Correcting Subseasonal Forecast Errors with an Explainable ANN to Understand Misrepresented Sources of Predictability of European Summer Temperatures

CHIEM VAN STRAATEN,[a,b] KIRIEN WHAN,[a] DIM COUMOU,[a,b] BART VAN DEN HURK,[b,c] AND MAURICE SCHMEITS[a,b]

[a] *Royal Netherlands Meteorological Institute (KNMI), De Bilt, Netherlands*
[b] *Institute for Environmental Studies (IVM), Vrije Universiteit Amsterdam, Amsterdam, Netherlands*
[c] *Deltares, Delft, Netherlands*

(Manuscript received 13 June 2022, in final form 3 April 2023, accepted 6 April 2023)

ABSTRACT: Subseasonal forecasts are challenging for numerical weather prediction (NWP) and machine learning models alike. Forecasting 2-m temperature (t2m) with a lead time of 2 or more weeks requires a forward model to integrate multiple complex interactions, like oceanic and land surface conditions leading to predictable weather patterns. NWP models represent these interactions imperfectly, meaning that in certain conditions, errors accumulate and model predictability deviates from real predictability, often for poorly understood reasons. To advance that understanding, this paper corrects conditional errors in NWP forecasts with an artificial neural network (ANN). The ANN postprocesses ECMWF extended-range summer temperature forecasts by learning to correct the ECMWF-predicted probability that monthly t2m in western and central Europe exceeds the climatological median. Predictors are objectively selected from ECMWF forecasts themselves, and from states at initialization, i.e., the ERA5 reanalysis. The latter allows the ANN to account for sources of predictability that are biased in the NWP model itself. We attribute ANN corrections with two explainable artificial intelligence (AI) tools. This reveals that certain erroneous forecasts relate to tropical western Pacific Ocean sea surface temperatures at initialization. We conjecture that the atmospheric teleconnection following this source of predictability is imperfectly represented by the ECMWF model. Correcting the associated conditional errors with the ANN improves forecast skill.

SIGNIFICANCE STATEMENT: We want to understand occasions in which a numerical weather prediction (NWP) model fails to forecast a predictable event existing in the real world. For forecasts of European summer weather more than 2 weeks in advance, real predictable events are rare. When misrepresented by the model, predicted future states become needlessly biased. We diagnose these missed opportunities with an explainable neural network. The neural network is aware of the initial state and learns to correct the NWP forecast on occasions when it misrepresents a teleconnection from the western tropical Pacific Ocean to Europe. The explainable architecture can be useful for other applications in which conditional model errors need to be understood and corrected.

KEYWORDS: Neural networks; Subseasonal variability; Model errors; Numerical weather prediction/forecasting; Postprocessing

1. Introduction

Subseasonal to seasonal (S2S) forecasts are made with a lead time of more than 2 weeks. S2S forecasts of variables like atmospheric temperature and precipitation are crucial in the anticipation of heatwaves and droughts (White et al. 2022). Often, however, the detailed temporal evolution of temperature and precipitation at lead times beyond 2 weeks is not predictable (e.g., Buizza and Leutbecher 2015; Zhang et al. 2019), as atmospheric motion is sensitive to small variations in initial conditions (Lorenz 1963). In these cases, ensemble members of numerical weather prediction (NWP) models show widely diverging possible states (Leutbecher and Palmer 2008).

Only in certain conditions can low-frequency internal variability of the atmosphere and interaction with persistent Earth system components create so-called windows of predictability (Mariotti et al. 2020). In such windows, a source of subseasonal predictability constrains the range of states that the atmosphere can visit (Palmer 1993; Toth and Buizza 2019). Figure 1a illustrates how atmospheric and oceanic sources of predictability can steer the probability of an event at a valid time more than 2 weeks into the future. Mandatory is that the sources of predictability are adequately represented in the NWP model (Fig. 1a).

Certain sources of subseasonal predictability are particularly important to represent. For instance, oceanic influences such as sea surface temperature (SST) patterns are known to interact with and steer the position of the North Atlantic jet stream in summer (Ossó et al. 2020; Osborne et al. 2020; Carvalho-Oliveira et al. 2022). Another example is land surface drought. Its feedback on the atmosphere during high pressure "blocking" systems (Kautz et al. 2022) makes high 2-m air temperatures t2m more likely (Quesada et al. 2012).

The existence of such time-dependent sources of predictability can in theory lead to better S2S forecasts of heatwaves and droughts (Hoskins 2013). In practice, this predictability is not achieved because NWP models are imperfect. A first reason is the mentioned sensitivity to initial conditions in the atmosphere and in other Earth system components. These are

Corresponding author: Chiem van Straaten, j.w.vanstraaten@vu.nl

DOI: 10.1175/AIES-D-22-0047.1 e220047

© 2023 American Meteorological Society. This published article is licensed under the terms of the default AMS reuse license. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

Unauthenticated | Downloaded 04/29/25 04:12 AM UTC

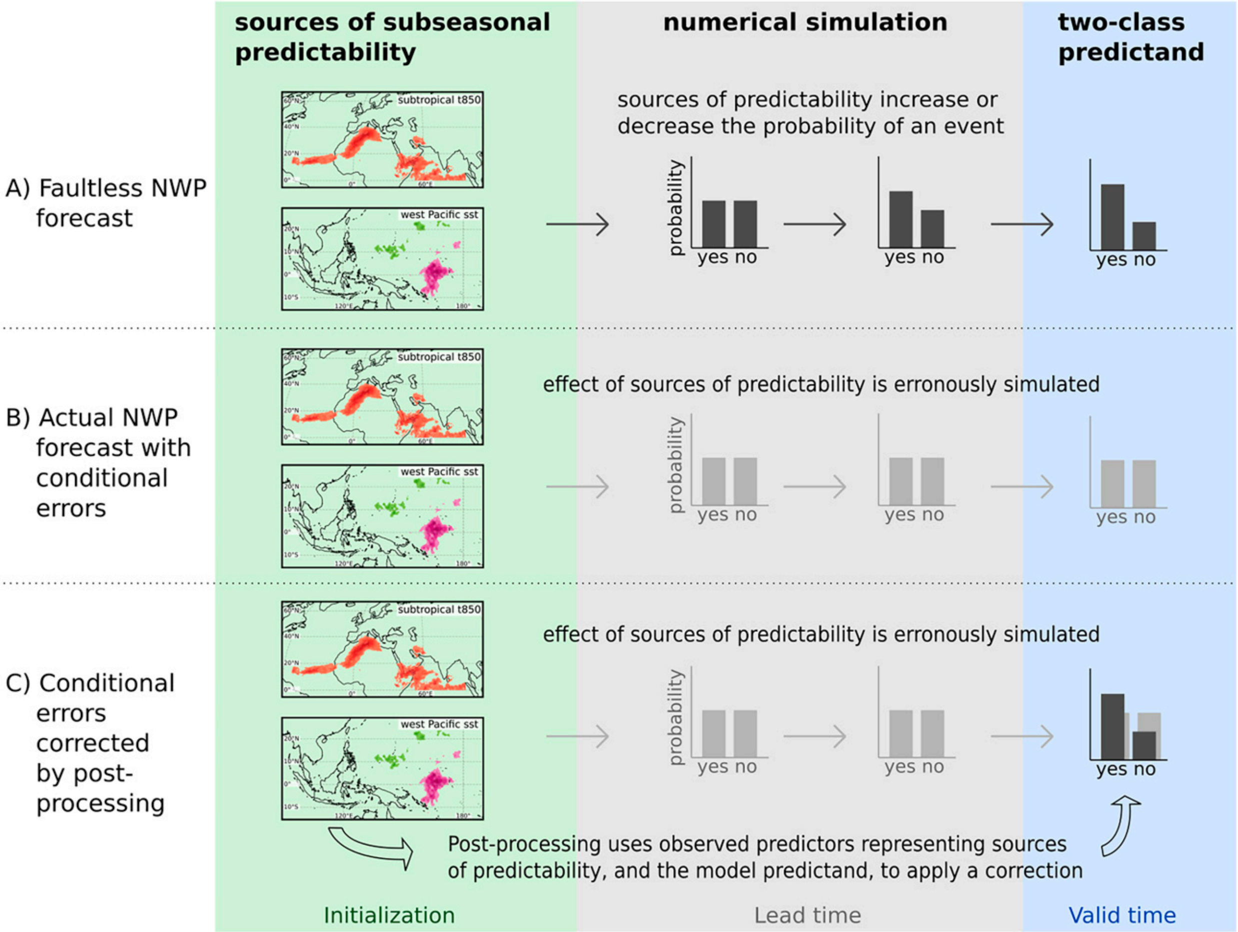

sources of subseasonal predictability
numerical simulation
two-class predictand
A) Faultless NWP forecast
subtropical t850
west Pacific sst
sources of predictability increase or decrease the probability of an event
probability
yes no
B) Actual NWP forecast with conditional errors
effect of sources of predictability is erronously simulated
C) Conditional errors corrected by post-processing
Post-processing uses observed predictors representing sources of predictability, and the model predictand, to apply a correction
Initialization
Lead time
Valid time

FIG. 1. Modeling time-dependent sources of predictability of an event-based, two-class (yes/no) predictand: (a) An ideal NWP model perfectly represents the sources of predictability and correctly simulates their effect from initialization until valid time. (b) Because of imperfect representations, actual NWP models make conditional errors. This situation generates errors in simulated states and results in erroneous predictions of the predictand (light-gray event probabilities at valid time). (c) As a solution, statistical postprocessing corrects the predictand after the model has run, utilizing the predictand itself and observed predictors that represent sources of predictability in their initial state. Note that the displayed sources of subseasonal predictability are examples of variables found to be important for forecasts of summer 2-m temperature in west and central Europe (discussed later in this study).

hard to initialize correctly (Merryfield et al. 2020), meaning that errors will grow and transfer to S2S time scales (Lorenz 1969). A second reason is that physical interactions need to be approximated numerically. The representations of sources of predictability are therefore imperfect and lead to conditional errors in the forecasts. Figure 1b illustrates how in an actual NWP model, errors from misrepresentations can affect simulated states and the predictand. Any subseasonal steering of the event probability at valid time, that is present in Fig. 1a, becomes biased or absent.

As an example, we know that the prediction of t2m in Europe in the ECMWF model is affected by too-low model soil moisture in spring and summer, which has consequences for the heating of the atmosphere (Dutra et al. 2021). We also know that the decay of Rossby wave packets (RWPs) above Europe is underestimated, resulting in too infrequent blocking (He et al. 2019; Quinting and Vitart 2019). The errors of NWP models can thus relate to highly specific conditions, like the arrival of an RWP or a soil moisture deficit in summer. This means that such conditional sources of predictability can potentially also predict the errors that their misrepresentation will result in. In this study, we hypothesize that conditional errors can, in principle, be predictable, correctable, and better understood when we relate them to the involved sources of predictability.

The idea of correcting errors is not new. Predictable NWP errors have been corrected with statistical methods since the 1970s (Glahn and Lowry 1972). In so-called statistical postprocessing, the estimate of a predictand of interest (in our case, European t2m, valid 2 weeks after the forecast is made) gets corrected after the NWP model has run [see Haupt et al. (2021) and Vannitsem et al. (2021) for current reviews of statistical postprocessing]. As S2S forecasts have attracted attention only recently (Vitart and Robertson 2018), most S2S

Unauthenticated | Downloaded 04/29/25 04:12 AM UTC

postprocessing has been simple, correcting only unconditional errors, i.e., applying the same (e.g., additive) bias correction to each and every forecast (Ferrone et al. 2017; Vigaud et al. 2017; Monhart et al. 2018; van Straaten et al. 2020; Graham et al. 2022).

To correct conditional errors, statistical postprocessing needs to be "weather dependent." This is generally achieved with predictors representing different weather conditions and with methods from machine learning (ML) that are capable of modeling nonlinear relations (e.g., Allen et al. 2019; Schulz and Lerch 2022; Hewson and Pillosu 2021). Postprocessing of S2S forecasts with ML often employs predictors representing large-scale patterns, such as upper-level geopotential height (in the case of Scheuerer et al. 2020; Fan et al. 2023), the El Niño–Southern Oscillation (ENSO) (in the case of Strazzo et al. 2019; Specq and Batté 2020), or large-scale atmospheric circulation patterns (in the case of Manzanas et al. 2018; Lavaysse et al. 2018; Richardson et al. 2020; Mastrantonas et al. 2022). Such predictors represent sources of predictability for a predictand defined at the surface.

The employed predictors are often model predictors, derived from NWP forecasts. For instance, simulated large-scale circulation is used to correct simulated t2m, both more than 2 weeks into the future. Since the predictor also comes from the NWP forecast, such a correction becomes problematic if conditional errors affect both predictor and predictand. In these cases, a better predictor might be found in the initial state of the atmosphere, ocean, and/or land. Figure 1c illustrates how such observed predictors could be used to correct the predictand at valid time.

In this study, we present a statistical postprocessing method that uses both observed predictors at initialization and model predictors at valid time. Relating predictors at initialization to errors at valid time is a novelty compared with many recent postprocessing studies. Most studies feed their ML-based correction method with forecast states only (Rasp and Lerch 2018; Strazzo et al. 2019; Scheuerer et al. 2020; Fan et al. 2023; Veldkamp et al. 2021). An exception is Grönquist et al. (2021), who used states at initialization and +24 h to correct a +48-h forecast. Such an approach has not been applied to correcting conditional forecast errors at S2S lead times. In this paper, we develop an artificial neural network (ANN) that relates predictors from initialization time to errors at valid time, to correct and understand conditional errors in subseasonal ECMWF forecasts of 2-m temperature in Europe during summer. Our ANN (section 3a) is based on the architecture proposed by Scheuerer et al. (2020). We adapt the architecture to directly relate learned conditional corrections to responsible predictors, using explainable artificial intelligence (AI) (XAI). These are tools used for interpreting complex ML models (Mueller et al. 2019; McGovern et al. 2019; Arrieta et al. 2020; Molnar et al. 2021; Toms et al. 2020; Clare et al. 2022) and have been successfully applied in the S2S range (Mayer and Barnes 2021; Gibson et al. 2021; van Straaten et al. 2022).

Relating each type of conditional error to predictors with XAI becomes especially potent in combination with observed predictors. When predictors from the initial state are found to be most predictive of accumulated NWP model errors at valid time, we can deduce that the sources of predictability that they stand for are crucial but imperfectly represented in the NWP model. In addition to improving skill, the postprocessing method presented here can help us understand conditional NWP errors and advance the numerical model representation of poorly understood sources of subseasonal predictability (Vitart and Robertson 2018; Vitart et al. 2019; Merryfield et al. 2020).

