Received: 12 February 2020 | Revised: 9 April 2020 | Accepted: 14 April 2020 | Published on: 13 May 2020

DOI: 10.1002/qj.3806

Quarterly Journal of the Royal Meteorological Society RMetS

RESEARCH ARTICLE

Recalibrating wind-speed forecasts using regime-dependent ensemble model output statistics

S. Allen | C. A. T. Ferro | F. Kwasniok

Department of Mathematics, University of Exeter, Exeter, UK

Correspondence
S. Allen, University of Exeter, Laver Building, North Park Road, Exeter, EX4 4QE, UK.
Email: sa495@exeter.ac.uk

Funding information
Natural Environment Research Council, Grant/Award Number: NE/N008693/1

Abstract

Raw output from deterministic numerical weather prediction models is typically subject to systematic biases. Although ensemble forecasts provide invaluable information regarding the uncertainty in a prediction, they themselves often misrepresent the weather that occurs. Given their widespread use, the need for high-quality wind-speed forecasts is well-documented. Several statistical approaches have therefore been proposed to recalibrate ensembles of wind-speed forecasts, including a heteroscedastic truncated regression approach. An extension to this method that utilises the prevailing atmospheric flow is implemented here in a quasigeostrophic simulation study and on Global Ensemble Forecasting System (GEFS) reforecast data, in the hope of alleviating errors owing to changes in the synoptic-scale atmospheric state. When the wind speed depends strongly on the underlying weather regime, the resulting forecasts have the potential to provide substantial improvements in skill relative to conventional post-processing techniques. This is particularly pertinent at longer lead times, where there is more improvement to be gained over current methods, and in weather regimes associated with wind speeds that differ greatly from climatology. In order to realise this potential, an accurate prediction of the future atmospheric regime is required.

KEYWORDS

probabilistic weather forecasting, statistical post-processing, weather regimes, wind

1 | INTRODUCTION

Numerical weather prediction (NWP) models aim to replicate the physical laws governing the atmosphere's trajectory. Due to the chaotic nature of the atmosphere, these models rely on a perfect formulation of their initial state. Since this is impossible to obtain in practice, the uncertainty in the model's initial conditions should be described using probability. The evolution of this probability distribution over time then provides information regarding the uncertainty of the atmosphere's future state (Epstein, 1969). Ensemble forecasts act as a Monte Carlo approximation to this, comprising multiple model runs from conditions sampled from the initial probability distribution, with the resulting forecasts assumed to be random draws from the distribution of the predictand (Leith, 1974).

This assumption is subject to two errors in particular: error when specifying the distribution for the initial conditions, and error due to the NWP model not fully capturing the small-scale effects and interactions that are present in the atmosphere. The ensemble forecasts therefore exhibit biases, typically being overconfident about their prediction. Statistical post-processing techniques have become an essential component of weather forecasts over the last couple of decades, due to their ability to correct for such errors (Vannitsem *et al.*, 2018).

Statistical post-processing refers broadly to any statistical procedure that is applied to forecasts after having obtained the numerical model output, examples of which include statistical downscaling and bias correction. It is used here to refer to forecast recalibration, a particular branch of post-processing that concerns issuing and calibrating probabilistic predictions, typically by specifying a parametric distribution for the response variable that depends on the raw ensemble forecast. The most prominent recalibration approaches are variations of Bayesian model averaging (BMA: Raftery *et al.*, 2005) and ensemble model output statistics (EMOS), also known as nonhomogeneous regression (Gneiting *et al.*, 2005).

High-quality forecasts of wind speed are particularly valuable, due to their application in decision-making in areas such as transportation, insurance, and renewable energy production. Therefore, several post-processing methods have been proposed to deal with systematic errors present in wind-speed forecasts. These include quantile regression (Bremnes, 2004, 2019), BMA using gamma component distributions (Sloughter *et al.*, 2010; Eide *et al.*, 2017), and implementations of EMOS with various choices of parametric family: truncated normal (Thorarinsdottir and Gneiting, 2010), gamma (Scheuerer *et al.*, 2015), and censored or truncated logistic distributions (Messner *et al.*, 2014; Scheuerer *et al.*, 2015), for example. Lerch and Thorarinsdottir (2013) and Baran and Lerch (2015) introduce regime-switching approaches that use linear combinations of predictive distributions to improve the upper tail of wind-speed forecasts, and more flexible combinations of EMOS models have also been found to outperform the component forecasts (Baran and Lerch, 2016, 2018).

Since the aim of forecast recalibration is to alleviate systematic biases in the dynamical model output, it is common to use only the ensemble forecast of the predictand as an input variable. Recently, however, techniques have been proposed that utilise more predictors, highlighting the potentially useful information that can be gained from other sources. Scheuerer (2014) and Scheuerer and Hamill (2015), for example, exploit predictions at neighbouring grid points when recalibrating precipitation forecasts, while Eide *et al.* (2017) employ wind direction as an additional predictor for wind speed. More data-driven approaches have also been proposed that can deal with a large set of possible inputs, and automatically select those most relevant for post-processing (Taillardat *et al.*, 2016; Messner *et al.*, 2017; Rasp and Lerch, 2018).

The underlying reason for adding predictors is that the additional variables provide helpful indications as to when the relationship between the forecast and the observation might vary. It may be the case that forecast accuracy is affected by the weather situation at hand. Weather forecasters often adjust their predictions depending on the prevailing large-scale flow (Roebber, 1998), and incorporating the flow directly into forecast recalibration methods serves as a way of automating this procedure. Synoptic-scale patterns in the atmosphere's circulation can also explain relationships between certain weather variables and locations. Integrating the circulation into post-processing therefore allows information from alternative variables to be utilised, without including them directly in the calibration.

Numerous studies have investigated how recurring weather patterns influence model biases in synoptic-scale forecasts, finding that errors are dependent on the underlying weather regime (Koch *et al.*, 1985; O'Lenic and Livezey, 1989; Stoss and Mullen, 1995). However, in comparison, limited work has examined how forecasts of smaller-scale variables rely on the flow. Lorenz (1969) remarks that the low-frequency circulation has a much larger range of predictability than shorter-scale flows, and hence predictions of the large-scale information could be used to assist forecasts of high-frequency, noisy events, such as the weather.

Atmospheric regimes are synoptic-scale patterns in the circulation that persist and recur at the same locations, defined dynamically as metastable equilibria in the atmosphere's phase space (Franzke *et al.*, 2008). They have been found to account for a large fraction of the atmosphere's low-frequency, or intraseasonal, variability, and commonplace statistical techniques are frequently used to estimate their occurrence (Horel, 1985; Cheng and Wallace, 1993; Kimoto and Ghil, 1993; Smyth *et al.*, 1999; Majda *et al.*, 2006). Scheuerer and Büermann (2014) suggested that these regimes could provide a suitable basis by which to train post-processing methods, and they have previously been used to calculate ensemble member weights in a consensus forecast (Greybush *et al.*, 2008), to calibrate and blend short-range precipitation forecasts (Kober *et al.*, 2014), and to extend probabilistic post-processing methods (Allen *et al.*, 2019).

The atmosphere's circulation is intimately connected to the Earth's winds and therefore forecasts of wind speed might be susceptible to improvements if this regime information were incorporated into the post-processing. Allen *et al.* (2019) introduced regime-dependent statistical post-processing, proposing that, if statistical techniques

can specify a probability model for the regime, then current post-processing methods can be conditioned on the underlying weather regime:

$$p(y|\boldsymbol{x}) = \sum_{r=1}^{R} p(y|\boldsymbol{x}, r)p(r), \quad (1)$$

where $p(r)$ is the probability of residing in regime r and $p(y|\boldsymbol{x})$ is the conditional distribution of the predictand y given the ensemble forecast $\boldsymbol{x} = (x_1, \ldots, x_M)$. The forecast in this case takes the form of a predictive distribution, also referred to as the forecast distribution. The probabilistic forecast is composed of predictive distributions, $p(y|\boldsymbol{x}, r)$, that depend on the prevailing weather regime and therefore, rather than specifying just one forecast distribution, a separate distribution must be specified for each regime.

Siegert *et al.* (2016) utilise the joint distribution of the verification and the forecast, arguing that the ensemble members should be treated as random quantities rather than known constants. This approach could similarly be extended, accounting for the fact that ensemble members may arise from different distributions depending on the atmospheric state:

$$p(y, \boldsymbol{x}) = \sum_{r=1}^{R} p(y, \boldsymbol{x}|r)p(r) = \sum_{r=1}^{R} p(y|\boldsymbol{x}, r)p(\boldsymbol{x}|r)p(r). \quad (2)$$

Improvements would then be expected if the forecast distribution, $p(y|\boldsymbol{x}, r)$, or the ensemble distribution, $p(\boldsymbol{x}|r)$, were to change between the regimes.

Allen *et al.* (2019) present a motivating example in which regime-dependent post-processing greatly outperforms current approaches when applied to simulated data from a highly idealised atmospheric model. Hence, the aim of this article is to investigate how the methods perform in more realistic settings. The post-processing framework is presented in the following section. A truncated normal (TN) EMOS approach is applied to wind-speed forecasts, along with extensions suitable for the regime paradigm.

The method is first implemented in a three-layer quasi-geostrophic (QG) model of the Northern Hemisphere in Section 3. The QG model used here is sufficiently realistic that it is capable of generating atmospheric patterns that are present in climate reanalyses, but simple enough that a large amount of data can be simulated, allowing an extensive investigation of regime-dependent approaches.

In Section 4, the same approach is trialled on retrospective wind-speed forecasts over the Euro-Atlantic region, taken from the National Centers for Environmental Prediction (NECP) Global Ensemble Forecasting System (GEFS: Hamill *et al.*, 2013). The GEFS reforecasts are generated from a higher resolution model than that used in the QG setting, yet still provide sufficient data with which to construct and assess regime-dependent forecast distributions reliably. Section 5 concludes and discusses the results.

2 | METHODOLOGY

2.1 | Statistical post-processing

To capture the relationship between the model and the atmosphere, statistical post-processing relies on a set of historical forecasts and observations, from which parameters can be estimated. This training set consists of pairs of data $(\boldsymbol{x}, y)$, where $\boldsymbol{x} = (x_1, \ldots, x_M)$ denotes an ensemble forecast comprised of M members, and y is the corresponding verification. Regime-dependent post-processing methods extend this, such that the training data pairs become triples of the form $(\boldsymbol{x}, y, \rho)$, where ρ represents some information regarding the atmospheric flow associated with that forecast and observation. This could be one weather regime, the probabilities of residing in each identified regime, or a continuous measure of the atmospheric flow, for example. Post-processing can then utilise this additional information. Allen *et al.* (2019) discuss ways of including the flow in forecast recalibration, arguing that partitioning the phase space into a discrete number of regimes can allow for more flexible forecast distributions.

