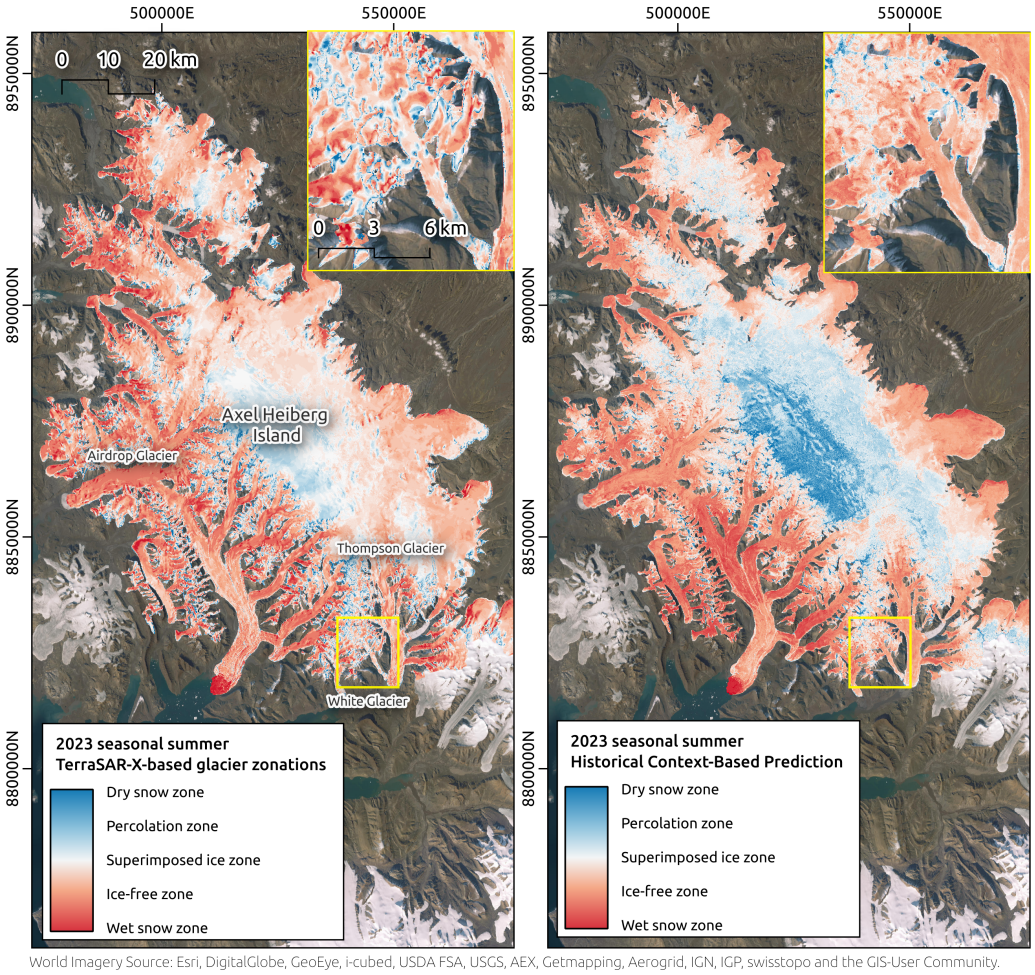


Figure 7.27.: Historical context-based seasonal prediction for summer. **Left:** HELIX-enriched reference map for the 2023 summer season, representing continuous glacier zone intensities across five facies (Dry Snow, Percolation, Superimposed Ice, Ice-Free, Wet Snow). **Right:** Model prediction for summer 2023 over AOI 1, generated using EO time-series data from 2023 for AOI 1, with a model trained solely on EO data from AOI 2 (2022) and the historical seasonal context vector ($\bar{L}_{\text{hist}}$), demonstrating a direct, unbiased spatio-temporal transfer. Both panels share the same continuous 1–5 scale, allowing direct comparison of predicted and reference facies tendencies.

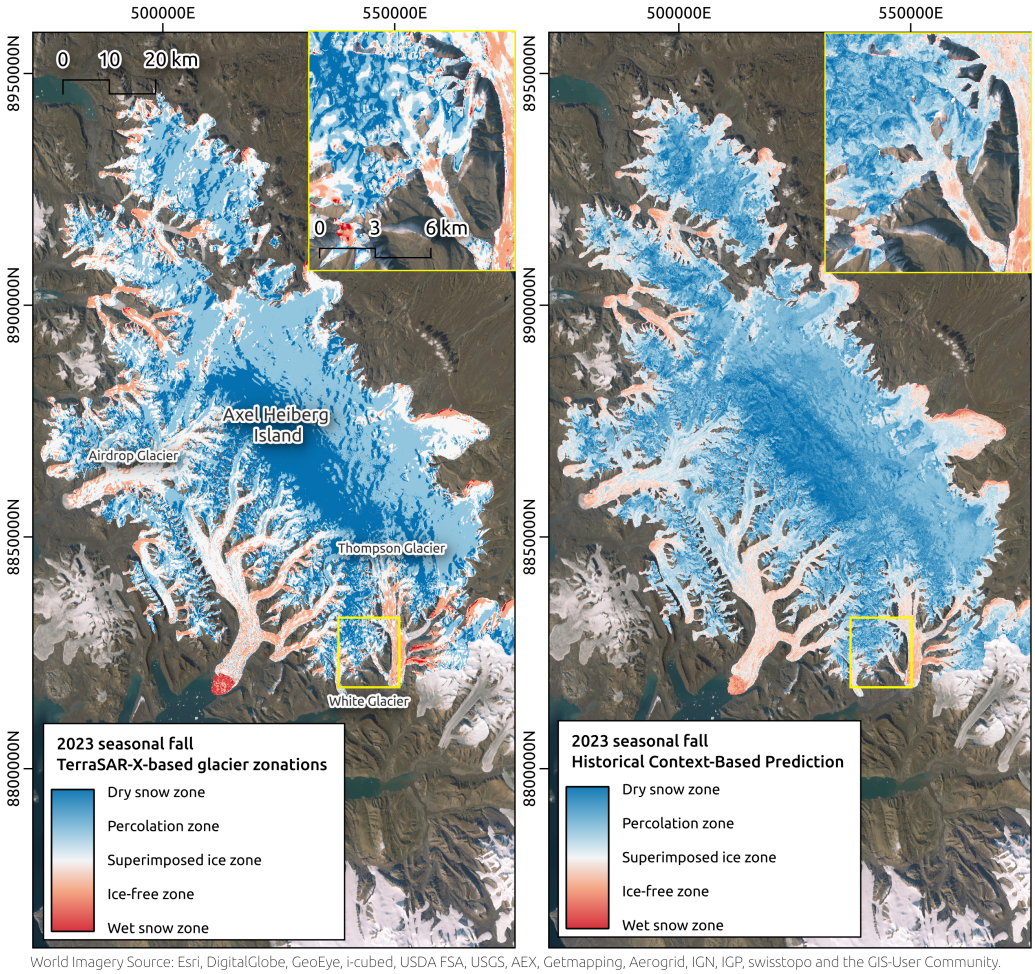


Figure 7.28.: Historical context-based seasonal prediction for fall. **Left:** HELIX-enriched reference map for the 2023 fall season, representing continuous glacier zone intensities across five facies (Dry Snow, Percolation, Superimposed Ice, Ice-Free, Wet Snow). **Right:** Model prediction for fall 2023 over AOI 1, generated using EO time-series data from 2023 for AOI 1, with a model trained solely on EO data from AOI 2 (2022) and the historical seasonal context vector ($\bar{L}_{\text{hist}}$), demonstrating a direct, unbiased spatio-temporal transfer. Both panels share the same continuous 1–5 scale, allowing direct comparison of predicted and reference facies tendencies.

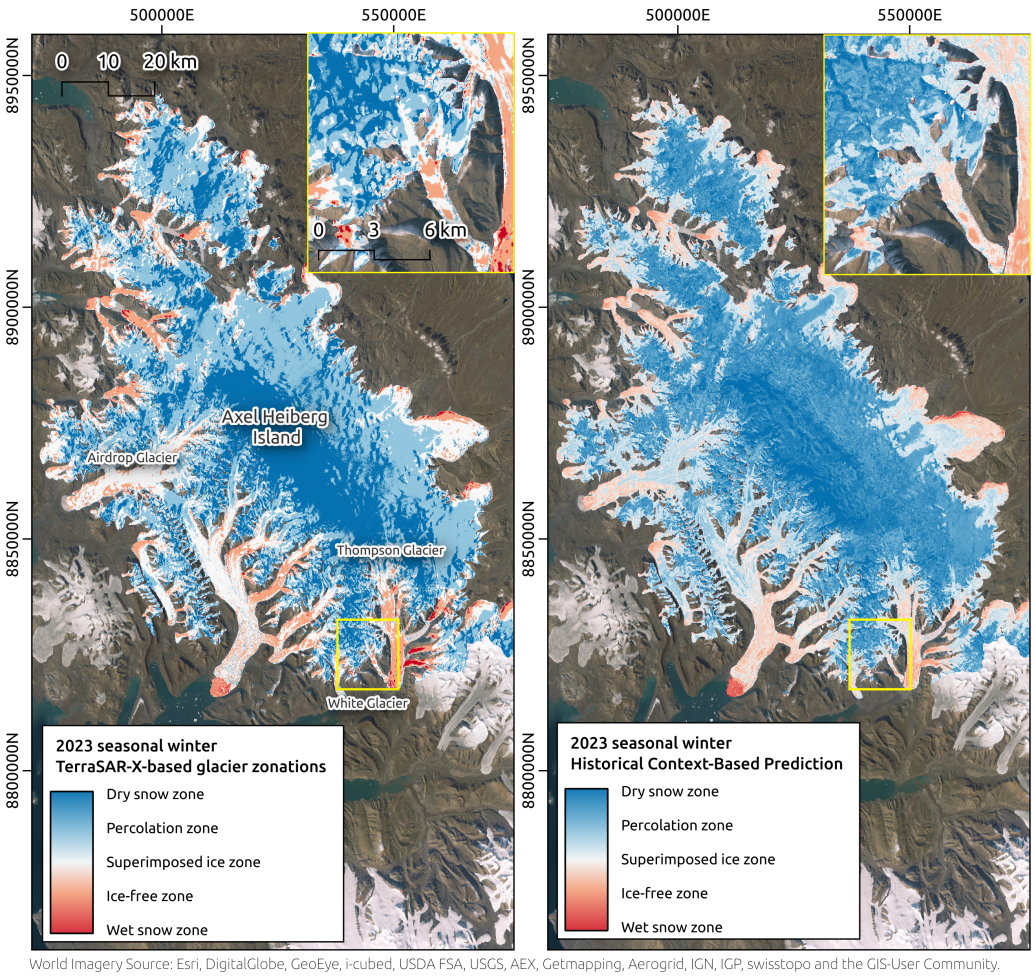


Figure 7.29.: Historical context-based seasonal prediction for winter. **Left:** HELIX-enriched reference map for the 2023 winter season, representing continuous glacier zone intensities across five facies (Dry Snow, Percolation, Superimposed Ice, Ice-Free, Wet Snow). **Right:** Model prediction for winter 2023 over AOI 1, generated using EO time-series data from 2023 for AOI 1, with a model trained solely on EO data from AOI 2 (2022) and the historical seasonal context vector ($\bar{L}_{\text{hist}}$), demonstrating a direct, unbiased spatio-temporal transfer. Both panels share the same continuous 1–5 scale, allowing direct comparison of predicted and reference facies tendencies.

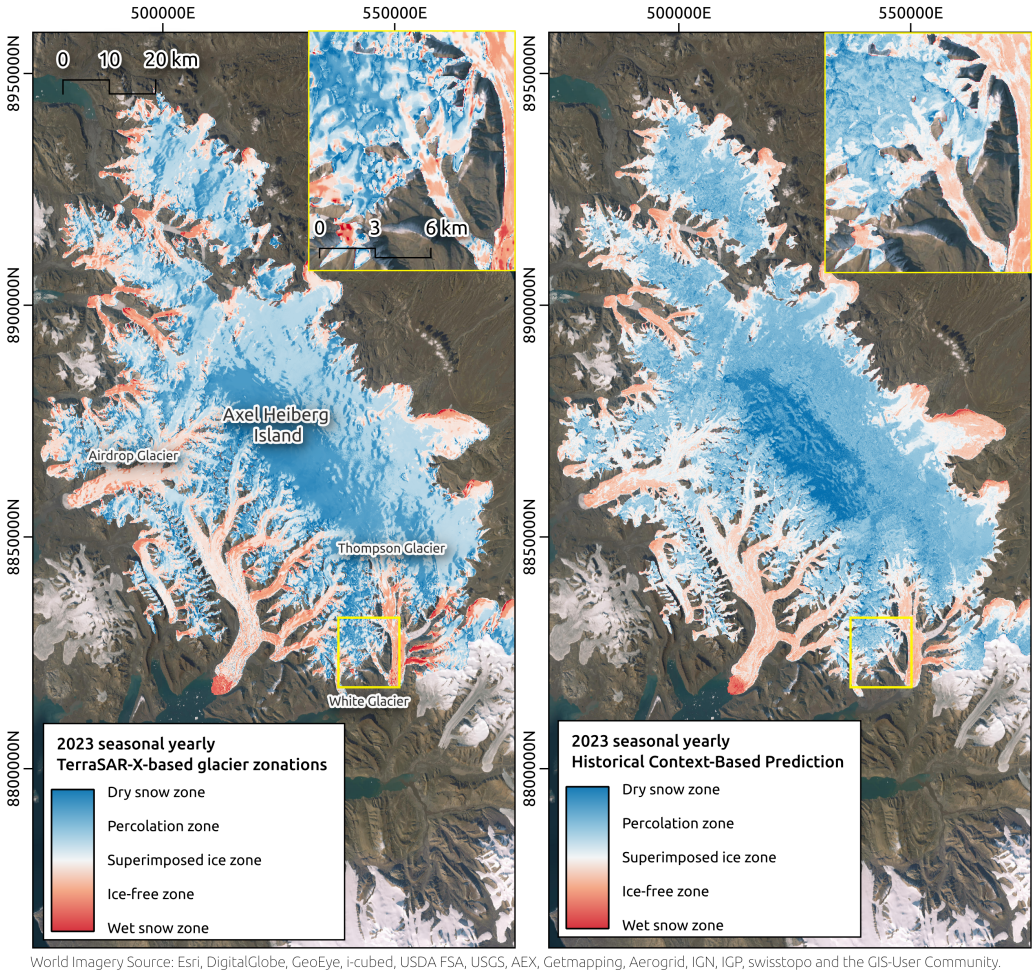


Figure 7.30.: Historical context-based seasonal prediction for the average glaciological year. **Left:** HELIX-enriched reference map for the 2023 whole season, representing continuous glacier zone intensities across five facies (Dry Snow, Percolation, Superimposed Ice, Ice-Free, Wet Snow). **Right:** Model prediction for the whole glaciological year 2023 over AOI 1, generated using EO time-series data from 2023 for AOI 1, with a model trained solely on EO data from AOI 2 (2022) and the historical seasonal context vector ($\bar{L}_{\text{hist}}$), demonstrating a direct, unbiased spatio-temporal transfer. Both panels share the same continuous 1–5 scale, allowing direct comparison of predicted and reference facies tendencies.

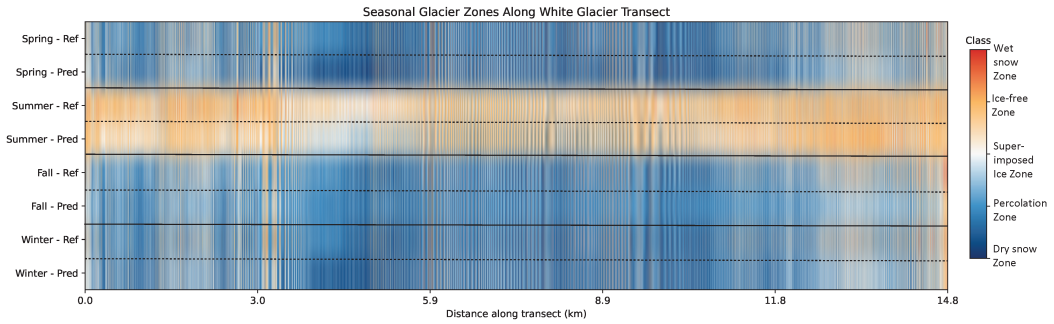


Figure 7.31.: Reference and predicted zonation classes along the *White Glacier* transect (14.8 km in length, 200 m in width), for all seasons.

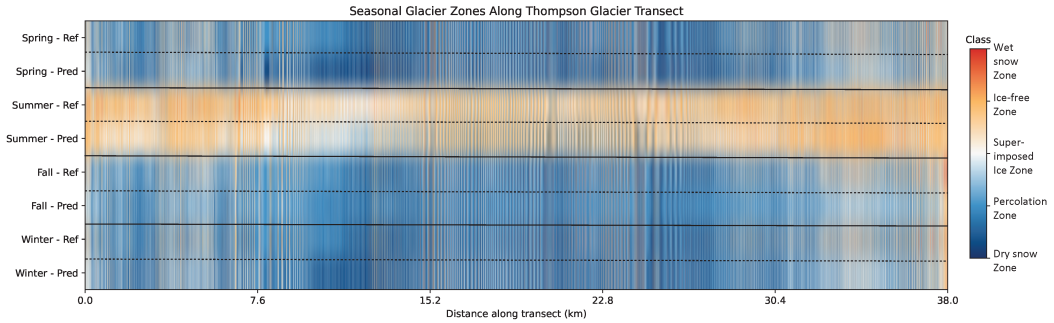


Figure 7.32.: Reference and predicted zonation classes along the *Thompson Glacier* transect (38.0 km in length, 200 m in width), for all seasons.

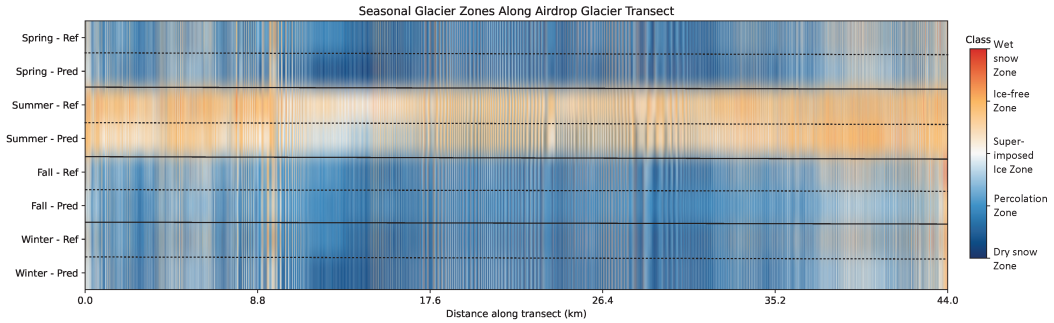


Figure 7.33.: Reference and predicted zonation classes along the *Airdrop Glacier* transect (44.0 km in length, 200 m in width), for all seasons.

studies rely on Ground Range Detected (GRD) products due to their simplicity and preprocessing availability. In contrast, the use of SLC data, especially in dual-polarized IW mode, remains relatively rare in glaciological applications, primarily due to its larger volume, greater processing demands, and the complexity of handling phase and amplitude information.

In this experiment, SLC scenes were systematically processed into polarimetric decomposition products via the Multi-SAR pipeline [38], enabling access to Kennaugh matrix elements at full resolution. Although the full SLC structure allows for rich polarimetric analysis, the regression experiment presented here relies solely on the K0 element, corresponding to total backscatter intensity, which can be equivalently derived from GRD products. This design choice reflects a focus on minimal yet physically grounded EO inputs. The fact that K0 alone enables meaningful glacier facies regression, particularly during the challenging summer season, underscores its sufficiency as a surface-sensitive indicator. Despite the availability of additional Kennaugh elements such as K5 or K8, K0 was found to carry most of the predictive signal, particularly in capturing wetness and structural transitions relevant to summer facies dynamics.

The full benefit of SLC data becomes apparent when combined with the hypercomplex fusion logic, which condenses the temporal series into seasonally representative bands. Unlike GRD-level composites or simple temporal means, the fused SLC-derived EO stack retains richer polarimetric and structural information, while remaining compact and computationally tractable for downstream modelling.

From a glaciological perspective, the use of temporally aggregated polarimetric decompositions derived from SLC data remains relatively novel. Unlike repeat-pass interferometry, which is well established for deformation monitoring, this approach focuses on the fusion of backscatter-based structural indicators (e.g., Kennaugh elements) across time to characterize seasonal surface transitions. This approach bridges the gap between traditional glaciology, which often relies on optical facies delineation or empirical zone mapping, and modern EO machine learning, by offering physically rich, temporally smoothed inputs that align with seasonal glacier dynamics. The successful application of this pipeline across years and AOIs thus not only confirms the viability of SLC-based modelling but positions it as a scalable, high-fidelity alternative to standard EO inputs in cryospheric research.

Having extracted this physically enriched input stack from SLC scenes, the modelling framework then opts for a remarkably simple yet effective feature: the fused total

backscatter intensity (K0). A defining characteristic of the proposed glacier zone prediction framework is its emphasis on simplicity, both in input design and modelling architecture. The decision to use only a single EO input band, Band 0 (K0), corresponding to the total SAR backscatter intensity, represents a deliberate trade-off between expressiveness and interpretability. Physically, K0 offers a compact yet robust representation of surface scattering behaviour, which is directly linked to glacier zone characteristics such as surface roughness, wetness, and layering. Empirical evaluations confirmed that this band consistently carried the strongest glacier-relevant signal across seasons and AOIs, while also exhibiting low noise and high temporal stability. This minimalistic design not only simplifies the modelling pipeline but also enhances its transparency. By reducing the input space to a single, interpretable EO feature, the model's decision behaviour remains tractable, reproducible, and better aligned with physical glacier processes. Such clarity is particularly advantageous in scientific contexts where explainability is paramount. Moreover, the use of a single input band supports seamless transferability across spatial and temporal domains, as it avoids overfitting to high-dimensional or site-specific patterns often encountered in multi-band fusion setups.

While alternative strategies involving multi-band EO stacks were tested, they resulted in only marginal gains in predictive performance, and at the cost of significantly increased model complexity, training time, and overfitting risk. The compact, physically grounded design centered on K0 thus emerges not as a constraint but as an elegant strength, striking a balance between scientific rigor, modelling efficiency, and practical deployability in remote sensing-based glaciological inference.

Effectiveness of HELIX-Enriched Labels

A central methodological innovation of this setup lies in the use of HELIX-enriched labels, which transform temporally discrete, categorical glacier facies maps into continuous, seasonally aggregated indicators. Rather than treating each label snapshot as an isolated, static class assignment, the HELIX approach aggregates class observations over time, within defined seasonal windows, into float-valued representations that capture the frequency and persistence of each glacier zone at the pixel level. This results in enriched label maps that express soft, probabilistic tendencies rather than hard assignments, thereby enabling a more physically meaningful and informative target signal for learning.

To empirically assess the effectiveness of the HELIX-enriched labels, a targeted regression experiment was conducted using the 2023 EO SAR time-series K0 of AOI 1 as input and summer glacier facies labels of 2023 as targets. The summer season was selected due to its high degree of glaciological variability and frequent facies transitions, which make it the most challenging and label-ambiguous period of the year. As such, it represents a stringent test case for evaluating the stability and informativeness of the HELIX label enrichment approach. A Ridge regression model was employed to quantify prediction accuracy under two supervision regimes: the HELIX-enriched seasonal label (summer), and a series of ten raw temporally discrete individual label scenes acquired throughout the same season. While the main predictive pipeline in this setup relies on XGBoost for glacier facies inference, Ridge regression was used in this comparative experiment to isolate the effect of label enrichment. Given the single-feature setup and the regression framing, Ridge provides a stable, interpretable baseline without introducing unnecessary model variance. Results are therefore attributable to differences in label structure rather than model expressiveness, and would be expected to generalize similarly across more complex learners. As shown in Figure 7.34, regression against the HELIX-enriched target resulted in the lowest overall MAE, indicating a more stable and learnable signal. This finding supports the notion that temporally contextualized label enrichment introduces an implicit regularization effect, particularly valuable during high-variability periods such as summer when glacier facies boundaries are most dynamic.

The implications of this transformation are substantial. First, it reformulates the task from multi-class classification to multivariate regression, a paradigm shift that affords greater modelling flexibility and sensitivity to inter-class gradients. In practice, this allows models to infer not only which zone is most likely at a given location, but also how confidently that prediction reflects temporal consistency (e.g., persistent dry snow vs. intermittent melt). This gradient structure is especially valuable in regions of transition, where facies boundaries are not sharply defined but evolve gradually in response to meteorological forcing.

Second, the smoothing and aggregation across seasons introduce a form of temporal regularization that improves model generalization across both time and space. By encoding typical seasonal behaviour, the enriched labels mitigate overfitting to idiosyncratic label noise or single-date anomalies, and instead guide the model toward learning the broader seasonal dynamics of glacier systems. As illustrated in Figures 7.21 and 7.22, these maps reflect not only intra-annual seasonal structure, but also inter-annual variability, as exemplary showcased across the Devon Ice Cap. The persistence and movement of

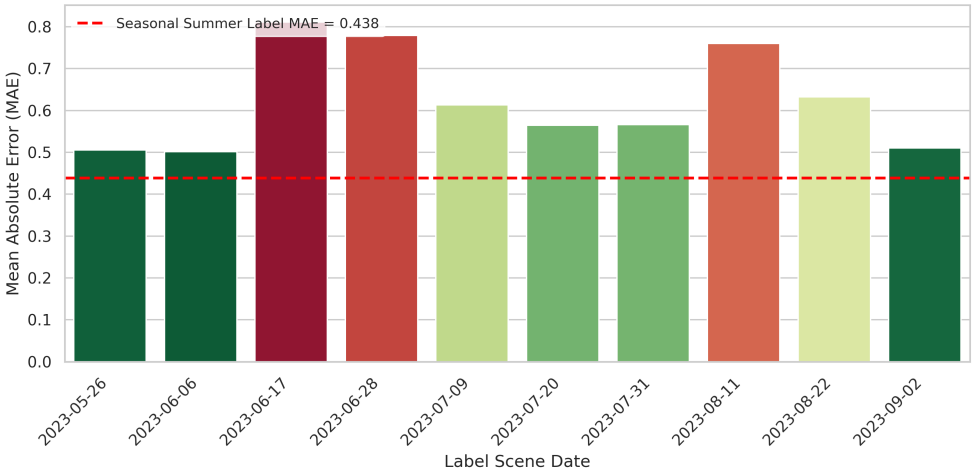


Figure 7.34.: Comparison of mean absolute error (MAE) for regression models predicting summer glacier facies using either the HELIX-enriched seasonal summer label or individual raw label scenes as targets. The HELIX-enriched summer label yields the lowest MAE overall, while individual labels display greater variation depending on acquisition date and scene conditions.

facies boundaries, particularly in regions such as the Cunningham Glaciers, highlight the physical relevance of the temporal signals embedded in the HELIX framework.

Nonetheless, this approach also introduces trade-offs. The act of temporal averaging inevitably discards high-frequency label variation, potentially smoothing over meaningful short-term transitions, particularly during abrupt melt or snowfall events. Moreover, the interpretability of the enriched labels, while conceptually intuitive, becomes less direct than categorical class maps and may pose challenges e.g., when expecting traditional facies delineation. Finally, while regression affords flexibility, it also complicates evaluation, as metrics like R^2 become sensitive to label variance and require careful interpretation alongside absolute measures such as MAE. These considerations suggest that while HELIX enrichment offers clear advantages for learning, it must be applied with awareness of its assumptions and implications for downstream use.

Choice of Base Estimator

The decision to use XGBoost as the core predictive model in this setup was driven by a balance of interpretability, performance, and suitability for the EO data structure. As a

gradient-boosted decision tree framework, XGBoost offers several advantages over deep learning models in the context of temporally aggregated EO data.

First, XGBoost inherently handles non-linearities and complex feature interactions, making it well-suited for learning relationships between SAR-derived backscatter intensities and glacier facies states. Unlike neural networks, it requires minimal preprocessing or normalization and can efficiently operate on tabular or structured data formats like the fused EO input stack used here. Importantly, it supports multi-output regression natively, aligning seamlessly with the prediction of seasonal facies vectors.

Second, its tree-based nature enables explicit feature importance analysis and residual diagnostics, fostering transparency and enabling easy debugging, crucial in scientific applications involving sensitive geophysical interpretations. This stands in contrast to recurrent neural networks (e.g., LSTMs), which, while powerful for raw time series, introduce substantial architectural complexity, require large datasets, and often function as black boxes. Given that this setup operates with seasonally aggregated inputs rather than dense sequential EO data, the temporal dependencies are already abstracted into the input, reducing the need for sequence-aware models.

Furthermore, XGBoost is computationally efficient, scalable across large spatial domains, and robust against overfitting through regularization, early stopping, and tree pruning. These properties make it particularly advantageous in scenarios where model deployment must be efficient and where training data volume is constrained by glaciological label availability.

While future work may explore deep models for spatio-temporal EO directly (e.g., CNN-LSTM hybrids or transformers), the current setup, fused EO inputs and HELIX-enriched label targets, benefits from the structured clarity and generalization capacity of XGBoost, especially when extended via residual refinement mechanisms.

Residual Hint Mechanism

The residual hint mechanism, introduced as a second-stage refinement strategy, proved to be one of the most impactful components of the modelling pipeline. Empirically, the addition of residual hints, computed as the pixel-wise difference between the predicted and actual seasonal labels from the base model, yielded a consistent and substantial performance gain. Across all seasons, the MAE dropped by approximately 30% relative to the baseline model, with notable improvements in more challenging periods such

as summer. These results highlight the mechanism's efficacy in capturing systematic prediction errors that the initial model alone could not resolve.

Conceptually, the residual hint approach functions as a lightweight correction layer. Instead of retraining the base model or adopting more complex temporal architectures, such as RNN or LSTM models, this method reuses the learned residual patterns as auxiliary features. This is both computationally efficient and interpretable: the refinement model operates on explicit error signals, learning where and when the base model tends to under- or over-predict, and adjusting accordingly. Importantly, this two-stage design preserves transparency, modularity, and training simplicity, qualities often sacrificed in end-to-end deep learning systems.

From a theoretical perspective, the residual hints may also be interpreted as a form of gradient re-weighting: by training a second model on residuals, the system implicitly learns a spatial-temporal error distribution over the glacier surface. This information allows the model to focus capacity on regions of high error or class ambiguity, effectively serving as an attention mechanism without the overhead of neural attention layers. Such targeted correction aligns well with the nature of EO-based glacier zone prediction, where certain regions (e.g., zone boundaries or melt zones) consistently exhibit higher uncertainty.

Nevertheless, one limitation of the residual hint approach is its reliance on the quality and representativeness of the initial predictions. If the base model fails to learn meaningful structure, e.g., due to poor input data or mislabelled targets, the residuals may be noisy and uninformative. Additionally, while the residual model improves accuracy, it introduces another modelling stage, potentially complicating deployment and calibration. Future work may explore integrating the residual estimation directly into a unified architecture, or leveraging learned residuals across time and space for meta-learning purposes.

Performance of Ensemble Models

To improve robustness and test the system's capacity for generalization, ensemble predictions were evaluated using combinations of independently trained models. The full ensemble, aggregating predictions from AOI 1 and AOI 2 models from both 2021 and 2022, demonstrated solid transferability to AOI 1 in 2023, a previously unseen year. Despite moderate reductions in R^2 (e.g., $R^2_{\text{year}} = 0.517$), the Year Mean MAE remained

low at 0.433, confirming that absolute prediction error was stable and within expected bounds for all seasons.

A second ensemble, composed solely of AOI 2 models (Devon Island), served as a spatio-temporal transfer benchmark. This setup excluded all prior exposure to the target region and still achieved a Year Mean MAE of 0.444. The ability of this ensemble to perform nearly as well as the full ensemble, despite lacking any spatial overlap with the evaluation AOI, demonstrates the portability of the learned representations and the effectiveness of glacier facies modelling via HELIX-enriched labels and radar-only EO input.

It is noteworthy, however, that while ensembles stabilized predictions, they did not yield the best performance overall. That distinction belonged to the historical-prior model, suggesting that ensembling across diverse years and locations introduces variance trade-offs that may not always improve precision. Nonetheless, ensembles remain valuable tools when retraining is infeasible or when performance must be sustained across varied environmental regimes.

Value of Historical Label Priors

A particularly novel aspect of the modelling approach was the use of temporally aggregated label histories as input priors. The historical label vector $\bar{L}_{\text{hist}}$, derived from per-pixel seasonal averages over 2017–2021, offers a purely label-side mechanism for injecting glaciological memory into the model. This innovation bypasses the need for consistent EO archives, enabling high-fidelity transfer learning even when EO observations from past years are unavailable or noisy.

Empirically, the inclusion of $\bar{L}_{\text{hist}}$ led to the strongest results of all tested configurations. A model trained with this auxiliary feature on AOI 2 (Devon Island) for 2022 achieved a Year Mean MAE of 0.076 and $R^2 = 0.968$ when evaluated on the same domain, outperforming both individual and ensemble baselines. More impressively, when transferred to AOI 1 in 2023, the model still delivered a Year Mean MAE of 0.399, better than either ensemble configuration, despite having no exposure to the region or year.

These results highlight the surprising effectiveness of embedding temporal context through the label domain. The spatial smoothness and seasonal consistency of the historical prior seem to provide a stabilizing effect, especially in ambiguous or low-signal regions. This strategy is especially promising for real-world applications in data-sparse

cryospheric regions, where access to multi-year EO time series is limited but high-quality zone classifications from prior campaigns may be available.

Performance Interpretation Across Metrics and Seasons: The Role of Label Variance and Melt-Season Ambiguity

Across all model configurations and evaluation settings, ranging from intra-AOI training to fully spatio-temporal ensemble generalization, a consistent observation emerged: while Mean Absolute Error (MAE) remained low and stable, the coefficient of determination (R^2) showed substantial seasonal and contextual fluctuation. Most notably, R^2 performance in the summer season was significantly lower compared to other periods, and even turned negative in some transfer settings. In contrast, the corresponding summer MAEs consistently remained within acceptable bounds (typically around 0.49), and Year Mean MAEs across all models never exceeded 0.52.

This discrepancy is not indicative of model failure, but rather reflects the nuanced behaviour of these two metrics under bounded, low-variance target distributions. The enriched seasonal labels, produced from 7-day cadence TSX-based glacier facies maps, are themselves radar-derived and temporally smoothed. Their values span a semi-discrete range of 1–5, representing continuous tendencies toward specific glacier zones (e.g., dry snow, wet snow, percolation). When projected into seasonal summaries, certain seasons, particularly summer and winter, exhibit label compression. In summer, widespread surface melting causes most glacier pixels to saturate toward class 5 (wet snow/melt), substantially reducing the variance of the label distribution.

Because R^2 is variance-sensitive by definition, small absolute deviations from a nearly constant ground truth can yield disproportionately large reductions in explained variance. This is most evident in the summer results: while MAE values in transfer settings (e.g., 0.491 for the full ensemble, 0.495 for the historical-prior model) indicate reliable subclass accuracy, R^2 plummets due to the compressed label space. Conversely, MAE, being a scale-invariant, unit-consistent error measure, provides a direct and interpretable indication of model fidelity, especially suitable for the 1–5 facies encoding. It effectively answers the question: "How far, on average, is the model off from the correct seasonal zone intensity?" An MAE < 0.5 means predictions differ on average by less than half a class. In a continuous 1–5 facies space, this typically implies soft misalignment (e.g., wet snow vs. ice-free) rather than true misclassification.

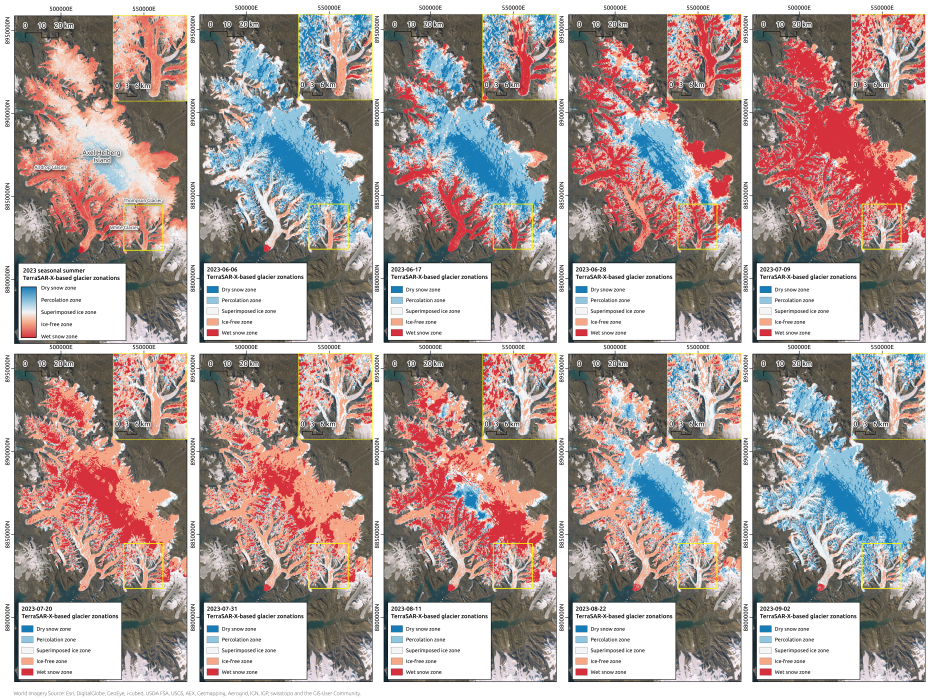


Figure 7.35.: Comparison of model prediction and original TSX-based glacier zonations for the 2023 summer season. **Top left:** HELIX-enriched seasonal mean glacier zones for summer 2023 (continuous target). **Remaining panels:** Individual TerraSAR-X derived glacier zone classifications used as input for the seasonal enrichment, spanning from June to early September 2023. Each classification shows a snapshot of dynamic glacier facies (Dry Snow, Percolation, Superimposed Ice, Ice-Free, Wet Snow) across Axel Heiberg Island. The sequence illustrates the variability and temporal compression characteristic of the melt season, which contributes to reduced variance in seasonal label distributions. The enriched target (top left) reflects the averaged signal of these temporally noisy observations, which the model successfully learns to predict with sub-class precision.

In practical terms, a MAE below 0.5 implies that the model's predictions deviate, on average, by less than half a facies class across all seasons. Given the continuous 1–5 encoding and the natural gradual transitions between glacier zones (e.g., from percolation to superimposed ice), this level of precision is well within acceptable bounds for both scientific interpretation and monitoring applications. Rather than indicating sharp misclassification, such deviations typically reflect minor shifts within the same facies family or adjacent categories. This highlights the robustness of the approach: even under full spatio-temporal transfer, the model maintains sub-class level alignment with glacier

zone patterns, reinforcing its suitability for dynamic, label-scarce environments (see Figure 7.35).

This counterintuitive behaviour is, in part, a direct consequence of the quality and structure of the HELIX-enriched labels themselves. By smoothing categorical class maps over time into continuous, seasonally stable indicators, the label variance is intentionally reduced, reflecting physical persistence in glacier facies rather than random fluctuation. While this promotes robust learning and transferability, it also compresses the target distribution in certain periods (notably summer), amplifying the sensitivity of R^2 to small residual errors. Thus, the observed drop in R^2 is not indicative of poor model generalization, but a known limitation of the metric when applied to low-variance, semi-discrete regression targets. In contrast, MAE remains a more appropriate and interpretable measure under these conditions, reliably capturing sub-class prediction fidelity across the full seasonal cycle.

The seasonal melt period introduces additional complexity. It is both short-lived and highly dynamic, with inter-annual variability in melt onset, duration, and spatial extent. EO responses during this period, particularly from SAR, are non-linear and ambiguous, influenced by changes in surface wetness, roughness, and dielectric properties. Given that both the input EO stack (Sentinel-1) and the supervisory label source (TSX-derived zonations) are radar-based, the learning task effectively becomes a SAR-to-SAR regression. While this cross-band setup ensures physical consistency in terms of observing microwave backscatter processes, it also introduces distinct challenges related to wavelength-dependent sensitivity. TSX X-band labels primarily capture near-surface wetness and fine-scale roughness, whereas Sentinel-1 C-band inputs are more influenced by subsurface scattering and broader-scale surface features. This discrepancy means that the model must learn to map between different depth sensitivities and scattering regimes, introducing both a risk of sensor-driven bias and an opportunity for physically meaningful cross-frequency learning. This risk is particularly acute during melt conditions, where signal saturation and ambiguity can occur in both input and label sources. There is a possibility that the model may inadvertently learn to reproduce sensor-specific artefacts or shared non-glaciological patterns, rather than true physical zonation changes. Although the system demonstrated robust generalization across AOIs and years, these results do not fully preclude the presence of frequency-induced bias, especially in low-variance or ambiguity-prone regimes. Future work may consider integrating non-radar auxiliary features, such as optical melt indicators or modelled melt energy, to further disentangle physical facies dynamics from wavelength-specific backscatter behaviour. Future work

could explore integrating non-radar auxiliary features, such as optical melt indicators (e.g., NDWI or NDSI from Sentinel-2) or physically modelled melt energy estimates (e.g., from surface energy balance models), to further disentangle genuine glacier facies dynamics from wavelength-specific backscatter artefacts. This step would help mitigate the risk of the model learning sensor-dependent noise patterns, especially during the melt season where both Sentinel-1 (C-band) and TSX (X-band) are simultaneously sensitive to surface wetness, roughness, and dielectric fluctuations but with different wavelength-dependent responses. Introducing such cross-modal information would enable the model to anchor its learning more directly in the physical surface processes that govern glacier zone evolution, rather than inadvertently overfitting to radar-specific signal behaviours.

Despite these challenges, all tested models, including the spatio-temporal ensembles and the historical-prior model, achieved high absolute performance. MAE values remained below 0.52 across all seasons and below 0.45 for the annual mean, even in unseen AOI–year combinations. This suggests that the learning framework successfully generalized seasonal zone tendencies, despite sensor ambiguity and label non-uniformity.

To better understand seasonal prediction behaviour and class-wise tendencies, confusion matrices were generated for each season and the annual mean. These visualize the mapping between reference labels and predicted classes based on a fuzzy discretization of the regression output. Instead of hard rounding, predictions in the continuous 1–5 range were softly binned into glacier facies classes, dry snow zone, percolation zone, superimposed ice zone, ice-free zone, and wet snow zone, using a tolerance-aware scheme. The analysis is based on the output of the historical-prior model, applied under full spatio-temporal transfer (trained on AOI2 in 2022, evaluated on AOI1 in 2023). The resulting matrices were normalized by row, reflecting per-class prediction accuracy as percentages. Figures 7.36–7.40 summarize these results for spring, summer, fall, winter, and the annual mean.

These normalized fuzzy confusion matrices confirm the expected class structure while revealing meaningful seasonal variations. Most prediction errors occur between neighbouring classes, such as the frequent mix-up between percolation and superimposed ice zones, or the tendency to conflate wet and ice-free zones during summer. This spatially local confusion is especially evident in transitional seasons like spring and fall, and reflects the physical blending of glacier facies rather than random misclassification. Prediction confidence, as reflected in the dominance of the diagonal entries, is highest in winter and lowest in summer, consistent with earlier observations on seasonal ambiguity, melting onset, and label compression.

Importantly, the use of fuzzy binning reinforces the interpretation that many apparent errors were not true misclassifications, but rather artifacts of discretization in gradient-like zones. The fuzzy matrices preserve class hierarchy, maintain high diagonal dominance across most seasons, and reflect a physically plausible pattern of localized uncertainty near facies boundaries. This behaviour further supports the use of regression-based output in glaciological applications, offering a balance between quantitative accuracy and qualitative interpretability, especially in data-sparse and transition-prone Arctic environments. Overall, the matrices serve not only as a validation tool but also as strong visual evidence of structure-aware generalization under domain transfer. When considered alongside the consistent MAE levels below 0.5, the fuzzy matrices substantiate that prediction errors are largely confined to adjacent glacier facies, not random misclassifications. This effectively answers the question posed earlier: "How far, on average, is the model off from the correct seasonal zone intensity?", demonstrating that an average deviation of half a class should not be seen as problematic. On the contrary, it reflects physically plausible transitions within a continuous facies gradient. These findings highlight that while traditional metrics like MAE and R^2 may appear ambiguous in semi-discrete regression tasks, a tolerance-aware interpretation confirms the model's fidelity in capturing meaningful seasonal structure. The fuzzy confusion matrices help interpret this further: they show that even when the average error is close to half a class, most deviations occur between neighbouring facies. This confirms that such errors are not indicative of structural model failure, but rather reflect physically meaningful transitions, supporting the validity of the regression framework in modelling continuous seasonal glacier dynamics.

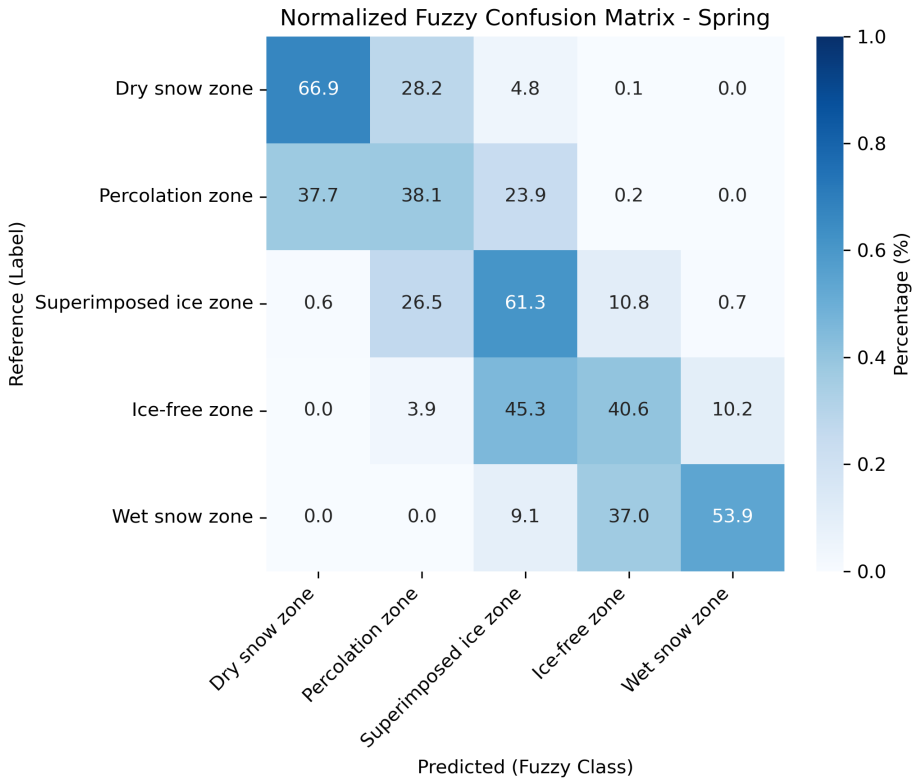


Figure 7.36.: Confusion Matrix – Spring. Reference (y-axis) vs. predicted (x-axis) classes after rounding regression outputs.

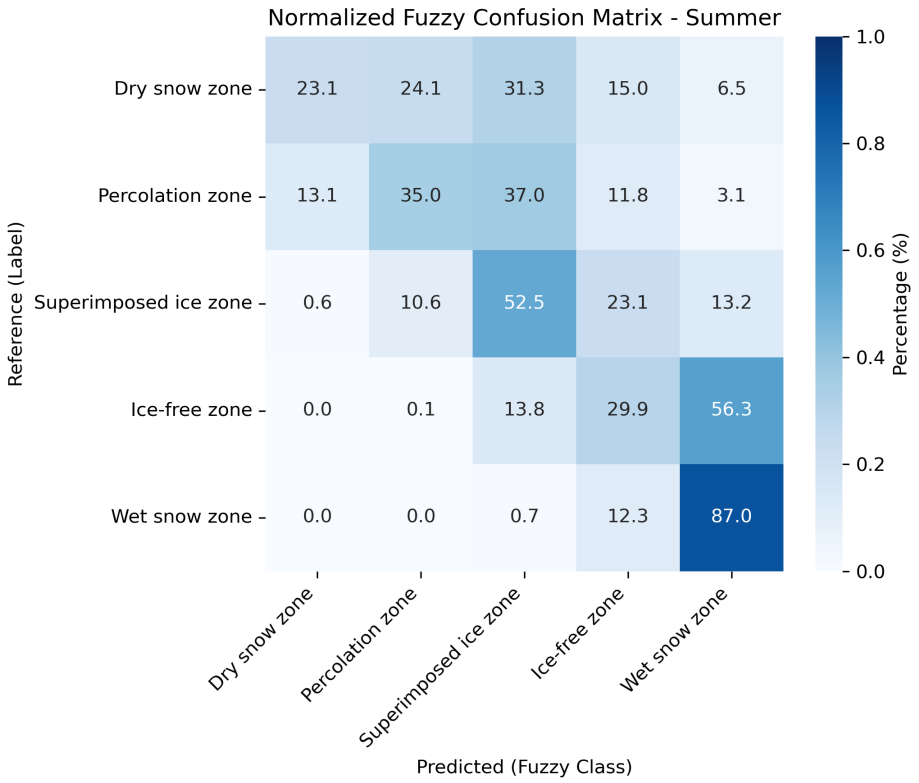


Figure 7.37.: Confusion Matrix – Summer. Reflects high overlap between melt-prone zones.

