Statistical Downscaling to Improve the Subseasonal Predictions of Energy-Relevant Surface Variables

Naveen Goutham, Riwal Plougonven, Hiba Omrani, Alexis Tantet, Sylvie Parey, Peter Tankov, Peter Hitchcock, Philippe Drobinski

Statistical Downscaling to Improve the Subseasonal Predictions of Energy-Relevant Surface Variables

NAVEEN GOUTHAM,[a,b] RIWAL PLOUGONVEN,[b] HIBA OMRANI,[a] ALEXIS TANTET,[b] SYLVIE PAREY,[a] PETER TANKOV,[c] PETER HITCHCOCK,[d] AND PHILIPPE DROBINSKI[b]

[a] *EDF Lab Paris-Saclay, Palaiseau, France*
[b] *Laboratoire de Météorologie Dynamique-IPSL, Ecole Polytechnique, Institut Polytechnique de Paris, ENS, PSL Research University, Sorbonne Université, CNRS, France*
[c] *CREST/ENSAE, Institut Polytechnique de Paris, Palaiseau, France*
[d] *Department of Earth and Atmospheric Sciences, Cornell University, Ithaca, New York*

(Manuscript received 22 June 2022, in final form 29 September 2022)

ABSTRACT: Owing to the increasing share of variable renewable energies in the electricity mix, the European energy sector is becoming more weather sensitive. In this regard, skillful subseasonal predictions of essential climate variables can provide considerable socioeconomic benefits to the energy sector. The aim of this study is therefore to improve the European subseasonal predictions of 100-m wind speed and 2-m temperature, which we achieve through statistical downscaling. We employ redundancy analysis (RDA) to estimate spatial patterns of variability from large-scale fields that allow for the best prediction of surface fields. We compare explanatory powers between the patterns obtained using RDA against those derived using principal component analysis (PCA), when used as predictors in multilinear regression models to predict surface fields, and show that the explanatory power of the former is superior to that of the latter. Subsequently, we employ the estimated relationship between RDA patterns and surface fields to produce statistical probabilistic predictions of gridded surface fields using dynamical ensemble predictions of RDA patterns. We finally demonstrate how a simple combination of dynamical and statistical predictions of surface fields significantly improves the accuracy of subseasonal predictions of both variables over a large part of Europe. We attribute the improved accuracy of these combined predictions to improvements in reliability and resolution.

KEYWORDS: Europe; Downscaling; Empirical orthogonal functions; Forecast verification/skill; Statistical forecasting; Subseasonal variability

1. Introduction

Subseasonal predictions refer to predictions beyond two weeks and up to two months (Robertson and Vitart 2018). These predictions are influenced by both atmospheric initial conditions and boundary forcings (Hoskins 2012). Predictability on subseasonal time scales is limited by the use of imperfect initial conditions and imperfect numerical formulations in prediction models (Lorenz 1963, 1982; Palmer et al. 2009; Leutbecher et al. 2016). Predictability of fine-scale atmospheric features on subseasonal time scales remains poor for fundamental reasons, because of the chaos inherent in the atmosphere (Lorenz 1965; Jifan 1989; Zhang et al. 2019a). However, the predictability of large-scale, low-frequency features in the ocean, over land, and in the cryosphere lasts well beyond two weeks (Vitart et al. 2012; Buizza and Leutbecher 2015; Toth and Buizza 2019). The key sources of subseasonal predictability are Madden–Julian oscillation (e.g., Jones et al. 2004a,b; Zheng et al. 2018), snow cover (e.g., Sobolowski et al. 2010; Lin and Wu 2011; Orsolini et al. 2013), stratosphere–troposphere interaction (e.g., Baldwin et al. 2003; Domeisen et al. 2020; Schwartz and Garfinkel 2020), land conditions (e.g., Koster et al. 2011; van den Hurk et al. 2012; Prodhomme et al. 2016; Seo et al. 2019), and ocean conditions (e.g., Woolnough et al. 2007; Fu et al. 2007; Subramanian et al. 2019). Predictions on subseasonal time scales, however, need to be averaged on large enough spatiotemporal scales to extract relevant and predictable components of the signal (Lorenz 1982; Zhu et al. 2014; Buizza and Leutbecher 2015). Since subseasonal predictions are beyond deterministic limits of predictability (i.e., about ten days), these predictions are produced as ensembles of numerical integrations, describing a range of possibilities instead of a unique best estimate of the future state. This shift from determinism to probabilism has been a major breakthrough in extending the predictability horizon of subseasonal predictions (Palmer 2012).

With a transition toward low carbon energy systems, the energy industry is going to be one of the most important end-users of subseasonal predictions (White et al. 2017). Skillful subseasonal predictions of essential climate variables such as wind speed, solar radiation, and surface temperature can inform the energy industry about expected renewable energy production and energy consumption, and further prepare the sector for any possible risks that may arise due to anomalies. A nonexhaustive list of applications in the energy sector for

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/MWR-D-22-0170.s1.

Corresponding author: Naveen Goutham, naveen.goutham@edf.fr

DOI: 10.1175/MWR-D-22-0170.1

which subseasonal predictions can be instrumental includes determining required reserve levels, maintenance scheduling, assessment of extreme risks, determining grid transmission capacity, and trading electricity in power markets.

Europe, being one of the world's largest energy consuming and greenhouse gas emitting regions, sits at the forefront of the energy transition (Liobikienė and Butkus 2017; Jonek-Kowalska 2022). In the European Union, wind power is becoming the largest renewable source of electricity (IEA 2020). Hence, we focus this study on subseasonal predictions of 100-m wind speed and 2-m temperature over Europe. Several studies have assessed the quality of subseasonal predictions of wind speed and surface temperature over Europe, and have found skillful predictions relative to climatology for weekly mean quantities beyond two weeks (e.g., Lynch et al. 2014; Monhart et al. 2018; Diro and Lin 2020; Dorrington et al. 2020; Goutham et al. 2022). Although the fundamental sources of subseasonal predictability have been identified (Vitart et al. 2012), the physical relationships between large-scale, low-frequency fields and surface fields (i.e., within the planetary boundary layer) are not well represented in subseasonal prediction models due to parameterizations (Palmer et al. 2009; Leutbecher et al. 2016; Robertson and Vitart 2018; Lledó and Doblas-Reyes 2020). In addition, the forecast errors of surface fields grow relatively faster than that of large-scale fields due to increased sensitivity of the former to model parameterizations (e.g., Buizza and Leutbecher 2015; Toth and Buizza 2019). Given the longer skill horizon of large-scale fields compared to surface fields (Buizza et al. 2015; Toth and Buizza 2019; Büeler et al. 2021), there is an opportunity to improve subseasonal surface-field predictions by accounting for the misrepresentations in physical relationships between large-scale and surface fields using historical data. In other words, the information contained in the prediction of large-scale fields is more reliable than that in surface fields, and statistical downscaling techniques can be implemented to correctly transfer this information from large-scale fields to surface fields (e.g., Scaife et al. 2014; Manzanas et al. 2018; Goutham et al. 2021).

The most popular statistical downscaling techniques are the linear methods due to their transparency and ease of interpretation (Benestad et al. 2008; Wilks 2019). Generally, linear statistical downscaling is done in three stages; one, choosing predictors which have physical relationships with the predictand; two, obtaining the linear relationship between predictors and the predictand; and finally, using future dynamical predictions of predictors to reconstruct the predictand. In a majority of studies focusing on statistical downscaling of surface variables over Europe, weather regimes obtained from dimension reduction or clustering of geopotential height at 500 hPa (Z500) are used as predictors, which are then regressed on surface variables (e.g., Grams et al. 2017; Alonzo et al. 2017; Ramon et al. 2021). The obtained coefficients are then employed on future predictions of weather regimes to reconstruct surface fields. Z500 has long been the variable of choice to determine weather regimes as it represents the mid-troposphere, making it easier to capture large-scale flow (Wallace and Gutzler 1981; Cheng and Wallace 1993; Wilby and Wigley 1997; Plaut and Simonnet 2001; Alonzo et al. 2017). Alonzo et al. (2017) have developed a methodology to estimate the distribution of surface wind speed over France based on the knowledge (or forecast) of the large-scale atmospheric state, the latter being summarized by the first few patterns obtained through principal component analysis (PCA). It was verified that these patterns or empirical orthogonal functions (EOFs) represent classical Euro-Atlantic weather regimes. Although each weather regime is associated with a set of surface meteorological conditions (van der Wiel et al. 2019), the main limitations of the use of classical weather regimes for predicting surface fields are that these weather regimes represent large-scale atmospheric variability independently of the predictand and that each surface climate variable responds differently to the same weather regime (Bloomfield et al. 2019). This calls for the development of new approaches to obtain large-scale spatial patterns of variability that take into account variability of the predictand itself (Bloomfield et al. 2019). One such approach is presented in Bloomfield et al. (2019) where they use k-means clustering to find "targeted circulation types" conditioned on the European power system. It is important to understand large-scale flow patterns that have the highest impact on surface variables as these patterns can be used to enhance the skill horizon of specific surface variables. The objectives of this research are therefore to identify spatial patterns of variability of Z500 conditioned on 100-m wind speed and 2-m temperature over Europe, and to use ensemble predictions of these patterns to improve subseasonal ensemble predictions of 100-m wind speed and 2-m temperature.

In this study, we employ a multivariate statistical technique called redundancy analysis (RDA) between the Z500 field and the surface fields to obtain patterns of Z500 that maximize explained variance of the surface variables (von Storch et al. 1999; Tippett et al. 2008; Wilks 2014, 2019). We then apply the estimated linear regression coefficients on ensemble dynamical predictions of a restricted number of RDA patterns of Z500 to obtain statistical ensemble predictions of surface variables. Several dimension reduction methods exist to summarize coupled variations of large-scale fields and surface fields using a few patterns and their corresponding coefficients (von Storch et al. 1999; Tippett et al. 2008; Wilks 2019). Among these methods, redundancy analysis distinguishes itself from classical multivariate techniques such as canonical correlation analysis or maximum covariance analysis by being asymmetric in the treatment of predictor and predictand (i.e., it distinguishes dependent and independent variables), as is the case with multilinear regression (von Storch et al. 1999; Wang and Zwiers 2001; Tippett et al. 2008; Wilks 2014). As far as the authors are aware, this is the first study to compare explanatory power between the patterns obtained using RDA of Euro-Atlantic Z500 against those derived using PCA, when used as predictors in a multilinear regression model to predict 100-m wind speed and 2-m temperature over Europe, and also the first to reveal the RDA patterns of Z500 conditioned on these two surface climate variables. In addition, contrary to several studies which have discarded dynamical predictions of surface variables completely in favor of statistical predictions (e.g., Alonzo et al. 2017; Ramon et al. 2021), we demonstrate how a simple combination of dynamical predictions of surface

variables with statistical predictions derived from redundancy analysis can enhance prediction skill. Although the idea of combining dynamical and statistical predictions has already been illustrated in some recent studies on seasonal time scales (e.g., Schepen et al. 2012, 2014, 2016; Strazzo et al. 2019), in this study, we demonstrate the value gained through a combination on subseasonal time scales. We also explore forecast quality attributes of different ensemble predictions to identify those that lead to differences in predictive quality between dynamical and combined (i.e., dynamical + statistical) predictions.

The article is organized as follows: Section 2 outlines the data used; section 3 describes redundancy analysis, the combination of dynamical and statistical predictions, and the metrics used to evaluate quality of predictions; section 4 presents the results in three parts: (i) compares and contrasts patterns obtained using RDA against those of PCA, (ii) compares the quality of different ensemble predictions, and (iii) takes a closer look at forecast attributes that contribute to differences in prediction quality between different ensemble predictions; and sections 5 and 6 are reserved for discussions and conclusions, respectively.

2. Data

a. Forecasts and reforecasts

The forecasts and retrospective forecasts (reforecasts) data used in this study originate from extended-range predictions (Vitart et al. 2017) of the European Centre for Medium-Range Weather Forecasts (ECMWF). The medium-range (i.e., up to two weeks) ocean–atmosphere coupled ensemble forecasts are extended to 46 days twice a week at 0000 UTC on Mondays and Thursdays to produce extended-range ensemble predictions (Vitart et al. 2019). The operational ensemble predictions consist of 51 members (50 perturbed + control). The perturbed members are obtained using singular vectors (Leutbecher 2005; Leutbecher and Palmer 2008) and ensemble data assimilation (Buizza et al. 2008; Isaksen et al. 2010). Stochastically perturbed parameterization tendencies (SPPT) scheme is used to represent model uncertainty (Buizza et al. 1999; Palmer et al. 2009; Leutbecher et al. 2016). These predictions are originally issued at a spatial resolution of Tco639L91 (~18 km) up to a lead time of 15 days, and at Tco319L91 (~36 km) after (Vitart et al. 2017, 2019).

The operational prediction model begins to drift significantly from reality after about ten days of coupled integrations. This drift can be attributed to inherent atmospheric unpredictability (Zhang et al. 2019b; Žagar and Szunyogh 2020), and the use of imperfect initial conditions and imperfect representation of physical processes in the numerical model (Palmer et al. 2009; Leutbecher et al. 2016). It is imperative to remove the drift before employing the model. The ECMWF produces reforecasts to estimate and remove the operational model drift (Vitart et al. 2008). A reforecast set consists of ensemble forecasts of 11 members (10 perturbed + control) issued for the same calendar day of the year as the operational forecast over each of the past 20 years. ERA5 reanalysis provides the initial conditions for the reforecasts. This reforecast set with 220 integrations (20 years × 11 members) allows for evaluation of the model climatology of operational forecasts.

We retrieve forecasts and the corresponding reforecasts of 2-m temperature (T2m), zonal and meridional components of 100-m wind speed, and geopotential at 500 hPa issued during boreal winter (DJF) on a global grid between December 2016 and February 2020. The retrieved spatial resolution is 0.9° and the temporal resolution is 6 h (instantaneous values at 0000, 0600, 1200, and 1800 UTC). The data are retrieved from the Meteorological Archival and Retrieval System (MARS) of the ECMWF. The 100-m wind speed (U100) is computed as the square root of the sum of squares of zonal and meridional components. The geopotential height (Z500) is computed by dividing the geopotential by Earth's gravitational acceleration g ($=9.806$ m s^{-2}). As the prediction model is undergoing periodic improvements, the dataset used in this study consists of forecasts and reforecasts from several versions (CY43R1, CY43R3, CY45R1, and CY46R1) (Vitart et al. 2019). Nevertheless, the differences in model formulation and hence the statistics between different versions are marginal (refer to appendix A in Goutham et al. 2022). We focus on boreal winter in this study as this season experiences high variability in wind energy production in addition to increased energy demand mainly for space heating. Furthermore, predictions are more skillful in winter compared with other seasons due to stronger boundary conditions (e.g., sea surface temperature gradients), reinforced coupling (e.g., stratosphere–troposphere), and enhanced memory of initial conditions such as soil moisture among others (Robertson and Vitart 2018). In this study, only the perturbed members of forecasts and reforecasts are used. The reader is referred to the data availability statement to learn about the missing control member. All the results shown in this study involving operational predictions rely on reforecasts for calibration as explained in appendix A.

b. Reference

Generally, the forecast quality is assessed by comparing against observations (Coelho et al. 2019; Wilks 2019). However, in the absence of a serially complete and spatially coherent observed dataset, reanalysis is used as a reference in forecast verification (Kalnay 2003). In this study, we use ERA5 reanalysis (Hersbach et al. 2020) as reference. ERA5 reanalysis is a fifth-generation high-resolution (hourly output, 31-km horizontal grid spacing) reanalysis produced using 4D-Var data assimilation and the CY41R2 version of the Integrated Forecast System of the ECMWF (Hersbach et al. 2020). We retrieve ERA5 reanalysis of T2m, zonal and meridional components of 100-m wind speed, and geopotential at 500 hPa on a global grid between January 1979 and January 2021 at the same spatial and temporal resolution as the forecasts. The data are retrieved from the Climate Data Store of the Copernicus Climate Change Services (Raoult et al. 2017). The 100-m wind speed and geopotential height are computed as previously described. Although ERA5 reanalysis shows cold biases

in representing surface temperature over the Iberian Peninsula and the Mediterranean (Johannsen et al. 2019), it represents the means and extremes well over most of Europe (e.g., Simmons et al. 2021; Velikou et al. 2022). Although ERA5 reanalysis severely underestimates the mean winds over complex terrain, it represents the variability of wind speed more realistically compared with other reanalysis datasets over Europe (e.g., Ramon et al. 2019; Jourdier 2020; Dörenkämper et al. 2020; Brune et al. 2021; Molina et al. 2021; Murcia et al. 2022). Inspite of the biases, the representation errors of ERA5 reanalysis are small, and hence acceptable for verification (Ramon et al. 2019; Velikou et al. 2022) and statistical modeling (Tarek et al. 2020). Accordingly, ERA5 reanalysis is used as a reference in forecast verification and as well as for training the statistical model in this study.