Thorarinsdottir and Gneiting (2010) introduce an EMOS approach that extends truncated regression models to include a nonconstant variance term. The method suggests that, given an ensemble forecast, the observed wind speed is a random variable y that follows a normal distribution truncated below at zero:

$$y|\boldsymbol{x} \sim N_0(\alpha + \beta\bar{x}, \gamma + \delta s^2). \quad (3)$$

The location and spread of the distribution are then linear functions of the ensemble mean $\bar{x}$ and ensemble variance s^2, respectively. This method was found here to outperform implementing BMA with gamma component distributions, and the resulting predictions were also of similar skill to the combination forecast approach of Baran and Lerch (2016).

We employ Equation 3 in the regime-dependent framework by using a mixture of truncated normal forecast distributions that depend on the coinciding weather regime:

$$y|\boldsymbol{x} \sim \sum_{r=1}^{R} w_r N_0(\alpha_r + \beta_r\bar{x}, \gamma_r + \delta_r s^2), \quad (4)$$

where w_r represents the probability of the atmosphere residing in regime r at the forecast validation time. This method involves estimating post-processing parameters $(\alpha, \beta, \gamma, \delta)$ for each of the regimes.

2.2 | Mixture-model weights

It is also necessary to define the mixture-model weights, w_r. The motivation for regime-dependent approaches assumes that there are differences in model biases that depend on the prevailing weather regime. Since the weather that materialises is dependent on the atmospheric state, the weights should provide probabilities that the atmosphere will reside in each regime at the forecast validation time. The weights can thus be thought of as predictions of the future atmospheric state.

Three choices for the weight are compared here. A first choice defines the weights by a persistence forecast: if s is the regime present at the forecast initialisation time, then $w_r = 1$ when $r = s$ and $w_r = 0$ when $r \neq s$. These will be called "initial regime" weights. The disadvantage of this approach is that, as the forecast horizon increases, so does the probability of transitioning to another regime. The initial regime would thus not be representative of the atmospheric conditions at the validation time. When the forecast lead time is long relative to the regime persistence times, model biases would not be expected to vary depending on the initial regime and hence the regime-dependent mixture model would revert back to the conventional truncated normal distribution in Equation 3, offering little improvement despite the added flexibility.

Since it is possible to determine which regime actually materialised for forecasts in the training data, a second choice is to find conditional probabilities of each regime occurring given the initial regime. That is, given a certain regime occurs at the initialisation time, one can calculate the proportion of instances in which each regime materialises at the validation time. An example of this for the GEFS reforecast data in Section 4 is shown in Table 1. The data and identified regimes are described in detail in Section 4.1. The initial regime weights assume that the probability of European blocking (EB) occurring after 2 days, given that it transpired at the initialisation time, is one, for example, whereas Table 1 suggests it is only 0.657 in reality. This, in theory, provides a more realistic probability of the regime at the forecast validation time. Such weights are called "conditional regime weights".

The ensemble members are themselves simulated trajectories of the atmosphere, and hence regimes can also be estimated from each ensemble member. The proportion of ensemble members that are assigned to a regime thus constitutes a probability of residing in that state at the forecast validation time. This third choice is called "ensemble regime weights". Allen *et al.* (2019) find that post-processing using ensemble regime weights outperforms the initial regime weights.

Since the weights define a probabilistic forecast of the future atmospheric state, they can be assessed by their ability to capture the regime that materialises. Figure 1 shows the Brier skill score (Brier, 1950) for the three different choices of weight, averaged across the four regimes identified in the reforecast data. The climatological frequencies of the different regimes are used as a reference forecast.

Although useful at very short lead times, initial regime weights become detrimental to forecasts relative to climatology after only a few days. Unsurprisingly, this scheme is particularly poor at predicting less persistent weather types. Conditional regime weights, on the other hand, are designed to be at least as good as climatology and hence always result in a positive skill score. However, they rely on information from the initial regime, and the skill score therefore tends to zero as the lead time increases. Output from the NWP model will also contain progressively less information as the forecast horizon increases, with studies highlighting model deficiencies in capturing the onset and decay of atmospheric blocking events (Tibaldi and Molteni, 1990). Nonetheless, the weights defined by the ensemble members offer considerably more skill than alternative approaches at all lead times considered here.

TABLE 1 Matrix of conditional probabilities of each regime occurring 48 hrs after a given initial regime

		True Reg.			
		NAO+	**NAO−**	**AR**	**EB**
Init. Reg.	NAO+	0.740	0.070	0.094	0.095
	NAO−	0.105	0.664	0.147	0.083
	AR	0.163	0.137	0.565	0.134
	EB	0.102	0.129	0.113	0.657

Note: As lead time increases, every row tends to the climatological frequencies of the regimes.

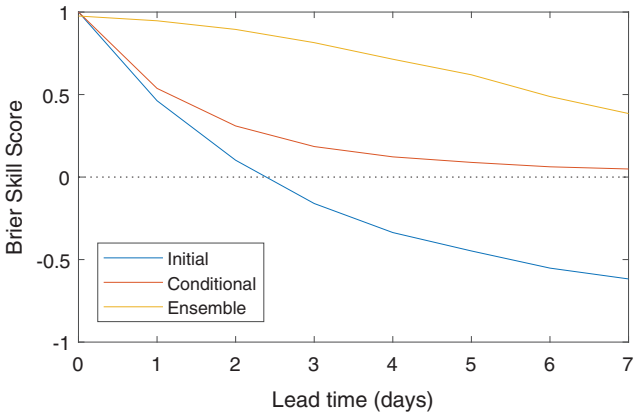

FIGURE 1 Brier skill score relative to climatology for different forecasts of the future weather regime

In a study such as this, where predictions are evaluated over a set of hindcasts, forecasts can be conditioned on perfect knowledge of the regime at the validation time. Although this information is not available a priori to forecasters, it is implemented here to obtain a rough upper bound on the improvements gained from regime-dependent post-processing. This is henceforth referred to as the "true regime". Furthermore, although conditional regime weights are found to offer better forecasts of the atmospheric state, the results are found to be similar to using the initial regime. In the subsequent analysis, results are therefore compared only for the initial regime, ensemble regime, and true regime weights.

Regime information is incorporated into post-processing via these mixture-model weights. Therefore, in order to obtain forecast distributions that utilise the regime information, the weights are estimated first, prior to fitting the regime-specific predictive distributions. Coefficients for these component distributions are then estimated conditional on the regime weights. Furthermore, since the regime-dependent weights considered here are functions of the atmospheric flow, rather than constant parameters, they can adapt to the current atmospheric conditions. This allows forecasts to be post-processed differently from one another, even when trained using the same data.

This is in contrast to alternative approaches that have been introduced to combine predictive distributions (Gneiting *et al.*, 2013). Baran and Lerch (2016), for example, estimate the mixture-model weights simultaneously to the post-processing parameters. Although the resulting wind-speed forecasts are found to exhibit significantly better calibration than the individual components, the corresponding parameter estimation step can result in optimisation problems that are complex and unstable, and thus computationally expensive. Baran and Lerch (2018) therefore investigate the use of forecast combination approaches that use a two-step procedure to estimate model parameters. The approaches discussed therein first fit two or more distinct EMOS models individually to all training data, and then find the optimal weights to combine the resulting forecast distributions. The method presented in this study similarly divides the parameter estimation into two stages, but distributional coefficients are instead estimated after having obtained the mixture-model weights. Doing so allows the component distributions to capture separate features of the training data that arise due to the occurrence of each weather regime.

2.3 | Parameter estimation

Gneiting and Raftery (2007) introduce the notion of optimum score estimation, which identifies the parameter values that optimise a proper score over the available training data. Maximum-likelihood estimation fits into this framework, since it is analogous to minimising the negative log-likelihood (NLL), or logarithmic, score. Another popular score in the forecasting literature is the continuous ranked probability score (CRPS), defined as

$$\mathrm{crps}(F, y) = \int_{-\infty}^{\infty} (F(u) - \mathbb{1}\{u \geq y\})^2 \, du, \tag{5}$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function, F is the forecast distribution, and y the verification (Matheson and Winkler, 1976).

However, Baran and Lerch (2016) note that the CRPS for mixture models such as that in Equation 4 cannot be evaluated analytically and hence must be calculated numerically. As a result, parameter estimation with the CRPS becomes computationally expensive. Although minimum CRPS estimation is generally regarded as a more robust choice than maximum likelihood for forecast recalibration, Gebetsberger *et al.* (2018) suggest that the estimators should generate similar results, provided the distributional assumptions are valid. Maximum likelihood is therefore chosen to estimate parameters in this study.

The density of a TN distribution $N_0(\mu, \sigma^2)$ is

$$f(y) = \frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right)\left[\Phi\left(\frac{\mu}{\sigma}\right)\right]^{-1}, \tag{6}$$

where $\phi(\cdot)$ is the probability density function, and $\Phi(\cdot)$ the cumulative distribution function, of the standard normal distribution. The density for a mixture of TN distributions, $\sum_{r=1}^{R} w_r N_0(\mu_r, \sigma_r^2)$, is then simply a weighted sum of the component densities:

$$g(y) = \sum_{r=1}^{R} \frac{w_r}{\sigma_r}\phi\left(\frac{y-\mu_r}{\sigma_r}\right)\left[\Phi\left(\frac{\mu_r}{\sigma_r}\right)\right]^{-1}. \tag{7}$$

The regime-dependent approaches estimate a set of parameters for each of the R identified regimes ($\alpha_r, \beta_r, \gamma_r, \delta_r$ for $r = 1, \ldots, R$) by maximising the likelihood of the mixture model in the training data, conditional on each choice of regime weights.

If the weight takes the form of an indicator function, as is the case for the initial and true regime weights, then the mixture-model forecast at a given time is equivalent to a truncated normal distribution with post-processing parameters that correspond to the predicted regime. The CRPS thus reduces to that for a truncated normal distribution, which is given in closed form in Equation 8. Nonetheless, to retain correspondence between the different methods, all statistical models are fitted using

maximum likelihood. In the case of indicator weights, each forecast–observation pair in the training data is assigned to exactly one regime, and the training data can be partitioned into R mutually exclusive, collectively exhaustive training subsets. Post-processing parameters for the truncated normal distribution associated with a regime are then estimated by maximising the likelihood only over the subset of data containing forecast–observation pairs allocated to that regime.

