Improving Subseasonal Forecasts of Central European Wintertime Temperatures and Cold-Wave Days Using Random Forests and Weather Regimes

Zur Erlangung des akademischen Grades eines
DOKTORS DER NATURWISSENSCHAFTEN(Dr. rer. nat.)
von der KIT-Fakultät für Physik des
Karlsruher Instituts für Technologie (KIT)

genehmigte

DISSERTATION

von

M.Sc. Selina Margarete Kiefer
aus Neustadt an der Weinstraße

Tag der mündlichen Prüfung:	13.12.2024
Referent:	Prof. Dr. Joaquim G. Pinto
Korreferent:	Prof. Dr. Peter Knippertz

Abstract

Skillful weather predictions on the subseasonal timescale (two to four weeks in advance) are crucial for many socio-economic applications. However, forecasting, especially extremes, on this timescale is very challenging as the information from initial conditions is gradually lost with increasing lead time. Therefore, data-driven methods are discussed as a cost-efficient alternative or complement in a postprocessing sense to numerical weather predictions (NWPs).

The aim of this research is to improve forecasts of wintertime 2-meter temperatures and the occurrence of cold-wave days in Central Europe at lead times of 14, 21 and 28 days. The main focus is thereby on the combination of meteorological knowledge and data-driven Random Forest (RF) models to either forecast these properties directly from predictors based on reanalysis fields or to use the RF models for postprocessing state-of-the-art NWPs. For those, European Centre for Medium-Range Weather Forecasts's (ECMWF's) Sub-seasonal to Seasonal (S2S) reforecasts are used. The (postprocessed) predictions are compared for the winters 2000/2001-2019/2020 to a climatological benchmark ensemble based on observational data. The evaluation is performed as full distribution predictions for continuous values using the Continuous Ranked Probability Skill Score and as binary categorical forecasts using the Brier Skill Score.

In case of the RF models based solely on reanalysis data, we achieve skillful forecasts of 2-meter temperatures compared to the climatological benchmark ensemble on lead times of 14, 21 and 28 days. Concerning the occurrence of cold-wave days, skillful forecasts are obtained by the RF models for lead times of 14 and 21 days. Therefore, these models are a suitable alternative to using a climatological model for subseasonal predictions. Furthermore, the application of Shapley Additive Explanations suggests that the RF models are able to learn physically known relationships in the data, as e.g. the relevance of the state of the stratospheric polar vortex during a severe cold wave.

In case of the RF-based postprocessing models, a higher skill for both the forecasts of 2-meter temperatures and the occurrence of cold-wave days is obtained compared to the predictions of the RF models based solely on reanalysis data. An Impurity Based Feature Importance reveals that especially the reforecasted 2-meter temperature and predictors representing the atmospheric large-scale flow are relevant for the models' predictions.

While the RF-based postprocessing models achieved the highest skill of the analyzed methods for the predictions of 2-meter temperatures at lead times of 14, 21 and 28 days and the predictions of the occurrence of cold-wave days at a lead time of 28 days, a simple lead-time-dependent mean bias correction of the original ECMWF's S2S reforecasts achieves the highest skill for forecasts of the occurrence of cold-wave days at lead times of 14 and 21 days. Furthermore, it improves the skill of the original ECMWF's S2S reforecasts at all lead times.

Besides improving forecast skill, we demonstrate the suitability of weather regimes (WRs) to assess forecast reliability. WRs represent the slowly varying, large-scale atmospheric background flow and therefore may contain relevant information about the subseasonal predictability of Central European weather. We find that the WR present at the initialization of the predictions influences the forecast skill and can therefore be used to assess forecast reliability. Nevertheless, due to a limited sample size, this has to be done with caution.

With the example of the mean bias corrected 14-day reforecasts of ECMWF for the occurrence of cold-wave days, we investigate possible reasons for the dependence of forecast skill on the WR present at initialization. Thereby, we specifically focus on reforecasts initialized during the WR Greenland Blocking (GL; characterized by a high pressure system over Greenland) and the WR Scandinavian Trough (ScTr; characterized by a low pressure system over Scandinavia) which show significant differences in their skill. In case of the reforecasts initialized during GL, which show a higher skill than reforecasts initialized during ScTr, we find WR successions which follow typical climatological patterns during the 14 days of forecasts more often. We suggest that this might be one of the main reasons for an increased forecast skill.

Furthermore, we analyze the WR successions for the best and worst predicted days within the observed cold waves of the winters 2000/2001-2019/2020. This is done independent from the WR present at initialization. We find that forecast skill at a lead time of 14 days is significantly higher, when the European Blocking WR (characterized by a high pressure system over the British Isles and southern Scandinavia) is present a few days before the target date of the forecast. Given a reasonably good forecast of WRs, this result can be used to assess the reliability of cold-wave day predictions.

The research presented in this thesis shows the potential of cost- and time-efficient RF models for both forecasting wintertime 2-meter temperatures and the occurrence of cold-wave days as well as postprocessing existing NWPs of these. Furthermore the suitability of WRs in assessing forecast reliability is demonstrated on the basis of a case study.

Zusammenfassung

Im Bereich vieler sozio-ökonomischer Anwendungen ist es von enomer Wichtigkeit das Wetter auf einer Zeitskala von zwei bis vier Wochen im Voraus (subsaisonal) korrekt vorherzusagen. Leider ist die Vorhersagbarkeit, insbesondere von Extremwetter, auf diesen Zeitskalen sehr schwierig, da die Informationen aus den Anfangsbedingungen der numerischen Gleichungen, aus denen ein Wettermodell aufgebaut ist, graduell abnehmen. Deswegen stehen daten-getriebene Modelle als Alternative zu numerischen Wettermodellen zur Diskussion.

Ziel dieser Arbeit ist die Verbesserung der Güte von Vorhersagen der 2-Meter Temperatur und des Auftretens von Kältewellentagen im mitteleuropäischen Winter 14, 21 und 28 Tage im Voraus. Der Hauptaspekt liegt dabei auf der Kombination von meteorologischem Sachverstand und daten-getriebenen Random Forest- (RF) Modellen, einer Art maschinellen Lernens. Diese Modelle werden einerseits eingesetzt um die Vorhersagegrößen direkt zu bestimmen, anderseits in einem Nachbereitungs-Kontext (Postprocessing). Im Falle ersteren werden anhand von meteorologischen Gesichtspunkten ausgewählte Prädiktoren genutzt, die auf Reanalyse-Daten beruhen. Im Falle letzteren werden die nachträglich vorhergesagten (re-vorhergesagten, reforecasted) Felder der subsaisonalen bis saisonalen (S2S) Re-Vorhersagen des Europäischen Zentrum für Mittelfristige Wettervorhersage (ECMWF) verwendet. Die Vorhersagen beider Varianten werden für die Winter 2000/2001-2019/2020 mit einem klimatologischen Ensemble basierend auf Beobachtungsdaten verglichen. Die Evaluation wird anhand von Vorhersagen der kompletten Verteilung der kontinuierlichen Werten mit dem kontinuierlichen Rank-Gütemaß (Continuous Ranked Probability Skill Score) vorgenommen und anhand von binären kategorischen Vorhersagen anhand des Brier Gütemaßes (Brier Skill Score).

Im Falle der RF-Modelle die ausschließlich auf Reanalyse-Daten beruhen werden bessere Vorhersagen der 2-Meter Temperatur für 14, 21 und 28 Tage im Voraus erreicht verglichen mit dem klimatologischen Ensemble. Bezüglich des Auftretens von Kältewellentagen werden bessere Vorhersagen 14 und 21 Tage im Voraus erreicht. Daher sind diese Modelle eine geeignete Alternative zu einer klimatologischen Vorhersage auf der subsaisonalen Zeitskala. Desweiteren, hat die Anwendung der Shapley Additive Explanations, einer Methode um die für Vorhersagen relevanten Prädiktoren herauszufinden, gezeigt, dass diese RF-Modelle in der Lage sind physikalisch bekannte Muster in den Eingabedaten zu lernen, wie beispielsweise die Relevanz des Zustands des stratosphärischen Polarwirbels im Winter während einer starken Kältewelle.

Obwohl gute Vorhersagen mit RF-Modellen, die nur auf Reanalyse-Daten beruhen, erreicht werden, so generieren RF-basierte Postprocessing-Modelle im Allgemeinen noch bessere Vorhersagen der 2-Meter Temperatur und des Auftretens von Kältewellentagen. Eine Impurity Based Feature Importance, eine Methode zur Bestimmung der wichtigsten Prädiktoren im Prozess des Lernens der Modelle, hat dabei gezeigt, dass vor allem die nachträglich vorhergesagt 2-Meter Temperatur und Prädiktoren, die den großskaligen atmosphärischen Fluss beschreiben, wichtig für die Vorhersagen der RF-basierten Postprocessing Modelle sind.

Eine einfache Postprocessing-Alternative zu den den RF-basierten Modellen ist die vorhersagedauerabhägige Korrektur des mittleren Fehlers (lead-time-dependent mean bias correction) der ECMWF S2S Re-Vorhersagen. Diese Methode verbessert die Vorhersage der unbearbeiteten ECMWF S2S Re- Vorhersagen für alle Vorhersagedauern. Wir zeigen, dass dieser Ansatz die beste Lösung für die Vorhersage von Kältewellentagen 14 und 21 Tage im Voraus ist. RF-basierte Postprocessing-Modelle erzielen die besten Vorhersagen der 2-meter Temperaturr 14, 21 und 28 Tage im Voraus sowie die besten Vorhersage der Kältewellentage 28 Tage im Voraus.

Neben der Verbesserung der Vorhersagegüte demonstrieren wir den Nutzen von Wetterregimen (WR) um die Verlässlichkeit der Vorhersagen einzuschätzen. WR repräsentieren die großskalige Luftbewegung in der Troposphäre und besitzen daher eventuell nützliche Informationen über die subsaisonale Vorhersagberkeit mitteleuropäischen Wetters. Wir sehen, dass das WR, das beim Vorhersagestart vorherrschend war, die Vorhersagegüte beeinflusst und daher zur Einschätzung der Verlässlichkeit von Vorhersagen genutzt werden kann. Allerdings muss dies aufgrund der geringen Datenlage mit Vorsicht geschehen.

Anhand des Beispiels der Vorhersagen von Kältewellentagen der mean bias corrected ECMWF S2S Re-Vorhersage mit einer Vorhersagedauer von 14 Tagen untersuchen wir mögliche Gründe für die Güteunterschiede der Vorhersagen abhängig von dem WR bei Initialisierung. Dabei konzentrieren wir uns speziell auf Vorhersagem die während einer blockierenden Wetterlage über Grönland (GL; charakterisiert durch ein Hochdrucksystem über Grönland) oder einer meridionalen Wetterlage über Skandinavian (ScTr; charakterisiert durch ein Tiefdrucksystem über Skandinavian) initialisiert werden, da diese signifikante Unterschiede in ihrer Güte zeigen. Vorhersagen, die während GL initialisiert werden, zeigen dabei das beste Brier-Gütemaß während Vorhersagen, die während ScTr initialisiert werden, das schlechteste Brier-Gütemaß aufweisen. Wir sehen, dass im Falle der Vorhersagen, die während GL initialisiert werden, die WR-Abfolgen häufiger typischen klimatologischen Mustern während der Vorhersagezeit folgen als im Falle der Vorhersagen, die während ScTr gestartet werden. Daher sehen wir es als wahrscheinlich an, dass dies einer der Gründe für eine bessere Vorhersagegüte ist.

Desweiteren analysieren wir, unabhängig vom WR das bei der Initialisierung der Vorhersagen vorherrscht, die WR-Abfolgen vor den am besten (schlechtesten) vorhergesagten Kältewellentagen der Winter 2000/2001-2019/2020. Wir zeigen, dass die Vorhersagegüte signifikant höher ist, wenn eine blockierende Wetterlage über Europa (charakterisiert durch ein Hochdrucksystem über den britischen Inseln und Süd-Skandinavien) in den Tagen vor dem Vorhersagedatum präsent ist. Unter der Annahme, dass eine entsprechend verlässliche Vorhersage der WRs existiert, kann dieses Ergebnis benutzt werden um die Verlässlichkeit der Vorhersagen von Kältewellen, die 14 Tage im Voraus vorhergesagt werden, einzuschätzen.

Die vorgestellte Arbeit zeigt das Potential von zeit- und kostengünstigen RF-Modellen für die Vorhersage der 2-Meter Temperatur und des Auftretens von Kältewellen in Mitteleuropa im Winter. Desweiteren wird die Nutzbarkeit von WR für die Einschätzung der Verlässlichkeit von Vorhersagen anhand einer Fallstudie demonstriert.

Preface

The PhD candidate confirms that the research presented in this thesis contains significant scientific contributions by herself. This thesis reuses material from the following publications:

> Kiefer, S.M., S. Lerch, P. Ludwig, and J.G. Pinto, 2023: Can Machine Learning Models Be a Suitable Tool for Predicting Central European Cold Winter Weather on Subseasonal to Seasonal Time Scales? *Artificial Intelligence for the Earth Systems*, **2 (4)**, e230 020, https://doi.org/10.1175/AIES-D-23-0020.1.
>
> Kiefer, S.M., S. Lerch, P. Ludwig, and J.G. Pinto, 2024a: Random Forests' Postprocessing Capability of Enhancing Predictive Skill on Subseasonal Timescales - A Flow-Dependent View on Central European Winter Weather. *Artificial Intelligence for the Earth Systems*, **3 (4)**, e240014, https://doi.org/10.1175/AIES-D-24-0014.1.
>
> Kiefer, S. M., P. Ludwig, S. Lerch, P. Knippertz, and J. G. Pinto, 2024b: The Role of Weather Regimes for Subseasonal Forecast Skill of Cold-Wave Days in Central Europe. *EGUsphere* [preprint, submitted to *Weather and Climate Dynamics*], 1-26, `https://doi.org/10.5194/egusphere-2024-2955`.

The concept of the study presented in the Kiefer et al. (2023) was developed by SMK, SL, PL and JGP. Data analysis and Figures were done by SMK. SMK wrote the original draft. All authors have contributed with methods, the discussion and revision of the original draft.

The concept of the study presented in the Kiefer et al. (2024a) was developed by SMK, SL, PL and JGP. Data analysis and Figures were done by SMK. SMK wrote the original draft. All authors have contributed with methods, the discussion and revision of the original draft.

The concept of the study presented in the Kiefer et al. (2024b) was developed by SMK, PL and JGP. Data analysis and Figures were done by SMK. SMK wrote the original draft with help of JGP. All authors have contributed with methods, the discussion and revision of the original draft.

The abstract, chapters 1, 2, 3, 5, 6 and 9 reuse material from Kiefer et al. (2023). ©American Meteorological Society. Used with permission.

The candidate wrote the text of this manuscript with advice from Prof. Dr. Joaquim G. Pinto, Prof. Dr. Peter Knippertz, Dr. Sebastian Lerch and Dr. Patrick Ludwig.

The research presented in this thesis has been accomplished using Python (available at: https://www.python.org/, last visited 27 Aug 2024) and Climate Data Operators (available at: http://www.mpimet.mpg.de/cdo, last visited 27 Aug 2024). The code developed by the PhD candidate is available at:

- https://github.com/selinakiefer/rfs_central_european_cold_winter_weather
  (Kiefer et al. (2023), results presented in chapter 6)
- https://github.com/selinakiefer/rfs_pp_subseasonal_european_winter_temperatures
  (Kiefer et al. (2024a), results presented in chapter 7 and 8)
- https://zenodo.org/records/14000544
  (Kiefer et al. (2024b), results presented in chapter 8, access is restricted before final publication but can be provided on request)

The research of the PhD candidate has been partly embedded within the subproject C8 "Stratospheric influence on predictability of persistent weather patterns" of the "Waves to Weather" Transregional Collaborative Research Center SFB/TRR 165 funded by the German Research Foundation (DFG). The research proposal has been written by Prof. Dr. Joaquim G. Pinto (Karlsruhe Institute of Technology), Dr. Hella Garny (German Aerospace Center) and Prof. Dr. Thomas Birner (Ludwig-Maximilians-University Munich). Further funding has been provided by the German Helmholtz Association within the “Changing Earth” program.

The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others. This copy has been supplied on the understanding that this is copyright material and that no quotation from the thesis may be published without proper acknowledgment.

List of Abbreviations

_all: nine meteorological fields and the month (as input for RF models)

_ens: ensemble information of reforecasts (as input for RF models)

_era5: reanalysis data (as input for RF models)

_pca: first ten PCs of fields (as input for RF models)

_s2s: reforecast data (as input for RF models)

_sel: three meteorological fields and the month (as input for RF models)

_sepf: all reforecast ensemble members (as input for RF models)

_stat: statistics of fields (as input for RF models)

4D-Var: Four Dimensional Variational Data Assimilation

ANN: Artificial Neural Network

AR: Atlantic Ridge

AT: Atlantic Trough

BLO±: positive/negative phase of blocking

BoM: (Australian) Bureau of Meteorology

BS: Brier Score

BSS: Brier Skill Score

CDF: Cumulative Distribution Function

CNN: Convolutional Neural Network

CRPS: Continuous Ranked Probability Score

CRPSS: Continuous Ranked Probability Skill Score

DT: Decision Tree

ECMWF: European Centre for Medium-Range Weather Forecasts

EDA: Ensemble Data Assimilation

ELU: Exponential Linear Unit

EMOS: Ensemble Model Output Statistics

ENSO: El Niño-Southern Oscillation

E-OBS: daily gridded observational dataset for Europe

EOF: Empirical Orthogonal Function

ERA5: ECMWF Reanalysis version 5

EuBL: European Blocking

Gini: generalized inequality

GL: Greenland Blocking

H850: specific humidity in 850 hPa height

hPa: hecto-Pascal (1 hPa = 100 Pa)

IFS: Integrated Forecasting System

I$_{WR}$: Weather Regime index

LRP: layer-wise relevance propagation

MJO: Madden-Julian Oscillation

ML: Machine Learning

MOS: Model Output Statistics

msl: mean sea level pressure

NAO±: positive/negative phase of the North Atlantic Oscillation

NGR: non-homogenous Gaussian regression

NCEP: (United States) National Centers for Environmental Prediction

No: No regime

NWP: Numerical weather prediction model

PCA: Principle Components Analysis

PC: Principle Components

PV: potential vorticity

QBO: Quasi-biennual Circulation

QRF: Quantile Regression Forest

RF: Random Forest

RFC: Random Forest Classifier

RWB: Rossby-Wave breaking

S2S: Subseasonal to Seasonal

ScBL: Scandinavian Blocking

ScTr: Scandinavian Trough

SHAP: SHapley Additive exPlanations

SST: sea surface temperature

SSW: Sudden Stratospheric Warming

SV: Singular Vectors

t2m: (reforecasted) 2-meter temperature

t850: temperature in 850 hPa height

tg: observed daily mean 2-meter temperature

u10: zonal wind in 10 hPa height

u300: zonal wind in 300 hPa height

UKMO: UK (United Kingdom) Met (Meteorological) Office

UNet: U-shaped CNN architecture

WMO: World Meteorological Organization

WR: Weather Regime

XAI: eXplainable Artificial Intelligence z100: geopotential in 100 hPa height

z250: geopotential in 250 hPa height

z300: geopotential in 300 hPa height

z500: geopotential in 500 hPa height

z850: geopotential in 850 hPa height

ZO: Zonal regime

Contents

1. Introduction

Wintertime temperature extremes, especially cold waves, can pose a major threat to society. In case of Central Europe, these occur when cold air masses originating from the Arctic or polar regions are transported over days into the area. Due to this shift in the atmospheric large-scale circulation, cold waves are a phenomena lasting several days which magnifies their impact. For example, 600 fatalities were reported across Europe due to extreme cold temperatures and heavy snowfall during the cold wave in February 2012 (DWD, 2012). Furthermore, transportation and energy supply were severely disrupted during this period. To minimize these and other negative impacts associated with cold wintertime temperatures, it is important to take timely measures. Therefore, it is of great interest to have skillful predictions of these well in advance. In fact, also many other decisions concerning e.g. agriculture, food and energy security, health care, and transportation are made on timescales of two weeks to three months into the future, which is also called the subseasonal to seasonal (S2S) timescale (Fig. 1.1 (a) and (b), DelSole et al. (2017)).

Although these decisions require accurate weather forecasts, the forecasting skill of traditional numerical weather prediction (NWP) models at that timescale is still low (Fig. 1.1 (a), White et al. (2017)). To a large extent, this is due to the fact that NWP models rely heavily on the initial conditions, e.g. the mid-tropospheric circulation state, which provide at best a practical predictability up to 10 days for local daily weather forecasts in the northern hemispheric mid-latitudes (White et al., 2017; Zhang et al., 2019). On the other hand, skillful forecasts can also be based on the boundary conditions, e.g. sea surface temperatures. But these are slowly changing variables which mainly contribute to the predictive skill at seasonal and longer lead times. In between, forecasting is very challenging as predictability is mostly derived from teleconnections, which are often non-stationary links (Knippertz et al., 2003) and still a subject of on-going research, and thus not represented well by the equations and parameterizations used in NWP models (Vitart and Robertson, 2018).

Teleconnections are phenomena which cause quasi-persistent, recurring large-scale circulation anomalies influencing the state of a region which is geographically not directly connected to the original location of the phenomena (Wallace and Gutzler, 1981). Since these anomalies need a certain time to propagate from their origin to the area of interest, they form a source of predictability at the S2S timescale (Baldwin et al., 2003; Tripathi et al., 2015). These scenarios, during which certain atmospheric conditions are present that foster the influence of teleconnections and feedbacks within the climate system, are also

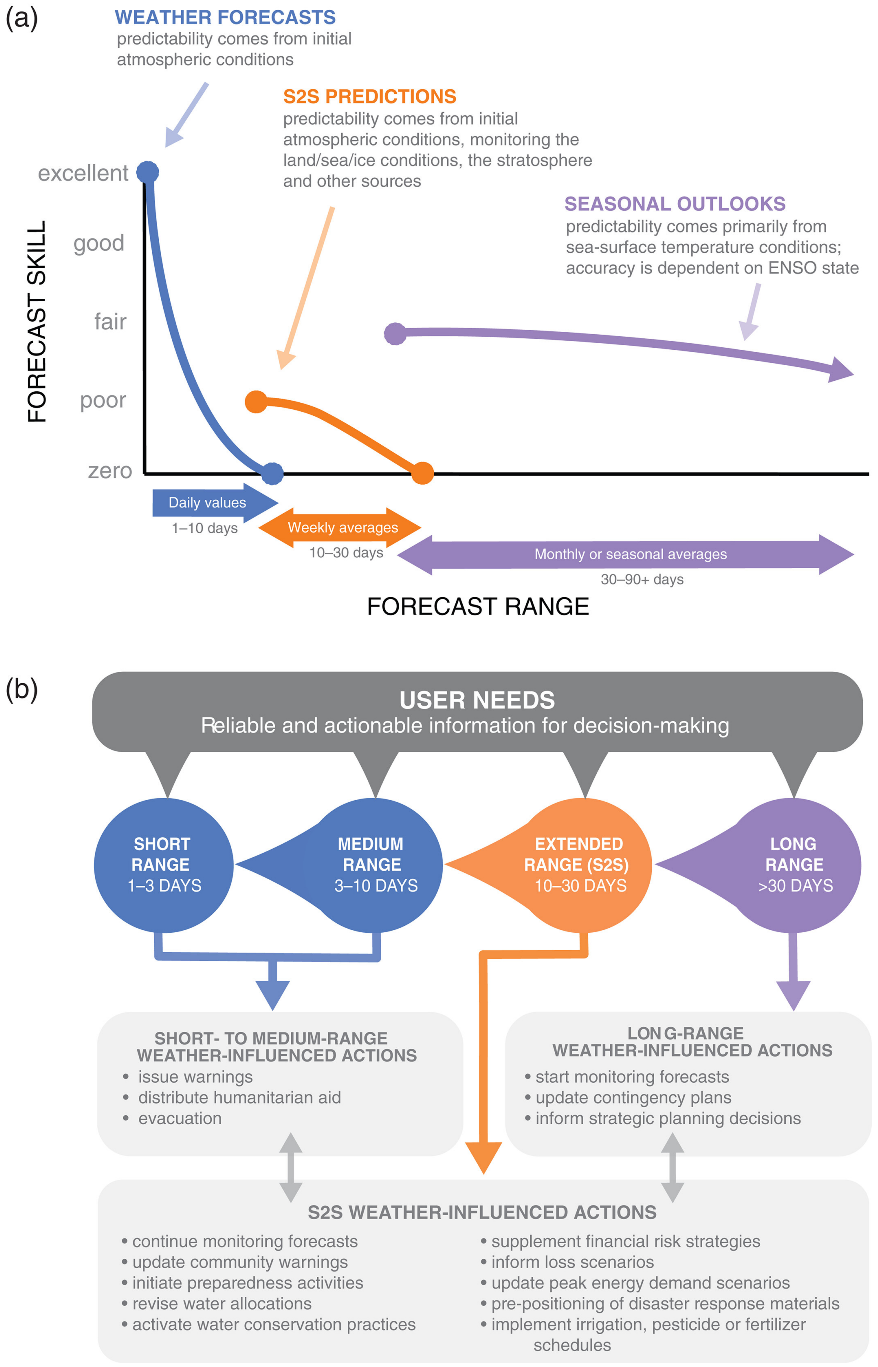

(a)
WEATHER FORECASTS
predictability comes from initial atmospheric conditions
S2S PREDICTIONS
predictability comes from initial atmospheric conditions, monitoring the land/sea/ice conditions, the stratosphere and other sources
SEASONAL OUTLOOKS
predictability comes primarily from sea-surface temperature conditions; accuracy is dependent on ENSO state
FORECAST SKILL
excellent
good
fair
poor
zero
Daily values
1–10 days
Weekly averages
10–30 days
Monthly or seasonal averages
30–90+ days
FORECAST RANGE
(b)
USER NEEDS
Reliable and actionable information for decision-making
SHORT RANGE 1–3 DAYS
MEDIUM RANGE 3–10 DAYS
EXTENDED RANGE (S2S) 10–30 DAYS
LONG RANGE >30 DAYS
SHORT- TO MEDIUM-RANGE WEATHER-INFLUENCED ACTIONS
• issue warnings
• distribute humanitarian aid
• evacuation
LON G-RANGE WEATHER-INFLUENCED ACTIONS
• start monitoring forecasts
• update contingency plans
• inform strategic planning decisions
S2S WEATHER-INFLUENCED ACTIONS
• continue monitoring forecasts
• update community warnings
• initiate preparedness activities
• revise water allocations
• activate water conservation practices
• supplement financial risk strategies
• inform loss scenarios
• update peak energy demand scenarios
• pre-positioning of disaster response materials
• implement irrigation, pesticide or fertilizer schedules

Figure 1.1.: (a) Qualitative estimate of forecast skill based on forecast range from short-range weather forecasts to long-range seasonal predictions, including potential sources of predictability. Relative skill is based on differing forecast averaging periods. (b) A schematic diagram highlighting the relationship between the subseasonal-to-seasonal (S2S) 'extended-range' forecast range and other prediction timescales, with examples of actionable information that can enable decision-making across sectors. Actions are examples only and are not exclusive to a forecast range. (a) Adapted by Elisabeth Gawthrop from an original figure by Tony Barnston, both International Research Institute for Climate and Society; edited and reproduced with permission. (b) Based on Meehl et al. (2001), Hurrell et al. (2009) and Goddard et al. (2014). Definitions are based on WMO meteorological forecasting ranges: `http://www.wmo.int/pages/prog/www/DPS/GDPS-Supplement5-AppI-4.html`. Figure and caption reprinted from White et al. (2017), their Fig. 1.

called "windows of opportunity" (Mariotti et al., 2020).

One example of these is the relation between the occurrence of Sudden Stratospheric Warming (SSW) events in winter and European surface weather in the subsequent weeks (Baldwin et al., 2003). Following the canonical behaviour proposed by Baldwin et al. (2003), temperature and wind anomalies form in the stratosphere and propagate downward. Especially over the North Atlantic sector, these anomalies sometimes penetrate through the tropopause and influence the large-scale tropsopheric flow which then determines Central European surface weather (Charlton-Perez et al., 2018). In general, physically known teleconnections, which also include tropical sources such as the El Niño Southern Oscillation (ENSO, Jiménez-Esteve and Domeisen (2018)), have in common that they influence Central European weather via changes in the tropospheric large-scale circulation over the North Atlantic Ocean and Europe.

This pathway can be used to create skillful forecasts on the subseasonal timescale in various ways. The European Centre for Medium-Range Weather Forecasts (ECMWF) predicts the large-scale flow as the probability of occurrence of four distinct, slowly evolving patterns of the tropospheric background flow with their NWP model (Grams et al., 2020). Two of these patterns represent thereby the North Atlantic Oscillation (NAO). The NAO considerably influences European weather in winter and is characterized by two centers of action, the Icelandic Low and the Azores High (Pinto and Raible, 2012). Depending on the strength of these pressure systems, the strength of the westerly winds over the eastern North Atlantic Ocean varies (Pinto and Raible, 2012) and the large-scale atmospheric flow is either more zonal (NAO+) or more meridional (NAO-) over the North Atlantic Ocean (Benedict et al., 2004). In case of the above mentioned teleconnection example of the influence of SSWs on surface weather, the likelihood of the establishment of NAO- conditions is enhanced after the occurrence of an SSW (Charlton-Perez et al., 2018). This, in turn, leads to an increased likelihood of the occurrence of cold waves in Central Europe.

Besides the NAO, the classification of the slowly evolving tropospheric background flow by the Euro-Atlantic weather regimes (WRs) proposed by Grams et al. (2017) is a useful tool to estimate forthcoming Central European surface weather. This classification consists of seven distinct WRs and a so-called "No" regime class, which collects all large-scale tropospheric flow patterns close to climatology. Four of the WRs are characterized by a meriodional large-scale flow over the Euro-Atlantic region usually leading to cool temperatures over Central Europe during winter. The other three regimes are characterized by a stronger than usual zonal flow transporting warm temperatures towards Central Europe. Both the seven Euro-Atlantic WRs and the NAO can be influenced by various teleconnections themselves.

This fact is one of the reasons why studies like e.g. Mayer and Barnes (2021) use machine learning (ML) models instead of NWP models to predict large-scale atmospheric flow patterns. The advantage of using ML instead of NWP models is that these do not depend on the classical initial or boundary conditions and can be capable of learning statistically relevant patterns in the data which correspond to e.g. teleconnections. ML models are generally very flexible, cost- and time-efficient. They can, similar to NWP models, be started at any given date, create forecasts of arbitrary lengths, resolutions and ensemble sizes, given that the ML model can be configured to produce ensemble forecasts. ML models can, as most NWP models, be computed globally (e.g. Nguyen et al., 2023; Chen et al., 2023). However, many state-of-the-art ML models for subseasonal predictions (two to four weeks in advance) are designed on local or regional scales for specific forecasting properties (e.g. He et al., 2021). A local approach is also used by van Straaten et al. (2022), who develop Random Forest- (RF) based models using predictors derived from reanalysis data for predicting high summer temperatures over Central and Western Europe. The advantage of using RFs in comparison to more complex ML models is their compatibility for parallel computing which saves time and therefore computational resources.

Although these studies show promising results in using ML instead of NWP models for forecasting on the subseasonal timescale, these models have also non-negligible disadvantages. ML models cannot extrapolate by design since they only "learn" structures in the given data, and therefore their ability to create skillful predictions is fully dependent on the quality of these. To overcome these limitations and still take advantage of the positive aspects, ML models can be applied in a postprocessing sense using the forecasts of NWP models as input (e.g. Horat and Lerch, 2024; Scheuerer et al., 2020). But also traditional postprocessing methods lead to an improvement of forecast skill on the subseasonal timescale (e.g. Hyvärinen et al., 2021; van Straaten et al., 2020). These are, besides a simple lead-time-dependent mean bias correction (Monhart et al., 2018), for example model output statistics (MOS) which use e.g. a linear or logistic regression between the observations and variables predicted by the model in the past, in order to correct current deterministic or ensemble forecasts (EMOS) (Vannitsem et al., 2021; Taillardat et al., 2016).

The aim of the thesis is to improve forecasts of wintertime 2-meter temperatures and the occurrence of cold-wave days in Central Europe at lead times of 14, 21 and 28 days. As depicted on Fig. 1.2 various strategies are explored to do so. The main focus of the research is on the combination of physical knowledge, including teleconnections and WRs, with RF-based ML models. As a benchmark model, a climatological ensemble is chosen. The developed RF-based models are used on the one hand as an alternative to NWP models for forecasting the 2-meter temperature and occurrence of cold-wave days directly. On the other hand, they are used as a complement to NWP models by postprocessing their predictions, here represented by ECMWF's S2S reforecasts. The reliability of forecasts is afterwards assessed using the seven WRs proposed by Grams et al. (2017).

The structure of this thesis is as follow: an overview over the relevant physical background is given in chapter 2. This includes a description of several WR classifications and a selection of teleconnections.

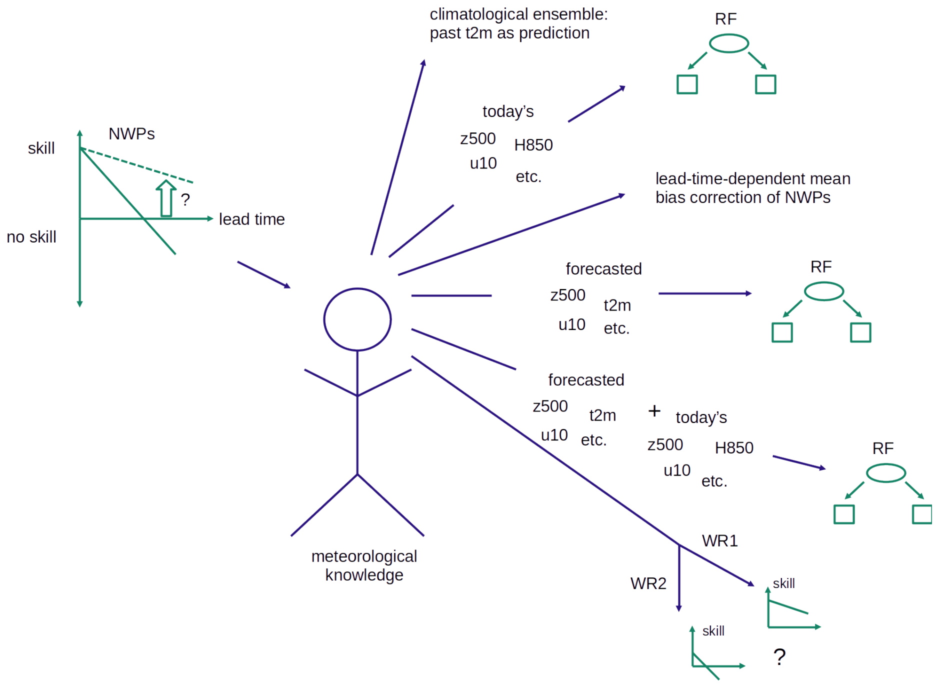


Figure 1.2.: Research strategies presented in this thesis. The question of how to improve subseasonal forecast skill (left side) is tackled by six approaches utilizing meteorological knowledge (right side). These are on the one hand using a climatological ensemble consisting of past observations as predictions (first one from top) or forecasts of RF models based on reanalysis data (denoted as predictors of "today", second from top) as alternatives to NWP forecasts. On the other hand, a lead-time-dependent mean bias correction of NWP forecasts (third from top) and RF-based postprocessing models based only on NWP predictions (denoted as "forecasted" predictors, forth from top) or additionally on reanalysis data (fifth from top) are used as complements to NWP forecasts. Furthermore, it is analyzed if subseasonal forecast skill is dependent on the WR present at initialization. The focus lies thereby on the prediction of wintertime 2-meter temperatures and the occurrence of cold-wave days in Central Europe.

Subsequently, in chapter 3, state-of-the-art approaches in subseasonal forecasting are presented including NWP and ML models as well as typical parametric postprocessing methods. Chapter 4 summarizes the research questions answered in this thesis and chapter 5 the data and methods used to do so. The presented results are divided into three parts. Chapter 6 describes the use of RF models based solely on reanalysis data for forecasting 2-meter temperatures and the occurrence of cold-wave days in Central Europe on lead times of 14, 21 and 28 days directly. Chapter 7 focuses on the postprocessing of ECMWF's S2S reforecasts of the same properties for the same lead times using RF-based models and a simple lead-time-dependent mean bias correction. Chapter 8 investigates how the reliability of the postprocessed forecasts can be assessed using WRs. This is investigated at the example of the mean bias corrected ECMWF's S2S reforecasts of the occurrence of cold-wave days at a lead time of 14 days. Lastly, chapter 9 provides an overall conclusion of the three research parts and in chapter 10 a short outlook is given.

2. The Meteorology of Central European Wintertime Surface Temperatures

In this chapter, the meteorological background of temperature is given. A special focus is thereby put on wintertime surface temperatures, measured at a height of two meters, in Central Europe and their relation to the tropospheric large-scale flow, investigated in form of the Euro-Atlantic WRs, and remote drivers, represented by a selection of teleconnections.

2.1. General Meteorological Background

Fundamentally, Earth's surface air temperature, which is measured at two meters above ground, is determined by the sun's irradiation. However, due to the Earth's atmosphere, the process of heating and cooling is intermitted. The atmosphere can be divided into several layers (Fig. 2.1) according to their temperature gradient. Concerning surface air temperature in the northern hemispheric mid-latitudes, especially the two lower layers are important. The upper one, the stratosphere, is characterized by an increasing temperatures with height (Fig. 2.2, Vallis (2017)). Here, the ozone layer of the atmosphere is located which absorbs a part of the solar irradiation (Vallis, 2017). The troposphere, the lower-most layer of the atmosphere, is characterized by decreasing air temperatures with height (Fig. 2.2, Vallis (2017)). The occurrence of clouds and high density of atmospheric gases in this layer leads to the reflection of a part of the incoming solar radiation as well as the reflection of a part of the terrestrial outgoing radiation (Spiridonov and Ćurić, 2021). In between the layers, a region with stable temperature, the so-called tropopause is located (Vallis, 2017). This intermits the exchange of mass and heat between the atmospheric layers.

Focusing on present day conditions, the tilt of planet Earth and the elliptical form of its orbit around the sun lead to a different amount of incoming solar radiation depending on the time of year and region of Earth. As described in Vallis (2017), at the equator a higher amount of incoming solar radiation is observed than at the poles. This leads to a surplus of heat at the equator. Therefore, several balancing movements take place in the atmosphere. Here, we focus on the relevant processes affecting the temperature in the northern hemispheric mid-latitudes in winter (the hemisphere belonging to the "winter pole" on Fig. 2.1). At the equator, warm air is rising until the tropopause is reached. The largest part of the heat is then redistributed at the tropopause level towards the poles. Thereby, the air cools and descends around 23°N forming the so-called Hadley cell (Fig. 2.1). A smaller part of the heat is transported

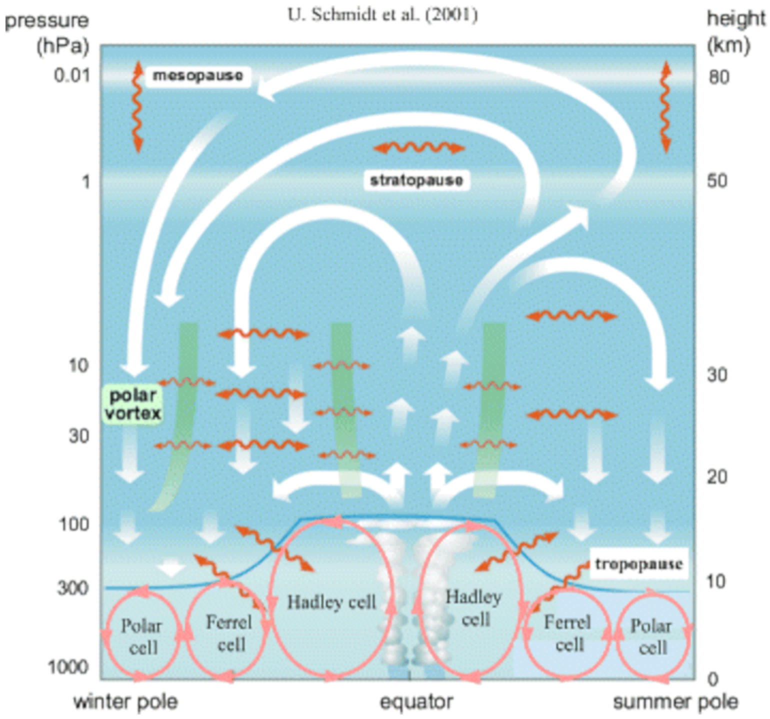


Figure 2.1.: The Earth's Brewer-Dobson circulation. The Hadley, Ferrel and Polar cells are visible in the figure for both hemispheres. The thick white arrows indicate the transformed Eulerian-mean mass stream function. The wavy orange arrows indicate two-way mixing processes. The thick green lines indicate the presence of stratospheric transport and mixing barriers. (Adapted from Bönisch et al. (2011)). Figure and caption reprinted from Proedrou et al. (2016), their Fig. 4, licensed under Creative Commons Attribution 4.0 International License (`http://creativecommons.org/licenses/by/4.0/.`)

further upwards and redistributed towards the winter pole trough the stratosphere and the atmospheric layer above (Fig. 2.1). This is called the Brewer-Dobson circulation. At the pole, the descending air is transported towards the mid-latitudes at the surface. During this movement the air is warming and therefore ascending around 60°N again forming the polar cell (Fig. 2.1). In between a third circulation cell is formed, the so-called Ferrel cell, which is driven by the descending air of the Hadley cell and the ascending air of the polar cell (Fig. 2.1).

On the "boundaries" of these cells strong winds, the so-called jet streams, are observed in the upper troposphere which are a balancing movement between the warm temperatures in the south and the cold temperatures in the north. Regarding Central European surface temperatures, mainly the jet stream between the polar and Ferrel cell is relevant. Due to the Coriolis force, which is a fictitious force induced by the Earth's rotation, the wind direction of the jet stream is generally westerly. However, due to orography, it has a wavy structure leading to the formation of ridges and troughs in the zonal flow. This wavy structure is a form of planetary-scale Rossby-Wave which can be explained by the concept of conservation of potential vorticity (PV). The PV can be formulated as (Kurz, 1990):

$$PV = -g\,(\zeta + \mathrm{f})\,\frac{\partial \Theta}{\partial p}, \tag{2.1}$$

$$\zeta = \nabla \times \vec{v},$$

$$\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right),$$

$$\Theta = T\left(\frac{p_0}{p}\right)^{\kappa}, \kappa = \frac{2}{7},$$

whereby g is the gravitational constant, f the Coriolis parameter, ζ the relative vorticity, $\vec{v}$ the wind speed, Θ the potential temperature and T the actual air temperature.

The Coriolis parameter defines the strength of the Coriolis force. It is dependent on the latitude and the closer to the pole, the stronger the Coriolis force, i.e. the stronger the deflection of an air parcel to the right on the northern hemisphere. An air parcel is thereby a theoretical concept which describes an amount of air with a constant amount of moisture and no addition of heat from the surrounding. The relative vorticity takes into account the deformation, stretching and shear of such an air parcel. The potential temperature of an unsaturated air parcel of dry air is the temperature which is achieved when the air parcel is brought without any exchange of mass or heat with the environment from a pressure level p to the standard pressure level $p_0 = 1000\,\mathrm{hPa}$. It can be used to "track" air parcels and thus reveal the wavy structure of the jet stream.

Due to this wavy structure of the jet stream, Central Europe is affected by air masses of different origins, as described in e.g. Kurz (1990) and Spiridonov and Ćurić (2021). In winter, mainly six categories of air masses are relevant (Fig. 2.3). Maritime Arctic air masses originating from the region of the Greenland Sea are humid and extreme cold leading to cold waves including potentially heavy snowfall in winter. Maritime polar air masses originating from the polar North Atlantic Ocean are less cold but still often lead to cold waves including snowfall. In contrast to that, maritime tropic air masses originating from the Central Atlantic Ocean are humid and warm leading to mild, wet and potentially stormy winter weather. Dry conditions and cold temperatures, potentially leading to cold waves, are observed during the occurrence of continental polar air masses originating from Central Asia. Extreme cold temperatures including cold waves accompany continental Arctic air masses from Siberia.

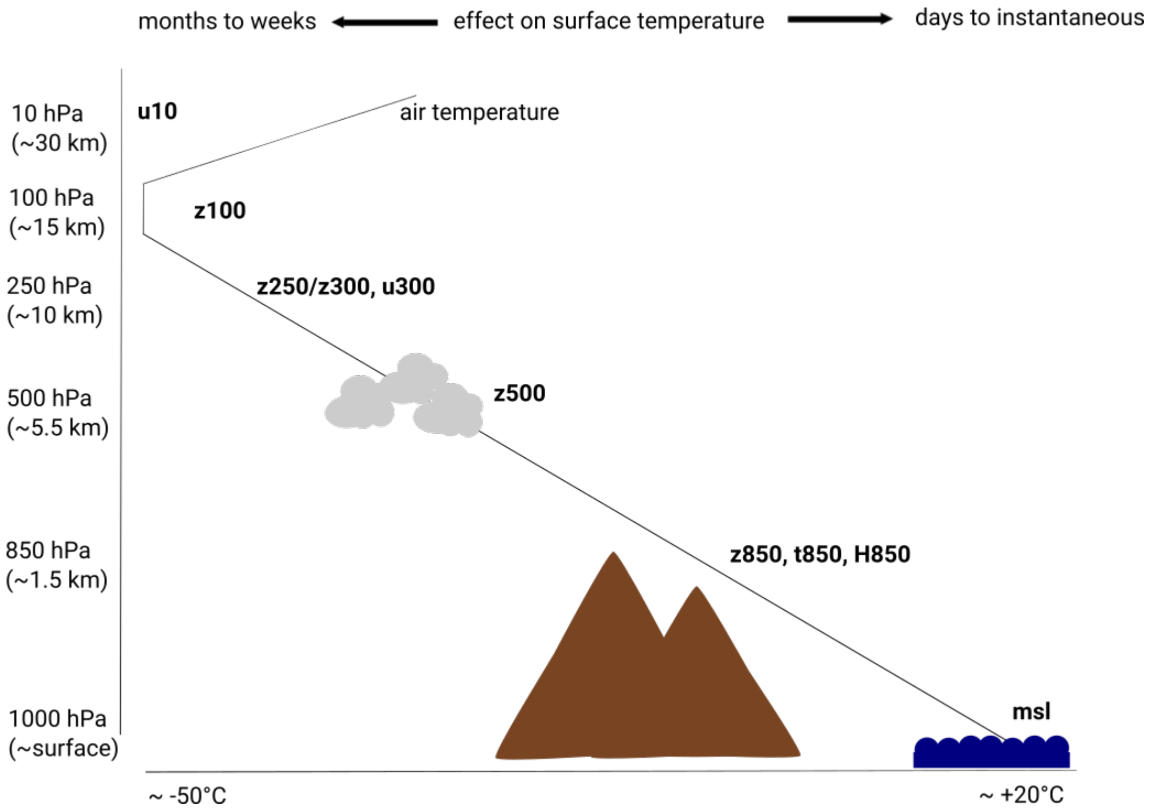


Figure 2.2.: Meteorological variables affecting Central European surface temperature. Depicted are the variables which are used as predictors of the ML models used in this thesis, as also described in section 5.8. In the background, the temperature profile in the troposphere (decreasing temperature up to ca. 100 hPa), at the tropopause (stable temperatures at ca. 100 hPa) and the lower stratosphere (increasing temperature starting from ca. 100 hPa) is indicated.

Besides the large-scale air masses, several more local processes characterize Central European surface temperatures, as formulated in the temperature tendency equation (Kurz, 1990):

$$\frac{\partial T}{\partial t} = -\underbrace{\vec{v}\cdot\nabla T}_{\text{I}} - \underbrace{(\gamma_t - \gamma)\,w}_{\text{II}} + \underbrace{\frac{1}{\mathrm{c_p}}\frac{\mathrm{d}Q}{\mathrm{d}t}}_{\text{III}}. \tag{2.2}$$

It describes the evolution of the temperature T of dry air with time t and can be divided into three terms as indicated. Term I describes the changes of temperature when air is advected. The nabla operator depicts the spatial deviations of temperatures which are advected horizontally with the wind speed $\vec{v}$. An advection of cold temperatures leads to a decrease of temperature over time. An advection of warm temperatures leads to an increase of temperatures. The vertical advection of air masses with the wind speed w is described in term II. γ and γ_t mark thereby the actual and the dry-adiabatic vertical temperature gradient. In case of the latter, only dry air without the exchange of heat or mass with its environment (= adiabatic) is considered. The temperature increases over time when air masses are descending. When air masses are ascending, the temperature decreases. The third process determining temperature changes over time is the accumulation or removal of heat via diabatic processes, described in term III. In this

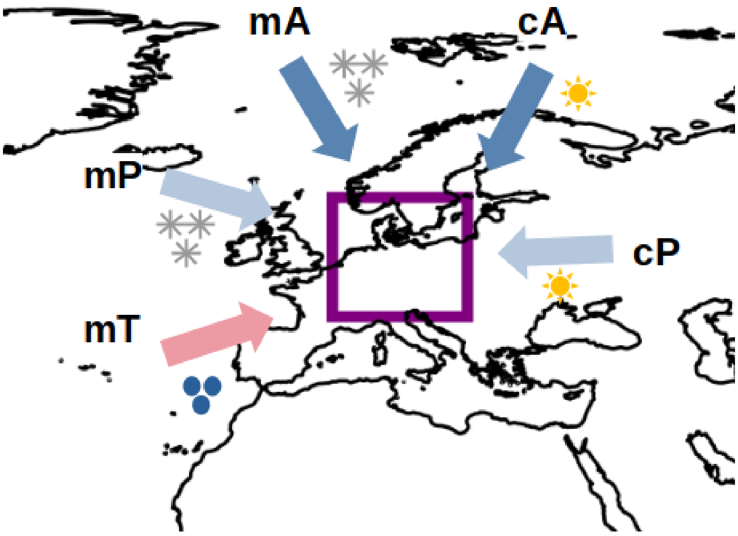


Figure 2.3.: Air masses influencing Central European surface weather in winter. The small letter "c" denotes continental air masses, the letter "m" maritime ones. "P" stands for polar, "T" for tropic and "A" for an Arctic origin of air masses. Colder than average temperatures are denoted by blue arrows (the darker color marks colder temperatures), warmer than usual temperatures by the red arrow. The symbol of the sun marks dry, the three blue circles rainy and the three snowflakes snowy conditions. The purple square shows the region of Central Europe. The map is created using the python package "Cartopy" (Met Office, 2010 - 2015).

case, the exchange of heat and/or mass of the air mass with its environment (= diabatic) is considered. Thereby, c_p denotes the specific heat at constant pressure and Q the amount of heat of the air. An increase in temperatures is observed when heat is accumulated diabatically, a decrease when heat is removed. In case of moist air, the temperature changes in form of latent heat release by condensation and deposition has to be additionally considered. Condensation and deposition processes lead to an increase in temperature, evaporation processes to a decrease.

Although all processes described by the temperature tendency equation (Eq. 2.2) are mostly relevant on timescales shorter than the subseasonal, the horizontal advection of air masses, described in term I, can be useful for subseasonal temperature predictions. Thereby, especially the origin of air masses as described above is of interest (Fig.2.3).

The different types of air masses are usually detected by the temperature ($t850$) and humidity ($H850$) in 850 hPa height (Fig. 2.2). High values of $t850$ are characteristic for air masses originating from the tropics which are transported from the south into Central Europe, low temperatures for air masses originating from the polar and Arctic regions. High values of $H850$ mark air masses which are transported over large water basins, such as the North Atlantic or Arctic Ocean before they reach Central Europe. Low values mark continental air masses. It has to be kept in mind though that the specific humidity can reach higher values below saturation for warmer temperatures. Therefore, air masses transported from the south into Central Europe have higher $H850$ values than air masses transported into Central Europe from the north, even when both air masses pass over the same amount of water beneath.

Humidity itself is also an important factor in determining local temperatures. The specific humidity describes the ratio of the mass of water vapor to the total mass of air. It can be formulated as (Vallis, 2017):

$$H = \frac{r}{1+r}, \text{ whereby} \tag{2.3}$$

$$r = \frac{0.622e}{p-e} \text{ and} \tag{2.4}$$

p the air pressure as well as e the pressure of water vapor. r is also called the mixing ratio. The air pressure is thereby directly proportional to the temperature as shown by the ideal gas equation (Vallis, 2017),

$$p = T\mathrm{R_l}\rho, \tag{2.5}$$

whereby $\mathrm{R_l}$ denotes the ideal gas constant and ρ the density of air.

In an idealized scenario without any humidity or friction, temperatures rise when pressure rises and vice versa. This is an almost instantaneous action. Therefore, the mean sea level pressure (*msl*) plays a non-negligible role in determining temperatures (Fig. 2.2). The same is true for the pressure at higher altitudes. However, in these cases, the geopotential relative to isobaric surfaces, measured in hPa, is usually used in NWP models instead of the pressure itself. The geopotential is defined as the potential energy of an air parcel relative to sea level pressure. It can be formulated as (Vallis, 2017):

$$\Phi = \int_0^z g\mathrm{d}z \simeq gz, \tag{2.6}$$

whereby z is the altitude of the air parcel. High values of geopotential on an isobaric surface denote high pressure systems, low values denote low pressure systems. In general, the geopotential is used as one way to describe the air-flow in the troposphere. Starting at the upper troposphere, the geopotential in 100 hPa ($z100$) depicts the atmospheric circulation in the tropopause region, the stable region between the troposphere and the stratosphere (Fig. 2.2). Here, it is determined whether stratospheric anomalies can propagate downward or not. Stratospheric anomalies which reach the tropopause in winter are usually connected to disruptions of the state of the polar vortex which is represented by the zonal wind in 10 hPa height ($u10$, Fig. 2.2). A downward propagation of these anomalies influences the tropospheric jet stream. Its position can be described using the geopotential in 250 ($z250$) or 300 hPa ($z300$), its velocity by using the zonal wind in 300 hPa height ($u300$, Fig. 2.2). Changes in the location and speed of the jet stream lead in turn to modifications of the large-scale atmospheric flow in the middle and lower troposphere, represented by e.g. the geopotential in 500 and 850 hPa ($z500$ and $z850$, Fig. 2.2). While changes in 500 hPa height need a few days until they influence surface temperatures, the signals detected in 850 hPa can be used to predict weather on a shorter time range.

Another factor determining temperature is the cloud cover, as described in Spiridonov and Ćurić (2021). Below clouds, temperatures are at night warmer than during clear sky conditions and at day colder. This is due to the fact that clouds reflect a large part of the incoming solar radiation on their top during day and a large part of the terrestrial outgoing radiation at their base at night.

Furthermore, in winter, large land-masses cool out faster than the water-bodies. This is due to the higher heat capacity of the ocean compared to land resulting in a higher thermal inertia and thus higher temperatures in winter. Land- or sea-masses that are covered with (bright) snow have a higher albedo, which is the ratio of the reflected to incoming solar radiation, than uncovered. Therefore, temperatures above these regions are also lower (Spiridonov and Ćurić, 2021).

Besides these large-scale variables, also local peculiarities play a huge role in determining the temperature at a specific location, as stated in Spiridonov and Ćurić (2021). Different ways of land use, e.g. forestry or agriculture, create a unique micro-climate. Urban areas with their sealed surfaces and buildings usually experience higher temperatures than rural areas due to differences in heat capacity. Places located in valleys can have regularly extreme cold temperatures at nighttime due to downbursts of cold air from mountain ridges. Places located at the sea benefit from the thermal inertia of the ocean and therefore warmer temperatures than places located in the midst of a continent where extreme cold temperatures can occur in winter.

Despite these, different phenomena in other parts of the globe can influence Central European Wintertime temperatures via teleconnections. These are described in section 2.3.

2.2. Euro-Atlantic Weather Regimes

Based on the demand to structure the atmospheric motions for a deeper understanding, WRs have been developed. By concentrating on deviations of the basic state of the atmospheric flow instead of daily fluctuations, WRs play an important role for weather predictions on the subseasonal timescale (e.g. Grams et al., 2020). WRs are defined as recurring large-scale, quasi-stationary and -persistent atmospheric circulation patterns over a certain area of interest (e.g. Hannachi et al., 2017). In case of Europe, this is the Euro-Atlantic sector.

In the following, three WR definitions for the Euro-Atlantic sector, which are frequently used in literature, are described. These are the regimes containing the two phases of the NAO and two phases of blocking, the classical four regimes forecasted at ECWMF and the year-round seven regimes proposed by Grams et al. (2017) (Grams et al., 2020).

