Can Machine Learning Models Be a Suitable Tool for Predicting Central European Cold Winter Weather on Subseasonal to Seasonal Time Scales?

SELINA M. KIEFER, a SEBASTIAN LERCH, b PATRICK LUDWIG, a AND JOAQUIM G. PINTO a

a Institute of Meteorology and Climate Research–Tropospheric Research, Karlsruhe Institute of Technology, Karlsruhe, Germany
b Institute of Economics, Karlsruhe Institute of Technology, Karlsruhe, Germany

ABSTRACT: Skillful weather prediction on subseasonal to seasonal time scales is crucial for many socioeconomic ventures. But forecasting, especially extremes, on these time scales is very challenging because the information from initial conditions is gradually lost. Therefore, data-driven methods are discussed as an alternative to numerical weather prediction models. Here, quantile regression forests (QRFs) and random forest classifiers (RFCs) are used for probabilistic forecasting of central European mean wintertime 2-m temperatures and cold wave days at lead times of 14, 21, and 28 days. ERA5 reanalysis meteorological predictors are used as input data for the machine learning models. For the winters of 2000/01–2019/20, the predictions are compared with a climatological ensemble obtained from E-OBS 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. We find skill at lead times up to 28 days in the 20-winter mean and for individual winters. Case studies show that all used machine learning models are able to learn patterns in the data beyond climatology. A more detailed analysis using Shapley additive explanations suggests that both random forest (RF)-based models are able to learn physically known relationships in the data. This underlines that RF-based data-driven models can be a suitable tool for forecasting central European mean wintertime 2-m temperatures and the occurrence of cold wave days.

SIGNIFICANCE STATEMENT: Because of the chaotic nature of weather, it is very complicated to make predictions with traditional numerical methods 2–4 weeks in advance. Therefore, we use alternative, interpretable methods that “learn” to find statistically relevant patterns in meteorological data that can be used for forecasting central European mean surface wintertime temperatures and cold wave days. These methods are part of the so-called machine learning methods that do not rely on the traditional numerical equations anymore. We test our methods for 20 winters between 2000/01 and 2019/20 against a static weather prediction consisting of the past 30 winters. For single winters and in a mean over the 20 predicted winters, we find improved predictions up to 4 weeks in advance.

KEYWORDS: Europe; Cold-air surges; Cool season; Forecasting; Subseasonal variability; Machine learning

1. Introduction

It is of great interest for society and various sectors of the economy to have a skillful prediction of wintertime temperatures and especially of potentially severe cold waves in advance. In fact, many decisions concerning, for example, agriculture, food and energy security, health care, and transportation are made at a longer time range, from 2 weeks to 3 months in advance, also called the subseasonal to seasonal (S2S) time scale (DelSole et al. 2017). But, as Vitart and Robertson (2018) state, forecasting on S2S time scales is very challenging. While the skill of short- and medium-range forecasts is governed by the initial conditions of the prediction, for example, the midtropospheric circulation state, these initial conditions are a minor contribution to the forecast skill on the S2S time range. The same applies to the boundary conditions of meteorological forecasts like sea surface temperatures, which contribute a lot to the predictive skill at seasonal and longer lead times. Instead, on the S2S time scales, teleconnections are important for predictability, which are still a subject of ongoing research. Known teleconnections influencing central European surface weather in winter are, for example, the stratospheric polar vortex (Domeisen et al. 2020) and the location and intensity of the jet stream over the North Atlantic Ocean (Hurrell et al. 2003; Kautz et al. 2020, 2022).

Although multiple databases offer S2S forecasts and reforecasts, the number of predictions is still scarce (Vigaud et al. 2018). Forecasts are typically initialized twice a week and contain fewer ensemble members than do short- to medium-range and seasonal forecasts (Vitart et al. 2017; Vigaud et al. 2018). Multimodel approaches that increase the ensemble size are not trivial to define, since the different operational weather centers initialize their forecasts at different times, with different initial and boundary conditions and different resolutions in time and space (Vitart et al. 2017).

A way to overcome these issues is the use of machine learning (ML) methods to directly forecast meteorological variables on the S2S time scales (Cohen et al. 2019). ML models do not depend on any initial or boundary conditions and can
be started at any given date. Furthermore, there is the possibility that an ML model “learns” data connections that have not yet been fully explored. Another advantage is that the ensemble size of the prediction is not limited, given that some ML models can be configured to produce ensemble forecasts. The input data for the ML models can be any type of meteorologically sensible data, for example, the output of numerical weather models (Whan and Schmeits 2018), climate indices like the North Atlantic Oscillation index (Baker et al. 2018), or observational data from satellites or radar measurements (McGovern et al. 2014). For the ground truth needed for forecast verification, the same data types are possible, depending on the exact forecast issue (e.g., Whan and Schmeits 2018; Vijverberg et al. 2020).

ML models can, as most traditional numerical weather prediction (NWP) models, be computed globally, as in Weyn et al. (2021). In their study, they generate 320 ensemble members using deep learning models to forecast different atmospheric variables globally on S2S time scales, for example, the 2-m temperature. That said, many state-of-the-art ML models for S2S prediction are designed on local or regional scales. For example, He et al. (2021) obtain skillful ML forecasts of 2-m temperature and precipitation over the contiguous United States on subseasonal time scales with respect to a climatological forecast. As input for their ML models, they use climate variables from the land, ocean, and atmosphere. Additionally, He et al. (2021) use Shapley additive explanations (SHAP; Lundberg et al. 2020), which show the contribution of each predictor to a model’s prediction in order to explain the predictions of their used ML models. Mayer and Barnes (2021) use only two predictor variables for training their ML model. These are the outgoing longwave radiation measured by satellite and the 500-hPa geopotential height field from reanalysis data. With these, their artificial neural network is able to create skillful binary predictions of the sign of the North Atlantic 500-hPa geopotential height anomalies 3 weeks in advance, which is a relevant source of predictability also for European weather. A direct prediction of European temperatures with ML models on S2S time scales is performed by van Straaten et al. (2022). In their study, the focus lies on forecasting high summer temperatures over western and central Europe. As input to their random forest (RF)-based model (Breiman 2001), they use a total of nine meteorological predictors from reanalysis data. Furthermore, they use different methods of interpretability, including SHAP, in order to find the most relevant predictors for their model.

