This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409

Gap Filling in Satellite Time Series: In Search of Simple Yet Reliable Reconstruction Methods

Andreas Schmitt and Sarah Hauser

***Abstract*—Complete and temporally consistent satellite image time series are essential for many Earth observation applications. This study formulates gap filling as a post-mask reconstruction problem and introduces two transparent operators that require neither external supervised training nor area-specific fine-tuning. Hypercomplex Bases (HCB) reconstruct missing observations directly from the spectro-temporal structure of the observed time series, whereas HELIX additionally exploits land-cover information through analytically estimated class-specific seasonal guidance. Reconstruction fidelity is first examined under spatially heterogeneous clustered cloud/shadow missingness, where unaffected observations from the same acquisition remain available. HCB and HELIX retain low spectral-angle deviations of 7.9° and 8.2°, respectively, together with comparatively limited spatial- and temporal-gradient changes and without the pronounced structural artefacts observed for several learned baselines. To isolate temporal reconstruction under common support conditions, a complementary complete-date benchmark removes entire acquisitions and systematically varies temporal support across seven study areas, all five main Köppen–Geiger climate groups, two years, 30 controlled gap scenarios, four spectral bands, and eleven land-cover classes. In this benchmark, HELIX achieves the lowest aggregate MAE (688.4 DN), MedAE (188 DN), RMSE (1636.4 DN), and mean SAM (7.38°), while HCB follows with an MAE of 879.4 DN, MedAE of 262 DN, RMSE of 1782.4 DN, and mean SAM of 7.88°. Both remain stable under interpolation-like and extrapolation-like configurations with one-sided temporal support. HCB provides strong reconstruction and spectral fidelity at low computational cost without ancillary semantic information, whereas HELIX complements it through semantic seasonal constraints that improve radiometric accuracy. The two approaches therefore represent complementary transparent operating points for robust and auditable reconstruction of masked optical satellite time series.**

***Index Terms*—Gap filling, satellite image time series, Sentinel-2, masked reconstruction, spectral-temporal reconstruction, land-cover constraints, spectral fidelity, transparent reconstruction.**

I. Introduction

OPTICAL satellite image time series are a key data source for land-cover monitoring, vegetation phenology, disturbance analysis, environmental assessment, urban climate studies, and model based decision support [1]–[3]. These applications require temporally consistent observations before, during, and after the event to be monitored. The utility of Earth observation (EO) archives therefore depends not only on spatial resolution and revisit time, but also on the continuity and reliability of the resulting data cubes.

In practice, optical time series often remain incomplete. Clouds and cloud shadows are the dominant source of invalid observations [4]; additional gaps arise from relief shadows, acquisition failures, quality issues, sensor artefacts, and product-specific masking rules. Higher revisit frequencies and multi-sensor constellations reduce this problem, but cannot eliminate it. For Sentinel-2, cloud-cover inhomogeneity can strongly reduce the number of usable observations [5], and even combined Landsat–Sentinel acquisition strategies cannot guarantee complete temporal coverage across regions and seasons [6], [7]. If such gaps are not handled consistently, they propagate into downstream products and may bias phenological timing, change detection, evapotranspiration estimates, land-surface temperature analyses, and other environmental indicators.

Modern EO processing increasingly separates invalid-observation detection from downstream analysis. Landsat Analysis Ready Data [8], [9], CEOS CARD4L specifications [10], Harmonized Landsat–Sentinel products [11], [12], daily mosaic products [13], and operational cloud and quality masks reduce the need for each application to solve cloud detection independently. Thus, cloud detection is not the methodological problem addressed here. This study assumes that invalid observations have already been identified by an external masking algorithm. The remaining task is post-mask reconstruction: estimating plausible spectral values at masked time steps from the available spectral-temporal support.

This distinction is central, because gap masks may originate from cloud screening, quality layers, acquisition failures, sensor artefacts, or deliberate validation masks. A general reconstruction method should therefore not depend on one cloud detector, one quality-layer convention, one auxiliary sensor, or one fixed temporal gap geometry. Instead, it should operate on co-registered multiband time series and externally defined masks, while preserving observed pixels unchanged.

Reliable reconstruction requires more than a low mean error. Reconstructed values should preserve temporal behaviour, remain spectrally consistent, and remain stable across regions, years, land-cover conditions, gap densities, and temporal gap positions. This includes interpolation-like cases, where missing dates are surrounded by valid observations, and extrapolation-like cases at the beginning or end of a sequence. Traceability is also required, because gap filled values often enter downstream modelling chains without man-

Manuscript received XX XX, 2026; revised XX XX, 2026.

Corresponding author: Andreas Schmitt.

Andreas Schmitt is with the Geoinformatics Department, Hochschule München University of Applied Sciences, Karlstraße 6, 80333 Munich, Germany, and with the Institute for Applications of Machine Learning and Intelligent Systems (IAMLIS), Lothstraße 34, 80335 Munich, Germany (e-mail: andreas.schmitt@hm.edu).

Sarah Hauser is with the Institute for Applications of Machine Learning and Intelligent Systems (IAMLIS), Lothstraße 34, 80335 Munich, Germany, with the Institute of Photogrammetry and Remote Sensing (IPF), Karlsruhe Institute of Technology (KIT), Englerstraße 7, 76131 Karlsruhe, Germany, and with the Faculty of Earth Sciences, Geography and Astronomy, Department of Geography and Regional Research, University of Vienna, 1010 Vienna, Austria (e-mail: sarah.hauser@univie.ac.at).

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409

ual inspection. Computational efficiency is likewise relevant because reconstruction may be applied repeatedly across large EO archives; avoiding unnecessary computational demand therefore concerns both operational scalability and the broader environmental footprint of EO processing [14]–[16].

This paper evaluates two transparent reconstruction operators that require no external supervised training or target-area fine-tuning for masked optical satellite time series. Hypercomplex Bases (HCB) exploit the spectral-temporal signature of the observed multiband time series. HELIX additionally uses land-cover information as a semantic reconstruction constraint. Both methods are applied with fixed settings to Sentinel-2 L2A 120 m Mosaic time series [17] from seven globally distributed study areas covering all main Köppen–Geiger climate groups [18], two years, 30 controlled gap scenarios, four spectral bands, and eleven ESA WorldCover classes [19]. The operator formulation requires only co-registered multiband data and an external gap mask.

The objective is to test whether these transparent operators can provide robust post-mask reconstruction across heterogeneous environmental and temporal conditions without external supervised training or target-area fine-tuning. HCB and HELIX are compared with linear interpolation, general-purpose time series models, and satellite-specific deep learning reconstruction methods. Reliability is evaluated through accuracy, robust dispersion, signed bias, scenario stability, land-cover sensitivity, and spectral fidelity, while computational efficiency is assessed separately through runtime and training requirements. The aim is not to replace specialised cloud-removal models, but to establish reproducible and auditable reconstruction operators in order to adequately fill already identified gaps in time series.

A. State of the Art in Gap Filling

Gap filling in optical satellite image time series spans deterministic interpolation, statistical smoothing, machine learning, deep learning, cross-sensor fusion, and uncertainty-aware reconstruction. These approaches differ in the information they use, the assumptions they impose, and the degree to which their behaviour remains traceable. For the post-mask setting considered here, the central question is not how invalid observations are detected, but how reliably already masked observations can be reconstructed under heterogeneous temporal gap geometries.

Classical methods remain relevant because they are efficient, reproducible, and inspectable. Linear interpolation, spline fitting, harmonic reconstruction, Whittaker–Eilers smoothing, Savitzky–Golay filtering, and asymmetric Gaussian functions impose temporal continuity or seasonal smoothness [20]–[22]. TIMESAT operationalises such curve-fitting strategies for vegetation-index time series [23]. Spatial extensions include kriging [24], [25], segmentation-based filling [26], [27], and similar-pixel approaches for Landsat scan-line corrector gaps [28]. Low-rank and tensor methods exploit redundancy across space, time, and spectra [29], [30].

Class-aware strategies provide a further transparent reconstruction route. Wang et al. [31] combine object-based segmentation and unsupervised classification to account for landscape heterogeneity in PlanetScope time-series gap filling, while CLEAR combines internally derived classes with class-based linear regression and iterative residual compensation for optical cloud-gap reconstruction [32]. These methods demonstrate that class structure can provide useful reconstruction support, although the classes are derived from the image data themselves. Classical and transparent approaches remain attractive, but can become fragile under long gaps, abrupt surface changes, heterogeneous landscapes, irregular sampling, and extrapolation-like gaps at the beginning or end of a sequence.

Machine learning methods increase flexibility by learning non-linear relations between missing observations and predictors. Random forests and related ensemble models have been used for soil-moisture reconstruction [33], vegetation-index gap filling [34], and thick-cloud Landsat reconstruction [35]. Such models can include temporal lags, spatial descriptors, meteorological variables, and land-cover information. However, they usually require feature engineering, fitted parameters, and recalibration when the sensor product, study region, predictor distribution, or gap regime changes. Semantic information can improve reconstruction, but often enters as an additional predictor rather than as an explicit and inspectable reconstruction constraint.

Deep learning has advanced cloud removal and optical time series reconstruction substantially. Convolutional networks [36]–[39], adversarial completion models [39]–[41], recurrent and attention-based temporal models [42]–[44], and spatio-temporal partial convolutions [45] exploit learned spatial and temporal context.

Related multitemporal EO research has also explored structured representations beyond direct reconstruction. Fourier-domain feature interaction has been used to suppress style-induced pseudochanges in bitemporal change detection [46], while wavelet-based frequency decoupling has been applied to separate and interact temporal feature components [47]. Prototype-aware representations and regularization have further been used to improve semantic consistency under noisy weak supervision [48]. These studies address change detection rather than missing-observation reconstruction, but demonstrate the broader utility of frequency-domain and semantic representations in multitemporal EO. Optical–SAR fusion further improves reconstruction when microwave observations are available [49]–[51]. Sentinel-2 studies have reconstructed vegetation-index trajectories [52]–[54], reflectance time series [55], and leaf-area-index time series across climatic and agricultural settings [56]. These methods demonstrate the strength of satellite-specific learning, but commonly rely on training data, auxiliary modalities, cloud-removal assumptions, or sensor-specific input conventions.

Recent satellite-specific models address reconstruction more directly. U-TILISE performs optical satellite time series cloud removal with a spatio-temporal sequence-to-sequence architecture and attention-based temporal encoding [57]. UnCRtainTS extends optical time series reconstruction with uncertainty estimation [58]. GANFilling uses a generative sequence-to-sequence design for spatio-temporal gap filling of Sentinel-

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409

2 visible and near-infrared reflectances [59]. RESTORE-DiT formulates satellite image time series reconstruction as multimodal conditional generation with a sequential diffusion transformer [60]. Mask-informed temporal embedding targets efficient long-term spectral-temporal reconstruction through temporal segments and explicit mask information [61]. Together, these studies show the rapid shift from generic temporal imputation toward satellite-specific spatial, temporal, multimodal, and uncertainty-aware reconstruction. They also define the relevant comparison context for the present benchmark, where transparent reconstruction operators are evaluated against both generic time series models and satellite-specific neural approaches.

General-purpose time series models provide a complementary benchmark. SAITS (Self-Attention-based Imputation for Time Series) is designed for multivariate time series imputation [62]. TimesNet models multi-periodic temporal variation for several time series tasks, including imputation [63]. MOMENT provides transferable representations for general time series analysis and reconstruction [64]. Such models test whether generic temporal modelling transfers to optical EO reconstruction. However, they are not specifically designed around satellite reflectance, spectral consistency, land-cover structure, complete-scene missingness, or EO-specific edge gaps.

Existing methods therefore occupy a clear trade-off space. Classical methods are transparent and efficient, but limited under long, structured, irregular, or temporally asymmetric gaps. Learned methods are more flexible and can exploit spatial, temporal, multimodal, and semantic context, but require training data, fitted parameters, higher computational effort, and careful domain-transfer validation. High-capacity EO models often depend on large benchmark and training infrastructures, such as EarthNet2021 [65], which are valuable for model development but bind performance assessment to represented spatial, climatic, sensor-specific, and temporal distributions. Their data volume and compute requirements also imply a non-negligible environmental footprint. The open question is whether transparent reconstruction operators without external supervised training can close part of this gap by retaining auditability while improving robustness under heterogeneous post-mask reconstruction scenarios.

B. Research Gap and Hypotheses

The research gap is not the absence of gap filling methods, but rather the limited evidence for reconstruction operators that are simultaneously transparent, transferable in formulation, computationally lightweight, independent of external supervised training or target-area fine-tuning, and robust under heterogeneous masked satellite time series. This gap emerges from the trade-off in the state of the art: classical methods are auditable but often too limited for difficult gap geometries, whereas learned methods depend on training data, learned parameters, computational resources, and domain-transfer constraints.

This gap matters because post-mask reconstruction is broader than cloud removal. In operational EO data cubes, masks may originate from cloud screening, quality layers, acquisition failures, sensor artefacts, or deliberate validation masks. A reconstruction method should therefore not depend on a specific cloud detector, quality-layer convention, auxiliary modality, training domain, band configuration, or product-specific metadata structure. It should operate once the gap mask is known.

The difficult cases are not only small missing patches. Entire acquisition dates may be unavailable or unusable, leaving no same-date spatial support for reconstruction. Gaps may also be irregular, clustered, or located at the temporal edges of the sequence. Middle gaps are closer to interpolation, whereas gaps at the beginning or end require extrapolation from one-sided temporal support. These configurations are critical for large-scale EO archives, yet they are less compatible with methods that assume dense temporal context, fixed sampling, local spatial completion, or sufficient training examples for the target distribution.

Reconstruction without external supervised training or target-area fine-tuning is important for this setting for three reasons. First, global or continental EO applications cannot assume that dedicated training data are available for every region, year, product, land-cover composition, and gap pattern. Second, analytically defined operators support reproducibility because identical inputs and methodological settings lead to reproducible outputs. Third, avoiding supervised training, fine-tuning, and repeated model selection reduces computational cost and the associated environmental footprint. This does not reduce the value of learned reconstruction models; it defines a complementary operating point for scalable and auditable post-mask reconstruction.

This study addresses this operating point through two reconstruction principles. *Feature-side* reconstruction uses information contained in the observed multiband time series itself, especially its spectral and temporal structure. *Label-side* reconstruction additionally uses external semantic information, here land-cover labels, as a reconstruction constraint. HCB represents the feature-side principle, whereas HELIX represents the label-side principle. Both approaches are deterministic for fixed inputs and methodological settings and require neither external supervised training nor target-area fine-tuning.

The evaluation compares linear interpolation, harmonic regression, HCB, HELIX, general-purpose time series models, and satellite-specific deep learning reconstructors within a common post-mask reconstruction protocol. The experimental design deliberately spans different forms and degrees of missing support, ranging from spatially heterogeneous within-acquisition gaps to complete-date missingness, different gap densities and temporal positions, and interpolation-like and extrapolation-like configurations. Robustness is further examined across five main Köppen–Geiger climate groups (Fig. 1) [18], all ESA WorldCover classes [19], multiple regions (Table I), two years, and four spectral bands. This design allows the two proposed reconstruction principles to be compared with classical and learned approaches not only in terms of point-wise reconstruction accuracy, but also with respect to spectral and structural fidelity, stability under changing support conditions, and computational requirements. Four

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409

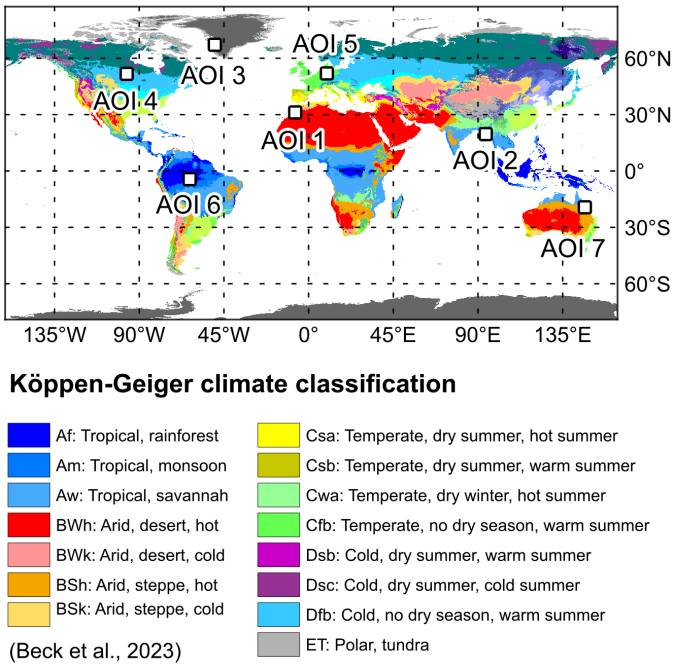


Fig. 1. Location of the seven AOIs with respect to the Köppen–Geiger climate zones [18]. The benchmark spans six continents and all five main climate groups.

TABLE I
OVERVIEW OF THE SEVEN BENCHMARK AOIS. KÖPPEN–GEIGER GROUPS ARE REPORTED AS MAIN CLIMATE GROUPS: A TROPICAL, B ARID, C TEMPERATE, D COLD CONTINENTAL, AND E POLAR [18]. DOMINANT LAND-COVER CLASSES REFER TO ESA WORLDCOVER 2020 AGGREGATED TO THE 120 M GRID [19].

AOI	Cont.	Location	Clim.	Dominant land-cover setting
1	Africa	Morocco, High Atlas	B/C/D	Bare / sparse vegetation, grassland
2	Asia	Myanmar, Rakhine State	A/C	Tree cover, mangroves, cropland
3	N. Atlantic	Greenland, ice margin	E	Snow and ice, moss and lichen, grassland
4	N. America	Canada, Manitoba	D	Tree cover, water bodies, grassland, shrubland
5	Europe	Germany, central lowlands	C	Cropland, tree cover, grassland
6	S. America	Brazil, central Amazon	A	Tree cover, water bodies, herbaceous wetland
7	Australia	Queensland, Ross River	A/C	Water bodies, tree cover, grassland

research hypotheses are proposed:

1) **Complexity of real world cloud scenarios:** Clouds can appear at any time in any place of the image leading to a very heterogeneous temporal signature of each pixel. These scenarios do not serve as comparable benchmark as the spatio-temporal variability causes hundreds degrees of freedom.
2) **Reconstruction robustness depends on gap geometry:** Reconstruction performance changes with the number, density, clustering, and temporal position of gaps. Reliable methods should remain stable not only for interpolation-like gaps, but also for complete-scene, irregular, and extrapolation-like configurations.
3) **Structural constraints improve reconstruction stability:** Information beyond local temporal interpolation can stabilise reconstruction under difficult gap geometries. HCB tests this through spectral-temporal structure in the observed multiband signal, whereas HELIX tests it through land-cover-constrained reconstruction.
4) **Reliable post-mask reconstruction requires accuracy, stability, and traceability:** Reconstruction operators without external supervised training or target-area fine-tuning can be competitive with learned baselines while remaining auditable and computationally lightweight. Reliability is therefore assessed not only by mean error, but also by robustness, scenario stability, regional and climatic consistency, land-cover sensitivity, spectral fidelity, and runtime.

The two novel approaches HCB and HELIX are subsequently evaluated against seven established reconstruction approaches under a common experimental protocol designed to test these hypotheses.

II. DATA AND EXPERIMENTAL DESIGN

The benchmark uses complete reference time series from which artificial gaps are generated in a controlled manner. This design is absolutely required because natural cloudy observations do not provide known gap-free reference values. Complete Sentinel-2 mosaic time series are therefore used as reference data, and artificial gaps are introduced by withholding valid acquisition dates. All methods reconstruct the same withheld observations and are evaluated against the same measured (not estimated) reference values.

Seven areas of interest (AOIs) are selected across six continents and all five main Köppen–Geiger climate groups [18]. The AOIs cover tropical, arid, temperate, cold continental, and polar environments, with land-cover settings including forests, croplands, sparse vegetation, snow and ice, wetlands, water bodies, and built-up areas. Fig. 1 shows their distribution, and Table I summarises their main environmental characteristics. For each AOI, all available Sentinel-2 L2A 120 m Mosaic observations [17] from 2019 and 2020 are used. Each AOI covers approximately 40 km $\times$ 40 km at 120 m spatial sampling. The product provides analysis-ready, co-registered surface-reflectance mosaics on a common grid, which reduces scene-level geometric and radiometric inconsistencies and enables comparable complete-date withholding experiments across all AOIs. The temporal sampling follows the available mosaic acquisition dates in each AOI–year rather than a fixed monthly aggregation. The experiments use four optical bands: blue (B02), green (B03), red (B04), and near infrared (B08). These visible–NIR bands form a common spectral core across all evaluated methods and support both reflectance-error and spectral-angle evaluation. Red-edge and SWIR bands are not included in this benchmark; they are left for extended multispectral experiments. Although the empirical benchmark is based on the Sentinel-2 L2A 120 m Mosaic product, the operator-based methods only require co-registered multiband time series and externally defined gap masks.

ESA WorldCover 2020 provides semantic land-cover information [19]. The original 10 m map is aggregated to the

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409

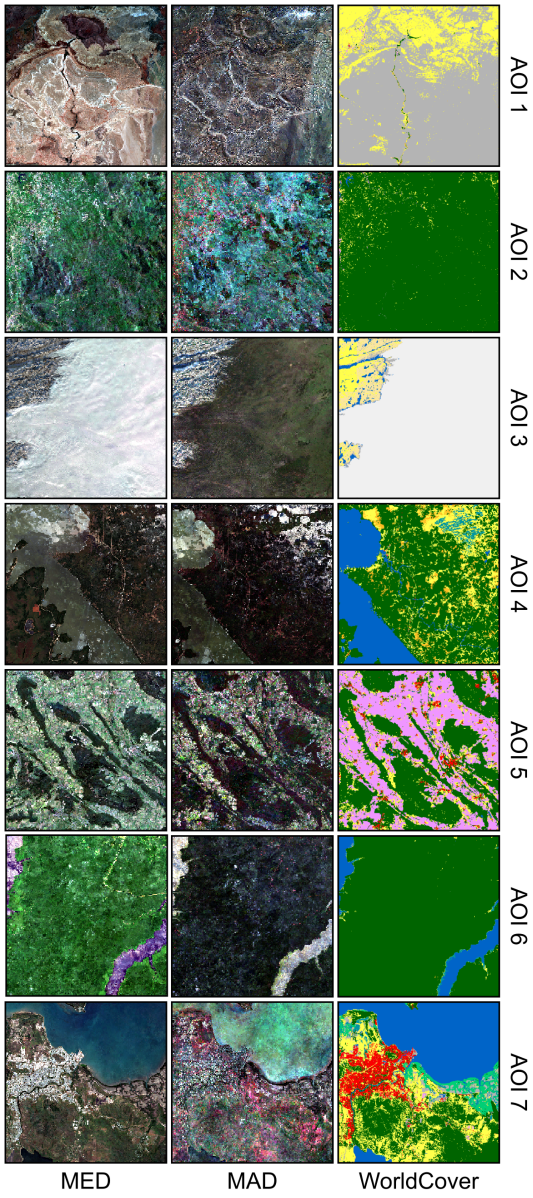


Fig. 2. Overview of the seven AOIs with yearly median reflectance, yearly median absolute deviation, and aggregated ESA WorldCover land-cover information [19].

120 m Sentinel-2 grid by majority vote, i.e., each 120 m cell is assigned the class occupying the largest fraction of its underlying 10 m pixels. The dominant-class share is stored as a semantic purity measure. For semantic reconstruction and land-cover-stratified evaluation, only pixels with a dominant WorldCover class share above 80 % are retained. Pixels below this threshold are treated as semantically ambiguous and are excluded from label-conditioned analyses, which reduces mixed-pixel effects and avoids assigning uncertain class priors. Across the seven AOIs, all eleven ESA WorldCover classes are represented. Fig. 2 illustrates the spatial and semantic composition of the benchmark AOIs.