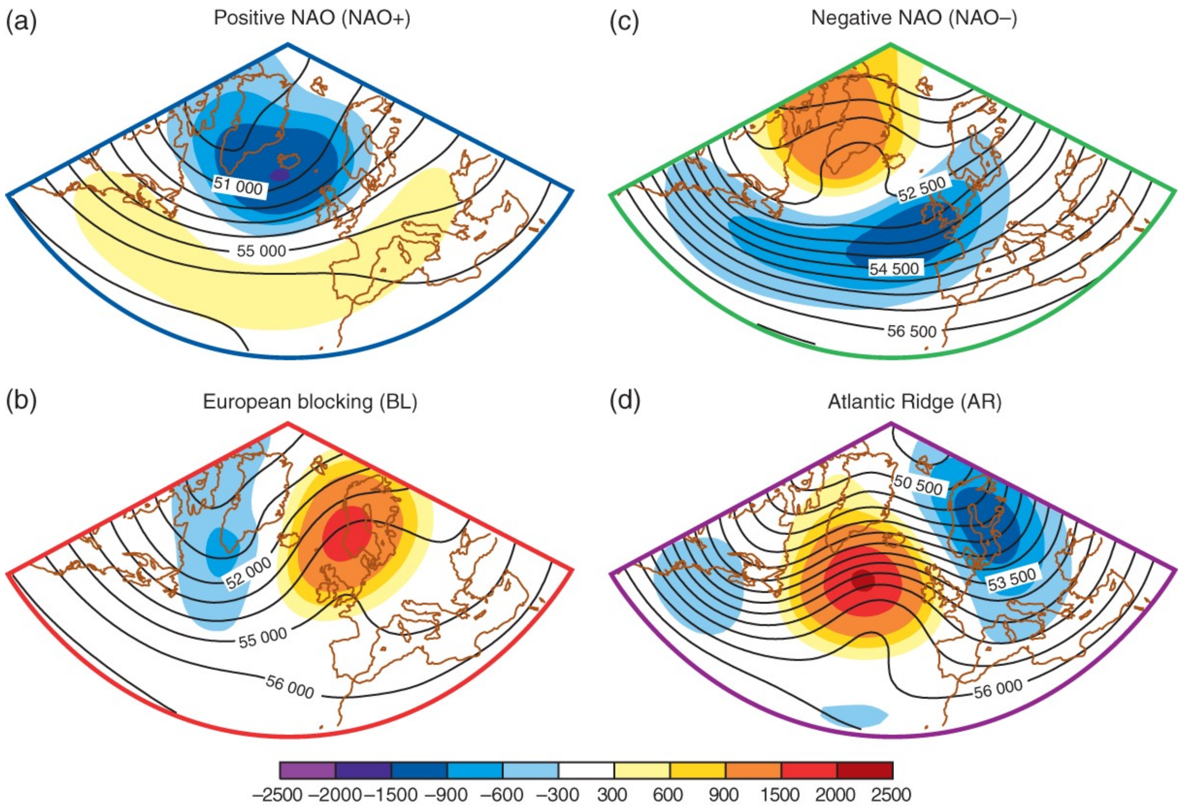


Figure 2.4.: (a–d) Geographical patterns of the four Euro-Atlantic climatological regimes (both anomalies and full fields) for the October to April cold season. The geopotential anomalies (colour shading) and geopotential (contours) at 500 hPa are shown. Figure and caption reprinted from Ferranti et al. (2015), their Fig. 1. ©Royal Meteorological Society.

The NAO is defined over the North Atlantic Ocean and reflects the strength of the two pressure systems located over Iceland and the Azores (Fig.2.4 (a) and (c), e.g. Pinto and Raible (2012)). If both pressure systems are stronger than usual, the meridional pressure gradient is enhanced leading to stronger than usual westerly winds over the North Atlantic Ocean (Pinto and Raible, 2012). This is called the positive state of the NAO (NAO+, Fig.2.4 (a)). If both pressure systems are weaker than usual, the pressure gradient is reduced and the westerly winds over the North Atlantic Ocean lose strength (Pinto and Raible, 2012). Often, the large-scale atmospheric flow has thereby a pronounced meridional component over Europe (Benedict et al., 2004). This configuration is called the negative state of the NAO (NAO- , Fig.2.4 (c)).

The development of the two phases of the NAO results from the breaking of Rossby-Waves (RWB) over the North Atlantic Ocean, as described in Benedict et al. (2004). Usually, atmospheric Rossby-Waves propagate from west to east in the northern hemispheric mid-latitudes. However, disturbances caused by e.g. orography lead to the formation of ridges and troughs in the zonal flow as described in section 2.1.

These disturbances sometimes strongly intensify leading to wave breaking. As described in Tamarin-Brodsky and Harnik (2024), RWB is characterized by an irreversible overturning of air parcels with the same PV leading to the inversion of the meridional PV gradient. The subsequent exchange of momentum with the surrounding of an air parcel can lead to changes in the jet stream's location and speed. As stated in Benedict et al. (2004), this can be detected in the potential temperature at the tropopause. It is used since RWB leads to a change in sign of the potential temperature and atmospheric large-scale patterns such as the NAO phases are depicted as a dipoles (Fig. 2.5 (d) for NAO+ and 2.6 (d) for NAO-).

Anticyclonic RWB close to the western coast of North America and subsequently over the subtropical North Atlantic Ocean starts the development of the NAO+ phase (Fig. 2.5 (a) and (b)). As a result of the first RWB, cold air is transported towards Greenland with the underlying westerly flow of the northern hemispheric mid-latitudes. During the second RWB, warm air is advected northward into the central North Atlantic Ocean (Fig. 2.5 (b) and (c)). The characteristic pattern of the NAO+ with cold air over the northern North Atlantic Ocean and warm air over the southern North Atlantic Ocean is established (Fig. 2.5 (d)). Consequently, the temperature and pressure gradients are enhanced which leads to an intensified zonal jet stream over the North Atlantic Ocean.

In contrast to the development of the NAO+ pattern, the evolution of the NAO- pattern starts by one cyclonic RWB over the North Atlantic Ocean (Fig. 2.6 (a) to (c)). During this, warm air is advected northwards and cold air southwards (Fig. 2.6 (d)). Since the advected air masses mix with the surrounding air, the north-south temperature and pressure gradients are weakened. The jet stream is decelerated and has a meridional component over Europe.

Besides the two phases of the NAO, the WR classification of Ferranti et al. (2019) uses the occurrence of a trough (BLO-) or ridge (BLO+, comparable to Fig. 2.4 (d)) in the Rossby-Wave structure over Scandinavia. During BLO+, which describes a so-called blocking situation caused by a quasi-stationary and -persistent ridge over Scandinavia, either cold polar air masses are transported southward or cold continental air masses are transported westward into Europe often leading to cold waves. During BLO-, mild winter temperatures prevail.

The four classical WRs which are predicted by the ECMWF on the subseasonal timescale are quite similar to the ones of Ferranti et al. (2019) but obtained in a different way. While Ferranti et al. (2019) simply use the attribution of the first and second Empirical Orthogonal Function (EOF), the classical four regimes are obtain by applying additionally an EOF-clustering. A description of how EOFs and clustering are obtained is provided in section 5.4.

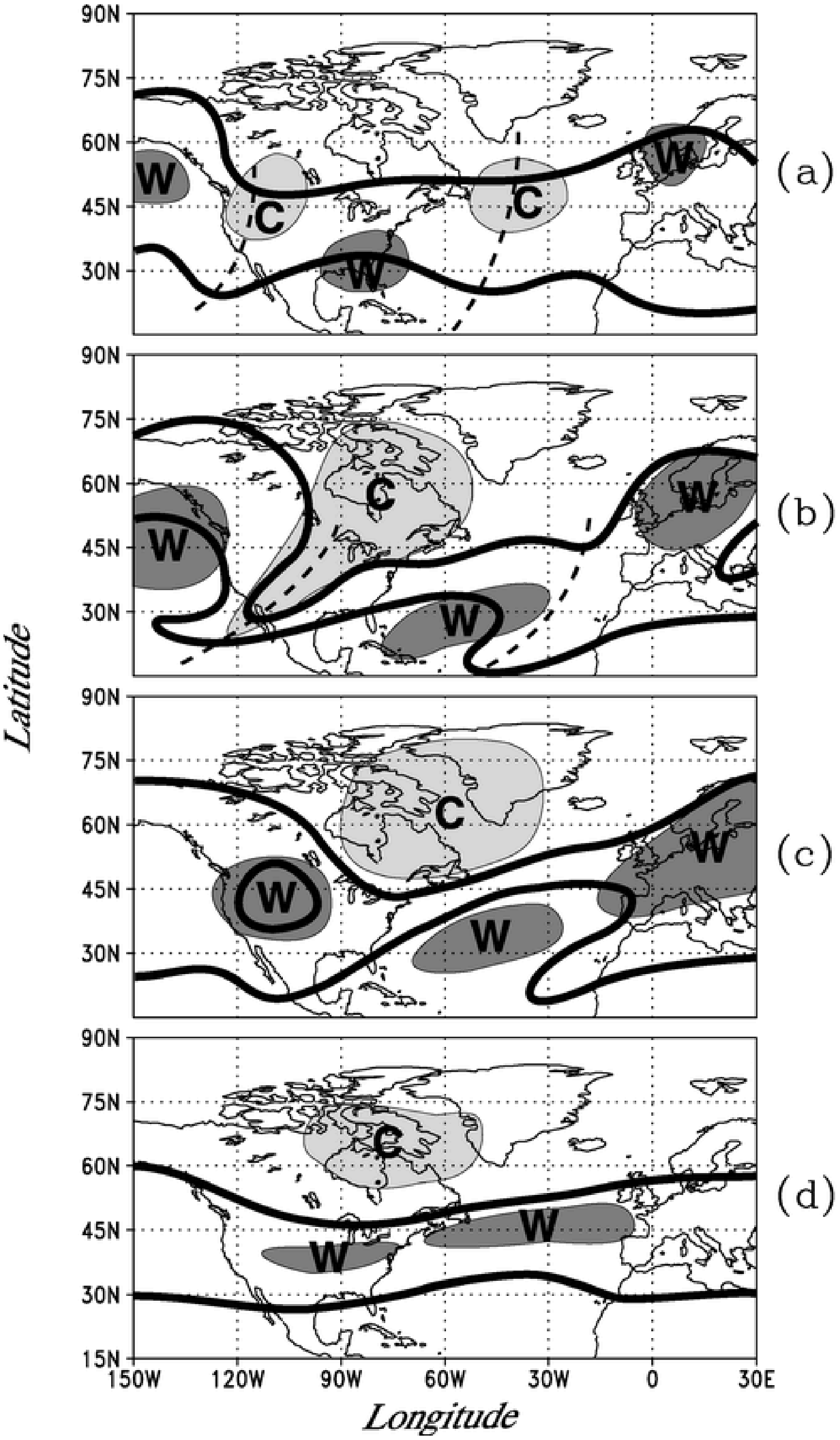


Figure 2.5.: (a)–(d) A schematic diagram depicting the generalized features and flow patterns of the positive NAO phase. Each frame captures the atmospheric features on 3–5-day increments, with (c) representing the zero-lag day. The thick contours are for the total flow, with the northern (southern) contour corresponding to ca. 305 K (ca. 335 K). The warm and cold air indicated in this figure correspond to anomalies. The dashed curves indicate the trough axes. Figure and caption reprinted from Benedict et al. (2004), their Fig. 7. ©American Meteorological Society. Used with permission.

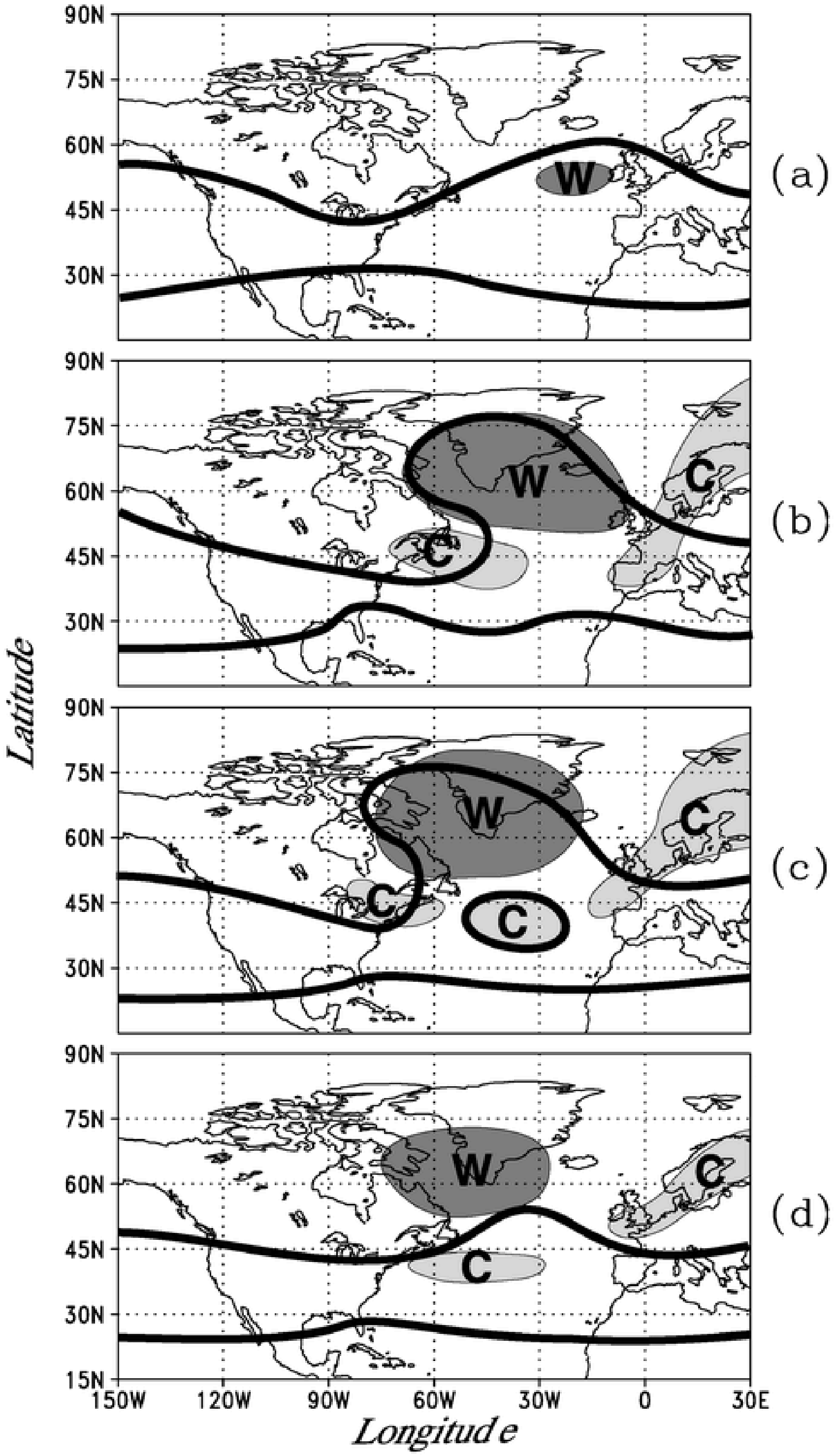


Figure 2.6.: (a)–(d) A schematic diagram depicting the generalized features and flow patterns of the negative NAO phase. As in Fig. 2.5, the time increment of each panel is approximately 3–5 days, with (c) again representing the atmospheric features corresponding to the zero-lag day. Figure and caption reprinted from Benedict et al. (2004), their Fig. 8. ©American Meteorological Society. Used with permission.

The four classical WRs include both phases of the NAO (Fig. 2.4 (a) and (c)), European blocking (EuBL) (Fig. 2.4 (b)) and the Atlantic Ridge (AR) regime (Fig. 2.4 (d)). The two latter are also part of the WR classification proposed by Grams et al. (2017).

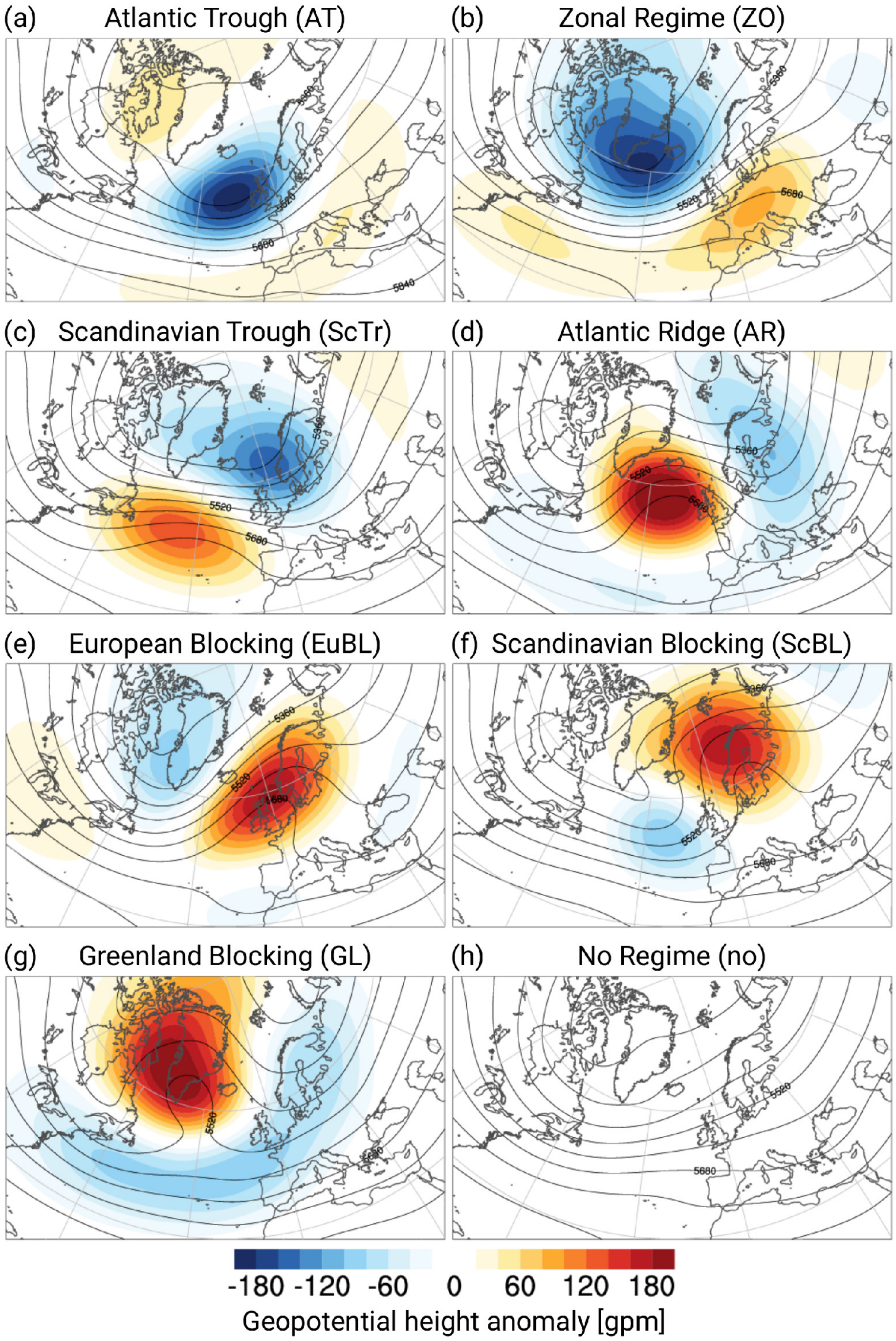

(a) Atlantic Trough (AT)
(b) Zonal Regime (ZO)
(c) Scandinavian Trough (ScTr)
(d) Atlantic Ridge (AR)
(e) European Blocking (EuBL)
(f) Scandinavian Blocking (ScBL)
(g) Greenland Blocking (GL)
(h) No Regime (no)
-180 -120 -60 0 60 120 180
Geopotential height anomaly [gpm]

Figure 2.7.: Cluster mean 500-hPa geopotential height (contours; gpm, i.e., geopotential meters) and corresponding anomalies (shading; gpm) of the seven year-round Atlantic–European weather regimes (defined based on ERA-Interim data between 1979 and 2018) and the "no regime" category. Figure and caption reprinted from Büeler et al. (2021), their Fig. 1.

This classification, described in detail in section 5.4, also uses EOFs and clustering to obtain the WRs. In contrast to the four classical WRs, seven WRs are obtained and the patterns projecting weakly onto the seven EOFs are summarized in a separate class called the "No" regime. The latter comprises thereby all cases which are closer to climatology than to any of the other seven WR patterns. The seven WRs are the Zonal regime (ZO), the Scandinavian Trough (ScTr), the Atlantic Trough (AT), AR, EuBL, Scandinavian Blocking (ScBL) and GL. The first three are called cyclonic regimes. As shown in Büeler et al. (2021), they are characterized by pronounced low pressure systems inside troughs and a zonal airflow across the North Atlantic Ocean (Fig. 2.7 (a)-(c)). In winter, relatively mild air masses are transported towards Europe when these regimes are present. During the ZO regime, the low pressure system is found between Iceland and southern Greenland (Fig. 2.7 (b)). During the ScTr regime it is located between Iceland and Scandinavia (Fig. 2.7 (c)), and in case of the AT regime (Fig. 2.7 (a)), it is found west of the British Isles. The other four regimes are called anticyclonic or blocked regimes. These are characterized by a pronounced high pressure system associated with a strong ridge and a meridional air-flow over the North Atlantic-European sector (Fig. 2.7 (d)-(g)). In winter, this leads to relatively cool air masses approaching Europe. During the AR regime, the high pressure system is found south of Iceland (Fig. 2.7 (d)). In case of EuBL, it is located over southern Scandinavia and the British Isles (Fig. 2.7 (e)). During the ScBL regime, the high pressure is system is located over northern Scandinavia (Fig. 2.7 (f)), and, in case of GL, over the Labrador Sea and south of Greenland (Fig. 2.7 (g)). The NAO itself is not included in the definition of the seven WRs but several of the WRs project onto it. These are in case of the NAO+ the ZO and ScTr regime and in case of the NAO- the GL regime (Domeisen et al., 2020).

2.3. Teleconnections

Teleconnections describe the lagged relation of phenomena occurring at different geographical locations which are often several thousand kilometers apart from each other. Thereby, one phenomenon causes recurring, quasi-persistent large-scale flow anomalies which influence the state of the atmosphere at a remote region via wave-propagation (Wallace and Gutzler, 1981). This influence is often non-stationary and stronger during some years than others (Knippertz et al., 2003). Since there are many teleconnections possibly affecting European surface temperatures in winter, only a selection of these is described in this section (as shown on Fig. 2.8).

A well-known teleconnection is the influence of the state of the stratospheric polar vortex on European surface weather in winter (e.g. Domeisen et al., 2020; Kautz et al., 2020) as mentioned in section 1 (light red arrow on Fig. 2.8). The stratospheric polar night jet which encircles the stratospheric polar vortex, is a circumpolar band of high wind speed occurring during polar night in the stratosphere. Under normal conditions, the wind direction is westerly and the stratospheric polar night jet located approximately circularly around the pole (Butler et al., 2017). However, as described in Charlton and Polvani (2007), roughly every second year a major disruption of the stratospheric climatological state occurs due to wave-breaking and subsequent redistribution of momentum and mass (Limpasuvan et al., 2004). This

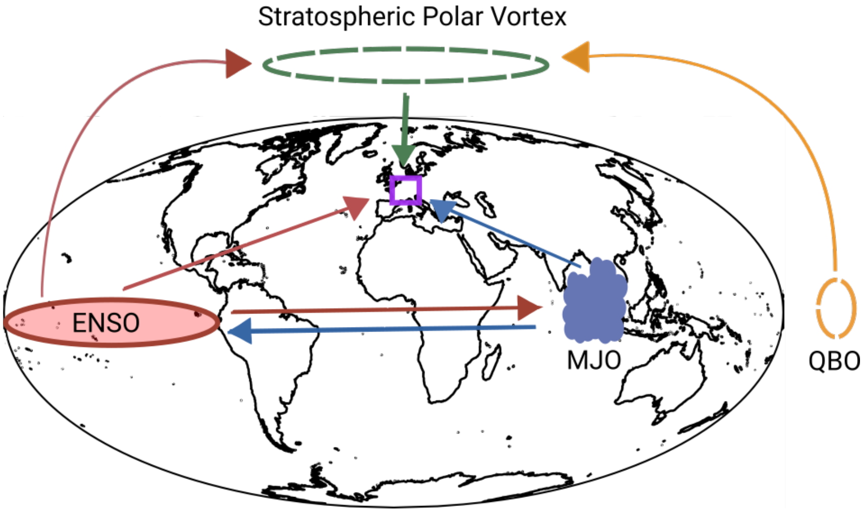


Figure 2.8.: Visualization of teleconnections affecting Central European surface temperatures in winter. The stratospheric polar vortex (represented by the circular broken green line) has a downward influence (green arrow) on Central European (purple square) surface temperatures. The El Niño-Southern Oscillation (ENSO, represented by the red circle) directly influences Central European surface temperatures and has additionally an influence on the stratospheric polar vortex and the Madden-Julian Oscillation (MJO) (red arrows). The MJO (represented by the blue clouds) influences in turn ENSO and European surface temperatures (blue arrows). The influence of the tropical phenomena on the stratosphere is modulated by the Quasi-Biennual-Oscillation (QBO, represented by the circular broken yellow line, yellow arrow). Note: the arrows are pointing qualitatively from the origin to the target of the teleconnection and are not necessarily representing the actual pathway. The map is created using the python package "Cartopy" (Met Office, 2010 - 2015).

manifests in an SSW event which is characterized by an increase of temperatures in the stratosphere by on average 7.8 K accompanied by a weakening and deformation of the stratospheric polar vortex. In extreme cases, the wind direction of the polar night jet is reversed to easterly and the vortex itself is either displaced from its concentric position around the pole or split up into two parts. The resulting anomalies can propagate downward in time until the tropopause is reached. Here, a downward influence of the stratospheric anomalies on the tropospheric large-scale circulation can take place when conditions are favorable. Thereby, as stated in Limpasuvan et al. (2004), the tropospheric large-scale flow is disturbed by changes in pressure which are induced by the stratospheric temperature anomalies according to the ideal gas equation (Eq. 2.5). Following Eq. 2.5, the pressure over the pole is increased due to the positive stratospheric temperature anomalies. Due to the developing pressure gradient between the pole (higher pressure) and the mid-latitudes (lower pressure), a compensating movement of cold polar air into the mid-latitudes is induced. Due to the downward movement of this relatively colder air in the mid-latitudes, temperatures at the surface decrease according to the temperature tendency equation (Eq. 2.2). The teleconnection of an SSW event to Central European surface temperatures is established. According to Domeisen et al. (2020), the lag between the surface response and the SSW event is thereby usually between ten and 60 days but a downward influence does not always occur.

As stated in Limpasuvan et al. (2004), the determining factor whether the stratospheric anomalies are able to influence the troposphere and thus set up a teleconnection is the tropospheric large-scale flow configuration. A coupling probably only takes place if the circulation anomalies in the lower stratosphere project onto existing large-scale flow pattern in the troposphere. This is due to the fact that the mass of the stratosphere is a lot less than the mass of the troposphere (at best one third of it in winter for the extratropical atmosphere). This leads to higher forcings in the troposphere which can easily super-seed any stratospheric influence. A particular sensitive area for the downward propagation of stratospheric anomalies into the troposphere is the region over the North Atlantic Ocean (Charlton-Perez et al., 2018). Since the resulting circulation pattern after an SSW resembles the NAO- over the North Atlantic Ocean, the stratospheric anomalies project well on the tropospheric large-scale configuration if NAO- conditions are present and a downward propagation of stratospheric anomalies with a subsequent influence on surface weather is possible (Limpasuvan et al., 2004).

As shown by Domeisen (2019), 2/3 of the SSW events observed between 1958 and 2014 are followed by an NAO- pattern in the troposphere. Nevertheless, as shown by Charlton-Perez et al. (2018), also the probability of a transition to an NAO- pattern is increased by 40-60% after a reduction of the strength of the stratospheric polar vortex. Following anomalously strong wind speeds of the stratospheric polar vortex, an increase in 10-30% of the probability of a transition to an NAO+ pattern is predicted by their logistic regression model.

Besides teleconnections from the stratosphere, influences from tropical phenomena on the state of the NAO are observed. One example therefore is the influence of ENSO (red arrows on Fig. 2.8). ENSO describes recurring sea surface temperature (SST) anomalies in the tropical Pacific Ocean on a varying timescale of two to seven years (Jiménez-Esteve and Domeisen, 2018). It is divided into three phases, one with warmer than usual SSTs, one with SSTs close to average and one featuring colder than average SSTs in the Central Pacific Ocean. According to Williams et al. (2023), mostly the warm phase of ENSO, El Niño, shows an influence on the NAO. The influence of the cold phase, La Niña, is weak. The teleconnection of ENSO to the NAO is accomplished via wave-propagation. Both phases of ENSO lead to a shift of the Tropical Pacific convection which is accompanied by the generation of Rossby-Waves. The planetary waves propagate poleward in the upper troposphere leading to disturbances in the mean westerly flow of the mid-latitudes. These disturbances then affect the development of an NAO pattern over the North Atlantic Ocean as describend in section 2.2. Following El Niño, NAO- is observed when a teleconnection is established. The teleconnection is thereby strongest in late winter. Following La Niña conditions, NAO+ conditions occur in case the forcing is strong enough (Jiménez-Esteve and Domeisen, 2018). In addition to the tropospheric teleconnection of ENSO to the NAO, also a two-way pathway via the stratospheric polar vortex exists (Fig. 2.8). According to Jiménez-Esteve and Domeisen (2018), the strength of the polar vortex is influenced by ENSO and thus also possible teleconnections to the NAO.

Thereby, during El Niño conditions a weakening of the stratospheric polar vortex is observed and during La Niña conditions a strengthening.

The teleconnection pathway of ENSO via the stratospheric polar vortex to the state of the NAO is additionally influenced by the Quasi-Biennual Oscillation (QBO, light purple arrow on Fig. 2.8) which describes the recurring change of the main wind direction in the equatorial stratosphere (Hansen et al., 2016). As stated in Hansen et al. (2016), the easterly phase of the QBO, i.e. when easterly winds are present in the equatorial stratosphere, in combination with El Niño conditions in the Central Pacific Ocean supports the establishment of an NAO- pattern over the North Atlantic Ocean. Accordingly, the combination of the westerly phase of the QBO in combination with La Niña conditions in the Central Pacific Ocean supports the development of an NAO+ pattern.

Another interaction of a tropical phenomenon which ultimately influences the state of the NAO is the interplay between ENSO and the Madden-Julian-Oscillation (MJO, blue arrows on Fig. 2.8). The MJO describes the large-scale eastward propagation of air along the equator, which manifests in a movement of convection between the Indian Ocean and the Central Pacific Ocean on timescales of 30 to 60 days (Lin et al., 2009). It is thereby divided into eight phases according to the position of the convection. As stated in Vitart et al. (2017), the probability of an NAO+ pattern is enhanced ten days after the MJO is in phase three (strong convection over the Indian Ocean) and decreased ten days after the MJO is in phase six (convection over the Western Pacific). In case of the NAO-, it is vice versa. This relation, however is additionally modulated by ENSO (Fig. 2.8). As shown in Lee et al. (2019), the connection between the MJO and NAO+ is strengthened during El Niño conditions and weakened during La Niña conditions. The connection between the MJO and NAO- is on the opposite weakened during El Niño conditions and strengthened during La Niña conditions. As in all described teleconnection processes in this section, the influence of the remote phenomenon on the NAO is achieved via Rossby-Wave propagation.

3. State-of-the-Art in Subseasonal Forecasting

In this chapter, state-of-the-art approaches for generating subseasonal weather forecasts are discussed. Besides the traditional NWP models, the focus lies on parametric and non-parametric postprocessing approaches as well as the application of ML models. Furthermore, the use of WRs as a source of subseasonal predictability is described.

3.1. Numerical Weather Prediction Models

NWP models represent the dynamics and physical processes of the atmosphere by discretized numerical equations and parameterizations (Owens and Hewson, 2018). In order to solve these numerical equations, NWP models rely heavily on the initial and boundary conditions given (Vitart and Robertson, 2018). As described by e.g. White et al. (2017) and Vitart and Robertson (2018) and shown on Fig. 1.1 (a), the initial conditions, e.g. today's mean sea level pressure or the mid-tropospheric circulation, comprise predictability of weather on the short- (up to three days lead time) to medium-range (up to ten days lead time). The boundary conditions, e.g. sea surface temperatures, provide predictability on the seasonal timescale (starting at 30 days lead time). However, both are not the relevant drivers for weather at the subseasonal timescale. Instead, these are teleconnections and large-scale feedbacks in the climate system which are usually not entirely represented by NWP models, sometimes also due to missing scientific knowledge.

As stated by Kautz et al. (2020) and shown on Fig. 3.1, teleconnections can lead to a shift in the probability of forecast outcomes. Instead of an equal probability of both positive and negative extremes, the probability of one of these is enhanced at a lead time around one month in the future. This means that the ensemble mean forecast is shifted from a neutral (climatological) forecast outcome towards one of the extremes and the forecast spread is narrowed. A probabilistic forecasting skill on the subseasonal timescale (called "extended-range" on Fig. 3.1) is achieved. For example, when a teleconnection from an SSW is established, the probability of colder than usual surface temperatures is enhanced. A further narrowing of the forecast spread and shift of the ensemble mean prediction is firstly observed when the predictive information in the initial conditions of the NWP model gets relevant for forecasting. This, however, is on timescales shorter than subseasonal.

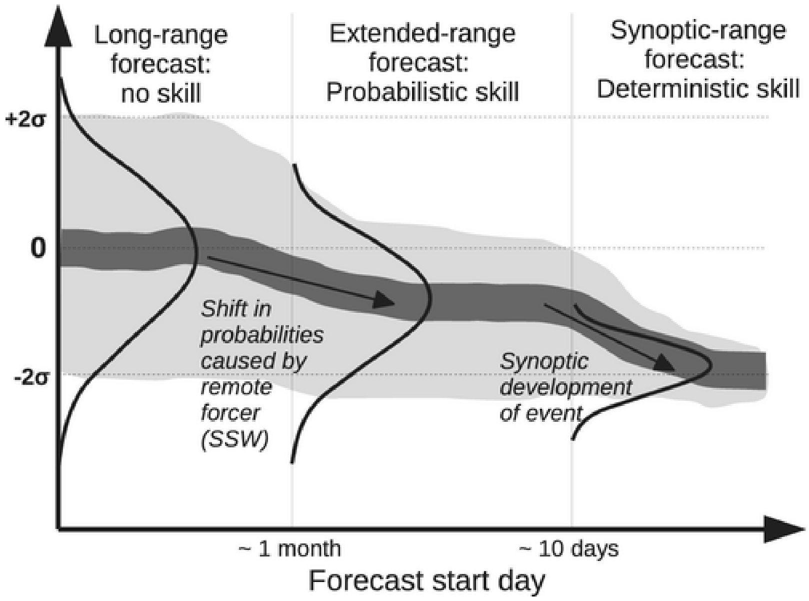

Long-range forecast: no skill
Extended-range forecast: Probabilistic skill
Synoptic-range forecast: Deterministic skill
+2σ
0
-2σ
Shift in probabilities caused by remote forcer (SSW)
Synoptic development of event
~ 1 month
~ 10 days
Forecast start day

Figure 3.1.: Schematic showing the predictability of a surface extreme event for different forecast ranges (from long-range forecasts with lead times of 1 month to synoptic-range forecasts with lead times of only a few days) under the influence of remote forcing (e.g., SSW) occurring during the extended range. Light grey shading indicates the ensemble spread (5th to 95th percentile), while dark grey shading indicates the ensemble mean. Figure and caption reprinted from Kautz et al. (2020), their Fig. 10.

Since teleconnections are are complicatedly interacting phenomena which depend on many factors (chapter 2.3), they pose a challenge for NWP models in both, representing them and extracting the information needed for correct subseasonal weather predictions. Since many teleconnections and large-scale climate feedbacks manifest themselves in changes in the atmosphere and ocean (e.g. Mariotti et al., 2020), most NWP models used for subseasonal forecasting are coupled atmosphere-ocean models and some also include a separate, coupled sea-ice model (Vitart et al., 2017).

In the following, a few state-of-the-art NWP models used operationally for subseasonal forecasting are described according to Vitart et al. (2017). Relevant changes in more recent years are indicated. As shown in Vitart et al. (2017), besides variations in model components, also forecasting frequencies and ensemble sizes differ substantially between the different weather services reaching from daily to weekly initializations and ensemble sizes between four and 51 ensemble members (as of 2017). While e.g. the UK Met Office (UKMO) releases subseasonal forecasts on a daily basis with an ensemble size of four, ECWMF started recently to provide subseasonal forecasts daily with an ensemble size of 101 (as of September 2024[1]). Also the horizontal and vertical resolution differs between models. The lowest resolution, which is $2^\circ \times 2^\circ$ horizontally and 17 vertical layers, is used at the Australian Bureau of Meteorology (BoM). The model with the highest resolution with $0.5^\circ \times 0.5^\circ$ horizontally and 91 vertical layers is used at ECMWF (in September 2024, the resolution is roughly 32 km and in the vertical, 137 layers are

[1] https://apps.ecmwf.int/datasets/data/s2s-realtime-instantaneous-accum-ecmf/levtype=sfc/type=pf/, last viewed 2 September 2024

used (Mladek, 2024)). Concerning the forecast length, BoM calculates the longest lead time of 62 days. However, meanwhile, forecasts with a lead time of 65 days are also available from a newer model not mentioned in Vitart et al. (2017)[2]. The shortest lead time with 32 days (as of September 2024[3]) is computed at the Italian Institute of Atmospheric Sciences and Climate of the National Research Council.

Besides the operational forecasts, the reforecasts (also called hindcasts) play an important role in subseasonal forecasting as they are used for model calibration. Reforecasts can be computed with a fixed model version or on the fly, which means that reforecasts are computed continuously for past dates with the current operational model version. The National Centers for Environmental Prediction (NCEP) e.g. uses a fixed model version to calculate several years of reforecasts. Their reference period as of 2017 are the years 1999-2010. At ECMWF, reforecasts are produced on the fly for the past 20 years. Similar to the operational subseasonal forecasts, also the frequency and ensemble size of the reforecasts differs between models reaching from daily to three times a month and from four ensemble members (as of September 2024[4]) to 33. For example, NCEP computes four reforecasts daily, ECMWF 11 but only bi-weekly and BoM 33 reforecasts six times a month.

In comparison to short- and medium-range forecasts, the amount of subseasonal (re-)forecasts is overall small (Vigaud et al., 2018). Therefore, multi-model approaches are a common tool to increase ensemble sizes although their creation is not trivial due to the different forecast initializations and resolutions used at the different weather services. For example, at the WMO Lead Centres for Sub-Seasonal Prediction Multi-Model Ensemble, operational multi-model subseasonal forecasts are offered (WMO, 2024). Also the North American Multimodel Ensemble is an operationally available multi-model ensemble which provides improved forecasts of the 3-month-mean probabilistic 2-meter temperature forecast, sea surface temperature and precipitation rate in comparison to the operational subseasonal forecast of NCEP (Climate Forecast System version 2) alone (Becker and Dool, 2016). Li et al. (2021) show that a multi-model ensemble of eight models yields better subseasonal forecasts of precipitation for the contiguous United States then each of the models individually. The skill of the multi-model ensemble is further improved when postprocessing, in their study EMOS, is applied.

Multi-model ensembles can be created in different ways. For example, Karpechko et al. (2018), calculate at first anomalies based on the climatology of each model, which is obtained from the respective reforecasts, separately and then combine each ensemble member into a large ensemble containing all models' ensemble members. The multi-model mean is then calculated as a weighted average to account for the

[2] https://apps.ecmwf.int/datasets/data/s2s-realtime-instantaneous-accum-anso/levtype=sfc/type=cf/

[3] https://apps.ecmwf.int/datasets/data/s2s-realtime-instantaneous-accum-isac/levtype=sfc/type=pf/, last viewed 2 September 2024

[4] e.g. https://apps.ecmwf.int/datasets/data/s2s-reforecasts-instantaneous-accum-kwbc/levtype=sfc/type=pf/, last viewed 2 September 2024

different ensemble sizes of the considered models. Another method to create a multi-model ensemble is using only the ensemble means of the different models. In this case, though, the ensemble size is smaller and the model spread not considered.

Important to keep in mind is that the skill of models used for multi-model ensembles can be substantially different from each other. When combining models with very different skill levels, the improvement in the multi-model ensemble can be relatively little compared to the skill of the best model alone. This is especially true, when the ensembles of the various models have been postprocessed before (WMO, 2021).

3.2. Weather-Regime Applications

Besides using the information inherent in teleconnections to achieve a probabilistic forecasting skill on the subseasonal timescale (Fig. 3.1), WRs can be utilized. As shown by e.g. Mayer and Barnes (2021) forecasting the large-scale atmospheric circulation can be a successful alternative to forecasting surface weather directly. Often, these large-scale pattern are more predictable than the individual fields since they are quasi-stationary and -persistent by definition (Osman et al., 2023). Furthermore, the relation of these pattern to surface weather is well established. For example Richardson et al. (2020) demonstrate that the subseasonal forecast of 30 common weather patterns for the UK and the subsequent inference of whether there is a drought or not is more accurate than forecasting the drought directly.

Robertson et al. (2020) show in a study using North American WRs that these can be used to find windows of opportunity where subseasonal forecast skill over North America is enhanced. This is the case, when certain WRs are predicted correctly well in advance. On the other hand, they also show that predictions of surface weather, in their study precipitation in California, can be very poorly in case the forecast of the WRs is incorrect.

In general, the predictability between various WRs differs, as shown by e.g. Osman et al. (2023), in their case for the Euro-Atlantic WRs. Their findings are based on the seven WRs proposed by Grams et al. (2017) and the predictions of three NWP models including ECMWF's S2S reforecasts. They find that the year-round subseasonal predictability is highest for the GL regime with an especially high predictability in winter and lowest for the EuBL and ScBL regime. The predictability is thereby determined by the number of days into the future for which skillful forecasts are possible.

Applying postprocessing methods to these forecasts can enhance predictability. For example Mockert et al. (2024) use an ensemble postprocessing technique to improve ECMWF's S2S reforecasts of the seven WRs proposed by Grams et al. (2017). By doing so, the reforecasts are able to predict the respective WR skillfully up to 14.5 days in the future.

Another approach of using postprocessing in combination with WRs is done by Allen et al. (2019). They show that postprocessing in dependence of the WR is leading to a higher forecast skill in comparison to postprocessing applied to all forecasts similarly. Thereby, they investigate lead times up to 15 days and use a highly idealized model which simulates the chaotic nature of the atmosphere. The used WRs are also calculated based on this model. Using the WRs present at the initialization lead to a slightly lower improvement of forecast skill than using the WRs present at the target date. Nevertheless, they add for consideration that the WR present at the target date is subject to forecast uncertainty whereas the WR present at initialization time is not. Furthermore, they suggest that their postprocessing approach can also significantly enhance weather predictions of "real" NWP models, especially, when WRs are clearly defined or in case of extreme weather events associated with specific WRs. This suggestion is supported by their following study, Allen et al. (2021), for the application of a weather-regime-dependent postprocessing of 10 m wind speed forecasts for the United Kingdom, shown for forecasts with lead times up to six days.

The results of these studies can be used by forecasters to improve the postprocessing of NWP forecasts and assess the reliability of predictions not only in terms of the forecasted WR but also in terms of the forecasted surface weather. As for now, operationally available as part of the ECMWF's subseasonal forecasts are only the predictions of the four classical WRs, which are described in section 2.2, but the use of several WR definitions in a complementary way is recommended (Grams et al., 2020).

3.3. Postprocessing Approaches for Numerical Weather Forecasts

The skill of subseasonal NWPs can be improved by postprocessing, as already mentioned e.g. in section 3.1. One reason therefore is the observation that model forecasts are drifting towards the model's climatology after a certain lead time. On the subseasonal timescale, this drift is significant (Owens and Hewson, 2018). Removing it from the predictions is one way of using postprocessing techniques to improve subseasonal forecast skill. Postprocessing is thereby done e.g. in the form of a lead-time-dependent mean bias correction. This technique uses observations to identify biases of past predictions which are then used for correcting the ensemble members of future predictions. The identification of biases can thereby be achieved in various ways.

Generally, the used methods are categorized as deterministic or probabilistic. Deterministic postprocessing approaches are used to bias-correct deterministic predictions or the ensemble mean of probabilistic forecasts. Probabilistic approaches take furthermore the ensemble information into consideration by e.g. postprocessing not only the mean of the ensemble but also its standard deviation. For example, Monhart et al. (2018) compare methods from both categories for bias-correcting subseasonal temperature and precipitation forecasts for Europe. As a deterministic approach, they use a local polynomial regression for mean bias identification which is equivalent to a moving average when the degree of the polynomial is zero. As the probabilistic approach, they use quantile mapping. Thereby, the statistical distribution of the forecasts is adjusted to match the statistical distribution of the observations. In their study, they find that generally the probabilistic postprocessing method leads to higher forecast improvements than the deterministic one.

Korhonen et al. (2020) use in their study a combination of a probabilistic lead-time-dependent mean bias correction of ECMWF's S2S reforecasts of surface temperatures in combination with an additive temperature term depending on the atmospheric conditions. Hereby, they focus on wintertime northern European 2-meter temperatures in the bi-weekly mean of weeks three and four as well as five and six.

Other postprocessing approaches use e.g. linear or logistic regressions to find relations between observations and predictions of the past which are then used to correct forecasts (Vannitsem et al., 2021; Taillardat et al., 2016). For example, Silini et al. (2022) yield improved forecasts of the MJO by applying a multiple linear regression to ECMWF's subseasonal MJO predictions. These and several other non-ML approaches are summarized as MOS in case of deterministic forecasts and EMOS (Gneiting et al., 2005) in case of ensemble forecasts. Hyvärinen et al. (2021) use MOS and EMOS on the subseasonal timescale to improve the skill of ECMWF's S2S reforecasts for wind speed predictions over Finland. Thereby, the EMOS method, also called non-homogenous Gaussian regression (NGR), is used to postprocess the weekly mean of the reforecasts. It assumes a Gaussian forecast distribution where the mean μ is given by

$$\mu = a + b\bar{x} \tag{3.1}$$

and the standard deviation σ by

$$\log\sigma = c + d\,s. \tag{3.2}$$

The parameters a, b, c and d are fitted by the NGR and $\bar{x}$ and s denote the reforecast mean and standard deviation, whereby the distribution of means is assumed to be Gaussian.

NGR is also used by van Straaten et al. (2020) in their study of subseasonal temperature forecasts for Europe. They show that ECMWF's S2S forecasts yield on average three more days of skillful predictions when postprocessing is applied.

A variant of NGR is developed by Dirkson et al. (2022) to postprocess subseasonal sea-ice forecasts for the Canadian Arctic. They use a doubly censored normal distribution which means that at specific thresholds the Gaussian distribution used by the NGR is replaced by a point mass. These thresholds are in their case zero and one referring to 0% and 100% sea-ice coverage.

Most of the studies described above use temporal and/or spatial aggregation of (re-)forecasts. It is shown by van Straaten et al. (2020) that temporal aggregation in the form of rolling means can be a useful tool to distinguish better predictable large-scale weather pattern from the highly fluctuating daily weather. This way, predictability in forecasts can be extended up to a lead time of four weeks, provided the persistence of predictable large-scale weather pattern. Vijverberg et al. (2020) increase the skill of subseasonal predictions of Eastern U.S. hot temperature events by using the mean of forecasts in fixed time windows of 15 days. Furthermore, van Straaten et al. (2020) show that besides a temporal aggregation of forecasts, also a spatial aggregation or a combination of both can increase forecast skill on the subseasonal timescale.

3.4. Machine Learning Applications

ML models are data-driven statistical models which predict a target variable based on given predictors, often from the past. In contrast to NWP models, ML models do not rely on physical equations which makes them a promising approach for forecasting on the subseasonal timescale where predictability is currently not captured by the numerical equations and parameterizations used in the NWP models (section 3.1).

Besides attempts to build ML-based weather forecasting models which aim to forecast the same amount of meteorological variables on the same, or even finer, time- and spatial scale as NWP models (e.g. Nguyen et al., 2023; Chen et al., 2023), ML models are also used for postprocessing existing NWP forecasts (e.g. Horat and Lerch, 2024; Scheuerer et al., 2020).

One example of an ML model is an ANN (Fig.3.2). As described in Scheuerer et al. (2020), an ANN is built up from different layers, an input layer, an arbitrary number or hidden layers and an output layer. Each layer consists thereby of several nodes which store the data. The number of nodes in the first layer, the input layer, is equal to the number of predictors since each predictor is stored in a separate node. The predictors are normalized beforehand in order to have the same magnitudes. This is important since the data is transferred between layers using weights which are optimized ("learned") in an iterative process. If the predictors have large differences in magnitude, the risk of either over-estimating the importance of a predictor or the weights converging to zero is increased. Both lead to a poorer model performance.

In the hidden layers the number of nodes is flexible whereby in the output layer the number of nodes equals the number of outputs. A simple ANN is thereby fully connected which means that every node of

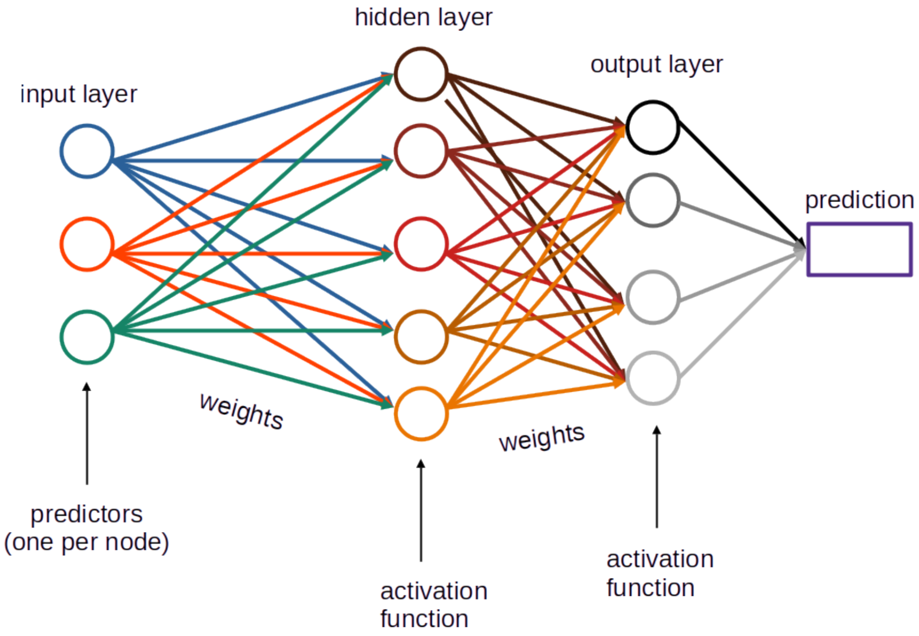


Figure 3.2.: Visualization of a fully-connected ANN. The colors mark the different nodes and their respective connections.

a layer is connected to every node in the following layer. In principle, a reduced number of connections is also possible but not discussed in detail here. In the nodes of the hidden layer and the output layer, the data is collected additively and then passed through a so-called "activation function". A typical activation function for hidden layer nodes, which is also used in the study of Scheuerer et al. (2020), are exponential linear units (ELUs). These can be formulated as (Clevert et al., 2016):

$$f(x) = \begin{cases} x, & \text{if } x > 0 \\ \alpha\,(\exp(x) - 1)\,, & \text{if } x \leqq 0. \end{cases} \tag{3.3}$$

Thereby, x denotes the respective predictor and $f(x)$ the value at the node in the hidden layer. α is set to one in the study of Scheuerer et al. (2020). While positive values are unchanged, negative values are substituted by a value between minus one and zero. This introduces non-linearity to the learning process of the ANN.

At the output layer another activation function is applied, the so-called "softmax" activation (Hastie et al., 2009; Scheuerer et al., 2020):

$$p_{\mathrm{i}} = \frac{\exp(x_{\mathrm{i}})}{\sum_{j=0}^{m} \exp(x_{\mathrm{j}})}, i = 0...m, \tag{3.4}$$

whereby $x_{\mathrm{i}}, x_{\mathrm{j}}$ denote the values of the nodes in the output layer of the ANN. The softmax activation results in positive values which sum up to one. In case of a deterministic prediction, the maximum value after the activation is usually taken as the prediction of the ANN.

One advantage of ANNs and ML models in general compared to the more traditional postprocessing methods (section 3.3) is the possibility to include further predictors besides the target variable and the observations. For example, Scheuerer et al. (2020) include geographic information and climatological probabilities to their ANN which is used to postprocess predictions of ECMWF's S2S reforecast ensemble of subseasonal precipitation accumulations over California.

An ANN is also applied by Silini et al. (2022) to postprocess ECMWF's subseasonal MJO predictions. Fan et al. (2023) use ANNs to postprocess NCEPs predictions of the mean 2-meter temperature and precipitation in the weeks three and four after initialization.

Another type of ML model often used are CNNs. In principal, a CNN is built up in the same way as an ANN with different layers and nodes (Fig.3.2). However, a CNN uses images as predictors whereby each image is stored in a separate node. Due to this, the data is processed differently (Horat and Lerch, 2024). Most prominently, the weights applied to the data when transferring it between the separate layers is stored in so-called filters, which cover a field of a certain number of pixels, e.g. an area of 3x3 pixels.

As described in Horat and Lerch (2024), at first, lower dimensional features of the input image are extracted. This is done by applying convolutional filters to each of the input images (Fig. 3.3 (c)). In this step, the weights of the different pixels in the image are learned by the model. This is done in a shared manner across pixels. For example, a 3x3 convolutional filter takes into consideration a field of 3x3 pixels. For the inner pixel, a dot product between the surrounding eight pixels is performed. Then, the filter is shifted by one pixel. Since the filters are larger than one pixel, the first computation is performed at the second pixel in the second row of the image and the last computation at the second-last pixel in the second-last row of the image. Thus, the resulting image is smaller than the original one when the filter has been passed through. This can be compensated by padding, e.g. copying the values at the margins of the resulting image.

Besides convolutional layers pooling layers are usually applied in an CNN (Horat and Lerch, 2024; Scheuerer et al., 2020). In a pooling layer, filters in form a square are passed through the images

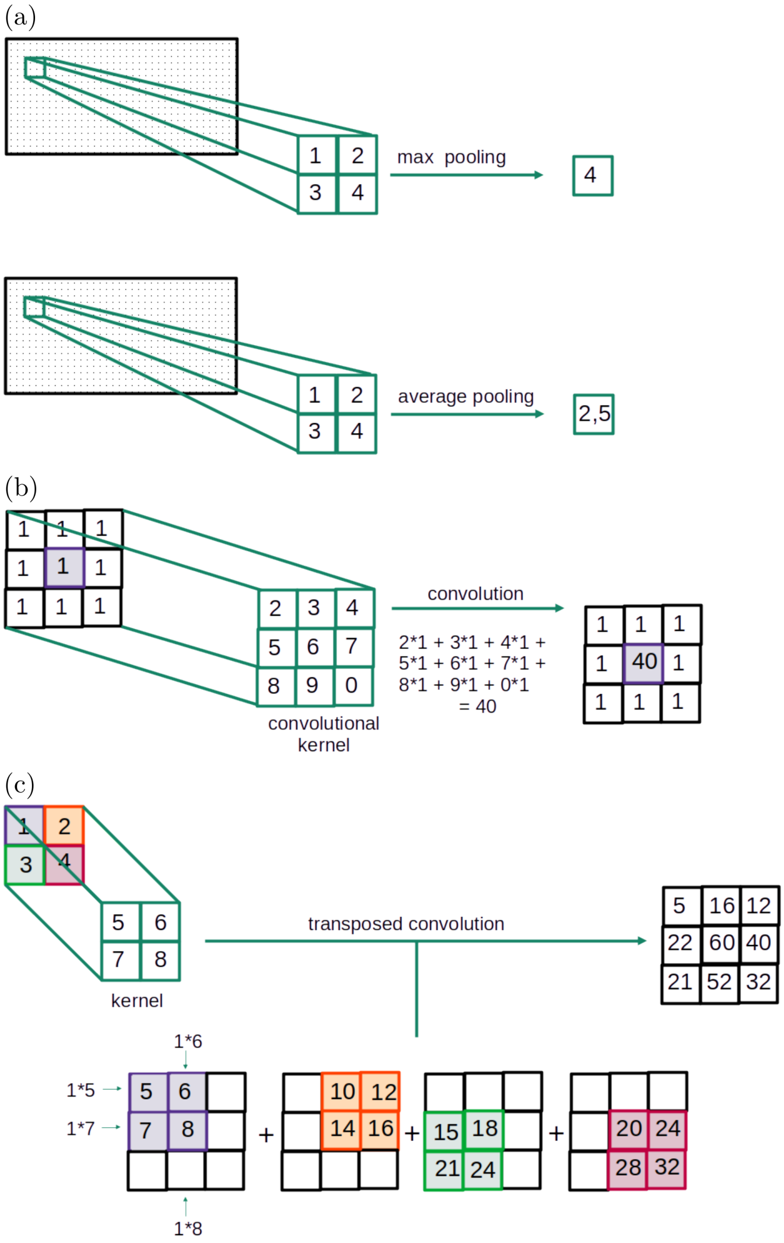


Figure 3.3.: Visualization of the concept of pooling (a), convolution (b) and transposed convolution (c) as used in CNNs.