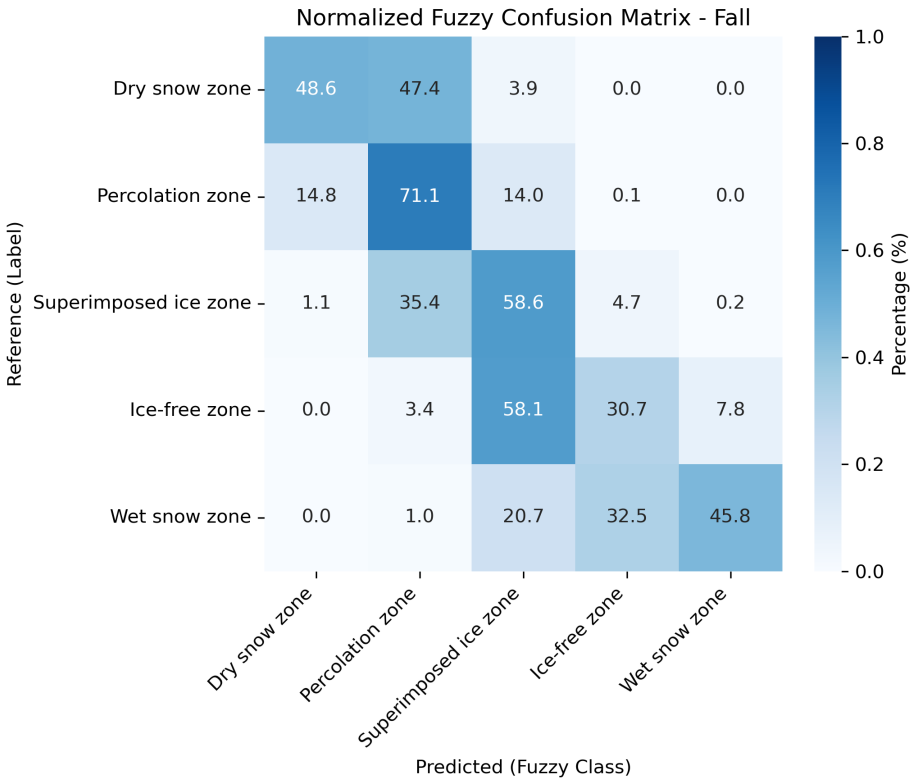


Figure 7.38.: Confusion Matrix – Fall. Shows transition behaviour and class overlap.

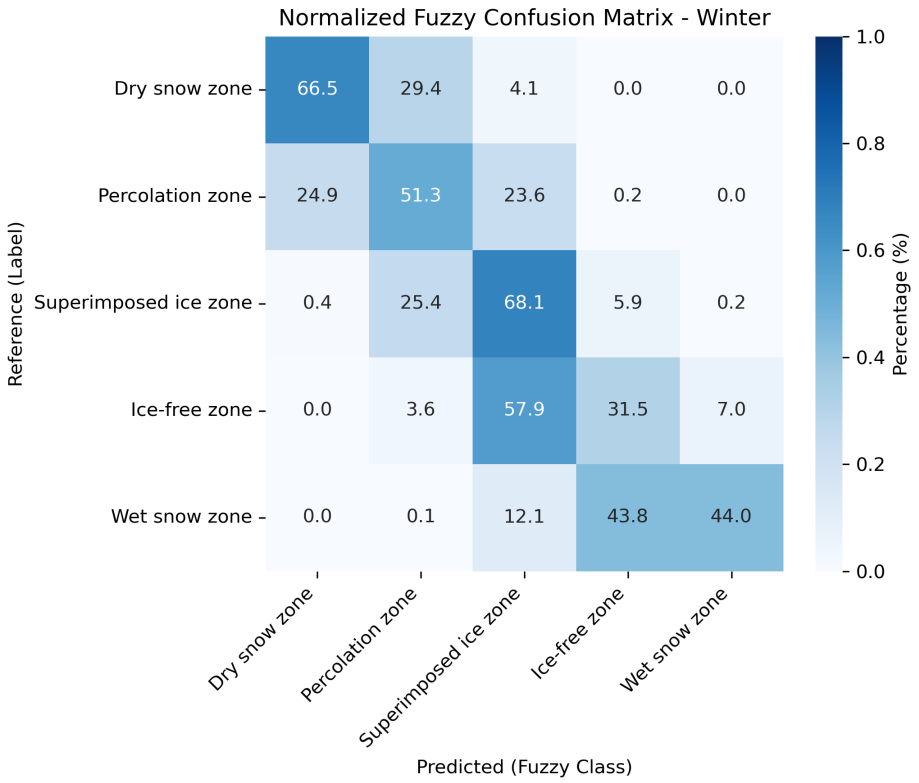


Figure 7.39.: Confusion Matrix – Winter. Indicates stronger class separation and higher prediction confidence.

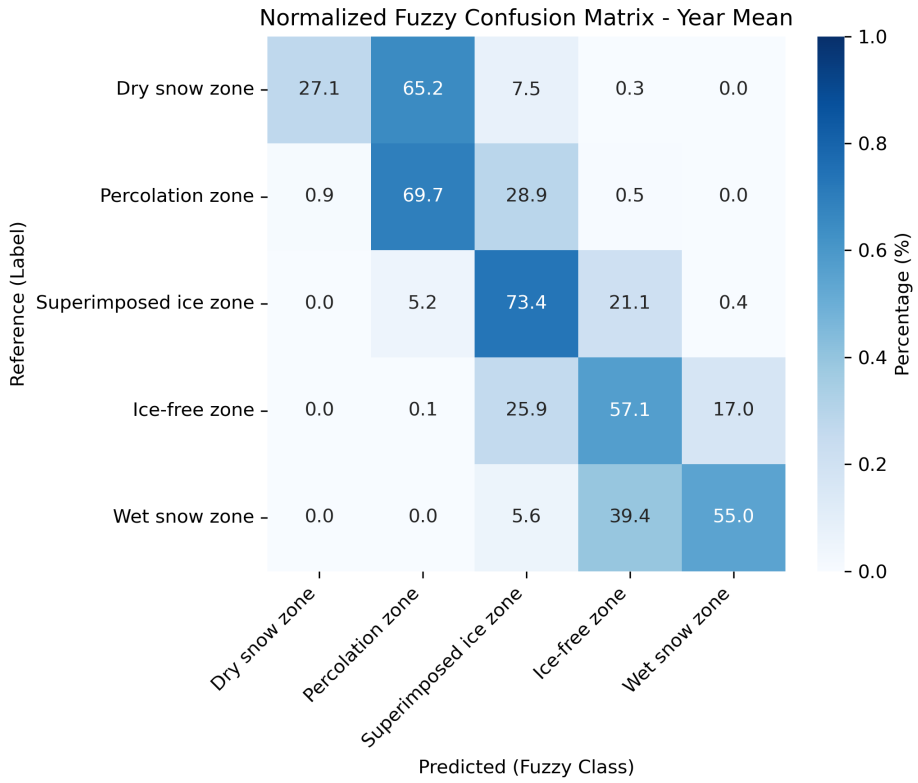


Figure 7.40.: Confusion Matrix – Year Mean. Aggregated across all seasons to reflect overall class-wise prediction behaviour.

To further assess the structural alignment between predicted outputs and HELIX-enriched labels, the *Kullback–Leibler (KL) divergence* was computed for each season. KL divergence quantifies the difference between two probability distributions. Formally, given a reference distribution P and an approximate distribution Q, the divergence is defined as:

$$\mathrm{KL}(P \parallel Q) = \sum_{i=1}^{N} P(i) \log \frac{P(i)}{Q(i)}, \tag{7.5}$$

where $P(i)$ and $Q(i)$ are the class probabilities (here: histogram-normalized frequencies) of class i in the label and prediction maps, respectively. In this expereiment, distributions were computed over five glacier facies classes (1–5), using soft-fuzzified label and prediction values, and normalized over valid pixels per season. KL divergence reflects how much information is "lost" when using Q (the prediction) to approximate P (the

label). A value of zero implies perfect agreement, while higher values indicate increasing distributional mismatch.

Results: Seasonal KL divergence values were consistently low:

- Spring: 0.047
- Summer: 0.072
- Fall: 0.066
- Winter: 0.040
- Annual mean: 0.024

These results indicate that, despite per-pixel deviations (e.g., MAE $\sim$ 0.49 in summer), the predicted facies distributions remain closely aligned with the true underlying class tendencies. In particular, the annual KL divergence of 0.024 demonstrates that, when aggregated spatially and temporally, the regression model preserves the expected structural distribution of glacier zones with high fidelity.

Importantly, this metric complements traditional regression performance indicators by evaluating not only *how far* predictions deviate numerically (as with MAE), but also *how well* they retain the physical and statistical structure of the facies distribution. In cases where labels represent temporally smoothed, semi-discrete probabilities, as in the HELIX framework, KL divergence provides a more semantically meaningful validation criterion than pointwise error alone.

Taken together with the fuzzy confusion matrices and per-class MAE analysis, these KL divergence results reinforce the conclusion that the regression model does not simply fit numeric values, but effectively learns the spatio-temporal logic embedded in glacier zone dynamics.

To further improve performance in melt-dominated conditions, future efforts may include melt-aware model components such as class-weighted training losses, melt-region masking, or the inclusion of additional auxiliary melt indices (e.g., NDWI from optical sources, modelled melt energy). Furthermore, adapting R^2 with respect to constrained label ranges, e.g., using adjusted R^2 or normalized explained variance, may yield more representative performance reporting in semi-discrete regression tasks.

Spatial Consistency and Seasonal Class Realism

In addition to aggregated accuracy metrics, the spatial and seasonal realism of the model predictions was evaluated, through comparing the predictions and the true-class variations, class distribution plots and longitudinal transect comparisons. As the Historical Context-Based Model, comparatively achieved the lowest error rates, the evaluations are focused on that model, which is also a true spatio-temporal transfer configuration, as the model was trained on AOI 2, using EO data from 2022 and the static historical context vector, and then evaluated on AOI 1 in 2023. This configuration constitutes a spatial and temporal transfer setting, with no direct access to either 2023 labels or EO history at inference time. The following analyses offer a qualitative and physically grounded complement to the confusion matrices, emphasizing not just class agreement but the geospatial structure and elevation-aligned logic of the outputs.

Figures 7.26–7.30 display the reference and predicted zonation maps at full spatial resolution, in the left panels the enriched seasonal reference zonations are shown, and the models predictions in the right panel across all four meteorological seasons and annually aggregated, for AOI 1 in 2023. These maps use a consistent colour scheme for facies types, ranging from Dry Snow to Wet Snow, to support direct visual comparison. Qualitatively, the model predictions exhibit strong spatial consistency with the reference data. Across Axel Heiberg Island, high-elevation regions such as at the Airdrop and Thompson Glaciers are correctly dominated by Dry Snow and Percolation zones, while lower elevations transition toward Superimposed Ice, Ice-Free, and Wet Snow zones. The preservation of these elevational gradients, even without any direct input from 2023 label data, illustrates the model's capacity for physically realistic facies reconstruction. Zoomed-in insets (highlighting White Glacier) reinforce this fidelity at finer spatial scales. While some local discrepancies are visible, especially in marginal or narrow zones, the broader structure remains intact. Importantly, mismatches are typically confined to transitions between adjacent facies classes, suggesting the model adheres to plausible glaciological behaviour even under strong spatio-temporal transfer.

Figure 7.25 presents the seasonal distribution of predicted and reference glacier facies across AOI 1, plotted on a logarithmic scale. The model successfully reproduces the dominant seasonal shifts in facies prevalence, with summer dominated by Ice-Free and Percolation zones, and winter showing strong presence of Dry Snow and Superimposed Ice. Slight overestimation of the Ice-Free zone in summer and modest underprediction of

the Wet Snow zone in transitional periods are evident, but overall seasonal structure is well captured.

To assess spatial consistency in detail, class profiles were extracted along three representative glacier transects, White Glacier, Thompson Glacier, and Airdrop Glacier, each spanning significant elevation gradients. Figures 7.31–7.33 show reference and predicted zonations across all seasons for each glacier. In each case, the model preserves the expected accumulation-to-ablation transition: Dry Snow dominates the upper elevations, gradually shifting through Percolation and Superimposed Ice toward Ice-Free and Wet Snow at the termini. Such coherence is particularly encouraging given the absence of glacier-specific tuning, suggesting that the combination of enriched labels and radar-based time-series fusion allows the model to internalize generalizable glaciological structure. Localized misclassifications appear mostly constrained to class boundaries, often between physically adjacent zones (e.g., Dry Snow $\leftrightarrow$ Percolation), and rarely violate the expected elevational logic. These results reinforce the interpretability and robustness of the model, demonstrating its ability to not only predict class proportions but to spatially reproduce glacier zonation patterns that align with physical processes and terrain structure. In real-world deployment scenarios, such fidelity is critical for ensuring scientific utility and user trust in glacier monitoring applications.

From High-Resolution Labels to Operational Prediction

A central practical trade-off in this framework arises from the use of TerraSAR-X (TSX)-based glacier facies classifications, available at 40 m resolution, as the reference for training models that ingest higher-resolution Sentinel-1 (S1) data (10 m). To mitigate this mismatch, all Sentinel-1 inputs were resampled to match the label resolution using nearest-neighbour alignment, ensuring consistent pixel correspondence during training. While the nominal resolution of the input remains finer, this controlled downscaling ensures that the model does not overfit to spatial details absent in the supervisory signal.

At first glance, this setup may appear counterintuitive: it violates the conventional expectation that higher-resolution data should be paired with equally high-resolution labels. Moreover, the TSX classifications themselves are derived from threshold-based heuristics, which introduces another layer of abstraction. Yet, the HELIX enrichment strategy, combined with a residual learning ensemble that includes historical label priors, enables the model to generalize beyond the limitations of its supervision source. Rather

than overfitting to the noise or rigid structure of individual label scenes, the framework extracts and distills the underlying seasonal signal, demonstrating that robust glacier facies prediction is feasible even when learning across sensors and resolutions.

Limitations

While the proposed framework demonstrates strong performance and broad generalization capability, several limitations merit discussion. First, the use of HELIX-enriched seasonal labels introduces assumptions about temporal persistence and facies regularity that may not fully capture rapid or short-lived glacier dynamics. Seasonal averaging smooths over high-frequency variations, potentially masking abrupt transitions such as melt onset or episodic snowfall, particularly in early summer. Moreover, the temporal windows used to define seasons are fixed and climatologically motivated, which may not align precisely with local inter-annual variability in glacier processes. Second, although the enriched labels provide a more informative learning signal than static classifications, they are ultimately derived from TSX-based label maps whose own accuracy and temporal density vary by year and region. Thus, label quality and availability remain a limiting factor, especially for extending the system to less-instrumented areas. Third, the interpretation of continuous facies scores, e.g., a predicted value of 3.7, remains abstract without post-hoc discretization or contextualization, which may limit their utility in field-based applications requiring clearly delineated zones. Furthermore, given the dependency on externally generated TSX-based classification products, label uncertainty and potential misclassifications propagate directly into the HELIX-enriched supervision signals. Errors in the original TSX labels, whether due to sensor noise, misclassification, or inconsistent temporal coverage, can distort the temporal statistics computed during HELIX kernel construction. This is especially problematic in periods of rapid surface change (e.g., melt onset), where even small temporal misalignments can lead to physically implausible class sequences in the target data. To mitigate this, future work could implement HELIX-based temporal consistency checks prior to training. Such a mechanism could, for example, flag or adjust label trajectories that violate known glaciological transition rules (e.g., sudden dry-to-wet reversals during late summer). Probabilistic smoothing or rule-based filters could be introduced at the label aggregation stage, enforcing physically plausible transition paths across the temporal kernel. By integrating such domain-aware consistency constraints into the label generation pipeline, the risk of learning from artefactual or glaciologically implausible targets can be reduced, further enhancing model robustness and interpretability. Finally, while the historical label prior vector proved highly effective

in both local and transfer settings, its utility depends on the availability of accurate multi-year label archives for the target region. In remote or data-sparse glaciers, this constraint may limit the generalizability of this component. Addressing these limitations through adaptive seasonal definitions, hybrid EO-label priors, and uncertainty-aware modelling remains a key direction for future work. Despite these limitations, the demonstrated performance across spatial and temporal domains supports the framework's viability for large-scale monitoring, while highlighting directions for future refinement.

Spatial and Temporal Transferability

A central challenge in EO-based glaciology is ensuring that predictive models trained on one location and time remain effective when applied elsewhere. The results presented here demonstrate that models trained on Devon Island (AOI 2) in 2021 and 2022 generalize successfully to Axel Heiberg Island (AOI 1) in 2023, a separate location and unseen glaciological year. Notably, this spatio-temporal transfer was achieved without retraining, fine-tuning, or access to historical EO or labels for the target region, emphasizing the strength of the modelling pipeline and the robustness of the HELIX-enriched label framework.

This robustness stems from two central design choices. First, the HELIX-enriched labels encode temporal structure and glaciological persistence, enabling the model to learn seasonal tendencies rather than overfitting to static maps or single-date snapshots. Second, the fused EO inputs, derived from SAR-based time series (Sentinel-1), aggregate surface dynamics over defined meteorological seasons. This reduces the sensitivity to outlier events or atmospheric noise, and helps the model detect seasonal phase transitions even in unseen terrain.

Implications and Practical Usage: Together, these features enable pre-trained models to be deployed across other Arctic glacier regions with minimal overhead. In practical terms, using an existing model requires only the following:

1. Acquisition of Sentinel-1 SAR data for the target glacier and year, processed into the same seasonal fusion format (Band 0 of the EO stack), using the freely leviable data-preprocessing (Sentinel-1) and adaptable data fusion algorithms [148].

2. Spatial masking of glacier-covered pixels, optionally based on static glacier outlines, e.g., by using the GLIMS [126] database.

3. Feeding the EO time series into the pre-trained model to predict the enriched seasonal label vector per pixel.

Crucially, no additional labels, field campaigns, or region-specific calibration are necessary. This makes the system not only generalizable, but highly scalable, capable of monitoring glacier zone evolution at continental scales with consistent accuracy.

Toward Arctic-Wide Monitoring: These findings suggest that models trained on a small number of well-characterized AOIs can support broad-scale glacier facies prediction across the Arctic, provided EO coverage is available. With the increasing availability of cloud-based EO processing platforms and global glacier masks, this approach could facilitate low-cost, repeatable assessments of glacier seasonal behaviour, supporting scientific research, climate monitoring, and risk assessment efforts in polar regions.

7.3.5 Conclusions

This study presents a compact, interpretable, and scalable framework for seasonal glacier zone prediction in Arctic regions, based on time-series analysis of radar EO data. The approach centres on a fused representation of multi-temporal Sentinel-1 observations, capturing the seasonal dynamics of glacier surface conditions in a physically grounded and computationally tractable way. Through a simplified yet expressive input design, each pixel is represented by a continuous-valued feature encoding seasonal backscatter evolution, which forms the basis for subsequent learning.

A defining innovation lies in the label-side enrichment strategy, inspired by the HELIX framework. By transforming per-date, discrete glacier facies classifications into seasonally aggregated, float-valued targets, this method introduces temporal semantics into otherwise static labels. This transformation is not cosmetic: it enables predictive modelling in the first place. Without such enrichment, labels remain too fragmented and inconsistent to serve as reliable supervision signals for seasonal inference. The HELIX process thus acts as a form of temporal grounding, allowing the model to learn not only which glacier zones are present, but how they evolve and persist across the glaciological cycle.

Crucially, the approach leverages glacier facies classifications derived from TSX, a radar system distinct from the Sentinel-1 input source. This cross-sensor setup introduces a form of informational triangulation, allowing the system to learn generalizable patterns from one sensor (TSX) and apply them to another (S1). By doing so, the pipeline effectively controls for sensor-specific biases and avoids overfitting to modality-specific features. The predictive capacity of the framework, especially in temporally and spatially unseen regions, hinges on this integration of external, physically grounded label information. It highlights a broader principle: that robust EO-based modelling can be achieved not only through data quantity, but through meaningful structural alignment between inputs and labels.

Empirical results across multiple years and glaciers demonstrate the feasibility and robustness of this method. The models achieved reliable seasonal predictions using minimal input features, and generalized well across both time and space. The introduction of a residual hint mechanism, training a lightweight second-stage model on the errors of the first, further improved accuracy, especially in ambiguous seasonal zones. Among all configurations, the model trained with historical seasonal priors ($\bar{L}_{\text{hist}}$) consistently outperformed others in both local and transfer settings. Across all configurations, Mean Absolute Error remained below 0.5 on the 1–5 facies scale, indicating sub-class prediction accuracy. This highlights the value of label-side temporal memory as a lightweight control mechanism for stabilizing predictions.

Importantly, the system operates without reliance on optical data, dense historical archives, or region-specific tuning. It runs entirely on open-access radar observations and pre-trained components, supporting operational deployment across large-scale cryospheric regions. Its scientific clarity, interpretability, and computational parsimony make it suitable not just for retrospective analysis but for near-operational forecasting in polar environments.

Taken together, this work contributes a novel framework for dynamic glacier zone modelling, one that links physical glacier processes with explainable ML. It systematically dissects and reconstructs the components needed for transferable, interpretable, and temporally coherent EO inference, demonstrating that carefully structured label enrichment and streamlined architectures may match and outperform more complex designs. By learning from one sensor modality (TSX-based Labels), predicting on another (Sentinel-1), and reasoning across time without recurrent inputs, the system effectively performs sensor-to-sensor knowledge transfer, while embedding its own form of internal control through temporally smoothed supervision. By leveraging external sensor intelligence,

embedding temporal memory directly in the labels, and generalizing without optical dependence, the framework defines a new paradigm for self-regularizing, sensor-to-sensor EO learning, scalable, interpretable, and ready for polar-scale deployment.

Lessons Learned

- **Seasonally enriched supervision enables temporally structured glacier zone modelling.** HELIX-style label aggregation transforms discrete facies maps into temporally expressive targets, enabling models to learn glacier behaviour over seasons rather than isolated snapshots.
- **Single-band radar inputs can support high-fidelity seasonal facies prediction.** Total SAR backscatter intensity (K0), extracted from Sentinel-1 time series, captured sufficient signal for accurate multi-season prediction, minimizing model complexity without sacrificing generalization.
- **Residual-based refinement improves predictive performance in ambiguous conditions.** A lightweight second-stage model trained on residuals of the base prediction enhanced accuracy, especially during transitional seasons with high zone variability.
- **Label-side historical priors provide strong generalization anchors.** A static seasonal prior ($\bar{L}_{\text{hist}}$) derived from previous years outperformed complex ensembles in unseen spatial-temporal settings, enabling reliable forecasting without historical EO input.
- **Sub-class prediction accuracy is achievable even in transfer scenarios.** Across all configurations, the model maintained MAE values below 0.5 on a 1–5 facies scale, indicating predictions deviated by less than one zone level on average, even when transferred across years and glacier regions.

Research Questions Revisited

RQ1: *Does seasonal label enrichment improve the model's ability to represent glacier facies transitions compared to discrete classification targets?*

Yes. HELIX-style aggregation introduced seasonal regularity into the supervision signal, which improved regression stability and reduced noise from per-date variability. Comparative results showed that models trained on enriched targets outperformed those using raw labels in both accuracy and interpretability.

RQ2: *How well do temporally fused Sentinel-1 features predict enriched seasonal glacier zone dynamics?*

Very well. A single temporally fused band (K0) derived from Sentinel-1 was sufficient to predict facies distributions across seasons with high accuracy, yielding Year Mean MAE < 0.45 and R^2 consistently > 0.5, even under transfer conditions.

RQ3: *Can a residual-based refinement stage enhance prediction accuracy and robustness across seasons?*

Yes. Incorporating residual hints reduced MAE by over 30% across all seasons. The approach provided modular, interpretable correction without the need for deep architectures or recurrent temporal modelling.

RQ4: *Does the inclusion of historical seasonal priors improve model generalization across glacier regions and years?*

Yes. The historical label vector ($\bar{L}_{\text{hist}}$) offered strong spatial memory and improved generalization to unseen years and locations. This prior-driven model achieved the best overall performance and proved effective even without concurrent EO data from the target domain.

Closing Remarks

This work introduces a streamlined framework for spatio-temporal glacier zone prediction based on radar EO and temporally enriched labels. By embedding seasonal memory into the label space and relying on fused single-band radar inputs, the system infers not only *where zones are*, but *how they evolve over time*.

The results demonstrate that regression-based prediction of glacier zones, supported by label-side temporal enrichment and residual correction, can achieve accurate, interpretable outputs using minimal EO features. This opens the door to efficient, Arctic-scale glacier facies monitoring using fully open-access data sources.

Through structural label design, SAR-to-SAR learning, and historical-prior generalization, this framework offers a reproducible, causally consistent path toward radar-based forecasting of seasonal glacier dynamics, one that is ready for real-world, large-scale cryospheric deployment.

7.4 Glacier Zone Change Forecasting from Polarimetrically and Spectrally Fused Sentinel-1 and Sentinel-2 Data with HELIX Temporal Supervision

This experiment advances the HELIX framework by applying label-side temporal enrichment to short-term glacier zone prediction. Given the sparse temporal coverage of in-situ glacier observations and the operational need for short-term surface condition forecasting (e.g., for hydrological models or field campaign planning), such EO-driven, trend-aware glacier zone forecasting approaches could close critical information gaps in polar regions. Rather than modelling static glacier facies classifications, the task is reframed as a regression over expected zonation change (class deltas) across a 5-week horizon. The predictive model is trained using fused EO features from Sentinel-1 (polarimetric SAR) and Sentinel-2 (spectral reflectance), along with recent glacier class trajectories, to estimate future zonation dynamics from a single EO acquisition date, to estimate future zonation dynamics from a single EO acquisition date *plus* recent class history. This formulation enables learning physically plausible surface transitions directly from satellite-observable properties, supporting trend-aware glacier monitoring in data-sparse polar environments.

The key innovation lies in the label formulation: future glacier zones are encoded as temporally smoothed deltas relative to the present class, enabling trend-aware learning without requiring time-series input features. This HELIX-inspired approach embeds temporal structure into the labels by computing future class deltas as training targets. During training, these enriched targets allow both base and residual models to learn expected short-term transitions. However, during inference, the model relies solely on EO features and past class dynamics, preserving causal integrity in all inputs.

This experiment addresses the following research questions:

RQ1: *Can temporally enriched supervision signals derived from HELIX-style label kernels support accurate learning of glacier zone evolution from EO mono-date data?*

RQ2: *Can glacier zone evolution be reliably inferred from mono-date EO features combined with recent zonation history?*

RQ3: *Is delta regression a suitable alternative to full class prediction in the context of glacier zone modelling?*

RQ4: *How effective is a two-stage regression architecture in capturing both dominant and residual glacier zonation dynamics?*

To assess these questions, a temporally causal modelling pipeline is implemented. It predicts class deltas from fused EO features (Sentinel-1 and Sentinel-2) and past HELIX-enriched glacier zonation. The pipeline is evaluated on held-out test sets using regression metrics (R^2, MAE) and directional classification accuracy. Additional evaluation explores cross-region generalization under consistent preprocessing and input formatting. At inference, the model uses only a single EO scene at time t, the current class map, and the past 5 weeks of class labels to forecast the expected mean future zonation class $\hat{Y}_{t+n}$.

7.4.1 Materials

This experiment builds upon the glacier zone annotations described in the preceding chapter (and Section 1.2.3), where high-resolution TSX imagery was processed into temporally resolved facies maps at 40,m spatial resolution [311]. For the predictive inputs, co-registered Sentinel-1 and Sentinel-2 acquisitions from mid-June 2021 over Axel Heiberg Island were selected to represent typical glacier surface conditions during the early ablation season. All EO data were spatially harmonized to a common 10 m grid and transformed into a fused spectral–polarimetric feature space following the methods outlined in Section 2.2.

Sentinel-1 Acquisition: The primary SAR input was obtained from the Sentinel-1A platform, on 16th of June 2021. This acquisition was selected as it aligns temporally with the optical data and offers minimal surface moisture interference. This acquisition, provided as a SLC product in Interferometric Wide Swath (IW) mode, offers full-resolution complex-valued backscatter information in dual-polarization (HH/VH). The dataset was processed using the Multi-SAR framework [38], including speckle filtering, radiometric calibration, terrain correction, and geocoding. Polarimetric information was then projected into Kennaugh elements k_0, k_1, k_5, k_8, as detailed in Section 2.2 and illustrated in Figure 6.19.

Sentinel-2 Acquisitions: To match the SAR observations, four Level-1C Sentinel-2 scenes were used, providing surface reflectance information across red, green, blue, and NIR bands:

- `S2A_MSIL1C_20210612T215051_T15XVK`
- `S2A_MSIL1C_20210612T215051_T15XWK`
- `S2B_MSIL1C_20210613T202849_T15XVJ`
- `S2B_MSIL1C_20210613T202849_T15XWJ`

All scenes were manually inspected and selected for minimal cloud contamination, using metadata to validate acquisition geometry and auxiliary file integrity. The four tiles were mosaicked to cover the full glacierized extent of Axel Heiberg Island. Each tile contributed critical multispectral bands at 10 m resolution, providing structural and surface condition cues from the visible and NIR domains.

These bands were transformed into a spectral-Kennaugh representation as described in Section 2.2, which decomposes the spectral signature into a hypercomplex basis. This transformation isolates spectral variation from brightness and reduces correlation between bands, enabling more interpretable fusion with polarimetric descriptors. Mathematically, this transformation is defined in Equation (2.5).

Spectro-Polarimetric Fusion (HCB): To integrate the optical and radar observations into a coherent feature space, the spectral and polarimetric Kennaugh representations were fused using the HCB transformation [289]. This approach yields an 8-dimensional feature vector:

$$\mathbf{K}_{\text{fused}} = [K_0, K_1, \ldots, K_7],$$

where K_0 denotes the total fused intensity, and K_1 through K_7 represent orthogonal spectro-polarimetric components capturing directionality, texture, and spectral gradients. The fusion is lossless, invertible, and semantically structured, as introduced in Figure 6.19 and detailed in Section 2.2. The final fused dataset is referred to as the `2021-06-13` HCB-fused stack, and is visualized in Figure 7.41. This 8-band feature space serves as the primary input for the regression models described in the following methods section. It

balances physical interpretability with compactness, providing a rich representation of glacier surface dynamics over multiple spectral and structural dimensions.

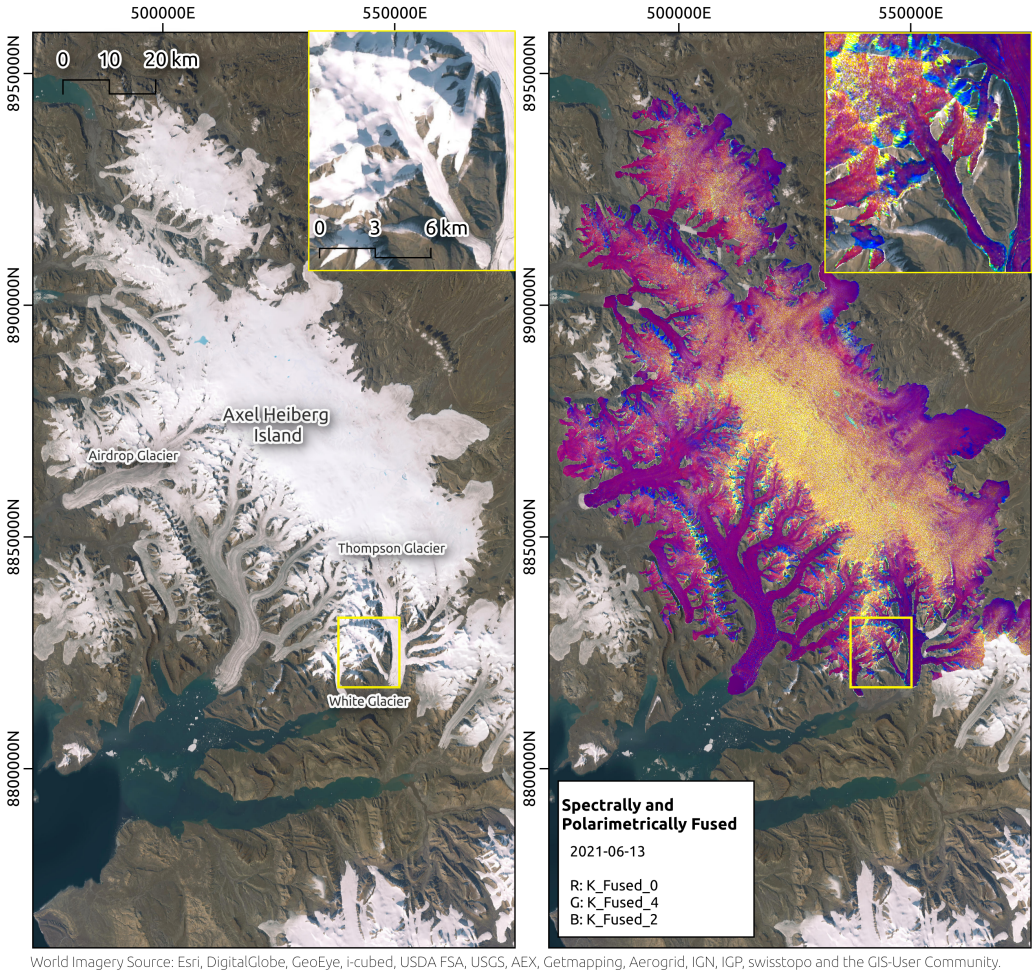


Figure 7.41.: Visualization over Axel Heiberg Island. **Left:** World Imagery [100] basemap showing the geographic extent of the study area. **Right:** Spectrally-polarimetrically fused Sentinel-1 SLC IW dataset from 2021-06-13, displayed in RGB, where Red represents $K_{\text{fused},0}$, Green represents $K_{\text{fused},4}$, and Blue represents $K_{\text{fused},2}$.

7.4.2 Methods

This section outlines the methodological framework for modelling short-term glacier zone evolution using temporally enriched supervision. Instead of learning to classify static glacier zones directly, the approach focuses on regressing the expected change in glacier class, derived from multi-week temporal label context. Inspired by the HELIX framework, temporal structure is injected into the labels rather than the input features, enabling a causally valid, EO-guided learning process.

Figure 7.42 provides a schematic overview of the full pipeline. EO inputs from Sentinel-1 and Sentinel-2 are aligned with TSX-derived zonation maps at time t. These labels are encoded and temporally enriched using past and future class trajectories to construct a per-pixel kernel. From this enriched label structure, a delta target $\Delta = \mu_{\text{future}} - \text{Class}(t)$ is computed, representing the smoothed zonation evolution over a short prediction horizon of 5 weeks (i.e., $t + n = t + 5\text{w}$).

This delta target is precomputed and held fixed during training, serving as a supervision signal for a two-stage regression model that learns to predict future glacier change from static EO inputs.

The following subsections detail each of these components: multi-scale temporal label enrichment, the delta regression formulation, and the two-stage learning architecture.

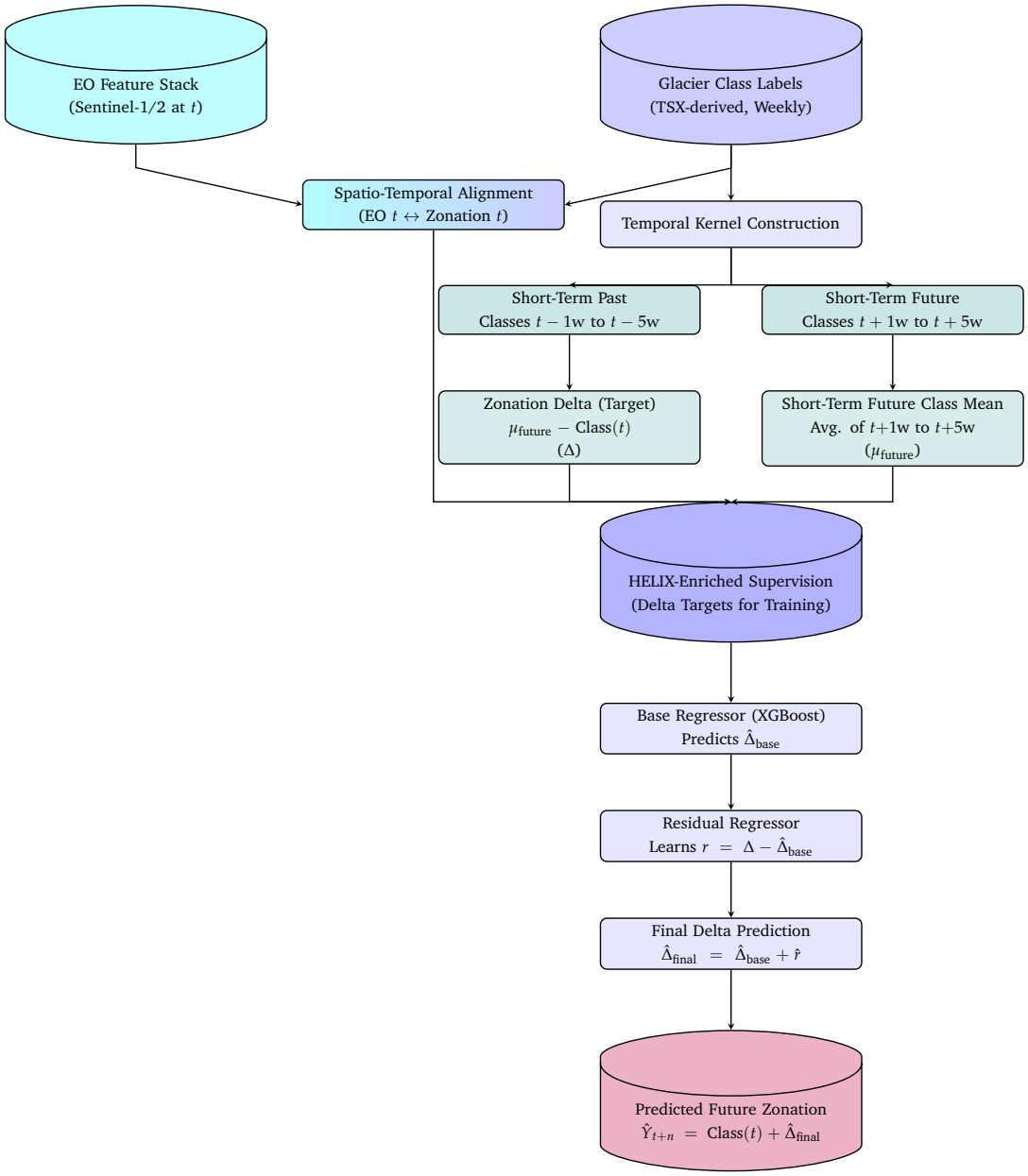


Figure 7.42.: End-to-end schematic of the HELIX-inspired glacier zonation modelling pipeline. Labels are temporally enriched using a structured kernel (past and future class history), from which a delta target $\Delta = \mu_{\text{future}} - \text{Class}(t)$ is derived. During training, EO features and zonation history are used to predict this delta via a two-stage regressor. The final predicted class $\hat{Y}_{t+n}$ is reconstructed by adding the predicted delta to the current class. All future information is used only for supervision and excluded at inference.

Multi-scale Temporal Label Context Enrichment for Dynamic Glacier Zoning

This experiment employs a label-side enrichment strategy inspired by the HELIX framework to embed temporal structure into static glacier classification labels. The resulting structure, referred to as a temporal kernel, provides a multi-scale, time-centric context for each pixel. This enriched supervision encodes not only the current zonation state but also its recent evolution and short-term future trajectory. The temporally structured kernel forms the foundation for learning glacier zone changes via delta regression.

Temporal Kernel Construction: At the core of HELIX-style enrichment is a per-pixel temporal kernel centred at time t. For each EO acquisition date, glacier zone classification rasters are extracted at weekly intervals before and after t from a TSX-derived archive (where w denotes weeks):

- Class label at time t
- Past labels: $t-1$ w to $t-5$ w
- Future labels: $t+1$ w to $t+5$ w
- Aggregated statistics: mean and mode across the past 5 and future 5 weeks

All temporal features are stacked into a single raster aligned to the base TSX feature grid (40 m resolution). This kernel encodes dynamic behaviour and class persistence trends for each pixel over time. This transformation converts discrete, static class labels into temporally structured supervision signals that enable regression-based learning of zonation trends.

Spatio-Temporal Alignment with EO Features: The enriched glacier labels are aligned to the Sentinel-1 and Sentinel-2 EO acquisitions, such that the reference glacier zone map at time t is co-registered to the nearest EO scene. This design choice reflects the fact that EO data, particularly cloud-free Sentinel-2 and high-quality Sentinel-1 backscatter, is temporally sparse and irregular. Since the model must ultimately operate on whichever EO scene is available at a given time, the supervision signal is constructed to match these EO timestamps. This alignment strategy enforces strict temporal causality and mirrors real-world deployment scenarios, where glacier zonation forecasting is based on single-date EO inputs.

Modelling Glacier Zone Dynamics with Enriched HELIX Labels

The modelling task is framed as a regression over glacier zone deltas, defined as:

$$\Delta = \mu_{\text{future}} - \text{Class}(t),$$

where $\mu_{\text{future}} \in \mathbb{R}$ is the mean glacier zone label from $t{+}1\,\text{w}$ to $t{+}5\,\text{w}$, and $\text{Class}(t) \in \mathbb{Z}$ is the current class. The delta $\Delta \in \mathbb{R}$ represents the expected short-term change in zonation, and forms the regression target. This future-derived target is only used during training. It is computed once from future class labels and then held fixed. No future EO or label information is available during inference.

Feature Composition: Each pixel is represented by a fused spatio-temporal feature vector:

- 8-dimensional HCB-fused Sentinel-1 and Sentinel-2 EO features at time t
- Glacier zone class at time t
- Past 5 weeks of glacier zone classes ($t{-}1\,\text{w}$ to $t{-}5\,\text{w}$)

Whereby, all EO features are resampled to the coarser 40 m label grid using spatial averaging, ensuring alignment with the HELIX-enriched label stack. This allows pixel-consistent learning, despite the resolution mismatch, and ensures that the temporal supervision kernel remains spatially coherent with the predictors.

For each downsampled pixel p in the 40 m label grid, EO features are computed by averaging over the corresponding high-resolution pixels:

$$\mathbf{x}_p^{(40\,\text{m})} = \frac{1}{N_p} \sum_{q \in \mathcal{N}(p)} \mathbf{x}_q^{(10\,\text{m})},$$

where $\mathcal{N}(p)$ denotes the set of 10 m EO pixels within the spatial extent of p, and $\mathbf{x}_q$ the EO feature vector at location q.

Two-Stage Delta Regression Architecture: The model learns to predict glacier zone evolution in two stages:

1. **Base Regressor (XGBoost)**
   Learns the dominant spatio-temporal patterns in class change by regressing $\hat{\Delta}_{\text{base}}$ from the input vector.

2. **Residual Regressor (Gradient Boosting)**
   Learns the residual error:
   $$r = \Delta_{\text{true}} - \hat{\Delta}_{\text{base}},$$
   which captures local variation not explained by the base model.

3. **Final Prediction**
   The corrected delta and future class mean are reconstructed as:

   $\hat{\Delta}_{\text{final}} = \hat{\Delta}_{\text{base}} + \hat{r}, \quad \hat{Y}_{t+n} = \text{Class}(t) + \hat{\Delta}_{\text{final}}.$ (interpreted as the expected mean class label over

This structure allows the model to learn both general trends and local deviations in glacier dynamics, improving interpretability and generalization.

Training and Evaluation: The model is trained using an 80/20 stratified split by $\text{Class}(t)$. Evaluation is conducted on held-out data using:

- R^2 and MAE for continuous delta regression
- Directional classification accuracy (up/stable/down)
- Final class classification accuracy after rounding $\hat{Y}_{t+n}$

Strict causality is enforced. All model inputs are limited to time t or earlier. Future class labels (e.g., μ_{future}) are never used as input, only as fixed targets during training. This ensures causal consistency and allows for real-world, forward-only deployment.

While the delta supervision is derived from future labels, these are used solely to construct the training targets and are excluded from all input features. The model thus learns to associate present EO patterns and past class dynamics with typical short-term evolution, generalizing from historical trajectories to unseen futures. This approach preserves strict causality and enables real-world deployment using single-date EO inputs.

7.4.3 Results

his section presents the model's performance in forecasting short-term glacier zonation, emphasizing both the directional accuracy of transitions and the final classification outcomes derived from predicted deltas.

To evaluate the model's ability to forecast short-term glacier zonation,first the enriched glacier class labels that serve as a reference and contextual foundation for training and assessment are visualized. Figure 7.44 displays the HELIX-derived enrichment results over Axel Heiberg Island using RGB composites of class information across time.

In the left panel, the RGB channels represent: **R** — the glacier zones at time t (2021-06-13), **G** — the short-term past glacier classes at $t - 5$w, and **B** — the short-term future glacier classes at $t + 5$w.

This highlights temporal class fluctuations by encoding class transitions as colour shifts.

The right panel extends this by using contextual means rather than single time points: **R** — glacier zones at time t, **G** — the mean glacier classes from $t - 1$w to $t - 5$w, and **B** — the mean glacier classes from $t + 1$w to $t + 5$w.

This representation offers a smoother, more aggregated view of short-term changes and forms the basis for downstream model training and evaluation.

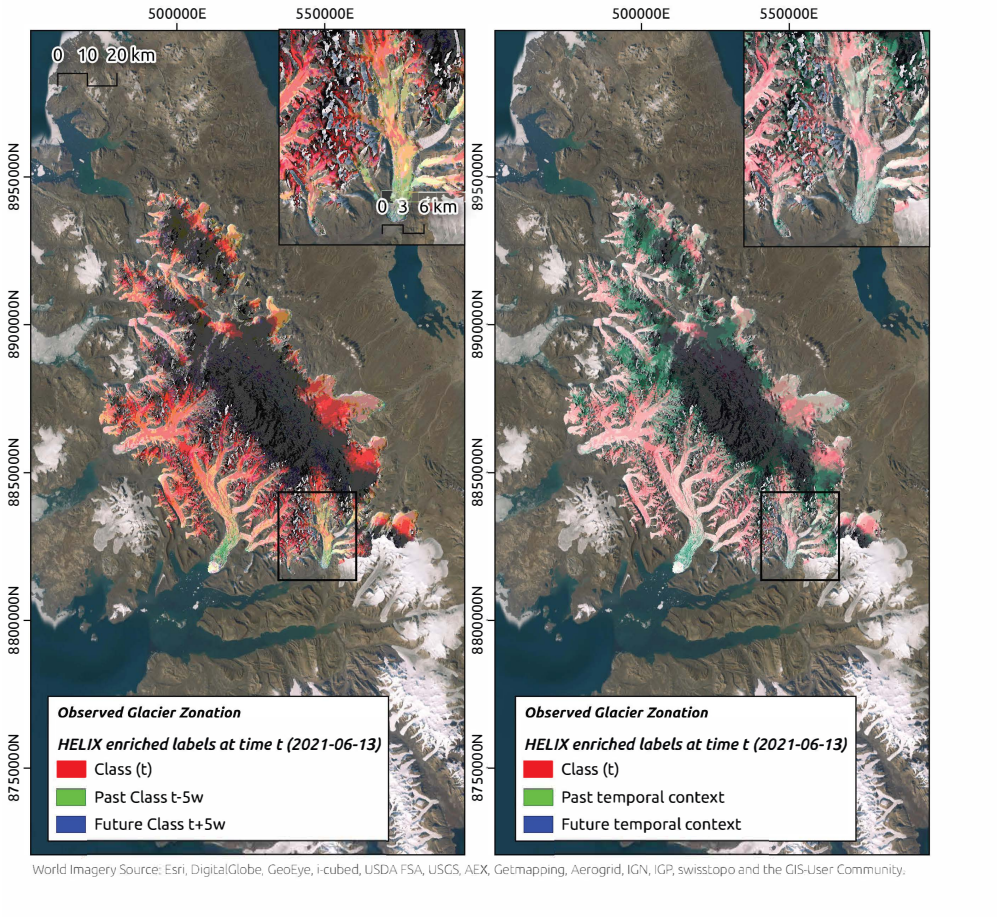


Figure 7.43.: HELIX-based enrichment of glacier zones over Axel Heiberg Island. Left: RGB encoding of glacier classes at t (R), $t-5$w (G), and $t+5$w (B), highlighting localised fluctuations. Right: RGB encoding of class means over t (R), $t-1$w to $t-5$w (G), and $t+1$w to $t+5$w (B), showing short-term zonal trends.

Zonal Delta Regression Performance To better understand the temporal dynamics of zonation changes, descriptive statistics of the class deltas between the current zonation class at time t and each of the future horizons ($t+1$ to $t+5$ weeks) were computed. This analysis revealed substantial variability across the prediction horizons. Notably, for $\Delta t+1$w, there was no change in class for any pixel, indicating a static response in the immediate short term. In contrast, for horizons $t+2$w to $t+5$w, more than half the pixels exhibited changes, with a growing proportion of transitions toward lower classes (e.g., wet to dry zones). These findings justify modelling each future step independently and emphasize the necessity of per-horizon evaluation metrics. The detailed statistics are presented in Table 7.15.

Table 7.15.: Summary statistics of class change (Δ Class) at future time steps, relative to the reference week (t). These distributions motivate the separate evaluation of each horizon.

Δ Step	Mean Δ	Std Δ	Median	% Δ ≠ 0	Unique Δs
Δt+1w	0.000	0.000	0.000	0.00%	[0]
Δt+2w	0.737	0.961	0.000	52.82%	[−4, . . . , 4]
Δt+3w	−0.847	0.968	−1.000	56.98%	[−4, . . . , 4]
Δt+4w	−0.799	0.971	−1.000	55.60%	[−4, . . . , 3]
Δt+5w	−0.921	1.035	−1.000	59.35%	[−4, . . . , 4]

Regression results for short-term zonal delta predictions revealed a near-perfect fit at the shortest time horizon. For $\Delta t + 1$w, where no class changes were observed, the model achieved $R^2 = 1.000$, MAE = 0, and residuals were uniformly zero. For horizons $\Delta t + 2$w to $\Delta t + 4$w, performance remained strong, with low MAE and high R^2 values, indicating the model's ability to capture short-term dynamics from EO and historical zone context.