3. Methodology

a. Redundancy analysis

RDA is a multivariate statistical technique that attempts to find lower-dimensional patterns of linear dependence between two multivariate datasets (i.e., between predictor and predictand) maximizing the coefficient of determination of linear regression (von Storch et al. 1999; Wang and Zwiers 2001; Tippett et al. 2008; Wilks 2014). There exist several other methods to find linearly coupled patterns between two multivariate datasets, notably canonical correlation analysis (CCA) and maximum covariance analysis (MCA). RDA, unlike CCA or MCA, is asymmetric in the treatment of two datasets as it identifies one as the predictor and the other as the predictand. This way, the patterns derived from RDA are specifically tailored for use in multilinear regression models:

$$\mathbf{P} = \begin{pmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,t} \\ p_{2,1} & \cdot & \cdots & p_{2,t} \\ \cdot & \cdot & \cdots & \cdot \\ p_{m,1} & \cdot & \cdots & p_{m,t} \end{pmatrix} \text{and } \mathbf{Q} = \begin{pmatrix} q_{1,1} & q_{1,2} & \cdots & q_{1,t} \\ q_{2,1} & \cdot & \cdots & q_{2,t} \\ \cdot & \cdot & \cdots & \cdot \\ q_{n,1} & \cdot & \cdots & q_{n,t} \end{pmatrix}.$$

Let $\mathbf{P}$ be the predictor anomaly matrix with each column representing an observation at each of the m grid points. Let $\mathbf{Q}$ be the predictand anomaly matrix with each column representing an observation at each of the n grid points. For illustration purpose, $\mathbf{P}$ can be thought of as Z500, and $\mathbf{Q}$ as U100. The elements of both $\mathbf{P}$ and $\mathbf{Q}$ are weighted by square root of cosine of latitude to equalize variance (von Storch et al. 1999; Wilks 2014, 2019). We use gridded Z500 weekly mean anomalies over the Euro-Atlantic (20°–80°N, 120°W–40°E) as the predictor and gridded T2m/U100 weekly mean anomalies over Europe (34°–74°, 13°W–40°E) as the predictand in this study. The choice of the predictor domain and its sensitivity to the predictand domain is discussed in section 5a. Regarding the calculation of anomalies, we tested lagging 15-yr as well as 20-yr mean climatology for computing anomalies. We observed that using lagging 15-year climatology (i.e., the most recent 15-yr period as climatology) performs better relative to using lagging 20-yr climatology (i.e., the most recent 20-yr period) in alleviating cold biases in temperature forecasts that are derived from the ongoing climate warming (not shown). This observation is consistent with the ones previously seen in the literature (e.g., Wilks 2013; Wilks and Livezey 2013; Wilks 2014). Therefore, we compute Z500, T2m, and U100 anomalies by removing lagging 15-year mean climatology from the observed weekly mean. The climatological data used to compute anomalies correspond to the same week and month of the year as the observation. Although U100 shows no particular trend, we retain the 15-yr period for computing U100 anomalies to make the inter-variable comparison consistent. We have more explanatory variables (i.e., grid points) than the number of observations in matrices $\mathbf{P}$ and $\mathbf{Q}$. Hence, to prevent over-determination and to lessen the computational burden, we perform PCA of matrices $\mathbf{P}$ and $\mathbf{Q}$ to obtain their corresponding principal components (PC) retaining 99% of the variance in the original data (von Storch et al. 1999; Wilks 2019). As the predictor and the predictand vectors are measured in different units, we normalize them by subtracting the gridpoint mean and dividing by the gridpoint standard deviation (Wilks 2019). The predictor and predictand PCs of the normalized variables $\mathbf{P}'$ and $\mathbf{Q}'$ are computed as $\mathbf{X} = \mathbf{E}_\mathbf{P}^\mathrm{T}\mathbf{P}'$ and $\mathbf{Y} = \mathbf{E}_\mathbf{Q}^\mathrm{T}\mathbf{Q}'$, respectively. Here, the matrices $\mathbf{E}_\mathbf{P}^\mathrm{T}$ and $\mathbf{E}_\mathbf{Q}^\mathrm{T}$ hold the predictor and predictand patterns, respectively. The superscript T denotes vector or matrix transpose. Using centered variables in place of normalized variables marginally degrades the results (not shown). The joint sample variance–covariance matrix of the leading predictor and predictand PCs is given by

$$\begin{aligned}\mathbf{S} &= (\mathbf{x}_1; \mathbf{x}_2; \ldots; \mathbf{x}_i; \mathbf{y}_1; \mathbf{y}_2; \ldots \mathbf{y}_j)(\mathbf{x}_1; \mathbf{x}_2; \ldots; \mathbf{x}_i; \mathbf{y}_1; \mathbf{y}_2; \ldots; \mathbf{y}_j)^\mathrm{T} \\ &= \begin{pmatrix} \mathbf{S}_\mathbf{XX} & \mathbf{S}_\mathbf{XY} \\ \mathbf{S}_\mathbf{YX} & \mathbf{S}_\mathbf{YY} \end{pmatrix}. \end{aligned} \tag{1}$$

Here, the $(i + j) \times t$ matrix $(\mathbf{x}_1\mathbf{x}_2 \ldots \mathbf{x}_i\mathbf{y}_1\mathbf{y}_2 \ldots \mathbf{y}_j)^\mathrm{T}$ is formed through concatenation of the leading i predictor PCs $\mathbf{x}_i$ and the leading j predictand PCs $\mathbf{y}_j$. Since we use standardized variables, $\mathbf{S}$ is in fact a correlation matrix (von Storch et al. 1999; Wilks 2019). Nonetheless, we use "covariance matrix" as a general terminology to describe the method. The covariance matrix of predictand PCs conditioned on predictor PCs is given by

$$\mathbf{S}_{\hat{\mathbf{Y}}\hat{\mathbf{Y}}} = \mathbf{S}_\mathbf{YX}\mathbf{S}_\mathbf{XX}^{-1}\mathbf{S}_\mathbf{XY}. \tag{2}$$

The eigen-decomposition of the square symmetric matrix in Eq. (2) yields orthonormal eigenvectors $\mathbf{B}$ and diagonal matrix $\mathbf{\Lambda}$ of positive eigenvalues λ, both sorted in descending order based on the values of λ. The columns of $\mathbf{B}$ consist in the patterns that account for the variance of predictand PCs when conditioned on predictor PCs. We can deduce the predictor patterns $\mathbf{A}$ through the equation:

$$\mathbf{A} = \frac{1}{\sqrt{\mathbf{\Lambda}}}\mathbf{S}_\mathbf{XX}^{-1}\mathbf{S}_\mathbf{XY}\mathbf{B}. \tag{3}$$

We can conveniently compute the predictor and predictand redundancy PCs using $\mathbf{V} = \mathbf{A}^\mathrm{T}\mathbf{X}$ and $\mathbf{W} = \mathbf{B}^\mathrm{T}\mathbf{Y}$, respectively. The regression coefficients of the linear relationship between

$\mathbf{V}$ and $\mathbf{W}$ are given by $\mathbf{R} = \sqrt{\mathbf{\Lambda}}$. The redundancy PCs $\mathbf{V}$ and $\mathbf{W}$ are linked through $\mathbf{W} = \mathbf{RV}$. For a given number of retained patterns, redundancy analysis guarantees that the coefficient of determination of the linear regression is maximized. In this study, we use 7-day rolling averages of ERA5 reanalysis of Z500, T2m, and U100 in a perfect prognosis framework for fitting the model (e.g., Hewitson and Crane 1996; Zorita and von Storch 1999; Ramon et al. 2021). We compute regression coefficients by fitting a separate model for each predictand. We choose the training period to be the boreal winter between December 1999 and February 2016 (i.e., 17 years).

b. Statistical and hybrid predictions

We can use redundancy regression coefficients ($\mathbf{R}$), predictor patterns ($\mathbf{A}$), predictand patterns ($\mathbf{B}$), and the relationship between predictor and predictand redundancy PCs to predict the predictand given a new set of predictors. Let $\mathbf{X_O}$ be a new set of predictor PCs computed from a new set of predictor anomaly matrix $\mathbf{P_O}$ (i.e., $\mathbf{X_O} = \mathbf{E_P^T P'_O}$ with $\mathbf{E_P^T}$ unchanged from the initial analysis). We can compute the predictand redundancy PCs as $\widehat{\mathbf{W}} = \mathbf{R^T V_O} = \mathbf{R^T A^T X_O}$. We can then obtain the standardized predictand vector anomalies using the following equation:

$$\widehat{\mathbf{Q_O}}' = \mathbf{E_Q}(\mathbf{B}^{\mathrm{T}})^{-1}\widehat{\mathbf{W}} = \mathbf{E_Q}(\mathbf{B}^{\mathrm{T}})^{-1}\mathbf{R}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{X_O}. \tag{4}$$

We consider an ensemble of weekly mean anomalies of Z500 operational extended-range predictions from ECMWF at any given lead time as $\mathbf{P_O}$. The operational predictions are bias-corrected using the mean and variance adjustment method (Torralba et al. 2017; Manzanas et al. 2019; Goutham et al. 2022) as described in appendix A. The prediction anomalies are computed in a similar way to the observed anomalies, but using a lagging 15-yr climatology derived from the reforecasts. We apply Eq. (4) on a *restricted* number of PCs ($\mathbf{X_O}$) of $\mathbf{P_O}$ to obtain an ensemble of predicted weekly mean anomalies of T2m or U100. The truncation of predictor PCs is a necessary step to optimize the accuracy of ensemble predictions, and it will be discussed further in the following sections. The predicted T2m or U100 anomalies are converted to absolute values by adding the lagging 15-yr climatology of the respective variable derived from reforecasts. Although we use dynamical predictions of Z500 to predict surface fields, we refer to the predicted vectors as statistical (ST) predictions, emphasizing the role of the statistical relationship between the predictor and the predictand. We then obtain a 100-member ensemble hybrid (HY) prediction by concatenating a 50-member statistical surface field prediction with a 50-member dynamical (DY) surface field prediction. All the ensemble members of the hybrid prediction receive equal weights.

c. Measures of prediction skill

Evaluation of probabilistic prediction skill involves measuring different aspects of prediction quality (Jolliffe and Stephenson 2003; Wilks 2019; Coelho et al. 2019). The most important attributes of a forecast/prediction quality are as follows:

- Accuracy: it measures the average distance between forecasts and observations;
- Association: it measures the strength of the relationship between forecasts and observations;
- Reliability: it measures calibration of the issued forecast probabilities;
- Resolution: it measures how the frequency of occurrence of an event varies as the issued forecast probability changes; and
- Sharpness: it measures the ability of forecasts to produce concentrated predictive distributions that are distinct from climatological probabilities.

Several scores have been proposed in the literature to assess probabilistic prediction skill taking into account these different forecast attributes (e.g., Jolliffe and Stephenson 2003). In this study, we employ the following metrics and diagnostic plots:

1) CONTINUOUS RANKED PROBABILITY SKILL SCORE (CRPSS)

The continuous ranked probability score (CRPS) measures the distance between the cumulative distribution functions (CDF) of a probabilistic prediction and an observation (Matheson and Winkler 1976; Unger 1985; Hersbach 2000). The CRPS is a negatively oriented score in that the smallest values indicate more accurate predictions. It is also a proper score as it rewards those predictions whose probabilities are concentrated around the observation (Gneiting and Raftery 2007). The CRPS has the same units as the physical quantity being verified. The CRPS can be decomposed into components consisting of reliability, resolution, and uncertainty (Hersbach 2000). The CRPSS compares the prediction skill of a given prediction system with that of a benchmark. In the absence of reliable forecasts for end-user applications, a common practice in the energy industry is to use observed climatology, a long-term average of observed weather (typically 35 years), as the expected weather. In this study, we use a 35-yr lagging observed climatology, derived from ERA5 reanalysis, as a 35-member ensemble benchmark prediction (CL). This climatological data corresponds to the same week and month of the year as the dynamical prediction but taken over the last 35 years. The choice of a 15-yr lagging climatology for computing anomalies, as described in section 3a, is solely to alleviate cold bias in statistical T2m predictions. Since the prediction systems compared in this study (i.e., dynamical, statistical, hybrid, and climatological) are composed of different ensemble sizes, we compute fair-CRPS (FCRPS) and fair-CRPSS (FCRPSS) to have an unbiased estimate of the scores (Ferro 2014). Skillful predictions should have FCRPSS greater than zero. The standard practice in forecast verification is to compute scores for reforecasts, and use these scores as an indication of the skill of the operational forecasts (Jolliffe and Stephenson 2003). However, Goutham et al. (2022) have shown that the skill of operational predictions on S2S time scales over Europe is higher than that of reforecasts. The improved skill of operational predictions is mainly attributed to their larger ensemble size relative to the reforecasts. Therefore, we compute all the scores and diagnostic plots for operational

dynamical predictions and the corresponding statistical and hybrid predictions in this study. In particular, we first calculate the FCRPS of weekly averaged dynamical, statistical, hybrid, and climatological predictions for each of the forecast issue dates and at each of the considered lead times. We then compute FCRPSS and its mean over all the forecast issue dates using climatological predictions as the benchmark. We apply the Wilcoxon signed-rank test (Wilcoxon 1945; Conover 1971; Wilks 2019) (see appendix B) to investigate the statistical significance of the differences of FCRPSS between hybrid and dynamical predictions.

2) Proportion of skillful forecasts (PSF)

As the mean of a distribution is sensitive to the existence of outliers, the mean-FCRPSS overemphasizes negative instances, and can therefore lead to underestimation of the prediction skill (Goutham et al. 2022). Therefore, we compute fair-proportion of skillful forecasts (FPSF) in addition to mean-FCRPSS. As the name suggests, the FPSF is a proportion of the number of predictions that have FCRPSS greater than zero to the total number of predictions considered.