2. Data

We postprocess NWP forecasts from ECMWF's Integrated Forecasting System. A single model version, namely, 45r1, is used to avoid changes in systematic model error that occur when different versions are utilized (Bauer et al. 2015). Cycle 45r1 ran operationally from 5 June 2018 to 10 June 2019 and produced a 51-member, 46-day ensemble forecast every Monday and Thursday. Each starting date ECMWF also produced 20 years of reforecasts, by starting additional runs from the same date, but in the 20 previous years. These additional years can be used for investigating model errors and fitting a statistical postprocessing method. As in the operational setting, reforecasts are initialized from an analysis, in this case a reanalysis. The control member starts from the analysis, and 10 other members start from slightly perturbed initial states, according to the estimated initial condition uncertainty (Leutbecher and Palmer 2008). At any valid time, the 11-member ensemble forms a sample from the distribution of possible future states. We merge 1 year of forecasts (after extracting the control member and 10 randomly sampled members) with 21 years of reforecasts to obtain a dataset spanning 22 years, from 1998 to 2019. We focus on summer only [June–August (JJA)] because we want to improve forecasts of anomalously warm conditions during that season.

The model predictand, i.e., the ECMWF forecast that will be postprocessed, is derived from gridded daily t2m. These forecasts are retrieved in a domain encompassing Europe and the North Atlantic, from 20° to 80°N and from −90° to 30°E, at a spatial resolution of 0.32° × 0.32°.

As data for model predictors, we retrieve forecasts of four variables that represent different sources of predictability, in the same domain as above, at a resolution of 1.5° × 1.5°. We retrieve geopotential at 300 hPa (z300) as a representation of the high-pressure "blocking" systems and quasi-stationary Rossby waves that relate to surface temperature extremes (Schaller et al. 2018; Wolf et al. 2018; Kautz et al. 2022). We extract SST because of its potential to influence the North Atlantic jet stream in summer (e.g., Ossó et al. 2020). Last, we extract shallow soil moisture in the top three model layers (0–100 cm) (swvl13) and deep soil moisture from layer four (100–289 cm) (swvl4). Both can influence t2m through the surface energy and water balance (Seneviratne et al. 2010; Miralles et al. 2019).

The data for the observed predictors are retrieved from the ERA5 and ERA5-Land reanalyses (Hersbach et al. 2020; Muñoz-Sabater et al. 2021). These reanalyses are based on assimilated observations and closely correspond to the states

Unauthenticated | Downloaded 04/29/25 04:12 AM UTC

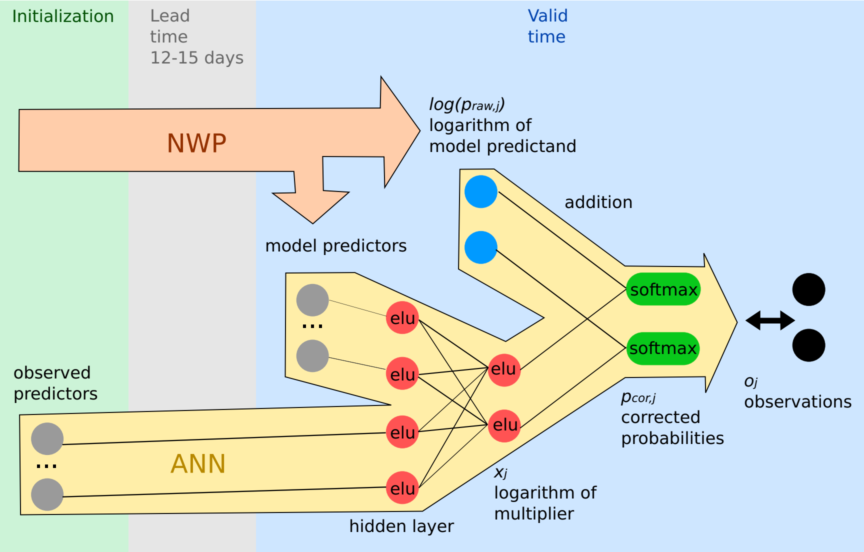


FIG. 2. ANN for postprocessing a two-class probability distribution. The fully connected network is forced to use the distribution forecast by the NWP model (blue) and learns to apply a multiplicative correction (red). The input layer is formed by a mixture of model predictors derived from forecasts at valid time and observed predictors representing the state at initialization (gray). The input layer is connected with single links for illustration purposes only; in reality, all predictors are fully connected to the four-node hidden layer. Activation functions are either elu or softmax and are annotated inside the nodes.

from which reforecasts are initialized. Similar to the model predictors, we retrieve daily gridded values of z300, SST, swvl13, and swvl4. The domain in which they are retrieved is larger (for a reason discussed in section 3d) and happens at a spatial resolution of 0.25° × 0.25° and 0.1° × 0.1°, for ERA5 and ERA5-Land, respectively. For additional observed predictors, we retrieve 850-hPa temperature (T850) and total cloud cover (tcc), respectively related to the low-level heating and clear-sky conditions in summertime blocking systems. We also extract sea ice concentration (siconc) and snow cover (snowc), as they can be relevant for the summertime jet stream position (Hall et al. 2017; Zhang et al. 2020). Last, we retrieve transpiration from vegetation (transp) as an extra indicator of the land surface–water balance. This selection of nine variables is the same as in an earlier study on S2S predictability of European summer temperatures (van Straaten et al. 2022).

From ERA5, we also extract gridded daily t2m, which will later form the "observation" that is used as truth in the training, validation, and verification of the postprocessing method.

For all gridded daily values described above, we subtract the local seasonal cycle. The climatological value per grid point and per day-in-the-year is computed by averaging values from the same day-in-the-year (±5 days) recorded in the 22 yr of the dataset. For the ECMWF model, the average value is computed by pooling all members, but is stratified per lead time to account for gradual drift in the model climatology, which is a known phenomenon in (sub)seasonal forecasts (Johnson et al. 2019). The result is a transformation to gridded daily anomalies relative to the mean seasonal cycle. As Manrique-Suñén et al. (2020) showed, this prevents seasonality to inflate any signals and will lead to a fairer assessment of subseasonal forecast skill.

3. Method

a. ANN-based postprocessing architecture

Our adaptation of the ANN architecture from Scheuerer et al. (2020) involves model predictors from valid time and observed predictors from initialization (gray nodes in Fig. 2). It postprocesses the model predictand at valid time (blue nodes in Fig. 2), which expresses the raw forecast probability that monthly average t2m does not ($p_{\mathrm{raw},0}$), or does ($p_{\mathrm{raw},1}$), exceed a threshold, in a period starting 12–15 days in the future (further details on the predictand in section 3b). These probabilities (blue nodes) are supplied to the neural network right before the final output layer (green nodes in Fig. 2). The correction of the model forecasts is learned in the fully connected part of the network (red nodes in Fig. 2). We now describe this formally.

Let x_j be one of the nodes in an output layer that predicts whether the observation will fall into the nonexceedance ($j = 0$) or exceedance class ($j = 1$). Softmax activation transforms the two nodes to two probabilities that sum to 1:

$$p_j = \frac{\exp(x_j)}{\sum_{k=0}^{1} \exp(x_k)}, \quad j = 0, 1. \tag{1}$$

As Scheuerer et al. (2020) demonstrate, and as is done in our ANN, we can take the logarithm of the model predictand

Unauthenticated | Downloaded 04/29/25 04:12 AM UTC

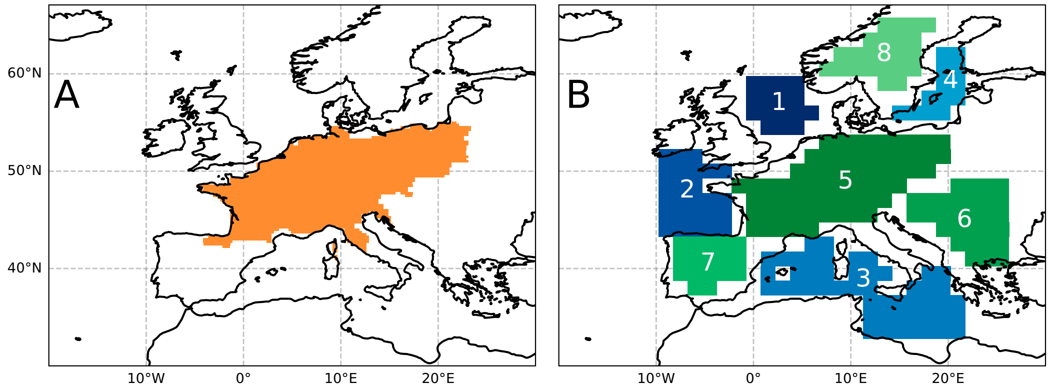


FIG. 3. (a) Region used for spatial averaging of gridded t2m anomalies to obtain the predictand. (b) Regions used to obtain model predictors at the valid time. One statistic is computed to summarize the spatiotemporal state in each region: either the ensemble mean SST anomaly in oceanic regions (in blue), or the ensemble mean swvl13 and swvl4 anomaly in terrestrial regions (in green). Cluster identifiers are annotated.

$p_{\mathrm{raw},j}$ and add it to x_j before softmax activation to obtain the following:

$$p_{\mathrm{cor},j} = \frac{\exp[x_j + \log(p_{\mathrm{raw},j})]}{\sum_{k=0}^{1} \exp[x_k + \log(p_{\mathrm{raw},k})]} = \frac{\exp(x_j)p_{\mathrm{raw},j}}{\sum_{k=0}^{1} \exp(x_k)p_{\mathrm{raw},k}}, \quad j = 0, 1. \tag{2}$$

In comparing Eq. (2) with Eq. (1), we see that the preactivation output x_j learned by the fully connected network (red in Fig. 2) can be interpreted as the logarithm of a weather-dependent multiplier of ECMWF's prior probability $p_{\mathrm{raw},j}$. After activation, one obtains the postprocessed distribution $p_{\mathrm{cor},j}$ (green in Fig. 2). The fact that multiplier x_j is a complex function of the predictors can be used to understand the physical circumstances leading up to each conditional correction and thus the weather dependence of forecast errors. In section 3h, we explain how the multiplier's value gets attributed with XAI.

On a sidenote, we concede that two output nodes can seem redundant for a two-class prediction, as the same might be achieved with one output node and sigmoid activation. But we prefer this general architecture because multiclass distributions can be obtained by adding more nodes.

b. *Predictand*

The choice of the predictand in S2S forecasting is a compromise between the desire for detail and the limited predictability in such details. Local daily values of t2m in Europe are hardly predictable at lead times beyond 2 weeks (van Straaten et al. 2020). This means that a certain amount of spatial and temporal aggregation is always needed (Shukla 1981; Roads 1986; Wheeler et al. 2017). Aggregated values represent the mean effect that sources of predictability have on daily t2m values that by themselves would be unpredictable. In this study, we limit ourselves to monthly (31 day) average t2m, averaged over a west and central European region (Fig. 3a), exceeding a certain quantile. We opt for relatively long, monthly averages, because an earlier analysis showed that for this region, t2m averaged over weeks 3, 4, 5, and 6 after initialization showed better predictability than a shorter-term average over weeks 3 and 4 (van Straaten et al. 2022). Apparently, the predictable process is better detectable on the longer time scale. We focus on a lead time range of 12–15 days before the start of the 31-day period, as combining multiple lead times gives us more samples to work with than a single lead time, and as 15 days is the maximum lead time in the 46-day ECMWF forecasts.

To derive the two-class probability estimate $p_{\mathrm{raw},j}$ from the ECMWF ensemble members, and the corresponding binary observation o_j from the ERA5 reanalysis, we apply three similar methodological steps to both datasets. First, the spatial average of gridded t2m anomalies in the region is taken. Second, 31-day temporal averaging is applied as a rolling window. Third, we estimate climatological quantile levels $q \in 0.5, 0.66, 0.75$, and 0.9, of which one serves as the exceedance threshold. Quantile estimation is done separately for forecasts and reanalysis and occurs per day-in-the-year (±15 days) to account for seasonality. The ±15-day window is larger than the ±5-day window used for computing anomalies (section 2) because estimating high quantiles is more susceptible to noise than estimating a mean. For the ECMWF model, the quantile estimation is also stratified per lead time, thereby implicitly correcting for drift in its climatological spread.

After threshold estimation, the 11-member ensemble ($M = 11$) is transformed to $p_{\mathrm{raw},1}$ by counting the number of members m above the threshold in a Tukey plotting position estimator (Wilks 2011):

$$p_{\mathrm{raw},1} = \frac{m + 1 - a}{M + 2 - 2a}, \tag{3}$$

where $a = 1/3$. Accordingly, $p_{\mathrm{raw},0} = 1 - p_{\mathrm{raw},1}$. We prefer the Tukey estimator over dividing m by M (also known as "democratic

Unauthenticated | Downloaded 04/29/25 04:12 AM UTC

voting"; $a = 1$), because the former leads to slightly higher forecast skill (Ferrone et al. 2017). It also prevents probabilities that are zero, which is required for transforming $p_{\mathrm{raw},j}$ to logarithms [Eq. (2)].

c. Model predictors

Model predictors are one of two types available to the ANN. The model predictors are derived from simulated SST, swvl13, swvl4, and z300 (section 2) at valid time: a period that starts 12–15 days into the future. For each of these variables, we capture distinct sources of predictability with tailored summary statistics, such that one predictor/statistic summarizes one spatiotemporal state that can relate to t2m in our target region (and to its forecast errors).

At valid time, predicted SST and swvl13 and swvl4 anomalies are simultaneous in time to the predictand. This means that they do not influence the predictand in a temporally lagged and spatially remote fashion, but that the dominant influence will be more local instead. As relevant spatiotemporal states, we thus extract 21- and 31-day temporal and spatial averages in regions close by or inside the predictand region (Fig. 3b). As summary statistic, we only extract the ensemble mean prediction, as the additional use of ensemble spread does not always improve postprocessing skill (Rasp and Lerch 2018; Schulz and Lerch 2022). With four regions for SST, four regions for swvl13 and swvl4 (Fig. 3b), and two time scales, this results in 24 predictors.

Predictors from z300 are obtained differently. The circulation at this upper-tropospheric level is known to exhibit recurring synoptic configurations over extended periods of time that have specific imprints on the surface weather (Hannachi et al. 2017; Casanueva et al. 2014; Grotjahn et al. 2016). Among these "regimes" are the ridges or blocks responsible for summer heat (Cassou et al. 2005; Pfahl 2014; Sousa et al. 2018; Kueh and Lin 2020; Kautz et al. 2022). Consequently, successful S2S prediction of surface weather is sometimes possible when forecast tropospheric circulation gets classified into appropriate regimes (Lavaysse et al. 2018; Richardson et al. 2020; Mastrantonas et al. 2022). Here, we obtain predictors summarizing the appropriate spatiotemporal states by classifying predicted z300 anomalies into four distinct regimes (details in appendix A). Recording the predicted frequency of each regime in a 21- and 31-day period results in 8 additional predictors.

d. Observed predictors

The second type of predictors that we supply to the ANN are observed predictors at initial time. These represent the initial state of atmosphere, ocean, and land, unaffected by NWP model errors.