The approach used in this study follows the same lines of thought. As input, up to nine different meteorological predictors are used that are known to influence central European wintertime temperatures and cold wave days, which are the focus of this study.

We use in our study quantile regression forests (QRFs; Meinshausen 2006) to predict continuous time series of central European mean 2-m temperatures and random forest classifiers (RFCs; Breiman 2001) to predict the occurrence of cold wave days over the same region for the extended winter season between November and April. The forecasts are assessed with respect to a climatological ensemble and interpreted at the example of two winters using SHAP. Specifically, we want to investigate 1) whether RF-based models are a suitable tool for predicting continuous wintertime central European mean 2-m temperatures and the occurrence of cold wave days, 2) which preprocessing of the input predictors (using all grid points of the meteorological fields, using only some spatial statistics of these fields, or using principal components of the fields) is suitable for doing so, and 3) which predictors are the most relevant for the models’ prediction.

In case our approach is successful, the ML models can either be used to forecast central European mean wintertime 2-m temperatures and the occurrence of cold wave days or to interpret or augment NWP model predictions on S2S time scales. Interpretation of NWP predictions can be done, for example, in the context of postprocessing by assigning different weights to the NWP ensemble members based on the predictions of the ML models.

Sections 2 and 3 of this paper give an overview of the used data and methods as well as the models’ architecture. In section 4, the results of this study are shown. Section 5 discusses and section 6 concludes this study.

2. Data

Observational data from the E-OBS, version 23.1e, dataset are used as the ground truth in this study (Cornes et al. 2018). This dataset consists of station observations over Europe that are interpolated to a regular 0.1° x 0.1° latitude–longitude grid. For this study, the daily mean 2-m temperature (tg) data of the extended winter season between November and April of the years 1950–2020 are retrieved for the region between 3° and 20°E and between 45° and 60°N (Fig. 1a). E-OBS data are only given for land grid points. To get a more homogeneous orography in the target area, the first grid points along the coasts and all mountainous grid points above 800 m above sea level (DWD 2023) are excluded (Fig. 1b). Over the resulting area, the 2-m temperature is spatially averaged. This region is referred to as central Europe.

ERA-5 reanalysis data (Bell et al. 2020a,b; Hersbach et al. 2020a,b) are taken as input data for the ML models. Nine different predictors are chosen based on meteorological domain knowledge. These are the zonal wind at 10 and 300 hPa; the geopotential at 100, 250, 500, and 850 hPa; the temperature and specific humidity at 850 hPa; and the mean sea level pressure. Every predictor is retrieved for the time period between 1950 and 2020 on a 1.5° x 1.5° latitude–longitude grid for the region between 60°W and 60°E and between 20° and 80°N (Fig. 1a). Only the data for the months October–April at 0000, 0600, 1200, and 1800 UTC are used and averaged daily afterward.

To account for the enhanced temporal uncertainty of subseasonal predictions, a 7-day running mean is used to smooth both the ground truth and the input data. The winters 2000/01 to 2019/20 are used to evaluate the ML models’ predictions. This time period is referred to as the evaluation period.
3. Methods

a. Climatological ensemble as a benchmark

As a benchmark, a climatological ensemble is created from the ground truth data. Using a climatological reference is a common choice in the field of S2S forecasting (e.g., Vijverberg et al. 2020). Since we chose to evaluate the ML models on the winters of 2000/01–2019/20, the climatological ensemble is created from the 30 winters before (1970/71–1999/2000). The decision to create the climatological ensemble from the 30 winters before the evaluation period follows the guidelines of the World Meteorological Organization (WMO) on the calculation of climatological normals (WMO 2020). The major difference here is that the data for the climatological ensemble are not averaged across the considered 30 winters. Instead, every winter is treated as a separate ensemble member, thus creating a climatological ensemble of 30 members (Fig. 1c). It is important to note here that the data are not detrended. For the central European mean, the winters from 1970/71 to 1999/2000 are approximately a mean of 0.8 K colder than the winters from 2000/01 to 2019/20. Therefore, the climatological ensemble has a slight advantage in predicting cold winter temperatures.

b. Choice of meteorological predictors as input for the ML models

The meteorological predictors and the input region used in this study are chosen based on physical knowledge. The used predictors are the zonal mean wind at 10-hPa height ($u_{10}$); the geopotential at 100 hPa ($z_{100}$), 250 hPa ($z_{250}$), 500 hPa ($z_{500}$), and 850 hPa ($z_{850}$); the temperature and specific humidity at 850 hPa ($t_{850}$ and $H_{850}$); and the mean sea level pressure (msl). The first predictor is selected since it represents the stratospheric state, which is an important source of predictability.
for central European winter weather (Domeisen et al. 2020). Furthermore, its pathway is a teleconnection between the northern hemispheric polar region in the input region and central Europe. Both z100 and z250 describe the upper-tropospheric state, which also influences central European surface weather (e.g., Hurrell et al. 2003; Kautz et al. 2022; Pinto et al. 2014). More details about the choice of predictors are given in section I of the online supplemental material.

We used u10, z100, and z250 as a small subset of predictors for the ML models. To include possibly relevant information from the medium-range time scale, the six remaining predictors are added. As a result, two different sets of predictors are created. To both, the month is added to account for the seasonality of temperatures.