Regime-dependent methods with indicator weights can therefore also be interpreted as analogue-based post-processing approaches, whereby a training data set is constructed from forecast–observation pairs that are believed to exhibit behaviour similar to the current forecast (Junk *et al.*, 2015). In this case, the assumption is that the forecast biases depend on the synoptic-scale behaviour of the atmosphere, which aligns with the motivation for using regime analogues in Barnes *et al.* (2019).

If the probabilities of residing in each regime are not strictly zero, or one then the training data consist of all available forecasts and observations. In this case, when estimating the post-processing parameters corresponding to a regime, the probability of each historical forecast–observation pair belonging to that regime determines the leverage it has in estimating the coefficients. In this case, all post-processing parameters are estimated simultaneously. Although this can be considerably more time-consuming than parameter estimation for conventional methods, it is not found to be prohibitively expensive.

Thorarinsdottir and Gneiting (2010) find a local EMOS approach, in which forecast recalibration occurs separately for each individual location, to perform better than aggregating training data across several spatial locations. Despite being more computationally demanding, this approach is implemented here, allowing the post-processing to account for local biases.

2.4 | Forecast verification

A forecast distribution is said to be calibrated if events materialise with the same frequency with which they are forecast, while sharpness refers to the concentration of the distribution. Forecasters have come to seek predictive distributions that are sharp subject to being calibrated (Gneiting *et al.*, 2007). The evaluation of forecasts must thus account for these two qualities, something that is achieved through the use of proper scoring rules (Gneiting and Raftery, 2007). The CRPS is used to verify forecasts in the following sections, though similar conclusions are drawn from the NLL score.

Thorarinsdottir and Gneiting (2010) provide the CRPS for a truncated normal predictive distribution in closed form:

$$\begin{aligned}&\mathrm{crps}\left[N_0(\mu,\sigma^2),y\right]\\&=\sigma\left[\Phi\left(\frac{\mu}{\sigma}\right)\right]^{-2}\left\{\frac{y-\mu}{\sigma}\Phi\left(\frac{\mu}{\sigma}\right)\left[2\Phi\left(\frac{y-\mu}{\sigma}\right)+\Phi\left(\frac{\mu}{\sigma}\right)-2\right]\right.\\&\left.+2\phi\left(\frac{y-\mu}{\sigma}\right)\Phi\left(\frac{\mu}{\sigma}\right)-\frac{1}{\sqrt{\pi}}\Phi\left(\frac{\sqrt{2}\mu}{\sigma}\right)\right\}.\end{aligned}\tag{8}$$

Since wind speed is non-negative, Equation 5 for a mixture-model forecast is evaluated using Gauss–Laguerre quadrature.

The CRPS is negatively oriented and hence larger values indicate poorer performance. To compare the ability of the TN and regime-dependent truncated normal (RDTN) frameworks, the continuous ranked probability skill score (CRPSS) is also applied, with the conventional TN approach as the reference, or baseline forecast. The CRPSS is defined as

$$\mathrm{CRPSS}=\frac{\langle\mathrm{crps}(F,y)\rangle-\langle\mathrm{crps}(G,y)\rangle}{\langle\mathrm{crps}(F,y)\rangle}=1-\frac{\langle\mathrm{crps}(G,y)\rangle}{\langle\mathrm{crps}(F,y)\rangle}.\tag{9}$$

F denotes the predictive distribution obtained from TN, G denotes that obtained from RDTN, y is the corresponding verification, and $\langle\mathrm{crps}(\cdot,y)\rangle$ is the average CRPS over forecasts in the test data set (Wilks, 2019). The skill score can be interpreted as the percentage improvement in score upon current post-processing methods, gained from regime-dependent post-processing. Skill scores are bounded above by one, and scores below zero indicate that the RDTN method is performing worse than its reference. Therefore, unlike the CRPS, high values of the CRPSS are desired.

3 | QUASIGEOSTROPHIC MODEL

3.1 | Data

We use here a spectral quasigeostrophic three-level atmospheric model of the Northern Hemisphere, truncated triangularly at wavenumber 21. The levels are located at 250, 500, and 750 hPa. The governing equations are

$$\frac{\partial q_i}{\partial t}+J(\Psi_i,q_i)=D_i+S_i,\qquad i=1,2,3.\tag{10}$$

Here, Ψ_i and q_i are the streamfunction and the potential vorticity at level i, respectively, and J denotes the Jacobian operator on the sphere. The dissipative terms D_i comprise Newtonian temperature relaxation at all levels, Ekman damping at the lowest level, and hyperviscosity

on the time-dependent part of the potential vorticity at all levels. The time-independent but spatially varying forcing terms S_i are diabatic sources of potential vorticity.

The model parameters and forcing are tuned in such a way that the model in a long-term integration exhibits a remarkably realistic mean state and variability pattern of streamfunction and potential vorticity. The model is integrated forward in time using the third-order Adams–Bashforth scheme with a constant time step of 1 hr. The details of the model configuration, parameter setting, parameter tuning procedure, and performance versus reanalysis data can be found in Kwasniok (2007) and Kwasniok (2019). The model configuration used here is exactly the same as described in Kwasniok (2019).

The streamfunction, $\boldsymbol{\Psi}$, represents the trajectory of particles in this model and hence the circulation of the atmosphere in the Northern Hemisphere for each of the vertical levels can be represented instantaneously by the streamfunction in 1024-dimensional space, comprised of grid-point values at 64 equally spaced longitudes and 16 Gaussian latitudes. Regimes are therefore located by searching for quasistationary equilibria in the streamfunction.

The system described above was first integrated forward in time for 50 years and the atmospheric regimes were identified using the resulting time series of daily streamfunction fields. To construct training and test data sets, the QG model was then run for a further 30 years, with both the streamfunction and wind speed at all locations recorded daily. Since this systems acts as a surrogate for the atmosphere, the recorded wind speeds are treated as observations, and the streamfunction field provides a best guess of the atmospheric state at that time. These "observed" states are then used as forecast analyses. An ensemble forecast comprised of ten exchangeable members was constructed by adding random perturbations from a $N(0, 0.00025^2)$ distribution to these analyses, expressed in spherical harmonics, and propagating the resulting initial conditions through time for 7 days, using a version of the quasigeostrophic model truncated at wavenumber 19. Perturbing the analyses reflects uncertainty in the initial forecast state, while a more severely truncated model is used to replicate an imperfect NWP model. The results were not dependent on the ensemble size, and post-processing separately at each location means that perturbations not necessarily being spatially independent should not have an adverse effect on the results.

The resulting data therefore include 30 years worth of daily forecast–observation pairs, for daily lead times up to one week ahead. Half of these data are used to train the post-processing methods and the remaining data are used to assess the resulting predictions. Both the training and test data sets thus consist of 5,475 ensemble forecasts of wind speed and their corresponding observations.

Quasigeostrophic models have previously been employed to investigate the behaviour of planetary-scale flow regimes (Marshall and Molteni, 1993; Majda *et al.*, 2006; Franzke *et al.*, 2008). Kondrashov *et al.* (2004), for example, used a similar model to study transitions between phases of the North Atlantic Oscillation and the Arctic Oscillation (or Northern Annular Mode), two dominant flow regimes in the Northern Hemisphere. One particular feature of the QG model is that it exhibits no seasonal cycle, residing perpetually in winter. This is the season in which the regime behaviour of the atmosphere is most pronounced, and therefore this system has the added benefit that it could produce more robust atmospheric states (Hannachi *et al.*, 2017).

Principal component analysis (PCA) is applied to the gridded streamfunction anomaly values at 500 hPa to reduce the dimension of the data. PCA works by finding orthonormal variables, $\boldsymbol{z}$, that are themselves linear combinations of the original variables, allowing a large proportion of the variation in the data to be represented by a comparatively small number of transformed variables (Wilks, 2019).

That is, rather than representing atmospheric circulation using a vector of streamfunction values at each gridded location

$$\boldsymbol{\Psi} = (\Psi_1, \Psi_2, ..., \Psi_{1024}), \tag{11}$$

PCA allows the flow to be described by just a few of the uncorrelated, transformed variables

$$\boldsymbol{z} = (z_1, z_2, ..., z_p), \tag{12}$$

which explain a relatively large proportion of the low-frequency variability in the atmosphere. In this study, the norm streamfunction metric is used in the PCA. The number of principal components chosen is $p \ll 1024$; the leading three principal components are retained here, explaining 22.0% of variation in the hemispherical streamfunction.

Barnes *et al.* (2019) define weather regimes as the leading principal components of mean sea-level pressure fields over a relevant spatial domain. In the work presented here, the synoptic-scale atmospheric state is similarly projected on to the leading principal components, but regimes are then identified by performing an additional clustering step in this reduced space. This nonlinear approach allows opposite phases of a mode of atmospheric variability to exhibit spatial asymmetries, as is the case for the North Atlantic Oscillation (NAO), for example (Cassou *et al.*, 2004).

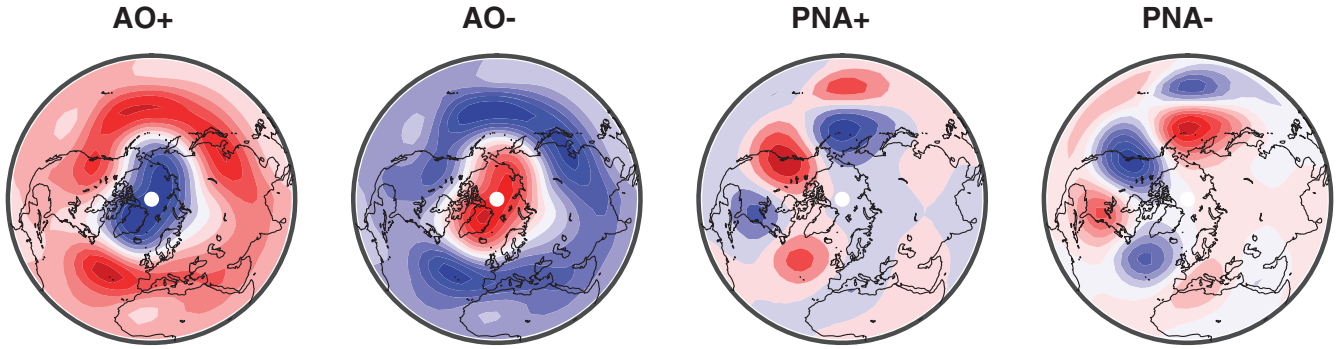

FIGURE 2 Regime centres when fitting a hidden Markov model to streamfunction anomalies from the quasigeostrophic model. Blue regions represent negative contours, while positive anomalies are shown in red. Contours are separated by intervals of 5×10^5 m^{2}s^{-1}

In particular, the time series consisting of 50 consecutive years worth of daily streamfunction anomalies is projected on to its leading three principal components, and it is from this sequence of 18,250 materialisations of $\boldsymbol{z}$ that the regimes are detected. These archived data are sequential, and so a hidden Markov model (HMM) is used to discern the regimes. Majda *et al.* (2006) first proposed the use of hidden Markov models in detecting atmospheric regimes, highlighting their ability to distinguish between distributions despite the leading principal components exhibiting nearly Gaussian statistics; HMMs are designed to detect more persistent regimes by exploiting the system's underlying dynamics.