The domains from which these predictors are obtained are larger than above because initialization occurs 12–15 days before our predictand period. In those two intervening weeks, midlatitude circulation can be influenced by short-lived heating from as far as the tropics (Branstator 2014). The lagged influence of a subseasonal source of predictability is not only local. Signals originating in tropical Pacific Ocean SST have long been known to affect the distribution of air masses elsewhere (Bjerknes 1969). The tropics are therefore included in the SST domain, and since such teleconnections not only originate from SST but also from other variables such as snow cover and sea ice (Hall et al. 2017; Zhang et al. 2020), nine reanalysis variables are taken into account (section 2). These variables and the associated domains are taken from van Straaten et al. (2022, their Fig. 1).

To capture the spatiotemporal states of the different sources of predictability in those domains, we first identify distinct regions in which gridded anomalies from the nine variables relate to our predictand of interest (Bello et al. 2015; Kretschmer et al. 2017). These regions resulted from the analysis of van Straaten et al. (2022), in which we computed the lagged correlation between the grid cells of each variable and the average t2m in the predictand region, starting 12–15 days later. Temporally, the correlation was performed on average states ranging from 1-day averages to 31-day averages of the daily gridded anomalies in order to extend the range of possibly useful time scales beyond the monthly time scale of the predictand. Spatially, groups of neighboring, significantly correlated grid cells were then clustered into feature regions. Assuming that each of those represented a distinct source of predictability, two statistics were extracted per region to summarize its spatiotemporal state: the spatial mean anomaly and a spatial covariance measuring the resemblance between the reanalysis state and the correlation pattern that was clustered (more details in van Straaten et al. 2022). With two statistics, a variable number of feature regions, nine variables, and eight scales of temporal averaging, the procedure results in 226 predictors.

As final observed predictors, we also include the state of the Madden–Julian oscillation (MJO), 12–15 days before our target period. This tropical oscillation is a dominant mode of subseasonal variability (Zhang 2013) and is known to conditionally affect atmospheric circulation over the Euro-Atlantic region (Cassou 2008; Lin et al. 2009). In NWP models, the amplitude of the connection is often found to be too weak (Vitart 2017). We use the daily two-component real-time multivariate MJO (RMM) index as supplied by the Australian Bureau of Meteorology (Wheeler and Hendon 2004). This adds two further predictors and results in a grand total of 260 (model and observed) predictors, which are subject to a predictor selection described in section 3g.

e. Benchmark models

With its mixture of predictors, the ANN corrects raw ECMWF probabilities p_{raw}. The skill of corrected probabilities p_{cor} is evaluated against p_{raw} itself and benchmarks that are based on ERA5 t2m data only.

Our first benchmark to compare against uses the threshold quantile level q. Exceedance of the 0.75 quantile happens (on average) in 25% of the data. Assuming stationarity, we create a constant probability forecast valid for each sample:

$$p_{\mathrm{constant},1} = 1 - q. \tag{4}$$

However, when climate is changing, a comparison against a stationary benchmark can artificially inflate skill (Hamill and

Unauthenticated | Downloaded 04/29/25 04:12 AM UTC

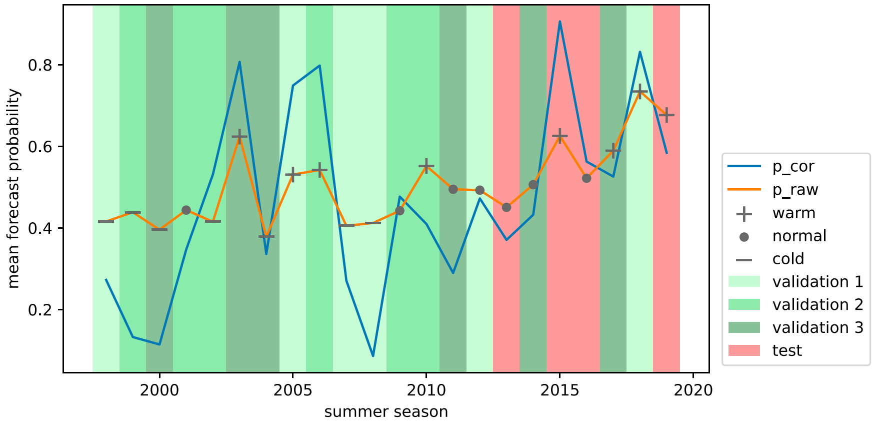


FIG. 4. Summer average probability that monthly temperature exceeds the 0.5 quantile, forecast with a lead time of 12–15 days before the start of the period. The dataset of 22 summer seasons is divided into subsets that balance the amount of warm, normal, and cold summers (tercile classes of p_{raw}; symbols). Test set summers are colored red, and the three fold cross-validated sets are colored green, of which two folds are used for training and the remaining one is used for optimizing the model. Raw ECMWF forecasts are indicated by the orange line; probabilities from an optimized postprocessing model are indicated by the blue line (p_{cor}, trained on all cross-validation folds).

Juras 2006; Manrique-Suñén et al. 2020; Wulff et al. 2022). Anthropogenic global warming makes exceedance of a stationary threshold less likely in the early part of the climatological period, and more likely at the end of it. A more competitive nonstationary benchmark takes that into account, which is why a second benchmark models the climate change–driven probability with logistic regression. p_{trend} is a function of time t (yearday) and regression coefficients β and γ:

$$p_{\mathrm{trend},1} = \frac{1}{1 + e^{-(\beta+\gamma t)}}. \tag{5}$$

Such a trend model signifies the climatologically expected probability of exceedance, not conditioned on any weather information. This means that we can introduce that conditioning by adjusting the expected probability up and down, based on a set of predictors, similar to adjustments made in our postprocessing of model-predicted probabilities. In the third benchmark model, we therefore supply $\log(p_{\mathrm{trend}})$ instead of $\log(p_{\mathrm{raw}})$ to the ANN as prior probability estimate (blue nodes in Fig. 2) and leave everything else the same. The resulting benchmark forecasts are called $p_{\mathrm{trend+ann}}$.

f. Performance metrics

We verify the probabilistic exceedance forecasts $p_{[\mathrm{raw,cor,constant,trend,trend+ann}],1}$ against binary observations o_1. With the Brier score (BS), we compute the mean squared error over the n forecast–observation pairs: $\mathrm{BS} = (1/n)\sum_{i=1}^{n}(p_{i,1} - o_{i,1})^2$ (Brier 1950). Additionally, we use reliability diagrams to visually assess the forecast reliability and resolution (Wilks 2011).

We also use the area under the receiver operating characteristic (ROC) curve (AUC). For this score, the forecasts are discretized for a range of probability thresholds, each providing a contingency table with true and false positives, and true and false negatives (respectively TP, FP, TN, and FN). This leads to a ROC diagram with false-alarm rates [FP/(FP + TN)] on the x axis and hit rates [TP/(TP + FN)] on the y axis. For each set of probability forecasts, AUC quantifies whether increases in p discriminate observed occurrences from nonoccurrences. AUC is, however, insensitive to the magnitude of the increase and therefore, only forms a measure of the forecast's potential usability (Kharin and Zwiers 2003). We also compute the Hanssen–Kuipers or Pierce score (PS), which is the hit rate minus the false-alarm rate. It expresses the maximal potential economic value a reliable forecast can have to users making yes-or-no decisions (Richardson 2000).

All numerical scores $S \in \{\mathrm{BS, AUC, PS}\}$ are transformed to skill scores [Brier skill score (BSS), AUC skill score (AUCSS), Pierce skill score (PSS)] by normalizing the difference between S and S_{constant}, the score computed for the constant benchmark model [Eq. (4)]:

$$\mathrm{SS} = \frac{S_{\mathrm{constant}} - S}{S_{\mathrm{constant}} - S_{\mathrm{perfect}}}, \tag{6}$$

where S_{perfect} is {0, 1, 1}, respectively.

g. Data partitioning and ANN tuning

We measure the performance of our ANN-based postprocessing method on independent test data. For tuning, the performance on a validation set is used. We split the set of 22 summers into four seasons for testing and 18 seasons for cross validation (Fig. 4). This means that the 18 seasons are further subdivided into three folds with 6 seasons each. Repeatedly, a model is trained on two of them and makes predictions for the third. After three repeats, the concatenated validation predictions are assessed for performance and used to select predictors and choose hyperparameter values.

Unauthenticated | Downloaded 04/29/25 04:12 AM UTC

The data splits are subject to two criteria. First, our observed predictors from initialization are based on clustered correlation patterns. These correlations were computed on data from the period 1981–2013 (more details in van Straaten et al. 2022). For independent testing, the test set should thus be after 2013. Second, all sets need to contain a balanced mix of cold, normal, and warm seasons, such that each is representative of the data distribution. Given that p_{raw} displays a warming trend (Fig. 4), an equal balance cannot be achieved with chronological splits. A training set from the period 1998–2004 would be dominated by lower-tercile or "cold" seasons (minus symbols in Fig. 4). So instead, we split the data such that each set contains seasons from all tercile classes (result visible in Fig. 4).

Note that although we use terciles to partition the data, the predictand is still a two-class variable, with a separate ANN being trained for each quantile exceedance threshold $q \in 0.5, 0.66, 0.75,$ and 0.9. For each of these ANNs, we select a set of l predictors based on combined validation performance. First, each of the 260 predictors was scaled to lie between 0 and 1 (validation and test sets remain unseen). Then, a greedy stepwise forward algorithm was applied. Suppose $l - 1$ predictors are already in use. Add an unused candidate, train the model on two cross-validation folds, make predictions for the remaining fold, and repeat the training and prediction to also cover the other folds. Compute the ranked probability score (RPS) $(1/n)\sum_{i=1}^{n}\sum_{j=0}^{1}(p_{i,j} - o_{i,j})^2$ over the n concatenated validation samples and the two classes j. Repeat the steps above three times to account for randomness due to weight initialization. Record the scores, switch the candidate predictor for a new one, and proceed from the start until all predictors are tested. Choose the one leading to the lowest mean RPS as the lth predictor. Continue to find the $l + 1$ most important predictor, but only if validation RPS decreases by more than 6%. This subjectively chosen stopping criterion resulted in 2 to 4 selected predictors per ANN.

Each ANN is trained with exponential linear unit (elu) activation functions, categorical cross-entropy loss, and the Adam algorithm for gradient descent (Kingma and Ba 2014). Tuning was required for (i) the number of hidden layers, (ii) the number of nodes per hidden layer, (iii) the learning rate, (iv) the batch size, and (v) the number of epochs for which we tolerate an increasing validation loss before stopping training early. Settings for these parameters were explored using a large random search by means of the SHERPA python package (Hertel et al. 2020). The number of trials was set to 200. Combinations leading to the lowest RPS over the three cross-validation folds were chosen (repeated eight times to mitigate randomness due to weight initialization). For all quantile thresholds, a simple model was the result, namely, with 1 hidden layer, 4 hidden-layer nodes, a learning rate of 0.0014, a batch size of 32, and a patience of 7 epochs before stopping training early (see the architecture in Fig. 2). With these hyperparameters and after predictor selection, each ANN is trained one final time on the combined cross-validation folds, with 33% of the data left out for the early stopping algorithm.

h. *Explainable AI*

Besides improving forecasts by postprocessing p_{raw}, we also want to understand the conditional NWP errors that the ANN corrects. Predictor contributions can be used to learn about the physical circumstances of each error.

After the ANNs have been trained, we apply two XAI techniques to attribute the learned corrections to the 2–4 predictors involved. Recall that the learned correction factor $\exp(x_1)$ is a weather-dependent multiplier of $p_{\mathrm{raw},1}$ [Eq. (2)]. Predictor contributions to this factor will thus vary from forecast to forecast, depending on the state of the sources of predictability that the observed and model predictors represent. Positive predictor contributions signify conditions in which $p_{\mathrm{raw},1}$ is usually an underestimation, meaning the ANN attempts to increase the probability of exceedance. Negative contributions signify conditions in which $p_{\mathrm{raw},1}$ is an overestimation.

The first method with which we quantify contributions is the model-agnostic Shapley additive value (SHAP) estimator KernelSHAP (Lundberg and Lee 2017). These values originate from game theory and are solutions to the problem of dividing a game's single payout over multiple contributing players. In statistical modeling, the equivalent situation is the division of a single predicted value over the contributions from each predictor. As these contributions are defined relative to a "normal" background, they can be negative or positive. Together with the background expectation, they will add up to the predicted value, in this case the multiplication factor. Determining a background from all samples is computationally demanding so we reduce the full dataset to 30 representative centroids with k-means and compute the background from those. SHAP values are then computed for each train and test sample.

The second XAI method is specific to differentiable models like neural networks. We quantify the gradient of the fully connected output x_1 with respect to the predictor values. Such a gradient is measured in the vicinity of the predicted value and represents a sensitivity. However, sensitivity does not always entail relevance. For instance, daytime t2m can be sensitive to cloud cover, with clearer skies leading to more solar irradiance and surface heating. But on a persistently cloudy day, the real cause of high t2m can be a process like warm air advection, to which it is also sensitive. Actual relevance is better approximated by multiplying the gradient by the predictor value, rather than using the gradient itself. This so-called input times gradient has shown good reconstruction of a known ground truth in a climate-like prediction problem (Mamalakis et al. 2022).

4. Results

a. *The benefit of postprocessing*

The ANN-based postprocessed forecast p_{cor} is supposed to improve upon the NWP forecast p_{raw} by using information from both initialization and valid time. Skill scores for the two, and other benchmarks, are presented in Fig. 5. Extremity of the predictand varies along the x axis, showing the quantile used as exceedance threshold in monthly temperature (section 3b). One ANN is fitted per quantile, to the combined lead times of 12–15 days before the start of the predictand period. A separate predictor selection is performed for each ANN.

Unauthenticated | Downloaded 04/29/25 04:12 AM UTC

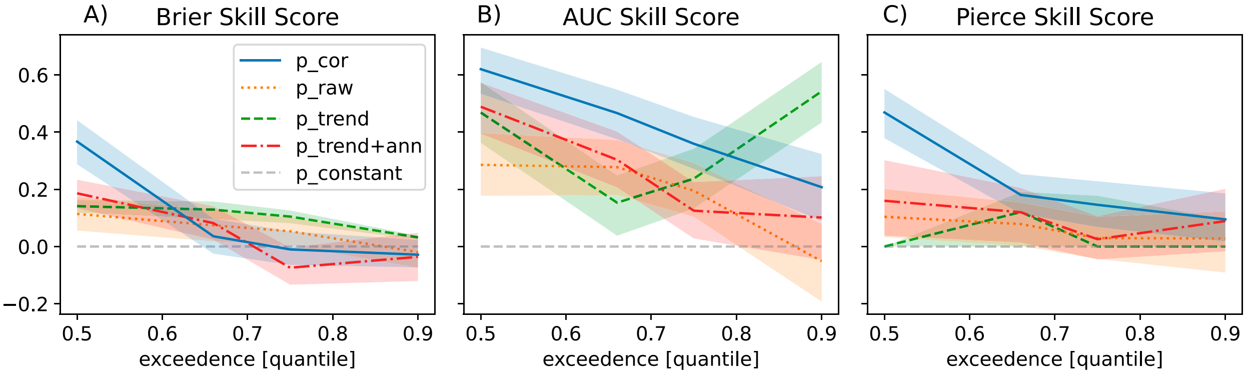


FIG. 5. Test set performance according to three probabilistic skill scores: (a) BSS, (b) AUCSS, and (c) PSS. The predictand is monthly average European temperature exceeding a quantile threshold (*x* axis). One ANN is fitted per exceedance threshold and to the combined lead times of 12–15 days before the start of the monthly period. The p_{cor} is the result of postprocessing p_{raw} with the ANN; $p_{\mathrm{trend+ann}}$ is the result of adjusting p_{trend}, also with the ANN. Comparison is made with the p_{raw} and p_{trend} themselves. Here, p_{constant} is the zero skill reference. Shading denotes the 5th–95th-percentile uncertainty bounds, as obtained by bootstrapping forecast–observation pairs before computation of the scores (1000 repeats).

