RESEARCH ARTICLE

Incorporating the North Atlantic Oscillation into the post-processing of MOGREPS-G wind speed forecasts

Sam Allen1 | Gavin R. Evans2 | Piers Buchanan2 | Frank Kwasniok1

1Department of Mathematics, University of Exeter, Exeter, UK
2Met Office, Exeter, UK

Abstract
Changes in the North Atlantic Oscillation (NAO) heavily influence the weather across the UK and the rest of Europe. Due to an incorrect representation of the polar jet stream and its associated physical processes, it is reasonable to believe that errors in numerical weather prediction models may also depend on the prevailing behaviour of the NAO. To address this, information regarding the NAO is incorporated into statistical post-processing methods through a regime-dependent mixture model, which is then applied to wind speed forecasts from the Met Office’s global ensemble prediction system, MOGREPS-G. The mixture model offers substantial improvements upon conventional post-processing methods when the local wind speed depends strongly on the NAO, but the additional complexity of the model can hinder forecast performance otherwise. A measure of regime dependency is thus defined that can be used to differentiate between situations when the numerical model output is, and is not, expected to benefit from regime-dependent post-processing. Implementing the regime-dependent mixture model only when this measure exceeds a certain threshold is found to further improve predictive performance, while also producing more accurate forecasts of extreme wind speeds.

KEYWORDS
North Atlantic Oscillation, statistical post-processing, weather forecasting, weather regimes, wind speed

1 INTRODUCTION

Numerical weather prediction (NWP) models are becoming progressively more complex, operating at higher resolutions and incorporating more intricate model physics. Nonetheless, systematic biases remain present in the model output due to incomplete knowledge of atmospheric dynamics, and errors when specifying the initial forecast state. Therefore, although an ensemble of model runs provides insight into the uncertainty surrounding the forecast, the comprising members are products of an imperfect prediction system. Raw ensemble output has thus repeatedly been found to exhibit less spread than desired, with the observed weather value falling outside the range of ensemble members more often than expected (e.g., Hamill and Colucci, 1997). Hence, to avoid misguided conclusions drawn directly from the model output, it has become imperative that the forecast undergoes some form of post-processing, or recalibration.

Over the past two decades, several statistical post-processing methods have been proposed that exploit relationships between historical forecasts and observations to...
correct for these systematic model errors. State-of-the-art techniques quantify the forecast uncertainty by issuing probabilistic forecasts in the form of statistical distributions. The benefits of such post-processing methods – most commonly variants of Bayesian Model Averaging (BMA; Raftery et al., 2005) and Ensemble Model Output Statistics (EMOS; Gneiting et al., 2005) – are now widely recognised, with the methods routinely implemented at operational forecasting centres.

Established post-processing methods often use only the ensemble members, or ensemble mean and variance, as predictors in the statistical models. However, recent studies have highlighted the potential benefits of using several sources of information when post-processing (Taillardat et al., 2016; Messner et al., 2017; Rasp and Lerch, 2018). Allen et al. (2019) suggest that additional predictors could be selected using meteorological intuition. In particular, the authors propose that errors in operational ensemble forecasts are dependent on synoptic-scale structures in the atmosphere, i.e. weather regimes, and hence this regime information should be utilised when post-processing. In doing so, the recalibration can remedy conditional biases owing to the occurrence of certain weather regimes, which might otherwise be ignored. The need for such regime-dependent post-processing methods has been noted in several recent studies (Scheuerer, 2014; Pantillon et al., 2018; Rodwell et al., 2018).

Indeed, weather regimes are closely tied to the atmosphere’s predictability (Thompson, 1957), and the effect they have on local weather systems is well documented (Hannachi et al., 2017). They are thus, in themselves, of considerable interest to weather forecasting centres. As such, it is common for these forecasting centres to monitor the occurrence and behaviour of circulation patterns over the local domain. The UK Met Office, for example, have identified 30 synoptic patterns that influence the UK weather, which are now used operationally in their Decider program (Neal et al., 2016). It should therefore be reasonably straightforward to utilise regime information within operational post-processing suites, and this article demonstrates how this could be achieved. To do so, wind speed forecasts from the Met Office’s global ensemble prediction system, MOGREPS-G (Walters et al., 2017; Porson et al., 2020), issued in the 2-year period between January 1, 2018 and December 31, 2019. No major model upgrades have been performed since this period, and hence the model configuration is very similar to that currently in operation. The model employs a 20 km horizontal resolution over the globe, with ensembles comprising 18 constituent members. We consider here lead times at 12-hr intervals up to 6 days ahead. The forecasts are initialised at 1200 UTC and thus validated at either midnight or midday, allowing the evaluation of both day- and night-time wind speed predictions.

Forecasts are evaluated at 106 stations over the UK and Ireland, pictured in Figure 1. Output from the grid-based MOGREPS-G prediction system is bi-linearly interpolated

2 | DATA

2.1 Forecasts and observations

This study utilises 10-m wind speed forecasts extracted from the Met Office’s global ensemble prediction system, MOGREPS-G (Walters et al., 2017; Porson et al., 2020), issued in the 2-year period between January 1, 2018 and December 31, 2019. No major model upgrades have been performed since this period, and hence the model configuration is very similar to that currently in operation. The model employs a 20 km horizontal resolution over the globe, with ensembles comprising 18 constituent members. We consider here lead times at 12-hr intervals up to 6 days ahead. The forecasts are initialised at 1200 UTC and thus validate at either midnight or midday, allowing the evaluation of both day- and night-time wind speed predictions.

Forecasts are evaluated at 106 stations over the UK and Ireland, pictured in Figure 1. Output from the grid-based MOGREPS-G prediction system is bi-linearly interpolated
2.2 Weather regimes

The synoptic-scale atmospheric flow can often be characterised by just a few circulation patterns, such that the continuous evolution of the atmosphere is represented by transitions between these distinct states (Franzke et al., 2011). These regimes recur at the same geographical locations and persist beyond the time scales of individual weather events, and can thus be thought of dynamically as quasi-stationary equilibria in the atmosphere’s phase space (Charney and DeVore, 1979). Although a dynamical phenomenon, atmospheric regimes are regularly identified using statistical methods that search for recurring patterns in archives of large-scale atmospheric variables (Hannachi et al., 2017).

The circulation patterns used here are those implemented in the Met Office’s Decider tool, described in detail in Neal et al. (2016). That is, a set of 30 recurring weather patterns over the Euro-Atlantic region, detected using a simulated annealing clustering technique applied to 154 years of mean sea-level pressure (MSLP) anomaly fields (Philipp et al., 2007). Eight larger-scale circulation types, hereafter referred to as regimes, are then constructed by objectively grouping similar weather patterns in such a way that patterns are matched with those that exhibit similar structural properties, so that zonal flow patterns, for example, across different seasons may be clustered together. Figure 2 depicts the eight MSLP anomaly fields corresponding to the centres of the weather regimes under consideration. The regime at any time is defined as that which minimises the distance between the regime centre and the instantaneous MSLP anomaly field. Descriptions of the eight regimes in relation to the flow over the UK are available in Neal et al. (2016).