(Fig. 3.3 (a) and (b)). Of the pixels inside the square, only one value is used in the resulting image which further reduces its size and condensates the information present in the image (Horat and Lerch, 2024). Depending on the kind of pooling filter this can either be e.g. the maximum (Fig. 3.3 (a)) or the average value (Fig. 3.3 (b)) of the pixels in the square. Analogously to the convolutional filters, the pooling filters are shifted through the image pixel by pixel. For example a 2x2 max-pooling filter takes only the maximum value of the four pixels considered.

Optionally, between convolutional and pooling filters, the images can be passed passed through an activation functions such as ELUs in order to add non-linearity to the learning process of the model, similar as being done at an ANN (Scheuerer et al., 2020). As a last step the image is flattened to create a one-dimensional prediction.

CNNs are applied by Scheuerer et al. (2020) to postprocess ECMWF's S2S reforecasts of the geopotential field in 500 hPa and total column water for forecasting precipitation over California. Horat and Lerch (2024) use them for postprocessing ECMWF's S2S reforecasts in the frame of the World Meteorological Organization's Sub-seasonal to seasonal (S2S) Artificial Intelligence Challenge (Vitart et al., 2022). This challenge aims specifically on the leverage of ML models for subseasonal predictions of bi-weekly averaged 2-meter temperatures and precipitation globally. Horat and Lerch (2024) use in their study besides the ECMWF's S2S reforecasts of the 2-meter temperature and precipitation further predictors from the ECMWF's S2S reforecasts suite. In case of the temperature forecasts, they include the mean sea level pressure as well as the geopotential in 500 and 850 hPa height. In case of the precipitation forecasts, total column water is added as a predictor.

Beside a classical CNN, Horat and Lerch (2024) also use so-called "UNets" (Fig. 3.4). This architecture is characterized by an encoder-decoder structure whereby the operations used in the encoder part (left side on Fig. 3.4) are used symmetrically in the decoder part (right side on Fig. 3.4). The only difference is the substitution of the pooling layers, which are used in the encoder part by transposed convolutions (Fig. 3.3 (d)) in the decoder part. Furthermore, at the front of every layer of the decoder, an image of the symmetric layer of the encoder part is concatenated to the resulting layer after the transposed convolutions. These "skip-connections" enable the decoder part to produce images of the same size as those obtained in the symmetric part of the encoder. In a last step, the output of the UNet is generated using a convolutional layer and a softmax-activation.

Apart from CNNs and U-Nets, also RF-based models have been successfully used for postprocessing ECMWF's S2S reforecasts in the frame of the World Meteorological Organization's Sub-seasonal to seasonal (S2S) Artificial Intelligence Challenge (Vitart et al., 2022). Although all of these models perform skillful in the global mean, the skill of forecasts is highly depending on the geographic region. Therefore, local ML-based models, as used e.g. in the study of Scheuerer et al. (2020), are a promising alternative for subseasonal forecasting. Besides a possibly more accurate forecasting skill for the target region, also the computational costs are reduced in comparison to calculating global forecasts.

In contrast to Scheuerer et al. (2020), e.g. van Straaten et al. (2022), Weyn et al. (2021) or Mayer and Barnes (2021) use ML models for direct subseasonal predictions instead of postprocessing existing subseasonal (re-)forecasts. Weyn et al. (2021) develop a 320 member ensemble created by CNNs to forecast

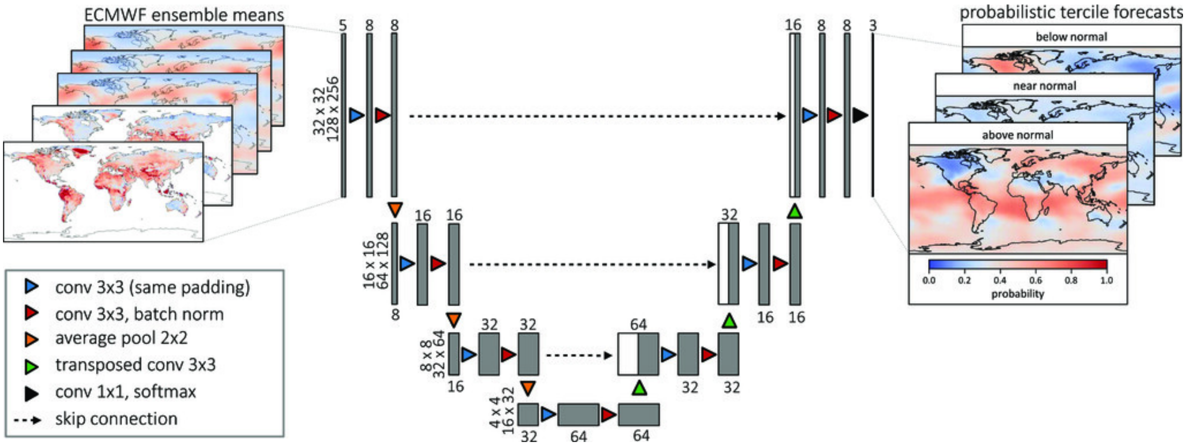


Figure 3.4.: The UNet architecture used for temperature (for precipitation, one additional predictor is used, resulting in six inputs). The number of global maps (channels) is denoted on top of each box, the spatial extent of these feature maps is given on the left side of each box for the patchwise training (left number) and the global training (right number). Figure and caption reprinted from Horat and Lerch (2024), their Fig. 2. ©American Meteorological Society. Used with permission.

different meteorological fields globally on timescales up to six weeks. The RF models developed by van Straaten et al. (2022) are based on reanalysis data. They use nine different predictors to forecast spatially averaged high summer 2-meter temperatures in Central and Western Europe. The atmospheric predictors, which are the geopotential in 300 hPa height, the temperature in 850 hPa height and the total cloud-cover fraction, are thereby complemented by oceanic and land-surface based predictors. A detailed description of the characteristics of RF models in general is given in section 5.10.

Mayer and Barnes (2021) use reanalysis values of the daily mean of the outgoing longwave radiation as well as the geopotential in 500 hPa height as input for an ANN to predict the sign of the geopotential anomalies in 500 hPa height over the North Atlantic Ocean. This information can then be used as a precursor for e.g. surface weather in Central Europe.

In addition to the pure forecasting task, van Straaten et al. (2022) and Mayer and Barnes (2021) put another focus on the explainability of their models' predictions. This is done one the one hand to enhance the trustworthiness of these models, which are unlike NWP models not solely based on human-understandable equations or parameterizations. On the other hand, these methods, summarized below the term eXplainable Artificial Intelligence (XAI), can help to discover new sources of predictability. Two XAI methods are used by van Straaten et al. (2022) for both cases. Permutation Feature Importance, which is closely related to the Impurity Based Feature Importance described in section 5.11.1, is used to explain which predictors are most relevant based on all model predictions. Similar to the Impurity Based Feature Importance, the Permutation Feature Importance measures the importance of the predictor via the overall model's performance. Thereby, in case of the Permutation Feature Importance, the values of the predictors are permuted, one after another, and used instead of the originally ordered values as input to the ML model. After every permutation, a prediction is made and the decrease in model performance assessed. Shapley Additive Explanations (SHAP), which are described in detail in section 5.11.2, are used to extract the contributions of the predictors to single forecasts.

While the Permutation Feature Importance is predominately used to identify which predictors are most relevant for the model (and to possibly minimize the number of predictors to reduce computational costs), SHAP can also be used to interpret the predictions of the model physically. However, only statements about correlations to existing physical knowledge can be made but causality not inferred. SHAP is used by van Straaten et al. (2022) to identify drivers of hot summer temperatures on a daily basis but also to assess the global importance of predictors by averaging over the resulting SHAP values. In a similar manner, Mayer and Barnes (2021) use the method of layerwise relevance propagation (LRP) to identify important sources of predictability of the 500 hPa geopotential height anomalies over the North Atlantic Ocean. As described in Mayer and Barnes (2021) and, in more detail, in Toms et al. (2020), LRP is done after the training of the ANN. Therefore, a single sample is used as input of the trained ANN. Usually, this is the sample of the output node with the highest value or class probability which also determines the prediction of the ANN. The value of this node is then propagated backwards through the ANN following a set of predefined rules. Since LRP is done sample-wise, sources of predictability for specific cases can be inferred. Nevertheless, only a correlation but not a causal link to e.g. physical connections can be made using this XAI method.

4. Research Questions

As outlined in the introduction, a skillful prediction of wintertime 2-meter temperature, especially the cold extremes, is highly desirable for decision-makers in the socio-economic sector. Unfortunately, as discussed in chapter 1, forecasting on this timescale is particularly challenging compared to forecasting on the short to medium and seasonal time range (White et al., 2017). Besides model errors and biases arising from measurement uncertainties, the data assimilation process and numerical approximations, which are relevant on all timescales, NWP models mainly suffer from three issues when predicting on the subseasonal timescale. These are the decreasing importance of initial conditions, the not-yet changing boundary conditions and the lack of realistic representations of teleconnections (e.g. White et al., 2017). Furthermore, a significant drift of the predictions to the model's climatology is observed on the subseasonal timescale (Owens and Hewson, 2018).

Improvements of NWP models are, although highly desirable, cost-intensive measures in terms of both computational resources and (research) time. One reason therefore is , as explained in chapter 2, that teleconnections often feature a complicatedly interwoven feedback within the climate system which is in turn challenging to represent in numerical equations suited to be run in an NWP model operationally.

As an alternative to adapting NWP models directly, previous work, as discussed in chapter 3, shows a range of possibilities to elevate the already available predictions to an improved forecast skill. One of these methods is creating an ensemble of predictions from multiple models instead of using only the predictions of a single NWP model (e.g. Becker and Dool, 2016; Li et al., 2021; Karpechko et al., 2018). Although this approach leads to an improvement in skill, several issues arise. First and foremost, the different model set-ups including different spatial and temporal resolutions, are not trivial to combine. Especially the different initialization times can be a major hurdle since ideally only timesteps with the same lead time should be combined. Second, depending on whether the multi-model ensemble consists of all ensemble members of the single models' predictions or only their ensemble mean, the amount of computational resources needed to provide such a forecast product can rise considerably.

A generally more cost-efficient method to improve subseasonal forecast skill is the postprocessing of predictions of a single NWP model. Postprocessing methods reach from rather simple mean bias corrections (e.g. Monhart et al., 2018), which remove the drift of the model's predictions to its climatology, to more complicated parametric (e.g. Silini et al., 2022; Hyvärinen et al., 2021; Dirkson et al., 2022) and non-parametric approaches (e.g. Horat and Lerch, 2024; Fan et al., 2023; Scheuerer et al., 2020). Thereby, it can be beneficial for the forecast skill if the postprocessing is done regime-dependent (Allen et al., 2019, 2021). As regimes, e.g. WRs can be used. These categorize the large-scale atmospheric flow into a number of common flow patterns. The patterns themselves are forecasted differently well (Osman et al., 2023) and since the large-scale atmospheric flow determines surface temperatures, WRs can potentially be used to assess forecast reliability.

WRs or the atmospheric flow directly can also be used as additional predictors for non-parametric postprocessing methods such as ANNs (Fan et al., 2023) or CNNs (Horat and Lerch, 2024; Scheuerer et al., 2020, e.g.), as described in chapter 3.4. This is an advantage in comparison to the parametric approaches since these additional predictors can yield further sources of predictability on the subseasonal timescales.

Another advantage of ML models is their ability to create subseasonal forecasts even without the inclusion of any NWP predictions. Especially RFs are thereby a substantially more cost-efficient method compared to NWP models and have shown to be able to create skillful subseasonal forecasts compared to a climatological benchmark (e.g. van Straaten et al., 2022).

In this thesis, the potential of RF models for improving forecasts of wintertime 2-meter temperatures and the occurrence of cold-wave days in Central Europe two to four weeks in advance is presented. Furthermore, an approach using WRs to assess forecast reliability is shown. The research is divided into three topical parts: the use of RF models as alternatives to NWP models, their use in a postprocessing sense as a complement to NWP models, and the use of WRs to assess forecast reliability. In the following, the research questions and approaches of each part are shortly summarized.

Chapter 6 describes the use of RF models based solely on reanalysis data for forecasting wintertime 2-meter temperature and the occurrence of cold-wave days in Central Europe at lead times of 14, 21 and 28 days. As predictors, only variables connected specifically to the forecasting task at hand, as described in section 2.1, are used. The forecasts are compared in a first step to a climatological benchmark ensemble. The following research questions are addressed:

Part 1: RF models as an alternative to NWP models

RQ 1.1 Can RF models using only reanalysis data as input provide skillful forecasts in comparison to a climatological benchmark?

RQ 1.2 Is the forecast skill equally good for periods with mild temperatures than periods with cold temperatures?

RQ 1.3 Which predictors are determining the models' predictions?

Improved predictions compared to a climatological benchmark ensemble can already be an important value when computational resources are severely limited or access to NWP forecasts is not provided. However, in Central Europe, the study area of this research, this is usually not the case. Therefore, in a second step, we compare the RF models' predictions to state-of-the-art subseasonal NWPs which are represented by ECMWF's S2S reforecast ensemble. This is discussed in chapter 7 which mainly focuses on the use of RF-based models for postprocessing. Besides these, also a simple lead-time-dependent mean bias correction of the numerical forecasts is analyzed for comparison. The following research questions are addressed:

Part 2: RF models as a complement to NWP models

RQ 2.1 Which postprocessing approaches perform particularly well in comparison to the original ECMWF's S2S reforecasts?

RQ 2.2 In case of the RF-based postprocessing models, are predictors based on the reforecasted meteorological fields at the target date more important for the models' predictions than the reanalyzed fields at the initialization?

Besides an improved forecast skill, decision makers are highly interested in the reliability of forecasts, especially when preventive measures are costly. Therefore, we utilize the seven WRs proposed by Grams et al. (2017), which are described in detail in section 5.4, as a measure for forecast reliability. Chapter 8 discusses this matter by addressing the following research questions:

Part 3: WRs for assessing forecast reliability

RQ 3.1 Does the forecasting skill of ECMWF's S2S reforecasts and the RF-based models depend on the WRs at initialization?

RQ 3.2 In how far can the WR succession during a forecast be linked to subseasonal forecast skill?

RQ 3.3 What are the differences in the WR successions before the best (worst) predicted days within cold waves?

Addressing the presented research questions will enable us to improve and assess subseasonal forecast skill in a cost-efficient manner. If used operationally, it can help socio-economic decision makers improve their subseasonal planning substantially.

5. Data and Methods

This chapter explains the data and methods used to answer the research questions described in chapter 4. In terms of data, these are observational, reanalysis and reforecast data as well as a categorical and continuous representation of the Euro-Atlantic WRs proposed by Grams et al. (2017). In terms of methods, the construction of a climatological benchmark ensemble, the definition of cold-wave days, the application of probabilistic RF models including the selection and preprocessing of predictors as well as two XAI-methods (Impurity Based Feature Importance and SHAP), forecast evaluation techniques and significance testing with Welch's t-test are described.

5.1. E-OBS Observational Data

We use observational data from the E-OBS dataset, version 23.1e (E-OBS, Cornes et al. (2018)), as the ground truth in this study. E-OBS is a dataset of land station data which is interpolated to a regular latitude-longitude grid. It consists of daily mean, minimum and maximum temperatures as well as precipitation for Europe. In our studies, we focus on daily mean temperatures. The data is provided from 1950 onwards. Since the number of stations is varying over time but generally increasing, the gridded data has tendentiously higher uncertainties for the earlier years. As of 2018, approximately 3700 temperature stations across Europe are included in the dataset. Here it is important to note, that the station density varies across Europe with a relatively high station density in Central Europa and Scandinavia and a relatively low station density in Southern and Eastern Europe. Another difference between stations is the way the daily mean temperature is calculated. At some stations, it is computed as the average of daily minimum and maximum temperatures, at other stations as the mean of (multi-) hourly readings. Both, as well as possible measurement uncertainties, are not considered in the creation of E-OBS.

We use the averaged daily mean temperature (*tg*) over Central Europe as the ground truth. We define Central Europe in our studies as the region between 3°E to 20°E and 45°N to 60°N (purple box on Fig. 5.1 (a)). To create a more homogeneous area, we exclude the first grid points along the coasts and all grid points with an altitude above 800 m, which are classified as mountainous regions by the German Weather Service (DWD (2023); Fig. 5.1 (b)). As also described in section 2.1, this is done since the coastal grid points are generally experiencing warmer temperatures due to their proximity to the sea which is generally warmer than land in winter. The mountainous grid points experience due to their altitude generally lower temperatures than grid points at lower heights.

We use daily mean instead of daily minimum or daily maximum temperatures partly due to the same reasons. Daily maximum temperatures are not used since during winter, cold temperatures play a more important role than mild temperatures for many socio-economic applications. When using daily maximum temperatures, temperatures below the frost point could be missed when the temperature rises due to sunny conditions during midday. But these temperatures below the frost point are highly relevant for e.g. the agricultural businesses.

Daily minimum temperatures, which capture the colder temperatures better than the milder, are not used since their variability is high among stations. For example, when a station is located in a valley, night-time temperatures can be regularly very cold due to descending cold air from the surrounding mountains leading to large differences in temperatures compared to stations located in flatlands (Spiridonov and Ćurić, 2021). Moreover, high fluctuations in daily minimum and maximum temperatures are observed depending on the cloud cover whereby during clear sky conditions the values are more extreme than during cloudy conditions (Spiridonov and Ćurić, 2021). When using daily mean temperatures, the differences between stations are smaller.

A spatial aggregation of temperatures is done to filter out remaining small-scale fluctuations and possibly enhance predictability (van Straaten et al., 2020). For the same reasons, we apply a 7-day running mean to the spatially averaged temperature timeseries. Since we only consider the extended winter season, the data is retrieved for the months November to April of the years 1950 until 2020 on a regular 0.1°×0.1° latitude-longitude grid.

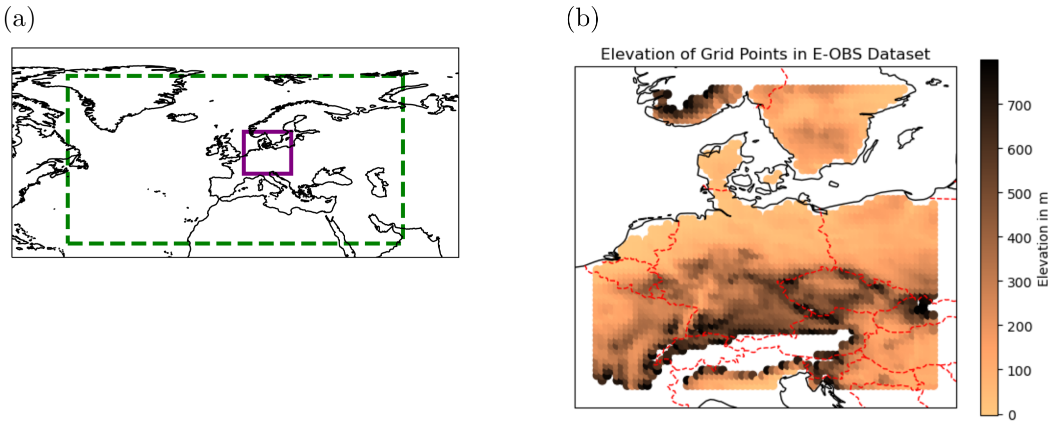


Figure 5.1.: Regions for which the input data and ground truth are retrieved. The input data of the RF models is retrieved from an area covering the North Atlantic Ocean and parts of Eurasia ((a), green dashed box). The ground truth is retrieved for Central Europe ((a), purple solid box) and then averaged over this area whereby only land grid points, excluding the first coastal grid points and all grid points above 800 m, are taken (b). The maps in (a) and (b) are created using the python package "Cartopy" (Met Office, 2010 - 2015).

5.2. ERA5 Reanalysis Data

As one part of the input of the RF models used in this thesis, the ECMWF Reanalysis version 5 (ERA5) data (Hersbach et al., 2020b,c; Bell et al., 2020a,b) is chosen. As described in Hersbach et al. (2020a), ERA5 is based on the model version CY41R2 of the Integrated Forecasting System (IFS) of ECMWF but instead of 91 vertical model levels 137 levels are used. The model top is located at 0.01 hPa. The horizontal resolution is approximately 32 km and the temporal frequency of the data 1 hour.

Besides the IFS-forecasts, observations are embedded in ERA5. This is done using the four-dimensional variational analysis (4D-Var) system in a 12 hourly window. As explained in Mladek (2024), the aim of 4D-Var is to find an optimal fit of all observations in the respective time window and to stay at the same time as close as possible to the IFS-forecast of that time window. Thereby, observations from over 200 satellite or conventional instruments are used including e.g. pressure, temperature and humidity measurements, rain rates, ozone concentration, soil moisture and snow depth (Hersbach et al., 2020a). The relative influence of the observations on the final reanalysis data is e.g. determined by the estimated accuracy of the measurements and the difference between the location and altitude of the measurement to the nearest grid point of the IFS-forecast.

We retrieve the msl, $z850$, $z500$, $z250$, $z100$, $t850$, $H850$, $u300$ and $u10$ for the region between 60°W to 60°E and 20°N to 80°N (green box on Fig. 5.1 (a)) at 00, 06, 12 and 18 UTC. The is retrieved only for the month October to April of the years 1950 to 2020 and averaged daily. To be compatible to ECMWF's S2S reforecasts, a resolution of $1.5^\circ \times 1.5^\circ$ latitude-longitude is used. The reanalysis data is smoothed by a 7-day running mean to consider the temporal uncertainty of subseasonal predictions.

5.3. ECWMF's S2S Reforecast Data

ECMWF's S2S reforecast ensemble, model version CY46R1, is used as predictions of a state-of-the-art NWP model. As described in Vitart et al. (2017) and ECMWF (2019), the reforecasts cover lead times up to 46 days and are initialized every Monday and Thursday at 00 UTC. They are computed for the same calendar day of the past 20 years and consist of 11 members (one unperturbed control forecast and ten perturbed forecasts). The same model as for the operational forecast, which is a coupled atmosphere-ocean model, is used. As stated by Mladek (2024), the horizontal resolution of the atmosphere is approximately 16 km until a lead time of 15 days and then reduced to 32 km. In the vertical, the atmospheric model is build up of 91 levels with the model top being at 0.01 hPa. The resolution of the ocean is a quarter of a degree horizontally and 75 levels vertically. Initial conditions for calculating the reforecast ensemble are based on ERA5 reanalysis data in case of the atmospheric model and Ocean Reanalysis System 5 data in case of the oceanic model. The perturbation is done using Singular Vectors (SV) in combination with Ensemble Data Assimilation (EDA) perturbations from ERA5 which are added in +/- pairs to the control forecast. SVs are used to identify perturbations for pressure, temperature and wind fields which maximize their hemispheric impact for a 48 h forecast (Owens and Hewson, 2018). EDA is used to estimate uncertainties in observations, boundary conditions and model set-up (Owens and Hewson, 2018).

We retrieve ECMWF's S2S reforecasts of the winters 2000/2001 until 2019/2020, which are used as the evaluation period in this study. The same predictors as from the ERA5 reanalysis dataset, $u10$, $u300$, $z100$, $z500$, $z850$, $t850$, $H850$, msl, with the exception of using the geopotential in 300 hPa ($z300$) instead of $z250$ due to reasons of availability, are retrieved for the area between 60°W to 60°E and 20°N to 80°N (green box on Fig. 5.1 (a)). The data is available for 00 UTC only. The 2-meter temperature (*t2m*) is retrieved for the region between 3°E to 20°E and 45°N to 60°N (purple box on Fig. 5.1 (a)) and averaged spatially over this area. Similarly as done for the E-OBS data (section 5.1), the first coastal grid points and all grid points with altitudes above 800 m are excluded before averaging. The horizontal resolution of the data is 1.5° × 1.5° latitude-longitude. We use the forecasts at lead times of 14 , 21 and 28 days for the month November to April.

Here, we use the predictions of a specific day instead of applying a 7-day running mean. This is done for three main reasons. First, the focus of this study is on predicting the temperatures or the occurrence of cold-wave days at a specific day since for socio-economic application this is often more relevant than a smoothed information, especially when it comes to extremes. Second, we aim at a fair comparison between the predictions of the RF models, which are created for a specific day of lead time, and the predictions of ECMWF's S2S reforecasts. And third, applying e.g. a 7-day running mean would imply to download the data of six additional lead times for every grid point which would heavily add to the computational costs.

5.4. Weather Regime Data

We use the seven WRs proposed by Grams et al. (2017) to investigate forecast skill dependent on various representative atmospheric large-scale flow circulations. These are calculated based on the geopotential height in 500 hPa over the North Atlantic-European region which spans across 80°W to 40°E and 30°N to 90°N. Similar as in Hauser et al. (2023), we use the WRs computed from ERA5 reanalysis data but at a three-hourly resolution. As described in their study, the first step for calculating the WRs is computing the geopotential height climatology at 500 hPa of the years 1979-2019 which is smoothed by a 90 d running mean (remark: the data has been extended meanwhile). This climatology is then used to obtain anomalies of the 500 hPa geopotential height which are smoothed in a next step by a 10 d low-pass filter and normalized. Then, the seven leading EOFs are determined which represent 74.4% of the variability in the data. As described in Hannachi et al. (2007), an EOF analysis splits the time-space field of meteorological data into separate representations of the time and space. Thereby, in case of the space field, orthogonal spatial patterns are derived. The orthogonality of these patterns is a property of the whole domain over which the EOF analysis is performed. Due to this, the spatial structures extracted cover large parts of the domain and local structures are neglected. Since WRs should represent the large-scale atmospheric flow over the whole chosen region, the space fields obtained by the EOF analysis are used further. To these, a k-means clustering resulting in seven clusters is applied to identify the most relevant structures (Hauser et al., 2023). As explained in Hannachi et al. (2017), k-means clustering is a method which sorts patterns into a pre-defined number of clusters. The clusters are defined by maximizing the ratio of variances between the clusters' centroid coordinates, which are the average coordinates of all cluster members. Thus, seven distinct clusters are obtained whereby each cluster mean represents one WR. In section 2.2 and Fig. 2.7, the characteristics of these WRs are described in detail.

To attribute the similarity of a weather pattern to the seven WRs, we use the continuous Weather Regime Index (I_{WR}). As shown in Hauser et al. (2023) and Büeler et al. (2021), it is defined as

$$I_{WR}(t) = \frac{P_{\mathrm{WR}}(t) - \overline{P_{\mathrm{WR}}}}{\sqrt{\frac{1}{\mathrm{NT}} \sum_{t=1}^{\mathrm{NT}} \left[P_{\mathrm{WR}}(t) - \overline{P_{\mathrm{WR}}}\right]^2}}, \text{ whereby} \tag{5.1}$$

$$P_{\mathrm{WR}}(t) = \frac{1}{\sum_{(\lambda,\varphi)\in\mathrm{NH}} \cos\varphi} \sum_{(\lambda,\varphi)\in\mathrm{NH}} \Phi(t,\lambda,\varphi)\Phi_{\mathrm{WR}}(\lambda,\varphi)\cos\varphi. \tag{5.2}$$

Thereby, $P_{\mathrm{WR}}(t)$ is the projection of the normalized, 10 d low-pass filtered geopotential height anomaly at 500 hPa at a given time t ($\Phi^L(t,\lambda,\varphi)$) onto the geopotential height patterns which determine the different WRs ($\Phi^L_{\mathrm{WR}}(\lambda,\varphi)$). λ and φ represent the latitude and longitude in the Northern Hemisphere (NH). $\overline{P_{\mathrm{WR}}}$ denotes the climatological mean of $P_{\mathrm{WR}}(t)$ and NT the number of timesteps of the climatological period.

Besides the continuous WR index, we use categorical WRs. For these, as explained in Grams et al. (2017) and Hauser et al. (2023), only WR index value above 1 are considered. A WR is chosen as the categorical WR of a timestep if its I_{WR} value is above 1 for more than five consecutive days and at least once during this time the highest I_{WR} value of all seven WRs. In case none of the WRs fulfills these criteria at a timestep, the categorical WR of that timestep is called the "No" (No) regime (31% of all days). This regime represents atmospheric states which are closer to the climatological atmospheric state than to any of the seven WRs.

Of the remaining seven WR, the most often occurring WRs are the ScTr and ScBL which occur on 11% of the days each. They are followed by the EuBL and GL regime which are present on 10% of the days each and the last three WRs, the ZO, AT and AR regime, which occur on respectively 9% of the days.

Since the number of days during which a WR is present is calculated here from the temporally extended data (11 January 1979 - 13 August 2022), the numbers differ from the values given in Büeler et al. (2021).

To be consistent to ECMWF's S2S reforecasts, we only use the categorical WRs and the continuous WR index at 00 UTC for the month November to April between 1 November 2000 and 30 April 2020 in our studies.

5.5. Climatological Benchmark Ensemble

In the field of subseasonal forecasting, climatological models are commonly used for benchmarking (e.g. Vijverberg et al., 2020; Scheuerer et al., 2020). Here, we use a climatological benchmark ensemble constructed from past winters of E-OBS observational data. Thereby, the daily mean 2-meter temperature timeseries of each winter between 1970/1971 to 1999/2000 is used as one ensemble member thus creating a 30-member ensemble (lightblue dashed lines on Fig. 5.2). We use the last 30 winters before our evaluation period for creating the climatological model as recommended by the World Meteorological Organization (WMO) for the computation of climatological normals (WMO, 2020). However, we omit calculating the mean over these last 30 winters in order to receive a probabilistic ensemble. This is done for two reasons.

Firstly, by using only the mean over the past 30 winters, either the information about extremes is lost (when there are warm and cold extremes on the same day of the year) or the daily ensemble mean is determined by one extreme value (when e.g. there is a extremely cold day in one winter while the same day of the year in all other winters is showing mild temperatures). In both cases, the mean does not represent the variability of winters well. The same applies to the use of the median of the past 30 winters, which is generally not exhibiting temperatures below the frost point for the whole extended winter season (blue solid line and darkgrey dashed line on Fig. 5.2).

The second reason for using a probabilistic climatological benchmark ensemble in our study is to allow for a fairer comparison with the probabilistic ECMWF's S2S reforecasts. Nevertheless, since we want to compare ECMWF's S2S reforecasts to the "best version" of the climatological benchmark ensemble, we use the whole 30 members instead of reducing the ensemble to the same number of ensemble members as the ECMWF's S2S reforecasts.

Temperatures are not detrended for the calculation of the climatological benchmark ensemble. A detrending of temperatures requires knowledge or assumptions about the winters following the climatological period. In our case, we know the temperatures of the winters 2000/2001-2019/2020. However, since we want to use the climatological benchmark model for predicting, it would not be fair to include this information since the model would then "know about the future".

Making an assumption of the global warming trend over Central Europe, our area of interest, is not trivial since regional warming is a non-linear process. Therefore, we decided to use the daily mean 2-meter temperatures without detrending to create the climatological benchmark ensemble. The approximately 0.8 K colder mean of the winters 1970/1971-1999/2000 compared to the winters 2000/2001-2019/2020 results in a slight advantage of the climatological benchmark ensemble in forecasting cold winter temperatures and a slight disadvantage in forecasting mild temperatures. The same applies to the prediction of cold-wave days. In the climatological benchmark ensemble, 21.6% of the days are classified as cold-wave days (the definition of cold-wave days is explained in section 5.6). In the evaluation period of the winters 2000/2001-2019/2020, only 15.3% of the days are classified as cold-wave days. Therefore, the climatological benchmark ensemble also has an advantage in predicting cold-wave days correctly and a disadvantage in forecasting non-cold-wave days.

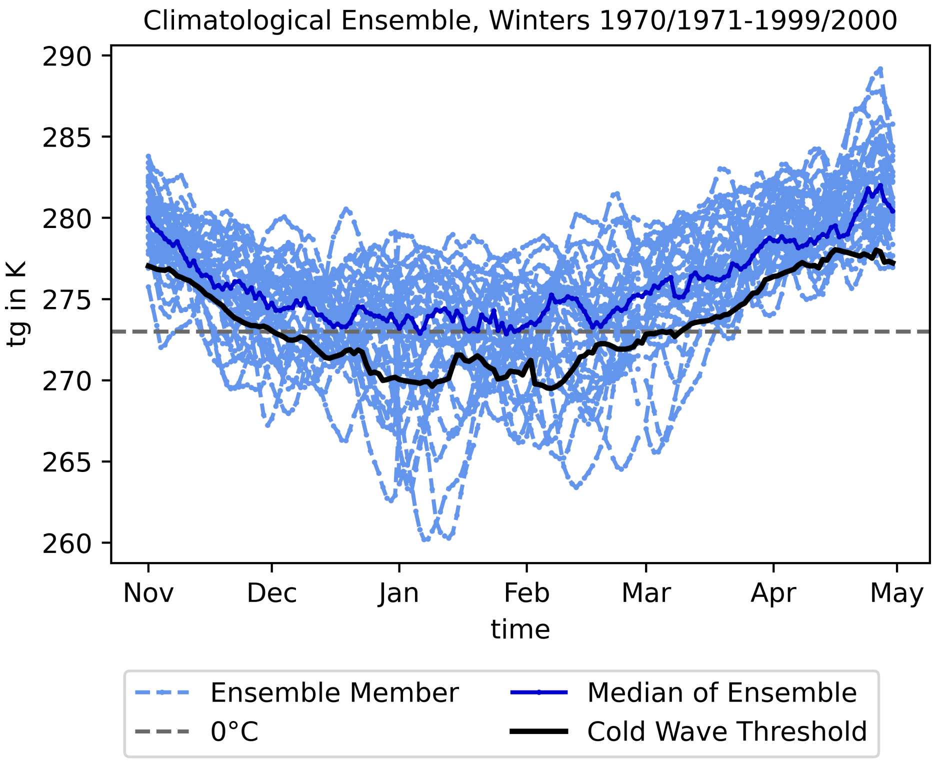


Figure 5.2.: Climatological benchmark ensemble and cold-wave thresholds. The climatological benchmark ensemble of continuous daily mean 2-meter temperatures (tg) is depicted as the dashed light blue lines, its median as the solid blue line. The daily cold-wave thresholds are shown by the solid black line which is located below the solid blue line. The 0°C-line is marked by the horizontal dashed darkgrey line.

5.6. Definition of Cold-wave Days

We define a day as a "cold-wave day" when it is part of a cold wave. Analagously, a day is labeled as a non-cold-wave day, when it is not part of a cold wave. A cold wave is thereby defined by the cold wave criterion of Smid et al. (2019) with small modifications. According to Smid et al. (2019), a cold wave consists of at least three consecutive days with temperatures below a certain threshold. This threshold is calculated by taking the daily multi-year 10^{th} percentile of daily minimum temperatures in a reference period. The temperature timeseries is thereby smoothed beforehand by a 31-day running mean. As the reference period, the years 1981-2010 are used.

In our study, we apply three modifications to this definition. Firstly, and most importantly, we calculate the cold-wave threshold for the Central European mean and not grid-point based. We do this, since our study is designed to forecast the occurrence of cold-wave days for the Central European mean and not for individual grid-points. Secondly, we use the winters 1970/1971-1999/2000 as the reference period since the original reference period is overlapping with our evaluation period. This is done in accordance with the WMO guidelines on how to calculate climatological normals (WMO, 2020). As a third modification, we use daily mean 2-meter temperatures from the E-OBS dataset to calculate the daily thresholds instead of the daily minimum 2-meter temperatures. This is done since the aim of this study is to forecast daily mean temperatures. Furthermore, as described in section 5.1, daily mean temperatures are generally more stable across stations than daily minimum temperatures. When using daily minimum temperatures in the definition of cold-wave days instead of daily mean temperatures, this would result in the detection of only the most extreme cold waves. Using daily mean temperatures to calculate cold-wave thresholds provide therefore more sensible values for the target region and practical applications in the socio-economic sector. The calculated thresholds lie thereby below the frost point between December and March and have their highest values around 4°C in April (Fig.5.2).

5.7. Lead-Time-Dependent Mean Bias Correction

According to ECMWF's "Forecast User Guide" (Owens and Hewson, 2018), starting ten days after initialization the drift of model calculations becomes significant. Since we analyze forecasts with longer lead times, we perform a lead-time-dependent mean bias correction (e.g. Monhart et al., 2018) to compensate the model drift inherent in ECMWF's S2S reforecasts. To do so, we use a bias correction based on the daily mean 2-meter temperature averaged over Central Europe from the E-OBS dataset. In a first step, the daily mean of the reforecast ensemble is calculated for every ensemble of the 20 winters in the evaluation period. From these daily ensemble means, the observational Central European mean 2-meter temperature is subtracted for the respective day. In the last step, we use a leave-one-(winter-) out approach to perform the bias correction. By doing so, the calculated difference is averaged over every day of winter for 19 out of the 20 winters to obtain a multi-year daily mean timeseries of temperature biases of ECMWF's S2S reforecasts. This timeseries is then subtracted from every ensemble member of the 2-meter temperature reforecasts of the remaining winter creating the mean bias corrected ECMWF's S2S reforecast ensemble. This is done for every lead time separately.

To illustrate, we describe the procedure with the example of the winter 2008/2009. Firstly, the daily ensemble mean of all reforecasts of the winters 2000/2001-2007/2008 and the reforecasts of the winters 2009/2010-2019/2020 is calculated. Then, from these ensemble means, the 2-meter temperatures of the E-OBS data are subtracted from the respective day, meaning that e.g. the observed temperature of the 1 January 2005 is subtracted from the reforecasts' ensemble mean of the 1 January 2005. This timeseries of temperature biases is then averaged over every day of winter leading to a timeseries starting with the mean of all 1st Novembers and ending with the mean of all 30th Aprils. In a last step, this resulting timeseries is subtracted from the 2-meter temperature reforecasts of the winter 2008/2009.

5.8. Selection of Meteorological Predictors

The selection of predictors used for the ML models is based on meteorological knowledge. This is done to reduce computational costs and comes with the advantage of an easier interpretability of the models since predictors are limited and understandable by humans.

Central European winter weather is essentially governed on the subseasonal timescale by large-scale atmospheric systems located over the North Atlantic Ocean and the European continent (Domeisen et al., 2020). We therefore constrain the predictor fields to the area between 60°W to 60°E and 20°N to 80°N (green box on Fig. 5.1). Since on the subseasonal timescale teleconnections play an important role for the predictability of European surface weather, as described in section 2.3, we use $u10$, which represents the stratospheric polar vortex (e.g. Domeisen et al., 2020), as well as $z100$, $z250$ (ERA5 reanalysis data) / $z300$ (ECMWF's S2S reforecasts) and $u300$ which describe the upper tropospheric state (e.g. Kautz et al., 2020; Pinto et al., 2014). Furthermore, the mid-tropospheric state, represented by $z500$, determines European surface weather on medium to subseasonal timescales and is itself influenced by teleconnections (e.g. Kautz et al., 2022; Büeler et al., 2021). The lower tropospheric predictors $z850$, $t850$ and $H850$ are relevant on short to medium timescales and are important indicators of the origin of air masses. On short timescales, *msl* is determining European surface weather and in case of persistent pressure systems, also influence weather on the subseasonal timescale. A detailed description of variables influencing temperatures in Central Europe is given in section 2.1.

Since Central European winter temperatures have a high dependency on the time of winter (Fig. 5.2), the month is included as another predictor. In case of the predictors derived from ECMWF's S2S reforecasts, also the predicted Central European mean 2-meter temperature is used. The complete list of predictors

Table 5.1.: Predictors used for the ML models. Predictors are retrieved from ERA5 reanalysis data (ERA5) at the time of model initialization and from ECMWF's S2S reforecasts (S2S) at the target date of the forecast.

Predictor	Height	Abbreviation	ERA5 (at initialization)	S2S (at target date)
Zonal Wind	10 hPa	$u10$	✓	✓
	300 hPa	$u300$	✓	✓
Geopotential	100 hPa	$z100$	✓	✓
	250 hPa	$z250$	✓	-
	300 hPa	$z300$	-	✓
	500 hPa	$z500$	✓	✓
	850 hPa	$z850$	✓	✓
Temperature	850 hPa	$t850$	✓	✓
	2-meter	$t2m$	-	✓
Specific Humidity	850 hPa	$H850$	✓	✓
Pressure	mean sea level	msl	✓	✓
Month	-	-	✓	✓

is summarized in Tab. 5.1 and depicted on Fig. 2.2.

Predictors originating from ERA5 reanalysis data are retrieved for the time of the model start. Therefore, predictors in higher altitudes are expected to contain more useful predictive information for the ML models to learn from. This is due to the fact that atmospheric signals emerging in higher altitudes need a certain amount of time, often more than ten days (Domeisen et al., 2020), to propagate down to the surface while signals emerging in the lower troposphere usually diminished within a few days. Nevertheless, predictors from the middle and lower tropospheric heights are also included since persistence can have a substantial impact on forecasts.

Predictors retrieved from ECMWF's S2S reforecasts are taken at the target date of forecasts. This is done to take advantage of the (although limited) physical knowledge about the evolution of the atmospheric flow on subseasonal timescales inherent in the equations of the NWP model. Furthermore, ECMWF's S2S reforecasts are initialized with ERA5 reanalysis data and therefore contain at initialization the same information as present in the predictors retrieved directly from the ERA5. Since forecasts are retrieved at the target date of the forecast, predictors in low tropospheric heights are expected to provide the most valuable information to learn from for the ML models. Especially the reforecasted 2-meter temperature, which is also the target variable (but from E-OBS observational data), is expected to be very relevant for the ML models' forecasts. However, according to Tripathi et al. (2015), even stratospheric anomalies are able to propagate downward to the surface within several days. Therefore, also predictors in higher altitudes are retrieved from ECMWF's S2S reforecasts.

5.9. Predictor Preprocessing

Two different types of preprocessing are applied to the predictor fields in order to enhance interpretability and further reduce computational costs. For comparison, also the non-preprocessed predictor fields are used in case of the predictors based on reanalysis data. The month is always used as a scalar predictor and not preprocessed. A summary of the different inputs is listed in Tab. 5.2.

In case of the non-preprocessed predictor fields, all grid points of the relevant fields are used in form of one vector per date. This approach yields the maximum amount of training data and is done for reanalysis data as input only. Here, we furthermore take advantage of the whole ERA5 timeseries which starts in 1950.

The first approach used to condense the information inherent in the predictor fields yields the minimum amount of data per date by taking only the spatial minimum, mean, maximum and variance of the meteorological fields and the month as predictors (denoted by "_stat"). Nevertheless, these statistics contain relevant information about the (evolution) of surface weather. For example, if the minimum, mean or even the maximum of the $u10$ field is negative, an SSW event is in place and there is an increased likelihood of cold winter weather in Central Europe in the following weeks, as described in section 2.3. When additionally the variance of the field is low, it is suggested that the SSW event has more probably an influence on the tropospheric state since its anomalies are located more certainly over the North Atlantic Ocean which is especially sensitive to be influenced by stratospheric anomalies (Limpasuvan et al., 2004). When the variance of the field is high, the stratospheric anomalies are concentrated in a small area which location cannot be determined by this approach.

The second approach uses a Principle Components Analysis (PCA, denoted by "_pca"). As described in Hannachi et al. (2007), the PCA analysis (which is also known as EOF analysis) is an approach for dimensionality reduction of data and, as stated in He et al. (2021), a suitable approach to extract features relevant for subseasonal weather forecasting. As already explained in section 5.4, the representation of the data is split into a spatial and temporal component. The aim of the PCA is to find the smallest amount of uncorrelated linear combinations of both that explain as much of the variance of the original timeseries of fields as possible. While the spatial component is represented by the calculated EOFs, the temporal component is represented by the principal components (PCs). The latter we use as input for a part of the ML models. We choose the number of PCs to be ten from tests with two, five, ten and 20 PCs. Furthermore, after testing, we choose to derive the PCs from the original meteorological fields instead of grid point-based normalized fields. With the exception of the $H850$ field, the first ten PCs derived from the original predictor fields explain over 75% of the variance of the predictor fields. For most predictors, the explained variance is greater than 90%. For calculating the PCs, we use the class "sklearn.decomposition.PCA" from the Python package "scikit-learn" (scikit-learn Developers, 2024).

Additionally, to using all nine meteorological fields and the month as predictors (denoted by "_all"), also a subsection of only three predictors ($u10$, $z100$ and $z250$) and the month (denoted by "_sel") is used in case of ML models using only reanalysis data as input.

Concerning the reforecast data, either the preprocessing is done for every ensemble member separately (denoted by "_sepf") or the information is used in a condensed form (denoted by "_ens"). In case of the latter, this means that in the "_stat" approach, only the minimum and variance of the minima of all ensemble members is taken as well as the mean and variance of the mean and variance of all ensemble members and the maximum and variance of the maximum of all ensemble members. From the 2-meter temperature, only the minimum, mean, maximum and variance is taken instead of all ensemble members separately. The same is done for the 2-meter temperature reforecasts when the "_pca" approach is used. Here, concerning the rest of the meteorological fields, the mean and variance of the PC values of the ensemble members is taken. The month is included as a scalar variable in all approaches.

5.10. Random Forest Models

As described in section 3.4, RFs are used as both alternatives to NWP models and complements to these in a postprocessing sense. As RFs are ML models, the first step before using them is preparing the data. Since Central European winters show highly variable temperature evolutions among themselves as demonstrated e.g. by the ensemble members of the climatological benchmark ensemble (lightblue lines on Fig. 5.2), we choose to evaluate our models on several winters instead of a single winter. The choice of using multiple winters for evaluating model forecasts is a common choice for meteorological applications (e.g Scheuerer et al., 2020). Since the developed models should be suitable for forecasting future winters, we select the 20 most recent winters before the start of our study to include any possible trends of climate change in recent years. The chosen winters are the winters 2000/2001-2019/2020, which we call the "evaluation period" as stated in section 5.3. We use this term rather than the term "validation period", which is often used for a part of data reserved for model optimization, since we use the data in this time period to assess model performance but not to optimize it. The term "test period" is also not fitting exactly since the data in this time period has to be strictly kept away from training and only used for the final model performance. But due to our use of a leave-one-(winter-)out cross-validation approach for training and a subsequent averaging over the forecasts of all winters for evaluating the final model performances, the "test period" of one model is included in the training period of another. Here, it becomes unclear whether the term "test period" can still be used in its originally meaning. To avoid any misconception, we therefore decide to use the term "evaluation period" instead.

Table 5.2.: Preprocessing of predictors for the use with ML models. The abbreviation shows the part of the model name referring to the data source which is either ERA5 reanalysis data (ERA5), ECWMF's S2S reforecasts (S2S) or both, and preprocessing type of the predictors. "min, mean, max, var" means that from every predictor field only the areal minimum, mean, maximum and variance are taken. In case of the "condensed" ensemble information of ECMWF's S2S reforecasts, from the calculated values per ensemble member only the minimum and variance in case of the minimum, the mean and variance in case of the mean and variance as well as the maximum and variance in case of the maximum are taken. "first 10 PCs" denotes that from every predictor field only the first ten PCs are taken. In case of the "condensed" information of the reforecast ensemble, from the calculated values per ensemble member only the mean and variance are taken. The month is added to every predictor set as a scalar. The reforecasted temperature is included in ECMWF's S2S reforecasts either with all ensemble members (in case of "all members") or with the ensemble minimum, mean, maximum and variance (in case of "condensed"). The numbers show the amount of predictors per date (#Per Date) and in the whole training set (#Training).

Abbreviation	Preprocessing	ERA5	S2S	# Per Date	# Training
_sel	-	✓	-	1081	13.518.986
_stat_sel	min, mean, max, var	✓	-	13	162.578
_pca_sel	first 10 PCs	✓	-	31	387.686
_all, _all_era5	-	✓	-	3241	40.531.946
_stat_all, _stat_all_era5	min, mean, max, var	✓	-	37	462.722
_pca_all, _pca_all_era5	first 10 PCs	✓	-	91	1.138.046
_stat_all_s2s_sepf	min, mean, max, var	-	all members	408	403.104
_stat_all_s2s_ens	min, mean, max, var	-	condensed	77	76.076
_pca_all_s2s_sepf	first 10 PCs	-	all members	1002	989.979
_pca_all_s2s_ens	first 10 PCs	-	condensed	185	182.780
_stat_all_s2s_sepf_era5	min, mean, max, var	✓	all members	444	438.672
_stat_all_s2s_ens_era5	min, mean, max, var	✓	condensed	113	111.644
_pca_all_s2s_sepf_era5	first 10 PCs	✓	all members	1092	1.078.896
_pca_all_s2s_ens_era5	first 10 PCs	✓	condensed	275	271.700

The use of a leave-one-(winter-)out cross-validation is common to effectively use the available data (e.g. Scheuerer et al., 2020). This means in our case, that for the 20 analyzed winters 20 separate ML models are trained whereby the winter to be predicted is the winter left out from training. For example, when the winter to be predicted is the winter 2008/2009, the training data of the RF models using reforecast data as input comprises the winters 2000/2001-2007/2008 and the winters 2009/2010-2019/2020. This leads to 988 dates (19 winters with 52 forecast dates) in the training dataset of each model. Models using only reanalysis data as input are trained on 12.506 dates (69 winters with 181 or 182 dates per winter, depending on whether the winter is part of a leap year), taking advantage of the longer available timeseries of reanalysis data starting in 1950 and daily values instead of only bi-weekly ones.

Independent of the kind of input data, the predictors are given as arrays into the model. Each row corresponds thereby to one date in the training data. Rows are generally build up in the format ”$predictor_1$, $predictor_2$, ... , $predictor_n$, *ground_truth_value*”, whereby n is the total number of predictors. Reanalysis predictors are taken from the "current date" (initialization), reforecast predictor from the target date of forecasts which are initialized at the "current date". The ground truth used is based on the daily mean 2-meter temperature observation at the target date. Predictors are not normalized since this preprocessing step does not affect the RF models’ performance.

We analyze 14 different types of RF-based models in this study for both the forecast of continuous 2-meter temperatures and the binary occurrence of cold-wave days. In case of the former, Quantile Regression Forests (QRFs) are used, in case of the latter, Random Forest Classifiers (RFCs) are used. Both yield probabilistic forecasts of the respective target. The respective model "types" feature thereby the same architecture, meaning that the QRF or RFC models have the same hyperparameter setting and thus are solely determined by their input variables (Tab. 5.3). This allows for a detailed analysis of the influence of predictor choices on model performance. The construction of the different input types of the models is explained in detail in section 5.9. In the following, we focus on the "meteorologically relevant" differences in inputs. Since the month is included in every predictor set it is not discussed further here.

The first set of models, denoted by "_sel" in the model name, takes only three meteorological variables ($u10$, $z100$, $z250$) from reanalysis data as input (first and seventh block in Tab. 5.3). These type of models contain the least meteorological information of all model types. The second set of models also relies fully on reanalysis data but takes nine ($u10$, $z100$, $z250$, $z500$, $z850$, $u300$, $t850$, $H850$, msl) instead of only three meteorological variables as input. These models are denoted by "_all" and the absence of "_s2s" in their model name (second and eighth block on Tab. 5.3). The third sets of models, denoted by "_s2s_sepf" and the absence of "_era5" in their model name, uses all ensemble members of ten meteorological variables ($u10$, $z100$, $z300$, $z500$, $z850$, $u300$, $t850$, $H850$, msl, $t2m$) from reforecasts as input (third and ninth block on Tab. 5.3). The fourth type of models uses the same meteorological input but instead of all ensemble members separately only the condensed ensemble information, as described in section 5.9, is considered (fourth and tenth block on Tab. 5.3). These models are denoted by "_s2s_ens" and the absence of "_era5" in their model name. The last two types of models use both, reanalysis ($u10$, $z100$, $z250$, $z500$, $z850$, $u300$, $t850$, $H850$, msl) and reforecast data ($u10$, $z100$, $z300$, $z500$, $z850$, $u300$, $t850$, $H850$, msl, $t2m$) as input. Thereby, either all ensemble members of the reforecast data are considered as input (fifth and eleventh block on Tab. 5.3) or only the condensed ensemble information of the reforecast data (sixth and twelfth block on Tab. 5.3). They are denoted by "_s2s_sepf_era5" and respectively "_s2s_ens_era5" in their model names. As summarized on Tab. 5.3, the different inputs are similar for both model types, the QRF and RFC models. In terms of data, only the ground truth is different between the QRF and RFC models.

5.10.1. Quantile Regression Forest Models

As explained in Meinshausen (2006), QRF models are designed to solve a regression task probabilistically. Similar to regular RF models, QRFs consist of several decision trees (DTs) which are run in parallel. A DT is thereby build up from several nodes which are connected by branches (Fig. 5.3). Analogously to a natural tree, a DT is build up from the root (Fig. 5.3 (b)). Here, at the so-called root node the first split of the data happens. For splitting the data, only a subset of the predictors is considered. This is done to obtain different thresholds for splitting the data learned by each individual DT inside the RF. From the chosen predictors, the DT learns to select the "best" predictor and threshold to split the data into two parts (Fig. 5.3 (a)). The "best" is thereby determined via a variance criterion applied to the data in the resulting nodes and found iteratively. Unlike the data considered for splitting at a node, which are the predictors from reanalysis and/or reforecast data ("x" on Fig. 5.3 (a)), the data used for determining the best split is the ground truth. More precisely, the ground truth data "belonging" to the remaining input data in the two resulting nodes ("y" on Fig. 5.3, here depicted in the final nodes). "Belonging" refers thereby to the ground truth value in the same row of the input data as the predictor used for splitting. In case of the reforecast predictors, this means that the date of the predictors and the ground truth is the same. In case of the reanalysis predictors, the ground truth date is taken at the target date of the forecast whereas the reanalysis predictors are taken at initialization time.

In order to find the best split, inside each of the resulting nodes, the variance among the ground truth values, which measures the spread of the values between their mean, is minimized (Molnar, 2022). The splitting criterion is thereby formulated in a binary way. Either the predictor used for splitting fulfills the criterion, then it is "transported" via the so-called "yes-branch" to one resulting node (denoted in blue on Fig. 5.3), or, it does not fulfill the criterion and is "transported" via the "no-branch" to the other resulting node (denoted in red on Fig. 5.3). At these nodes, the procedure of finding the "best" splitting criterion is repeated and the data split again. This is done until either in the resulting node a single data point is present or until a predefined number of samples is reached inside the node. These final nodes are

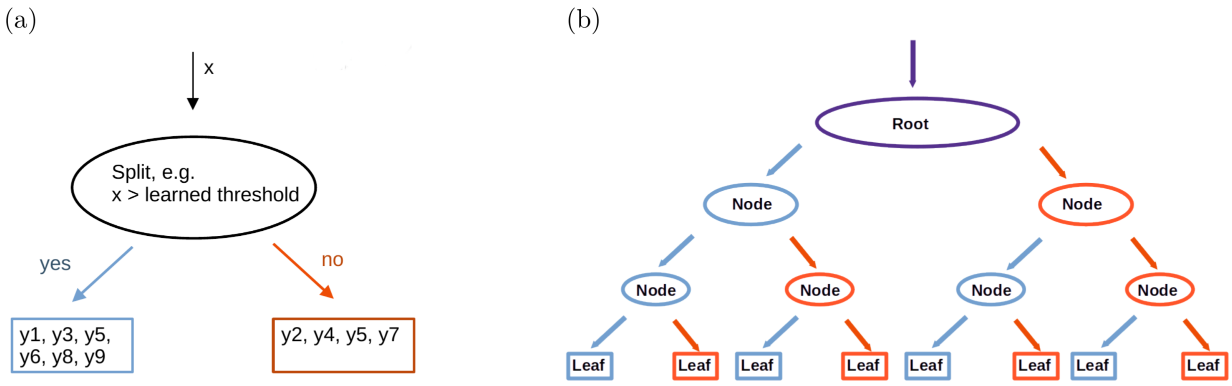


Figure 5.3.: Visualization of a node of a DT (a) and the structure of a DT with a depth of three (b). A node of the DT is represented by an ellipse, the branches by arrows ("yes-branch" in blue, "no-branch" in red) and the leaves by squares ((a) and (b)). The first node of a DT where the data has not been partitioned, yet is called the "root" (b). x denotes the predictors and y the ground truth values (a).

called the "leaves" of the DT (squares on Fig. 5.3). Inside these, all ground truth values which reach the respective leaf, are collected. From here, the prediction of the DT is derived. To predict, the predictors of the relevant days of the future without the ground truth (which is not available in case of predictions of the future) are passed through the DT. At every node, the predictors "moves" via the respective branch determined by whether they fulfilled the splitting criterion or not, until the leaf is reached. The prediction of the DT is then made from the ground truth values collected in this leaf during training. In case of a regular DT, this is simply done by taking the average of these values. The prediction of the whole RF is then obtained by calculating the mean of the predictions of all DTs belonging to the RF. In case of a QRF model, the ground truth values in the final leaves of the DTs are taken together as a cumulative distribution function (CDF). From this function, a predefinded number of quantiles is taken which serves as the probabilistic forecast of the QRF model.