3) Anomaly correlation coefficient (ACC)

The ACC is a deterministic score that measures the linear association as Pearson's correlation coefficient between the anomalies of the ensemble mean predictions and observations (Namias 1952; Wilks 2019). ACC is complementary to CRPS as it is insensitive to forecast errors in accuracy. Accordingly, a forecast can be skillful (based on ACC) if it has some temporal association with the observations, irrespective of the magnitude of its accuracy.

4) Reliability diagram

A reliability diagram is a diagnostic plot to understand the full joint distribution of predictions and observations for probabilistic predictions of a binary predictand (Sanders 1963; Jolliffe and Stephenson 2003; Wilks 2019). It can be used to measure reliability, resolution, and sharpness. In this study, we plot reliability diagrams for upper and lower terciles of weekly mean predictions averaged over a geographical domain. Geographical domain averaging, wherever applicable, is computed as the mean of cosine-latitude weighted gridpoint values.

In this study, we compute all the scores and diagnostic plots for weekly averaged quantities at four subseasonal lead times. More specifically, lead week 3 corresponds to the weekly average between days 14–20, week 4 between days 21–27, week 5 between days 28–34, and week 6 between days 35–41. We employ leave-one-out cross validation to estimate the optimum number of predictor patterns (i.e., truncation) in statistical predictions. The truncation is carried out with the criterion to optimize the median of FCRPSS of target predictions (i.e., statistical or hybrid) over the domain. This means that the number of predictor patterns retained in the best statistical predictions and in statistical predictions which form a component of hybrid predictions is different. Since hybrid predictions retain a large part of the information from dynamical predictions, the statistical predictions which form a component of hybrid predictions require deep truncation, i.e., only a small number of predictor patterns are sufficient. On the other hand, a large number of predictor patterns are required to obtain optimum statistical-only predictions. We prefer, as a truncation criterion, the median of FCRPSS to the mean as the mean is relatively more sensitive to extreme values. The number of predictor principal components required to obtain optimum U100 and T2m predictions, and the sensitivity of these predictions to the number of retained principal components will be discussed in the following section.

4. Results

As a first step, we test the efficiency of patterns obtained using PCA on one hand, to those obtained using RDA on the other, to provide information on surface fields when used as predictors in a multilinear regression model. Subsequently, we analyze the differences in prediction quality between dynamical, statistical, hybrid, and climatological predictions as well as between U100 and T2m. Finally, we explore and compare several forecast quality attributes between hybrid and dynamical predictions to understand the reasons for the differences in skill between the two.

a. How do the patterns derived using redundancy analysis differ from those obtained using principal component analysis?

We compare the differences between EOFs of Z500 anomalies over the Euro-Atlantic derived using PCA, against those derived using RDA conditioned on U100 and T2m over Europe in Fig. 1. The EOFs of the Z500 field in PCA are chosen independently of the predictand to represent the maximum possible variability contained in the predictor itself. In contrast, the EOFs of the Z500 field in RDA are chosen to maximize the explained variance of the predictand. Alonzo et al. (2017) have verified that the principal components obtained through PCA of Z500 over the Euro-Atlantic represent the classical Euro-Atlantic weather regimes.

The first three patterns shown in Fig. 1a resemble classical weather regimes of North Atlantic Oscillation, Scandinavian regime, and Atlantic regime, respectively (Alonzo et al. 2017; Bloomfield et al. 2019; van der Wiel et al. 2019; Garrido-Perez et al. 2020). Please note that the patterns in Fig. 1 are sign indefinite, and that the color bars have no units as the units are carried by the corresponding PCs. The imprints of weather regimes on 10-m wind speed and T2m can be obtained from Bloomfield et al. (2019), van der Wiel et al. (2019), and Garrido-Perez et al. (2020). The first observation that can be made from Z500 patterns in Fig. 1 is that the centers of action of RDA (top rows in Figs. 1b and 1c) are shifted toward or onto the European domain. This is logical as RDA patterns are conditioned on U100 and T2m over the European domain. Besides, we can notice variations in the strengths of troughs and ridges between PCA and RDA Z500 patterns. The Z500 patterns obtained using RDA further display intervariable differences that may be attributed to the behavior of the conditioned variable itself. Some of the Z500 patterns obtained using RDA may be seen as perturbations of those

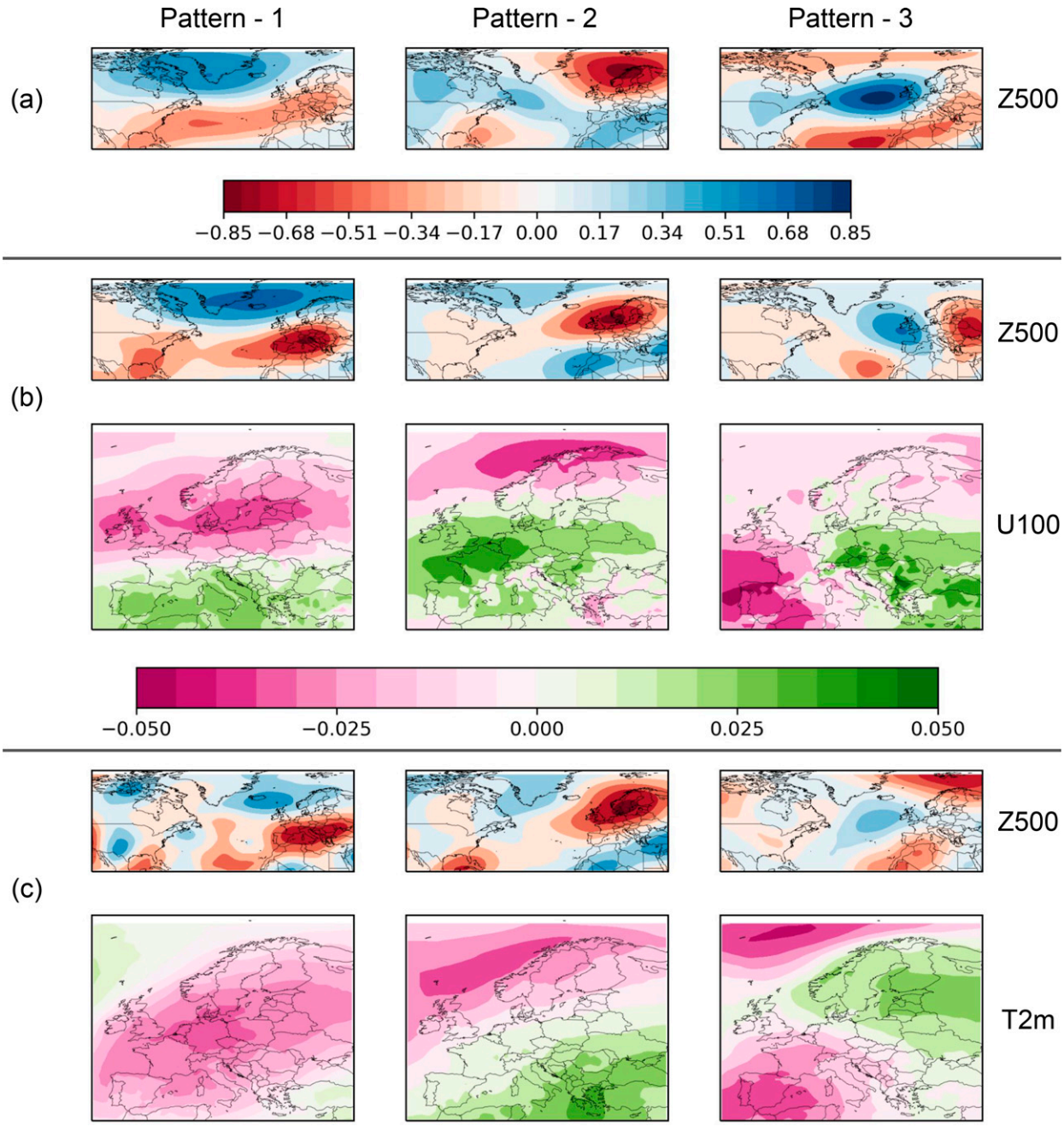

FIG. 1. (a) The first three patterns or empirical orthogonal functions of Z500 anomalies over the Euro-Atlantic computed through principal component analysis. (b) The first three paired patterns of redundancy analysis of Z500 anomalies conditioned on U100 anomalies. (c) The first three paired patterns of redundancy analysis of Z500 anomalies conditioned on T2m anomalies. The top rows in (b) and (c) represent Z500 RDA patterns, and the corresponding bottom rows represent imprints of U100 and T2m, respectively. Please note that the color bars have no units, and that the signs are arbitrary. The patterns shown are unrotated.

obtained using PCA, but with major changes in the relative importance of patterns for surface field prediction.

The imprints of classical weather regimes on 10-m wind speed and T2m are illustrated in Fig. 2 in Bloomfield et al. (2019). Although RDA surface field patterns in Fig. 1 in this work and the surface responses of weather regimes in Fig. 2 in Bloomfield et al. (2019) are not measured in the same units, they can still be compared assuming an equivalent multiplication factor to U100 and T2m imprints in Fig. 1. Overall, it is conspicuous that the surface imprints of RDA patterns are stronger and more concentrated over Europe compared with the surface responses of classical weather regimes. The imprints

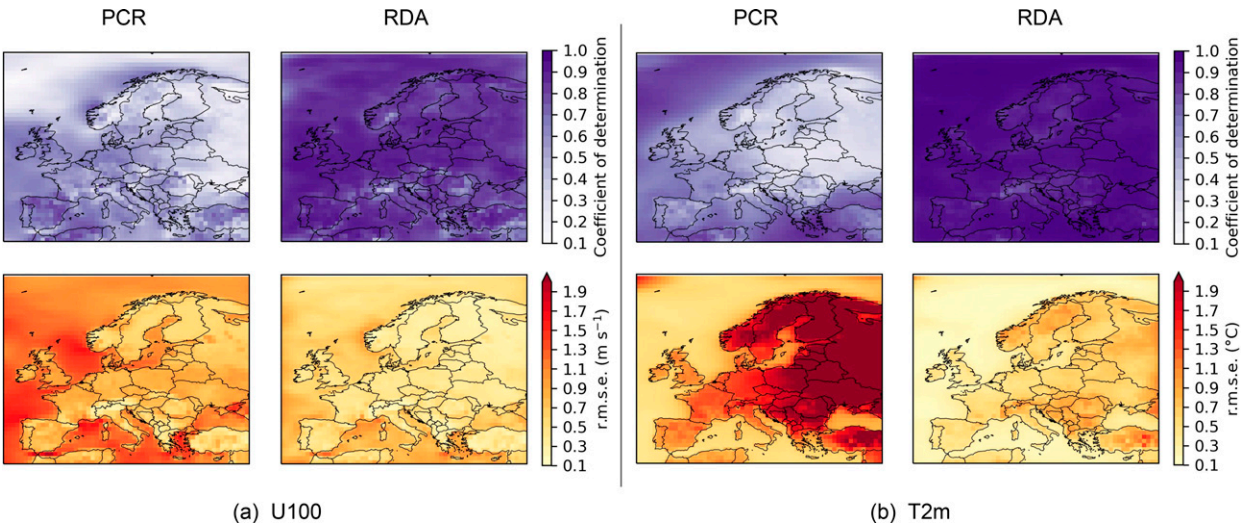

FIG. 2. Comparison of regression performance between principal component regression (PCR) and redundancy analysis (RDA) models. Performance is measured using coefficient of determination (R^2) of the linear fit between the predictor and the predictand, and root-mean-squared error (rmse) between the predicted values and ERA5 reanalysis. The models are fit for weekly averages for the boreal winter between December 1999 and February 2016. (a) U100 and (b) T2m.

of RDA patterns on U100 in Fig. 1 show anomalous meridional and zonal dipoles which are originally absent in the surface responses of classical weather regimes on 10-m wind speed [Fig. 2 in Bloomfield et al. (2019)]. Although there are similarities in the first two imprints of weather regimes and RDA patterns on T2m, the center of the anomaly of the imprint corresponding to the first RDA pattern is shifted toward the southwest, while the imprint corresponding to the second RDA pattern shows a stronger dipole with a stretched northern anomaly center relative to the responses of weather regimes on T2m. The imprint of the third RDA pattern on T2m is significantly different from that of the Atlantic regime in Fig. 2 of Bloomfield et al. (2019) and shows pronounced variations along with a tripole. The subsequent patterns are not shown in this work, but they present similar characteristics. Overall, the patterns in Fig. 1 indicate stronger surface imprints of RDA patterns compared with weather regimes on both U100 and T2m.

Having understood the differences in surface imprints between weather regimes and RDA patterns, we now compare the statistical explanatory power between PCA and RDA Z500 patterns, when used as predictors in a multilinear regression model, to accurately reconstruct surface fields. In Fig. 2, we compare the performance of regression between principal component regression (PCR) and RDA models. Particularly, we use coefficient of determination (R^2) and root mean squared error (rmse) of the linear fit between the predictor and the predictand as evaluation metrics. The R^2 measures the proportion of variation of the predictand that is accounted for by regression. Accordingly, the higher the R^2, the more is the predictand explained by the predictor. Kindly note that the R^2 presented in Fig. 2 shows the explained variance of individual grid points, while redundancy analysis maximizes the R^2 averaged over the domain. In PCR, we first compute the PCs of the Euro-Atlantic Z500 anomaly field through principal component analysis, and then use these PCs as predictors to predict U100 and T2m over Europe via standard linear regression (Wilks 2019). In Fig. 2, we retain the same number of PCs (i.e., all) for both PCR and RDA methods to facilitate comparison. From Fig. 2, it is conspicuous that the R^2 of RDA, with domain averages for U100 and T2m being 0.83 and 0.94, respectively, is substantially higher than that of PCR (domain averages of U100 and T2m being 0.46 and 0.56, respectively) for both the variables. Consequently, the rmse of the linear fit between the predictor and the predictand of RDA, with domain averages for U100 and T2m being 0.53 m s^{-1} and 0.46°C, respectively, is lower than that of PCR (domain averages of U100 and T2m being 0.96 m s^{-1} and 1.27°C, respectively). The spatial variations of rmse of U100 and T2m resemble that of the interannual variability of the respective variables (see appendix C). The R^2 of RDA for T2m is relatively high compared with U100. Despite RDA models having higher R^2, the R^2 for U100 drops to values below 0.5 over mountainous and other regions where the local effects are considerable. In general, using RDA Z500 patterns conditioned on the targeted predictand is advantageous over the use of Z500 patterns derived using PCA.

b. *How do the different types of ensemble predictions compare?*

In this section, we compare the skill of dynamical, statistical, hybrid, and climatological predictions in predicting U100 and T2m over Europe. To begin, we consider one case for illustrative purposes: the temporal evolution of ensemble probability density functions (PDFs) of dynamical, statistical, hybrid, and climatological U100 predictions over southern Scandinavia initialized on 6 February 2017 is illustrated in Fig. 3. The chosen domain, i.e., southern Scandinavia, spans 52.0°–61.0°N, 4.4°–19.0°E (see appendix C), and the PDFs are computed for weekly means. The domain averaging is

