

Verification of European Subseasonal Wind Speed Forecasts

KIERAN J. LYNCH

Department of Meteorology, University of Reading, Reading, United Kingdom

DAVID J. BRAYSHAW

Department of Meteorology, University of Reading, and National Centre for Atmospheric Science, Reading, United Kingdom

ANDREW CHARLTON-PEREZ

Department of Meteorology, University of Reading, Reading, United Kingdom

(Manuscript received 23 October 2013, in final form 5 April 2014)

ABSTRACT

Analysis of the forecasts and hindcasts from the ECMWF 32-day forecast model reveals that there is statistically significant skill in predicting weekly mean wind speeds over areas of Europe at lead times of at least 14–20 days. Previous research on wind speed predictability has focused on the short- to medium-range time scales, typically finding that forecasts lose all skill by the later part of the medium-range forecast. To the authors' knowledge, this research is the first to look beyond the medium-range time scale by taking weekly mean wind speeds, instead of averages at hourly or daily resolution, for the ECMWF monthly forecasting system. It is shown that the operational forecasts have high levels of correlation (~ 0.6) between the forecasts and observations over the winters of 2008–12 for some areas of Europe. Hindcasts covering 20 winters show a more modest level of correlation but are still skillful. Additional analysis examines the probabilistic skill for the United Kingdom with the application of wind power forecasting in mind. It is also shown that there is forecast “value” for end users (operating in a simple cost/loss ratio decision-making framework). End users that are sensitive to winter wind speed variability over the United Kingdom, Germany, and some other areas of Europe should therefore consider forecasts beyond the medium-range time scale as it is clear there is useful information contained within the forecast.

1. Introduction

Traditionally, studies of wind speed predictability have focused on the short- to medium-range time scales, as seasonal forecast systems (Arribas et al. 2011) have shown little skill in predicting large-scale features such as the North Atlantic Oscillation (NAO). A recent review of wind speed and power forecasting techniques by Foley et al. (2012) discussed many statistical and dynamical techniques with a focus on nowcasting to the medium-range time scale. While they acknowledged the

importance of statistical models in estimating the monthly/seasonal mean wind speed, there was no mention of forecast information on these time scales from dynamical models. To our knowledge, no published peer-reviewed literature currently exists demonstrating that there is wind speed predictive skill from dynamical forecasts beyond the medium-range time scale.

Prediction of hour-to-hour variations in wind at long lead times is difficult. Pinson and Hagedorn (2012) show that hourly forecast data from the European Centre for Medium-Range Weather Forecasts (ECMWF) medium-range forecast loses almost all skill after 5–6 days. There is, however, more opportunity for prediction over longer time-averaging windows. For example, Rodwell and Doblas-Reyes (2006) explain that by taking a weekly time average over the meteorological variable of interest, the unpredictable short-term fluctuations are reduced and predictive skill arises from slow changes in the boundary forcing. Weekly averaging for monthly forecasts of key

 Denotes Open Access content.

Corresponding author address: Kieran J. Lynch, Department of Meteorology, University of Reading, Earley Gate, P.O. Box 243, Reading RG6 6BB, United Kingdom.
E-mail: k.j.lynch@pgr.reading.ac.uk

DOI: 10.1175/MWR-D-13-00341.1

meteorological variables such as temperature and geopotential height at 500 hPa have been shown to produce skillful forecasts (Vitart 2004; Vitart et al. 2008; Weigel et al. 2008; Hudson et al. 2011), suggesting that the monthly forecast system is able to capture some predictability of the large-scale flow. Here we extend the evaluation of the monthly forecasting system to surface wind speeds.

Dynamical monthly forecasting, otherwise known as subseasonal prediction, is relatively new and lies at the traditional “gap” between weather and climate. Common forecasting options at this lead time have therefore been “persistence” or “climatology.” A persistence forecast uses the most recent weather observations up to the point of the forecast being issued to project forward in time (assuming autocorrelation exists in the weather variable of interest). A climatological forecast uses only prior information from a long-term set of observations (often 30 years) for the forecast period of interest. More advanced statistical or dynamical models can therefore be benchmarked in terms of performance against the climatology or persistence methods. The monthly time scale is influenced by both the initial atmospheric conditions and boundary forcings from slower moving components. It is known that sources of predictability come from the stratosphere (Baldwin and Dunkerton 2001), the Madden–Julian oscillation (MJO; Vitart and Jung 2010), soil moisture anomalies (Koster et al. 2010), and sea surface temperatures (SSTs; Hu et al. 2011). Vitart (2004) shows that the ECMWF monthly forecast for days 12–18 can outperform persistence of the previous week (days 5–11) from the same forecast. Since then, improvements in the forecast system have enhanced forecast skill (Vitart 2014).

Wind speed predictability has a range of economic and societal impacts. Forecasts better than climatology or persistence can add value to the operations of many end users. An end user is defined as someone who is exposed to weather-related risk and is therefore faced with making decisions based on a forecast and the potential impacts. Examples of monthly forecast end users include the energy sector, supermarket chains (as the sales volume of certain products is dependent upon temperature), agricultural industries, and financial institutions (trading commodities that are impacted by weather). For instance, in the energy sector, an increasing amount of wind power is being built (McGinn et al. 2013). Therefore, the need for accurate forecasts becomes increasingly important because of the intermittent nature of wind power generation and the need to match supply with demand at all times (Füss et al. 2013). While previous studies have examined the forecasting of short-term fluctuations in wind speed/power for

electricity grid management and market participation (Pinson 2013), there has been little investigation on the use of information contained in forecasts at the weekly resolution and the implications for risk management in the power system.

This research examines the skill in the ECMWF monthly forecast system for wind speed at the weekly time scale over Europe, with the aim of giving forecasters an overview of the current limits of predictability. One of the reasons a weekly average was chosen is that it matches the length of commonly traded future contracts for commodities such as power and gas. The European region was chosen because of a significant penetration of wind energy and therefore a strong interest in longer-term wind speed predictability, particularly for countries such as the United Kingdom, Germany, Denmark, and Spain (McGinn et al. 2013). The special application of this to quantitative risk management applications in the power sector is a subject of ongoing research and will be presented in subsequent papers.

To quantify the predictability of monthly wind speed forecasts, the ECMWF monthly forecast system was chosen because of its known skill in the extratropics (Vitart et al. 2008). We focused on the winter months of December, January, and February, as the winter season was found to have the strongest predictability. In addition to this, larger wind speed variability during winter (Sinden 2007) could make forecasts more useful for some end users. Section 2 briefly outlines the data sources and methodology used. Section 3 examines the seasonality of predictive skill in the monthly forecast system and the differences between the forecast and hindcasts in terms of the correlation between the ensemble mean and the observations. Section 4 examines the probabilistic predictability of the operational forecasts using verification scores and a number of common diagnostics. In section 5, the value of the forecasts to the end user is demonstrated in terms of cost/loss ratios. The findings are then discussed with potential applications in section 6, and conclusions are presented in section 7.