Spatially heterogeneous cloud- and shadow-induced missingness is first examined for AOI 5 (Germany). The selected range of cloud occurrence is motivated by observations over Germany, where 61 % of 358 Sentinel-2 acquisitions were reported as fully cloud-covered, including shadows, while only 13 % were completely clear-sky and snow-free [66]. Spatially heterogeneous cloud fields are generated using clustered geometries, with target cloud fractions ranging from 10–90 %. Corresponding shadow regions are obtained by spatial displacement of the cloud fields and combined with the cloud masks to define the withheld observations used for reconstruction and evaluation. For partially masked acquisitions, however, unaffected observations from the same date remain available and can provide spatial or spatio-temporal support. Both the amount and spatial configuration of this contemporaneous information vary with cloud fraction and geometry, which makes it difficult to isolate reconstruction from temporal information alone. As discussed in Sec. I, incomplete temporal coverage remains relevant even at high revisit frequencies and may also result from acquisition failures, quality masking, or other causes that render an acquisition unusable. Building on this distinction, the main benchmark adopts complete-date missingness as the basis for a large-scale and systematically controlled evaluation. Entire acquisitions are withheld so that no same-date spatial observations remain available and all methods must reconstruct the missing acquisition from the remaining temporal support. This establishes common support conditions across methods and enables reconstruction behaviour to be assessed systematically across AOIs, years, gap frequencies, and temporal gap arrangements without confounding the comparison by cloud-mask morphology or residual same-date observations. Reduced observation availability and extended periods of missing data are known to increase the difficulty of optical time-series reconstruction and modelling [67]. Complete-date withholding therefore constitutes a deliberately restrictive temporal-support stress test while simultaneously representing a relevant acquisition-level limiting case. The original withheld acquisitions remain available as measured, rather than estimated, reference observations for evaluation. For each time series, 30 scenarios combine six gap frequencies, namely 5, 10, 15, 20, 25, and 30 withheld dates, with five temporal arrangements: (i) gaps at the beginning of the sequence, (ii) gaps at the end, (iii) gaps at both temporal edges, (iv) gaps concentrated in the middle, and (v) approximately equally spaced gaps. The resulting scenario matrix is shown in Fig. 3. This design covers interpolation-like cases, where missing dates are surrounded by temporal support, and extrapolation-like cases, where gaps occur at the beginning or end of the sequence. Withheld reflectance values and statistics derived from them are excluded from reconstruction and used only as reference values for evaluation. High-frequency gap scenarios provide increasingly restrictive stress tests with strongly reduced temporal support. The same complete-date scenario design is applied to all AOIs and both years. The benchmark is therefore stratified by region, climate, year, spectral band, land-cover class, gap frequency, and temporal gap position. This enables reconstruction behaviour to be analysed systematically as a function of both environmental setting and temporal gap geometry.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409



Fig. 3. Artificial gap-scenario matrix applied independently to each AOI and year, shown here for 2020. Rows represent scenarios and columns represent acquisition dates. Green cells indicate retained temporal support observations, whereas grey cells denote acquisition dates withheld for reconstruction. The withheld observations remain available as reference data for validation.

III. Methodology

This section defines the reconstruction methods applied in the benchmark. The proposed HCB and HELIX operators are introduced first, followed by the benchmark configurations used for comparison. The study is formulated as post-mask reconstruction: cloud detection, quality screening, and mask generation are assumed to have been completed before gap filling begins. The reconstruction methods therefore receive a co-registered multiband time series with distinct gaps defined in an external support mask (Fig. 3).

This formulation separates the reconstruction problem from the spatial morphology and origin of the mask. Missing observations may result from spatially heterogeneous cloud and shadow masks, isolated invalid pixels, quality filtering, acquisition gaps, or complete missing dates. HCB and HELIX do not require cloud-object geometry, connected gap regions, or mask-boundary information. The complete-date gaps used in the benchmark are therefore a controlled temporal-support stress test, not a methodological restriction. For HELIX, semantic information enters separately through the land-cover label of each pixel; it constrains plausible seasonal reflectance behaviour but does not define the spatial shape of the gap.

A. HCB Approach

The HCB approach assumes that the ordered temporal reflectance sequence contains lower-order spectro-temporal rank than represented by the orthogonal basis. The method does not require strict periodicity; instead, it uses the basis representation to separate coarse temporal structure from higher-frequency deviations under incomplete temporal support. The novelty of this work lies in a missingness-aware coefficient estimation, in which positive and negative basis contributions are treated separately and averaged over the available temporal support. This sign-separated, support-conditioned averaging enables a consistent and most stable reconstruction even in the presence of incomplete observations.

The key finding is a rearrangement of the Hadamard matrix according to the respective scale. Starting with the formulation of $H^{\{2\}}$ (complex basis) the Hadamard matrices with dimensions of powers of two can be calculated by recursive Kronecker multiplication [68], [69].

$$H^{\{2\}} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$
$$H^{\{n\}} = H^{\{2p\}} = H^{\{2\}} \otimes H^{\{p\}} \tag{1}$$

Some others, e.g. $H^{\{12\}}$, do also exist. Hadamard transformation matrices share the following characteristics:

1) They are constructed from sign matrices, that only consist of $+1$ and -1. Hence, all entries adopt the same absolute value to avoid any imbalance.
2) They are orthogonal, which means that they merely rotate and reflect the feature space without distorting it through non-uniform scaling or shearing.
3) The factor $\frac{1}{\sqrt{n}}$ turns the orthogonal matrix into an orthonormal matrix, i.e., the absolute value of the determinant is always 1 which means that even uniform scaling is prevented.
4) Being orthogonal the transposed matrix is equal to the inverse matrix $H^{-1} = H^T$, which means that the inverse matrix always exists on the one hand, and on the other hand, that numerical errors in the calculation of the inverse matrix can be ruled out.
5) In the case of powers of two, the Hadamard matrices are even self-inverse, i.e., they are symmetric, so that

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409

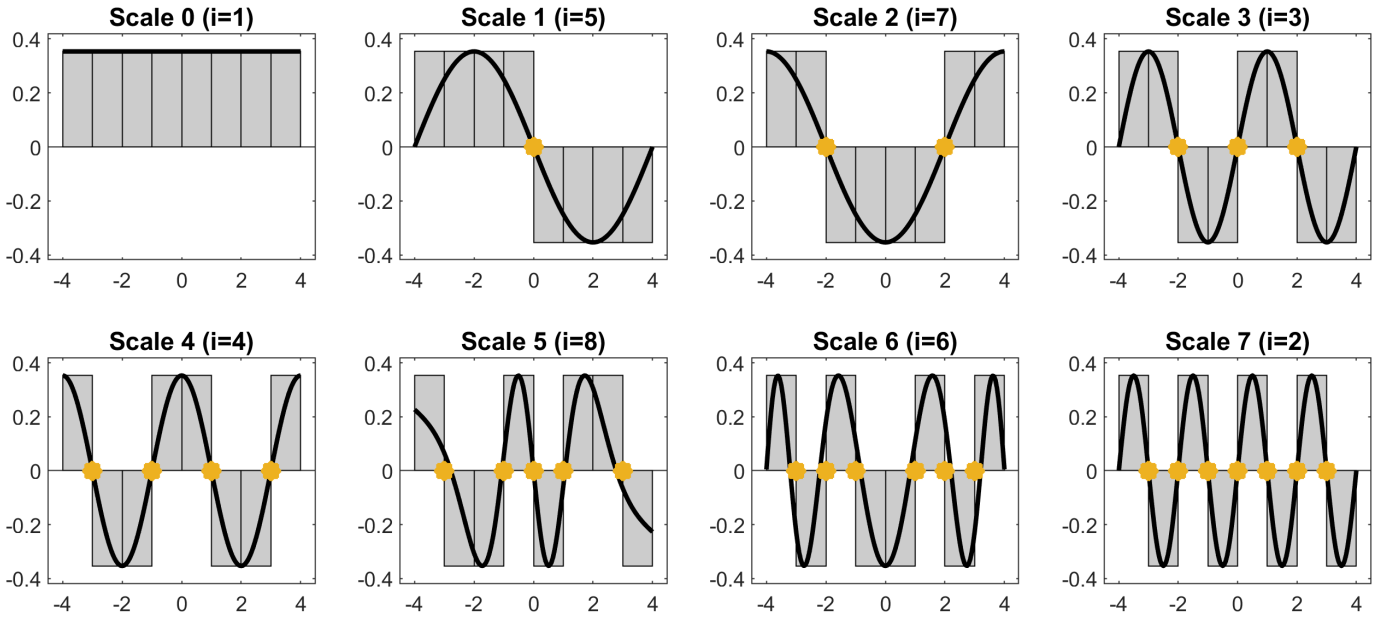


Fig. 4. Octonion base elements of the hypercomplex bases and their associated continuous waves. The number of zero crossings corresponds to the scale in Eq. 2: $i = 1$ represents the coarsest mean component, $i = 5$ the temporal slope, $i = 7$ the temporal curvature, and $i = 2$ the finest mean change between neighbouring dates.

$H = H^T = H^{-1}$ and therewith, the identical matrix can be used for forward and backward transformation in favour of maximum stability.

Each row i of the above defined matrix (Eq. 1) corresponds to a certain wavelength indicated by the number of zero crossings which is exemplified in the following for the Hadamard matrix $H^{\{8\}}$ with eight dimensions, i.e. also eight elements. A visualisation of the individual elements and their respective waves can be found in Fig. 4.

$$H_{i,t}^{\{8\}} = \frac{1}{\sqrt{8}} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{bmatrix} \quad (2)$$

$$\left\langle \left| \overrightarrow{\Delta} H_{i,t}^{\{8\}} \right| \right\rangle_t = \frac{1}{\sqrt{8}} \begin{bmatrix} 0 & 7 & 3 & 4 & 1 & 6 & 2 & 5 \end{bmatrix}^T$$

The $H_{i,t}^{\{8\}}$ matrix can now be rearranged in order to represent the spectral-temporal scales in ascending frequency order according to Eq. 2 and Fig. 4. The row index i of $H_{i,t}^{\{8\}}$ is switched to s to express the scale dependency of $H_{s,t}^{*\{8\}}$. The number of zero crossings now obviously increases monotonically.

$$H_{s,t}^{*\{8\}} = \frac{1}{\sqrt{8}} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \end{bmatrix} \quad (3)$$

$$\left\langle \left| \overrightarrow{\Delta} H_{s,t}^{*\{8\}} \right| \right\rangle_t = \frac{1}{\sqrt{8}} \begin{bmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{bmatrix}^T$$

In the case of no gaps, the transform unfolds to a simple multiplication of the reflectance vector R with the corresponding Hadamard matrix H^* resulting in the Kennaugh element vector K [70]. The rank of the matrix $\{8\}$ will no longer be specified in what follows, because the statements are generally valid. The sum over all columns, i.e. acquisition dates is denoted by angle brackets $\langle\ \rangle_t$.

$$K = H^* \cdot R \ \rightarrow \ K_s = \left\langle H_{s,t}^* \cdot R_t \right\rangle_t \quad (4)$$

Each Kennaugh element composes of an equal number of positive contributions ($t+ \in H_{s,t} > 0$) and negative contributions ($t- \in H_{s,t} < 0$), except for the first element that represents the total intensity and thus, only contains positive entries. If not all R_t are available due to gaps, the respective sums can be replaced by the expectation value over all available measurements t_m. For the total intensity in K_0 this simply unfolds to:

$$K_0 = \frac{\langle R_t \rangle_t}{n} = \sqrt{n} \cdot \frac{\langle R_t \rangle_t}{\langle t \rangle_t} = \sqrt{n} \cdot \frac{\langle R_t \rangle_{t_m}}{\langle t \rangle_{t_m}} \quad (5)$$

$$K_0 = \sqrt{n} \cdot \mathbb{E}\left(R_{t_m}\right) \qquad t_m \in \text{measured}$$

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409

For all other K_i the sum can be decomposed into the difference of the expectation values of the positive contributions $t+$ and negative contributions $t-$ as derived in Eq. 6 . This sign-separated averaging constitutes the central extension from the HCB decomposition to the HCB gap filling that has never been shown before.

$$\begin{aligned} K_s &= \langle H_{s,t} \cdot R_t \rangle_{t+} + \langle H_{s,t} \cdot R_t \rangle_{t-} \qquad s \in [1, n-1] \\ &= \frac{1}{\sqrt{n}} \left(\langle R_t \rangle_{t+} - \langle R_t \rangle_{t-} \right) \\ &= \frac{\sqrt{n}}{2} \left(\frac{\langle R_t \rangle_{t+}}{\langle t \rangle_{t+}} - \frac{\langle R_t \rangle_{t-}}{\langle t \rangle_{t-}} \right) \\ &= \frac{\sqrt{n}}{2} \left(\mathbb{E}\left(R_{t_+}\right) - \mathbb{E}\left(R_{t_-}\right) \right) \end{aligned} \tag{6}$$

With the above formulation in Eq. 6 we gained two arithmetic means by simple algebraic manipulation. Mean values link the purely algebraic formulation to statistics: the arithmetic mean can be interpreted as expectation values that can be derived from one possibly reduced statistical population, i.e., also in the presence of gaps in the time series, which would not be possible in the original transform. The statistically estimated K_s thus unfold to:

$$\begin{aligned} K_s &= \frac{\sqrt{n}}{2} \left(\mathbb{E}\left(R_{t_m+}\right) - \mathbb{E}\left(R_{t_m-}\right) \right) \quad s \in [1, n-1] \\ &= \frac{\sqrt{n}}{2} \left(\frac{\langle R_{t_m} \rangle_{t+}}{\langle t_m \rangle_{t+}} - \frac{\langle R_{t_m} \rangle_{t-}}{\langle t_m \rangle_{t-}} \right) \quad t_m \in \text{measured} \end{aligned} \tag{7}$$

It is important to mention that the balance between positive and negative measurements in Eq.7 has to be assured, i.e., positive and negative entries must not be mixed up. Doing so, n Kennaugh elements are calculated from m reflectance measurements with $m < n$. A simple math example evaluating Eqs 5 and 7 is provided in the Appendix A. Information multiplication being not possible, the number of meaningful Kennaugh elements must be reduced to an optimum number s_{opt} according to the available measurements and — most important — to the described oscillation in the multi-scale representation with $s \in [1, n]$ using the following cost function that is proportional to the mean absolute deviation (MAE).

$$s_{\text{opt}} = \arg\min_s \left\langle \left| R_t - \langle H_{t,s} \cdot K_s \rangle_s \right| \right\rangle_{t_m} \tag{8}$$

As the reconstruction even proposes enhanced values for the available reflectances R_a, the mean absolute error over known measurements serves as a robust multi-temporal cost function to be minimised in a multi-scale approach starting from the mean value and progressively adding increasing frequencies until the optimal scale s_{opt} is reached. The scale adopts $s_{\text{opt}} = 1$, if only one measurement is known and reaches $s_{\text{opt}} = n$, if all reflectance values are measured $m = n$. In the usual case of $1 < m < n$, the algorithm intrinsically selects the optimal scale according to the available data. For better comprehension, Appendix A derives the multi-scale reconstruction and visualises the optimal scale for nine selected signals.

$$R_t^* = \left\langle H_{t,s}^* \cdot K_s \right\rangle_{s \in [1, s_{\text{opt}}]} \tag{9}$$

This methodology can be seen as self-optimizing technique, because no (hyper-)parameter tuning is necessary. The algorithm autonomously searches for the optimal representation on hypercomplex bases according to the above defined cost function. It further introduces an unprecedented flexibility in the application of Hadamard transforms: if the number of measurements in a time series does not match the dimensions for which Hadamard matrices are known, the next larger Hadamard can be selected and the unnecessary samples can simply be treated as gaps. As the self-optimization only fits to the known measurements, these artificial gaps have no influence at all.

The resulting full reflectance vector R^* thus incorporates n elements, but only s_{opt} wavelengths. The minimum $\text{MAE}_{s_{\text{opt}}}$ additionally provides a goodness of fit measure for the time series reconstruction. Now, it is up to the user whether the measured values are replaced by reconstructed and thus optimised values in the sense of a self-correcting process or kept unchanged as fault-free measurements. Regarding the subsequent evaluation, this is negligible as only the reconstructed originally missing values are taken into account.

The support mask with gaps only defines which temporal entries of a pixel- and band-specific reflectance vector are seen as missing for the coefficient estimation in Eq. 7 and the cost function optimization in Eq. 8. No term in the HCB reconstruction uses cloud-object geometry, spatial connected components, or gap-boundary information. A complete-date gap and a spatially heterogeneous cloud mask therefore enter the operator in the same way: as different support sets over the ordered reflectance sequence.

B. HELIX Approach

The HELIX approach uses land-cover information as a deterministic semantic prior for post-mask reflectance reconstruction. Its use of class-wise support is conceptually related to the non-local means principle, in which information is aggregated according to similarity rather than spatial proximity alone [71]. HELIX transfers this idea to the semantic–temporal domain: instead of relying only on neighbouring reflectances, it decomposes the local signal into a class-specific seasonal expectation derived from semantically related observations and a local residual around this expectation [72]. The land-cover class therefore constrains the seasonal reference, while the measured temporal support preserves local deviations from the class prototype.

In the present benchmark, a semantic prior is given by the ESA WorldCover class [19] denoted as subscript c in the following. It enters the operator through class-wise seasonal prototypes and plausibility ranges. The method is deterministic for fixed inputs, labels, gap configuration, and parameter settings, and does not require external training data, learned weights, or fine-tuning.

The processing principle is illustrated in Fig. 5. Measured observations define the local residual history, the land-cover label defines the seasonal prototype and plausibility rails, and only target values are reconstructed. The operator is applied independently to each spectral band of the multiband time

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited, content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409

series. The equations below describe the operation for one spectral band at one location assigned to land-cover class c. Let R_t denote the local reflectance of that band at temporal position t, with t expressed on a common 365-day annual phase. The notation distinguishes the target temporal position t from measured support positions t_m.

For each gap scenario, observations designated as missing are excluded from all HELIX parameter estimation and reconstruction support, including class-wise prototype fitting and residual and spread statistics. Their reflectance values are used only after reconstruction as reference values for evaluation.

For each land-cover class c, HELIX derives a smooth seasonal prototype $\mu_{t,c}$ for the currently processed spectral band. The prototype is fitted to the class-wise median reflectance observed at each temporal position, using only valid measured observations. Following the common use of harmonic functions to represent smooth annual behaviour in optical EO time series [20], [21], [23], the prototype is represented by a compact Fourier model. The Fourier order F specifies the number of harmonic components included. In the primary HELIX configuration, $F = 2$, yielding the following second-order model:

$$\begin{aligned}\mu_{t,c} = \beta_{0,c} &+ \alpha_{1,c}\cos\left(\frac{2\pi t}{365}\right) + \gamma_{1,c}\sin\left(\frac{2\pi t}{365}\right) \\ &+ \alpha_{2,c}\cos\left(\frac{4\pi t}{365}\right) + \gamma_{2,c}\sin\left(\frac{4\pi t}{365}\right).\end{aligned} \tag{10}$$

The coefficients $\beta_{0,c}$, $\alpha_{1,c}$, $\gamma_{1,c}$, $\alpha_{2,c}$, and $\gamma_{2,c}$ are obtained by least-squares fitting of Eq. 10 to $y_{t,c}$. The coefficient $\beta_{0,c}$ defines the mean reflectance level of the seasonal curve. The coefficients $\alpha_{1,c}$ and $\gamma_{1,c}$ describe the annual component, whereas $\alpha_{2,c}$ and $\gamma_{2,c}$ describe the semi-annual component. The $F = 2$ configuration therefore contains one annual and one semi-annual oscillation and defines one smooth seasonal prototype for each land-cover class and spectral band.

Each measured observation is decomposed into the class-level expectation and a local residual:

$$r_{t_m,c} = R_{t_m,c} - \mu_{t_m,c} \tag{11}$$

The residual $r_{t_m,c}$ captures how the local reflectance trajectory deviates from the fitted semantic prototype at measured temporal positions. Missing values are therefore not filled by class averages alone. Instead, reconstruction is based on locally fused residual information that is re-anchored to the class-specific seasonal expectation.

After fitting the seasonal prototype, a class-specific tolerance around this curve is required. The Fourier model represents the smooth central seasonal expectation of class c but the observed class-level signal is not perfectly harmonic. It can deviate from the fitted curve because of within-class variability, phenological phase differences, residual atmospheric effects or even mixed pixels. It therefore estimated a robust spread from this mismatch between the observes class-wise median curve and the fitted prototype. Let $y_{t_m,c}$ observed class wise median reflectance over the valid measurement locations assigned to a class c at measured temporal position t_m. The robust spread is then calculated as

$$\sigma_c = \max\left(1,\ 1.4826 \cdot \text{median}\,|y_{t_m,c} - \mu_{t_m,c}|\right) \tag{12}$$

The median term measures the typical absolute deviation of the observed class-level seasonal signal from its smooth harmonic approximation. The factor 1.4826 is the normal-consistency factor that converts a median absolute deviation into a standard-deviation-equivalent robust spread [73]. The lower bound prevents degenerate zero-width intervals.

This spread defines class-specific plausibility rails:

$$L_{t,c} = \mu_{t,c} - 2\sigma_c \qquad U_{t,c} = \mu_{t,c} + 2\sigma_c \tag{13}$$

The lower rail $L_{t,c}$ and upper rail $U_{t,c}$ provide a robust semantic reference range for plausible reflectance values of class c at temporal position t. For a target position t, residual support is taken from measured neighbouring observations, from within a temporal search window W:

$$|t_m - t| \le W. \tag{14}$$

Temporal proximity within this support is encoded by

$$w_m = \exp\left(-\frac{|t_m - t|}{\tau}\right). \tag{15}$$

The temporal weighting scale is set to

$$\tau = \max\left(20,\ \frac{W}{2}\right)\ \text{days}. \tag{16}$$

Nearby observations therefore receive larger weights than distant observations. The lower bound of 20 days avoids overly steep weighting for short windows, whereas $\frac{W}{2}$ adapts the weighting to wider temporal searches. The target residual is estimated as the weighted median of the measured residuals, that satisfy Eq. 14:

$$\hat{r}_{t,c} = \text{wmed}\left(r_{t_m,c};\, w_m\right). \tag{17}$$

The weighted median is taken over all measured temporal positions t_m within the search window W. This reduces the sensitivity to outlying support dates while keeping the reconstruction local to the measured temporal trajectory. The preliminary reconstruction is obtained by re-anchoring the fused residuals to the class-specific prototype:

$$Q_{t,c} = \mu_{t,c} + \hat{r}_{t,c}. \tag{18}$$

If Q_t lies within the plausibility interval $[L_{t,c}, U_{t,c}]$, it is used unchanged. Values below the lower rail are moved by $\lambda = 0.25$ toward $L_{t,c}$, and values above the upper rail are moved by the same factor toward $U_{t,c}$. This soft adjustment avoids hard clipping while reducing implausible deviations from the class-specific seasonal range. The final reconstructed value is denoted by R_t^*.

If the residual support set in Eq. 14 is empty, the temporal search window is expanded up to $W_{\max} = 120$ days. Remaining unsupported pixels are handled by a temporal nearest-neighbour fallback, which assigns the closest valid measured observation in time. Locations outside the valid spatial support are kept unchanged and are not reconstructed.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409

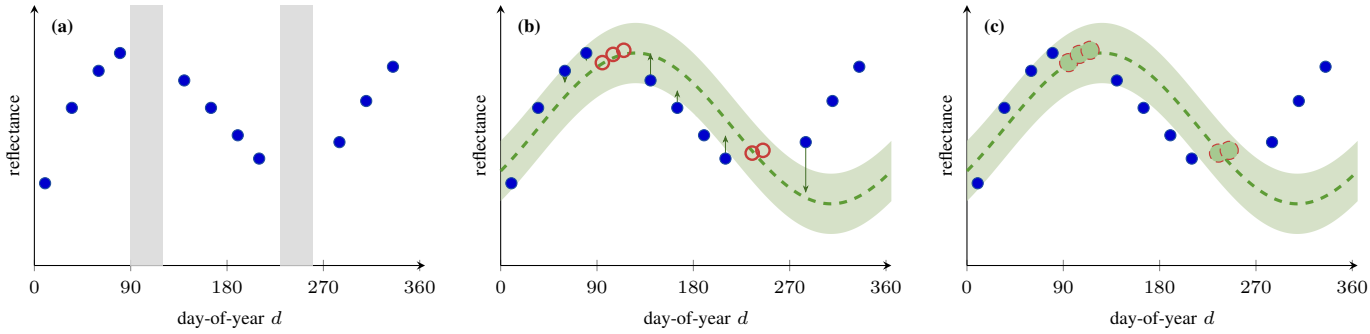


Fig. 5. Conceptual overview of the HELIX label-side reconstruction strategy for one spectral band and one land-cover class. (a) Observed EO time series with withheld target dates indicated by gray bars. (b) A class-specific seasonal prototype $\mu_c(d)$ is derived from retained observations and combined with plausibility rails; red open circles mark target dates. (c) Final reconstruction after re-anchoring fused residual information to the class-specific prototype: retained observations remain unchanged, and only target dates are filled.