At $\Delta t + 5$w, however, predictive accuracy declined ($R^2 = 0.875$, MAE = 0.241), reflecting greater uncertainty over longer horizons. This shift likely results from temporal decoupling between predictors and outcomes, driven by external factors not captured in the EO data (e.g., meteorology, ice dynamics). Nonetheless, the model retained substantial predictive value even at this stage.

Table 7.16.: Regression performance summary across future time horizons.

Time Step	R^2	MAE	Residual Test
$\Delta t + 1$w	1.000	0.000	NaN
$\Delta t + 2$w	0.998	0.052	$p < 10^{-20}$
$\Delta t + 3$w	0.992	0.078	$p < 10^{-20}$
$\Delta t + 4$w	0.985	0.102	$p < 10^{-20}$
$\Delta t + 5$w	0.875	0.241	$p < 10^{-20}$
ΔMean	0.988	0.048	$p < 10^{-20}$

Residual tests refer to non-parametric significance testing (Kruskal–Wallis) on model residuals vs. reference deltas. The extremely small p-values ($p < 10^{-20}$) indicate that residuals, while numerically small (see MAE), are statistically distinguishable from zero due to the large sample size.

Directional Transition Accuracy. To assess whether the model correctly captured the direction (i.e., upward = transition to a higher class, such as from firn to dry snow; downward = e.g., from wet snow to firn) of glacier zone transitions (i.e., upward, downward, or stable class shifts), the predicted delta values were discretized into three directional categories. The results are summarized in Table 7.17.

Table 7.17.: Directional transition accuracy (Down / Stable / Up) for each delta horizon.

Time Step	Accuracy	Macro F1	Observation
$\Delta t + 1$w	1.00	0.33	All predicted as Stable
$\Delta t + 2$w	0.53	0.40	Directional separation begins
$\Delta t + 3$w	0.59	0.35	Good recall, weak balance
$\Delta t + 4$w	0.57	0.55	Stronger directional balance
$\Delta t + 5$w	0.56	0.57	Highest directional F1

While the earliest prediction horizon (Δt+1 w) shows perfect accuracy, this is due to a degenerate case where all predictions default to the majority "Stable" class. As prediction horizons increase, directional separation improves—reflected in both reduced overall accuracy and rising macro F1 scores. The macro F1 metric, which averages F1 scores across all classes equally, highlights this gain in balance. At Δt+5 w, the model achieves its best directional performance, suggesting that temporally enriched supervision supports robust learning of medium-term zonation trends, even from mono-date inputs.

Discrete Accuracy of Mean Future Zonation. The final predicted zonal class means, obtained by summing the modelled delta and the current class and rounding to the nearest discrete class, were benchmarked against the HELIX-derived future mean zonation. Table 7.18 shows the classification results for each glacier zone class.

Finally, for the aggregated average future class delta (ΔMean), the model achieved excellent predictive capacity with $R^2 = 0.988$, MAE = 0.048, RMSE = 0.073, 97.00% classification accuracy, and a Cohen's $\kappa = 0.958$, indicating strong agreement beyond chance. These metrics demonstrate the model's robustness in capturing longer-range trends despite individual week-to-week uncertainty.

Table 7.18.: Classification performance of predicted vs. reference future glacier zones.

Class	Precision	Recall	F1-score	Support
Dry snow zone	0.975	0.988	0.982	292,460
Percolation zone	0.981	0.974	0.977	824,578
Superimposed ice zone	0.978	0.945	0.961	1,066,013
Ice-free zone	0.951	0.990	0.970	946,077
Wet snow zone	0.984	0.947	0.965	41,495
Overall Accuracy	0.970 (on 3,170,623 pixels)			

The predicted mean zonation achieves a global accuracy of 97%, with precision and recall consistently above 95% across all classes. These results confirm that despite potential error accumulation in intermediate deltas, the model successfully learns stable surface behaviour and produces reliable spatial forecasts of future glacier zones.

Figure 7.44 shows a side-by-side comparison of the reference and predicted future glacier zones. The class-wise performance summary is provided in Table 7.18.

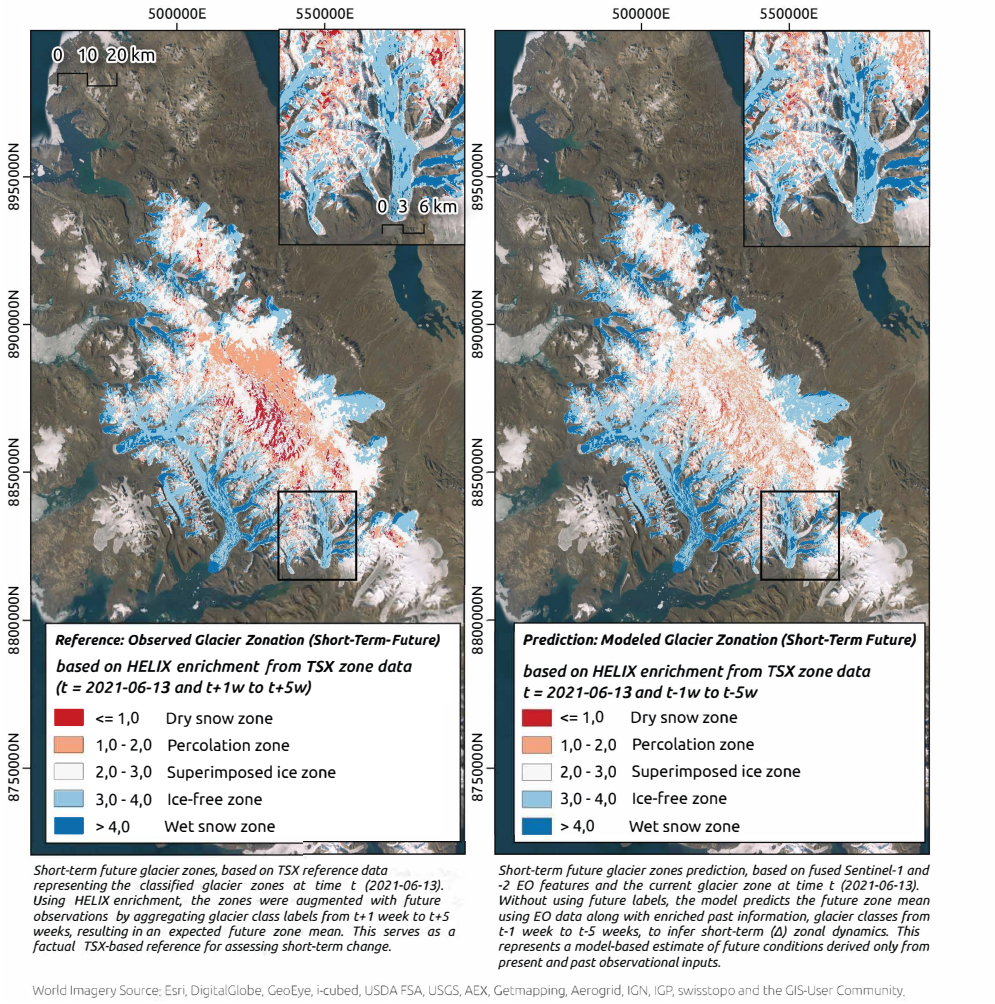


Figure 7.44.: Comparison of short-term future glacier zones. Left: Reference based on TSX-classified zones at time t (2021-06-13), enriched via HELIX with labels from $t+1$ to $t+5$ weeks. Right: Prediction based on fused Sentinel-1/-2 features and prior glacier classes ($t-5$w to t), without using future label input.

7.4.4 Discussion

This section reflects on the rationale, performance, and broader implications of the proposed HELIX-enriched glacier zone forecasting framework. We revisit the key design choices, label-side temporal enrichment, delta-based regression, and causal modelling, and evaluate how these components contribute to the framework's predictive accuracy, physical interpretability, and real-world deployability. The discussion also examines

model behaviour across forecast horizons, assesses the plausibility of misclassifications, and outlines both current limitations and future extensions.

Label-Side Temporal Enrichment as Supervision Strategy The HELIX framework introduces a novel form of temporal supervision by enriching the label space rather than the feature inputs. Each training target is constructed using both past and future glacier class trajectories (plus their mean and mode), forming a temporally structured kernel per pixel. This stands in contrast to conventional temporal modelling, which injects time into the feature stack, often at the cost of causal validity.

This design is conceptually aligned with glaciological dynamics. Surface zone transitions occur gradually, driven by thermal and hydrological processes with inertia. Consequently, regions on the verge of transition often exhibit premonitory signals in EO features, such as emerging melt signatures or reflectance anomalies. By encoding expected future evolution directly into the training target, the model learns to associate present-day EO cues with short-term class changes, without requiring future inputs at inference.

Causal Learning of Glacier Zone Evolution from Present Observables The model exploits the lag between observable surface signals and their eventual manifestation in zonation class change. Sentinel-2 reflectance and Sentinel-1 backscatter jointly capture surface texture, wetness, and snow cover, attributes that correlate strongly with upcoming transitions (e.g., dry snow to percolation, firn to ice).

The learning process is designed to respect strict temporal causality. During training, the model sees only EO and label data up to time t; the future is accessed solely to compute a stable target: the mean glacier class label across the next five weeks (μ_{future}). This delta target ($\Delta = \mu_{\text{future}} - \text{Class}(t)$) smooths short-term noise and emphasizes persistent evolution. During inference, only current and past inputs are available, ensuring that the model forecasts from truly observable conditions.

Delta-Based Supervision for Trend-Focused Forecasting By shifting the learning objective from discrete classification to regression over expected class change, the model focuses on trend detection rather than state replication. This enhances both performance and interpretability. Predicting a continuous delta:

$$\Delta = \mu_{\text{future}} - \text{Class}(t),$$

allows the model to represent directional trends (e.g., +1: melt onset; 0: stability; –1: accumulation), rather than forcing a hard classification. The two-stage architecture, a base regressor followed by residual correction, further improves the ability to capture both dominant and local behaviours in glacier evolution.

This structure makes the model robust to noise in weekly class labels and better suited to representing smooth transitions inherent in physical glacier processes. Moreover, the final output ($\hat{Y}_{t+n} = \text{Class}(t) + \hat{\Delta}$) provides an interpretable, physically grounded forecast.

Interpreting Predictive Accuracy The model achieves near-perfect predictive performance across short-term horizons. For forecast windows of up to 4 weeks, R^2 approaches 1.0 and MAE remains near zero (in zonation units), indicating high accuracy. This reflects both the persistence of glacier zones over weekly timescales and the fact that EO data at time t captures precursors to future change.

Notably, these results do not stem from data leakage. Rather, they emerge from (1) the use of temporally aggregated targets that emphasize stable trends, and (2) the physical observability of glacier state transitions. As forecast horizons increase, performance gradually degrades (e.g., $R^2 = 0.875$ at 5 weeks), suggesting that EO-only inputs are increasingly insufficient in the absence of explicit meteorological forcing.

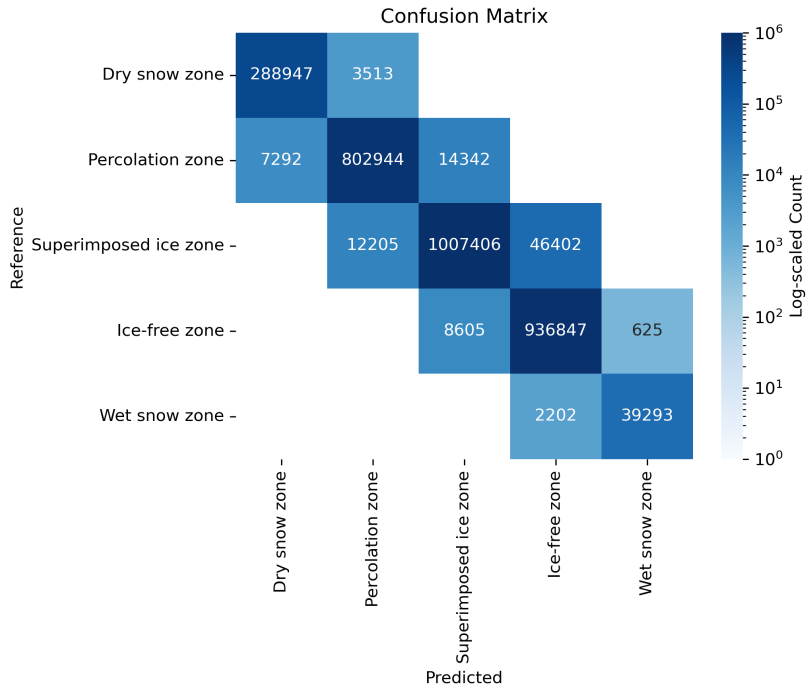


Figure 7.45.: Confusion matrix of predicted vs. reference glacier zone classes. Values are log-scaled counts.

On the Physical Plausibility of Misclassifications Figure 7.45 shows a strong diagonal in the predicted-vs-true class matrix, indicating overall high agreement. These confusion patterns are physically plausible and reflect the continuum-like nature of glacier surface zones, where transitions are often gradual and not strictly bounded. The use of a log-scaled colour bar (10^0 to 10^6) ensures that both dominant and subtle misclassifications are visible, avoiding saturation in large classes while still surfacing minority errors. Where misclassifications occur, they are glaciologically interpretable:

- **Percolation zone** is confused with **Dry snow** or **Superimposed ice**, reflecting their intermediate surface states.
- **Superimposed ice zone** overlaps with **Ice-free areas**, especially near debris-covered margins or exposed bare ice.
- **Wet snow** is misclassified as **Ice-free** in some cases, likely due to transient drying or optical ambiguity under cloud-affected scenes.

Such errors reflect the continuum-like nature of glacier surfaces and reinforce that the model's behaviour is physically reasonable.

Deployment and Transfer Potential Because the model relies only on current EO inputs and prior label history, it can be deployed operationally for any glacier where:

- Recent Sentinel-1 and Sentinel-2 EO data are available,
- A zonation map at time t can be inferred (e.g., from TSX or proxy models),
- Past 5 weeks of zonation history are available or predicted.

Transferability depends on the similarity of glacier regimes and preprocessing consistency, but the use of EO-visible features suggests strong generalization potential across mid- to high-latitude glaciers with similar dynamics.

Advantages of the HELIX-Enriched Framework The proposed approach offers several conceptual and operational benefits:

- **Physically informed targets:** Delta regression aligns with the continuous, inertia-driven nature of glacier change.
- **Causally sound inference:** All features used are observable at time t, enabling real-time deployment.
- **Label-side temporal supervision:** Encodes dynamics without contaminating inputs, improving generalization and reducing noise.
- **Interpretability:** Continuous outputs allow analysis of change direction and magnitude, beyond classification accuracy.

Limitations and Future Directions Despite its strengths, the method has several important limitations:

1. **Dependence on upstream class labels:** Errors in TSX-derived zonation maps propagate into supervision, potentially biasing learning.
2. **Uniform temporal weighting:** The current temporal kernel treats all past/future weeks equally; adaptive weighting could improve fidelity.

3. **Missing physical drivers:** The model does not incorporate explicit meteorological or topographic inputs, which may be critical under extreme or rapidly changing conditions.

4. **Smooth-transition assumption:** Delta regression assumes continuity, which may under-represent rare abrupt events (e.g., calving, rockfall).

5. **Domain generalizability untested:** While promising for other applications (e.g., permafrost, vegetation change), HELIX-style label enrichment has yet to be evaluated outside glacier zoning.

6. **Seasonal stability bias:** Model performance was evaluated on early summer (mid-June) scenes, when glacier zones are relatively stable. Performance under more dynamic seasonal conditions (e.g., spring melt onset or autumn refreezing) remains untested and may degrade in the presence of rapid surface transitions.

Future work could address these gaps by integrating meteorological forecasts, testing adaptive temporal kernels, and extending the method to other domains with dynamic class transitions. Deep learning architectures, such as temporal attention models, may further enhance performance by learning when and where past dynamics matter most.

Outlook: Toward Physically Guided EO Forecasting This work demonstrates that glacier evolution is not only observable, but learnable, if supervised with labels that reflect physical processes and temporal dynamics. By rethinking how temporal context is encoded and aligning model structure with glaciological reasoning, the HELIX framework opens a path toward causal, interpretable, and deployable forecasting from Earth Observation. Extending such approaches beyond glaciology may help bridge the gap between EO-based monitoring and proactive environmental decision-making.

7.4.5 Conclusions

This study reframes short-term glacier zonation modelling as a delta regression task grounded in label-side temporal enrichment. By drawing temporal structure into the supervision signal, rather than the inputs, it is demonstrated that glacier surface evolution is learnable from mono-date EO imagery and past class trends.

Lessons Learned

- **Label-side temporal enrichment offers a powerful alternative to time-series feature modelling.** By embedding past and future class information into the labels, rather than the predictors, we enable models to learn temporally structured trends without compromising causal validity.
- **EO mono-date data contain sufficient signal to infer short-term glacier surface evolution.** Sentinel-1 and Sentinel-2 data at a single point in time, combined with past label context, support accurate prediction of zonal class changes over several weeks.
- **Delta-based targets align with glacier dynamics.** While we did not explicitly compare delta vs. full-class modelling, learning class deltas reflects the physical reality of gradual surface transitions and allows the model to estimate directional change more naturally.
- **A two-stage residual learning architecture improves predictive accuracy.** Capturing dominant trends with a base regressor and refining errors through a residual model leads to higher fidelity predictions, especially in complex terrain or transitional zones.
- **The HELIX framework enables structured, interpretable glacier modelling.** Its use of temporal kernels, aggregated supervision, and causal input formatting supports reliable, real-world application without relying on future EO data.

Research Questions Revisited

RQ1: *Can temporally enriched supervision signals derived from HELIX-style label kernels support accurate learning of glacier zone evolution from EO mono-date data?*

Yes. The HELIX-inspired label enrichment embeds short-term temporal context into the training signal, enabling models to learn expected zonation trends from static EO inputs. This setup yielded strong regression results ($R^2 > 0.98$ up to 4-week horizons), demonstrating that label-side temporal structure can compensate for the lack of temporal EO stacks during training.

RQ2: *Can glacier zone evolution be reliably inferred from mono-date EO features combined with recent zonation history?*

Yes. Despite the absence of future EO data, the combination of mono-date Sentinel-1/-2 features and short zonation history enables accurate short-term forecasting of glacier zone changes. The results show high predictive performance across multiple horizons and confirm that EO features at a single timestamp contain sufficient signal when combined with recent class trends.

RQ3: *Is delta regression a suitable alternative to full class prediction in the context of glacier zone modelling?*

Yes. Modelling the expected change (Δ) rather than absolute future classes allows the model to focus on trends rather than discrete transitions. This improves learning stability, temporal generalization, and interpretability, especially given the gradual nature of glacier surface evolution.

RQ4: *How effective is a two-stage regression architecture in capturing both dominant and residual glacier zonation dynamics?*

Highly effective. The base regressor captures broad spatio-temporal patterns, while the residual model improves fine-grained accuracy in complex regions. This modular approach boosts performance relative to single-stage baselines and provides more interpretable error structure.

Closing Remarks

This work introduces a HELIX-inspired approach to modelling short-term glacier zone dynamics by using temporally enriched supervision and delta regression. By shifting temporal information to the label space and relying on EO mono-date features, the model learns to infer not just *what is*, but *what will* be.

The results demonstrate that delta-based supervision, fused EO features, and residual learning can jointly support robust, temporally aware glacier forecasts. The proposed framework offers a reproducible, causally sound foundation for cryospheric EO modelling, and opens new paths for forecasting rather than simply classifying glacier surface change.

8 Conclusions and Outlook

> *All truths are easy to understand once they are discovered; the point is to discover them.*
>
> **— Galileo Galilei**
> Astronomer, Physicist

The preceding chapters have explored the methodological and conceptual components involved in translating EO data into ecologically grounded inference. Rather than treating EO as a purely technical exercise, the work has positioned it as a framework requiring careful alignment between observation, representation, and interpretation. With the empirical findings and methodological proposals now laid out, the concluding chapter seeks to synthesise the main contributions of the thesis. It is intended to offer an integrative perspective on how the combination of feature fusion, enriched supervision, and benchmarking protocols may support a more robust and transferable use of EO in environmental modelling. In doing so, the chapter reflects on the broader implications of the research and considers potential avenues for future work.

8.1 Overview and Reflections on the Research Journey

This thesis was motivated by a foundational question: *how can EO data be translated into ecologically meaningful, spatially transferable, and operationally robust insight?* At its core, this inquiry recognises EO not simply as a passive data acquisition process, but as a system of environmental inference, a medium through which the state and dynamics of the Earth can be measured, modelled, and ultimately understood. The ISPRS defines remote sensing as "*the science and technology of capturing, processing and analysing imagery, in conjunction with other physical data of the Earth and the planets, from sensors in space, in the air and on the ground.*" While accurate, this definition conceals a more fundamental challenge: remote sensing imagery alone does not yield understanding. The

pixel records energy and location, not meaning. Meaning must be constructed through informed transformation, contextualisation, and learning.

This thesis has argued that addressing environmental questions with EO requires more than sophisticated models or high-resolution inputs. It demands conceptual alignment: between *what* is observed, *how* it is structured, and *what* it seeks to represent. EO–ML pipelines, if they are to yield actionable knowledge, must encode not just surface reflectance but ecological structure; not just predictions, but interpretability; not just signal, but uncertainty. Environmental phenomena, whether glacier melt, forest disturbance, or karst-related sinkhole activity, are inherently spatial, temporally dynamic, and often ambiguous. This demands EO–ML approaches that move beyond naive pixel-wise classification or regression, instead aligning features, labels, and models structurally across scales and time.

The work undertaken in this thesis builds a conceptual and methodological pipeline that integrates EO feature fusion, label enrichment, and learning system design. This was not approached as a monolithic system but as a modular logic: each component, fusion strategies, enriched labels, residual learning, was independently examined and collectively integrated. Through case studies spanning arid, temperate, and cryospheric domains, the thesis demonstrated that robustness and generalisation in EO–ML increase not by stacking complexity, but by respecting the structure of the environmental processes under study.

Three intersecting insights structured this research journey:

1. **EO–ML performance is bottlenecked less by model architecture than by the structure and semantics of inputs and supervision.** Across multiple case studies, from forest structure estimation to seasonal glacier zonation, results consistently showed that meaningful gains came from improving the alignment and quality of input features and supervisory labels. Marginal benefits from deeper models paled in comparison to those from well-structured supervision and temporally aware fusion.

2. **Fusion must be ecologically informed, not merely technically feasible.** The integration of different EO modalities, Sentinel-1 SAR, Sentinel-2 optical, and legacy systems such as TSX or ALOS, was most successful when guided by ecological logic. This included phenological awareness (e.g., dry-season SAR vs. peak-vegetation optical) and process sensitivity (e.g., vegetation decline signalling subsidence). The

CDVI index operationalises this principle by explicitly encoding contrast between ecologically informative time points.

3. **Interpretability, uncertainty, and transferability are not secondary concerns, they are central to trustworthy EO.** The HELIX was designed to embed spatial and temporal context into labels, enabling standard models to reason structurally even under sparse or noisy supervision. Residual-aware feedback architectures allowed the model to learn from its own errors, surfacing epistemic uncertainty. The Wald5Dplus benchmarking study provided systematic insight into which model–modality combinations generalise best, and why.

These insights were not pursued in isolation. They shaped the design logic of new indices (e.g., CDVI), the structure of label descriptors (HELIX), and the organisation of benchmarking protocols (Wald5Dplus). They also underpin the central argument of this thesis: that EO-based environmental modelling requires a move from superficial correlation to structural alignment, between the data collected, the processes studied, and the systems built to make sense of them.

This research journey has reframed EO not as a pipeline of data, but as a grammar of environmental representation, one that requires carefully aligned syntax (features), semantics (labels), and inference (models). The chapters that follow in this conclusion will examine each core contribution in turn, offering an integrated synthesis of how fusion, supervision, and benchmarking can jointly elevate EO from imagery to insight.

8.2 Evaluating Multi-Modal and Multi-Temporal EO Predictive Capacity

A core research objective was to evaluate the predictive capacity of individual, multi-modal, and multi-temporal EO inputs. To this end, multiple dataset configurations were tested, each designed to isolate the contributions of specific sensing modalities, fusion strategies, and temporal context:

Sentinel-1 and -2 (Spectral, Polarimetric, and Temporal Hypercomplex Fusion): The most comprehensive configuration, combining full temporal sequences, polarimetric SAR, and spectral data through hypercomplex fusion. This setup captures both seasonal variability and sensor complementarity.

Sentinel-1 + Sentinel-2 (Hypercomplex Fusion): A multi-modal fusion setup that integrates structural (SAR) and spectral (optical) information usinghypercomplex algebraic framework. Operates in a mono-temporal regime to isolate fusion effects without time-series input.

Sentinel-2: A spectral-only baseline assessing the predictive power of optical data, including a Kennaugh-like transformation of Sentinel-2 inputs to maintain comparability. Represents a mono-date, mono-sensor reference.

Sentinel-1: A radar-only baseline that examines model performance under mono-temporal conditions using standard Sentinel-1 polarimetric inputs. Highlights structural sensitivity absent in optical datasets.

TerraSAR-X and ALOS-2: A cross-sensor benchmark comparing Sentinel-1 with alternative high-resolution SAR systems, focusing on the influence of SAR-specific acquisition characteristics such as resolution, incidence angle, and polarization mode. Operates in mono-temporal configurations.

The results presented in this thesis suggest that predictive performance in EO-based, e.g., forest parameter modelling, does not depend solely on increasing model complexity or relying on any single sensor modality. Instead, it appears that accuracy and generalization are most effectively achieved through thoughtful combinations of diverse data types and modelling strategies. The analyses reported in the foundational benchmarking chapter support the view that multi-modal and multi-temporal fusion, when paired with robust ensemble learning, can substantially elevate the predictive capacity of remote sensing workflows.

The study did not treat modality, preprocessing, and modelling architecture as isolated parameters, but rather investigated their interaction through a structured experimental design. By evaluating over 500 unique configurations across different feature types, filtering regimes, and learning algorithms, the work identified which combinations support both high-fidelity local predictions and robust spatial transfer. The consistently strong performance of SAR features for vertical forest metrics, the improved generalization introduced by spectral and structural transformations, and the stability gained from temporal stacking all point to the value of compositional rather than singular design logics in EO modelling.

The findings further indicate that multi-temporal integration contributes not only additional information but also functional redundancy. This redundancy proved instrumental

in mitigating the impact of temporal acquisition noise and label aging, particularly in ecologically diverse landscapes. Similarly, stacked ensemble methods, especially those based on RF, offered a practical and interpretable mechanism to integrate spatially diverse model predictions, outperformed more specialised or deeper alternatives under most transfer scenarios. These ensembles consistently improved performance under domain shift, suggesting their applicability to operational contexts where training data may be sparse, misaligned, or outdated. These findings challenge the prevailing emphasis on ever-deeper models and point instead to the importance of input design and supervision fidelity.

From a methodological perspective, the importance of data preprocessing emerged as a recurrent theme. Conservative filtering thresholds, outlier mitigation, and ecologically motivated transformations were shown to improve both accuracy and stability across experiments. This reinforces the notion that attention to data structure, rather than model tuning alone, is critical when developing predictive pipelines in EO.

Taken together, the benchmarking framework developed in this thesis provides an empirical foundation for evaluating the trade-offs between model specificity, data richness, and operational feasibility. It also demonstrates that scalable, generalizable EO–ML systems are achievable, provided that the design of inputs, fusion strategies, and evaluation logic is guided by both ecological relevance and computational discipline. The systematic evaluation of EO modality–model interactions thus confirmed a core hypothesis of the thesis: that predictive performance and ecological validity in EO–ML are more strongly governed by how features are constructed and aligned than by model choice alone. This insight underpins the subsequent development of temporally aware indices and contextually enriched labels, forming the foundation for a structurally coherent EO–ML pipeline.

This research highlights the potential of structured EO fusion, in both the spectral–structural and temporal domains, as a foundational component of next-generation environmental monitoring systems. The predictive capacity of remote sensing, it is argued, can be significantly enhanced when modelling pipelines are treated not as fixed algorithms but as modular, ecologically aligned systems. This perspective invites further exploration of how such systems might evolve toward higher autonomy, spatial scalability, and decision-support relevance in forest management and beyond.

8.3 Temporal Fusion Strategies and Design

A central proposition of this thesis was that temporal fusion in EO–ML pipelines should not be reduced to the simple stacking of index sequences. Rather, time must be treated as an ecological axis, structured, asymmetric, and often divergent across sensor modalities. This re-framing guided the development of fusion strategies that go beyond conventional time-series modelling, anchoring them instead in the ecological logic of the processes under observation.

Chapter 5 of this thesis investigated how temporal heterogeneity in sensor responses could be harnessed to design more informative and contextually aligned feature representations. This involved differentiating between *intra-seasonal* and *cross-seasonal* fusion strategies. Intra-seasonal fusion emphasises temporal densification within a defined phenological phase, e.g., capturing variability during the peak vegetation period. By contrast, cross-seasonal fusion intentionally combines acquisitions from ecologically distinct periods, such as dry-season SAR observations with peak-season optical data. While the former approach supports high-resolution trend detection, the latter introduces functional complementarity into the feature space.

This principle was operationalised through the development of the CDVI, a bi-temporal metric designed to quantify structural and phenological divergence across SAR and optical sensors. CDVI does not assume that change is symmetric or monotonic; instead, it formalises ecological non-equivalence as an informative feature. For example, in karstic landscapes affected by sinkholes and soil moisture anomalies, Sentinel-1 imagery acquired during the dry season captures subsurface and structural features invisible to optical sensors. When paired with peak-vegetation Sentinel-2 data, this contrast reveals latent ecological heterogeneity that is otherwise masked by surface greenness.

The sinkhole mapping case study exemplifies this logic: dry-season SAR inputs identified sub-surface depressions and soil instability, while concurrent optical data captured overlying vegetation health. Their divergence, quantified via CDVI, yielded a fused representation more closely aligned with geomorphological risk patterns than either modality alone. Importantly, this was not a case of mere temporal averaging, but of targeted temporal contrast.

More broadly, these findings reinforce a central claim of this thesis: that *time in EO is not a neutral axis*. It is an ecological signal in itself, carrying phase-specific, process-sensitive information that, if ignored or collapsed into averaged indices, leads to the erosion of

contextual interpretability. Designing temporal fusion strategies must therefore involve decisions about when, not just how often, sensors are queried.

This view challenges the prevalent assumption that denser time-series necessarily yield automatically better models. Rather, it suggests that temporally sparse but ecologically complementary observations may be valuable, especially when fused across modalities and aligned with structural changes in the target environment. In this way, temporal fusion becomes a modelling decision rather than a passive input accumulation.

Finally, the temporal design logic established here feeds directly into downstream benefits observed in predictive performance and interpretability. Whether in ecological anomaly detection or structural forest mapping, temporally aware features, particularly bi-temporal indices such as the CDVI, supported not only more accurate but also more traceable modelling outcomes. This points to an important future direction: the explicit integration of phenological and ecological reasoning into the architectural design of EO–ML systems.

8.4 Structuring Supervision - The HELIX Framework for Label Enrichment

A key insight developed through the research aim of this thesis is that in EO–ML pipelines, label design often constitutes a greater bottleneck than model architecture. While substantial attention is typically paid to refining input features and tuning learning algorithms, the quality, structure, and semantics of supervision remain under-addressed. This oversight limits not only the accuracy of predictions but also their ecological validity and generalisation capacity.

The HELIX framework, introduced and operationalised across Chapters 3 and 4 as well as Chapter 7 was developed to address this challenge. It offers a systematic approach to label enrichment, embedding spatial, temporal, and epistemic descriptors into the supervision layer of EO–ML models. Rather than treating labels as fixed scalar values, HELIX reconceptualises them as context-aware, process-aligned constructs that more faithfully reflect environmental dynamics.

Spatially, HELIX integrates local multi-scale statistics into the label structure. In the bark beetle outbreak case study, raw disturbance labels were augmented with descriptors

such as local mean, and variance across varying neighbourhood sizes. This provided a structured account of spatial autocorrelation and heterogeneity, enabling the learning model to distinguish between isolated noise and ecologically coherent patterns of spread. This form of enrichment was crucial in disentangling genuine ecological signals from artefacts introduced by classification errors or cloud masking.

Temporally, HELIX employs convolutional kernel filters to extract phase-sensitive dynamics from EO time series. In the glacier change analysis, raw elevation differences were transformed into temporal descriptors encoding onset timing, persistence, and rate-of-change signatures. These derived features captured not just the magnitude but the trajectory of cryospheric processes, allowing the model to differentiate between transient artefacts (e.g., seasonal snow) and structurally meaningful glacier change. This temporal logic extended the representational capacity of supervision beyond static snapshots.

Epistemically, HELIX introduces residual-based descriptors derived from prior model fits. By analysing *where* and *how* predictions diverge from ground truth, the framework quantifies model uncertainty in an operationally meaningful way. These residuals are not treated merely as error but as *feedback signals*, indicating areas of label misalignment, signal ambiguity, or structural model bias. This feedback loop allows for iterative supervision refinement, closing the gap between training objectives and ecological process understanding.

Collectively, these design elements shift the role of supervision from a passive target to an active component of model design. Supervision, in this enriched form, encodes domain structure, ecological dynamics, and epistemic uncertainty. It becomes a co-evolving layer in the EO–ML system, one that adapts not only to changing inputs but also to the interpretive goals of the model. This structural view challenges the conventional input–label dichotomy and calls for supervision-aware architectures that treat learning as a joint optimisation of representation, context, and feedback.

The HELIX framework thus contributes a practical methodology and a conceptual lens through which supervision can be re-engineered to better reflect the complexities of environmental monitoring. Its applications in this thesis demonstrate that model performance, interpretability, and transferability all benefit from supervision that is not only accurate but structurally informative. As EO–ML systems scale toward broader operational deployment, such enriched supervision strategies may prove indispensable in bridging the gap between data abundance and decision relevance.

8.5 Interpretation of Kennaugh Elements

The varied roles and effects of Kennaugh elements and Hadamard-based HCB framework across the experimental pipelines in this thesis reveal both the promise and complexity of multi-modal EO for environmental inference. By exploring five distinct Kennaugh families, (1) polarimetric, (2) spectral (Kennaugh-like), (3) fused spectral–polarimetric, (4) temporal–polarimetric, and (5) spectral–polarimetric–temporal, within this thesis, it demonstrates how different fusion strategies interact with model architectures and environmental attributes. Each family offers a unique epistemic window into the landscape, contributing differentially to generalisation, interpretability, and ecological relevance.

Spectral and Polarimetric Kennaugh Elements

The application of Kennaugh elements, both polarimetric (SAR-based) and spectral (optical-based), proved instrumental in enhancing the interpretability and effectiveness of remote sensing-based forest parameter modelling. Across the experimental setups, these feature representations supported the extraction of meaningful and transferable information from Sentinel-1 and Sentinel-2 observations, respectively, with tangible benefits for a range of forest structural attributes.

Polarimetric Kennaugh elements from Sentinel-1 C-band SAR provided physically grounded descriptors of canopy structure and scattering behaviour. The total backscatter intensity (k_0), polarization contrast (k_1), and the real and imaginary parts of the correlation (k_5, k_8) enabled nuanced characterizations of canopy density, surface roughness, and volume scattering. These properties were especially effective in modelling variables such as crown area, tree count, and mean tree height, where radar's sensitivity to canopy penetration and dielectric contrast proved advantageous. The relatively stable spatial generalization performance of these elements, particularly in height and count estimations, also underlined their value for operational forest monitoring, especially in persistently cloud-covered regions where optical data is limited.

Spectral Kennaugh-like elements, derived through hypercomplex decomposition of Sentinel-2 reflectance bands, enabled a structured and orthogonal reformulation of optical signals. By separating overall brightness from chromatic contrasts, this representation emphasized physiologically meaningful variation in vegetation (e.g., biomass-related reflectance vs. species-driven spectral contrast). In tasks involving crown area prediction

and vegetation coverage estimation, these transformed features consistently outperformed or matched the predictive accuracy of raw spectral bands. Moreover, the improved robustness of spectral Kennaugh-like features during spatial transfer suggested enhanced generalizability due to reduced spectral redundancy and collinearity.

Crucially, both types of Kennaugh elements facilitated the construction of a harmonized, modality-agnostic feature space suitable for later fusion scenarios. Their shared mathematical structure and interpretability support transparent modelling workflows, reduce dependence on domain-specific preprocessing, and provide a consistent basis for integrating optical and SAR data. Taken together, the use of Kennaugh elements not only improved model performance but also fostered a deeper understanding of how different EO modalities encode forest structural traits, making them a valuable asset for scalable, interpretable, and transferable forest monitoring systems.

Spectral–Polarimetric Fusion

The fusion of Sentinel-1 polarimetric and Sentinel-2 spectral information through hypercomplex bases (HCB) yielded a semantically rich and mathematically orthogonal eight-dimensional feature space. By combining structurally interpretable SAR Kennaugh elements (k_0, k_1, k_5, k_8) with their spectral counterparts (transformed bands B2, B3, B4, B8), the fused representation captured shared and complementary information across sensor modalities. Each resulting channel within the fused vector $K_{\text{fused},0-7}$ corresponds to a unique combination of backscatter physics and spectral reflectance properties, providing enhanced input semantics for modelling forest structure.

From a theoretical standpoint, the fusion transformation, based on the Hadamard matrix Q, ensures orthogonality and lossless integration. It preserves modality-specific information while simultaneously enhancing contrast, structural interpretability, and noise robustness. Empirically, the fused Kennaugh elements demonstrated improved regression performance across key vegetation parameters, particularly in crown area, tree counts, and height metrics. Models trained on fused features consistently outperformed those using only SAR or optical inputs, with the most notable gains observed in:

- **Tree counts and height estimation**: The fused representation improved mean absolute errors (MAE) across deciduous and coniferous tree counts and mean tree height, benefiting from both the vertical sensitivity of SAR and the biochemical differentiation of optical data.

- **Robustness under spatial transfer**: While domain shifts still caused performance degradation, fused models exhibited better resilience compared to unimodal baselines, especially for vertical structural variables.
- **Semantic diversity**: Each $K_{\text{fused},i}$ channel was empirically linked to distinct physical or ecological properties, such as NDVI sensitivity, edge structure, habitats, supporting interpretable and variable-specific learning.

Despite these advantages, crown volume and summed crown area remained sensitive to spatial heterogeneity, showing only limited generalization gains under fusion. This suggests that, while hypercomplex data fusion enhances the information basis for modelling, complex canopy metrics remain intrinsically harder to capture without localized calibration.

The spectral–polarimetric Kennaugh fusion delivers a compact, interpretable, and orthogonal feature space that strengthens EO-based modelling of forest structure. It aligns with operational goals of scalability and transferability, enabling more resilient predictions across diverse landscapes. Its structure-aware design provides both physical interpretability and statistical robustness, positioning it as a foundational feature representation for multi-sensor forest monitoring, as shown in this thesis.

The interpretative profiles of the fused Kennaugh elements are grounded both in the theoretical framework of the hypercomplex fusion design and empirical findings from this thesis. Supporting evidence includes correlation analyses with key vegetation indices (NDVI and NDWI), habitat classifications from the Bavarian Forest National Park dataset [203], and spatial texture patterns characterized by GLCM metrics. The habitats considered span a diverse range of ecological types, including mature coniferous and deciduous forests, mixed stands, shrublands, meadows, wetlands, rocky outcrops, and anthropogenic areas such as residential zones, roads, and clear-cuts. This comprehensive habitat spectrum enables robust evaluation of the fused feature space across natural and human-influenced landscapes.

Table 8.1.: Interpretative summary of fused Kennaugh elements $K_{\text{fused},0}$ to $K_{\text{fused},7}$, derived from spectral–polarimetric fusion. Interpretations integrate symbolic formulations, NDVI/NDWI correlations, texture metrics (GLCM), and observed habitat contrast patterns. Note: high average intensity does not imply class separability; interpretations emphasize discriminative contrast and statistical evidence.

Band (Formula)	Statistical Pattern	Habitat Contrast Pattern	Interpretation
$K_{\text{fused},0}$ $(k_0 + B_2 + B_3 + B_4 + B_8)$	Very high GLCM contrast; moderate NDVI/NDWI correlation	Strong separation between bare/urban vs vegetated classes	Texture-dominant band capturing surface roughness and spatial heterogeneity. Reflects total intensity but has moderate vegetation/water specificity. Useful for heterogeneous land cover.
$K_{\text{fused},1}$ $(k_1 + B_2 - B_3 + B_4 - B_8)$	Moderate contrast; moderate NDVI/NDWI correlation	Separates urban edges, ecotones	Edge-sensitive band highlighting spectral–polarimetric imbalances. Detects fragmented landscapes and transitions between natural and anthropogenic zones.
$K_{\text{fused},2}$ $(k_5 + B_2 + B_3 - B_4 - B_8)$	Moderate texture; strong NDVI/NDWI correlation	High contrast across meadows, shrublands, and early growth types	Vegetation-sensitive band. Captures mid-biomass states and moisture gradients. Valuable for early phenological stages or disturbed vegetative cover.
$K_{\text{fused},3}$ $(k_8 + B_2 - B_3 - B_4 + B_8)$	Moderate contrast; strong NDVI/NDWI correlation	Relatively uniform across most classes	Supplementary vegetation/moisture band. Adds depth to structural interpretations but limited class separability.
$K_{\text{fused},4}$ $(k_0 - B_2 - B_3 - B_4 - B_8)$	Very low contrast; very strong NDWI/NDVI correlation	High separability of water bodies and wet habitats	Hydrologically responsive band. Excels in delineating aquatic features and consistently saturated zones due to its strong spectral–polarimetric suppression of vegetation.
$K_{\text{fused},5}$ $(k_1 - B_2 + B_3 - B_4 + B_8)$	Very low contrast; strong NDVI/NDWI correlation	Separates riparian, wetland fringes from dry zones	Gradient-sensitive band. Responds to transitional moisture conditions at water–land boundaries. Enhances wetland classification.
$K_{\text{fused},6}$ $(k_5 - B_2 - B_3 + B_4 + B_8)$	Very low contrast; strong NDVI/NDWI correlation	Differentiates vegetated and semi-vegetated covers (e.g., meadows)	Fine-scale ecological gradient band. Sensitive to canopy layering and terrain-induced vegetation variation.
$K_{\text{fused},7}$ $(k_8 - B_2 + B_3 + B_4 - B_8)$	Moderate contrast; weaker NDVI/NDWI correlation	Separates manmade structures from natural covers	Anthropogenic texture band. Detects grid-like, linear, and artificial patterns (e.g., roads, roofs).

1. $K_{\textbf{fused},0}$ Represents the total radar backscatter and broad optical reflectance intensity. It is highly sensitive to surface texture and heterogeneity, capturing structural complexity such as canopy roughness, bare soil, and urban surfaces. Although its vegetation correlation is moderate, it contains rich information for canopy biomass and overall surface brightness. Strongly separates vegetated and non-vegetated

regions, particularly in heterogeneous landscapes. Based on GLCM contrast (64) and NDVI/NDWI correlations ($\rho \approx \pm 0.6$).

2. $K_{\mathbf{fused},1}$ Highlights transitions between land cover types, especially human–natural boundaries. Moderate GLCM contrast and NDVI/NDWI correlations indicate its sensitivity to edge zones, roads, ecotones, and fragmented land use. Though not highly specific to vegetation, it provides valuable contextual detail. Shows contrast-based class separability in urban margins.

3. $K_{\mathbf{fused},2}$ Encodes mid-level biomass and spectral–polarimetric vegetation patterns. Strong correlations with NDVI/NDWI ($\rho \approx \pm 0.65$) and moderate texture suggest suitability for discriminating meadows, scrub, and early growth stages. Supports detection of moisture-linked vegetative stress and seasonal gradients.

4. $K_{\mathbf{fused},3}$ Complementary to $K_{\text{fused},2}$, this band contributes to vegetation structure interpretation but shows lower class-level separability. NDVI/NDWI correlations remain strong, but moderate texture and habitat uniformity suggest a supporting rather than leading role in classification.

5. $K_{\mathbf{fused},4}$ Dominated by spectral–polarimetric suppression, it yields very low GLCM contrast and extremely high NDWI correlation ($\rho \approx +0.94$). Most effective for detecting water bodies and persistently moist substrates, with strong contrast against dry or vegetated areas. Statistically the most discriminative hydrological band.

6. $K_{\mathbf{fused},5}$ Closely aligned with $K_{\text{fused},4}$, but with slightly shifted spectral–polarimetric balance. Useful for characterizing transitional zones like wetland edges, showing strong NDWI correlation and uniform texture. Enhances mapping of moist gradients near water–land interfaces.

7. $K_{\mathbf{fused},6}$ Highly homogeneous with strong NDVI/NDWI correlation ($\rho \approx \pm 0.94$). Best suited for detecting subtle changes in vegetative layering, moisture variation, and topography–vegetation interactions. Differentiates semi-open meadows and lightly forested areas from denser covers.

8. $K_{\mathbf{fused},7}$ Displays moderate texture and relatively weak NDVI/NDWI correlation. Primarily responds to anthropogenic features, roads, roofs, agricultural patterns, rather than natural ecological gradients. Useful for mapping manmade structures or patterned land cover.

Each of these fused elements contributes a semantically distinct descriptor to the eight-dimensional HCB feature space. Their design ensures mutual orthogonality, allowing downstream models to leverage uncorrelated and interpretable axes of information. From a landscape perspective, they provide a compact yet information-rich basis for classification, structural estimation, and multi-temporal analysis.

Spectral–Polarimetric–Temporal Fusion

Across time, 64 such fused images are aligned and decomposed using a 64×64 Hadamard matrix applied along the temporal axis. This yields a 512-dimensional feature vector per pixel, comprising temporally transformed versions of each of the 8 fused channels. The result is a compact yet information-rich structure that captures not only spectral and polarimetric content, but also the temporal evolution of those features, disentangling stability, gradual change, and episodic variation within a shared space.

From an interpretive perspective, the temporal decomposition acts as a frequency-like filter bank. Its orthogonal basis vectors resemble low-frequency and high-frequency wavelet components, enabling the disentanglement of complex temporal signals embedded in the fused spectral–polarimetric data. Each resulting temporal mode exhibits a distinct semantic signature:

- $\mathbf{tK_0} \sim [+,+,+,+]$: Captures the temporal mean of each feature, effectively summarizing persistent surface properties that are stable over the annual cycle. This includes dense canopy cover, consistent soil reflectance, and urban impervious surfaces. Because it aggregates over time, it is robust to transient noise and sensor variability, providing a reliable baseline characterization of the landscape.
- $\mathbf{tK_1} \sim [+,-,+,-]$: Isolates short-term oscillations with a characteristic alternating pattern. This mode is sensitive to rapid environmental fluctuations such as vegetation phenology on a weekly or monthly scale, soil moisture changes driven by rainfall events, and transient atmospheric effects like clouds impacting optical data. The alternation in sign allows it to emphasize contrasts between successive acquisitions, highlighting ephemeral changes.
- $\mathbf{tK_2} \sim [+,+,-,-]$: Reflects longer-term seasonal gradients and inter-half-year trends. It captures the progression of phenological stages such as leaf emergence and senescence, gradual degradation from stressors like pests or drought, and other

slow-moving ecological transitions. This mode facilitates monitoring ecosystem health and productivity shifts across seasons.

- $\mathbf{tK_3} \sim [+, -, -, +]$: Amplifies episodic or localized events, including harvesting, insect outbreaks, fire scars, and snow-melt episodes. Its pattern enables it to highlight spatially or temporally confined anomalies against the background variability.

This decomposition's resemblance to Haar wavelets and discrete Fourier transform underscores its role as a compact, orthogonal basis well suited to capturing multi-scale temporal dynamics.

The first four temporal Hadamard components, $\mathbf{tK_0}$ through $\mathbf{tK_3}$, represent the dominant modes of temporal variability in the fused feature space. The zero-frequency component $\mathbf{tK_0}$ encodes the mean reflectance or backscatter over the entire time series, providing a stable baseline characterization of surface properties. The subsequent components act as temporal filters, progressively capturing higher-frequency signals: $\mathbf{tK_1}$ isolates short-term oscillations and abrupt changes; $\mathbf{tK_2}$ corresponds to seasonal trends and mid-term shifts; while $\mathbf{tK_3}$ highlights localized or transient events.

Beyond these primary components, higher-order temporal modes ($\mathbf{tK_4}, \mathbf{tK_5}, \ldots$) capture increasingly subtle and complex temporal variations, including fine-scale phenological shifts, irregular disturbance events, and noise patterns, thereby enriching the temporal resolution and expressive power of the fused feature space.

The orthogonality ensures minimal redundancy among modes, facilitating clearer statistical separation of overlapping phenomena, such as distinguishing between seasonal phenology and disturbance events, even when these occur simultaneously. Because the transformation is applied to a physically grounded, fused spectral–polarimetric representation, the temporal components retain coherent semantic meaning across sensor modalities, supporting interpretable change detection and robust ecological inference.

Empirically, the model results presented in Sections 7.1 and 6.1 confirm the expressive potential of this representation. Performance improvements across forest structural variables, including mean tree height, deciduous tree count, crown area, and volume, validate the utility of combining spectral, polarimetric, and temporal information in this form. Notably, feature importance analysis revealed a striking flattening of contribution scores: no single date or fused band dominated model performance. Instead, predictive power was distributed broadly across time and channels, highlighting the robustness and expressiveness of the fused temporal representation. This balance reflects the dual

design intent: redundancy across time reduces the impact of noise or artifacts, while orthogonality ensures that change-related signals remain distinguishable. As a result, the fused cube enables the model to simultaneously learn static traits (e.g., canopy extent), gradual shifts (e.g., phenological curves), and unexpected transitions (e.g., damage or clearing).