The regimes in Figure 2 are displayed in order of decreasing frequency, and their mean persistence times during the two year period considered here range from 1.2 to 2.8 days, indicating fairly transient behaviour that corroborates findings in Neal et al. (2016). The distributions of wind speeds across all locations are displayed for each regime in Figure 3. There is some clear dependence on the regimes, particularly between Regimes 1 and 2, which represent, respectively, the negative and positive phase of the North Atlantic Oscillation (NAO). The NAO is known to have a profound effect on the UK (and European)
weather, with more extreme wind speeds related to the occurrence of positive NAO events (Hurrell, 1995; Hurrell and Deser, 2009).

Furthermore, if there is insufficient data available on which to reliably train the post-processing methods, then they may overfit the training data, resulting in poor out-of-sample predictions. By including additional predictors, regime-dependent post-processing methods are less parsimonious than conventional approaches, and hence more susceptible to overfitting. This is especially pertinent in operational settings, where the amount of available data is limited due to continual model upgrades. To circumvent this, the eight regimes are condensed into just three. The opposite phases of the NAO each have a significant effect on the observed wind speeds, whereas the variation between the remaining six regimes is comparatively weak. Moreover, some of the latter regimes occur fairly infrequently during the time period under examination. Therefore, the remaining six regimes are grouped together; the three regimes considered here are thus the NAO- and NAO+ regimes, as well as a third regime combining Regimes 3-8 of Figure 2. In the 2-year time period of interest, the NAO- regime occurs on 22.1% of the 730 days, the NAO+ regime occurs on 16.3%, and the third regime on the remaining 61.6%. It is possible to utilise all eight regimes when post-processing, but such an approach is found here to induce parameter uncertainty in the recalibration methods that outweighs the benefits of including the regime information (not shown).

Although Figure 3 demonstrates that the wind speed depends on the prevailing phase of the NAO, it may be the case that the MOGREPS-G ensemble forecasts capture this dependence, rendering regime-dependent post-processing methods superfluous. To assess whether this is the case, we evaluate the calibration of the forecasts in each regime. For this purpose, we employ the reliability index, defined as

\[
\Delta = \sum_{m=1}^{M+1} \left| \rho_m - \frac{1}{M+1} \right|
\]  

(Delle Monache et al., 2006), where \(M\) is the size of the ensemble, and \(\rho_m\) denotes the relative frequency with which the observed wind speed exceeds \(m - 1\) of the ensemble members. If the ensemble prediction system is calibrated, then the verifying wind speed should be equally likely to fall between any two members of the ensemble, meaning \(\rho_m = 1/(M + 1)\), for all \(m = 1, \ldots, M + 1\). The reliability index thus measures the divergence between the relative frequencies and their ideal value, with a more reliable prediction system attaining a lower index.

To evaluate whether biases exist in the MOGREPS-G ensemble forecasts owing to the occurrence of certain weather regimes, we calculate this reliability index separately for forecasts associated with each regime under consideration. Figure 4 illustrates how this index varies for forecasts corresponding to the two phases of the NAO. If the prevailing regime does not influence the errors in the prediction system, then the reliability index should not change depending on the phase of the NAO, meaning points would lie along the dotted line of equality in Figure 4. The index is substantially greater than zero in both the NAO- and NAO+, reflecting the underdispersive nature of the forecasts, and, although there is weak positive correlation between \(\Delta\) in the two regimes (0.369), substantial deviation from the diagonal advocates the implementation of regime-dependent post-processing methods.

Moreover, the reliability index is not distributed evenly about the diagonal in Figure 4, with larger forecast errors
Typically occurring in the positive phase of the NAO than in the NAO-. Regime-dependent post-processing may therefore be more beneficial for forecasts associated with this cyclonic regime. The miscalibration also varies considerably with the spatial location, suggesting localised post-processing methods are desirable, while the variation in the reliability index between the two regimes also depends substantially on the station, even over the relatively small spatial domain considered here. Hence, although there is evidence to suggest regime-dependent post-processing is advantageous in this instance, it may not be necessary at all locations. To better understand the spatial properties of the forecast errors, the weather stations in Figure 3 have been classified into coastal, inland and mountainous locations, though it appears that regime-dependent biases in the MOGREPS-G output occur for all types of locations at this lead time.

3 | STATISTICAL POST-PROCESSING

3.1 | Training and testing

Statistical post-processing seeks to exploit systematic errors in previous forecasts to address those that might occur in the current prediction. To do so, a training set of historical forecasts and observations is required, where the choice of training data should reflect the biases expected to manifest in the current forecast. Training windows that adapt in the presence of new data assume that the error in the current forecast will likely behave similarly to that of recent forecasts. Rolling, or sliding, windows, for example, use only the most recent forecast-observation pairs available to train the post-processing model. The length of the window is a compromise between using enough data to obtain reliable parameter estimates, and not using too much data, so as to capture the recent behaviour of the model biases.

The limited amount of data used in a rolling window can often result in unstable parameter estimates (Scheuerer, 2014; Lang et al., 2019). Although the window length can be extended to avoid this, continually increasing the length contradicts the motivation for using a rolling window. Once the window can no longer account for the seasonal cycle in the data, it may be more sensible to use a larger amount of data, spanning several seasons or years. This data set could be fixed to obviate the need to update parameter estimates for every new forecast (Pinson and Girard, 2012). In this study, since the observed wind speeds do not exhibit a strong annual cycle, the additional training data afforded by a fixed window is found to outweigh the adaptive nature of a rolling window (not shown). Therefore, the coefficients of all post-processing models described in this section are estimated over a fixed window between January 1, 2018 and December 31, 2018. The remaining year of data is used as a test data set, on which to verify the resulting forecasts.

Although there is little seasonal variability in the model biases, Figure 4 illustrates that there is a considerable spatial dependency. Therefore, to account for locally varying biases, post-processing is performed separately at each station under consideration. Despite being more computationally expensive, this method performs significantly better than a global post-processing approach, in which one set of model parameters is estimated for every location, and the site-specific post-processing approach is thus favoured here.

Just as the post-processing methods are trained using station observations, the resulting forecasts are also verified against wind speed measurements at the locations of interest. The resulting forecast distributions are assessed using the continuous ranked probability score (CRPS). For a forecast with predictive cumulative distribution function (CDF) \( F \) and corresponding observation \( y \), the CRPS is defined as

\[
\text{CRPS}(F, y) = \int_{-\infty}^{\infty} [F(u) - I(u \geq y)]^2 du
\]

(Matheson and Winkler, 1976), where \( I(\cdot) \) denotes the indicator function, which takes the value one when the statement inside is true, and zero otherwise. The CRPS is negatively oriented, so that a lower score indicates a more accurate forecast, and, as a proper score, it assesses both the reliability and sharpness of the forecast (Gneiting and Raftery, 2007). The total CRPS is then taken to be the average CRPS over all forecasts in the test data. The continuous ranked probability skill score (CRPSS) is also used to measure the skill of the regime-dependent approaches relative to conventional post-processing. The CRPSS can be interpreted in terms of the percentage improvement in the forecast accuracy relative to a baseline forecast, and hence larger values are desired (Wilks, 2019).

The CRPS and its skill score are commonplace in weather forecasting. In the following section, rank and Probability Integral Transform (PIT) histograms (Dawid, 1984; Hamill and Colucci, 1997; Gneiting et al., 2007), as well as the coverage of 90% prediction intervals, are also used to assess the reliability of forecasts. The sharpness, or resolution of the predictive distribution is then considered using the width of the 90% prediction intervals. A smaller width indicates a sharper, or more refined forecast, but is desirable only subject to calibration (Gneiting et al., 2007).