In Fig. 5, skill scores generally decrease with increasing exceedance threshold. The forecasts p_{cor} (solid blue) outperform all benchmarks in all three scores for median exceedance, and only in AUCSS and PSS for the 0.66 and 0.75 quantile. For the 0.9 quantile, the AUCSS of the trend benchmark drastically increases (dashed green, Fig. 5b). That behavior might be caused by AUC's trapezoidal approximation, leading to distorted results for extremes (Ben Bouallègue and Richardson 2022). Comparing the scores, BSS can seem overly conservative. However, the Brier score is a proper score (Gneiting and Raftery 2007) and it is common that skill for 90th percentile events does not extend beyond 2 weeks after initialization (Lavaysse et al. 2019). Absence of skill can come from ANN's inability to learn the right conditional corrections in a limited set of samples, or from an intrinsic lack of predictability in these extremes.

For median exceedances, it is clear that p_{cor} provides the most skillful forecasts (solid blue in Fig. 5), also when comparing with $p_{\mathrm{trend+ann}}$ (dash–dotted red). The latter ANN uses p_{trend} as prior probability estimate instead of p_{raw} (p_{raw} is, in that case, supplied as a regular predictor, but never selected). The difference between the two does not stem from their trend awareness, as p_{raw} is trended as well (Fig. 4) (see also Shao et al. 2022). Also, both are equally weak when unadjusted (dotted orange and dashed green, Fig. 5a). The skill gained by p_{cor} over $p_{\mathrm{trend+ann}}$ therefore shows that p_{raw} forecasts from ECMWF contain a predictive value that emerges when its shortcomings are corrected. In the remainder of the paper, we focus on the ANN-based p_{cor} forecasts for median exceedance.

A more complete look at the performance of p_{cor} is given by reliability diagrams (Fig. 6). On the training set, the reliability achieved by postprocessing is near perfect (solid blue in Fig. 6a). This is obvious because the ANN trains to mimic the observations, until it is stopped early by increasing validation loss. The early stopping is visible in the fact that p_{cor} is not concentrated at 0 and 1 (which would be a perfect mimicry), but that the issued probabilities of exceedance cover a range in between (solid blue in Fig. 6b). The ANN is able to issue sharper forecasts (i.e., more probabilities close to 0 and 1) than p_{raw} (dotted orange in Fig. 6b). These improvements transfer to the unseen test data, as p_{cor} lies closer to the 1:1 line and results in more probabilities close to 1 than p_{raw} (Figs. 6c,d).

b. *Selected predictors*

The ANN-based forecast presented in Fig. 6 is made with a set of three predictors, besides the model predictand p_{raw} that is being corrected. Figure 7 depicts the top 20 predictors that the forward predictor selection found. Only the top 3 decreased RPS on the combined validation folds by more than 6% and were therefore included in the ANN (marked blue dots in bottom row of Fig. 7). The figure shows whether the predictors are observed or model predicted (top row), whether they relate to atmospheric, oceanic, or land surface variability (second row), and at what time scale they are defined (third row).

The highest-ranked predictor is 21-day SST from ERA5. It represents the state of SST in the western equatorial Pacific at initialization (see the bottom source of predictability mapped in Fig. 1). In the construction of observed predictors, both the mean anomaly and spatial covariance were extracted, but it is the spatial-covariance statistic that is selected here (we will discuss what it represents in section 4c). The primacy of this predictor is confirmed by two other methods of predictor selection in which it also appeared first (not shown). The second predictor is the model-predicted 31-day average SST in the North Sea (region 1, Fig. 3b), and therefore comes from valid time. The third predictor is ERA5 850-hPa temperature at initialization, specifically the mean 21-day anomaly in a subtropical region stretching from the Atlantic, over the Sahel to the Indian Ocean (see the upper source of predictability mapped in Fig. 1).

Unauthenticated | Downloaded 04/29/25 04:12 AM UTC

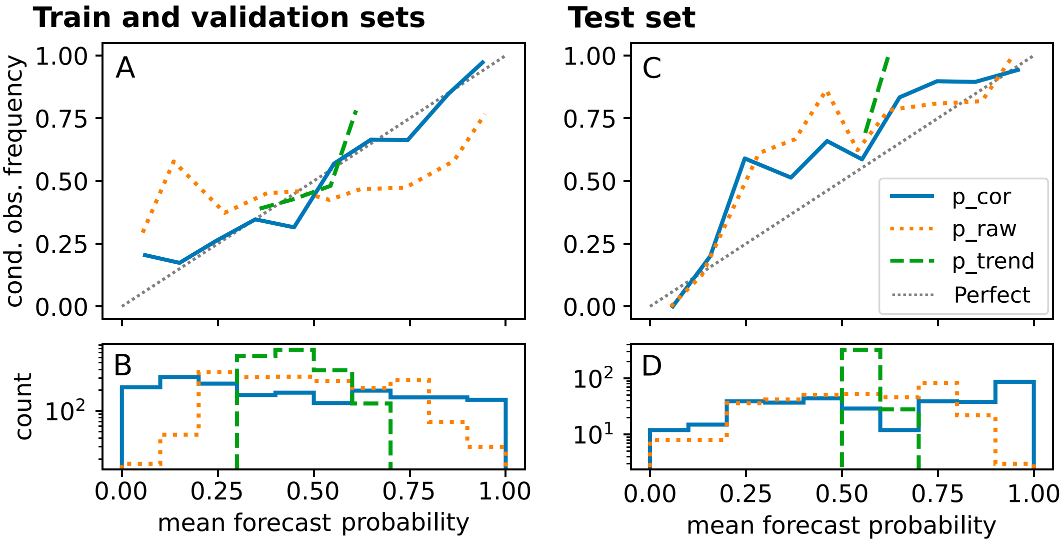


FIG. 6. Reliability diagrams of ANN-based postprocessing of monthly t2m exceeding q0.5 (blue), measured against raw NWP forecasts (orange) and the trended benchmark (green). Lead time is 12–15 days before the start of the monthly period. (left) Performance on the combined cross-validation sets (green summers in Fig. 4). (right) Test set. Shown are (a),(c) reliability diagrams of the conditional observed frequency per bin vs binned forecast probabilities (the dotted 1:1 line shows perfect reliability) and (b),(d) histograms of binned forecast probabilities, where concentration close to 0 and 1 hints at high forecast sharpness.

Two of the predictors responsible for the corrections represent an observed state at initialization time. The relation of these sources of predictability to NWP errors suggests that their effects are potentially misrepresented by the NWP model but can be corrected. Only one predictor, namely, North Sea SST, is a model-predicted state at valid time. Remarkably, model-predicted candidates like soil moisture and circulation regime are not selected. Their influence on t2m is either already properly resolved in the ECMWF model, meaning they do not relate to systematic t2m errors compared with reanalysis, or they themselves are biased or lack predictability at these lead times. Either reason would render them useless for postprocessing (this topic is further discussed in appendix A).

c. *Understanding forecast errors*

Next, we aim at understanding conditional NWP errors, by attributing the ANN's corrections to the three predictors. Figure 8a displays the three predictors and their Spearman rank correlation with p_{raw}, the correction ($p_{\text{cor}} - p_{\text{raw}}$), and p_{cor}. The tropical west Pacific SST predictor shows a weak positive correlation with p_{raw} and a strong positive correlation with the correction and p_{cor} (Fig. 8a, first row). This amplification by the ANN indicates that the connection in the ECMWF model between this initial SST pattern and our monthly predictand is weaker than it should be.

Of the three predictors, North Sea SST is most positively correlated with p_{raw} (Fig. 8a, first column). This is understandable given the geographical proximity of the North Sea to the predictand region and the fact that simulated cold or warm spells are likely to extend over both (Fig. 3). Direct correlation between this predictor and the correction is close to zero, meaning that the role of North Sea SST in the ANN is likely nonlinear and not captured by a monotonic correlation metric. Average absolute SHAP values (interpretable as the average relevance over all samples) show that it is an important predictor (Fig. 8b). It is more important than subtropical

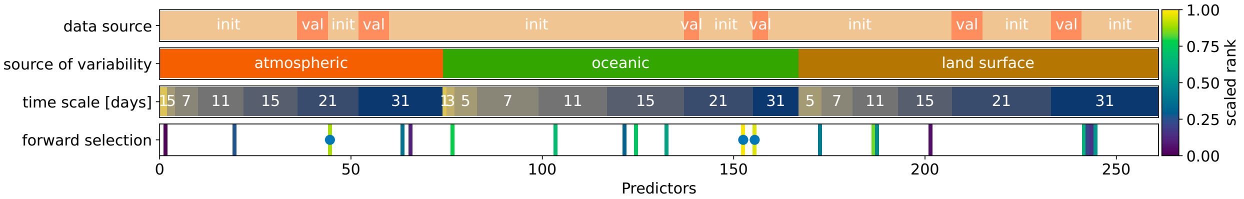


FIG. 7. Characteristics of predictors (*x* axis) selected for the ANN-based postprocessing of monthly temperature exceeding q0.5 (lead time of 12–15 days before the start of the monthly period). The bottom row shows the rank of the top-20 forward selected predictors. Marked by blue dots are the top three predictors that decrease RPS by more than 6%. The top row shows the data source [whether observed at initialization ("init") or model predicted at valid time ("val")]. The two middle rows respectively show the source of variability and the time scale of the predictors (i.e., the average in days before the lead-time gap for observed predictors, and the average in days after the gap for model predictors).

Unauthenticated | Downloaded 04/29/25 04:12 AM UTC

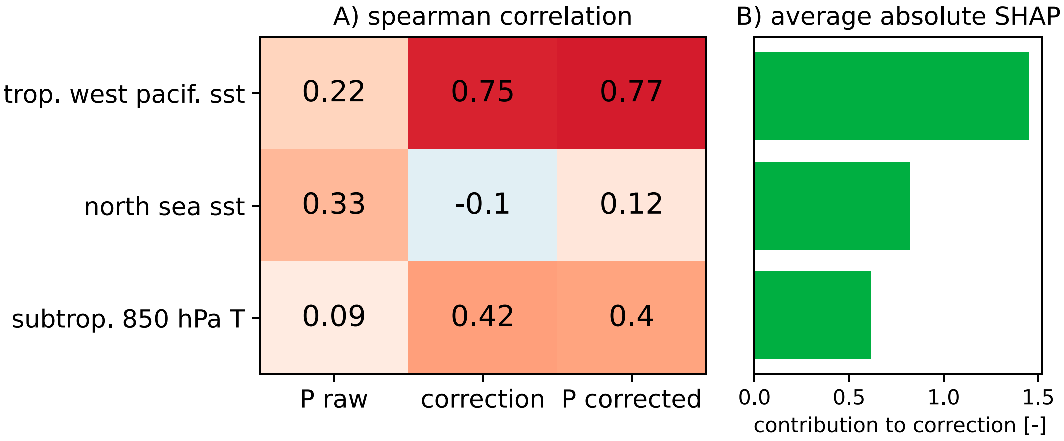


FIG. 8. Summary of the role of the three selected predictors in postprocessing monthly temperature > q0.5 with a lead time of 12–15 days before the start of the monthly period: (a) Spearman rank correlation with p_{raw} from ECMWF, with the correction by the ANN ($p_{\mathrm{cor}} - p_{\mathrm{raw}}$), and with postprocessed probability (p_{cor}). (b) Importance for the learned multiplicative correction as measured with the absolute SHAP value over all samples.

T850, even though the latter shows higher correlation with the correction (Fig. 8a, second column). We will see later that this is because predicted North Sea SST plays a modulating, conditional role.

Figure 9 shows the contributions from each predictor per corrected forecast. The applied correction is in the top row, with SHAP contributions and input times gradient contributions below. As we are interested in the physical circumstances of different errors, we reorder the samples along the x axis. The order is the so-called leaf order of a hierarchical clustering algorithm (applied to the SHAP values with a Euclidian distance metric). This order refers to the lowest level in the hierarchy, where each sample is still its own cluster and the algorithm has shuffled similar samples close to each other. We perform this grouping because NWP errors, and therefore, the learned corrections, are highly conditional. This means that the physical circumstances can only be understood when looking at similar corrections that are applied for the same reason. To achieve this, samples should not be grouped on predictor values only, as that would weigh west Pacific SST and T850 equally, even though their importance in the correction differs (Fig. 8b). Neither can we group samples by their correction factor only, as that would mix samples with similar corrections that are applied for very different reasons. Hierarchical clustering of SHAP values reconciles these two approaches, as predictor values are replaced by predictor contributions to the

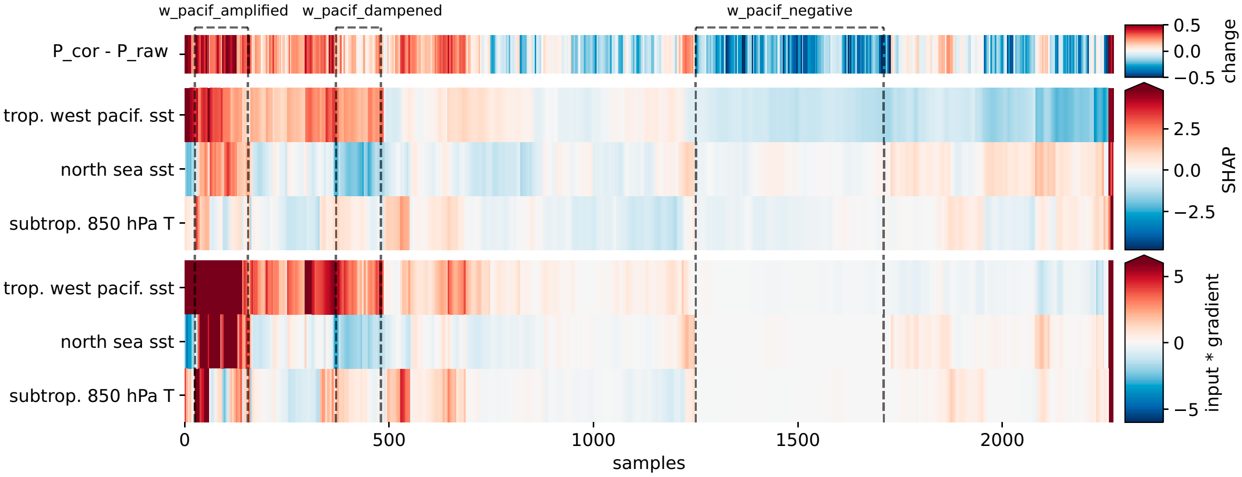


FIG. 9. Contributions of the three predictors to the ANN-based correction of forecasts of monthly temperature above the median, made with a lead time of 12–15 days before the start of the period. The top row displays the change in probability applied by the ANN ($p_{\mathrm{cor}} - p_{\mathrm{raw}}$). The next three rows show contributions as quantified by SHAP (summing up to the multiplicative correction factor). The bottom three rows are contributions as quantified by input times gradient (inputs are standardized instead of minimum–maximum scaled between 0 and 1). Samples (x axis) are sorted by the leaf order that results from a hierarchical clustering of SHAP values so that situations requiring similar corrections for the same reason are close to each other.