However, challenges remain. Crown volume and other context-sensitive metrics continued to suffer from spatial domain shift, suggesting that temporal structure, while helpful, cannot fully resolve spatial generalization barriers. Still, the temporal fusion substantially improved resilience, particularly for features related to vegetation height, type, and density, across highly varied AOIs.

The fusion pipeline developed here integrates spectral, polarimetric, and temporal signals into a unified, orthogonally transformed 512-dimensional space. It preserves pixel-level integrity, enhances interpretability, and distributes predictive capacity across multiple fused channels. These properties make it a strong candidate for downstream modelling pipelines, especially when extended with attention mechanisms, uncertainty quantification, or domain-adaptive strategies for large-scale environmental inference.

Temporal–Polarimetric Fusion

This experiment explored the use of multi-temporal Sentinel-1 SAR acquisitions for glacier monitoring through the lens of polarimetric decomposition. Specifically, four Kennaugh elements, k_0, k_1, k_5, and k_8, were derived per acquisition and evaluated for their ability to reveal consistent spatial and temporal patterns across glaciated terrain. The aim was to assess whether stacking these polarimetric descriptors over time could enhance surface characterization and glacier boundary delineation, especially in regions where optical data are unreliable due to cloud cover or seasonal snow.

Among the four elements, only k_0, the total backscatter power, emerged as a consistently useful and interpretable signal. It exhibited strong spatial contrast between glacier tongues, accumulation zones, and adjacent terrain, and its values remained robust across acquisition dates. Tongue regions and crevassed zones were characterized by high k_0 intensities, reflecting radar-bright surfaces with high roughness or dielectric contrast. In contrast, accumulation areas and smoother snow-covered slopes showed persistently low k_0 values, consistent with radar-dark, volume-scattering-dominated returns. These

patterns were stable across the time series, enabling temporally reinforced interpretation of glacier morphology.

The remaining polarimetric elements, k_1, k_5, and k_8, showed limited utility in this context. While they theoretically capture polarization asymmetry, correlation, and phase relationships, they proved less spatially consistent and more noise-prone over glacier surfaces. Their contribution to class separation or structural interpretation was negligible compared to k_0, and no clear temporal enhancement was observed when stacking them. This outcome underscores the dominance of intensity-based contrasts in glacier SAR interpretation at C-band and supports the use of reduced polarimetric models in operational settings.

From a temporal perspective, k_0 allowed for capturing subtle changes indicative of melt processes, surface transitions, or radar incidence effects. Localized changes in backscatter over time highlighted evolving surface roughness or wetness, particularly in tongue regions. However, the majority of the glacier signal remained stable, suggesting that multi-temporal k_0 does more to reinforce spatial segmentation than to reveal short-term dynamics. Still, this temporal stacking improved robustness by mitigating single-scene noise and emphasizing persistent structures.

In conclusion, while full polarimetric decomposition was applied, only k_0 consistently delivered interpretable and stable results for glacier monitoring. Its temporal fusion enhanced glacier delineation and surface characterization, making it a valuable radar-based proxy under data-sparse or cloud-obscured conditions. The findings support the prioritization of k_0 in future large-scale or automated SAR-based glacier observation frameworks, particularly where simplicity, interpretability, and temporal robustness are desired.

8.6 Limitations and Methodological Caveats

While the thesis presents a series of methodological advancements and demonstrates the value of structurally enriched EO–ML pipelines, several limitations and caveats must be acknowledged to contextualise its findings.

First, although multi-modal and multi-temporal fusion strategies yielded consistent gains in accuracy and generalisation, the approach remains constrained by the availability, alignment, and quality of input data. Sentinel-based systems provide globally consistent

coverage, but their spatial, spectral, and temporal characteristics impose limits on the resolution and specificity of ecological phenomena that can be captured. For example, SAR-derived vertical structure proxies may be affected by terrain-induced distortions or moisture variability, while optical features are susceptible to cloud contamination, atmospheric interference, and phenological mismatches. Despite extensive preprocessing and transformation, such noise sources cannot be fully eliminated and may influence model outputs in subtle but consequential ways.

Second, the challenge of temporal misalignment between input features and reference labels emerged as a recurrent limitation in both the forest structure and outbreak forecasting contexts. In the forest structure analysis, models trained on EO data from 2020–2021 were supervised using inventory-derived labels from 2016–2018, introducing an inherent discrepancy between observed canopy conditions and ground-truth structure. Similarly, in the HELIX-based modelling framework, historical bark beetle outbreak polygons were used to construct spatial descriptors that served as targets for prediction or forecasting using more recent EO imagery. Although ensemble learning and temporally enriched input features helped mitigate these discrepancies, such forms of lagged supervision nonetheless introduced epistemic uncertainty, particularly in dynamic or disturbance-prone landscapes where rapid ecological change can outpace static labels. The HELIX framework partially addressed this by smoothing sparse, binary labels into continuous descriptors that encode temporally diffuse spatial structure, effectively allowing models to learn from historical dynamics embedded in current EO signals. This transformation from sparse labels to enriched, learnable descriptors reduced the impact of label aging by incorporating both lagged and co-temporal outbreak information. However, the broader challenge remains: as EO–ML systems increasingly aim for operational deployment, particularly in forecasting and early warning applications, the reliance on temporally misaligned or outdated supervision becomes a bottleneck. These findings suggest that future pipelines should prioritise dynamic label updating strategies and explicitly model temporal uncertainty, either through uncertainty-aware loss functions, spatio-temporal regularisation, or generative supervision frameworks like HELIX that enable learning from context rather than direct correspondence. Addressing temporal supervision misalignment is likely to be a key enabler for achieving resilient, generalisable EO-based ecological monitoring systems.

Third, the modular benchmarking framework used in the foundational analysis, while systematic, represents a controlled environment that does not capture all operational constraints. The experiments were designed to isolate specific variable interactions,

such as the influence of modality or preprocessing, under idealised conditions. In real-world deployments, additional constraints such as data latency, sensor availability, computational limits, and integration with policy frameworks may shape pipeline design in non-trivial ways. Thus, while the findings are informative at a methodological level, their operational translation must be evaluated case-by-case.

Fourth, although RF and their ensembles performed consistently well, this choice also imposed certain architectural constraints. RF, by design, lack native mechanisms for capturing sequential or spatial autocorrelation beyond input feature engineering. While HELIX and multi-date inputs compensated for this to some extent, DL architectures, if appropriately regularised and interpretable, may offer additional capacity for learning spatio-temporal structure directly. Their inconsistent performance in this study suggests a need for further work on training regimes, interpretability tools, and data augmentation strategies to unlock this potential.

Finally, this thesis focused predominantly on pixel-level modelling strategies. While this granularity offers precision and flexibility, it also exposes the system to residual uncertainties related to geolocation accuracy, sensor resolution mismatch, and ecological misalignment between pixels and reference units (e.g., plot or stand scale). Aggregation strategies or object-based approaches may help mitigate these discrepancies and should be further explored as complementary paths to robust inference.

In summary, the methodologies developed and validated here show strong promise for structurally aligned EO–ML, yet their broader adoption requires careful attention to sensor limitations, label dynamics, operational constraints, and representational granularity. Acknowledging these caveats ensures that the thesis remains both critically grounded and open to iterative improvement in future research and practice.

8.7 Future Work and Research Directions

The methodologies and findings presented in this thesis suggest several avenues for future research, aimed at enhancing the scalability, interpretability, and operational relevance of EO–ML systems. While many components, such as fusion design, descriptor construction, and ensemble learning, have reached a high level of methodological maturity, their integration into dynamic, policy-aligned, and adaptive systems remains an open challenge.

Four thematic directions are outlined below, each addressing critical next steps for the evolution of structurally aligned EO–ML pipelines.

HELIX Automation and Probabilistic Descriptors

The HELIX framework introduced a generalisable logic for constructing contextually enriched supervision signals from sparse and often binary ecological labels. Yet its implementation still required manual descriptor design, thresholding choices, and domain-specific intuition. Future work should aim to formalise and automate these steps, potentially through meta-learning, generative modelling, or Bayesian descriptor inference. Embedding true probabilistic structure into HELIX descriptors would allow the pipeline not only to express confidence intervals over supervision but also to reflect epistemic uncertainty in sparse-label regimes. Such probabilistic descriptors could, in turn, inform model calibration, active learning loops, and uncertainty-aware decision-making.

Dynamic Labels and Model–Label Co-Design

A recurring insight across experiments was the fragility of EO–ML performance under static or lagged supervision. This suggests that more adaptive forms of label construction, such as rolling updates from inventory, participatory monitoring, or feedback from predictive residuals, may be essential for resilient ecological forecasting. Future EO–ML pipelines could benefit from co-designed supervision architectures, where models not only learn from labels but contribute to their refinement. This would require the integration of model diagnostics, temporal anomaly detection, and even causal feature attributions into the supervision workflow, turning labels from fixed inputs into co-evolving elements of the modelling system. The interplay between model structure and label design thus emerges as a critical and under-explored axis of system-level performance.

Policy Alignment and SDG Integration

To translate methodological advances into societal impact, future EO–ML systems must align more directly with policy agendas and sustainability monitoring frameworks. In particular, the modelling logics and output semantics developed in this thesis could be mapped onto key indicators within the Sustainable Development Goals (SDGs), such as

forest cover change (SDG 15.2), disaster risk (SDG 13.1), or biodiversity status (SDG 15.5). However, such integration would require not only technical interoperability, e.g., spatially explicit, scalable, and transparent predictions, but also the capacity to interface with institutional workflows. This includes uncertainty quantification, scenario analysis, and explainability features that allow outputs to be interpreted within decision-making contexts. Research into standardised data contracts, indicator-specific model adaptation, and stakeholder feedback mechanisms will be essential for operationalising EO–ML within policy cycles.

Edge-AI and Causal Inference

As RS data volumes continue to grow, computational and energy constraints will increasingly shape the viability of modelling approaches. One promising direction is the deployment of lightweight, explainable models directly on satellite or ground-based sensor platforms, enabling so-called edge AI for near-real-time processing and alert generation. At the same time, the complexity of ecological systems demands stronger causal reasoning capabilities. Existing EO–ML models primarily capture statistical associations; future systems must advance toward encoding causal structures, testing counterfactuals, and differentiating between drivers and correlates of observed phenomena. Advances in causal discovery, graph-based learning, and simulation-based inference may help bridge this gap, enabling EO–ML systems to transition from predictive accuracy to mechanistic insight.

Importantly, the modelling strategies developed in this thesis intentionally prioritized lightweight, interpretable architectures over resource-intensive DL alternatives. Approaches such as delta regression, HELIX-style label-side temporal enrichment, and residual-based refinement were designed to balance predictive performance with computational efficiency and transparency. This design logic stands in deliberate contrast to more opaque, energy-intensive solutions like LSTMs or large transformer-based time-series models. By focusing on causally sound, label-driven temporal structuring, rather than brute-force sequence learning, this work demonstrates that robust EO-based forecasting can be achieved with parsimonious models. This reduces both training costs and operational energy footprints, while enhancing interpretability and deployment potential, especially in resource-constrained or field-based settings.

On Model Complexity and Responsible EO–ML Design

A deliberate design philosophy underpins this thesis: that depth of learning architecture is not a universal proxy for scientific quality, generalizability, or ecological relevance in EO-based environmental modelling. While much of the current RS and ML literature gravitates toward increasingly deep, computationally intensive architectures, ranging from LSTMs to transformer-based sequence models, this work takes a counter-position. It argues that robust environmental inference is often less constrained by model complexity and more by data structure, label quality, and the ecological coherence of inputs and targets.

Heavyweight models applied to noisy, misaligned, or weakly structured EO datasets risk amplifying artefacts, learning sensor-specific noise, and obscuring causal relationships behind layers of parametrized abstraction. In contrast, the lightweight, interpretable, and label-enriched architectures deployed here, including delta-state regression, HELIX-informed supervision, and residual-based refinement, prioritize causal integrity, computational efficiency, and interpretability without compromising predictive accuracy.

This modelling stance is both technically and ethically motivated: computational responsibility demands that EO–ML systems minimize energy footprint where possible, especially given the climate-sensitive contexts they often aim to monitor. More fundamentally, environmental decision-making requires models that are not only performant but explainable, systems capable of offering not just answers, but reasoning pathways. By demonstrating that meaningful forecasting and spatial transferability can be achieved with lightweight, causally aligned methods, this thesis challenges the default assumption that "*more layers*" inherently equates to "*more insight*."

8.8 Final Reflections - Learning from and with EO

As EO matures from a data acquisition enterprise into a full-fledged system of environmental inference, the question shifts from whether we can observe change to how deeply we can understand it. This thesis has advanced a structural and methodological argument: that understanding environmental change through EO requires not just more data or deeper models, but more meaningful relationships between *what is sensed*, *how it is modelled*, and *what it represents*. Yet beyond this technical scaffolding lies a broader philosophical insight: that EO is not merely a monologue delivered from orbit, but a

dialogue between systems, between sensors and landscapes, between algorithms and ecosystems, and ultimately, between human and planetary knowledge.

To "*understand*" environmental change, in this vision, means more than identifying spatial patterns or predicting temporal trends. It involves situating observations within ecological processes, recognising causality where possible, and tracing uncertainty where necessary. The contributions of this thesis, particularly in label enrichment, temporal structuring, and interpretability, aim to move EO–ML closer to this ideal. By embedding ecological reasoning into the very architecture of learning, they invite models to reflect not only the surface of the Earth but also the logic of its transformations.

Interpretability, in this context, becomes more than a convenience for human users; it becomes a form of ecological literacy. A model that explains why it predicts a bark beetle outbreak or a shift in glacier zonation is not merely transparent, it is participatory. It allows stakeholders, scientists, and systems to learn with the model, not just from it. This orientation could redefine the role of EO in sustainability science: from that of a passive watcher to an active interlocutor, capable of supporting adaptive decision-making in complex and uncertain environments.

As the technological infrastructure of EO accelerates, through new sensors, faster processing, and larger archives, the demand for thoughtful, structure-aware, and ethically grounded modelling will only grow. The work presented here is but a step in that direction. It proposes that future EO–ML systems must not only be accurate and scalable but must also carry a deeper structural fidelity to the systems they aim to understand. Such systems will be judged not by how closely they match a validation set, but by how well they facilitate learning across domains, disciplines, and generations.

Ultimately, to learn from and with EO is to recognise that the Earth is not merely a dataset, but a dialogue partner. In this light, modelling becomes an act of listening, one that is technical, ecological, and profoundly human. And in keeping with Galileo's reminder that "*all truths are easy to understand once they are discovered; the point is to discover them,*" the journey of EO remains one of discovery, of both our planet and the ways we choose to model it.

Declaration of Supportive Resources

During the implementation of the algorithms presented in this thesis and throughout the writing process, various external software packages and tools were utilized to support both the technical development and the preparation of this document. The following section outlines the key resources used and their specific roles.

Software Packages and Toolboxes:

1. **SNAP** [101]: Used for Sentinel-1 data preprocessing. The processing chain included implementations based on [148].
2. **Anaconda** [16]: Provided the working environment for Python-based data analysis and algorithm development.
3. **Scikit-learn** [254, 236]: Used for implementing, training, and evaluating machine learning models.
4. **Numba** [200]: Used as a JIT (Just-In-Time) compiler to accelerate numerically intensive Python code and enhance processing performance.
5. **Joblib [176]**: Used for efficient parallel computing and caching during the execution of machine learning workflows.

Tools Supporting the Writing Process:

1. **DeepL** [80]: Utilized for translating selected sections of the text from German into English.
2. **Grammarly** [128]: Employed for grammar checking, spelling correction, and stylistic improvements to enhance clarity and readability.

3. **HAWKI ChatGPT** [243]: Provided by the AI Lab of Munich University of Applied Sciences and based on GPT-4 and GPT-4.o models. The language model was used exclusively for grammar, wording, and stylistic suggestions. Only self-authored text was provided as input, and all AI-generated suggestions were critically evaluated and selectively incorporated by the author.

All content, ideas, and conclusions presented in this thesis are solely the intellectual work of the author.

Bibliography

[1] T. A. W. Aaron E. Maxwell and F. Fang. Implementation of machine-learning classification in remote sensing: an applied review. *International Journal of Remote Sensing*, 39(9):2784–2817, 2018. doi: 10.1080/01431161.2018.1433343.

[2] S. Abdikan, F. B. Sanli, F. Sunar, and M. E. and. A comparative data-fusion analysis of multi-sensor satellite images. *International Journal of Digital Earth*, 7(8):671–687, 2014. doi: 10.1080/17538947.2012.748846. URL https://doi.org/10.1080/17538947.2012.748846.

[3] S. P. Abercrombie and M. A. Friedl. Improving the consistency of multitemporal land cover maps using a hidden markov model. *IEEE Transactions on Geoscience and Remote Sensing*, 54(2):703–713, 2016. doi: 10.1109/TGRS.2015.2463689.

[4] R. Aberle, E. Enderlin, S. O'Neel, C. Florentine, L. Sass, A. Dickson, H.-P. Marshall, and A. Flores. Automated snow cover detection on mountain glaciers using spaceborne imagery and machine learning. *The Cryosphere*, 19(4):1675–1693, 2025.

[5] K. Ahmadi, B. Kalantar, V. Saeidi, E. K. G. Harandi, S. Janizadeh, and N. Ueda. Comparison of machine learning methods for mapping the stand characteristics of temperate forests using multi-spectral sentinel-2 data. *Remote Sensing*, 12(18), 2020. ISSN 2072-4292. URL https://www.mdpi.com/2072-4292/12/18/3019.

[6] S. Aigner, S. Hauser, and A. Schmitt. Pattern-based sinkhole detection in arid zones using open satellite imagery: A case study within kazakhstan in 2023. *Sensors*, 25(3), 2025. ISSN 1424-8220. doi: 10.3390/s25030798. URL https://www.mdpi.com/1424-8220/25/3/798.

[7] V. Akbari, A. P. Doulgeris, and T. Eltoft. Monitoring glacier changes using multitemporal multipolarization sar images. *IEEE Transactions on Geoscience and Remote Sensing*, 52(6):3729–3741, 2014. doi: 10.1109/TGRS.2013.2275203.

[8] E. Akça, S. Aydemir, S. Kadir, M. Eren, C. Zucca, H. Günal, F. Previtali, P. Zdruli, A. Çilek, M. Budak, A. Karakeçe, S. Kapur, and E. A. FitzPatrick. *Calcisols and*

Leptosols, pages 139–167. Springer International Publishing, Cham, 2018. ISBN 978-3-319-64392-2. doi: 10.1007/978-3-319-64392-2_10. URL https://doi.org/10.1007/978-3-319-64392-2_10.

[9] K. M. Akhmedenov, D. Z. Iskaliev, and V. P. Petrishev. Karst and pseudokarst of the west kazakhstan (republic of kazakhstan). *Int. J. Geosci.*, 2014, 2014.

[10] S. Alam, M. S. Ayub, S. Arora, and M. A. Khan. An investigation of the imputation techniques for missing values in ordinal data enhancing clustering and classification analysis validity. *Decision Analytics Journal*, 9:100341, 2023. ISSN 2772-6622. doi: https://doi.org/10.1016/j.dajour.2023.100341. URL https://www.sciencedirect.com/science/article/pii/S2772662223001819.

[11] C. M. Albrecht, F. Marianno, and L. J. Klein. Autogeolabel: Automated label generation for geospatial machine learning. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 1779–1786, 2021. doi: 10.1109/BigData52589.2021.9672060.

[12] I. Ali, F. Greifeneder, J. Stamenkovic, M. Neumann, and C. Notarnicola. Review of machine learning approaches for biomass and soil moisture retrievals from remote sensing data. *Remote Sensing*, 7(12):16398–16421, 2015.

[13] M. Allen, F. Dorr, J. A. Gallego-Mejia, L. Martínez-Ferrer, A. Jungbluth, F. Kalaitzis, and R. Ramos-Pollán. Fewshot learning on global multimodal embeddings for earth observation tasks, 2023. URL https://arxiv.org/abs/2310.00119.

[14] C. I. Alvarez-Mendoza, D. Guzman, J. Casas, M. Bastidas, J. Polanco, M. Valencia-Ortiz, F. Montenegro, J. Arango, M. Ishitani, and M. G. Selvaraj. Predictive modeling of above-ground biomass in brachiaria pastures from satellite and uav imagery using machine learning approaches. *Remote Sensing*, 14(22), 2022. ISSN 2072-4292. doi: 10.3390/rs14225870. URL https://www.mdpi.com/2072-4292/14/22/5870.

[15] N. Amiri, P. Krzystek, M. Heurich, and A. Skidmore. Classification of tree species as well as standing dead trees using triple wavelength als in a temperate forest. *Remote Sensing*, 11(22), 2019. ISSN 2072-4292. doi: 10.3390/rs11222614. URL https://www.mdpi.com/2072-4292/11/22/2614.

[16] *Anaconda Software Distribution, Version 2-2.4.0*. Anaconda Inc., 2016. URL https://anaconda.com. Accessed: 2025-06-28.

[17] D. K. H. and. Remote sensing applications to hydrology; imaging radar. *Hydrological Sciences Journal*, 41(4):609–624, 1996. doi: 10.1080/02626669609491528. URL https://doi.org/10.1080/02626669609491528.

[18] R. Andersen. Nonparametric methods for modeling nonlinearity in regression analysis. *Annual Review of Sociology*, 35(1):67–85, 2009.

[19] V. Andrejchuk and A. Klimchouk. Mechanisms of karst breakdown formation in the gypsum karst of the fore-ural region, russia (from observations in the kungurskaja cave). implication of speleological studies for karst subsidence hazard assessment. *Int. J. Speleol*, (31):89–114, 2002.

[20] J. C. Angel, D. O. Nelson, and S. V. Panno. Comparison of a new gis-based technique and a manual method for determining sinkhole density: An example from illinois' sinkhole plain. *J. Cave Karst Stud.*, 66:9–17, 2004.

[21] P. M. Atkinson and A. R. L. Tatnall. Introduction neural networks in remote sensing. *International Journal of Remote Sensing*, 18(4):699–709, 1997. doi: 10.1080/014311697218700. URL https://doi.org/10.1080/014311697218700.

[22] Atlogis. TMS Data Service—CRS: 3857, 2016. URL http://www.atlogis.de. Available online.

[23] N. Audebert, B. Le Saux, and S. Lefèvre. Segment-before-detect: Vehicle detection and classification through semantic segmentation of aerial images. *Remote Sensing*, 9(4):368, 2017.

[24] M. Baatz and M. Schäpe. Multiresolution segmentation – an optimization approach for high quality multi-scale image segmentation. In J. Strobl, T. Blaschke, and G. Griesebner, editors, *Angewandte Geographische Informations-Verarbeitung XII*, pages 12–23. Wichmann Verlag, Karlsruhe, 2000.

[25] X. Bai, C. Liu, P. Ren, J. Zhou, H. Zhao, and Y. Su. Object classification via feature fusion based marginalized kernels. *IEEE Geoscience and Remote Sensing Letters*, 12 (1):8–12, 2015. doi: 10.1109/LGRS.2014.2322953.

[26] B. Balachander and D. Dhanasekaran. Comparative study of image fusion techniques in spatial and transform domain. *ARPN Journal of Engineering and Applied Sciences*, 11(9):5779–5783, 2016.

[27] A. Baltiyeva, E. Orynbassarova, M. Zharaspaev, and R. Akhmetov. Studying sinkholes of the earth's surface involving radar satellite interferometry in terms of zhezkazgan field, kazakhstan. *Min. Miner. Depos*, 17:61–74, 2023.

[28] J. L. Bamber, R. M. Westaway, B. Marzeion, and B. Wouters. The land ice contribution to sea level during the satellite era. *Environmental Research Letters*, 13(6): 063008, 2018.

[29] R. Barella, M. Callegari, C. Marin, C. Klug, R. Sailer, S. Galos, R. Dinale, M. Gianinetto, and C. Notarnicola. Combined use of sentinel-1 and sentinel-2 for glacier mapping: An application over central east alps. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pages 1–1, 05 2022. doi: 10.1109/JSTARS.2022.3179050.

[30] B. Barzycka, M. Błaszczyk, M. Grabiec, and J. Jania. Glacier facies of vestfonna (svalbard) based on sar images and gpr measurements. *Remote Sensing of Environment*, 221:373–385, 2019. ISSN 0034-4257. doi: https://doi.org/10.1016/j.rse.2018.11.020. URL https://www.sciencedirect.com/science/article/pii/S0034425718305297.

[31] B. Barzycka, M. Grabiec, J. Jania, M. Błaszczyk, F. Pálsson, M. Laska, D. Ignatiuk, and G. Aðalgeirsdóttir. Comparison of three methods for distinguishing glacier zones using satellite sar data. *Remote Sensing*, 15(3), 2023. ISSN 2072-4292. doi: 10.3390/rs15030690. URL https://www.mdpi.com/2072-4292/15/3/690.

[32] E. Basaeed, H. Bhaskar, and M. Al-Mualla. Beyond pan-sharpening: Pixel-level fusion in remote sensing applications. pages 139–144, 03 2012. ISBN 978-1-4673-1100-7. doi: 10.1109/INNOVATIONS.2012.6207718.

[33] A. Basso, E. Bruno, M. Parise, and M. Pepe. Morphometric analysis of sinkholes in a karst coastal area of southern apulia (italy). *Environ. Earth Sci.*, 70:2545–2559, 2013.

[34] M. Bayat, M. Ghorbanpour, R. Zare, A. Jaafari, and B. T. Pham. Application of artificial neural networks for predicting tree survival and mortality in the hyrcanian forest of iran. *Computers and Electronics in Agriculture*, 164:104929, 2019.

[35] M. Belgiu and L. Drăguţ. Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114:24–31, 2016. doi: 10.1016/j.isprsjprs.2016.01.011. ISPRS Journal of Photogrammetry and Remote Sensing.

[36] C. S. Benson. Stratigraphic studies in the snow and firn of the greenland ice sheet. Technical Report Technical Report 70, U.S. Army Snow, Ice and Permafrost Research Establishment, Corps of Engineers, 1996.

[37] C. Berger, M. Voltersen, C. Schmullius, and S. Hese. Robust mapping of urban structure types using high resolution geospatial data. *gis.Science – Die Zeitschrift für Geoinformatik*, (2):47–59, 2018.

[38] A. Bertram, A. Wendleder, A. Schmitt, and M. Huber. Long-term monitoring of water dynamics in the sahel region using the multi-sar-system. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 41:313–320, 2016.

[39] S. Bhat and R. V. Babu. Prior2posterior: Model prior correction for long-tailed learning. In *IEEE/CVF Winter Conference on Applications of Computer Vision, Tucson, Arizona, 28 February - 4 March*, pages 1296–1305, 2025.

[40] B. Bigdeli, F. Samadzadegan, and P. Reinartz. Fusion of hyperspectral and lidar data using decision template-based fuzzy multiple classifier system. *International Journal of Applied Earth Observation and Geoinformation*, 38:309–320, 2015. ISSN 1569-8432. doi: https://doi.org/10.1016/j.jag.2015.01.017. URL https://www.sciencedirect.com/science/article/pii/S0303243415000306.

[41] A. Billi, L. Fillippis, P. P. Poncia, P. Sella, and C. Faccenna. Hidden sinkholes and karst cavities in travertine plateau of the highly-populated geothermal seismic territory (tivoli, central italy). *Geomorphology*, 255:63–80, 2016.

[42] C. M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer-Verlag New York, Inc., New York, 2006.

[43] T. Blaschke. Object based image analysis for remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(1):2–16, 2010. ISSN 0924-2716. doi: 10.1016/j.isprsjprs.2009.06.004. URL https://www.sciencedirect.com/science/article/pii/S0924271609000884.

[44] H. Blatter. Stagnant ice at the bed of white glacier, axel heiberg island, n.w.t., canada. *Annals of Glaciology*, 9:35–38, 1987. doi: 10.3189/S0260305500200724. 1987b.

[45] P. Blum, M. Reinke, and J. Reh. Kulturlandschaftsgliederung bayern – neue wege für naturschutz und planung, March 2011. URL https://www.lfu.bayern.de. Informationsdienst Weihenstephan, Bayerisches Landesamt für Umwelt.

[46] S. Bojinski, M. Verstraete, T. C. Peterson, C. Richter, A. Simmons, and M. Zemp. The concept of essential climate variables in support of climate research, applications, and policy. *Bulletin of the American Meteorological Society*, 95(9):1431–1443, 2014.

[47] J.-B. Bosson, M. Huss, S. Cauvy-Fraunié, J.-C. Clément, G. Costes, M. Fischer, J. Poulenard, and F. Arthaud. Future emergence of new ecosystems caused by glacial retreat. *Nature*, 620(7974):562–569, 2023.

[48] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. URL http://dx.doi.org/10.1023/A%3A1010933404324.

[49] R. Brinkmann, K. Wilson, N. Elko, L. D. Seale, L. Florea, and H. L. Vacher. Sinkhole distribution based on pre-development mapping in urbanized pinellas county, florida. *USA, , 279*:5–11, 2007. *Geol. Soc. Lond. Spec. Publ.*

[50] R. Brinkmann, M. Parise, and D. Dye. Sinkhole distribution in a rapidly developing urban environment: Hillsborough county. Tampa Bay area, Florida*, , 99:169–184, 2008.* Eng. Geol.

[51] M. Buchhorn. Copernicus global land service: Land cover 100m: Collection 3: Epoch 2019: Globe, Sept. 2020. URL https://doi.org/10.5281/zenodo.3939050. Also available at: https://doi.org/10.1007/978-3-030-84017-4.

[52] Y. Cai, K. Guan, J. Peng, S. Wang, C. Seifert, B. Wardlow, et al. A high-performance and in-season classification system of field-level crop types using time-series landsat data and a machine learning approach. *Remote Sens. Environ.*, 210:35–47, 2018. doi: 10.1016/j.rse.2018.02.045.

[53] M. Callegari, L. Carturan, C. Marin, C. Notarnicola, P. Rastner, R. Seppi, and F. Zucca. A pol-sar analysis for alpine glacier classification and snowline altitude retrieval. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(7):3106–3121, 2016. doi: 10.1109/JSTARS.2016.2587819.

[54] M. Callegari, C. Marin, and C. Notarnicola. Multi-temporal and multi-source alpine glacier cover classification. In *2017 9th International Workshop on the*

Analysis of Multitemporal Remote Sensing Images (MultiTemp), pages 1–3, 2017. doi: 10.1109/Multi-Temp.2017.8035233.

[55] G. Camps-Valls, L. Bruzzone, J. Rojo-Alvarez, and F. Melgani. Robust support vector regression for biophysical variable estimation from remotely sensed images. *IEEE Geoscience and Remote Sensing Letters*, 3(3):339–343, 2006. doi: 10.1109/LGRS.2006.871748.

[56] J. Cao, K. Liu, L. Zhuo, L. Liu, Y. Zhu, and L. Peng. Combining uav-based hyperspectral and lidar data for mangrove species classification using the rotation forest algorithm. *International Journal of Applied Earth Observation and Geoinformation*, 102:102414, 2021. ISSN 1569-8432. doi: https://doi.org/10.1016/j.jag.2021.102414. URL https://www.sciencedirect.com/science/article/pii/S0303243421001215.

[57] Y. Cao, X. Zhou, Y. Yu, S. Rao, Y. Wu, C. Li, et al. Forest fire prediction based on time series networks and remote sensing images. *Forests*, 15(7):1221, 2024. doi: 10.3390/f15071221.

[58] Z. Cao, L. Jiang, P. Yue, J. Gong, X. Hu, S. Liu, H. Tan, C. Liu, B. Shangguan, and D. Yu. A large scale training sample database system for intelligent interpretation of remote sensing imagery. *Geo-spatial Information Science*, 0(0):1–20, 2023. doi: 10.1080/10095020.2023.2244005. URL https://doi.org/10.1080/10095020.2023.2244005.

[59] D. Carbonel, V. Rodríguez, F. Gutiérrez, J. P. McCalpin, R. Linares, C. Roqué, M. Zarroca, J. Guerrero, and I. Sasowsky. Evaluation of trenching, ground penetrating radar (gpr), and electrical resistivity tomography (ert) for sinkhole characterization. *Earth Surf. Processes Landforms*, , *39*:214–227, 2014.

[60] D. Carbonel, V. Rodríguez-Tribaldos, F. Gutiérrez, J. P. Galve, J. Guerrero, M. Zarroca, C. Roqué, R. Linares, J. P. McCalpin, and E. Acosta. Investigating a damaging buried sinkhole cluster in an urban area integrating multiple techniques: Geomorphological surveys, dinsar, gpr, ert, and trenching. *Geomorphology*, 229:3–16, 2015.

[61] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021. doi: 10.1109/ICCV48922.2021.00951.

[62] B. Chen, J. Li, and Y. Jin. Deep learning for feature-level data fusion: Higher resolution reconstruction of historical landsat archive. *Remote Sensing*, 13(2), 2021. ISSN 2072-4292. doi: 10.3390/rs13020167. URL https://www.mdpi.com/2072-4292/13/2/167.

[63] G. Chen, G. J. Hay, L. Carvalho, and M. A. Wulder. Object-based change detection. *International Journal of Remote Sensing*, 33(4):1232–1256, 2019.

[64] J. Chen, I. Dowman, S. Li, Z. Li, M. Madden, J. Mills, N. Paparoditis, F. Rottensteiner, M. Sester, C. Toth, J. Trinder, and C. Heipke. Information from imagery: Isprs scientific vision and research agenda. *ISPRS Journal of Photogrammetry and Remote Sensing*, 115:3–21, 2016. ISSN 0924-2716. doi: https://doi.org/10.1016/j.isprsjprs.2015.09.008. URL https://www.sciencedirect.com/science/article/pii/S092427161500218X. Theme issue 'State-of-the-art in photogrammetry, remote sensing and spatial information science'.

[65] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[66] T. Chen, Z. Mai, R. Li, and W. Chao. Segment anything model (SAM) enhanced pseudo labels for weakly supervised semantic segmentation. *CoRR*, abs/2305.05803, 2023. doi: 10.48550/ARXIV.2305.05803. URL https://doi.org/10.48550/arXiv.2305.05803.

[67] K. Cheng, J. Wang, and X. Yan. Mapping forest types in China with 10 m resolution based on spectral–spatial–temporal features. *Remote Sens.*, 13(5):973, 2021. doi: 10.3390/rs13050973.

[68] Q. Cheng, H. Liu, H. Shen, P. Wu, and L. Zhang. A spatial and temporal nonlocal filter-based data fusion method. *IEEE Transactions on Geoscience and Remote Sensing*, 55(8):4476–4488, 2017. doi: 10.1109/TGRS.2017.2692802.

[69] M. Choi, R. Y. Kim, M.-R. Nam, and H. O. Kim. Fusion of multispectral and panchromatic satellite images using the curvelet transform. *IEEE Geoscience and Remote Sensing Letters*, 2(2):136–140, 2005. doi: 10.1109/LGRS.2005.845313.

[70] K. Clasen, L. Hackel, T. Burgert, G. Sumbul, B. Demir, and V. Markl. reBEN: Refined BigEarthNet Dataset for Remote Sensing Image Analysis. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2025.

[71] S. Cloude and E. Pottier. An entropy based classification scheme for land applications of polarimetric sar. *IEEE Transactions on Geoscience and Remote Sensing*, 35 (1):68–78, 1997. doi: 10.1109/36.551935.

[72] J. Cogley, R. Hock, L. Rasmussen, A. Arendt, A. Bauder, R. Braithwaite, P. Jansson, G. Kaser, M. Möller, L. Nicholson, et al. *Glossary of Glacier Mass Balance and Related Terms*. IHP-VII Technical Documents in Hydrology No. 86, IACS Contribution No. 2. International Hydrological Programme of the United Nations Educational, Scientific and Cultural Organization, Paris, France, 2011. URL https://unesdoc.unesco.org/ark:/48223/pf0000192525. Accessed on 13 December 2021.

[73] J. G. Cogley, W. P. Adams, M. A. Ecclestone, F. Jung-Rothenhäusler, and C. S. L. Ommanney. Mass balance of white glacier, axel heiberg island, n.w.t., canada, 1960–91. *Journal of Glaciology*, 42(142):548–563, 1996. doi: 10.3189/S0022143000003531.

[74] K. Coleman, J. Müller, and C. Kuenzer. Remote sensing of forests in bavaria: A review. *Remote Sensing*, 16(10), 2024. ISSN 2072-4292. doi: 10.3390/rs16101805. URL https://www.mdpi.com/2072-4292/16/10/1805.

[75] A. H. Cooper. Airborne multispectral scanning of subsidence caused by permian gypsum dissolution at ripon, north yorkshire. *Q. J. Eng. Geol.*, 22:219–229, 1989.

[76] L. Cui, B. Pang, G. Zhao, C. Ban, M. Ren, D. Peng, D. Zuo, and Z. Zhu. Assessing the sensitivity of vegetation cover to climate change in the yarlung zangbo river basin using machine learning algorithms. *Remote Sensing*, 14(7), 2022. ISSN 2072-4292. doi: 10.3390/rs14071556. URL https://www.mdpi.com/2072-4292/14/7/1556.

[77] B. B. Damodaran, R. Flamary, V. Seguy, and N. Courty. An entropic optimal transport loss for learning deep neural networks under label noise in remote sensing images. *Computer Vision and Image Understanding*, 191:102863, 2020. ISSN 1077-3142. doi: 10.1016/j.cviu.2019.102863.

[78] R. C. Daudt, A. Chan-Hon-Tong, B. Le Saux, and A. Boulch. Learning to understand earth observation images with weak and unreliable ground truth. *IGARSS*, 2019 - IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan: 5602–5605, 2019. doi: 10.1109/IGARSS.2019.8898563.

[79] G. De'ath and K. E. Fabricius. Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, 81(11):3178–3192, 2000.

[80] DeepL SE. Deepl. URL https://www.deepl.com/de/translator. Accessed: 2025-06-28.

[81] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[82] S. Dersch, A. Schöttl, P. Krzystek, and M. Heurich. Towards complete tree crown delineation by instance segmentation with mask r–CNN and DETR using UAV-based multispectral imagery and lidar data. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 8:100037, apr 2023. doi: 10.1016/j.ophoto.2023.100037. URL https://doi.org/10.1016%2Fj.ophoto.2023.100037.

[83] T. G. Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

[84] N. Dionelis, C. Fibaek, L. Camilleri, A. Luyts, J. Bosmans, and B. L. Saux. Evaluating and benchmarking foundation models for earth observation and geospatial ai, 2024. URL https://arxiv.org/abs/2406.18295.

[85] G. A. C. (DLR). Sentinel-2 msi - level 2a (maja tiles) - germany, (2019). URL https://geoservice.dlr.de/data-assets/ifczsszkcp63.html.

[86] A. Dogan, D. Birant, and A. Kut. Multi-target regression for quality prediction in a mining process. In *2019 4th international conference on computer science and engineering (UBMK)*, pages 639–644. IEEE, 2019.

[87] W. Dorigo, S. Dietrich, F. Aires, L. Brocca, S. Carter, J.-F. Cretaux, D. Dunkerley, H. Enomoto, R. Forsberg, A. Güntner, et al. Closing the water cycle from observations across scales: Where do we stand? *Bulletin of the American Meteorological Society*, 102(10):E1897–E1935, 2021.

[88] J. Dou, X. Li, A. P. Yunus, U. Paudel, K.-T. Chang, Z. Zhu, and H. Pourghasemi. Automatic detection of sinkhole collapses at finer resolutions using a multi-component remote sensing approach. *Nat. Hazards*, 78:1021–1044, 2015.

[89] J. Dozier. Spectral signature of alpine snow cover from the landsat thematic mapper. *Remote Sensing of Environment*, 28:9–22, 1989. ISSN 0034-4257. doi: https://doi.org/10.1016/0034-4257(89)90101-6. URL https://www.sciencedirect.com/science/article/pii/0034425789901016.

[90] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik. Support vector regression machines. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996. URL https://proceedings.neurips.cc/paper_files/paper/1996/file/d38901788c533e8286cb6400b40b386d-Paper.pdf.

[91] C. O. Dumitru, G. Schwarz, and M. Datcu. *AL4SLEO: An Active Learning Solution for the Semantic Labelling of Earth Observation Satellite Images—Part 1*, pages 105–118. Springer Nature Singapore, 2023. ISBN 9789819939701. doi: 10.1007/978-981-99-3970-1_7. URL http://dx.doi.org/10.1007/978-981-99-3970-1_7.

[92] M. Ehlers, S. Klonus, P. J. Åstrand, and P. R. and. Multi-sensor image fusion for pansharpening in remote sensing. *International Journal of Image and Data Fusion*, 1(1):25–45, 2010. doi: 10.1080/19479830903561985. URL https://doi.org/10.1080/19479830903561985.

[93] M. Eineder, T. Fritz, et al. TerraSAR-X Ground Segment, Basic Product Specification Document. Technical Report TX-GS-DD-3302, German Aerospace Center (DLR), 2013. URL https://sss.terrasar-x.dlr.de/docs/TX-GS-DD-3302.pdf. Accessed March 2025.

[94] S. A. Elmasry, W. A. Awad, and S. A. Abd El-hafeez. Review of different image fusion techniques: Comparative study. In A. Z. Ghalwash, N. El Khameesy, D. A. Magdi, and A. Joshi, editors, *Internet of Things—Applications and Future*, pages 41–51, Singapore, 2020. Springer Singapore. ISBN 978-981-15-3075-3.

[95] Environment Canada. National climate data and information archive. http://climate.weatheroffice.gc.ca/climatenormals, 2025. Last access: 9 June 2025.

[96] ESA. Sentinel-2 user handbook, 2015. URL https://pubs.usgs.gov/fs/2008/3061/pdf/fs2008-3061.pdf. Available online: https://sentinel.esa.int/documents/247904/685211/Sentinel-2_User_Handbook (accessed on 10 January 2025).

[97] Moderate Resolution Imaging Spectroradiometer (MODIS) Overview. 2008. Available online: .

[98] E. S. A. (ESA). Copernicus digital elevation model (glo-30). https://doi.org/10.5270/ESA-c5d3d65, 2020. Accessed 2024.

[99] ESA Climate Office. Earth observation for climate monitoring, 2023. URL https://climate.esa.int/en/news-events/harnessing-earth-observation-for-climate-action/. Remarks by Prof. Jim Skea at COP28.

[100] Esri. World imagery, 2024. URL https://www.arcgis.com/home/item.html?id=10df2279f9684e4a9f6a7f08febac2a9. Accessed: 2025-04-09.

[101] *SNAP – ESA Sentinel Application Platform v12.0.0*. European Space Agency (ESA), 2025. URL http://step.esa.int. Accessed: 2025-06-28.

[102] FAO. *The State of the World's Forests 2022: Forest Pathways for Green Recovery and Building Inclusive, Resilient and Sustainable Economies*. FAO, Rome, 2022. URL https://doi.org/10.4060/cb9360en.

[103] D. Farinotti, M. Huss, J. J. Fürst, J. Landmann, H. Machguth, F. Maussion, and A. Pandit. A consensus estimate for the ice thickness distribution of all glaciers on earth. *Nature Geoscience*, 12(3):168–173, 2019.

[104] F. E. Fassnacht, J. C. White, M. A. Wulder, and E. Næsset. Remote sensing in forestry: current challenges, considerations and directions. *Forestry: An International Journal of Forest Research*, 97(1):11–37, 2024.

[105] M. Fauvel, J. Chanussot, and J. Benediktsson. Decision fusion for the classification of urban remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 44(10):2828–2838, 2006. doi: 10.1109/TGRS.2006.876708.

[106] S. Feng, J. M. Cook, A. M. Anesio, L. G. Benning, and M. Tranter. Long time series (1984–2020) of albedo variations on the greenland ice sheet from harmonized landsat and sentinel 2 imagery. *Journal of Glaciology*, 69(277):1225–1240, 2023.

[107] Y. Feng, J. Zhu, R. Song, and X. Wang. S2eft: Spectral-spatial-elevation fusion transformer for hyperspectral image and lidar classification. *Knowledge-Based Systems*, 283:111190, 2024. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2023.111190. URL https://www.sciencedirect.com/science/article/pii/S0950705123009401.

[108] V. Festa, A. Fiore, M. Parise, and A. Siniscalchi. Sinkhole evolution in the Apulian karst of southern Italy: A case study, with some considerations on sinkhole hazards. *j. cave karst stud.* **2012**. *74*, 137–147,, 2012.

[109] G. F. Ficetola, S. Marta, A. Guerrieri, I. Cantera, A. Bonin, S. Cauvy-Fraunié, R. Ambrosini, M. Caccianiga, F. Anthelme, R. S. Azzoni, et al. The development of terrestrial ecosystems emerging after glacier retreat. *Nature*, 632(8024):336–342, 2024.

[110] L. Fonseca, L. Namikawa, E. Castejon, L. Carvalho, C. Pinho, and A. Pagamisse. Image fusion for remote sensing applications. In Y. Zheng, editor, *Image Fusion and Its Applications*, chapter 9. IntechOpen, Rijeka, 2011. doi: 10.5772/22899. URL https://doi.org/10.5772/22899.

[111] G. M. Foody, A. Mathur, C. Sanchez-Hernandez, and D. S. Boyd. Training set size requirements for the classification of a specific class. *Remote Sensing of Environment*, 104(1):1–14, 2006. ISSN 0034-4257. doi: https://doi.org/10.1016/j.rse.2006.03.004. URL https://www.sciencedirect.com/science/article/pii/S0034425706001234.

[112] B. A. for Digitalisation. High-speed internet and surveying, 2024. URL https://ldbv.bayern.de/produkte/3dprodukte/gelaende.html. 3D Products—Terrain Models.

[113] S. Fragou, K. Kalogeropoulos, N. Stathopoulos, P. Louka, P. K. Srivastava, S. Karpouzas, D. P. Kalivas, and G. P. Petropoulos. Quantifying land cover changes in a mediterranean environment using landsat tm and support vector machines. *Forests*, 11(7):750, 2020.

[114] M. Franquesa, A. M. Rodriguez-Montellano, E. Chuvieco, and I. Aguado. Reference data accuracy impacts burned area product validation: The role of the expert analyst. *Remote Sensing*, 14(17), 2022. ISSN 2072-4292. URL https://www.mdpi.com/2072-4292/14/17/4354.

[115] M. A. Friedl, D. Sulla-Menashe, B. Tan, A. Schneider, N. Ramankutty, A. Sibley, and X. Huang. Modis collection 5 global land cover: Algorithm refinements and characterization of new datasets. *Remote Sensing of Environment*, 114(1):168–182, 2010. doi: 10.1016/j.rse.2009.08.016.

[116] M. Fromm, M. Schubert, G. Castilla, J. Linke, and G. McDermid. Automated detection of conifer seedlings in drone imagery using convolutional neural networks. *Remote Sensing*, 11(21), 2019. ISSN 2072-4292. doi: 10.3390/rs11212585. URL https://www.mdpi.com/2072-4292/11/21/2585.

[117] A. Frumkin, M. Ezersky, A. Al-Zoubi, E. Akkawi, and A.-R. Abueladas. The dead sea sinkhole hazard: Geophysical assessment of salt dissolution and collapse. *Geomorphology*, 134:102–117, 2011.

[118] J. P. Galve, C. Castañeda, F. Gutiérrez, and G. Herrera. Assessing sinkhole activity in the ebro valley mantled evaporite karst using advanced dinsar. *Geomorphology*, 229:30–44, 2015.

[119] P. Gamba and J. Chanussot. Foreword to the special issue on data fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 46(5):1283–1288, 2008. doi: 10.1109/TGRS.2008.919761. URL https://hal.science/hal-00348849.

[120] F. Gao, J. Masek, M. Schwaller, and F. Hall. On the blending of the landsat and modis surface reflectance: predicting daily landsat surface reflectance. *IEEE Transactions on Geoscience and Remote Sensing*, 44(8):2207–2218, 2006. doi: 10.1109/TGRS.2006.872081.

[121] F. Gao, T. Hilker, X. Zhu, M. Anderson, J. Masek, P. Wang, and Y. Yang. Fusing landsat and modis data for vegetation monitoring. *IEEE Geoscience and Remote Sensing Magazine*, 3(3):47–60, 2015. doi: 10.1109/MGRS.2015.2434351.

[122] P. Ghamisi, B. Rasti, N. Yokoya, Q. Wang, B. Hofle, L. Bruzzone, F. Bovolo, M. Chi, K. Anders, R. Gloaguen, P. M. Atkinson, and J. A. Benediktsson. Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, 7(1):6–39, 2019. doi: 10.1109/MGRS.2018.2890023.

[123] R. Gharbia, A. T. Azar, A. E. Baz, and A. E. Hassanien. Image fusion techniques in remote sensing, 2014. URL https://arxiv.org/abs/1403.5473.

[124] H. Ghassemian. Multi-sensor image fusion using multirate filter banks. In *Proceedings 2001 International Conference on Image Processing (Cat. No.01CH37205)*, volume 1, pages 846–849 vol.1, 2001. doi: 10.1109/ICIP.2001.959178.

[125] H. Ghassemian. A review of remote sensing image fusion methods. *Information Fusion*, 32:75–89, 2016. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2016.03.003. URL https://www.sciencedirect.com/science/article/pii/S1566253516300173.

[126] GLIMS and National Snow and Ice Data Center. Global land ice measurements from space glacier database. https://doi.org/10.7265/N5V98602, 2005. Updated 2018.