Pantillon et al. (2018) suggest that if extreme wind speeds, or wind gusts, can be linked to the occurrence
of certain weather patterns, then regime-dependent post-processing methods may be capable of improving forecasts of these high-impact weather events. Therefore, predictive performance in the upper tail of the forecast distribution is evaluated using the threshold-weighted continuous ranked probability score (twCRPS). The twCRPS is defined as

$$\text{twCRPS}(F, y) = \int_{-\infty}^{\infty} [F(u) - I(u \geq y)]^2 \omega(u) \, du,$$

for some non-negative weight function $\omega(u)$ (Gneiting and Ranjan, 2011). As in Lerch and Thorarinsdottir (2013), interest lies in the upper tail of the forecast distribution, and hence the weight function used here is $\omega(u) = I(u \geq t)$, for some threshold $t$. When such a weight function is used, the twCRPS is a strictly locally proper scoring rule (Holzmann and Klar, 2017) that considers only the performance of the forecast distribution above this threshold, and it thus concerns predictions of the upper-tail behaviour.

### 3.2 Ensemble Model Output Statistics

Ensemble Model Output Statistics (EMOS), or non-homogeneous regression, is possibly the most frequently implemented post-processing approach, owing largely to its simplicity and the ease with which it can be modified for use in different situations. EMOS assumes that the weather variable to be forecast, referred to as the predictor or response variable, follows a statistical distribution that depends on the raw ensemble output. The choice of predictive distribution is determined by the weather variable of interest: wind speed, for example, is non-negative, and hence a sensible choice would be a forecast distribution with a positive support, such as a gamma (Sloughter et al., 2010), truncated normal (Thorarinsdottir and Gneiting, 2010), or truncated logistic (Messner et al., 2014; Scheuerer and Möller, 2015) distribution. The latter is found here to outperform alternative options.

We therefore assume that the future wind speed is a random variable, $Y$, that follows a truncated logistic distribution with location and scale that depend on the ensemble mean, $\bar{x}$, and variance, $s^2$, respectively:

$$Y|\mathbf{x} \sim L_0(\alpha + \beta \bar{x}, \sqrt{\gamma + \delta s^2}),$$

where $\mathbf{x}$ denotes the vector of ensemble members, and $L_0(\mu, \sigma)$ denotes the logistic distribution that has been truncated below at zero, with location parameter $\mu$ and scale parameter $\sigma$. Hence, the square of the scale is expressed as a linear function of the ensemble variance, while truncation at zero ensures the resulting forecast distributions assign mass only to non-negative wind speed values. The post-processing parameters ($\alpha, \beta, \gamma$ and $\delta$) are estimated by finding those that minimise the total CRPS over the training data, and, to ensure positive variance components, optimisation is performed using $\xi = \sqrt{\gamma}$ and $\kappa = \sqrt{\delta}$ rather than $\gamma$ and $\delta$ directly.

The CRPS for a truncated logistic distribution with location $\mu$ and scale $\sigma$ can be given in closed form:

$$\text{CRPS}(L_0(\mu, \sigma), y) = \sigma \left( \log \frac{1 - p_0}{1 - p_y} - \frac{p_0^2}{(1 - p_0)^2} \log p_0 \right) + \frac{1}{1 - p_0} \log p_0 - \frac{1}{1 - p_0} \right).$$

This expression is identical to, but more compact than, that derived by Scheuerer and Möller (2015). In keeping with the notation therein, $p_0$ and $p_y$ are the values of the logistic, $L(\mu, \sigma)$, CDF evaluated at 0 and $y$, respectively. The total CRPS is then the average CRPS across all forecast-observation pairs.

The twCRPS can similarly be derived in closed form. Our result, verified against numerical integration, is that

$$\text{twCRPS}(L_0(\mu, \sigma), y) = \frac{\sigma(p_1 - 1 - \log p_t)}{(1 - p_0)^2}$$

if $y \leq t$, and

$$\text{twCRPS}(L_0(\mu, \sigma), y) = \sigma \left( \log \frac{1 - p_t}{1 - p_y} - \frac{p_t^2}{(1 - p_t)^2} \log p_t \right) - \frac{1}{1 - p_0} \log p_0 - \frac{1}{1 - p_0} \right).$$

if $y > t$. Here, $p_t$ is the logistic CDF evaluated at the threshold $t$. Again, the twCRPS values of the individual observations are averaged to give the total twCRPS of the data set.

Although interest here is on the logistic distribution truncated below at zero, we remark that Equations (5)–(7) can be generalised to any truncation point $l$ of the logistic distribution by replacing $p_0$ with $p_l$, the CDF of the logistic distribution at $l$. As $l \to -\infty$ and $p_l \to 0$, the logistic distribution $L(\mu, \sigma)$ is recovered from the truncated logistic distribution. Correspondingly, Equation (5) tends to the well-known CRPS for the logistic distribution,

$$\text{CRPS}(L(\mu, \sigma), y) = -\sigma[1 + \log p_t + \log(1 - p_y)] = y - \mu - \sigma - 2\sigma \log p_y,$$

given by Taillardat et al. (2016) and Jordan et al. (2018). Similarly, the twCRPS of the logistic distribution can be
recovered from Equations (6) and (7):

\[ \text{twCRPS}(L(\mu, \sigma), y) = \sigma(p_t - 1 - \log p_t) \] (9)

if \( y \leq t \), and

\[ \text{twCRPS}(L(\mu, \sigma), y) = \sigma \left( p_t - 1 + \log \frac{1 - p_t}{1 - p_y} - \log p_y \right) \] (10)

if \( y > t \), a result that, to our knowledge, has not previously been given in the literature.

3.3 Regime-dependent mixture model

Mixture models have previously been shown to be an effective way of combining information from several sources in a weather forecasting context (Wilks, 2002; Gneiting and Ranjan, 2013; Baran and Lerch, 2016, 2018). Allen et al. (2019) therefore propose a mixture model to include regime information when post-processing. Mathematically, this involves extending Equation (4) to

\[ Y|\mathbf{x} \sim \sum_{r=1}^{R} w_r L_0(\alpha_r + \beta_r \mathbf{x}, \sqrt{\gamma_r + \delta_r s^2}), \] (11)

where \( R \) is the number of regimes under consideration: in this case, \( R = 3 \). A separate forecast distribution is associated with each identified regime, and, making the same distributional assumptions about the predictive distribution in each regime, there are \( 4R \) parameters to estimate, one set for each regime. The weights associated with each regime, \( w_r \), allow the model to account for uncertainty present when attributing the forecast to a regime. It is important to note that the weights in this case are functions of time, rather than parameters as such, highlighting that the weights will change depending on the prevailing behaviour of the atmosphere. Potential approaches to calculate the mixture model weights are discussed in Allen et al. (2020), where it is found that, in comparison with alternative choices, the regimes predicted by high-resolution numerical weather models provide a reasonable estimate of the future synoptic-scale state.

As well as recording the regime that manifests, the Met Office store the daily regime that is forecast by their global deterministic model, up to 6 days in advance. Therefore, the mixture model weight used here is an indicator function that takes the value one only for the regime that is forecast by this deterministic model. In this case, the mixture model reverts to a truncated logistic distribution with the appropriate coefficients based on the regime that is forecast. Parameter estimation and forecast verification can thus again be performed using the CRPS and twCRPS as given in Equations (5)–(7).