A HMM assumes that the transformed variables in each regime follow a multivariate normal distribution, $\boldsymbol{z} \sim N(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$, and hence a mean vector, $\boldsymbol{\mu}$, and covariance matrix, $\boldsymbol{\Sigma}$, corresponding to each state must be estimated. A transition matrix, documenting the probability of transitioning between regimes, is also estimated. This is implemented using maximum likelihood via the Baum–Welch algorithm, a variant of the expectation-maximisation (EM) algorithm (Dempster *et al.*, 1977).

The centres of the four regimes identified by fitting a hidden Markov model to the archived data are depicted in Figure 2. The number of clusters is chosen to be four, due to the similarity of the resulting patterns to recognised atmospheric regimes: the positive and negative phases of both the Pacific North American (PNA+, PNA−) and Arctic Oscillation (AO+, AO−) patterns. The positive (negative) phase of the AO is synonymous with a strong (weak) polar vortex over the Arctic Circle, surrounded by a band of above (below) average streamfunction anomalies in the midlatitudes. The AO thus represents a zonally symmetric seesaw in streamfunction, or pressure anomalies between the Arctic basin and the extratropics (Thompson and Wallace, 1998). The positive (negative) PNA pattern, on the other hand, consists of below (above) average streamfunction anomalies over the Aleutian Islands and areas of high (low) anomalies over the Pacific basin and the northwestern United States (Wallace and Gutzler, 1981). Mean persistence times can be calculated from the Viterbi path, the most likely regime sequence over the data set, which can readily be determined from the HMM. The AO− regime (which occurs 24.9% of the time) has the longest mean persistence time, 10.7 days, followed by the AO+ regime (26.0%), which lasts for 9.6 days on average. The PNA patterns are comparatively less persistent, with the positive mode (29.2%) lasting for 6.1 days on average and the negative mode (19.9%) only 5.6 days, making it the least stable.

3.2 | Assigning forecasts to regimes

As mentioned above, HMMs assume a statistical distribution for the transformed streamfunction variables conditional on each underlying regime. As a result, having projected the streamfunction anomaly field on to the leading three principal components, Bayes' theorem can be used to calculate posterior probabilities of the atmosphere residing in each regime given the streamfunction values. The probability of residing in regime r given the reduced circulation, $\boldsymbol{z}$, is

$$p(r|\boldsymbol{z}) = \frac{p(\boldsymbol{z}|r)p(r)}{p(\boldsymbol{z})} = \frac{p(\boldsymbol{z}|r)p(r)}{\sum_{j=1}^{R} p(\boldsymbol{z}|j)p(j)}. \tag{13}$$

Here, $p(\boldsymbol{z}|r)$ is the likelihood of seeing the observed or predicted streamfunction values given that the atmosphere resides in regime r, and can be calculated from the multivariate normal density with mean vector and covariance matrix associated with that regime, $\boldsymbol{\mu}_r$ and $\boldsymbol{\Sigma}_r$. The climatology frequency of regime r is denoted by $p(r)$.

When the forecast must be assigned to exactly one regime, that with the highest posterior probability is chosen. Therefore, the initial and true regimes can be

determined by finding the regime that maximises the posterior probability given the observed streamfunction anomaly field at the forecast reference time and validation time, respectively. Similarly, the predicted streamfunction fields from the ensemble members can be used to allocate each member to a regime.

Obtaining a probabilistic distribution for the regime accounts for some of the inherent uncertainty present when identifying the latent atmospheric state. However, Bayes' theorem as given in Equation 13 does not make use of the estimated transition matrix and hence does not utilise the HMM's dependence on the system dynamics perfectly. A HMM produces a time series of posterior probabilities for each state given all data in the sequence. Calculating the Viterbi path, the most probable sequence of hidden states, then allows exactly one regime to be identified at each point in time. In a forecast setting, if a window of recent values were available prior to the current forecast, then the initial regime could be determined from the Viterbi path over this window, rather than the static posterior probability. A similar procedure is often implemented in data assimilation.

This would exploit the dynamics of the underlying states, and would therefore be particularly useful when the spatial structures of the regimes were similar, so that the temporal behaviour was more important when distinguishing between states. The regimes here differ considerably in space, and hence the estimated regime is not sensitive to the choice of method (not shown). Bayes' theorem, however, can be applied more easily to determine the future regime from each ensemble member when the preceding states are also unknown. Therefore, for ease of implementation, Bayes' theorem is used here to evaluate the regime given a streamfunction anomaly field.

3.3 | Results

The spatial domain of the QG model consists of 64 longitudes and 16 latitudes in the Northern Hemisphere and statistical post-processing is implemented at every grid point, yielding calibrated forecasts at 1024 locations. No spatial aggregation is performed and hence forecasts at each site are recalibrated using only previous forecast–observation pairs at the same location.

Figure 3 displays the CRPS for the TN approach, plotted on a map of the Northern Hemisphere at a lead time of 6 days. Forecast accuracy is worst towards the centres of the Pacific and Atlantic oceans, areas which correspond to well-known storm tracks. Maps of the CRPSS for the three regime-dependent methods, assessed using TN as reference, are also shown in Figure 3 for the same lead time. RDTN-init denotes the regime-dependent truncated normal approach conditioned on the initial regime, RDTN-ens is that using the ensemble member weights to predict the regime, and RDTN-true is dependent on the true weather regime at the forecast validation time. At locations far removed from the centres of the weather regimes, the improvements unsurprisingly fluctuate around zero. However, when the regimes affect the local wind speeds strongly, the RDTN-true method produces noticeable improvements in forecast skill. Both the RDTN-init and RDTN-ens methods appear considerably less effective than using the true regime.

In the Northern Hemisphere, wind travels counterclockwise around large-scale low-pressure systems and clockwise around high pressure, with the strengths of the winds related to the north–south pressure gradient (Hurrell and Deser, 2009). The improvements gained from the RDTN-true approach in Figure 3 are therefore concentrated in the North Atlantic and Pacific basins, and over northwest Canada: these regions surround the regime centres of action, so that the wind direction and intensity vary substantially depending on the prevailing regime. The wind speeds at these locations are thus influenced more heavily by the different regimes, resulting in an increased need for regime-dependent post-processing methods.

Historical observations at any given location can be grouped depending on the coinciding regime. At this location, the average wind speed given a regime can be described by the sample median of the observations in the relevant group. The variance of the sample medians then measures the spread of the average wind speed among the regimes. This spread quantifies the effect the regimes have on the wind speed at this location. A scatter plot is displayed in Figure 4, showing this metric against the CRPSS for all locations at a lead time of one week. Table 2 records the associated correlation between the spread of the regimes and the improvements gained from regime-dependent post-processing at lead times up to 7 days. Although the correlation is initially fairly low, the improvements gained when the true regime is used in post-processing become highly correlated with the spread of the average wind speeds at longer lead times. This suggests that there is more potential for improvement upon current post-processing approaches for forecasts further in advance. Neither the initial regime nor the ensemble members capture this behaviour.

To highlight the potential improvements, we focus now on results at one location in the west of the Atlantic Ocean. The marginal distribution of the wind speed, shown in Figure 5 when the atmosphere resides in each regime, indicates that the local wind speed is dependent on the prevailing state. At this location, AO+ corresponds to a strong negative meridional gradient in streamfunction anomalies, in turn producing high zonal wind speeds. Conversely,

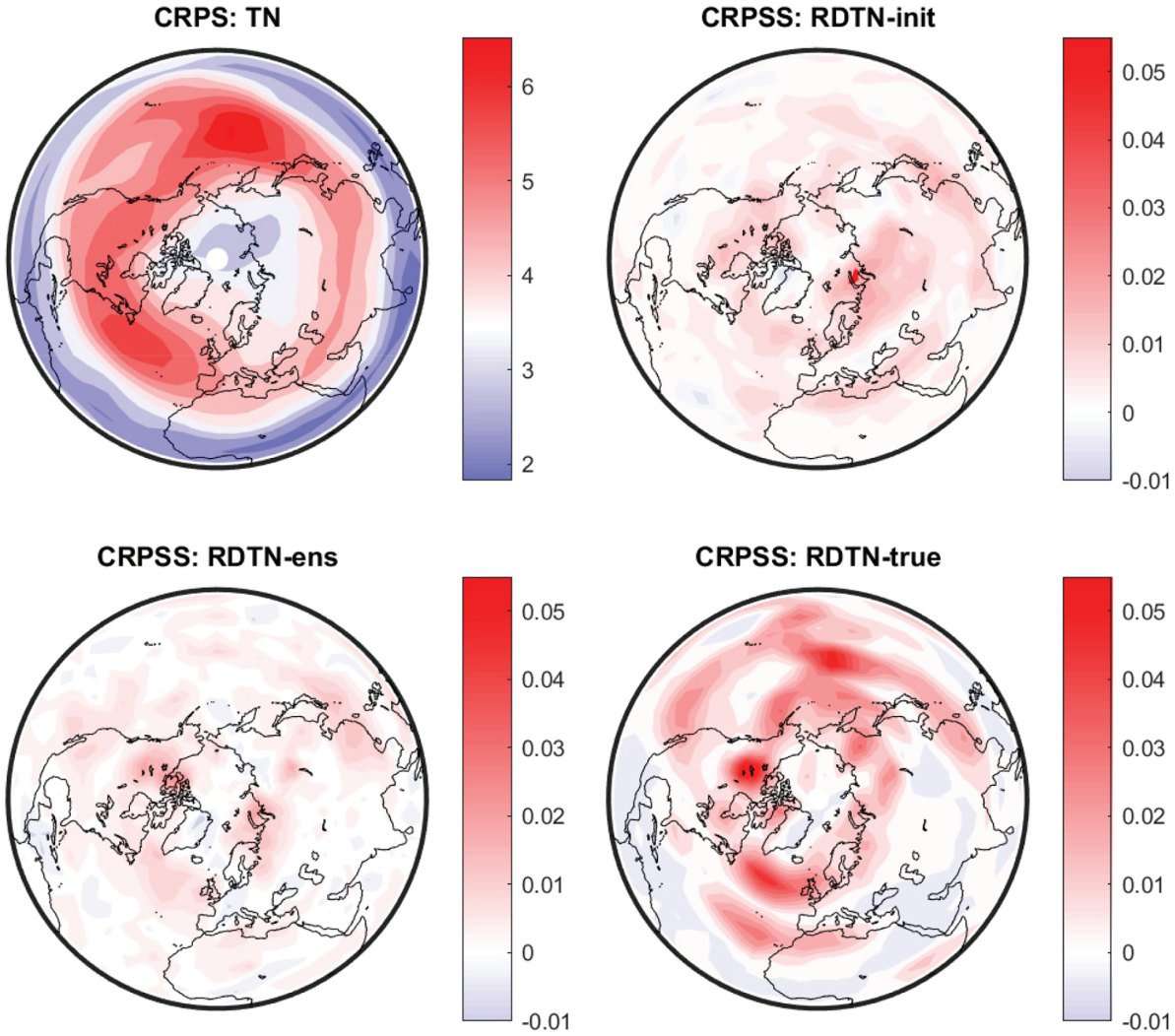

FIGURE 3 Map of the CRPS for the TN approach and the CRPSS for RDTN methods using TN as the reference forecast, at a lead time of 6 days. Standard errors in the CRPS are typically around 1% of the score's magnitude

the negative phase of the Arctic Oscillation is synonymous with low wind speeds in this area. The PNA patterns have less influence at this location, though wind speeds that are slightly lower than average occur in the negative phase.