Unauthenticated | Downloaded 04/29/25 04:12 AM UTC

output and are in units that sum up to the applied correction. Lundberg et al. (2020) showed that this makes the clustering supervised and weighs variables according to their impact.

The hierarchical reordering enables us to distinguish different groups by eye. We distinguish and annotate three groups (Fig. 9). In "west Pacific amplified" circumstances are such that the tropical west Pacific and North Sea SST predictors contribute positively to the correction (both red), with a large increase from p_{raw} to p_{cor} as a consequence. In "west Pacific dampened," the west Pacific contributes positively, but North Sea SST contributes negatively (blue). The eventual applied correction is close to zero. In "west Pacific negative," the west Pacific contributes negatively, while other predictors remain neutral, leading to distinct negative corrections (more visible in SHAP than in input times gradient).

Note that the sign of the T850 contribution varies within the delineated groups, which means that the hierarchical algorithm considers T850 of lesser importance for finding similar physical circumstances. Quantitatively, it is also a less important predictor than west Pacific and North Sea SST (Fig. 8b). In the remainder of the paper, we therefore focus on the two SST predictors.

The three identified correction groups point to different physical circumstances that we investigate by creating a composite of the samples in each group, for the variables z300, SST, swvl13, and t2m. Figure 10 displays these variables (in rows) at different moments in time (columns), namely, the analysis at initialization time (first column), the analysis at valid time (second column), and the ensemble mean forecast at valid time (third column).

In "west Pacific amplified," the development of high pressure (and high t2m) over west and central Europe is underestimated by the ECMWF forecasts. The ERA5 data show that z300 from initialization onward (Fig. 10a) develops into a zonal quasi-stationary wave pattern, with a core of high pressure over western Europe, flanked by two low pressure regions (Fig. 10b). The raw forecasts misplace the high pressure core over the North Atlantic (Fig. 10c). In ERA5, the heat situated over the Iberian Peninsula expands into west and central Europe (Figs. 10j,k). Comparatively, the forecast of t2m in the target region is too cold (Fig. 10l). Since the ECMWF model retains a correct soil moisture pattern (Figs. 10g–i), it seems that the missed or misplaced atmospheric wave is the prime reason for the ANN to apply a strong upward correction in these situations (top row, Fig. 9). As noted, the west Pacific SST predictor contributes positively to this correction. As a whole, tropical Pacific SST resembles an El Niño state (Figs. 10d,e).

An informative contrast appears with "west Pacific negative," where Pacific SST shows an opposite, La Niña–like pattern (Figs. 10ab,ac). Again, an initial z300 pattern develops into a wavelike pattern with strong anomalies (Figs. 10y,z). Low pressure is now situated over the British Isles and high pressure over western Russia (Fig. 10z). For the predictand region, this means that soils wetten (Figs. 10ae,af) and that heat accumulates east of it, over western Russia (Fig. 10ai). The ECMWF model does not capture this pattern, as it places the wettening of soils around the Mediterranean (Fig. 10ag). Moreover, the low pressure system is too weak and located too far to the west over the Atlantic (Fig. 10aa). The slightly cool t2m anomaly in the predictand region and the warm anomaly over Russia (Fig. 10ai) are missing in the forecasts (Fig. 10aj), which is why the ANN learns to decrease the p_{raw} exceedance probability in this group of samples (Fig. 9, top row).

In the physical circumstances of "west Pacific amplified" and "west Pacific negative," the ECMWF model thus appears unable to simulate a developing atmospheric wave, associated with west/central European heat in the former and Russian heat in the latter. The ANN learns that the necessary, opposing corrections are predicted by west Pacific SST. Figure 11 summarizes the NWP forecast errors for the three groups ("west Pacific dampened" will be discussed later) and the state of the west Pacific predictor. As a covariance predictor, it measures correspondence between anomalies (red–blue shading) and correlation pattern (green–purple contours) in the grid cells that correlate significantly to European t2m (see also section 3d). Positively correlated cells lie at the western edge of the Niño-4 region (0°, 160°E) (Trenberth and Stepaniak 2001). Negatively correlated cells lie at 10° and 20°N (green contours in Fig. 11a). In "west Pacific amplified," the positive–negative correlation dipole is appearing as a hot–cold dipole in the anomalies. This leads to positive covariance (0.027; Fig. 11a). In "west Pacific negative," the hot–cold dipole in the anomalies is inverted, leading to negative covariance (−0.033; Fig. 11c). The (inversion of the) pattern predicts the under- (over)estimation in p_{raw}, as compared with the observed frequency of exceedance (Figs. 11e,g).

Based on the composite SST pattern, we infer that the west Pacific hot–cold dipole consists of an ENSO-like source of predictability, with a contrast between SST anomalies in the central tropical Pacific and those in the Maritime Continent. It is a source of predictability for the development of an atmospheric wave in the reanalysis and errors at valid time in the ECMWF model. Some aspect of the predictable physical pathway is thus misrepresented. Indeed, tropical convection patterns, closely related to ENSO, are known to influence Euro-Atlantic summer circulation (Ding et al. 2011; O'Reilly et al. 2018; Ma and Franzke 2021). Extratropical circulation gets affected by patterns of tropical convection, leading to diabatic heating and upper-level divergence, which forces Rossby waves (Bjerknes 1969; Sardeshmukh and Hoskins 1988; Ting 1994; Trenberth et al. 1998). Aspects of such teleconnection responses, like the propagation of Rossby waves, are known to be misrepresented in NWP models (Beverley et al. 2019) [see O'Reilly et al. (2018) and Strazzo et al. (2019) for misrepresentation of ENSO-like teleconnections].

Our results, however, also reveal that the NWP error is conditional. Corrections in the "west Pacific dampened" group suggest that the error can be absent. In these samples, the west Pacific SST dipole is in a positive state (covariance of 0.033, Fig. 11b), but predicted p_{raw} is already close to the observed frequency of exceedance (Fig. 11f). Looking at Fig. 10, we see that a strong Scandinavian blocking is present at initialization (Fig. 10m), causing high t2m anomalies in the predictand region (Fig. 10v). This situation persists over the

Unauthenticated | Downloaded 04/29/25 04:12 AM UTC

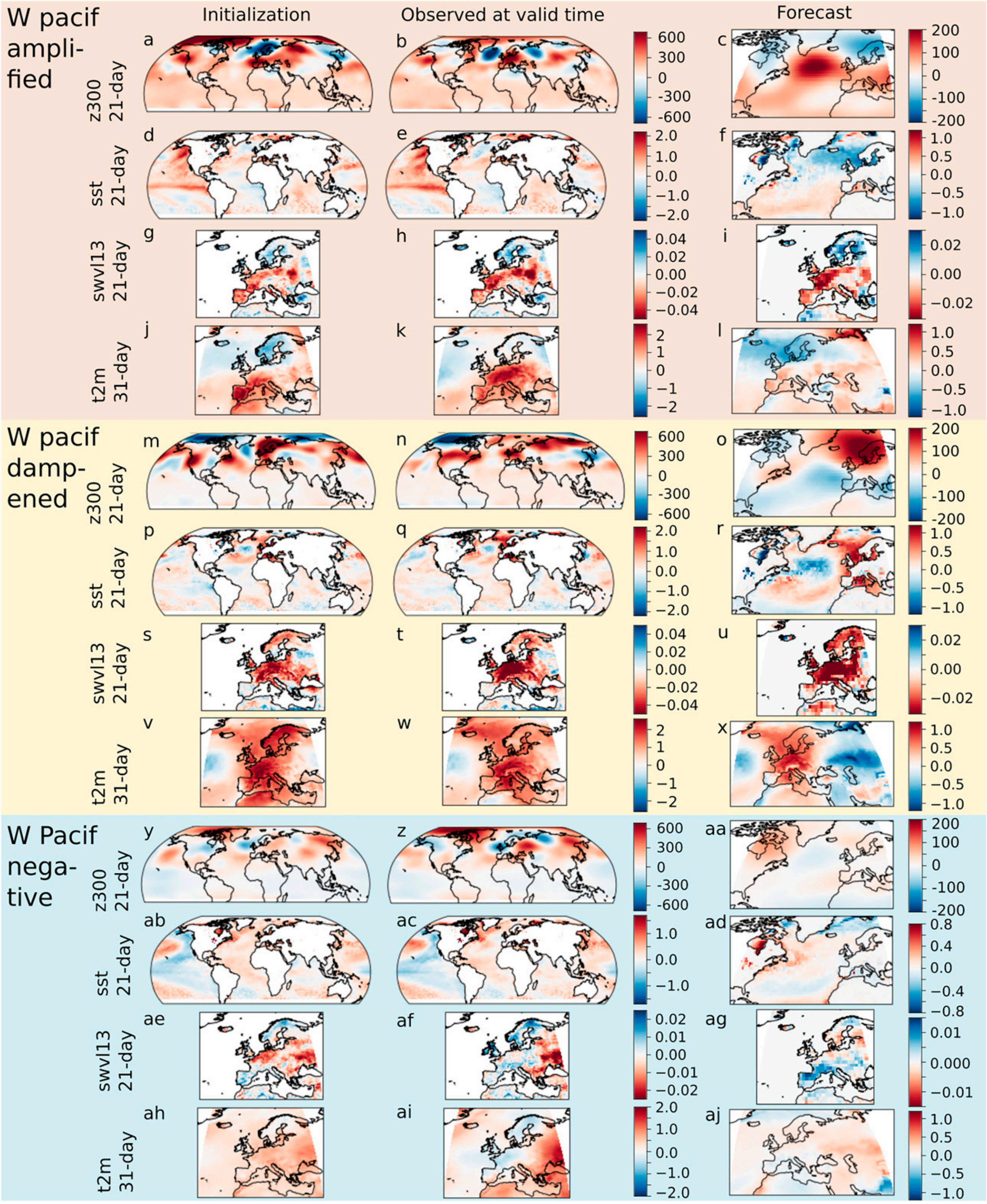
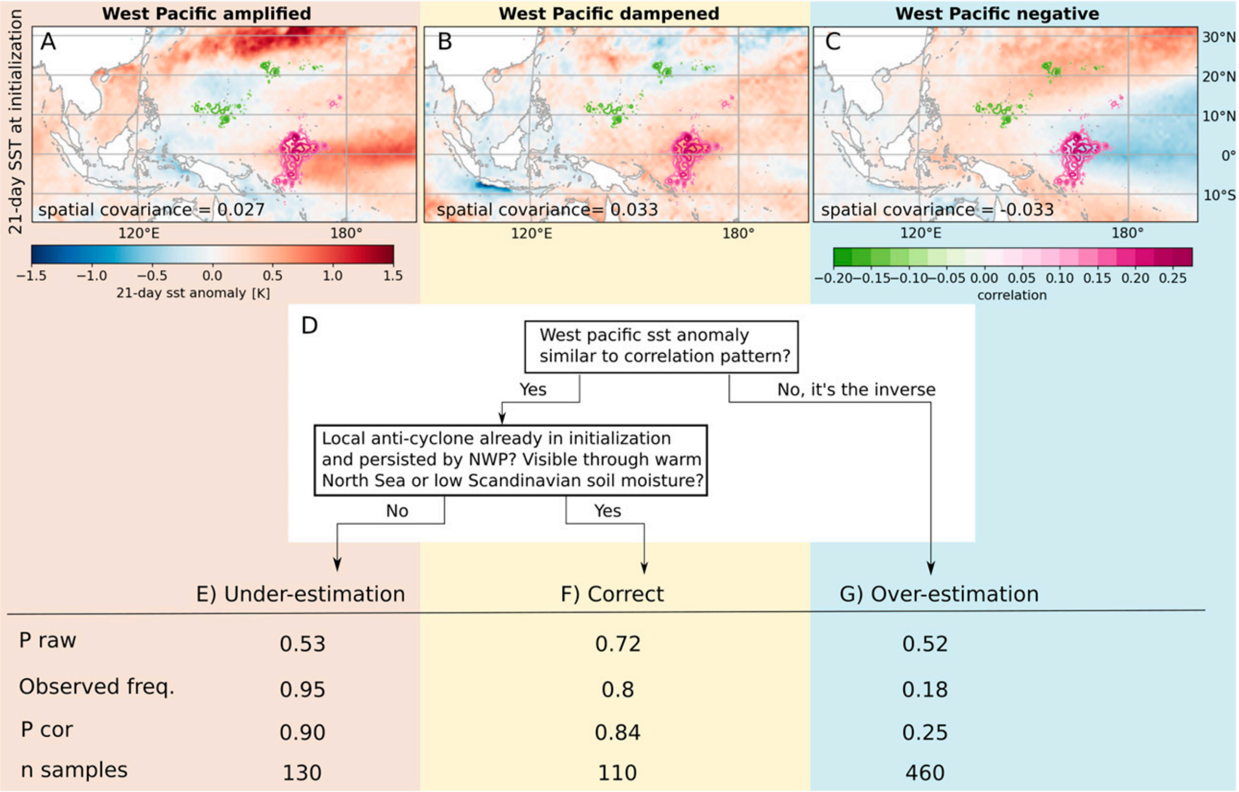


FIG. 10. Composite anomalies for samples belonging to the three groups delineated in Fig. 9. Rows display different variables [z300 ($m^2 s^{-2}$), SST (K), swvl13 ($m^3 m^{-3}$), and t2m (K)]; columns display different moments in time. (left) The 21- or 31-day reanalyzed state before initialization. (center) The 21- or 31-day reanalyzed state during our predictand period (starting 2 weeks later, at valid time). (right) The 21- or 31-day ECMWF ensemble mean forecast over the Euro-Atlantic domain (also at valid time). Note that the color-bar scale for forecasts is reduced because we take the ensemble mean. The time scale for t2m is 31 days, to overlap fully with the t2m-based predictand. The time scale for z300, SST, and swvl13 is 21 days to overlap fully with our most important SST predictor from initialization.