In our study, we use 100 quantiles including the minimum and maximum of the resulting CDF as predictions. The QRF contains thereby 1000 DTs. The number of DTs is determined from experiments using 100, 500, and 1000 trees in combination with different minimal node sizes. The latter determines the minimal amount of data points inside a resulting node after splitting. Here, we try minimal node sizes of one, five, 10, 15, 30 and 100 and decide to set the minimal node size to five. These hyperparameter experiments are done using the QRF model "QRF_stat_all" at a lead time of 14 days and an evaluation of the predictive skill in the 20 winter mean. For saving data and computational costs, we opted against reserving data only for hyperparameter experiments and constrained ourselves to using only one model. This, however, might lead to an overestimation of the performance of the used model. But since RFs are according to Breiman (2001) not prone to overfitting, we assume this to be unproblematic. Overfitting happens, when the RF model approximates the ground truth data during training too closely and therefore fails to generalize well to the unseen data in the evaluation period. One reason why RFs usually don't overfit (Breiman, 2001) is the fact that the DTs consider only a subset of predictors for splitting at each node. This subset consists of a number of predictors equal to the rounded down square root of the total number of predictors. Using only these randomly selected subsets of predictors for splitting the data at a node leads to different thresholds learned by the DT at each node and to different thresholds among the various trees. This diversity prevents the RF from overfitting.

In our study, we use the default QRF implementation of the python package "skranger" (Flynn, 2021) with a few adaptions. These are, as stated above, the number of trees which we set to 1000 instead of 100, the minimal node size which is 5 instead of zero and the enabling of the quantile predictions. The exact set-up of the model is listed in Tab. A.1 in appendix A.

5.10.2. Random Forest Classifier Models

RFC models are used to solve classification tasks (Breiman, 2001). Essentially, they are build up, similarly to QRF models, from several DTs. The main differences are the way data is split at the nodes of the DTs and how the predictions are generated. In case of RFC models, the data is also split according to an optimal threshold learned by the model. However, the "best" split is determined according to the generalized inequality (Gini) index. The Gini index estimates the probability of a datapoint being incorrectly classified by a DT (Suthaharan, 2016). It is minimized when the number of ground truth values belonging to the same class inside a node is maximal. The predictions at the leaves are obtained by taking the fraction of ground truth values belonging to each class from every DT and then averaging across all DTs.

For a better comparison of which predictors are relevant in case of continuous forecasts of wintertime 2-meter temperature versus the occurrence of cold-wave days, we use the same hyperparameters as for the QRF models. The model implementation is again the default RFC model set-up from the python package "skranger" (Flynn, 2021) with the modifications of the number of trees (1000 instead of 100) and the minimal node size (five instead of one). All settings of the models are summarized in Tab. A.2 in appendix A.

Table 5.3.: RF model types developed in this thesis. The RF model types are determined by their inputs. The prefix "QRF" of the model name indicates that this model is a QRF model issuing continuous temperature forecasts. The prefix "RFC" of the model name indicates that this model is an RFC model issuing binary forecasts of the occurrence of cold-wave days. The number of variables shows how many different meteorological fields (before preprocessing) are taken from reanalysis data (# Var. ERA5) and reforecast data (# Var. S2S). Additionally, the month is added once to each model input as a scalar value. The "S2S Type" shows whether all members of the reforecast ensemble are considered separately or if only the condensed ensemble information is taken into account.

Model name	Input Type	# Var. ERA5	# Var. S2S	S2S Type
QRF_sel	all grid points	3	-	-
QRF_stat_sel	statistics	3	-	-
QRF_pca_sel	PCs	3	-	-
QRF_all, QRF_all_era5	all grid points	9	-	-
QRF_stat_all, QRF_stat_all_era5	statistics	9	-	-
QRF_pca_all, QRF_pca_all_era5	PCs	9	-	-
QRF_stat_all_s2s_sepf	statistics	-	10	all members
QRF_pca_all_s2s_sepf	PCs	-	10	all members
QRF_stat_all_s2s_ens	statistics	-	10	condensed
QRF_pca_all_s2s_ens	PCs	-	10	condensed
QRF_stat_all_s2s_sepf_era5	statistics	9	10	all members
QRF_pca_all_s2s_sepf_era5	PCs	9	10	all members
QRF_stat_all_s2s_ens_era5	statistics	9	10	condensed
QRF_pca_all_s2s_ens_era5	PCs	9	10	condensed
RFC_sel	all grid points	3	-	-
RFC_stat_sel	statistics	3	-	-
RFC_pca_sel	PCs	3	-	-
RFC_all, RFC_all_era5	all grid points	9	-	-
RFC_stat_all, RFC_stat_all_era5	statistics	9	-	-
RFC_pca_all, RFC_pca_all_era5	PCs	9	-	-
RFC_stat_all_s2s_sepf	statistics	-	10	all members
RFC_pca_all_s2s_sepf	PCs	-	10	all members
RFC_stat_all_s2s_ens	statistics	-	10	condensed
RFC_pca_all_s2s_ens	PCs	-	10	condensed
RFC_stat_all_s2s_sepf_era5	statistics	9	10	all members
RFC_pca_all_s2s_sepf_era5	PCs	9	10	all members
RFC_stat_all_s2s_ens_era5	statistics	9	10	condensed
RFC_pca_all_s2s_ens_era5	PCs	9	10	condensed

5.11. Explainable Artificial Intelligence

XAI-methods are used to reveal how ML models generate their predictions. Thereby, the relevance of predictors is assessed either for single predictions or the whole model training. The former is used to identify which predictors govern the forecast of a specific case. This information gives insights about possible drivers of the state of the target variable in that particular case which could be very different from the mean state of the target variable. The latter is used to identify which predictors are most often used by the model during training and therefore provide a suggestion which predictors might be drivers for the mean state of the target variable. Furthermore, this information can be utilized when the amount of predictors should be minimized to reduce computational costs. In both cases, physical relationships or causality cannot be inferred since physical equations are (with the exception of some physics-informed ML models) neither incorporated in the ML models nor the XAI methods.

5.11.1. Impurity Based Feature Importance

One way to explain which predictors are generally most useful for a given ML model's predictions is using a feature importance approach. This method measures how much the overall model performance is dependent on the predictors used during model training (Janitza et al., 2018). We use the Impurity Based Feature Importance. As described in Janitza et al. (2018), it is based on the assumption, that a predictor used often for splitting the data at the nodes of the respective DTs is more important than a seldomly used predictor (Fig. 5.4 (a)). The model using the original predictors is thereby considered as "pure". Is a predictor considered often for splitting the data at the nodes, left out, the overall number of predictors used for splitting the data inside the DTs increases which makes the model "impure". Accordingly, the model's forecasting skill decreases in comparison to the skill of the pure model. This results in a high feature importance value. The higher the feature importance value, the more important is the predictor for the model's forecasting skill.

The decrease in forecasting skill is thereby measured by the changes in variances in case of the QRF models and by the changes of the Gini index in case of the RFC models (Wright and Ziegler, 2017; Suthaharan, 2016). In case of QRFs, the optimal split is characterized by the lowermost variances inside the resulting nodes. Therefore, predictors which are used often, i.e. are important for the model's learning process, lead to a lower overall variance inside the nodes of the DTs than less important ones. In case of RFCs, important predictors lead to high decrease in the Gini index. As described in Hastie et al. (2009), after every split of the data at a node of an RFC, the Gini index of the two following nodes is less than the Gini index of the original node. The sum of the decreases of the Gini index in a DT gives an estimate of how important the variable is.

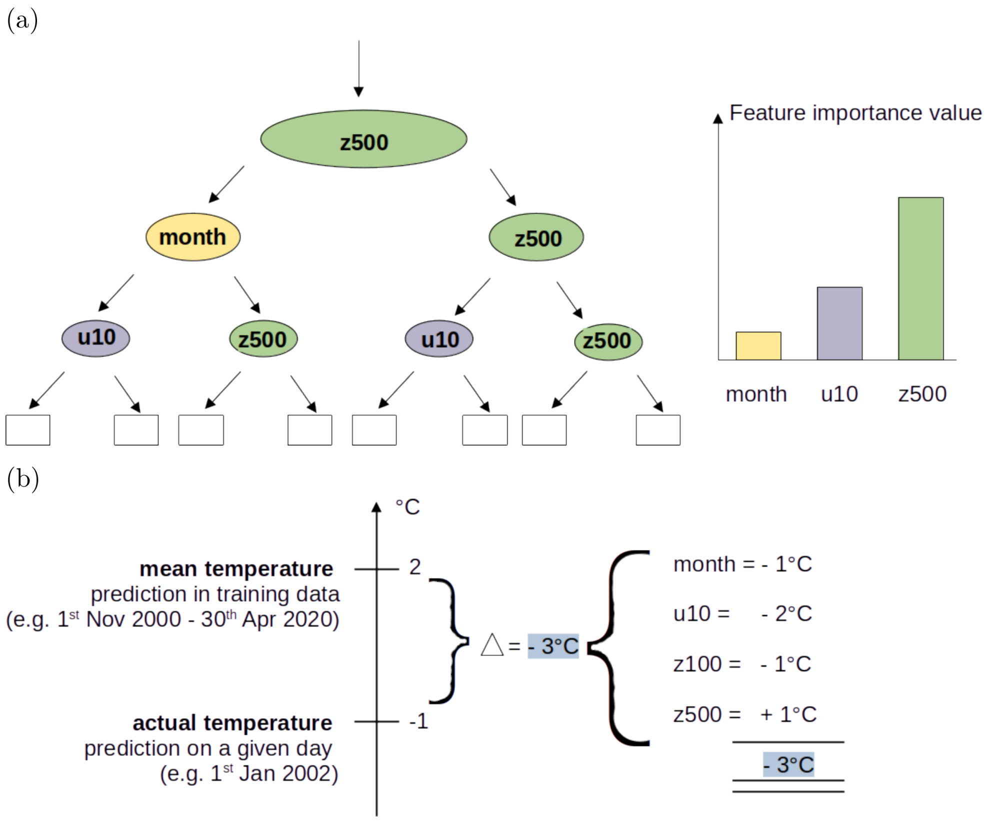


Figure 5.4.: Visualization of the concepts of feature importance and SHAP. The basic concept of feature importance is shown exemplarily for a DT (a). In each node (ellipses) the predictor used for splitting the data is shown (left). The number of occurrence of the respective predictor determines its feature importance value (right). The basic concept of SHAP is shown exemplarily for a temperature forecast (b). SHAP investigates the difference between the mean prediction in the training data and the actual prediction on a given day (left). It reveals the contribution of each predictor to this difference (right).

In our studies, we use the "skranger" package (Flynn, 2021) to directly obtain the feature importance values during training. Then, we average the values calculated for each model trained during the leave-one-(winter-)out cross-validation to create the feature importance values in a 20-winter mean. We do this to smooth possible peculiarities of certain winters obtaining a more representative picture of the most important predictors for forecasting an "average" winter.

5.11.2. Shapley Additive Explanations

In contrast to the Impurity Based Feature Importance, which detects the most important predictors during model training, we use SHAP to detect which predictors are most relevant for certain predictions (Fig. 5.4 (b)). According to Molnar (2022), SHAP computes, among other things, Shapley values. These values explicitly state to which amount each predictor contributes to the final prediction. Thereby, the contribution to the difference of the actual prediction to the mean prediction during training is shown. For example, when the mean prediction of the training set is 10°C and the actual prediction for the day of interest is 4°C, the Shapley values of the different predictors add up to the difference of 6°C. Originally, Shapley values are designed for the use with deterministic forecasts. Therefore, we apply some adaptions to our probabilistic predictions. At first, we create an average mean prediction of the training dataset. In case of the QRFs, this means creating daily averages over the predicted quantiles and then taking the mean over all days in the training dataset. In case of the RFCs, this means taking the mean over all predicted days of a respective class of the training dataset. This yields two mean predictions of the training data. One for the occurrence of cold-wave days and one for the occurrence of non-cold-wave days. The Shapley values are then calculated for the class of interest. In a second step, we compute the actual prediction of the day(s) of interest. Thereby the averaging process (taking the daily mean of the predicted quantiles or taking the class of interest) is analogously done to the averaging of the mean prediction of the training dataset. Since we analyze time periods in this study rather than days, we furthermore average over all days of interest.

Three things are important to note when interpreting Shapley values. First, Shapley values only show how much each predictor contributes to the difference of the actual prediction to the mean prediction of the training dataset. It does not give an estimate of how accurate the prediction is. Second, although maybe correlations of the determined highly contributing predictors to physically known relations can be made, causality cannot be inferred from this method. And third, due to the averaging processes we apply to our probabilistic forecasts, the calculated Shapley values do not (necessarily) add up exactly to the difference of the actual prediction to the mean prediction of the training dataset as they would in case the forecasts were deterministic.

To calculate the Shapley values, we use the implementation for SHAP by Lundberg et al. (2020) which is implemented in the python package "shap" (Lundberg, 2018). It can be applied directly to the python package "skranger" (Flynn, 2021) which is used for training the RF models.

5.12. Forecast Evaluation

Forecasts are evaluated using proper scoring rules. As described in Gneiting et al. (2007), proper scoring rules take the sharpness as well as the calibration of probabilistic forecasts into account and are therefore a suitable measure to compare forecasts of different models with each other. The sharpness of an ensemble forecast is determined via the distributions of the individual ensemble members. The more concentrated the distributions are (i.e. the more similar the mean and standard deviation), the sharper the ensemble forecast. Forecasts are calibrated, when the statistical measures of their distributions (i.e. the mean and standard deviation) are equal to the statistical measures of the distribution of the observations.

5.12.1. Continuous Ranked Probability Score

To evaluate the continuous forecasts 2-meter temperature forecasts against the continuous ground truth (the observation), we use the Continuous Ranked Probability Score (CRPS). The CRPS can be formulated as (Gneiting et al., 2007)

$$CRPS(F,x) = \int_{-\infty}^{\infty} (F(y) - \mathbf{1}\{y \geq x\})^2 \, \mathrm{d}y, \tag{5.3}$$

where $F(y)$ is the predictive CDF and x the observation. Alternativly, the CRPS can be expressed as

$$CRPS(F,x) = \mathbf{E}_F|X - x| - \frac{1}{2}\mathbf{E}_F|X - X'|, \tag{5.4}$$

where $\mathbf{E}_F$ is the expectation of a random variable with predictive CDF F, X as well as X' two independent random variables with CDF F and a finite first moment and x the observation. The CRPS is given in the same unit as the forecast and observation, and corresponds to the mean absolute error in case of a deterministic forecast. Since the CRPS is a proper scoring rule, the forecasts are the more similar to the observation the lower the CRPS value and a perfect forecast has a CRPS value of zero.

5.12.2. Brier Score

For evaluating the binary forecasts of the occurrence of cold-wave days against the binary ground truth, we use the Brier Score (BS). The BS can be expressed as (Brier, 1950):

$$BS_{\mathrm{F}}(y) = (F(1) - \mathbf{1}\{y = 1\})^2, \tag{5.5}$$

$$F(1) = \hat{P}(y = 1),$$

where $F(1)$ is the predicted probability of an event. The BS is a dimensionless measure, always positive and the lower the better. A perfect forecast has a BS value of zero.

5.12.3. Skill Scores

To compare the forecast scores to a reference score, obtained by the climatological benchmark ensemble, skill scores are used (e.g. Taillardat et al., 2016). Skill scores are computed from the mean of the scores over the respective winters. From these 20 skill scores, a 20-winter mean skill score is additionally computed. In case of the continuous forecasts, the Continuous Ranked Probability Skill Score (CRPSS) is used which can be formulated as

$$CRPSS = 1 - \frac{CRPS_{\text{model, averaged}}}{CRPS_{\text{benchmark, averaged}}}. \tag{5.6}$$

In case of the binary forecasts, the Brier Skill Score (BSS) is used which can be expressed analogously as

$$BSS = 1 - \frac{BS_{\text{model, averaged}}}{BS_{\text{benchmark, averaged}}}. \tag{5.7}$$

The values of skill scores range between $-\infty$ and one. Negative values show that the model's forecasts are less skillful compared to the forecasts of the benchmark model. A skill score of zero denotes that both the predictions of the model and the benchmark model are equally skillful. Positive values show that the model's forecasts are more skillful in comparison to the benchmark model's predictions whereby a perfect model reaches a skill score of one.

5.13. Significance Testing with Welch's T-Test

We use Welch's t-test (Welch, 1947) to estimate possible significances in skill differences of various subsets of predictions based on their CRPS or BS values. Welch's t-test is a modification of Student's t-test to be applicable to subsamples with different variances. As described in Moser and Stevens (1992), it is tested whether the two subsamples have (sufficiently) equal mean values of their distributions which is the null hypothesis. To do so, the test statistics of Welch's t-test is calculated as

$$t = \frac{(\bar{x} - \bar{y})}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}, \tag{5.8}$$

where $\bar{x}$ is the mean of the first sample with its corresponding standard deviation s_1 and sample size n_1 as well as $\bar{y}$ is the mean of the second sample with standard deviation s_2 and sample size n_2. The null hypothesis is rejected if the calculated statistic t is greater than the statistics t^* of a predefined function. In this case, Student's t-function is used. In a next step, the probability of t being greater than t^* is computed. The resulting p-value gives an estimate whether the null hypothesis should be rejected or not according to a chosen significance level α. As commonly used, we set α to 0.05. If the calculated p-value is smaller than α, the null hypothesis is rejected and thus the differences between the two subsamples significant. We perform the calculation of the p-values using the Python package "scipy" (The SciPy community, 2024).

6. Random Forest Models as an Alternative to Numerical Weather Prediction Models

In the first part of our research we focus on the use of RF models as an alternative to NWP models for the direct forecasting of 2-meter temperatures and the occurrence of cold-wave days at lead times of 14, 21 and 28 days. The predictions are provided for the Central European mean as described in section 5.1. The set-up of the used RF models is described in section 5.10. ERA5 reanalysis data, as described in section 5.2, is used as input for the models. The ground truth is taken from E-OBS observational data, described in section 5.1. For now, the ML models are only evaluated with respect to a climatological benchmark ensemble, which is described in section 5.5. This gives a first insight about the suitability of the RF models for the forecasting task at hand.

6.1. Properties of Predictions of the Climatological Benchmark Ensemble and the RF Models

To demonstrate the different properties of the predictions of the climatological benchmark ensemble and RF models, two winters are chosen as case studies. These are the winters 2011/2012 and 2013/2014. Both differ considerably in terms of their 2-meter temperature evolution. Most distinctively, in the absence of cold waves during the winter 2013/2014 and the presence of a severe cold wave in the winter 2011/2012 starting rapidly at the end of January (Fig. 6.1).

The climatological benchmark ensemble is a static ensemble created from the winters 1970/1971-1999/2000 (section 5.5). Therefore, its predictions are the same for each winter. The temperature range of these is between roughly 8 and 15 K at each day of the winter, always comprising both temperatures below and above the frost point (red shaded area on Fig. 6.1). Its maximum temperature of roughly 287 K is observed at the end of April, the minimum temperature of approximately 263 K at the beginning of January.

In contrast to the climatological benchmark ensemble, the predictions of the RF models differ for each forecasted winter (blue dotted shading Fig. 6.1). This is shown exemplary for one (arbitrary) RF model and the lead time of 14 days. In case of the winter 2011/2012, the maximum value of the predictions is circa 287.5 K and found at the end of April . The minimum value of 266 K occurs at the end of January and in the beginning of March (Fig. 6.1 (a)). Thereby, the temperature range at each day varies between

6.5 K and 15 K. In case of the winter 2013/2014, the maximum predicted temperature value is also around 287.5 K at the end of April but the minimum temperature of roughly 266 K is predicted only at the end of January. The temperature range at each day is similar to the one seen for the winter 2011/2012, varying between 6.5 K and 15 K (Fig. 6.1 (b)).

The similar maximum and minimum temperatures as well as the same temperature range of forecasts as the climatological benchmark ensemble, show that the climatological state plays a large role in the predictions of the RF model. Nevertheless, the different position of the maximum and minimum temperatures as well as the maximum and minimum of the daily temperature range indicate that the RF model is able to learn relations in the data beyond climatology. The same is true for the binary prediction of cold-wave days.

For an exemplary visualization, the predictions of one RFC model are shown at a lead time of 14 days. For consistency, the RFC model using the same input as the chosen model for continuous temperature forecasting is selected. By design, a fixed but daily varying fraction of members of the climatological benchmark ensemble predicts a cold-wave day (Fig. 6.2). The highest ratio of ensemble members predicting a cold-wave day is thereby found in mid February with roughly 40%. The lowest ratio with 0% of the ensemble members predicting a cold-wave day at the end of March and in April. Since these predictions are binary, a range of predicted values as seen in case of the continuous temperature forecasts is not predicted. Concerning the predictions of the RFC model, the daily fraction varies between winters (Fig. 6.2). This shows that the RFC is basing its predictions on different predictors every day and between winters instead of relying only on the same predictors and producing the same forecast value every time. In case of the winter 2011/2012, the highest fraction of ensemble members predicting a cold-wave day is found, similarly to the climatological benchmark ensemble, in mid February reaching a value of slightly above 40% (Fig. 6.2 (a)). The lowest ratio with around 1% is found in the beginning of January. In case of the winter 2013/2014, the highest ratio of ensemble members predicting a cold-wave day is seen in mid April reaching a value around 53%. The lowest ratio with a value of approximately 2% is found at the end of December (Fig. 6.2 (b)). Although the predictions of the climatological benchmark ensemble and the RFC model differ, they are still roughly in the same range indicating that also in case of the binary predictions the climatological state plays a non-negligible role in the ML-forecasts (Fig. 6.2).

6.2. Comparison of Predictions from Random Forest Models to the Climatological Benchmark Ensemble

The naive expectation is that ML models can outperform a climatological benchmark ensemble at all analyzed lead times and days. This would imply a positive skill in the 20-winter mean and individual winters. The predictive skill of the RF models is thereby assessed in term of the CRPSS for the continu-

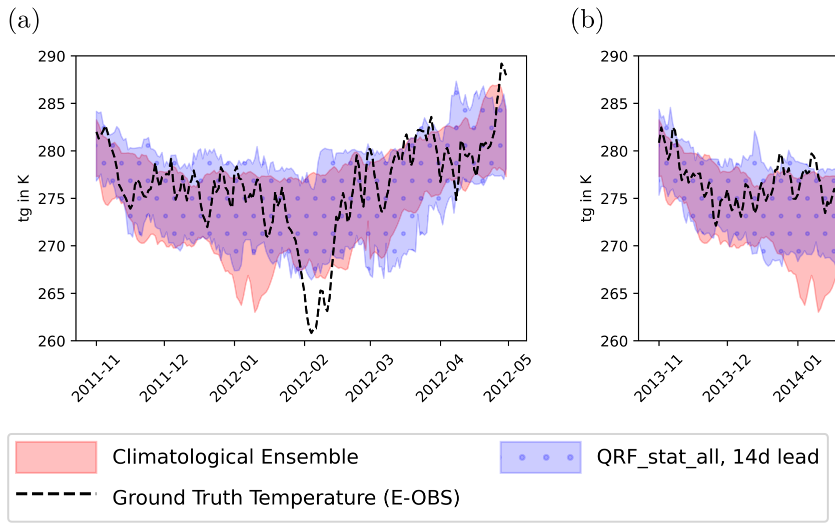


Figure 6.1.: Continuous winter temperature forecasts of the climatological benchmark ensemble and a chosen QRF model. The predictions of the climatological benchmark ensemble and a QRF model trained on the minimum, mean, maximum and variance of nine meteorological predictor fields as well as the month are shown for the winters 2011/2012 (a) and 2013/2014 (b).

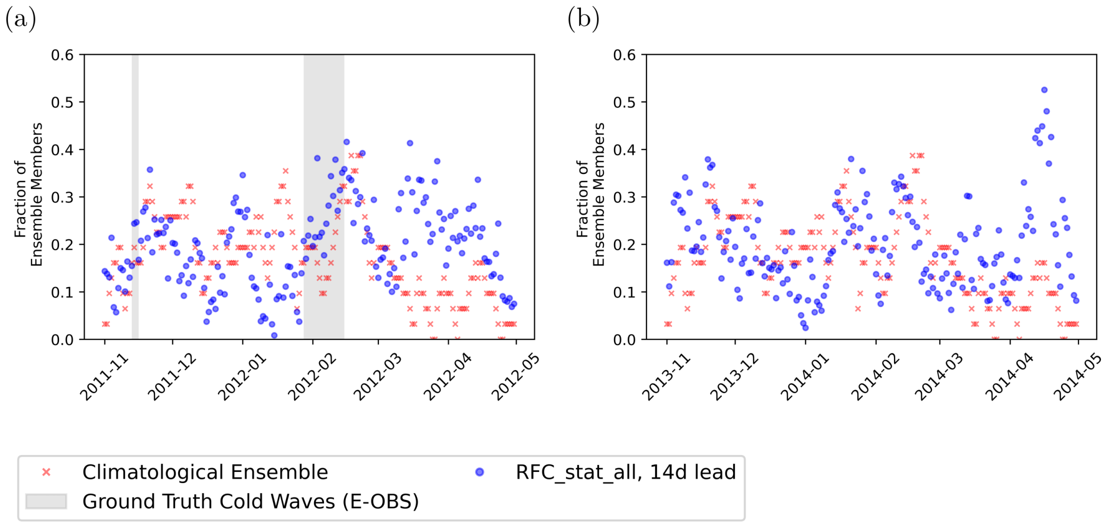


Figure 6.2.: Forecasts of the occurrence of cold-wave days of the climatological benchmark ensemble and a chosen RFC model. The predictions of the climatological benchmark ensemble and an RFC model trained on the minimum, mean, maximum and variance of nine meteorological predictor fields as well as the month are shown for the winters 2011/2012 (a) and 2013/2014 (b).

ous temperature forecasts and the BSS for the binary predictions, both with respect to the climatological benchmark ensemble.

6.2.1. Skill of 2-Meter Temperature Predictions in the 20-Winter Mean

To obtain the 20-winter mean CRPSS of model predictions, firstly, the winterwise CRPSS is calculated for the whole evaluation period. To do so, the CRPSS is calculated from the CRPS values of the RF model and the climatological benchmark ensemble, average temporally over the whole winter. The resulting 20 values are then again averaged to create the 20-winter mean. The 20-winter mean CRPSS shows at all lead times only small deviation from zero but a large standard deviation (Fig. 6.3). This indicates that the QRF models' performance strongly depends on the exact winter. The QRF models using all grid points of the meteorological fields as input (#1+2) show the highest mean CRPSS values per lead time followed by the models using the statistics of the fields as input (#3+4). The QRF models trained with nine meteorological predictors and the month (#1,3,5) show a better performance than the models trained with three meteorological predictors and the month (#2,4,5). In general, the skill of all QRF models decreases with increasing lead times.

At a lead time of 14 d, the best performing models are the QRF models trained on all grid points of all predictor fields and only selected predictor fields as well as the QRF models trained on the statistics of all predictor fields (#1-3 on Fig. 6.3 (a)). All other QRF models show negative mean CRPSS values (Fig. 6.3 (a)). For a lead time of 21 d, all QRF models except the one trained on all grid points of all predictors fields (#1) show negative mean CRPSS values (Fig. 6.3 (b)). The same picture is seen at a lead time of 28 d (Fig. 6.3 (c)).

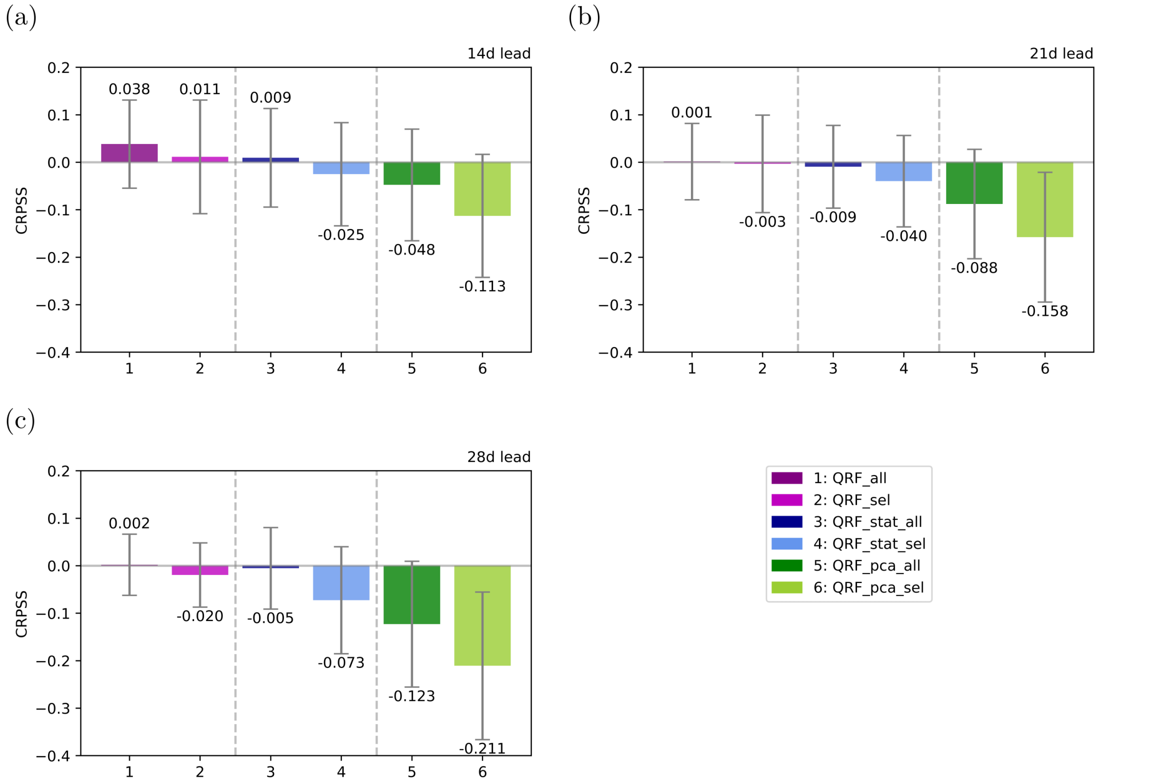


Figure 6.3.: CRPSS of the QRF models with respect to the climatological benchmark ensemble. The 20-winter mean (Winter 2000/2001 - Winter 2019/2020) CRPSS is shown for a lead time of 14 days (a), 21 days (b) and 28 days (c). The whiskers show the 5th and 95th percentiles of the wintermean CRPSS values over the 20 winters, the printed values the mean CRPSS value, which corresponds to the height of the bar.

6.2.2. Skill of 2-Meter Temperature Predictions for Two Individual Winters

As shown in subsection 6.2.1, the large standard deviations in CRPSS values between the winters at all lead times imply that the skill is not equally distributed between winters. Thus, depending on the winter, the model set-ups may perform differently well and the optimal set-up varies. Therefore, we look at the CRPSS values of two individual winters. In order to take the large variability between winters into account, we use the two winters described in section 6.1 since their temperature evolution varies substantially.

At a lead time of 14 days, for both analyzed winters, four QRF models show positive mean CRPSS values (Fig. 6.4 (a) and (b)). In case of the winter 2011/2012, the best performing model is the QRF trained on the statistics of the selected predictor fields (#4 on Fig. 6.4 (a)). In case of the winter 2013/2014, it is the model trained on all grid points of all predictor fields (#1 on Fig. 6.4 (b)). At a lead time of 21 days, two models show positive CRPSS values in case of the winter 2011/2012, but none in case of the winter 2013/2014. This shows how difficult forecasting on the subseasonal timescale can be (Fig. 6.4 (c) and (d)). The best performing model in predicting the 2-meter temperature of the winter 2011/2012 is the QRF trained on all grid points of the selected predicted fields (#2 on Fig. 6.4 (c)). When looking at a lead time of 28 days, for both winters all QRF models show only negative mean CRPSS values (Fig. 6.4 (e) and (f)).

6.2.3. Skill of Predictions of the Occurrence of Cold-Wave Days in the 20-Winter Mean

The skill of the RFC models predicting the binary occurrence of cold-wave days is measured in term of the BSS. The 20-winter mean BSS is thereby calculated analogously to the 20-winter mean CRPSS.

When looking at the predictive skill of these RFC models at a lead time of 14 d, four models show positive mean BSS values in the 20 winter mean (Fig. 6.5 (a)). These are the RFC models trained on all predictor fields (#1,3,5) and the RFC model trained on the statistics of only the selected predictor fields (#4). Two of these models, the RFCs trained on the statistics of all predictor fields and the RFCs trained on the first ten PCs of all predictor fields, show also positive BSS values at a lead time of 21 days (#3+5 on Fig. 6.5 (b)). This is an improvement compared to the forecast of the continuous 2-meter temperature, for which only one of the QRF models shows skill at a lead time of 21 days (Fig. 6.3 (b)). At a lead time of 28 d, in the 20 winter mean none of the RFC models shows a positive mean BSS value (Fig. 6.5 (c)). The standard deviation of all models is again high, suggesting that a detailed look at case studies is useful. For consistency and comparability of both approaches, the same case studies as determined for the continuous temperature forecasts are used.

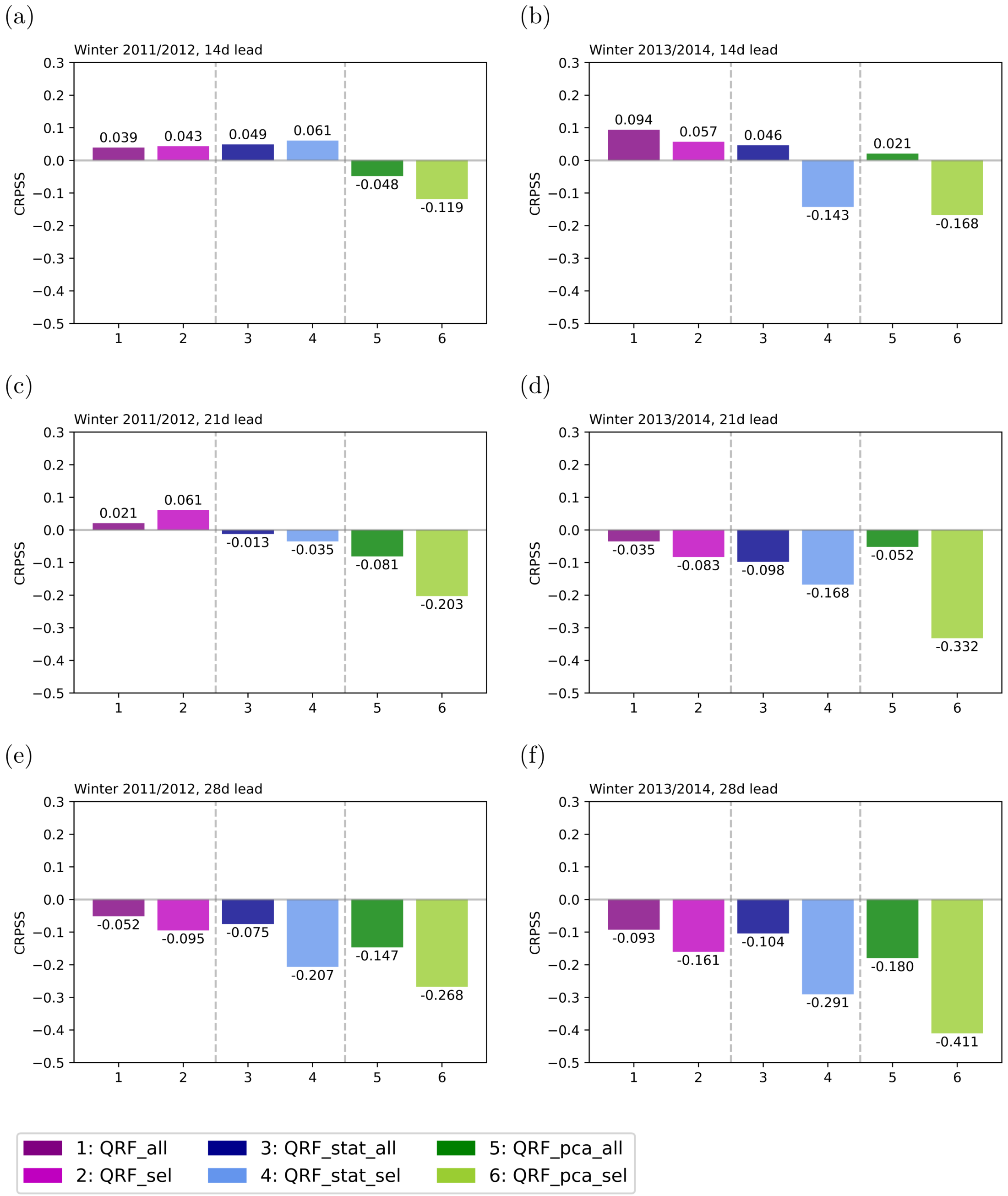


Figure 6.4.: Wintermean CRPSS of the QRF models with respect to the climatological benchmark ensemble. Shown for the winters 2011/2012 ((a), (c), (e)) and 2013/2014 ((b),(d),(f)) for lead times of 14 days ((a),(b)), 21 days ((c),(d)) and 28 days ((e),(f)).

6.2.4. Skill of Predictions of the Occurrence of Cold-Wave Days for Two Individual Winters

When assessing the skill of the RFC models with respect to the climatological benchmark ensemble for individual winters, it is striking that all RFC models show positive BSS values for the winter 2011/2012 at a lead time of 14 d (Fig. 6.6 (a)). The best performing model is thereby the RFC using the statistics of all predictor fields as input (#3). In the case of the winter 2013/2014, only two models show positive BSS

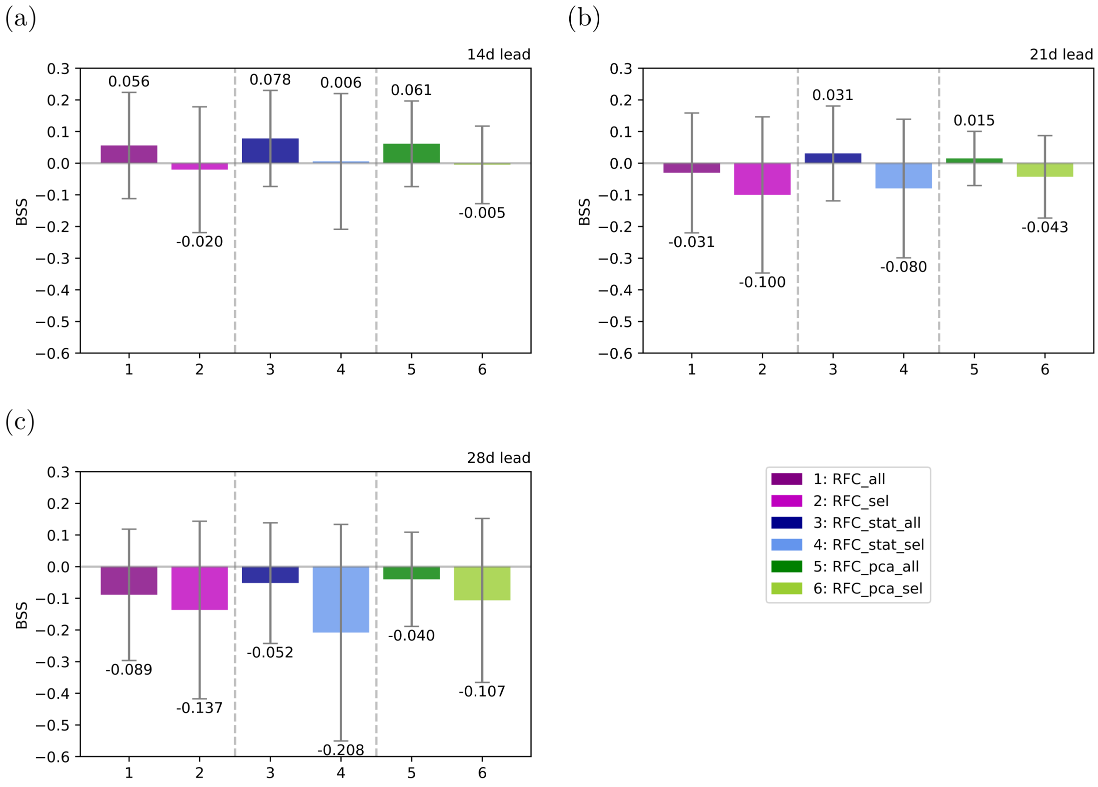


Figure 6.5.: BSS of the RFC models with respect to the climatological benchmark ensemble. The 20-winter mean (Winter 2000/2001 - Winter 2019/2020) BSS is shown for a lead time of 14 days (a), 21 days (b) and 28 days (c). The whiskers show the 5th and 95th percentiles of the wintermean BSS values over the 20 winters, the printed values the mean BSS value, which corresponds to the height of the bar.

values (Fig. 6.6 (b)). These are the RFC trained on the statistics of all predictor fields (#3) and the first ten PCs of all predictor fields (#5). At lead times of 21 and 28 days, none of the RFC models shows positive BSS values for the winter 2013/2014 (Fig. 6.6 (d) and (f)). In case of the winter 2011/2012, though, skill is also found for some RFC models on these longer lead times, whereby the RFCs trained on the first ten PCs of both, all predictors fields and only the selected predictor fields, perform best (Fig. 6.6 (c) and (e)).

6.3. Distribution of Daily Forecast Skill at the Example of Two Individual Winters

In section 6.2, we focus on the mean skill of predictions. Now, we investigate the daily skill distribution across two individual winters. This is done to investigate a possible dependence of the models' performances of the temperature or time in winter. For consistency, we choose the winters 2011/2012 and 2013/2014 again as case studies. The characteristics of these winters are described in section 6.1. Since the used skill scores are mean values, we now use the daily difference in scores. In case of the continuous forecasts this is the CRPS difference and in case of the binary forecasts it is the BS difference, both between the respective RF model's predictions and the predictions of the climatological benchmark

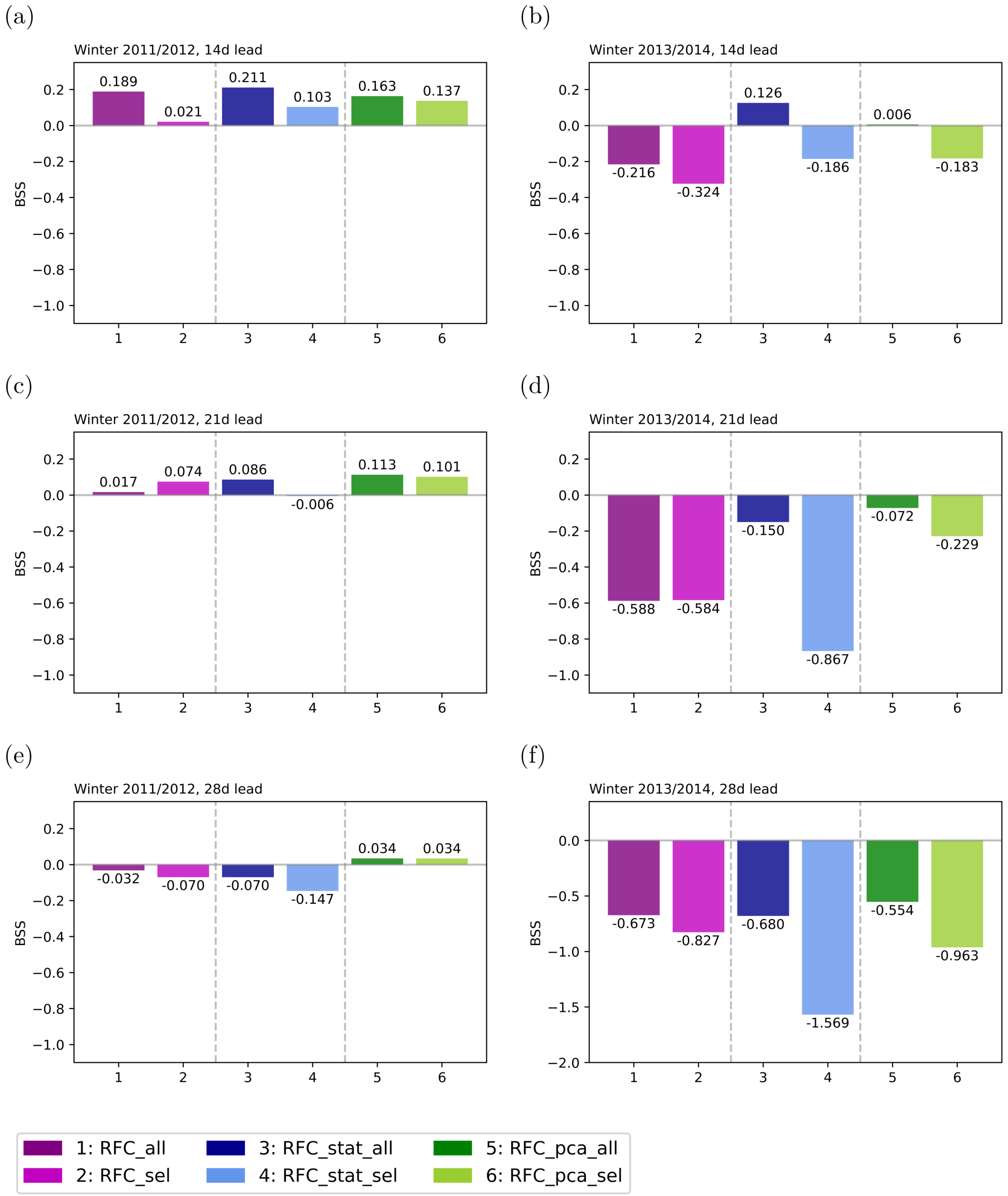


Figure 6.6.: Wintermean BSS of the RFC models with respect to the climatological benchmark ensemble. Shown for the winters 2011/2012 ((a), (c), (e)) and 2013/2014 ((b),(d),(f)) for lead times of 14 days ((a),(b)), 21 days ((c),(d)) and 28 days ((e),(f)). Note the different scale on (f).

ensemble. Positive values of these score differences denote that the RF model's predictions are more skillful than the forecasts of the climatological benchmark ensemble at that day.

6.3.1. Evolution of Skill of 2-Meter Temperature Forecasts

When looking at the daily CRPS difference of the winter 2011/2012, it is striking that at lead times of 14 and 21 days, the QRF models perform especially well during the time of the mid-winter cold wave

(Fig. 6.7 (a) and (c)). At a lead time of 14 days, it is already evident from the spread of the predictions of the climatological benchmark ensemble that it is not able to capture the severe cold wave in January and February (Fig.6.1 (a)). On the contrary, the smaller variations in temperatures around the winter-mean are captured well.

At a lead time of 28 days and for the winter 2013/2014, a dependence of the ML models' performance on temperature is not seen (Fig. 6.7 (b), (d), (e) and (f)). This might be due to the fact that during this winter the temperatures lie in the range of the climatological benchmark ensemble at almost all days (Fig.6.1 (b)). The described evolution of skill with a better performance of the RF models during long-lasting severe cold waves are also seen for the remaining winters between 2000/2001 and 2019/2020 (Fig. B1 to B6 in the appendix B).

6.3.2. Evolution of Skill of Forecasts of the Occurrence of Cold-Wave Days

A detailed look on the daily BS difference shows that, similar to the QRFs, the RFCs perform especially well during long-lasting and mid-winter cold waves (Fig. 6.8 (a), (c) and (e) as well as Fig. B1 - B6 in the appendix B). However, as seen for the fraction of ensemble members predicting a cold-wave day during the winter 2011/2012, the highest predicted probability of a cold-wave day is below 0.5 (Fig. 6.2 (a)) This shows that the RFC is not certain in predicting these cold-wave days. Nevertheless, the severe cold wave in mid-winter and the short cold wave in the beginning of the winter are better captured than by the climatological benchmark ensemble. Interestingly, in the case of the cold-wave-free winter 2013/2014, the highest predicted probability of a cold-wave day is around 0.55 (Fig. 6.2 (b)). Generally, during times with mild temperatures, a clear pattern in RFC performance is not found (Fig. 6.8 (b), (d) and (f) as well as Fig. B1 - B6 in appendix B)). This is also true for comparably short cold-wave periods.

6.4. Quantification of Predictor Contributions to the Random Forest Models' Predictions Using Shapley Additive Explanations

In order to investigate what governs the performance of the RF models during the different periods with cold and mild temperatures, the predictions of the models for the winters 2011/2012 and 2013/2014 are analyzed in more detail. To do so, the winter 2011/2012 is split into periods with cold waves and periods with warm temperatures. The periods with warm temperatures are thereby defined as days not belonging to cold waves (all areas around the grey boxes on Fig. 6.2 (a)). The winter 2013/2014 is analyzed as one since it does not contain any cold waves.

In this study, SHAP is used to examine the contributions of predictors to the RF models' forecasts. Ideally, ML models learn physically relevant pattern in the data and use them for predicting. Independent of the method used to reveal the importance of predictors for the forecasts of ML models, only correlation

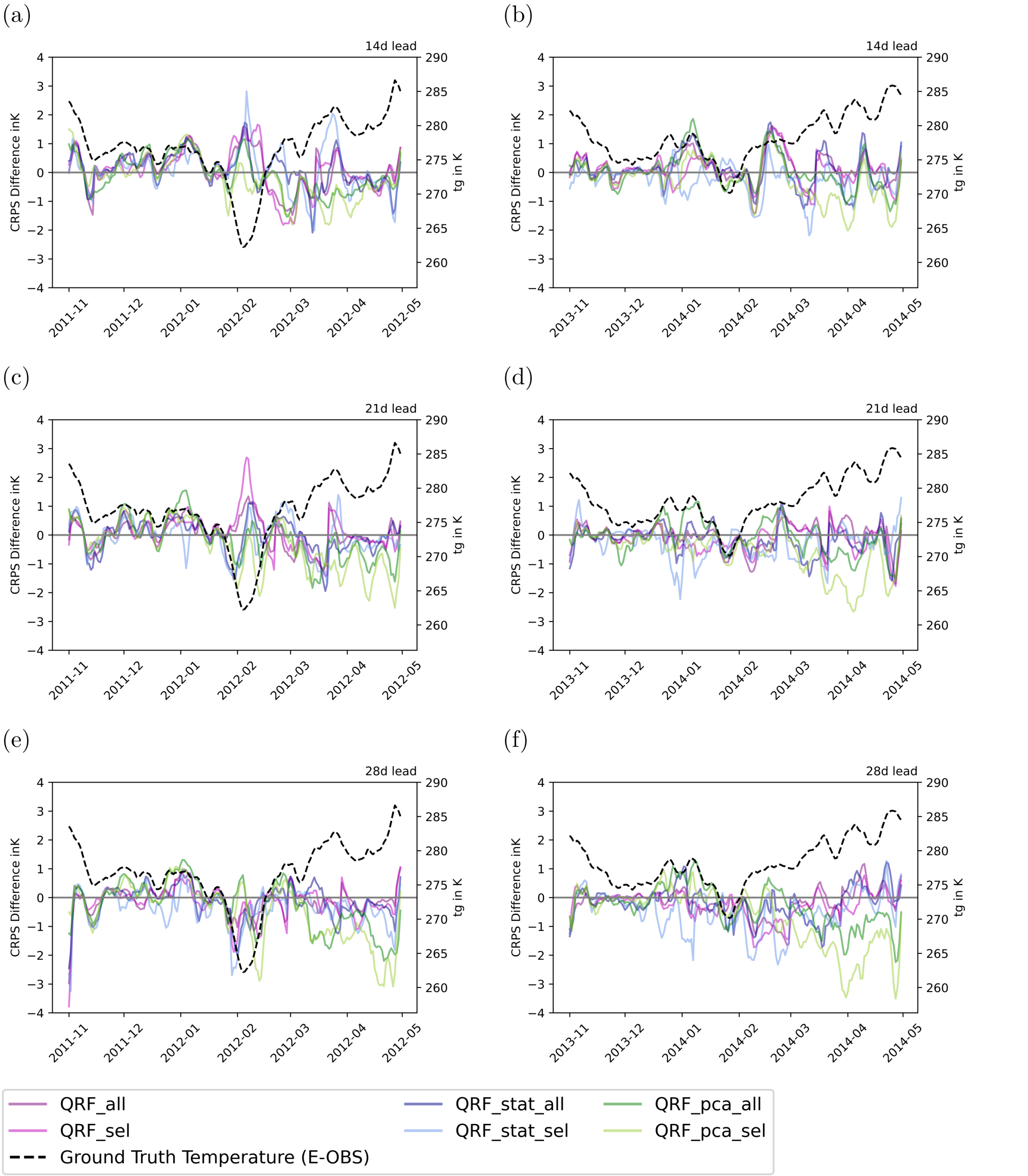


Figure 6.7.: Daily CRPS difference of the QRF models and the climatological benchmark ensemble. The daily CRPS difference is shown for the winters 2011/2012 ((a), (c), (e)) and 2013/2014 ((b),(d),(f)) for lead times of 14 days ((a),(b)), 21 days ((c),(d)) and 28 days ((e),(f)).

but no causality can be inferred.

Due to the aggregation of forecasts, the mean value and the spread of SHAP values is shown. Furthermore, we concentrate on the top five most positively and negatively contributing predictors to the forecasts. SHAP values are explained in detail in section 5.11.2. The models used for the analysis are the QRF and RFC models with a lead time of 14 days using the statistics of all predictor fields and the

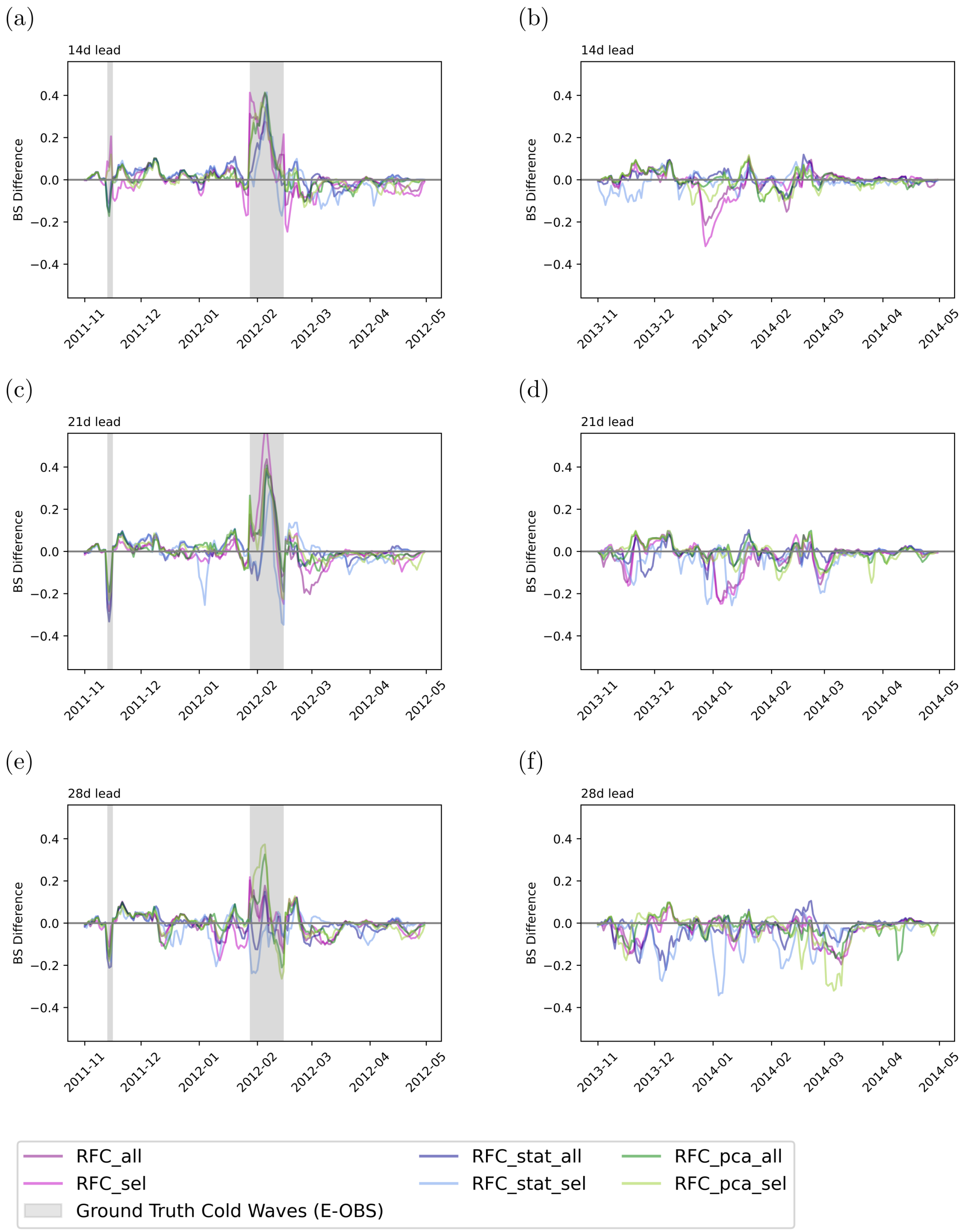


Figure 6.8.: Daily BS difference of the RFC models and the climatological benchmark ensemble. The daily BS difference is shown for the winters 2011/2012 ((a), (c), (e)) and 2013/2014 ((b),(d),(f)) for lead times of 14 days ((a),(b)), 21 days ((c),(d)) and 28 days ((e),(f)).

month as input. These are used since their predictors are easily understandable by humans and since they are limited to four per predictors field, at least three different meteorological variables have to be part of the top ten predictors contributing to the models' forecast. In case of the RF models using the first ten PCs of the meteorological fields as input, theoretically, a single meteorological variable could appear in all of the top ten predictors. In case of the RF models using all grid points of the meteorological fields as input, the relevance of a single value at a grid point is not meaningful to a human interpreter.