[127] M. Golipour, H. Ghassemian, and F. Mirzapour. Integrating hierarchical segmentation maps with mrf prior for classification of hyperspectral images in a bayesian framework. *IEEE Transactions on Geoscience and Remote Sensing*, 54(2):805–816, 2016. doi: 10.1109/TGRS.2015.2466657.

[128] Grammarly. Grammarly. URL https://app.grammarly.com/. Accessed: 2025-06-28.

[129] Y. Gu and B. K. Wylie. Developing a 30-m grassland productivity estimation map for central nebraska using 250-m modis and 30-m landsat-8 observations. *Remote Sensing of Environment*, 171:291–298, 2015.

[130] F. Guo, Q. Meng, Z. Li, G. Ren, L. Wang, J. Zhang, R. Xin, and Y. Hu. Multisource feature embedding and interaction fusion network for coastal wetland classification with hyperspectral and lidar data. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024. doi: 10.1109/TGRS.2024.3367960.

[131] J. Guo, H. Sun, J. Han, B. Song, Y. Chi, and B. Song. Multitask fine-grained feature mining for multilabel remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–17, 2024. doi: 10.1109/TGRS.2024.3426473.

[132] Z. Guo, R. Xu, C.-C. Feng, and Z. Zeng. Pif-net: A deep point-image fusion network for multimodality semantic segmentation of very high-resolution imagery and aerial point cloud. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024. doi: 10.1109/TGRS.2023.3342477.

[133] V. Gupta and S. Mehra. Image fusion techniques - a comparative study. *International Journal of Engineering Trends and Technology*, 32:113–118, 2 2016. doi: 10.14445/22315381/IJETT-V32P220.

[134] F. Gutiérrez, J. P. Galve, P. Lucha, C. Castañeda, J. Bonachea, and J. Guerrero. Integrating geomorphological mapping, trenching, insar, and gpr for the identification and characterization of sinkholes in the mantled evaporite karst of the ebro valley (ne spain). *Geomorphology*, 134:144–156, 2011.

[135] W. Ha, P. H. Gowda, and T. A. Howell. A review of potential image fusion methods for remote sensing-based irrigation management: part ii. *Irrigation*

Science, 31:851–869, 2013. doi: 10.1007/s00271-012-0340-6. URL https://doi.org/10.1007/s00271-012-0340-6.

[136] W. Haeberli and C. Whiteman. Chapter 1 - snow and ice-related hazards, risks, and disasters: A general framework. In J. F. Shroder, W. Haeberli, and C. Whiteman, editors, *Snow and Ice-Related Hazards, Risks, and Disasters*, Hazards and Disasters Series, pages 1–34. Academic Press, Boston, 2015. ISBN 978-0-12-394849-6. doi: https://doi.org/10.1016/B978-0-12-394849-6.00001-9. URL https://www.sciencedirect.com/science/article/pii/B9780123948496000019.

[137] W. Hagg. *Gletscherkunde und Glazialgeomorphologie*. Springer Spektrum Berlin, Heidelberg, 1 edition, 2020. ISBN 978-3-662-61993-3. doi: 10.1007/978-3-662-61994-0. 208 pages, 25 b/w and 92 color illustrations.

[138] D. L. Hall. *Mathematical Techniques in Multisensor Data Fusion*. Artech House, Boston, MA, 1992.

[139] A. B. Hamida, A. Benoit, P. Lambert, and C. B. Amar. 3-d deep learning approach for remote sensing image classification. *IEEE Transactions on geoscience and remote sensing*, 56(8):4420–4434, 2018.

[140] P. Hansen and J. Schjoerring. Reflectance measurement of canopy biomass and nitrogen status in wheat crops using normalized difference vegetation indices and partial least squares regression. *Remote Sensing of Environment*, 86(4):542–553, 2003. ISSN 0034-4257. doi: https://doi.org/10.1016/S0034-4257(03)00131-7. URL https://www.sciencedirect.com/science/article/pii/S0034425703001317.

[141] C. Hao, T. Oguchi, and P. Wu. A semi-automatic model for sinkhole identification in a karst area of zhijin county, China. 2015. doi: 10.1117/12.2207433.

[142] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud, and A. Hussain. Interpreting black-box models: A review on explainable artificial intelligence. *Cognitive Computation*, 16:45–74, 2024. doi: 10.1007/s12559-023-10179-8.

[143] S. Hauser and A. Schmitt. Glacier retreat in iceland mapped from space: Time series analysis of geodata from 1941 to 2018. *PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 89(3):273–291, Jun 2021. ISSN 2512-2819. doi: 10.1007/s41064-021-00139-y. URL https://doi.org/10.1007/s41064-021-00139-y.

[144] S. Hauser and A. Schmitt. Forest5dplus: An open benchmark data set for the estimation of forest parameters from sentinel-1 and -2 time series with machine learning methods. In *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, pages 1700–1703, 2023. doi: 10.1109/IGARSS52108.2023.10282042.

[145] S. Hauser, A. Wendleder, and R. Pesch. Monitoring of the glaciers on axel heiberg island from space. *FORUM*, pages 4–14, 2021. Article in German.

[146] S. Hauser, A. Wendleder, A. Roth, W. van Wychen, and L. Thompson. Glacier zonation and velocity estimations on axel heiberg island using terrasar-x data. In *42nd Canadian Symposium on Remote Sensing*, 2021.

[147] S. Hauser, M. Ruhhammer, A. Schmitt, and P. Krzystek. An open benchmark dataset for forest characterization from sentinel-1 and -2 time series. *Remote Sens.*, 16(3):488, 2024. doi: 10.3390/rs16030488.

[148] S. Hauser, A. Schmitt, P. Krzystek, and M. Ruhhammer. Wald5dplus (1.0. 0)[data set], 2024.

[149] S. Hauser, L. Augner, and A. Schmitt. Perfect labelling: A review and outlook of label optimization techniques in dynamic earth observation. *Remote Sensing*, 17(7):1246, 2025. doi: 10.3390/rs17071246. URL https://doi.org/10.3390/rs17071246.

[150] L. Heidrich, S.-E. Bae, S. Levick, S. Seibold, W. Weisser, P. Krzystek, P. Magdon, T. Nauss, P. Schall, A. Serebryanyk, S. Wollauer, C. Ammer, C. Bassler, I. Doerfler, M. Fischer, M. M. Gossner, M. Heurich, T. Hothorn, K. Jung, H. Kreft, E.-D. Schulze, N. Simons, S. Thorn, and J. Muller. Heterogeneity-diversity relationships differ between and within trophic levels in temperate forests. *Nature Ecology & Evolution*, 4:1431–1431, 2020.

[151] V. Henrich, G. Krauss, C. Götze, and C. Sandow. IDB - www.indexdatabase.de, Entwicklung einer Datenbank für Fernerkundungsindizes. Presented at AK Fernerkundung, 2012. Bochum, 4–5 October 2012.

[152] M. Heurich, B. Beudert, H. Rall, and Z. Křenová. National parks as model regions for interdisciplinary long-term ecological research: The bavarian forest and šumava national parks underway to transboundary ecosystem research. In F. Müller, C. Baessler, H. Schubert, and S. Klotz, editors, *Long-Term Ecological Research: Between Theory and Applications*, pages 327–344. Springer, Amsterdam, 2010.

[153] M. Heurich, T. Ochs, T. Andresen, and T. Schneider. Object-orientated image analysis for the semi-automatic detection of dead trees following a spruce bark beetle (ips typographus) outbreak. *European Journal of Forest Research*, 129: 313–324, 2010.

[154] T. Hilker, M. A. Wulder, N. C. Coops, J. Linke, G. McDermid, J. G. Masek, F. Gao, and J. C. White. A new data fusion model for high spatial- and temporal-resolution mapping of forest disturbance based on landsat and modis. *Remote Sensing of Environment*, 113(8):1613–1627, 2009. ISSN 0034-4257. doi: https://doi.org/10.1016/j.rse.2009.03.007. URL https://www.sciencedirect.com/science/article/pii/S003442570900087X.

[155] N. V. Hoai, N. M. Dung, and S. Ro. Sinkhole detection by deep learning and data association. In *Proceedings of the 2019 Eleventh International Conference on Ubiquitous and Future Networks (ICUFN)*, 2019.

[156] P. Hochreuther, N. Neckel, N. Reimann, A. Humbert, and M. Braun. Fully automated detection of supraglacial lake area for northeast greenland using sentinel-2 time-series. *Remote Sensing*, 13(2):205, 2021.

[157] S. M. Hochstuhl, N. Pfeffer, A. Thiele, S. Hinz, J. Amao-Oliva, R. Scheiber, A. Reigber, and H. Dirks. Pol-insar-island – a benchmark dataset for multi-frequency pol-insar data land cover classification. *Karlsruhe Institute of Technology, KITopen-DOI*, 10(5445):2023–06, 2023. doi: 10.35097/1450. (KITopen-DOI) 10.5445/IR/1000159469.

[158] R. Hock, A. Bliss, B. Marzeion, R. H. Giesen, Y. Hirabayashi, M. Huss, V. Radić, and A. B. Slangen. Glaciermip–a model intercomparison of global-scale glacier mass-balance models and projections. *Journal of Glaciology*, 65(251):453–467, 2019.

[159] S. Holzwarth, F. Thonfeld, P. Kacic, S. Abdullahi, S. Asam, K. Coleman, C. Eisfelder, U. Gessner, J. Huth, T. Kraus, et al. Earth-observation-based monitoring of forests in germany—recent progress and research frontiers: a review. *Remote Sensing*, 15 (17):4234, 2023.

[160] M. J. Hopwood, D. Carroll, T. Dunse, A. Hodson, J. M. Holding, J. L. Iriarte, S. Ribeiro, E. P. Achterberg, C. Cantoni, D. F. Carlson, et al. How does glacier discharge affect marine biogeochemistry and primary production in the arctic? *The Cryosphere*, 14(4):1347–1383, 2020.

[161] F. Hu, G.-S. Xia, J. Hu, and L. Zhang. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. *Remote Sensing*, 7(11):14680–14707, 2015.

[162] Y. Hu. *Automated Extraction of Digital Terrain Models, Roads and Buildings Using Airborne Lidar Data*. phdthesis, University of Calgary (Canada), October 2003. URL https://ui.adsabs.harvard.edu/abs/2004PhDT.....212H/abstract.

[163] Y. Hu, Y. Han, and Y. Zhang. Land desertification and its influencing factors in kazakhstan. *Journal of Arid Environments*, 180:104203, 2020. doi: 10.1016/j.jaridenv.2020.104203. URL https://doi.org/10.1016/j.jaridenv.2020.104203. Accessed: 2024-11-21.

[164] M. Huber, W. Hummelbrunner, J. Raggam, D. Small, and D. Kosmann. Technical aspects of envisat-asar geocoding capability at dlr. In *ENVISAT and ERS Symposium*, Salzburg, Austria, 2004.

[165] P. J. Huber. Robust regression: Asymptotics, conjectures and monte carlo. *The Annals of Statistics*, 1(5):799–821, 1973. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/2958283.

[166] M. Huss and R. Hock. Global-scale hydrological response to future glacier mass loss. *Nature Climate Change*, 8(2):135–140, 2018.

[167] D. Ienco, R. Gaetano, C. Dupaquier, and P. Maurel. Land cover classification via multitemporal spatial data by deep recurrent neural networks. *IEEE Geoscience and Remote Sensing Letters*, 14(10):1685–1689, 2017.

[168] I. P. G. Ikonos. 2010, url = https://earth.esa.int/eogateway/documents/20142/37627/IKONOS-Imagery-Product-Guide.pdf, urldate = 2025-01-10,.

[169] Intergovernmental Panel on Climate Change (IPCC). Sixth assessment report (ar6). https://www.ipcc.ch/assessment-report/ar6/, 2023. Accessed: 2025-06-17.

[170] B. Janga, G. P. Asamani, Z. Sun, and N. Cristea. A review of practical ai for remote sensing in earth sciences. *Remote Sensing*, 15(16), 2023. ISSN 2072-4292. doi: 10.3390/rs15164112. URL https://www.mdpi.com/2072-4292/15/16/4112.

[171] L. Ji and A. J. Peters. Lag and seasonality considerations in evaluating avhrr ndvi response to precipitation. *Photogrammetric Engineering and Remote Sensing*, 71 (9):1053–1061, 2005. doi: 10.14358/PERS.71.9.1053.

[172] S. Ji, C. Zhang, A. Xu, Y. Shi, and Y. Duan. 3d convolutional neural networks for crop classification with multi-temporal remote sensing images. *Remote Sensing*, 10 (1):75, 2018.

[173] X. Jia, A. Khandelwal, G. Nayak, J. Gerber, K. Carlson, P. West, and V. Kumar. Incremental dual-memory lstm in land cover prediction. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 867–876, 2017.

[174] P. Jian, Y. Ou, and K. Chen. Uncertainty-aware graph self-supervised learning for hyperspectral image change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–19, 2024. doi: 10.1109/TGRS.2024.3363886.

[175] D. Jiang, D. Zhuang, and Y. Huang. Investigation of image fusion for remote sensing application. In Q. Miao, editor, *New Advances in Image Fusion*, chapter 1. IntechOpen, Rijeka, 2011. doi: 10.5772/56946. URL https://doi.org/10.5772/56946.

[176] Joblib Development Team. Joblib: running python functions as pipeline jobs, 2025. URL https://joblib.readthedocs.io/.

[177] G. Kabzhanova, R. Arystanova, A. Bissembayev, A. Arystanov, J. Sagin, B. Nasiyev, and A. Kurmasheva. Remote sensing applications for pasture assessment in kazakhstan. *Agronomy*, 15(3):526, 2025.

[178] S. Kahraman and A. Ertürk. A comprehensive review of pansharpening algorithms for gÖktÜrk-2 satellite images. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-4/W4:263–270, 2017. doi: 10.5194/isprs-annals-IV-4-W4-263-2017. URL https://isprs-annals.copernicus.org/articles/IV-4-W4/263/2017/.

[179] Y. Karimi, S. Prasher, A. Madani, S. Kim, et al. Application of support vector machine technology for the estimation of crop biophysical parameters using aerial hyperspectral observations. *Can. Biosyst. Eng*, 50(7):13–20, 2008.

[180] G. Kaufmann and D. Romanov. Geophysical investigation of a sinkhole in the northern harz foreland (north germany). *Environ. Geol.*, 59:401–405, 2009.

[181] M. Kautz, J. Feurer, and P. Adler. Early detection of bark beetle (ips typographus) infestations by remote sensing – a critical review of recent research. *Forest Ecology*

and Management, 556:121595, 2024. ISSN 0378-1127. doi: https://doi.org/10.1016/j.foreco.2023.121595. URL https://www.sciencedirect.com/science/article/pii/S0378112723008290.

[182] A. A. Khan, A. Jamil, D. Hussain, M. Taj, G. Jabeen, and M. K. Malik. Machine-learning algorithms for mapping debris-covered glaciers: The hunza basin case study. *IEEE Access*, 8:12725–12734, 2020. doi: 10.1109/ACCESS.2020.2965768.

[183] Y. J. Kim, B. H. Nam, and H. Youn. Sinkhole detection and characterization using lidar-derived dem with logistic regression. *Remote Sens.*, 11, 2019.

[184] A. Klimchouk and V. Andrejchuk. Karst breakdown mechanisms from observations in the gypsum caves of the western ukraine: Implications for subsidence hazard assessment. *Environ. Geol.*, (48):336–359, 2005.

[185] K. Knauer, U. Gessner, R. Fensholt, and C. Kuenzer. An estarfm fusion framework for the generation of large-scale time series in cloud-prone and heterogeneous landscapes. *Remote Sensing*, 8(5), 2016. ISSN 2072-4292. doi: 10.3390/rs8050425. URL https://www.mdpi.com/2072-4292/8/5/425.

[186] M. Kneib, E. S. Miles, S. Jola, P. Buri, S. Herreid, A. Bhattacharya, C. Watson, T. Bolch, D. Quincey, and F. Pellicciotti. Mapping ice cliffs on debris-covered glaciers using multispectral satellite images. *Remote Sensing of Environment*, 253: 112201, 2021.

[187] A. Kobler, N. Pfeifer, P. Ogrinc, L. Todorovski, K. Oštir, and S. Džeroski. Repetitive interpolation: A robust algorithm for dtm generation from aerial laser scanner data in forested terrain. *Remote Sens. Environ.*, 108:9–23, 2007.

[188] N. Kolarik, N. Shrestha, T. Caughlin, and J. Brandt. Leveraging high resolution classifications and random forests for hindcasting decades of mesic ecosystem dynamics in the landsat time series. *Ecological Indicators*, 158:111445, 2024. ISSN 1470-160X. doi: https://doi.org/10.1016/j.ecolind.2023.111445. URL https://www.sciencedirect.com/science/article/pii/S1470160X2301587X.

[189] V. Kolluru, R. John, J. Chen, M. Jarchow, R. Goljani, V. Giannico, S. Saraf, K. Jain, M. Kussainova, and J. Yuan. Untangling the impacts of socioeconomic and climatic changes on vegetation greenness and productivity in kazakhstan. *Environmental Research Letters*, 2022. doi: 10.1088/1748-9326/ac8c59. URL https://doi.org/10.1088/1748-9326/ac8c59. Accessed: 2024-11-21.

[190] Komal and R. Dewan. Energy based wavelet image fusion. *International Journal for Innovative Research in Science & Technology*, 1(6), 2014.

[191] A. Koshim, A. Sergeyeva, R. Bexeitova, and A. Aktymbayeva. Landscape of the mangystau region in kazakhstan as a geomorphotourism destination: A geographical review. *Geoj. Tour. Geosites*, 29:385–397, 2020.

[192] S. Kotsiantis and P. Pintelas. Mixture of expert agents for handling imbalanced data sets. *Annals of Mathematics, Computing & Teleinformatics*, 1(1):46–55, 2003.

[193] P. Krzystek, A. Serebryanyk, C. Schnörr, J. Červenka, and M. Heurich. Large-scale mapping of tree species and dead trees in šumava national park and bavarian forest national park using lidar and multispectral imagery. *Remote Sensing*, 12 (4):661, feb 2020. doi: 10.3390/rs12040661. URL https://doi.org/10.3390%2Frs12040661.

[194] M. Kufer, A. Schmitt, A. Wendleder, P. Krzystek, and M. Heurich. Estimation of forest parameters from polarimetric l-band images. In *EUSAR 2022*, pages 126–131, 2022. URL https://elib.dlr.de/189672/.

[195] L. Kulanuwat, C. Chantrapornchai, M. Maleewong, P. Wongchaisuwat, S. Wimala, K. Sarinnapakorn, et al. Anomaly detection using a sliding window technique and data imputation with machine learning for hydrological time series. *Water*, 13 (13):1862, 2021. doi: 10.3390/w13131862.

[196] A. Kulshrestha, L. Chang, and A. Stein. Sinkhole scanner: A new method to detect sinkhole-related spatio-temporal patterns in insar deformation time series. *Remote Sens.*, 13, 2021.

[197] C. Kwan, B. Chou, J. Yang, D. Perez, Y. Shen, J. Li, and K. Koperski. Fusion of landsat and worldview images. In I. Kadar, E. P. Blasch, and L. L. Grewe, editors, *Signal Processing, Sensor/Information Fusion, and Target Recognition XXVIII*, volume 11018, page 1101816. International Society for Optics and Photonics, SPIE, 2019. doi: 10.1117/12.2518949. URL https://doi.org/10.1117/12.2518949.

[198] S. König, F. Thonfeld, M. Förster, O. Dubovyk, and M. H. and. Assessing combinations of landsat, sentinel-2 and sentinel-1 time series for detecting bark beetle infestations. *GIScience & Remote Sensing*, 60(1):2226515, 2023. doi: 10.1080/15481603.2023.2226515. URL https://doi.org/10.1080/15481603.2023.2226515.

[199] S. König, F. Thonfeld, M. Förster, O. Dubovyk, and M. Heurich. Assessing combinations of landsat, sentinel-2 and sentinel-1 time series for detecting bark beetle infestations. *GIScience & Remote Sensing*, 60(1):2226515, 2023. doi: 10.1080/15481603.2023.2226515. URL https://doi.org/10.1080/15481603.2023.2226515.

[200] S. K. Lam, A. Pitrou, and S. Seibert. Numba: A llvm-based python jit compiler. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, pages 1–6, 2015.

[201] K. Langley, S.-E. Hamran, K. A. Hogda, R. Storvold, O. Brandt, J. Kohler, and J. O. Hagen. From glacier facies to sar backscatter zones via gpr. *IEEE Transactions on Geoscience and Remote Sensing*, 46(9):2506–2516, 2008. doi: 10.1109/TGRS.2008.918648.

[202] D. J. Lary, G. K. Zewdie, X. Liu, D. Wu, E. Levetin, R. J. Allee, N. Malakar, A. Walker, H. Mussa, A. Mannino, and D. Aurin. *Machine Learning Applications for Earth Observation*, volume 15 of *ISSI Scientific Report Series*. Springer, Cham, 2018. doi: 10.1007/978-3-319-65633-5_8.

[203] H. Latifi, S. Holzwarth, A. Skidmore, J. Brůna, J. Červenka, R. Darvishzadeh, M. Hais, U. Heiden, L. Homolová, P. Krzystek, T. Schneider, M. Starý, T. Wang, J. Müller, and M. Heurich. A laboratory for conceiving essential biodiversity variables (ebvs)—the "data pool initiative for the bohemian forest ecosystem.". *Methods in Ecology and Evolution*, 12:2073–2083, 2021. doi: 10.1111/2041-210X.13695.

[204] M. Lebedeva-Verba and M. Gerasimova. Micromorphology of takyrs and the desert "papyrus" of southwestern turkmenia. *Eurasian Soil Sci.*, 43:1220–1229, 2010.

[205] M. Lechner, A. Dostálová, M. Hollaus, C. Atzberger, and M. Immitzer. Combination of sentinel-1 and sentinel-2 data for tree species classification in a central european biosphere reserve. *Remote Sensing*, 14(11), 2022. ISSN 2072-4292. doi: 10.3390/rs14112687. URL https://www.mdpi.com/2072-4292/14/11/2687.

[206] H. Lee, J. Wang, and B. Leblon. Using linear regression, random forests, and support vector machine with unmanned aerial vehicle multispectral images to predict canopy nitrogen weight in corn. *Remote Sensing*, 12(13), 2020. ISSN 2072-4292. doi: 10.3390/rs12132071. URL https://www.mdpi.com/2072-4292/12/13/2071.

[207] J.-S. Lee, M. Grunes, T. Ainsworth, L.-J. Du, D. Schuler, and S. Cloude. Unsupervised classification using polarimetric decomposition and the complex wishart classifier. *IEEE Transactions on Geoscience and Remote Sensing*, 37(5):2249–2258, 1999. doi: 10.1109/36.789621.

[208] T. Leichtle, S. Helgert, M. Müller, J. Handschuh, T. Erbertseder, M. Wurm, and H. Taubenböck. Opposing land surface and air temperatures from remote sensing and citizen science for quantification of the urban heat island effect. In *2023 Joint Urban Remote Sensing Event (JURSE)*, pages 1–5, 2023. doi: 10.1109/JURSE57346.2023.10144135.

[209] T. Leichtle, M. Kühnl, A. Droin, C. Beck, M. Hiete, and H. Taubenböck. Quantifying urban heat exposure at fine scale - modeling outdoor and indoor temperatures using citizen science and vhr remote sensing. *Urban Climate*, 49: 101522, 2023. ISSN 2212-0955. doi: 10.1016/j.uclim.2023.101522. URL https://www.sciencedirect.com/science/article/pii/S2212095523001165.

[210] T. M. Lenton, J. F. Abrams, A. Bartsch, S. Bathiany, C. A. Boulton, J. E. Buxton, A. Conversi, A. M. Cunliffe, S. Hebden, T. Lavergne, B. Poulter, A. Shepherd, T. Smith, D. Swingedouw, R. Winkelmann, and N. Boers. Remotely sensing potential climate change tipping points across scales. *Nature Communications*, 15(1):343, 2024. ISSN 2041-1723. doi: 10.1038/s41467-023-44609-w. URL https://doi.org/10.1038/s41467-023-44609-w.

[211] G. Lesins, T. Duck, and J. Drummond. Climate trends at eureka in the canadian high arctic. *Atmosphere-Ocean*, 48:59–80, 2010. doi: 10.3137/AO1103.2010.

[212] J. Li, X. Huang, and J. Gong. Deep neural network for remote-sensing image interpretation: status and perspectives. *National Science Review*, 6(6):1082–1086, 05 2019. ISSN 2095-5138. doi: 10.1093/nsr/nwz058. URL https://doi.org/10.1093/nsr/nwz058.

[213] J. Li, L. Meng, B. Yang, C. Tao, L. Li, and W. Zhang. Labelrs: An automated toolbox to make deep learning samples from remote sensing images. *Remote Sensing*, 13(11), 2021. ISSN 2072-4292. doi: 10.3390/rs13112064. URL https://www.mdpi.com/2072-4292/13/11/2064.

[214] S. Li, B. Yang, and J. Hu. Performance comparison of different multi-resolution transforms for image fusion. *Information Fusion*, 12(2):74–84, 2011. ISSN 1566-

2535. doi: https://doi.org/10.1016/j.inffus.2010.03.002. URL https://www.sciencedirect.com/science/article/pii/S1566253510000382.

[215] Y. Li, H. Zhang, and Q. Shen. Spectral–spatial classification of hyperspectral imagery with 3d convolutional neural network. *Remote Sensing*, 9(1):67, 2017.

[216] Y. Li, M. Li, C. Li, and Z. Liu. Forest aboveground biomass estimation using landsat 8 and sentinel-1a data with machine learning algorithms. *Scientific reports*, 10(1): 9952, 2020.

[217] B. Liang, S. Han, W. Li, G. Huang, and R. He. Spatial-temporal alignment of time series with different sampling rates based on cellular multi-objective whale optimization. *Inf. Process. Manag.*, 60(1):103123, 2023. doi: 10.1016/j.ipm.2022.103123. URL https://www.sciencedirect.com/science/article/pii/S0306457322002242.

[218] W. Liao, A. Pižurica, R. Bellens, S. Gautama, and W. Philips. Generalized graph-based fusion of hyperspectral and lidar data using morphological features. *IEEE Geoscience and Remote Sensing Letters*, 12(3):552–556, 2015. doi: 10.1109/LGRS.2014.2350263.

[219] D. A. Lilien, N. F. Nymand, T. A. Gerber, D. Steinhage, D. Jansen, L. Thomson, M. Myers, S. Franke, D. Taylor, P. Gogineni, and et al. Potential to recover a record of holocene climate and sea ice from müller ice cap, canada. *Journal of Glaciology*, 70:e72, 2024. doi: 10.1017/jog.2024.75.

[220] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

[221] E. R. Lines, M. Allen, C. Cabo, K. Calders, A. Debus, S. W. D. Grieve, M. Miltiadou, A. Noach, H. J. F. Owen, and S. Puliti. Ai applications in forest monitoring need remote sensing benchmark datasets. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 4528–4533, 2022. doi: 10.1109/BigData55660.2022.10020772.

[222] E. R. Lines, F. J. Fischer, H. J. F. Owen, and T. Jucker. The shape of trees: Reimagining forest ecology in three dimensions with remote sensing. *Journal of Ecology*, 110(8):1730–1745, 2022. doi: https://doi.org/10.1111/1365-2745.

13944. URL https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/1365-2745.13944.

[223] W. Liu, R. Li, T. Wu, X. Shi, X. Wu, L. Zhao, G. Hu, J. Yao, J. Ma, S. Wang, et al. Preliminary simulation of spatial distribution patterns of soil thermal conductivity in permafrost of the arctic. *International Journal of Digital Earth*, 16(2):4512–4532, 2023.

[224] A. Loew, W. Bell, L. Brocca, C. E. Bulgin, J. Burdanowitz, X. Calbet, R. V. Donner, D. Ghent, A. Gruber, T. Kaminski, J. Kinzel, C. Klepp, J.-C. Lambert, G. Schaepman-Strub, M. Schröder, and T. Verhoelst. Validation practices for satellite-based earth observation data across communities. *Reviews of Geophysics*, 55:779–817, 2017. doi: 10.1002/2017RG000562.

[225] L. Loncan, L. B. de Almeida, J. M. Bioucas-Dias, X. Briottet, J. Chanussot, N. Dobigeon, S. Fabre, W. Liao, G. A. Licciardi, M. Simões, J.-Y. Tourneret, M. A. Veganzones, G. Vivone, Q. Wei, and N. Yokoya. Hyperspectral pansharpening: A review. *IEEE Geoscience and Remote Sensing Magazine*, 3(3):27–46, 2015. doi: 10.1109/MGRS.2015.2440094.

[226] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez. Convolutional neural networks for large-scale remote-sensing image classification. *IEEE Transactions on geoscience and remote sensing*, 55(2):645–657, 2016.

[227] K. Maharana, S. Mondal, and B. Nemade. A review: Data pre-processing and data augmentation techniques. *Glob. Transit. Proc.*, 3(1):91–99, 2022. doi: 10.1016/j.gltp.2022.04.020.

[228] G. Mai, W. Huang, J. Sun, S. Song, D. Mishra, N. Liu, S. Gao, T. Liu, G. Cong, Y. Hu, C. Cundy, Z. Li, R. Zhu, and N. Lao. On the opportunities and challenges of foundation models for geoai (vision paper). *ACM Trans. Spatial Algorithms Syst.*, 10(2), 2024. ISSN 2374-0353. doi: 10.1145/3653070. URL https://doi.org/10.1145/3653070.

[229] L. Mandl and S. Lang. Uncovering early traces of bark beetle induced forest stress via semantically enriched sentinel-2 data and spectral indices. *PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 91(3):211–231, 2023. doi: 10.1007/s41064-023-00240-4. URL https://doi.org/10.1007/s41064-023-00240-4.

[230] J. Mardian, C. Champagne, B. Bonsal, and A. Berg. A machine learning framework for predicting and understanding the canadian drought monitor. *Water Resources Research*, 59(8):e2022WR033847, 2023.

[231] B. Marzeion, R. Hock, B. Anderson, A. Bliss, N. Champollion, K. Fujita, M. Huss, W. W. Immerzeel, P. Kraaijenbrink, J.-H. Malles, et al. Partitioning the uncertainty of ensemble projections of global glacier mass change. *Earth's Future*, 8(7): e2019EF001470, 2020.

[232] J. Mascaro, G. P. Asner, D. E. Knapp, T. Kennedy-Bowdoin, R. E. Martin, C. Anderson, M. Higgins, and K. D. Chadwick. A tale of two "forests": Random forest machine learning aids tropical forest carbon mapping. *PloS one*, 9(1):e85993, 2014.

[233] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa. Pansharpening by convolutional neural networks. *Remote Sensing*, 8(7):594, 2016.

[234] A. Mathew, P. Amudha, and S. Sivakumari. Deep learning techniques: An overview. In A. Hassanien, R. Bhatnagar, and A. Darwish, editors, *Advanced Machine Learning Technologies and Applications*, volume 1141 of *Advances in Intelligent Systems and Computing*, pages 609–618. Springer, Singapore, 2021. doi: 10.1007/978-981-15-3383-9_54. AMLTA 2020.

[235] V. Mazzia, A. Khaliq, and M. Chiaberge. Improvement in land cover and crop classification based on temporal features learning from sentinel-2 data using recurrent-convolutional neural network (r-cnn). *Applied Sciences*, 10(1), 2020. ISSN 2076-3417. doi: 10.3390/app10010238. URL https://www.mdpi.com/2076-3417/10/1/238.

[236] W. McKinney. Data structures for statistical computing in python. In S. van der Walt and J. Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56–61, 2010.

[237] R. E. McRoberts, S. V. Stehman, G. C. Liknes, E. Næsset, C. Sannier, and B. F. Walters. The effects of imperfect reference data on remote sensing-assisted estimators of land cover class proportions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 142:292–300, 2018. ISSN 0924-2716. doi: https://doi.org/10.1016/j.isprsjprs.2018.06.002. URL https://www.sciencedirect.com/science/article/pii/S0924271618301655.

[238] X. Miao, X. Qiu, S.-S. Wu, J. Luo, D. R. Gouzie, and H. Xie. Developing efficient procedures for automated sinkhole extraction from lidar dems. *Photogramm. Eng. Remote Sens.*, 79:545–554, 2013.

[239] D. H. T. Minh, D. Ienco, R. Gaetano, N. Lalande, E. Ndikumana, F. Osman, and P. Maurel. Deep recurrent neural networks for winter vegetation quality mapping via multitemporal sar sentinel-1. *IEEE Geoscience and Remote Sensing Letters*, 15 (3):464–468, 2018.

[240] S. Mohan. Radar remote sensing for earth and planetary science. *International Journal of Scientific and Engineering Research*, 4:212, 12 2013.

[241] G. Mountrakis, J. Im, and C. Ogole. Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66:247–259, 05 2011. doi: 10.1016/j.isprsjprs.2010.11.001.

[242] F. Mumuni and A. Mumuni. Segment anything model for automated image data annotation: empirical studies using text prompts from grounding dino, 2024. URL https://arxiv.org/abs/2406.19057.

[243] Munich University of Applied Sciences AI Lab. Hawki chatgpt. URL https://ai.lab.hm.edu/login. Accessed: 2025-06-28.

[244] F. Müller. Zonation in the accumulation area of the glaciers of axel heiberg island, n.w.t., canada. *Journal of Glaciology*, 4(33):302–311, 1962. doi: 10.3189/S0022143000027623.

[245] J. Müller, H. Bußler, and T. Kneib. Saproxylic beetle assemblages related to silvicultural management intensity and stand structures in a beech forest in southern germany. *Journal of Insect Conservation*, 12:107–124, April 2008. ISSN 1572-9753. doi: 10.1007/s10841-006-9065-2. URL https://doi.org/10.1007/s10841-006-9065-2.

[246] F. Nencini, A. Garzelli, S. Baronti, and L. Alparone. Remote sensing image fusion using the curvelet transform. *Information Fusion*, 8(2):143–156, 2007. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2006.02.001. URL https://www.sciencedirect.com/science/article/pii/S1566253506000340. Special Issue on Image Fusion: Advances in the State of the Art.

[247] L. Niu, X. Tang, S. Yang, Y. Zhang, L. Zheng, and L. Wang. Detection of antarctic surface meltwater using sentinel-2 remote sensing images via u-net with attention

blocks: A case study over the amery ice shelf. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023. doi: 10.1109/TGRS.2023.3275076.

[248] A. Nowakowski, M. P. Del Rosso, P. Zachar, D. Spiller, G. Gab, D. Barretta, K. Kalinowska, K. Choromański, A. Wilkowski, A. Sebastianelli, P. Kupidura, K. Osińska-Skotak, and S. Ullo. Transfer learning in earth observation data analysis: A review. *IEEE Geoscience and Remote Sensing Magazin*, pages 2–33, 2024. doi: 10.1109/MGRS.2024.3494673.

[249] H. Ouchra, A. Belangour, and A. Erraissi. Machine learning algorithms for satellite image classification using google earth engine and landsat satellite data: Morocco case study. *IEEE Access*, 11:71127–71142, 2023. doi: 10.1109/ACCESS.2023.3293828.

[250] S. V. Panno and D. E. Luman. Mapping palimpsest karst features on the illinois sinkhole plain using historical aerial photography. *Carbonates Evaporites*, 28: 201–214, 2013.

[251] M. Parise. A procedure for evaluating the susceptibility to natural and anthropogenic sinkholes. *Georisk*, (9):272–285, 2015.

[252] G. Parrella, I. Hajnsek, and K. P. Papathanassiou. Model-based interpretation of polsar data for the characterization of glacier zones in greenland. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:11593–11607, 2021. doi: 10.1109/JSTARS.2021.3126069.

[253] W. S. B. Paterson. *The Physics of Glaciers*. Elsevier Science Ltd., Oxford, 3 edition, 1994.

[254] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

[255] C. Pelletier, G. I. Webb, and F. Petitjean. Temporal convolutional neural network for the classification of satellite image time series. *Remote Sensing*, 11(5), 2019. ISSN 2072-4292. doi: 10.3390/rs11050523. URL https://www.mdpi.com/2072-4292/11/5/523.

[256] N. Pettorelli et al. Satellite remote sensing. Technical Report 4, Conservation Technology Series WWF-UK, WWF-UK, 2018.

[257] B. Pogoda, S. Hauser, M. Rothe, F. Bakker, T. Hausen, B. Colsoul, K. Heinicke, and R. Pesch. GIS-based suitability modelling for the European oyster within the German exclusive zone of the North Sea [gis-basierte modellierung von eignungsflächen für die wiederansiedlung der europäischen auster in der awz der nordsee]. *gis.Science - Die Zeitschrift für Geoinformatik*, 2022(2):47–62, 2022.

[258] B. Pogoda, T. Hausen, M. Rothe, F. Bakker, S. Hauser, B. Colsoul, M. Dureuil, J. Krause, K. Heinicke, C. Pusch, S. Eisenbarth, A. Kreutle, C. Peter, and R. Pesch. Come, tell me how you live: Habitat suitability analysis for restoration. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 33(7):678–695, 2023. doi: https://doi.org/10.1002/aqc.3928. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/aqc.3928.

[259] C. Pohl and J. L. Van Genderen. Multisensor image fusion in remote sensing: Concepts, methods, and applications. *Int. J. Remote Sens.*, 19(5):823–854, 1998. doi: 10.1080/014311698215748.

[260] M. T. Pratola, H. A. Chipman, J. R. Gattiker, D. M. Higdon, R. McCulloch, and W. N. Rust. Parallel bayesian additive regression trees. *Journal of Computational and Graphical Statistics*, 23(3):830–852, 2014. ISSN 10618600. URL http://www.jstor.org/stable/43304924.

[261] C. Prieur, A. Rabatel, J.-B. Thomas, I. Farup, and J. Chanussot. Machine learning approaches to automatically detect glacier snow lines on multi-spectral satellite images. *Remote Sensing*, 14(16), 2022. ISSN 2072-4292. doi: 10.3390/rs14163868. URL https://www.mdpi.com/2072-4292/14/16/3868.

[262] R. Princess, S. Kumar, and R. Begum. Comprehensive and comparative study of different image fusion techniques. *International Journal of Advanced Research in Electrical Electronics and Instrumentation Engineering*, 3:11800–11806, 09 2014. doi: 10.15662/ijareeie.2014.0309015.

[263] O. Pueyo-Anchuela, A. M. Casas-Sainz, M. A. Soriano, and A. Pocoví-Juan. A geophysical survey routine for the detection of doline areas in the surroundings of zaragoza, ne spain. *Eng. Geol.*, 114:382–396, 2010.

[264] T. Purevdorj, R. Tateishi, T. Ishiyama, and Y. Honda. Relationships between percent vegetation cover and vegetation indices. *International journal of remote sensing*, 19(18):3519–3535, 1998.

[265] C. Qiu, X. Zhang, X. Tong, N. Guan, X. Yi, K. Yang, J. Zhu, and A. Yu. Few-shot remote sensing image scene classification: Recent advances, new baselines, and future trends. *ISPRS Journal of Photogrammetry and Remote Sensing*, 209:368–382, 2024. ISSN 0924-2716. doi: 10.1016/j.isprsjprs.2024.02.005. URL https://www.sciencedirect.com/science/article/pii/S0924271624000509.

[266] C. Rao, J. Rao, A. Kumar, D. Jain, and V. Dadhwal. Satellite image fusion using fast discrete curvelet transforms. In *2014 IEEE International Advance Computing Conference (IACC)*, pages 952–957, 2014. doi: 10.1109/IAdCC.2014.6779451.

[267] F. Rau. *Schneeeigenschaften und Gletscherzonen der Antarktischen Halbinsel im Radarbild*. PhD thesis, Albert-Ludwigs-Universität Freiburg, Freiburg i. Brsg., 2004.

[268] F. Rau et al. Radar glacier zones and their boundaries as indicators of glacier mass balance and climatic variability. In *Proceedings of EARSeL-SIG-Workshop Land Ice and Snow*, volume 1, page 317, Dresden, Germany, 2000.

[269] B. H. Raup, A. Racoviteanu, S. J. S. Khalsa, C. Helm, R. Armstrong, and Y. Arnaud. The glims geospatial glacier database: A new tool for studying glacier change. *Global and Planetary Change*, 56(1-2):101–110, 2007. doi: 10.1016/j.gloplacha.2006.07.018.

[270] G. Reiersen, D. Dao, B. Lütjens, K. Klemmer, K. Amara, A. Steinegger, C. Zhang, and X. Zhu. Reforestree: A dataset for estimating tropical forest carbon stock with deep learning and aerial imagery. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36:12119–12125, 06 2022. doi: 10.1609/aaai.v36i11.21471.

[271] J. Reitberger, C. Schnörr, P. Krzystek, and U. Stilla. 3d segmentation of single trees exploiting full waveform lidar data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 64(6):561–574, 2009.

[272] RGI Consortium. Randolph glacier inventory—a dataset of global glacier outlines: Version 6.0, 2017. URL https://doi.org/10.7265/4m1f-gd79. Version 6.0.

[273] M. Rieger. Classification and mapping of the devon ice cap based on terrasar-x data from 2017 to 2020, 2020.

[274] F. Rottensteiner, G. Sohn, M. Gerke, and J. D. Wegner. *ISPRS Test Project on Urban Classification and 3D Building Reconstruction*. ISPRS - Commission III - Photogrammetric Computer Vision and Image Analysis, Working Group III / 4 - 3D Scene Analysis.

[275] D. R. Rounce, R. Hock, F. Maussion, R. Hugonnet, W. Kochtitzky, M. Huss, E. Berthier, D. Brinkerhoff, L. Compagno, L. Copland, et al. Global glacier change in the 21st century: Every increase in temperature matters. *Science*, 379(6627): 78–83, 2023.

[276] T. Roßberg and M. Schmitt. Dense ndvi time series by fusion of optical and sar-derived data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:7748–7758, 2024. doi: 10.1109/JSTARS.2024.3379838.

[277] M. Rug. Evaluating ai-based approaches for doline detection. Master's thesis, Munich University of Applied Sciences, Munich, Germany, 2024. Bachelor's Thesis.

[278] M. Ruhhammer, S. Hauser, A. Schmitt, and A. Wendleder. Forest parameter estimation from dual-frequency polarimetric sar. In *Proceedings of the 15th European Conference on Synthetic Aperture Radar (EUSAR 2024)*, pages 966–971, Munich, Germany, 2024. ISBN 978-3-8007-6287-3.

[279] M. Rußwurm and M. Korner. Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 11–19, 2017.

[280] M. Rybakov, Y. Rotstein, B. Shirman, and A. Al-Zoubi. Cave detection near the dead sea: A micromagnetic feasibility study. *Lead. Edge*, 6:585–590, 2005.

[281] T. L. Saaty. A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology*, 15(3):234–281, 1977. ISSN 0022-2496. doi: https://doi.org/10.1016/0022-2496(77)90033-5. URL https://www.sciencedirect.com/science/article/pii/0022249677900335.

[282] F. Samadzadegan, A. Toosi, F. Dadrass Javan, and A. Stein. Decision-based fusion of pansharpened vhr satellite images using two-level rolling self-guidance filtering and edge information. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, X-4/W1-2022:691–698, 2023. doi: 10.5194/isprs-annals-X-4-W1-2022-691-2023. URL https://isprs-annals.copernicus.org/articles/X-4-W1-2022/691/2023/.

[283] N. Saxena, G. Saxena, N. Khare, and M. H. Rahman. Pansharpening scheme using spatial detail injection–based convolutional neural networks. *IET Image Processing*, 16(9):2297–2307, 2022. doi: https://doi.org/10.1049/ipr2.12384. URL https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/ipr2.12384.

[284] K. Schindler. An overview and comparison of smooth labeling methods for land-cover classification. *IEEE Transactions on Geoscience and Remote Sensing*, 50(11): 4534–4545, 2012. doi: 10.1109/TGRS.2012.2192741.

[285] A. Schmitt. Multiscale and multidirectional multilooking for sar image enhancement. *IEEE Transactions on Geoscience and Remote Sensing*, 54(9):5117–5134, 2016. doi: 10.1109/TGRS.2016.2555624.

[286] A. Schmitt and A. Wendleder. Sar-sharpening in the kennaugh framework applied to the fusion of multi-modal sar and opticle images. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-1:133–140, 2018. doi: 10.5194/isprs-annals-IV-1-133-2018. URL https://isprs-annals.copernicus.org/articles/IV-1/133/2018/.

[287] A. Schmitt, B. Wessel, and A. Roth. An innovative curvelet-only-based approach for automated change detection in multi-temporal sar imagery. *Remote Sensing*, 6 (3):2435–2462, 2014. ISSN 2072-4292. doi: 10.3390/rs6032435. URL https://www.mdpi.com/2072-4292/6/3/2435.

[288] A. Schmitt, A. Wendleder, and S. Hinz. The kennaugh element framework for multi-scale, multi-polarized, multi-temporal and multi-frequency sar image preparation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 102:122–139, 2015. ISSN 0924-2716. doi: 10.1016/j.isprsjprs.2015.01.007. URL https://www.sciencedirect.com/science/article/pii/S0924271615000246.

[289] A. Schmitt, A. Wendleder, R. Kleynmans, M. Hell, A. Roth, and S. Hinz. Multi-source and multi-temporal image fusion on hypercomplex bases. *Remote Sensing*, 12(6), 2020. ISSN 2072-4292. doi: 10.3390/rs12060943. URL https://www.mdpi.com/2072-4292/12/6/943.

[290] M. Schmitt and X. Zhu. Data fusion and remote sensing: An ever-growing relationship. *IEEE Geoscience and Remote Sensing Magazine*, 4(4):6–23, 2016. ISSN 2473-2397. doi: 10.1109/MGRS.2016.2561021. Publisher Copyright: © 2013 IEEE.

[291] M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu. Sen12ms–a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion. In *PIA19: Photogrammetric Image Analysis*, volume IV2/W7 of *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, pages 153–160, 2019. URL https://elib.dlr.de/133280/.

[292] M. Schmitt, S. A. Ahmadi, Y. Xu, G. Taşkin, U. Verma, F. Sica, and R. Hänsch. There are no data like more data: Datasets for deep learning in earth observation. *IEEE Geoscience and Remote Sensing Magazine (GRSM)*, 11(3):63–97, August 2023. URL https://elib.dlr.de/200091/.

[293] L. Scholleer, A. Schmitt, A. Wendleder, and S. Rogginger. Schritthaltende baufallerkundung aus dem all mit frei verfügbaren satellitenaufnahmen. *ZFV - Zeitschrift für Geodasie, Geoinformation und Landmanagement*, (3):168–180, März 2022. URL https://elib.dlr.de/187221/.

[294] A. Schörgenhumer, M. Kahlhofer, P. Chalupar, P. Grünbacher, and H. A. Mössenböck. Framework for preprocessing multivariate, topology-aware time series and event data in a multi-system environment. In *IEEE 19th International Symposium on High Assurance Systems Engineering (HASE)*, pages 115–122, 2019. doi: 10.1109/HASE.2019.00026.

[295] K. Schürholt, D. Taskiran, B. Knyazev, X. G. i Nieto, and D. Borth. Model zoos: A dataset of diverse populations of neural network models, 2022.

[296] A. Sebastianelli, A. Nowakowski, E. Puglisi, M. P. D. Rosso, J. Mifdal, F. Pirri, P. P. Mathieu, and S. L. Ullo. Spatio-temporal sar-optical data fusion for cloud removal via a deep hierarchical model, 2022. URL https://arxiv.org/abs/2106.12226.

[297] R. Sedona, G. Cavallaro, J. Jitsev, A. Strube, M. Riedel, and J. A. Benediktsson. Remote sensing big data classification with high performance distributed deep learning. *Remote Sensing*, 11(24), 2019. ISSN 2072-4292. doi: 10.3390/rs11243056. URL https://www.mdpi.com/2072-4292/11/24/3056.

[298] M. R. Segal, J. D. Barbour, and R. M. Grant. Relating hiv-1 sequence variation to replication capacity via trees and forests. *Statistical applications in genetics and molecular biology*, 3(1), 2004.

[299] M. P. Sesmero, A. I. Ledezma, and A. Sanchis. Generating ensembles of heterogeneous classifiers using stacked generalization. *WIREs Data Mining and Knowledge Discovery*, 5(1):21–34, 2015. doi: https://doi.org/10.1002/widm.1143. URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1143.

[300] J. J. Sharma, I. Hajnsek, K. P. Papathanassiou, and A. Moreira. Polarimetric decomposition over glacier ice using long-wavelength airborne polsar. *IEEE Transactions on Geoscience and Remote Sensing*, 49(1):519–535, 2011. doi: 10.1109/TGRS.2010.2056692.

[301] M. Sharp, D. O. Burgess, F. Cawkwell, L. Copland, J. A. Davis, E. K. Dowdeswell, et al. Remote sensing of recent glacier changes in the canadian arctic. In J. S. Kargel, M. P. Bishop, A. Kääb, and B. Raup, editors, *Global Land Ice Measurements from Space*, Springer Praxis Books, pages 205–228. Springer, Dordrecht, 2013.

[302] M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, and S. Homayouni. Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:6308–6325, 2020. doi: 10.1109/JSTARS.2020.3026724.

[303] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, pages 802–810, 2015. URL https://proceedings.neurips.cc/paper/2015/file/07563a3fe3bbe7e3ba84431ad9d055af-Paper.pdf.

[304] E. Siero. Resolving soil and surface water flux as drivers of pattern formation in turing models of dryland vegetation: A unified approach. *Physica D: Nonlinear Phenomena*, 414:132695, 2020. ISSN 0167-2789. doi: https://doi.org/10.1016/j.physd.2020.132695. URL https://www.sciencedirect.com/science/article/pii/S0167278919306505.