The shape of the empirical wind-speed distributions also undergoes noticeable changes between the regimes: in the AO− regime the wind speeds are far more positively skewed than in the AO+ pattern. Although the formulations of the mixture model in Equations 1 and 2 allow separate forecast distributions to be issued depending on the regime, the truncated normal distribution is able to adapt for such changes.

The skill of the TN post-processing approach can be evaluated using the CRPS. Figure 6 displays the breakdown of the CRPS depending on the weather regime that occurs at the forecast validation time. There is a clear difference in forecast performance depending on the prevailing weather type. Scores are largest when the AO+ regime, in which extremely high wind speeds occur more frequently, materialises, while the lower wind speeds in the AO− regime are more predictable.

Figure 7 exhibits the skill of the regime-dependent TN predictive distributions relative to the conventional TN approach, assessed using the CRPSS. The uncertainty in the skill score is described by errors bars representing 95% confidence intervals, obtained via nonparametric bootstrap resampling. Although the improvements for all methods are initially negligible, RDTN-true forecasts become substantially more skilful at longer lead times; wind-speed forecasts at this location improve by almost 5% by including the synoptic-scale information. The RDTN-init approach, on the other hand, fails to make any meaningful contribution to the forecast. The CRPSS for RDTN-ens is significantly larger than zero for forecasts 5 and 6 days in advance, though the magnitude of the improvement in both cases is small. Figure 6 suggests

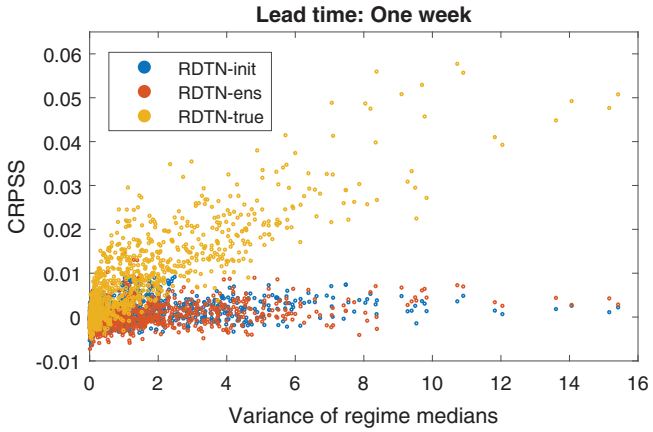

FIGURE 4 Scatter plot showing the variance of the average wind speeds among the regimes against the CRPSS at each location at a lead time of 7 days

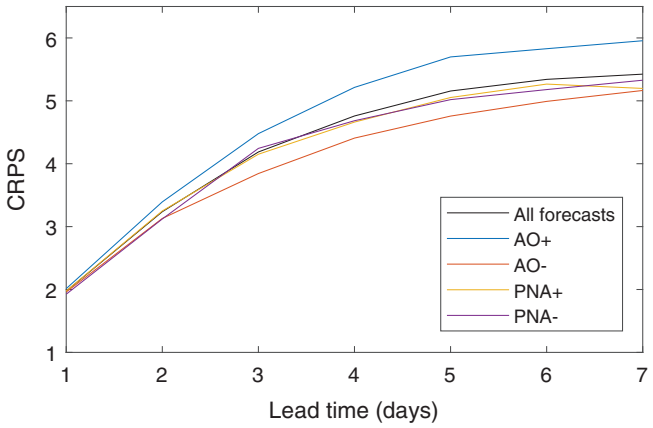

FIGURE 6 CRPS for the TN method against lead time when the atmosphere resides in each regime at the forecast validation time in the QG study

that forecast biases are initially relatively insensitive to the underlying regime and hence incorporating regimes would only be expected to benefit forecasts at longer lead times. However, by the time the biases become dependent on the weather regime, the ability of the mixture-model weights to recognise the true regime deteriorates. The skill score for the RDTN-init and RDTN-ens methods therefore consistently remains close to zero.

It is possible to decompose the CRPSS for the RDTN-true approach into the constituent regimes, as displayed in Figure 8. Although wind speeds are initially most predictable in AO−, improvements are also largest in this regime, reaching 12% for forecasts one week ahead. Predictions of the higher wind speeds in AO+ also improve, becoming up to 6% more skilful than when regime information is neglected. The PNA patterns have much less influence on the wind speed here and hence there is little benefit to including information in these states. Nonetheless, the improvements in forecasts in the AO regimes indicate that regime-dependent approaches may be more capable of forecasting events that deviate substantially from the local climatology, including extreme weather events.

TABLE 2 The correlation between the variance of the average wind speeds among the regimes and the CRPSS, calculated over all locations

	1	3	5	7
RDTN-init	0.047	0.367	0.357	0.275
RDTN-ens	0.022	0.432	0.500	0.298
RDTN-true	0.056	0.646	0.816	0.845

Note: Results are shown for the three regime-dependent methods at lead times of 1, 3, 5, and 7 days.

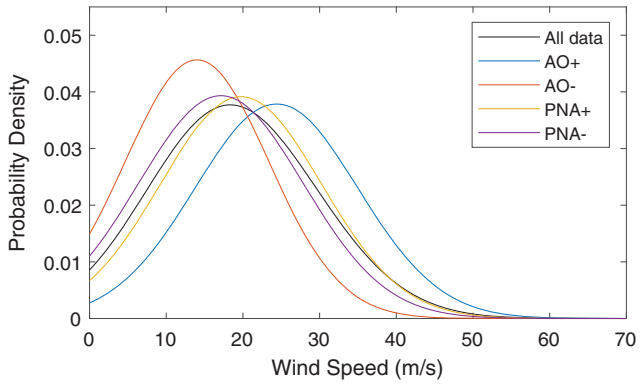

FIGURE 5 Empirical distribution of wind-speed observations at the location of interest when the atmosphere resides in each regime in the QG setting

For sufficiently large lead times, the raw forecast becomes uninformative, containing no information about the predictand. In this case, the statistical post-processing methods should forecast the marginal distribution of the weather variable of interest. It is believed that if the atmospheric regime could be forecast perfectly then the improvement gained from regime-dependent post-processing would be present even this far in advance, since the regime-dependent post-processing will issue the marginal distribution of the wind speed in each regime. The additional flexibility of the mixture model therefore allows it to capture more complex features that arise due to the different regimes, such as multimodality of the marginal distribution.

Figure 9 displays the relative frequency with which the observation assumes each rank when pooled with the ensemble members, at a lead time of 5 days. Rank histograms are a commonly used tool for assessing

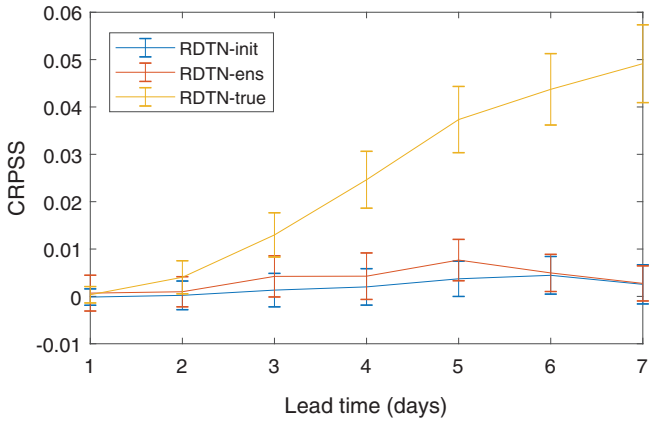

FIGURE 7 CRPSS for all three regime-dependent post-processing models in the QG study against lead time, with TN as the baseline. Error bars show 95% confidence intervals at each lead time

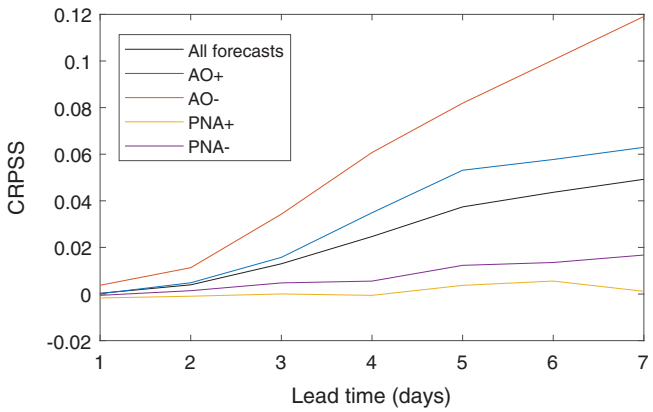

FIGURE 8 CRPSS for RDTN-true forecasts against lead time when the atmosphere resides in each regime at the forecast validation time in the QG study, with TN as the baseline

the calibration of ensemble forecasts, with uniform histograms denoting reliable predictions (Anderson, 1996; Hamill and Colucci, 1997; Talagrand, 1997). Clearly, however, the raw ensemble forecast is underdispersed, with observations falling outside the range of ensemble members more frequently than would be expected if the forecast were calibrated, regardless of the underlying regime. Also shown in Figure 9 are probability integral transform (PIT) histograms, the continuous analogue of the rank histogram. The PIT histograms evaluate the TN and RDTN-true forecast distributions at the verification, and record the rate at which the resulting probabilities fall into each of a number of equally sized bins. There are 11 possible positions of the verification when pooled with the 10 ensemble members, and hence the number of bins is also chosen to be 11.