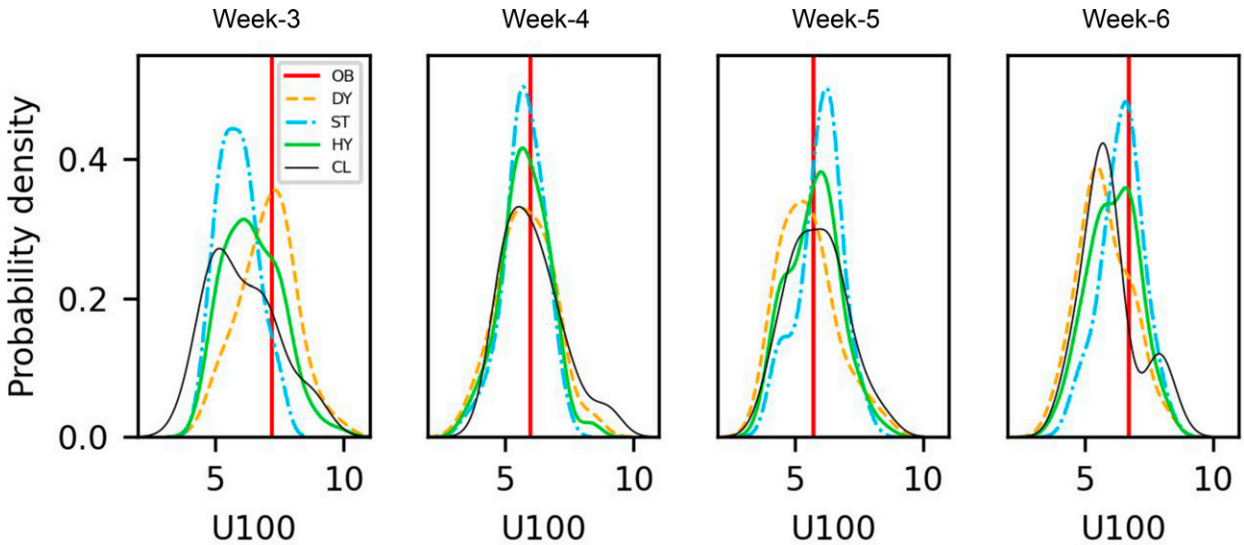

FIG. 3. Illustration of the temporal evolution of ensemble probability density functions (PDFs) of dynamical (DY), statistical (ST), hybrid (HY), and climatological (CL) U100 predictions. The PDFs are computed as kernel density estimates (Gaussian kernel) using ensemble members of weekly mean values averaged over southern Scandinavia (52.0°–61.0°N, 4.4°–19.0°E). The gridpoint values are weighted by the cosine of their respective latitude before computing domain average. This illustration corresponds to dynamical predictions initialized on 6 Feb 2017. The red vertical line in each of the panels indicates the observed weekly mean (OB).

computed as the mean of the cosine of latitude weighted gridpoint values. Besides having a longer skill horizon of subseasonal U100 predictions compared to other European regions (Goutham et al. 2022), southern Scandinavia is one of the most important regions for the wind energy industry in Europe (WindEurope 2022). Hence, we consider southern Scandinavia for illustration purposes. This specific forecast (initiated on 6 February 2017) was chosen as it is qualitatively representative of the overall results. In Fig. 3, the climatological predictions correspond to the same week and month of the year as the dynamical predictions but taken over each of the previous 35 years.

In week 3 in Fig. 3, the PDF of the dynamical prediction (μ = 7.03 m s^{-1} and σ = 1.08 m s^{-1}) is closer to the observation (=7.22 m s^{-1}), and therefore it appears to be more accurate than statistical prediction (μ = 5.88 m s^{-1} and σ = 0.76 m s^{-1}). However, dynamical predictions begin to converge toward their model climatology starting week 4. The statistical predictions are usually sharper compared with dynamical predictions at short lead times. This is attributed to slower evolution of large-scale fields compared to surface fields (e.g., Buizza and Leutbecher 2015; Robertson and Vitart 2018). Beyond week 4, statistical predictions typically carry more valuable information relative to their dynamical counterparts and thus contribute greatly to hybrid prediction accuracy. To illustrate, the week-6 statistical prediction (μ = 6.45 m s^{-1} and σ = 0.82 m s^{-1}) is closer to observation (=6.73 m s^{-1}) compared with dynamical prediction (μ = 5.79 m s^{-1} and σ = 0.99 m s^{-1}). The statistical predictions are not perfect, and they are indeed only as good as the skill of large-scale fields in dynamical predictions. They fail when dynamical predictions fail, for instance when dynamical predictions are initialized during days closer to sudden stratospheric warming events (e.g., Gerber et al. 2009; Tripathi et al. 2015). For curious readers, some additional examples of comparison of different predictions are illustrated in the online supplemental material. Overall, the PDFs in Fig. 3 suggest that the hybrid predictions may be more accurate than either dynamical or statistical predictions beyond week 3.

We understood the behavior of different ensemble predictions for one particular forecast in Fig. 3. In this section, we look at an overall assessment of U100 forecasts initialized in the boreal winter between December 2016 and February 2020 over southern Scandinavia. The comparison of the temporal evolution of fair-CRPSS between dynamical, statistical, and hybrid U100 predictions over southern Scandinavia is shown in Fig. 4. These violin plots are produced by aggregating the domain-averaged fair-CRPSS of all the forecasts initiated in boreal winter between December 2016 and February 2020. In week 3, the ocean–atmosphere coupled dynamical predictions still carry important information about U100 over southern Scandinavia. The dynamical predictions, with a mean of FCRPSS of −0.0008 and a median of FCRPSS of 0.06, perform better than statistical predictions (mean = −0.10 and median = −0.04). The PDF of statistical predictions is heavily skewed toward negative values. The hybrid predictions, with a mean of 0.02 and a median of 0.07, perform better than either dynamical or statistical predictions. In week 3, dynamical predictions have a major contribution (compared with statistical predictions) to the improved skill of hybrid predictions. Contrary to dynamical predictions, the statistical predictions become more skillful with increasing lead time. Hence, statistical predictions contribute considerably to the improved skill of hybrid predictions at longer leads compared to shorter leads. The interquartile range (IQR) of hybrid predictions decreases with increasing lead time. This can be attributed to the decreasing IQR of statistical predictions with lead time.

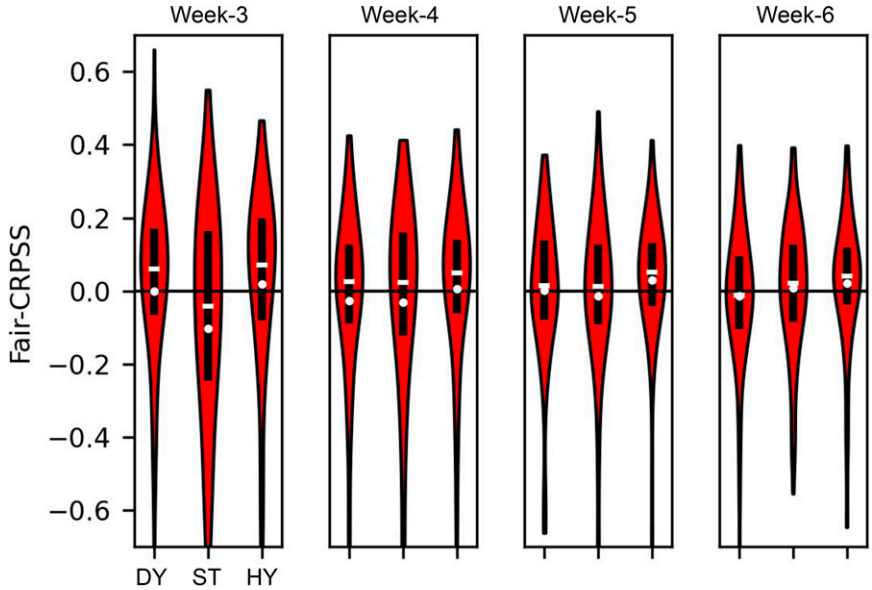

FIG. 4. Illustration of the temporal evolution of Fair-CRPSS of dynamical (DY), statistical (ST), and hybrid (HY) U100 predictions averaged over southern Scandinavia. In these standard violin plots, horizontal white dashes indicate the median, white circles indicate the mean, black boxes indicate the first and third quartiles, and black curves symmetric about the vertical (enclosing the red region) indicate the probability density of the fair-CRPSS. The left, middle, and right violin plots in each of the panels correspond to dynamical, statistical, and hybrid predictions, respectively. Values above zero indicate skillfulness of the respective predictions relative to climatology.

Beyond week-3, the improvements of hybrid predictions relative to their dynamical counterparts are statistically significant with p values ≤ 0.005 based on a Wilcoxon signed-rank test. Overall, the means and medians of FCRPSS of hybrid predictions, taking advantage of the strengths of the component prediction systems, are both positive and higher than either dynamical or statistical predictions at all lead times.

For a more general assessment and to understand spatial variations of skill, we now compare the skill of dynamical and hybrid predictions at the scale of grid points over Europe. Figure 5 shows the comparison of the weekly evolution of mean-FCRPSS and FPSF between dynamical and hybrid U100 predictions across Europe. Although the hybrid prediction skill in week 3 is marginally poorer relative to that of dynamical predictions over southern Europe, the former has a relatively more positive mean-FCRPSS over northern Europe. Generally, the dynamical predictions, with an exception over and around the North Sea, are hardly skillful beyond week 3. The added value of the information from the slowly evolving large-scale fields through statistical predictions can be clearly noticed starting week 5. In week 6, the average of FPSF over Europe for dynamical and hybrid predictions are 49.6% and 54.1%, respectively. This indicates that the hybrid predictions outperform both the dynamical and statistical predictions (not shown) over a large part of Europe. Similar to the results presented in Ramon et al. (2021), the improvements brought in by the hybrid predictions are more pronounced over northern Europe than southern Europe. The number of patterns retained in statistical predictions that form a component of optimum U100 hybrid predictions increases slightly with lead time. The optimum U100 hybrid prediction skill is achieved when statistical predictions are produced using 8–11 patterns on average, representing between 88% and 92% of the explained variance, depending on the lead time. As hybrid predictions are constructed by concatenating the ensemble members of dynamical and statistical predictions, the poor hybrid prediction skill over southern Europe may be attributed to the poor skill of dynamical U100 predictions as well as low R^2 of the linear fit between Z500 and U100 (Fig. 2). Overall, hybrid predictions are more skillful than dynamical predictions at all lead times over a large part of Europe. Since a major proportion of the European wind farms are concentrated in and around the North Sea (WindEurope 2022), the wind energy industry could greatly benefit from improved hybrid predictions over this region.

Analogous to U100, Fig. 6 compares the temporal evolution of mean-FCRPSS and FPSF between dynamical and hybrid T2m predictions over Europe. While hybrid predictions are typically more skillful than their dynamical counterparts over central, northern, and eastern Europe at all lead times, their skill is marginally degraded over southwestern Europe. Since we use a statistical model trained on the predictor and predictand *anomalies*, we notice the presence of cold biases in

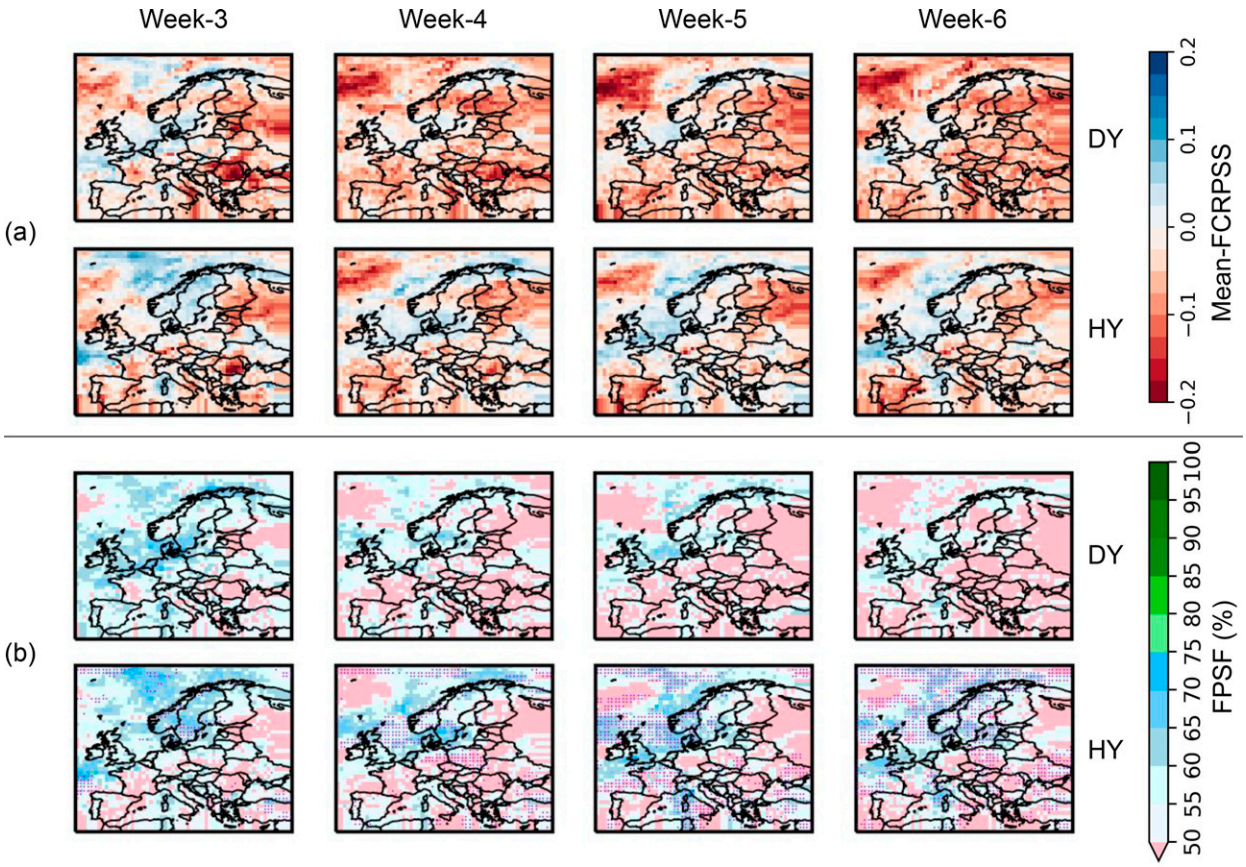

FIG. 5. Comparison of the temporal evolution of skill between dynamical (DY) and hybrid (HY) U100 predictions across Europe. (a) Mean-FCRPSS. (b) Fair-proportion of skillful forecasts (FPSF). In (a) and (b), the top rows correspond to dynamical predictions (DY), and the bottom rows correspond to hybrid predictions (HY). Values above zero in (a) and above 50% in (b) indicate skillful predictions relative to climatology. Violet dots in hybrid predictions in (b) correspond to regions with statistically significant improvements at a significance level of $p \leq 0.05$ based on a Wilcoxon signed-rank test (see appendix B).

statistical T2m predictions attributed to the ongoing climate warming. This observation is consistent with the literature (e.g., Wilks 2013; Wilks and Livezey 2013; Wilks 2014). The presence of cold biases in statistical predictions translates to biased hybrid T2m predictions, and the poor hybrid prediction skill over certain regions in Fig. 6 can be attributed to these biases. Additional postprocessing of hybrid predictions may be required for them to be useful for practical applications, and this will be discussed in section 4c of the manuscript. Contrary to U100, the optimum T2m hybrid prediction skill is achieved when statistical predictions are produced using *three* patterns on average, representing about 75% of the explained variance. The number of patterns retained in statistical predictions that form a component of optimum T2m hybrid predictions is virtually insensitive to lead time. The differences in the number of retained patterns between U100 and T2m can be attributed to the complexity of the fields themselves. Overall, hybrid predictions benefit from both the skillful dynamical predictions of surface fields at shorter leads, and the longer skill horizon of large-scale fields and their statistical relationship with surface fields at longer leads, and hence are usually more skillful than either dynamical or statistical predictions.

c. *Which forecast quality attribute(s) improves the hybrid prediction accuracy?*

The previous section has shown how the dynamical and statistical predictions complement each other to make the hybrid ensemble predictions more accurate than their components. In this section, we explore the differences in other forecast quality attributes such as association, reliability, resolution, and sharpness between dynamical and hybrid predictions, to understand the reasons for differences in accuracy between the two.