Although the large-scale atmospheric state is considerably more predictable than the more turbulent weather, this forecast of the regime may itself exhibit biases. Table 1 presents the number of instances that each regime is predicted and observed in the 730 days under consideration, shown at a lead time of 6 days. Even this far in advance, the numerical weather model only once mistakes the phase of the NAO, with the deterministic global forecast correctly estimating the future regime on roughly 70% of the 730 days. It is possible to address this error by applying a simple form of post-processing to these regime forecasts: the mixture model weight corresponding to a particular forecast could be given by the relative frequency of each regime occurring given the regime that is forecast. For example, if the NAO-regime is forecast to occur in 6 days time, then Table 1 suggests that the mixture model weight corresponding to this regime should be 0.657 (111/169), whereas those for the NAO+ and ‘Other’ regimes should be 0.006 (1/169) and 0.337 (57/169), respectively. Since this weight is not an indicator function, the post-processing coefficients associated with each regime would be estimated simultaneously.

The ability of these two weight functions to forecast the future weather regime is assessed using the Brier score (Brier, 1950) in Figure 5. Results are also shown for a persistence forecast constructed from the regime at the forecast initialisation time, and the climatological regime frequencies. From Figure 5, we see that the initial regime becomes no better than climatology after just 2 days, highlighting the transient behaviour of the regimes considered here. The deterministic regime forecast, however, maintains skill until 6 days in advance, at which point the recalibration becomes beneficial. The recalibrated regime forecast is constrained to be always as good as climatology, and hence will, in theory, always have non-negative skill, even after the deterministic forecast on which it is based is completely uninformative.

Nonetheless, the forecasts resulting from the two choices of weight perform comparably (not shown), and hence results in the following section are for the simpler, deterministic regime forecast. Since we are post-processing wind speed forecasts issued by a global

| Observed regime | NAO− | NAO+ | Other | Total |
|----------------|------|------|-------|-------|
| NAO−           | 111  | 1    | 57    | 169   |
| Forecast regime|      |      |       |       |
| NAO+           | 0    | 72   | 47    | 119   |
| Other          | 49   | 45   | 348   | 442   |
| Total          | 160  | 118  | 452   | 730   |
ensemble prediction system, it would also be possible to derive a forecast of the future regime from each of the ensemble members, and use the proportion of members predicting each regime to define the mixture model weights. However, these regime forecasts have not been stored for the prediction system and date range under consideration, and hence are not readily available. Therefore, although the distribution of the ensemble members would capture some of the flow-dependent uncertainty in the future regime, making them a sensible option when post-processing in real time (Allen et al., 2020), the regime predicted by the deterministic global model is used to define the weights in this study. Using a weight that is an indicator function means parameters can be estimated separately by stratifying the training data into regime-dependent subsets depending on the regime that is forecast (Allen et al., 2019). This is more numerically stable than estimating all parameters simultaneously, and hence also less computationally expensive (Baran and Lerch, 2018).

3.4 A hybrid approach

The regime-dependent mixture model is more complex than the truncated logistic forecast distribution, and is thus more susceptible to overfitting in the presence of limited training data. Therefore, the mixture model, although offering more flexibility, could potentially hinder forecast performance when the model biases do not depend on the regimes under consideration. A more robust approach might involve identifying situations in which regime-dependent post-processing is expected to be most beneficial, and implementing it only on such occasions. This emulates ideas in Lerch and Thorarinsdottir (2013), whereby a separate post-processing model is applied depending on the prevailing circumstances. Lerch and Thorarinsdottir (2013), and later Baran and Lerch (2015), implement a post-processing model that switches between two predictive distributions depending on whether the ensemble median exceeds a predefined threshold. In doing so, the forecast can better adapt to the biases expected of the current model output.

It remains to identify circumstances in which regime-dependent post-processing should be beneficial. Allen et al. (2020) consider the spread of the average wind speeds between regimes at a variety of locations, and show that improvements gained by including the regime information are correlated with this between-regime spread. That is, locations whose wind speeds are more heavily influenced by changes in the weather regime tend to benefit most from regime-dependent post-processing. Therefore, we calculate here a similar measure of regime dependency from the observations in the training data, and record whether it exceeds a certain threshold. The measure of regime dependency is taken to be the between-group component of the empirical law of total variance applied to the wind speed observations in the training data:

$$\sum_{r=1}^{R} \frac{n_r}{n} (\bar{y}_r - \bar{y})^2,$$

where $n_r/n$ is the proportion of days on which regime $r$ occurs in the training data, $\bar{y}_r$ is the average wind speed given regime $r$, and $\bar{y}$ is the overall mean wind speed. The measure is shown for each location in Figure 6. The stations have been divided into coastal, inland, and mountainous locations, from which it can be seen that, just as locations on the west coast of the UK and Ireland were earlier identified as being associated with higher average wind speeds, they are also particularly affected by the phase of the North Atlantic Oscillation.

The measure of regime dependency introduced here is similar to that used in Allen et al. (2020), but can better account for the different climatological frequencies of the regimes. Larger values suggest a stronger regime influence and hence advocate the use of regime-dependent post-processing, whereas if this quantity lies below the chosen threshold, then the truncated logistic distribution is issued as the forecast, and the regime information is ignored. Of course, there are several alternative ways to measure the regime dependency that may be more indicative of situations where regime-dependent post-processing is desirable. A brief comparison of possible measures is performed in the following section.

Lerch and Thorarinsdottir (2013) estimate the threshold for their combination method by finding that which minimises the CRPS in the training data. This approach
Theregimedependency ($m^2 \cdot s^{-2}$; Equation 12) of the day-time wind speeds at each station on the domain under consideration. The points have been classified into coastal (C), inland (IL), and mountainous (M) locations [Colour figure can be viewed at wileyonlinelibrary.com]

fails here since the motivation for pooling together the two post-processing methods is to avoid situations in which the mixture model overfits the training data; in the training data, the mixture model almost always performs better than the truncated logistic method, and hence the optimal threshold in terms of the CRPS is zero, suggesting regime-dependent post-processing should always be implemented. However, this does not necessarily translate to better out-of-sample forecasts, and an alternative method of choosing the threshold is thus required. The threshold used here is the same for all lead times, and is chosen subjectively such that the regime-dependent mixture model is applied to roughly 50% of forecasts.

4 | RESULTS

Firstly, consider the calibration of the post-processed forecasts. Figure 7 shows rank and probability integral transform (PIT) histograms for the ensemble forecast, and for the various post-processing methods. The raw ensemble is again found to be underdispersed, with the observation falling either above or below all ensemble members with a disproportionately high frequency, whereas the post-processing methods yield PIT histograms that are approximately uniform, indicating well-calibrated forecasts. There appears, however, to be some systematic deviation from uniformity in the tails of the post-processed predictive distributions, though this is not remedied when using alternative families of parametric distributions.