Land-cover information enters the operator through the class-specific prototype $\mu_{t,c}$ and the plausibility interval $[L_{t,c}, U_{t,c}]$. The filled value is therefore not a class average. It remains anchored in the local residual history whenever temporal support is available, while the class label controls the seasonal reference and the range of plausible values [74]. A complete-date gap and a spatially heterogeneous cloud mask therefore enter the operator in the same way: as different temporal support sets over the local reflectance sequence. The main limitations follow directly from this design: sparse temporal support, mixed or outdated land-cover labels, and surface changes not represented by the categorical layer can reduce the reliability of the semantic constraint.

C. Benchmark Approaches

Linear interpolation is used as the deterministic temporal baseline. It reconstructs missing dates from neighbouring valid observations and fills edge gaps by extending the nearest available temporal gradient. The method is included because it is transparent, computationally negligible, and fully reproducible.

Harmonic regression is included as a stronger deterministic seasonal baseline. A fixed second-order annual Fourier model ($F = 2$) is estimated independently for each pixel and spectral band from the valid observations of each gap scenario, using acquisition day of year and a 365-day period. Withheld reflectance values are excluded from parameter estimation and used only for evaluation. The method operates purely in the temporal domain and uses neither land-cover information nor spatial context. Its inclusion provides a direct reference for assessing whether improvements over linear interpolation arise from seasonal modelling alone or from the additional reconstruction mechanisms introduced by HELIX.

The learned reference approaches cover generic temporal models, a pre-trained time-series foundation model, and satellite-specific optical reconstruction methods. **SAITS** [62] and **TimesNet** [63] represent generic multivariate time-series architectures. **MOMENT** [64] is included as a pre-trained time-series foundation model to assess the transfer of generic temporal representations to masked EO reconstruction. **U-TILISE** [57] and **UnCRtainTS** [58] represent satellite-specific optical image time-series reconstruction approaches. Their adaptation regimes are summarised in Table II; the complete implementation protocol is provided in Table VIII in the Appendix.

TABLE II
BENCHMARK ROLE AND ADAPTATION REGIME OF THE REFERENCE RECONSTRUCTION APPROACHES.

Approach	Benchmark role	Adaptation in this study
Linear	Deterministic temporal baseline	Fixed interpolation; no training or target-dependent parameter estimation.
Harmonic	Deterministic seasonal baseline	Second-order annual Fourier model fitted separately for each pixel, band, and scenario from the valid target-series observations; no spatial or semantic information.
SAITS	Generic multivariate imputation	Trained from scratch separately for each AOI–year–scenario on target-area pixel sequences.
TimesNet	Generic deep time-series modelling	Trained from scratch separately for each AOI–year–scenario on target-area pixel sequences.
MOMENT	Pre-trained time-series foundation model	Released MOMENT-1-base weights; no target-area fine-tuning; band-wise reconstruction.
U-TILISE	Satellite-specific optical reconstruction	Released EarthNet2021 checkpoint; no target-area fine-tuning [65].
UnCRtainTS	Satellite-specific optical reconstruction	Released optical no-SAR checkpoint; no target-area fine-tuning.

SAITS and TimesNet receive pixel-wise sequences with 32 time steps and four spectral bands and are trained separately for each AOI–year–scenario. The dates withheld by the respective scenario are set to missing before model fitting and their reflectance values are therefore unavailable during training. SAITS samples up to 200,000 eligible pixel sequences per scenario, limited by the number of available pixels, whereas TimesNet uses 60,000 sampled sequences. No separate benchmark validation subset is supplied to the fitting routine; evaluation is performed exclusively on the withheld acquisition dates. Training and evaluation can therefore share spatial pixel locations, but the withheld reflectance values themselves do not enter model fitting.

MOMENT is evaluated from the released MOMENT-1-base weights without fine-tuning and is applied band-wise to uni-

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited, content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409

variate temporal sequences using the common scenario mask. U-TILISE and UnCRtainTS are likewise evaluated from released pre-trained checkpoints without target-area fine-tuning. U-TILISE uses the released four-band EarthNet2021 configuration [57], [65], whereas UnCRtainTS uses the released optical no-SAR configuration adapted to the four benchmark bands. Model specific scaling, masking, temporal handling, checkpoint selection, and hyperparameters are reported in Table VIII.

All reconstructors operate on the same 32-date acquisition grid, use the identical scenario-specific missing-date masks, and are evaluated on the same spatial samples. The $> 80\,\%$ dominant-WorldCover purity criterion is applied once, and the resulting set of retained pixels is used for all compared methods. Across the 30 controlled scenarios, withheld dates are systematically distributed over the temporal sequence, providing the same gap positions and available temporal support to every method. No method receives additional acquisitions or access to withheld reflectance values; these values are excluded from training, parameter estimation, and reconstruction support and are used only for evaluation. Method-specific input transformations are reversed before scoring so that all predictions are compared on the common Sentinel-2 reflectance scale. Retained observations are restored before evaluation, and accuracy is computed exclusively at withheld positions.

The approaches nevertheless differ in how the same available target-series support is exploited. HELIX performs analytical target-series parameter estimation from the valid observations remaining under the respective scenario mask. SAITS and TimesNet are trained from scratch separately for each AOI–year–scenario using target-area pixel sequences after application of the same mask. MOMENT, U-TILISE, and UnCRtainTS are applied from released pre-trained weights without benchmark-specific fine-tuning. Thus, the benchmark covers analytical target-series estimation, target-specific learned fitting, and inference-only application of pre-trained models. Because the temporal grid, gap masks, and distribution of withheld dates are identical across methods, these adaptation regimes do not provide different access to temporal observations; they differ in how the available reconstruction support is processed. The comparison therefore evaluates complete reconstruction approaches under their respective operating configurations rather than isolating algorithmic design from adaptation strategy. The corresponding training, adaptation, masking, and implementation protocols are detailed in Table VIII in the Appendix.

Runtime measurements quantify the computational effort required to execute each evaluated workflow under the respective experimental configuration. Because the approaches differ fundamentally in their training requirements and were not executed under a strictly hardware-normalised timing protocol, the reported runtimes are interpreted as implementation-specific computational costs rather than as a controlled hardware-level speed ranking. For pretrained models, the reported inference times exclude the upstream computational cost of model training, whereas methods trained within the benchmark include the corresponding training effort. Training and inference times, together with the execution devices and configurations, are reported in Table VIII.

IV. Results

The results address the two experimental settings introduced in Sec. II. Across both settings, Spectral Angle Mapper (SAM) is used to quantify spectral-shape fidelity through the angular difference between reconstructed and reference spectral vectors. The spatially heterogeneous cloud/shadow experiment is evaluated in Sec. IV-A from three additional and complementary perspectives: radiometric change, spatial-gradient change, and temporal-gradient change. Radiometric change quantifies deviations in the reconstructed spectral values themselves, whereas spatial-gradient change assesses whether local spatial transitions and image structure are preserved. Temporal gradient change analogously quantifies deviations in the temporal transitions between consecutive acquisitions. These structural diagnostics are particularly informative for spatially heterogeneous gaps, because a reconstruction may remain close to the local or temporal mean while suppressing spatial contrast, dampening temporal dynamics, or introducing locally inconsistent structures. Such behaviour can therefore appear comparatively moderate under point-wise radiometric errors while still representing a substantial loss or alteration of the information contained in the original observation.

The complete-date benchmark instead addresses a broader reconstruction question: how reliably can an entire missing acquisition be recovered when no same-date spatial support remains and the available temporal support is systematically varied? Its purpose is therefore not primarily to diagnose individual spatial artifacts, but to compare reconstruction accuracy and stability across the full experimental design comprising AOIs, years, spectral bands, land-cover classes, gap frequencies, and temporal gap positions. For this large-scale cross-scenario comparison, radiometric performance is characterised using mean error (ME), mean absolute error (MAE), median absolute error (MedAE), and root mean square error (RMSE), complemented by SAM as the common measure of spectral fidelity. Together, these metrics provide consistently aggregatable measures of systematic radiometric bias, typical and average absolute error, sensitivity to large reconstruction failures, and spectral-shape preservation. ME captures the direction and magnitude of systematic radiometric bias. MedAE represents the central robust absolute error, whereas MAE captures the average absolute error magnitude and is more sensitive to broadly distributed deviations. RMSE places additional weight on large errors through quadratic penalisation and therefore emphasises severe reconstruction failures.

Before considering the individual result dimensions, Fig. 6 provides an exploratory post-hoc overview of multimetric reconstruction behaviour across the complete-date benchmark using Uniform Manifold Approximation and Projection (UMAP) [75]. Such metric embeddings can support model behaviour analysis by organising complementary performance descriptors into local regimes that may not be apparent from aggregate scores alone [76]. Each point represents one

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409

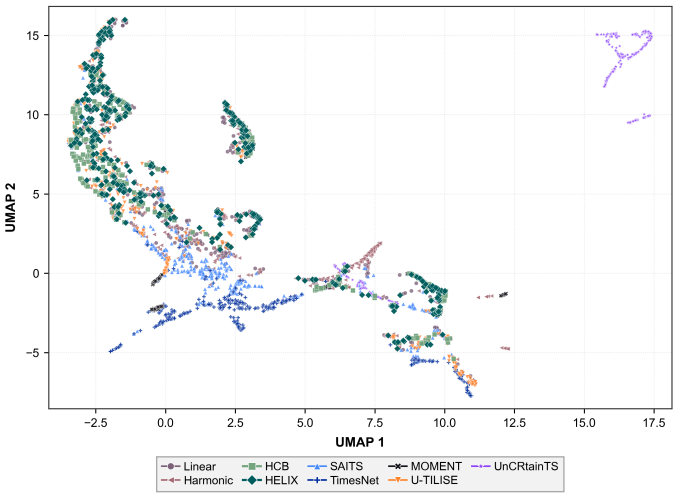


Fig. 6. Exploratory post-hoc UMAP embedding of multimetric reconstruction behaviour. Each point represents one method–AOI–year–scenario profile described by MedAE, MAE, ME, $Q90_{\mathrm{AE}}$, $Q95_{\mathrm{AE}}$, and SAM, with radiometric descriptors aggregated over spectral bands. Neighbourhood structure rather than axis magnitude or global distance is interpreted. Extreme MOMENT points are omitted from the zoomed display only; their underlying profiles remain part of the embedding.

method–AOI–year–scenario profile, with the radiometric descriptors aggregated over the four spectral bands. The descriptor vector combines MAE, MedAE, ME, and SAM with two upper-tail descriptors, $Q90_{\mathrm{AE}}$ and $Q95_{\mathrm{AE}}$, corresponding to the 90th and 95th percentiles of the absolute-error distribution. The two upper-tail descriptors capture comparatively large reconstruction errors that are not represented by central metrics alone.

All non-negative descriptors are $\log_{10}$-transformed before projection to reduce the influence of strongly skewed error distributions. ME is retained on its signed scale because its sign carries information on under- and overestimation. All descriptors are subsequently standardised to zero mean and unit variance. UMAP is then applied using Euclidean distance, $n_{\mathrm{neighbors}} = 30$, $\mathrm{min\_dist} = 0.10$, and a fixed random state of 42. The resulting axes are unitless and have no direct physical interpretation; the embedding is therefore interpreted through local neighbourhood structure rather than axis magnitude or global distance.

The projection reveals method-dependent structures superimposed on several shared performance regimes. HCB and HELIX frequently co-locate with Linear and U-TILISE across multiple neighbourhoods, indicating similar multimetric behaviour under subsets of the evaluated conditions. SAITS partly overlaps these regions but also extends into distinct portions of the embedding. Harmonic, TimesNet, and MOMENT show more method-specific branches, whereas UnCRtainTS exhibits the clearest separation, including several isolated regions. These patterns reflect different combinations of central error, upper-tail behaviour, radiometric bias, and spectral distortion. The embedding is used only as an exploratory summary; quantitative interpretation is based on the metric-specific analyses in the following subsections.

A. Real World Cloud Scenario

The real world cloud scenario reveals most interesting insights. Amongst the selected acquisitions in Fig. 7, the deterministic approaches Linear interpolation and Harmonic regression are able fill the cloud gaps adequately, sometimes with a slight radiometric distortion, e.g. Harmonic of 02 Aug and 23 Sep. HCB and HELIX also guarantee a complete reconstruction with similar, almost identical results. Only the HELIX reconstruction of 02 Aug characterises a lower sharpness in comparison to the reference. SAITS completely fails. Not even the values that are known in themselves are reproduced correctly. The entirely missing dates 02 Aug and 22 Sep are reconstructed, but with a greenish tint (02 Aug) and a yellowish tint (22 Sep). TimesNet leaves these two acquisitions blank. MOMENT only reproduces the edges. U-TILISE works quite well from April to June. However, the filling of thicker clouds is connected with a loss of sharpness, see 12 Jun and 02 Jul. In 02 August, a completely missing scene, the largest part of the image is adequately reconstructed with a loss of brightness at the upper and lower edges. The only thing that really stands out is a cloud-like structure in the lower half filled with a regular gray pattern. The pattern is intensified and extended on 22 Sep, where it covers almost the whole image. The upper and lower boundaries show spectral distortions. Only some parts (around the cloud of 02 Jul) are more or less correctly reproduced. UnCRtainTS fills the clouds with uniform (12 Apr and 02 May) and graduated (12 Jun and 02 Jul) pink, inserted square tiles or leaves it blank (02 Aug and 21 Sep). In a nutshell, the results of the deep learning approaches are highly variable but rarely reliable. This is also evident in the numerical evaluation.

The deterministic methods provide a reliable reconstruction both in radiometry and spectrometry indicated by zero radiometric change (in median) and very low spectral angle deviations of of maximum of five degrees (again in median), see Table III. The respective mean values are higher because of some strong outliers. The learned approaches have significantly higher deviations, except for U-TILISE. As it is known from the visual inspection of Fig.7 that U-TILISE lowers the sharpness or fill with artificial patterns, the spatial and temporal sharpness is additionally addressed. The spatial gradient change (gradient in the reconstruction vs. gradient in the reference) is zero (in median) for the deterministic methods and at least -50 for the learned models, which refers to a loss of sharpness. Only Linear interpolation and Harmonic regression show positive mean values due to some stronger gradients in extrapolated parts of the image. The same applies to the temporal gradient change, although the temporal gradient is in general lower than the spatial gradient. Linear interpolation and UnCRtainTS reach positive mean values meaning a amplification of the temporal gradient. To sum up, in this real world cloud scenario, the learned methods do not provide reliable results. In order to explore the behaviour in detail, a more complex scenario with a varying number of gap and varying temporal distribution is adopted in the following section.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409

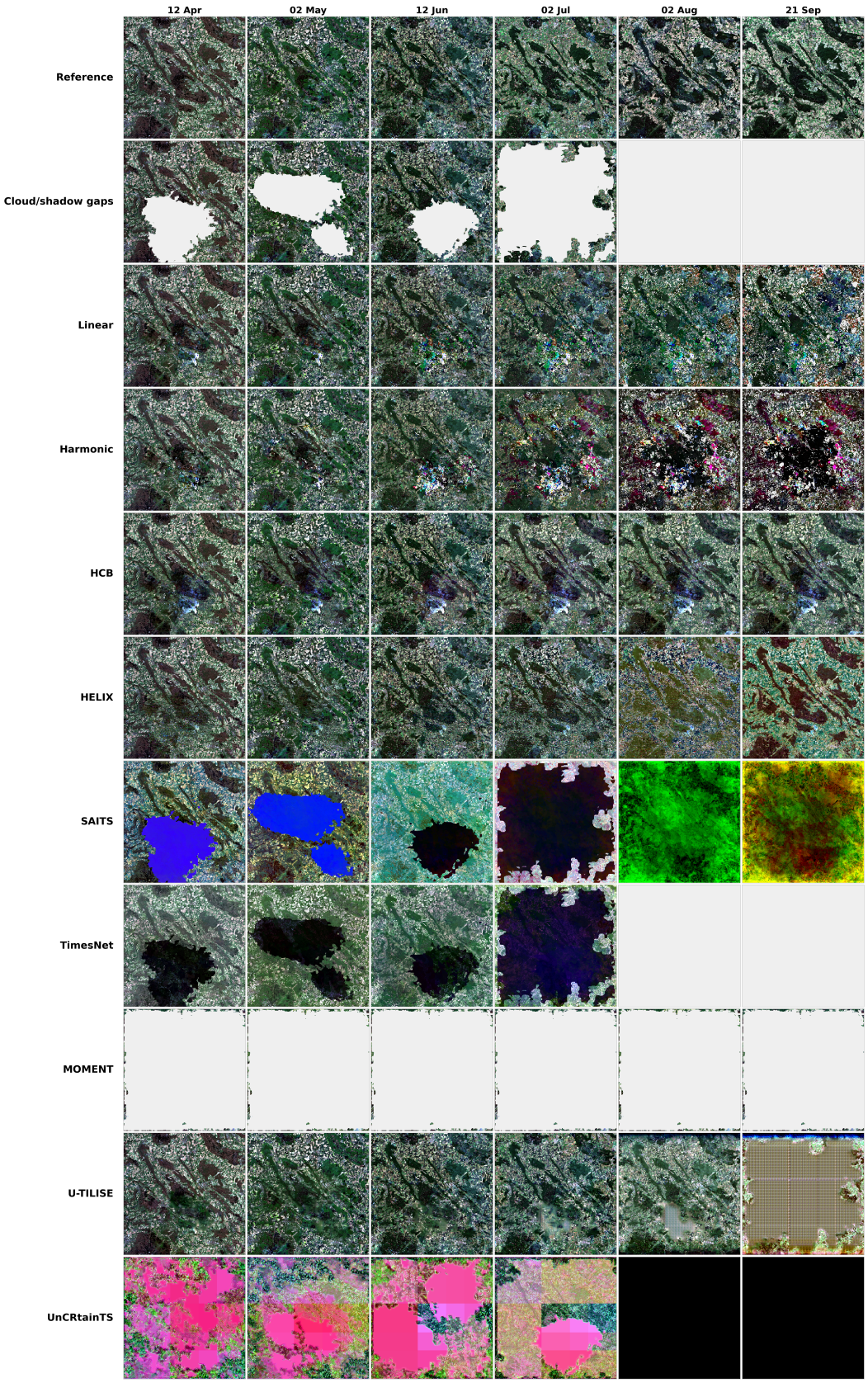


Fig. 7. Qualitative RGB comparison for selected acquisitions of the clustered cloud/shadow experiment in AOI 5. The first two rows show the original reference observations and the corresponding masked inputs, followed by the reconstructions of the nine evaluated methods. For visualization only, each panel is independently contrast-stretched using the 2nd–98th percentiles of its valid RGB values (B04/B03/B02); all quantitative analyses are performed on the original unstretched DN data.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409

TABLE III
NUMERICAL EVALUATION OF THE REAL WORLD CLOUD SCENARIO DEPICTED IN FIG. 7: THE MEAN AND MEDIAN RADIOMETRIC BIAS, THE ABSOLUTE SPATIAL AND THE ABSOLUTE TEMPORAL GRADIENT ARE REPORTED. THE HALLUCINATIONS AND VOIDS OBSERVED IN THE RESPECTIVE FIGURE SHOW UP IN MAJOR DEVIATIONS BETWEEN THE RECONSTRUCTED IMAGES AND THE REFERENCE.

Method ↓	*Mean*	*Median*	*Mean*	*Median*	*Mean*	*Median*	*Mean*	*Median*
Linear	120.8	0	10.7	3	366.9	0	5.5	0
Harmonic	235.3	0	18.8	5	825.6	0	-8.3	0
Proposed HCB	**-72.5**	**0**	**7.9**	**5**	**-156.1**	**0**	**-18.3**	**0**
Proposed HELIX	**-168.4**	**0**	**8.2**	**4**	**-151.1**	**0**	**-29.6**	**0**
SAITS	-237.4	-166	23.3	17	-564.8	-292	-40.9	-13
TimesNet	-715.1	-781	39.3	38	-438.8	-173	-30.2	-11
MOMENT	-1191.3	-1180	10.3	7	-567.5	-301	-23.8	-6
U-TILISE	-150.0	-2	7.1	3	-291.4	-50	-25.1	-1
UnCRtainTS	2815.0	2121	43.5	44	-360.3	-183	72.3	-1
Statistics and Unit →	Radiometric Change [DN]		Spectral Change [°]		Spatial Gradient Change [DN]		Temporal Gradient Change [DN]	

TABLE IV
OVERALL CONTROLLED-GAP RECONSTRUCTION ACCURACY AND RECORDED PROCESSING TIME. BIAS, MAE, MEDAE, RMSE, AND SAM ARE SUMMARISED OVER ALL WITHHELD OBSERVATIONS. RUNTIME REFLECTS THE RESPECTIVE BENCHMARK CONFIGURATION AND IS NOT HARDWARE-NORMALISED; FOR SAITS AND TIMESNET IT INCLUDES SCENARIO-SPECIFIC TRAINING AND INFERENCE, WHEREAS PRE-TRAINED MODELS REQUIRE INFERENCE ONLY. DETAILED EXECUTION SETTINGS AND TRAINING/INFERENCE DECOMPOSITION ARE GIVEN IN TABLE VIII.

	Gap filling accuracy					**Processing time**			
Method ↓	*Bias*	*MAE*	*MedAE*	*RMSE*	*SAM*	*Min*	*Mean*	*Max*	*Gap-dependent*
Linear	+392.9	1319.5	288	4151.4	10.60	0.4	0.5	1.0	no
Harmonic	+2188.8	3823.8	595	11177.5	16.85	17.5	18.5	20.9	yes
Proposed HCB	**+20.1**	**879.4**	**262**	**1782.4**	**7.88**	**0.9**	**1.0**	**1.7**	**no**
Proposed HELIX	**-0.6**	**688.4**	**188**	**1636.4**	**7.38**	**38.9**	**42.7**	**47.6**	**yes**
SAITS	+98.0	1260.6	712	2139.8	20.30	7019.9	7716.9	10056.5	yes
TimesNet	-1206.8	2187.2	1412	3247.2	52.05	4170.8	4476.1	7246.0	yes
MOMENT	-1203.9	1612.0	588	2948.5	19.77	573.7	582.8	595.7	yes
U-TILISE	-822.0	1117.3	262	2377.1	13.16	1216.8	1229.4	1241.4	yes
UnCRtainTS	+3946.3	4223.1	4712	4641.0	32.62	862.8	878.0	899.4	yes
Unit →	[DN]	[DN]	[DN]	[DN]	[°]	[s]	[s]	[s]	[yes/no]

Note: Radiometric errors are reported in DN; divide by 10,000 to obtain surface-reflectance units.

B. Total Reconstruction Accuracy

Table IV summarises aggregate controlled-gap reconstruction performance and recorded processing time under the respective benchmark configurations as defined in Fig. 3. HELIX achieves the lowest aggregate MAE (688.4 DN), MedAE (188 DN), RMSE (1636.4 DN), and mean SAM (7.38°), together with a near-zero bias of -0.6 DN. At the aggregate level, this corresponds to a 47.8 % lower MAE than Linear and a 38.4 % lower MAE than U-TILISE, the strongest satellite-specific learned reference in this benchmark. HCB reaches the second-lowest aggregate MAE (879.4 DN) and a similarly low SAM of 7.88°, while requiring approximately 1 s per benchmark run. Harmonic reconstruction shows substantially larger radiometric errors and positive bias, despite remaining computationally inexpensive relative to the learned approaches. The aggregate results therefore indicate a pronounced accuracy–efficiency trade-off: HELIX provides the lowest reconstruction errors, whereas HCB retains low error at substantially lower computational cost.

The aggregate differences in Table IV were assessed using non-parametric inference based on AOI–year-level accuracy summaries. The primary analysis comprised 14 paired AOI–year observations, while the 420 AOI–year–scenario combinations were retained as a scenario-resolved sensitivity analysis. Runtime was analysed separately using the directly measured execution times.

The global Friedman [77] and Kruskal–Wallis [78] tests identify significant differences among the nine methods for all benchmark parameters (with $p < 0.001$). For accuracy, the primary AOI–year analysis shows strong concordance ($W \in [0.727, 0.854]$), whereas the scenario-resolved sensitivity analysis confirms the same global conclusion ($W \in [0.505, 0.632]$). Runtime likewise exhibits a strong overall method effect ($\epsilon^2 = 0.939$).