Tropical predictors are not included directly in this study, but their indirect effects are, for example, via an enhanced likelihood of sudden stratospheric warming events during certain phases of the quasi-biennial circulation (Tripathi et al. 2015).

c. Preprocessing of meteorological predictors

In addition to the two subsets of predictors, three different types of input data are used in this study. First, all grid points of the meteorological predictor fields are converted into a vector. To keep as much spatial information as possible, the longitude and latitude of each grid point are kept and also used as predictors. This approach retains the maximum amount of data used in this study. The downside, however, is that this kind of model will not be easily interpretable later since its input consists merely of numbers on gridpoint level. Therefore, two additional types of input data that can be more easily physically interpreted are also considered. One data type consists only of spatial statistics across the whole input region—here, the minimum, mean, maximum, and variance—of the predictor fields. The last type of input data relates to the first 10 principal components of each predictor field. According to He et al. (2021), principal component analysis (PCA) is an appropriate method to extract features for weather forecasting on S2S time scales. The number of principal components derived from the original predictor field has been chosen as suitable from tests with 2, 5, 10, and 20 principal components for the original predictor fields and a gridpoint-based normalization of these (not shown). The first 10 principal components of each predictor field explain, with the exception of H850, greater than 75% of the variance of these fields, and for most fields, even greater than 90%. The month is not preprocessed further and is given as is in all three kinds of input data. Since there are two sets of predictor variables as described in the section above, in total, six different sets of input data are created.

d. Quantile regression forest model

In this study, a QRF model is used to solve a regression task. Proposed by Meinshausen (2006), a QRF is an extension of a regular RF model. An RF is a collection of several regression trees (RTs) in case of a regression task, or several decision trees (DTs) in case of a classification task, which are run in parallel. Both are built from several nodes. Every node is created by splitting the corresponding input data into two parts. The goal is that the ground truth data belonging to the input data in each part are as similar as possible and that the two parts are as different as possible. To find the best splitting criterion to reach this goal, for a regression task, the variance inside the two resulting parts is minimized and the variance between the two resulting parts maximized by an optimization algorithm. The data used for splitting at the first node are a subset of the training data obtained by bootstrapping with replacement. For every following node, all data remaining in the resulting node are used. In addition, at every node, only a subset of predictors, chosen by the algorithm, is considered for each split. The number of predictors in this subset is equal to the rounded-down square root of the total number of predictors. By using this, different split variables and thresholds are obtained at each node and in each tree. The process of growing a tree is stopped when the minimum node size is reached. This hyperparameter can be tuned by setting the minimal number of data points in a node. Alternatively, a minimal required number of data for splitting can be set. In the default setup, the minimum node size is 1, meaning that the number of remaining data points in this node is 1. This node is then a terminal node, also called a leaf.

To obtain a forecast with an RT, the validation or test data are sorted by the RT into the respective nodes. When using a regular RT, the forecast is then the weighted mean of all training ground truth data in the final leaf, and the mean of all RTs results in the forecast of the RF. When using a QRF, the ground truth data in the final leaf of each individual RTs are sorted by value. This is then considered a cumulative distribution function (CDF). The forecast of the QRF is created by calculating quantiles of the combined CDFs of the individual RTs. The combination of CDFs is done directly by the QRF model. In that way, an ensemble of predictions can be created easily.

In this study, 100 equidistant quantiles, including the minimum and maximum of the combined CDFs, are used as predictors, whereby each quantile serves as an ensemble member. The used QRF model is the default model from the Python package skranger, with the only changes being that the number of trees is set to 1000, the minimal node size is set to five, and quantile regression is enabled (Flynn 2021). The change of the model architecture is based on additional experiments with different hyperparameters settings of 100, 500, and 1000 trees as well as minimal node sizes of 1, 5, 10, 15, 30, and 100 and different combination of these two hyperparameters. The decision about the number of trees and the minimal node size is based on the results of the 20-winter mean for a QRF model trained with the spatial statistics of all predictor fields at a lead time of 14 days. For data-saving reasons, a separate validation set is not created. Therefore, it might be possible that the model used for selecting the hyperparameters is overestimated in its performance. Nevertheless, since RFs are not prone to overfitting, we see this as unproblematic. A detailed overview of all hyperparameters of the model is provided in the appendix (Table A1).

For the SHAP analysis, additional information about the underlying RTs is saved.
For training the ML models, the three different kinds of inputs are used, resulting in three different kind of models (see section 3c). In total, considering also the two subsets of predictors, six different ML models are trained for each lead time (Table 1). A leave-one (winter)-out cross-validation approach is performed in both cases to increase the amount of training data. This means that the training data contain the winters 1950/51–1999/2000 and 19 of the 20 winters from winter 2000/01 to 2019/20. The winter left out is the winter to be predicted (e.g., when using the winter 2011/12 as the test data to evaluate the model prediction, the training data contain the winters of 1950/51–2010/11 and 2012/13–2019/20).

e. Random forest classifier model

To forecast cold wave days instead of continuous winter-time temperatures, RFC models are used. In contrast to the QRF models, the RFC models perform a binary classification task and, therefore, the ground truth data are converted into a binary time series. The input remains the same as for the QRF models (Table 1). The converted ground truth is a binary list of cold wave days, where 1 indicates that the given day is characterized as a “one cold wave day” and 0 indicates that the given day is characterized as a “no cold wave day.” The cold wave definition of Smid et al. (2019) is used to classify days with cold waves. A cold wave is characterized by at least three consecutive days with temperatures below a threshold obtained from climatology. The original reference period used by Smid et al. (2019) is 1981–2010. Since the evaluation period chosen for this study overlaps with this time period, the same time period as used for the climatological ensemble, the winters of 1970/71–1999/2000 are taken as the reference period instead. Over this reference period, the daily mean 2-m temperature time series is smoothed by a 31-day running mean and, afterward, the multiyear 10th percentile of this time series is taken. The value of this percentile is taken as the daily threshold for cold waves (Fig. 1d).