2. Data and methodology

a. Data

The wind speed forecasts used in this analysis were from the ECMWF monthly forecast model. The model is an extension of the medium-range forecasting system, and the configuration analyzed runs by coupling to an ocean model after day 10 (with persisted SST anomalies prior to that). For a full description of the model, see Vitart et al. (2008). Since March 2008, operational 51-member ensemble monthly forecasts have been issued

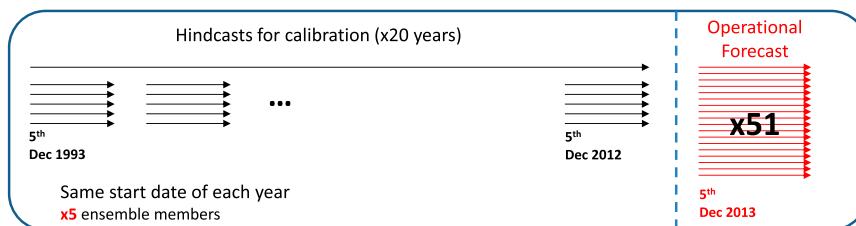


FIG. 1. Schematic explaining the operational method currently used at the ECMWF for producing the monthly forecasts. Prior to the operational forecast being generated, 20 hindcast simulations are run over the previous 20 years from the same start date of each year. Each hindcast run consists of five ensemble members. The total set of hindcasts is then used to generate the model climatology and to calibrate the operational 51-member ensemble forecast.

every Thursday (and since October 2011, every Monday and Thursday). Prior to each Thursday's forecast release, a set of calibration hindcasts is run using the same version of the model from the same day of the year (e.g., 5 December) over the previous 20 years with only five members each time (see Fig. 1).¹ There are therefore two sets of data that we can use to evaluate the forecast performance: the 51 member operational forecasts (2008–13) and the five-member hindcasts (1992–2012, 20 years). As the operational forecast system began to produce forecasts twice a week (Mondays as well as the original Thursday forecasts) from October 2011, all Monday forecasts have been removed from the quantitative analysis in order to maintain statistical independence of the results. We also verified that using the Monday forecasts, instead of the Thursday forecasts, from October 2011 did not significantly affect the results.

The hindcasts and forecasts have advantages and disadvantages. Analysis of the hindcasts, as retrospective forecasts, allows a large period of climatic variability (20 years) to be sampled from but is limited by only having five ensemble members per hindcast, which makes reliable probabilistic predictions difficult. On the other hand, the operational forecasts have 51 ensemble members, allowing a better estimation of the uncertain future atmospheric state but covering a relatively short period of climatic variability (winters of 2008–13). Any conclusions drawn on this data alone suffer from the fact that the winters of 2008–13 may have been more (or less) predictable than usual. This research therefore examines 10-m wind speed predictability using both the forecasts and hindcasts.

To verify the forecasts, the Interim ECMWF Reanalysis (ERA-Interim) dataset (Dee et al. 2011) was

used as the observational “truth” as it covers the full period of the forecasts and hindcasts. Although the reanalysis data are subject to error, Decker et al. (2012) show that ERA-Interim has low RMSE; therefore, once weekly averaging is applied to wind speeds, the random error contribution should be minimized. We also confirmed prior to this research that the error between reanalysis wind speeds and meteorological mast data is small for offshore sites. For onshore sites, it was shown that by averaging over a larger number of meteorological mast stations with anemometers at 10 m covering an area of 300–400 km results in a large reduction in the random error term. This validation of U.K. wind speeds and wind power in reanalysis data is discussed in more detail in Cannon et al. (2014, manuscript submitted to *Renewable Energy*). From these results, we concluded that the error in the reanalysis (when compared to the “true” observations) is small relative to the error of the monthly forecast model at longer lead times.

b. Methodology

A bias correction calibration was applied to the 10-m wind speed data from the ECMWF monthly forecast to correct for the model drift. This model drift occurs because of the model having its own natural climatology that it tends to over time, leading to biases relative to the observed climate. The forecasts/hindcasts and reanalysis data were processed on a grid point by grid point basis, over a regular N128 Gaussian grid covering the North Atlantic and most of Europe. The forecasts were analyzed at a time step resolution of 6 h (0000, 0600, 1200, and 1800 UTC) running out to 32 days. As one of the potential applications at these time scales is producing wind power forecasts, ideally 100-m wind speeds should be the height of choice to match the typical hub height of wind turbines. However, the ECMWF has not been archiving 100-m wind speeds for as long as the 10-m wind speeds. As a larger number of operational forecasts in the analysis is important, 10-m winds were chosen. Moreover, there is little difference in the verification scores between

¹ On a technical note, Monday operational forecasts are calibrated using a weighted combination of the two Thursday hindcasts on either side.

the 10 and 100 m winds (approximately 0.02 difference in terms of correlation when forecasting 14 days ahead). All verification scores were applied to the forecast anomalies (i.e., measuring the ability of the model to forecast departures from the seasonal and diurnal cycle) rather than the absolute wind speed values, as there is clear seasonality and diurnality in wind speeds. The precise details of how to calculate the model and observational climatologies and convert to wind speed anomalies is listed below.

1) ERA-INTERIM (OBSERVATIONS)

To estimate the seasonal cycle, a wind speed climatology for every sixth hour of the day, for each day in the year, is estimated by averaging over the whole ERA-Interim record spanning 1979–2013. Any particular hour of the year, such as 1800 UTC on 5 December, consists of only 34 samples (one for each year 1979–2012), and therefore, some noise remains because of sampling uncertainty. As an additional noise reduction step, a moving average was applied to the climatology to isolate the low-frequency seasonal variability. The 31-day moving average was applied separately for each of the hour of the day (0000, 0600, 1200, and 1800 UTC) to avoid artificial skill arising from simulating the diurnal cycle. A time series of wind speed anomalies is then calculated by removing the climatology from the original ERA-Interim wind speeds, giving the observational truth.

2) OPERATIONAL FORECASTS

The model climatology is estimated as a function of lead time because of the model drift in the monthly forecast model at longer lead times. Removing the model climatology from the forecasts therefore minimizes the model bias. The operational forecast ensemble members are converted to anomalies by removing this mean model climatology calculated at every time step from the 100 hindcast runs. These forecast anomalies are then compared with the “observed” anomalies from ERA-Interim. Temporal smoothing of the model climate is not applied in this case because of there being a larger number of realizations (100) of the model “climate” relative to 33–34 realizations in the ERA-Interim record. This method is currently the same that is used at the ECMWF (F. Vitart 2013, personal communication) to calculate the operational forecast anomalies.

Wind speed terciles refer to the lower or upper one-third of the climatological distribution. To calculate the tercile thresholds, the model climate (again estimated from the respective hindcasts) was used. There was no significant difference in terms of skill scores if either the model climate or observed climate (ERA-Interim) was used as the threshold to estimate the terciles forecast probabilities (as discussed in [section 4](#)).

3) HINDCASTS

Although the hindcasts are used to calibrate the operational forecasts (see above), they can also be used as “retrospective forecasts.” These retrospective forecasts are generated from the hindcasts, even though they are only used in an operational setting to calibrate the real-time forecasts. For a set of hindcasts run from the same start day of each year over the past 20 years, 19 years of the hindcast data are used to calculate the model climatology (same method as above). To calculate the forecast anomalies, the model climatology (estimated from the 19 years) is removed from the remaining year, which is considered as the retrospective forecast. This process is repeated over each of the 20 years, with the forecast year always being removed from the model climatology calculation. The 20 years of hindcast anomalies can then be compared with the ERA-Interim anomalies.