To understand the differences in association, we compare the temporal evolution of ACC between dynamical and hybrid predictions of U100 and T2m over Europe in Fig. 7. The ACC of dynamical predictions of T2m is typically higher than that of U100 at all lead times. For U100 dynamical predictions, the ACC values drop below 0.4 starting week 4, whereas, for T2m, the ACC values drop below 0.4 starting week 5. Overall, the differences in ACC between dynamical

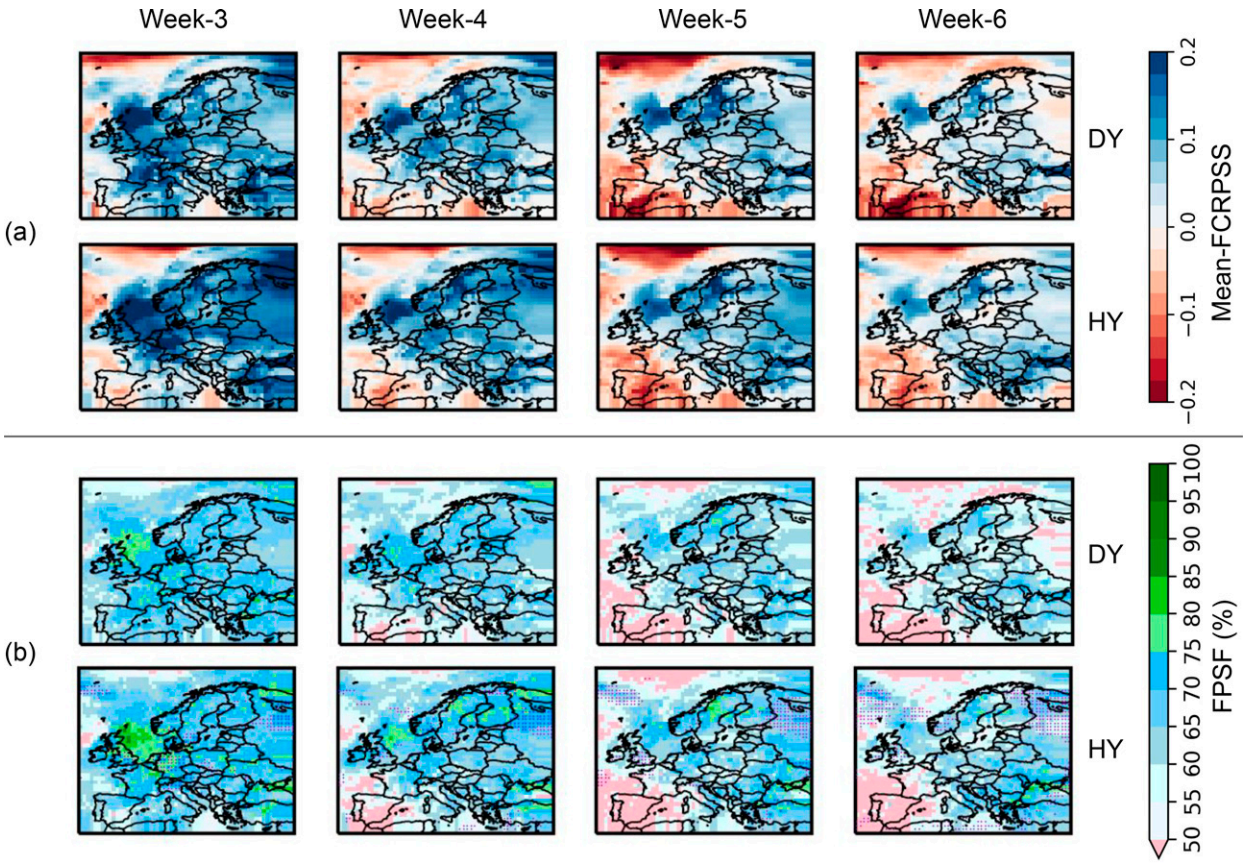

FIG. 6. As in Fig. 5, but for T2m.

and hybrid predictions are marginal. The ACC of week-3 hybrid U100 predictions, relative to dynamical predictions, is lower over southern Europe and stronger over northern Europe. However, the differences in ACC between week-4 hybrid and dynamical U100 predictions are marginal over the European domain. There are only marginal differences in ACC between hybrid and dynamical T2m predictions over continental Europe in week 3 and week 4. Similar to Figs. 5 and 6, the improvements in ACC brought in by hybrid predictions are noticeable starting week 5. Although the ACC remains poor (i.e., ≤ 0.4) for U100 predictions starting week 5, the hybrid predictions marginally improve ACC over a large part of the domain. The ACC of T2m hybrid predictions is also marginally improved over more than two-thirds of continental Europe starting week 5.

We now compare the differences in reliability, resolution, and sharpness between dynamical and hybrid predictions with the help of reliability diagrams. We recall that reliability is a measure of calibration of the issued forecast probabilities. In a reliability diagram, the reliability component can be measured as the weighted average of the squared difference between the points and the diagonal line. The number of forecasts in each bin is used as weights. The smaller the vertical distance between the points and the diagonal line, the more reliable are the predictions. In other words, perfectly reliable predictions have forecast probabilities essentially equal to observed frequencies, and hence all the points fall on the 45° diagonal line. The climatological line is the vertical or horizontal line drawn at the theoretical climatological probability of occurrence of the event considered (e.g., the climatological probability for a tercile is 1/3). Resolution measures the variations in the frequency of occurrence of an event as a function of the issued forecast probability. The resolution component can be measured as the weighted average of the squared difference between the points and the horizontal climatological line. The larger the vertical distance between the points and the horizontal climatological line, the higher the resolution. If the line connecting the points shows persistent offset from the 45° diagonal line, it indicates the presence of unconditional biases. Sharpness is a measure of the ability of forecasts to produce concentrated predictive distributions that are distinct from climatological probabilities. In a reliability diagram, the larger the horizontal distance between the climatological probability bin and the bin containing the maximum number of forecast instances, the sharper the predictions. The *no skill line* is the line located midway between the perfect reliability line and the horizontal climatological line. Accordingly, the points located within the gray region bounded by the vertical climatological line and the no skill line contribute positively to skill. For a detailed description of

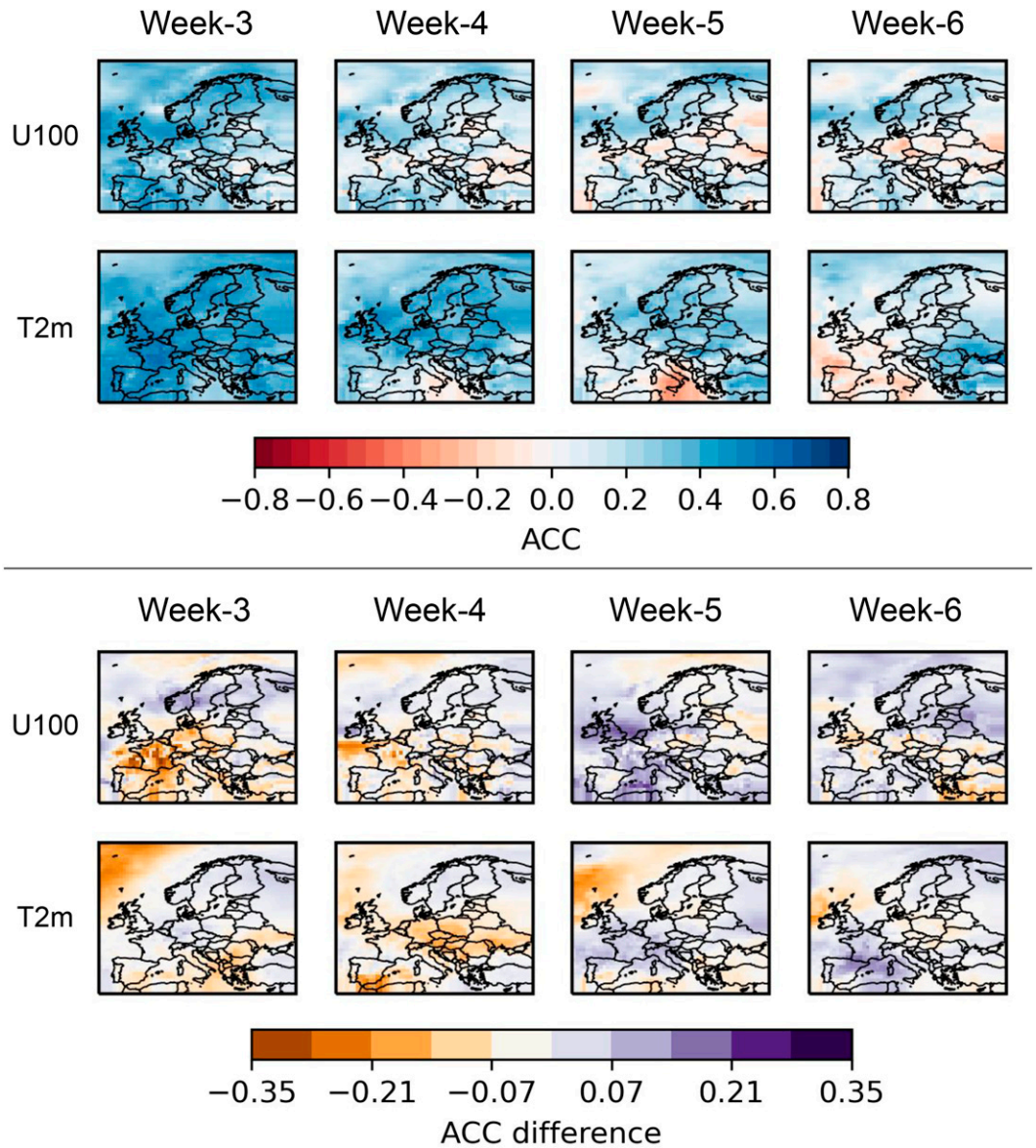

FIG. 7. (top) Temporal evolution of anomaly correlation coefficient (ACC) of dynamical predictions of U100 and T2m over Europe. (bottom) Temporal evolution of the difference of ACC between hybrid and dynamical predictions (i.e., $ACC_{HY} - ACC_{DY}$) over Europe.

reliability diagrams with examples, the reader is directed to Wilks (2019).

Figure 8 compares reliability diagrams between dynamical and hybrid models for upper and lower terciles of weekly mean U100 predictions for week 4 averaged across southern Scandinavia. We choose this particular domain and week for illustration purposes as the mean-FCRPSS of hybrid predictions is higher than that of dynamical predictions over this domain and during this week (Fig. 5). For the upper tercile, hybrid predictions with a reliability component of 0.009 are more reliable than dynamical predictions (reliability = 0.016). Both the dynamical and hybrid predictions for upper tercile have a similar resolution (~0.007). However, for the lower tercile, both the reliability and resolution components of hybrid predictions are better than that of dynamical predictions. The difference in sharpness between dynamical and hybrid predictions is marginal. Reliability diagrams for the other lead times yield similar conclusions to the one obtained here: hybrid predictions are more reliable and have a better resolution than dynamical predictions (Fig. D1 in appendix D).

Figure 9 compares the reliability diagrams for upper and lower terciles of weekly mean T2m predictions for week 4 averaged across Germany (47.3°–55°N, 6.3°–15.4°E) (see Fig. C1 in appendix C) between dynamical and hybrid predictions. For the upper tercile, the reliability component of hybrid

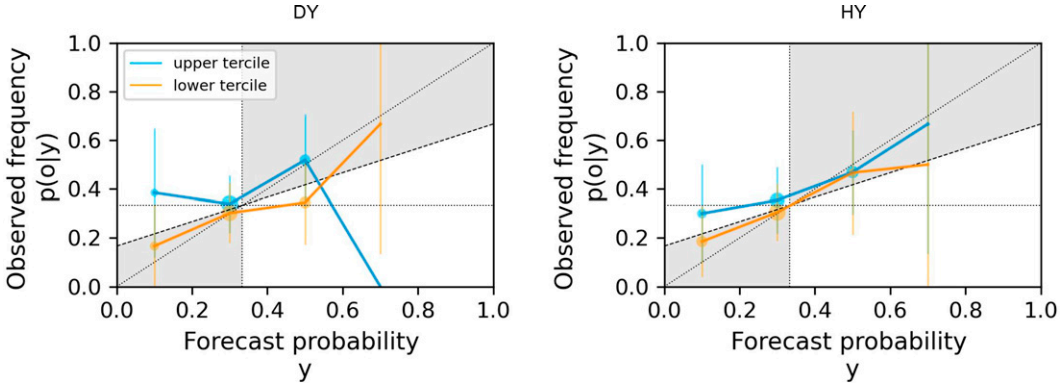

FIG. 8. Reliability diagrams for upper and lower terciles of weekly mean U100 predictions for week 4 averaged across southern Scandinavia (52.0°–61.0°N, 4.4°–19.0°E) (see Fig. C1 in appendix C). The forecasts are stratified into five bins of equal width. The size of the points is proportional to the number of forecasts in the respective bins. The vertical bars refer to the 95% confidence intervals computed through the standard parametric approach by assuming a normal distribution for the underlying data (Machin et al. 2013). The vertical and horizontal dotted lines indicate the climatological tercile probabilities (theoretically, the value is 1/3) in the forecasts and observations, respectively. Perfectly reliable predictions fall on the dotted diagonal line (45°) connecting the points (0, 0) and (1, 1). The points located within the gray area contribute positively to skill. DY is dynamical predictions; HY is hybrid predictions.

predictions (0.019) is higher than that of dynamical predictions (0.010), with both predictions having a similar resolution (~0.02). Although the reliability of upper tercile hybrid predictions looks similar to that of dynamical predictions at first sight, the hybrid predictions are in fact degraded due to the introduction of cold bias through the statistical model as a result of the warming climate. On the other hand for the lower tercile, the reliability component of hybrid predictions (0.013) is lower than that of dynamical predictions (0.018), with both predictions having a similar resolution (~0.025). The sharpness of the upper tercile hybrid predictions is marginally degraded relative to that of dynamical predictions.