Pairwise differences were subsequently assessed using two-sided Wilcoxon signed-rank tests for the four accuracy parameters and Mann–Whitney U tests for runtime. Holm correction was applied across the 36 method pairs within each parameter [79]. Figure 8 shows the corresponding rank-biserial effect sizes r_{rb} [80]. Green stands for a better performance and orange for a poorer performance of the method in the row versus the method in the column. Grey cells denote comparisons that are not significant after Holm correction. Only unique method pairs are shown of the originally skew-symmetric matrix.

The pairwise results in Fig. 8 reveal distinct strengths of the two proposed reconstruction approaches. HELIX achieves significantly lower MAE, MedAE, and Relative MAE than HCB, with $r_{\mathrm{rb}} = 0.96$, 0.96, and 1.00, respectively, whereas their SAM difference is not significant after Holm correction. HELIX also significantly outperforms Linear and Harmonic

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409

TABLE V
GLOBAL NON-PARAMETRIC COMPARISON OF THE NINE BENCHMARK METHODS. ACCURACY METRICS ARE ASSESSED USING FRIEDMAN TESTS ON 14 PAIRED AOI–YEAR OBSERVATIONS, WITH KENDALL'S W AS EFFECT SIZE; THE SCENARIO-RESOLVED SENSITIVITY ANALYSIS RETAINS ALL 420 AOI–YEAR–SCENARIO COMBINATIONS. RUNTIME IS ASSESSED USING A KRUSKAL–WALLIS TEST ON THE DIRECTLY MEASURED EXECUTION TIMES, WITH ϵ^2 AS EFFECT SIZE.

		Primary / runtime analysis				Scenario-resolved sensitivity		
Metric	*Test*	n	*Statistic* χ^2_F	p	*Effect size* W	n	*Statistic* χ^2_F	*Effect size* W
MAE	Friedman	14	81.562	< 0.001	0.728	420	1704.592	0.507
MedAE	Friedman	14	81.438	< 0.001	0.727	420	1884.534	0.561
Relative MAE	Friedman	14	87.410	< 0.001	0.780	420	1698.227	0.505
SAM	Friedman	14	95.619	< 0.001	0.854	420	2122.574	0.632
Runtime	Kruskal–Wallis	3020	$H = 2836.377$	< 0.001	$\epsilon^2 = 0.939$	–	–	–

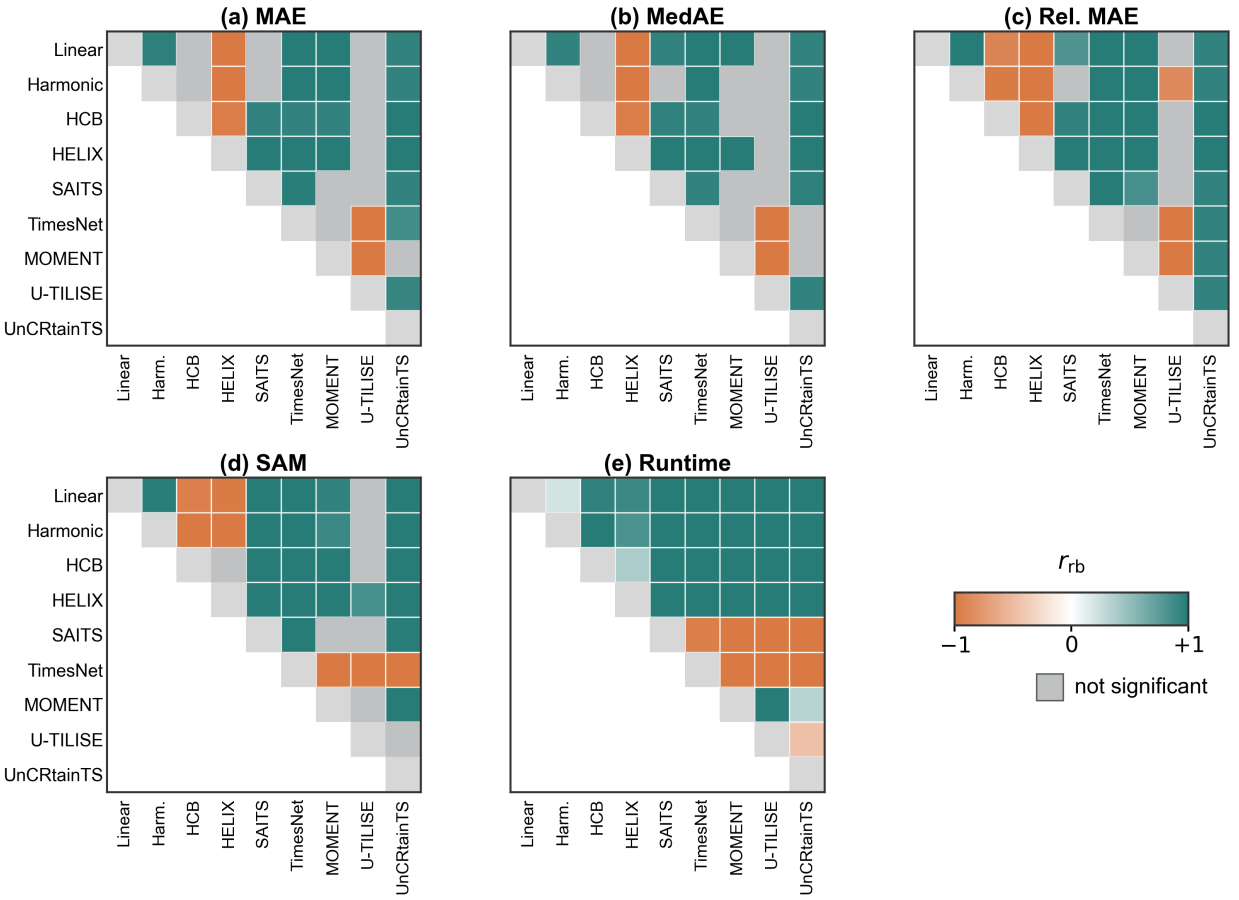


Fig. 8. Pairwise comparison of the nine benchmark methods for (a) MAE, (b) MedAE, (c) Relative MAE, (d) SAM, and (e) runtime. Colour encodes the signed rank-biserial effect size r_{rb} on the common range $[-1, +1]$. Positive values (green) indicate a better performance and negative values (orange) indicate a poorer performance of the method written in the row in relation to the method in the column. Grey cells denote comparisons that are not significant after Holm correction. Accuracy panels are based on paired Wilcoxon signed-rank tests; runtime comparisons use Mann–Whitney U tests.

across all four accuracy parameters. Relative to U-TILISE, the radiometric differences are not significant, while HELIX retains a significant SAM advantage ($r_{\mathrm{rb}} = 0.85$).

HCB exhibits a different performance profile. Compared to Linear, it achieves significantly lower Relative MAE ($r_{\mathrm{rb}} = 0.90$) and SAM ($r_{\mathrm{rb}} = 0.94$), while the MAE and MedAE differences are not significant. HCB is likewise not significantly separated from U-TILISE across the four accuracy parameters. The pairwise structure therefore indicates metric-dependent differences rather than uniform dominance.

The scenario-resolved sensitivity analysis further resolves the HCB–HELIX relationship for spectral fidelity. HELIX yields the lower SAM in 62.4 % of the 420 evaluation contexts, whereas HCB is lower in 37.6 %, corresponding to $r_{\mathrm{rb}} = -0.27$ for HCB relative to HELIX. Thus, the non-significant SAM difference after AOI–year aggregation does not imply identical behaviour. Both approaches provide context-dependent spectral advantages, with a stronger overall tendency towards HELIX. Together with the radiometric results, this supports complementary rather than redundant behaviour between HCB and HELIX.

Runtime adds a second dimension to this complementarity. As shown in Table IV and Fig. 8, Linear and HCB occupy the lowest-runtime range, followed by Harmonic and HELIX,

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409

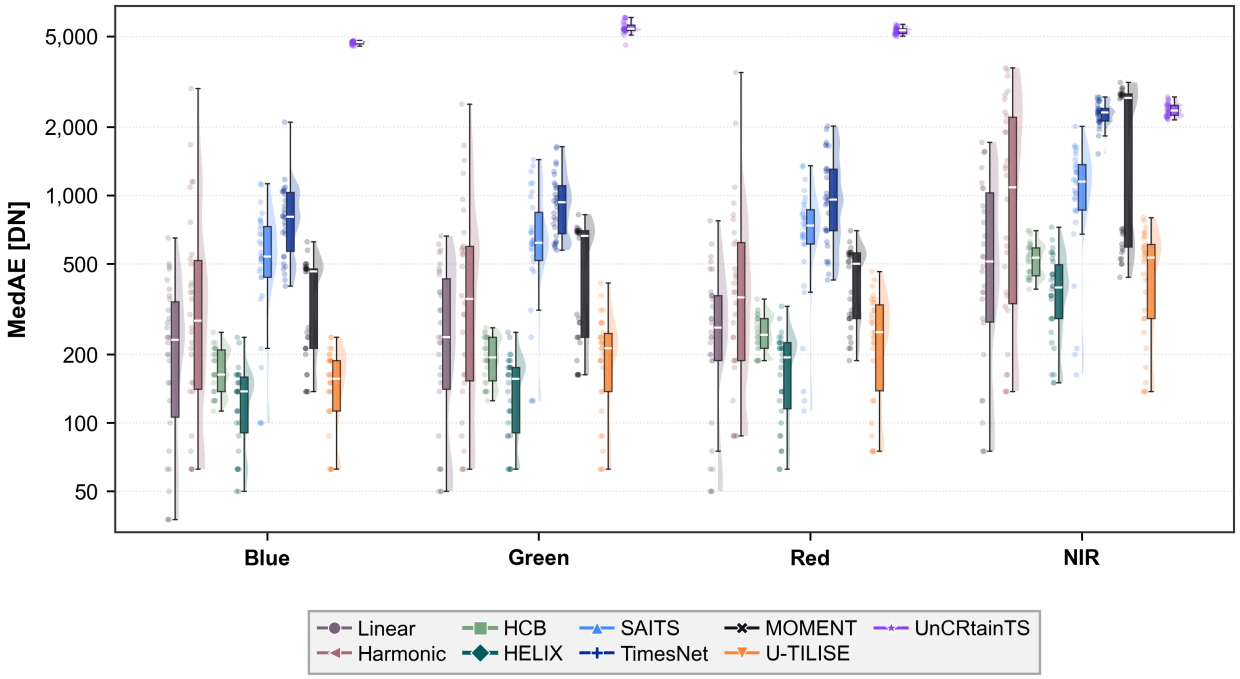


Fig. 9. Band-wise reconstruction error distributions for the nine benchmark methods. Values represent MedAE on withheld observations and are shown on a logarithmic scale. Distributions summarise variation across AOIs, years, and controlled gap scenarios.

while the learned approaches require substantially greater computational effort under the benchmark configurations. This is particularly relevant relative to U-TILISE: HELIX achieves lower aggregate MAE, MedAE, RMSE, and SAM while reducing mean runtime from 1229.4 s to 42.7 s, whereas HCB achieves lower aggregate MAE, RMSE, and SAM, the same MedAE, and a mean runtime of only 1.0 s. Hence, U-TILISE does not provide a compensating accuracy advantage for its higher computational demand in this benchmark. Taken together, HELIX and HCB occupy complementary positions in the accuracy–efficiency space: HELIX provides the lowest aggregate reconstruction errors, whereas HCB retains statistically comparable accuracy to U-TILISE at more than three orders of magnitude lower runtime and provides context-specific gains in spectral fidelity.

Figure 9 shows the corresponding band-wise MedAE distributions. HCB and HELIX exhibit consistently low errors across the visible bands, with HELIX generally showing the lowest median values and HCB remaining within a similar low-error range. Linear and U-TILISE overlap this range in several bands, while Harmonic shows higher and more variable errors, particularly towards the NIR. The transition to NIR increases MedAE for most methods and produces a clearer separation of the error distributions. HCB, HELIX, and U-TILISE remain within the lower-error range, whereas SAITS, TimesNet, MOMENT, and Harmonic show stronger increases. UnCRtainTS exhibits markedly higher errors across the visible bands and remains within a high-error regime in NIR.

To complement the full band-wise error distributions in Fig. 9, Fig. 10 characterises the direction and variability of systematic reconstruction error across the benchmark. For each method and spectral band, the marker denotes the median mean error across the evaluated AOI–year–scenario combinations, while the horizontal interval spans the corresponding 5th–95th percentiles. Negative values indicate systematic underestimation and positive values overestimation. The intervals therefore describe the empirical central 90% range of context-wise bias and should not be interpreted as probabilistic confidence intervals.

The bias profiles reveal differences that are not captured by absolute-error magnitude alone. Linear, HCB, HELIX, and SAITS remain comparatively well centred around zero across the four bands, with HELIX showing particularly narrow ranges. Harmonic also has central values close to zero but exhibits substantially wider context-dependent excursions, particularly towards positive bias. TimesNet and MOMENT are predominantly shifted towards negative values, indicating systematic underestimation across substantial parts of the benchmark, while U-TILISE shows a weaker negative tendency. In contrast, UnCRtainTS exhibits pronounced positive bias, especially in the visible bands. Radiometric bias therefore varies systematically with both reconstruction method and spectral band.

Figure 11 evaluates MedAE by AOI and year. Absolute reconstruction error varies strongly between regions, while the relative performance patterns remain similar across the two evaluated years in most AOIs. AOI 3 represents the most challenging regional setting, with substantially higher errors for all methods; HELIX nevertheless attains the lowest MedAE in both years. U-TILISE achieves the lowest errors in AOI 4 and remains within the low-error range in AOI 2, indicating strong performance in selected regional settings. HCB and HELIX show comparatively small differences between 2019 and 2020 in most AOIs, whereas Harmonic and several learned

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409

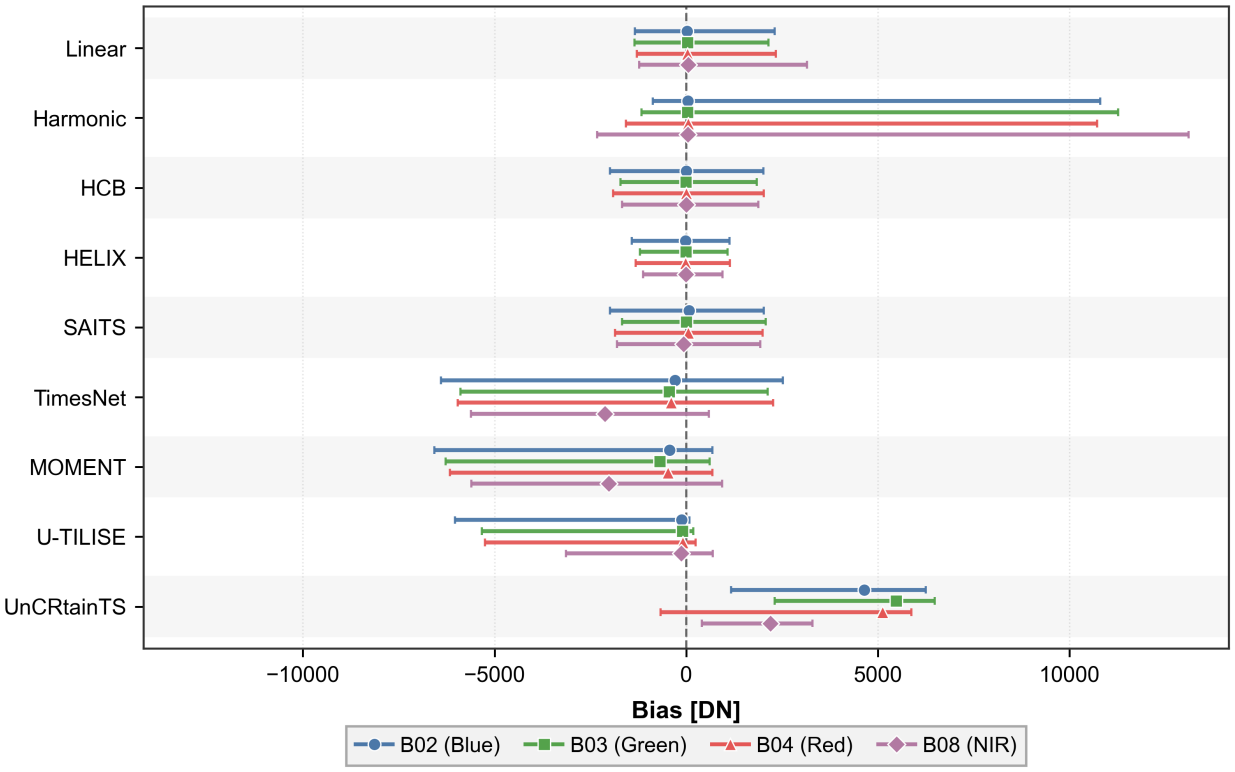


Fig. 10. Empirical bias ranges by spectral band. Markers denote the median mean error across the evaluated AOI–year–scenario combinations, while horizontal intervals span the corresponding 5th–95th percentiles. Negative values indicate underestimation and positive values overestimation; the dashed vertical line denotes zero bias.

approaches exhibit larger regional and interannual variation.

C. Land-Cover and Gap-Scenario Sensitivity

The AOI–year comparison indicates pronounced regional differences in reconstruction difficulty. The following diagnostics therefore examine whether these differences are systematically associated with semantic surface type and temporal gap geometry.

Figure 12 shows the land-cover-stratified MedAE distributions. Reconstruction difficulty varies substantially across ESA WorldCover classes [19] and spectral bands. For most vegetated and managed classes, including trees, grassland, cropland, and herbaceous wetlands, errors increase from the visible bands towards NIR. HCB and HELIX remain within the lowest-error range across most classes, with HELIX generally attaining the lowest or near-lowest median errors. Linear and U-TILISE are competitive in several classes, whereas Harmonic typically occupies an intermediate error range and shows stronger degradation for snow and ice. Snow and ice constitute a particularly distinct regime, where HCB and HELIX retain comparatively low errors while several other methods deteriorate markedly. Permanent water and moss and lichen also exhibit class-specific behaviour with a reduced separation between some methods. SAITS, TimesNet, MOMENT, and UnCRtainTS show larger class- and band-dependent errors in several regimes, with UnCRtainTS remaining among the highest-error methods for most land-cover classes.

To retain the temporal gap dimension, Fig. 13 contrasts cropland with snow and ice across the 30 controlled scenarios. Recurring scenario-dependent error increases occur in both classes, while their magnitude differs substantially between methods and surface types, indicating that reconstruction difficulty depends jointly on temporal gap configuration and land-cover regime. To examine whether reconstruction difficulty is seasonally structured across surface types, Figs. 14 and 15 relate reconstruction error to ESA WorldCover class and the seasonal position of the withheld observations. Seasonal position is represented by the median day of year (DOY) of the missing acquisitions for each gap scenario, enabling comparison across the climatically heterogeneous AOIs without imposing fixed phenological phase boundaries. The analysis includes all WorldCover classes represented on the common evaluation support and distinguishes the visible (VIS) and near-infrared (NIR) spectral ranges. The seasonal profiles reveal pronounced method- and class-dependent differences. Linear and Harmonic exhibit the strongest seasonal excursions, with several land-cover classes showing marked error increases for specific gap positions. In contrast, HCB and HELIX generally maintain flatter seasonal profiles, although systematic class differences remain. Among the learned approaches, SAITS and TimesNet show comparatively moderate seasonal variation, whereas MOMENT, U-TILISE, and UnCRtainTS exhibit stronger class-specific offsets or localized seasonal responses.

D. Spectral Fidelity and Gap-Scenario Robustness

Absolute band-wise reconstruction errors partly reflect the radiometric magnitude of the underlying signal. Figure 16

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited, content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409

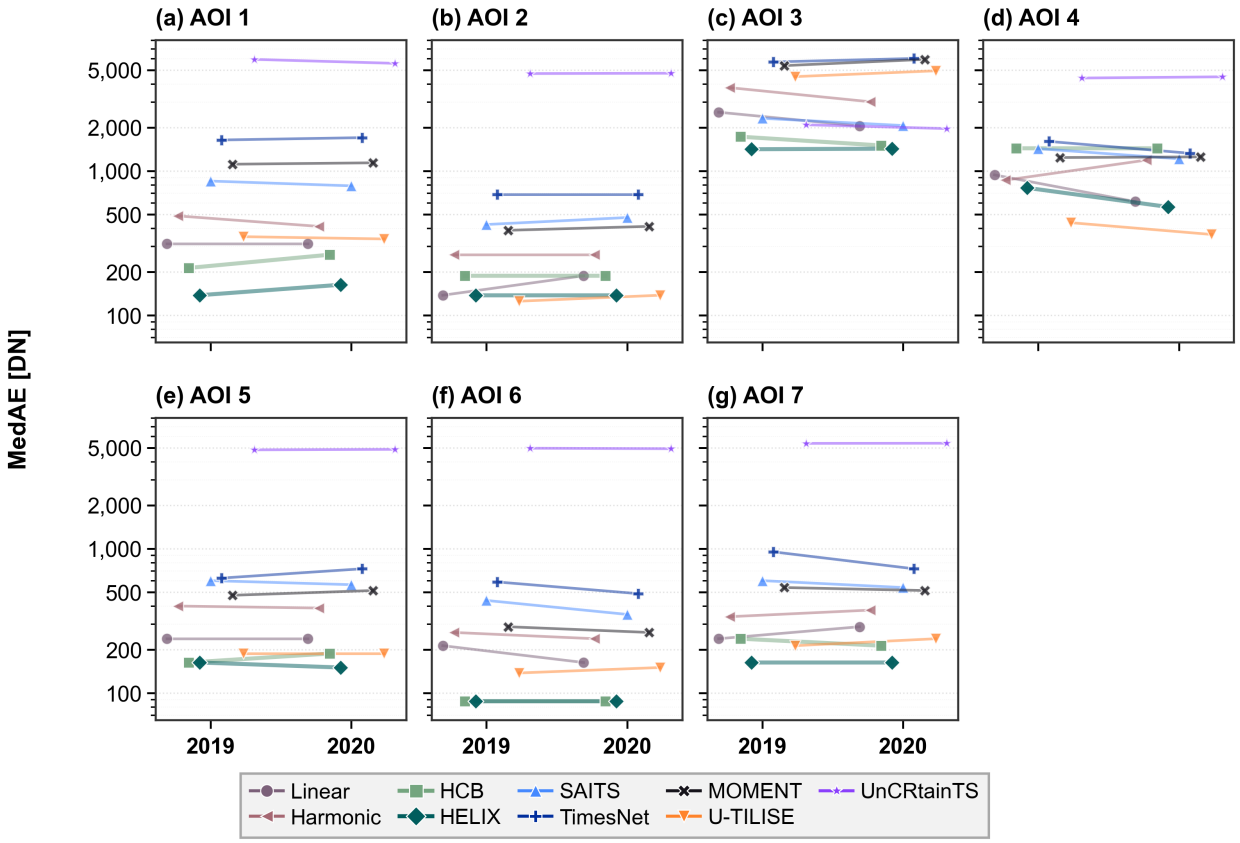


Fig. 11. AOI–year reconstruction accuracy for the nine benchmark methods. Values represent MedAE on withheld observations, aggregated over spectral bands and controlled gap scenarios for each AOI and year. The logarithmic scale highlights regional differences in reconstruction error and changes between the two evaluated years.

therefore contrasts absolute MAE with MAE normalized by the mean reference reflectance of the corresponding AOI–year–scenario–band evaluation unit. In absolute terms, B08 exhibits higher errors than the visible bands for most methods. After normalization, this separation decreases substantially, and relative NIR errors are comparable to or lower than those of the visible bands for several methods. The elevated absolute NIR errors therefore cannot be interpreted as uniformly poorer reconstruction performance. SAM complements the radiometric error measures by quantifying spectral-shape fidelity. Figure 17 relates SAM to the signed brightness ratio $\log_{10}(\hat{b}/b)$, where b and $\hat{b}$ denote the Euclidean norms of the reference and reconstructed four-band spectral vectors. Negative and positive values indicate brightness under- and overestimation, respectively, while zero denotes balanced spectral brightness. HCB and HELIX remain within the lowest SAM range across the central brightness-ratio interval. Linear and U-TILISE show stronger increases towards larger brightness deviations, Harmonic occupies an intermediate range, and SAITS, TimesNet, MOMENT, and UnCRtainTS exhibit higher spectral-angle errors over substantial parts of the ratio domain. Low spectral-angle errors therefore occur predominantly where reconstructed spectral brightness remains close to the reference. Figure 18 resolves spectral fidelity across the 30 controlled gap scenarios. HCB and HELIX maintain comparatively low SAM across most scenarios, whereas Linear, Harmonic, and U-TILISE exhibit more pronounced scenario-specific peaks. SAITS, MOMENT, and UnCRtainTS occupy higher or more variable SAM ranges, while TimesNet remains consistently elevated. Spectral reconstruction quality therefore depends on the temporal configuration of the missing observations in addition to radiometric error magnitude.