Although the PIT histogram for the TN approach over all forecast–observation pairs is suitably uniform at this location, the parameters estimated over the entire training set produce a model that does not fit the data well when the system resides in the Arctic Oscillation patterns. In particular, oppositely skewed PIT histograms indicate that the observed wind speed falls into the upper tail of the forecast distribution when the AO+ regime occurs, and the lower tail when the AO− pattern materialises. The TN approach is thus not calibrated conditional on the regime. The RDTN-true approach, on the other hand, accounts for the varying model biases in the regimes, and the corresponding PIT histograms are close to uniform in all of the four regimes.

4 | GEFS REFORECASTS

4.1 | Data

Previous occasions in which similar atmospheric behaviour has occurred will likely lead to similar model biases. Therefore, regime-dependent statistical post-processing would be particularly well-suited to use with reforecast data, where a large set of hindcasts from a frozen operational model are available. These hindcasts, spanning several years or decades, can be used to train statistical post-processing methods (Hamill *et al.*, 2004). In this section, the regime-dependent approaches are implemented on data from version 2 of the National Oceanic and Atmospheric Administration's Reforecast project (Hamill *et al.*, 2013). Forecasts are taken from a recent version of the National Centers for Environmental Prediction's (NCEP) Global Ensemble Forecasting System (GEFS), and although Hamill *et al.* (2013) note that they may also be prone to model biases, the reanalyses are used as a best guess for the observed wind-speed values. To negate these biases, the control member run from the reanalysis is omitted, resulting in a forecast of 10 statistically interchangeable ensemble members.

Since regime behaviour is most prominent during winter, the data set covers the 34 cold seasons (November–March inclusive) between 1985 and 2019. Results in the previous section established that locations heavily affected by the identified weather regimes are likely to improve as a result of regime-dependent post-processing. Therefore, it is hoped that defining localised regimes over a smaller spatial domain will have a larger effect on the recalibration. Following Ferranti *et al.* (2015), the atmospheric regimes here are detected using k-means clustering over the Euro-Atlantic sector (80°W–40°E, 30–90°N). k-means clustering partitions data into a prespecified number of groups by assigning data points to clusters such that the distance between the point

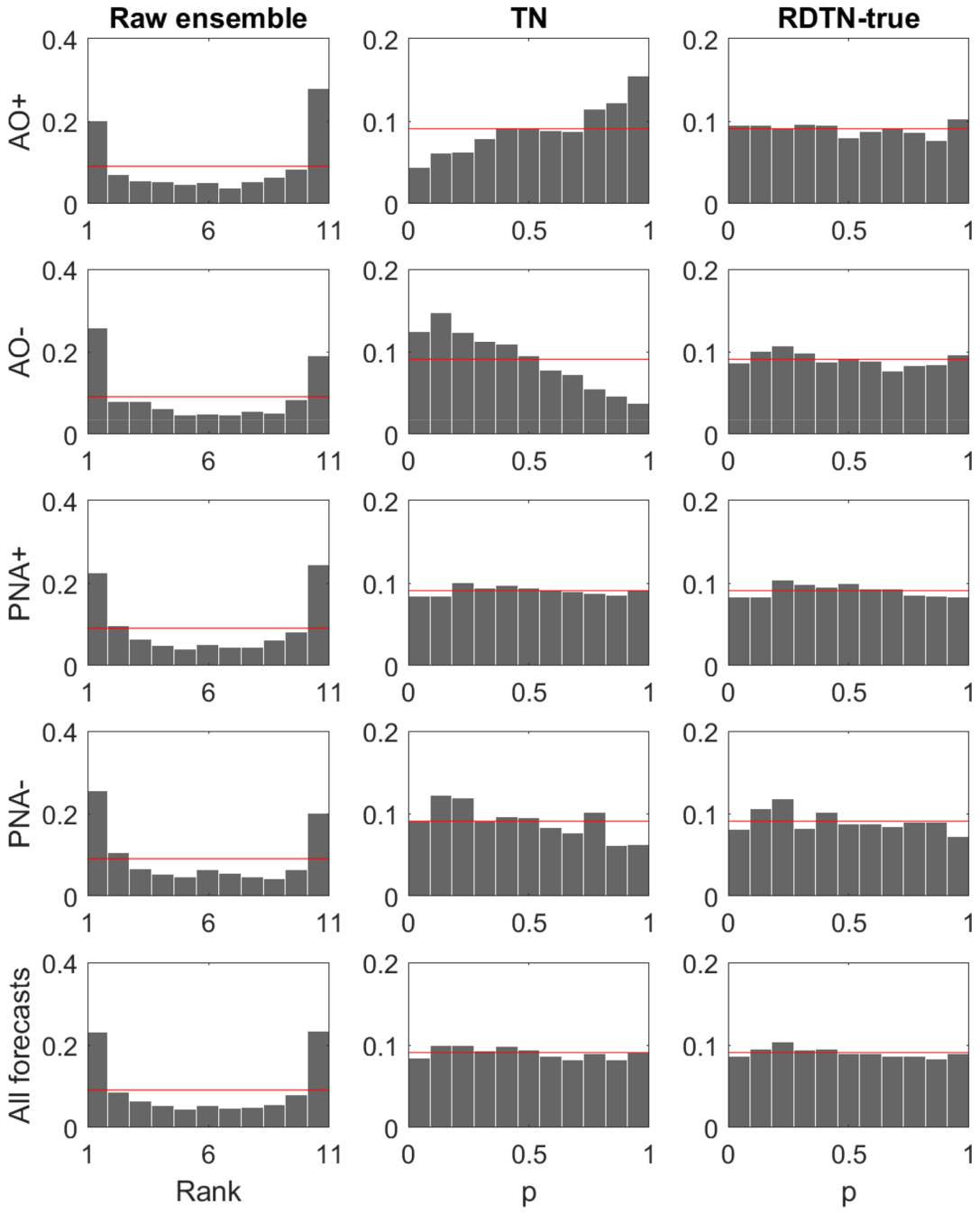

FIGURE 9 Rank and PIT histograms showing the relative frequency of each bin for the raw ensemble forecasts, the TN and RDTN-true post-processing methods at a lead time of 5 days. Histograms are shown for forecasts grouped by the atmospheric regime at the forecast validation time. A horizontal line is added in red at 0.091 to indicate perfect uniformity across the bins

and the allocated cluster centroid is minimised (Michelangeli *et al.*, 1995; Wilks, 2019). The number of clusters, k, must be chosen prior to implementing the algorithm. Four regimes are again used here due to the similarity of the resulting patterns to those identified in numerous studies of regime behaviour over this domain (Cassou *et al.*, 2004).

Reanalyses of 500-hPa geopotential height anomaly fields are used to represent the atmosphere's circulation in this domain. PCA is then applied, using the Euclidean metric in grid-point space, to these anomaly fields and clustering is performed in this reduced space. The leading three principal components, which explain 48% of the variation in the flow, are chosen. Figure 10 shows the geopotential height anomalies that correspond to the regime, or cluster, centres identified using k-means clustering. Despite fewer principal components being used than in Ferranti *et al.* (2015), there is similar evidence

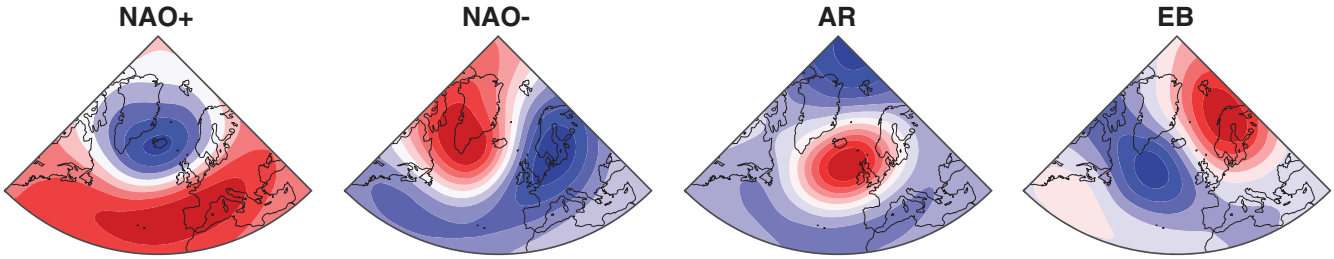

FIGURE 10 Regime centres identified by applying *k*-means clustering to the reforecast geopotential height anomalies. Anomalies are displayed at 25-hPa intervals, with blue regions indicating negative contours and red regions representing positive anomalies

to support the use of four regimes, which resemble the positive and negative phases of the NAO as well as European Blocking (EB) and an Atlantic Ridge (AR). It has been proposed that the NAO corresponds to the same mode of circulation variability as the Arctic Oscillation described in Section 3 (Hurrell and Deser, 2009). The NAO thus constitutes a north–south dipole, characterised by negatively correlated height anomalies between Iceland and the Azores, although opposite phases of the NAO do not exhibit identical spatial structures. The AR pattern represents an anticyclonic regime over the eastern North Atlantic Ocean, while European Blocking consists of a dipole with positive geopotential height anomalies over Scandinavia and negative anomalies to the south of Greenland. The atmosphere resides in the NAO+ regime 30.8% of the time, making it the most frequently occurring regime. The NAO− and EB regimes occur similarly often (24.2 and 24.3% respectively), while AR materialises least often (20.7%).

The data are noncontiguous and hence a hidden Markov model cannot readily be applied to the data. The main benefit of *k*-means clustering, on the other hand, is that forecasts can be assigned to a regime with ease. Franzke *et al.* (2008) remark that, although clustering approaches find the states that have the highest probability of occurring, the resulting regimes do not necessarily exhibit persistence. As a result, the mean persistence times are much lower for these regimes than for the patterns found using a HMM in the QG framework: AR events persist for only 3.8 days on average, EB for 4.9 days and the NAO− and NAO+ regimes for 5.4 and 6.1 days on average, respectively.

Forecasts are assigned to one of these four patterns by finding the regime for which the Euclidean distance between the associated cluster centroid and the reduced geopotential height anomaly field is minimised. As before, the initial and true regimes make use of the observed geopotential height anomaly field at the forecast reference time and validation time, respectively, while output from the ensemble forecast is used to allocate each member to a regime. This approach fails to account for the inherent uncertainty when assigning a forecast to a regime; every point allotted to a cluster is assumed to exhibit the same biases and systematic errors, regardless of its distance to the cluster centre, and hence a method that provides the probability of residing in the different regimes, or a degree of membership, would be more informative in this respect.