[305] J. Sigurdsson, S. E. Armannsson, M. O. Ulfarsson, and J. R. Sveinsson. Fusing sentinel-2 and landsat 8 satellite images using a model-based method. *Remote Sensing*, 14(13), 2022. ISSN 2072-4292. doi: 10.3390/rs14133224. URL https://www.mdpi.com/2072-4292/14/13/3224.

[306] V. P. Singh, P. Singh, and U. K. Haritashya, editors. *Encyclopedia of Snow, Ice and Glaciers*. Encyclopedia of Earth Sciences Series. Springer Dordrecht, 1 edition, 2011. ISBN 978-90-481-2641-5. doi: 10.1007/978-90-481-2642-2. 1253 pages, 231 b/w and 428 color illustrations.

[307] T. Slater, I. R. Lawrence, I. N. Otosaka, A. Shepherd, N. Gourmelen, L. Jakob, P. Tepes, L. Gilbert, and P. Nienow. Earth's ice imbalance. *The Cryosphere*, 15(1): 233–246, 2021.

[308] H. Song, B. Huang, Q. Liu, and K. Zhang. Improving the spatial resolution of landsat tm/etm+ through fusion with spot5 images via learning-based super-

resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 53(3):1195–1204, 2015. doi: 10.1109/TGRS.2014.2335818.

[309] H. Song, H. Xie, Y. Duan, X. Xie, F. Gan, W. Wang, and J. Liu. Pure data correction enhancing remote sensing image classification with a lightweight ensemble model. *Scientific Reports*, 15(1):5507, 2025. ISSN 2045-2322. doi: 10.1038/s41598-025-89735-1. URL https://doi.org/10.1038/s41598-025-89735-1.

[310] Z. Song, Y. Lu, Z. Ding, D. Sun, Y. Jia, and W. Sun. A new remote sensing desert vegetation detection index. *Remote Sensing*, 15(24), 2023. ISSN 2072-4292. doi: 10.3390/rs15245742. URL https://www.mdpi.com/2072-4292/15/24/5742.

[311] M. Spasari. Terrasar-x basierte klassifizierung von gletscherzonen: Monitoring von axel heiberg, devon ice cap und manson icefield zwischen 2017 und 2023, 2024. Bachelor's Thesis.

[312] R. I. P. Specifications. 2016, url = https://sg.geodatenzentrum.de/public/gdz/dokumentation/deu RapidEye, urldate = 2025-01-10,.

[313] E. Sreehari and S. Srivastava. Prediction of climate variable using multiple linear regression. In *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, pages 1–4, 2018. doi: 10.1109/CCAA.2018.8777452.

[314] C. Strobl, J. Malley, and G. Tutz. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods*, 14(4):323, 2009.

[315] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 5901–5904, 2019. doi: 10.1109/IGARSS.2019.8900532.

[316] Z. Sun, L. Di, and H. Fang. Using long short-term memory recurrent neural network in land cover classification on landsat and cropland data layer time series. *International journal of remote sensing*, 40(2):593–614, 2019.

[317] C. M. Surdu, C. R. Duguay, and D. Fernández Prieto. evidence of recent changes

in the ice regime of lakes in the canadian high arctic from spaceborne satellite observations. *The Cryosphere*, 10(3):941–960, 2016. doi: 10.5194/tc-10-941-2016. URL https://tc.copernicus.org/articles/10/941/2016/.

[318] U. S. G. Survey. Landsat—earth observation satellites. In *U. S. Geological Survey Fact Sheet 2015–3081*. Reston, VA, USA, , note = 2015, doi = 10.3133/fs20153081„ ; U. S. Geological Survey.

[319] Svevind Energy Group. Reference data for sinkhole mapping: Provided shapefile data and georeferenced imagery of known sinkholes, 2023. Personal communication.

[320] V. Syrris, O. Pesek, and P. Soille. Satimnet: Structured and harmonised training data for enhanced satellite imagery classification. *Remote Sensing*, 12(20):3358, October 2020. ISSN 2072-4292. doi: 10.3390/rs12203358. URL http://dx.doi.org/10.3390/rs12203358.

[321] M. Taviani, L. Angeletti, E. Campiani, A. Ceregato, F. Foglini, V. Maselli, M. Morsilli, M. Parise, and F. Trincardi. Drowned karst landscapes offshore the apulian margin (southern adriatic sea, italy). *J. Cave Karst Stud.*, 74:197–212, 2012.

[322] M. M. Taye. Understanding of machine learning with deep learning: Architectures, workflow, applications and future directions. *Computers*, 12(5), 2023. ISSN 2073-431X. doi: 10.3390/computers12050091. URL https://www.mdpi.com/2073-431X/12/5/91.

[323] B. Theilen-Willige, H. Ait Malek, A. Charif, B. Fatima, and M. Chaibi. Remote sensing and gis contribution to the investigation of karst landscapes in nw-morocco. *Geosciences*, 4, 2014.

[324] J. Tigges, T. Lakes, and P. Hostert. Urban vegetation classification: Benefits of multitemporal rapideye satellite data. *Remote Sensing of Environment*, 136:66–75, 2013. doi: 10.1016/j.rse.2013.04.004. URL https://www.sciencedirect.com/science/article/pii/S0034425713001429.

[325] C. J. Tucker. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, 8(2):127–150, 1979. doi: 10.1016/0034-4257(79)90013-0.

[326] T. Ullmann, A. Schmitt, A. Roth, J. Duffe, S. Dech, H.-W. Hubberten, and R. Baumhauer. Land cover characterization and classification of arctic tundra environments by means of polarized synthetic aperture x- and c-band radar (polsar) and landsat 8 multispectral imagery — richards island, canada. *Remote Sensing*, 6(9):8565–8593, 2014. ISSN 2072-4292. doi: 10.3390/rs6098565. URL https://www.mdpi.com/2072-4292/6/9/8565.

[327] United Nations. Transforming our world: The 2030 agenda for sustainable development. https://digitallibrary.un.org/record/1654217, 2015. Resolution adopted by the UN General Assembly on 25 September 2015, A/RES/70/1.

[328] U.S. Department of the Army. *Soviet Topographic Map Symbols*. Arlington, VA, 1958. URL https://books.google.de/books?id=aX8_AAAAIAAJ. Available online.

[329] M. Ustuner, F. B. Sanli, S. Abdikan, M. T. Esetili, and Y. Kurucu. Crop type classification using vegetation indices of rapideye imagery. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 40 (7):195–199, 2014. doi: 10.5194/isprsarchives-XL-7-195-2014.

[330] R. J. B. Vaibhav R. Pandit. Image fusion in remote sensing applications: A review. *International Journal of Computer Applications*, 120(10):22–32, June 2015. ISSN 0975-8887. doi: 10.5120/21263-3846. URL https://ijcaonline.org/archives/volume120/number10/21263-3846/.

[331] W. O. van der Knaap, J. F. N. van Leeuwen, L. Fahse, S. Szidat, T. Studer, J. Baumann, M. Heurich, and W. Tinner. Vegetation and disturbance history of the bavarian forest national park, germany. *Vegetation History and Archaeobotany*, 29(2):277–295, March 2020. doi: 10.1007/s00334-019-00742-5. URL https://doi.org/10.1007/s00334-019-00742-5.

[332] J. van der Zee, D. Marcos, and E. Siero. Estimating parameters of a spatial dryland vegetation model from time series of satellite images using differentiable programming. In *ESA Living Planet Symposium 2025*, Vienna, Austria, June 2025. Oral presentation.

[333] J. van Genderen and C. Pohl. Image fusion : issues, techniques and applications : a selected bibliography on image fusion. In *EARSeL workshop on intelligent image fusion : 11 September 1994, Strasbourg, France, 12 p.*, 1994. In: EARSeL workshop on intelligent image fusion : 11 September 1994, Strasbourg, France, 12 p.

[334] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1 edition, 1995. ISBN 978-1-4757-2440-0. doi: 10.1007/978-1-4757-2440-0. URL https://doi.org/10.1007/978-1-4757-2440-0. Springer Science+Business Media New York. Published online: April 17, 2013.

[335] O. G. Varga, Z. Kovács, L. Bekő, P. Burai, Z. Csatáriné Szabó, I. Holb, S. Ninsawat, and S. Szabó. Validation of visually interpreted corine land cover classes with

spectral values of satellite images and machine learning. *Remote Sensing*, 13(5), 2021. ISSN 2072-4292. doi: 10.3390/rs13050857. URL https://www.mdpi.com/2072-4292/13/5/857.

[336] M. Varma, S. Jyothi, and A. Sibyala. Data preprocessing in multi-temporal remote sensing data for deforestation analysis data preprocessing in multi-temporal remote sensing data for deforestation analysis strictly as per the compliance and regulations of: Data preprocessing in multi-temporal remote sensing data for deforestation analysis. *Global Journal of Computer Science and Technology Software and Data Engineering*, 13, 01 2013.

[337] M. Volpi and D. Tuia. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2):881–893, 2016.

[338] K. Von Schuckmann, A. Minière, F. Gues, F. J. Cuesta-Valero, G. Kirchengast, S. Adusumilli, F. Straneo, M. Ablain, R. P. Allan, P. M. Barker, et al. Heat stored in the earth system 1960–2020: where does the energy go? *Earth System Science Data*, 15(4):1675–1709, 2023.

[339] L. Wald. Some terms of reference in data fusion. *Geoscience and Remote Sensing, IEEE Transactions on*, 37:1190 – 1193, 06 1999. doi: 10.1109/36.763269.

[340] A. C. Waltham and P. G. Fookes. Engineering classification of karst ground conditions. *Q. J. Eng. Geol. Hydrogeol*, 36:101–118, 2003.

[341] J. T. Walton. Subpixel urban land cover estimation. *Photogrammetric Engineering & Remote Sensing*, 74(10):1213–1222, 2008.

[342] H. Wang, H. Li, W. Qian, W. Diao, L. Zhao, J. Zhang, et al. Dynamic pseudo-label generation for weakly supervised object detection in remote sensing images. *Remote Sens.*, 13(8):1461, 2021. doi: 10.3390/rs13081461.

[343] H. Wang, Y. Cheng, X. Liu, and X. Wang. Reinforcement learning based markov edge decoupled fusion network for fusion classification of hyperspectral and lidar. *IEEE Transactions on Multimedia*, PP:1–13, 01 2024. doi: 10.1109/TMM.2024.3360717.

[344] L. Wang, M. J. Sharp, B. Rivard, S. Marshall, and D. Burgess. Melt season duration on canadian arctic ice caps, 2000–2004. *Geophysical Research Letters*, 32(19),

2005. doi: https://doi.org/10.1029/2005GL023962. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2005GL023962.

[345] S. Wang, C. Hou, Y. Chen, Z. Liu, Z. Zhang, and G. Zhang. Classification of hyperspectral and lidar data using multi-modal transformer cascaded fusion net. *Remote Sensing*, 15(17), 2023. ISSN 2072-4292. doi: 10.3390/rs15174142. URL https://www.mdpi.com/2072-4292/15/17/4142.

[346] W. Wang, W. Zeng, Y. Huang, X. Ding, and J. Paisley. Deep blind hyperspectral image fusion. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4149–4158, 2019. doi: 10.1109/ICCV.2019.00425.

[347] W. Wang, C. Li, P. Ren, X. Lu, J. Wang, G. Ren, and B. Liu. Dual-branch feature fusion network based cross-modal enhanced cnn and transformer for hyperspectral and lidar classification. *IEEE Geoscience and Remote Sensing Letters*, 21:1–5, 2024. doi: 10.1109/LGRS.2024.3367171.

[348] Z. Wang, W. Yan, and T. Oates. Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International joint conference on neural networks (IJCNN)*, pages 1578–1585. IEEE, 2017.

[349] J. Wei, X. Liu, and B. Zhou. Sensitivity of vegetation to climate in mid-to-high latitudes of asia and future vegetation projections. *Remote Sensing*, 15(10), 2023. ISSN 2072-4292. doi: 10.3390/rs15102648. URL https://www.mdpi.com/2072-4292/15/10/2648.

[350] A. Wendleder, A. Schmitt, T. Erbertseder, P. d'Angelo, C. Mayer, and M. Braun. Seasonal evolution of supraglacial lakes on baltoro glacier from 2016 to 2020. *Frontiers in Earth Science*, 9:1–16, Dezember 2021. URL https://elib.dlr.de/145929/.

[351] A. Wendleder, A. Schmitt, T. Erbertseder, P. D'Angelo, C. Mayer, and M. H. Braun. Seasonal evolution of supraglacial lakes on baltoro glacier from 2016 to 2020. *Frontiers in Earth Science*, 9, 2021. ISSN 2296-6463. doi: 10.3389/feart.2021.725394. URL https://www.frontiersin.org/journals/earth-science/articles/10.3389/feart.2021.725394.

[352] A. Wendleder, V. Mix, and A. Schmitt. The glacier zone index applied on the manson icefield. In *EUSAR 2022; 14th European Conference on Synthetic Aperture Radar*, pages 1–5, 2022.

[353] A. Wendleder, A. Schmitt, T. Erbertseder, P. d'Angelo, C. Mayer, and M. Braun. Mapping the seasonal evolution of supraglacial lakes on baltoro glacier from 2016 to 2020. In *Living Planet Symposium 2022*, page 1, 2022. URL https://elib.dlr.de/187216/.

[354] W. Wright. *Fast image fusion with a Markov random field*, pages 557–561. 1999. doi: 10.1049/cp:19990384. URL https://digital-library.theiet.org/doi/abs/10.1049/cp%3A19990384.

[355] Z. Wu, Y. Huang, and K. Zhang. Remote sensing image fusion method based on pca and curvelet transform. *Journal of the Indian Society of Remote Sensing*, 46, 01 2018. doi: 10.1007/s12524-017-0736-0.

[356] M. Xiao, Y. Wu, G. Zuo, S. Fan, H. Yu, Z. A. Shaikh, Z. Wen, and D. M. Shafiq. Addressing overfitting problem in deep learning-based solutions for next generation data-driven networks. *Wirel. Commun. Mob. Comput.*, 2021, Jan. 2021. ISSN 1530-8669. doi: 10.1155/2021/8493795. URL https://doi.org/10.1155/2021/8493795.

[357] Y. Yang, J. Qu, W. Dong, T. Zhang, S. Xiao, and Y. Li. Tmcfn: Text-supervised multidimensional contrastive fusion network for hyperspectral and lidar classification. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024. doi: 10.1109/TGRS.2024.3374372.

[358] H. Yessou, G. Sumbul, and B. Demir. A comparative study of deep learning loss functions for multi-label remote sensing image classification. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Waikola, HI, USA, 26 September-2 October*, pages 1349–1352, 2020. doi: 10.1109/IGARSS39084.2020.9323583.

[359] J. Yin, J. Dong, N. A. Hamm, Z. Li, J. Wang, H. Xing, and P. Fu. Integrating remote sensing and geospatial big data for urban land use mapping: A review. *International Journal of Applied Earth Observation and Geoinformation*, 103:102514, 2021. ISSN 1569-8432. doi: 10.1016/j.jag.2021.102514. URL https://www.sciencedirect.com/science/article/pii/S030324342100221X.

[360] R. Zangl, S. Hauser, and A. Schmitt. Guide to the practical use of image data fusion in remote sensing. *GIS Sci.*, 4:123–147, 2022.

[361] J. Zhang. Multi-source remote sensing data fusion: Status and trends. *International Journal of Image and Data Fusion*, 1:5–24, 03 2010. doi: 10.1080/19479830903561035.

[362] L. Zhang and L. Zhang. Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities. *IEEE Geoscience and Remote Sensing Magazine*, 10(2):270–294, 2022. doi: 10.1109/MGRS.2022.3145854.

[363] Q. Zhang, Q. Yuan, C. Zeng, X. Li, and Y. Wei. Missing data reconstruction in remote sensing image with a unified spatial–temporal–spectral deep convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 56(8): 4274–4288, 2018.

[364] Y. Zhang, J. Liu, and W. Shen. A review of ensemble learning algorithms used in remote sensing applications. *Applied Sciences*, 12(17), 2022. ISSN 2076-3417. doi: 10.3390/app12178654. URL https://www.mdpi.com/2076-3417/12/17/8654.

[365] L. Zhao, Y. Zhou, W. Zhong, C. Jin, B. Liu, and F. Li. A spatio-temporal deep learning model for automatic arctic sea ice classification with sentinel-1 sar imagery. *Remote Sensing*, 17(2), 2025. ISSN 2072-4292. doi: 10.3390/rs17020277. URL https://www.mdpi.com/2072-4292/17/2/277.

[366] Q. Zhao, S. Yu, F. Zhao, L. Tian, and Z. Zhao. Comparison of machine learning algorithms for forest parameter estimations and application for forest quality assessments. *Forest Ecology and Management*, 434:224–234, 2019.

[367] S. Zhao, X. Chen, S. Wang, J. Li, and W. Yang. A new method of remote sensing image decision-level fusion based on support vector machine. pages 91 – 96, 12 2003. ISBN 0-7803-8142-4. doi: 10.1109/RAST.2003.1303889.

[368] Z. Zheng, S. Ermon, D. Kim, L. Zhang, and Y. Zhong. Changen2: Multi-temporal remote sensing generative change foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(2):725–741, 2025. doi: 10.1109/TPAMI.2024.3475824.

[369] B. Zhong, L. Yang, X. Luo, J. Wu, and L. Hu. Extracting shrubland in deserts from medium-resolution remote-sensing data at large scale. *Remote Sensing*, 16(2), 2024. ISSN 2072-4292. doi: 10.3390/rs16020374. URL https://www.mdpi.com/2072-4292/16/2/374.

[370] J. Zhu, T. P. Taylor, J. C. Currens, and M. M. Crawford. Improved karst sinkhole mapping in kentucky using lidar techniques: A pilot study in floyds fork watershed. *J. Cave Karst Stud.*, 76:207–216, 2014.

[371] X. Zhu, J. Chen, F. Gao, X. Chen, J. G. Masek, and E. F. Vermote. An enhanced spatial and temporal adaptive reflectance fusion model for complex heterogeneous regions. *Remote Sens. Environ.*, 114(11):2610–2623, 2010. doi: 10.1016/j.rse.2010.05.032.

[372] X. X. Zhu, J. Hu, C. Qiu, Y. Shi, J. Kang, L. Mou, H. Bagheri, M. Haberle, Y. Hua, R. Huang, L. Hughes, H. Li, Y. Sun, G. Zhang, S. Han, M. Schmitt, and Y. Wang. So2sat lcz42: A benchmark data set for the classification of global local climate zones [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 8 (3):76–89, 2020. doi: 10.1109/MGRS.2020.2964708.

[373] M. Łucka, R. Hejmanowski, and W. Witkowski. Potential of employing a machine learning model for glacier motion monitoring. In *EGU General Assembly 2024*, Vienna, Austria, 2024. doi: 10.5194/egusphere-egu24-19059. URL https://doi.org/10.5194/egusphere-egu24-19059.

List of Figures

1.1 Overview of the Study Area in Southwestern Kazakhstan. 28
1.2 Overview of Steigerwald (AOI 1) and it's subdivisions. 38
1.3 Overview of the AOI 2 and it's subdivisions. **(a)** Raster tiles NP_D01–D06 covering deadwood variables in the Bavarian Forest National Park; **(b)** Raster tiles NP_T00–T11 covering structural forest variables; **(c)** Field transects 1–4 corresponding to ground reference measurements. 39
1.4 Map of study sites displaying the three designated areas of interest 40
1.5 Exemplary aggregation results of the tree segments onto the 10 m grid of the raster data, displaying the tree class (**l.**) and the crown volume (**r.**) of a subset in the Bavarian Forest National Park test site (AOI 2). 45
1.6 Ground truth from Field campaigns in the three designated study sites (a) Steigerwald, (b) Bavarian Forest National Park, and (c) Kranzberg Forest. . 46
1.7 Schematic representation of glaciological snow zones based on Paterson [253] and corresponding radar glacier zones based on Rau et al. [268]. . . . 51
1.8 Overview of the Canadian High Arctic glacier study areas, comprising Axel Heiberg Island and Devon Island. 53
1.9 Thesis Structure and Thematic Flow: The thesis is grounded in the systemic integration of EO features, reference labels, and ML methodologies. Chapter I establishes this foundation. Chapters II and III address the two core data pillars: multi-sensor EO features and robust labelling strategies. Chapter IV introduces the HELIX framework for dynamic label enrichment, while Chapter V explores temporal fusion and feature behaviour. Chapter VI systematically benchmarks EO modality–model interactions for continuous label prediction, while Chapter VII extends this analysis by evaluating the impact of context-aware label enrichment and multi-scale learning using the HELIX. The thesis concludes in Chapter VIII with a synthesis of findings across all experimental dimensions. 61

2.1 Stepwise example of hypercomplex spectral–polarimetric fusion. Sentinel-1 dual-pol inputs are first transformed into four polarimetric Kennaugh elements (S_1), while Sentinel-2 reflectances (shown here as example bands B2–B8) are transformed into four spectral Kennaugh-like elements (S_2) using the quaternion-based Hadamard matrix Q. Note that the optical input is not limited to these four bands; additional or alternative spectral channels can also be used following the same principle. Finally, both feature vectors (S_1, S_2) are orthogonally fused into the 8-dimensional spectral–polarimetric feature vector $\mathbf{F}$, as described in Equation (2.7). 82

2.2 Temporal extension of hypercomplex data fusion. Starting from the spectral–polarimetric fused feature vector $\mathbf{F}$ ($K_{\text{fused},0-7}$), a Hadamard-based temporal transform (Q_T) is applied across all T time steps. This results in a temporally enriched $T \times 8$-dimensional dataset ($K_{*,0-(T-1)}$), capturing both persistent and dynamic modes of variation. The temporal Hadamard matrix $Q_T \in \mathbb{R}^{T\times T}$ follows the same family of orthogonal hypercomplex bases (e.g., C, Q, O, S) as used in the spectral–polarimetric fusion, now extended to the temporal domain. The illustrated example shows $T = 4$; for the Wald5Dplus application, $T = 64$ acquisitions were used. 85

3.1 Labels from terrestrial surveys in the rural community of Hochstadt (Bavaria, Germany) in comparison to different EO data sources: (top-left) land use labels as provided by the *Bavarian Surveying Administration (Bayerische Vermessungsverwaltung)*—www.geodaten.bayern.de (accessed on 1 February 2025) (top-right) digital orthophoto 20 cm (DOP20 by the *Bavarian Surveying Administration (Bayerische Vermessungsverwaltung)*—www.geodaten.bayern.de (accessed on 1 February 2025); (bottom-left) Sentinel-2 (©ESA (2023)) true colour image (TCI); and (bottom-right) Sentinel-1 (©ESA (2023)) total intensity (K0). The figure elucidates the impact of image resolution and geometric co-registration on the usability of labels. On the one hand, the DOP20 shows much more details than the labels require; on the other, the satellite images are too coarse to capture the relatively narrow polygons of (e.g.) the traffic class. Regarding Sentinel-1, the signatures of high-rise objects like the buildings or trees are spatially overlaid with neighbouring polygons. 107

3.2 Labels from an airborne PolInSAR flight campaign over the German Wadden See around the island of Baltrum (Lower Saxony, Germany) in comparison to multi-temporal spaceborne optical acquisitions in the visible and near-infrared spectral range: (top-left) land cover labels [157] (accessed on 1 February 2025) with digital orthophoto 20 cm in the background (LGLN (2024)), and Colour Infrared (CIR) images by Sentinel-2 on September 2nd, 4th, and 7th (©ESA (2023)) as multi-temporal features. The figure impressively visualizes the high temporal variability of features acquired by spaceborne EO sensors due to the immanent tidal range opposite the temporally stable land cover classes. 108
3.3 Labels from airborne LiDAR in the Bavarian Forest National Park conditioned for use with spaceborne sensors: (top-left) single tree polygons derived from point clouds that contain the tree geometry and further characteristics as attributes. These labels concerning the Bavarian Forest National Park were provided by the Bavarian National Park Research under the Bohemian Forest Datapool Initiative [203] (accessed on 29 February 2024); (top-right) the 10 m pixel grid of the satellite data; (bottom-left) tree characteristics aggregated on the grid by the Wald5Dplus project [147] for use as labels; and (bottom-right) Kennaugh elements 1 to 3 of the 512 bands included in the Analysis-Ready Data (ARD) cube provided by Wald5Dplus [148] (accessed on 1 February 2025) for use as features. The figure addresses the two main labelling problems of Wald5Dplus: first, the gridded labels represent geospatial statistics instead of single tree characteristics; second, the multi-temporal EO features contain structures that are not visible in the mono-temporal labels and vice versa. 109

3.4 Labels from airborne photography: (top-left) manually drawn wind-throw areas after storm Kyrill in January 2007 categorized in single-tree, group, and areal wind-throw in the Bavarian Forest National Park. The labels concerning the Bavarian Forest National Park were provided by the Bavarian National Park Research under the Bohemian Forest Datapool Initiative [203] (accessed on 29 February 2024), with the ESRI World Topo Map in the background. The other sub-figures show Landsat True Colour Images (TCI) taken from space in the years 2007, 2009, and 2020 in parts (top-right sub-figure) with some clouds (Landsat 5 and 8 images courtesy of the U.S. Geological Survey). The reference data consist of overlapping polygons, which inhibits the assignment of clear label to pixels. Although the satellite image from summer 2007 takes up the structures of the labels, many more areas appear very similar to the mapped wind-throw areas, which underlines the necessity of multi-temporal features and/or the inclusion of static labels. The image from 2009 indicates clearing after the storm, whereas the image from 2020 reveals regrowth. 110

3.5 Multi-temporal labels from airborne photography: yearly deadwood after barkbeetle infestation mapped by a human interpreter based on stereoscopic images acquired during yearly airborne flight campaigns. The polygons delineate dead trees, categorized by the last date on which they were classified as healthy. Labels concerning the Bavarian Forest National Park were provided by the Bavarian National Park Research under the Bohemian Forest Datapool Initiative [203] (accessed on 29 February 2024). The raster image in the background contains the multi-temporal Normalized Difference Vegetation Index (red: NDVI in spring; green: NDVI in summer; blue: NDVI in autumn) from Sentinel-2 images (©ESA (2018, 2020, 2022, 2024)). The brightness shows the healthiness of the vegetation, whereas the hue shows its temporal variation, e.g., red stands for high photosynthetic activity in the spring and reduced photosynthetic activity in the summer and autumn. Dark areas stand for low-to-negligible photosynthetic activity throughout the year. This figure illustrates the challenges of temporal alignment; some upcoming deadwood areas are already visible in the space-borne time series, even though they are still classified as healthy by the yearly manual assessment. Thus, the image from 2024 (bottom-right) shows a composition of deadwood and regrowth areas that only partially match the reference polygons due to the increasing time lag. 111

4.1 Conceptual framework of the proposed HELIX framework, illustrating the integration and preprocessing pipeline for static and dynamic labels. 121
4.2 Illustration of STRtree hierarchy: Leaf nodes store bounding boxes of polygons (MBRs), which are grouped into internal nodes. The root node covers all inputs. Queries first test against parent MBRs to prune unnecessary comparisons. 125
4.3 STRtree spatial filtering: the raster cell C_{ij} (blue) queries polygon candidates by overlapping bounding boxes (dashed), reducing the need for unnecessary geometric operations. 125
4.4 **The helical window around a grid cell $C_{ij}^{(t)}$.** Here, $C_{ij}^{(t)}$ represents the target cell at position (i, j) and time step t. The helical window aggregates data from spatial neighbours within a radius at each time step across a symmetric temporal window. This produces contextual statistics for both space and time around each target cell, enabling robust modelling of local dynamics such as gradual deforestation or vegetation change. 130
4.5 Comparison of Fourier encodings. The joint embedding preserves seasonal smoothness across time. 131
4.6 Temporal window width adjusts to label sparsity, improving robustness in low-density intervals. 132
4.7 Cross-time comparisons encode temporal transitions, critical for dynamic event understanding. 133
4.8 Fractional spatial probability estimation. Two polygonal labels (Label A and Label B) overlap a shared grid. For each cell, the fraction $P_{ij}^{(L)}$ is computed by dividing the intersected area $A_{ij}^{(L)}$ by the cell area A_{cell}, allowing soft, probabilistic label assignment. 134
4.9 Illustration of local variance estimation. The variance value computed for the centre cell $P_{ij}^{(t)}$ reflects the degree of heterogeneity within its surrounding neighbourhood window $\mathcal{N}_{ij}^{(t)}$. Shading indicates the underlying spatial distribution of neighbour values contributing to the calculation: red for highly variable (unstable) neighbors, green for homogeneous (stable) areas. 135
4.10 Visual example of how confidence scores could be derived from label agreement. Multiple overlapping shapes represent input labels from different sources. Higher overlap implies higher agreement and thus higher label confidence. 140

5.1 Statistical test results (t-Test, Pearson, Kendall, Spearman, ANOVA) for all fusion configurations across bands $K_{\text{fused},0}$ to $K_{\text{fused},7}$. CDVI (SAR August + Optical March) yields consistently superior performance across metrics, especially in bands $K_{\text{fused},2}$, $K_{\text{fused},4}$, $K_{\text{fused},5}$, and $K_{\text{fused},6}$. 152
5.2 Distribution of CDVI values for dominant land cover classes in the study area: sinkholes, takyr surfaces, dense vegetation, sparse vegetation, and bare ground. Boxplots show distinct value ranges, highlighting strong class separability. 160
5.3 Spatial visualisation of CDVI performance: **(a)** CDVI values overlaid on World Imagery [100] for a sinkhole with sparse vegetation; **(b)** the same area without overlay for comparison. The CDVI distinguishes structural and vegetative features clearly. 161
5.4 Comparison of CDVI and Proba-V LC100 [51] vegetation cover fractions across quantiles (Q1–Q3). CDVI correlates strongly with Proba-V Total Vegetation ($r = 0.67$), with slightly weaker correlation for Grass and Shrub fractions. 162
5.5 Spatial differences between smoothed CDVI vegetation cover and Proba-V LC100 [51] Grass, Shrub, and Total Vegetation layers. Brown hues indicate higher CDVI values; green hues indicate higher Proba-V values. 163
5.6 CDVI overlay on Sentinel-2 (©ESA 2023) imagery. Green areas indicate vegetation, red areas mark detected sinkholes, and orange outlines show GPS-based sinkhole references [319]. 164

6.1 RF architecture, each decision tree is trained on a bootstrapped subset of the input data and outputs predictions for all targets. Final output is obtained through averaging or majority voting. 175
6.2 SVR architecture. A kernel function transforms the input space, and a regression function is fitted with an ε-insensitive margin (gray band). Only support vectors outside this tube contribute to the loss function. 176
6.3 1D-CNN architecture with convolutional layers showing filters, kernel sizes, and activations. 177
6.4 End-to-end fusion pipeline: Sentinel-1 and Sentinel-2 inputs are transformed into compatible Kennaugh representations and fused spectrally, structurally, and temporally using the Hypercomplex Bases method. 182
6.5 Typical spectral signatures of deciduous, coniferous, and dead trees [147]. . 184

6.6 Workflow for Stacked RF Meta-Model Development and Prediction. The pipeline includes EO and reference data preprocessing, sub-AOI tiling, feature engineering, individual model training, ensemble stacking, and final prediction through a meta-learner. 186
6.7 Mean Absolute Error (MAE) distributions for different preprocessing configurations predicting Sum crown volume of coniferous trees (m^2). Each configuration varies in mask threshold (M), Z-score filtering (Z), and aggressive filtering (A). Results are shown for two values of the number of estimators in the model: 50 and 150. Lower MAE values indicate better model performance. 189
6.8 Mean Absolute Error (MAE) distributions for different preprocessing configurations predicting Sum crown volume of deciduous trees (m^2). Each configuration varies in mask threshold (M), Z-score filtering (Z), and aggressive filtering (A). Results are shown for two values of the number of estimators in the model: 50 and 150. Lower MAE values indicate better model performance. 190
6.9 Mean Absolute Error (MAE) distributions for different preprocessing configurations predicting Count of deciduous trees. Each configuration varies in mask threshold (M), Z-score filtering (Z), and aggressive filtering (A). Results are shown for two values of the number of estimators in the model: 50 and 150. Lower MAE values indicate better model performance. 190
6.10 Mean Absolute Error (MAE) distributions for different preprocessing configurations predicting Count of coniferous trees. Each configuration varies in mask threshold (M), Z-score filtering (Z), and aggressive filtering (A). Results are shown for two values of the number of estimators in the model: 50 and 150. Lower MAE values indicate better model performance. 191
6.11 Mean Absolute Error (MAE) distributions for different preprocessing configurations predicting Tree area coverage (%). Each configuration varies in mask threshold (M), Z-score filtering (Z), and aggressive filtering (A). Results are shown for two values of the number of estimators in the model: 50 and 150. Lower MAE values indicate better model performance. 191
6.12 Mean Absolute Error (MAE) distributions for different preprocessing configurations predicting Sum crown volume (m^3). Each configuration varies in mask threshold (M), Z-score filtering (Z), and aggressive filtering (A). Results are shown for two values of the number of estimators in the model: 50 and 150. Lower MAE values indicate better model performance. 192

6.13 Mean Absolute Error (MAE) distributions for different preprocessing configurations predicting mean crown base height (m). Each configuration varies in mask threshold (M), Z-score filtering (Z), and aggressive filtering (A). Results are shown for two values of the number of estimators in the model: 50 and 150. Lower MAE values indicate better model performance. 192
6.14 Mean Absolute Error (MAE) distributions for different preprocessing configurations predicting Mean tree height (m). Each configuration varies in mask threshold (M), Z-score filtering (Z), and aggressive filtering (A). Results are shown for two values of the number of estimators in the model: 50 and 150. Lower MAE values indicate better model performance. 193
6.15 RF regression results of a spectrally, polarimetrically, and temporally fused Sentinel-1 & -2 (© ESA, 2021) dataset in the Steigerwald study site (AOI 1), displaying predicted values using $K_{0,*}$ against actual reference values for each target variable, based on the best overall RF configuration (general-purpose setup, see Section 6.1.3). Point density is shown as a logarithmic count; the red dashed line represents perfect agreement. (a) sum crown area of deciduous trees [m^2], (b) sum crown area of coniferous trees [m^2], (c) count deciduous trees [amount], (d) count coniferous trees [amount], (e) tree area coverage [%], (f) sum crown volume [m^3], (g) mean tree height [m], and (h) mean crown base height [m]. 194
6.16 RF regression results of a spectrally, polarimetrically, and temporally fused Sentinel-1 & -2 (©ESA, 2021) dataset in the Steigerwald study site (AOI 1), displaying the predicted values using $K_{0,*}$ against the actual values (i.e., reference data in Table 1.2) for each present target variable including the point density as logarithmic count and the perfect conditions (red dashed line); (a) sum crown area of deciduous trees [m^2], (b) sum crown area of coniferous trees [m^2], (c) count deciduous trees [amount], (d) count coniferous trees [amount], (e) tree area coverage [%], (f) sum crown volume [m^3], (g) mean tree height [m] and (h) mean crown base height [m]. 196

6.17 RF regression results of a spectrally, polarimetrically, and temporally fused Sentinel-1 & -2 (©ESA, 2021) dataset in the Bavarian Forest National Park study site (AOI 2), displaying the predicted values using $K_{0,*}$ against the actual values (i.e., reference data in Table 1.2) for each present target variable including the point density as logarithmic count and the perfect conditions (red dashed line); (a) sum crown area of deciduous trees [m^2], (b) sum crown area of coniferous trees [m^2], (c) sum crown area of dead trees [m^2], (d) count deciduous trees [amount], (e) count coniferous trees [amount], (f) count dead trees [amount], (g) tree area coverage [%], (h) Sum crown volume [m^3], (i) mean tree height [m] and (j) mean crown base height [m]. 198
6.18 RF regression results of a spectrally, polarimetrically, and temporally fused Sentinel-1 & -2 (©ESA, 2021) dataset in the Kranzberg Forest (AOI 3), displaying the predicted values using $K_{0,*}$ against the actual values (i.e., reference data in Table 1.2) for each present target variable including the point density as logarithmic count and the perfect conditions (red dashed line); (a) sum crown area of deciduous trees [m^2], (b) sum crown area of coniferous trees [m^2], (c) count deciduous trees [amount], (d) count coniferous trees [amount], (e) tree area coverage [%], (f) sum crown volume [m^3], (g) mean tree height [m] and (h) mean crown base height [m]. . . . 199
6.19 Fusion pipeline up to the combination of Sentinel-1 and Sentinel-2 Kennaugh representations using the Hypercomplex Bases (HCB) method. 211
6.20 Parallel feature extraction pipelines from Sentinel-2 MAJA data. The 10 m resolution Blue, Green, Red, and Near-Infrared (NIR) bands were used directly as raw input features, and alternatively transformed into Kennaugh-like elements capturing brightness and spectral structure. Both data streams are derived independently from Sentinel-2 acquisitions (2 July 2020 and 3 July 2021) over AOI 1 (Wald5Dplus project), supporting comparative modelling experiments. 217
6.21 Overview of evaluated scenarios based on ALOS-2 L-band and TSX X-band PolSAR data, including mono-frequency evaluations and dual-frequency fusion strategies. 238
6.22 Radar plot showing multi-criteria performance comparison of different Earth Observation modality-model strategies across four evaluation dimensions: In-Domain Accuracy, Cross-AOI Transferability, Computational Simplicity, and Interpretability. 253

7.1 Overview of the Helix-based modelling pipeline. A baseline model is trained on fused EO predictors and static Wald5Dplus labels to generate initial predictions and residuals. These residuals are reintroduced as uncertainty-aware features, while the original labels are enriched via multi-scale contextual averaging (Spatio-contextual Situation). The resulting context-enriched feature and target stacks are used to train a second-stage model, which outputs refined forest structure predictions. 261
7.2 Hexbin density plots comparing predicted versus true values for the variable *Sum crown area of deciduous trees (m^2)* under the baseline model (left) and the Helix+ model (right). Each subplot visualizes the prediction distribution using log-scaled point density. The Helix model achieves lower error across all reported metrics (MAE, RMSE, MAD, STD), with visibly tighter clustering along the 1:1 diagonal, indicating improved predictive accuracy and reduced residual variance. 264
7.3 Hexbin density plots comparing predicted versus true values for the variable *Sum crown area of coinferous trees (m^2)* under the baseline model (left) and the Helix+ model (right). Each subplot visualizes the prediction distribution using log-scaled point density. The Helix model achieves lower error across all reported metrics (MAE, RMSE, MAD, STD), with visibly tighter clustering along the 1:1 diagonal, indicating improved predictive accuracy and reduced residual variance. 265
7.4 Hexbin density plots comparing predicted versus true values for the variable *Sum crown area of dead trees (m^2)* under the baseline model (left) and the Helix+ model (right). Each subplot visualizes the prediction distribution using log-scaled point density. The Helix model achieves lower error across all reported metrics (MAE, RMSE, MAD, STD), with visibly tighter clustering along the 1:1 diagonal, indicating improved predictive accuracy and reduced residual variance. 265
7.5 Hexbin density plots comparing predicted versus true values for the variable *Count of deciduous trees* under the baseline model (left) and the Helix+ model (right). Each subplot visualizes the prediction distribution using log-scaled point density. The Helix model achieves lower error across all reported metrics (MAE, RMSE, MAD, STD), with visibly tighter clustering along the 1:1 diagonal, indicating improved predictive accuracy and reduced residual variance. 266

7.6 Hexbin density plots comparing predicted versus true values for the variable *Count of coniferous trees* under the baseline model (left) and the Helix+ model (right). Each subplot visualizes the prediction distribution using log-scaled point density. The Helix model achieves lower error across all reported metrics (MAE, RMSE, MAD, STD), with visibly tighter clustering along the 1:1 diagonal, indicating improved predictive accuracy and reduced residual variance. 266
7.7 Hexbin density plots comparing predicted versus true values for the variable *Count of dead trees* under the baseline model (left) and the Helix+ model (right). Each subplot visualizes the prediction distribution using log-scaled point density. The Helix model achieves lower error across all reported metrics (MAE, RMSE, MAD, STD), with visibly tighter clustering along the 1:1 diagonal, indicating improved predictive accuracy and reduced residual variance. 267
7.8 Hexbin density plots comparing predicted versus true values for the variable *Tree area coverage (%)* under the baseline model (left) and the Helix+ model (right). Each subplot visualizes the prediction distribution using log-scaled point density. The Helix model achieves lower error across all reported metrics (MAE, RMSE, MAD, STD), with visibly tighter clustering along the 1:1 diagonal, indicating improved predictive accuracy and reduced residual variance. 267
7.9 Hexbin density plots comparing predicted versus true values for the variable *Sum crown volume (m^3)* under the baseline model (left) and the Helix+ model (right). Each subplot visualizes the prediction distribution using log-scaled point density. The Helix model achieves lower error across all reported metrics (MAE, RMSE, MAD, STD), with visibly tighter clustering along the 1:1 diagonal, indicating improved predictive accuracy and reduced residual variance. 268
7.10 Hexbin density plots comparing predicted versus true values for the variable *Mean tree height (m)* under the baseline model (left) and the Helix+ model (right). Each subplot visualizes the prediction distribution using log-scaled point density. The Helix model achieves lower error across all reported metrics (MAE, RMSE, MAD, STD), with visibly tighter clustering along the 1:1 diagonal, indicating improved predictive accuracy and reduced residual variance. 268

7.11 Hexbin density plots comparing predicted versus true values for the variable *Mean crown base height (m)* under the baseline model (left) and the Helix+ model (right). Each subplot visualizes the prediction distribution using log-scaled point density. The Helix model achieves lower error across all reported metrics (MAE, RMSE, MAD, STD), with visibly tighter clustering along the 1:1 diagonal, indicating improved predictive accuracy and reduced residual variance. 269
7.12 Top 20 most influential features across the first 10 Helix+ target regressors. 270
7.13 **Helix spatial descriptors for 2020 outbreak structure at kernel scale** $s = 1$. The left panel shows the Helix `mean` band (`helix_mean_y2020_s1`), which encodes localized outbreak density, values range from 0 (no outbreak presence) to 1 (full neighbourhood affected). Red polygons denote bark beetle outbreak annotations for 2020. The right panel shows the Helix `variance` band (`helix_var_y2020_s1`), which captures spatial heterogeneity within each kernel neighbourhood. Variance values range from 0 (homogeneous areas, either all outbreak or all unaffected) to 0.25 (maximum spatial fragmentation, typically at outbreak boundaries). Mid-range values ($\sim$0.125) suggest partial infestation or edge zones. 281
7.14 ROC curves for synthetic classification task across spatial kernels ($s1$, $s2$, $s3$). Each subplot compares input configurations (EO only, EO+mean, EO+var, EO+mean+var). 291
7.15 Aggregate ROC curves for predicting high outbreak density. Models include EO combined with raw label, and EO with Helix mean at increasing kernel scales. 292
7.16 Grouped bar plots of mean absolute error (MAE) for all predicted Helix bands, organized by year. Within each year, bands are sorted by MAE from low to high. Blue tones represent mean bands; red tones represent variance bands. 295
7.17 Visual comparison of 2021 Helix outbreak intensity (mean descriptor, spatial kernel $s = 2$). **Left:** Ground-truth Helix mean band derived from labelled outbreak polygons. **Right:** Predicted Helix mean band obtained via ensemble forecasting model using only pre-2021 data. Both maps are normalized to the range [0, 1], where 0 denotes low or no outbreak intensity and 1 indicates maximal inferred damage. 300

7.18 Sentinel-1 SLC IW data over Axel Heiberg Island. **Left:** Single-date SAR image showing Kennaugh element K_0 derived from a Sentinel-1 SLC IW scene acquired on 2021-06-21. **Right:** Temporally fused dataset, generated from multiple Sentinel-1 SLC IW acquisitions within the defined seasonality periods, showing K_0. 312
7.19 Label enrichment via HELIX. **Left:** Static glacier facies classification from TSX for a single date (2021-06-21), representing categorical zone labels: Dry Snow, Percolation, Superimposed Ice, Ice-Free, and Wet Snow. **Right:** HELIX-based seasonal enrichment for the 2021 summer season, shown as continuous-valued regression targets capturing intra-seasonal glacier zone tendencies. 317
7.20 Seasonal label enrichment visualization. **Left:** RGB composite of HELIX-enriched seasonal facies for 2021, where Red = Spring, Green = Summer, Blue = Fall. **Right:** Alternate RGB encoding showing Red = Fall, Green = Winter, Blue = Annual Mean. The continuous colouring reveals spatial persistence and transition dynamics across glacier zones. 318
7.21 HELIX-enriched seasonal facies over the Devon Ice Cap. The four panels show: (1) a basemap view of the Devon Ice Cap from World Imagery; (2) HELIX-enriched seasonal glacier facies for 2017; (3) for 2018; and (4) for 2019. In panels 2–4, colour channels represent meteorological seasons: red = spring, green = summer, and blue = fall. 319
7.22 HELIX-enriched seasonal glacier facies on the Devon Ice Cap from 2020 to 2023. Each panel visualizes one year's seasonal pattern using RGB colour channels: red = spring, green = summer, and blue = fall. Panels from left to right correspond to the years 2020, 2021, 2022, and 2023. The imagery highlights inter-annual variability in glacier zone distribution. 319
7.23 HELIX-based supervised regression architecture for glacier facies prediction. Temporally fused Sentinel-1 inputs (Band 0) are combined with HELIX-enriched labels and optional historical label priors, feeding into two parallel modelling pathways with residual refinement. Models are evaluated across four generalization settings. 323
7.24 Relative importance of individual EO bands for glacier zone prediction using the full temporally-fused Sentinel-1 input stack. Band 0 (K0), representing total SAR backscatter intensity, demonstrated the highest predictive power across multiple model runs, justifying its selection as the sole input feature for subsequent modelling. 324

7.25 Seasonal class distributions in AOI 1, plotted on a logarithmic pixel-count scale. Reference and predicted distributions are shown for each season, excluding the annual mean. The model correctly reflects dominant seasonal trends, with slight overestimation of the Ice-Free zone in summer and consistent detection of Dry Snow and Superimposed Ice zones in transitional seasons. 329
7.26 Historical context-based seasonal prediction for spring. **Left:** HELIX-enriched reference map for the 2023 spring season, representing continuous glacier zone intensities across five facies (Dry Snow, Percolation, Superimposed Ice, Ice-Free, Wet Snow). **Right:** Model prediction for spring 2023 over AOI 1, generated using EO time-series data from 2023 for AOI 1, with a model trained solely on EO data from AOI 2 (2022) and the historical seasonal context vector ($\bar{L}_{\text{hist}}$), demonstrating a direct, unbiased spatio-temporal transfer. Both panels share the same continuous 1–5 scale, allowing direct comparison of predicted and reference facies tendencies. 331
7.27 Historical context-based seasonal prediction for summer. **Left:** HELIX-enriched reference map for the 2023 summer season, representing continuous glacier zone intensities across five facies (Dry Snow, Percolation, Superimposed Ice, Ice-Free, Wet Snow). **Right:** Model prediction for summer 2023 over AOI 1, generated using EO time-series data from 2023 for AOI 1, with a model trained solely on EO data from AOI 2 (2022) and the historical seasonal context vector ($\bar{L}_{\text{hist}}$), demonstrating a direct, unbiased spatio-temporal transfer. Both panels share the same continuous 1–5 scale, allowing direct comparison of predicted and reference facies tendencies. . . 332
7.28 Historical context-based seasonal prediction for fall. **Left:** HELIX-enriched reference map for the 2023 fall season, representing continuous glacier zone intensities across five facies (Dry Snow, Percolation, Superimposed Ice, Ice-Free, Wet Snow). **Right:** Model prediction for fall 2023 over AOI 1, generated using EO time-series data from 2023 for AOI 1, with a model trained solely on EO data from AOI 2 (2022) and the historical seasonal context vector ($\bar{L}_{\text{hist}}$), demonstrating a direct, unbiased spatio-temporal transfer. Both panels share the same continuous 1–5 scale, allowing direct comparison of predicted and reference facies tendencies. 333

7.29 Historical context-based seasonal prediction for winter. **Left:** HELIX-enriched reference map for the 2023 winter season, representing continuous glacier zone intensities across five facies (Dry Snow, Percolation, Superimposed Ice, Ice-Free, Wet Snow). **Right:** Model prediction for winter 2023 over AOI 1, generated using EO time-series data from 2023 for AOI 1, with a model trained solely on EO data from AOI 2 (2022) and the historical seasonal context vector ($\bar{L}_{\text{hist}}$), demonstrating a direct, unbiased spatio-temporal transfer. Both panels share the same continuous 1–5 scale, allowing direct comparison of predicted and reference facies tendencies. 334
7.30 Historical context-based seasonal prediction for the average glaciological year. **Left:** HELIX-enriched reference map for the 2023 whole season, representing continuous glacier zone intensities across five facies (Dry Snow, Percolation, Superimposed Ice, Ice-Free, Wet Snow). **Right:** Model prediction for the whole glaciological year 2023 over AOI 1, generated using EO time-series data from 2023 for AOI 1, with a model trained solely on EO data from AOI 2 (2022) and the historical seasonal context vector ($\bar{L}_{\text{hist}}$), demonstrating a direct, unbiased spatio-temporal transfer. Both panels share the same continuous 1–5 scale, allowing direct comparison of predicted and reference facies tendencies. 335
7.31 Reference and predicted zonation classes along the *White Glacier* transect (14.8 km in length, 200 m in width), for all seasons. 336
7.32 Reference and predicted zonation classes along the *Thompson Glacier* transect (38.0 km in length, 200 m in width), for all seasons. 336
7.33 Reference and predicted zonation classes along the *Airdrop Glacier* transect (44.0 km in length, 200 m in width), for all seasons. 336
7.34 Comparison of mean absolute error (MAE) for regression models predicting summer glacier facies using either the HELIX-enriched seasonal summer label or individual raw label scenes as targets. The HELIX-enriched summer label yields the lowest MAE overall, while individual labels display greater variation depending on acquisition date and scene conditions. 340