3. Ensemble mean forecast skill

This section focuses on the ensemble mean forecast, evaluating the forecasts in terms of the correlation coefficient of the anomalies (ACC) for both the hindcasts and the operational forecasts on a grid point by grid point basis. The resulting maps therefore give an indication of the variation in prediction skill across Europe. As the correlation is invariant to a change in the mean or rescaling, it can be seen as the potential model skill obtainable with good calibration.

a. Prediction skill by season

[Figure 2](#) shows the ACC in the 51-member operational forecasts for each season of the year. There is a clear difference in the level of prediction skill between the different seasons when forecasting with a lead time of 14 days (i.e., wind speeds averaged over days 14–20). The strongest correlation that is statistically significant occurs in the months of December–February (DJF) across areas of the North Atlantic, United Kingdom, and northern Europe. The months of June–August (JJA) and September–November (SON; [Figs. 2c,d](#)) have very limited correlation over the United Kingdom and Europe. The months of March–May (MAM) have a high correlation in some areas using the operational forecast ([Fig. 2b](#)), but, if the analysis is repeated using the hindcasts, the correlation is much reduced and comparable to SON (not shown).

There are potentially a number of reasons that may explain the larger correlation during the winter months: these include larger SST gradients, stronger coupling between the stratosphere and troposphere, and influence from the MJO. It is beyond the scope of this study

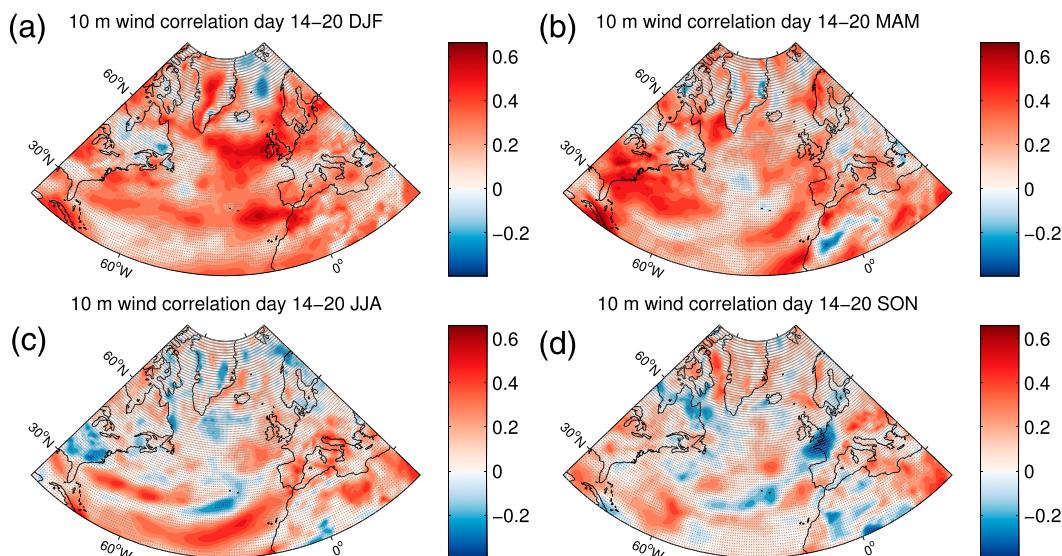


FIG. 2. Seasonal variation in the ACC between 10-m wind speed ensemble mean operational forecasts and ERA-Interim. A lead time of 14 days is taken (and the subsequent week is averaged over day 14–20 before calculating the ACC). The four plots show the differences in predictability between the seasons. (a) DJF shows the strongest signal over the United Kingdom and some of northern Europe. (b) MAM, (c) JJA, and (d) SON show little to no predictability over most of Europe. The stippling (gray dots) mask the regions that are not statistically significant at the 95% level.

to address these in detail, but the results are consistent with Vitart (2004), which also shows Northern Hemisphere predictability is the strongest in winter for 2-m temperature. As it is clear that the strongest predictability exists in DJF, all subsequent analysis focuses on the winter season DJF.

b. Differences between forecasts and hindcasts

Figures 3a,c clearly show the operational forecasts and hindcasts have positive skill over large areas of the North Atlantic and Europe at lead times of 14 days. During the five winters of operational forecasts (Fig. 3a), correlations of approximately 0.6 are found over the United Kingdom between the forecast and reanalysis at a lead time of 14 days. For the hindcasts (Fig. 3c), the correlation is small but still statistically significant. For the longer lead time of 22 days, the operational forecasts have limited to no correlation. The majority of Europe (Fig. 3b) is covered in gray stippling (dots), indicating that the correlation is not statistically significant for the ECMWF forecasting system at lead times of 22 days. However, given that the patterns of correlation are similar to those at shorter lead times and that there is some small positive correlation over Europe, a larger number of operational forecasts may reveal statistically significant skill. It therefore appears that there is skill over many areas of Europe at lead times of 14 days, with suggestions of potential skill at 22 days but not enough evidence to currently be confident.

There are two potential reasons why the forecasts have higher levels of correlation than the hindcasts. The first is that a greater number of ensemble members in the forecasts leads to a better estimation of the future atmospheric state. The second is that the five winters covered by the operational forecasts were for some reason more predictable than usual. To confirm that the hindcast skill is not just coming from the years overlapping the operational forecasts (2008–12), the hindcast correlation scores have also been calculated only for the years prior to the operational forecasts (i.e., before December 2008) and shown in Fig. 3d. From this it is clear that significant correlation scores also exist prior to December 2008 in the hindcasts. It is slightly weaker over the United Kingdom and northern Europe in Fig. 3d than in the full hindcast period (Fig. 3c), but it confirms that the correlation is not just from predictability over the shorter period covering that of the recent operational forecasts (December 2008 onward).

In summary, the ensemble mean weekly average wind forecast has statistically significant levels of correlation at lead times of 14 days for many parts of Europe. The patterns of correlation are similar in both the forecasts and the hindcasts. The hindcasts do have much lower levels of correlation than the forecasts, but this correlation exists across the full 20 years and not just the five winters that the operational forecasts cover (2008–12).

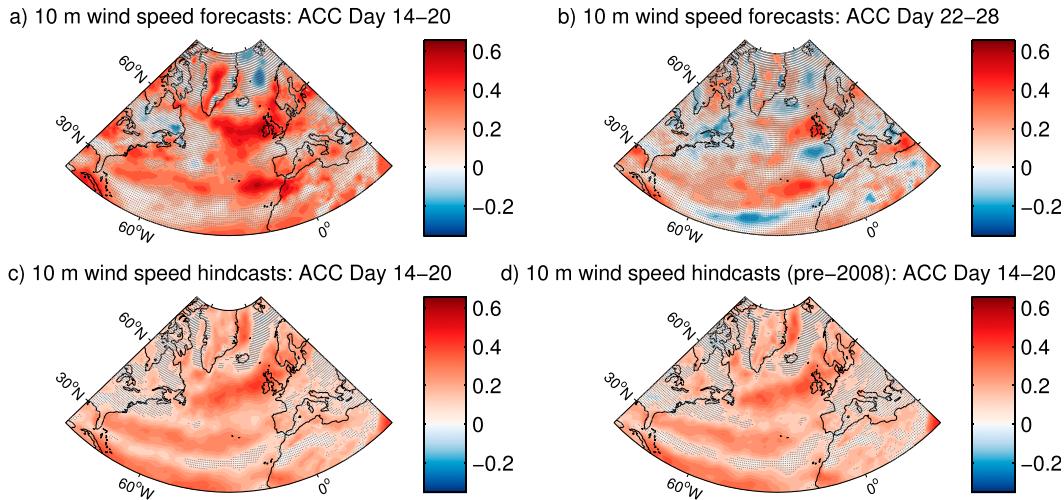


FIG. 3. ECMWF monthly (ensemble mean) forecast for DJF 10-m wind speed ACC between weekly mean forecast (or hindcast) and observations for operational forecasts from December 2008 to January 2013 at a lead time of (a) 14 days and (b) 22 days; (c) hindcasts (most recent model version) from December 1992 to February 2012 at a lead time of 14 days, and (d) as in (c), but only the years prior to the operational forecast (i.e., pre-2008). The stippling (gray dots) mask the regions that are not statistically significant at the 95% level.