To assess whether the scenario-level spectral errors translate into seasonal structure at the acquisition-date level, Fig. 19 resolves SAM by day of year for AOI 5 in 2020. For each acquisition date, SAM is aggregated over all controlled scenarios in which that observation was withheld and reconstructed. HCB and HELIX remain within a comparatively low SAM range throughout the annual cycle, whereas SAITS shows stronger date-dependent fluctuations and TimesNet remains consistently elevated. Harmonic, Linear, MOMENT, U-TILISE, and UnCRtainTS exhibit intermediate or more variable temporal behaviour. The date-resolved example therefore shows that seasonal spectral-error structure is method dependent rather than being expressed uniformly across reconstruction approaches.

V. Discussion

The aggregated and stratified results in Tables III and IV as well as Figs. 9–18 show that post-mask reconstruction performance is governed not only by the amount of missing data but also by the temporal configuration of the gaps,

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409

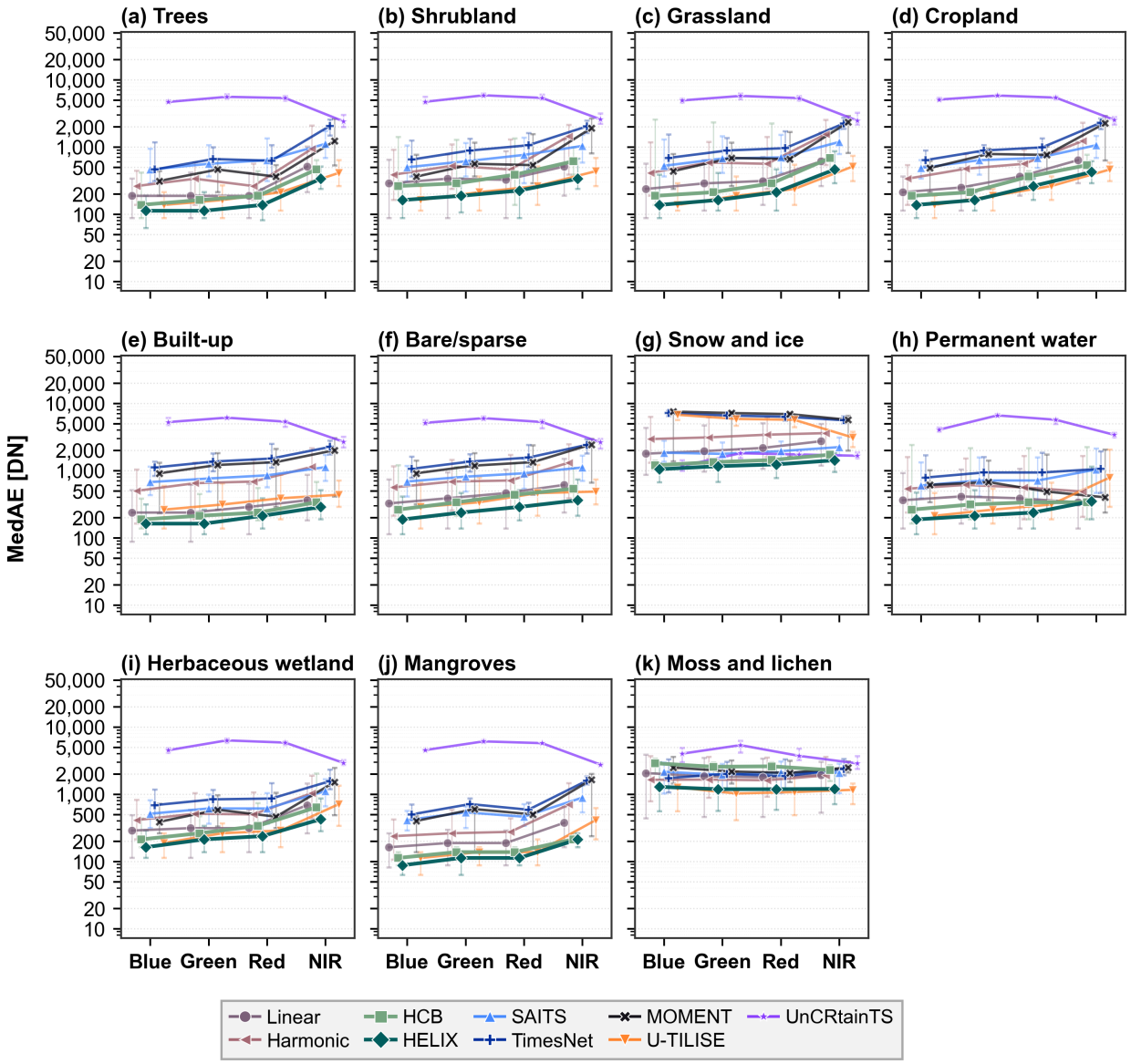


Fig. 12. Land-cover-stratified reconstruction error for all evaluated methods. Values represent MedAE on withheld observations and are grouped by ESA WorldCover class [19] and spectral band. The logarithmic scale highlights class-dependent reconstruction difficulty and band-dependent error behaviour.

spectral band, land-cover class, regional surface regime, and reconstruction strategy. Four main conclusions emerge. First, heterogeneous spatio-temporal cloud distribution leads to very heterogeneous reconstruction results and therewith, the distortion of spatial and temporal patterns in the image that are hard to numerically assess. Second, reconstruction difficulty depends strongly on temporal support geometry rather than on missing-data fraction alone. Third, explicit structural constraints can stabilise reconstruction under complete-date gaps, either through feature-side temporal structure in HCB or semantic seasonal guidance in HELIX. And fourth, model complexity does not imply superior reconstruction: reliable gap filling requires a balanced combination of radiometric accuracy, spectral fidelity, computational efficiency, and interpretable behaviour.

The sophisticated in-deep interpretation however requires a clear idea of how noise evolves in remote sensing images. According to recent noise models, e.g. the perturbation based model [81], [82], the apparent signal contains two noise components: an additive noise floor and a multiplicative noise component proportional to the signal's amplitude. In consequence, low scattering targets like water typically are primarily influenced by the noise floor and thus, often hard to predict (Fig. 15h). High scattering targets have a stronger impact of the multiplicative noise component. Additionally, both components are not independent, but intrinsically linked to each other [83]. This explains two observations: first, snow being a very bright target shows a higher MedAE than other land cover classes (Fig. 12). Second, the NIR channel with a higher overall backscattering is more strongly influenced by the multiplicative component. This becomes obvious in Fig. 16, where the MAE is generally higher for NIR, but the relative MAE is equal to or even lower than the MAE of the visible channels. It is important to keep these basic points in

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409

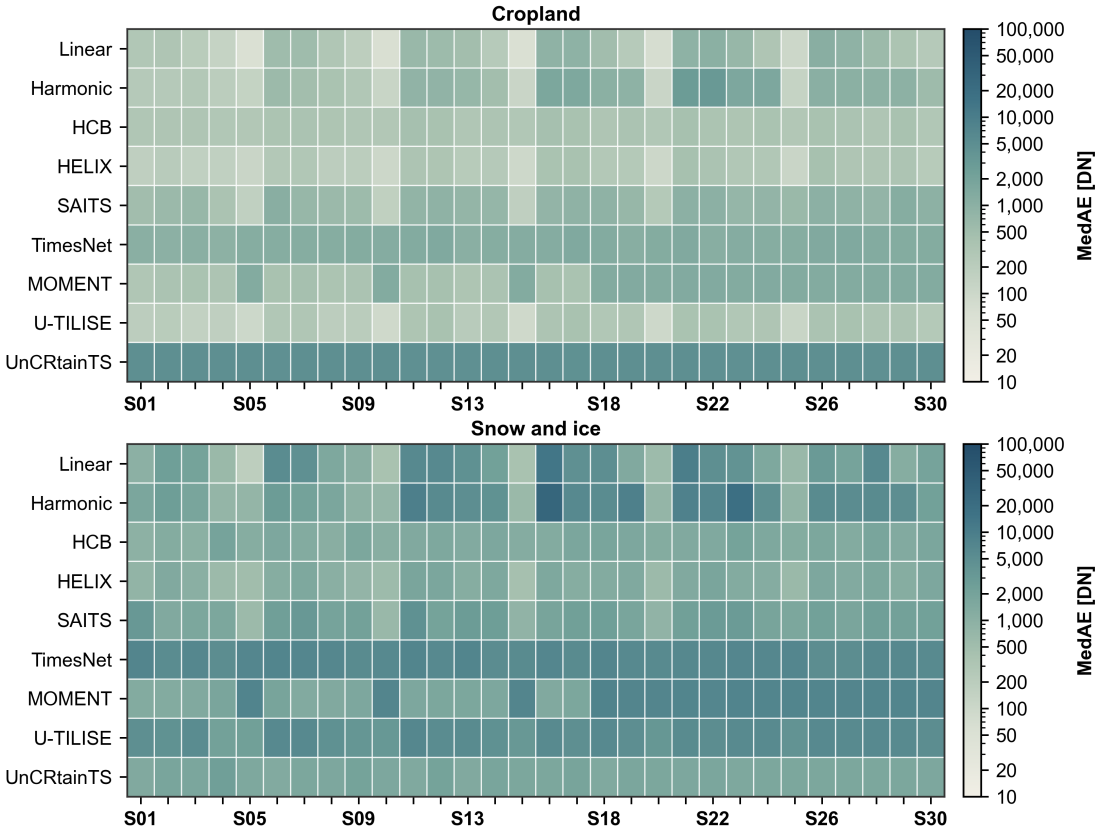


Fig. 13. Scenario-wise MedAE for cropland and snow and ice across the 30 controlled gap scenarios. Rows correspond to the nine benchmark methods. The two classes illustrate how temporal gap configuration interacts with contrasting land-cover regimes.

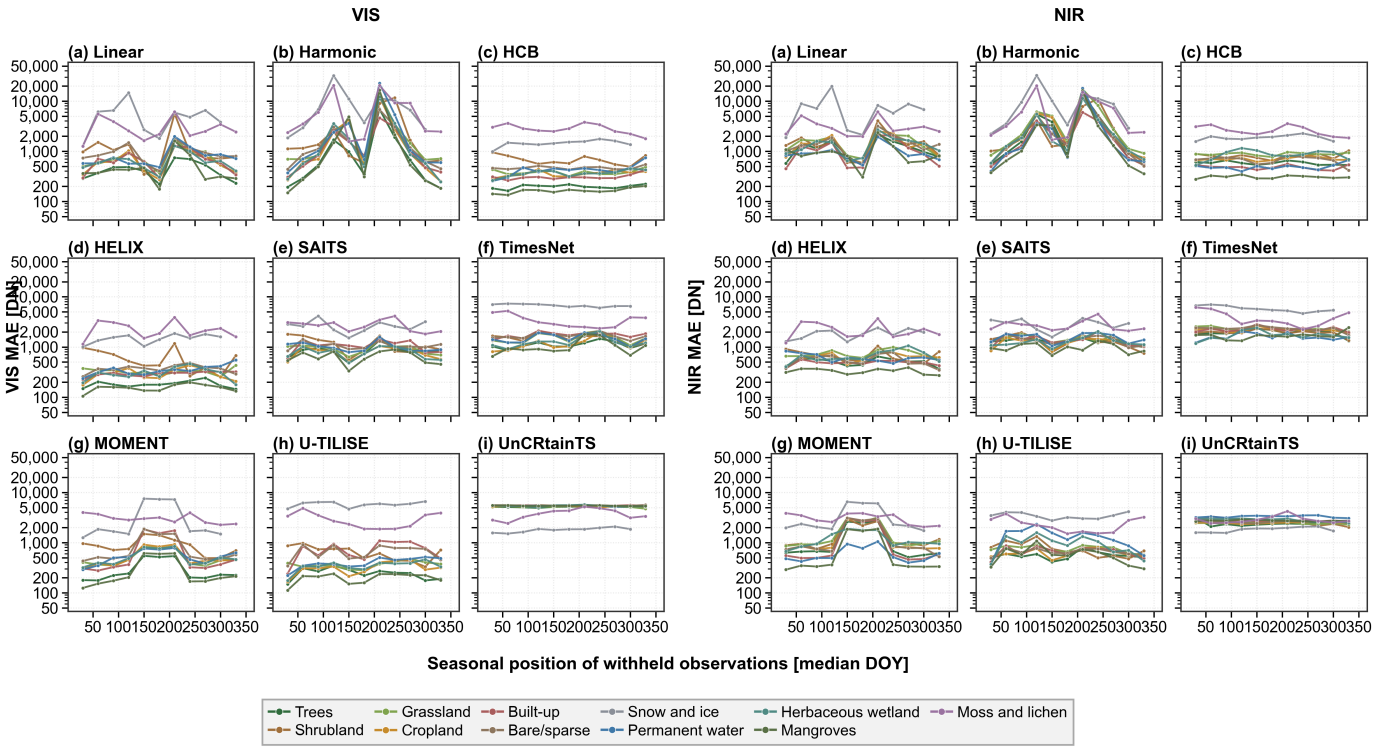


Fig. 14. Seasonal land-cover dependence of absolute reconstruction error for the nine benchmark methods. VIS MAE is aggregated over B02–B04, while NIR MAE corresponds to B08. Curves represent ESA WorldCover classes and the x-axis denotes the median DOY of the withheld observations for each gap scenario. Values are aggregated across AOIs and years on the common evaluation support. VIS and NIR use the same logarithmic error scale.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409

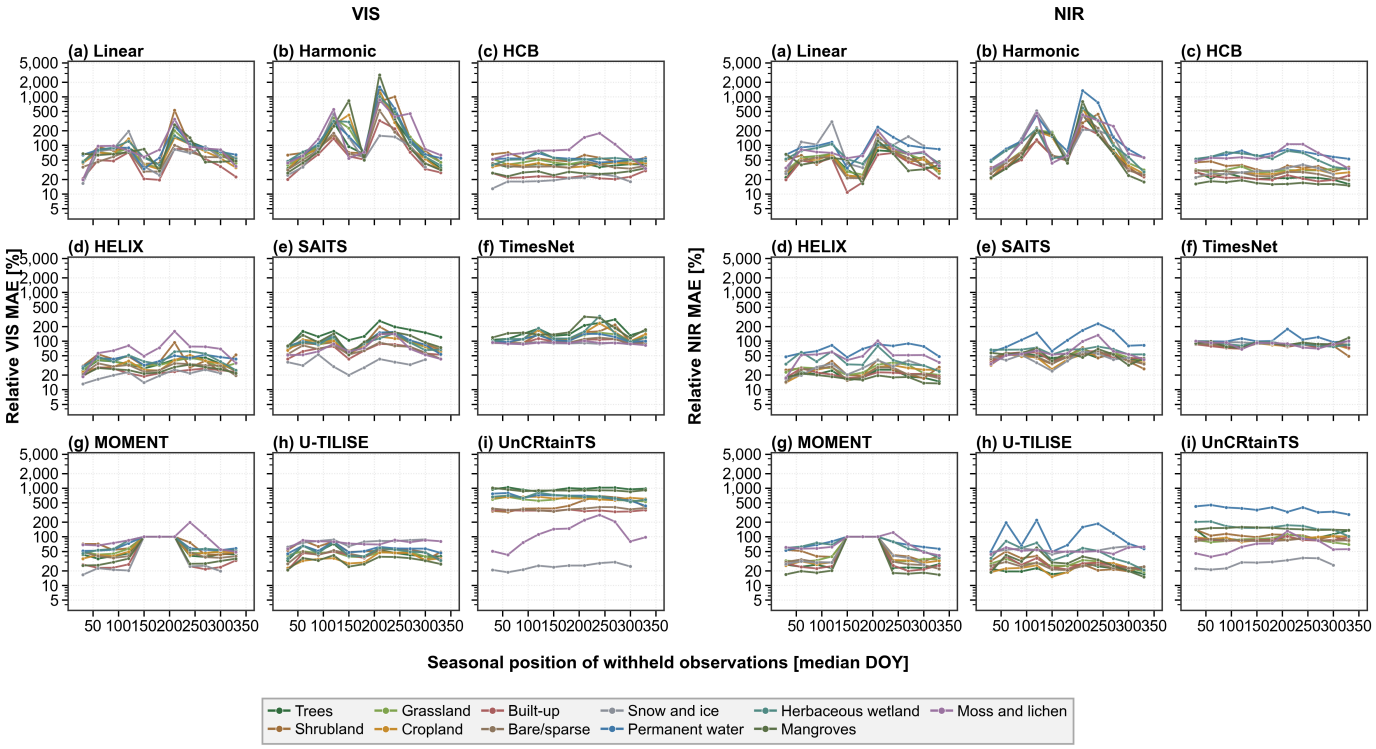


Fig. 15. Seasonal land-cover dependence of relative reconstruction error for the nine benchmark methods. Relative VIS MAE is aggregated over B02–B04, while relative NIR MAE corresponds to B08. Curves represent ESA WorldCover classes and the x-axis denotes the median DOY of the withheld observations for each gap scenario. Errors are normalized by the corresponding mean reference reflectance and aggregated across AOIs and years on the common evaluation support. VIS and NIR use the same logarithmic error scale.

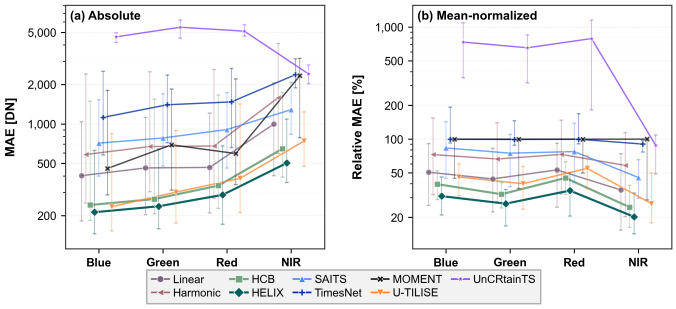


Fig. 16. Band-wise absolute and relative reconstruction error. Markers denote the median and whiskers the interquartile range across AOI–year–scenario combinations. Panel (a) shows absolute MAE in reflectance DN; panel (b) shows MAE normalized by the mean reference reflectance of the corresponding evaluation unit and expressed as percentage. Both panels use logarithmic y-axes.

mind when reading the following explanations.

A. Sensitivity to Spatially Heterogeneous Cloud/Shadow Masks

The spatially heterogeneous cloud/shadow experiment provides direct evidence for the first research hypothesis. Fig. 7 and Table III demonstrate that reconstructions with comparable radiometric behaviour can differ substantially in spectral consistency, local spatial structure, and temporal dynamics. This distinction is particularly relevant for methods that tend towards smooth, averaged, or otherwise regularised solutions or hallucinations: such reconstructions may remain numerically close to the reference while attenuating spatial contrast or temporal transitions, whereas other methods may preserve structural characteristics despite a systematic radiometric offset. The gradient-based diagnostics therefore complement conventional radiometric errors by indicating whether spatial and temporal structure is retained rather than merely whether reconstructed values are numerically close to the reference. HCB and HELIX are particularly consistent in this respect. Despite relying on different reconstruction principles, both retain comparatively coherent spectral, spatial, and temporal structure under clustered cloud/shadow missingness. HCB achieves this solely from the observed spectro-temporal signal, whereas HELIX additionally exploits class-specific seasonal information; their behaviour therefore supports the interpretation of feature-side and label-side constraints as complementary mechanisms rather than interchangeable variants of the same reconstruction strategy. The experiment also demonstrates why spatially heterogeneous missingness constitutes a distinct reconstruction setting from complete-date gaps. Partially masked acquisitions retain contemporaneous spatial information that can be exploited to varying degrees, whereas complete-date masking removes this support entirely and isolates reconstruction from the remaining temporal context. Taken together, the two experimental settings indicate that robust post-mask reconstruction should be evaluated under both heterogeneous spatial support and restrictive temporal-support conditions. The first hypothesis is therefore supported:

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409

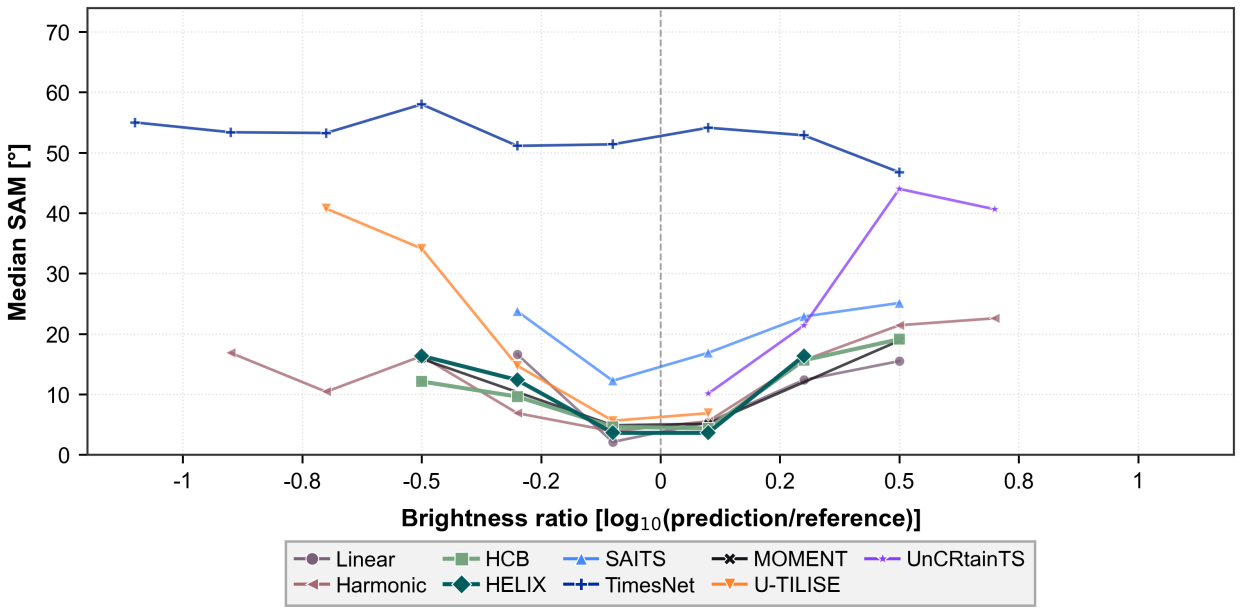


Fig. 17. Spectral fidelity as a function of signed brightness ratio. Lines show binned median SAM for all evaluated methods on withheld observations.

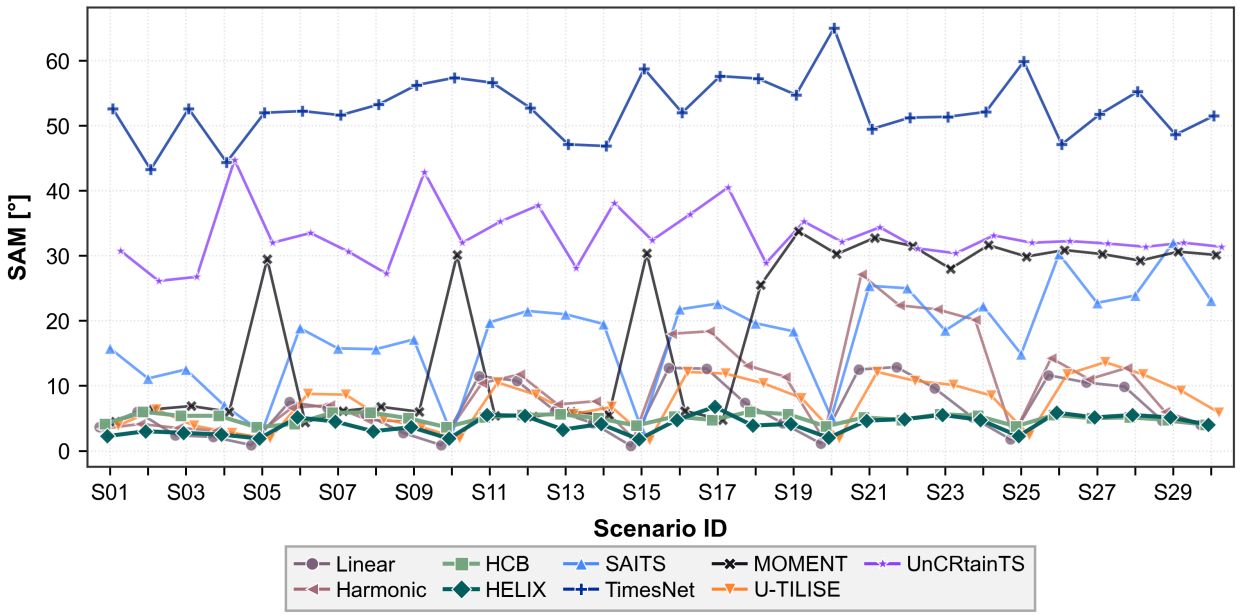


Fig. 18. Scenario-wise spectral fidelity for all evaluated methods. Values represent SAM aggregated over AOIs, years, and withheld observations. Scenario IDs correspond to the controlled gap matrix in Fig. 3.

real world scenarios are far too complex to establish a reliable benchmark dataset for the evaluation of gap filling algorithms by simple radiometric error measures. The present cloud/shadow experiment therefore remains controlled and is limited to one AOI and one clustered spatial configuration; naturally occurring cloud and shadow masks across additional regions, seasons, and cloud regimes thus remain an important direction for further validation.