Table 2 displays the average coverage and width of the 90% prediction intervals obtained from the various post-processing methods. The truncated logistic distribution generates a coverage that is close to, but slightly larger than, the optimal 90% coverage, and the regime-dependent approaches improve on this slightly. We see, however, that the standard post-processing method tends to be overdispersed in the more predictable NAO- regime, but underdispersed in the NAO+, suggesting it does not fully capture the changes in predictability arising due to the regimes. The regime-dependent approaches decrease the spread of the forecast distribution in the NAO- regime and increase the spread in the NAO+ regime, thus yielding a coverage closer to the optimal value. Being a combination of the two, the hybrid method exhibits prediction intervals that are a compromise between those of the truncated logistic distribution and the mixture model.

Looking now at the improvement in skill gained by including regime information, Figure 8 shows the continuous ranked probability skill score (CRPSS) against forecast horizon for the two regime-dependent methods, with the truncated logistic forecast used as a baseline. The score is averaged over all locations and forecast instances in the test data set. Neither method performs consistently worse than the standard approach, though the improvements are always relatively small. This is to a lesser extent when the hybrid method is used, despite the CRPSS being constrained to equal zero at a large proportion of locations. The improvements are largest at short lead times, before receding to zero as the forecast horizon increases. As is apparent in Figure 5, the quality of the regime forecasts deteriorates with lead time, and hence the biases in the forecast become less dependent on the regime that is predicted by the deterministic weather model, meaning there is less benefit to post-processing using this definition of the regime. At these longer lead times, it is found that significant improvements are available if post-processing is conditioned on the true regime at the forecast validation time (not shown; Allen et al., 2020); the forecast accuracy across all locations improves by almost 1% in this case. Although this future regime is unknown in practice, this result highlights that a more accurate prediction of the future regime would be informative when post-processing medium- and long-range weather forecasts.

Despite its simplicity, the threshold for switching between the truncated logistic distribution and the mixture model appears to distinguish well between situations when regime-dependent post-processing is, and is not, desirable. The skill score for the hybrid approach is shown in Figure 9 for a range of possible threshold choices at a lead time of 5 days. A threshold of zero corresponds
FIGURE 7  Rank and Probability Integral Transform (PIT) histograms for the raw ensemble forecast, along with the three post-processing methods, shown at lead time of 4 days. A horizontal red line is drawn at 1/19, indicating perfect calibration [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 2  Average coverage (%) and width (m s\(^{-1}\)) of 90% prediction intervals derived from the truncated logistic (TL), mixture model (MM), and hybrid forecasts at a lead time of 24 hr

|              | NAO−   | NAO+   | Other  | Overall |
|--------------|--------|--------|--------|---------|
| Coverage     | TL     |        |        |         |
|              | 92.03  | 87.69  | 91.56  | 91.01   |
|              | MM     | 90.26  | 89.86  | 90.17   | 90.14   |
|              | Hybrid | 91.02  | 88.47  | 91.16   | 90.67   |
| Width        | TL     |        |        |         |
|              | 5.61   | 6.09   | 5.76   | 5.78    |
|              | MM     | 5.37   | 6.66   | 5.50    | 5.67    |
|              | Hybrid | 5.43   | 6.28   | 5.66    | 5.71    |

FIGURE 8  CRPSS for the regime-dependent forecast (RDPP), as well as the hybrid approach, relative to the conventional truncated logistic forecast distribution. Error bars indicate 95% confidence intervals for the skill score, obtained via non-parametric bootstrap resampling [Colour figure can be viewed at wileyonlinelibrary.com]

The effect of changing the threshold can also be seen in Figure 10. Since the post-processing methods are trained using a fixed, site-specific window, each location is associated with only one value of the regime dependency (Equation 12). The top panel of Figure 10 shows to always implementing the regime-dependent mixture model, and, conversely, as the threshold increases, this is equivalent to always enforcing the truncated logistic forecast, and hence never utilising regime information. As a result, the CRPSS tends to zero for higher thresholds. A threshold of 0.4 was used here, which appears suitable at this lead time.

The forecast skill for the mixture model at each location under consideration, plotted against this measure of regime dependency. A vertical line is drawn at the chosen threshold. As anticipated, locations whose wind speeds are highly dependent on the prevailing NAO phase tend to exhibit larger skill scores, with improvements at individual locations reaching 4% at this lead time. A distinction has
also been made between coastal, inland and mountainous sites, from which it can be seen that regime-dependent post-processing is typically most beneficial at locations on the coast of the UK and Ireland for this lead time. However, there are several negative skill scores to the left of the chosen threshold in Figure 10, suggesting the mixture model tends to do more harm than benefit at sites where the regime dependency is weak.

The lower panel of Figure 10 shows the analogous plot for the hybrid approach. Below the threshold, the skill scores are zero since the same forecast is being issued as the reference. This removes the largely negative effect the more complex model has at these locations. Conversely, the larger, positive skill scores that tend to occur at stations heavily influenced by the regimes are still present. This approach thus combines the benefits of the established post-processing with the more complex, yet more flexible, regime mixture model. Note, however, that some negative skill scores still occur for the hybrid approach, and some positive skill scores below the threshold are nullified. This could be avoided if an alternative way to choose the threshold exists that can better recognise when regime-dependent post-processing is desirable.

With this in mind, we look to compare various choices of the measure of regime dependency. Six different measures are evaluated, all of which assume the following form:

$$\sum_{r=1}^{R} \frac{N_r}{n} (\bar{z}_r - \bar{z})^2,$$

(13)

where $\bar{z}$ denotes the mean of a quantity $z$ calculated over the entire data set, while $\bar{z}_r$ is the average of the same quantity estimated only from forecast–observation pairs associated with regime $r$. Hence, all of these measures of regime dependency represent the amount of variation in a metric that can be explained by changes in the regime. Clearly, Equation (12) is a particular example of this with $z$ equal to the observed wind speed. In addition to the regime-dependent variation of the wind speed itself, we also consider the variation in the bias of the ensemble mean forecast, the squared error of the ensemble mean forecast, the absolute error of the ensemble mean forecast, the ensemble variance and the reliability index of the raw ensemble forecast (Equation 1). These are summarised in Table 3.

For use within the hybrid method proposed here, a measure is desired that generates a monotonic relationship between the regime dependency and the improvement gained by the regime-dependent mixture model relative to the conventional truncated logistic EMOS approach. Hence, to assess the utility of the various measures, we report Spearman’s rank correlation (Wilks, 2019) between each measure and the improvement, as quantified using the CRPSS; a higher absolute correlation indicates a more
monotonic relationship. To maintain comparison with Figure 10, Table 3 displays the results at a lead time of 5 days. The measures of regime dependency have been calculated over the training data, since this information would be available to practitioners at the time of forecasting. Although post-processing is concerned with forecast errors rather than the observations themselves, the improvements are most strongly correlated with variations in the observed wind speed owing to the regimes. This highlights the extent to which climatological information is used by the post-processing model to recalibrate the ensemble forecasts. This is particularly pertinent at longer forecast horizons, where less information is contained in the ensemble prediction system, and hence post-processing methods should decrease the influence that the numerical forecasts assert on the resulting predictive distribution and use instead more information from the climatological distribution of the wind speeds.