6.4.1. Most Important Contributing Predictors to 2-Meter Temperature Forecasts

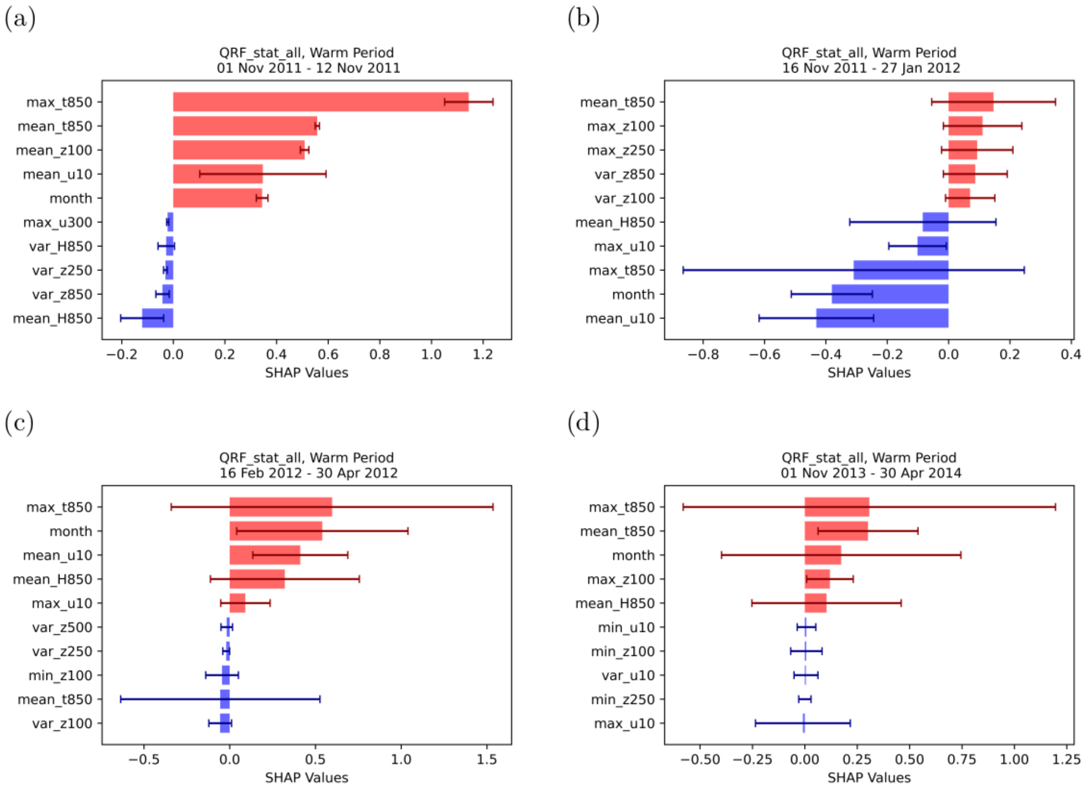
In case of the warm periods of both winters, the top five contributing predictors according to the SHAP values differ for every period. This shows that the predictions of the QRF model is not solely based on a fixed set of predictors with a fixed contribution of each. Common candidates are either the mean or maximum of the $t850$ which are contributing positively to the prediction.

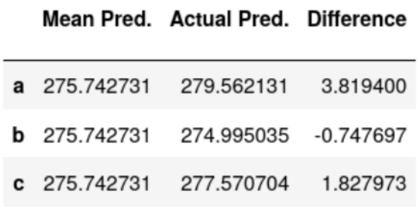
During the first warm period in November 2012, the main driver for the higher than average temperature prediction is the maximum of $t850$ (Fig. 6.9 (a)). Although it is not the main contributor to the forecast value, the month also contributes positively to the prediction, underlining the suggestion of the dependence of the model performance on the time of season. The month is furthermore one of the most positively contributing predictors of the warm period between February and April 2012 (Fig. 6.9 (c)). During the warm period from mid-November 2011 to end of January 2012, the months contributes negatively to the forecast (Fig. 6.9 (b)). This also underlines the time dependence of the forecast. It is interesting to note here that the maximum and mean of $u10$ also contribute negatively to the prediction. When looking at the winter 2013/2014, this is also visible in the mean contribution of the mean, maximum and variance of $u10$, but the standard deviation spans across positive and negative SHAP values (Fig. 6.9 (d)). The same is true for the month and the maximum of $t850$ but with a positive mean SHAP value.

For both cold and warm period predictions in mid-winter, the month contributes negatively to the predicted temperature, otherwise positively. This underlines the time dependence of the forecast and is physically plausible. In case of the cold periods of the winter 2011/2012, common candidates are furthermore $t850$ and, additionally, $u10$. Depending on the cold wave period, the contribution is either positive or negative. A pattern is not obvious. During the first cold wave in November, $t850$ contributes positively to the prediction and the mean of $u10$ negatively (Fig. 6.10 (a)). Interestingly, on the other hand, the most positively contributing predictor to the prediction of the severe cold wave in mid-winter is the mean of $u10$ (Fig. 6.10 (b)). The maximum of $t850$ and the month are the most negative contributors.

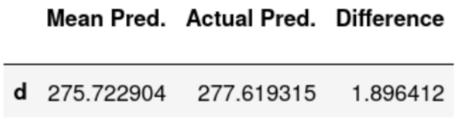
6.4.2. Most Important Contributing Predictors to the Forecasts of the Occurrence of Cold-Wave Days

When looking at the SHAP analysis of the RFC models, during the warm periods of the winter 2011/2012, either the mean, maximum or variance of $t850$ contributes positively to the forecast (Fig. 6.11). This is similar to the QRF models. During the warm period in November 2011, the most positively contributing





	Mean Pred.	Actual Pred.	Difference
a	275.742731	279.562131	3.819400
b	275.742731	274.995035	-0.747697
c	275.742731	277.570704	1.827973

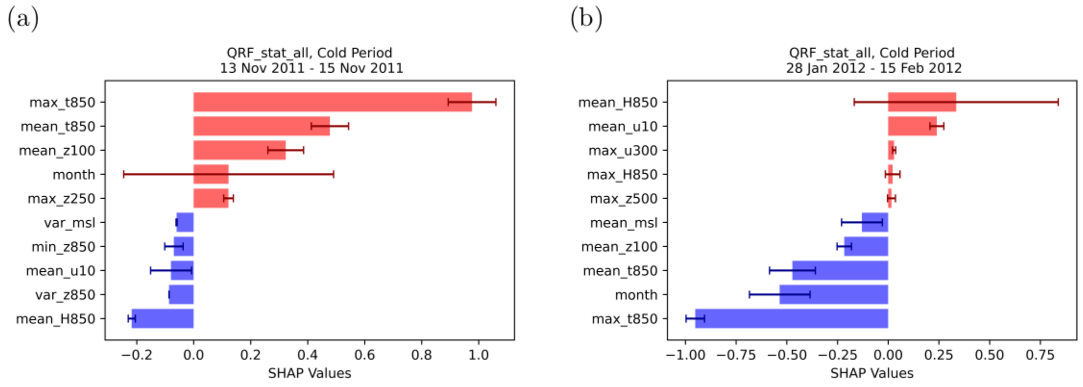


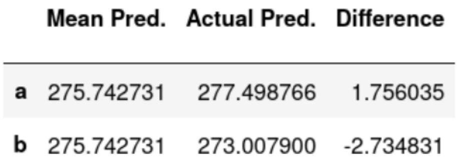
	Mean Pred.	Actual Pred.	Difference
d	275.722904	277.619315	1.896412

Figure 6.9.: SHAP Values for the continuous forecasts during the warm periods during the winters 2011/2012 and 2013/2014. The top five positive and negative mean SHAP values of the QRF_stat_all models at a lead time of 14 days for the warm periods during the winter 2011/2012 ((a), (b), and (c)) and 2013/2014 ((d)) are shown. The whiskers show the standard deviation of the SHAP values during the respective time period.

predictors are the maximum and mean of $t850$ (Fig. 6.11 (a)). The most negatively contributing predictors the variance of $z100$, $u10$ and the maximum of $u10$. The month is contributing positively. The month is also contributing positively to the predictions during the warm period occurring between February and April 2012 but negatively during the warm period between November and January 2012 (Fig. 6.11 (b) and (c)). This leads to the assumption that, similar as in case of the QRF models, the model's predictions depend on the time of season.

During the warm period between November and January 2012, the most positively contributing predictors are the variance and mean of $t850$ and the most negatively contributing predictor the mean of $u10$ (Fig. 6.11 (b)). In case of the warm period between February and April 2012, it is the variance of $z100$ (Fig. 6.11 (c)). The most positively contributing predictors to the forecasts are the maximum of $t850$





	Mean Pred.	Actual Pred.	Difference
a	275.742731	277.498766	1.756035
b	275.742731	273.007900	-2.734831

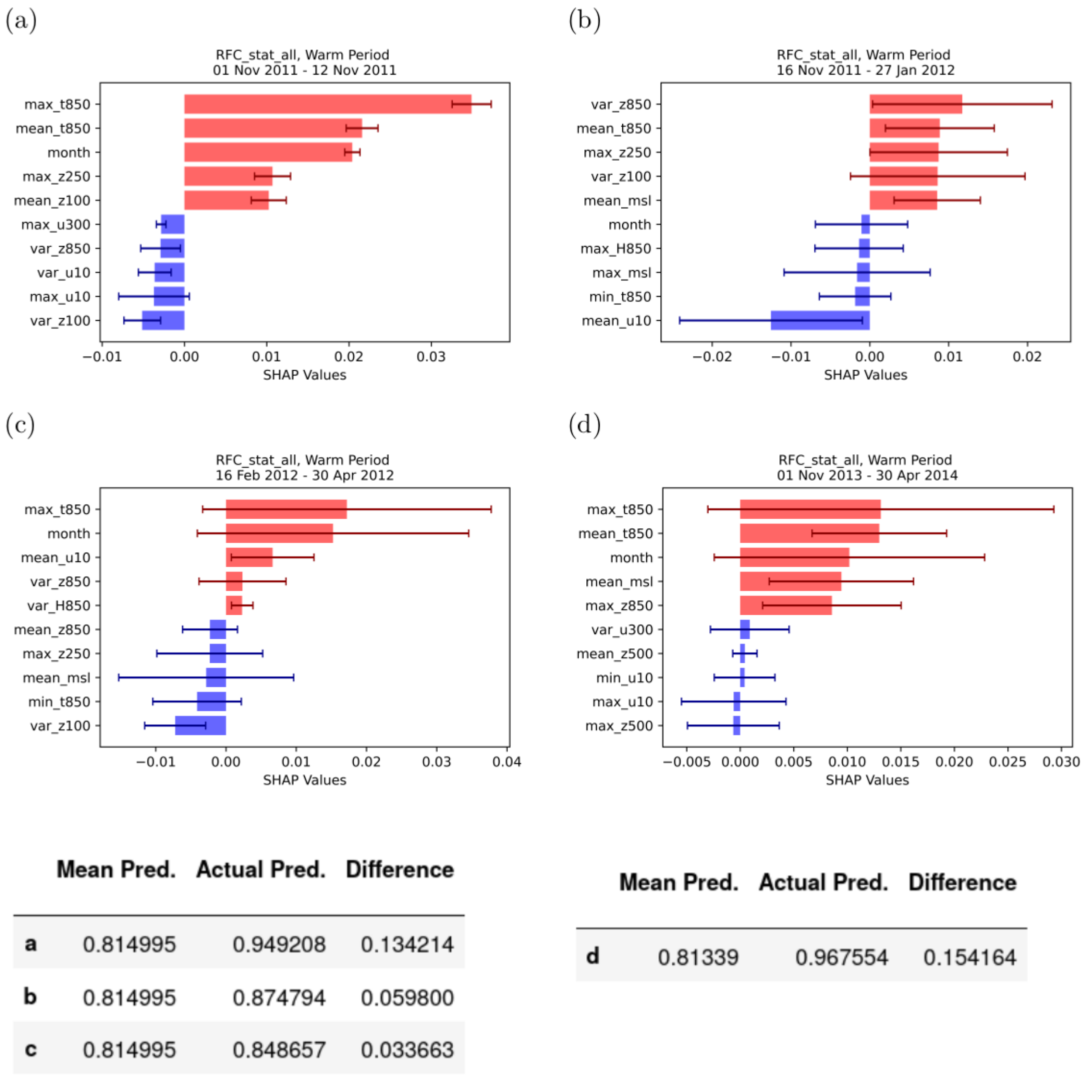
Figure 6.10.: SHAP Values for continuous forecasts during the cold periods during the Winter 2011/2012. The top five positive and negative mean SHAP values of the QRF_stat_all models at a lead time of 14 days for the cold periods during the winter 2011/2012 ((a) and (b)) are shown. The whiskers show the standard deviation of the SHAP values during the respective time period.

and the month. This also the case during the winter 2013/2014 where additionally the mean of $t850$ is among the most positively contributing predictors to the forecasts (Fig. 6.11 (d)). The most negatively contributing predictors are the maximum of $u10$ and $z500$ but their absolute contribution is a lot smaller than the contribution of the positively contributing predictors.

In case of the prediction of the cold wave in November 2011, the most positively contributing predictors are the variance of $z100$ and the maximum of $z500$ (Fig. 6.12 (a)). The most negatively contributing predictors are the maximum and mean of $t850$. In this case, the negative contributions are considerably higher than the positive ones. Concerning the severe mid-winter cold wave, the positive contributions are higher than the negative ones (Fig. 6.12 (b)). Interestingly, the mean and maximum of msl is one of the most positively contributing predictors besides the maximum and variance of $t850$. The most negatively contributing predictors are the mean and maximum of $H850$.

6.5. Summary and Discussion

In this part of our research we demonstrate that a rather simple combination of physical knowledge about the weather forecasting task at hand and RF models can lead to improved subseasonal forecasts compared to a climatological benchmark ensemble. As described in section 3.4, ML models used for direct forecasting on the subseasonal timescale are able to improve predictions compared to climatological benchmarks. One study, van Straaten et al. (2022), uses thereby RF models for forecasting summertime

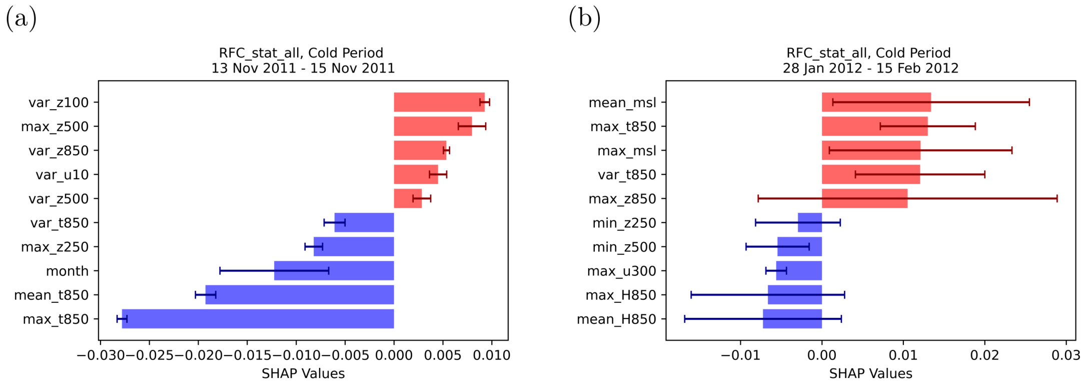


	Mean Pred.	Actual Pred.	Difference
a	0.814995	0.949208	0.134214
b	0.814995	0.874794	0.059800
c	0.814995	0.848657	0.033663

	Mean Pred.	Actual Pred.	Difference
d	0.81339	0.967554	0.154164

Figure 6.11.: SHAP Values for the binary forecasts during the warm periods during the winters 2011/2012 and 2013/2014. The top five positive and negative mean SHAP values of the RFC_stat_all models at a lead time of 14 days for the warm periods during the winter 2011/2012 ((a), (b) and (c)) and 2013/2014 ((d)) are shown. The whiskers show the standard deviation of the SHAP values during the respective time period.

hot temperature events in Europe. As predictors, they use meteorological variables from ERA5 reanalysis. In our study, we follow a similar approach but tailored to our forecasting task which is the prediction of the 2-meter temperature and occurrence of cold-wave days in Central Europe at lead times of 14, 21 and 28 days. Similar to van Straaten et al. (2022), we also use RF models. QRFs are used in case of probabilistic forecasts of 2-meter temperatures and RFCs in case of predictions of the probability of occurrence of cold-wave days. Since we focus on wintertime instead of summertime temperatures, our chosen predictors differ from the ones selected by van Straaten et al. (2022). Using predictors chosen on physical knowledge is a straight-forward approach to limit the training data to the most sensible values and thereby optimize the use of computational resources. Furthermore, the architecture of the RF models



	Mean Pred.	Actual Pred.	Difference
a	0.185005	0.109061	-0.075944
b	0.185005	0.278979	0.093973

Figure 6.12.: SHAP Values of the binary forecasts during the cold periods during the winter 2011/2012. The top five positive and negative mean SHAP values of the RFC_stat_all models at a lead time of 14 days for the cold periods during the winter 2011/2012 ((a) and (b)) are shown. The whiskers show the standard deviation of the SHAP values during the respective time period.

needs not to be modified in order to include physical knowledge into the models' predictions.

In terms of predictors we concentrate on the month as a representative of the seasonality of temperatures and nine meteorological predictors known to affect Central European surface weather in winter. Thereby, we limit the area from which the predictors are retrieved from globally to a smaller region covering the North Atlantic Ocean and parts of Eurasia. By doing so, we try to find a balance between the area necessary to include as many physically known drivers relevant on the subseasonal timescale as possible and the increasing amount of computational resources needed to process these. As a result, we find that the most useful input in terms of performance, computational efficiency and interpretability consists of the month and the minimum, mean, maximum as well as the variance of $u10$, $z100$, $z250$, $z500$, $z850$, $t850$, $H850$, $u300$ and msl. This holds for both, the QRFs and RFCs.

Similar to van Straaten et al. (2022), we predict a regional average of temperatures instead of a value at each grid point. This is done to increase subseasonal forecast skill by spatial aggregation, as shown in van Straaten et al. (2020). Following the results of this study, we use the mean over Central Europe as a whole instead of smaller areas. Furthermore, we use a 7-day running mean for filtering out synoptic-scale disturbances which might decrease forecasting skill on the subseasonal timescale. This temporal aggregation is done for the ground truth, the meteorological fields used as input for the RF models, and the climatological benchmark ensemble.

Similar to many studies, e.g. van Straaten et al. (2022), we start our research by answering the following question:

RQ 1.1 Can RF models using only reanalysis data as input provide skillful forecasts in comparison to a climatological benchmark?

- Yes, in the 20-winter mean, QRF models can provide skillful forecasts compared to a climatological benchmark ensemble at lead times of 14, 21 and 28 days. RFC models can provide skillful forecasts on lead times of 14 and 21 days.
- However, for both QRF and RFC models, the variability of forecast skill between winters is high.

The variability between the different model types is visible but not as high as the variability in skill across winters. The models with the optimal input (the month and the statistics of all nine meteorological predictors fields, as stated above) show a positive skill at a lead time of 14 days, and, in the case of the RFCs, also at a lead time of 21 days. Only one QRF model, the one trained on all grid points of all predictor fields, shows a higher CRPSS value at a lead time of 14 days and also positive CRPSS values at lead times of 21 and 28 days. However, the differences in the mean CRPSS value are small and additional challenges arise making the kind of input not optimal. Besides the increased use of computational resources due to the higher amount of training data, models trained on all grid points of predictor fields are not easy to interpret which is on the other hand crucial when trying to build trust in the models' forecasts.

As stated above, the skill of the QRF and RFC models depends on the winter to be predicted. When comparing the winter-mean skill of the winter 2011/2012, which features a severe and long-lasting cold wave in February, with the skill of the winter 2013/2014, which is characterized by the absence of cold waves, we find that in case of the winter 2011/2012 skillful forecasts of the occurrence of cold-wave days are achieved by two RFC models at all analyzed lead times and skillful predictions of the 2-meter temperature by at least one RF model at lead times of 14 and 21 days. In case of the winter 2013/2014, skillful forecasts of both properties are only found at a lead time of 14 days. Since especially the presence, respectively absence, of cold waves determines the difference between these two winters, we answer in a second step the following question:

RQ 1.2 Is the forecast skill equally good for periods with mild temperatures than periods with cold temperatures?

- No, the skill of the individual models depends on the time of winter whereby the models perform better during long-lasting and mid-winter cold waves than at the margins of winters where temperatures are usually milder.

We base this statement not only on the two named winters but on all winters in the time period between 2000 and 2020. As a measure, we use the daily difference in scores of the climatological benchmark ensemble and the RF models' forecasts.

Especially during the mid-winter severe cold waves, the ML models perform peculiarly well compared to the climatological benchmark ensemble. One explanation therefore might be the choice of the meteorological predictors. For example, a weak polar vortex, represented by low $u10$ values, can lead to strong stratospheric anomalies which can influence the geopotential height field at tropopause level, represented by $z100$, and then lead to cold waves over Central Europe. The timing of these stratospheric anomalies is well in the range of the subseasonal timescale.

Although we include predictors like $z500$ and $z850$, which both represent the large-scale tropospheric flow and origin of air masses leading to cold as well as to mild temperatures, the RF models perform worse during periods with mild temperatures, such as during the whole winter 2013/2014, than for periods with colder temperatures, such as in the winter 2011/2012.

Until now, the relevance of predictors is done rather speculatively. This is changed in a next step using a SHAP analysis of the two named winters. Thereby, we answer the following research question:

RQ 1.3 Which predictors are determining the models' predictions?

- The most common predictor is $t850$ for warm periods for the analyzed QRF and RFC model.
- During the mid-winter cold wave in 2011/2012, common predictors of the analyzed QRF and RFC model which contribute to forecasting colder temperatures and cold-wave days are $t850$ and msl.

Interestingly, the most common predictor for both, mild and cold temperatures, is $t850$, which is also one of the most important predictors for hot summertime temperatures found by van Straaten et al. (2022). According to them, this is not expected since the air masses in 850 hPa are usually known for their influence on shorter lead times than subseasonal. Nevertheless, persistence of these can play a role on the subseasonal timescale. The same is true for msl.

In case of the severe cold wave in the winter 2011/2012, the SHAP analysis supports our suggestions that stratospheric and upper tropospheric predictors play a large role in forecasting. It shows that $u10$ and $z100$ are among the most contributing predictors to the QRFs forecast during this period. Furthermore, the month is an important predictor suggesting that the model performance is depending on the time of seasons. At the margins of the winter, it contributes positively to the temperature forecast and in mid-winter, it contributes negatively. Depending on the study, during December and January (Tomczyk et al., 2019) or January and February (Lhotka and Kyselý, 2015) major cold waves over Central Europe occur most often.

Although the physical expectations are not always met, the fact that both model types agree on common predictors underlines the suitability of these models to generate profound skillful forecasts on the subseasonal timescale instead of producing skillful forecasts by coincidence. Furthermore, with the reduced input based on meteorological knowledge, the developed RF models are computationally more efficient than traditional NWP models.

We assume that in case of mild wintertime temperatures, the possible number of drivers is higher and generally more diverse than in the case of severe cold waves. This would lead to less clear patterns in the input data and increased complexity for the RF models to learn from and thus leading to worse forecasts of mild temperatures in comparison to predicting cold wintertime temperatures. Knowing this is helpful in assessing the RF models' forecasts reliability. You may trust a forecast of severe cold winter temperatures more than one of mild temperatures. Especially in winter, we think this is not a major caveat since the extreme cold temperatures have a generally larger impact on the planning on the subseasonal timescale than mild temperatures.

7. Random Forest Models as a Complement to Numerical Weather Prediction Models

In chapter 6, the use of RF models as an alternative to NWP models is discussed. Now, we analyze the application of RF models in a postprocessing sense as a complement to NWP models. As discussed in section 3.3, postprocessing methods can be a useful tool to improve the skill of subseasonal NWPs. From the large variety of possible approaches, we select two for our study. These are a lead-time-dependent mean bias correction and RF-based postprocessing models. The RF models use thereby the same setup as the RF models described in chapter 6, which is explained in detail in section 5.10. The different models vary only in terms of their input, using either predictors based on reforecasted fields at the target date and/or reanalyzed fields at initialization. Since the RF models are used in a postprocessing sense, all models take the reforecasted 2-meter temperature as input.

7.1. Comparisons of Predictions of Random Forest Models to ECMWF's S2S Reforecast Ensemble

To obtain a broad overview over the performance of both, the S2S reforecasts ensemble of ECMWF and our RF-based models, we analyze the 20-winter mean skill scores with respect to the climatological benchmark ensemble. This approach is similar to the one used in section 6.2.

7.1.1. Skill of 2-Meter Temperature Predictions in the 20-Winter Mean

At a lead time of 14 days, all selected models except the one with the first ten PCs of the reanalysis fields as input (#13) show a positive skill (Fig. 7.1 (a)). It is remarkable that the QRF-based models used to postprocess the S2S reforecasts (#7-10) are the only ones showing solely positive CRPSS values in their spread. While ECMWF's S2S reforecasts (#1) outperform all ML models using only ERA5 reanalysis data as input (#11-13), they show a lower skill for the average 20-winter mean than the QRF-based postprocessing models (#3-10) (Fig. 7.1 (a)). In comparison with five of the eight QRF-based postprocessing models, this is also true for the mean bias corrected ECMWF's S2S reforecasts (#2).

At a lead time of 21 days, the mean bias corrected ECMWF's S2S reforecasts (#2) show a lower average 20-winter mean CRPSS than all QRF-based postprocessing models (#3-10) but in contrast to the original reforecasts ensemble (#1), still a positive value (Fig. 7.1 (b)). At a lead time of 28 days, also the mean bias corrected ECMWF's S2S reforecasts (#2) only have a negative CRPSS value in the 20-winter mean

(Fig. 7.1 (c)). With the exception of the QRF models using all grid points of ERA5 reanalysis fields as input (#11), this is also valid for the ML models using only reanalysis data as input (#12+13). On the contrary, all QRF-based postprocessing models (#3-10) show skill in the 20-winter mean compared to the climatological benchmark ensemble. At all three considered lead times, the best performing model is the QRF model using the ensemble information of the statistics of the S2S reforecasts fields as input (#5).

7.1.2. Skill of Predictions of the Occurrence of Cold-Wave Days in the 20-Winter Mean

The original ECWMF's S2S reforecasts show only skill at a lead time of 14 days in the 20-winter mean (#1 on Fig. 7.2 (a)). The spread between winters is large, reaching from negative to positive winter-mean BSS values for all three lead times (Fig. 7.2). However, the mean bias corrected ECMWF's S2S reforecasts (#2) have positive 20-winter mean BSS values at all considered lead times with only positive values in the spread at a lead time of 14 days (Fig. 7.2).

With the exception of four RFC-based postprocessing models (#3, 5, 7+9), the latter is also true for all other analyzed models. All three RFC models using only reanalysis data as input (#11-13) are able to produce skillful forecasts at a lead time of 14 days but are outperformed by all other models including both ECMWF's S2S reforecasts (#1+2) (Fig. 7.2 (a)). The best performing model at this lead time is the mean bias corrected ECMWF's S2S reforecast ensemble (#2 on Fig. 7.2 (a)).

At a lead time of 21 days, only two RFC models with solely reanalysis data as input (#12+13) show positive BSS values and produce better forecasts than ECMWF's S2S reforecasts (#1) (Fig. 7.2 (b)). Nevertheless, all RFC-based postprocessing models (#3-10) and the mean bias corrected S2S reforecasts (#2) have a higher skill than these two. The highest BSS value shows the latter (#2) and the RFC model using the ensemble mean of the first ten PCs of the reforecasts and the first ten PCs of the reanalysis fields (#10) as input.

At a lead time of 28 days, a similar RFC model set-up, with the difference that instead of the ensemble mean, the first ten PCs of the reforecasts and reanalysis fields are used directly, has the best skill (#8 on Fig. 7.2 (c)). Besides the RFC-based postprocessing models (#3-10) and the mean bias corrected ECMWF's S2S reforecasts ensemble (#2), all other models only show negative BSS values in the 20-winter mean.

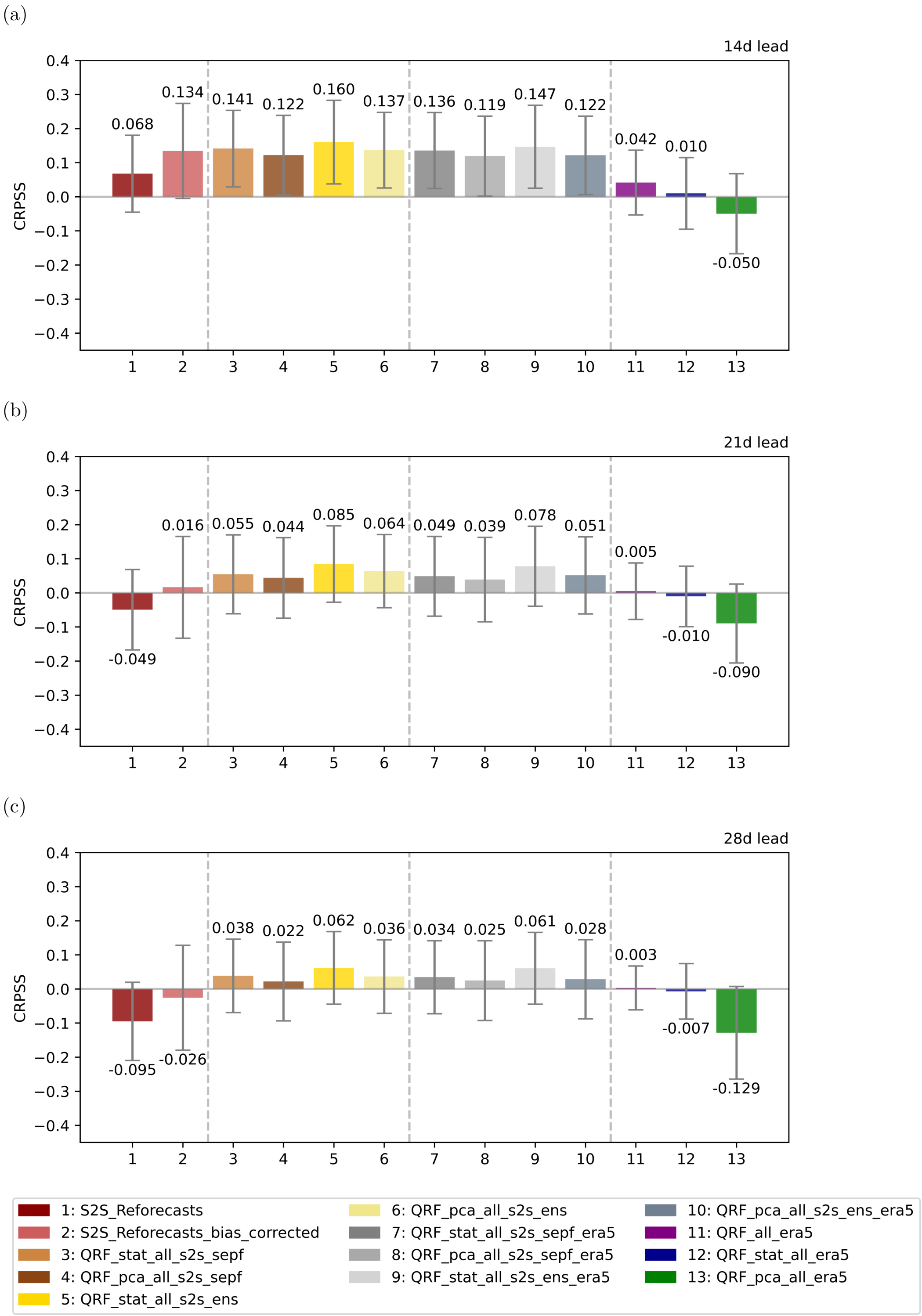


Figure 7.1.: 20-winter mean CRPSS values. The CRPSS values are computed with respect to the climatological benchmark ensemble for a lead time of 14 days (a), 21 days (b) and 28 days (c). The whiskers show the standard deviation of the CRPSS between winters, the printed values the 20-winter mean CRPSS.

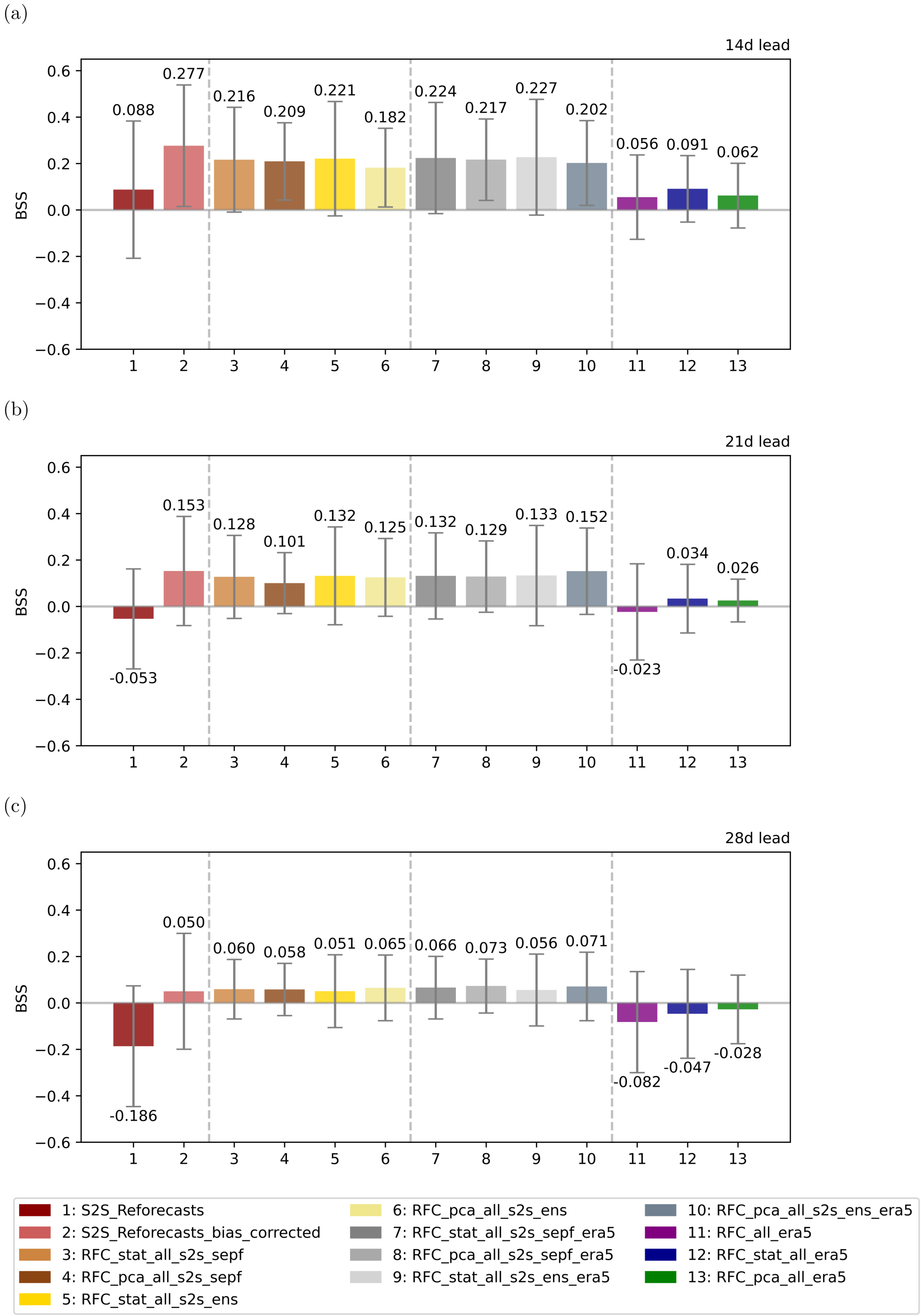


Figure 7.2.: 20-winter mean BSS values. The BSS values are computed with respect to the climatological benchmark ensemble for a lead time of 14 days (a), 21 days (b) and 28 days (c). The whiskers show the standard deviation of the BSS between winters, the printed values the 20-winter mean BSS.

7.2. Predictor Importance of the Random Forest Models' Forecasts Using Impurity Based Feature Importance

In order to determine which predictors are most relevant for the models' predictions, we apply an Impurity Based Feature Importance exemplary for the two RF-based postprocessing models using the ensemble information of the statistics of the reforecasted fields and the statistics of the reanalyzed fields as input (Fig. 7.3, #9 on Fig. 7.1 and 7.2). These models are chosen since their predictors are the easiest to interpret. Furthermore, we concentrate on the impurity feature importance values in the 20-winter mean as they show the overall importance of the predictors and average out possible peculiarities of single winters.

The analysis reveals that the reforecasted mean 2-meter temperature is the most important predictor for both the analyzed QRF and RFC model at all lead times (Fig. 7.3). In case of the QRF model, other predictors derived from the reforecasted fields play a more important role than the predictors from the reanalysis data at all lead times (Fig. 7.3 (a), (c) and (e)). This is quite the opposite in case of the RFC models. Apart from the reforecasted 2-meter temperature predictors, the RFC models rely mostly on the reanalyzed predictors at lead times of 21 and 28 days (Fig. 7.3 (d) and (f)). At a lead time of 14 days, the reforecasted and reanalyzed predictors, besides the reforecasted 2-meter temperature, are approximately equally important (Fig. 7.3 (b)).

Independent of the type of predictors, the variables that represent the large-scale flow, e.g. $z500$, provide the second most important predictors of the ten most important ones for both model types at all lead times (Fig. 7.3).

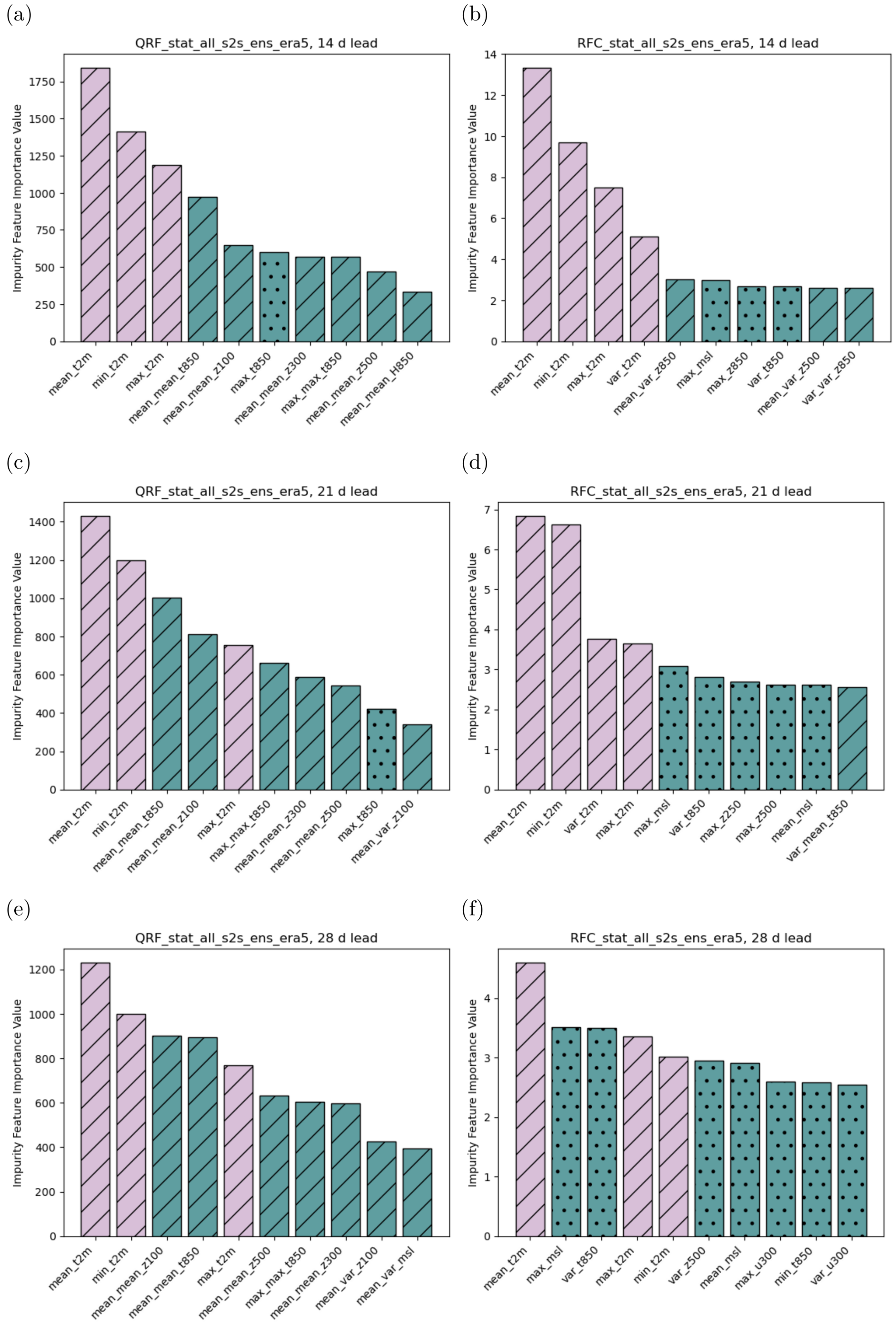


Figure 7.3.: Impurity based feature importance. The ten most important predictors according to the impurity based feature importance, averaged over all 20 analyzed winters, are shown for a representative QRF (left) and RFC model (right) for lead times of 14 ((a) and (b)), 21 ((c) and (d)) and 28 days ((e) and (f)). Dotting indicates reanalysis predictors, hatching reforecasts predictors, the light purple color marks predictors containing the 2-meter temperature and the dark teal color marks all other predictors.

7.3. Summary and Discussion

In this part of our research we show that postprocessing methods are a useful tool to improve the forecasting skill of ECWMF's S2S reforecasts. Analogously to the analysis shown in chapter 6, we focus on the prediction of 2-meter temperatures and the occurrence of cold-wave days in Central Europe at lead times of 14 , 21 and 28 days.

With the exception of the forecast of 2-meter temperatures at a lead time of 28 days, the predictions of the postprocessed ECMWF's S2S reforecasts yield a better skill in the 20-winter mean than the predictions of the RF models based only on reanalysis data. This is true for both the continuous 2-meter temperature forecasts and the binary predictions of the occurrence of cold-wave days. Furthermore, all postprocessing approaches lead to an improvement in the respective skill score compared to the original ECMWF's S2S reforecast ensemble. However, the performance of the individual approaches differ and not all of them provide skillful forecasts for both properties at all lead times. Therefore, we answer in a first step the following question:

RQ 2.1 Which postprocessing approaches perform particularly well in comparison to the original ECMWF's S2S reforecasts?

- In case of the 2-meter temperature predictions, the best performing postprocessing model is the QRF using the ensemble information of the statistics of the reforecasted meteorological fields and the month as input in addition to the statistics of the reforecasted 2-meter temperature.
- In case of the predictions of the binary occurrence of cold-wave days, the lead-time-dependent mean bias corrected ECMWF's S2S reforecast yield the best skill at lead times of 14 and 21 days. At a lead time of 28 days, the best performing RFC model is the one using the ensemble information of the first ten principle components of the reforcasted meteorological fields, the month, the ten principle components of the reanalyzed meteorological fields and the statistics of the reforecasted 2-meter temperature as input.
- The variation in mean skill scores between the different postprocessing models is small at all lead times for both forecasted properties.

As expected, the vast majority of the analyzed models shows in the 20-winter mean a decrease in skill with increasing lead time. We find that all QRF-based postprocessing models are able to increase forecasting skill at lead times of 21 and 28 days in case of the continuous 2-meter temperature forecasts in comparison to the lead-time-dependent mean bias corrected ECMWF's S2S reforecasts. At a lead time of 14 days, the QRF-based postprocessing model have a comparable skill. RFC-based postprocessing models show a similar skill as the lead-time-dependent mean bias corrected ECMWF's S2S reforecasts for forecasts of the occurrence of cold-wave days on lead times of 21 and 28 days and a slightly lower skill for forecasts with a lead time of 14 days.

Regarding the use of every member of the reforecast ensemble as a separate predictor in comparison to using only the condensed ensemble information, we cannot see a substantial difference. This is in accordance with Horat and Lerch (2024) who state based on a literature research that the information of the variability of the ensemble is little compared to the information inherent in the ensemble mean.

Although all RF-based postprocessing models are trained on a smaller number of training samples (bi-weekly data from the winters 2000/2001 to 2019/2020) than RF models using the same kind of input but only from reanalysis fields (daily data, winters 1950/1951 to 2019/2020), they create substantially better forecasts of both properties at all lead times. Therefore, the amount of training data seems not to be crucial for a good model performance. Due to the high annual variability between winters concerning temperatures and cold-wave days, as shown in chapter 6, we doubt that the restriction to the last 20 years of the training data, e.g. due to possible trends, is the major contributor to a better forecasting skill. Instead, we argue that the reforecasted meteorological fields, especially of the 2-meter temperature, might be of major importance due to the possibly enhanced predictive skill. These fields contain physically consistent information about the modeled future, which is missing in the reanalysis data at initialization. To validate these assumption, we answer the following question using an Impurity Based Feature Importance approach:

RQ 2.2 In case of the RF-based postprocessing models, are predictors based on the reforecasted meteorological fields at the target date more important for the models' predictions than the reanalyzed fields at the initialization?

- For forecasting both 2-meter temperatures and the occurrence of cold-wave days, the most important predictor is the reforecasted 2-meter temperature.
- In case of the predictions of the 2-meter temperatures, predictors based on the reforecasted fields at the target date are more important at all lead times than predictors based on the reanalyzed fields at initialization time.
- In case of the predictions of the occurrence of cold-wave days at a lead time of 14 days, besides the reforecasted 2-meter temperature, predictors based on the reanalyzed fields at initialization are equally important than predictors based on the reforecasted fields at the target date. At lead times of 21 and 28 days, these predictors are even more important.

The Impurity Based Feature Importance analysis is done exemplary for the two RF-based postprocessing models using the ensemble information of the reforecasted fields and the statistics of the reanalysis fields as input. The high relevance of predictors based on reanalysis fields in case of the forecasts of the occurrence of cold-wave days can be explained under the assumption that ECMWF's subseasonal model is tuned to produce better forecasts of mean temperatures than extremes. Thus, the reforcasted fields are more useful in predicting "averaged" temperatures, which are present most of the time during a winter, than extremes. But it is the correct prediction of extremes which is important when it comes to cold-wave days since they are characterized by lower than usual temperatures.

Independent of the type of predictors, the variables that represent the large-scale flow, e.g. the geopotential in 500 hPa height, provide the second most important predictors of the ten most important ones for both analyzed model types at all lead times.

8. Weather Regimes for Assessing Forecast Reliability

As shown in chapter 7, the variables that represent the large-scale flow, e.g. $z500$, provide the second most important predictors besides the reforecasted 2-meter temperature for the predictions of the RF-based postprocessing models. Following this observation and previous studies (e.g. Grams et al., 2020), we analyze if we can further use the information of the large-scale flow present at model initialization for a conditional improvement of forecasts and thus assessing forecast reliability. As shown in section 2.2, different studies (e.g Osman et al., 2023) find clear differences in the skill of subseasonal forecasts of WRs. Here, we investigate if we also find differences in subseasonal forecasts of 2-meter temperatures and the occurrence of cold-wave days in Central Europe depending on the WR at initialization and in how far these can be linked to the WR successions during the forecast. Thereby we use, similar as Büeler et al. (2021) and Osman et al. (2023), the seven WRs proposed by Grams et al. (2017). These are described in detail in section 2.2. Furthermore, we investigate independent of the WR present at initialization possible differences in WR successions before the best and worst predicted days within cold waves.

8.1. Comparisons of Forecasts Initialized During Different Weather Regimes

As a first step, we cluster the predictions of the 2-meter temperatures and the occurrence of cold-wave days by the WR present at initialization. We select ECMWF's mean bias corrected S2S reforecasts ensemble (#2) and the best performing RF model from every category in the 20-winter mean across lead times to investigate the forecast skill dependent on the large-scale flow present at initialization. To assess the significance of possible differences in skill, the p-values of Welch's t-test, described in section 5.13, are calculated based on the CRPS, respectively BS, distribution of the subsamples of forecasts initialized during the respective WRs.

8.1.1. Flow Dependence of the Skill of 2-Meter Temperature Forecasts

The chosen QRF models for investigating a possible skill dependence on the WR at initialization are the QRF using the ensemble information of the statistics of the reforecasted fields (#5), the QRF using the ensemble information of the statistics of the reforecasted fields and the statistics of the reanalysis fields (#9) as well as the QRF using all grid points of the reanalysis fields as input (#11). With the exception of the latter category, which shows high fluctuations in skill between models, the chosen QRF models are seen as representative for their category. As already observed in the 20-winter mean, the models in the two former categories produce similarly skillful forecasts among themselves across regimes and lead times (#5+9 on Fig. 7.1 and #5+9 on Fig. 8.1). These QRF-based postprocessing models show, independent from the WR present at initialization, higher CRPSS values than the original ECMWF's S2S reforecasts (#1) at all analyzed lead times (#5+9 on Fig. 8.1). With the exception of forecast with a lead time of 14 days initialized during the AT and ScTr regime, this is also true in case of the mean bias corrected ECMWF's S2S reforecasts (#2). The forecasts of the QRF-based postprocessing models are thereby almost equally skillful and have CRPSS values above zero for all regimes and lead times.

The QRF model trained solely on reanalysis data (#11) shows at a lead time of 14 days positive CRPSS values only when initialized during the AT, ScTr, ZO, EuBL and "No" regime. At this lead time, the mean bias corrected ECMWF's S2S reforecasts ensemble (#2) produces better forecasts during all WRs at initialization than this QRF model using only reanalysis data as input (#11). The gap between the mean bias corrected ECMWF's S2S reforecasts ensemble and the best performing QRF-based postprocessing models is smallest when started during the EuBL and GL regime and largest during the ZO regime at initialization (#2, 5 and 9 on Fig. 8.1 (a)). Forecasting skill is generally better for all models when initialized during the AT, EuBL and "No" regime and worse during ZO, GL and ScBL regimes, in case of the QRF model trained solely on reanalysis data (#11), also for the AR regime. The differences in skill between the named regimes are thereby not significant (Fig. 8.2 (a) and (b)).

At a lead time of 21 days, the QRF model using only reanalysis data as input (#11) has positive CRPSS values when the AT, ZO, ScBL or "No" regime is present at the forecast start. In case of the mean bias corrected ECMWF's S2S reforecasts ensemble, positive CRPSS values are found when initialized during the ScTr, ZO, ScBL regime or the "No" regime (#2 on Fig. 8.1 (b)). These are also the regimes at the start of the forecasts during which the QRF-based postprocessing models (#5+9) show the best skill. The least skillful forecasts of these models are produced when the GL regime is present at initialization, whereby only the difference in CRPS values to the forecasts initialized during the ScBL regime are significant (Fig. 8.2 (d)). In case of the mean bias corrected ECMWF's S2S reforecasts, the least skillful forecasts are created when the EuBL regime is present at initialization. Thereby, the differences between forecasts initialized during the EuBL regime and forecasts initialized during the ScBL or AR regime are significant (Fig. 8.2 (c)). The largest improvement in forecasting skill is found for the QRF-based postprocessing models (#5+9) in comparison with the mean bias corrected ECMWF's S2S reforecast data (#2) when initialized during the EuBL regime, the smallest when initialized during the ScTr regime.

At a lead time of 28 days, the best forecasts in case of the QRF-based postprocessing models are produced when started during the EuBL regime, followed by the ZO regime (#5+9 on Fig. 8.1 (c)). The least skill is found for forecasts initialized during the AT, AR and GL regime, whereby the differences to the most skillful forecasts are not significant (Fig. 8.2 (f)). The QRF-based postprocessing models

(#5+9) show positive CRPSS values during all regimes at initialization while the mean bias corrected ECMWF's S2S reforecasts ensemble (#2) shows only positive CRPSS values during the ZO, EuBL and ScBL regime at the forecast start. The QRF model trained solely on reanalysis data (#11) shows positive CRPSS values when the ZO, EuBL, ScBL or "No" regime is present at the start of the forecast. The difference in CRPSS between the mean bias corrected ECMWF's S2S reforecast ensemble (#2) and the QRF-based postprocessing models (#5+9) is largest when the AT regime is present at the forecast start and smallest when the ScBL regime is present.

With the exception of the QRF model using the statistics of the reanalysis fields as input (#12), which shows slightly higher (but still negative) mean CRPSS values at a lead time of 28 days than of 21 days, the forecast skill generally decreases with increasing lead time for all models in the 20-winter mean (Fig. 7.1) (b) and (c)).

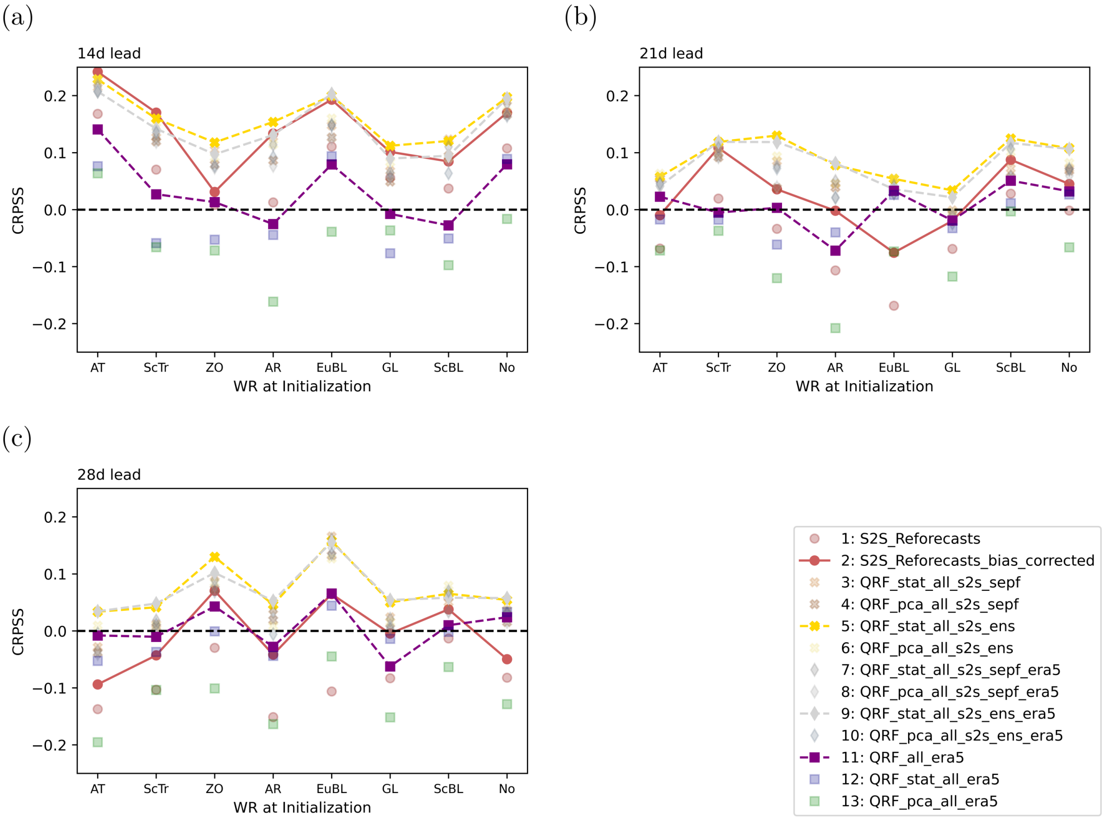


Figure 8.1.: Flow-dependent skill. The skill is shown in terms of CRPSS values for lead times of 14 (a), 21 (b) and 28 days (c).