Hamill *et al.* (2013) remark that the method for constructing the forecast analyses in the GEFS changes in February 2011, and the forecast skill consequently improves. Therefore, to maintain the similarity of the model biases in the study, only forecasts in the 25 cold seasons from November 1985–March 2010 are considered. Although this model change also affects the observed geopotential height anomalies, it provides a more informed estimate of the atmospheric state and hence data after this change are still utilised when detecting the regimes. The resulting regimes are found to be more robust when these additional data are included.

Although techniques have recently been proposed that include cyclic functions to remove seasonal model errors (Dabernig *et al.*, 2017; Lang *et al.*, 2019), parameter estimation is typically performed operationally using a training window that consists only of the most recently available forecast–observation pairs. These rolling windows account for the recent behaviour of forecast errors, and alleviate biases owing to changes in the NWP model. The size of this window is clearly a compromise that requires having enough data to obtain reliable parameter estimates without using too much, so as to capture the recent behaviour of the atmosphere.

The CRPS is found here to decrease as the amount of training data available increases, but is generally insensitive to the window length (not shown). In fact, at the majority of locations tested, more skilful forecasts are issued when a fixed training window is used, containing several years of past data. Therefore, for both the standard and regime-dependent post-processing methods, the first 15 cold seasons (those beginning in 1985–1999) are used as a fixed training window, while the remaining 10

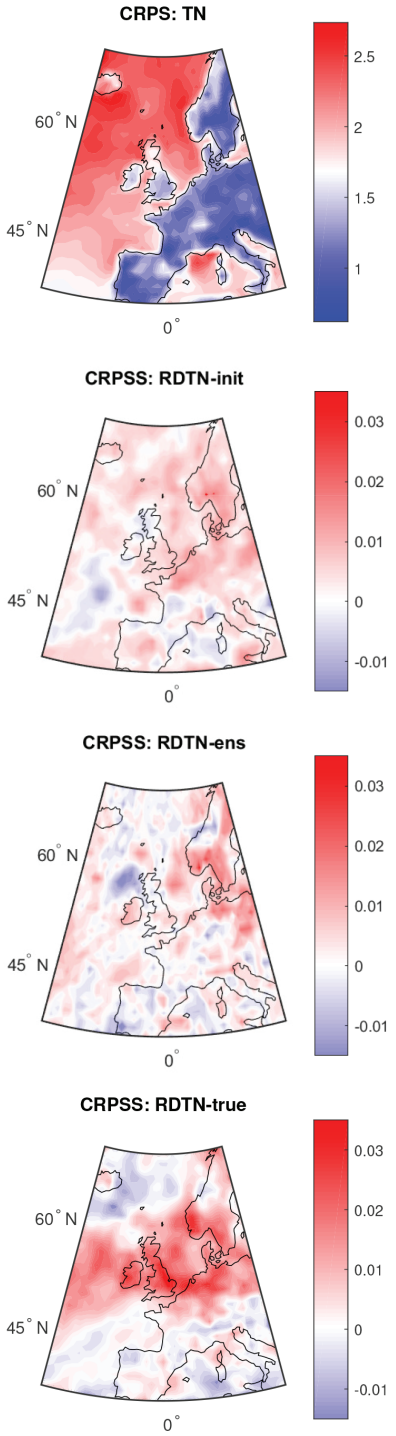

FIGURE 11 CRPS for the TN approach and CRPSS for the RDTN-init, RDTN-ens, and RDTN-true extensions at a lead time of 7 days, plotted on a map of the spatial domain under consideration. Standard errors typically lie between 1.5 and 2.5% of the CRPS itself

(2000–2009) are used as test data. The advantages of using a rolling window diminish in this case since, there are no changes in the prediction system, and investigating only cold seasons accounts for some of the seasonality in the biases.

TABLE 3 The correlation between the variance of the average wind speeds among the regimes and the CRPSS, calculated over all locations

	1	3	5	7
RDTN-init	0.010	0.034	0.044	−0.017
RDTN-ens	−0.055	−0.059	0.040	0.010
RDTN-true	−0.065	−0.007	0.212	0.569

Note: Results are shown for the three regime-dependent methods at lead times of 1, 3, 5, and 7 days.

4.2 | Results

Post-processing is performed here on a subset of the spatial domain under consideration, which consists of 1,353 locations over western Europe and the east of the North Atlantic ocean (21°W–19°E, 37–69°N). Locations are separated by 1° of longitude and latitude.

The CRPS for the TN post-processing approach is displayed in Figure 11. Wind-speed forecasts are significantly more skilful over land than sea and forecasts at locations close to Iceland are particularly poor, since this corresponds to a mode of North Atlantic storm-track variability (Serreze *et al.*, 1997). The CRPSS for RDTN-init, RDTN-ens, and RDTN-true are also displayed in Figure 11 at a lead time of one week. At this longer lead time, the CRPSS for RDTN-init and RDTN-ens remains close to zero, though large improvements are seen when the true regime is used, particularly at locations surrounding the North Sea.

We postulate that the spatial structure of the improvements in Figure 11 is again linked to how air flows around large-scale pressure systems. The regime centres in Figure 10 suggest that the regions between the modes of high and low pressure often intersect the area surrounding the North Sea, and therefore the wind speeds at neighbouring grid points are more dependent on the prevailing weather type. Calibrating forecasts separately in each regime thus produces larger improvements at these locations.

Table 3 shows that the improvements are again correlated with the spread of the average wind speeds between the regimes. The magnitude of the spread tends to be much smaller than in the QG study, suggesting that the regimes have less effect on the wind speeds in this sectorial domain than over the entire hemisphere. This is consistent with results in Tibaldi and Molteni (1990), which show that relatively intense blocking events occur more over the Pacific than the Euro-Atlantic.

More detailed results are provided for one location close to Bergen, on the west coast of Norway. The quality of the raw ensemble forecast can be assessed using

TABLE 4 CRPS for raw ensemble forecasts initialised in each regime, at a lead time of 3 days

	NAO+	NAO−	AR	EB	Total
Raw ensemble	1.44	1.18	1.06	1.24	1.23

the CRPS to understand how the raw model errors change with the regime. Table 4 displays the skill of the raw ensemble forecasts initialised in each regime. Forecasts initialised in the NAO+ regime, which coincides with more extreme wind speeds, exhibit considerably less skill than those in the other regimes. Differences in the skill of forecasts among the regimes indicate that conditioning the statistical post-processing on the prevailing regime may therefore be expected to yield more skilful forecasts.

Figure 11 suggests that using the true regime at this location provides relatively large improvements that are not present when conditioning forecasts on the regime at the initialisation time. The CRPSS is shown for all lead times in Figure 12, with 95% confidence intervals at each lead time estimated using nonparametric bootstrap resampling. A pattern similar to that seen previously emerges: scores for the initial regime recede to zero as lead time increases, while there appears to be more room for improvement at longer lead times, something that is exploited when using the true regime. The RDTN-ens approach performs significantly better than the TN method, and is comparable to RDTN-true, for forecasts up to 4 days ahead, but its skill declines as lead time increases. The fact that RDTN-ens produces a larger CRPSS than RDTN-true at early lead times could indicate that the ensemble regime may be exploiting a feature of the data that is not picked up by the true regime, though it is more likely a result of sampling variation. Due to the large amounts of training data available from reforecasts, no methods perform worse than when regimes are not included in the post-processing, despite the increased number of parameters.

Figure 13 shows the empirical distributions of the wind speed when the atmosphere resides in each regime. Positive NAO indices are linked to more intense and frequent storms in the Norwegian Sea (Serreze *et al.*, 1997), and hence wind speeds here are largest in the NAO+ regime. Comparatively low wind speeds are associated with the NAO− regime, while the EB and AR regimes do not have much effect on the wind speed at this location. As a result, the improvements gained from regime-dependent post-processing are dominated by improvements in the two phases of the NAO patterns. Figure 14 shows the CRPSS for the RDTN-true approach for forecasts corresponding to each regime at the forecast validation time. In

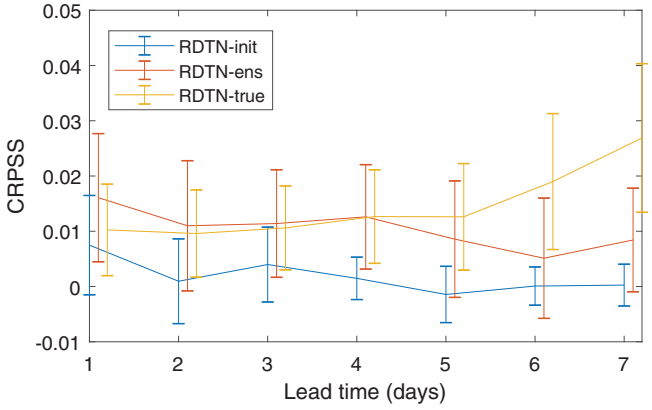

FIGURE 12 CRPSS against lead time for each of the three regime-dependent methods at a location close to Bergen, Norway. Scores for RDTN-ens have been offset by 0.1 days and those for RDTN-true by 0.2 days, to visualise 95% confidence intervals around the scores

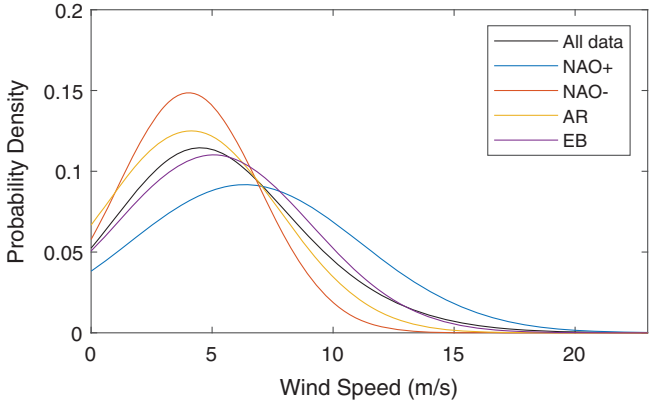

FIGURE 13 Empirical distributions of wind speed at a location close to Bergen, Norway, when each identified regime occurs in the reforecast setting

particular, Figure 14 suggests that improvements at short lead times occur primarily in the NAO− regime, while at longer lead times forecasts in the NAO+ regime improve upon the conventional TN approach by as much as 4%. Since the positive phase of the NAO is associated with particularly high wind speeds at this location and the negative phase with low wind speeds, these results reinforce the idea that if the regime at the forecast validation time is correctly identified then regime-dependent post-processing can provide better forecasts of extremely high and low wind speeds. Rank and PIT histograms for the various post-processing methods display features similar to those shown in Figure 9: the observation falls into the upper tail of the TN forecast distribution more frequently than expected during NAO+ events, and less frequently during NAO− events, while the RDTN-true method appears calibrated conditional on the regimes. However, since the

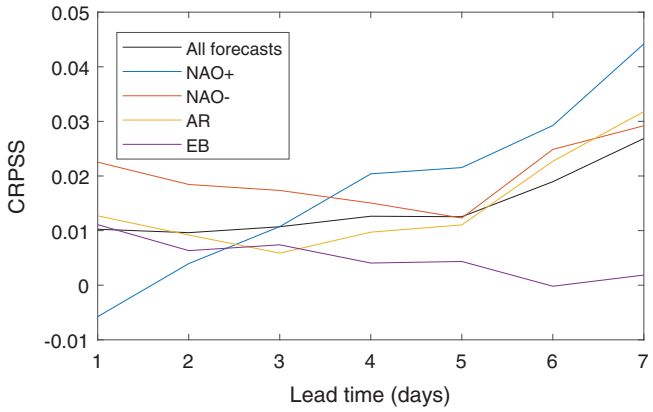

FIGURE 14 Skill score for RDTN-true forecasts partitioned by the true regime at the forecast validation time, shown at all lead times. The TN method is used as a reference forecast

improvements are smaller here, deviations from uniformity are less pronounced.