Moreover, since Equation (12) is not dependent on the forecasts, it would be possible to calculate this quantity over a larger set of historical observations, rather than only the training data. This is not possible if forecast biases were used to calculate this measure of regime dependency instead. In any case, the bias in the ensemble mean forecast exhibits a lower correlation with the CRPSS, and this decreases further when considering the squared or absolute forecast error. Similar conclusions are drawn for other lead times. Surprisingly, the correlation is lowest between the improvements and the reliability index of the raw ensembles. One reason for this is that the index does not distinguish between the direction of any errors. For example, the reliability index generated by a negatively biased prediction system would be the same as that produced by forecasts exhibiting a positive bias of the same magnitude, behaviour that has previously been observed for wind speed forecasts associated with different regimes (Allen et al., 2020), particularly when the regimes correspond to the same mode of synoptic-scale variation, as is the case here. More complex measures of regime dependency could be derived to exploit this, but this is not considered here.

We now focus further on the behaviour of the hybrid approach. Figure 11 displays the CRPS for the truncated logistic EMOS and the hybrid methods, evaluated separately for days associated with each regime. The corresponding skill score of the hybrid approach relative to the truncated logistic forecast is also shown. Note first that the fluctuations between lead times indicate forecasts verifying during the night are generally more accurate than those made for midday. There is also a noticeable structure to the improvements. Although there is little difference in the CRPS between the approaches in the NAO- and “Other” regimes, in the NAO+ the hybrid approach improves upon the conventional method by as much as 3%, even when averaged over all locations. This agrees with results in Allen et al. (2020), whereby forecasts assigned to regimes that differ most from climatology are those most likely to improve. As seen in Figure 3, the NAO+ is associated with higher wind speeds, suggesting the regime-dependent approaches may help to forecast the occurrence of high-impact wind speed events.

The improvements in the NAO+ reach 3% across all locations, but the CRPSS at particular stations can be much larger than this. Figure 12 displays the CRPS and its skill score for the hybrid approach relative to the truncated logistic forecast at one location on the south-west coast of Wales. When neither the NAO- nor NAO+ regime occurs, the improvements are negligible and tend to fluctuate around zero for all lead times. During NAO+ events, on the other hand, the regime-dependent approach greatly outperforms the standard post-processing method, with improvements reaching almost 20% at a lead time of 36 hrs. Large CRPSS values are also observed in the NAO-regime, though the improvement in this regime depends strongly on the lead time of interest. Overall, the improvement gained from regime-dependent post-processing at this location remains above 5% until 3 days in advance, before decreasing somewhat for longer lead times.

These results suggest that the regime-dependent approaches better capture the upper-tail behaviour of the observed wind speeds. This is confirmed in Figure 13, where the skill score of the twCRPS is displayed for the two regime-dependent approaches relative to the truncated logistic EMOS forecasts. The twCRPS in Figure 13 is calculated using a threshold equal to 8 m/s, roughly corresponding to the 90th percentile of the wind speed

| WS | B | SE | AE | EV | RI |
|----|---|----|----|----|----|
| z  | y | \(\bar{x} - y\) | \((\bar{x} - y)^2\) | \(\bar{x} - y\) | s^2 | \(\sum_{m=1}^{M} |\rho_m - \frac{1}{M+1}|\) |
| SRCC | 0.503 | 0.321 | 0.305 | 0.243 | 0.305 | 0.072 |

Note: Results are shown at a lead time of 5 days. The measures considered are the variations in the following quantities arising because of changes in the regime (Equation 13): the wind speed observations (WS), the bias of the ensemble mean forecast (B), the squared error of the ensemble mean forecast (SE), the absolute error of the ensemble mean forecast (AE), the ensemble variance (EV) and the reliability index of the raw ensemble forecasts (RI).
FIGURE 11 Top: CRPS for the hybrid approach (solid) and the truncated logistic EMOS forecasts (dashed) corresponding to each regime. Bottom: CRPSS for the hybrid approach relative to the truncated logistic EMOS forecasts corresponding to each regime. 95% confidence intervals for the overall skill score are shown in grey. Skill scores above the dotted line at zero indicate positive skill relative to the baseline forecasts [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE 12 Same as in Figure 11 but for one location on the south-west coast of Wales [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE 13 Threshold-weighted CRPSS for the regime-dependent mixture model (MM), as well as the hybrid approach, relative to the conventional truncated logistic EMOS forecasts. Error bars indicate 95% confidence intervals for the skill score, obtained via non-parametric bootstrap resampling. The threshold used in Equation (3) is 8 m·s⁻¹, roughly corresponding to the 90th percentile of the marginal distribution of the observed wind speeds [Colour figure can be viewed at wileyonlinelibrary.com]

observations across all locations. The improvements now reach almost 3% when averaged over all locations and all regimes at a lead time of 36 hr, indicating that the upper tail of the forecast distribution displays a more pronounced improvement than that observed when considering the entire predictive distribution. Similar results are also found using thresholds of 10 and 12 m·s⁻¹, which are respectively close to the 95th and 98th percentiles of the distribution of observations.

5 | DISCUSSION

This work investigates how weather types can be used to calibrate ensembles of weather forecasts, focusing in particular on how such approaches can be applied in an operational setting. A mixture model approach to include regime information into statistical post-processing methods has previously been proposed (Allen et al., 2019, 2020), though these studies have utilised large sets of data, which are not always readily available to forecasting centres. To circumvent this lack of data, this work combines a conventional ensemble model output statistics approach with a regime-dependent method, producing a prediction system that can adapt to issue the most relevant forecast given the current circumstances.

For example, always implementing a regime-dependent mixture model is found here to outperform
a contemporary post-processing method when averaged over all locations, though at several locations at which the regime dependency is small, the mixture model results in a less accurate forecast. The reason for this is that the small benefits that would be obtained by regime-dependent post-processing at these locations are outweighed by the additional parameter uncertainty induced by the more complex approach. To alleviate this issue, the study presented here has looked at implementing regime-dependent post-processing models only when it is expected to be beneficial. A measure of regime dependency is calculated over the training data, and a regime-based approach is applied only if this measure exceeds a certain threshold. If not, then standard post-processing is implemented.

This hybrid approach offers consistent improvements over a conventional truncated logistic-based EMOS approach, with the forecast accuracy at particular locations, measured using the continuous ranked probability score, improving by over 5%. These locations tend to be stations on the west coast of the UK and Ireland, which are heavily affected by the movement of air masses across the Atlantic Ocean. This improvement becomes larger yet when considering forecasts linked to particular regimes. The regimes considered here are the opposite phases of the North Atlantic Oscillation, along with a third group corresponding to when neither of these NAO regimes occur. If the positive phase of the NAO, generally associated with wind storms and more extreme wind speeds, is predicted, then the regime-based methods yield forecasts that are 3% more accurate than those generated using the conventional EMOS approach when averaged over all locations, with improvements at individual stations reaching 20%. Similar results are also observed when the threshold-weighted CRPS is used to assess the upper tail of the forecast distributions, demonstrating that regime information can benefit predictions of high-impact wind speed events, which, despite their obvious significance, are commonly overlooked when implementing conventional post-processing methods (Williams et al., 2014; Friederichs et al., 2018; Pantillon et al., 2018). This result also has implications for multivariate post-processing approaches, since compound weather events – combinations of multiple weather hazards – are often associated with the occurrence of certain atmospheric regimes (Zscheischler et al., 2020).

The NAO regimes considered in this study are utilised operationally in the Met Office’s Decider tool, described in detail in Neal et al. (2016). As discussed in Section 2.2, this product in fact consists of eight weather regimes, which are themselves constructed by clustering together a further 30 weather patterns. However, the regime-dependent mixture model estimates a separate predictive distribution corresponding to each regime. Therefore, for the relatively small amount of data used here, post-processing with the set of eight regimes induces parameter uncertainty that outweighs the benefits of including these additional regimes. As a result, if the mixture model of Section 3.3 is applied using the eight regimes, then the resulting forecasts are found to perform worse than those discussed here (not shown).