In this study, the 2-m daily mean temperature is taken instead of the 2-m minimum temperature, as done in Smid et al. (2019), for calculating the cold wave thresholds. This is done for two main reasons. First, the aim of this paper is to forecast daily mean temperatures. Therefore, we think a comparison with a cold wave threshold based on daily mean temperatures is fairer. Daily mean temperatures are less dependent on the small-scale terrain and cloudiness than daily minimum temperatures and provide, therefore, a more sensible average over the target region. Second, when using the minimum instead of the daily mean temperatures, the number of cold wave days is reduced, which leads to a more severe class imbalance of events.

The training and leave-one (winter)-out cross-validation for the RFC models are done analogously to the QRF models. As a prediction, the fraction of ground-truth-training examples in the respective terminal leaves are taken. Besides this, the most important difference between the QRF and RFC models is the split rule. While for the QRF models “variance” is taken as the split rule, for the RFC models it is “gini.” The gini impurity is a measure of how likely it is that a data point is classified incorrectly (Suthaharan 2016). This probability is minimized during training. The RFC model used in this study is the default model from the Python package sklearn (Flynn 2021) with the hyperparameter adjustments for the number of trees (set to 1000) and the minimal node size (set to 5). A detailed overview of all hyperparameters of the model is shown in the appendix (Table A2).

f. Forecast evaluation with the continuous ranked probability score

Forecasts are made for every day of winter during the evaluation period. These probabilistic forecasts are compared with the deterministic ground truth and evaluated with the continuous ranked probability score (CRPS). As shown in Gneiting et al. (2007), the CRPS can be written as

\[
\text{CRPS}(F, x) = \int_{-\infty}^{\infty} (F(y) - 1(y \geq x))^2 \, dy,
\]

where \(F(y)\) is the predictive CDF and \(x\) is the observation; here \(1(y \geq x)\) is the indicator function that takes the value 1 if \(y \geq x\) and 0 otherwise. An alternative representation of the CRPS is

\[
\text{CRPS}(F, x) = E_F[X - x] - \frac{1}{2} E_F[X - X'],
\]

where \(E_F\) is the expectation of a random variable with predictive CDF \(F\), \(x\) is the observation, and \(X\) as well as \(X'\) are two independent random variables with CDF \(F\) and a finite first
moment. The unit of the CRPS is the same as the unit of the observation and forecasts. Its value corresponds directly to the mean absolute error of the forecast in case the forecast is deterministic. The CRPS is always positive and the lower the better.

g. Forecast evaluation with the Brier score

In addition to the prediction of the continuous time series of predictions, binary predictions are also made. In this study, the predictions are sorted into two categories according to the calculated cold wave thresholds. The Brier score (BS) is used to evaluate binary categorical forecasts. It is the same as the ranked probability score (RPS) for binary categories and can be written as

\[ BS_F(y) = RPS_{\text{binary}}(y) = (F(1) - \mathbf{1}(y = 1))^2, \]

where \( F(1) \) is the predicted probability of an event (Brier 1950). The BS is always positive and lower values indicate a better forecast. The integral over the BS is equal to the CRPS [Eq. (1)].

h. Comparison with a benchmark model using skill scores

A comparison of the skill of a model in reference to a benchmark model is done using skill scores. These are calculated from the respective mean score of the models. For the CRPS, the respective skill score can be written as (Taillardat et al. 2016)

\[ \text{CRPSS} = 1 - \frac{\text{CRPS}_{\text{model}}}{\text{CRPS}_{\text{benchmark}}}. \]

Here, the mean CRPS over a longer time period of both models is used. In this study, these are a whole winter and the mean over 20 winters. The respective skill score for the BS (BSS) is, analogously,

\[ \text{BSS} = 1 - \frac{\text{BS}_{\text{model}}}{\text{BS}_{\text{benchmark}}}. \]

Positive values of the skill score indicate that the model is able to skillfully forecast with respect to the benchmark model. This means that the forecasts of the model are more skillful than the forecasts of the benchmark model. A perfect model shows a skill score of 1. If the skill score is 0, the model is as good as the benchmark model. If the model is not able to predict more skillfully than the benchmark model, the skill score is smaller than 0. In this study, the benchmark model is set to be the climatological ensemble.

i. Shapley additive explanations

To interpret the predictions of the ML models, we decided to perform a SHAP analysis. SHAP values are based on Shapley values and have a similar interpretation (Molnar 2022). In this study, we use the implementation by Lundberg et al. (2020) in the Python package shap (Lundberg 2018). The so-called TreeExplainer, which calculates the SHAP values for RT- and DT-based models, can be used directly in combination with the Python package skranger, from which we create the RF-based ML models. According to Molnar (2022), for the interpretation of Shapley values, two predictions are important. The first is the mean prediction of the training dataset. In the case of the QRFs, we obtain this prediction by first averaging over all predicted quantiles and then over all predicted days. In the case of the RFCs, we use the probabilistic prediction and then select the class that corresponds to the days we want to analyze with SHAP. In the next step, we average over all predicted days. The second prediction value needed is the actual prediction by the model. This is the prediction of the days we want to analyze with SHAP, again averaged to one value, as shown for the mean prediction of the training data. SHAP values show how much each predictor of the input data contributes to the difference between the mean prediction of the training data and the actual prediction of the model. For deterministic forecasts, all SHAP values add up to this difference. It is important to note that this difference between the mean prediction of the training data and the actual prediction does not show whether the actual prediction is accurate or not, since SHAP does not take the ground truth into consideration. We use SHAP to analyze the most positive and most negative contributing predictors to our models’ prediction and try to connect it to physically known relations on S2S time scales. However, with this method, it is not possible to show causality or prove, ultimately, that a model has learned the physically relevant pattern in the given data that one can interpret into the SHAP values.

4. Results

a. Continuous temperature forecasts

1) 20-WINTER MEAN

The naïve expectation is that ML models can outperform a climatological ensemble at all analyzed lead times and, ideally, also all forecast times. This would imply a positive skill in the 20-winter mean and an equally distributed skill over all winters.