4. Probabilistic forecast skill

So far, discussion of the forecasts and hindcasts has focused on the correlation coefficient, which is a deterministic score that only considers the ensemble mean. However, there is often more value in the forecast if the full predictive distribution is analyzed and taken into account when an end user is faced with making a decision (Gneiting 2011). This section uses the operational probabilistic forecast to assess the past performance of the system in terms of the continuous ranked probability score (CRPS) and the characteristics of the forecast resolution, reliability, and spread–skill relationship. Each of the forecast verification metrics/diagnostics is introduced in turn with the results stated alongside. All the following results in this section are for predictability over the United Kingdom (i.e., the weekly 10-m wind speed averaged over a box spanning 49°–60°N and 10°W–4°E). By taking a United Kingdom–wide average, the noise is reduced, enhancing the predictability. A United Kingdom–wide wind speed forecast would be useful for U.K. total wind power output.

a. CRPS skill score

The CRPS is a commonly used probabilistic score that evaluates the predictive skill of the full probability distribution. The CRPS becomes the mean absolute error if a deterministic forecast is used (and may therefore be thought of as the probabilistic version of the mean absolute error). The CRPS (Jolliffe and Stephenson 2011) is defined as the integral of the squared difference between

the cumulative density forecast $[P(x)]$ and the observation (x_a):

$$\text{CRPS}(P, x_a) = \int_{-\infty}^{+\infty} [P(x) - H(x - x_a)]^2 dx, \quad (1)$$

where the Heaviside function H denotes the cumulative density function for the observation x_a :

$$H(x - x_a) = \begin{cases} 0 & (x - x_a) < 0 \\ 1 & (x - x_a) \geq 0 \end{cases}. \quad (2)$$

The CRPS can be converted into a skill score, measuring the performance of a forecast relative to some benchmark forecast (i.e., climatology, persistence, or another forecast system). In this case, we define the skill score $\text{CRPS}_{\text{skill}}$ by normalizing it relative to a climatological forecast, $\text{CRPS}_{\text{clim}}$:

$$\text{CRPS}_{\text{skill}} = 1 - \frac{\text{CRPS}}{\text{CRPS}_{\text{clim}}}. \quad (3)$$

Skill scores below 0 are therefore defined as unskillful, those equal to 0 are equal to the climatology forecast, and anything above 0 is an improvement upon climatology, up to 1, which indicates a “perfect” forecast. Figure 4 shows the $\text{CRPS}_{\text{skill}}$ for the weekly mean 51-member operational forecast. It can be seen that there is positive skill at all lead times over the United Kingdom. There is high predictability (of weekly average wind speeds) for the short lead times, as one would expect. This forecast

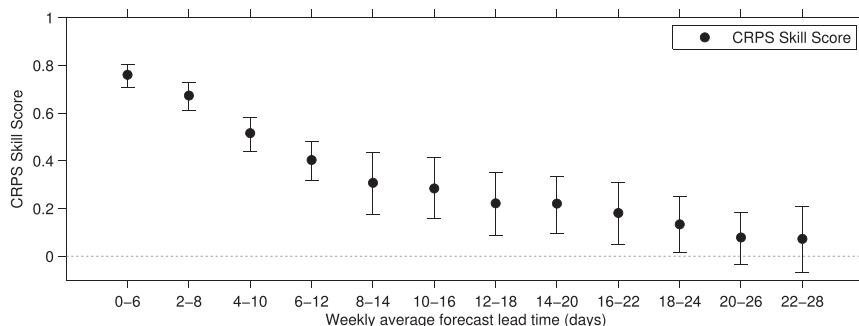


FIG. 4. Operational ECMWF monthly forecast for 10-m wind speed averaged across the United Kingdom (for weekly averages). CRPS skill score as a function of operational forecast lead time (error bars show the 95% confidence intervals). A skill score of 1 indicates a perfect forecast; 0 is equal to climatology, and therefore anything better than 0 is an improvement on climatology.

skill drops off quickly but remains positive for lead times up to 18 days (i.e., forecasting for a lead time of 18 days and taking an average wind speed over days 18–24). The forecast skill is statistically significant at the 95% level.² At longer lead times, the forecast skill is no longer statistically significant.

b. ROC, reliability, and the spread–skill relationship

The quality of a probabilistic forecast system can be analyzed in terms of a range of attributes such as resolution, reliability, and the spread–skill relationship. Each of these are considered now.

The empirical relative operating characteristic (ROC) curve is commonly used to assess the forecast skill for a binary forecast (i.e., 0/1 or no/yes) and indicates forecast resolution. In this case, the binary forecast variable is defined as the occurrence of the wind in the lower (or upper) tercile. The hit rate and false alarm rate can be calculated for the probabilistic forecast by applying a probability threshold γ to the predictive probability distribution allowing classification of a forecast into a binary event (if the threshold probability of occurrence γ is exceeded) or nonevent (if γ is not exceeded). For a set probability threshold, the forecasts can be evaluated in terms of hit rate and false alarm rate using the contingency matrix (Table 1), where the hit rate H and false alarm rate F are calculated respectively as

$$H(\gamma) = \frac{a}{a + c} \quad \text{and} \quad (4)$$

$$F(\gamma) = \frac{b}{b + d}. \quad (5)$$

By changing the threshold probability γ of the event occurring from 0 to 1 over all values, the respective hit rate can be plotted against the false alarm rate on the ROC diagram with the line going from (1, 1) (when γ is equal 0) to (0, 0) (when γ is equal 1). A convex curve, above the diagonal one-to-one line, occurs when there is forecast skill. This is this case in Fig. 5a, where both the lower and upper terciles have positive forecast skill. The area under the upper tercile curve is larger than the lower tercile area, indicating better forecast performance for the upper tercile. The ROC curves in Fig. 5a indicate good forecast resolution for both terciles, that is, the ability of the forecast system to discriminate between the two types of events: occurrence and non-occurrence.

A reliable probabilistic forecast is also a desirable attribute: to illustrate, if a forecast issues a 70% chance of rain, in terms of long-run statistics, on average it should rain 70% of the time (for all forecasts that issue a 70% chance of rain). In this case, the reliability of forecasting the weekly average wind speeds occurring in the lower or upper tercile is examined.³ For a perfectly reliable forecast, the points should lie along the one-to-one diagonal line in Fig. 5b. The upper tercile points fall close to the diagonal line, indicating good reliability; for

²Confidence intervals are estimated using a nonparametric bootstrap method. Correlation is calculated for 10 000 resamples with replacement. Because of serial correlation between weekly wind speeds, a block-based sampling method was used with a block length of 6 weeks. It should be pointed out that if the five winters covered by the operational forecasts are not representative of the long-term climatic predictability, then the confidence intervals may also not be as representative.