8.1.2. Lead-Time Dependence of the Skill of 2-Meter Temperature Forecasts

In addition to the skill dependence on the WR present at initialization, we now analyze the skill evolution across lead times for the different WRs. We find that forecasting skill decreases with increasing lead time when forecasts are initialized during the AT, ScTr and "No" regime (Fig. 8.3 (b), (c) and (h)). When

(a)

14d lead	AT	ScTr	ZO	AR	EuBL	GL	ScBL	No
AT	1	0.43	0.06	0.85	0.54	0.2	0.37	0.17
ScTr	0.43	1	**0.01**	0.52	0.17	**0.03**	0.08	**0.01**
ZO	0.06	**0.01**	1	**0.04**	0.2	0.41	0.25	0.34
AR	0.85	0.52	**0.04**	1	0.42	0.13	0.25	0.09
EuBL	0.54	0.17	0.2	0.42	1	0.55	0.81	0.56
GL	0.2	**0.03**	0.41	0.13	0.55	1	0.7	0.93
ScBL	0.37	0.08	0.25	0.25	0.81	0.7	1	0.72
No	0.17	**0.01**	0.34	0.09	0.56	0.93	0.72	1

(b)

14d lead	AT	ScTr	ZO	AR	EuBL	GL	ScBL	No
AT	1	0.28	0.28	0.45	0.67	0.25	0.71	0.35
ScTr	0.28	1	**0.05**	0.77	0.17	**0.03**	0.15	**0.03**
ZO	0.28	**0.05**	1	0.09	0.54	0.95	0.46	0.7
AR	0.45	0.77	0.09	1	0.27	0.06	0.27	0.07
EuBL	0.67	0.17	0.54	0.27	1	0.53	0.93	0.73
GL	0.25	**0.03**	0.95	0.06	0.53	1	0.44	0.71
ScBL	0.71	0.15	0.46	0.27	0.93	0.44	1	0.62
No	0.35	**0.03**	0.7	0.07	0.73	0.71	0.62	1

(c)

21d lead	AT	ScTr	ZO	AR	EuBL	GL	ScBL	No
AT	1	0.64	0.81	0.08	0.21	0.49	0.06	0.72
ScTr	0.64	1	0.81	0.21	0.1	0.24	0.17	0.36
ZO	0.81	0.81	1	0.12	0.14	0.35	0.1	0.52
AR	0.08	0.21	0.12	1	**0.01**	**0.01**	0.86	**0.01**
EuBL	0.21	0.1	0.14	**0.01**	1	0.5	**0**	0.28
GL	0.49	0.24	0.35	**0.01**	0.5	1	**0.01**	0.66
ScBL	0.06	0.17	0.1	0.86	**0**	**0.01**	1	**0.01**
No	0.72	0.36	0.52	**0.01**	0.28	0.66	**0.01**	1

(d)

21d lead	AT	ScTr	ZO	AR	EuBL	GL	ScBL	No
AT	1	0.89	0.5	**0.03**	0.44	0.36	0.1	0.68
ScTr	0.89	1	0.43	**0.03**	0.53	0.46	0.09	0.81
ZO	0.5	0.43	1	0.14	0.16	0.11	0.29	0.22
AR	**0.03**	**0.03**	0.14	1	**0.01**	**0**	0.75	**0**
EuBL	0.44	0.53	0.16	**0.01**	1	0.94	**0.02**	0.62
GL	0.36	0.46	0.11	**0**	0.94	1	**0.01**	0.52
ScBL	0.1	0.09	0.29	0.75	**0.02**	**0.01**	1	**0.02**
No	0.68	0.81	0.22	**0**	0.62	0.52	**0.02**	1

(e)

28d lead	AT	ScTr	ZO	AR	EuBL	GL	ScBL	No
AT	1	0.61	0.07	0.32	0.66	0.86	0.4	0.48
ScTr	0.61	1	**0.04**	0.17	0.39	0.5	0.22	0.25
ZO	0.07	**0.04**	1	0.41	0.21	0.07	0.38	0.16
AR	0.32	0.17	0.41	1	0.63	0.38	0.92	0.67
EuBL	0.66	0.39	0.21	0.63	1	0.77	0.71	0.87
GL	0.86	0.5	0.07	0.38	0.77	1	0.47	0.57
ScBL	0.4	0.22	0.38	0.92	0.71	0.47	1	0.77
No	0.48	0.25	0.16	0.67	0.87	0.57	0.77	1

(f)

28d lead	AT	ScTr	ZO	AR	EuBL	GL	ScBL	No
AT	1	0.35	0.18	0.52	0.8	0.52	0.88	0.62
ScTr	0.35	1	**0.04**	0.14	0.27	0.68	0.42	0.15
ZO	0.18	**0.04**	1	0.5	0.34	**0.04**	0.14	0.29
AR	0.52	0.14	0.5	1	0.73	0.19	0.43	0.81
EuBL	0.8	0.27	0.34	0.73	1	0.39	0.7	0.87
GL	0.52	0.68	**0.04**	0.19	0.39	1	0.63	0.21
ScBL	0.88	0.42	0.14	0.43	0.7	0.63	1	0.51
No	0.62	0.15	0.29	0.81	0.87	0.21	0.51	1

Figure 8.2.: P-values of Welch's t-test for the continuous temperature forecasts. The p-values are shown for the forecasts of the mean bias corrected ECMWF's S2S reforecast ensemble for lead times of 14 days (a), 21 days (c) and 28 days (e) as well as for the forecasts of the QRF-based postprocessing model using the ensemble information of the statistics of the reforecasted fields as input for lead times of 14 days (b), 21 days (d) and 28 days (f).

forecasts are initialized during the EuBL regime, the forecast skill is higher at a lead time of 28 days than 21 days for all models but the difference in CRPS values at these two lead times is not significant

(Fig. 8.3 (g) and Tab. 8.1). For some ML models, this is also true when initialized during the AR and GL regimes (#11 on Fig. 8.3 (a) and #5+9 on Fig. 8.3 (e)). The same is seen for the ZO regime but here the forecast skill is rather steady across lead times (Fig. 8.3 (d)). Interestingly, forecasts initialized during the ScBL regime show, with the exception of ECWMF's S2S reforecasts (#1), the highest skill at a lead time of 21 days (Fig. 8.3 (f)). In case of the QRF model using the statistics of the S2S reforecasts fields as input (#5), the difference in CRPS values between forecasts at this lead time and a lead time of 28 days is significant, but not the difference to forecasts with a lead time of 14 days (Tab. 8.1).

Table 8.1.: P-values of Welch's t-test calculated for differences in CRPS values between lead times of forecasts. The calculations are done for every WR separately for ECMWF's mean bias corrected S2S reforecasts (abbreviated as "S2S") and the QRF_stat_all_s2s_ens model (#5, abbreviated as "QRF"). Corresponding figure is Fig. 8.3 which shows the CRPSS values of the forecasts sorted after the WR at initialization.

	14 d vs. 21 d		14 d vs.28 d		21 d vs. 28 d	
Regime at Initialization	S2S	QRF	S2S	QRF	S2S	QRF
AR	0.37	0.83	**0.02**	**0.05**	0.10	0.08
AT	**0.03**	0.11	**0**	0.06	0.39	0.78
ScTr	**0.01**	**0.01**	**0.00**	**0.00**	0.10	0.30
ZO	0.89	0.79	0.44	0.55	0.47	0.71
GL	0.08	0.17	0.06	0.16	0.98	1.00
ScBL	0.63	0.60	0.20	0.11	0.08	**0.04**
EuBL	**0.01**	0.09	0.08	0.34	0.39	0.48
No regime	**0.05**	0.12	0.06	0.40	0.92	0.48

8.1.3. Flow Dependence of the Skill of Forecasts of the Occurrence of Cold-Wave Days

Analogously as done for the 2-meter temperature forecasts, the best performing RFC model of each category in the 20-winter mean across all lead times is used as a representative model to investigate a possible skill dependence of forecasts on the WR present at initialization. The selected models are the RFC using the ensemble information of the statistics of the reforecasted fields (#5), the RFC using the ensemble information of the first ten PCs of the reforecasted fields and the first ten PCs of the reanalysis fields (#10) as input as well as the RFC using the statistics of the reanalysis fields as input (#12).

In contrast to the continuous forecasts, RFC-based postprocessing models (#5+10) do generally not produce the most skillful binary predictions of the occurrence of cold-wave days compared to the other analyzed models (Fig. 8.4). Additionally, with the exception of the lead time of 28 days, the gaps between BSS values of the different models are smaller. At a lead time of 14 days, the RFC-based postprocessing models show the best BSS values only when the ZO, GL or "No" regime are present at the forecast start (#5+10 on Fig. 8.4 (a)). The smallest BSS values are found during the ScTr and ScBL regime at initialization, whereby only the differences in BS values between forecasts initialized under the GL and the ScTr regime are significant (Fig. 8.5 (b)). When the AT and ScTr regime are present at the start of

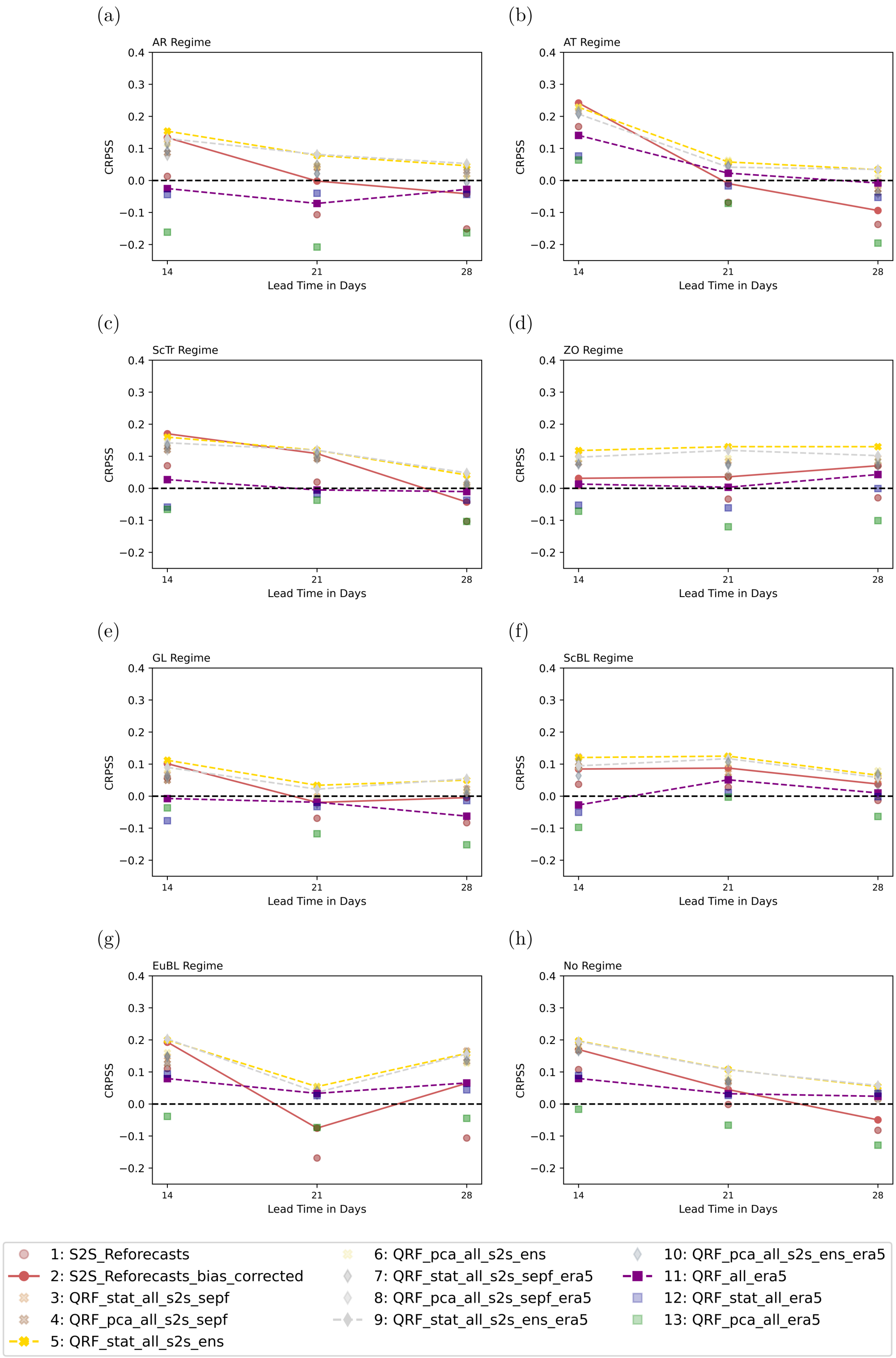


Figure 8.3.: Flow-dependent skill evolution across lead times. The skill is shown as CRPSS values for lead times of 14, 21 and 28 days for the AT (a), AR (b), ScTr (c), ZO (d), GL (e), ScBL (f), EuBL (g) and "No" regime (h).

the forecasts, the skill of these models is comparable to the skill of the RFC model using the statistics of the reanalysis fields as input (#12). The mean bias corrected ECMWF's S2S reforecasts (#2) show the highest BSS values of all analyzed models when initialized during the GL regime. The least skillful forecasts are produced by this model when initialized during the ScTr regime. The differences in BS values in comparison to the forecasts initialized during the GL regime are thereby significant (Fig. 8.5 (a)). While the mean bias corrected ECMWF's S2S reforecasts (#2) show a relatively high variation in skill between the different WRs present at initialization, the RFC-based postprocessing models (#5+10) have a more similar skill. With the exception of the ScBL regime, this is also true for the RFC using the statistics of the reanalysis fields as input (#12).

At a lead time of 21 days, the mean bias corrected ECMWF's S2S reforecasts show the highest BSS values when initialized during the ScTr and ScBL regime (#2 on Fig. 8.4 (b)). These are also the regimes at initialization, during which the RFC-based postprocessing models (#5+10) produce the best forecasts. The worst forecasts are produced by all models, excluding the mean bias corrected ECMWF's S2S reforecasts, during the AR regime at initialization. The mean bias corrected ECMWF's S2S reforecasts (#2) produce the least skillful forecasts when initialized during the ZO regime. The differences in BS values of the most and least skillful forecasts are not significant for both model types (Fig. 8.5 (c) and (d)). Besides when the AR regime is present at the forecast start, a few RFC-based postprocessing models (#5+10) also show negative BSS values during AT, ZO and EuBL regime at initialization. Nevertheless, the RFC-based postprocessing models are the model types showing the highest BSS values during all WRs at the forecast start except the GL regime, where two RFC models using reanalysis fields as input (#12) show higher BSS values. The latter regime, together with the ScTr and ScBL regime, are the WRs at initialization during which the RFC models using only reanalysis fields as input (#12) perform best.

At a lead time of 28 days, at least one of the RFC models using solely reanalysis data as input (#12) has higher or comparable BSS values than the mean bias corrected ECMWF's S2S reforecasts ensemble (#2) during each WR at initialization except the AT regime (Fig. 8.4 (c)). Generally, the difference in BSS values between the analyzed ML models (#5, 10+12) and the mean bias corrected ECMWF's S2S reforecasts (#2) is smallest during the AR, ScTr, EuBL, GL and ScBL regime at initialization and largest during the ZO regime. RFC-based postprocessing models (#5+10) and the RFC model using only reanalysis data as input (#12) perform comparably across all WRs at initialization. The best forecasts are thereby produced under the EuBL regime at the forecast start, the worst during the AT regime. The difference in BS values of forecasts initialized during these two regimes is thereby significant (Fig. 8.5 (f)). The mean bias corrected ECMWF's S2S reforecasts ensemble (#2) performs best during the EuBL and ScBL regime at the forecast start and worst when initialized under the ZO regime. In this case, the differences in BS values of the forecasts initialized during the different regimes is not significant (Fig. 8.5

(e)).

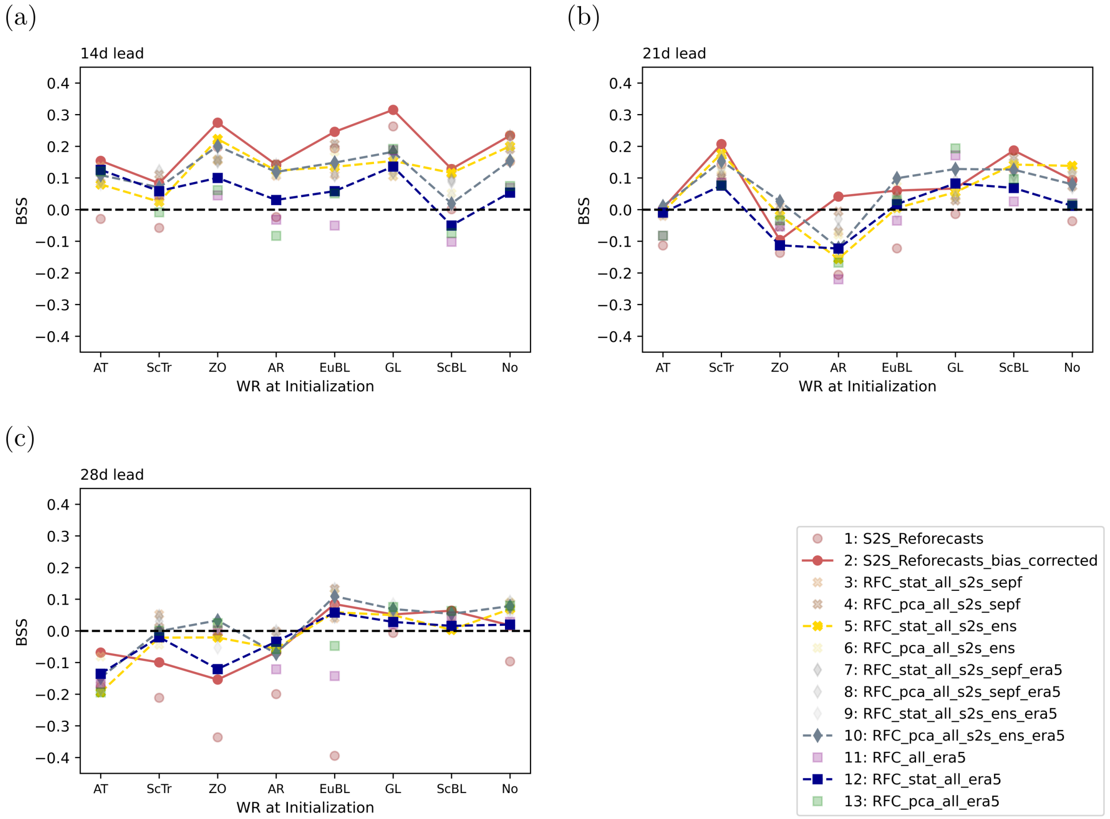


Figure 8.4.: Flow-dependent skill. The skill is shown in terms of BSS values for lead times of 14 (a), 21 (b) and 28 days (c).

8.1.4. Lead-Time Dependence of the Skill of Forecasts of the Occurrence of Cold-Wave Days

Analogously to the forecast of continuous 2-meter temperatures, also the forecast skill of cold-wave days shows a dependence on the lead time and large-scale flow at initialization. With a few exceptions, the same pattern is seen as it is for the continuous temperature forecasts. Forecast skill is also decreasing under the AT and "No" regime at initialization but not under the ScTr regime (Fig. 8.6 (b), (c) and (h)). In contrast to the continuous forecasts, skill is also decreasing for all models during the GL regime at initialization (Fig. 8.6 (e)). During the ScTr and the ScBL regime at the start of the forecasts, the highest skill is found for all models at a lead time of 21 days (Fig. 8.6 (c) and (f)). For the latter, this is also seen in case of the continuous temperature forecasts. During the rest of the WRs at initialization, the skill evolution depends on the model type. Again, similar as for the continuous temperatures, these are the AR and the ZO regime, whereby during the latter at initialization, skill is, with the exception of the mean bias corrected ECMWF's S2S reforecasts (#2), rather steady at the two longer lead times (Fig. 8.6 (a) and (d)). When the EuBL regime is present at initialization, all models but ECMWF's S2S reforecasts

(a)

14d lead	AT	ScTr	ZO	AR	EuBL	GL	ScBL	No
AT	1	0.65	0.73	0.51	0.47	**0.02**	0.44	0.93
ScTr	0.65	1	0.42	0.24	0.23	**0**	0.21	0.62
ZO	0.73	0.42	1	0.78	0.71	0.07	0.69	0.63
AR	0.51	0.24	0.78	1	0.91	0.1	0.88	0.37
EuBL	0.47	0.23	0.71	0.91	1	0.16	0.98	0.36
GL	**0.02**	**0**	0.07	0.1	0.16	1	0.16	**0.01**
ScBL	0.44	0.21	0.69	0.88	0.98	0.16	1	0.32
No	0.93	0.62	0.63	0.37	0.36	**0.01**	0.32	1

(b)

14d lead	AT	ScTr	ZO	AR	EuBL	GL	ScBL	No
AT	1	0.55	0.6	0.54	0.33	**0**	0.29	0.93
ScTr	0.55	1	0.26	0.2	0.13	**0**	0.1	0.4
ZO	0.6	0.26	1	0.96	0.63	**0.01**	0.6	0.59
AR	0.54	0.2	0.96	1	0.65	**0.01**	0.61	0.51
EuBL	0.33	0.13	0.63	0.65	1	0.07	0.99	0.31
GL	**0**	**0**	**0.01**	**0.01**	0.07	1	0.06	**0**
ScBL	0.29	0.1	0.6	0.61	0.99	0.06	1	0.25
No	0.93	0.4	0.59	0.51	0.31	**0**	0.25	1

(c)

21d lead	AT	ScTr	ZO	AR	EuBL	GL	ScBL	No
AT	1	0.99	0.63	0.69	0.67	**0.01**	0.97	0.97
ScTr	0.99	1	0.61	0.69	0.66	**0**	0.98	0.98
ZO	0.63	0.61	1	0.35	0.98	**0.02**	0.6	0.54
AR	0.69	0.69	0.35	1	0.43	**0**	0.71	0.64
EuBL	0.67	0.66	0.98	0.43	1	**0.04**	0.65	0.61
GL	**0.01**	**0**	**0.02**	**0**	**0.04**	1	**0**	**0**
ScBL	0.97	0.98	0.6	0.71	0.65	**0**	1	1
No	0.97	0.98	0.54	0.64	0.61	**0**	1	1

(d)

21d lead	AT	ScTr	ZO	AR	EuBL	GL	ScBL	No
AT	1	0.82	0.92	0.86	0.75	**0.01**	0.82	0.96
ScTr	0.82	1	0.89	0.94	0.92	**0.02**	1	0.82
ZO	0.92	0.89	1	0.94	0.82	**0.01**	0.89	0.94
AR	0.86	0.94	0.94	1	0.86	**0.01**	0.94	0.87
EuBL	0.75	0.92	0.82	0.86	1	**0.04**	0.92	0.75
GL	**0.01**	**0.02**	**0.01**	**0.01**	**0.04**	1	**0.02**	**0**
ScBL	0.82	1	0.89	0.94	0.92	**0.02**	1	0.82
No	0.96	0.82	0.94	0.87	0.75	**0**	0.82	1

(e)

28d lead	AT	ScTr	ZO	AR	EuBL	GL	ScBL	No
AT	1	0.61	0.5	0.51	0.09	0.1	0.48	0.93
ScTr	0.61	1	0.24	0.9	**0.03**	0.28	0.82	0.5
ZO	0.5	0.24	1	0.18	0.32	**0.02**	0.18	0.47
AR	0.51	0.9	0.18	1	**0.02**	0.33	0.91	0.39
EuBL	0.09	**0.03**	0.32	**0.02**	1	**0**	**0.02**	**0.04**
GL	0.1	0.28	**0.02**	0.33	**0**	1	0.43	**0.05**
ScBL	0.48	0.82	0.18	0.91	**0.02**	0.43	1	0.37
No	0.93	0.5	0.47	0.39	**0.04**	**0.05**	0.37	1

(f)

28d lead	AT	ScTr	ZO	AR	EuBL	GL	ScBL	No
AT	1	0.92	0.12	0.69	**0.04**	0.21	0.59	0.51
ScTr	0.92	1	0.15	0.61	**0.05**	0.17	0.51	0.58
ZO	0.12	0.15	1	0.06	0.57	**0**	**0.04**	0.24
AR	0.69	0.61	0.06	1	**0.02**	0.4	0.89	0.27
EuBL	**0.04**	**0.05**	0.57	**0.02**	1	**0**	**0.01**	0.07
GL	0.21	0.17	**0**	0.4	**0**	1	0.48	**0.03**
ScBL	0.59	0.51	**0.04**	0.89	**0.01**	0.48	1	0.2
No	0.51	0.58	0.24	0.27	0.07	**0.03**	0.2	1

Figure 8.5.: P-values of Welch's t-test for the binary forecasts of the occurrence of cold-wave days. The p-values are shown for the forecasts of the mean bias corrected ECMWF's S2S reforecast ensemble for lead times of 14 days (a), 21 days (c) and 28 days (e) as well as for the forecasts of the RFC-based postprocessing model using the ensemble information of the first ten PCs of the reforecasted fields and the first ten PCs of the reanalysis fields as input for lead times of 14 days (b), 21 days (d) and 28 days (f).

(#1) show a higher skill at a lead time of 28 days than of 21 days (Fig. 8.6 (g)). However, the shown differences are not significant with the exception of forecasts with lead times of 14 and 28 days during the ScTr regime (Tab. 8.2).

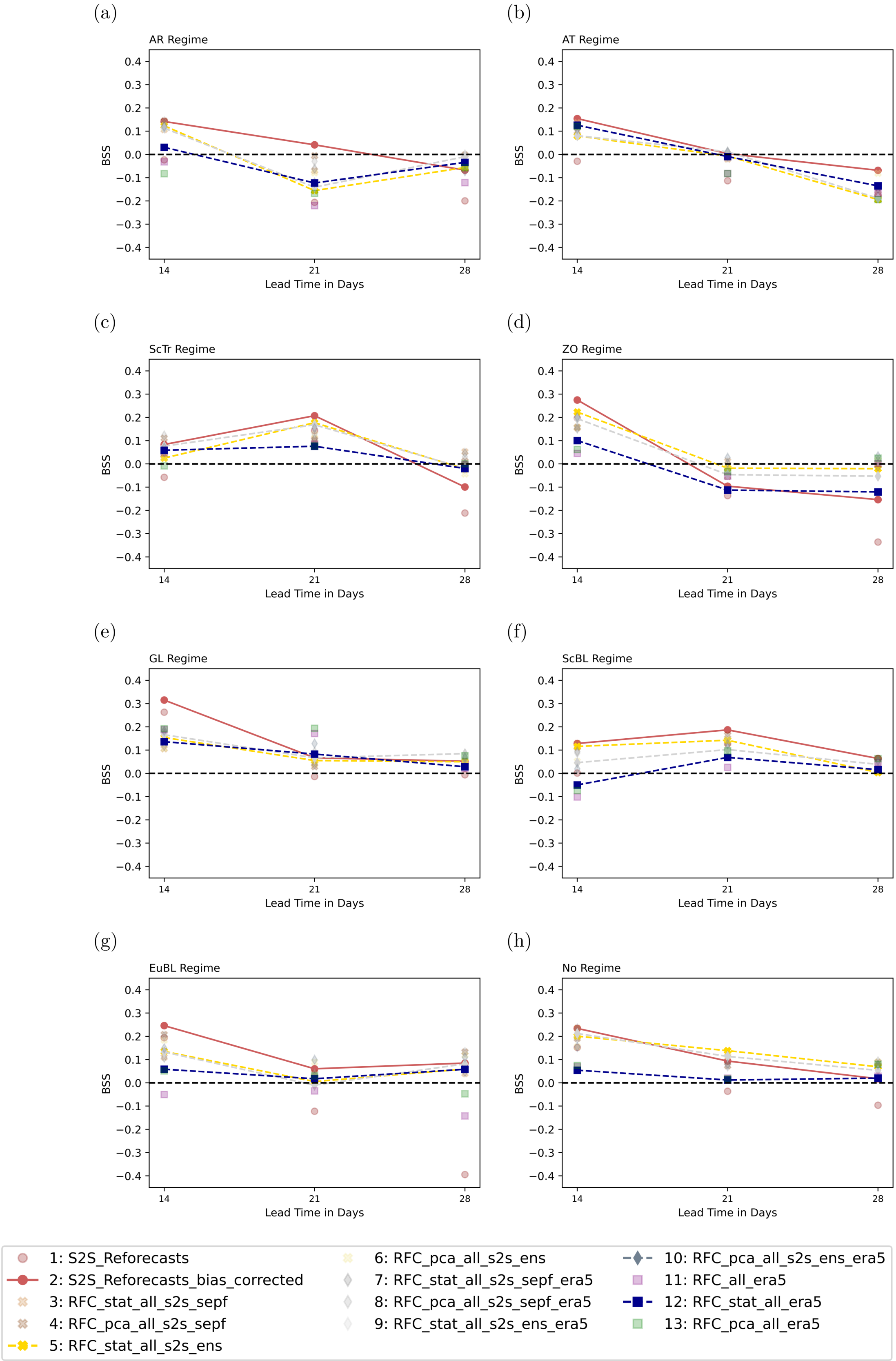


Figure 8.6.: Flow-dependent skill evolution across lead times. The skill is shown as BSS values for lead times of 14, 21 and 28 days for the AT (a), AR (b), ScTr (c), ZO (d), GL (e), ScBL (f), EuBL (g) and "No" regime (h).

Table 8.2.: P-values of Welch's t-test calculated for differences in BS values between lead times of forecasts. The calculations are done for every WR separately for ECMWF's mean bias corrected S2S reforecasts (abbreviated as "S2S" and the RFC_pca_all_s2s_ens_era5 model (#10, abbreviated as "RFC"). Corresponding figure is Fig.8.6 which shows the BSS values of the forecasts sorted after the WR at initialization.

	14 d vs. 21 d		14 d vs.28 d		21 d vs. 28 d	
Regime at Initialization	S2S	RFC	S2S	RFC	S2S	RFC
AR	0.66	0.96	0.22	0.25	0.09	0.27
AT	0.54	0.66	0.26	0.20	0.60	0.42
ScTr	0.25	0.20	**0.03**	0.07	0.28	0.62
ZO	0.47	0.98	0.93	0.42	0.52	0.41
GL	0.18	0.93	0.48	0.78	0.54	0.73
ScBL	0.85	0.71	0.29	0.44	0.21	0.25
EuBL	0.77	0.83	0.20	0.09	0.14	0.14
No regime	0.28	0.53	0.08	0.53	0.46	0.75

8.2. Linkage of WR Successions During the Forecast to Subseasonal Forecast Skill

In section 8.1, we find among other results that the mean bias corrected ECMWF's S2S reforecasts with a lead time of 14 days show a significantly better skill in the 20-winter mean when initialized during the GL regime in comparison to the ScTr regime. In the following, these differences are investigated further in terms of the WR successions observed during the time of the forecast. In order to increase the sample size, we use in the following all days of the winters 2000/2001-2019/2020 which show either the GL or ScTr regime instead of only the days where ECMWF's S2S reforecasts are initialized. This leads to a mixture of "hypothetical" forecasts (since no reforecasts are initialized at that date) and "real" forecasts. For better reading, we use only the term "forecasts" in the following. We assume that the number of days on which each regime is present, 757 in case of GL and 713 in case of ScTr, are similar enough to make a fair comparison. We chose the example of the mean bias corrected ECMWF's S2S reforecasts since these predictions rely on the physics of the atmosphere which are also represented by the WRs.

8.2.1. WR Successions During Two Illustrative Winters

To illustrate how the skill evolution of subseasonal forecasts might depend on the WR present at initialization, we perform two case studies. Since we focus on forecasts initialized during the GL and ScTr regime, we select two winters which contain both regimes. These are, among others, the winters of 2010/2011 and 2017/2018.

In the former, GL is the dominant regime at initialization for forecasts predicting the occurrence of (non-) cold-wave days between mid-November and mid-January (Fig. 8.7 (a)). Except for a single day, the BS difference is always positive during that time indicating that forecasts of the mean bias corrected ECMWF's S2S reforecasts initialized during the GL regime are more skillful than the climatological benchmark during this time, which features many cold-wave days. The ScTr regime is present at initialization in the beginning of November and mid-March. During these periods, the BS difference is either slightly positive or negative.

In case of the winter 2017/2018, the ScTr regime is present at initialization five times, roughly once every month (Fig. 8.7 (b)). Except for forecasts with target dates in the beginning of February, where the BS difference is slightly positive, the BS difference is either close to zero or slightly negative. The GL regime is present three times at initialization during this winter, once for forecasts predicting the occurrence of (non-) cold-wave days in November and twice in February/March. Except for the last one, the BS difference is positive or close to zero.

A striking difference between the GL regimes and the ScTr regimes in the analyzed two winters is their duration. The GL regimes tend to be more persistent such that the number of regime successions during the lead time of the forecast is smaller, which is a possible explanation for the skill differences. For computational simplicity, we measure this as the number of transitions between regimes (parameter one). During the GL regime, the WR index has higher values than during the ScTr regimes and there is a tendency for the other regimes to show more negative values of the WR index during the GL regimes than during the ScTr regimes. This means that the number of active WRs at each day of lead time is less for the GL regime, a second possible explanation for the skill differences. We measure this as the overall number of active WRs at each day of lead time during all forecasts of the winters 2000/2001-2019/2020 (parameter two). For simplicity, we only take the most prominent WRs into consideration as represented by the categorical WR of the respective day. The third difference between the two regimes at initialization is the subsequent regime that occurs. In case of GL, either the AR or AT regime is following. In case of ScTr, the subsequent regime is either the GL, EuBL, AT or ZO. Thus, the third parameter considered is the actual succession of WRs during the forecast.

In the following, we analyze the three named parameters for all forecasts of the winters 2000/2001-2019/2020 initialized during the GL or ScTr regime.

8.2.2. Number of Regime Transitions During the Forecasts

Regime changes might be more difficult to forecast than persistence. Therefore, we compare the number of regime changes depending on the WR at initialization and the target date. The fraction of forecasts reaching a specific WR at the target date is different depending on the WR at initialization. Thus, for every set of forecasts, we weigh the median regime changes per WR present at the target date by the fraction of forecasts reaching the respective WR.

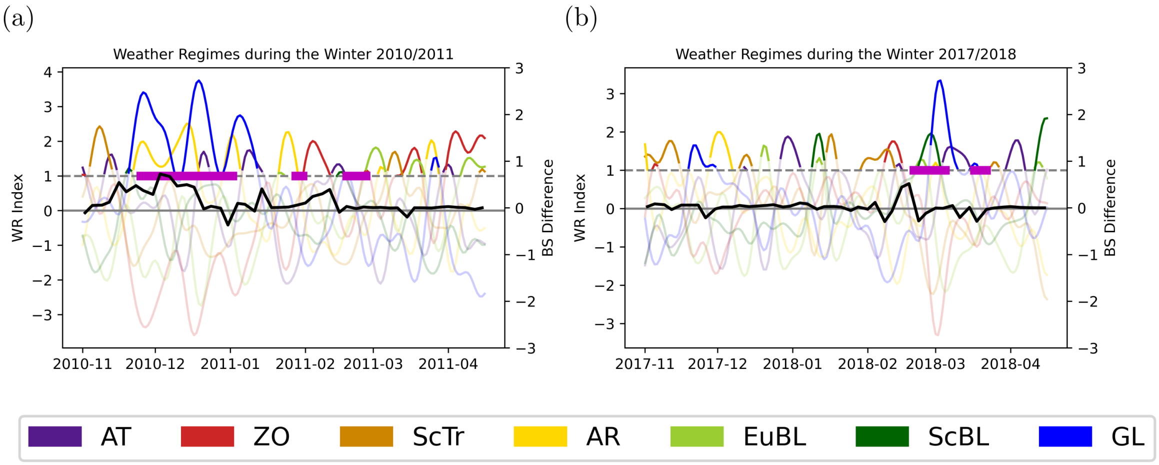


Figure 8.7.: WR index during two winters. The WR index is shown with a lag of 14 days (= WR at initialization) for the winters 2010/2011 (a) and 2017/2018 (b). The thick black line shows the difference in BS values of the mean bias corrected ECMWF's S2S reforecasts and the climatological benchmark ensemble. Positive values denote a better performance of the reforecasts. The magenta dashed line shows the occurrence of cold-wave days.

The weighted median of regime changes during the lead time of 14 days is 1.21 for forecasts initialized during the GL regime and 1.22 for forecasts initialized during the ScTr regime (Fig. 8.8). This means that for forecasts initialized during the GL or ScTr regime, on average, one or two regime transitions occur during the 14-day forecast period. Both regimes show a maximum of four regime changes during the forecast. The interquartile range between the number of regime changes and the number of outliers differ between the regimes present at the target date. However, overall the differences are small.

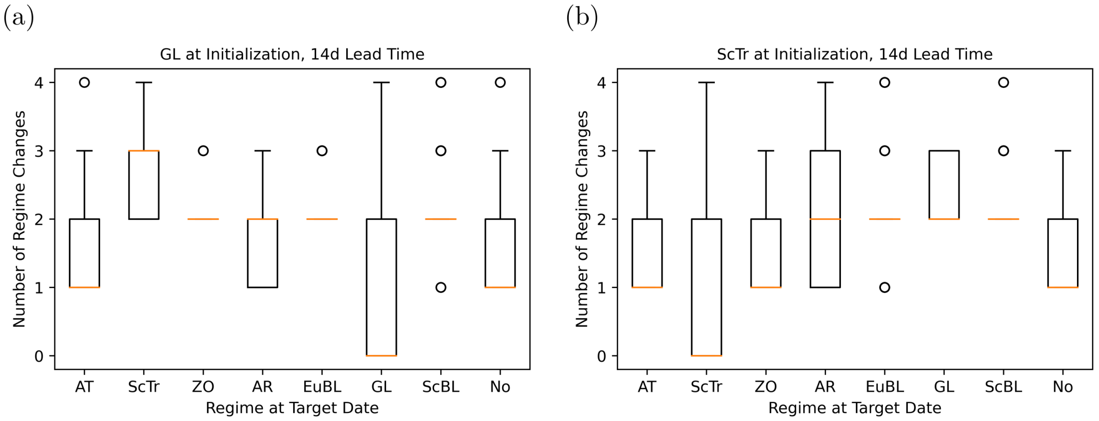


Figure 8.8.: Number of regime transitions during the forecast. The number of regime changes are calculated for all forecasts of the winters 2000/2001-2019/2020 initialized during the GL (a) and ScTr regime (b) for a lead time of 14 days. The orange line marks the median of regime changes, the boxes the interquartile range and the open circles the 1.5 times interquartile range.

8.2.3. Overall Number of Active WRs per Day of Lead Time

As a next step, we investigate the WR successions during the forecasts. Here, we do this in a cumulative way focusing on the number of active WRs per day of lead time summed over all forecasts (Fig. 8.9 and 8.10). As an example, if on one day of the lead time either the AT or AR regime is present in all considered forecasts, the overall number of active WRs at his day of the lead time is two.

We hypothesize a higher predictability when the overall number of active WRs per day of lead time is low. Furthermore, we hypothesize a higher predictability if the fraction of a small number of WRs (e.g. two) is high (e.g. occuring in more than 50% of the forecasts) in comparison to the fraction of the other WRs. These numbers are determined for the forecasts of the winters 2000/2001-2019/2020, sorted by the WR present at initialization (either GL or ScTr) and target date (all WRs).

In 31.6% of forecasts initialized during the GL regime, the "No" regime is reached at the target date (Fig. 8.9 (e)). Thereby, at each day during the forecasts, in more than 60% of the cases either the GL or "No" regime is present. The second most often occurring regime at the target date is the GL regime itself (Fig. 8.9 (d)). Here, in more than 60% of the cases the GL regime is present at each day of the forecasts. Together, these two account for roughly half of all forecasts initialized during the GL regime. In another roughly 30% of all forecasts, namely the ones with the AR, AT and ZO regime at the target date, mainly two regimes occur during the forecasts but with varying fractions (Fig. 8.9 (a), (b) and (h)). The remaining cases show a more diverse picture with up to seven regimes per forecast day (Fig. 8.9 (c), (f) and (g)).

Likewise, for forecasts initialized during the ScTr regime, the "No" regime occurs most often at the target date (Fig. 8.10). Here, in more than 2/3 of the cases, either the ScTr or the "No" regime is present at each day of the forecasts (Fig. 8.10 (e)). The second most often occurring regime at the target date is the ScTr regime with the ScTr regime also present at each day of the lead time in a little less than 50% of the cases (Fig. 8.10 (g)) and the third most often occurring regime at the target date is the ZO regime with the ScTr or ZO regime present at each day during the forecasts in 70% of the cases (Fig. 8.10 (h)). The latter two regimes occur roughly equally often and make up approximately 60% of all forecasts together with the ones with the "No" regime at the target date. During the other regimes at target date, the fraction of the regimes during each day of the forecasts is more diverse (Fig. 8.10 (a), (b), (c) (d) and (f)).

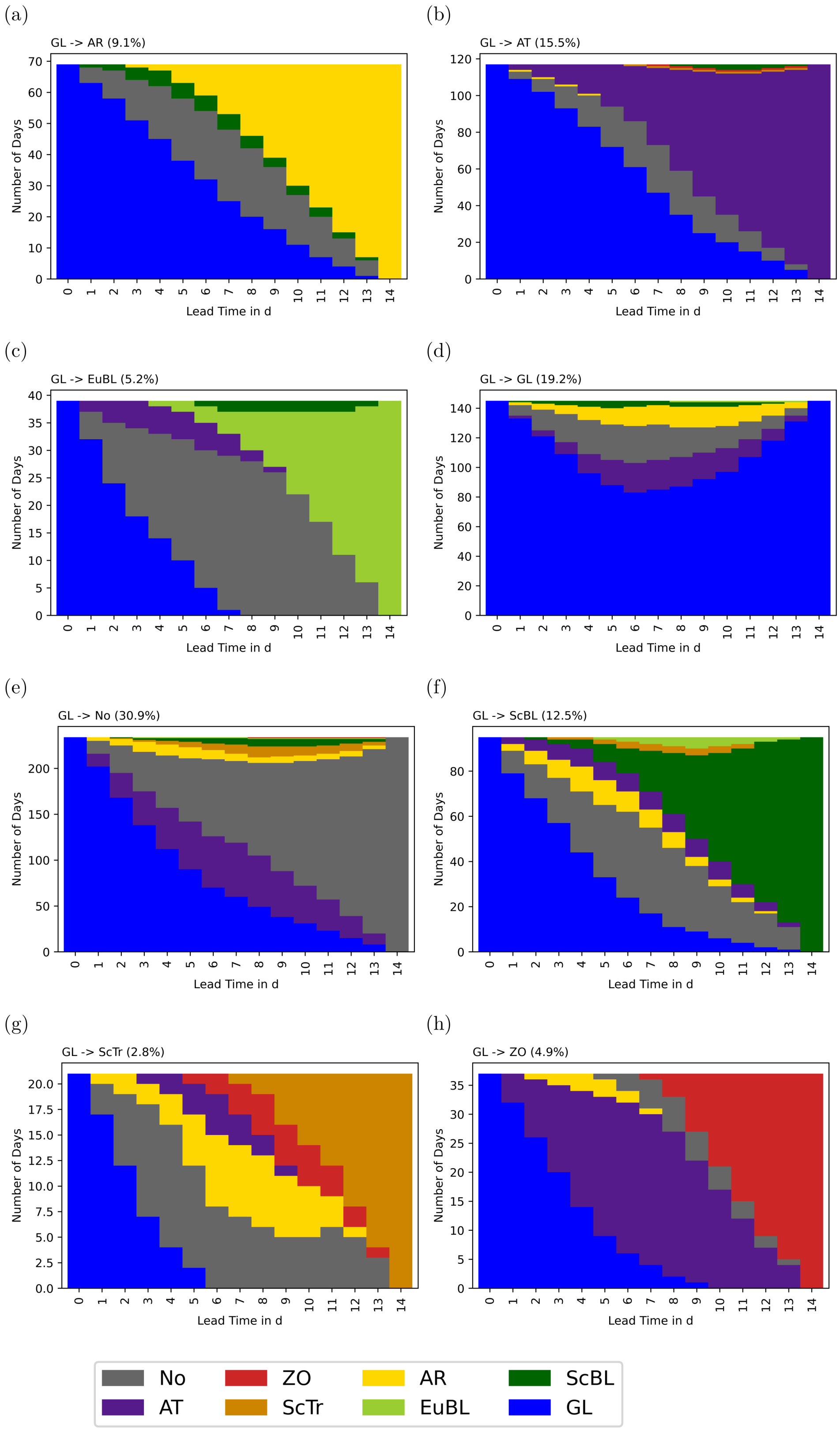


Figure 8.9.: Frequency of WRs for forecasts initialized during the GL regime with a lead time of 14 days. The frequencies of WRs are shown for forecasts with the AR (a), AT (b), EuBL (c), GL (d), No (e), ScBL (f), ScTr (g) and ZO (h) regime present at the target date of the forecast. The fraction of forecasts reaching the different WRs at their target date are given in the titles above the subplots.

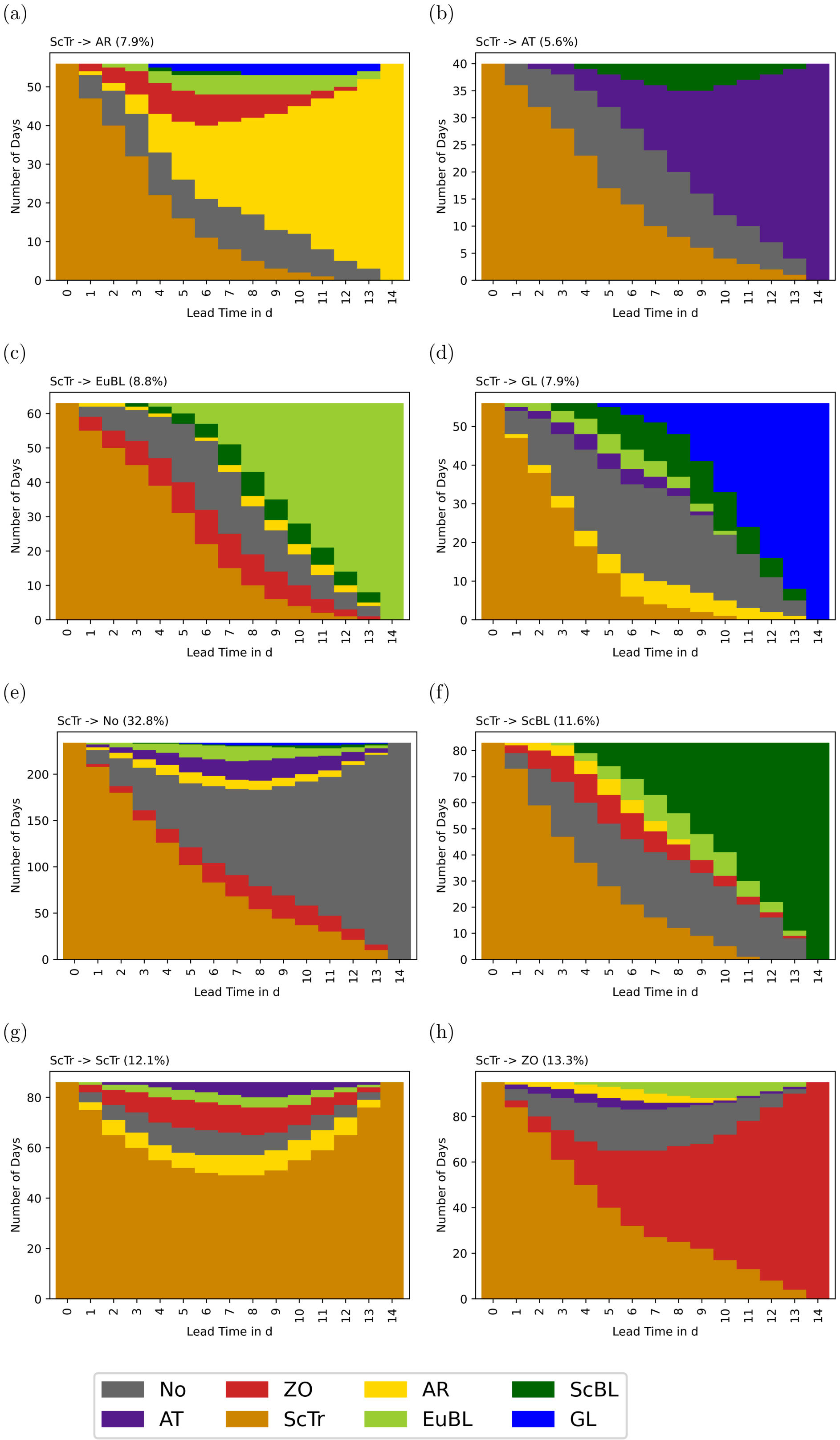


Figure 8.10.: Frequency of WRs for forecasts initialized during the ScTr regime with a lead time of 14 days. The frequencies of WRs are shown for forecasts with the AR (a), AT (b), EuBL (c), GL (d), No (e), ScBL (f), ScTr (g) and ZO (h) regime present at the target date of the forecast. The fraction of forecasts reaching the different WRs at their target date are given in the titles above the subplots.

8.2.4. Actual WR Successions During the Forecasts

Besides the overall number of active WRs at each day during the forecast, we investigate the role of the actual WR successions. This means that we are now focusing on the single WR successions (e.g. "GL → AT") of the forecasts. Possibly, some WR successions are easier to forecast than others. We assume that this is especially the case when WR succession follow typical climatological patterns. To test this, we first analyze all WR successions occurring during the winters 2000/2001 - 2019/2020. Then, we split the regime sequence during a forecast into parts of two WRs each (e.g. "GL → AT → No" is split into "GL → AT" and "AT → No"). If all parts of the regime sequence during a forecast follow typical climatological patterns, the whole WR sequence is considered to do so. During the analysis, we focus on the sequence of regimes as a whole without taking persistence of the individual WRs per se into account. However, when only few WRs are present, the persistence of at least one WR must be high. If one WR is persistent during all 14 days of the forecast, it is treated as a WR succession of only one WR for simplicity of analysis even if it is technically not a "succession".

The most often occurring actual WR succession during the winters 2000/2001-2019/2020 is the transition from ScTr to the "No" regime (Fig. 8.11). It is followed by the successions "No → EuBL" and "AT → No". The least often occurring WR successions are "ScTr → ScBL", "ScBL → ScTr", "GL → ScBL", "ZO → ScBL", "AT → EuBL", "ScTr → EuBL", "EuBL → GL", and "GL → ScTr".

We define the top 11 successions present during the winters 2000/2001 - 2019/2020 as the "typical climatological patterns". These are "ScTr → No", "No → EuBL", "AT → No", "No → ScTr", "AR → No", "GL → No", "No → ZO", "No → ScBL", "No → AT", "ZO → No", and "No → AR". We consider the top 11 instead of the top ten successions because the tenth and 11th most often occurring WR succession have the same frequency. The "No" regime is present in every typical WR succession which is expected since according to Grams et al. (2017), the "No" regime summarizes the atmospheric flow configurations which are closer to the climatology than to any of the other seven WRs.

For the forecasts initialized during the GL regime, the number of possible WR successions varies between three for ZO and EuBL at the target date and 13 for the "No" regime (Fig. 8.12). In case of the latter, two of these 13 options make up the majority of successions (Fig. 8.12 (e)). These are "GL → No" and "GL → AT → No", whereby the former is following typical climatological patterns. In case of the ZO regime at the target date, the most important is "GL → AT → ZO" which is not following typical climatological patterns (Fig. 8.12 (h)). When EuBL is present at the target date, the most often actual WR succession is "GL → No → EuBL" which is following typical climatological patterns (Fig. 8.12 (c)). When the ScBL regime is present at the target date, the most often occurring WR succession is "GL → No → ScBL" (following typical climatological patterns), and when the AT regime is present at the

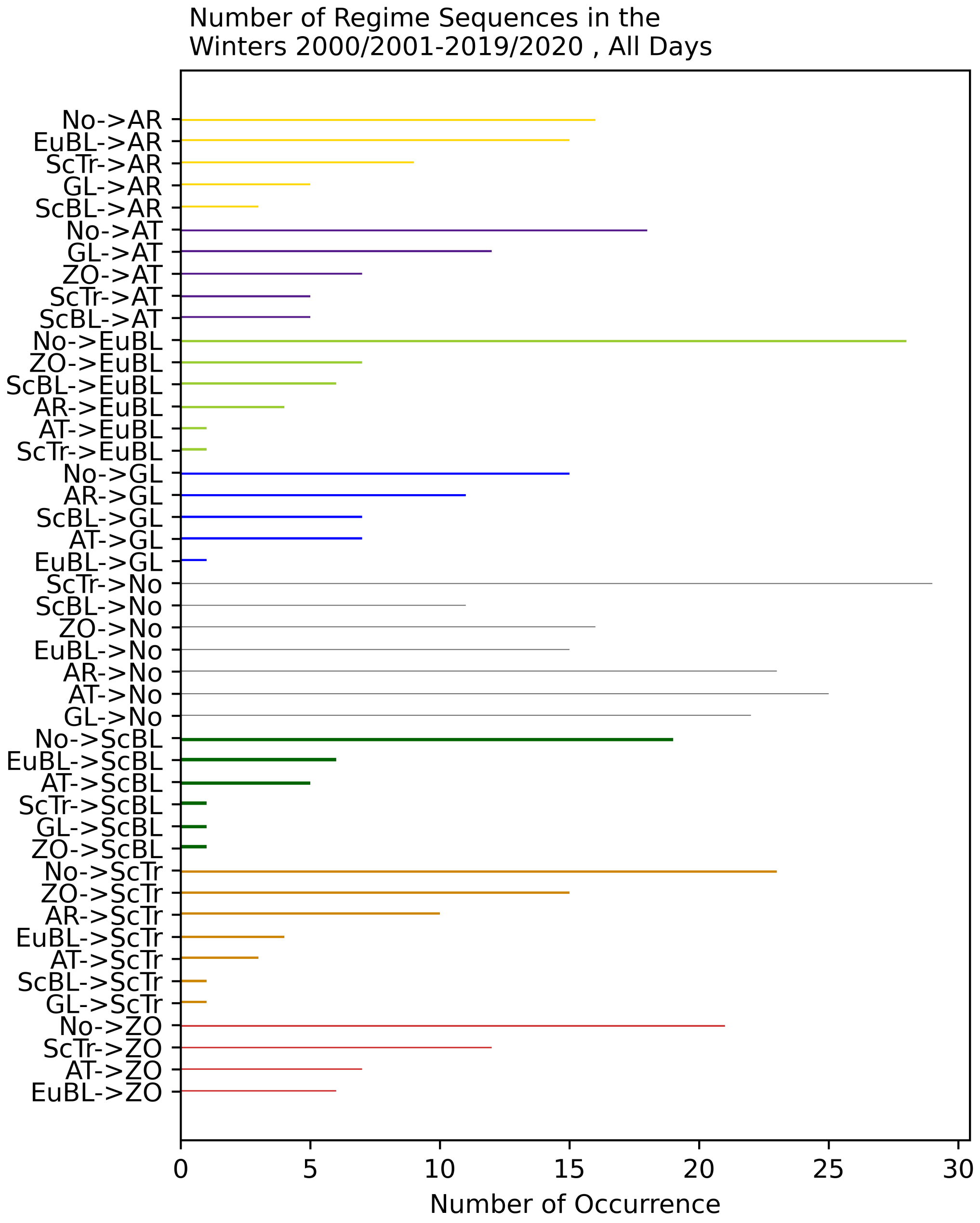


Figure 8.11.: Preferred actual WR successions during the winters 2000/2001 - 2019/2020.

target date is is "GL → AT" (Fig. 8.12 (b) and (f)). For forecasts with AR at the target date, the successions "GL → No → AR" and "GL → AR" are most often found and equally likely but only the former is following typical climatological patterns (Fig. 8.12 (a)). In case of the GL present at initialization and target date, it is persistence (Fig. 8.12 (d)) and in case of the ScTr regime found at the target date, there is no clearly preferred succession (Fig. 8.12 (g)).

Persistence is also the most common for forecasts with the ScTr regime present at the target date when initialized during the ScTr regime (Fig. 8.13 (g)). Clear common actual WR successions are found for forecasts with the "No", ScBL and ZO regime present at the target date (Fig. 8.13 (e), (f) and (h)). These are "ScTr → No", "ScTr → No → ScBL" and "ScTr → ZO". With the exception of the latter, these actual WR successions follow typical climatological patterns. A clear preferred WR succession is not seen for forecasts with the AR, AT, EuBL and GL regime at the target date (Fig. 8.13 (a), (b), (c) and (d)).

Considering only the most often occurring actual WR successions per WR at the target date of the forecasts initialized during the GL and ScTr regime, we find that 61.6% of the forecasts initialized during the GL regime show actual WR successions that follow typical climatological patterns. Additionally, 18.1% of the forecasts show persistence. In case of forecasts initialized during the ScTr regime, 53.9% of the forecasts show actual WR successions following typical climatological patterns and 13.5% of the forecasts persistence. This supports our hypothesize that actual WR successions following typical climatological patterns might be easier to forecast and thus leading to an increased forecast skill. Furthermore, persistence of WRs seems to be a factor for an increase in skill.