The predictive skill of the QRF models is evaluated in terms of the continuous ranked probability skill score (CRPSS). First, the daily CRPSS is calculated for the whole evaluation period. Then, these values are averaged winterwise, resulting in 20 values for the 20 analyzed winters. These 20 values are then averaged again to create the 20-winter mean CRPSS. This 20-winter mean CRPSS shows at all lead times only a small deviation in the mean from 0 but a large standard deviation (Fig. 2). 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 show the highest mean CRPSS values per lead time, followed by the models using the statistics of the fields as input. The QRF models trained with nine meteorological predictors show a better performance than the models trained with three meteorological predictors. In general, the skill of all QRF models decreases with increasing lead times.

At a lead time of 14 days, 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 (Fig. 2a). All other QRF models show negative mean CRPSS values (Fig. 2a). For a lead time of 21 days, all QRF models except the one trained on all grid points of all predictor fields show negative mean CRPSS values (Fig. 2b). The same picture is seen at a lead time of 28 days (Fig. 2c). As mentioned before, the large standard deviations in CRPSS values between the winters at all lead times imply that the skill is not equally distributed between winters. Therefore, as a next step, we look at case studies.

2) SELECTION OF CASE STUDIES

As case studies, two fundamentally different winters are chosen. As the first case study, the winter 2011/12 is selected. This winter is characterized by a severe cold wave in February (Fig. 3a). As a second case study, the winter 2013/14 is selected. This winter is characterized by mild temperatures throughout the winter (Fig. 3b). For additional information see section II of the online supplemental material.

3) CASE STUDIES OF WINTERS 2011/12 AND 2013/14

For analyzing the chosen case studies in more detail, the CRPSS is calculated for each winter separately. This allows us to investigate which set of input data is best suited for predicting the respective winter. Additionally, the daily difference of CRPSS values between the climatological ensemble and each QRF model is obtained (Fig. 5). This is done to investigate a possible dependence of the models’ performances of the temperature or time in winter.

At a lead time of 14 days, for both analyzed winters, four QRF models show positive mean CRPSS values (Figs. 4a,b). In case of the winter 2011/12, the best-performing model is the QRF trained on the statistics of the selected predictor fields (Fig. 4a). In the case of the winter 2013/14, it is the model trained on all grid points of all predictor fields (Fig. 4b). At a lead time of 21 days, two models show positive CRPSS values in the case of the winter 2011/12, but none in the case of the winter 2013/14. This shows how difficult forecasting on S2S time scales can be (Figs. 4c,d). The best-performing model in predicting the 2-m temperature of the winter 2011/12 is the QRF trained on all grid points of the selected predicted fields (Fig. 4c). When looking at a lead time of 28 days, for both winters, all QRF models show only negative mean CRPSS values (Figs. 4e,f).

When looking at the daily CRPS difference, it is striking that at lead times of 14 and 21 days, the ML models perform especially well during the time of the cold wave (Figs. 5a,c). At a lead time of 28 days and for the winter 2013/14, a dependence of the ML models’ performance on temperature is not seen (Figs. 5b,d–f).
4) SHAP FOR CASE STUDIES OF WINTERS 2011/12 AND 2013/14

To investigate which predictors influence the prediction in which way, SHAP values are analyzed. Thereby, the suggested time dependence on the QRF models’ performance on the time of winter, represented by the month in the input data, is also checked. Here, the winter 2011/12 is split into periods with cold waves and periods with warm temperatures. The periods with warm temperatures are defined here as days without cold waves (all areas around the gray boxes in Fig. 7a). The winter 2013/14 is analyzed as one since it does not contain any cold waves. As for the analyzed QRF model type, the QRF_stat_all model with a lead time of 14 days is chosen. This is done since it shows positive CRPSS values in the 20-winter mean (Fig. 2a) and for both case studies (Figs. 4a,b). A detailed discussion of the SHAP analysis can be found in section IV of the online supplemental material.

In the case of the warm periods and the winter 2013/14, the top five predictors according to the SHAP values differ for every period. This shows that the predictions of the QRF model are not solely based on a fixed set of predictors with a fixed contribution of each. Common candidates are either the mean or maximum of t850, which are contributing positively to the prediction. In the case of the cold periods of the winter 2011/12, common candidates are also t850 and, additionally, u10. Depending on the cold wave period, the contribution is either positive or negative. A clear pattern is not seen. For both cold and warm period predictions in midwinter, the month contributes negatively to the predicted temperature, otherwise it contributes positively. This underlines the time dependence of the forecast and is physically plausible. Depending on the study, major cold waves over central Europe occur most often during December and January (Tomczyk et al. 2019) or January and February (Lhotka and Kyselý 2015) (Fig. 1).

b. Binary cold wave day forecasts

1) 20-WINTER MEAN

Concerning central Europe, arguably the most relevant temperature extremes in winter are cold waves. Therefore, we also try to directly forecast days that fulfill a cold wave criterion based on Smid et al. (2019), using binary RFCs. The skill of these models is measured with 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 days, four models show positive mean BSS values in the 20-winter mean (Fig. 6a). These are the RFC models trained with all predictor fields and the RFC model trained on the statistics of only the selected predictor fields. Two of these models, the RFCs trained on the statistics of all predictor fields and the RFCs trained on the first 10 principal components of all predictor fields, show also positive BSS values at a lead time of 21 days (Fig. 6b). This is an improvement relative to the forecast of the continuous 2-m temperature, for which only one of the QRF models shows skill at a lead time of 21 days (Fig. 2b). At a lead time of 28 days, for the 20-winter mean, none of the RFC models shows a positive mean BSS value (Fig. 6c). 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 are used that we used to determine the continuous temperature forecast.