Unauthenticated | Downloaded 04/29/25 04:12 AM UTC


West Pacific amplified
West Pacific dampened
West Pacific negative
21-day SST at initialization
A
B
C
spatial covariance = 0.027
spatial covariance= 0.033
spatial covariance = -0.033
120°E
180°
30°N
20°N
10°N
0°
10°S
−1.5 −1.0 −0.5 0.0 0.5 1.0 1.5
21-day sst anomaly [K]
−0.20 −0.15 −0.10 −0.05 0.00 0.05 0.10 0.15 0.20 0.25
correlation
D
West pacific sst anomaly similar to correlation pattern?
Yes
No, it's the inverse
Local anti-cyclone already in initialization and persisted by NWP? Visible through warm North Sea or low Scandinavian soil moisture?
No
Yes
E) Under-estimation
F) Correct
G) Over-estimation

	E) Under-estimation	F) Correct	G) Over-estimation
P raw	0.53	0.72	0.52
Observed freq.	0.95	0.8	0.18
P cor	0.90	0.84	0.25
n samples	130	110	460

FIG. 11. (a)–(c) Composite of initial 21-day SST anomalies (K), in the three groups of Fig. 9. Contours display correlation values for grid cells that correlate significantly with t2m in the predictand region. The west Pacific predictor measures correspondence between anomalies and correlation pattern. Its composite value is annotated. (d) Human summary of learned weather-dependent corrections. (e)–(g) Table with associated forecast errors and their correction: mean p_{raw}, observed frequency of exceedance, mean p_{cor}, and number of samples in each group.

region until valid time (Figs. 10n,w) and is accompanied by further depletion of soil moisture and warming of North Sea SST (Figs. 10t,q), all of which is correctly captured by the ECMWF forecasts (Figs. 10r,u,x). Only the predicted low pressure over the Iberian Peninsula is not observed (Figs. 10n,o). The model-predicted warm North Sea and depleted Scandinavian soil moisture especially distinguish this group of correct forecasts from the cold North Sea and wet Scandinavia in "west Pacific amplified" (Figs. 10f,i). This suggests that subseasonal prediction of median t2m exceedance is more successful when local anticyclonic circulation is already present in the initial conditions (including derivatives like a warm North Sea) (also noted in Vitart et al. 2019). Because there is no need to apply a correction, the ANN modulates contributions from the west Pacific predictor (which would lead to an upward correction) with a second predictor, namely, predicted North Sea SSTs, such that the overall correction is minor. Figure B1 in appendix B further illustrates that, statistically, the ANN dampens the west Pacific influence when model-predicted North Sea SST is high. The importance of such conditional modulation is further demonstrated by the comparison in Table C1 of appendix C, in which the ANN performs better than a logistic regression, which is incapable of modeling conditional interactions. Overall, we can summarize ECMWF's conditional errors and the learned corrections with the decision tree in Fig. 11d.

5. Discussion

Our ANN-based postprocessing technique corrected forecasts with a mixture of observed predictors from initialization (ERA5) and model predictors from valid time (ECMWF forecasts). Subsequent application of XAI highlighted a connection between west Pacific SST anomalies at initialization and west/central European t2m more than 2 weeks later. The atmospheric wave associated with the west Pacific dipole is absent in the ECMWF model (Fig. 10), meaning that its forecast (p_{raw}) lacks the observed shift to high and low probabilities of median exceedance (Figs. 11e,g). Postprocessing applies this shift through conditional correction, effectively strengthening the relation between the west Pacific SST pattern and exceedance probability (first row, Fig. 8a). The robustness of this correction is supported by increased skill on unseen test data and by the fact that the ANN used only three predictors besides p_{raw}. This makes it unlikely that the ANN exploited spurious interactions between predictors for their coincidental alignment with variability in the predictand.

Unauthenticated | Downloaded 04/29/25 04:12 AM UTC

The discovered relation between the initial west Pacific state and forecast errors highlights the importance of correctly representing tropical to extratropical teleconnections in the NWP model. Previous studies have also found that tropical Pacific sources of predictability are important for European summer weather (Ding et al. 2011; O'Reilly et al. 2018; Ma and Franzke 2021). The particular source of predictability that our predictor represents appears to be related to ENSO, as correction occurs during El Niño (Figs. 10d,e) and La Niña (Figs. 10ab,ac) conditions. Also, when SHAP values are ordered by time (Fig. B2 in appendix B), we see that the "west Pacific negative" condition predominantly occurs in the 1999, 2008, and end of 2010/beginning of 2011 summer seasons, which are years with a persisting La Niña (Jong et al. 2020). This implies that ENSO's interannual variability can be important for subseasonal forecasts, and forms a cross–time scale connection (Liu and Alexander 2007; Hoskins 2013). ENSO is known to display predictable month-to-month evolution (Chapman et al. 2021) and to be capable of modulating shorter-term variability like summer monsoons (Di Capua et al. 2020). Surprisingly, the relation of MJO to forecast errors seems weak. We find that the RMM MJO index gets disregarded in the predictor selection (and does not improve scores when added manually).

The current analysis related west Pacific SSTs at initialization directly to errors at valid time. This enabled the ANN-based postprocessing to use sources of predictability whose states in the NWP model are potentially biased. A consequence was, however, that our interpretation of NWP errors lacked intermediate predictors, making it difficult to diagnose which exact part of the pathway is imperfectly represented. Shortcomings in the representation of teleconnections can reside in many model components. Shortcomings in tropical convection often relate to multiple physical parameterizations (Kim et al. 2018). Diabatic modification of Rossby waves over the midlatitudes is often poorly represented (Gray et al. 2014). Future studies should investigate the implied pathway in more detail, particularly the information shared between ENSO and the west Pacific predictor.

Nonetheless, the ANN architecture combined with XAI proved highly insightful for understanding conditional NWP errors and their physical circumstances. Three aspects should be noted for future applications. First, balanced train–validation–test splits are required. Without training on samples from the full p_{raw} range (and the global warming trend it contained), robust corrections of p_{raw} could not be learned. One avenue to prevent out-of-range values can be to train on simulations of the future climate. This will require "perfect model" assumptions to generate pseudo-observations. Second, the method cannot be immediately applied in operational settings. To summarize the initial state of relevant sources of predictability, we based our observed predictors on multiday anomalies from reanalyses, which are not available in real time. Operational analyses usually only stretch for a few hours, meaning that multiple analyses would need to be concatenated. Third, in operational practice, there is also a risk that unprecedented events lead to out-of-range input values. Thorough knowledge about the ANN's failure conditions would be required.

6. Conclusions

This study demonstrated that ANN-based postprocessing improves probabilistic forecasts of monthly summer temperature exceeding the median, with a lead time of 12–15 days before the start of the period. Raw ECMWF probability forecasts did not always outperform the climatological and climate change trend benchmark, but the ANN-corrected predictions outperformed both. The ANN bases its corrections on a shallow neural network architecture and three predictors, besides the raw ECMWF temperature forecast. One of these predictors represents a source of subseasonal predictability with influence on the Euro-Atlantic circulation. Using the state of this predictor at initialization, the ANN is able to correct conditional errors that are made when forecasts lack an atmospheric response.

Detailed explanations were obtained with XAI, specifically "input times gradient" and SHAP, which quantifies the contributions to each ANN-based correction. Hierarchical clustering subsequently groups the forecasts into groups with similar corrections for the same physical reasons. These analyses revealed that SSTs in the tropical west Pacific were predictive of ensuing errors in the NWP forecast. The pattern appears to represent an ENSO-like tropical variability, for which other studies have also shown that NWP models do not perfectly represent the teleconnections. It also appeared that the NWP error is conditional, because raw ECMWF forecasts did not need to be corrected when local anticyclonic conditions were already present in the initial conditions.

The ANN architecture developed in this paper corrects two-class NWP probability forecasts of monthly temperature exceedance for 2-m temperature over Europe during summer. The ANN learns conditional corrections that can be explained with XAI, demonstrating its ability to identify and understand conditional errors in NWP models.

Acknowledgments. This study is part of the open research program Aard- en Levenswetenschappen, Project ALWOP.395, which is financed by the Dutch Research Council (NWO). We thank maintainers and funders of the BAZIS cluster at VU Amsterdam for computational resources. Author Coumou acknowledges funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant 101003469. We thank three anonymous reviewers for their insightful comments.

Data availability statement. The ERA5 and ERA5-Land reanalysis can be obtained from the Copernicus Climate Data Store (https://cds.climate.copernicus.eu). ECMWF (re)forecasts from cycle 45r1 can be obtained through the ECMWF archive (login required at https://apps.ecmwf.int/mars-catalogue). The RMM MJO index can be accessed online (http://www.bom.gov.au/climate/mjo/graphics/rmm.74toRealtime.txt). Python code used to conduct this study is available at GitHub (https://github.com/chiemvs/Hybrid).

Unauthenticated | Downloaded 04/29/25 04:12 AM UTC

APPENDIX A

Regime Classification in z300

We distinguish four regimes in the Euro-Atlantic region. Four is commonly seen as the minimum amount needed in this region (Michelangeli et al. 1995; Zampieri et al. 2017; Falkena et al. 2020). We use the same domain as Cassou et al. (2005)—namely, from 20° to 80°N and from −90° to 30°E—and derive regimes from daily ERA5 z300 anomalies spanning from May to August. First, the ERA5 fields are regridded to the resolution of 1.5° × 1.5° at which we extracted the ECMWF forecasts. Then we linearly detrend all grid cells at once to account for thermodynamic expansion of the air column due to global warming, disregarding gridcell-specific trends like the North Atlantic warming hole (Chemke et al. 2020). We follow the approach of Ferranti et al. (2015) and Michelangeli et al. (1995) in reducing the phase space to 10 leading empirical orthogonal functions (EOFs) and computing 4 *k*-means clusters.

After computation, we assign the daily reanalysis fields to the closest centroid, measured in terms of Euclidean distance in EOF space. If a daily distance to all *k*-means centroids is found to be larger than the median distance of all samples to that centroid, the pattern is labeled "unclassified." This unclassified regime populates about 12% of reanalysis states. Forecast z300 anomalies are detrended similarly and assigned to the ERA5 centroids. Each member at each lead time gets classified as one of the four, or the unclassified regime. At a 1-day lead time, about 10% of forecasts are unclassified, against 18% at +40 days. The NWP model thus shows a drift toward unclassified flow patterns.

Because daily regime forecasts can be subject to timing errors (e.g., blocking develops a day too late), the relation to monthly t2m (section 3b) is stronger if we quantify a period's tendency toward a certain regime. To that end, we count the relative frequency of each regime over the 11 members and a succession of lead times [similar to Lavaysse et al. (2018) and Cortesi et al. (2021)]. We extract frequencies for 21- and 31-day periods.

In the end, these predictors were not selected as part of the top predictors (Fig. 7), even though NWP-predicted regimes were found to be useful in other S2S studies (Richardson et al. 2020; Mastrantonas et al. 2022). It might relate to our simplistic regime classification (for improvements, see Grams et al. 2017; Falkena et al. 2020; Dorrington and Strommen 2020). Alternatively, we know that summer circulation is continuous in phase space and less "regime like" than in winter (Rousi et al. 2021). This can make our division into four classes arbitrary, harder to predict, and of little use for S2S postprocessing. The limited use of ECMWF regime states is, however, consistent with our suggestion that misrepresented sources of predictability affect a multitude of model-predicted states at valid time, among which is atmospheric circulation (Fig. 1b). Results indeed indicate the missed development of an atmospheric wave (Figs. 10c,aa). With NWP models lacking such connections, it is not surprising that skill in regime predictions often does not extend beyond the third forecast week (Cortesi et al. 2021; Büeler et al. 2021).

APPENDIX B

Additional XAI Plots

Figure B1 shows statistically that the ANN dampens the west Pacific influence when model-predicted North Sea SST is high. Time-ordered SHAP values are used in Fig. B2 to show that the "west Pacific negative" condition predominantly occurs in years with a persisting La Niña: in the summer seasons of 1999 and 2008, in late summer 2010, and in early summer 2011, the contribution from west Pacific SSTs is negative and that of the other predictors is neutral.

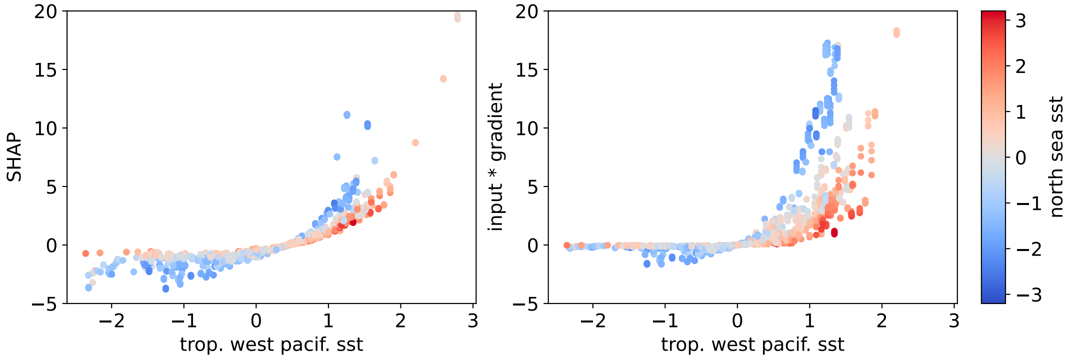


FIG. B1. Effect of the initial west Pacific SST pattern on the ANN-based correction, in terms of (left) SHAP value and (right) input × gradient. The *x* axis is the standardized value of the west Pacific predictor, and the *y* axis is the contribution to the multiplicative correction factor. The color scale shows the standardized value of predicted North Sea SST. The curves show that high west Pacific values (positive correspondence between initial anomalies and correlation pattern) result in positive corrections and negative values result in negative corrections, but the impact is modulated by North Sea SST (west Pacific contributions are brought closer to zero when predicted North Sea SST is high).

Unauthenticated | Downloaded 04/29/25 04:12 AM UTC

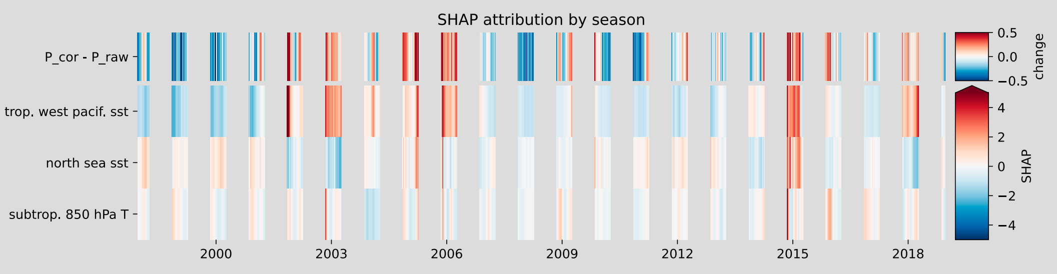


FIG. B2. As in Fig. 9, but with samples ordered by time instead of by the leaf order from hierarchical clustering, and with SHAP values only.

APPENDIX C

Additional Logistic Regression Benchmarks

Table C1 shows a Brier score comparison on the test set of ANN-based postprocessed forecasts with two simple logistic regression models.