7.35 Comparison of model prediction and original TSX-based glacier zonations for the 2023 summer season. **Top left:** HELIX-enriched seasonal mean glacier zones for summer 2023 (continuous target). **Remaining panels:** Individual TerraSAR-X derived glacier zone classifications used as input for the seasonal enrichment, spanning from June to early September 2023. Each classification shows a snapshot of dynamic glacier facies (Dry Snow, Percolation, Superimposed Ice, Ice-Free, Wet Snow) across Axel Heiberg Island. The sequence illustrates the variability and temporal compression characteristic of the melt season, which contributes to reduced variance in seasonal label distributions. The enriched target (top left) reflects the averaged signal of these temporally noisy observations, which the model successfully learns to predict with sub-class precision. 345
7.36 Confusion Matrix – Spring. Reference (y-axis) vs. predicted (x-axis) classes after rounding regression outputs. 349
7.37 Confusion Matrix – Summer. Reflects high overlap between melt-prone zones. 350
7.38 Confusion Matrix – Fall. Shows transition behaviour and class overlap. . . . 351
7.39 Confusion Matrix – Winter. Indicates stronger class separation and higher prediction confidence. 352
7.40 Confusion Matrix – Year Mean. Aggregated across all seasons to reflect overall class-wise prediction behaviour. 353
7.41 Visualization over Axel Heiberg Island. **Left:** World Imagery [100] basemap showing the geographic extent of the study area. **Right:** Spectrally-polarimetrically fused Sentinel-1 SLC IW dataset from 2021-06-13, displayed in RGB, where Red represents $K_{\text{fused},0}$, Green represents $K_{\text{fused},4}$, and Blue represents $K_{\text{fused},2}$. 367
7.42 End-to-end schematic of the HELIX-inspired glacier zonation modelling pipeline. Labels are temporally enriched using a structured kernel (past and future class history), from which a delta target $\Delta = \mu_{\text{future}} - \text{Class}(t)$ is derived. During training, EO features and zonation history are used to predict this delta via a two-stage regressor. The final predicted class $\hat{Y}_{t+n}$ is reconstructed by adding the predicted delta to the current class. All future information is used only for supervision and excluded at inference. 369
7.43 HELIX-based enrichment of glacier zones over Axel Heiberg Island. Left: RGB encoding of glacier classes at t (R), $t-5$w (G), and $t+5$w (B), highlighting localised fluctuations. Right: RGB encoding of class means over t (R), $t-1$w to $t-5$w (G), and $t+1$w to $t+5$w (B), showing short-term zonal trends. . 374

7.44 Comparison of short-term future glacier zones. Left: Reference based on TSX-classified zones at time t (2021-06-13), enriched via HELIX with labels from $t+1$ to $t+5$ weeks. Right: Prediction based on fused Sentinel-1/-2 features and prior glacier classes ($t-5$w to t), without using future label input. 378
7.45 Confusion matrix of predicted vs. reference glacier zone classes. Values are log-scaled counts. 381

A.1 Residual distributions for predicted Helix descriptors. 537
A.2 Residual distributions for predicted Helix descriptors. 538
A.3 Residual distributions for predicted Helix descriptors. 539
A.4 Residual distributions for predicted Helix descriptors. 540
A.5 Residual distributions for predicted Helix descriptors. 541
A.6 Residual distributions for predicted Helix descriptors. 542
A.7 Residual distributions for predicted Helix descriptors. 543
A.8 Residual distributions for predicted Helix descriptors. 544
A.9 Residual distributions for predicted Helix descriptors. 545

List of Tables

1.1 Overview of AOIs, Subdivisions, and Associated EO Data Years 39
1.2 Characteristics of the examined Areas of Interest 41
1.3 Rasterized single-tree polygon bands capturing key forest attributes in a 10 m grid format. 43
1.4 Thresholds for Glacier Zones Derived by Histogram Peaks Method [311] . . 57

2.1 Classification of image data fusion approaches according to various publications. The most common and consistently described classification in the literature is by pixel, feature, and decision level.This table is adapted from [360]. 67

3.1 ML methods to reduce dependency from exhaustive labelled datasets. 92
3.2 Example of structured reference labels for tree polygons in an exemplary attribute table with the respective scale level assigned in the last row. 94
3.3 Challenges in reference data collection and their implications for ML models. 98
3.4 Label engineering techniques and their effects in EO workflows. 101
3.5 EO data dimensions that serve as a source of both input features and reference labels. 104
3.6 Temporal dynamic labelling methods in EO. 116
3.7 spatio-temporal dynamic labelling methods in EO. 118

5.1 Overview of the six temporal fusion configurations tested, including acquisition dates with respective seasons and cloud cover for Sentinel-2. 150
5.2 Roles of individual components in the CDVI formula. 159
5.3 Total Proba-V [51] vegetation cover fraction (Grass and Shrubs) and corresponding CDVI vegetation statistics across quantiles. 163

6.1 Summary of RF model performance for all forest variables. Best intra-AOI configuration and corresponding cross-AOI performance shifts (ΔMAE, ΔRMSE) are reported. 187

6.2 Summary of best-performing model configurations across all forest structural attributes using Sentinel-1 Kennaugh features. For each variable, the model yielding the lowest validation error (based on MAE) is reported, along with its specific preprocessing parameters . . . 231
6.3 Summary of best-performing model configurations across all forest structural attributes using Sentinel-1 Kennaugh features in the cross-validation scenario (Delta Metrics). Reported error values reflect improvements relative to the respective baseline setup for that attribute. . . . 233
6.4 Evaluated Scenarios for Forest Parameter Estimation . . . 238

7.1 Performance comparison between the baseline RF model and the context-enriched Helix+ model on all 10 forest structure variables. Each entry reports MAE, RMSE, MAD, and STD for both models, with the improvement (Δ) shown beneath each metric. . . . 263
7.2 Reduction in MAE for each forest structure variable between the baseline model and the Helix model. Absolute (Δ) and relative (%) improvements are reported. . . . 264
7.3 Top 10 most influential features across the first 10 Helix+ target regressors. Mean and standard deviation of feature importances are computed across models. Residual-based features rank prominently, highlighting the relevance of feedback signals. . . . 270
7.4 ROC-AUC and F1 scores for each spatial kernel scale and feature configuration. 290
7.5 Top Helix Bands from 2020 (Current Year) Sorted by RMSE . . . 292
7.6 Top Helix Bands from 2019 (Past Year) Sorted by RMSE . . . 292
7.7 Top Helix Bands from 2022 (Future Year) Sorted by RMSE . . . 293
7.8 Summary of fuzzy evaluation metrics for selected Helix bands. . . . 295
7.9 Stage 1 reconstruction metrics for Helix descriptors derived from 2020 EO data. . . . 297
7.10 Stage 2 ensemble regression metrics for predicting 2020 Helix mean and variance bands. . . . 298
7.11 Stage 3 Forecast Accuracy for 2021 Mean Helix Bands . . . 299
7.12 Mean Absolute Error (MAE) per season across all models and ensembles. Intra-AOI and cross-AOI scenarios are shown separately, with Year Mean summarizing average performance. . . . 326
7.13 Model Performance with Historical Label Vector ($\bar{L}_{\text{hist}}$) on AOI 2 . . . 327

7.14 Performance of Historical Label Prior Model (Trained on AOI 2, Evaluated on AOI 1, 2023) . . . 328
7.15 Summary statistics of class change (Δ Class) at future time steps, relative to the reference week (t). These distributions motivate the separate evaluation of each horizon. . . . 375
7.16 Regression performance summary across future time horizons. . . . 375
7.17 Directional transition accuracy (Down / Stable / Up) for each delta horizon. 376
7.18 Classification performance of predicted vs. reference future glacier zones. . 377

8.1 Interpretative summary of fused Kennaugh elements $K_{\text{fused},0}$ to $K_{\text{fused},7}$, derived from spectral–polarimetric fusion. Interpretations integrate symbolic formulations, NDVI/NDWI correlations, texture metrics (GLCM), and observed habitat contrast patterns. Note: high average intensity does not imply class separability; interpretations emphasize discriminative contrast and statistical evidence. . . . 398

A.1 Accuracy assessment of the RF regression for AOI 1 (Steigerwald). . . . 487
A.2 Accuracy assessment of the RF regression for AOI 2 (Bavarian Forest National Park). . . . 488
A.3 Accuracy assessment of the RF regression for AOI 3 (Kranzberg Forest). . . . 488
A.4 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Sum crown area of deciduous trees (m^2) . . . 489
A.5 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Sum crown area of coniferous trees (m^2) . . . 489
A.6 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Count of deciduous trees . . . 490
A.7 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Count of coniferous trees . . . 490
A.8 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Tree area coverage (%) . . . 491
A.9 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Sum crown volume (m^3) . . . 491
A.10 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Mean tree height (m) . . . 492
A.11 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Mean crown base height (m) . . . 492

A.12 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Sum crown area of deciduous trees (m^2) . . . 493
A.13 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Sum crown area of coniferous trees (m^2) . . . 493
A.14 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Count of deciduous trees 494
A.15 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Count of coniferous trees 494
A.16 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Tree area coverage (%) 495
A.17 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Sum crown volume (m^3) 495
A.18 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Mean tree height (m) 496
A.19 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Mean crown base height (m) 496
A.20 Transfer performance of a RF model trained on SW_1_2020, evaluated on SW_2_2020 and the Bavarian Forest NP (2021) across all 8-band variables. 497
A.21 Transfer performance of a RF model trained on NP_T10_2020, evaluated on Bavarian Forest NP (2021) and SW_2_2020 across all 8-band variables. . . . 498
A.22 Performance of the RF-stacked ensemble model trained on Steigerwald sub-AOIs SW_1, SW_2, SW_4, SW_5, and SW_6 (8-band input) in intra-AOI (SW_2) and transfer-AOI (SW_3) domains. Each variable shows MAE, RMSE, MAD, and STD for both intra and transfer settings, with corresponding delta values (Δ) below each metric. 499
A.23 Performance of the RF-stacked ensemble model trained on D03, D04, SW_1, SW_2 (Steigerwald) (2020 and 2021) and T10 (NP 2020) with the full 10-band input in the complete NP in 2020 and a temporal transfer setting in 2021. Each variable shows MAE, RMSE, MAD, and STD for both intra and transfer scenarios, with corresponding delta values (Δ) below each metric. . 500
A.24 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Sum crown area of deciduous trees (m^2) 501
A.25 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Sum crown area of coniferous trees (m^2) 502
A.26 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Count of deciduous trees 502

A.27 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Count of coniferous trees 503
A.28 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Tree area coverage (%) 503
A.29 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Sum crown volume (m^3) 504
A.30 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Mean tree height (m) 504
A.31 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Mean crown base height (m) 505
A.32 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Sum crown area of deciduous trees (m^2) . . . 506
A.33 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Sum crown area of coniferous trees (m^2) . . . 506
A.34 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Count of deciduous trees 507
A.35 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Count of coniferous trees 507
A.36 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Tree area coverage (%) 508
A.37 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Sum crown volume (m^3) 508
A.38 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Mean tree height (m) 509
A.39 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Mean crown base height (m) 509
A.40 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Sum crown area of deciduous trees (m^2) 510
A.41 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Sum crown area of coniferous trees (m^2) 511
A.42 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Count of deciduous trees 511
A.43 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Count of coniferous trees 512
A.44 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Tree area coverage (%) 512

A.45 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Sum crown volume (m^3) . . . 513
A.46 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Mean tree height (m) . . . 513
A.47 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Mean crown base height (m) . . . 514
A.48 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Sum crown area of deciduous trees (m^2) . . . 515
A.49 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Sum crown area of coniferous trees (m^2) . . . 515
A.50 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Count of deciduous trees . . . 516
A.51 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Count of coniferous trees . . . 516
A.52 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Tree area coverage (%) . . . 517
A.53 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Sum crown volume (m^3) . . . 517
A.54 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Mean tree height (m) . . . 518
A.55 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Mean crown base height (m) . . . 518
A.56 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Sum crown area of deciduous trees (m^2) . . . 519
A.57 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Sum crown area of coniferous trees (m^2) . . . 520
A.58 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Count of deciduous trees . . . 520
A.59 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Count of coniferous trees . . . 521
A.60 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Tree area coverage (%) . . . 521
A.61 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Sum crown volume (m^3) . . . 522
A.62 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Mean tree height (m) . . . 522

A.63 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Mean crown base height (m) . . . 523
A.64 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Sum crown area of deciduous trees (m^2) . . . 524
A.65 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Sum crown area of coniferous trees (m^2) . . . 524
A.66 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Count of deciduous trees . . . 525
A.67 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Count of coniferous trees . . . 525
A.68 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Tree area coverage (%) . . . 526
A.69 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Sum crown volume (m^3) . . . 526
A.70 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Mean tree height (m) . . . 527
A.71 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Mean crown base height (m) . . . 527
A.72 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Sum crown area of deciduous trees (m^2) . . . 528
A.73 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Sum crown area of coniferous trees (m^2) . . . 529
A.74 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Count of deciduous trees . . . 529
A.75 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Count of coniferous trees . . . 530
A.76 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Tree area coverage (%) . . . 530
A.77 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Sum crown volume (m^3) . . . 531
A.78 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Mean tree height (m) . . . 531
A.79 Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Mean crown base height (m) . . . 532
A.80 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Sum crown area of deciduous trees (m^2) . . . 533

A.81 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Sum crown area of coniferous trees (m^2) . . . 533
A.82 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Count of deciduous trees 534
A.83 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Count of coniferous trees 534
A.84 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Tree area coverage (%) 535
A.85 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Sum crown volume (m^3) 535
A.86 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Mean tree height (m) 536
A.87 Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Mean crown base height (m) 536
A.88 Per-band classification metrics using EO + individual helix features 546

List of Abbreviations

1D-CNN One-Dimensional Convolutional Neural Network

2D-CNN Two-Dimensional Convolutional Neural Network

3D-CNN Three-Dimensional Convolutional Neural Network

AI Artificial Intelligence

ANN Artificial Neural Network

AOI Area of Interest

AUC Area Under the Curve

CART Classification and Regression Trees

CDVI Combined Doline Vegetation Index

CNN Convolutional Neural Network

CRS Coordinate Reference System

DEM Digital Elevation Model

DETR Detection Transformer (object detection model)

DL Deep Learning

DLR German Aerospace Center (Deutsches Zentrum für Luft- und Raumfahrt)

DTM Digital Terrain Model

ELA Equilibrium Line Altitude

EPSG European Petroleum Survey Group (coordinate system codes)

ERT Extreme Random Tree

EVI Enhanced Vegetation Index

FM Foundation Model

F1 F1 Score (harmonic mean of precision and recall)

GLCM Gray Level Co-occurrence Matrix

GPR Ground-Penetrating Radar

HCB Hypercomplex Bases

HMM Hidden Markov Models

HSV Hue Saturation Value (color space)

IoU Intersection over Union

IW Interferometric Wide

LSTM Long Short-Term Memory

LiDAR Light Detection and Ranging

MAE Mean Absolute Error

MAJA MACCS-ATCOR Joint Algorithm (Atmospheric correction processor)

MAD Mean Absolute Deviation

ML Machine Learning

MSI Multispectral Instrument

NDVI Normalized Difference Vegetation Index

NDBI Normalized Difference Built-up Index

NDSI Normalized Difference Snow Index

NDWI Normalized Difference Water Index

RF Random Forest

RMSE Root Mean Square Error

RNN Recurrent Neural Network

ROC Receiver Operating Characteristic

RS Remote Sensing

SAR Synthetic Aperture Radar

SLC Single Look Complex

SDG Sustainable Development Goals

STD Standard Deviation

SVM Support Vector Machine

SVR Support Vector Regression

TempCNN Temporal Convolutional Neural Network

UTM Universal Transverse Mercator (coordinate system)

XAI Explainable Artificial Intelligence

XGBoost Extreme Gradient Boosting

YOLO You Only Look Once (object detection algorithm)

Appendix

A

A.1 Additional Data and Tables

A.1.1 Foundational Analysis of EO Modality–Model Interactions

Polarimetrically, Spectrally and Temporally Fused Sentinel-1 and Sentinel-2 Data

Wald5Dplus

Table A.1.: Accuracy assessment of the RF regression for AOI 1 (Steigerwald).

Variable	MAD	MAE	STD	Unit
Sum crown area of deciduous trees	4.120	5.195	6.768	m^2
Sum crown area of coniferous trees	3.510	4.326	5.433	m^2
Count deciduous trees	0.200	0.248	0.317	amount
Count coniferous trees	0.090	0.133	0.193	amount
Tree area coverage	0.710	1.075	1.602	%
Sum crown volume	24.330	31.140	40.601	m^3
Mean tree height	0.709	0.962	1.220	m
Mean crown base height	0.430	0.530	0.692	m

Table A.2.: Accuracy assessment of the RF regression for AOI 2 (Bavarian Forest National Park).

Variable	MAD	MAE	STD	Unit
Sum crown area of deciduous trees	5.765	6.249	7.564	m^2
Sum crown area of coniferous trees	4.630	5.238	6.525	m^2
Sum crown area of dead trees	3.115	3.811	4.745	m^2
Count deciduous trees	0.240	0.281	0.344	amount
Count coniferous trees	0.190	0.221	0.262	amount
Count dead trees	0.090	0.125	0.161	amount
Tree area coverage	1.345	2.133	3.202	%
Sum crown volume	52.130	64.858	83.534	m^3
Mean tree height	0.910	1.246	1.608	m
Mean crown base height	1.045	1.249	1.488	m

Table A.3.: Accuracy assessment of the RF regression for AOI 3 (Kranzberg Forest).

Variable	MAD	MAE	STD	Unit
Sum crown area of deciduous trees	7.800	6.917	7.513	m^2
Sum crown area of coniferous trees	4.700	6.759	8.663	m^2
Count deciduous trees	0.170	0.156	0.186	amount
Count coniferous trees	0.040	0.084	0.115	amount
Tree area coverage	1.320	2.955	5.423	%
Sum crown volume	117.410	140.235	134.399	m^3
Mean tree height	0.540	0.976	1.417	m
Mean crown base height	1.250	1.332	1.552	m

Model Setup Comparison per Variable

Table A.4.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Sum crown area of deciduous trees (m^2)

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 1, Z=1, Aggressive=False	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.044	5.636	3.489	5.635
Mask> 1, Z=3, Aggressive=False	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.433	5.622	34.006	0.562
Mask> 1, Z=None, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.435	5.637	34.427	5.636
Mask> 1, Z=None, Aggressive=True	Linear Regression		7.711	9.704	6.475	0.970
Mask> 1, Z=1, Aggressive=False	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	69.974	8.618	6.018	8.618

Table A.5.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Sum crown area of coniferous trees (m^2)

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 0, Z=1, Aggressive=False	Linear Regression		0.805	10.828	5.956	10.828
Mask> 0, Z=3, Aggressive=True	SVR	`C=1` `kernel=rbf`	0.824	13.346	4.101	12.466
Mask> 0, Z=1, Aggressive=False	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	0.872	11.190	3.865	11.190
Mask> 0, Z=3, Aggressive=False	Linear Regression		8.060	10.833	5.952	10.833
Mask> 0, Z=3, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	74.938	9.580	34.153	9.580

Table A.6.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Count of deciduous trees

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 1, Z=None, Aggressive=False	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	0.315	39.002	2.727	39.002
Mask> 0.1, Z=1, Aggressive=False	Linear Regression		0.594	7.591	48.910	0.759
Mask> 0.1, Z=None, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	0.596	7.671	48.786	0.767
Mask> 0, Z=3, Aggressive=True	SVR	`C=1` `kernel=rbf`	5.820	7.574	46.347	7.565
Mask> 1, Z=1, Aggressive=False	SVR	`C=1` `kernel=rbf`	42.861	0.557	0.341	5.559

Table A.7.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Count of coniferous trees

Experiment	Model	Params	MAE	RMSE	MAD	STD
NaN=mean, ZeroRowFilter=True, Norm=MinMax_1 D	1D CNN	`epochs=20` `batch_size=32` `lr=0.001`	0.198	0.299	0.108	0.298
Mask> 0.1, Z=None, Aggressive=False	SVR	`C=1` `kernel=rbf`	0.268	38.911	17.632	3.830
Mask> 0.1, Z=None, Aggressive=True	SVR	`C=1` `kernel=rbf`	0.268	38.911	17.632	3.830
Mask> 0, Z=None, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	12.237	1.750	0.637	17.496
Mask> 0.1, Z=1, Aggressive=False	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	30.346	38.137	20.952	38.136

Table A.8.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Tree area coverage (%)

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 1, Z=3, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	0.973	14.071	4.964	14.070
Mask> 0.1, Z=None, Aggressive=False	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.197	29.198	4.182	0.292
Mask> 0.1, Z=None, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.197	29.198	4.182	0.292
Mask> 0, Z=None, Aggressive=False	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	11.975	3.052	3.910	3.051
Mask> 0.1, Z=1, Aggressive=False	Linear Regression		27.910	5.338	17.926	5.338

Table A.9.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Sum crown volume (m^3)

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 0.1, Z=None, Aggressive=False	Linear Regression		0.072	9.752	57.477	0.975
Mask> 0.1, Z=None, Aggressive=True	Linear Regression		0.072	9.752	57.477	0.975
Mask> 0.1, Z=3, Aggressive=False	Linear Regression		0.072	9.752	57.477	0.975
Mask> 0, Z=1, Aggressive=False	Linear Regression		7.538	10.281	0.591	10.281
NaN=mean, ZeroRowFilter=True, Norm=MinMax$\setminus_1 D$	1D CNN	`epochs=20` `batch_size=32` `lr=0.001`	80.726	110.411	62.434	108.666

Table A.10.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Mean tree height (m)

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 0.1, Z=1, Aggressive=False	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.088	1.196	0.634	11.963
Mask> 1, Z=1, Aggressive=False	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	0.107	1.334	9.007	13.338
Mask> 0, Z=3, Aggressive=True	SVR	`C=1` `kernel=rbf`	0.237	30.798	18.989	30.766
Mask> 0.1, Z=3, Aggressive=False	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	8.569	11.674	6.129	1.167
Mask> 0, Z=1, Aggressive=False	SVR	`C=1` `kernel=rbf`	23.158	30.491	17.883	30.448

Table A.11.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Mean crown base height (m)

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 0, Z=1, Aggressive=False	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.107	13.887	8.497	13.887
Mask> 0.1, Z=None, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	0.224	27.891	19.931	2.789
Mask> 0.1, Z=3, Aggressive=False	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	0.224	27.891	19.931	2.789
Mask> 0.1, Z=None, Aggressive=False	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	9.673	1.306	7.342	13.058
Mask> 0, Z=1, Aggressive=False	SVR	`C=1` `kernel=rbf`	27.655	35.703	2.230	3.554

Model Setup Comparison Cross-Validation per Variable

Table A.12.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Sum crown area of deciduous trees (m^2)

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 1, Z=1, Aggressive=False	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.419	−3.768	6.722	13.042
Mask> 1, Z=3, Aggressive=False	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	14.412	−3.719	−32.985	18.475
Mask> 1, Z=None, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	14.217	12.949	−34.326	−3.778
Mask> 1, Z=None, Aggressive=True	Linear Regression		−6.847	1.709	−0.151	0.044
Mask> 1, Z=1, Aggressive=False	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	−68.497	10.333	4.704	10.331

Table A.13.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Sum crown area of coniferous trees (m^2)

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 0, Z=1, Aggressive=False	Linear Regression		2.700	−6.228	19.343	−6.254
Mask> 0, Z=3, Aggressive=True	SVR	`C=1` `kernel=rbf`	27.749	17.723	3.900	−11.246
Mask> 0, Z=1, Aggressive=False	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	25.411	−8.306	2.909	0.746
Mask> 0, Z=3, Aggressive=False	Linear Regression		26.644	34.809	19.012	−6.294
Mask> 0, Z=3, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	−46.674	21.250	−27.123	2.736

Table A.14.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Count of deciduous trees

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 1, Z=None, Aggressive=False	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	6.225	−30.962	3.017	−31.032
Mask> 0.1, Z=1, Aggressive=False	Linear Regression		2.837	−3.510	−30.599	30.783
Mask> 0.1, Z=None, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	7.620	2.615	−48.267	−0.689
Mask> 0, Z=3, Aggressive=True	SVR	`C=1` `kernel=rbf`	−5.120	1.387	−40.728	0.520
Mask> 1, Z=1, Aggressive=False	SVR	`C=1` `kernel=rbf`	−36.395	7.401	5.202	2.076

Table A.15.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Count of coniferous trees

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
NaN=mean, ZeroRowFilter=True, Norm=MinMax_1 D	1D CNN	`epochs=20` `batch_size=32` `lr=0.001`	0.088	0.153	0.030	0.131
Mask> 0.1, Z=None, Aggressive=False	SVR	`C=1` `kernel=rbf`	5.950	−31.272	12.946	45.396
Mask> 0.1, Z=None, Aggressive=True	SVR	`C=1` `kernel=rbf`	5.950	−31.272	12.946	45.396
Mask> 0, Z=None, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	−2.983	8.497	22.697	26.805
Mask> 0.1, Z=1, Aggressive=False	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	−22.380	−28.978	3.827	−33.600

Table A.16.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Tree area coverage (%)

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 1, Z=3, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	23.711	−9.193	8.887	33.633
Mask> 0.1, Z=None, Aggressive=False	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	23.885	−28.750	10.500	44.383
Mask> 0.1, Z=None, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	23.885	−28.750	10.500	44.383
Mask> 0, Z=None, Aggressive=False	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	−9.559	1.274	10.711	1.267
Mask> 0.1, Z=1, Aggressive=False	Linear Regression		−27.748	15.993	−16.760	13.145

Table A.17.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Sum crown volume (m^3)

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 0.1, Z=None, Aggressive=False	Linear Regression		0.316	−9.260	−57.451	2.824
Mask> 0.1, Z=None, Aggressive=True	Linear Regression		0.316	−9.260	−57.451	2.824
Mask> 0.1, Z=3, Aggressive=False	Linear Regression		0.316	−9.260	−57.451	2.824
Mask> 0, Z=1, Aggressive=False	Linear Regression		10.529	−8.021	−0.466	−8.416
NaN=mean, ZeroRowFilter=True, Norm=MinMax_1D	1D CNN	`epochs=20` `batch_size=32` `lr=0.001`	37.953	62.446	17.377	63.833

Table A.18.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Mean tree height (m)

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 0.1, Z=1, Aggressive=False	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	18.367	1.164	15.436	11.408
Mask> 1, Z=1, Aggressive=False	RF Regressor	max_depth=10 max_features=log2 min_samples_leaf=2	19.022	23.266	6.513	−11.058
Mask> 0, Z=3, Aggressive=True	SVR	C=1 kernel=rbf	1.497	−28.460	−17.715	−28.435
Mask> 0.1, Z=3, Aggressive=False	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	9.966	11.548	9.470	1.127
Mask> 0, Z=1, Aggressive=False	SVR	C=1 kernel=rbf	−3.749	−30.246	−2.012	−28.004

Table A.19.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Mean crown base height (m)

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 0, Z=1, Aggressive=False	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	27.325	19.587	9.478	12.299
Mask> 0.1, Z=None, Aggressive=True	RF Regressor	max_depth=10 max_features=log2 min_samples_leaf=2	22.638	0.386	−17.901	−0.132
Mask> 0.1, Z=3, Aggressive=False	RF Regressor	max_depth=10 max_features=log2 min_samples_leaf=2	22.638	0.386	−17.901	−0.132
Mask> 0.1, Z=None, Aggressive=False	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	−7.309	27.638	12.137	−10.413
Mask> 0, Z=1, Aggressive=False	SVR	C=1 kernel=rbf	−3.107	−32.627	17.426	25.929

Ensemble

Table A.20.: Transfer performance of a RF model trained on SW_1_2020, evaluated on SW_2_2020 and the Bavarian Forest NP (2021) across all 8-band variables.

Variable	Experiment	MAE	RMSE	MAD	STD
Sum crown area of deciduous trees (m²)	SW_2_2020	16.92	19.79	8.71	15.84
Sum crown area of deciduous trees (m²)	Bavarian Forest NP 2021	41.9	46.4	19.7	41.1
Sum crown area of coniferous trees (m²)	SW_2_2020	12.52	14.29	4.77	1.17
Sum crown area of coniferous trees (m²)	Bavarian Forest NP 2021	27.03	34.54	24.0	31.3
Count of deciduous trees	SW_2_2020	0.87	1.10	0.64	0.97
Count of deciduous trees	Bavarian Forest NP 2021	1.68	1.90	0.64	1.00
Count of coniferous trees	SW_2_2020	0.42	0.48	0.16	0.38
Count of coniferous trees	Bavarian Forest NP 2021	0.98	1.26	0.86	1.14
Tree area coverage (%)	SW_2_2020	2.70	3.50	1.00	3.22
Tree area coverage (%)	Bavarian Forest NP 2021	23.10	27.70	18.60	25.71
Sum crown volume (m³)	SW_2_2020	86.42	108.63	71.99	108.54
Sum crown volume (m³)	Bavarian Forest NP 2021	262.26	315.49	251.41	299.49
Mean tree height (m)	SW_2_2020	4.02	4.96	2.70	3.93
Mean tree height (m)	Bavarian Forest NP 2021	8.25	10.09	7.10	8.64
Mean crown base height (m)	SW_2_2020	4.41	5.79	3.11	4.66
Mean crown base height (m)	Bavarian Forest NP 2021	3.42	5.05	1.72	4.80

Table A.21.: Transfer performance of a RF model trained on NP_T10_2020, evaluated on Bavarian Forest NP (2021) and SW_2_2020 across all 8-band variables.

Variable	Experiment	MAE	RMSE	MAD	STD
Sum crown area of deciduous trees (m^2)	Bavarian Forest NP 2021	32.01	36.15	20.10	32.90
Sum crown area of deciduous trees (m^2)	SW_2_2020	68.83	71.19	10.11	18.17
Sum crown area of coniferous trees (m^2)	Bavarian Forest NP 2021	23.75	29.03	20.72	28.73
Sum crown area of coniferous trees (m^2)	SW_2_2020	34.80	36.63	5.57	12.78
Count of deciduous trees	Bavarian Forest NP 2021	0.73	0.84	0.44	0.70
Count of deciduous trees	SW_2_2020	3.21	3.37	0.66	1.01
Count of coniferous trees	Bavarian Forest NP 2021	0.89	1.09	0.78	1.08
Count of coniferous trees	SW_2_2020	1.18	1.24	0.19	0.41
Tree area coverage (%)	Bavarian Forest NP 2021	16.30	20.10	10.95	20.09
Tree area coverage (%)	SW_2_2020	13.20	14.02	2.72	5.00
Sum crown volume (m^3)	Bavarian Forest NP 2021	225.96	268.02	216.62	267.71
Sum crown volume (m^3)	SW_2_2020	226.67	23.62	77.95	118.70
Mean tree height (m)	Bavarian Forest NP 2021	6.94	8.24	6.33	8.00
Mean tree height (m)	SW_2_2020	6.56	7.63	3.18	4.43
Mean crown base height (m)	Bavarian Forest NP 2021	3.78	4.60	2.36	4.57
Mean crown base height (m)	SW_2_2020	4.41	5.86	3.14	4.90

Table A.22.: Performance of the RF-stacked ensemble model trained on Steigerwald sub-AOIs SW_1, SW_2, SW_4, SW_5, and SW_6 (8-band input) in intra-AOI (SW_2) and transfer-AOI (SW_3) domains. Each variable shows MAE, RMSE, MAD, and STD for both intra and transfer settings, with corresponding delta values (Δ) below each metric.

Variable	MAE	RMSE	MAD	STD
Sum crown area of deciduous trees (m^2) Δ: -7.162	17.014 / 9.852	19.775 / 13.530	8.492 / 6.464	15.800 / 11.924
Sum crown area of coniferous trees (m^2) Δ: -4.543	11.992 / 7.449	13.446 / 9.555	3.570 / 5.908	11.415 / 8.176
Count of deciduous trees Δ: -0.433	0.870 / 0.437	1.112 / 0.629	0.655 / 0.291	0.985 / 0.592
Count of coniferous trees Δ: -0.149	0.388 / 0.239	0.435 / 0.306	0.118 / 0.192	0.365 / 0.260
Sum crown volume (m^3) Δ: -31.406	85.238 / 53.832	106.854 / 76.627	70.873 / 35.999	106.536 / 76.626
Tree area coverage (%) Δ: -0.928	2.529 / 1.601	3.402 / 2.567	1.084 / 0.757	3.273 / 2.537
Mean tree height (m) Δ: -3.233	5.054 / 1.821	6.070 / 2.671	2.987 / 1.180	4.137 / 2.577
Mean crown base height (m) Δ: -3.553	5.041 / 1.487	6.563 / 2.396	3.193 / 0.823	4.865 / 2.334

Table A.23.: Performance of the RF-stacked ensemble model trained on D03, D04, SW_1, SW_2 (Steigerwald) (2020 and 2021) and T10 (NP 2020) with the full 10-band input in the complete NP in 2020 and a temporal transfer setting in 2021. Each variable shows MAE, RMSE, MAD, and STD for both intra and transfer scenarios, with corresponding delta values (Δ) below each metric.

Variable	MAE	RMSE	MAD	STD
Sum crown area of deciduous trees (m^2) Δ: 3.26	25.60 / 28.86	31.25 / 35.38	17.02 / 18.33	30.41 / 34.79
Sum crown area of coniferous trees (m^2) Δ: 4.38	20.65 / 25.03	24.94 / 31.14	19.61 / 22.95	24.29 / 29.20
Sum crown area of dead trees (m^2) Δ: 0.94	7.50 / 8.44	8.33 / 9.29	1.75 / 2.41	4.26 / 4.54
Count of deciduous trees Δ: 0.05	0.33 / 0.38	0.34 / 0.40	0.07 / 0.10	0.21 / 0.23
Count of coniferous trees Δ: 0.06	0.57 / 0.63	0.68 / 0.76	4.20 / 4.26	0.68 / 0.76
Count of dead trees Δ: 0.08	0.81 / 0.89	0.10 / 1.12	0.72 / 0.81	0.98 / 1.07
Tree area coverage (%) Δ: 3.73	13.24 / 16.97	16.54 / 20.08	7.50 / 10.37	16.54 / 19.88
Sum crown volume (m^3) Δ: 62.29	184.37 / 246.66	223.00 / 293.29	165.70 / 222.13	222.00 / 278.95
Mean tree height (m) Δ: 1.28	5.84 / 7.12	7.06 / 8.54	5.33 / 6.60	7.04 / 8.47
Mean crown base height (m) Δ: 0.10	3.30 / 3.40	4.67 / 4.95	2.20 / 2.02	4.38 / 4.60

Model Setup Comparison per Variable

Table A.24.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Sum crown area of deciduous trees (m^2)

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	3.285	4.065	2.672	4.064
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	3.320	4.121	3.013	4.120
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	3.392	4.245	2.761	4.244
Mask> 1, Z=1, Aggressive=False	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	8.892	11.096	7.578	11.096
Mask> 0.1, Z=None, Aggressive=True	Linear Regression		17.226	22.116	13.841	22.116

Table A.25.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Sum crown area of coniferous trees (m^2)

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	2.960	3.660	2.787	3.660
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	2.969	3.621	2.627	3.620
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	2.997	3.664	2.579	3.664
Mask> 1, Z=None, Aggressive=False	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	7.442	9.321	6.237	9.321
Mask> 0.1, Z=None, Aggressive=False	Linear Regression		13.414	17.097	9.796	17.097

Table A.26.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Count of deciduous trees

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	0.148	0.190	0.122	0.190
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.149	0.191	0.115	0.191
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.150	0.193	0.121	0.193
Mask> 0.1, Z=1, Aggressive=True	Linear Regression		0.433	0.521	0.394	0.521
Mask> 0, Z=None, Aggressive=True	Linear Regression		0.817	1.051	0.661	1.051

Table A.27.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Count of coniferous trees

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 0, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.090	0.118	0.056	0.118
Mask> 0, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.090	0.117	0.057	0.117
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.113	0.142	0.096	0.142
Mask> 1, Z=1, Aggressive=False	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	0.263	0.335	0.202	0.335
Mask> 0.1, Z=None, Aggressive=True	Linear Regression		0.423	0.559	0.280	0.559

Table A.28.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Tree area coverage (%)

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.374	0.522	0.274	0.522
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.375	0.528	0.247	0.528
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	0.392	0.540	0.258	0.540
Mask> 0, Z=3, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	1.629	2.497	0.704	2.497
Mask> 0.1, Z=None, Aggressive=False	Linear Regression		2.807	5.959	1.195	5.959

Table A.29.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Sum crown volume (m^3)

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	21.896	27.793	17.935	27.789
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	22.853	28.291	20.131	28.290
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=10 max_features=log2 min_samples_leaf=2	23.045	29.163	20.545	29.160
Mask> 1, Z=None, Aggressive=True	RF Regressor	max_depth=10 max_features=log2 min_samples_leaf=2	60.632	77.782	49.101	77.782
Mask> 0.1, Z=None, Aggressive=False	SVR	C=1 kernel=rbf	96.205	126.921	77.957	126.915

Table A.30.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Mean tree height (m)

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	0.506	0.656	0.381	0.655
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	0.507	0.670	0.393	0.669
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=10 max_features=log2 min_samples_leaf=2	0.530	0.681	0.418	0.681
Mask> 0.1, Z=1, Aggressive=True	Linear Regression		1.552	1.887	1.367	1.887
Mask> 0, Z=1, Aggressive=False	Linear Regression		3.088	3.896	2.570	3.896

Table A.31.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Mean crown base height (m)

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.640	0.786	0.501	0.786
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	0.646	0.798	0.533	0.798
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.648	0.789	0.542	0.789
Mask> 0.1, Z=1, Aggressive=True	Linear Regression		1.651	1.973	1.495	1.973
Mask> 0, Z=1, Aggressive=False	Linear Regression		3.450	4.341	2.794	4.341

Model Setup Comparison Cross-Validation per Variable

Table A.32.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Sum crown area of deciduous trees (m^2)

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	13.757	17.375	11.409	17.352
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	13.719	17.400	11.013	17.360
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	13.680	17.412	11.622	17.356
Mask> 1, Z=1, Aggressive=False	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	6.540	9.449	5.405	8.114
Mask> 0.1, Z=None, Aggressive=True	Linear Regression		16.834	16.770	−1.817	−3.300

Table A.33.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Sum crown area of coniferous trees (m^2)

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	8.880	10.620	6.622	10.209
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	8.695	10.487	6.531	10.119
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	8.692	10.494	6.619	10.141
Mask> 1, Z=None, Aggressive=False	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	3.460	3.944	1.862	3.760
Mask> 0.1, Z=None, Aggressive=False	Linear Regression		7.769	7.646	−1.091	−4.225

Table A.34.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Count of deciduous trees

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	0.606	0.722	0.514	0.707
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.592	0.707	0.519	0.695
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.595	0.705	0.509	0.684
Mask> 0.1, Z=1, Aggressive=True	Linear Regression		1.034	1.126	0.119	0.278
Mask> 0, Z=None, Aggressive=True	Linear Regression		0.738	0.700	−0.062	−0.201

Table A.35.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Count of coniferous trees

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 0, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.180	1.241	0.246	0.365
Mask> 0, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.190	1.251	0.231	0.365
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.260	0.359	0.178	0.347
Mask> 1, Z=1, Aggressive=False	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	0.117	0.162	0.087	0.154
Mask> 0.1, Z=None, Aggressive=True	Linear Regression		0.367	0.358	0.015	−0.091

Table A.36.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Tree area coverage (%)

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	2.189	4.678	1.273	4.411
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	2.194	4.668	1.312	4.412
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=10 max_features=log2 min_samples_leaf=2	2.168	4.633	1.330	4.383
Mask> 0, Z=3, Aggressive=True	RF Regressor	max_depth=10 max_features=log2 min_samples_leaf=2	0.776	2.516	0.667	2.172
Mask> 0.1, Z=None, Aggressive=False	Linear Regression		0.486	−0.473	0.216	−1.488

Table A.37.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Sum crown volume (m^3)

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	67.836	92.518	35.251	87.868
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	66.766	92.872	38.095	88.592
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=10 max_features=log2 min_samples_leaf=2	67.186	91.593	33.526	86.416
Mask> 1, Z=None, Aggressive=True	RF Regressor	max_depth=10 max_features=log2 min_samples_leaf=2	56.394	64.402	3.955	39.726
Mask> 0.1, Z=None, Aggressive=False	SVR	C=1 kernel=rbf	−12.576	−10.990	−27.594	−12.972

Table A.38.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Mean tree height (m)

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.439	1.841	1.291	1.691
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.486	1.875	1.365	1.686
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	1.414	1.842	1.282	1.677
Mask> 0.1, Z=1, Aggressive=True	Linear Regression		0.256	0.389	0.163	0.388
Mask> 0, Z=1, Aggressive=False	Linear Regression		−0.990	−1.154	−0.916	−1.424

Table A.39.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Mean crown base height (m)

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.340	1.687	1.153	1.687
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	1.322	1.642	1.066	1.635
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.303	1.645	1.046	1.646
Mask> 0.1, Z=1, Aggressive=True	Linear Regression		0.599	0.795	0.531	0.661
Mask> 0, Z=1, Aggressive=False	Linear Regression		−0.115	−0.304	−0.804	−1.522

Reflectance Bands and Spectral Kennaugh-like Elements from Sentinel-2 Data

Raw Sentinel-2 Data

Model Setup Comparison per Variable

Table A.40.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Sum crown area of deciduous trees (m^2)

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	3.861	4.834	3.271	4.834
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	3.894	4.893	3.306	4.893
Mask> 0, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	4.347	5.633	3.342	5.633
Mask> 1, Z=None, Aggressive=False	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	10.335	12.973	8.559	12.973
Mask> 0.1, Z=None, Aggressive=True	SVR	`C=1` `kernel=rbf`	18.315	23.815	14.515	23.619

Table A.41.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Sum crown area of coniferous trees (m^2)

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 0, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	3.230	4.236	1.939	4.235
Mask> 0, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	3.254	4.289	1.986	4.288
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	3.272	4.131	2.698	4.131
Mask> 1, Z=None, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	8.426	10.600	6.894	10.600
Mask> 0.1, Z=None, Aggressive=True	SVR	`C=1` `kernel=rbf`	13.837	18.712	10.301	17.973

Table A.42.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Count of deciduous trees

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.185	0.227	0.156	0.227
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.186	0.232	0.159	0.232
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	0.236	0.285	0.218	0.285
Mask> 1, Z=None, Aggressive=False	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	0.464	0.583	0.388	0.583
Mask> 0, Z=None, Aggressive=False	Linear Regression		0.851	1.087	0.695	1.087

Table A.43.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Count of coniferous trees

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 0, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.095	0.126	0.059	0.126
Mask> 0, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.096	0.127	0.061	0.127
Mask> 0.1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.127	0.160	0.098	0.160
Mask> 1, Z=1, Aggressive=False	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	0.298	0.378	0.217	0.378
Mask> 0.1, Z=None, Aggressive=False	Linear Regression		0.432	0.569	0.292	0.569

Table A.44.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Tree area coverage (%)

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.678	0.891	0.464	0.891
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.685	0.895	0.485	0.895
Mask> 0.1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.768	1.071	0.416	1.070
Mask> 0, Z=None, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.757	4.112	0.618	4.111
Mask> 0.1, Z=None, Aggressive=False	Linear Regression		2.889	6.009	1.262	6.009

Table A.45.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Sum crown volume (m^3)

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	27.308	33.203	24.128	33.203
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	27.437	33.267	24.479	33.266
Mask> 0.1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	29.299	35.951	25.019	35.951
Mask> 1, Z=None, Aggressive=False	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	69.452	87.838	58.485	87.838
Mask> 0.1, Z=None, Aggressive=True	SVR	`C=1` `kernel=rbf`	97.294	128.491	78.267	128.480

Table A.46.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Mean tree height (m)

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.649	0.808	0.548	0.808
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.653	0.813	0.511	0.813
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	0.796	0.975	0.678	0.975
Mask> 0, Z=None, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.729	2.272	1.334	2.272
Mask> 0, Z=None, Aggressive=True	SVR	`C=1` `kernel=rbf`	3.268	4.118	2.733	4.117

Table A.47.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Mean crown base height (m)

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.720	0.880	0.601	0.880
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.727	0.878	0.661	0.878
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	0.876	1.046	0.794	1.046
Mask> 0, Z=1, Aggressive=False	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.956	2.531	1.536	2.530
Mask> 0, Z=None, Aggressive=True	SVR	`C=1` `kernel=rbf`	3.570	4.580	2.928	4.543

Model Setup Comparison Cross-Validation per Variable

Table A.48.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Sum crown area of deciduous trees (m^2)

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	12.329	15.564	9.719	15.497
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	12.341	15.370	10.085	15.351
Mask> 0, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	40.080	42.961	9.418	14.056
Mask> 1, Z=None, Aggressive=False	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	3.452	4.411	2.706	4.348
Mask> 0.1, Z=None, Aggressive=True	SVR	`C=1` `kernel=rbf`	7.121	7.117	−3.934	−5.682

Table A.49.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Sum crown area of coniferous trees (m^2)

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 0, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	33.429	34.601	6.095	8.587
Mask> 0, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	33.593	34.732	6.218	8.554
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	7.383	9.022	5.479	8.886
Mask> 1, Z=None, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	2.176	2.086	0.930	1.489
Mask> 0.1, Z=None, Aggressive=True	SVR	`C=1` `kernel=rbf`	9.912	7.907	−3.143	−5.933

Table A.50.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Count of deciduous trees

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	0.556	0.669	0.476	0.651
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	0.543	0.652	0.443	0.638
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=10 max_features=log2 min_samples_leaf=2	0.512	0.618	0.417	0.601
Mask> 1, Z=None, Aggressive=False	RF Regressor	max_depth=10 max_features=log2 min_samples_leaf=2	0.147	0.170	0.122	0.156
Mask> 0, Z=None, Aggressive=False	Linear Regression		−0.152	−0.197	−0.192	−0.337

Table A.51.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Count of coniferous trees

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 0, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	1.206	1.254	0.231	0.333
Mask> 0, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	1.208	1.255	0.214	0.330
Mask> 0.1, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	0.919	0.985	0.176	0.305
Mask> 1, Z=1, Aggressive=False	RF Regressor	max_depth=10 max_features=log2 min_samples_leaf=2	0.057	0.074	0.059	0.072
Mask> 0.1, Z=None, Aggressive=False	Linear Regression		0.100	0.125	−0.016	−0.108

Table A.52.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Tree area coverage (%)

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.934	4.376	1.088	4.119
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.968	4.386	1.118	4.149
Mask> 0.1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.895	4.181	1.169	3.932
Mask> 0, Z=None, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.286	0.963	1.237	0.953
Mask> 0.1, Z=None, Aggressive=False	Linear Regression		−0.000	−1.200	0.030	−1.238

Table A.53.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Sum crown volume (m^3)

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	56.529	82.792	26.016	81.031
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	57.024	82.845	28.805	80.619
Mask> 0.1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	57.515	85.085	28.285	84.510
Mask> 1, Z=None, Aggressive=False	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	37.150	45.466	20.169	40.880
Mask> 0.1, Z=None, Aggressive=True	SVR	`C=1` `kernel=rbf`	−17.534	−14.348	−27.506	−14.553

Table A.54.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Mean tree height (m)

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	1.454	1.925	1.045	1.679
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	1.441	1.913	1.102	1.682
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=10 max_features=log2 min_samples_leaf=2	1.323	1.778	0.963	1.520
Mask> 0, Z=None, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	0.533	0.742	0.281	0.568
Mask> 0, Z=None, Aggressive=True	SVR	C=1 kernel=rbf	−1.051	−1.316	−0.995	−1.750

Table A.55.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Mean crown base height (m)

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	1.649	2.051	1.423	1.916
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	1.603	1.996	1.362	1.891
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=10 max_features=log2 min_samples_leaf=2	1.449	1.819	1.187	1.715
Mask> 0, Z=1, Aggressive=False	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	0.587	0.646	0.573	0.367
Mask> 0, Z=None, Aggressive=True	SVR	C=1 kernel=rbf	−1.143	−1.612	−0.952	−1.960

Model Setup Comparison per Variable

Table A.56.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Sum crown area of deciduous trees (m^2)

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	3.630	4.637	3.017	4.636
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	3.642	4.649	3.116	4.649
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	3.923	4.939	3.426	4.939
Mask> 1, Z=None, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	10.230	12.895	8.541	12.895
Mask> 0.1, Z=3, Aggressive=False	Linear Regression		17.644	22.533	14.239	22.533

Table A.57.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Sum crown area of coniferous trees (m^2)

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	3.176	4.027	2.482	4.027
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	3.193	3.981	2.562	3.981
Mask> 0, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	3.234	4.272	1.960	4.272
Mask> 1, Z=1, Aggressive=False	RF Regressor	max_depth=10 max_features=log2 min_samples_leaf=2	8.467	10.649	6.800	10.649
Mask> 0.1, Z=None, Aggressive=False	Linear Regression		13.647	17.351	10.156	17.351

Table A.58.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Count of deciduous trees

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	0.175	0.216	0.142	0.216
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	0.178	0.220	0.148	0.220
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=10 max_features=log2 min_samples_leaf=2	0.199	0.244	0.175	0.244
Mask> 1, Z=3, Aggressive=False	RF Regressor	max_depth=10 max_features=log2 min_samples_leaf=2	0.461	0.580	0.382	0.580
Mask> 0, Z=3, Aggressive=False	Linear Regression		0.831	1.068	0.678	1.068

Table A.59.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Count of coniferous trees