B. Impact of Temporal Gap Geometry

The second hypothesis is supported by the evaluation of the world-wide scenarios (Figs 2 and 3): reconstruction accuracy depends not only on the number of withheld dates, but also on their temporal arrangement. The scenario-wise error and SAM diagnostics in Figs. 13 and 18 show that gap configurations with comparable missing-data fractions can produce markedly different reconstruction errors. Gaps located within the observed sequence retain two-sided temporal support and are therefore closer to interpolation, whereas gaps near the beginning or end increasingly require one-sided reconstruction. Complete-date gaps further remove all observations at the affected acquisition time, making temporal support particularly critical.

Linear interpolation illustrates this dependence directly.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409

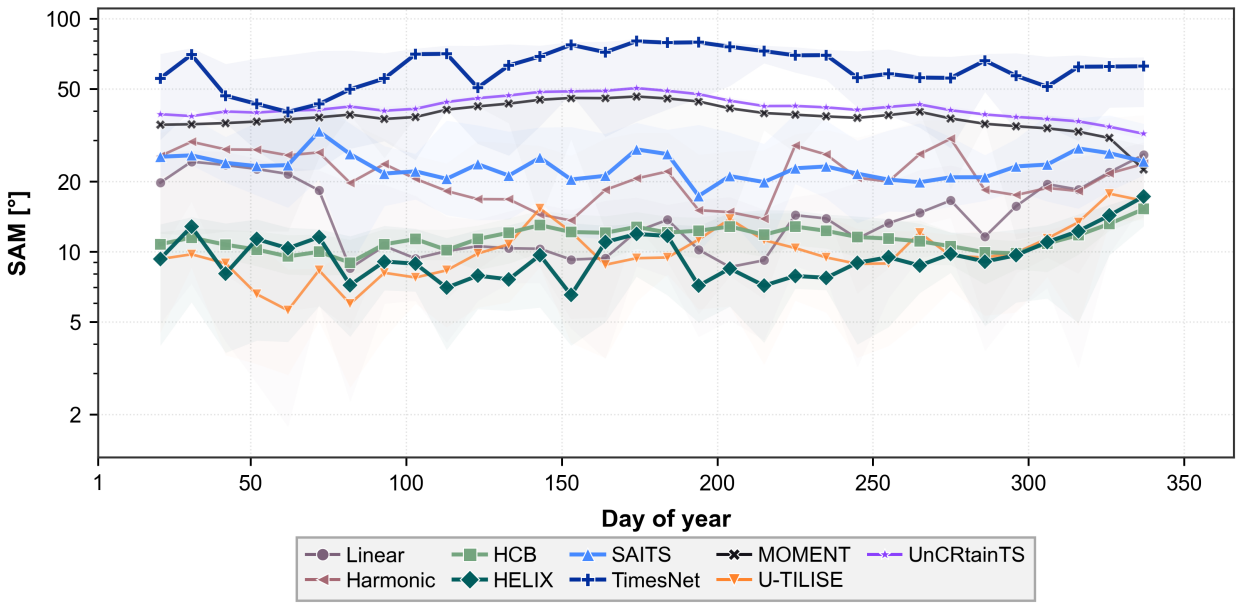


Fig. 19. Date-resolved spectral reconstruction error for AOI 5 in 2020. For each method and physical acquisition date, markers denote the median scenario-specific mean SAM across all controlled gap scenarios in which that acquisition was withheld and reconstructed; shaded intervals show the interquartile range. The horizontal axis represents the actual acquisition day of year.

It remains competitive when valid observations bracket the missing dates but deteriorates when temporal anchors are sparse or unfavourably positioned. Harmonic reconstruction is less dependent on immediate neighbours but exhibits larger scenario-specific excursions, indicating that a global periodic representation alone does not ensure robustness under irregular support. HCB and HELIX are less sensitive to unfavourable gap placement because they exploit additional structure beyond adjacent observations: HCB uses the spectro-temporal structure of the observed sequence, whereas HELIX combines local temporal residuals with land-cover-conditioned seasonal expectations.

The SAM analysis in Fig. 18 shows that the same dependence extends to spectral fidelity. HCB and HELIX maintain comparatively low SAM across most scenarios, whereas several baselines exhibit stronger scenario-specific increases. The geometry of the temporal gap therefore affects not only the magnitude of the radiometric error but also the preservation of the spectral angle. The controlled complete-date scenarios consequently provide a stringent test of temporal support geometry rather than merely varying the overall missing-data fraction.

C. Spectral, Semantic, and Regional Error Structure

The third hypothesis is also supported, but with an important qualification. Systematic constraints improve reconstruction stability, yet the observed error structure cannot be attributed to land-cover alone. The band-wise and land-cover-stratified results in Figs. 9 and 12 show a consistent increase in error from the visible bands towards NIR for many vegetated and managed classes. This pattern is linked partly to the larger dynamic range of NIR reflectance and to its stronger sensitivity to seasonal vegetation dynamics as well as the higher noise content due to the higher reflectance values over healthy vegetation.

Land-cover effects are nevertheless visible. Snow and ice, moss and lichen, and permanent water form more difficult or more distinct reconstruction regimes than many vegetation and built-up classes. These classes differ not only semantically, but also radiometrically and seasonally. Snow and ice are affected by melt timing, mixed snow–water states, and abrupt transitions. Permanent water has low NIR reflectance and different spectral dynamics from vegetated surfaces. Moss and lichen occur in environments where sparse vegetation, snow persistence, and low reflectance contrast can complicate reconstruction. These effects explain why the difficulty of class-level reconstruction is not uniform.

At the same time, the scenario-wise land-cover checks in Fig. 13 indicate that semantic class does not fully determine reconstruction behaviour. Similar scenario-dependent structures occur across different classes. Land-cover mainly shifts the absolute error level and the spectral-band response, whereas the relative difficulty of individual gap configurations remains broadly controlled by temporal support geometry. This is important for interpreting the controlled complete-date design. The scenario effect is not merely a land-cover artefact; it reflects the temporal information available to the reconstruction method.

The AOI–year results in Fig. 11 add a regional layer to this interpretation. AOI 3 is consistently difficult, and AOI 4 also shows a strong dependence on the method. These differences cannot be explained by a single WorldCover class alone, because the same class can occur under different climate regimes, seasonal amplitudes, snow conditions, phenological cycles, and surface mixtures. The study areas therefore span the major Köppen climate groups and a broad range of environmental settings. Reconstruction difficulty is thus better

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited, content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409

understood as a joint function of regional surface regime, temporal dynamics, spectral magnitude, and gap geometry. The comparatively consistent behaviour of HCB and HELIX across the evaluated AOIs and years indicates that their structural constraints remain effective across contrasting environmental settings represented in the benchmark, although absolute error levels remain scene dependent. For HELIX, this raises the additional question of how strongly reconstruction depends on the internal seasonal representation and the semantic prior.

Tables VI and VII address this question through component ablation, harmonic-order sensitivity, and controlled perturbation of the WorldCover prior. The component ablation demonstrates the contribution of local residual fusion to the semantic seasonal prototype. Adding the residual term reduces MAE from 578.5 to 301.4 DN and SAM from 7.82° to 3.02°, while the subsequent plausibility constraints have only a minor effect on aggregate accuracy.

The harmonic-order experiment supports the use of a low-order seasonal representation. Because each prototype spans one annual cycle, $F = 1$ captures the fundamental annual component, while $F = 2$ additionally represents a semi-annual component and can therefore accommodate asymmetric or multi-phase seasonal trajectories without substantially increasing model complexity. Higher orders introduce progressively shorter-period oscillations that are less well supported by irregular and partially missing annual observations and generally increase reconstruction error. The AOI-wise results nevertheless show that no single order is optimal in every regional setting. The polar, snow- and ice-dominated AOI 3 favours $F = 1$, consistent with a comparatively simple dominant annual cycle, whereas the cold-continental and more heterogeneous AOI 4 benefits more clearly from $F = 2$, plausibly reflecting a more complex combination of vegetation, water, and snow-related seasonal dynamics (Table I). The selected $F = 2$ therefore represents a parsimonious fixed compromise across the benchmark rather than a universally optimal harmonic order.

The controlled WorldCover perturbation experiment shows similarly limited sensitivity to moderate semantic-label corruption. Even at 30 % corruption, changes remain small relative to the respective regional baseline errors, including in the more difficult AOIs 3 and 4. HELIX therefore benefits from class-specific seasonal information without being strongly determined by moderate random errors in the semantic prior. This robustness should not be extrapolated directly to systematic temporal land-cover change, where a static semantic prior may remain persistently inconsistent with the contemporary surface state. Such deviations can be partly absorbed by the local residual once supported by valid observations, but temporally inconsistent semantic priors remain a distinct limitation.

D. Accuracy, Spectral Fidelity, and Computational Parsimony

The aggregated accuracy and runtime results in Table IV, together with the pairwise analysis in Fig. 8, support the fourth hypothesis with respect to accuracy, spectral fidelity, and computational efficiency. HELIX achieves the lowest aggregate MAE, MedAE, RMSE, and mean SAM, together with near-zero signed bias. Its radiometric advantage over HCB is statistically significant for MAE, MedAE, and Relative MAE, whereas the two approaches are not significantly separated in SAM at the primary AOI–year level. HCB nevertheless reaches the second-lowest aggregate MAE and a similarly low mean SAM while requiring approximately one second per benchmark execution and no external semantic information. The two methods therefore occupy distinct but complementary operating points: HELIX prioritises radiometric accuracy through semantic seasonal guidance, whereas HCB provides comparable spectral fidelity with substantially greater computational parsimony.

TABLE VI
TARGETED HELIX COMPONENT ABLATION AND HARMONIC-ORDER SENSITIVITY. RESULTS ARE POOLED ACROSS SEVEN AOIS AND THREE REPRESENTATIVE GAP SCENARIOS (`sc5`, `sc15`, AND `sc25`) FOR 2020.

Configuration ↓	*F*	*Bias*	*MAE*	*MedAE*	*RMSE*	*SAM*
Prototype	2	-34.9	578.5	234	1105.0	7.82
Prototype + residual	2	-15.7	301.4	99	672.7	3.02
HELIX	1	-9.7	305.1	99	694.9	3.10
HELIX	2	-22.9	305.3	101	672.5	3.11
HELIX	3	-4.9	334.0	108	717.8	3.29
HELIX	4	+0.4	348.1	115	742.2	3.32
Unit →	[–]	[DN]	[DN]	[DN]	[DN]	[°]

The learned baselines do not dominate under the controlled post-mask protocol. U-TILISE is the strongest satellite-specific learned reference and remains competitive across several AOIs, classes, and spectral conditions. Its radiometric differences relative to HELIX are not significant after correction, while HELIX retains a significant SAM advantage. HCB is likewise not significantly separated from U-TILISE across the four primary accuracy parameters. Considering runtime jointly with accuracy, however, neither comparison favours the substantially more computationally demanding U-TILISE implementation: HELIX attains lower aggregate errors with a mean runtime of 42.7 s compared with 1229.4 s for U-TILISE, while HCB achieves statistically comparable primary accuracy at approximately 1.0 s. Harmonic remains computationally inexpensive but exhibits substantially higher aggregate radiometric and spectral errors, illustrating that low runtime alone is insufficient.

The remaining learned approaches show stronger compromises between error, spectral fidelity, and computational demand. SAITS and TimesNet require scenario-specific training and substantially longer runtimes, while MOMENT and UnCRtainTS also occupy higher-error or higher-SAM regimes in several diagnostics. These differences should be interpreted in the context of the respective adaptation paradigms rather than as evidence of intrinsic algorithmic superiority. HELIX estimates its seasonal prototypes and residual statistics directly from valid observations of the target time series, whereas the pre-trained U-TILISE and UnCRtainTS models are applied without target-area fine-tuning. Differences in source domain, temporal context, and input representation may therefore contribute to the observed performance.

Under the present complete-date post-mask setting, model complexity alone does not guarantee superior reconstruction.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409

TABLE VII
AOI-WISE SENSITIVITY OF HELIX TO HARMONIC ORDER AND CONTROLLED PERTURBATION OF THE WORLDCOVER SEMANTIC PRIOR. VALUES DENOTE MAE [DN] AGGREGATED ACROSS THE THREE REPRESENTATIVE GAP SCENARIOS `sc5`, `sc15`, AND `sc25` FOR 2020.

	Harmonic order				*WorldCover corruption*			
AOI ↓	$F = 1$	$F = 2$	$F = 3$	$F = 4$	0%	10%	20%	30%
AOI 1	176.1	175.5	176.3	195.0	175.5	175.9	176.0	175.8
AOI 2	101.7	100.9	102.5	116.7	100.9	101.3	101.5	101.7
AOI 3	887.8	943.9	1021.1	985.9	943.9	954.3	963.9	973.9
AOI 4	536.0	482.1	598.6	669.2	482.1	488.4	491.0	494.7
AOI 5	168.5	167.5	169.3	175.3	167.5	168.0	168.2	168.4
AOI 6	114.0	114.6	117.0	128.9	114.6	116.3	117.6	118.1
AOI 7	165.2	166.8	169.4	180.9	166.8	166.4	165.6	165.3
Unit →	[DN]	[DN]	[DN]	[DN]	[DN]	[DN]	[DN]	[DN]

Linear interpolation is computationally minimal but becomes less reliable under unfavourable temporal support, while HCB retains low computational cost with markedly stronger aggregate accuracy and spectral fidelity. HELIX requires additional computation for semantic prototype construction and scenario-dependent reconstruction but remains substantially less demanding than the learned references. The results therefore support transparent, structurally constrained operators as practical alternatives for scalable post-mask reconstruction where reproducibility, computational efficiency, and traceable behaviour are required.

The runtime results in Table IV reinforce this interpretation. Linear interpolation is fastest, but remains more sensitive to unfavourable temporal support. HCB is nearly as efficient while achieving significantly lower Relative MAE and SAM and substantially lower aggregate errors overall. HELIX requires more computation than HCB because semantic prototypes and scenario-dependent reconstruction steps are involved, but remains moderate compared to the learned baselines. SAITS and TimesNet require substantially longer processing times because scenario-specific training is included. For large EO archives, these differences are operationally relevant because repeated training, inference, and model selection increase computational cost and limit scalability. The results therefore support transparent, structurally constrained operators as practical alternatives for scalable post-mask reconstruction, particularly where reproducibility, auditability, and computational efficiency are required.

E. Accuracy–Efficiency Trade-offs

Reconstruction quality should not be considered independently of the computational effort required to obtain it. HCB and HELIX achieve competitive reconstruction accuracy without external supervised training or area-specific fine-tuning, with HCB additionally exhibiting a particularly low operational runtime. HCB requires no model fitting, whereas HELIX performs analytical target-series parameter estimation by fitting class-specific seasonal prototypes and associated statistics from the available observations. This distinction is relevant for repeated processing of large EO archives, where comparatively small differences in per-run computational demand can accumulate across extensive spatial and temporal workloads.

The reported runtimes quantify only the computation performed within the present benchmark and therefore do not represent the complete computational cost of all approaches. The evaluated methods differ fundamentally in their adaptation regime. SAITS [62] and TimesNet [63] require benchmark-specific model training, whereas MOMENT [64], U-TILISE [57], and UnCRtainTS [58] use previously pre-trained models and are evaluated here at the deployment stage. HCB requires neither learned model training nor target-series model fitting. HELIX likewise requires no external supervised training or pre-trained model weights, but estimates its analytical parameters directly from the target time series. Harmonic reconstruction similarly involves analytical parameter fitting rather than learned model training. These regimes are therefore not directly comparable through inference time alone.

This distinction is increasingly emphasised in machine learning and computational science, where predictive performance is considered alongside computational requirements, energy demand, and transparent reporting of resource use [14], [16], [84]. Similar considerations apply to EO processing, for which large-scale data handling and repeated computation can contribute non-negligible environmental impacts [15]. Energy consumption and carbon emissions were not measured directly; sustainability implications are therefore restricted to computational requirements and to the distinction between analytical, benchmark-trained, and pre-trained approaches.

U-TILISE provides an illustrative example of this distinction. The evaluated implementation uses released pre-trained weights based on the large-scale EarthNet2021 framework [65]; its reported benchmark runtime therefore characterises deployment rather than the upstream computation required to obtain the model. No corresponding estimate of training energy or carbon emissions is attempted. By contrast, HCB has no upstream model training requirement, while HELIX replaces learned model training with target-series-dependent analytical parameter estimation. Their computational advantage therefore extends beyond the measured reconstruction stage in terms of avoided supervised training and fine-tuning, although the associated energy or environmental benefit cannot be quantified from the present experiments.

F. Limitations and Operational Implications

Several limitations follow directly from the experimental design. First, both evaluation settings rely on controlled miss-

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited, content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409

ingness generated from complete reference time series. This is essential to the quantitative evaluation because naturally invalid observations do not provide known reference data. The spatially heterogeneous cloud/shadow experiment evaluates reconstruction fidelity when partial same-date spatial support remains available, whereas the complete-date scenarios provide a stringent temporal stress test by removing this support entirely. The former one is restricted to one AOI–year and one clustered synthetic mask configuration and therefore, does not reproduce the full morphology, detection uncertainty, or boundary effects of naturally occurring cloud and shadow masks.

Second, the empirical evaluation is based on Sentinel-2 L2A 120 m Mosaic data and four spectral bands [17]. The operator formulations are not restricted to this product, but the reported numerical performance is product-specific. Higher spatial resolution, red-edge bands, SWIR bands, SAR inputs, spectral indices, or biophysical variables may alter the relative behaviour of the methods. The visible–NIR four-band setting should therefore be interpreted as a controlled spectral core rather than a complete multispectral assessment on the one hand. On the other hand, this band combination complies with the aerial photos, which are becoming increasingly freely available as Open Data.

Third, the seven AOIs provide broad environmental coverage across the major Köppen climate groups and contrasting land-cover regimes, but one or a small number of AOIs per climatic setting and two observation years do not provide exhaustive global spatial or temporal replication. The results therefore support applicability across the environmental conditions represented in the benchmark rather than unrestricted geographic generalisation. Further investigations, could extent the spatial coverage of the AOIs e.g. to fit a regular pattern all over the Earth.

Fourth, HELIX depends on semantic information. The benchmark reduces mixed-pixel ambiguity by using majority-aggregated ESA WorldCover labels and retaining high-purity pixels [19]. However, land-cover maps can still be outdated, uncertain, or too coarse for specific applications. The controlled perturbation experiment indicates limited sensitivity to moderate random label corruption, but does not reproduce systematic temporal inconsistency between the semantic prior and the contemporary surface state. Semantic priors can stabilise reconstruction only when the label information remains sufficiently representative. HCB avoids this dependency and is therefore attractive for label-sparse, rapidly changing, or semantically uncertain settings.

Fifth, abrupt land-surface transitions can be retraced only to the extent reported by the remaining temporal observations. A transition occurring entirely within an observation gap cannot be uniquely identified or temporally located from the available support. HELIX can accommodate persistent observed deviations through its local residual component, but may additionally be affected by a static semantic prior that no longer represents the contemporary surface state. HCB avoids this semantic-prior mismatch because it does not use land-cover labels, but likewise cannot infer an otherwise unobserved transition. Conversely, once observations of a changed state become available, a marked deviation from the reconstructed temporally coherent trajectory can provide an indication of abrupt surface change or anomalous behaviour. Such reconstruction residuals may therefore support subsequent change or outlier screening, but this diagnostic use is not evaluated here and should not be interpreted as a validated change-detection capability.

Sixth, the benchmark does not isolate algorithmic design from adaptation strategy. HELIX performs analytical target-series parameter estimation from the available observations, whereas SAITS and TimesNet are trained separately for each AOI–year–scenario using target-area samples. MOMENT, U-TILISE, and UnCRtainTS are applied from released pre-trained weights without benchmark-specific fine-tuning. Learned approaches may therefore behave differently under alternative training regimes, auxiliary inputs, or target-specific adaptation. The results consequently establish comparative performance under the evaluated operational configurations rather than a universal ranking of reconstruction architectures independent of adaptation strategy.

Seventh, for operational EO data cubes, the main implication is that gap filling should not be evaluated only by aggregate radiometric error. Reconstruction quality also depends on whether spectral characteristics, spatial structure, and temporal dynamics are retained, and on the support and assumptions underlying the reconstructed value. A useful reconstruction method should therefore document how the value was reconstructed, which temporal and spatial support was available, whether semantic information was used, and how reliable the filled value is expected to be. HCB and HELIX are suitable building blocks for such traceable reconstruction because their assumptions are explicit and their outputs are deterministic for fixed inputs, masks, labels, and parameters. Future work should extend this framework with calibrated reliability indicators, provenance layers, and validated selectors between feature-side and label-side reconstruction.

G. Complementarity of Feature- and Label-Side Approach

The UMAP representation in Fig. 6 indicates that HCB and HELIX occupy similar regions of multi-metric performance space despite their fundamentally different reconstruction principles: feature-side reconstruction for HCB and label-side reconstruction for HELIX. HELIX provides the strongest aggregate radiometric accuracy, whereas HCB is preferable in a non-negligible subset of evaluation contexts and operates at substantially lower computational cost. This difference follows directly from their formulations. HCB reconstructs each target series from its observed spectro-temporal structure without ancillary semantic information, whereas HELIX additionally derives class-specific seasonal signatures from the available semantic classes. The latter can stabilise reconstruction when the semantic prior is representative, but also introduces sensitivity to label quality and temporal consistency. Their differing strengths therefore motivate the investigation of validated context-dependent selection or fusion strategies.

Figure 20 compares both methods at the level of shared AOI–year–scenario contexts for spectral fidelity. Each point

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409

represents one evaluation context, with HCB-SAM on the horizontal axis and HELIX-SAM on the vertical axis. Points below the one-to-one line indicate lower spectral-angle error for HELIX, whereas points above the line indicate lower error for HCB.

The distribution is asymmetric but not one-sided. For MedAE, HELIX yields the lower error in 70.5 % of contexts, HCB in 17.0 %, and 12.4 % are ties under the applied tolerance. For SAM, HELIX is lower in 62.4 % of contexts and HCB in 37.6 %. The primary AOI–year analysis does not identify a significant SAM difference between the two approaches after Holm correction, while the scenario-resolved comparison shows that HCB nevertheless provides lower spectral distortion in more than one-third of the evaluated contexts. At the same time, HCB requires approximately 1 s per benchmark execution compared with approximately 43 s for HELIX. HCB therefore retains spectral and computational advantages in specific contexts that are not reflected by aggregate radiometric rankings alone. The diagnostic best-of-two envelope is consequently interpreted only as an estimate of the potential benefit of a future validated context-selection strategy, not as an operational ensemble result.

This complementarity has a direct methodological basis. HELIX benefits when land-cover information is reliable and class-wise seasonal behaviour is representative for the target pixel. HCB does not use external semantics and can therefore be advantageous where the observed spectral-temporal structure is informative but semantic priors are less decisive or potentially imperfect. The two methods therefore represent different reconstruction principles: feature-side temporal structure and label-side semantic constraint. This supports a future combined framework, but only if the selector or fusion rule is validated explicitly and accompanied by uncertainty and provenance indicators.