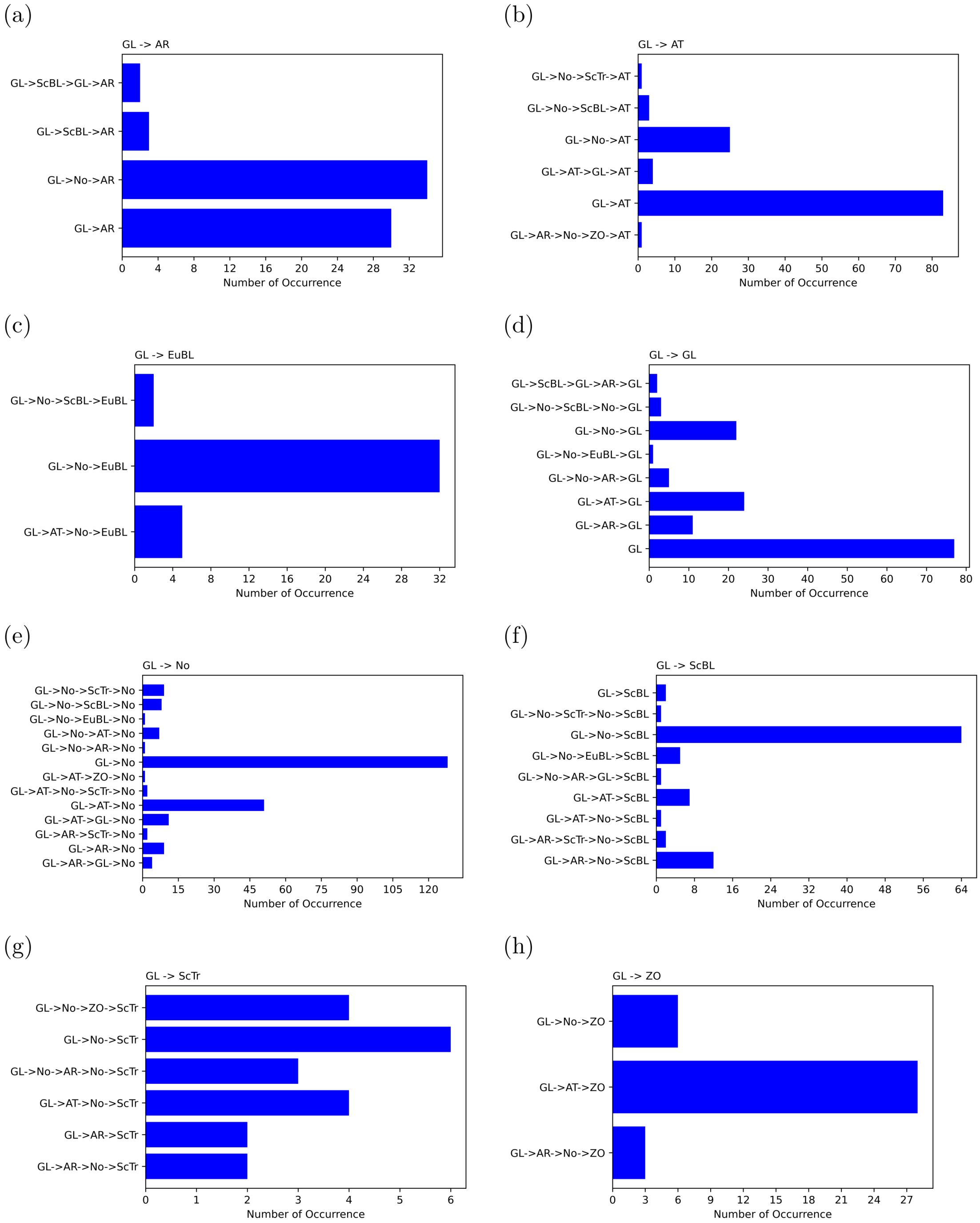


Figure 8.12.: Successions of WRs for forecasts of the winters 2000/2001-2019/2020 with a lead time of 14 days initialized during the GL regime. The frequencies of WR successions are shown for forecasts with the AR (a), AT (b), EuBL (c), GL (d), No (e), ScBL (f), ScTr (g) and ZO (h) regime present at the target date of the forecast.

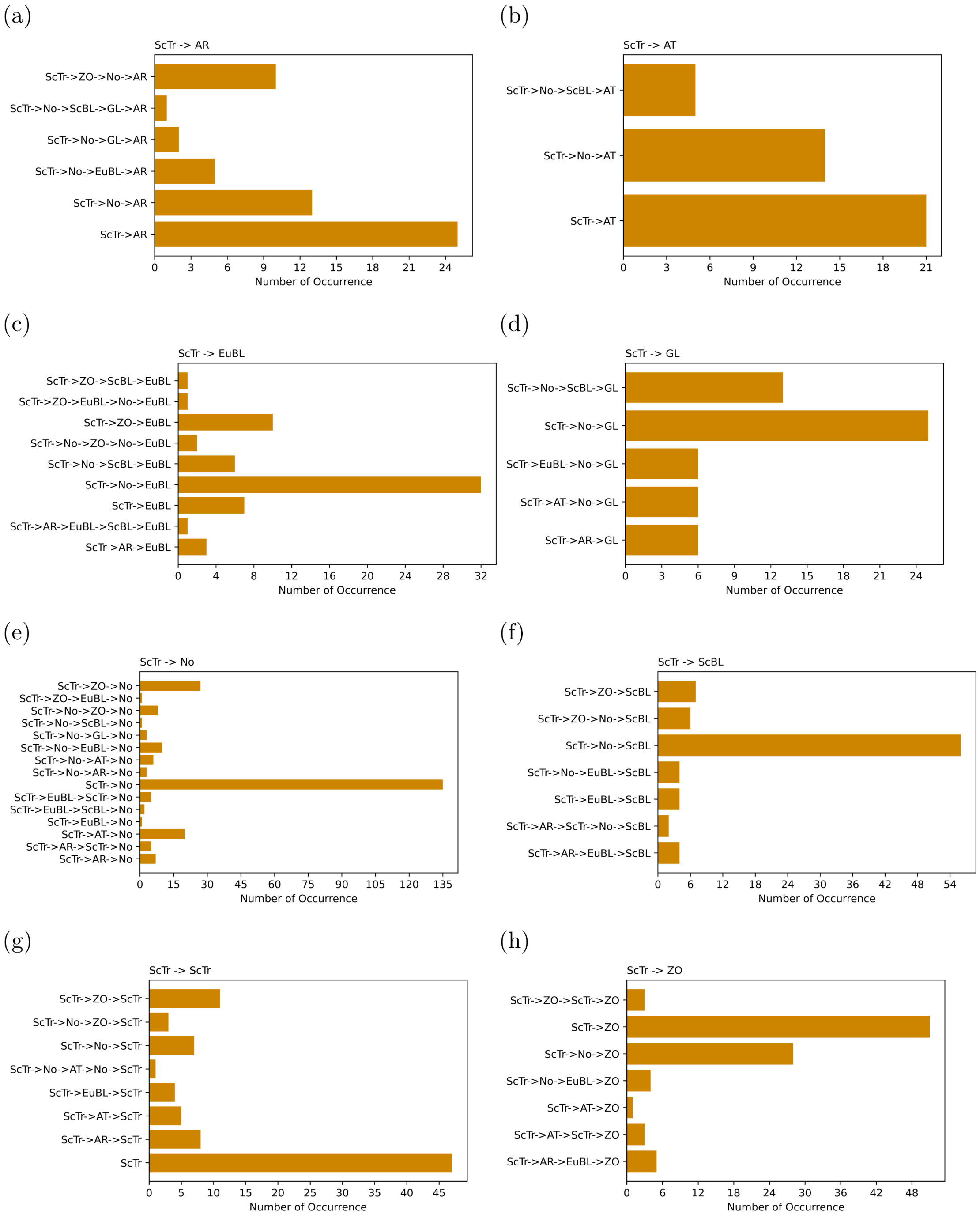


Figure 8.13.: Successions of WRs for forecasts of the winters 2000/2001-2019/2020 with a lead time of 14 days initialized during the ScTr regime. The frequencies of WR successions are shown for forecasts with the AR (a), AT (b), EuBL (c), GL (d), No (e), ScBL (f), ScTr (g) and ZO (h) regime present at the target date of the forecast.

8.3. Differences in WR Successions Before the Best (Worst) Predicted Days Within Cold Waves

Analogously as done for to the occurrence of (non-) cold-wave days, we investigate the WR characteristics during days within cold waves. This is done since arguable days within cold waves, which feature very often temperatures below the frost point (Fig. 5.2), are for most practical applications, e.g. agriculture and transportation, more important to forecast correctly than days with mild winter temperatures. In contrast to Section 8.2, in this section only the "real" S2S reforecasts from ECWMF are considered since comparisons based on forecast skill are made.

8.3.1. WR-Characteristic of All Cold-Wave Days During the Winters 2000/2001 - 2019/2020

The most frequent WR present at the start of a cold wave is the AR regime, which is found at roughly 30% of the cold wave starts during the winters 2000/2001-2019/2020 (Fig. 8.14 (b)). The GL regime is present at the start of approximately 1/5 of the cold waves, followed by the "No" regime, which is found in roughly 17% of the cases. All other regimes are less often present at the start of cold waves, whereby the EuBL and ScBL regime are found in roughly 11% of the cases each. The AT regime is not found at all at the start of the cold waves during the winters 2000/2001 - 2019/2020.

Up to four regime transition are observed during cold waves, whereby the median number of regime transitions is one (Fig. 8.14 (c)). The interquartile range is thereby between zero and two. Most often, a regime transition occurs in the second half of the cold wave after approximately 55% of the days (Fig. 8.14 (d)). Here, the interquartile range is between 35% of the days and the end of the cold wave.

Concerning the actual WR successions during cold waves, no clearly preferred successions are found (Fig. 8.15 (a)). Most often observed is the persistence of AR followed by the WR succession "GL → No". However, these occur only during five, respectively three, cold waves of the winters 2000/2001-2019/2020. All other WR successions occur only once or twice.

Since air masses need a certain time to reach and influence Central European temperatures from their origin, the actual WR successions observed in the week before the cold wave starts are considered. However, this provides an even less clear picture (Fig. 8.15 (b)). Before four cold waves each, either the GL or ScBL regime is persistent. Other WR successions occur again only once or twice before the cold waves of the winters 2000/2001-2019/2020. When looking at the WR index of the week before the cold waves instead of the categorical WRs, in the mean, the AR and GL regime show the highest values until one day before the cold wave start. However, the large spread of the WR indices shows that there is no clearly preferred WR (Fig. 8.14 (a)).

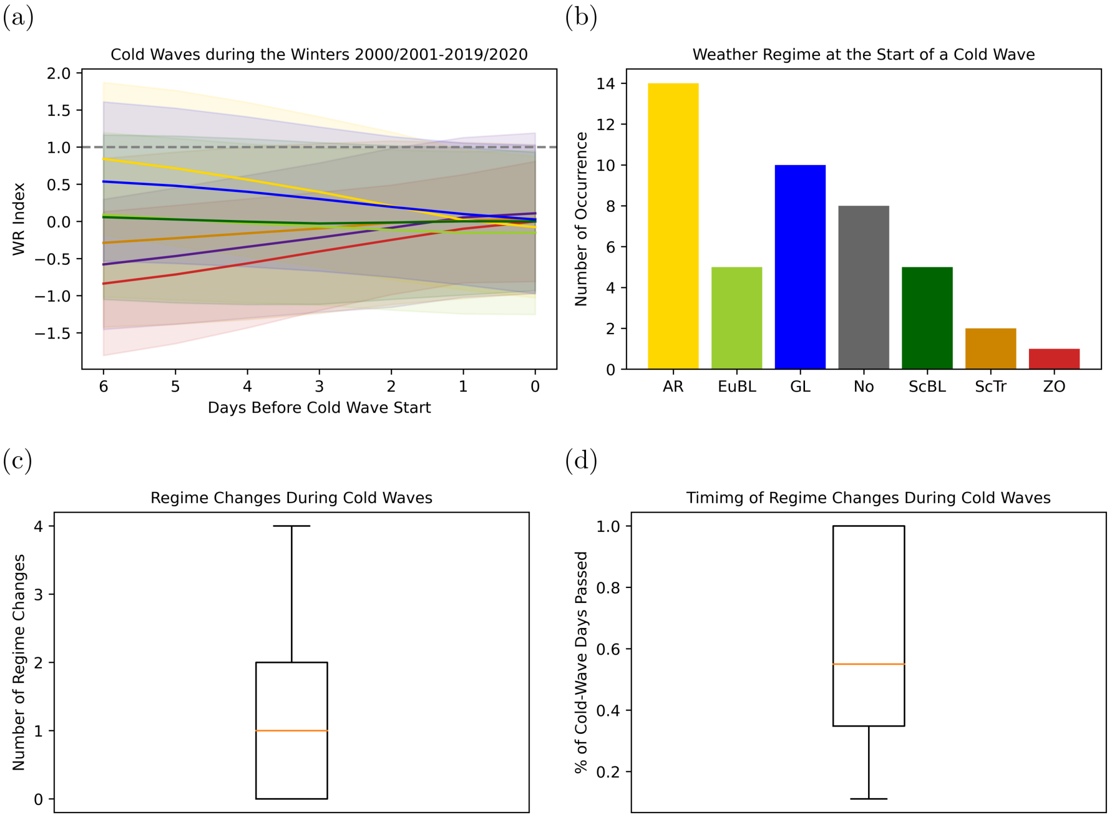


Figure 8.14.: WRs at the start of cold waves, number and timing of WR transitions during cold waves during the winters 2000/2001-2019/2020. The median WR index (solid lines) and its spread (shading) is shown for the week before the cold wave start ((a), colors as in (b)). The categorical WR at the start of the cold wave is depicted on panel (b). Furthermore, the number of regime changes during a cold wave (c) and the timing of the regime changes (d) are shown.

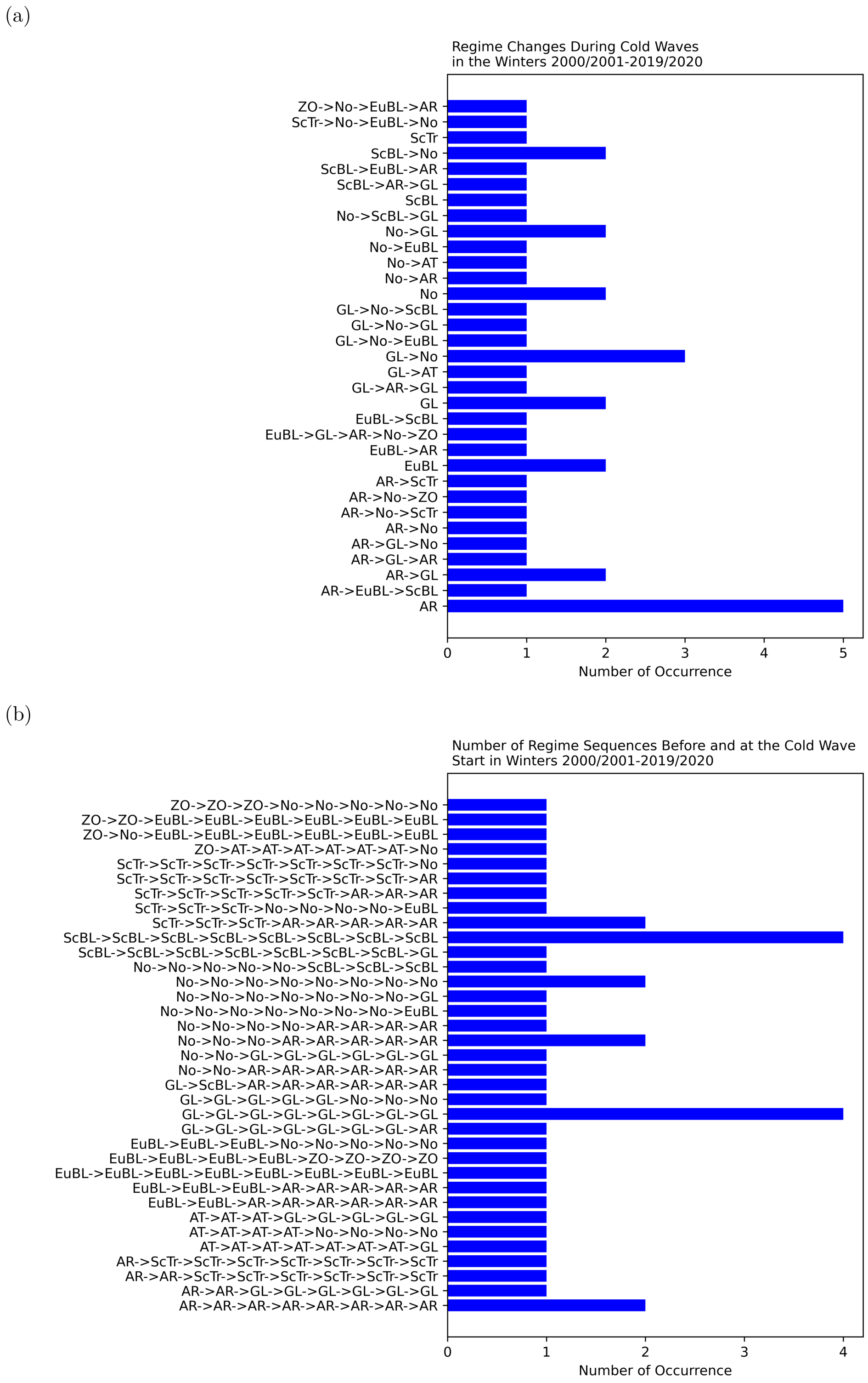


Figure 8.15.: WR successions before cold-wave days and WR transitions during cold waves during the winters 2000/2001-2019/2020. The WR transitions during cold waves (a) and the WR successions in the week before the cold wave start (b) are shown.

8.3.2. WR Index of Best and Worst Predicted Cold-Wave Days

In order to analyze what distinguishes the best from the worst predicted cold-wave days, we use the continuous WR index instead of the categorical WRs. This is done, since a large proportion of the days is usually classified as "No" regime. Nevertheless, also during the "No" regime the atmospheric conditions can vary substantially, which is depicted by the continuous WR index.

We compare the WR indices of the 10% best and worst predicted cold-wave days of the winters 2000/2001-2019/2020 independent of the WR present at the initialization of the respective forecasts. These are determined via the BS difference between mean bias corrected ECMWF's S2S reforecasts and the climatological benchmark. In total, we analyze the 13 best predicted cold-wave days, of which 12 are intermediate days and one the last day of a cold wave. The days are scattered over eight different winter periods, whereby at most two days belong to the same winter and cold wave. Of the 13 worst predicted cold-wave days, eight are intermediate days, three end and two start days of cold waves. These days are spread over nine different winters, whereby at most three days belong to the same winter and two to the same cold wave.

We concentrate our analysis on the WR index values above one, which indicate the presence of a well defined regime, and take the daily mean of those. In case of the best predicted cold-wave days, the AR regime is the most prominent regime one day before initialization until eight days after initialization (Fig. 8.16 (a)). The GL regime also has a high contribution during this time. Between six and three days before the cold wave starts, the EuBL regime is most prominent. In the remaining days before the cold wave, the ScTr regime is most pronounced.

For the 10% worst predicted cold-wave days, the GL regime is most prominent between 17 and 11 days before the cold-wave day occurs (Fig. 8.16 (b)). Then, the AR regime is the most pronounced WR until seven days prior to the onset of the cold wave followed by the ScTr regime being most prominent until two days before the cold wave begins. In the last two days prior to the cold wave, both the GL and AR regime are most pronounced.

According to a Welch's t-test, the difference of the AR and EuBL regimes between the best and worst predicted cold waves is significant. Thus, we suggest that the occurrence of the EuBL regime in the week before the predicted cold-wave day increases forecast skill. Furthermore, the timing of the occurrence of the AR regime seems to play a role.

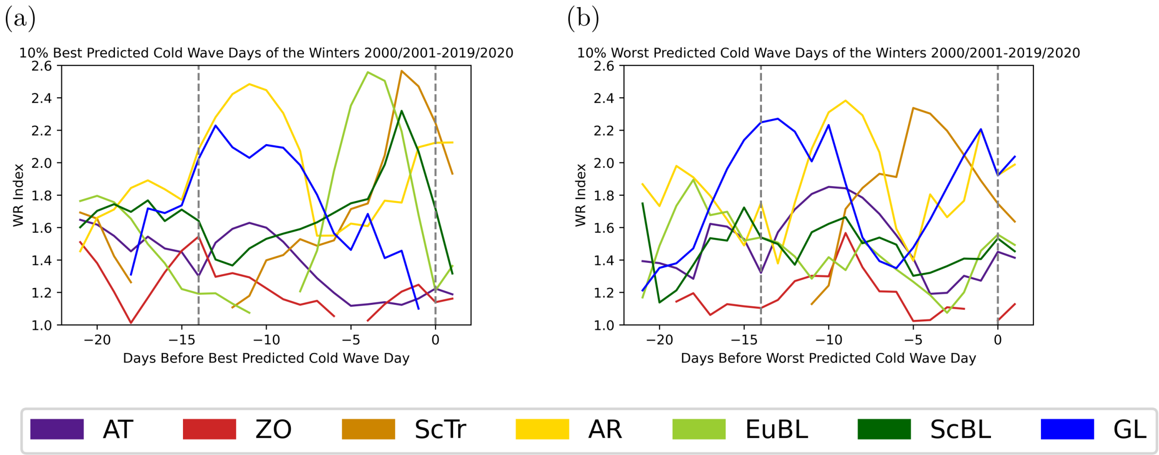


Figure 8.16.: Timeseries of the WR index of the best and worst predicted cold-wave days during the winters 2000/2001 - 2019/2020. The timeseries of the WR index of the 10% best predicted cold-wave days is shown on the left (a) and the timeseries of the WR index of the 10% worst predicted cold-wave days on the right (b). Only the mean of values above 1 are shown. The vertical dashed lines show the initialization of the forecasts and the respective cold-wave day.

8.3.3. Illustrative Case Studies of Two Cold Waves

Given the considerable variability between individual cold waves in Central Europe, we select two case studies to show the practical relevance of our results and to illustrate the characteristics of such events. These cold waves feature a best (worst) predicted cold-wave day and the WRs present are close to the dominant WRs in the mean of the 10% best (worst) predicted cold-wave days.

One of the cold waves featuring a best predicted cold-wave day of the winters 2000/2001-2019/2020 is the one occurring in January and February 2006 (Fig. 8.17 (a)). In this case, the ZO regime is present until two days before the initialization of the forecast and then followed by the "No" regime, which persists until four days after initialization. During that time, the strong zonal flow which leads to stormy and warm winter weather is continuously weakened. The former strong Icelandic Low weakens, while simultaneously a strong high pressure system over the British Isles and southern Scandinavia evolves and the EuBL regime is established. The longer the regime persists, the better the forecast skill gets.

The cold wave in January 2019 features one of the worst predicted cold-wave days of the considered 20 winters (Fig. 8.17 (a)). The atmospheric large-scale circulation in the week before the forecast initialization is almost equally close to both the EuBL and AR regime. The strong high pressure system is located between Iceland and Scandinavia propagating to the north-west while the AR regime is getting more pronounced in the seven days after the forecast initialization. At the same time, a moderate Icelandic Low and a moderately strong high pressure system over the North Atlantic Ocean form during the ScTr regime. Possibly, this is due to wave breaking and an easterly flow of air masses. This would explain the cold temperatures occurring in Europe although the atmospheric large-scale circulation is closest to the ScTr regime which is usually characterized by a zonally transport of warm air masses across the Atlantic Ocean towards Central Europe.

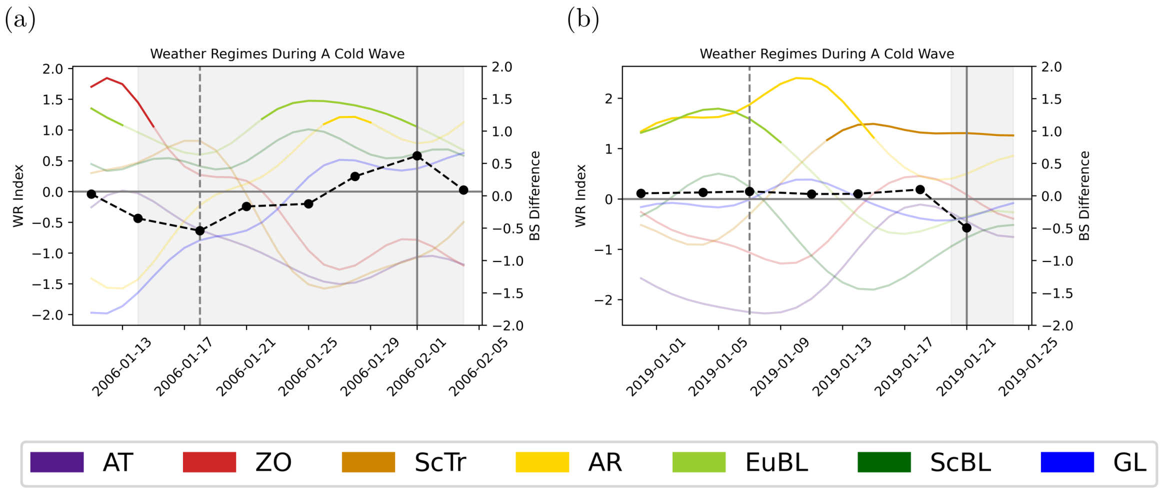


Figure 8.17.: WR index before and during two exemplary cold waves. The WR index is shown for a cold wave with a one of the best predicted days in the 20-winter mean on the left (a) and for a cold wave with one of the worst predicted days in the 20-winter mean on the right (b). The black dashed line with the markers shows the BS difference of the forecast days. The vertical dashed show the initialization of the forecasts and the solid black line the best/worst predicted cold-wave day.

8.3.4. Influence of the Timing of a Day Inside a Cold Wave on the Forecast Skill

Since we use a temporal criterion in the definition of cold waves, which is that at least three consecutive days experience temperatures below a certain threshold, we assume that days in the middle or at the end of cold waves are generally better forecasted since the atmospheric conditions leading to the fulfillment of this criterion are established longer.

To validate this assumption, we sort all days belonging to the cold waves in the winters 2000/2001-2019/2020 into the three categories "start days", "end days", and "intermediate days". Depending on the length of the cold wave, the number of intermediate days varies. In total, we analyze 20 start days, 106 intermediate and 15 end days. The different number of start and end days results from the bi-weekly initialization of ECMWF's S2S reforecasts, which not necessarily falls together with the start or end of a cold wave with respect to the lead time of 14 days. We find that the mean bias corrected ECMWF's S2S reforecasts predict days during and at the end of a cold wave better than days at the beginning (Fig. 8.18 (a)). On average, intermediate and end days of cold waves are predicted better in comparison to the climatological ensemble while days in the beginning of cold waves are not. However, these differences are not statistically significant.

For a comparison, we additionally analyze the forecasts of an RFC-based postprocessing model. In this case, only the intermediate days of cold waves are predicted better than by the climatological ensemble (Fig. 8.18 (b)). Their distribution of BS differences to the distribution belonging to the start days of the forecasts is statistically significant. However, the start and end days of cold waves are predicted less skillful than by the climatological ensemble.

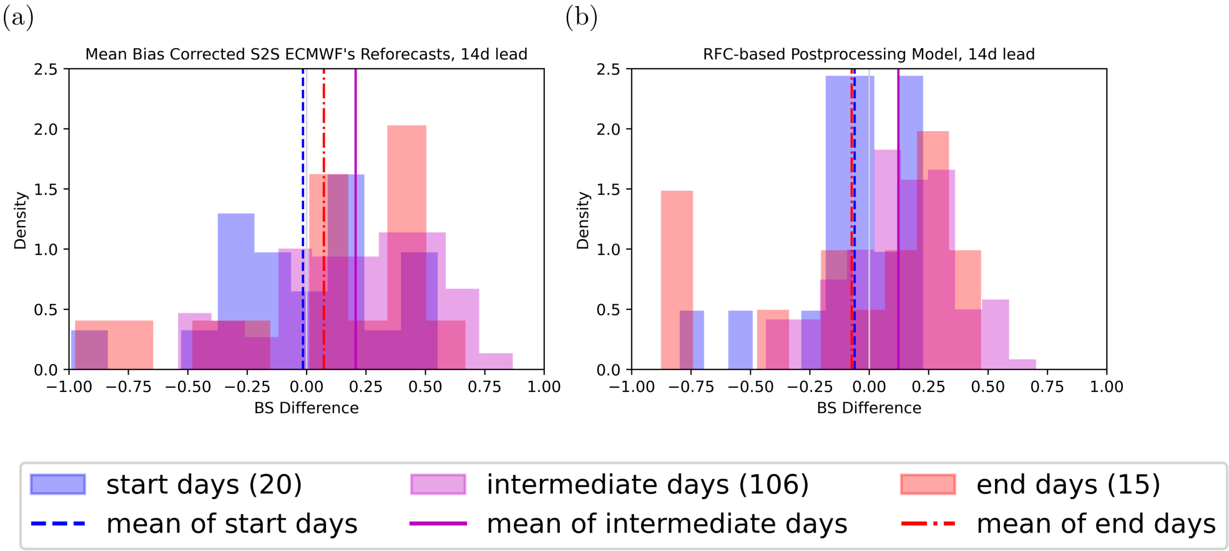


Figure 8.18.: BS differences of forecasts of cold-wave days. The daily BS differences of the mean bias corrected ECMWF's S2S reforecasts (a) and the RFC-based postprocessing model (b) to the climatological benchmark ensemble are shown. The vertical grey line marks the BS difference at which both models predict equally well. A positive BS difference shows that the model performs better than the climatological benchmark ensemble.

8.4. Summary and Discussion

This part of our research explores how the information about the WR present at the forecast start and the WR successions during the forecast can be used to assess forecast reliability. Analogously to chapter 6 and 7, we focus on the subseasonal forecasts of wintertime 2-meter temperatures and the occurrence of cold-wave days in Central Europe.

As a first step we analyze if the WRs present at initialization can be used for assessing forecast reliability. Thereby, we answer the following question:

RQ 3.1 Does the forecasting skill of ECMWF's S2S reforecasts and the RF-based models depend on the WRs at initialization?

- Although the skill of all models depends on the WR present at initialization, only few differences in skill between the forecasts initialized during the various WRs are significant.

- The skill evolution across lead times varies between forecasts initialized during the different WRs but the deviations are most of the time not significant.

We suggest that in cases where the differences in skill are significant, the forecasts initialized during one WR have a conditionally improved skill in comparison to forecasts initialized during another WR. This is especially relevant when forecasts initialized during one WR have the highest skill score values of all forecasts and forecasts initialized during the other WR the lowest. Concerning the forecasts of continuous 2-meter temperature, this is true in two cases. The mean bias corrected ECMWF's S2S reforecasts initialized during GL show a conditionally improved skill compared to the reforecasts initialized during ScTr at a lead time of 14 days. The skill of predictions of a representative QRF-based postprocessing model is conditionally improved when initialized during the ScBL regime in comparison to when initialized during the GL regime at a lead time of 21 days. Binary cold-wave day predictions of the mean bias corrected ECMWF's S2S reforecasts are conditionally improved at a lead time of 14 days when initialized during the GL regime in comparison to forecasts initialized during the ScTr regime. At a lead time of 28 days, binary cold-wave day forecasts of a representative RFC-based postprocessing model are conditionally improved when started during the EuBL regime in comparison to predictions started when the AT regime is present.

Depending on the various regimes at initialization, also the evolution of skill across lead times varies. The differences in skill between forecasts of different lead times are mostly non-significant, though.

As a next step, we analyze the predictions of the mean bias corrected ECMWF's S2S reforecasts at a lead time of 14 days initialized during the GL and ScTr regime in more detail. We focus especially on these predictions since their differences in skill are significant and the predicting model is relying on physics which are also represented by the WRs. We start by answering the following research question:

RQ 3.2 In how far can the WR succession during a forecast be linked to subseasonal forecast skill?

- The WR successions during a forecast can be linked via their observed occurrences to subseasonal forecast skill. Forecasts showing a higher skill feature more WR succession which follow typical climatological patterns than forecasts with a lower forecast skill.
- The overall number of active WRs per day of forecast, accumulated for all forecasts of the winters 2000/2001-2019/2020 initialized during the same WR, is not a main driver in increasing forecast skill.
- Also, the number of regime transitions during the forecast cannot be used to explain an increase in forecast skill.

We find that a similar number of regime transitions for both reforecasts initialized during GL and ScTr. Therefore, we suggest that the number of regime transitions is not the main reason for an increased or decreased skill of the predictions. Furthermore, during both regimes at initialization a similar amount of forecasts shows persistence in the WR which is hypothesized to be easier to forecast than a large number of regime transitions.

The overall number of active WRs at each day during the forecasts is also not the main driver of an increase in forecasting skill since the number is similar between the two WRs present at initialization.

However, we find that a higher number of actual WR successions follow typical climatological patterns in case of the reforecasts with the higher skill. Therefore, we hypothesize that WR successions following typical climatological patterns and persistent WRs are easier to forecast and thus lead to an increase in forecast skill.

Interestingly, this finding does not hold for the predictions of the RFC-based postprocessing model which shows significant differences in the predictions of the occurrence of cold-wave days at a lead time of 28 days. The fact that the RFC-based postprocessing models rely more on large-scale flow predictors derived from reanalysis data than from reforecasts might be an explanation for the irrelevance of the actual WR successions for an increased skill in case of these predictions.

Besides the forecasts of (non-) cold-wave days, we specifically investigate the skill of predictions of days within cold waves in a next step. In case of Central Europe in winter, arguably a correct forecast of a cold-wave day is more important than the correct forecast of a non-cold-wave day. We investigate the 45 cold waves observed during the winters 2000/2001-2019/2020 separately. It is important to keep in mind here that this is a rather small number of events, such that general statements cannot be made. Due to the bi-weekly initialization of ECMWF's S2S reforecasts, only 20 start days, 15 end days and 106 intermediate days of the cold waves are included in this analysis.

We find that the timing of a cold-wave day inside a cold wave is not significantly influencing (a positive) forecast skill. Nevertheless, we find that intermediate days and days at the end of cold waves are generally better forecasted than days at the beginning of a cold wave.

Two thirds of the analyzed cold waves start when either the AR regime, the GL or the "No" regime is present at initialization. The median number of regime transitions during the cold waves itself is one and happens in the second half of the cold wave. Most often, the persistence of the AR regime or the transition from the GL to the "No" regime is found during a cold wave. During the AR and GL regime, cold polar air masses are advected towards Central Europe but during the "No" regime, warmer air masses are prevailing. It is important to kept in mind here that the change of air masses need a certain time to influence the surface and thus the response in surface temperatures is delayed.

Therefore, we consider additionally the WR successions in the week prior to the cold wave. However, we cannot find dominating WR successions here. Nevertheless, we find that the WR successions in the week before the cold waves start are following typical climatological patterns more often when initialized during the GL regime than the ScTr regime. Additionally, the occurrence of a persistent GL regime during the forecast is more often observed than a persisting ScTr regime. This might be another factor explaining the higher forecast skill of the mean bias corrected ECMWF's S2S reforecasts when initialized during the GL regime in comparison to an initialization when ScTr is present.

In a next step, we compare independent of the WR present at initialization the 10% best and worst predicted days within cold waves of the winters 2000/2001-2019/2020 and answer the following research question:

RQ 3.3 What are the differences in the WR successions before the best (worst) predicted days within cold waves?

- The most important difference is the presence of the EuBL regime in the days before the target date in case of the best predicted days within cold waves instead of the ScTr regime which is found in case of the worst predicted days within cold waves.
- The differences in the WR index values of EuBL between the best and worst predicted days within cold waves are significant.

We find that the most important difference in the WR successions before the best predicted days within cold waves is the WR present in the days before the target date. In case of the best predicted cold-wave days, it is the EuBL regime. In case of the worst predicted cold-wave days, the ScTr regime is present. The difference in the distributions of the WR index of the EuBL regime of the two forecast groups is thereby significant. Interestingly, according to Osman et al. (2023), EuBL has a skill horizon of roughly 13 days and the presence of EuBL in case of the 10% best predicted cold-wave days lies at the end of this time period.

At the example of two illustrative case studies we show that the discussed average results can also be found for individual cases. It has to be kept in mind though that Central European cold waves have a high variability in terms of WR successions as shown for the winters 2000/2001-2019/2020.

9. Conclusions

The aim of the research presented in this thesis is to improve forecasts of wintertime 2-meter temperatures and the occurrence of cold-wave days in Central Europe two to four weeks in advance. To that end, we combine physical knowledge with RF models to develop on the one hand side a cost-efficient alternative to NWP models and on the other hand side a postprocessing model for NWPs. The input variables are thereby based on meteorological knowledge. QRF models are trained to forecast the continuous 2-meter temperatures and RFC models to forecast the binary occurrence of cold-wave days. We evaluate our results for lead times of 14, 21 and 28 days against a climatological benchmark ensemble and compare them to the original and lead-time-dependent mean bias corrected ECMWF's S2S reforecasts. The considered evaluation period consists of the extended winter seasons (November to April) of the winters 2000/2001-2019-2020. The climatological benchmark ensemble is constructed using the timeseries of observed daily mean 2-meter temperatures from the E-OBS dataset of the 30 preceding winters of the evaluation period. It is evaluated against the daily mean 2-meter temperatures during the evaluation period from the E-OBS dataset which is taken as the ground truth in this research. To obtain the occurrence of cold-wave days, the temperature of each day is compared to a percentile-based threshold obtained from climatology. If three or more consecutive days show temperatures below this threshold, these days are considered as cold-wave days. In case of ensemble predictions, this is done for every ensemble member separately. The continuous temperature forecasts are evaluated using the CRPS, the binary cold-wave day forecasts using the BS. To subsequently assess the reliability of forecasts, the WRs present at the initialization and the WR successions during the forecast are taken into account.

Since the current study deals with forecasts with lead times at the subseasonal timescale, skill is generally low (as outlined in section 3.1) and the skill improvements found in this study accordingly small. Additionally, since the results are obtained for only one evaluation period of 20 consecutive winters, these cannot necessarily be generalized.

Our research is structured into three topical parts: the use of RFs as alternatives to NWP models (chapter 6), the use of RFs in a postprocessing sense as a complement to NWP models (chapter 7), and the use of WRs to assess forecast reliability (chapter 8). A detailed summary and discussion of the results of each research part is given in chapters 6 to 8. In the following, the main findings of the three research parts are briefly recapped and evaluated following the research questions defined in chapter 4. Afterwards, we highlight overarching results and their relevance for research and practical applications.

Conclusions of part 1: RF models as an alternative to NWP models

RQ 1.1 Can RF models using only reanalysis data as input provide skillful forecasts in comparison to a climatological benchmark?

In the 20-winter mean, QRF models provide skillful forecasts compared to a climatological benchmark ensemble at lead times of 14 , 21 and 28 days. In case of the RFC models, skillful forecasts are provided on lead times of 14 and 21 days. However, for both QRF and RFC models, the variability of forecast skill between winters is high.

RQ 1.2 Is the forecast skill equally good for periods with mild temperatures than periods with cold temperatures?

The skill of the individual models depends on the time of winter whereby the models perform better during long-lasting and mid-winter cold waves than at the margins of winters where temperatures are usually milder.

RQ 1.3 Which predictors are determining the models' predictions?

The most common predictor is $t850$ for warm periods for the analyzed QRF and RFC model. During the mid-winter cold wave in 2011/2012, common predictors of the analyzed QRF and RFC model which contribute to forecasting colder temperatures and cold-wave days are $t850$ and msl.

In this first part of our research we demonstrate that RF models based solely on reanalysis data provide skillful forecasts of wintertime 2-meter temperatures and the occurrence of cold-wave days in Central Europe compared to a climatological benchmark ensemble. Especially for decision makers who do not have access to NWPs, our developed models are often a better alternative to using climatological predictions (Fig.9.1). The value of our models is even enhanced when focusing on the prediction of unusually cold temperatures and cold-wave days. And most often it is the cold temperature extremes in winter which are highly relevant to be known in advance for decision-makers.

Conclusions of part 2: RF models as a complement to NWP models

RQ 2.1 Which postprocessing approaches perform particularly well in comparison to the original ECMWF's S2S reforecasts?

In case of the 2-meter temperature predictions, the best performing postprocessing model is the QRF using the ensemble information of the statistics of the reforecasted meteorological fields and the month in addition to the statistics of the reforecasted 2-meter temperature as input. In case of the predictions of the binary occurrence of cold-wave days, the lead-time-dependent mean bias corrected ECMWF's S2S reforecasts yield the best skill at lead times of 14 and 21 days. At a lead time of 28 days, the best performing RFC model is the one using the ensemble information of the first ten principle components of the reforcasted meteorological fields, the month, the ten principle components of the reanalyzed meteorological fields and the statistics of the reforecasted 2-meter temperature as input. The variation in mean skill scores between the different postprocessing models is small at all lead times for both forecasted properties.

RQ 2.2 In case of the RF-based postprocessing models, are predictors based on the reforecasted meteorological fields at the target date more important for the models' predictions than the reanalyzed fields at the initialization?

For forecasting both 2-meter temperatures and the occurrence of cold-wave days, the most important predictor is the reforecasted 2-meter temperature. In case of the predictions of the 2-meter temperatures, predictors based on the reforecasted fields at the target date are more important at all lead times than predictors based on the reanalyzed fields at initialization time. In case of the predictions of the occurrence of cold-wave days, besides the reforecasted 2-meter temperature, predictors based on the reanalyzed fields at initialization are equally important than predictors based on the reforecasted fields at the target date at a lead time of 14 days and more important at the lead times of 21 and 28 days.

In this second part of our research we show the strength of RF models in postprocessing forecasts of NWP models. Especially when considering the forecast of continuous 2-meter temperatures, RF-based postprocessing models yield the best predictions (Fig. 9.1). Therefore, our developed postprocessing models provide more skillful predictions for decision makers than ECMWF's S2S reforecast ensemble alone. Another advantage of the RF models is their cost-efficiency compared to more complex ML models such as CNNs. Our research therefore promotes the use of the still underrepresented simple ML models in both research and practical applications in order to spare computational resources and thus energy. Furthermore, we also show that in case of the prediction of the occurrence of cold-wave days at lead times of 14 and 21 days an even simpler and more cost-efficient postprocessing method, namely a lead-time-dependent mean bias correction, leads to the best results (Fig. 9.1). This finding advocates the consideration of basic methods, which spare computational resources, whenever possible.

Conclusions of part 3: WRs for assessing forecast reliability

RQ 3.1 Does the forecasting skill of ECMWF's S2S reforecasts and the RF-based models depend on the WRs at initialization?

Although the skill of all models depends on the WR present at initialization, only few differences in skill between the forecasts initialized during the various WRs are significant. The skill evolution across lead times varies between forecasts initialized during the different WRs but the deviations are most of the time not significant.

RQ 3.2 In how far can the WR succession during a forecast be linked to subseasonal forecast skill?

The WR successions during a forecast can be linked via their observed occurrences to subseasonal forecast skill. Forecasts showing a higher skill show more WR succession which follow typical climatological patterns than forecasts with a lower forecast skill. The overall number of active WRs per day of forecast, accumulated for all forecasts of the winters 2000/2001-2019/2020 initialized during the same WR, is not a main driver in increasing forecast skill. Also, the number of regime transitions during the forecast cannot be used to explain an increase in forecast skill.

RQ 3.3 What are the differences in the WR successions before the best (worst) predicted days within cold waves?

The most important difference is the presence of the EuBL regime in the days before the target date in case of the best predicted days within cold waves instead of the ScTr regime which is found in case of the worst predicted days within cold waves. The differences in the WR index values of EuBL between the best and worst predicted days within a cold wave are significant.

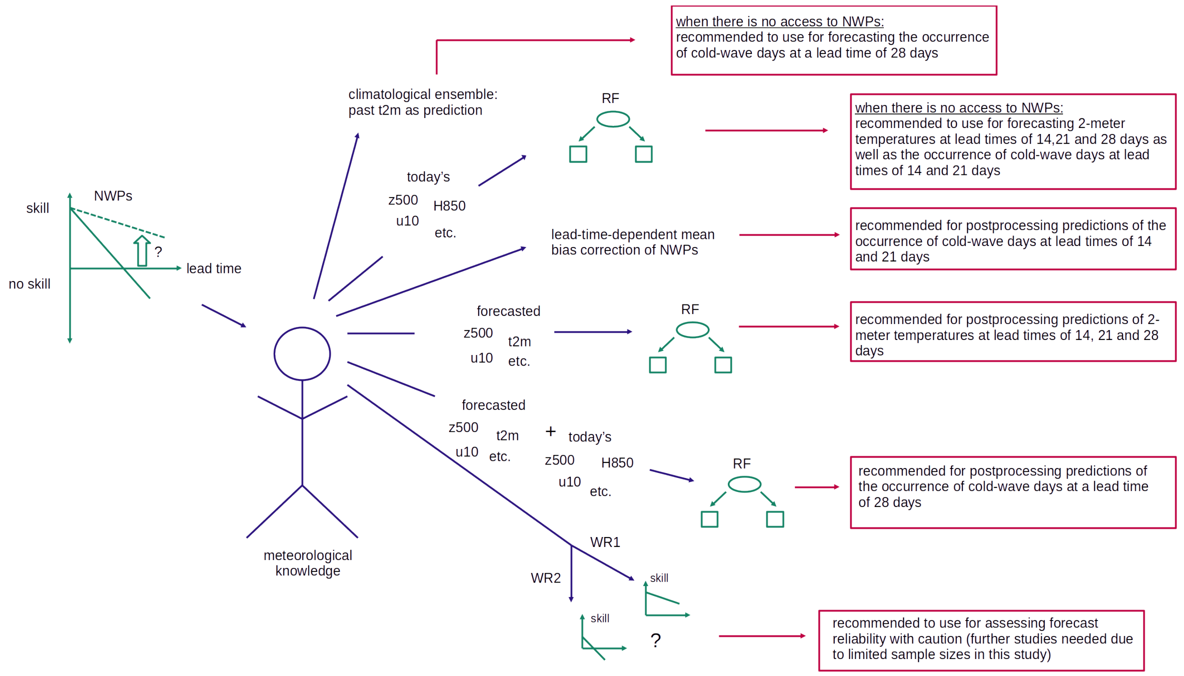
In this third part of our research, we demonstrate at a significantly relevant example the suitability of WRs for assessing forecast reliability. We show that depending on the WR at initialization, forecast skill varies, although the differences in forecasts with different WRs at initialization are mostly non-significant. Furthermore, we find that the presence of particular WRs in the days before a cold-wave day enhance the predictive skill of this day. These results can be used in combination with the expected forecast skill of the prediction of a WR (see e.g. Osman et al., 2023) to assess the reliability of subseasonal forecast of the occurrence of cold-wave days (Fig. 9.1). This is especially relevant for decision makers since the occurrence of cold-wave days in winter usually demands for a particular amount of planning.

Overarching conclusions

In the following, the main conclusions based on the three topical parts described above are summarized and set into context:

- postprocessing models outperform RF models based only on reanalysis data at all lead times and for both, the prediction of 2-meter temperatures and the occurrence of cold-wave days. Nevertheless, an advantage of RF models based only on reanalysis data is their cost efficiency.
- RF-based postprocessing models lead to an improvement of forecast skill of predictions of the 2-meter temperature and the occurrence of cold-wave days on lead times of 14, 21 and 28 days compared to a climatological benchmark ensemble and the original ECMWF's S2S reforecasts. However, in case of the predictions of the occurrence of cold-wave days at lead times of 14 and 21 days, the mean bias corrected ECMWF's S2S reforecasts outperform the RF-based postprocessing models and have furthermore the advantage of a higher cost efficiency. The differences between all postprocessing models at all lead times are small.
- the WR present at initialization can be used as a parameter to assess forecast reliability. Thereby, predictions featuring WR successions following typical climatological patterns during their forecast time are more skillful. However, the differences between the WR at initialization leading to the best (worst) forecast skill are most often non-significant and in the current research only a case study is analyzed.
- the WR present in the days before a cold-wave day occurs can be used to determine forecast reliability of days within cold waves. However, as shown in literature, the skill of the forecast of the WR itself is a limiting factor, especially on the longer lead times.

The presented research shows the potential of RF models in improving subseasonal forecasts of Central European mean wintertime 2-meter temperature as well as the occurrence of cold-wave days and the use of WRs in assessing the reliability of these forecasts. In combination, these can be used to aid decision makers in their planning, especially when cost- and time-consuming measures have to be taken. Detailed recommendations for the use of the different methods applied in this thesis are summarized on Fig. 9.1. Decision makers without access to NWP products can use RF models based on reanalysis data to obtain most often more skillful predictions compared to climatology. Decision makers with access to NWP products can use either the very cost-efficient method of a simple lead-time-dependent mean bias correction of NWPs or an RF-based postprocessing model. Depending on the forecast task at hand and the desired lead time, both yield skillful forecasts compared to a climatological benchmark ensemble and the original NWPs. Taking furthermore the WR present at initialization into account, the reliability of forecasts can be assessed since forecast skill depends on the WR present at initialization. However, due



when there is no access to NWPs: recommended to use for forecasting the occurrence of cold-wave days at a lead time of 28 days

when there is no access to NWPs: recommended to use for forecasting 2-meter temperatures at lead times of 14,21 and 28 days as well as the occurrence of cold-wave days at lead times of 14 and 21 days

recommended for postprocessing predictions of the occurrence of cold-wave days at lead times of 14 and 21 days

recommended for postprocessing predictions of 2-meter temperatures at lead times of 14, 21 and 28 days

recommended for postprocessing predictions of the occurrence of cold-wave days at a lead time of 28 days

recommended to use for assessing forecast reliability with caution (further studies needed due to limited sample sizes in this study)

Figure 9.1.: Recommendations for improving subseasonal forecast skill based on the research strategies presented in this thesis. The question of how to improve subseasonal forecast skill (left side) is tackled by six approaches utilizing meteorological knowledge (right side). These are on the one hand using a climatological ensemble consisting of past observations as predictions (first one from top) or forecasts of RF models based on reanalysis data (denoted as predictors of "today", second from top) as alternatives to NWP forecasts. On the other hand, a lead-time-dependent mean bias correction of NWP forecasts (third from top) and RF-based postprocessing models based only on NWP predictions (denoted as "forecasted" predictors, forth from top) or additionally on reanalysis data (fifth from top) are used as complements to NWP forecasts. Furthermore, it is analyzed if subseasonal forecast skill is dependent on the WR present at initialization. Recommendations for using the different approaches are given for the prediction of wintertime 2-meter temperatures and the occurrence of cold-wave days in Central Europe.

to the often not relevant significance, caution needs to be taken in order to not overestimate the reliability of forecasts.

Overall, this work contributes to improving weather predictions on the subseasonal timescale. Although our research was conducted specifically for the prediction of wintertime 2-meter temperatures and the occurrence of cold-wave days in Central Europe, the methods used are transferable to other regions and forecasting tasks. When doing so, the predictors need to be changed accordingly and the WRs chosen to be suitable for the target region. In order to enable researcher and users to tailor our developed models to their use, our code is publicly available. And last but not least, by using cost-efficient ML models and postprocessing methods, we make a contribution to the attempt of saving as many computational resources as possible and thus having a positive impact on sustainability.

10. Outlook

The presented research explores the possibilities of RF based models as alternative forecasting systems on the one hand and their use for postprocessing existing NWPs on the other hand. It gives an overview how these models can be applied to improve predictions of wintertime 2-meter temperatures and the occurrence of cold-wave days in Central Europe at lead times of 14, 21 and 28 days. Furthermore, it provides a first insight of how the reliability of forecasts can be assessed using WRs.

Using a combination of physical knowledge and cost-efficient ML models is a promising approach for improving subseasonal forecasts operationally. This way, computational resources are minimized which saves energy, time and money. Ideally, the chosen models are paired with an operational WR forecast to assess forecast reliability.

The developed RF models based solely on reanalysis data as input can be used as a fast, low-cost option to generate skillful subseasonal forecasts compared to climatology. Another advantage of these models compared to state-of-the-art NWPs on the subseasonal timescale is the possibility to create forecasts for any day (in contrast to e.g. bi-weekly initializations) and the greater ensemble size. The downside is that the RF models are not able to extrapolate and therefore to adjust to possible extremes not yet occurred in the recent climate. However, further studies can be conducted to test whether this disadvantage can be diminished by including climate projections as predictors. Additionally, further studies are needed to create an optimized version of these models for use in operational forecasting in terms of minimizing the number of predictors and still yielding skillful forecasts, especially when including climate projections.

The inclusion of climate projections to the RF-based postprocessing models might also be beneficial, especially when aiming for a use in the more distant future. RF-based postprocessing models are an alternative to parametric postprocessing approaches and yield better results in many cases than a lead-time-dependent mean bias correction of NWPs. Further research can be done by applying the RF-based postprocessing models to a multi-model ensemble of subseasonal forecasts which might be beneficial to further reduce possible model biases. Moreover, more advanced ML models, such as CNNs, can be applied to the forecast problem at hand. However, the more advanced ML models as well as multi-model approaches come with the disadvantage of needing more computational resources.

In this study, the seven WRs proposed by Grams et al. (2017) are used to assess the reliability of the postprocessed forecasts. When aiming to use the concept of utilizing WRs for assessing forecast reliability, it would be beneficial to conduct further studies with operationally forecasted WRs, such as the classical four WRs forecasted by ECMWF. As shown by Allen et al. (2021), it can also be beneficial to apply postprocessing dependent on the WR at initialization. This approach can be tested for the forecasts used in this study. Furthermore, the uncertainty in WR forecasts (e.g. Osman et al., 2023), depending on the WR at initialization, can be investigated further and potentially included in the assessment of forecast reliability.

Besides forecast improvement the predictions of our developed RF models can also be used for causal discovery. To do so, different XAI methods have to be applied in future studies and compared against each other. This can potentially reveal unknown drivers of e.g. severe cold waves.

Although we conduct the research specifically for the prediction of 2-meter temperature and the occurrence of cold-wave days in Central Europe, the used methods and concepts are easily transferable to other regions and forecasting task. However, the predictor selection and WR definition has to be suitable for the region and forecasting task. For example, when wanting to predict summertime hot days in North America, one should exchange predictors from the northern hemispheric stratosphere, which are mostly relevant in winter, by predictors such as soil moisture (Vijverberg et al., 2020). Furthermore, the Euro-Atlantic WRs should be replaced by North American WRs (Lee et al., 2023).

In summary, the presented research in this thesis demonstrates a cost- and time-efficient versatile method to improve subseasonal forecasts. Besides applications in the socio-economic sector to help decision makers with their planning, also further applications in research such as causal discovery are possible.

Acknowledgments

Pursuing a PhD turned out to be much more than just completing the academic training of becoming a meteorologist. Along the way, I grew as a scientist and a person. I am grateful for having had this opportunity.

This would have not been possible without Prof. Dr. Joaquim G. Pinto, my first advisor. Thankfully, he offered me the opportunity to combine research on a topic I was eager to learn more about with the continuation of working in the C8-team. Joaquim always contributed with helpful ideas to the design of the studies performed and the optimization of the written manuscripts.

Many thanks go to Prof. Dr. Peter Knippertz who agreed to be my second advisor. Even before this agreement has been made, Peter contributed considerably to the direction of my research by giving valuable feedback at my various talks in the institute's seminar.

One of the most valuable people who supported me scientifically is Dr. Sebastian Lerch. Although there were no official ties he included me warmly into his weekly group meetings where I learned a lot about the topic of machine learning which was totally new to me before starting my PhD. Sebastian supported me in every question I had to the topic which I am very grateful for. Furthermore, he provided me with the possibility to participate in the "WMO Prize Challenge to Improve Subseasonal to Seasonal Predictions Using Artificial Intelligence" in the team of Nina Horat which has given me a lot of insights to the state-of-the-art ML models used for subseasonal forecasting.

Many thanks go to Dr. Patrick Ludwig. Our meetings with Joaquim and Sebastian always improved the concepts of my studies and their description in our published papers.

Prof. Dr. Christian Grams and Fabian Mockert are gratefully acknowledged for providing the weather regime data which has been used extensively in my research.

Besides the scientific assistance I received, I am grateful for having had the opportunity to participate in teaching. This experience showed me my passion for sharing scientific knowledge. I want to thank Joaquim for trusting me to teach in the exercises accompanying the bachelor's "Climatology" and the master's "Methods of Data Analysis" lectures. Especially the re-organization of the latter together with

Nina Horat has been a great task during which I learned a lot.

Sharing knowledge among each other has also been a key factor of my decision to join the newly-established PhD council of the institute. It has been a great pleasure to work together with highly motivated fellow PhD students on topics relevant for daily work.

Reaching way beyond my time as a PhD student, I am especially grateful for the continuing support of my family and friends.

A. Appendix to chapter 5

Hyperparameters of QRF and RFC Models

In this appendix, the hyperparameter configurations of the used QRF models are shown in Tab.A.1. The used hyperparameters of the RFC models are summarized in Tab. A.2.