We treat the cold bias of hybrid predictions by adjusting the warming trend through a simple procedure. First, we compute the observed climatology for any given week and year under consideration by aggregating T2m of the same week and the two adjacent weeks over each of the previous 15 years from ERA5 reanalysis. Then, we assume the trend to be linear and fit a trend line to this climatological data. As we have chosen a 15-yr period for climatology, the linearity assumption for the warming trend stays approximately valid (e.g., Wilks 2013; Wilks and Livezey 2013). Finally, we extrapolate the trend to the year under consideration and add it to statistical predictions. We then obtain trend-adjusted hybrid predictions (HY_{TrAd}) by concatenating dynamical predictions to trend-adjusted statistical predictions. Adjusting for the trend in this way improves the reliability component of HY_{TrAd} for the upper (0.004) tercile without degrading the resolution. Since

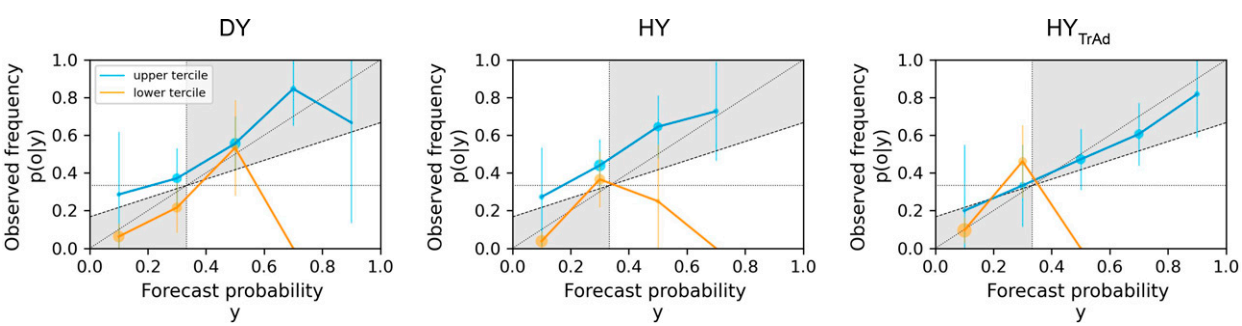

FIG. 9. Reliability diagrams for upper and lower terciles of weekly mean T2m predictions for week 4 averaged across Germany (47.3°–55°N, 6.3°–15.4°E) (see appendix C). The forecasts are stratified into five bins of equal width. The size of the points is proportional to the number of forecasts in the respective bins. The vertical bars refer to the 95% confidence intervals computed through the standard parametric approach by assuming a normal distribution for the underlying data (Machin et al. 2013). The vertical and horizontal dotted lines indicate the climatological tercile probabilities (theoretically, the value is 1/3) in the forecasts and observations, respectively. Perfectly reliable predictions fall on the dotted diagonal line (45°) connecting the points (0, 0) and (1, 1). The points located within the gray area contribute positively to skill. DY is dynamical predictions; HY is hybrid predictions; HY_{TrAd} is trend-adjusted hybrid predictions.

adjusting for the trend in this way shifts the entire distribution to the right, it may degrade both reliability and resolution for the lower tercile. Nevertheless, trend-adjusted hybrid predictions show improved sharpness for both the terciles. (The reliability diagrams for the remaining lead times are presented in Fig. D2 in appendix D.) The intensity of the cold bias in the statistical model is different in different regions within Europe. There are other more sophisticated ways to deal with trend in a nonstationary climate, some of which are described in Wilks (2013), Wilks and Livezey (2013), and Wilks (2014). An alternative way to deal with the trend is to train on detrended T2m anomalies in the statistical model. Nevertheless, exploring the efficiencies of different trend adjustment methods will depend on the targeted application, and is hence beyond the scope of this research. Overall, the presented reliability diagrams show that the improved accuracy of hybrid predictions of surface fields relative to the corresponding dynamical predictions (Figs. 4–6) stems from improved reliability and resolution.

5. Discussion

a. How sensitive are the statistical predictions to the choice of predictor domain?

The assessment in the previous section shows promising potential for extracting more skillful information from subseasonal predictions for surface variables using the methodology described in section 3. In this section, we discuss some of the choices and perspectives regarding this methodology.

We tested several predictor domains by varying size and geographical location to investigate the sensitivity of statistical prediction quality to the choice of domain. The predictor domain is tailored to describe the large-scale circulation, and more precisely features of this circulation that impact surface fields over Europe. Hence, it is logical to choose a predictor domain that is larger than the predictand domain. Given the midlatitude circulation and dynamics (westerly flow, eastward traveling perturbations), it is expected that a domain shifted westward (i.e., upstream) should be best (Alonzo 2018). We recall that the predictor domain retained in this study is the Euro-Atlantic (20°–80°N, 120°W–40°E). Displacing the predictor domain eastward (i.e., between 90°W and 70°E) without changing the size yields similar results to the retained Euro-Atlantic domain. However, choosing the predictor domain to be the same as that of the predictand domain (i.e., 34°–74°N, 13°W–40°E) considerably degrades statistical prediction quality. Shifting the predictand domain-sized predictor domain to the west between 67° and 14°W marginally (but identifiably) degrades statistical prediction quality with respect to the case when the predictand domain-sized predictor domain is centered over the predictand domain. This confirms our hypothesis that the statistical model captures prominent information from the large-scale, midtropospheric westerly flow.

b. How to further improve the hybrid prediction quality?

In this study, we used a single predictor, i.e., Z500 over the Euro-Atlantic, to improve the quality of surface field predictions. Other fields, tapping into other sources of subseasonal predictability (e.g., ocean, soil moisture, cryosphere), could be used in complement to Z500 to further improve the quality of surface field predictions (e.g., Seo et al. 2019; Domeisen et al. 2020). The redundancy analysis model can also be used as a tool to investigate physical as well as time-lag relationships between different predictors and predictands, such as in the case of assessing impacts of MJO on extratropical weather (Zheng et al. 2018).

An additional way to improve hybrid predictions involves a more clever combination of ensembles from dynamical and statistical predictions. We recall that in this study, we concatenate dynamical and statistical predictions with each having an ensemble size of 50 to obtain a 100-member equally weighted ensemble hybrid prediction. Nonetheless, combining ensembles through concatenation induces redundancy as some of the information that statistical prediction brings in may already be present in dynamical prediction. With the addition of other predictors, the redundancy of ensemble members of hybrid predictions increases substantially.

As a first attempt to put proper weights on statistical and dynamical ensemble members of hybrid predictions, we classified statistical predictions into skillful or otherwise based on the values of observed redundancy PCs with respect to their climatological distribution when the dynamical predictions are initiated. In other words, we examined whether the observed redundancy PC values being in the lower or upper extreme quantiles with respect to their climatology at the time when dynamical predictions are initialized leads to the improved or degraded skill of statistical predictions. However, the results showed no detectable relationship between the two, and hence this path was not pursued further. A promising way forward here is through a linear inverse model approach as realized in Albers and Newman (2021), and the authors plan to implement this in a future study.

There exist several alternative techniques to reasonably select ensemble members by minimizing redundancy. The most popular method for combining ensembles is Bayesian model averaging (BMA) (e.g., Schepen et al. 2012, 2014, 2016; Strazzo et al. 2019). BMA combines different models by giving different weights to the ensemble members from each model based on their performance. Another method for combining models that is gaining attention is the optimal transport distance or Wasserstein distance (e.g., Peyré and Cuturi 2019; Cumings-Menon and Shin 2020). The authors plan to investigate and compare these two methods in a future study.

6. Conclusions

With increasing de-carbonization of the energy sector (IEA 2021), the energy industry requires accurate predictions of essential climate variables such as surface temperature and 100-m wind speed across a continuum of time scales. Having accurate predictions of essential climate variables on subseasonal time scales enables the energy industry to anticipate and prepare contingency plans in the face of anomalies in wind energy production and consumption (White et al. 2017). This calls for ways to improve the skill horizon of predictions of essential climate variables.

Surface variables such as wind speed and temperature are essential for many applications, yet in the forecast models, surface variables are not the most realistic. They are indeed strongly affected by small-scale, local features, and are heavily sensitive to parameterizations, which always introduce strong uncertainties. The skill horizon of predictions of surface variables is thus limited by errors in the representation of initial conditions, model formulations, and the use of restricted spatial resolution in subseasonal prediction models. Large-scale, low-frequency fields have the advantage of being more skillful than surface fields, and in addition, they drive a large part of the variability of surface fields. In this study, we have proposed a novel methodology to improve predictions of surface variables by tapping into the large-scale, more reliable variables (e.g., Z500), and relating these to a surface variable of interest by training on the observationally derived historical data (i.e., ERA5 reanalysis). Generally, across Europe, weather regimes have been commonly used to provide a compact summary of the large-scale configuration of the atmosphere. For instance, Alonzo (2018) used weather regimes to summarize the large-scale atmospheric state and infer the likely surface wind speed distribution from these regimes using nonlinear regression. Although weather regimes are powerful tools to anticipate surface conditions, the main limitations of the use of classical weather regimes for deducing surface fields are that these weather regimes represent large-scale atmospheric variability independently of the surface fields, and that each surface climate variable responds differently to the same weather regime. This calls for the development of new approaches to obtain large-scale spatial patterns of variability which take into account the variability of the targeted surface variable.

In this study, we have employed redundancy analysis to carry out a dimension reduction of the large-scale field (i.e., Z500). Redundancy analysis provides large-scale patterns specifically designed to capture the variability of a surface field of interest. We have compared the coefficients of determination between patterns obtained using principal component analysis against those derived using redundancy analysis when used as predictors in a multilinear regression model to reconstruct surface fields. We have then employed the relationship between patterns obtained using redundancy analysis and surface fields on the subseasonal dynamical predictions of patterns to obtain statistical probabilistic predictions of surface fields. Subsequently, we have combined statistical and dynamical predictions of surface fields through a simple concatenation of the respective ensemble members. From the results presented, the following conclusions can be drawn:

1) The large-scale patterns obtained using redundancy analysis better capture surface fields over Europe compared to patterns derived using principal component analysis.
2) The added value of statistical predictions increases with lead time, and so does their contribution to the improved skill of hybrid predictions.
3) Combining dynamical and statistical predictions through a simple concatenation improves the skill of surface field predictions significantly over a large part of Europe at all lead times.
4) The improved accuracy of hybrid predictions relative to dynamical predictions stems from improved reliability and resolution. No significant changes are observed in association and sharpness between dynamical and hybrid predictions.
5) The combination of dynamical and statistical predictions can certainly be improved. Depending on the initial state and/or the forecast evolution, one may have an a priori estimate of predictability, which could inform a more efficient combination of dynamical and statistical predictions.

The redundancy analysis model employed in this study can be used to identify spatial patterns of variability that impact surface conditions at a particular location of interest such as a wind farm. Wind farms, in addition to wind speed, require information about wind direction for operational purposes, e.g., to take into account the effect of wakes. In this regard, the redundancy analysis model can be employed for wind components separately. As a perspective, the redundancy analysis model could be deployed to identify spatial patterns of variability for other climate variables such as solar radiation and precipitation, and as well as for energy variables such as renewable energy production and consumption. The skill of statistical predictions realized in this study can be decomposed into two components: one, the skill of relevant large-scale Z500 patterns in the dynamical predictions, and two, the skill of regression. Since the patterns derived using redundancy analysis differ from those obtained using principal component analysis, in terms of both spatial structure and explanatory power, it would be interesting to understand the differences in skill horizon between these patterns in dynamical predictions. As hybrid predictions developed in this study remain skillful even at a lead time of six weeks for both variables, it would be interesting to see their added value on seasonal time scales. These research questions along with the addition of other sources of predictability, and employment of more efficient ensemble selection are objectives for future studies.

Acknowledgments. Naveen Goutham would like to thank Association Nationale de la Recherche et de la Technologie (ANRT) for the Convention Industrielle de formation par la recherché (CIFRE) fellowship for his Ph.D. This work contributes to the Energy4Climate Interdisciplinary Centre (E4C) of Institut Polytechnique de Paris and Ecole des Ponts ParisTech, supported by 3rd Programme d'Investissements d'Avenir (ANR-18-EUR-0006-02). The authors acknowledge the data center Ensemble de services pour la recherche à l'Institut Pierre-Simon Laplace (ESPRI) for their help in storing and accessing the data. Finally, the authors thank the editor and the two reviewers for carefully reviewing the work.

Data availability statement. The archived ECMWF extended-range forecasts and reforecasts are published under Creative Commons Attribution 4.0 International (CC BY 4.0). However, a data access fee may be applicable. For further information, kindly refer to https://www.ecmwf.int/. The ECMWF ERA5 reanalysis data are publicly available and can be accessed through the Climate Data Store of the Copernicus

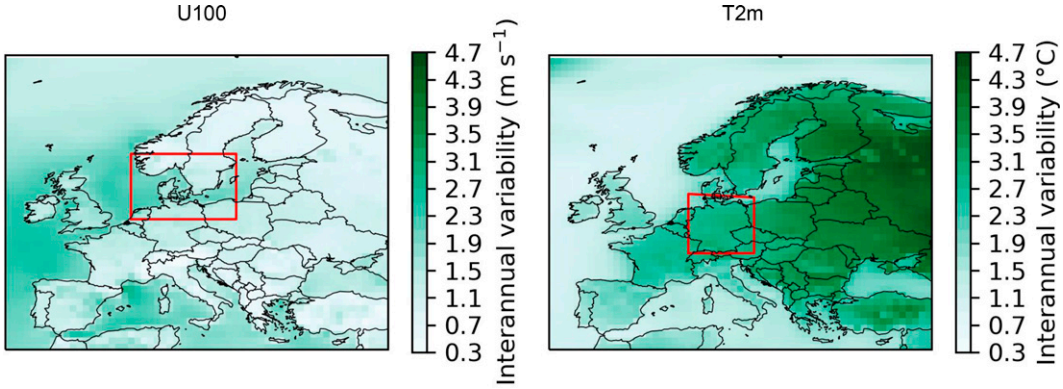

FIG. C1. Illustration of the interannual variability of boreal winter ERA5 reanalysis of 100-m wind speed (U100) and 2-m temperature (T2m) over Europe. The interannual variability is computed as the standard deviation of boreal winter weekly means between December 1999 and February 2016. The red colored rectangles in U100 and T2m correspond to southern Scandinavia (52.0°–61.0°N, 4.4°–19.0°E) and Germany (47.3°–55°N, 6.3°–15.4°E), respectively.

Climate Change Services upon registration. The control of the ensemble should be treated as another indistinguishable ensemble member. However, due to the unavailability of the control member in the internal database of Ensemble de services pour la recherche à l'Institut Pierre-Simon Laplace (ESPRI) as a result of an unintentional man-made error, we had to use only the perturbed members.

APPENDIX A

Forecast Bias Adjustment

The bias-adjusted ensemble member x_k^* for any forecast at a given lead time is

$$x_k^* = (x_k - \overline{x}_{\mathrm{cli}})\frac{\sigma_{\mathrm{ref}}}{\sigma_{\mathrm{cli}}} + \overline{o}_{\mathrm{ref}}, \tag{A1}$$

where x_k is the raw member, $\overline{x}_{\mathrm{cli}}$ and σ_{cli} are the mean and standard deviation, respectively, of all the members of the reforecast set corresponding to the forecast, $\overline{o}_{\mathrm{ref}}$ and σ_{ref} are the mean and standard deviation, respectively, of ERA5 reanalysis corresponding to the reforecasts.

APPENDIX B

Wilcoxon Signed-Rank Test for Statistical Significance

The Wilcoxon signed-rank test is a nonparametric hypothesis test that is used to investigate whether the considered data samples are derived from the same population or generating process (Wilcoxon 1945; Conover 1971; Wilks 2019). The test assesses for possible differences in location (i.e., rank) between members of a paired dataset. Here, the test statistic is based on ranks rather than numerical values of the data.

In this study, we consider a total of 103 forecasts initialized in the boreal winter between December 2016 and February 2020. We define the null hypothesis as "there is no difference in fair-CRPSS between dynamical and hybrid predictions," and the alternate hypothesis as "hybrid predictions have higher fair-CRPSS than dynamical predictions."