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 0, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.095	0.126	0.059	0.126
Mask> 0, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.096	0.128	0.060	0.128
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.124	0.154	0.102	0.154
Mask> 0, Z=1, Aggressive=False	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	0.292	0.410	0.138	0.410
Mask> 0.1, Z=None, Aggressive=True	Linear Regression		0.430	0.567	0.288	0.567

Table A.60.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Tree area coverage (%)

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.612	0.802	0.419	0.802
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.616	0.801	0.441	0.801
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	0.662	0.866	0.477	0.866
Mask> 0, Z=None, Aggressive=False	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.705	4.043	0.605	4.042
Mask> 0.1, Z=None, Aggressive=True	Linear Regression		2.834	5.974	1.258	5.974

Table A.61.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Sum crown volume (m^3)

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	27.255	33.533	23.763	33.532
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	27.318	33.789	24.845	33.788
Mask> 0, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	28.740	35.201	25.006	35.201
Mask> 1, Z=1, Aggressive=False	RF Regressor	max_depth=10 max_features=log2 min_samples_leaf=2	66.915	85.270	55.527	85.268
Mask> 0.1, Z=3, Aggressive=False	SVR	C=1 kernel=rbf	96.126	126.738	78.220	126.725

Table A.62.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Mean tree height (m)

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	0.630	0.793	0.529	0.793
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	0.639	0.805	0.514	0.805
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=10 max_features=log2 min_samples_leaf=2	0.706	0.878	0.580	0.878
Mask> 0, Z=1, Aggressive=False	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	1.707	2.227	1.333	2.227
Mask> 0, Z=1, Aggressive=False	SVR	C=1 kernel=rbf	3.199	4.020	2.688	4.020

Table A.63.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Mean crown base height (m)

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.736	0.889	0.682	0.889
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.739	0.896	0.642	0.896
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	0.812	0.972	0.740	0.972
Mask> 0, Z=None, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.919	2.487	1.493	2.487
Mask> 0, Z=1, Aggressive=False	Linear Regression		3.611	4.518	2.990	4.518

Model Setup Comparison Cross-Validation per Variable

Table A.64.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Sum crown area of deciduous trees (m²)

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	12.835	16.255	9.421	16.148
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	12.422	16.029	9.067	15.846
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=10 max_features=log2 min_samples_leaf=2	12.151	15.687	8.694	15.515
Mask> 1, Z=None, Aggressive=True	RF Regressor	max_depth=10 max_features=log2 min_samples_leaf=2	3.256	4.722	2.562	4.650
Mask> 0.1, Z=3, Aggressive=False	Linear Regression		1.309	2.177	−3.333	−4.350

Table A.65.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Sum crown area of coniferous trees (m²)

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	7.738	9.548	5.400	9.543
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	7.748	9.690	5.446	9.682
Mask> 0, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	30.407	31.630	5.506	8.267
Mask> 1, Z=1, Aggressive=False	RF Regressor	max_depth=10 max_features=log2 min_samples_leaf=2	2.434	2.486	0.376	1.936
Mask> 0.1, Z=None, Aggressive=False	Linear Regression		−0.136	0.239	−2.915	−4.920

Table A.66.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Count of deciduous trees

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.590	0.717	0.492	0.713
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.592	0.713	0.495	0.709
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	0.575	0.696	0.466	0.693
Mask> 1, Z=3, Aggressive=False	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	0.181	0.228	0.129	0.222
Mask> 0, Z=3, Aggressive=False	Linear Regression		0.010	−0.021	−0.154	−0.283

Table A.67.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Count of coniferous trees

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 0, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.145	1.200	0.223	0.345
Mask> 0, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.141	1.195	0.219	0.341
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.236	0.330	0.164	0.318
Mask> 0, Z=1, Aggressive=False	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	0.586	0.581	0.153	0.058
Mask> 0.1, Z=None, Aggressive=True	Linear Regression		0.108	0.128	−0.022	−0.113

Table A.68.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Tree area coverage (%)

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	1.945	4.258	1.073	4.076
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	1.970	4.276	1.131	4.104
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=10 max_features=log2 min_samples_leaf=2	1.874	4.176	0.978	3.982
Mask> 0, Z=None, Aggressive=False	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	0.904	0.310	0.947	0.310
Mask> 0.1, Z=None, Aggressive=True	Linear Regression		−0.414	−1.541	0.120	−1.556

Table A.69.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Sum crown volume (m^3)

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	56.440	79.351	33.343	79.051
Mask> 1, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	57.457	79.675	35.227	79.490
Mask> 0, Z=1, Aggressive=True	RF Regressor	max_depth=None max_features=log2 min_samples_leaf=2	56.253	85.741	29.225	83.826
Mask> 1, Z=1, Aggressive=False	RF Regressor	max_depth=10 max_features=log2 min_samples_leaf=2	34.401	45.025	17.533	41.774
Mask> 0.1, Z=3, Aggressive=False	SVR	C=1 kernel=rbf	−15.468	−13.449	−24.972	−13.847

Table A.70.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Mean tree height (m)

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.349	1.765	1.024	1.537
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.325	1.750	1.015	1.489
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	1.309	1.719	0.995	1.479
Mask> 0, Z=1, Aggressive=False	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.448	0.570	0.290	0.412
Mask> 0, Z=1, Aggressive=False	SVR	`C=1` `kernel=rbf`	−1.294	−1.602	−1.142	−1.756

Table A.71.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Mean crown base height (m)

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.390	1.772	1.144	1.707
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.395	1.796	1.143	1.678
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	1.313	1.689	1.079	1.624
Mask> 0, Z=None, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.581	0.689	0.472	0.426
Mask> 0, Z=1, Aggressive=False	Linear Regression		−1.377	−1.722	−1.321	−2.067

Model Setup Comparison per Variable

Table A.72.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Sum crown area of deciduous trees (m^2)

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	3.889	4.857	3.565	4.855
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	3.933	4.933	3.491	4.933
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	4.264	5.241	3.924	5.241
Mask> 1, Z=None, Aggressive=False	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	12.074	15.136	10.019	15.136
Mask> 0.1, Z=3, Aggressive=False	Linear Regression		20.278	25.866	15.319	25.866

Table A.73.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Sum crown area of coniferous trees (m^2)

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	3.386	4.168	3.050	4.167
Mask> 0, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	3.400	4.397	2.028	4.397
Mask> 0, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	3.427	4.465	2.082	4.464
Mask> 1, Z=None, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	9.627	12.028	8.031	12.028
Mask> 0.1, Z=1, Aggressive=False	Linear Regression		15.699	19.573	10.461	19.573

Table A.74.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Count of deciduous trees

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.168	0.212	0.140	0.212
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.169	0.213	0.140	0.213
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	0.176	0.222	0.151	0.222
Mask> 1, Z=None, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	0.482	0.616	0.391	0.616
Mask> 0, Z=None, Aggressive=False	Linear Regression		0.905	1.159	0.736	1.159

Table A.75.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Count of coniferous trees

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 0, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.098	0.128	0.062	0.128
Mask> 0, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.099	0.129	0.062	0.129
Mask> 0.1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.129	0.163	0.098	0.163
Mask> 1, Z=1, Aggressive=False	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	0.334	0.427	0.249	0.427
Mask> 0.1, Z=1, Aggressive=False	Linear Regression		0.488	0.629	0.294	0.629

Table A.76.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Tree area coverage (%)

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.463	0.619	0.310	0.619
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.469	0.622	0.309	0.622
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	0.499	0.652	0.337	0.652
Mask> 1, Z=3, Aggressive=True	SVR	`C=1` `kernel=rbf`	1.822	3.192	0.800	2.962
Mask> 0.1, Z=1, Aggressive=False	Linear Regression		3.028	6.289	0.842	6.289

Table A.77.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Sum crown volume (m^3)

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	23.014	29.622	17.294	29.617
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	23.176	29.824	18.632	29.823
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	23.659	30.503	18.874	30.503
Mask> 1, Z=None, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	70.981	91.409	57.070	91.409
Mask> 0.1, Z=1, Aggressive=False	SVR	`C=1` `kernel=rbf`	97.009	128.410	77.984	128.399

Table A.78.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Mean tree height (m)

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.610	0.792	0.495	0.792
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.625	0.806	0.478	0.806
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	0.697	0.884	0.559	0.884
Mask> 1, Z=None, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	2.027	2.538	1.691	2.538
Mask> 0, Z=None, Aggressive=False	Linear Regression		3.227	4.068	2.670	4.068

Table A.79.: Performance Summary of Top 3, Median, and Worst Models in the Original Spatial Domain: Mean crown base height (m)

Experiment	Model	Params	MAE	RMSE	MAD	STD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.711	0.861	0.651	0.861
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.716	0.864	0.655	0.864
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	0.762	0.917	0.658	0.917
Mask> 0, Z=3, Aggressive=False	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.827	2.394	1.390	2.394
Mask> 0, Z=None, Aggressive=False	Linear Regression		3.510	4.433	2.848	4.433

Model Setup Comparison Cross-Validation per Variable

Table A.80.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Sum crown area of deciduous trees (m^2)

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	13.590	16.760	10.271	16.749
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	13.568	16.681	10.149	16.661
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	13.199	16.346	9.829	16.327
Mask> 1, Z=None, Aggressive=False	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	5.886	6.991	3.077	6.349
Mask> 0.1, Z=3, Aggressive=False	Linear Regression		7.056	8.075	−1.937	−4.953

Table A.81.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Sum crown area of coniferous trees (m^2)

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	8.666	10.231	5.530	9.557
Mask> 0, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	35.000	36.350	6.574	9.234
Mask> 0, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	34.964	36.284	6.566	9.196
Mask> 1, Z=None, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	3.495	3.565	0.770	2.076
Mask> 0.1, Z=1, Aggressive=False	Linear Regression		6.650	6.587	−1.992	−5.974

Table A.82.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Count of deciduous trees

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.577	0.699	0.484	0.675
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.574	0.692	0.493	0.665
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	0.566	0.685	0.468	0.660
Mask> 1, Z=None, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	0.285	0.320	0.241	0.270
Mask> 0, Z=None, Aggressive=False	Linear Regression		0.251	0.237	−0.106	−0.284

Table A.83.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Count of coniferous trees

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 0, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.243	1.298	0.227	0.355
Mask> 0, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.244	1.298	0.226	0.356
Mask> 0.1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	0.975	1.039	0.162	0.314
Mask> 1, Z=1, Aggressive=False	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	0.087	0.083	0.046	0.068
Mask> 0.1, Z=1, Aggressive=False	Linear Regression		0.334	0.323	−0.001	−0.148

Table A.84.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Tree area coverage (%)

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	2.207	4.577	1.412	4.321
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	2.179	4.547	1.448	4.314
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	2.167	4.535	1.388	4.298
Mask> 1, Z=3, Aggressive=True	SVR	`C=1` `kernel=rbf`	0.885	2.270	0.606	1.985
Mask> 0.1, Z=1, Aggressive=False	Linear Regression		−0.401	−1.433	0.748	−1.450

Table A.85.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Sum crown volume (m^3)

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	64.300	90.299	40.743	88.174
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	63.914	90.952	37.629	89.255
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	64.019	90.653	39.290	89.133
Mask> 1, Z=None, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	17.998	31.846	−2.762	31.114
Mask> 0.1, Z=1, Aggressive=False	SVR	`C=1` `kernel=rbf`	−18.340	−15.491	−29.678	−15.709

Table A.86.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Mean tree height (m)

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.381	1.733	1.206	1.659
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.391	1.743	1.240	1.675
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	1.321	1.669	1.135	1.598
Mask> 1, Z=None, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	0.303	0.347	0.358	0.208
Mask> 0, Z=None, Aggressive=False	Linear Regression		−0.984	−1.340	−0.787	−1.442

Table A.87.: Performance Shifts (Δ) of Top 3, Median, and Worst Models Between Original and Cross-Validation Regions: Mean crown base height (m)

Experiment	Model	Params	ΔMAE	ΔRMSE	ΔMAD	ΔSTD
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.186	1.576	0.873	1.527
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.206	1.599	0.905	1.538
Mask> 1, Z=1, Aggressive=True	RF Regressor	`max_depth=10` `max_features=log2` `min_samples_leaf=2`	1.167	1.549	0.966	1.496
Mask> 0, Z=3, Aggressive=False	RF Regressor	`max_depth=None` `max_features=log2` `min_samples_leaf=2`	1.183	1.229	1.227	0.882
Mask> 0, Z=None, Aggressive=False	Linear Regression		−0.739	−0.983	−0.444	−1.373

A.1.2 Context-Aware Label Enrichment and Multi-Scale Learning with the HELIX Framework

Forest Disturbance Forecasting from Fused Sentinel-1 and Sentinel-2 Data with Helix-Based Spatio-Temporal Label Enrichment

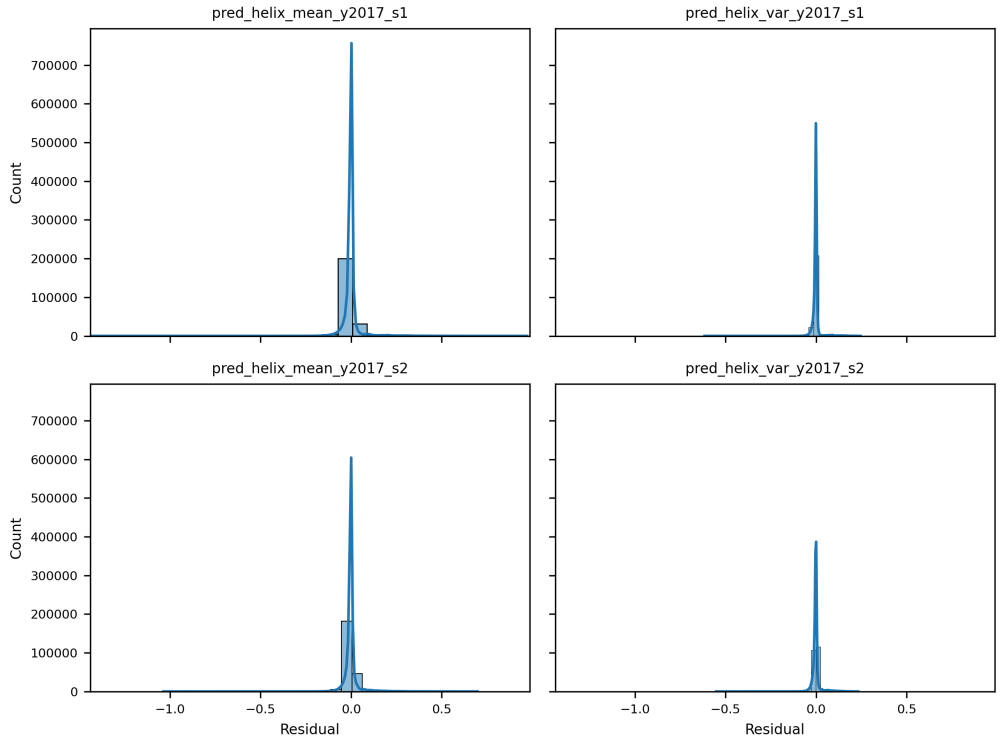


Figure A.1.: Residual distributions for predicted Helix descriptors.

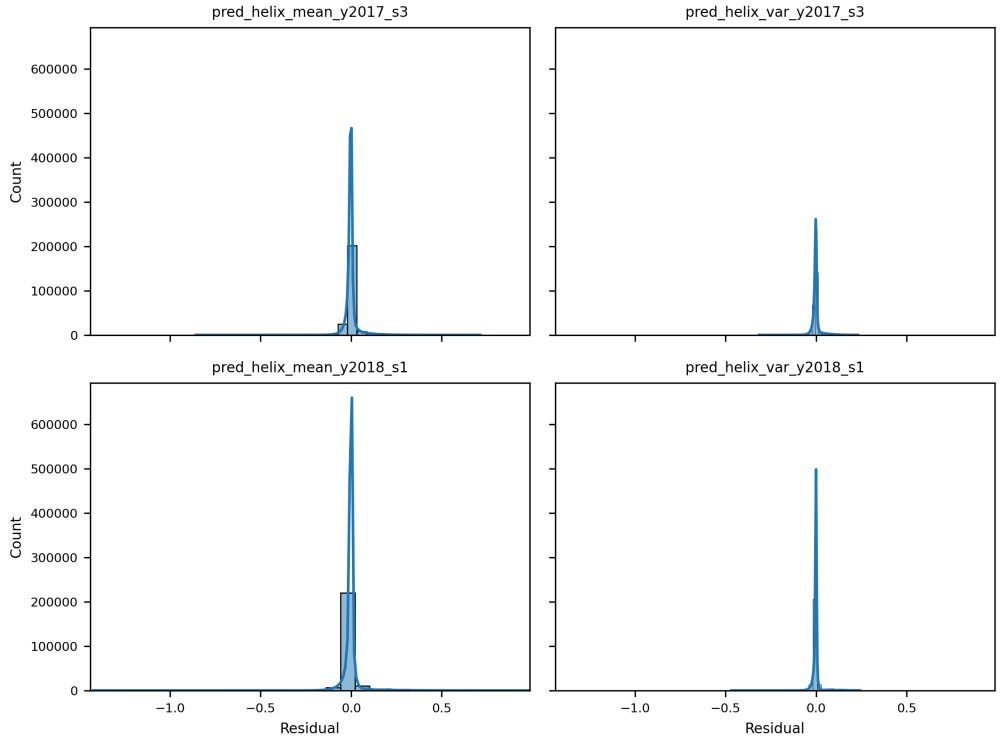


Figure A.2.: Residual distributions for predicted Helix descriptors.

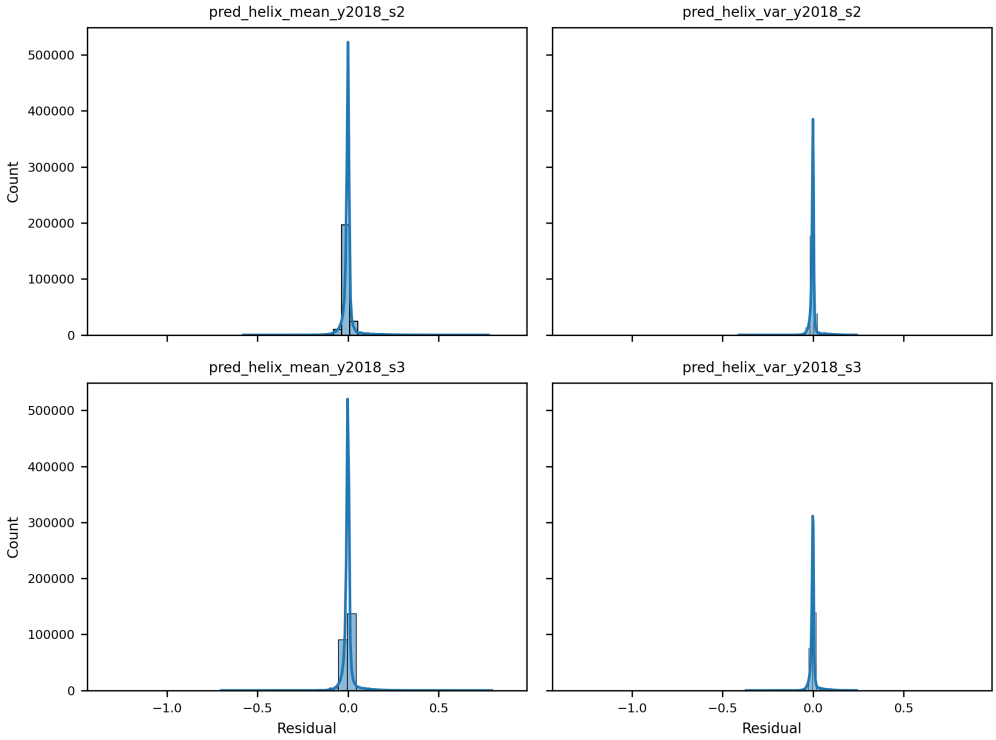


Figure A.3.: Residual distributions for predicted Helix descriptors.

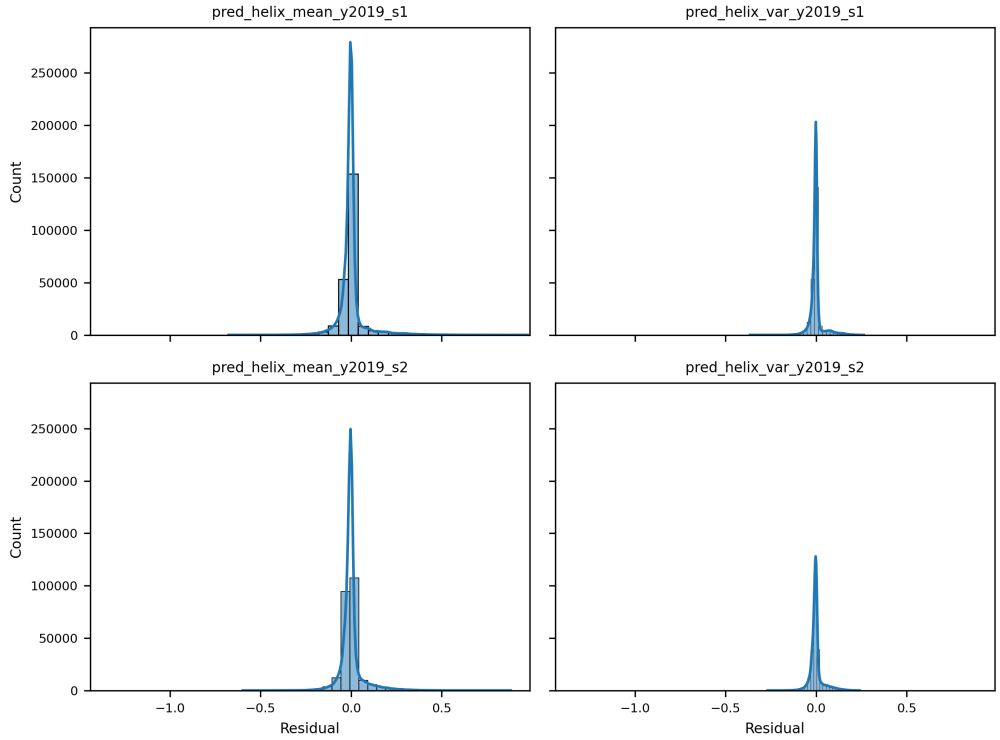


Figure A.4.: Residual distributions for predicted Helix descriptors.

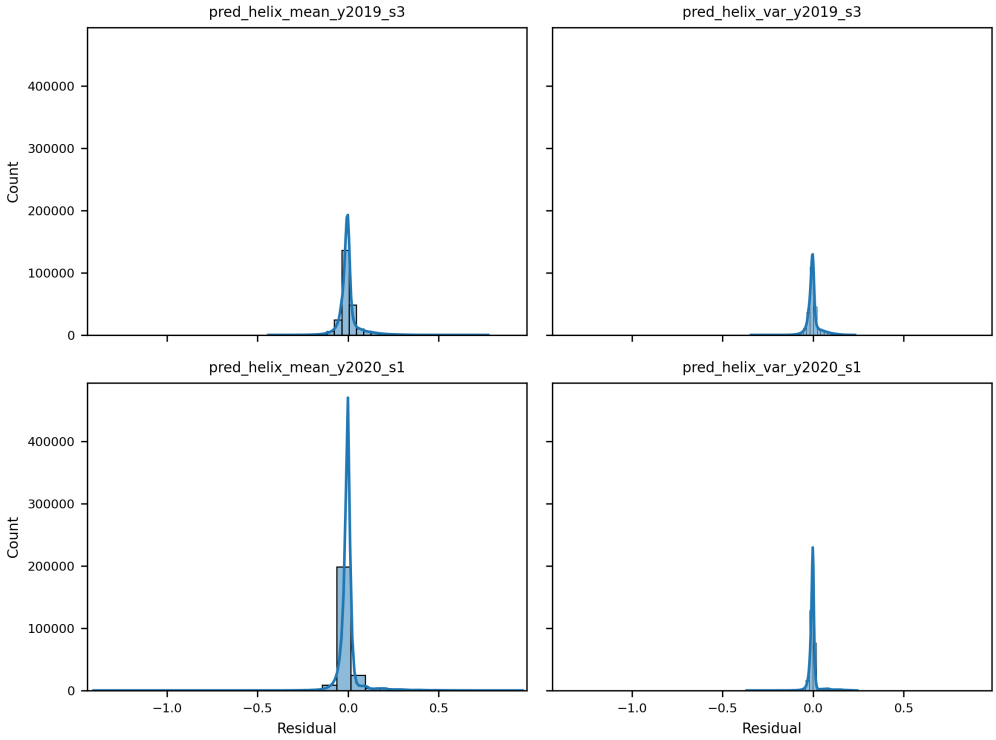


Figure A.5.: Residual distributions for predicted Helix descriptors.

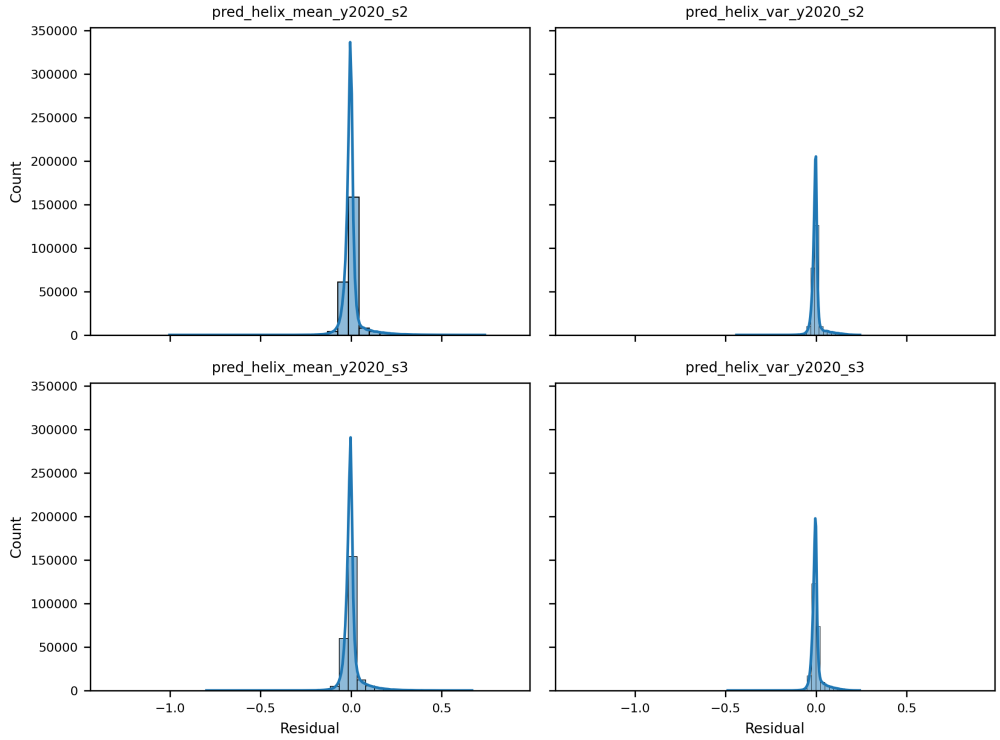


Figure A.6.: Residual distributions for predicted Helix descriptors.

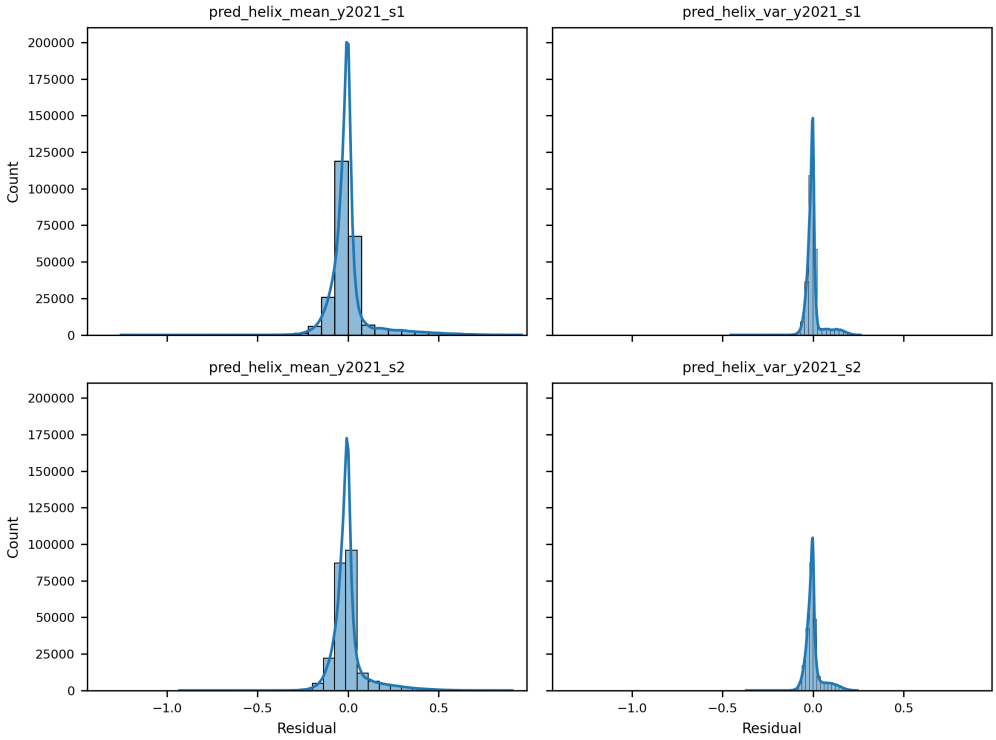


Figure A.7.: Residual distributions for predicted Helix descriptors.

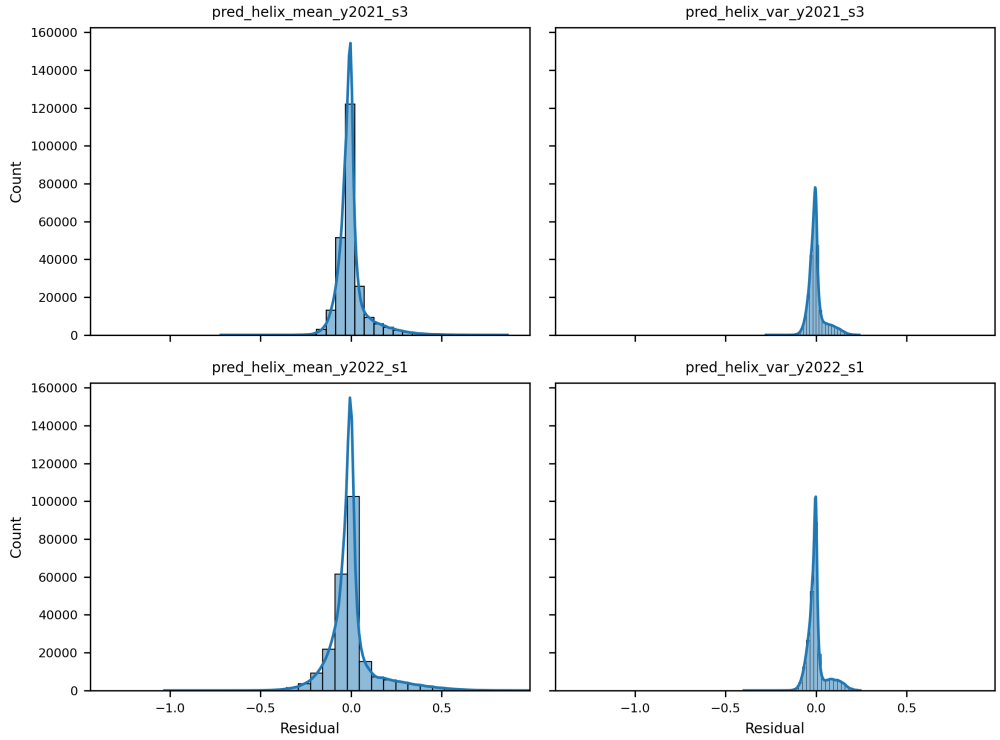


Figure A.8.: Residual distributions for predicted Helix descriptors.

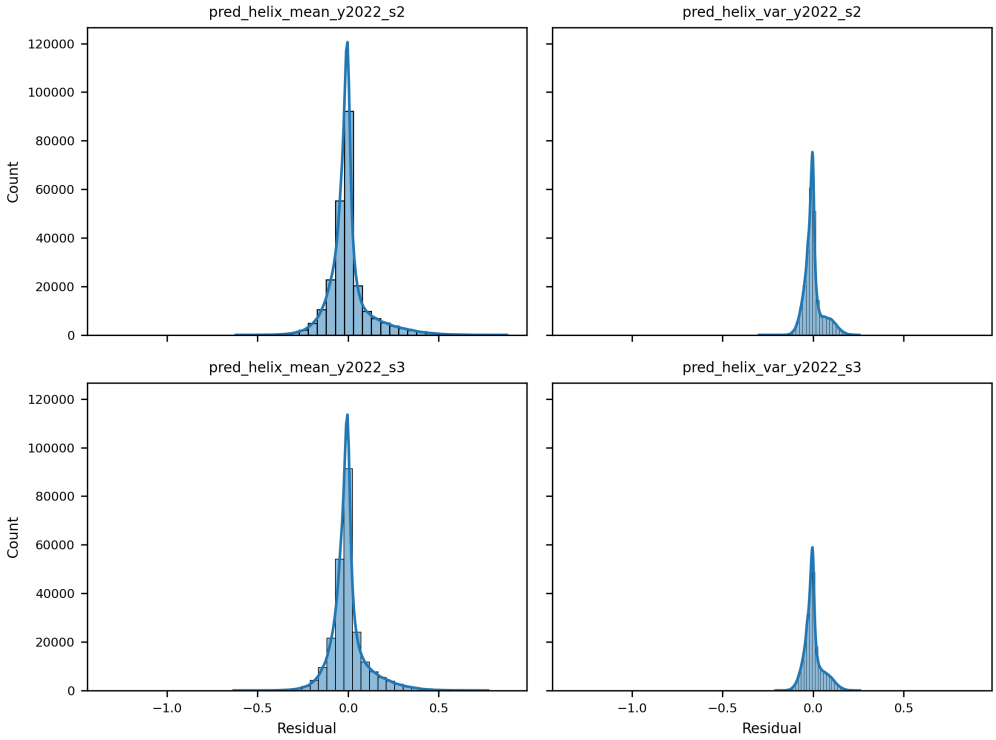


Figure A.9.: Residual distributions for predicted Helix descriptors.

Table A.88.: Per-band classification metrics using EO + individual helix features

Band	Precision	Recall	F1 Score	AUC
helix_mean_y2021_s1	0.8242	0.8206	0.8224	0.9941
helix_mean_y2021_s2	0.7832	0.7000	0.7392	0.9886
helix_var_y2021_s1	0.7823	0.6964	0.7368	0.9773
helix_var_y2021_s2	0.7655	0.6836	0.7222	0.9832
helix_mean_y2021_s3	0.7677	0.6468	0.7021	0.9846
helix_var_y2021_s3	0.7247	0.6740	0.6984	0.9827
helix_mean_y2022_s1	0.7905	0.4778	0.5956	0.9179
helix_mean_y2020_s1	0.7988	0.4742	0.5951	0.9119
helix_var_y2022_s1	0.8020	0.4726	0.5948	0.9131
helix_var_y2022_s2	0.7962	0.4746	0.5947	0.9190
helix_var_y2019_s3	0.7966	0.4726	0.5933	0.9150
helix_mean_y2018_s1	0.7926	0.4702	0.5903	0.9104
helix_var_y2020_s2	0.7873	0.4718	0.5901	0.9157
helix_var_y2022_s3	0.7903	0.4698	0.5893	0.9175
helix_mean_y2019_s3	0.7856	0.4714	0.5893	0.9133
helix_var_y2020_s3	0.7878	0.4702	0.5889	0.9210
helix_mean_y2018_s2	0.7894	0.4686	0.5881	0.9100
helix_mean_y2022_s2	0.7915	0.4670	0.5874	0.9189
helix_mean_y2020_s2	0.7919	0.4666	0.5872	0.9150
helix_var_y2019_s2	0.7952	0.4654	0.5872	0.9116
helix_mean_y2017_s3	0.7875	0.4678	0.5870	0.9104
helix_var_y2017_s1	0.7863	0.4674	0.5863	0.9103
helix_mean_y2017_s1	0.7839	0.4682	0.5863	0.9109
helix_var_y2017_s3	0.7936	0.4638	0.5855	0.9099
helix_var_y2019_s1	0.7877	0.4654	0.5851	0.9075
helix_var_y2018_s3	0.7870	0.4650	0.5846	0.9107
helix_mean_y2019_s2	0.7850	0.4654	0.5844	0.9108
helix_var_y2017_s2	0.7835	0.4654	0.5840	0.9107
helix_mean_y2019_s1	0.7752	0.4670	0.5829	0.9117
helix_mean_y2018_s3	0.7863	0.4630	0.5829	0.9096
helix_var_y2018_s2	0.7881	0.4622	0.5827	0.9091
helix_var_y2020_s1	0.7954	0.4583	0.5815	0.9119
helix_mean_y2020_s3	0.7905	0.4598	0.5815	0.9213
helix_mean_y2022_s3	0.7827	0.4618	0.5809	0.9237
helix_var_y2018_s1	0.7791	0.4622	0.5802	0.9099
helix_mean_y2017_s2	0.7786	0.4622	0.5801	0.9074

A.2 Code and Algorithms

A.2.1 Python-based implementation of the HCB Fusion

```
1
2  import os
3  import numpy as np
4  import glob
5  import re
6  from osgeo import gdal, gdal_array
7  import xarray as xr
8  import rasterio
9  from datetime import datetime, timedelta
10 import math
11
12 % Folders
13 input_folder = ""
14 processed_rasters_folder_name = "_fused"
15
16 # Create the full path for the new output folder within the input
       path
17 fused_output_folder = os.path.join(input_folder,
       processed_rasters_folder_name)
18 # Create the fused_output_folder if it doesn't exist
19 if not os.path.isdir(fused_output_folder):
20     os.makedirs(fused_output_folder)
21
22
23 def process_raster(input_path, fused_output_path):
24     src = rasterio.open(input_path)
25     dat = src.read()
26
27     # Validate if input raster has 8 bands
28     if dat.shape[0] != 8:
29         raise ValueError(f"The␣input␣raster␣must␣have␣8␣bands,␣but
              ␣it␣has␣{dat.shape[0]}␣bands.")
30
31     # Get metadata from the source dataset
32     meta = src.meta
```

```
33
34      ## DEFINITION OF ORTHOGONAL MATRICES
35      C = np.array([[1,1],[1, -1]]) # KOMPLEX
36      C = np.divide(C, 2)
37
38      Q = np.block([[C,C],[C, -C]]) # QUATERNION
39      Q = np.divide(Q, 2)
40
41      # Define the "no data" value -> stacked no-data value e.g.
            65535 | change if another no-data value was used!
42      no_data_value = 65535
43      # Check if at least one band at a pixel contains a "no data"
            value (e.g. 65535)
44      invalid_condition = np.any(dat == no_data_value, axis=0)
45      # Set all bands to 0 for pixels where at least one band
            contains a "no data" value
46      for band in range(dat.shape[0]):
47          dat[band, invalid_condition] = 0
48
49      ## Remove spatial dimensions
50      data = np.double(dat.reshape(dat.shape[0], dat.shape[1]* dat.
            shape[2])) # 0: 8 Bands from Input; 1:  X; 2: Y
51
52      ## SENTINEL-1 & SENTINEL-2 INVALIDS
53      invi = np.where(np.min(data, axis = 1)<1)
54
55      # SENTINEL-1 | K-SAR
56      data_sar = data[0:4, :]
57
58      data_sar[data_sar == 1] = 2 # search indices from INVALIDS
59      data_sar[data_sar == 65535] = 65534
60      data_uint = np.uint16(data_sar)
61      data_double = np.double(data_sar)
62
63      ksar = (data_double[0:4, :]-32768)/32767 # Conversion of
            Sentinel-1 from DNs to normalised Kennaughs
64
65      ksar[0,:] = np.divide((1+ksar[0,:]), np.dot((1-ksar[0,:]),(np.
            sqrt(2)))) # Conversion of the normalised intensity to
            linear
66
```

```
67      for k in range(1, 4):
68
69          ksar[k,:] = np.multiply(ksar[k,:],ksar[0,:]); #
                Conversion of the normalised polarimetric Kennaughs
                to linear
70
71
72      # Prepare Export K-SAR
73      ksar_out = np.uint16(np.dot(ksar, 32767 )+ 32768)
74      ksar_out = np.double(ksar_out.reshape(4, dat.shape[1], dat.
            shape[2]))
75
76      # Identify the border pixels based on the earlier condition
            that input data pixels have a value of 65535 are set to 0
77      input_data = dat[0]
78      # Identify the border pixels based on the earlier condition
            that input data pixels have a value of 65535 are set to 0
79      border_condition = (input_data == 0)
80      ksar_out[:, border_condition] = 0
81
82      # Generate output filenames based on input filename (e.g.
            include the date)
83      filename_ksar = os.path.basename(input_path)
84      fused_output_filename_ksar = os.path.join(fused_output_path, f
            "{os.path.splitext(filename_ksar)[0]}_k_sar.tif")
85
86      # Export fused and fused_n data with 8 bands in metadata
87      meta['count'] = 4
88      meta['nodata'] = 0
89
90      with rasterio.open(fused_output_filename_ksar, 'w', **meta) as
            dst:
91          dst.write(ksar_out)
92
93
94      ## SENTINEL-2 | K-OPT
95      data_opt = data[4:8, :]
96
97      bw = np.array([66, 36, 31, 106]) # Bandwidth adjustment
98      bwadj = 60 / bw / 10000
99      data_opt = (data_opt.T* bwadj).T
```

```
100
101     kopt = np.zeros(data[4:8,:].shape)
102     kopt = np.dot(Q, data_opt) # Conversion to linear Kennaughs
103
104
105     # Prepare Export K-OPT
106     kopt_out = np.uint16(np.dot(kopt, 32767 )+ 32768)
107     kopt_out = np.double(kopt_out.reshape(4, dat.shape[1], dat.
            shape[2]))
108
109     kopt_out[:, border_condition] = 0
110
111     # Generate output filenames based on input filename (e.g.
            include the date)
112     filename_kopt = os.path.basename(input_path)
113     fused_output_filename_kopt = os.path.join(fused_output_path, f
            "{os.path.splitext(filename_kopt)[0]}_k_opt.tif")
114
115     # Export fused and fused_n data with 8 bands in metadata
116     meta['count'] = 4
117     meta['nodata'] = 0
118
119     with rasterio.open(fused_output_filename_kopt, 'w', **meta) as
             dst:
120         dst.write(kopt_out)
121
122
123     ## DATAFUSION | K-FUS
124     ksar_vstack = np.vstack([ksar, ksar])
125     kopt_vstack = np.vstack([kopt, -kopt])
126
127     kfus = np.add(ksar_vstack, kopt_vstack)
128
129     ## NORMALISATION  | K-NOR
130     knor = np.zeros(kfus.shape)
131
132     refi = 0.2
133     knor[0,:] = np.divide((kfus[0,:] - refi), (kfus[0,:] + refi))
            # Normalisation of the total intensity
134
135     for k in range(1,8):
```

```
136         knor[k,:] = np.divide(kfus[k,:],kfus[0,:]); #
                Normalisation of the poalrimetric-spectrometric
                Kennaughs
137
138     ## Preparation for output
139     kndn = np.uint16(np.dot(knor, 32767 )+ 32768)
140     kndn = np.double(kndn.reshape(dat.shape[0], dat.shape[1], dat.
            shape[2])) # recreate spatial dimensions
141
142     kndn[:, border_condition] = 0
143
144     # Generate output filenames based on input filename (e.g.
            include the date)
145     filename = os.path.basename(input_path)
146     fused_output_filename = os.path.join(fused_output_path, f"{os.
            path.splitext(filename)[0]}_fused.tif")
147
148     # Export fused and fused_n data with 8 bands in metadata
149     meta['count'] = 8
150     meta['nodata'] = 0
151
152     with rasterio.open(fused_output_filename, 'w', **meta) as dst:
153         dst.write(kndn)
154
155 # Loop over each raster file and process it
156 for input_path in input_raster_paths:
157     # Call the process_raster function with the input path and the
            shared output folder path
158     process_raster(input_path, fused_output_folder)
159
160 def is_power_of_two(n):
161     return (math.ceil(math.log2(n)) == math.floor(math.log2(n))) #
            check for power of 2
162
163 def read_tif_folder(folder_path):
164     """
165     Read multiple .tif files from a folder and returns it as a
            DataArray object.
166     """
167     # Get list of all .tif files in folder
168     file_list = glob.glob(folder_path + "/*_fused.tif")
```

```
169
170     # Check if the number of files is a power of two
171     num_files = len(file_list)
172     if not is_power_of_two(num_files):
173         raise ValueError("The number of TIF files is not a power
                of two. Found {} files.".format(num_files))
174
175     # Get list of dates from file names using regex
176     dates = []
177     for f in file_list:
178         match = re.search(r'(?<=.{0})\d{8}', f) # date = first 8
                letters according to output IMAGEFUSION YYYYMMDD
179         if match:
180             dates.append(datetime.strptime(match.group(), '%Y%m%d'
                    ).date()) # date format
181
182     # Sort files and dates by date
183     file_list = [f for _, f in sorted(zip(dates, file_list))]
184     dates = sorted(dates)
185
186     # Read in the first .tif file to get metadata
187     with rasterio.open(file_list[0]) as src:
188         transform = src.transform
189         crs = src.crs
190         meta = {
191             'driver': src.driver,
192             'height': src.height,
193             'width': src.width,
194             'count': len(file_list),
195             'dtype': src.dtypes[0],
196             'nodata': src.nodata,
197             'crs': crs,
198             'transform': transform
199         }
200
201     # Read in .tif files as a list of numpy arrays
202     data_arrays = []
203     for f in file_list:
204         raster = gdal.Open(f)
205         data = raster.ReadAsArray()
206         data_arrays.append(data)
```

```
207
208     ds = xr.DataArray(np.stack(data_arrays), dims=['time', 'band',
            'y', 'x'], coords={'time': dates})
209
210     return ds, meta
211
212 ds, meta = read_tif_folder(fused_output_folder) # execute the
        function 'read_tif_folder'
213
214 length = ds.shape[0]
215 tdim = int(2 ** np.ceil(np.log2(length)))
216 ddim = ds.shape[1]
217
218 T = np.array([[1, 1], [1, -1]]) / 2
219 while T.shape[0] < tdim:
220     T = np.block([[T, T], [T, -T]]) / 2
221
222 # Flatten the spatial dimensions ('y', 'x') to a single 'pixels'
        dimension
223 data_reshaped = ds.stack(pixels=('y', 'x'))
224 # Convert to numpy
225 data_reshaped_numpy = data_reshaped.values
226
227 # Find and mask no-data values [0]
228 data_reshaped_numpy_zero = (data_reshaped_numpy == 0)
229
230 # INVALIDS
231 data_reshaped_numpy[data_reshaped_numpy == 65535] = 65534
232 data_reshaped_numpy[data_reshaped_numpy < 2] = 2
233
234 data_reshaped_numpy = data_reshaped_numpy.astype(np.float64)
235
236 # Conversion to linear
237 kens = (data_reshaped_numpy - 32768) / 32767
238 # First band linear conversion
239 refi = 0.2
240 kens[:, 0, :] = refi * ((1+kens[:, 0, :]) / (1-kens[:, 0, :])) #
        Conversion of the normalised intensity to linear
241
242 # Polarimetric Kennaughs conversion for remaining bands
243 for k in range(1, 8):
```

```
    kens[:, k, :] = kens[:, k, :] * kens[:, 0, :]   # Conversion
        of the normalised polarimetric Kennaughs to linear

kens_tp = kens.transpose(2, 1, 0)

kens_tf = np.zeros(kens_tp.shape)

# Perfrorm temporal fusion
for s in range(ddim):
    kens_tf[:, s, :] = np.dot(kens_tp[:, s, :], T)

# Perform normalization
for z in range(tdim):  # z goes from 0 to tdim-1
    for s in range(ddim):  # s goes from 0 to ddim-1
        if not (z == 0 and s == 0):  # Perform normalization
            except for the first element
            kens_tf[:, s, z] = kens_tf[:, s, z] / kens_tf[:, 0, 0]
                # Normalisation of the polarimetric-spectrometric
                characteristics

# Normalize the first element
kens_tf[:, 0, 0] = (kens_tf[:, 0, 0] - refi) / (kens_tf[:, 0, 0] +
    refi)

# convert back to Uint-16
kndn = (kens_tf * 32767 + 32768).astype(np.uint16)

# set no-data values (e.g. image border to 0)
kndn_transposed = kndn.transpose(2, 1, 0)
kndn_transposed[data_reshaped_numpy_zero] = 0
kndn = kndn_transposed.transpose(2, 1, 0)

# Restore spatial dimensions (pixel to 'y', 'x')
kfus = kndn.reshape(ds.values.shape[2], ds.values.shape[3], kndn.
   shape[1] * kndn.shape[2])
kfus = np.transpose(kfus, (2, 0, 1))

# convert to Uint-8
masked_kfus = (kfus == 0)
kfus_u8 = (kfus.astype(float)-32768)/32767
kfus_u8[masked_kfus] = 0
```

```
kfus_u8 = np.tanh(np.arctanh(kfus_u8)*5)
kfus_u8 = np.round(kfus_u8*127+128)
kfus_u8[masked_kfus] = 0
kfus_u8 = np.uint8(kfus_u8)
```

Listing A.1: Fusion of Sentinel-1/2 raster data to hypercomplex basis representations.

Colophon

This thesis was typeset with LaTeX 2_ε. It uses the *Clean Thesis* style developed by Ricardo Langner. The design of the *Clean Thesis* style is inspired by user guide documents from Apple Inc. Download the *Clean Thesis* style at http://cleanthesis.der-ric.de/.