TABLE C1. Brier score comparison on the test set (lower is better) of ANN-based postprocessed forecasts (p_{cor}) with two simple logistic regression models using the same forward selected predictors as the ANN, with either p_{raw} or $\log(p_{\mathrm{raw}})$ as extra predictor. ANN characteristics such as the amount of hidden layers and the amount of predictors are varied over rows. Benchmarks p_{trend} and p_{raw} have a BS of 0.214 and 0.224, respectively.

	p_{cor}	Logistic + p_{raw}	Logistic + $\log(p_{\mathrm{raw}})$
3 predictors; 1 hidden layer	0.168	0.188	0.190
3 predictors; 0 hidden layers	0.171	0.188	0.190
1 predictor; 1 hidden layer	0.188	0.195	0.196
1 predictor; 0 hidden layers	0.188	0.195	0.196

REFERENCES

Allen, S., C. A. Ferro, and F. Kwasniok, 2019: Regime-dependent statistical post-processing of ensemble forecasts. *Quart. J. Roy. Meteor. Soc.*, **145**, 3535–3552, https://doi.org/10.1002/qj.3638.

Arrieta, A. B., and Coauthors, 2020: Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion*, **58**, 82–115, https://doi.org/10.1016/j.inffus.2019.12.012.

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, https://doi.org/10.1038/nature14956.

Bello, G. A., M. Angus, N. Pedemane, J. K. Harlalka, F. H. Semazzi, V. Kumar, and N. F. Samatova, 2015: Response-guided community detection: Application to climate index discovery. *Machine Learning and Knowledge Discovery in Databases*, A. Appice et al., Eds., Lecture Notes in Computer Science, Vol. 9285, Springer, 736–751, https://doi.org/10.1007/978-3-319-23525-7_45.

Ben Bouallègue, Z., and D. S. Richardson, 2022: On the ROC area of ensemble forecasts for rare events. *Wea. Forecasting*, **37**, 787–796, https://doi.org/10.1175/WAF-D-21-0195.1.

Beverley, J. D., S. J. Woolnough, L. H. Baker, S. J. Johnson, and A. Weisheimer, 2019: The Northern Hemisphere circumglobal teleconnection in a seasonal forecast model and its relationship to European summer forecast skill. *Climate Dyn.*, **52**, 3759–3771, https://doi.org/10.1007/s00382-018-4371-4.

Bjerknes, J., 1969: Atmospheric teleconnections from the equatorial Pacific. *Mon. Wea. Rev.*, **97**, 163–172, https://doi.org/10.1175/1520-0493(1969)097<0163:ATFTEP>2.3.CO;2.

Branstator, G., 2014: Long-lived response of the midlatitude circulation and storm tracks to pulses of tropical heating. *J. Climate*, **27**, 8809–8826, https://doi.org/10.1175/JCLI-D-14-00312.1.

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

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. *Quart. J. Roy. Meteor. Soc.*, **147**, 4283–4309, https://doi.org/10.1002/qj.4178.

Buizza, R., and M. Leutbecher, 2015: The forecast skill horizon. *Quart. J. Roy. Meteor. Soc.*, **141**, 3366–3382, https://doi.org/10.1002/qj.2619.

Carvalho-Oliveira, J., L. F. Borchert, E. Zorita, and J. Baehr, 2022: Self-organizing maps identify windows of opportunity for seasonal European summer predictions. *Front. Climate*, **4**, 844634, https://doi.org/10.3389/fclim.2022.844634.

Casanueva, A., C. Rodríguez-Puebla, M. Frías, and N. González-Reviriego, 2014: Variability of extreme precipitation over Europe and its relationships with teleconnection patterns.

Unauthenticated | Downloaded 04/29/25 04:12 AM UTC

Hydrol. Earth Syst. Sci., **18**, 709–725, https://doi.org/10.5194/hess-18-709-2014.

Cassou, C., 2008: Intraseasonal interaction between the Madden–Julian oscillation and the North Atlantic Oscillation. *Nature*, **455**, 523–527, https://doi.org/10.1038/nature07286.

——, L. Terray, and A. S. Phillips, 2005: Tropical Atlantic influence on European heat waves. *J. Climate*, **18**, 2805–2811, https://doi.org/10.1175/JCLI3506.1.

Chapman, W. E., A. C. Subramanian, S.-P. Xie, M. D. Sierks, F. M. Ralph, and Y. Kamae, 2021: Monthly modulations of ENSO teleconnections: Implications for potential predictability in North America. *J. Climate*, **34**, 5899–5921, https://doi.org/10.1175/JCLI-D-20-0391.1.

Chemke, R., L. Zanna, and L. M. Polvani, 2020: Identifying a human signal in the North Atlantic warming hole. *Nat. Commun.*, **11**, 1540, https://doi.org/10.1038/s41467-020-15285-x.

Clare, M. C. A., M. Sonnewald, R. Lguensat, J. Deshayes, and V. Balaji, 2022: Explainable artificial intelligence for Bayesian neural networks: Towards trustworthy predictions of ocean dynamics. arXiv, 2205.00202v1, https://doi.org/10.48550/arxiv.2205.00202.

Cortesi, N., V. Torralba, L. Lledó, A. Manrique-Suñén, N. Gonzalez-Reviriego, A. Soret, and F. J. Doblas-Reyes, 2021: Yearly evolution of Euro-Atlantic weather regimes and of their sub-seasonal predictability. *Climate Dyn.*, **56**, 3933–3964, https://doi.org/10.1007/s00382-021-05679-y.

Di Capua, G., J. Runge, R. V. Donner, B. van den Hurk, A. G. Turner, R. Vellore, R. Krishnan, and D. Coumou, 2020: Dominant patterns of interaction between the tropics and mid-latitudes in boreal summer: Causal relationships and the role of timescales. *Wea. Climate Dyn.*, **1**, 519–539, https://doi.org/10.5194/wcd-1-519-2020.

Ding, Q., B. Wang, J. M. Wallace, and G. Branstator, 2011: Tropical–extratropical teleconnections in boreal summer: Observed interannual variability. *J. Climate*, **24**, 1878–1896, https://doi.org/10.1175/2011JCLI3621.1.

Dorrington, J., and K. Strommen, 2020: Jet speed variability obscures Euro-Atlantic regime structure. *Geophys. Res. Lett.*, **47**, e2020GL087907, https://doi.org/10.1029/2020GL087907.

Dutra, E., F. Johannsen, and L. Magnusson, 2021: Late spring and summer subseasonal forecasts in the Northern Hemisphere midlatitudes: Biases and skill in the ECMWF model. *Mon. Wea. Rev.*, **149**, 2659–2671, https://doi.org/10.1175/MWR-D-20-0342.1.

Falkena, S. K., J. de Wiljes, A. Weisheimer, and T. G. Shepherd, 2020: Revisiting the identification of wintertime atmospheric circulation regimes in the Euro-Atlantic sector. *Quart. J. Roy. Meteor. Soc.*, **146**, 2801–2814, https://doi.org/10.1002/qj.3818.

Fan, Y., V. Krasnopolsky, H. van den 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. *Wea. Forecasting*, **38**, 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. *Quart. J. Roy. Meteor. Soc.*, **141**, 916–924, https://doi.org/10.1002/qj.2411.

Ferrone, A., D. Mastrangelo, and P. Malguzzi, 2017: Multimodel probabilistic prediction of 2 m-temperature anomalies on the monthly timescale. *Adv. Sci. Res.*, **14**, 123–129, https://doi.org/10.5194/asr-14-123-2017.

Gibson, P. B., W. E. Chapman, A. Altinok, L. Delle Monache, M. J. DeFlorio, and D. E. Waliser, 2021: Training machine learning models on climate model output yields skillful interpretable seasonal precipitation forecasts. *Commun. Earth Environ.*, **2**, 159, https://doi.org/10.1038/s43247-021-00225-4.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor. Climatol.*, **11**, 1203–1211, https://doi.org/10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2.

Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.*, **102**, 359–378, https://doi.org/10.1198/016214506000001437.

Graham, R. M., J. Browell, D. Bertram, and C. J. White, 2022: The application of sub-seasonal to seasonal (S2S) predictions for hydropower forecasting. *Meteor. Appl.*, **29**, e2047, https://doi.org/10.1002/met.2047.

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. *Nat. Climate Change*, **7**, 557–562, https://doi.org/10.1038/nclimate3338.

Gray, S. L., C. M. Dunning, J. Methven, G. Masato, and J. M. Chagnon, 2014: Systematic model forecast error in Rossby wave structure. *Geophys. Res. Lett.*, **41**, 2979–2987, https://doi.org/10.1002/2014GL059282.

Grönquist, P., C. Yao, T. Ben-Nun, N. Dryden, P. Dueben, S. Li, and T. Hoefler, 2021: Deep learning for post-processing ensemble weather forecasts. *Philos. Trans. Roy. Soc.*, **A379**, 20200092, https://doi.org/10.1098/rsta.2020.0092.

Grotjahn, R., and Coauthors, 2016: North American extreme temperature events and related large scale meteorological patterns: A review of statistical methods, dynamics, modeling, and trends. *Climate Dyn.*, **46**, 1151–1184, https://doi.org/10.1007/s00382-015-2638-6.

Hall, R. J., J. M. Jones, E. Hanna, A. A. Scaife, and R. Erdélyi, 2017: Drivers and potential predictability of summer time North Atlantic polar front jet variability. *Climate Dyn.*, **48**, 3869–3887, https://doi.org/10.1007/s00382-016-3307-0.

Hamill, T. M., and J. Juras, 2006: Measuring forecast skill: Is it real skill or is it the varying climatology? *Quart. J. Roy. Meteor. Soc.*, **132**, 2905–2923, https://doi.org/10.1256/qj.06.25.

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

Haupt, S. E., W. Chapman, S. V. Adams, C. Kirkwood, J. S. Hosking, N. H. Robinson, S. Lerch, and A. C. Subramanian, 2021: Towards implementing artificial intelligence post-processing in weather and climate: Proposed actions from the Oxford 2019 workshop. *Philos. Trans. Roy. Soc.*, **A379**, 20200091, https://doi.org/10.1098/rsta.2020.0091.

He, B., P. Liu, Y. Zhu, and W. Hu, 2019: Prediction and predictability of Northern Hemisphere persistent maxima of 500-hPa geopotential height eddies in the GEFS. *Climate Dyn.*, **52**, 3773–3789, https://doi.org/10.1007/s00382-018-4347-4.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, https://doi.org/10.1002/qj.3803.

Hertel, L., J. Collado, P. Sadowski, J. Ott, and P. Baldi, 2020: Sherpa: Robust hyperparameter optimization for machine learning. *SoftwareX*, **12**, 100591, https://doi.org/10.1016/j.softx.2020.100591.

Unauthenticated | Downloaded 04/29/25 04:12 AM UTC

Hewson, T. D., and F. M. Pillosu, 2021: A low-cost post-processing technique improves weather forecasts around the world. *Commun. Earth Environ.*, **2**, 132, https://doi.org/10.1038/s43247-021-00185-9.

Hoskins, B., 2013: The potential for skill across the range of the seamless weather-climate prediction problem: A stimulus for our science. *Quart. J. Roy. Meteor. Soc.*, **139**, 573–584, https://doi.org/10.1002/qj.1991.

Johnson, S. J., and Coauthors, 2019: SEAS5: The new ECMWF seasonal forecast system. *Geosci. Model Dev.*, **12**, 1087–1117, https://doi.org/10.5194/gmd-12-1087-2019.

Jong, B.-T., M. Ting, R. Seager, and W. B. Anderson, 2020: ENSO teleconnections and impacts on U.S. summertime temperature during a multiyear La Niña life cycle. *J. Climate*, **33**, 6009–6024, https://doi.org/10.1175/JCLI-D-19-0701.1.

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. *Wea. Climate Dyn.*, **3**, 305–336, https://doi.org/10.5194/wcd-3-305-2022.

Kharin, V. V., and F. W. Zwiers, 2003: On the ROC score of probability forecasts. *J. Climate*, **16**, 4145–4150, https://doi.org/10.1175/1520-0442(2003)016<4145:OTRSOP>2.0.CO;2.

Kim, H., F. Vitart, and D. E. Waliser, 2018: Prediction of the Madden–Julian oscillation: A review. *J. Climate*, **31**, 9425–9443, https://doi.org/10.1175/JCLI-D-18-0210.1.

Kingma, D. P., and J. Ba, 2014: Adam: A method for stochastic optimization. arXiv, 1412.6980v9, https://doi.org/10.48550/arXiv.1412.6980.

Kretschmer, M., J. Runge, and D. Coumou, 2017: Early prediction of extreme stratospheric polar vortex states based on causal precursors. *Geophys. Res. Lett.*, **44**, 8592–8600, https://doi.org/10.1002/2017GL074696.

Kueh, M.-T., and C.-Y. Lin, 2020: The 2018 summer heatwaves over northwestern Europe and its extended-range prediction. *Sci. Rep.*, **10**, 19283, https://doi.org/10.1038/s41598-020-76181-4.

Lavaysse, C., J. Vogt, A. Toreti, M. L. Carrera, and F. Pappenberger, 2018: On the use of weather regimes to forecast meteorological drought over Europe. *Nat. Hazards Earth Syst. Sci.*, **18**, 3297–3309, https://doi.org/10.5194/nhess-18-3297-2018.

——, G. Naumann, L. Alfieri, P. Salamon, and J. Vogt, 2019: Predictability of the European heat and cold waves. *Climate Dyn.*, **52**, 2481–2495, https://doi.org/10.1007/s00382-018-4273-5.

Leutbecher, M., and T. N. Palmer, 2008: Ensemble forecasting. *J. Comput. Phys.*, **227**, 3515–3539, https://doi.org/10.1016/j.jcp.2007.02.014.

Lin, H., G. Brunet, and J. Derome, 2009: An observed connection between the North Atlantic Oscillation and the Madden–Julian oscillation. *J. Climate*, **22**, 364–380, https://doi.org/10.1175/2008JCLI2515.1.

Liu, Z., and M. Alexander, 2007: Atmospheric bridge, oceanic tunnel, and global climatic teleconnections. *Rev. Geophys.*, **45**, RG2005, https://doi.org/10.1029/2005RG000172.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.*, **20**, 130–141, https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

——, 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, **21A**, 289–307, https://doi.org/10.3402/tellusa.v21i3.10086.

Lundberg, S. M., and S.-I. Lee, 2017: A unified approach to interpreting model predictions. *Proc. 31st Int. Conf. on Neural Information Processing Systems*, Long Beach, CA, Association for Computing Machinery, 4768–4777, https://dl.acm.org/doi/10.5555/3295222.3295230.

——, and Coauthors, 2020: From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.*, **2**, 56–67, https://doi.org/10.1038/s42256-019-0138-9.

Ma, Q., and C. L. Franzke, 2021: The role of transient eddies and diabatic heating in the maintenance of European heat waves: A nonlinear quasi-stationary wave perspective. *Climate Dyn.*, **56**, 2983–3002, https://doi.org/10.1007/s00382-021-05628-9.

Mamalakis, A., I. Ebert-Uphoff, and E. A. Barnes, 2022: Neural network attribution methods for problems in geoscience: A novel synthetic benchmark dataset. *Environ. Data Sci.*, **1**, e8, https://doi.org/10.1017/eds.2022.7.