2) CASE STUDIES OF WINTERS 2011/12 AND 2013/14

For visualizing the predictions of the climatological ensemble and a chosen RFC model, the model trained on the statistics of all predictor fields (RFC_stat_all) at a lead time of 14 days is used as an example. This is done to be consistent with the chosen model for continuous temperature forecasting. Similar to the continuous temperature forecasts, it is seen that the
predictions of the climatological ensemble and the RFC model differ, but not largely (Figs. 3 and 7). In the case of the winter 2011/12, the highest predicted probability of a cold wave day is below 0.5, showing that the RFC is not certain in predicting cold waves (Fig. 7a). Nevertheless, the severe cold wave in midwinter and the short cold wave in the beginning of the winter are better captured than by the climatological ensemble. Interestingly, in the case of the cold-wave-free winter 2013/14, the highest predicted probability of a cold wave day is around 0.55 (Fig. 7b). This, among the other differences in the predicted probabilities of cold waves between the two winters, 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.

When assessing the skill of the RFC models with respect to the climatological ensemble, it is striking that all RFC models show positive BSS values for the winter 2011/12 at a lead time of 14 days (Fig. 8a). The best-performing model is the RFC_stat_all. In the case of the winter 2013/14, only two models show positive BSS values (Fig. 8b). These are the RFC trained on the statistics of all predictor fields and the first 10 principal components of all predictor fields. At lead times of 21 and

---

**Fig. 4.** Winter mean CRPSS of the QRF models with respect to the climatological ensemble for the winters of (left) 2011/12 and (right) 2013/14 for lead times of (a),(b) 14; (c),(d) 21; and (e),(f) 28 days.
28 days, none of the RFC models shows positive BSS values for the winter 2013/14 (Figs. 8d,f). In the case of the winter 2011/12, though, skill is also found for some RFC models on these longer lead times, whereby the RFCs trained on the first 10 principal components of both all predictor fields and only the selected predictor fields perform best (Figs. 8c,e). This suggests that the RFC models might be better suited for forecasting winters with cold waves than winters without, when compared with the climatological ensemble. A more detailed look shows that the RFCs perform especially well during long-lasting and midwinter cold waves (Figs. 9a,c,e, along with Figs. S1–S6 in section III of the online supplemental material). During times of mild temperatures, a clear pattern in RFC performance is not found (Figs. 9b,d,f, along with Figs. S1–S6). This is also true for short cold wave periods.

3) SHAP FOR CASE STUDIES OF WINTERS 2011/12 AND 2013/14

To remain consistent with the SHAP analysis for the QRF models, the RFC model with the same input (RFC_stat_all) is...
used. A detailed analysis can be found in section IV of the online supplemental material. When looking at the SHAP analysis of the warm periods of the RFC models, either the mean or maximum of $t_{850}$ is a common predictor that contributes positively to the forecast. Interestingly, $msl$ is one of the most positively contributing predictors to the forecast of the severe midwinter cold wave. Usually, $msl$ is expected to play only a minor role on S2S time scales since this is a generally fast-varying predictor. However,
the msl pattern can be persistent over many days up to weeks (e.g., in the case of blocking situations); thus, msl can also play a relevant role on the S2S time scales.

5. Discussion

In this study, we show that a simple combination of physical knowledge about the weather forecasting task at hand and RF-based ML models can lead to improved subseasonal forecasts relative to a climatological benchmark. We chose to incorporate our physical knowledge directly into the predictors. This is not only a simple approach, since the architecture and loss functions of the RF-based models are not modified, but it also leads to a large reduction in the amount of training data and subsequently the amount of necessary computational resources. We concentrated on nine meteorological predictors known to affect central European surface weather in winter. Since we focus on S2S forecasting, we can limit the area from which the predictors are retrieved from globally to a smaller region covering roughly one-fourth of the Northern Hemisphere. Thereby, we tried to find a balance between

![Fig. 8. Winter mean BSS of the RFC models with respect to the climatological ensemble for the winters of (left) 2011/12 and (right) 2013/14 for lead times of (a),(b) 14; (c),(d) 21; and (e),(f) 28 days. Note the different scale in (f).](image-url)
the area necessary to include as many physically known drivers relevant on that time scale as possible and the increasing amount of computational resources needed to process these. We found that the most promising combination of predictors is the minimum, mean, maximum, and variance of $u_{10}$, $z_{100}$, $z_{250}$, $z_{500}$, $z_{850}$, $t_{850}$, $H_{850}$, $u_{300}$, and $msl$ over a region covering the North Atlantic Ocean, Europe, and western-most Asia. This holds for both the QRFs and RFCs.

In the 20-winter mean, the models using these predictors as input are able to produce skillful forecasts 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, exclusively at a lead time of 14 days. But the difference in the mean CRPSS value is small and additional challenges arise. Models trained on all grid points of predictor fields are not

![Figure 9](image-url)
easy to interpret, which is crucial when trying to build trust in the models’ forecasts.

For all model types, the spread in CRPSS and BSS values is high, showing a strong dependence of the model performance on the specific winter to be forecast. This is also true for the climatological ensemble. In case of continuous temperatures, skill is found in the 20-winter mean up to a lead time of 28 days when using a QRF trained on all grid points of all predictor fields. In the case of the binary cold-wave-day forecasts, skill at a lead time of 28 days cannot be found in the 20-winter mean.

To find out under which conditions the RF-based ML models perform best, we chose two case studies for a more detailed analysis. The chosen winters are very different. The first winter, the winter 2011/12, is characterized by a severe, abruptly starting, and long-lasting cold wave in January and February. Interestingly, especially during the time of this cold wave, the ML models perform especially well relative to the climatological ensemble. One explanation, therefore, might be the choice of meteorological predictors. For example, a weak polar vortex, represented by low u10 values, can lead to strong stratospheric anomalies that can influence the geopotential height field at tropopause level, represented by z100, and then lead to cold spells over central Europe. The timing of these stratospheric anomalies is well in the range of the S2S time scale. This suggestion is supported by the SHAP analysis, which shows that u10 and z100 are among the most contributing predictors to the QRFs forecast during this period.