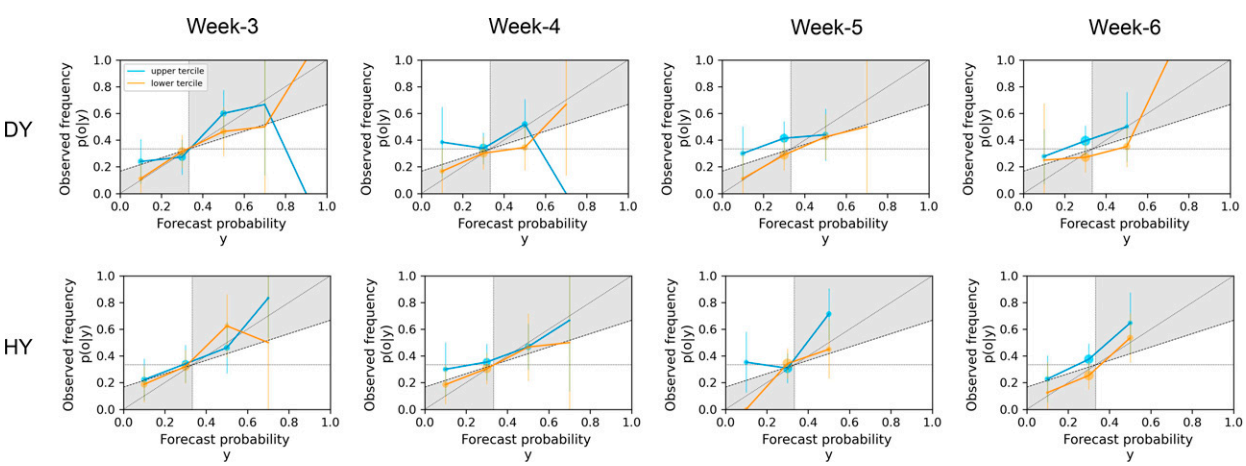

FIG. D1. As in Fig. 8, but includes additional lead times.

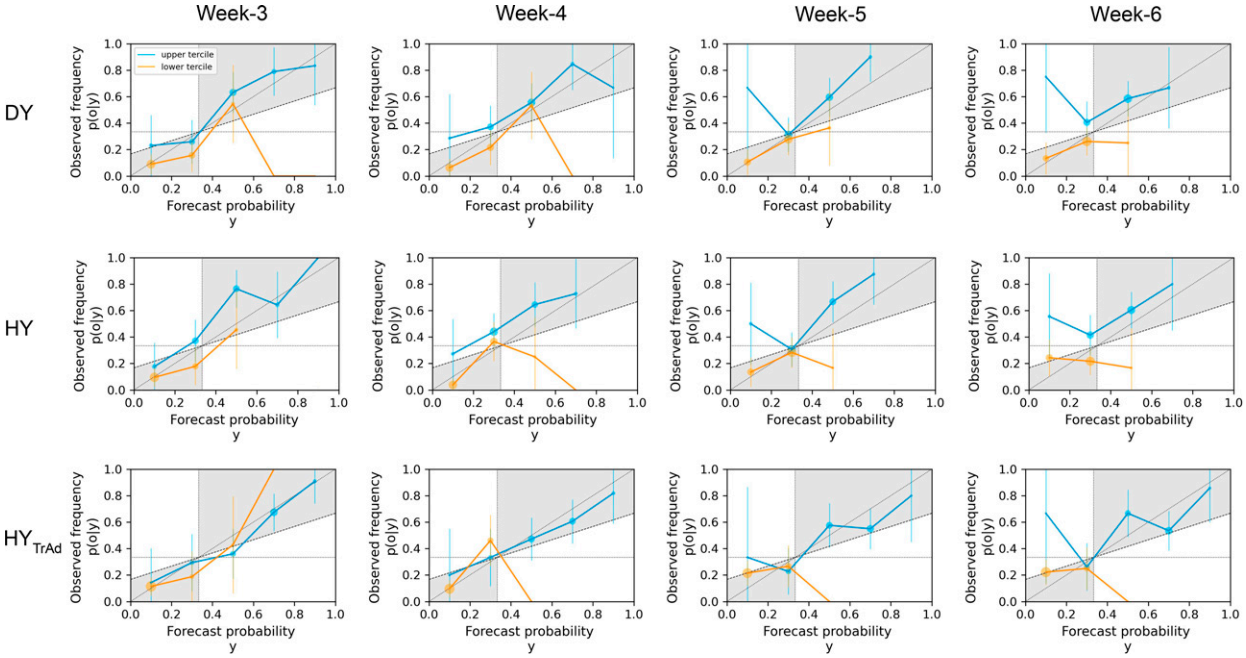

FIG. D2. As in Fig. 9, but includes additional lead times.

We treat *ties* in the paired data using the method described in Pratt (1959).

APPENDIX C

Interannual Variability of U100 and T2m over Europe

Figure C1 shows the interannual variability of boreal winter ERA5 reanalysis of 100-m wind speed (U100) and 2-m temperature (T2m) over Europe. The interannual variability is computed as the standard deviation of boreal winter weekly means between December 1999 and February 2016.

APPENDIX D

Additional Reliability Diagrams

Figures D1 and D2 show the reliability diagrams for upper and lower terciles of weekly mean predictions of 100-m wind speed across southern Scandinavia and 2-m temperature across Germany, respectively.

REFERENCES

Albers, J. R., and M. Newman, 2021: Subseasonal predictability of the North Atlantic Oscillation. *Environ. Res. Lett.*, **16**, 044024, https://doi.org/10.1088/1748-9326/abe781.

Alonzo, B., 2018: Seasonal forecasting of wind energy resource and production in France and associated risk. Ph.D. thesis, Université Paris-Saclay, 142 pp.

——, H.-K. Ringkjob, B. Jourdier, P. Drobinski, R. Plougonven, and P. Tankov, 2017: Modelling the variability of the wind energy resource on monthly and seasonal timescales. *Renew. Energy*, **113**, 1434–1446, https://doi.org/10.1016/j.renene.2017.07.019.

Baldwin, M. P., D. B. Stephenson, D. W. J. Thompson, T. J. Dunkerton, A. J. Charlton, and A. O'Neill, 2003: Stratospheric memory and skill of extended-range weather forecasts. *Science*, **301**, 636–640, https://doi.org/10.1126/science.1087143.

Benestad, R. E., D. Chen, and I. Hanssen-Bauer, 2008: *Empirical-Statistical Downscaling*. World Scientific Publishing Company, 228 pp.

Bloomfield, H. C., D. J. Brayshaw, and A. J. Charlton-Perez, 2019: Characterizing the winter meteorological drivers of the European electricity system using targeted circulation types. *Meteor. Appl.*, **27**, e1858, https://doi.org/10.1002/met.1858.

Brune, S., J. Keller, and S. Wahl, 2021: Evaluation of wind speed estimates in reanalyses for wind energy applications. *Adv. Sci. Res.*, **18**, 115–126, https://doi.org/10.5194/asr-18-115-2021.

Büeler, D., L. Ferranti, L. Magnusson, J. F. Quinting, and C. M. Grams, 2021: Year-round subseasonal 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.

——, M. Milleer, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **125**, 2887–2908, https://doi.org/10.1002/qj.49712556006.

——, M. Leutbecher, and L. Isaksen, 2008: Potential use of an ensemble of analyses in the ECMWF Ensemble Prediction System. *Quart. J. Roy. Meteor. Soc.*, **134**, 2051–2066, https://doi.org/10.1002/qj.346.

——, ——, and A. Thorpe, 2015: Living with the butterfly effect: A seamless view of predictability. *ECMWF Newsletter*, No. 145, ECMWF, Reading, United Kingdom, 18–23, https://

www.ecmwf.int/sites/default/files/elibrary/2015/17265-living-butterfly-effect-seamless-view-predictability.pdf.

Cheng, X., and J. M. Wallace, 1993: Cluster analysis of the Northern Hemisphere wintertime 500-hPa height field: Spatial patterns. *J. Atmos. Sci.*, **50**, 2674–2696, https://doi.org/10.1175/1520-0469(1993)050<2674:CAOTNH>2.0.CO;2.

Coelho, C. A. S., B. Brown, L. Wilson, M. Mittermaier, and B. Casati, 2019: Forecast verification for S2S timescales. *Subseasonal to Seasonal Prediction: The Gap between Weather and Climate Forecasting*, A. Robertson and F. Vitart, Eds., Elsevier, 337–361.

Conover, W., 1971: *Practical Nonparametric Statistics*. Wiley, 608 pp.

Cumings-Menon, R., and M. Shin, 2020: Probability forecast combination via entropy regularized Wasserstein distance. *Entropy*, **22**, 929, https://doi.org/10.3390/e22090929.

Diro, G. T., and H. Lin, 2020: Subseasonal forecast skill of snow water equivalent and its link with temperature in selected SubX models. *Wea. Forecasting*, **35**, 273–284, https://doi.org/10.1175/WAF-D-19-0074.1.

Domeisen, D. I. V., and Coauthors, 2020: The role of the stratosphere in subseasonal to seasonal prediction: 2. Predictability arising from stratosphere-troposphere coupling. *J. Geophys. Res. Atmos.*, **125**, e2019JD030923, https://doi.org/10.1029/2019JD030923.

Dörenkämper, M., and Coauthors, 2020: The making of the new European wind atlas—Part 2: Production and evaluation. *Geosci. Model Dev.*, **13**, 5079–5102, https://doi.org/10.5194/gmd-13-5079-2020.

Dorrington, J., I. Finney, T. Palmer, and A. Weisheimer, 2020: Beyond skill scores: Exploring subseasonal forecast value through a case-study of French month-ahead energy prediction. *Quart. J. Roy. Meteor. Soc.*, **146**, 3623–3637, https://doi.org/10.1002/qj.3863.

Ferro, C. T., 2014: Fair scores for ensemble forecasts. *Quart. J. Roy. Meteor. Soc.*, **140**, 1917–1923, https://doi.org/10.1002/qj.2270.

Fu, X., B. Wang, D. E. Waliser, and L. Tao, 2007: Impact of atmosphere–ocean coupling on the predictability of monsoon intraseasonal oscillations. *J. Atmos. Sci.*, **64**, 157–174, https://doi.org/10.1175/JAS3830.1.

Garrido-Perez, J. M., C. Ordóñez, D. Barriopedro, R. García-Herrera, and D. Paredes, 2020: Impact of weather regimes on wind power variability in Western Europe. *Appl. Energy*, **264**, 114731, https://doi.org/10.1016/j.apenergy.2020.114731.

Gerber, E. P., C. Orbe, and L. M. Polvani, 2009: Stratospheric influence on the tropospheric circulation revealed by idealized ensemble forecasts. *Geophys. Res. Lett.*, **36**, L24801, https://doi.org/10.1029/2009GL040913.

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.

Goutham, N., and Coauthors, 2021: Using machine-learning methods to improve surface wind speed from the outputs of a numerical weather prediction model. *Bound.-Layer Meteor.*, **179**, 133–161, https://doi.org/10.1007/s10546-020-00586-x.

——, R. Plougonven, H. Omrani, S. Parey, P. Tankov, A. Tantet, P. Hitchcock, and P. Drobinski, 2022: How skillful are the European subseasonal predictions of wind speed and surface temperature? *Mon. Wea. Rev.*, **150**, 1621–1637, https://doi.org/10.1175/MWR-D-21-0207.1.

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.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570, https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

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

Hewitson, B., and R. G. Crane, 1996: Climate downscaling: Techniques and application. *Climate Res.*, **7**, 85–95, https://doi.org/10.3354/cr007085.

Hoskins, B., 2012: Predictability beyond the deterministic limit. *WMO Bull.*, **61**, 33.

IEA, 2020: European Union 2020: Energy Policy Review. Tech. Rep., Internal Energy Agency, 310 pp., https://www.iea.org/reports/european-union-2020.

——, 2021: Net Zero by 2050: A Roadmap for the Global Energy Sector. Tech. Rep., Internal Energy Agency, 224 pp., https://www.iea.org/reports/net-zero-by-2050.

Isaksen, L., M. Bonavita, R. Buizza, M. Fisher, J. Haseler, M. Leutbecher, and L. Raynaud, 2010: Ensemble of data assimilations at ECMWF. ECMWF Tech. Memo. 636, 45 pp., https://www.ecmwf.int/node/10125.

Jifan, C., 1989: Predictability of the atmosphere. *Adv. Atmos. Sci.*, **6**, 335–346, https://doi.org/10.1007/BF02661539.

Johannsen, F., S. Ermida, J. P. A. Martins, I. F. Trigo, M. Nogueira, and E. Dutra, 2019: Cold bias of ERA5 summertime daily maximum land surface temperature over Iberian Peninsula. *Remote Sens.*, **11**, 2570, https://doi.org/10.3390/rs11212570.

Jolliffe, I., and D. Stephenson, 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley, 304 pp.

Jonek-Kowalska, I., 2022: Towards the reduction of CO_2 emissions. Paths of pro-ecological transformation of energy mixes in European countries with an above-average share of coal in energy consumption. *Resour. Policy*, **77**, 102701, https://doi.org/10.1016/j.resourpol.2022.102701.

Jones, C., D. E. Waliser, K. M. Lau, and W. Stern, 2004a: Global occurrences of extreme precipitation and the Madden–Julian oscillation: Observations and predictability. *J. Climate*, **17**, 4575–4589, https://doi.org/10.1175/3238.1.

——, ——, ——, and ——, 2004b: The Madden–Julian oscillation and its impact on Northern Hemisphere weather predictability. *Mon. Wea. Rev.*, **132**, 1462–1471, https://doi.org/10.1175/1520-0493(2004)132<1462:TMOAII>2.0.CO;2.

Jourdier, B., 2020: Evaluation of ERA5, MERRA-2, COSMO-REA6, NEWA and AROME to simulate wind power production over France. *Adv. Sci. Res.*, **17**, 63–77, https://doi.org/10.5194/asr-17-63-2020.

Kalnay, E., 2003: *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, 368 pp.

Koster, R. D., and Coauthors, 2011: The second phase of the Global Land–Atmosphere Coupling Experiment: Soil moisture contributions to subseasonal forecast skill. *J. Hydrometeor.*, **12**, 805–822, https://doi.org/10.1175/2011JHM1365.1.

Leutbecher, M., 2005: On ensemble prediction using singular vectors started from forecasts. ECMWF Tech. Memo. 462, 11 pp., https://www.ecmwf.int/node/10732.

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

——, and Coauthors, 2016: Model uncertainty representations in the IFS. *ECMWF/WWRP Workshop: Model Uncertainty*, ECMWF, Reading, United Kingdom, 2 pp., https://www.ecmwf.int/node/16369.

Lin, H., and Z. Wu, 2011: Contribution of the autumn Tibetan Plateau snow cover to seasonal prediction of North American winter temperature. *J. Climate*, **24**, 2801–2813, https://doi.org/10.1175/2010JCLI3889.1.

Liobikienė, G., and M. Butkus, 2017: The European Union possibilities to achieve targets of Europe 2020 and Paris agreement climate policy. *Renew. Energy*, **106**, 298–309, https://doi.org/10.1016/j.renene.2017.01.036.

Lledó, L., and F. J. Doblas-Reyes, 2020: Predicting daily mean wind speed in Europe weeks ahead from MJO status. *Mon. Wea. Rev.*, **148**, 3413–3426, https://doi.org/10.1175/MWR-D-19-0328.1.

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.

——, 1965: A study of the predictability of a 28-variable atmospheric model. *Tellus*, **17**, 321–333, https://doi.org/10.3402/tellusa.v17i3.9076.