Table A.1.: Hyperparameters of QRF-models. These are the used hyperparameters of the "Ranger Random Forest Regression" model (Flynn, 2021).

hyperparameter	name in class "skranger.ensemble.RangerForestRegressor"	value
number of trees	$n_{estimators}$	**1000**
verbose logging	*verbose*	False
number of features to split on each node	*mtry*	**0**
feature importance	*importance*	’impurity’ for QRF_stat_all_s2s_ens_era5
minimal node size	*min_node_size*	**5**
maximal number of nodes per tree	*max_depth*	**0**
replacement for bootstrapping sampling	*replace*	True
fraction of training data for sampling	*sample_fraction*	None
save how often data points are in-bag	*keep_inbage*	False
list of the in-bag counts	*inbag*	None
rule to define splitting of data	*split_rule*	**variance**
number of random splits for the split rule	*number_random_splits*	1
significance threshold for maxstat split rule	*alpha*	0.5
Lower quantile of covariate distribution for maxstat	*minprop*	0.1
weight vector for selecting features for splitting	*split_select_weights*	None
use always certain features for splitting	*always_split_features*	None
list of categorical features	*categorical_features*	None
consider categorical features	*respect_categorical_features*	None
permutation feature importance	*scale_permutation_importance*	False
local permutation feature importance	*local_importance*	False
vector with regularization values for features	*regularization_factor*	None
consider depth in regularization	*regularization_usedepth*	False
Hold-out samples with case weight 0	*holdout*	False
enables quantile regression after fitting	*quantiles*	**True**
out-of-bag prediction error	*oob_error*	False
number of threads	*n_jobs*	**-1**
reduce speed of computation to save memory	*save_memory*	False
random seed value	*seed*	**42**
detailing the underlying decision trees	*enable_tree_details*	True for QRF_stat_all

Table A.2.: Hyperparameters of RFC-models. These are the used hyperparameters of the // "Ranger Random Forest Classifier" model (Flynn, 2021).

hyperparameter	name in class "skranger.ensemble.RangerForestClassifier"	value
number of trees	$n_{estimators}$	**1000**
verbose logging	*verbose*	False
number of features to split on each node	*mtry*	**0**
feature importance	*importance*	’impurity’ for RFC_stat_all_s2s_ens_era5
minimal node size	*min_node_size*	**5**
maximal number of nodes per tree	*max_depth*	**0**
replacement for bootstrapping sampling	*replace*	True
fraction of training data for sampling	*sample_fraction*	None
save how often data points are in-bag	*keep_inbage*	False
list of the in-bag counts	*inbag*	None
rule to define splitting of data	*split_rule*	**gini**
number of random splits for the split rule	*number_random_splits*	1
consider categorical features	*respect_categorical_features*	None
permutation feature importance	*scale_permutation_importance*	False
local permutation feature importance	*local_importance*	False
vector with regularization values for features	*regularization_factor*	None
consider depth in regularization	*regularization_usedepth*	False
Hold-out samples with case weight 0	*holdout*	False
out-of-bag prediction error	*oob_error*	False
number of threads	*n_jobs*	**-1**
reduce speed of computation to save memory	*save_memory*	False
random seed value	*seed*	**42**
detailing the underlying decision trees	*enable_tree_details*	True for RFC_stat_all

B. Appendix to chapter 6

Plots of CRPS and BS Difference of Winters at a Lead Time of 14 Days

In this appendix, the CRPS and BS differences of the respective QRF and RFC models in comparison to the climatological benchmark ensemble are shown for all analyzed winters except the winters 2011/2012 and 2013/2014 which are shown in section 6.3. A positive score difference denotes that the RF models provide more skillful forecasts at that day while a negative difference shows that the predictions of the climatological benchmark ensemble are more skillful. A difference of zero means that the forecasts of both models are equally skillful.

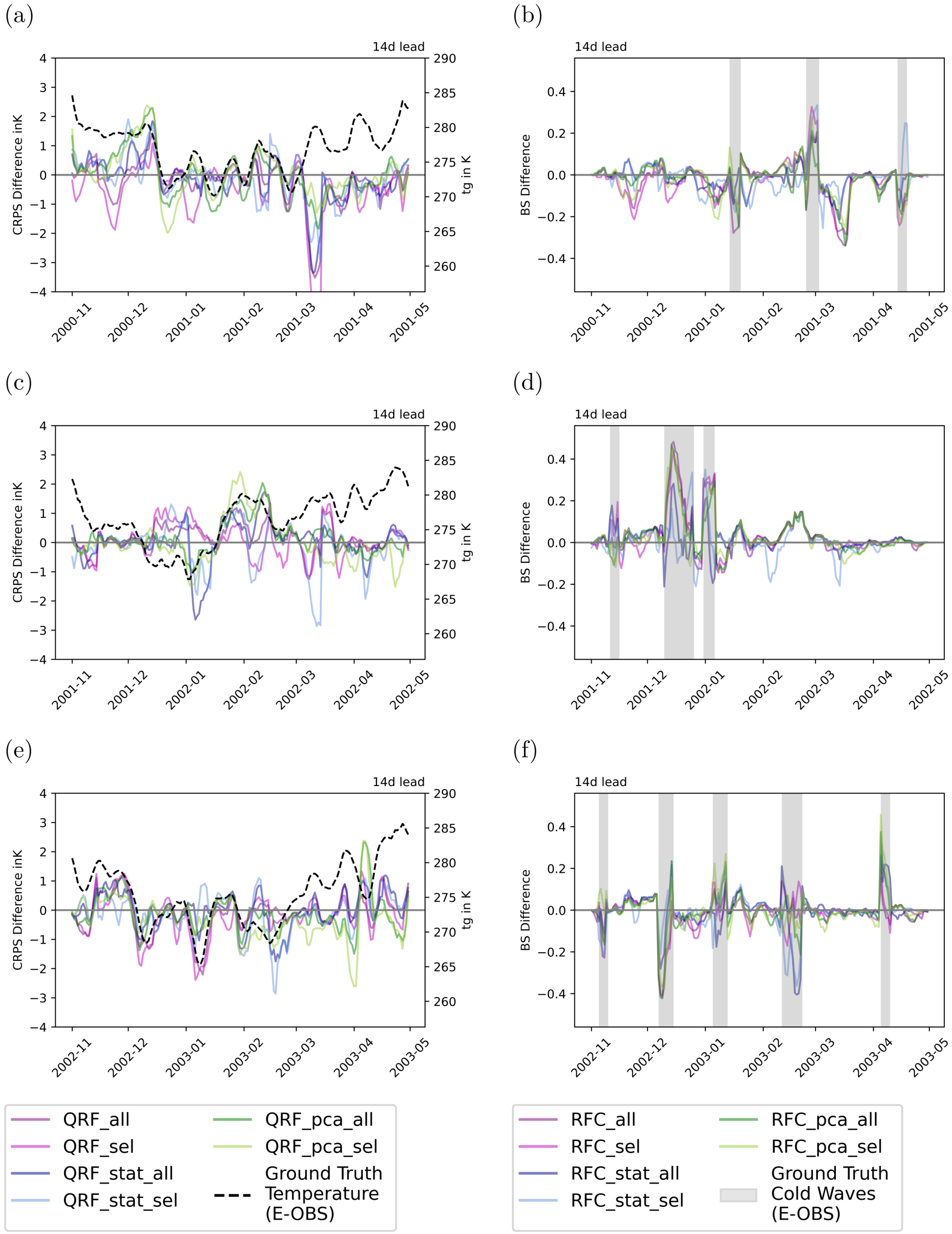


Figure B.1.: CRPS and BS difference at a lead time of 14 days for winters 2000/2001, 2001/2002 and 2002/2003. The CRPS differences are shown on the left column, the BS differences on the right column.

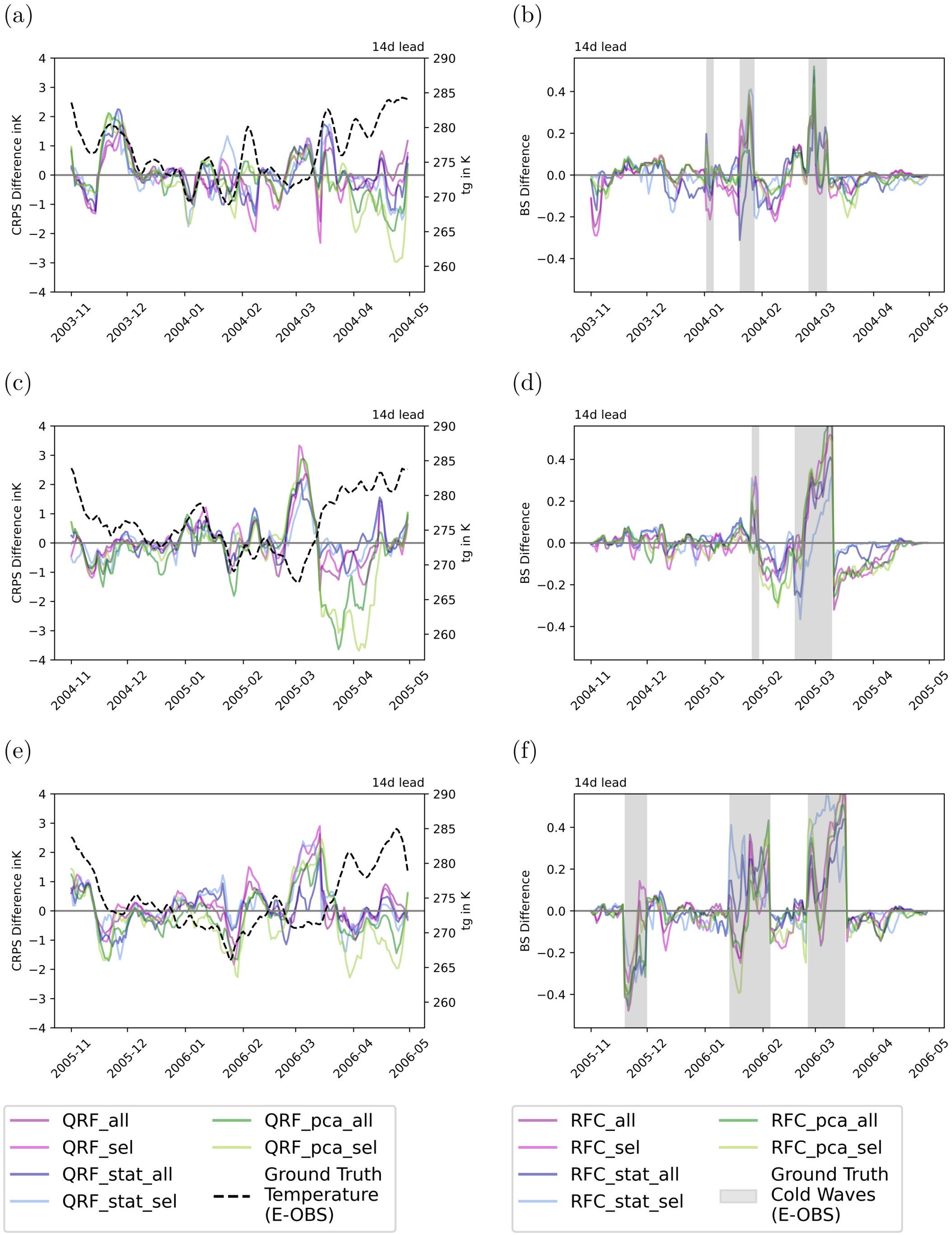


Figure B.2.: CRPS and BS difference at a lead time of 14 days for Winters 2003/2004, 2004/2005 and 2005/2006. The CRPS differences are shown on the left column, the BS differences on the right column.

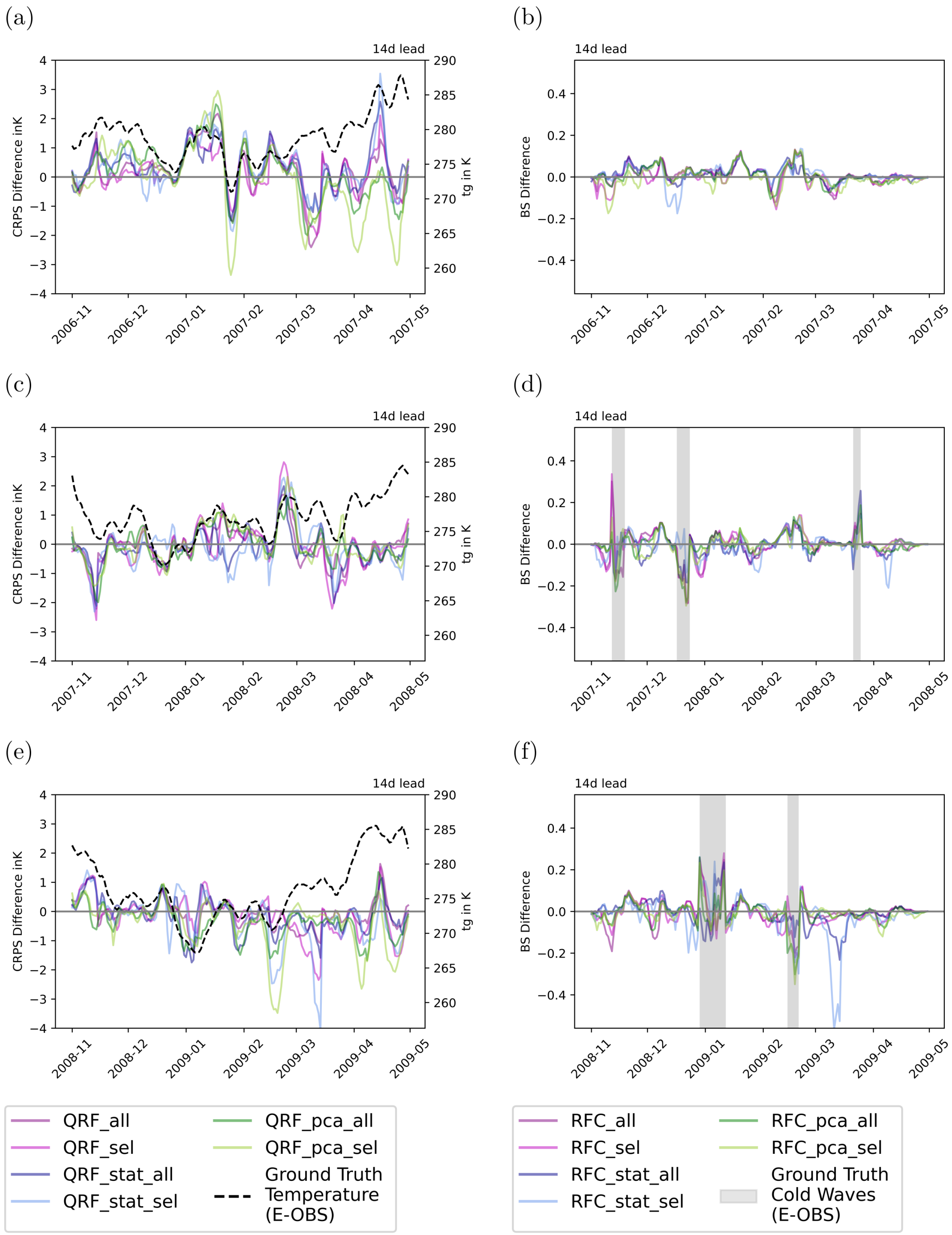


Figure B.3.: CRPS and BS difference at a lead time of 14 days for Winters 2006/2007, 2007/2008 and 2008/2009. The CRPS differences are shown on the left column, the BS differences on the right column.

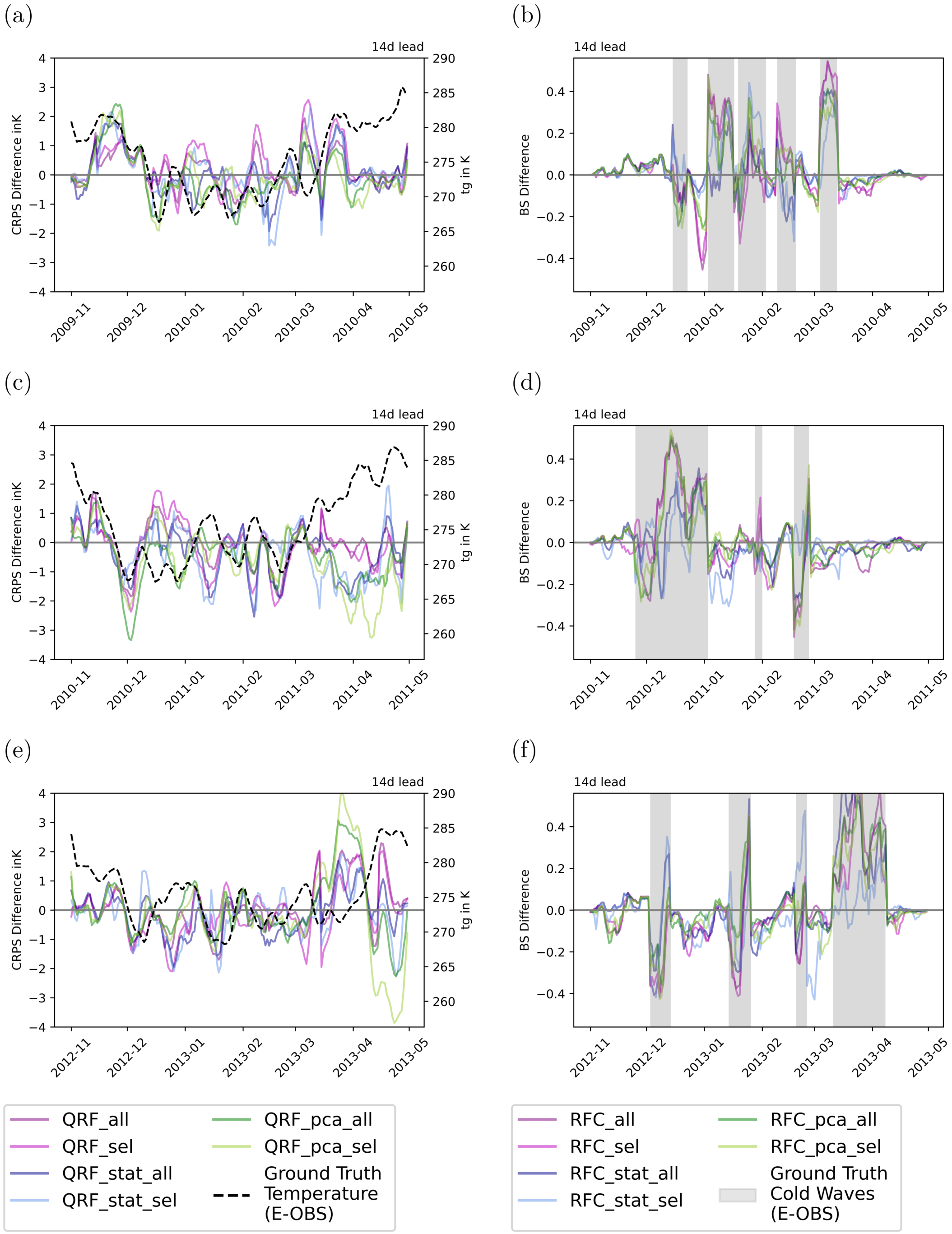


Figure B.4.: CRPS and BS difference at a lead time of 14 days for Winters 2009/2010, 2010/2011 and 2012/2013. The CRPS differences are shown on the left column, the BS differences on the right column.

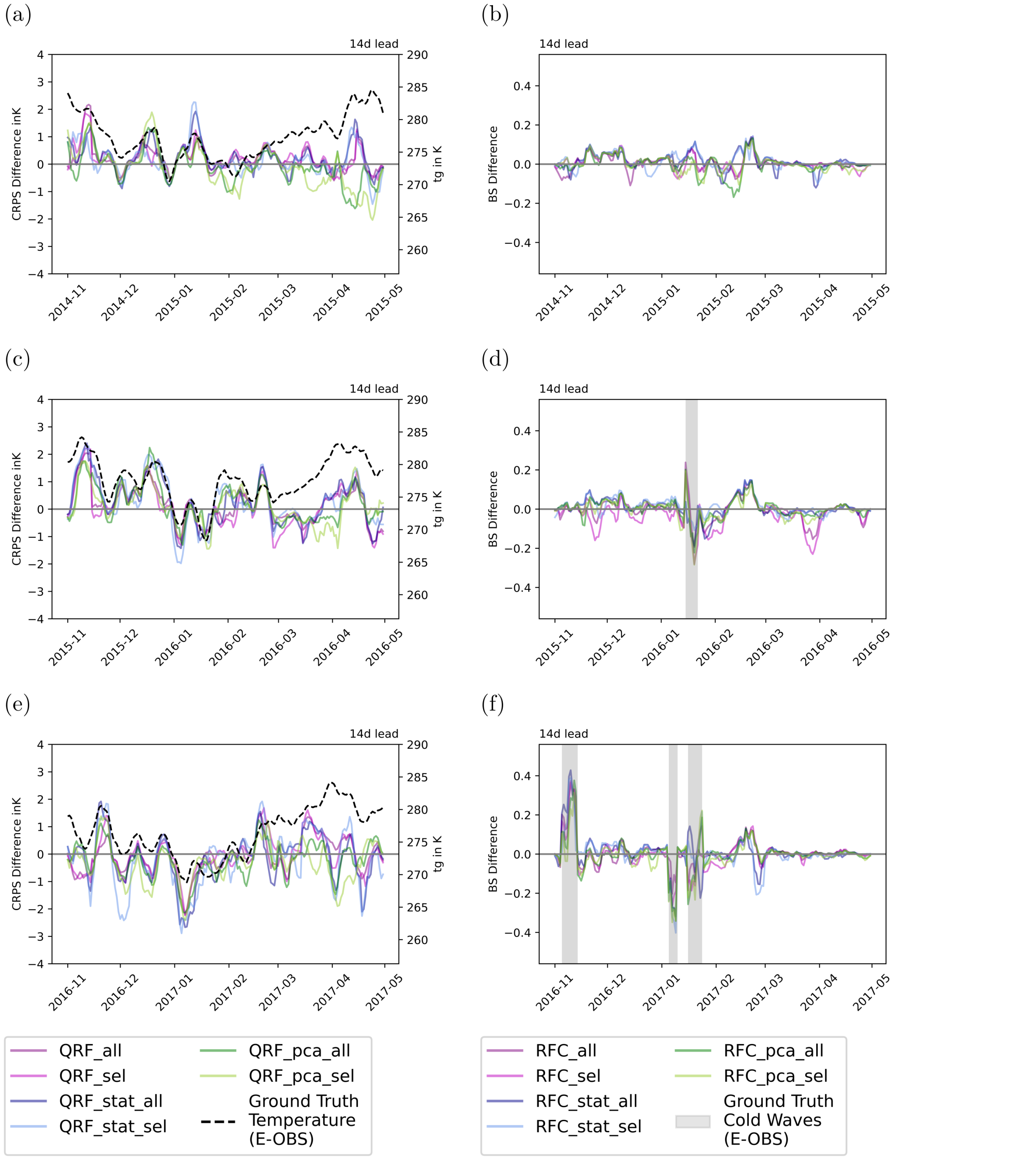


Figure B.5.: CRPS and BS difference at a lead time of 14 days for Winters 2014/2015, 2015/2016 and 2016/2017. The CRPS differences are shown on the left column, the BS differences on the right column.

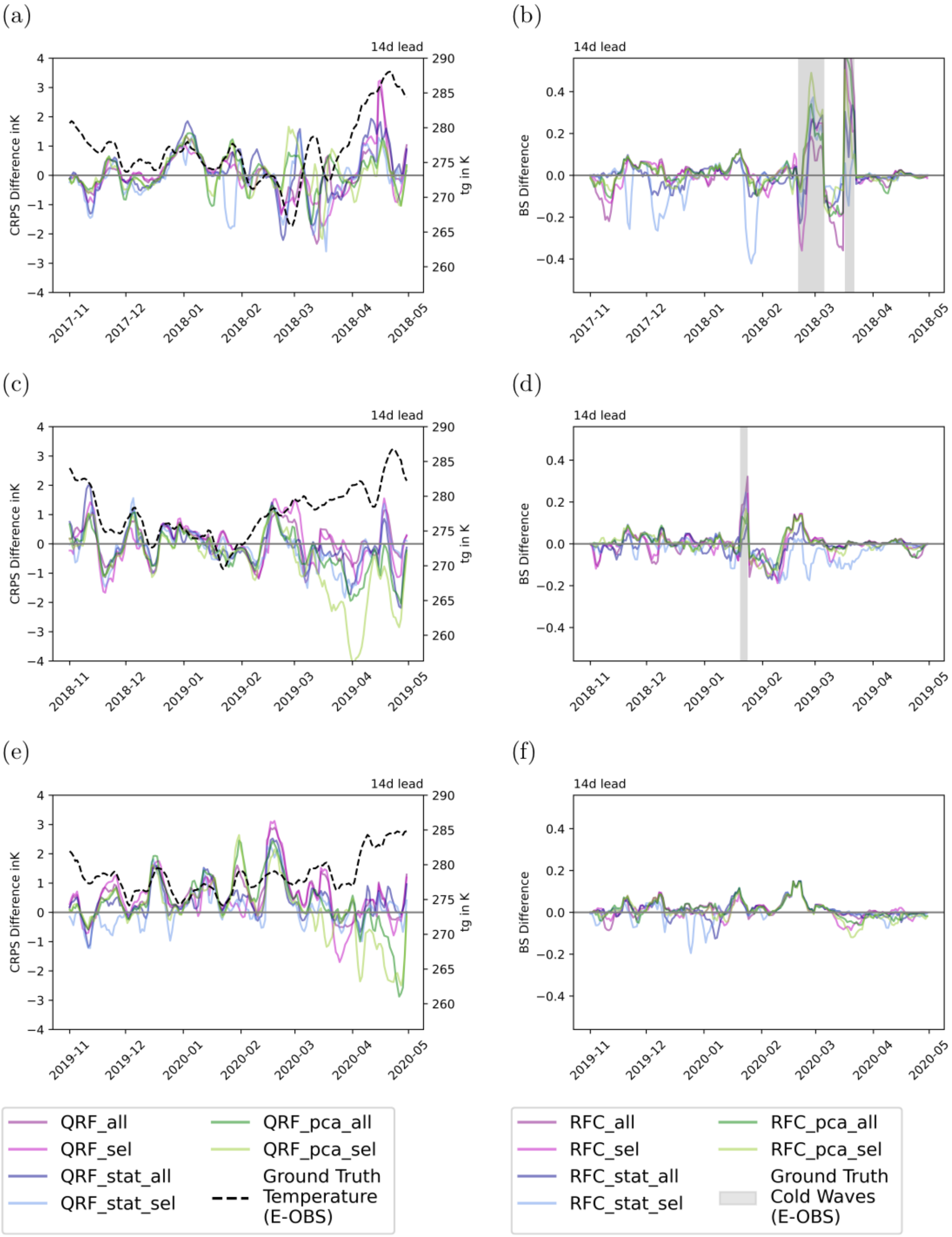


Figure B.6.: CRPS and BS difference at a lead time of 14 days for Winters 2017/2018, 2018/2019 and 2019/2020. The CRPS differences are shown on the left column, the BS differences on the right column.

C. Bibliography

Allen, S., G. R. Evans, P. Buchanan, and F. Kwasniok, 2021: Incorporating the North Atlantic Oscillation into the post-processing of MOGREPS-G wind speed forecasts. *Quarterly Journal of the Royal Meteorological Society*, **147 (735)**, 1403–1418, https://doi.org/10.1002/qj.3983.

Allen, S., C. A. T. Ferro, and F. Kwasniok, 2019: Regime-dependent statistical post-processing of ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, **145 (725)**, 3535–3552, https://doi.org/10.1002/qj.3638.

Baldwin, M. P., D. W. J. Thompson, E. F. Shuckburgh, W. A. Norton, and N. P. Gillett, 2003: Weather from the Stratosphere? *Science*, **301 (5631)**, 317–319, https://doi.org/10.1126/science.1085688.

Becker, E., and H. v. d. Dool, 2016: Probabilistic Seasonal Forecasts in the North American Multimodel Ensemble: A Baseline Skill Assessment. *Journal of Climate*, **29 (8)**, 3015–3026, https://doi.org/10.1175/JCLI-D-14-00862.1.

Bell, B., H. Hersbach, P. Berrisford, P. Dahlgren, A. Horányi, J. Muñoz Sabater, J. Nicolas, R. Radu et al., 2020a: ERA5 hourly data on pressure levels from 1950 to 1978 (preliminary version). *Copernicus Climate Change Service (C3S) Climate Data Store (CDS).*

Bell, B., H. Hersbach, P. Berrisford, P. Dahlgren, A. Horányi, J. Muñoz Sabater, J. Nicolas, R. Radu et al., 2020b: ERA5 hourly data on single levels from 1950 to 1978 (preliminary version). *Copernicus Climate Change Service (C3S) Climate Data Store (CDS).*

Benedict, J. J., S. Lee, and S. B. Feldstein, 2004: Synoptic View of the North Atlantic Oscillation. *Journal of the Atmospheric Sciences*, **61 (2)**, 121–144, https://doi.org/10.1175/1520-0469(2004)061<0121:SVOTNA>2.0.CO;2.

Breiman, L., 2001: Random Forests. *Machine Learning*, **45 (1)**, 5–32, https://doi.org/10.1023/A:1010933404324.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78 (1)**, 1–3, https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Butler, A. H., J. P. Sjoberg, D. J. Seidel, and K. H. Rosenlof, 2017: A sudden stratospheric warming compendium. *Earth System Science Data*, **9 (1)**, 63–76, https://doi.org/10.5194/essd-9-63-2017.

Bönisch, H., A. Engel, T. Birner, P. Hoor, D. W. Tarasick, and E. A. Ray, 2011: On the structural changes in the Brewer-Dobson circulation after 2000. *Atmospheric Chemistry and Physics*, **11 (8)**, 3937–3948, https://doi.org/10.5194/acp-11-3937-2011.

Büeler, D., L. Ferranti, L. Magnusson, J. F. Quinting, and C. M. Grams, 2021: Year-round sub-seasonal forecast skill for Atlantic–European weather regimes. *Quarterly Journal of the Royal Meteorological Society*, **147 (741)**, 4283–4309, https://doi.org/10.1002/qj.4178.

Charlton, A. J., and L. M. Polvani, 2007: A New Look at Stratospheric Sudden Warmings. Part I: Climatology and Modeling Benchmarks. *Journal of Climate*, **20 (3)**, 449–469, https://doi.org/10.1175/JCLI3996.1.

Charlton-Perez, A. J., L. Ferranti, and R. W. Lee, 2018: The influence of the stratospheric state on North Atlantic weather regimes. *Quarterly Journal of the Royal Meteorological Society*, **144 (713)**, 1140–1151, https://doi.org/10.1002/qj.3280.

Chen, L., X. Zhong, J. Wu, D. Chen, S. Xie, Q. Chao, C. Lin, Z. Hu et al., 2023: FuXi-S2S: An accurate machine learning model for global subseasonal forecasts. arXiv, URL `http://arxiv.org/abs/2312.09926`.

Clevert, D.-A., T. Unterthiner, and S. Hochreiter, 2016: Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). arXiv, https://doi.org/10.48550/arXiv.1511.07289.

Cornes, R. C., G. van der Schrier, E. J. M. van den Besselaar, and P. D. Jones, 2018: An ensemble version of the E-OBS temperature and precipitation data sets. *Journal of Geophysical Research: Atmospheres*, **123 (17)**, 9391–9409, https://doi.org/https://doi.org/10.1029/2017JD028200.

DelSole, T., L. Trenary, M. K. Tippett, and K. Pegion, 2017: Predictability of week-3-4 average temperature and precipitation over the contiguous united states. *Journal of Climate*, **30 (10)**, https://doi.org/10.1175/JCLI-D-16-0567.1.

Dirkson, A., B. Denis, W. J. Merryfield, K. A. Peterson, and S. Tietsche, 2022: Calibration of subseasonal sea-ice forecasts using ensemble model output statistics and observational uncertainty. *Quarterly Journal of the Royal Meteorological Society*, **148 (747)**, 2717–2741, https://doi.org/10.1002/qj.4332.

Domeisen, D., 2019: Estimating the Frequency of Sudden Stratospheric Warming Events From Surface Observations of the North Atlantic Oscillation. *Journal of Geophysical Research: Atmospheres*, **124 (6)**, 3180–3194, https://doi.org/10.1029/2018JD030077.

Domeisen, D. I. V., C. M. Grams, and L. Papritz, 2020: The role of North Atlantic–European weather regimes in the surface impact of sudden stratospheric warming events. *Weather and Climate Dynamics*, **1 (2)**, 373–388, https://doi.org/10.5194/wcd-1-373-2020.

DWD, 2012: Wetter und Klima - Deutscher Wetterdienst - Our services - Cold spell in Europe and Asia in late winter 2011/2012. URL `https://www.dwd.de/EN/ourservices/specialevents/temperature/20120315_cold_europe_february_2012_en.html`.

DWD, 2023: Wetter und Klima - Deutscher Wetterdienst - Glossar - B - Bergland und Tiefland. URL `https://www.dwd.de/DE/service/lexikon/Functions/glossar.html?lv2=100310&lv3=100414`.

ECMWF, 2019: *IFS Documentation CY46R1 - Part V: Ensemble Prediction System*. 5, https://doi.org/10.21957/38yug0cev.

Fan, Y., V. Krasnopolsky, H. v. d. Dool, C.-Y. Wu, and J. Gottschalck, 2023: Using Artificial Neural Networks to Improve CFS Week-3–4 Precipitation and 2-m Air Temperature Forecasts. *Weather and Forecasting*, **38 (5)**, 637–654, https://doi.org/10.1175/WAF-D-20-0014.1.

Ferranti, L., S. Corti, and M. Janousek, 2015: Flow-dependent verification of the ECMWF ensemble over the Euro-Atlantic sector. *Quarterly Journal of the Royal Meteorological Society*, **141 (688)**, 916–924, https://doi.org/10.1002/qj.2411.

Ferranti, L., L. Magnusson, F. Vitart, and D. Richardson, 2019: A new product to flag up the risk of cold spells in Europe weeks ahead. URL `https://www.ecmwf.int/en/newsletter/158/meteorology/new-product-flag-risk-cold-spells-europe-weeks-ahead`.

Flynn, C., 2021: skranger documentation [Python package]. URL `https://skranger.readthedocs.io/en/stable/`.

Gneiting, T., F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69 (2)**, 243–268, https://doi.org/10.1111/j.1467-9868.2007.00587.x.

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation. *Monthly Weather Review*, **133 (5)**, 1098–1118, https://doi.org/10.1175/MWR2904.1.

Goddard, L., W. E. Baethgen, H. Bhojwani, and A. W. Robertson, 2014: The International Research Institute for Climate & Society: why, what and how. *Earth Perspectives*, **1 (1)**, 10, https://doi.org/10.1186/2194-6434-1-10.

Grams, C., L. Ferranti, and L. Magnusson, 2020: How to make use of weather regimes in extended-range predictions for Europe. URL `https://www.ecmwf.int/en/newsletter/165/meteorology/how-make-use-weather-regimes-extended-range-predictions-europe`.

Grams, C. M., R. Beerli, S. Pfenninger, I. Staffell, and H. Wernli, 2017: Balancing Europe's wind-power output through spatial deployment informed by weather regimes. *Nature Climate Change*, **7 (8)**, 557–562, https://doi.org/10.1038/nclimate3338.

Hannachi, A., I. T. Jolliffe, and D. B. Stephenson, 2007: Empirical orthogonal functions and related techniques in atmospheric science: A review. *International Journal of Climatology*, **27 (9)**, 1119–1152, https://doi.org/10.1002/joc.1499.

Hannachi, A., D. M. Straus, C. L. E. Franzke, S. Corti, and T. Woollings, 2017: Low-frequency nonlinearity and regime behavior in the Northern Hemisphere extratropical atmosphere. *Reviews of Geophysics*, **55 (1)**, 199–234, https://doi.org/10.1002/2015RG000509.

Hansen, F., K. Matthes, and S. Wahl, 2016: Tropospheric QBO–ENSO Interactions and Differences between the Atlantic and Pacific. *Journal of Climate*, **29 (4)**, 1353–1368, https://doi.org/10.1175/JCLI-D-15-0164.1.

Hastie, T., R. Tibshirani, and J. Friedman, 2009: *The Elements of Statistical Learning*. Springer Series in Statistics, Springer, New York, NY, https://doi.org/10.1007/978-0-387-84858-7.

Hauser, S., F. Teubler, M. Riemer, P. Knippertz, and C. M. Grams, 2023: Towards a holistic understanding of blocked regime dynamics through a combination of complementary diagnostic perspectives. *Weather and Climate Dynamics*, **4 (2)**, 399–425, https://doi.org/10.5194/wcd-4-399-2023.

He, S., X. Li, T. DelSole, P. Ravikumar, and A. Banerjee, 2021: Sub-seasonal climate forecasting via machine learning: Challenges, analysis, and advances. *Proceedings of the AAAI Conference on Artificial Intelligence*, **35 (1)**, 169–177.

Hersbach, H., B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey et al., 2020a: The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, **146 (730)**, 1999–2049, https://doi.org/10.1002/qj.3803.

Hersbach, H., B. Bell, P. Berrisford, G. Biavati, A. Horányi, J. Muñoz Sabater, J. Nicolas, C. Peubey et al., 2020b: ERA5 hourly data on pressure levels from 1979 to present. *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*, https://doi.org/10.24381/cds.bd0915c6.

Hersbach, H., B. Bell, P. Berrisford, G. Biavati, A. Horányi, J. Muñoz Sabater, J. Nicolas, C. Peubey et al., 2020c: ERA5 hourly data on single levels from 1979 to present. *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*, https://doi.org/10.24381/cds.adbb2d47.

Horat, N., and S. Lerch, 2024: Deep Learning for Postprocessing Global Probabilistic Forecasts on Subseasonal Time Scales. *Monthly Weather Review*, **152 (3)**, 667–687, https://doi.org/10.1175/MWR-D-23-0150.1.

Hurrell, J., G. A. Meehl, D. Bader, T. L. Delworth, B. Kirtman, and B. Wielicki, 2009: A Unified Modeling Approach to Climate System Prediction. *Bulletin of the American Meteorological Society*, **90 (12)**, 1819–1832, https://doi.org/10.1175/2009BAMS2752.1.

Hyvärinen, O., T. K. Laurila, O. Räty, N. Korhonen, A. Vajda, and H. Gregow, 2021: Winter Subseasonal Wind Speed Forecasts for Finland from ECMWF. *Advances in Science and Research*, Vol. 18, 127–134, https://doi.org/10.5194/asr-18-127-2021.

Janitza, S., E. Celik, and A.-L. Boulesteix, 2018: A computationally fast variable importance test for random forests for high-dimensional data. *Advances in Data Analysis and Classification*, **12 (4)**, 885–915, https://doi.org/10.1007/s11634-016-0276-4.

Jiménez-Esteve, B., and D. I. V. Domeisen, 2018: The Tropospheric Pathway of the ENSO–North Atlantic Teleconnection. *Journal of Climate*, **31 (11)**, 4563–4584, https://doi.org/10.1175/JCLI-D-17-0716.1.

Karpechko, A. Y., A. Charlton-Perez, M. Balmaseda, N. Tyrrell, and F. Vitart, 2018: Predicting Sudden Stratospheric Warming 2018 and Its Climate Impacts With a Multimodel Ensemble. *Geophysical Research Letters*, **45 (24)**, 13,538–13,546, https://doi.org/10.1029/2018GL081091.

Kautz, L.-A., O. Martius, S. Pfahl, J. G. Pinto, A. M. Ramos, P. M. Sousa, and T. Woollings, 2022: Atmospheric blocking and weather extremes over the Euro - Atlantic sector - a review. *Weather and Climate Dynamics*, **3 (1)**, 305–336, https://doi.org/10.5194/wcd-3-305-2022.

Kautz, L.-A., I. Polichtchouk, T. Birner, H. Garny, and J. G. Pinto, 2020: Enhanced extended-range predictability of the 2018 late-winter Eurasian cold spell due to the stratosphere. *Quarterly Journal of the Royal Meteorological Society*, **146 (727)**, 1040–1055, https://doi.org/10.1002/qj.3724.

Kiefer, S. M., S. Lerch, P. Ludwig, and J. G. Pinto, 2023: Can Machine Learning Models Be a Suitable Tool for Predicting Central European Cold Winter Weather on Subseasonal to Seasonal Time Scales? *Artificial Intelligence for the Earth Systems*, **2 (4)**, e230 020, https://doi.org/10.1175/AIES-D-23-0020.1.

Kiefer, S. M., S. Lerch, P. Ludwig, and J. G. Pinto, 2024a: Random Forests' Postprocessing Capability of Enhancing Predictive Skill on Subseasonal Time Scales—A Flow-Dependent View on Central European Winter Weather. *Artificial Intelligence for the Earth Systems*, **3 (4)**, e240 014, https://doi.org/10.1175/AIES-D-24-0014.1.

Kiefer, S. M., P. Ludwig, S. Lerch, P. Knippertz, and J. G. Pinto, 2024b: The Role of Weather Regimes for Subseasonal Forecast Skill of Cold-Wave Days in Central Europe. *EGUsphere*, 1–26, https://doi.org/10.5194/egusphere-2024-2955.

Knippertz, P., U. Ulbrich, F. Marques, and J. Corte-Real, 2003: Decadal changes in the link between El Niño and springtime North Atlantic oscillation and European–North African rainfall. *International Journal of Climatology*, **23 (11)**, 1293–1311, https://doi.org/10.1002/joc.944.

Korhonen, N., O. Hyvärinen, M. Kämäräinen, D. S. Richardson, H. Järvinen, and H. Gregow, 2020: Adding value to extended-range forecasts in northern Europe by statistical post-processing using stratospheric observations. *Atmospheric Chemistry and Physics*, **20 (14)**, 8441–8451, https://doi.org/10.5194/acp-20-8441-2020.

Kurz, M., 1990: *Synoptische Meteorologie*. 2nd ed., No. 8, Leitfäden für die Ausbildung im Deutschen Wetterdienst, Selbstverlag des Deutschen Wetterdienstes, Offenbach am Main.

Lee, R. W., S. J. Woolnough, A. J. Charlton-Perez, and F. Vitart, 2019: ENSO Modulation of MJO Teleconnections to the North Atlantic and Europe. *Geophysical Research Letters*, **46 (22)**, 13 535–13 545, https://doi.org/10.1029/2019GL084683.

Lee, S. H., M. K. Tippett, and L. M. Polvani, 2023: A New Year-Round Weather Regime Classification for North America. *Journal of Climate*, **36 (20)**, 7091–7108, https://doi.org/10.1175/JCLI-D-23-0214.1.

Lhotka, O., and J. Kyselý, 2015: Characterizing joint effects of spatial extent, temperature magnitude and duration of heat waves and cold spells over Central Europe. *International Journal of Climatology*, **35 (7)**, 1232–1244, https://doi.org/10.1002/joc.4050.

Li, Y., D. Tian, and H. Medina, 2021: Multimodel Subseasonal Precipitation Forecasts over the Contiguous United States: Skill Assessment and Statistical Postprocessing. *Journal of Hydrometeorology*, **22 (10)**, 2581–2600, https://doi.org/10.1175/JHM-D-21-0029.1.

Limpasuvan, V., D. W. J. Thompson, and D. L. Hartmann, 2004: The Life Cycle of the Northern Hemisphere Sudden Stratospheric Warmings. *Journal of Climate*, **17 (13)**, 2584–2596, https://doi.org/10.1175/1520-0442(2004)017<2584:TLCOTN>2.0.CO;2.

Lin, H., G. Brunet, and J. Derome, 2009: An Observed Connection between the North Atlantic Oscillation and the Madden–Julian Oscillation. *Journal of Climate*, **22 (2)**, 364–380, https://doi.org/10.1175/2008JCLI2515.1.

Lundberg, S., 2018: SHAP documentation [Python package]. URL `https://shap.readthedocs.io/en/latest/`.

Lundberg, S. M., G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb et al., 2020: From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, **2 (1)**, 2522–5839.

Mariotti, A., C. Baggett, E. A. Barnes, E. Becker, A. Butler, D. C. Collins, P. A. Dirmeyer, L. Ferranti et al., 2020: Windows of Opportunity for Skillful Forecasts Subseasonal to Seasonal and Beyond. *Bulletin of the American Meteorological Society*, **101 (5)**, E608–E625, https://doi.org/10.1175/BAMS-D-18-0326.1.

Mayer, K. J., and E. A. Barnes, 2021: Subseasonal forecasts of opportunity identified by an explainable neural network. *Geophysical Research Letters*, **48 (10)**, https://doi.org/10.1029/2020GL092092.

Meehl, G. A., R. Lukas, G. N. Kiladis, K. M. Weickmann, A. J. Matthews, and M. Wheeler, 2001: A conceptual framework for time and space scale interactions in the climate system. *Climate Dynamics*, **17 (10)**, 753–775, https://doi.org/10.1007/s003820000143.

Meinshausen, N., 2006: Quantile Regression Forests. *Journal of Machine Learning Research*, **7**, 983 – 999.

Met Office, 2010 - 2015: *Cartopy: a cartographic Python library with a Matplotlib interface*. Exeter, Devon, URL `https://scitools.org.uk/cartopy`.

Mladek, R., 2024: ECMWF Model - S2S - ECMWF Confluence Wiki. URL `https://confluence.ecmwf.int/display/S2S/ECMWF+Model`.

Mockert, F., C. M. Grams, S. Lerch, M. Osman, and J. Quinting, 2024: Multivariate post-processing of probabilistic sub-seasonal weather regime forecasts. *Quarterly Journal of the Royal Meteorological Society*, 1–17, https://doi.org/10.1002/qj.4840.

Molnar, C., 2022: *Interpretable Machine Learning*. URL `https://christophm.github.io/interpretable-ml-book/`.

Monhart, S., C. Spirig, J. Bhend, K. Bogner, C. Schär, and M. A. Liniger, 2018: Skill of Subseasonal Forecasts in Europe: Effect of Bias Correction and Downscaling Using Surface Observations. *Journal of Geophysical Research: Atmospheres*, **123 (15)**, 7999–8016, https://doi.org/10.1029/2017JD027923.

Moser, B. K., and G. R. Stevens, 1992: Homogeneity of Variance in the Two-Sample Means Test. *The American Statistician*, **46 (1)**, 19–21, https://doi.org/10.2307/2684403.

Nguyen, T., J. Brandstetter, A. Kapoor, J. K. Gupta, and A. Grover, 2023: ClimaX: A foundation model for weather and climate. arXiv, URL `http://arxiv.org/abs/2301.10343`.

Osman, M., R. Beerli, D. Büeler, and C. M. Grams, 2023: Multi-model assessment of sub-seasonal predictive skill for year-round Atlantic–European weather regimes. *Quarterly Journal of the Royal Meteorological Society*, **149 (755)**, 2386–2408, https://doi.org/10.1002/qj.4512.

Owens, R., and T. Hewson, 2018: ECMWF Forecast User Guide. https://doi.org/10.21957/M1CS7H.

Pinto, J. G., I. Gómara, G. Masato, H. F. Dacre, T. Woollings, and R. Caballero, 2014: Large-scale dynamics associated with clustering of extratropical cyclones affecting Western Europe. *Journal of Geophysical Research: Atmospheres*, **119 (24)**, 13,704–13,719, https://doi.org/10.1002/2014JD022305.

Pinto, J. G., and C. C. Raible, 2012: Past and recent changes in the North Atlantic oscillation. *WIREs Climate Change*, **3 (1)**, 79–90, https://doi.org/10.1002/wcc.150.

Proedrou, E., K. Hocke, and P. Wurz, 2016: The middle atmospheric circulation of a tidally locked Earth-like planet and the role of the sea surface temperature. *Progress in Earth and Planetary Science*, **3 (1)**, 22, https://doi.org/10.1186/s40645-016-0098-1.

Richardson, D., H. J. Fowler, C. G. Kilsby, R. Neal, and R. Dankers, 2020: Improving sub-seasonal forecast skill of meteorological drought: a weather pattern approach. *Natural Hazards and Earth System Sciences*, **20 (1)**, 107–124, https://doi.org/10.5194/nhess-20-107-2020.

Robertson, A. W., N. Vigaud, J. Yuan, and M. K. Tippett, 2020: Toward Identifying Subseasonal Forecasts of Opportunity Using North American Weather Regimes. *Monthly Weather Review*, **148 (5)**, 1861–1875, https://doi.org/10.1175/MWR-D-19-0285.1.

Scheuerer, M., M. B. Switanek, R. P. Worsnop, and T. M. Hamill, 2020: Using Artificial Neural Networks for Generating Probabilistic Subseasonal Precipitation Forecasts over California. *Monthly Weather Review*, **148 (8)**, 3489–3506, https://doi.org/10.1175/MWR-D-20-0096.1.

scikit-learn Developers, 2024: PCA. URL `https://scikit-learn/stable/modules/generated/sklearn.decomposition.PCA.html`.

Silini, R., S. Lerch, N. Mastrantonas, H. Kantz, M. Barreiro, and C. Masoller, 2022: Improving the prediction of the Madden–Julian Oscillation of the ECMWF model by post-processing. *Earth System Dynamics*, **13 (3)**, 1157–1165, https://doi.org/10.5194/esd-13-1157-2022.

Smid, M., S. Russo, A. Costa, C. Granell, and E. Pebesma, 2019: Ranking European capitals by exposure to heat waves and cold waves. *Urban Climate*, **27**, 388–402, https://doi.org/10.1016/j.uclim.2018.12.010.

Spiridonov, V., and M. Ćurić, 2021: *Fundamentals of Meteorology*. Springer International Publishing, Cham, https://doi.org/10.1007/978-3-030-52655-9.

Suthaharan, S., 2016: Decision Tree learning. *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*, S. Suthaharan, Ed., Integrated Series in Information Systems, Springer US, Boston, MA, 237–269, https://doi.org/10.1007/978-1-4899-7641-3_10.

Taillardat, M., O. Mestre, M. Zamo, and P. Naveau, 2016: Calibrated ensemble forecasts using Quantile Regression Forests and Ensemble Model Output Statistics. *Monthly Weather Review*, **144 (6)**, 2375–2393, https://doi.org/10.1175/MWR-D-15-0260.1.

Tamarin-Brodsky, T., and N. Harnik, 2024: The relation between Rossby wave-breaking events and low-level weather systems. *Weather and Climate Dynamics*, **5 (1)**, 87–108, https://doi.org/10.5194/wcd-5-87-2024.

The SciPy community, 2024: SciPy API — SciPy v1.13.0 Manual. URL `https://docs.scipy.org/doc/scipy/reference/index.html`.

Tomczyk, A. M., E. Bednorz, and A. Sulikowska, 2019: Cold spells in Poland and Germany and their circulation conditions. *International Journal of Climatology*, **39 (10)**, https://doi.org/10.1002/joc.6054.

Toms, B. A., E. A. Barnes, and I. Ebert-Uphoff, 2020: Physically Interpretable Neural Networks for the Geosciences: Applications to Earth System Variability. *Journal of Advances in Modeling Earth Systems*, **12 (9)**, e2019MS002 002, https://doi.org/10.1029/2019MS002002.

Tripathi, O. P., M. Baldwin, A. Charlton-Perez, M. Charron, S. D. Eckermann, E. Gerber, R. G. Harrison, D. R. Jackson et al., 2015: The predictability of the extratropical stratosphere on monthly time-scales and its impact on the skill of tropospheric forecasts. *Quarterly Journal of the Royal Meteorological Society*, **141 (689)**, 987–1003, https://doi.org/10.1002/qj.2432.

Vallis, G. K., 2017: *Atmospheric and Oceanic Fluid Dynamics: Fundamentals and Large-Scale Circulation*. 2nd ed., Cambridge University Press, Cambridge, https://doi.org/10.1017/9781107588417.

van Straaten, C., K. Whan, D. Coumou, B. van den Hurk, and M. Schmeits, 2020: The influence of aggregation and statistical post-processing on the subseasonal predictability of European temperatures. *Quarterly Journal of the Royal Meteorological Society*, **146 (731)**, 2654–2670, https://doi.org/10.1002/qj.3810.

van Straaten, C., K. Whan, D. Coumou, B. van den Hurk, and M. Schmeits, 2022: Using explainable machine learning forecasts to discover subseasonal drivers of high summer temperatures in Western and Central Europe. *Monthly Weather Review*, **150 (5)**, 1115–1134, https://doi.org/10.1175/MWR-D-21-0201.1.

Vannitsem, S., J. B. Bremnes, J. Demaeyer, G. R. Evans, J. Flowerdew, S. Hemri, S. Lerch, N. Roberts et al., 2021: Statistical Postprocessing for Weather Forecasts: Review, Challenges, and Avenues in a Big Data World. *Bulletin of the American Meteorological Society*, **102 (3)**, E681–E699, https://doi.org/10.1175/BAMS-D-19-0308.1.

Vigaud, N., M. K. Tippett, and A. W. Robertson, 2018: Probabilistic skill of subseasonal precipitation forecasts for the East Africa - West Asia Sector during September - May. *Weather and Forecasting*, **33 (6)**, 1513–1532, https://doi.org/10.1175/WAF-D-18-0074.1.

Vijverberg, S., M. Schmeits, K. van der Wiel, and D. Coumou, 2020: Subseasonal statistical forecasts of Eastern U.S. hot temperature events. *Monthly Weather Review*, **148 (12)**, 4799–4822, https://doi.org/10.1175/MWR-D-19-0409.1.

Vitart, F., and A. W. Robertson, 2018: The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events. *npj Climate and Atmospheric Science*, **1 (1)**, 3, https://doi.org/10.1038/s41612-018-0013-0.

Vitart, F., C. Ardilouze, A. Bonet, A. Brookshaw, M. Chen, C. Codorean, M. Déqué, L. Ferranti et al., 2017: The Subseasonal to Seasonal (S2S) prediction project database. *Bulletin of the American Meteorological Society*, **98 (1)**, 163–173, https://doi.org/10.1175/BAMS-D-16-0017.1.

Vitart, F., A. W. Robertson, A. Spring, F. Pinault, R. Roškar, W. Cao, S. Bech, A. Bienkowski et al., 2022: Outcomes of the WMO Prize Challenge to Improve Subseasonal to Seasonal Predictions Using Artificial Intelligence. *Bulletin of the American Meteorological Society*, **103 (12)**, E2878–E2886, https://doi.org/10.1175/BAMS-D-22-0046.1.

Wallace, J. M., and D. S. Gutzler, 1981: Teleconnections in the Geopotential Height Field during the Northern Hemisphere Winter. *Monthly Weather Review*, **109 (4)**, 784–812, https://doi.org/10.1175/1520-0493(1981)109<0784:TITGHF>2.0.CO;2.

Welch, B. L., 1947: The generalization of 'Student's' problem when several different population variances are involved. *Biometrika*, **34 (1-2)**, 28–35, https://doi.org/10.1093/biomet/34.1-2.28.

Weyn, J. A., D. R. Durran, R. Caruana, and N. Cresswell-Clay, 2021: Sub-Seasonal Forecasting With a Large Ensemble of Deep-Learning Weather Prediction Models. *Journal of Advances in Modeling Earth Systems*, **13 (7)**, e2021MS002 502, https://doi.org/10.1029/2021MS002502.

White, C. J., H. Carlsen, A. W. Robertson, R. J. Klein, J. K. Lazo, A. Kumar, F. Vitart, E. Coughlan de Perez et al., 2017: Potential applications of subseasonal-to-seasonal (S2S) predictions. *Meteorological Applications*, **24 (3)**, 315–325, https://doi.org/10.1002/met.1654.

Williams, N. C., A. A. Scaife, and J. A. Screen, 2023: Underpredicted ENSO Teleconnections in Seasonal Forecasts. *Geophysical Research Letters*, **50 (5)**, e2022GL101 689, https://doi.org/10.1029/2022GL101689.

WMO, 2020: WMO climatological normals. URL `https://community.wmo.int/wmo-climatological-normals`.

WMO, 2021: *Guidelines on Ensemble Prediction System Postprocessing*, Vol. 978-92-63-11254-5. 2021st ed., WMO, Geneva.

WMO, 2024: WMO Lead Centre for Sub-seasonal forecast multi-model ensemble. URL `https://charts-dev.ecmwf.int/wmo/charts`.

Wright, M. N., and A. Ziegler, 2017: ranger : A Fast Implementation of Random Forests for High Dimensional Data in *C++* and *R*. *Journal of Statistical Software*, **77 (1)**, https://doi.org/10.18637/jss.v077.i01.

Zhang, F., Y. Q. Sun, L. Magnusson, R. Buizza, S.-J. Lin, J.-H. Chen, and K. Emanuel, 2019: What Is the Predictability Limit of Midlatitude Weather? *Journal of the Atmospheric Sciences*, **76 (4)**, 1077–1091, https://doi.org/10.1175/JAS-D-18-0269.1.