³The reliability diagram has been conditioned on the forecast frequency rather than the observed frequency. The rationale for this is that an end user wants to know, given that the forecast suggests the probability of an event occurring is x , has this event occurred with the same frequency x over all past forecasts?

TABLE 1. The contingency table for the four different possibilities (a , b , c , and d) for some event, that is, wind speed in the upper tercile.

		Event occurs	
		Yes	No
Event forecast	Yes	a	b
	No	c	d

instance, when a 50% chance of the wind being in the upper tercile is forecast, the observed event occurred close to 50% of the time. The lower tercile has a systematic low bias, resulting in a higher frequency of observed events relative to those forecast.

The spread–error relationship can be used to assess whether there is useful information contained in the ensemble spread. Forecasts with a wider spread, should, on average, be less accurate. The forecasts are binned by ensemble spread (i.e., standard deviation) and the error for all forecast/observation pairs calculated (Fig. 5). Because of there only being 59 forecast/observation pairs, the data were stratified into three bins (of approximate equal size) based on the standard deviation of the ensemble. For each bin, the RMSE and CRPS is calculated. Figure 5c shows a positive spread–skill relationship, not just for the deterministic RMSE measure, but also for the probabilistic CRPS score. A forecast issued with a smaller spread therefore gives the end user more confidence in the forecast.

c. Discussion of forecast skill

The week 3 wind speed forecast, for a grid box over the United Kingdom, results in a CRPS score greater than climatology. The CRPS score is a harder metric to outperform climatology on than simple scores such as the Brier score for terciles. It should be noted that the wind speed climatology used as a benchmark only has a varying time-mean component (seasonal cycle) and assumes constant variance for the 3 months of winter (DJF). There is clearly seasonality in the variance of the wind speeds across the United Kingdom, larger in the winter and smaller in summer. However, the reduction in the CRPS skill score from including a time-varying variance in the climatology will be minimal. This is because the difference in wind speed variance during the months of DJF is found to be relatively small.

Reliability is a requirement for predictability, but resolution is also needed to be able to correctly differentiate between events and nonevents. Reliability can also be corrected for by a posterior calibration, whereas the resolution cannot be improved (Jolliffe and Stephenson 2011) by calibration. Provided a large enough set of previous forecasts exists, it is possible to correct the reliability.

TABLE 2. The cost/loss contingency table for the four different possibilities for some event, that is, wind speed in the upper tercile.

		Event occurs	
		Yes	No
Action taken	Yes	C	C
	No	L	0

A reliability correction could be applied to future forecasts, based on the previous operational forecasts performance. However, the ECMWF monthly forecast model is updated every year (or sometimes more often), and it is unknown whether any reliability correction would still hold under a new version of the model with slightly different model physics.

The spread–skill relationship (Fig. 5c) shows that there is useful information contained in the variance of the forecast pdf. There is therefore more value in the forecast if the end user considers not just the ensemble mean but also the ensemble variance. Although the three bins, plotted in Fig. 5c have an approximately linear relationship, it should be noted that the error of any individual forecast may be substantially different from the average error expected from a forecast that has been binned by its spread.

The spread–skill relationship was also analyzed for the hindcasts, as there is a larger set of forecast–observation pairs. However, despite there being over 200 independent forecasts, there was no spread–skill relationship. This suggests that only having five ensemble members does not allow effective sampling of the flow-dependent atmospheric uncertainty. This suggests that the hindcasts are only able to provide predictability of the mean atmospheric state, whereas the operational forecasts are also able to provide information on the uncertainty that is conditional on the forecast probability distribution.

5. Forecast value

A range of end-user applications require the conversion of a probabilistic forecast into a yes/no decision for protection against an event; an example given in Richardson (2000) is the gritting of roads to protect against ice. The rationale is that an end user has some risk, such that if an event occurs they incur some loss L . They may choose to take preventative action to avoid the loss at a cost C . They are therefore faced with the contingency matrix outlined in Table 2. Given a probabilistic forecast of the event (say, lower tercile winds), the user therefore wishes to determine an optimal probability threshold γ above which they will pay the cost C to avoid the loss L . Below the threshold, they take no action. The threshold γ can therefore be chosen to

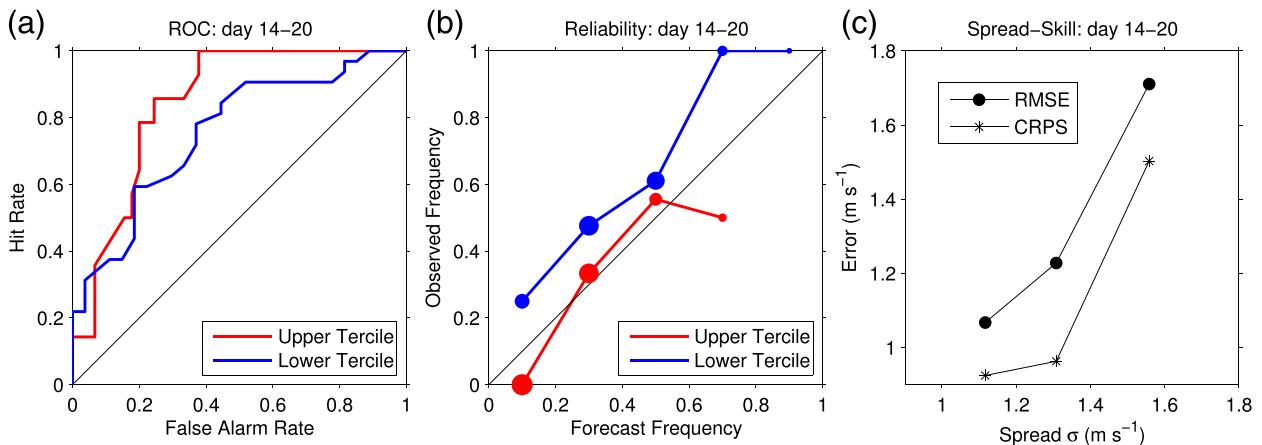


FIG. 5. Operational ECMWF monthly forecast for 10-m wind speed averaged across the United Kingdom (for weekly averages). (a) ROC for upper and lower terciles (days 14–20); convex curves above the diagonal indicate forecast skill for both terciles. (b) Reliability diagram for upper and lower terciles (days 14–20), where the size of circles is proportional to the number of forecast cases in each bin. For perfect reliability, the points should fall on the diagonal line. The upper tercile forecast is reliable; however, the lower tercile probabilities are systematically too low. (c) Spread–skill relationship for the ECMWF operational forecasting system (during DJF) for the United Kingdom. The 59 individual forecasts issued every Thursday during the winter (from December 2008 to January 2013) have been binned by their ensemble spread (with sizes of 20, 20, and 19 for the lower, middle, and upper bins, respectively). The binned ensemble spread is plotted against the RMSE for the respective forecasts in each bin. The plot shows a positive spread–skill relationship, which thus gives the forecaster more confidence in the accuracy of the ensemble mean if the forecasted ensemble spread is smaller.