Manrique-Suñén, A., N. Gonzalez-Reviriego, V. Torralba, N. Cortesi, and F. J. Doblas-Reyes, 2020: Choices in the verification of S2S forecasts and their implications for climate services. *Mon. Wea. Rev.*, **148**, 3995–4008, https://doi.org/10.1175/MWR-D-20-0067.1.

Manzanas, R., A. Lucero, A. Weisheimer, and J. M. Gutiérrez, 2018: Can bias correction and statistical downscaling methods improve the skill of seasonal precipitation forecasts? *Climate Dyn.*, **50**, 1161–1176, https://doi.org/10.1007/s00382-017-3668-z.

Mariotti, A., and Coauthors, 2020: Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond. *Bull. Amer. Meteor. Soc.*, **101**, E608–E625, https://doi.org/10.1175/BAMS-D-18-0326.1.

Mastrantonas, N., L. Magnusson, F. Pappenberger, and J. Matschullat, 2022: What do large-scale patterns teach us about extreme precipitation over the Mediterranean at medium- and extended-range forecasts? *Quart. J. Roy. Meteor. Soc.*, **148**, 875–890, https://doi.org/10.1002/qj.4236.

Mayer, K. J., and E. A. Barnes, 2021: Subseasonal forecasts of opportunity identified by an explainable neural network. *Geophys. Res. Lett.*, **48**, e2020GL092092, https://doi.org/10.1029/2020GL092092.

McGovern, A., R. Lagerquist, D. John Gagne II, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, 2019: Making the black box more transparent: Understanding the physical implications of machine learning. *Bull. Amer. Meteor. Soc.*, **100**, 2175–2199, https://doi.org/10.1175/BAMS-D-18-0195.1.

Merryfield, W. J., and Coauthors, 2020: Current and emerging developments in subseasonal to decadal prediction. *Bull. Amer. Meteor. Soc.*, **101**, E869–E896, https://doi.org/10.1175/BAMS-D-19-0037.1.

Michelangeli, P.-A., R. Vautard, and B. Legras, 1995: Weather regimes: Recurrence and quasi stationarity. *J. Atmos. Sci.*, **52**, 1237–1256, https://doi.org/10.1175/1520-0469(1995)052%3C1237:WRRAQS%3E2.0.CO;2.

Miralles, D. G., P. Gentine, S. I. Seneviratne, and A. J. Teuling, 2019: Land–atmospheric feedbacks during droughts and heatwaves: State of the science and current challenges. *Ann. N. Y. Acad. Sci.*, **1436**, 19–35, https://doi.org/10.1111/nyas.13912.

Molnar, C., G. Casalicchio, and B. Bischl, 2021: Interpretable machine learning—A brief history, state-of-the-art and challenges. *ECML PKDD 2020 Workshops*, I. Koprinksa et al., Eds., Communications in Computer and Information Science, Vol. 1323, Springer, 417–431, https://doi.org/10.1007/978-3-030-65965-3_28.

Monhart, S., C. Spirig, J. Bhend, K. Bogner, C. Schär, and M. Liniger, 2018: Skill of subseasonal forecasts in Europe: Effect of bias correction and downscaling using surface observations.

Unauthenticated | Downloaded 04/29/25 04:12 AM UTC

J. Geophys. Res. Atmos., **123**, 7999–8016, https://doi.org/10.1029/2017JD027923.

Mueller, S. T., R. R. Hoffman, W. J. Clancey, A. Emrey, and G. Klein, 2019: Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. arXiv, 1902.01876v1, https://doi.org/10.48550/arXiv.1902.01876.

Muñoz-Sabater, J., and Coauthors, 2021: ERA5-land: A state-of-the-art global reanalysis dataset for land applications. *Earth Syst. Sci. Data*, **13**, 4349–4383, https://doi.org/10.5194/essd-13-4349-2021.

O'Reilly, C. H., T. Woollings, L. Zanna, and A. Weisheimer, 2018: The impact of tropical precipitation on summertime Euro-Atlantic circulation via a circumglobal wave train. *J. Climate*, **31**, 6481–6504, https://doi.org/10.1175/JCLI-D-17-0451.1.

Osborne, J. M., M. Collins, J. A. Screen, S. I. Thomson, and N. Dunstone, 2020: The North Atlantic as a driver of summer atmospheric circulation. *J. Climate*, **33**, 7335–7351, https://doi.org/10.1175/JCLI-D-19-0423.1.

Ossó, A., R. Sutton, L. Shaffrey, and B. Dong, 2020: Development, amplification, and decay of Atlantic/European summer weather patterns linked to spring North Atlantic sea surface temperatures. *J. Climate*, **33**, 5939–5951, https://doi.org/10.1175/JCLI-D-19-0613.1.

Palmer, T. N., 1993: Extended-range atmospheric prediction and the Lorenz model. *Bull. Amer. Meteor. Soc.*, **74**, 49–66, https://doi.org/10.1175/1520-0477(1993)074<0049:ERAPAT>2.0.CO;2.

Pfahl, S., 2014: Characterising the relationship between weather extremes in Europe and synoptic circulation features. *Nat. Hazards Earth Syst. Sci.*, **14**, 1461–1475, https://doi.org/10.5194/nhess-14-1461-2014.

Quesada, B., R. Vautard, P. Yiou, M. Hirschi, and S. I. Seneviratne, 2012: Asymmetric European summer heat predictability from wet and dry southern winters and springs. *Nat. Climate Change*, **2**, 736–741, https://doi.org/10.1038/nclimate1536.

Quinting, J., and F. Vitart, 2019: Representation of synoptic-scale Rossby wave packets and blocking in the S2S Prediction Project Database. *Geophys. Res. Lett.*, **46**, 1070–1078, https://doi.org/10.1029/2018GL081381.

Rasp, S., and S. Lerch, 2018: Neural networks for postprocessing ensemble weather forecasts. *Mon. Wea. Rev.*, **146**, 3885–3900, https://doi.org/10.1175/MWR-D-18-0187.1.

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **126**, 649–667, https://doi.org/10.1002/qj.49712656313.

——, H. J. Fowler, C. G. Kilsby, R. Neal, and R. Dankers, 2020: Improving sub-seasonal forecast skill of meteorological drought: A weather pattern approach. *Nat. Hazards Earth Syst. Sci.*, **20**, 107–124, https://doi.org/10.5194/nhess-20-107-2020.

Roads, J. O., 1986: Forecasts of time averages with a numerical weather prediction model. *J. Atmos. Sci.*, **43**, 871–893, https://doi.org/10.1175/1520-0469(1986)043<0871:FOTAWA>2.0.CO;2.

Rousi, E., F. Selten, S. Rahmstorf, and D. Coumou, 2021: Changes in North Atlantic atmospheric circulation in a warmer climate favor winter flooding and summer drought over Europe. *J. Climate*, **34**, 2277–2295, https://doi.org/10.1175/JCLI-D-20-0311.1.

Sardeshmukh, P. D., and B. J. Hoskins, 1988: The generation of global rotational flow by steady idealized tropical divergence. *J. Atmos. Sci.*, **45**, 1228–1251, https://doi.org/10.1175/1520-0469(1988)045<1228:TGOGRF>2.0.CO;2.

Schaller, N., J. Sillmann, J. Anstey, E. Fischer, C. Grams, and S. Russo, 2018: Influence of blocking on northern European and western Russian heatwaves in large climate model ensembles. *Environ. Res. Lett.*, **13**, 054015, https://doi.org/10.1088/1748-9326/aaba55.

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. *Mon. Wea. Rev.*, **148**, 3489–3506, https://doi.org/10.1175/MWR-D-20-0096.1.

Schulz, B., and S. Lerch, 2022: Machine learning methods for postprocessing ensemble forecasts of wind gusts: A systematic comparison. *Mon. Wea. Rev.*, **150**, 235–257, https://doi.org/10.1175/MWR-D-21-0150.1.

Seneviratne, S. I., T. Corti, E. L. Davin, M. Hirschi, E. B. Jaeger, I. Lehner, B. Orlowsky, and A. J. Teuling, 2010: Investigating soil moisture–climate interactions in a changing climate: A review. *Earth-Sci. Rev.*, **99**, 125–161, https://doi.org/10.1016/j.earscirev.2010.02.004.

Shao, Y., Q. J. Wang, A. Schepen, and D. Ryu, 2022: Introducing long-term trends into sub-seasonal temperature forecasts through trend-aware post-processing. *Int. J. Climatol.*, **42**, 4972–4988, https://doi.org/10.1002/joc.7515.

Shukla, J., 1981: Dynamical predictability of monthly means. *J. Atmos. Sci.*, **38**, 2547–2572, https://doi.org/10.1175/1520-0469(1981)038<2547:DPOMM>2.0.CO;2.

Sousa, P. M., R. M. Trigo, D. Barriopedro, P. M. Soares, and J. A. Santos, 2018: European temperature responses to blocking and ridge regional patterns. *Climate Dyn.*, **50**, 457–477, https://doi.org/10.1007/s00382-017-3620-2.

Specq, D., and L. Batté, 2020: Improving subseasonal precipitation forecasts through a statistical–dynamical approach: Application to the southwest tropical Pacific. *Climate Dyn.*, **55**, 1913–1927, https://doi.org/10.1007/s00382-020-05355-7.

Strazzo, S., D. C. Collins, A. Schepen, Q. J. Wang, E. Becker, and L. Jia, 2019: Application of a hybrid statistical–dynamical system to seasonal prediction of North American temperature and precipitation. *Mon. Wea. Rev.*, **147**, 607–625, https://doi.org/10.1175/MWR-D-18-0156.1.

Ting, M., 1994: Maintenance of northern summer stationary waves in a GCM. *J. Atmos. Sci.*, **51**, 3286–3308, https://doi.org/10.1175/1520-0469(1994)051<3286:MONSSW>2.0.CO;2.

Toms, B. A., E. A. Barnes, and I. Ebert-Uphoff, 2020: Physically interpretable neural networks for the geosciences: Applications to Earth system variability. *J. Adv. Model. Earth Syst.*, **12**, e2019MS002002, https://doi.org/10.1029/2019MS002002.

Toth, Z., and R. Buizza, 2019: Weather forecasting: What sets the forecast skill horizon? *Sub-Seasonal to Seasonal Prediction*, A. W. Robertson and F. Vitart, Eds., Elsevier, 17–45, https://doi.org/10.1016/B978-0-12-811714-9.00002-4.

Trenberth, K. E., and D. P. Stepaniak, 2001: Indices of El Niño evolution. *J. Climate*, **14**, 1697–1701, https://doi.org/10.1175/1520-0442(2001)014<1697:LIOENO>2.0.CO;2.

——, G. W. Branstator, D. Karoly, A. Kumar, N.-C. Lau, and C. Ropelewski, 1998: Progress during TOGA in understanding and modeling global teleconnections associated with tropical sea surface temperatures. *J. Geophys. Res.*, **103**, 14 291–14 324, https://doi.org/10.1029/97JC01444.

Vannitsem, S., and Coauthors, 2021: Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big

Unauthenticated | Downloaded 04/29/25 04:12 AM UTC

data world. *Bull. Amer. Meteor. Soc.*, **102**, E681–E699, https://doi.org/10.1175/BAMS-D-19-0308.1.

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. *Quart. J. Roy. Meteor. Soc.*, **146**, 2654–2670, https://doi.org/10.1002/qj.3810.

——, ——, ——, ——, and ——, 2022: Using explainable machine learning forecasts to discover sub-seasonal drivers of high summer temperatures in western and central Europe. *Mon. Wea. Rev.*, **150**, 1115–1134, https://doi.org/10.1175/MWR-D-21-0201.1.

Veldkamp, S., K. Whan, S. Dirksen, and M. Schmeits, 2021: Statistical postprocessing of wind speed forecasts using convolutional neural networks. *Mon. Wea. Rev.*, **149**, 1141–1152, https://doi.org/10.1175/MWR-D-20-0219.1.

Vigaud, N., A. W. Robertson, and M. K. Tippett, 2017: Multimodel ensembling of subseasonal precipitation forecasts over North America. *Mon. Wea. Rev.*, **145**, 3913–3928, https://doi.org/10.1175/MWR-D-17-0092.1.

Vitart, F., 2017: Madden–Julian oscillation prediction and teleconnections in the S2S database. *Quart. J. Roy. Meteor. Soc.*, **143**, 2210–2220, https://doi.org/10.1002/qj.3079.

——, 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.

——, and Coauthors, 2019: Extended-range prediction. ECMWF Tech. Memo. 854, 60 pp., https://doi.org/10.21957/pdivp3t9m.

Wheeler, M. C., and H. H. Hendon, 2004: An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. *Mon. Wea. Rev.*, **132**, 1917–1932, https://doi.org/10.1175/1520-0493(2004)132%3C1917:AARMMI%3E2.0.CO;2.

——, H. Zhu, A. H. Sobel, D. Hudson, and F. Vitart, 2017: Seamless precipitation prediction skill comparison between two global models. *Quart. J. Roy. Meteor. Soc.*, **143**, 374–383, https://doi.org/10.1002/qj.2928.

White, C. J., and Coauthors, 2022: Advances in the application and utility of subseasonal-to-seasonal predictions. *Bull. Amer. Meteor. Soc.*, **103**, E1448–E1472, https://doi.org/10.1175/BAMS-D-20-0224.1.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. Elsevier, 676 pp.

Wolf, G., D. J. Brayshaw, N. P. Klingaman, and A. Czaja, 2018: Quasi-stationary waves and their impact on European weather and extreme events. *Quart. J. Roy. Meteor. Soc.*, **144**, 2431–2448, https://doi.org/10.1002/qj.3310.

Wulff, C. O., F. Vitart, and D. I. Domeisen, 2022: Influence of trends on subseasonal temperature prediction skill. *Quart. J. Roy. Meteor. Soc.*, **148**, 1280–1299, https://doi.org/10.1002/qj.4259.

Zampieri, M., A. Toreti, A. Schindler, E. Scoccimarro, and S. Gualdi, 2017: Atlantic multi-decadal oscillation influence on weather regimes over Europe and the Mediterranean in spring and summer. *Global Planet. Change*, **151**, 92–100, https://doi.org/10.1016/j.gloplacha.2016.08.014.

Zhang, C., 2013: Madden–Julian oscillation: Bridging weather and climate. *Bull. Amer. Meteor. Soc.*, **94**, 1849–1870, https://doi.org/10.1175/BAMS-D-12-00026.1.

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? *J. Atmos. Sci.*, **76**, 1077–1091, https://doi.org/10.1175/JAS-D-18-0269.1.

Zhang, R., C. Sun, J. Zhu, R. Zhang, and W. Li, 2020: Increased European heat waves in recent decades in response to shrinking Arctic Sea ice and Eurasian snow cover. *npj Climate Atmos. Sci.*, **3**, 7, https://doi.org/10.1038/s41612-020-0110-8.

Unauthenticated | Downloaded 04/29/25 04:12 AM UTC