This distinction is especially relevant under land-cover change or label uncertainty. If a pixel undergoes a real transition, for example from forest to built-up land, while the ancillary land-cover layer still represents the previous class, HELIX may regularise the reconstruction toward an outdated seasonal prototype. The effect is most critical during or shortly after transitions, when the observed pixel trajectory and the semantic prior are no longer consistent. HCB is less exposed to this failure mode because it does not use class labels, but it also cannot exploit class-specific seasonal constraints. The relative advantage of the two operators therefore depends on whether semantic information is reliable, temporally consistent, and representative of the target pixel.

H. Future Directions

The role of spectro-temporal variability remains only partly resolved. The results show class-, climate-, band-, and AOI-specific differences, but these effects are difficult to separate from reflectance magnitude, seasonal phase, and temporal gap geometry. Future work should therefore quantify spectral magnitude and temporal variability explicitly, for example by relating reconstruction error to class-wise seasonal amplitude, intra-annual variance, phenological transition rates, and abrupt surface transitions.

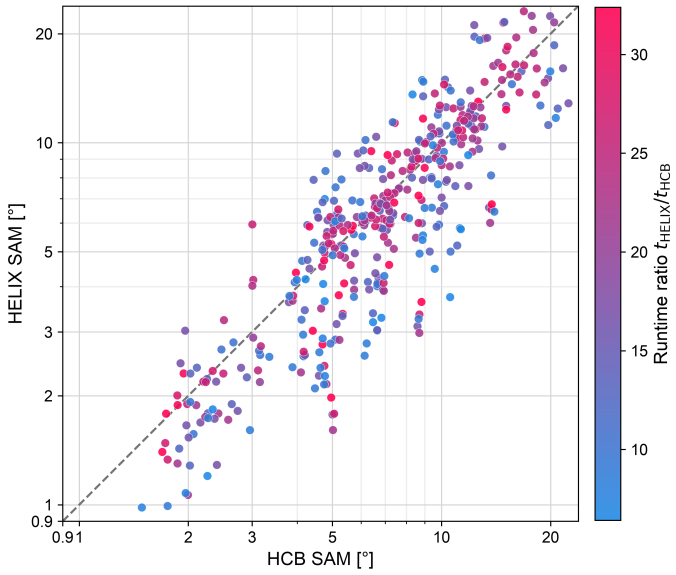


Fig. 20. Context-level complementarity between HCB and HELIX. Each point represents one shared AOI–year–scenario context, with HCB-SAM on the horizontal axis and HELIX-SAM on the vertical axis. Points below the one-to-one line indicate lower spectral-angle error for HELIX, whereas points above the line indicate lower error for HCB. Colour encodes the ratio of HELIX to HCB runtime and is shown as an additional computational descriptor rather than as an explanatory variable for reconstruction accuracy.

A second direction concerns more realistic spatially heterogeneous missingness. The controlled cloud/shadow experiment complements the complete-date benchmark by evaluating reconstruction when partial same-date spatial support remains available and demonstrates that structural fidelity cannot be inferred from radiometric error alone. However, the present experiment is restricted to one AOI–year and one clustered synthetic mask configuration. Future work should therefore evaluate naturally occurring cloud and shadow masks, cloud-edge effects, spatially varying mask uncertainty, and mixed missingness regimes across additional environmental settings.

A third direction is the extension beyond the visible–NIR spectral core evaluated in this benchmark. The underlying HCB formulation is not specific to optical reflectance and builds on a Hadamard/Kennaugh representation framework established in radar remote sensing, while the present work extends it to missingness-aware temporal reconstruction. Likewise, the HELIX operator is formulated for a temporally sampled variable through a class-specific seasonal expectation and local residuals rather than through optical-specific terms. The present experiments nevertheless validate the reconstruction performance only for visible and NIR reflectance. Red-edge bands, SWIR bands, SAR backscatter, spectral indices, and biophysical variables may exhibit different temporal dynamics, noise properties, and physical constraints and therefore require dedicated reconstruction experiments.

The complementarity between HCB and HELIX also motivates further methodological development. HELIX acts as a semantic reconstruction operator: it is stabilised by class-specific seasonal prototypes, but depends on the availability and reliability of ancillary land-cover information. HCB provides a complementary label-free operator that captures

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409

temporal structure directly from the observed reflectance sequence and is therefore attractive for large-scale or label-sparse settings. A validated context selector or combined feature–semantic operator could exploit both principles while preserving interpretability and deterministic behaviour. One feasible solution would be the transition from external semantic classes to internal syntax classes in order to prevent any deterioration caused by external factors.

A further direction is task-level validation. The present evaluation characterises reconstruction fidelity through radiometric and robust errors, spectral angle, spatial- and temporal-gradient behaviour, and computational requirements. Future studies should assess how these reconstruction differences propagate into downstream workflows such as vegetation monitoring, land-cover classification, change detection, and biophysical parameter retrieval. Such evaluations should account for the target variable, temporal aggregation, model type, and the sensitivity of each task to radiometric bias, spectral distortion, and altered spatial or temporal structure.

The self-correction capability of transparent reconstruction operators also deserves further investigation. In the present benchmark, observed values are preserved by design so that all methods operate under a strict post-mask protocol. In operational settings, however, masks may be incomplete: cloud-contaminated, shadow-affected, or otherwise invalid observations can remain marked as usable. A controlled extension could therefore test whether reconstruction residuals, support inconsistency, or prototype deviations can flag and correct such residual mask errors. This would move from gap filling alone toward joint reconstruction and quality refinement, while requiring careful safeguards to avoid replacing valid but unusual surface observations.

Finally, future reconstruction systems should include explicit reliability and provenance as embeddings. Each reconstructed value should document the available temporal support, gap position, reconstruction source, semantic-prior use, and expected uncertainty. This information would support operational quality control and downstream use in classification, change detection, and environmental monitoring. That kind of information would be a prerequisite for future scaling of transparent post-mask reconstruction from controlled benchmarks to national, continental, and global EO data cubes.

VI. Conclusion

Complete and temporally consistent satellite image time series are essential for environmental monitoring, modelling, and downstream data-driven analysis. This study addressed optical Earth observation gap filling as a post-mask reconstruction problem and introduced two transparent reconstruction approaches: the feature-side Hypercomplex Bases approach (HCB) and the label-side HELIX approach. Both were evaluated against Linear and Harmonic reconstruction, SAITS, TimesNet, MOMENT, U-TILISE, and UnCRtainTS under two complementary experimental settings: spatially heterogeneous cloud/shadow missingness and a large-scale complete-date benchmark. The latter comprises 30 controlled missingness scenarios derived from complete Sentinel-2 L2A 120 m Mosaic time series covering seven AOIs across all five Köppen–Geiger main climate groups and all eleven ESA WorldCover classes [19].

The results show that reconstruction difficulty is not governed by gap fraction alone. Temporal gap distribution, spectral band, regional surface regime, and land-cover context jointly control the error structure. Edge gaps and temporally clustered gaps are more difficult than gaps with surrounding temporal support, while the NIR band and dynamic snow-, vegetation-, and water-related regimes show stronger deviations. The spatially heterogeneous cloud/shadow experiment further demonstrates that radiometric error alone does not fully characterise reconstruction fidelity: spatial- and temporal-gradient deviations reveal smoothing, altered local structure, and temporally damped or otherwise inconsistent reconstructions that may remain comparatively inconspicuous under point-wise error measures. Reconstruction reliability therefore depends not only on absolute radiometric agreement, but also on the preservation of spectral, spatial, and temporal structure.

Across the complete-date benchmark, HCB and HELIX form the leading accuracy group while requiring neither external supervised training nor area-specific fine-tuning. HELIX achieves the lowest aggregate MAE (688.4 DN), MedAE (188 DN), RMSE, and mean SAM (7.38°). HCB follows with the second-lowest aggregate MAE (879.4 DN), a MedAE of 262 DN, and a similarly low mean SAM of 7.88°, while requiring only approximately 1 s mean recorded runtime and no ancillary semantic information. The rank-based and pairwise statistical analyses reinforce this joint position: HCB and HELIX consistently occupy the leading ranks across the primary reconstruction metrics and show statistically significant advantages over several classical and learned baselines. Within this leading pair, HELIX provides the stronger radiometric reconstruction, with significantly lower MAE, MedAE, and Relative MAE than HCB, whereas their SAM difference is not significant after correction, confirming the similarly strong spectral fidelity of both approaches. Their performance is not restricted to interpolation-like gaps between valid observations, but remains stable under temporally asymmetric, extrapolation-like configurations in which temporal support is available only from one side. HCB therefore provides a particularly favourable computationally parsimonious and label-independent operating point, whereas HELIX complements it through semantic seasonal guidance that prioritises radiometric reconstruction accuracy. The two approaches consequently represent complementary rather than interchangeable reconstruction strategies.

The formulations of HCB and HELIX are not restricted to visible–NIR reflectance. HCB operates directly on the observed spectro-temporal signal through a missingness-aware representation and requires neither ancillary semantic information nor target-area parameter tuning. HELIX combines class-specific seasonal expectations with local temporal residual information and thereby introduces an additional semantic constraint when suitable land-cover information is available. Their mathematical formulations hence permit the transfer to other temporally sampled EO variables, although the present experiments validate reconstruction performance only for

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409

Sentinel-2 visible and NIR reflectance. A dedicated evaluation still has to be done for red-edge and SWIR bands, SAR backscatter, spectral indices, biophysical variables, and other sensor modalities.

The clustered cloud/shadow experiment shows that the favourable behaviour of HCB and HELIX extends beyond complete-date missingness and also highlights the importance of evaluating spatial and temporal structure in addition to conventional radiometric errors. Future work should extend the controlled spatial-mask analysis to naturally occurring cloud and shadow masks across additional regions and years, investigate validated feature–semantic selection or fusion strategies, and incorporate explicit reliability and provenance information for reconstructed observations. These developments would further support scalable, transparent, and traceable post-mask reconstruction in operational EO data cubes.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited, content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409

APPENDIX
HCB ESTIMATION IN THE PRESENCE OF GAPS

The following equations derive the concrete formulas enabling the estimation of hypercomplex bases from sparse time series data by a H^4 transform without limiting its general applicability. Eq. 19 shows the gap-free case as already stated in Eq.6 in the methodology section.

$$K = H_s \cdot R = \frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \frac{1}{2}\begin{bmatrix} a+b+c+d \\ a+b-c-d \\ a-b-c+d \\ a-b+c-d \end{bmatrix} = \begin{bmatrix} 2 \cdot \mathbb{E}(a,b,c,d) \\ \mathbb{E}(a,b) - \mathbb{E}(c,d) \\ \mathbb{E}(a,d) - \mathbb{E}(b,c) \\ \mathbb{E}(a,c) - \mathbb{E}(b,d) \end{bmatrix} \tag{19}$$

With one gap, the expectation values refer to a reduced sample size. The following Eq. 20 describes the four possible cases with a, b, c, or d missing. As the four K_s are estimated from only three measurements, interdependency is unavoidable.

$$\begin{aligned} a = \emptyset \rightarrow K &= \begin{bmatrix} 2 \cdot \mathbb{E}(b,c,d) \\ \mathbb{E}(b) - \mathbb{E}(c,d) \\ \mathbb{E}(d) - \mathbb{E}(b,c) \\ \mathbb{E}(c) - \mathbb{E}(b,d) \end{bmatrix} = \begin{bmatrix} \frac{2}{3}(b+c+d) \\ b - \frac{c+d}{2} \\ c - \frac{b+c}{2} \\ d - \frac{b+d}{2} \end{bmatrix}, \quad b = \emptyset \rightarrow K = \begin{bmatrix} 2 \cdot \mathbb{E}(a,c,d) \\ \mathbb{E}(a) - \mathbb{E}(c,d) \\ \mathbb{E}(a,d) - \mathbb{E}(c) \\ \mathbb{E}(a,c) - \mathbb{E}(d) \end{bmatrix} = \begin{bmatrix} \frac{2}{3}(a+c+d) \\ a - \frac{c+d}{2} \\ \frac{a+d}{2} - c \\ \frac{a+c}{2} - d \end{bmatrix} \\ c = \emptyset \rightarrow K &= \begin{bmatrix} 2 \cdot \mathbb{E}(a,b,d) \\ \mathbb{E}(a,b) - \mathbb{E}(d) \\ \mathbb{E}(a,d) - \mathbb{E}(b) \\ \mathbb{E}(a) - \mathbb{E}(b,d) \end{bmatrix} = \begin{bmatrix} \frac{2}{3}(a+b+d) \\ \frac{a+b}{2} - d \\ \frac{a+d}{2} - b \\ a - \frac{b+d}{2} \end{bmatrix}, \quad d = \emptyset \rightarrow K = \begin{bmatrix} 2 \cdot \mathbb{E}(a,b,c) \\ \mathbb{E}(a,b) - \mathbb{E}(d) \\ \mathbb{E}(a) - \mathbb{E}(b,c) \\ \mathbb{E}(a,c) - \mathbb{E}(b) \end{bmatrix} = \begin{bmatrix} \frac{2}{3}(a+b+c) \\ \frac{a+d}{2} - d \\ a - \frac{b+c}{2} \\ \frac{a+c}{2} - b \end{bmatrix} \end{aligned} \tag{20}$$

The identical approach also applies in the presence of two gaps. Two cases are distinguished in the following: two consecutive gaps in Eq. 21 and two separated gaps with one measurement in between in Eq. 22. The linear dependence between the estimated K_s now becomes obvious in identical or mirrored entries.

$$\begin{matrix} a = \emptyset \\ b = \emptyset \end{matrix} \rightarrow \begin{bmatrix} 2 \cdot \mathbb{E}(c,d) \\ \emptyset - \mathbb{E}(c,d) \\ \mathbb{E}(d) - \mathbb{E}(c) \\ \mathbb{E}(c) - \mathbb{E}(d) \end{bmatrix} = \begin{bmatrix} c+d \\ 0 \\ d-c \\ c-d \end{bmatrix}, \begin{matrix} b = \emptyset \\ c = \emptyset \end{matrix} \rightarrow \begin{bmatrix} 2 \cdot \mathbb{E}(a,d) \\ \mathbb{E}(a) - \mathbb{E}(d) \\ \mathbb{E}(a,d) - \emptyset \\ \mathbb{E}(a) - \mathbb{E}(d) \end{bmatrix} = \begin{bmatrix} a+d \\ a-d \\ 0 \\ a-d \end{bmatrix}, \begin{matrix} c = \emptyset \\ d = \emptyset \end{matrix} \rightarrow \begin{bmatrix} 2 \cdot \mathbb{E}(a,b) \\ \mathbb{E}(a,b) - \emptyset \\ \mathbb{E}(a) - \mathbb{E}(b) \\ \mathbb{E}(a) - \mathbb{E}(b) \end{bmatrix} = \begin{bmatrix} a+b \\ 0 \\ a-b \\ a-b \end{bmatrix} \tag{21}$$

$$\begin{matrix} a = \emptyset \\ c = \emptyset \end{matrix} \rightarrow \begin{bmatrix} 2 \cdot \mathbb{E}(b,d) \\ \mathbb{E}(b) - \mathbb{E}(d) \\ \mathbb{E}(d) - \mathbb{E}(b) \\ \emptyset - \mathbb{E}(b,d) \end{bmatrix} = \begin{bmatrix} b+d \\ b-d \\ d-b \\ 0 \end{bmatrix}, \begin{matrix} b = \emptyset \\ d = \emptyset \end{matrix} \rightarrow \begin{bmatrix} 2 \cdot \mathbb{E}(a,c) \\ \mathbb{E}(a) - \mathbb{E}(c) \\ \mathbb{E}(a) - \mathbb{E}(c) \\ \mathbb{E}(a) - \emptyset \end{bmatrix} = \begin{bmatrix} a+c \\ a-c \\ a-c \\ 0 \end{bmatrix}, \begin{matrix} a = \emptyset \\ d = \emptyset \end{matrix} \rightarrow \begin{bmatrix} 2 \cdot \mathbb{E}(b,c) \\ \mathbb{E}(b) - \mathbb{E}(c) \\ \emptyset - \mathbb{E}(b,c) \\ \mathbb{E}(c) - \mathbb{E}(b) \end{bmatrix} = \begin{bmatrix} b+c \\ b-c \\ 0 \\ c-b \end{bmatrix} \tag{22}$$

Even the extreme case involving three gaps is covered by the identical approach. According to Eq.23, only one available measurement allows for the estimation of only one absolute value without any variation, which is reasonable.

$$\begin{matrix} a = \emptyset \\ b = \emptyset \\ c = \emptyset \end{matrix} \rightarrow K = \begin{bmatrix} 2 \cdot \mathbb{E}(d) \\ \emptyset - \mathbb{E}(d) \\ \mathbb{E}(d) - \emptyset \\ \emptyset - \mathbb{E}(d) \end{bmatrix} = \begin{bmatrix} 2d \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{matrix} a = \emptyset \\ b = \emptyset \\ d = \emptyset \end{matrix} \rightarrow K = \begin{bmatrix} 2c \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{matrix} a = \emptyset \\ c = \emptyset \\ d = \emptyset \end{matrix} \rightarrow K = \begin{bmatrix} 2b \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{matrix} b = \emptyset \\ c = \emptyset \\ d = \emptyset \end{matrix} \rightarrow K = \begin{bmatrix} 2a \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{23}$$

APPENDIX
RECONSTRUCTION FROM ESTIMATED HCBS

The estimated HCBs lack independence because they are calculated from a lower number of reflectance values due to gaps. In consequence a complete reconstruction would treat gaps as zero-measurements and thus, reconstruct the value 0. Therefore,

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited, content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409

the reconstruction is based on the multi-scale multi-looking approach [69], i.e., it starts with the coarsest scale and consecutively adds finer scales (Eq. 24) until the optimal representation in the appropriate scale is reached.

$$
\begin{aligned}
R_0^* &= H_0 \cdot K = \frac{1}{2}\begin{bmatrix}1\\1\\1\\1\end{bmatrix} \cdot \begin{bmatrix}K_0\end{bmatrix} = \frac{1}{2}\begin{bmatrix}K_0\\K_0\\K_0\\K_0\end{bmatrix} \\
R_1^* &= H_1 \cdot K = \frac{1}{2}\begin{bmatrix}1 & 1\\1 & 1\\1 & -1\\1 & -1\end{bmatrix} \cdot \begin{bmatrix}K_0\\K_1\end{bmatrix} = \frac{1}{2}\begin{bmatrix}K_0+K_1\\K_0+K_1\\K_0-K_1\\K_0-K_1\end{bmatrix} \\
R_2^* &= H_2 \cdot K = \frac{1}{2}\begin{bmatrix}1 & 1 & 1\\1 & 1 & -1\\1 & -1 & -1\\1 & -1 & 1\end{bmatrix} \cdot \begin{bmatrix}K_0\\K_1\\K_2\end{bmatrix} = \frac{1}{2}\begin{bmatrix}K_0+K_1+K_2\\K_0+K_1-K_2\\K_0-K_1-K_2\\K_0-K_1+K_2\end{bmatrix} \\
R_3^* &= H_3 \cdot K = \frac{1}{2}\begin{bmatrix}1 & 1 & 1 & 1\\1 & 1 & -1 & -1\\1 & -1 & -1 & 1\\1 & -1 & 1 & -1\end{bmatrix} \cdot \begin{bmatrix}K_0\\K_1\\K_2\\K_3\end{bmatrix} = \frac{1}{2}\begin{bmatrix}K_0+K_1+K_2+K_3\\K_0+K_1-K_2-K_3\\K_0-K_1-K_2+K_3\\K_0-K_1+K_2-K_3\end{bmatrix}
\end{aligned}
\tag{24}
$$

The appropriate scale is determined by the cost function $\langle |R_t - R_t^*| \rangle_{t_m}$ as defined in Eq. 8. Of course, the mean absolute error can only be evaluated for the measured values although the resulting appropriate scale depends both on the gaps and the signal shape. For easier understanding, Fig. 21 refers to the case of no gaps and shows the appropriate scale just in dependence on the signal itself neglecting possible gaps. If the measurements are almost equal, the reconstruction on scale 0, i.e., a constant value is already sufficient (Fig. 21a). With one step in the middle of the time series (Fig. 21b+c), scale 1 is appropriate. Scale 2 optimally describes a signal with a curvature as depicted in Fig. 21d+e+f. High frequency contributions (Fig. 21g+h+f) require a full reconstruction until scale 3. Thanks to the multi-scale multi-looking approach, this high frequency and therewith, rather unstable Ks are only evaluated if a sufficient number of measurements is available.

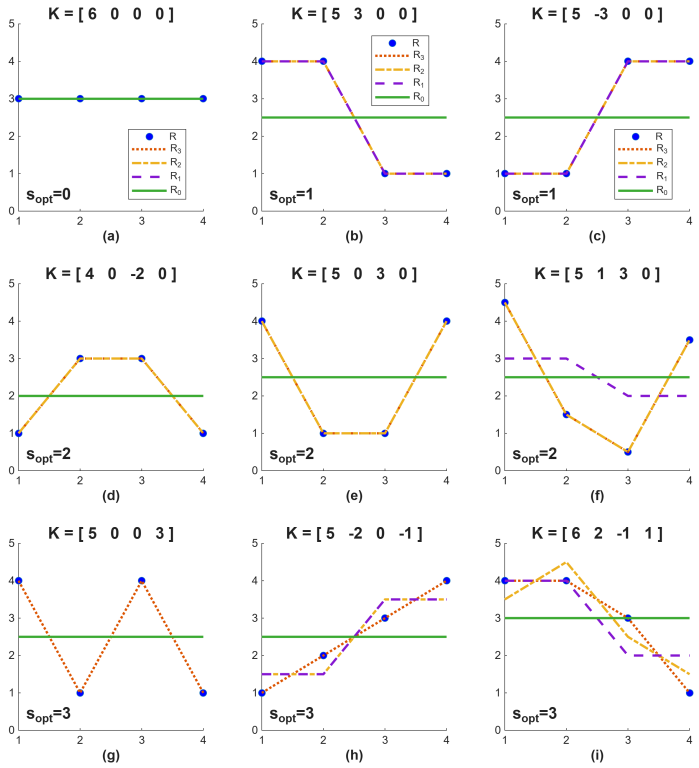


Fig. 21. Time series reconstruction from estimated hypercomplex bases. The original measurements R are plotted as blue points, whereas the reconstructed measurements R_0, R_1, R_2, and R_3 are plotted as lines. The number indicates the scale of the reconstruction according to Eq. 24. The respective values of K are given in the title of each plot. It is easy to see that the sign causes a reflection along a horizontal axis. The optimal scale of the reconstruction is shown on the bottom left of each plot. Obviously, the optimal scale depends on the signal shape. In the presence of gaps, the cost function (Eq. 8) controls the optimal fit.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409

APPENDIX
BASELINE IMPLEMENTATION PROTOCOL

The detailed baseline training, adaptation, masking, and computational settings are summarised in Table VIII. All benchmark approaches operate on the same 32-date acquisition grid and identical scenario-specific missing-date masks. Across the 30 controlled scenarios, withheld dates are systematically distributed over the temporal sequence, and evaluation is performed on the same withheld observations.

TABLE VIII
DETAILED TRAINING, ADAPTATION, MASKING, AND IMPLEMENTATION PROTOCOL FOR THE LEARNED AND PRE-TRAINED BENCHMARK APPROACHES.