——, 1982: Atmospheric predictability experiments with a large numerical model. *Tellus*, **34**, 505–513, https://doi.org/10.3402/tellusa.v34i6.10836.

Lynch, K. J., D. J. Brayshaw, and A. Charlton-Perez, 2014: Verification of European subseasonal wind speed forecasts. *Mon. Wea. Rev.*, **142**, 2978–2990, https://doi.org/10.1175/MWR-D-13-00341.1.

Machin, D., T. Bryant, D. Altman, and M. Gardner, 2013: *Statistics with Confidence: Confidence Intervals and Statistical Guidelines*. Wiley, 283 pp.

Manzanas, R., J. Gutiérrez, J. Fernández, E. van Meijgaard, S. Calmanti, M. Magariño, A. Cofiño, and S. Herrera, 2018: Dynamical and statistical downscaling of seasonal temperature forecasts in Europe: Added value for user applications. *Climate Serv.*, **9**, 44–56, https://doi.org/10.1016/j.cliser.2017.06.004.

——, ——, J. Bhend, S. Hemri, F. J. Doblas-Reyes, V. Torralba, E. Penabad, and A. Brookshaw, 2019: Bias adjustment and ensemble recalibration methods for seasonal forecasting: A comprehensive intercomparison using the C3S dataset. *Climate Dyn.*, **53**, 1287–1305, https://doi.org/10.1007/s00382-019-04640-4.

Matheson, J. E., and R. L. Winkler, 1976: Scoring rules for continuous probability distributions. *Manage. Sci.*, **22**, 1087–1096, https://doi.org/10.1287/mnsc.22.10.1087.

Molina, M. O., C. Gutiérrez, and E. Sánchez, 2021: Comparison of ERA5 surface wind speed climatologies over Europe with observations from the HadISD dataset. *Int. J. Climatol.*, **41**, 4864–4878, https://doi.org/10.1002/joc.7103.

Monhart, S., C. Spirig, J. Bhend, K. Bogner, C. Schär, and M. A. Liniger, 2018: Skill of subseasonal forecasts in Europe: Effect of bias correction and downscaling using surface observations. *J. Geophys. Res. Atmos.*, **123**, 7999–8016, https://doi.org/10.1029/2017JD027923.

Murcia, J. P., M. J. Koivisto, G. Luzia, B. T. Olsen, A. N. Hahmann, P. E. Sørensen, and M. Als, 2022: Validation of European-scale simulated wind speed and wind generation time series. *Appl. Energy*, **305**, 117794, https://doi.org/10.1016/j.apenergy.2021.117794.

Namias, J., 1952: The annual course of month-to-month persistence in climatic anomalies. *Bull. Amer. Meteor. Soc.*, **33**, 279–285, https://doi.org/10.1175/1520-0477-33.7.279.

Orsolini, Y. J., R. Senan, G. Balsamo, F. J. Doblas-Reyes, F. Vitart, A. Weisheimer, A. Carrasco, and R. E. Benestad, 2013: Impact of snow initialization on subseasonal forecasts. *Climate Dyn.*, **41**, 1969–1982, https://doi.org/10.1007/s00382-013-1782-0.

Palmer, T. N., 2012: Towards the probabilistic Earth-system simulator: A vision for the future of climate and weather prediction. *Quart. J. Roy. Meteor. Soc.*, **138**, 841–861, https://doi.org/10.1002/qj.1923.

——, R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G. J. Shutts, M. Steinheimer, and A. Weisheimer, 2009: Stochastic parametrization and model uncertainty. ECMWF Tech. Memo. 598, 42 pp., https://www.ecmwf.int/node/11577.

Peyré, G., and M. Cuturi, 2019: Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, M. Jordan and R. Tibshirani, Eds., Now Publishers, 272 pp.

Plaut, G., and E. Simonnet, 2001: Large-scale circulation classification, weather regimes, and local climate over France, the Alps and western Europe. *Climate Res.*, **17**, 303–324, https://doi.org/10.3354/cr017303.

Pratt, J. W., 1959: Remarks on zeros and ties in the Wilcoxon Signed Rank procedures. *J. Amer. Stat. Assoc.*, **54**, 655–667, https://doi.org/10.1080/01621459.1959.10501526.

Prodhomme, C., F. Doblas-Reyes, O. Bellprat, and E. Dutra, 2016: Impact of land-surface initialization on subseasonal to seasonal forecasts over Europe. *Climate Dyn.*, **47**, 919–935, https://doi.org/10.1007/s00382-015-2879-4.

Ramon, J., L. Lledó, V. Torralba, A. Soret, and F. J. Doblas-Reyes, 2019: What global reanalysis best represents near-surface winds? *Quart. J. Roy. Meteor. Soc.*, **145**, 3236–3251, https://doi.org/10.1002/qj.3616.

——, ——, P.-A. Bretonnière, M. Samsó, and F. J. Doblas-Reyes, 2021: A perfect prognosis downscaling methodology for seasonal prediction of local-scale wind speeds. *Environ. Res. Lett.*, **16**, 054010, https://doi.org/10.1088/1748-9326/abe491.

Raoult, B., C. Bergeron, A. L. Alós, J.-N. Thépaut, and D. Dee, 2017: Climate service develops user-friendly data store. *ECMWF Newsletter*, No. 151, Reading, United Kingdom, 22–27, https://www.ecmwf.int/en/newsletter/151/meteorology/climate-service-develops-user-friendly-data-store.

Robertson, A., and F. Vitart, 2018: *Subseasonal to Seasonal Prediction: The Gap between Weather and Climate Forecasting*. Elsevier, 588 pp.

Sanders, F., 1963: On subjective probability forecasting. *J. Appl. Meteor. Climatol.*, **2**, 191–201, https://doi.org/10.1175/1520-0450(1963)002<0191:OSPF>2.0.CO;2.

Scaife, A. A., and Coauthors, 2014: Skillful long-range prediction of European and North American winters. *Geophys. Res. Lett.*, **41**, 2514–2519, https://doi.org/10.1002/2014GL059637.

Schepen, A., Q. J. Wang, and D. E. Robertson, 2012: Combining the strengths of statistical and dynamical modeling approaches for forecasting Australian seasonal rainfall. *J. Geophys. Res.*, **117**, D20107, https://doi.org/10.1029/2012JD018011.

——, ——, and ——, 2014: Seasonal forecasts of Australian rainfall through calibration and bridging of coupled GCM outputs. *Mon. Wea. Rev.*, **142**, 1758–1770, https://doi.org/10.1175/MWR-D-13-00248.1.

——, ——, and Y. Everingham, 2016: Calibration, bridging, and merging to improve GCM seasonal temperature forecasts in Australia. *Mon. Wea. Rev.*, **144**, 2421–2441, https://doi.org/10.1175/MWR-D-15-0384.1.

Schwartz, C., and C. I. Garfinkel, 2020: Troposphere-stratosphere coupling in subseasonal-to-seasonal models and its importance for a realistic extratropical response to the Madden-Julian Oscillation. *J. Geophys. Res. Atmos.*, **125**, e2019JD032043, https://doi.org/10.1029/2019JD032043.

Seo, E., and Coauthors, 2019: Impact of soil moisture initialization on boreal summer subseasonal forecasts: Mid-latitude surface air temperature and heat wave events. *Climate Dyn.*, **52**, 1695–1709, https://doi.org/10.1007/s00382-018-4221-4.

Simmons, A., and Coauthors, 2021: Low frequency variability and trends in surface air temperature and humidity from ERA5 and other datasets. ECMWF Tech. Memo. 881, 99 pp., https://www.ecmwf.int/node/19911.

Sobolowski, S., G. Gong, and M. Ting, 2010: Modeled climate state and dynamic responses to anomalous North American snow cover. *J. Climate*, **23**, 785–799, https://doi.org/10.1175/2009JCLI3219.1.

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.

Subramanian, A. C., and Coauthors, 2019: Ocean observations to improve our understanding, modeling, and forecasting of subseasonal-to-seasonal variability. *Front. Mar. Sci.*, **6**, 427, https://doi.org/10.3389/fmars.2019.00427.

Tarek, M., F. P. Brissette, and R. Arsenault, 2020: Evaluation of the ERA5 reanalysis as a potential reference dataset for hydrological modeling over North America. *Hydrol. Earth Syst. Sci.*, **24**, 2527–2544, https://doi.org/10.5194/hess-24-2527-2020.

Tippett, M. K., T. DelSole, S. J. Mason, and A. G. Barnston, 2008: Regression-based methods for finding coupled patterns. *J. Climate*, **21**, 4384–4398, https://doi.org/10.1175/2008JCLI2150.1.

Torralba, V., F. J. Doblas-Reyes, D. MacLeod, I. Christel, and M. Davis, 2017: Seasonal climate prediction: A new source of information for the management of wind energy resources. *J. Appl. Meteor. Climatol.*, **56**, 1231–1247, https://doi.org/10.1175/JAMC-D-16-0204.1.

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

Tripathi, O. P., and Coauthors, 2015: The predictability of the extratropical stratosphere on monthly time-scales and its impact on the skill of tropospheric forecasts. *Quart. J. Roy. Meteor. Soc.*, **141**, 987–1003, https://doi.org/10.1002/qj.2432.

Unger, D. A., 1985: A method to estimate the Continuous Ranked Probability Score. *Ninth Conf. on Probability and Statistics in Atmospheric Sciences*, Virginia Beach, VA, Amer. Meteor. Soc., 206–213.

van den Hurk, B., F. Doblas-Reyes, G. Balsamo, R. D. Koster, S. I. Seneviratne, and H. Camargo, 2012: Soil moisture effects on seasonal temperature and precipitation forecast scores in Europe. *Climate Dyn.*, **38**, 349–362, https://doi.org/10.1007/s00382-010-0956-2.

van der Wiel, K., H. C. Bloomfield, R. W. Lee, L. P. Stoop, R. Blackport, J. A. Screen, and F. M. Selten, 2019: The influence of weather regimes on European renewable energy production and demand. *Environ. Res. Lett.*, **14**, 094010, https://doi.org/10.1088/1748-9326/ab38d3.

Velikou, K., G. Lazoglou, K. Tolika, and C. Anagnostopoulou, 2022: Reliability of the ERA5 in replicating mean and extreme temperatures across Europe. *Water*, **14**, 543, https://doi.org/10.3390/w14040543.

Vitart, F., and Coauthors, 2008: The new VAREPS-monthly forecasting system: A first step towards seamless prediction. *Quart. J. Roy. Meteor. Soc.*, **134**, 1789–1799, https://doi.org/10.1002/qj.322.

——, A. W. Robertson, and D. L. Anderson, 2012: Subseasonal to seasonal prediction project: Bridging the gap between weather and climate. *WMO Bull.*, **61**, 23.

——, and Coauthors, 2017: The Subseasonal to Seasonal (s2s) Prediction Project database. *Bull. Amer. Meteor. Soc.*, **98**, 163–173, https://doi.org/10.1175/BAMS-D-16-0017.1.

——, and Coauthors, 2019: Extended-range prediction. ECMWF Tech. Memo. 854, 60 pp., https://www.ecmwf.int/node/19286.

von Storch, H., F. Zwiers, and C. U. Press, 1999: *Statistical Analysis in Climate Research*. Cambridge University Press, 995 pp.

Wallace, J. M., and D. S. Gutzler, 1981: Teleconnections in the geopotential height field during the Northern Hemisphere winter. *Mon. Wea. Rev.*, **109**, 784–812, https://doi.org/10.1175/1520-0493(1981)109<0784:TITGHF>2.0.CO;2.

Wang, X. L., and F. W. Zwiers, 2001: Using redundancy analysis to improve dynamical seasonal mean 500 hPa geopotential forecasts. *Int. J. Climatol.*, **21**, 637–654, https://doi.org/10.1002/joc.638.

White, C. J., and Coauthors, 2017: Potential applications of Subseasonal-to-Seasonal (S2S) predictions. *Meteor. Appl.*, **24**, 315–325, https://doi.org/10.1002/met.1654.

Wilby, R. L., and T. M. Wigley, 1997: Downscaling general circulation model output: A review of methods and limitations. *Prog. Phys. Geogr.*, **21**, 530–548, https://doi.org/10.1177/030913339702100403.

Wilcoxon, F., 1945: Individual comparisons by ranking methods. *Biom. Bull.*, **1**, 80–83, https://doi.org/10.2307/3001968.

Wilks, D. S., 2013: Projecting "normals" in a nonstationary climate. *J. Appl. Meteor. Climatol.*, **52**, 289–302, https://doi.org/10.1175/JAMC-D-11-0267.1.

——, 2014: Comparison of probabilistic statistical forecast and trend adjustment methods for North American seasonal temperatures. *J. Appl. Meteor. Climatol.*, **53**, 935–949, https://doi.org/10.1175/JAMC-D-13-0294.1.

——, 2019: *Statistical Methods in the Atmospheric Sciences*. 4th ed. Elsevier, 840 pp.

——, and R. E. Livezey, 2013: Performance of alternative "normals" for tracking climate changes, using homogenized and nonhomogenized seasonal U.S. surface temperatures. *J. Appl. Meteor. Climatol.*, **52**, 1677–1687, https://doi.org/10.1175/JAMC-D-13-026.1.

WindEurope, 2022: Wind energy in Europe: 2021 statistics and the outlook for 2022-2026. Tech. Rep., Rue Belliard 40, B-1040, 40 pp., https://windeurope.org/intelligence-platform/product/wind-energy-in-europe-2021-statistics-and-the-outlook-for-2022-2026/.

Woolnough, S. J., F. Vitart, and M. A. Balmaseda, 2007: The role of the ocean in the Madden–Julian oscillation: Implications for MJO prediction. *Quart. J. Roy. Meteor. Soc.*, **133**, 117–128, https://doi.org/10.1002/qj.4.

Žagar, N., and I. Szunyogh, 2020: Comments on "What is the predictability limit of midlatitude weather?" *J. Atmos. Sci.*, **77**, 781–785, https://doi.org/10.1175/JAS-D-19-0166.1.

Zhang, F., Y. Q. Sun, L. Magnusson, R. Buizza, S.-J. Lin, J.-H. Chen, and K. Emanuel, 2019a: What is the predictability limit of midlatitude weather? *J. Atmos. Sci.*, **76**, 1077–1091, https://doi.org/10.1175/JAS-D-18-0269.1.

——, ——, ——, ——, ——, ——, and ——, 2019b: What is the predictability limit of midlatitude weather? *J. Atmos. Sci.*, **76**, 1077–1091, https://doi.org/10.1175/JAS-D-18-0269.1.

Zheng, C., E. K.-M. Chang, H.-M. Kim, M. Zhang, and W. Wang, 2018: Impacts of the Madden–Julian oscillation on storm-track activity, surface air temperature, and precipitation over North America. *J. Climate*, **31**, 6113–6134, https://doi.org/10.1175/JCLI-D-17-0534.1.

Zhu, H., M. C. Wheeler, A. H. Sobel, and D. Hudson, 2014: Seamless precipitation prediction skill in the tropics and extratropics from a global model. *Mon. Wea. Rev.*, **142**, 1556–1569, https://doi.org/10.1175/MWR-D-13-00222.1.

Zorita, E., and H. von Storch, 1999: The analog method as a simple statistical downscaling technique: Comparison with more complicated methods. *J. Climate*, **12**, 2474–2489, https://doi.org/10.1175/1520-0442(1999)012<2474:TAMAAS>2.0.CO;2.