Alternatively, more parsimonious approaches could be applied to add the regime information into post-processing that do not involve specifying an entirely new forecast distribution for each regime. Since the focus here is on the North Atlantic Oscillation, one example is applied (not shown) that employs a NAO index as an additional predictor in the EMOS model (along with an interaction with the ensemble mean). Such an approach requires fewer model parameters and is thus yet more suitable for use with limited data sets, but is found to be less informative than the mixture model described herein. Instead, data-driven post-processing methods may be an effective way to model more complex relationships between the weather regimes and the existing predictors (e.g., Rasp and Lerch, 2018). For example, the influence that weather regimes assert on the biases of ensemble prediction systems may itself depend on the time of the year, and regime-dependent post-processing methods should be capable of addressing this. We anticipate that this would be more prevalent for weather variables that exhibit a particularly pronounced seasonal cycle, such as temperature, than for wind speed or precipitation, though more data-rich studies may wish to investigate this. The seasonal dependence of the weather regimes could also be incorporated into post-processing using mixture model-based approaches by utilising regimes that are more closely linked to particular seasons (Grams et al., 2017), thereby assuming that seasonality in the weather variable of interest can be attributed to the occurrence of particular atmospheric regimes (Scheuerer, 2014).

Nonetheless, in this study, it is found that wind speed forecasts benefit more from larger training data sets than from the inclusion of seasonal information. Hence, the regime-dependent approach proposed here uses weather regimes as a basis from which to select the training data for the post-processing method, in place of more contemporary approaches that rely on the season. It is demonstrated that operational weather forecasts may exhibit biases that depend on the prevailing weather regime, and hence it is necessary to investigate the presence of conditional biases prior to post-processing. If such biases are identified, as is the case for the MOGREPS-G prediction system considered here, then regime-dependent approaches are necessary to remove them. As such, work is now ongoing to integrate the regime-dependent hybrid approach.
proposed here into IMPROVER (Evans et al., 2020), a library of algorithms implemented by the Met Office that utilise Rose and Cylc suites (Oliver et al., 2018; Oliver et al., 2019) to post-process and verify weather forecasts (https://github.com/metoppv/improver).

ACKNOWLEDGEMENTS
Sam Allen was supported during this work by a NERC Industrial CASE studentship under grant reference NE/N008693/1. Comments from Ken Mylne, Sebastian Lerch and an anonymous reviewer have significantly improved the quality of this work, as have discussions with Rob Neal concerning the Met Office’s weather regimes. The authors are grateful also to Ric Crocker for maintaining an archive of the regimes forecast by the Met Office’s global deterministic model.

ORCID
Sam Allen https://orcid.org/0000-0003-1971-8277
Gavin R. Evans https://orcid.org/0000-0002-5935-3965
Frank Kwasiak https://orcid.org/0000-0003-1421-4010

REFERENCES
Allen, S., Ferro, C.A.T. and Kwasniok, F. (2019) Regime-dependent statistical post-processing of ensemble forecasts. Quarterly Journal of the Royal Meteorological Society, 145, 3535–3552.
Allen, S., Ferro, C.A.T. and Kwasniok, F. (2020) Recalibrating wind speed forecasts using regime-dependent ensemble model output statistics. Quarterly Journal of the Royal Meteorological Society, 146, 2576–2596.
Baran, S. and Lerch, S. (2015) Log-normal distribution based ensemble model output statistics models for probabilistic wind-speed forecasting. Quarterly Journal of the Royal Meteorological Society, 141, 2289–2299.
Baran, S. and Lerch, S. (2016) Mixture EMOS model for calibrating ensemble forecasts of wind speed. Environmetrics, 27, 116–130.
Baran, S. and Lerch, S. (2018) Combining predictive distributions for the statistical post-processing of ensemble forecasts. International Journal of Forecasting, 34, 477–496.
Brier, G.W. (1950) Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78, 1–3.
Charney, J.G. and DeVore, J.G. (1979) Multiple flow equilibria in the atmosphere and blocking. Journal of the Atmospheric Sciences, 36, 1205–1216.
Dawid, A.P. (1984) Present position and potential developments: some personal views statistical theory the frequentist approach. Journal of the Royal Statistical Society: Series A (General), 147, 278–290.
Delle Monache, L., Hacker, J.P., Zhou, Y., Deng, X. and Stull, R.B. (2006) Probabilistic aspects of meteorological and ozone regional ensemble forecasts. Journal of Geophysical Research, 111(D24), http://dx.doi.org/10.1029/2005jd006917.
Evans, G.R., Ayliffe, B.A., Sandford, C., Rust, F.M., Jones, C., Fitzpatrick, B.J.R., Moseley, S.R., Hopkinson, A.R., Pillinger, T., Baker, M., Crosswaite, N., Howard, K., Beard, L., Worsfold, M., Abernethy, P., Trzeciak, T.M., Gale, T., Jackson, S.D., Booton, A., Smith, E., Kinoshita, B.P., Sampson, C., Hume, T., Canvin, J., Friedrich, M., Roberts, N.M., Wright, B.J. and Nic Ghulidhir, M. (2020). metoppv/improver: IMPROVER: a library of algorithms for meteorological post-processing. (version 0.10.0). https://doi.org/10.5281/zenodo.4017773.
Feldmann, K., Richardson, D.S. and Gneiting, T. (2019) Grid-versus station-based postprocessing of ensemble temperature forecasts. Geophysical Research Letters, 46, 7744–7751.
Franzke, C., Woolings, T. and Martius, O. (2011) Persistent circulation regimes and preferred regime transitions in the North Atlantic. Journal of the Atmospheric Sciences, 68, 2809–2825.
Friederichs, P., Wahl, S. and Buschow, S. (2018). Postprocessing for extreme events, Vannitsem, S., Wilks, D.S. & Messner, J., Statistical Postprocessing of Ensemble Forecasts (pp. 127–154): Amsterdam, Elsevier.
Gneiting, T., Balabdaoui, F. and Raftery, A.E. (2007) Probabilistic forecasts, calibration and sharpness. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69, 243–268.
Gneiting, T. and Raftery, A.E. (2007) Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102, 359–378.
Gneiting, T., Raftery, A.E., Westveld, I. and Goldman, T. (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Monthly Weather Review, 133, 1098–1118.
Gneiting, T. and Ranjan, R. (2011) Comparing density forecasts using threshold- and quantile-weighted scoring rules. Journal of Business and Economic Statistics, 29, 411–422.
Gneiting, T. and Ranjan, R. (2013) Combining predictive distributions. Electronic Journal of Statistics, 7, 1747–1782.
Grams, C.M., Beerli, R., Pfenninger, S., Staffell, I. and Wernli, H. (2017) Balancing Europe’s wind-power output through spatial deployment informed by weather regimes. Nature Climate Change, 7, 557–562.
Hamill, T.M. (2018). Practical aspects of statistical postprocessing., Vannitsem, S., Wilks, D.S. & Messner, J., Statistical Postprocessing of Ensemble Forecasts, pp. 187–217: Amsterdam: Elsevier.
Hamill, T.M. and Colucci, S.J. (1997) Verification ofEta–RSM short-range ensemble forecasts. Monthly Weather Review, 125, 1312–1327.
Hannachi, A., Straus, D.M., Franzke, C.L., Corti, S. and Woolings, T. (2017) Low-frequency nonlinearity and regime behavior in the northern hemisphere extratropical atmosphere. Reviews of Geophysics, 55, 199–234.
Holzmann, H. and Klar, B. (2017) Focusing on regions of interest in forecast evaluation. The Annals of Applied Statistics, 11, 2404–2431.
Hurrell, J.W. (1995) Decadal trends in the North Atlantic Oscillation: regional temperatures and precipitation. Science, 269, 676–679.
Hurrell, J.W. and Deser, C. (2009) North Atlantic climate variability: the role of the North Atlantic Oscillation. Journal of Marine Systems, 79, 231–244.
Jordan, A., Krüger, F. and Lerch, S. (2019) Evaluating probabilistic forecasts with scoringRules. Journal of Statistical Software, 90(12), http://dx.doi.org/10.18637/jss.v090.i12.
Lang, M.N., Lerch, S., Mayr, G.J., Simon, T., Stauffer, R. and Zeileis, A. (2019) Remember the past: a comparison of time-adaptive training schemes for non-homogeneous regression. Nonlinear Processes in Geophysics, 27, 23–34.
Lerch, S. and Thorarinsdottir, T.L. (2013) Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. Tellus A: Dynamic Meteorology and Oceanography, 65, 21206.