maximize the economic value of the forecast for a specific cost/loss ratio ($\alpha = C/L$). To calculate the forecast value, Richardson (2000) shows that the hit rate H [Eq. (4)], false alarm rate F [Eq. (5)], climatological frequency s of the event, and cost/loss ratio α are needed:

$$V = \frac{\min(\alpha, s) - F(1-s)\alpha + Hs(1-\alpha) - s}{\min(\alpha, s) - s\alpha}. \quad (6)$$

Negative forecast value indicates that the end user is better using the climatological forecast, and positive forecast value offers more financial gain than climatology and is bounded at one. For a specific probability threshold [$\Pr(U)$], the forecast value can be calculated using Eq. (6) for all cost/loss ratios α . An example is given by the solid black line in Fig. 6a, where the value is shown using a forecast threshold [$\Pr(U)$] of 0.3 and therefore insuring against an event when any forecast exceeds the threshold. The optimal probability threshold γ is shown in Figs. 6c and 6d by the colored dots. For a specific cost/loss ratio, the color of the dot corresponds to the optimal (long run) probability threshold that the end user should use to convert the probability forecast into a yes/no decision.

As the cost of protecting against the event increases (high cost/loss ratio), a higher threshold γ is required as more certainty is needed that the event will happen. The upper tercile has more value to the end user than the lower tercile for cost/loss ratios below approximately 0.6. The highest value obtainable from the lower tercile

is for a user with a cost/loss ratio equal to the climatological frequency ($1/3$) of lower tercile occurring. This is the same for the upper tercile; however, this maximum value achievable from the forecast also occurs for cost/loss ratios below the climatological frequency. In general, there is a large range over which the forecast has value.

Overall, we have shown that there is forecast value for binary decision making relating to the upper and lower terciles across all cost/loss ratios. This is beneficial, as typically the forecast value lies in a narrower range of cost/loss ratios (see Vitart 2004; Vitart et al. 2008). Forecast value is likely to exist for other thresholds over a range of cost/loss ratios. Further analysis (not shown), revealed that in the hindcasts, for a range of different thresholds, the area under the ROC curve was greater than 0.5, which indicates an improvement on climatology and positive forecast value (for at least some cost/loss ratios), given the link between the ROC curve and forecast value in Eq. (6). Moreover, the forecast value is present for all cost/loss ratios from zero and one, as is shown in Figs. 6c and 6d.

6. Discussion

We have shown that there is statistically significant probabilistic forecast skill over the United Kingdom for weekly averaged 10-m wind speeds during DJF, at lead times of up to 3 weeks and potentially beyond. This was

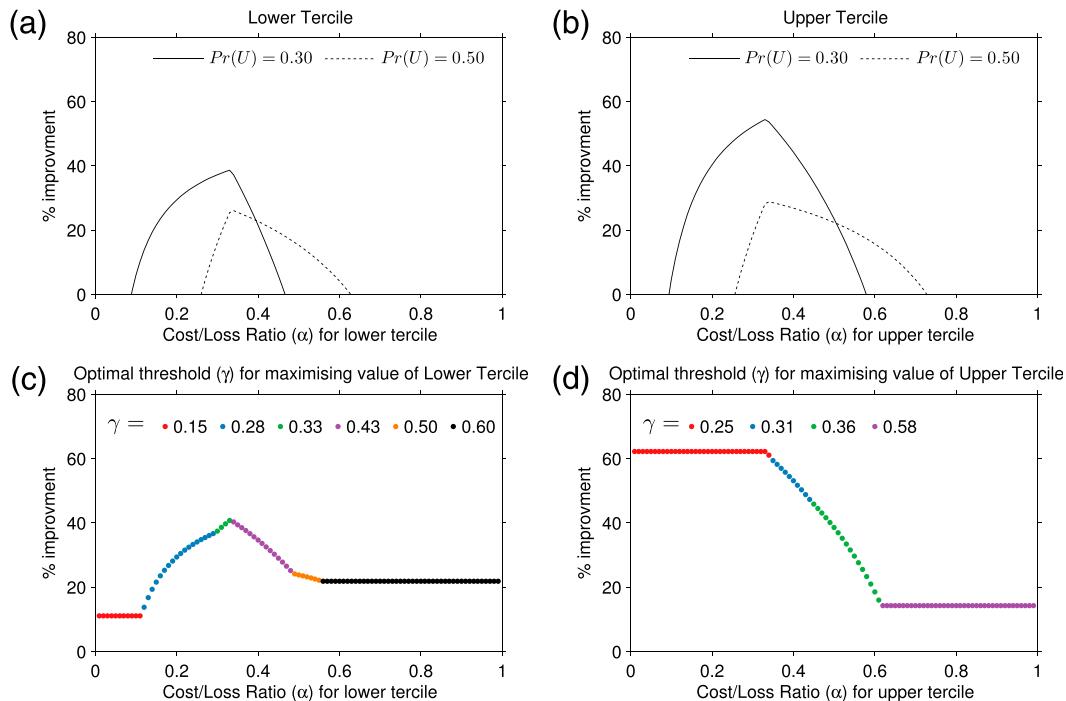


FIG. 6. End user value for using the ECMWF monthly forecast (during DJF) instead of climatology for the United Kingdom when exposed to some risk associated with the wind occurring in the (left) lower tercile and (right) upper tercile. The value obtained is expressed in terms of percentage improvement upon climatology, where 0% is equal to climatology, anything greater than 0% is an improvement, and 100% is obtained from using a perfect forecast. (a),(b) Each black line shows the value obtained by choosing to act (i.e., pay the cost) when a specific forecast probability threshold P_u is exceeded (for the wind being in the lower/upper tercile). The solid black line corresponds to a probability threshold P_u of 0.30, and the dotted black line corresponds to a P_u of 0.50. (c),(d) The colored dots show the optimal P_u an end user should choose (given their cost/loss ratio α) in order to maximize the value of the forecast.

achieved by taking weekly averaged wind speeds instead of hourly/daily wind speeds used in studies such as Pinson and Hagedorn (2012). There is therefore the possibility that monthly forecasts may provide useful wind speed forecasts, helping extend the limit of predictability beyond the medium-range time scale.

As this is the first study to quantify wind speed predictability in the monthly forecast (to the best of our knowledge), we can only compare with previous literature that has focused on 2-m temperature. As wind speed is generally harder to predict than temperature [see CRPS skill differences between Pinson and Hagedorn (2012) and Hagedorn et al. (2012)], it would be expected that the wind speed skill scores are lower than that in previous verification studies for temperature. The model skill shown in Vitart (2004) is for an old version of the ECMWF monthly forecast that was only run 45 times each with 51 ensemble members. The ROC score for 2-m temperature (averaged across all land points in the extratropics) at similar lead times was less than that found for the 10-m wind speeds averaged across the United Kingdom in Fig. 5. There was also no 2-m temperature

skill for the area over the United Kingdom. As this was an old version of the model, subsequent improvements have led to an increase in the skill (Bechtold et al. 2008). Weigel et al. (2008) used a more recent version of the ECMWF monthly forecast model and show that there is some small skill over the United Kingdom for spring temperatures. The ROC diagrams shown in Vitart (2004), Vitart et al. (2008), and Vitart and Molteni (2010) for 2-m temperature in the extratropics and/or Europe at lead times (similar to week 3) have less area under the curve than the wind speed ROC curves shown in Fig. 5a. Many factors such as the different geographical areas, lead times, number of ensemble members, and model version make a direct comparison between the literature and the ROC curves (Fig. 5a) in this paper impossible. However, the key point is that the 10-m wind speed ROC curve area is larger than anything in the literature for similar lead times when forecasting 2-m temperature.