Method	Training / adaptation	Input / masking / scaling	Configuration	Execution
SAITS	Trained from scratch separately for each AOI–year–scenario. Up to 200,000 eligible target-area pixels are sampled without replacement using seed 123; the effective number is limited by the available spatial samples. No separate benchmark validation subset is supplied to the fitting routine.	Pixel-wise 32×4 sequences (B02, B03, B04, B08). Scenario-withheld dates are set to missing before fitting and inference. Reflectance is divided by 10 000 before model input and returned to the original scale after reconstruction. Target-date reflectances are unavailable during fitting.	PyPOTS SAITS; 2 layers; $d_{\text{model}} = 128$; 4 attention heads; $d_k = d_v = 32$; $d_{\text{ffn}} = 256$; dropout $= 0$; attention dropout $= 0$; diagonal attention mask enabled; batch size 1024; maximum 30 epochs; patience 5.	RTX A6000. Training and inference per AOI–year–scenario. Mean per-scenario training/inference: 252.0/5.1 s.
TimesNet	Trained from scratch separately for each AOI–year–scenario. 60,000 eligible target-area pixels are sampled without replacement using seed 123. No separate benchmark validation subset is supplied to the fitting routine.	Pixel-wise 32×4 sequences. Scenario-withheld dates are set to missing before fitting and inference. Reflectance is divided by 10 000 before model input and returned to the original scale after reconstruction. Target-date reflectances are unavailable during fitting.	PyPOTS TimesNet; 2 layers; top-$k = 3$; $d_{\text{model}} = 64$; $d_{\text{ffn}} = 128$; 4 kernels; dropout $= 0$; non-stationary normalisation disabled; batch size 512; maximum 12 epochs; patience 4.	RTX A6000. Training and inference per AOI–year–scenario. Mean per-scenario training/inference: 143.6/5.6 s.
MOMENT	Released `AutonLab/MOMENT-1-base` pre-trained weights; reconstruction task; no benchmark-specific training or fine-tuning.	Each spectral band is processed as a separate univariate sequence with the common 32-date scenario mask. Missing dates are represented through the temporal input mask. No external reflectance rescaling is applied; outputs are returned in the input scale.	MOMENT reconstruction pipeline with one input channel; patch length 8; stride 8; sequence padding to the required patch structure; pre-fill denoising disabled.	RTX A6000; inference only.
U-TILISE	Released EarthNet2021 checkpoint; no training or fine-tuning on the seven benchmark AOIs.	B02/B03/B04/B08 input; reflectance divided by 10 000 and clipped to $[0, 1]$. Entire withheld acquisitions are masked. Day-of-year information is retained. Processing uses 128×128 tiles; predictions are rescaled to the original reflectance scale.	Official repository implementation; EarthNet2021 four-channel "bgr-nir" configuration; temporal window 10; released `utilise_earthnet2021.pth` checkpoint.	CPU, 16 threads; inference only.
UnCRtainTS	Released optical no-SAR checkpoint and configuration; no training or fine-tuning on the benchmark AOIs.	Reflectance divided by 10 000 and clipped to $[0, 1]$. B02/B03/B04/B08 are inserted into the corresponding positions of the 13-band optical input; unused channels are zero-filled. Three nearest available observations provide temporal context. Processing uses 128×128 tiles.	Official repository implementation; optical no-SAR checkpoint; temporal context length 3; batch size 1; uncertainty output excluded from the primary accuracy comparison.	CPU, 16 threads; inference only.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409

REFERENCES

[1] M. Gašparović, I. Pilaš, D. Radočaj, and D. Dobrinić, "Monitoring and prediction of land surface phenology using satellite earth observations—a brief review," *Applied Sciences*, vol. 14, no. 24, p. 12020, 2024.

[2] J. Transon, R. D'Andrimont, A. Maugnard, and P. Defourny, "Survey of hyperspectral earth observation applications from space in the sentinel-2 context," *Remote Sensing*, vol. 10, no. 2, p. 157, 2018.

[3] Z. Mitraka and N. Chrysoulakis, "Earth observation for urban climate monitoring: Surface cover and land surface temperature," in *Multi-purposeful Application of Geospatial Data*, R. B. Rustamov, S. Hasanova, and M. H. Zeynalova, Eds. Rijeka: IntechOpen, 2018, ch. 7, p. 1.

[4] D. Frantz, E. Haß, A. Uhl, J. Stoffels, and J. Hill, "Improvement of the fmask algorithm for sentinel-2 images: Separating clouds from bright surfaces based on parallax effects," *Remote Sensing of Environment*, vol. 215, pp. 471–481, 2018.

[5] M. Sudmanns, D. Tiede, H. Augustin, and S. Lang, "Assessing global sentinel-2 coverage dynamics and data availability for operational earth observation (eo) applications using the eo-compass," *International Journal of Digital Earth*, vol. 13, no. 7, pp. 768–784, 2020.

[6] J. Li and D. P. Roy, "A global analysis of sentinel-2a, sentinel-2b and landsat-8 data revisit intervals and implications for terrestrial monitoring," *Remote Sensing*, vol. 9, no. 9, p. 902, 2017.

[7] R. Fischer, N. Piatkowski, C. Pelletier, G. I. Webb, F. Petitjean, and K. Morik, "No cloud on the horizon: Probabilistic gap filling in satellite image series," in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, 2020, pp. 546–555.

[8] J. L. Dwyer, D. P. Roy, B. Sauer, C. B. Jenkerson, H. K. Zhang, and L. Lymburner, "Analysis ready data: Enabling analysis of the landsat archive," *Remote Sensing*, vol. 10, no. 9, p. 1363, 2018.

[9] Z. Zhu, "Science of landsat analysis ready data," *Remote Sensing*, vol. 11, no. 18, p. 2166, 2019.

[10] A. Siqueira, A. Lewis, M. Thankappan, Z. Szantoi, P. Goryl, S. Labahn, J. Ross, S. Hosford, S. Mecklenburg, T. Tadono, A. Rosenqvist, and J. Lacey, "Ceos analysis ready data for land – an overview on the current and future work," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019, pp. 5536–5537.

[11] M. Claverie, J. Ju, J. G. Masek, J. L. Dungan, E. F. Vermote, J.-C. Roger, S. V. Skakun, and C. Justice, "The harmonized landsat and sentinel-2 surface reflectance data set," *Remote Sensing of Environment*, vol. 219, pp. 145–161, 2018.

[12] J. Ju, Q. Zhou, B. Freitag, D. Roy, H. Zhang, M. Sridhar, J. Mandel, S. Arab, G. Schmidt, C. Crawford, F. Gascon, P. Strobl, J. Masek, and C. Neigh, "The harmonized landsat and sentinel-2 version 2.0 surface reflectance dataset," *Remote Sensing of Environment*, vol. 324, p. 114723, 07 2025.

[13] D. P. Roy, H. Huang, R. Houborg, and V. S. Martins, "A global analysis of the temporal availability of planetscope high spatial resolution multi-spectral imagery," *Remote Sensing of Environment*, vol. 264, p. 112586, 2021.

[14] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green ai," *Commun. ACM*, vol. 63, no. 12, p. 54–63, Nov. 2020. [Online]. Available: https://doi.org/10.1145/3381831

[15] R. Wilkinson, M. Mleczko, R. Brewin, K. Gaston, M. Mueller, J. Shutler, X. Yan, and K. Anderson, "Environmental impacts of earth observation data in the constellation and cloud computing era," *Science of The Total Environment*, vol. 909, p. 168584, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0048969723072121

[16] L. Lannelongue, H. E. G. Aronson, A. Bateman *et al.*, "GREENER principles for environmentally sustainable computational science," *Nature Computational Science*, vol. 3, pp. 514–521, 2023. [Online]. Available: https://doi.org/10.1038/s43588-023-00461-y

[17] European Space Agency, "Sentinel-2 L2A 120m Mosaic," Registry of Open Data on AWS, 2026, processed by Sinergise; contains modified Copernicus Sentinel data. Accessed 27 February 2026. [Online]. Available: https://registry.opendata.aws/sentinel-s2-l2a-mosaic-120

[18] H. E. Beck, T. R. McVicar, N. Vergopolan, A. Berg, N. J. Lutsko, A. Dufour, Z. Zeng, X. Jiang, A. I. J. M. van Dijk, and D. G. Miralles, "High-resolution (1 km) Köppen-Geiger maps for 1901–2099 based on constrained CMIP6 projections," *Scientific Data*, vol. 10, p. 724, 2023.

[19] D. Zanaga, R. Van De Kerchove, D. Daems, W. De Keersmaecker, C. Brockmann, G. Kirches, J. Wevers, O. Cartus, M. Santoro, S. Fritz, M. Lesiv, M. Herold, N.-E. Tsendbazar, P. Xu, F. Ramoino, and O. Arino, "Esa worldcover 10 m 2020 v100," 10 2020.

[20] S. Kandasamy, F. Baret, A. Verger, P. Neveux, and M. Weiss, "A comparison of methods for smoothing and gap filling time series of remote sensing observations – application to modis lai products," *Biogeosciences*, vol. 10, no. 6, pp. 4055–4071, 2013.

[21] Y. Julien and J. A. Sobrino, "Optimizing and comparing gap-filling techniques using simulated ndvi time series from remotely sensed global data," *International Journal of Applied Earth Observation and Geoinformation*, vol. 76, pp. 93–111, 2019.

[22] J. P. Musial, M. M. Verstraete, and N. Gobron, "Technical note: Comparing the effectiveness of recent algorithms to fill and smooth incomplete and noisy time series," *Atmospheric Chemistry and Physics*, vol. 11, no. 15, pp. 7905–7923, 2011.

[23] P. Jönsson and L. Eklundh, "Timesat—a program for analyzing time-series of satellite sensor data," *Computers & geosciences*, vol. 30, no. 8, pp. 833–845, 2004.

[24] R. E. Rossi, J. L. Dungan, and L. R. Beck, "Kriging in the shadows: geostatistical interpolation for remote sensing," *Remote Sensing of Environment*, vol. 49, no. 1, pp. 32–40, 1994.

[25] C. Zhang, W. Li, and D. J. Travis, "Restoration of clouded pixels in multispectral remotely sensed imagery with cokriging," *International Journal of Remote Sensing*, vol. 30, no. 9, pp. 2173–2195, 2009.

[26] F. Bédard, G. Reichert, R. Dobbins, and I. Trépanier, "Evaluation of segment-based gap-filled landsat etm+ slc-off satellite data for land cover classification in southern saskatchewan, canada," *International Journal of Remote Sensing*, vol. 29, no. 7, pp. 2041–2054, 2008.

[27] S. Maxwell, G. L. Schmidt, and J. C. Storey, "A multi-scale segmentation approach to filling gaps in landsat etm+ slc-off images," *International Journal of Remote Sensing*, vol. 28, no. 23, pp. 5339–5356, 2007.

[28] J. Chen, X. Zhu, J. E. Vogelmann, F. Gao, and S. Jin, "A simple and effective method for filling gaps in landsat etm+ slc-off images," *Remote sensing of environment*, vol. 115, no. 4, pp. 1053–1064, 2011.

[29] L. Zhang, W. Wei, Q. Shi, C. Shen, A. van den Hengel, and Y. Zhang, "Accurate tensor completion via adaptive low-rank representation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 10, pp. 4170–4184, 2020.

[30] Y. Zhao and T. Cui, "Tensor-train methods for sequential state and parameter learning in state-space models," 2024. [Online]. Available: https://arxiv.org/abs/2301.09891

[31] J. Wang, C. K. Lee, X. Zhu, R. Cao, Y. Gu, S. Wu, and J. Wu, "A new object-class based gap-filling method for planetscope satellite image time series," *Remote Sensing of Environment*, vol. 280, p. 113136, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425722002504

[32] H. Guo, Y. Zheng, H. Xu, and L. Bruzzone, "Clear: A fast cloud removal method for optical remote sensing images combining class-based linear regression and iterative residual compensation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 64, pp. 1–16, 2026.

[33] P. Mi, C. Zheng, L. Jia, and Y. Bai, "Reconstruction of global long-term gap-free daily surface soil moisture from 2002 to 2020 based on a pixel-wise machine learning method," *Remote Sensing*, vol. 15, no. 8, p. 2116, 2023.

[34] B. van Jaarsveld, S. M. Hauswirth, and N. Wanders, "Machine learning and global vegetation: random forests for downscaling and gap filling," *Hydrology and Earth System Sciences*, vol. 28, no. 11, pp. 2357–2374, 2024.

[35] Q. Wang, L. Wang, X. Zhu, Y. Ge, X. Tong, and P. M. Atkinson, "Remote sensing image gap filling based on spatial-spectral random forests," *Science of Remote Sensing*, vol. 5, p. 100048, 2022.

[36] H. Su, V. Jampani, D. Sun, O. Gallo, E. Learned-Miller, and J. Kautz, "Pixel-adaptive convolutional neural networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 11 158–11 167.

[37] Y. Chen, Q. Weng, L. Tang, X. Zhang, M. Bilal, and Q. Li, "Thick clouds removing from multitemporal landsat images using spatiotemporal neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[38] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[39] Y. Wang, X. Zhou, Z. Ao, K. Xiao, C. Yan, and Q. Xin, "Gap-filling and missing information recovery for time series of modis data using deep learning-based methods," *Remote Sensing*, vol. 14, no. 19, p. 4692, 2022.

[40] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of the 28th International Conference on Neural Information*

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409

Processing Systems - Volume 2, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, p. 2672–2680.

[41] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2536–2544.

[42] Y.-S. Chang, H.-T. Chiao, S. Abimannan, Y.-P. Huang, Y.-T. Tsai, and K.-M. Lin, "An lstm-based aggregated model for air pollution forecasting," *Atmospheric Pollution Research*, vol. 11, no. 8, pp. 1451–1463, 2020.

[43] Y. Zhou, X. Yang, L. Feng, W. Wu, T. Wu, J. Luo, X. Zhou, and X. Zhang, "Superpixel-based time-series reconstruction for optical images incorporating sar data using autoencoder networks," *GIScience & Remote Sensing*, vol. 57, no. 8, pp. 1005–1025, 2020.

[44] Y. Zhou, S. Wang, T. Wu, L. Feng, W. Wu, J. Luo, X. Zhang, and N. Yan, "For-backward lstm-based missing data reconstruction for time-series landsat images," *GIScience & Remote Sensing*, vol. 59, no. 1, pp. 410–430, 2022.

[45] M. Appel, "Efficient data-driven gap filling of satellite image time series using deep neural networks with partial convolutions," *Artificial Intelligence for the Earth Systems*, vol. 3, no. 2, p. 220055, 2024.

[46] Y. Chen, S. Feng, C. Zhao, N. Su, W. Li, R. Tao, and J. Ren, "High-resolution remote sensing image change detection based on fourier feature interaction and multiscale perception," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024.

[47] Y. Tang, S. Feng, C. Zhao, Y. Fan, Q. Shi, W. Li, and R. Tao, "An object fine-grained change detection method based on frequency decoupling interaction for high-resolution remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2024.

[48] Y. Chen, S. Feng, C. Zhao, Y. Tang, W. Li, and R. Tao, "A prototype-aware learning and dual-view regularization network for weakly supervised change detection in vhr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–15, 2025.

[49] W. Li, Y. Li, and J. C.-W. Chan, "Thick cloud removal with optical and sar imagery via convolutional-mapping-deconvolutional network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 4, pp. 2865–2879, 2020.

[50] Y. Zhao, S. Shen, J. Hu, Y. Li, and J. Pan, "Cloud removal using multimodal gan with adversarial consistency loss," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[51] A. Meraner, P. Ebel, M. Schmitt, and X. X. Zhu, "Cloud removal in sentinel-2 imagery using a deep residual neural network and sar-optical data fusion," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 333–346, Juli 2020.

[52] J. Li, C. Li, W. Xu, H. Feng, F. Zhao, H. Long, Y. Meng, W. Chen, H. Yang, and G. Yang, "Fusion of optical and sar images based on deep learning to reconstruct vegetation ndvi time series in cloud-prone regions," *International Journal of Applied Earth Observation and Geoinformation*, vol. 112, p. 102818, 2022.

[53] Y. Chen, R. Cao, S. Liu, L. Peng, X. Chen, and J. Chen, "A new deep learning-based model for reconstructing high-quality ndvi time-series data in heavily cloudy areas: fusion of sentinel 1 and 2 data," *International Journal of Digital Earth*, vol. 17, no. 1, p. e2407941, 2024.

[54] I. Tsardanidis, A. Koukos, V. Sitokonstantinou, T. Drivas, and C. Kontoes, "Cloud gap-filling with deep learning for improved grassland monitoring," *Computers and Electronics in Agriculture*, vol. 230, p. 109732, 2025.

[55] S. Liu, H. Chen, K. Tang, Y. Chen, H. Shu, T. Zan, Y. Xue, and J. Chen, "Innovative sar-optical data fusion for reflectance time series reconstruction in vegetation-covered regions," *International Journal of Applied Earth Observation and Geoinformation*, vol. 140, p. 104567, 2025.

[56] J. Li, D. Ienco, R. Gaetano, M. Ciotola, B. Mateo-Herrera, Z. Xiao, and M. Weiss, "A novel deep learning framework combining optical-sar data to generate gap-free 20 m lai time series for crop monitoring," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2026.

[57] C. Stucker, V. S. F. Garnot, and K. Schindler, "U-TILISE: A sequence-to-sequence model for cloud removal in optical satellite time series," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, 2023.

[58] P. Ebel, V. S. F. Garnot, M. Schmitt, J. Wegner, and X. X. Zhu, "UnCRtainTS: Uncertainty Quantification for Cloud Removal in Optical Satellite Time Series," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2023. [Online]. Available: "https://openaccess.thecvf.com/content/CVPR2023W/EarthVision/papers/Ebel_UnCRtainTS_Uncertainty_Quantification_for_Cloud_Removal_in_Optical_Satellite_Time_CVPRW_2023_paper.pdf"

[59] M. Gonzalez-Calabuig, M. Ángel Fernández-Torres, and G. Camps-Valls, "Generative networks for spatio-temporal gap filling of sentinel-2 reflectances," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 220, pp. 637–648, 2025.

[60] Q. Shu, X. Zhu, S. Xu, Y. Wang, and D. Liu, "Restore-dit: Reliable satellite image time series reconstruction by multimodal sequential diffusion transformer," *Remote Sensing of Environment*, vol. 328, no. 114872, 2025.

[61] J. Yang and X. Huang, "Efficient spectral-temporal reconstruction of long-term satellite time series via temporal segments and mask-informed embedding," *Remote Sensing of Environment*, vol. 342, p. 115470, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425726002403

[62] W. Du, D. Côté, and Y. Liu, "Saits: Self-attention-based imputation for time series," *Expert Systems with Applications*, vol. 219, p. 119619, 2023.

[63] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long, "Timesnet: Temporal 2d-variation modeling for general time series analysis," in *International Conference on Learning Representations 2023*, 2023, p. n.a., https://iclr.cc/virtual/2023/poster/11976.

[64] M. Goswami, K. Szafer, A. Choudhry, Y. Cai, S. Li, and A. Dubrawski, "Moment: a family of open time-series foundation models," in *Proceedings of the 41st International Conference on Machine Learning*, ser. ICML'24. JMLR.org, 2024.

[65] C. Requena-Mesa, V. Benson, M. Reichstein, J. Runge, and J. Denzler, "Earthnet2021: A large-scale dataset and challenge for earth surface forecasting as a guided video prediction task," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1132–1142, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:233306922

[66] J. Rieser, M. Veste, M. Thiel, and S. Schönbrodt-Stitt, "Coverage and rainfall response of biological soil crusts using multi-temporal sentinel-2 data in a central european temperate dry acid grassland," *Remote Sensing*, vol. 13, no. 16, 2021. [Online]. Available: https://www.mdpi.com/2072-4292/13/16/3093

[67] L. Yan and D. P. Roy, "Spatially and temporally complete landsat reflectance time series modelling: The fill-and-fit approach," *Remote Sensing of Environment*, vol. 241, p. 111718, 2020.

[68] A. Schmitt, A. Wendleder, R. Kleynmans, M. Hell, A. Roth, and S. Hinz, "Multi-source and multi-temporal image fusion on hypercomplex bases," *Remote Sens.*, vol. 12, no. 6, p. 943, 2020.

[69] 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*, vol. 102, pp. 122–139, 2015.

[70] A. Schmitt and A. Wendleder, "Sar-sharpening in the kennaugh framework applied to the fusion of multi-modal sar and optical images," in *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, ser. ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., vol. IV-1, 2018, pp. 133–140.

[71] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, 2005, pp. 60–65 vol. 2.

[72] S. Hauser, "HELIX-Trust: A model-agnostic trust framework for spatio-temporal geospatial predictions," *gis.Science - Die Zeitschrift fur Geoinformatik*, vol. 39, no. 2, pp. 52 – 61, 2026.

[73] P. J. Rousseeuw and C. Croux, "Alternatives to the median absolute deviation," *Journal of the American Statistical Association*, vol. 88, no. 424, pp. 1273–1283, 1993. [Online]. Available: http://www.jstor.org/stable/2291267

[74] S. Hauser, S. Dachsberger, A. Schmitt, and S. Hinz, "Calibrated U-Net with HELIX-based label enrichment for ageing-aware spatio-temporal urban change detection," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XI-3-2026, pp. 959–966, 2026. [Online]. Available: https://isprs-annals.copernicus.org/articles/XI-3-2026/959/2026/

[75] L. McInnes, J. Healy, N. Saul, and L. Grossberger, "UMAP: Uniform manifold approximation and projection," *Journal of Open Source Software*, vol. 3, no. 29, p. 861, 2018.

[76] S. Mählkvist, P. Netzell, T. Helander, and K. Kyprianidis, "Trust, but verify—post-hoc analysis of industrial machine learning via interpretability metric embedding and surrogate mapping," *Sensors*, vol. 26, no. 10, 2026. [Online]. Available: https://www.mdpi.com/1424-8220/26/10/3232

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2026.3741409

[77] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the American Statistical Association*, vol. 32, no. 200, pp. 675–701, 1937. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/01621459.1937.10503522

[78] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583–621, 1952. [Online]. Available: https://doi.org/10.1080/01621459.1952.10483441

[79] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979. [Online]. Available: http://www.jstor.org/stable/4615733

[80] D. S. Kerby, "The simple difference formula: An approach to teaching nonparametric correlation1," *Comprehensive Psychology*, vol. 3, p. 11.IT.3.1, 2014. [Online]. Available: https://journals.sagepub.com/doi/abs/10.2466/11.IT.3.1

[81] A. Schmitt, "Multiscale and multidirectional multilooking for sar image enhancement," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 9, pp. 5117–5134, 2016.

[82] ——, "Änderungserkennung in multitemporalen und multipolarisierten radaraufnahmen [change detection in multitemporal and multipolarized radar acquisitions]," Ph.D. dissertation, Karlsruhe Institute of Technology (KIT), 2012.

[83] ——, "Perturbation-based significance estimation for multi-temporal and/or multi-polarized sar descriptors," in *EUSAR 2014; 10th European Conference on Synthetic Aperture Radar*, 2014, pp. 1–4.

[84] L. Lannelongue, J. Grealey, and M. Inouye, "Green algorithms: Quantifying the carbon footprint of computation," *Advanced Science*, vol. 8, no. 12, p. 2100707, 2021. [Online]. Available: https://doi.org/10.1002/advs.202100707



Andreas Schmitt was awarded the Diploma degree "Dipl.-Ing." in Geodesy and Geoinformatics from the University of Karlsruhe (TH), Germany, in 2008, and subsequently the Ph.D. degree "Dr.-Ing." from the Karlsruhe Institute of Technology (KIT), Germany, in 2011. From 2008 until 2016, he was employed by the German Remote Sensing Data Center (DFD) within the German Aerospace Center (DLR), which is located in Oberpfaffenhofen, Germany. During the course of his doctoral and postdoctoral research, the subject further developed the radar data preparation and multisource data integration of the TerraSAR-X ground segment. In 2016, he was appointed to the position of Professor of Applied Geodesy at Munich University of Applied Sciences (HM), Germany. Since 2020, he has been a founding member of the Institute for Applications of Machine Learning and Intelligent Systems (IAMLIS). At the Munich Center for Digitization and Artificial Intelligence (MUC.DAI), he is responsible for the Geodata Science study programme. Professor Schmitt's special interests concern the methodology of integrating multi-dimensional geodata from satellite remote sensing and various other data sources in one consistent framework for the subsequent fully-automated interpretation and validation by reliable and retraceable machine learning methods. The applications of his methodological developments span from settlement analysis via forest assessment to glacier monitoring (i.a.).



Sarah Hauser received the B.Eng. and M.Sc. degrees in geoinformation from Munich University of Applied Sciences, Munich, Germany, in 2019, and Jade University of Applied Sciences, Oldenburg, Germany, in 2021, respectively, and the Dr.-Ing. degree in engineering sciences from the Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany, in 2025. From 2022 to 2025, she worked as a Research Associate in the Department of Geoinformatics and the Institute for Applications of Machine Learning and Intelligent Systems at Munich University of Applied Sciences (e.g., the Wald5Dplus project), in parallel with her doctoral research at KIT. Following her doctorate, she worked with the Institute of Photogrammetry and Remote Sensing, KIT. She is currently an E-STEEM Postdoctoral Fellow with the Department of Geography and Regional Research, Faculty of Earth Sciences, Geography and Astronomy, University of Vienna, Vienna, Austria. Her research focuses on trustworthy Earth observation using multisensor and multitemporal data, feature and label enhancement, with emphasis on quality-aware machine learning and interpretable AI models, and methods for quantifying predictive trust. Her primary scientific contributions include the development of the HELIX Framework.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/