Matheson, J.E. and Winkler, R.L. (1976) Scoring rules for continuous probability distributions. Management Science, 22, 1087–1096.

Messner, J.W., Mayr, G.J., Wilks, D.S. and Zeileis, A. (2014) Extending extended logistic regression: extended versus separate versus ordered versus censored. Monthly Weather Review, 142, 3003–3014.

Messner, J.W., Mayr, G.J. and Zeileis, A. (2017) Nonhomogeneous boosting for predictor selection in ensemble postprocessing. Monthly Weather Review, 145, 137–147.

Neal, R., Fereday, D., Crocker, R. and Comer, R.E. (2016) A flexible approach to defining weather patterns and their application in weather forecasting over Europe. Meteorological Applications, 23, 389–400.

Oliver, H., Shin, M., Matthews, D., Sanders, O., Bartholomew, S., Clark, A., Fitzpatrick, B., van Haren, R., Hut, R. and Drost, N. (2019) Workflow automation for cycling systems: the Cycl workflow engine. Computing in Science and Engineering.

Oliver, H.J., Shin, M. and Sanders, O. (2018) Cylc: a workflow engine for cycling systems. Journal of Open Source Software, 3, 737.

Pantillon, F., Lerch, S., Knippertz, P. and Corsmeier, U. (2018) Forecasting wind gusts in winter storms using a calibrated convection-permitting ensemble. Quarterly Journal of the Royal Meteorological Society, 144, 1864–1881.

Philipp, A., Della-Marta, P.-M., Jacobit, J., Fereday, D.R., Jones, P.D., Moberg, A. and Wanner, H. (2007) Long-term variability of daily North Atlantic–European pressure patterns since 1850 classified by simulated annealing clustering. Journal of Climate, 20, 4065–4095.

Pinson, P. and Girard, R. (2012) Evaluating the quality of scenarios of short-term wind power generation. Applied Energy, 96, 12–20.

Porson, A.N., Carr, J.M., Hagelin, S., Darvell, R., North, R., Walters, D., Mylne, K.R., Mittermaier, M.P., Willington, S. and Macpherson, B. (2020) Recent upgrades to the Met Office convective-scale ensemble: an hourly time-lagged 5-day ensemble. Quarterly Journal of the Royal Meteorological Society.

Raftery, A.E., Gneiting, T., Baladadaoui, F. and Polakowski, M. (2005) Using Bayesian model averaging to calibrate forecast ensembles. Monthly Weather Review, 133, 1155–1174.

Rasp, S. and Lerch, S. (2018) Neural networks for postprocessing ensemble weather forecasts. Monthly Weather Review, 146, 3885–3900.

Rodwell, M.J., Richardson, D.S., Parsons, D.B. and Wernli, H. (2018) Flow-dependent reliability: a path to more skillful ensemble forecasts. Bulletin of the American Meteorological Society, 99, 1015–1026.

Scheuerer, M. (2014) Probabilistic quantitative precipitation forecasting using ensemble model output statistics. Quarterly Journal of the Royal Meteorological Society, 140, 1086–1096.

Scheuerer, M. and Möller, D. (2015) Probabilistic wind speed forecasting on a grid based on ensemble model output statistics. The Annals of Applied Statistics, 9, 1328–1349.

Slaughter, J.M., Gneiting, T. and Raftery, A.E. (2010) Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. Journal of the American Statistical Association, 105, 25–35.

Taillardat, M., Mestre, O., Zamo, M. and Naveau, P. (2016) Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. Monthly Weather Review, 144, 2375–2393.

Thompson, P.D. (1957) Uncertainty of initial state as a factor in the predictability of large scale atmospheric flow patterns. Tellus, 9, 275–295.

Thorarinsdottir, T.L. and Gneiting, T. (2010) Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression. Journal of the Royal Statistical Society: Series A (Statistics in Society), 173, 371–388.

Walters, D., Brooks, M., Boutle, I., Melvin, T., Stratton, R., Vosper, S., Wells, H., Williams, K., Wood, N., Allen, T., Bushell, A., Copsey, D., Earnshaw, P., Edwards, J., Gross, M., Hardiman, S., Harris, C., Heming, J., Klingeman, N., Levine, R., Manners, J., Martin, G., Milton, S., Mittermaier, M., Morcrette, C., Riddick, T., Roberts, M., Sanchez, C., Selwood, P., Stirling, A., Smith, C., Suri, D., Tennant, W., Vidale, P.L., Wilkinson, J., Willett, M., Woolnough, S. and Xavier, P. (2017) The Met Office unified model global atmosphere 6.0/6.1 and JULES global land 6.0/6.1 configurations. Geoscientific Model Development, 10, 1487–1520.

Wilks, D.S. (2002) Smoothing forecast ensembles with fitted probability distributions. Quarterly Journal of the Royal Meteorological Society, 128, 2821–2836.

Wilks, D.S. (2019) Statistical Methods in the Atmospheric Sciences. Amsterdam: Elsevier.

Williams, R., Ferro, C.A.T. and Kwarsnik, F. (2014) A comparison of ensemble post-processing methods for extreme events. Quarterly Journal of the Royal Meteorological Society, 140, 1112–1120.

Zscheischler, J., Martius, O., Westra, S., Bevacqua, E., Raymond, C., Horton, R.M., van den Hurk, B., AghaKouchak, A., Jézéquel, A., Mahecha, M.D., Maraun, D., Ramos, A.M., Ridder, N.N., Thiery, W. and Vignotto, E. (2020) A typology of compound weather and climate events. Nature Reviews Earth and Environment, 1, 333–347.

How to cite this article: Allen S, Evans GR, Buchanan P, Kwarsnik F. Incorporating the North Atlantic Oscillation into the post-processing of MOGREPS-G wind speed forecasts. Q J R Meteorol Soc. 2021;147:1403–1418. https://doi.org/10.1002/qj.3983