One possible explanation for this high level of predictability over the last five winters (2008–12) is from sampling a period of enhanced predictability. It is possible that the last five winters may not have been representative

of the long-term climatic variability in flow-dependent predictability. This issue has been addressed in [section 3](#), where the both the hindcasts and the operational forecasts have been compared side by side. This analysis confirmed that the week 3 predictive skill was robust and not simply an artifact of an unusually predictable period. Nevertheless, it should be noted that [Jung et al. \(2010\)](#) showed that the winter of 2009/10 was highly predictable in the ECMWF monthly forecast system, with skill 3–4 weeks ahead forecasting the onset of an extremely negative NAO and maintaining its persistence thereafter in subsequent forecasts. The winter of 2010/11 was also predominantly a negative NAO, and medium-range forecasts had high levels of predictability ([Langland and Maue 2012](#)). It is therefore likely that enhanced predictability, due to more predictable large-scale flow conditions, may have contributed to high levels of forecast skill in the operational forecasts at times over the winters of 2008–12.

It is accepted that having a larger ensemble will generally lead to an improvement in the forecast skill ([Buizza et al. 1998](#)). Consistent with this, the operational forecast tends to outperform the hindcasts over Europe, so an additional explanation for the high levels of skill in the operational forecasts is having 51 ensemble members, relative to the 5 in the hindcasts. The impact of sampling a period of enhanced predictability and having more ensemble members in the operational forecasts are therefore both likely to contribute to high forecast skill in the operational forecasts. However, it is not possible to quantify the relative contributions from these two sources (given the limited sample size of the available forecast data) as the operational forecast model configuration is continually changing relative to the most recent hindcasts used in the analysis.

Without a full dynamical analysis, we can only suggest what might be causing this signal of predictability over Europe. There is literature showing the MJO ([Vitart and Molteni 2010](#); [Cassou 2008](#); [Lin et al. 2010](#)), ENSO ([Ineson and Scaife 2009](#); [Bell et al. 2009](#)), and the stratosphere ([Sigmond et al. 2013](#)) can have an impact on large-scale extratropical circulation at these time scales. It is shown in [Vitart and Molteni \(2010\)](#) that forecasts that were initialized with a strong MJO in the initial conditions had a greater Brier skill score and ROC score. This was most evident for lead times of 12–19 days. It is therefore possible that the skill in the week 3 wind speed forecasts is coming from the model's ability to represent the MJO- and ENSO-related teleconnections.

A common criticism of subseasonal and seasonal forecasts is the low levels of skill in the extratropics. Typical meteorological forecast end users prefer point forecasts (i.e., a single value) for the variable of interest.

With such low levels of skill, the ensemble mean estimate can often be of little to no practical use. Methodologies therefore need to be developed to take advantage of the full probability density forecast and to utilize the information effectively in a decision-making context.

To give an example, vessel hire for offshore wind farm maintenance can result in large losses/costs depending on the decisions made when to book a vessel. For instance, the loss of revenue for a 100-MW wind farm that was offline for a month was estimated to be approximately £3 million ([Turner et al. 2013](#)). Boat hire can be in the region of £270 000 per day for a vessel that can install a substation ([BVG Associates 2012](#)). It therefore becomes clear that applying a cost/loss ratio decision framework to the problem of whether to hire the boat or wait another week could result in significant savings for the business.

7. Conclusions

Traditionally, wind speed forecasts were thought to contain no skill beyond the medium-range time scales. Having skillful forecasts at longer lead times may enable new risk management strategies for wind energy producers. We have shown that there is 10-m wind speed forecast skill at lead times of 14 days when averaging the wind speed over a weekly period (day 14–20). This skill is found to exist over the United Kingdom and other areas of Europe during the winter months of December–February. It was found in the operational forecasts that correlations of 0.6 between the forecast and observations existed at lead times of 14 days (i.e., weekly averaged wind speed over days 14–20) for some regions of Europe, particularly the United Kingdom. At lead times of 14 days, there was also probabilistic skill for the operational forecasts over the United Kingdom, with a positive CRPS skill score. It was shown that the end user could gain potential economic value from using the forecast instead of climatology, if faced with a cost/loss binary decision problem.

Very high levels of predictability were found in the operational forecasts for the recent period (2008–13). This comes, in part, from a large ensemble size (51 members), but also partly as a result of an anomalously predictable period in the large-scale weather flows. It is not possible, with the limited sample size of the 51 ensemble member forecasts, to quantify the relative contributions of these two factors in the current analysis. Nevertheless, the hindcast analysis, covering a longer period (20 years) also has small but statistically significant levels of predictability, enhancing our confidence in the ability of the model to forecast wind speeds at lead times of 14 days. It should be noted that the predictability in

week 3 was found to be the strongest during the winter months and did not cover all areas of Europe; by week 4 skill in the forecast was not statistically significant. As this was only a statistical analysis, further work would be needed to investigate the dynamical processes resulting in this predictability. Despite these limitations, it is possible the forecasts may offer useful information for a range of end-user applications.

This research has shown that skillful weekly average wind speed predictions beyond the medium-range time scales are possible. Having demonstrated that there is statistically significant wind speed skill for week 3 forecasts over the United Kingdom (and other parts of Europe) during winter, novel techniques utilizing the probabilistic information effectively to enhance the value of operational business decisions can be developed. Ongoing work to be published will demonstrate the application of these forecasts for quantitative risk management in the energy sector. Probabilistic skill scores could also be calculated for areas of Europe other than the United Kingdom that also have high levels of correlation, for instance, Germany.

Acknowledgments. We thank the U.K. Natural Environment Research Council (NERC) for funding Lynch's Ph.D. research. We thank the ECMWF for allowing use of the monthly forecasts and ERA-Interim data and Frédéric Vitart for providing the details on the monthly forecast calibration procedure. Finally, we would also like to thank Centrica for funding the CASE partnership on Lynch's Ph.D. and, in particular, Chris Durman, James Richards, and Matt Dodwell for their invaluable and ongoing discussion with regards to this research.