The second winter, the winter 2013/14, is characterized by the absence of cold waves. Instead, it was a mild and very stormy winter. During this winter, not only the climatological ensemble’s forecasts perform worse than for the winter 2011/12, the same is true for the ML models. This, again, might be due to the choice of meteorological predictors, although we tried to include predictors like z500, which is used to detect blocking, and z850, which represents the large-scale air masses. The latter is also detected by the SHAP analysis as one of the most contributing predictors during warm periods in winter.

We assume that in the cases of mild and stormy winter weather, the possible number of drivers is higher and generally more diverse than in the case of severe cold waves. This would lead to a less clear pattern in the input data and increased complexity for the ML models to learn how to forecast mild winter temperatures, leading to worse forecasts. In case of the severe, long-lasting cold waves, three main drivers are well known. This is a more manageable amount of pattern for the ML models to learn and, therefore, results in better forecasts. Knowing this is helpful in assessing the ML models’ forecast 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 S2S time scales than do mild temperatures.

Summary

We found the following in this study:

- The most useful input in terms of performance, computational efficiency, and interpretability consists of the month and the minimum, mean, maximum, and variance of u10, z100, z250, z500, z850, t850, H850, u300, and msl.
- In the 20-winter mean, QRF models show skill at lead times of 14, 21 and 28 days; RFC models show skill at lead times of 14 and 21 days; for both model types, the variability between winters is high.
- The skill of the individual models depends on the time of winter, whereby the models perform better during long-lasting and midwinter cold waves than at the margins of winters.
- According to the SHAP analysis of case studies, the most common predictor is t850 for warm periods for both models.
- During the midwinter cold wave in 2011/12, common predictors of both models that contribute to a forecast of colder temperatures and cold wave days are also t850 and msl. Especially the latter is not expected from a physical point of view.

6. Conclusions

In this study, we find that QRF and RFC models are suitable tools for forecasting central European mean 2-m temperature and the occurrence of cold wave days on S2S time scales. The analysis of SHAP values for two case studies shows that in the case of mild winter temperatures, these ML models are able to select predictors as most relevant that are consistent with well-known physical relationships in the data. In the case of a severe midwinter cold wave, for both models, the relevant predictors in common are not expected from physical knowledge. 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 skillful forecasts on S2S time scales with reduced input data based on meteorological knowledge. This makes these kinds of models computationally more efficient than traditional NWP models. While, in general, the forecast skill with respect to the climatological ensemble decreases with increasing lead time, the variability between winters is high. Skill is found in the 20-winter mean at lead times of 14, 21, and 28 days in case of the QRF models, and at lead times of 14 and 21 days in case of the RFC models.

For now, the ML models are only evaluated with respect to a climatological ensemble. This gives a first insight into the suitability of these models to forecast central European mean 2-m temperatures and the occurrence of cold wave days. In a next step, these models can be evaluated with respect to state-of-the-art NWP models that forecast on S2S time scales. Furthermore, more advanced ML models, such as deep learning models or convolutional neural networks, will be applied to the forecast problem at hand. Another promising research direction is the use of ML models for postprocessing S2S forecasts of state-of-the-art NWP models, as already shown for neural networks in the framework of WMO’s S2S AI Challenge (Vitart et al. 2022).

Acknowledgments. The research leading to these results was partly embedded within the subproject C8 “Stratospheric
influence on predictability of persistent weather patterns” of the Transregional Collaborative Research Center SFB/TRR 165 “Waves to Weather” (http://www.wavestoweeather.de), funded by the German Research Foundation (DFG). Authors Kiefer and Ludwig were funded by the German Helmholtz Association (“Changing Earth” program). Author Lerch gratefully acknowledges support by the Vector Stiftung through the Young Investigator Group “Artificial Intelligence for Probabilistic Weather Forecasting.” Author Pinto thanks the AXA Research Fund for support. We thank the two anonymous reviewers for their very valuable comments. Special thanks are given to Hendrik Feldman and the ClimXtreme team for the courtesy of providing the E-OBS data used in this study.

Data availability statement. The code used for this study is available online (https://github.com/selinakiefer/rfs_central_european_cold_winter_weather). We acknowledge the E-OBS dataset (Cornes et al. 2018) from the EU-FP6 project UERRA (http://www.uerra.eu) and the Copernicus Climate Change Service, and the data providers in the ECA&D project (https://www.ecad.eu). Bell et al. (2020a,b) and Hersbach et al. (2020a,b) data were downloaded from the Copernicus Climate Change Service (C3S) Climate Data Store. The results contain modified Copernicus Climate Change Service information for 2021. Neither the European Commission nor ECMWF is responsible for any use that may be made of the Copernicus information or data it contains.

APPENDIX

Hyperparameters of QRF and RFC Models

This appendix provides a detailed overview of all hyperparameters of the QRF (Table A1) and RFC (Table A2) models. The hyperparameters that are highlighted with boldface type in the appendix tables are the most important ones for defining the models’ architecture.

Table A1. Hyperparameters of QRF models. These are the used hyperparameters of the ranger random forest regression model (Flynn 2021). Here “maxstat” indicates the maximally selected rank statistics splitting rule.