REFERENCES

- Arribas, A., and Coauthors, 2011: The GloSea4 ensemble prediction system for seasonal forecasting. *Mon. Wea. Rev.*, **139**, 1891–1910, doi:10.1175/2010MWR3615.1.
- Baldwin, M. P., and T. J. Dunkerton, 2001: Stratospheric harbingers of anomalous weather regimes. *Science*, **294**, 581–584, doi:10.1126/science.1063315.
- Bechtold, P., K. Martin, T. Jung, F. Doblas-Reyes, M. Leutbecher, M. J. Rodwell, F. Vitart, and G. Balsamo, 2008: Advances in simulating atmospheric variability with the ECMWF model: From synoptic to decadal time-scales. *Quart. J. Roy. Meteor. Soc.*, **134**, 1337–1351, doi:10.1002/qj.289.
- Bell, C., L. Gray, A. Charlton-Perez, M. Joshi, and A. Scaife, 2009: Stratospheric communication of El Niño teleconnections to European winter. *J. Climate*, **22**, 4083–4096, doi:10.1175/2009JCLI2717.1.
- Buizza, R., T. Petroliaigis, T. Palmer, J. Barkmeijer, M. Hamrud, A. Hollingsworth, A. Simmons, and N. Wedi, 1998: Impact of model resolution and ensemble size on the performance of an Ensemble Prediction System. *Quart. J. Roy. Meteor. Soc.*, **124**, 1935–1960, doi:10.1002/qj.49712455008.
- BVG Associates, 2012: A guide to an offshore wind farm. Tech. Rep., Crown Estate, 71 pp. [Available online at www.thecrownestate.co.uk/media/211144/guide_to_offshore_windfarm.pdf.]
- Cassou, C., 2008: Intraseasonal interaction between the Madden-Julian Oscillation and the North Atlantic Oscillation. *Nature*, **455**, 523–527, doi:10.1038/nature07286.
- Decker, M., M. A. Brunke, Z. Wang, K. Sakaguchi, X. Zeng, and M. G. Bosilovich, 2012: Evaluation of the reanalysis products from GSFC, NCEP, and ECMWF using flux tower observations. *J. Climate*, **25**, 1916–1944, doi:10.1175/JCLI-D-11-00004.1.
- Dec, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, doi:10.1002/qj.828.
- Foley, A. M., P. G. Leahy, A. Marvuglia, and E. J. McKeogh, 2012: Current methods and advances in forecasting of wind power generation. *Renewable Energy*, **37**, 1–8, doi:10.1016/j.renene.2011.05.033.
- Füss, R., S. Mahringer, and M. Prokopczuk, 2013: Electricity derivatives pricing with forward-looking information. Working Papers on Finance 2013/17, University of St. Gallen, St. Gallen, Switzerland, 65 pp., doi:10.2139/ssrn.2219855.
- Gneiting, T., 2011: Quantiles as optimal point forecasts. *Int. J. Forecasting*, **27**, 197–207, doi:10.1016/j.ijforecast.2009.12.015.
- Hagedorn, R., R. Buizza, T. M. Hamill, M. Leutbecher, and T. N. Palmer, 2012: Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts. *Quart. J. Roy. Meteor. Soc.*, **138**, 1814–1827, doi:10.1002/qj.1895.
- Hu, Z.-Z., A. Kumar, B. Huang, Y. Xue, W. Wang, and B. Jha, 2011: Persistent atmospheric and oceanic anomalies in the North Atlantic from summer 2009 to summer 2010. *J. Climate*, **24**, 5812–5830, doi:10.1175/2011JCLI4213.1.
- Hudson, D., O. Alves, H. H. Hendon, and A. G. Marshall, 2011: Bridging the gap between weather and seasonal forecasting: Intraseasonal forecasting for Australia. *Quart. J. Roy. Meteor. Soc.*, **137**, 673–689, doi:10.1002/qj.769.
- Ineson, S., and A. Scaife, 2009: The role of the stratosphere in the European climate response to El Niño. *Nat. Geosci.*, **2**, 32–36, doi:10.1038/ngeo381.
- Jolliffe, I., and D. Stephenson, Eds., 2011: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. 2nd ed. John Wiley, 292 pp.
- Jung, T., M. J. Miller, and T. N. Palmer, 2010: Diagnosing the origin of extended-range forecast errors. *Mon. Wea. Rev.*, **138**, 2434–2446, doi:10.1175/2010MWR3255.1.
- Koster, R. D., and Coauthors, 2010: Contribution of land surface initialization to subseasonal forecast skill: First results from a multi-model experiment. *Geophys. Res. Lett.*, **37**, L02402, doi:10.1029/2009GL041677.
- Langland, R. H. and R. N. Maue, 2012: Recent Northern Hemisphere mid-latitude medium-range deterministic forecast skill. *Tellus*, **64A**, 17531, doi:10.3402/tellusa.v64i0.17531.
- Lin, H., G. Brunet, and J. S. Fontecilla, 2010: Impact of the Madden-Julian Oscillation on the intraseasonal forecast skill of the North Atlantic Oscillation. *Geophys. Res. Lett.*, **37**, L19803, doi:10.1029/2010GL044315.
- McGinn, D., and Coauthors, 2013: Wind power. Renewables 2013: Global Status Report. REN21 Tech. Rep., 49–52. [Available online at www.ren21.net/Portals/0/documents/Resources/GSR/2013/GSR2013_lowres.pdf.]
- Pinson, P., 2013: Wind energy: Forecasting challenges for its operational management. *Stat. Sci.*, **28**, 564–585, doi:10.1214/13-STS445.

- , and R. Hagedorn, 2012: Verification of the ECMWF ensemble forecasts of wind speed against analyses and observations. *Meteor. Appl.*, **19**, 484–500, doi:10.1002/met.283.
- Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **126**, 649–667, doi:10.1002/qj.49712656313.
- Rodwell, M. J., and F. J. Doblas-Reyes, 2006: Medium-range, monthly, and seasonal prediction for Europe and the use of forecast information. *J. Climate*, **19**, 6025–6046, doi:10.1175/JCLI3944.1.
- Sigmond, M., J. F. Scinocca, V. V. Kharin, and T. G. Shepherd, 2013: Enhanced seasonal forecast skill following stratospheric sudden warmings. *Nat. Geosci.*, **6**, 98–102, doi:10.1038/ngeo1698.
- Sinden, G., 2007: Characteristics of the UK wind resource: Long-term patterns and relationship to electricity demand. *Energy Policy*, **35**, 112–127, doi:10.1016/j.enpol.2005.10.003.
- Turner, G., S. Roots, M. Wiltshire, J. Trueb, S. Brown, G. Benz, and M. Hegelbach, 2013: Profiling the risks in solar and wind: A case for new risk management approaches in the renewable energy sector. Tech. Rep., Swiss Re, 18 pp. [Available online at <https://www.bnef.com/InsightDownload/8090/pdf/>.]
- Vitart, F., 2004: Monthly forecasting at ECMWF. *Mon. Wea. Rev.*, **132**, 2761–2779, doi:10.1175/MWR2826.1.
- , 2014: Evolution of ECMWF sub-seasonal forecast skill scores. *Quart. J. Roy. Meteor. Soc.*, doi:10.1002/qj.2256, in press.
- , and T. Jung, 2010: Impact of the Northern Hemisphere extratropics on the skill in predicting the Madden Julian Oscillation. *Geophys. Res. Lett.*, **37**, L23805, doi:10.1029/2010GL045465.
- , and F. Molteni, 2010: Simulation of the Madden–Julian Oscillation and its teleconnections in the ECMWF forecast system. *Quart. J. Roy. Meteor. Soc.*, **136**, 842–855, doi:10.1002/qj.623.
- , and Coauthors, 2008: The new VarEPS-monthly forecasting system: A first step towards seamless prediction. *Quart. J. Roy. Meteor. Soc.*, **134**, 1789–1799, doi:10.1002/qj.322.
- Weigel, A. P., D. Baggenstos, M. A. Liniger, F. Vitart, and C. Appenzeller, 2008: Probabilistic verification of monthly temperature forecasts. *Mon. Wea. Rev.*, **136**, 5162–5182, doi:10.1175/2008MWR2551.1.