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Name in class “skranger.ensemble.RangerForestRegressor”</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of trees</td>
<td>n_estimators</td>
<td>1000</td>
</tr>
<tr>
<td>Verbosc logging</td>
<td>verbose</td>
<td>False</td>
</tr>
<tr>
<td>Number of features to split on each node</td>
<td>mtry</td>
<td>0</td>
</tr>
<tr>
<td>Feature importance</td>
<td>importance</td>
<td>None</td>
</tr>
<tr>
<td>Minimal node size</td>
<td>min_node_size</td>
<td>5</td>
</tr>
<tr>
<td>Maximal number of nodes per tree</td>
<td>max_depth</td>
<td>0</td>
</tr>
<tr>
<td>Replacement for bootstrapping sampling</td>
<td>replace</td>
<td>True</td>
</tr>
<tr>
<td>Fraction of training data for sampling</td>
<td>sample_fraction</td>
<td>None</td>
</tr>
<tr>
<td>Save how often data points are in bag</td>
<td>keep_inbage</td>
<td>False</td>
</tr>
<tr>
<td>List of the in-bag counts</td>
<td>inbag</td>
<td>None</td>
</tr>
<tr>
<td>Rule to define splitting of data</td>
<td>split_rule</td>
<td>Variance</td>
</tr>
<tr>
<td>Number of random splits for the split rule</td>
<td>number_random_splits</td>
<td>1</td>
</tr>
<tr>
<td>Significance threshold for maxstat split rule</td>
<td>alpha</td>
<td>0.5</td>
</tr>
<tr>
<td>Lower quantile of covariate distribution for maxstat</td>
<td>minprop</td>
<td>0.1</td>
</tr>
<tr>
<td>Weight vector for selecting features for splitting</td>
<td>split_select_weights</td>
<td>None</td>
</tr>
<tr>
<td>Use always certain features for splitting</td>
<td>always_split_features</td>
<td>None</td>
</tr>
<tr>
<td>List of categorical features</td>
<td>categorical_features</td>
<td>None</td>
</tr>
<tr>
<td>Consider categorical features</td>
<td>respect_categorical_features</td>
<td>None</td>
</tr>
<tr>
<td>Permutation feature importance</td>
<td>scale_permutation_importance</td>
<td>False</td>
</tr>
<tr>
<td>Local permutation feature importance</td>
<td>local_importance</td>
<td>False</td>
</tr>
<tr>
<td>Vector with regularization values for features</td>
<td>regularization_factor</td>
<td>None</td>
</tr>
<tr>
<td>Consider depth in regularization</td>
<td>regularization_usedepth</td>
<td>False</td>
</tr>
<tr>
<td>Hold out samples with case weight 0</td>
<td>holdout</td>
<td>False</td>
</tr>
<tr>
<td>Enables quantile regression after fitting</td>
<td>quantiles</td>
<td>True</td>
</tr>
<tr>
<td>Out-of-bag prediction error</td>
<td>oob_error</td>
<td>False</td>
</tr>
<tr>
<td>Number of threads</td>
<td>n_jobs</td>
<td>-1</td>
</tr>
<tr>
<td>Reduce speed of computation to save memory</td>
<td>save_memory</td>
<td>False</td>
</tr>
<tr>
<td>Random seed value</td>
<td>seed</td>
<td>42</td>
</tr>
<tr>
<td>Detailing the underlying decision trees</td>
<td>enable_tree_details</td>
<td>True</td>
</tr>
</tbody>
</table>
**Table A2. Hyperparameters of RFC models.** These are the used hyperparameters of the ranger random forest classifier model (Flynn 2021).  

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Name in class “skranger.ensemble.RangerForestClassifier”</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of trees</strong></td>
<td>n_estimators</td>
<td>1000</td>
</tr>
<tr>
<td>Verbose logging</td>
<td>verbose</td>
<td>False</td>
</tr>
<tr>
<td><strong>Number of features to split on each node</strong></td>
<td>mtry</td>
<td>0</td>
</tr>
<tr>
<td>Feature importance</td>
<td>importance</td>
<td>None</td>
</tr>
<tr>
<td><strong>Minimal node size</strong></td>
<td>min_node_size</td>
<td>5</td>
</tr>
<tr>
<td><strong>Maximal number of nodes per tree</strong></td>
<td>max depth</td>
<td>0</td>
</tr>
<tr>
<td>Replacement for bootstrapping sampling</td>
<td>replace</td>
<td>True</td>
</tr>
<tr>
<td>Fraction of training data for sampling</td>
<td>sample_fraction</td>
<td>None</td>
</tr>
<tr>
<td>Save how often data points are in bag</td>
<td>keep_inbag</td>
<td>False</td>
</tr>
<tr>
<td>List of the in-bag counts</td>
<td>inbag</td>
<td>None</td>
</tr>
<tr>
<td><strong>Rule to define splitting of data</strong></td>
<td>split_rule</td>
<td>gini</td>
</tr>
<tr>
<td>Number of random splits for the split rule</td>
<td>number_random_splits</td>
<td>1</td>
</tr>
<tr>
<td>Consider categorical features</td>
<td>respect_categorical_features</td>
<td>None</td>
</tr>
<tr>
<td>Permutation feature importance</td>
<td>scale_permutation_importance</td>
<td>False</td>
</tr>
<tr>
<td>Local permutation feature importance</td>
<td>local_importance</td>
<td>False</td>
</tr>
<tr>
<td>Vector with regularization values for features</td>
<td>regularization_factor</td>
<td>None</td>
</tr>
<tr>
<td>Consider depth in regularization</td>
<td>regularization_useddepth</td>
<td>False</td>
</tr>
<tr>
<td>Hold out samples with case weight 0</td>
<td>holdout</td>
<td>False</td>
</tr>
<tr>
<td>Out-of-bag prediction error</td>
<td>oob_error</td>
<td>False</td>
</tr>
<tr>
<td><strong>Number of threads</strong></td>
<td>n_jobs</td>
<td>–1</td>
</tr>
<tr>
<td>Reduce speed of computation to save memory</td>
<td>save_memory</td>
<td>False</td>
</tr>
<tr>
<td><strong>Random seed value</strong></td>
<td>seed</td>
<td>42</td>
</tr>
<tr>
<td>Detailing the underlying decision trees</td>
<td>enable_tree_details</td>
<td>True for RFC_stat_all</td>
</tr>
</tbody>
</table>

**REFERENCES**


---


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