5 | DISCUSSION

This article builds upon previous work on the regime-dependent statistical post-processing of ensemble forecasts (Allen *et al.*, 2019). It is suggested that NWP models exhibit biases that change depending on the concurrent atmospheric regime and hence conditioning current statistical calibration methods on these regimes can enhance forecasts. Wind speed is closely connected to the movement of air in the atmosphere and is hence dependent on the prevailing regime behaviour.

Regime-dependent extensions of nonhomogeneous regression are proposed that utilise a weighted mixture of truncated normal predictive distributions. Mixture models of this form provide a more flexible forecast distribution that accounts for biases owing to large-scale changes in the atmosphere's circulation. The weights represent the probability of residing in a number of identified weather regimes, and results are presented here for three ways of defining them. The first is an indicator function that depends on the regime at the forecast initialisation time, the second is the proportion of ensemble members predicting each regime at the validation time, and the regime that actually materialises at the validation time is also implemented. Although the latter approach is not applicable in practice, it is regarded here as an upper bound for the improvement and hence provides a useful comparison. It could also be argued that if the true regime were known then it might be more useful to condition on both the true regime and the forecast regime; if the forecast predicted one regime yet the actual regime was known to be different, then the biases would be larger than if the forecast and atmosphere agreed on the regime.

The regime-dependent approaches are implemented in two scenarios: in a quasigeostrophic model of the Northern Hemisphere and on GEFS retrospective forecasts over the Euro-Atlantic sector. Regimes are identified by projecting the large-scale flow, represented by the streamfunction or geopotential height anomalies at all spatial locations, on to the leading three principal components, before detecting patterns in the resulting variables. A hidden Markov model is fitted in the QG setting, while k-means clustering is applied to the reforecast data. The retrospective forecasts are generated from a higher resolution NWP model than that studied in the QG framework, but a data-rich simulation study is also helpful when trialling a new method, since conclusions can be made that are more resistant to sampling variation. The results found in the reforecast setting corroborate those in the QG study.

If a probabilistic approach is used to define the regimes at the initialisation time, as is done in the QG study, then it is possible to use these posterior probabilities as weights in Equation 1. Results indicate that accounting for this uncertainty improves the skill score for the RDTN-init approach slightly. These results were not included to maintain comparison between the results in the different settings: such a probabilistic approach is not possible when using k-means clustering to identify regimes. It would also be possible to use the training data to obtain conditional probabilities of each regime occurring given the ensemble member regime weights. This could itself be thought of as post-processing applied to the forecast of the regime.

Hamill *et al.* (2004) conclude that running operational models from historical analyses, or reanalyses, is a more efficient use of computational resources than increasing the model resolution. Recalibration methods can then utilise a considerably larger amount of data. Although this is highly computationally demanding at first, it obviates the need to re-estimate post-processing parameters for every forecasting occasion. The use of more data helps to refine forecasts of more extreme weather events, whilst also reducing parameter uncertainty in the post-processing models considerably. These benefits have been reinforced here, where a fixed training window consisting of several cold seasons worth of previous forecast–observation pairs was found to outperform a rolling training window at the majority of locations.

The improvements gained from regime-dependent post-processing were found to be positively correlated with the spread of the average wind speed between the regimes. That is, the larger the effect of the weather regimes on the local wind speed, the larger the expected improve-

ment. This allowed us to provide an example for both studies that highlighted the potential developments from the regime-dependent approach.

Forecasters have noted that knowing the prevailing synoptic behaviour of the atmosphere at the initialisation time can help to predict the forecast accuracy. It is found here that, in order to benefit post-processing at longer lead times, it is not enough to know the behaviour at the initialisation time; instead, a good estimate of the behaviour at the validation time is required.

Using the regime defined at the forecast initialisation time contains little information regarding the true regime at longer lead times and therefore, although there were minor improvements for forecasts at small lead times, they were significantly less pronounced for longer forecast horizons. Using the ensemble members to predict the regime offered more skill than using the initial regime, though skill scores again reduced to zero as lead time increased. The upper bound on the skill score, on the other hand, appeared to increase with lead time, suggesting that larger relative improvements over conventional post-processing methods are potentially available for forecasts further in advance.

A more accurate NWP model would likely be more adept at identifying the regime correctly at the forecast validation time. However, if the NWP model is used to identify the regime, then as the model produces more skilful forecasts of the large-scale circulation (from which the regimes can be identified) it may also provide better forecasts of other, smaller-scale variables, such as wind speed or temperature. The available improvements upon standard post-processing methods would therefore decrease as the biases in the model become smaller and less varied between the different regimes. This intuition also explains why the potential improvements of regime-dependent post-processing are particularly small at short lead times; the magnitudes of model biases are generally smaller and hence the differences between the regimes become insignificant.

Nonetheless, Ferranti *et al.* (2015) show that high-resolution ensemble prediction systems still exhibit biases that depend on atmospheric regimes, and hence there is still reason to believe that regime-dependent approaches will be useful when calibrating these more accurate forecasts. The GEFS reforecasts here were verified against model analyses, which may be subject to the same limitations as the prediction system. Since the NWP model may not simulate the spatial and temporal characteristics of the observed weather regimes correctly, evaluating forecasts against station observations may result in larger regime-dependent biases.

It may be the case that the choice of predictive distribution should vary with the regime and hence future work could investigate which distributions are most appropriate for certain weather types or situations. The numbers of regimes used in this study were chosen subjectively, using results from previous studies as guidance. Whether there exists a statistical procedure to estimate these regimes such that they are optimal for use in post-processing is also a topic for further research.

Furthermore, it may be the case that the optimal regimes, or number of regimes, changes depending on the location or predictand under consideration. The extent to which a regime affects the wind speed at a certain location was found here to depend on its proximity to the regime centres of action. Each regime thus provided valuable information at some locations, but not others. If interest lies only in one location, then it may be preferable to estimate more localised, or even site-specific, regimes, which could also vary for each predictand being forecast.

The regimes considered here are advantageous because they are physically meaningful, which may not be the case for regimes estimated separately at every location for each variable. As a result, considerable work has been devoted to studying their dynamical and statistical properties, and such studies can be used to identify situations where the inclusion of regimes may be most beneficial. Previous work, for example, has noted the impact they have on local weather systems, and how they can account for the dependence between meteorological variables and multiple locations. They thus naturally lend themselves to use with post-processing in a spatial or multivariate context.

We therefore argue that the appropriate regimes and number of regimes, should be investigated prior to post-processing, utilising previous studies of low-frequency variability in the domain under consideration. The number of regimes to use also depends on the amount of data available. Using a large number of weather types can result in overfitting of the training data, leading to less informative out-of-sample predictions. In the study reported in this article, estimating four times as many parameters as the original truncated normal approach did not induce any problems of this sort.

Alternatively, atmospheric circulation could be incorporated into post-processing approaches without discretisation into a finite number of regimes. It is found that improvements are only likely to be seen for regimes in which wind speeds differ severely from the average wind speed at a location. If the local weather depends strongly on one or two known regimes, then the continuous indices for these patterns, if such indices exist, could be incorporated as additional covariates in the post-processing model. Since this requires fewer parameters to be estimated, it would be more feasible to implement with a rolling training window when reforecast data were not available.

We expect improvements from regime-dependent post-processing to be largest in winter, since this is when the regime behaviour of the atmosphere is most pronounced. The regime-dependent model biases may themselves be dependent on the season. For example, blocking episodes are associated with heat waves in summer and cold snaps in winter and therefore temperature biases may be inconsistent between separate occurrences of the same regime. If all seasons are considered at once, then these could be treated as separate regimes, despite corresponding to the same large-scale mode of variability. If a large number of regimes, or even smaller-scale weather patterns, were used, then other latent variable methods, such as hierarchical models, may be more appropriate.

Moreover, results here suggest that regime-dependent post-processing is particularly adept at calibrating forecasts corresponding to regimes in which the weather differs greatly from the local climatology. Further investigation into the use of regime-dependent approaches when forecasting extreme events would therefore complement previous comparison studies (Williams *et al.*, 2014).

ACKNOWLEDGEMENTS

Sam Allen was supported during this work by a NERC Industrial CASE studentship under grant reference NE/N008693/1. We thank Gavin Evans, Piers Buchanan, and Theo Economou for helpful comments and suggestions throughout this study. Feedback from two anonymous reviewers has also been greatly appreciated.

ORCID

S. Allen https://orcid.org/0000-0003-1971-8277
C. A. T. Ferro https://orcid.org/0000-0002-9830-9270
F. Kwasniok 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.

Anderson, J.L. (1996) A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *Journal of Climate*, 9, 1518–1530.

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.

Barnes, C., Chandler, R.E. and Brierley, C.M. (2019) New approaches to postprocessing of multi-model ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 145, 3479–3498.

Bremnes, J.B. (2004) Probabilistic wind power forecasts using local quantile regression. *Wind Energy*, 7, 47–54.

Bremnes, J.B. (2019) Constrained quantile regression splines for ensemble postprocessing. *Monthly Weather Review*, 147, 1769–1780.

Brier, G.W. (1950) Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78, 1–3.

Cassou, C., Terray, L., Hurrell, J.W. and Deser, C. (2004) North Atlantic winter climate regimes: spatial asymmetry, stationarity with time, and oceanic forcing. *Journal of Climate*, 17, 1055–1068.

Cheng, X. and Wallace, J.M. (1993) Cluster analysis of the northern hemisphere wintertime 500-hPa height field: spatial patterns. *Journal of the Atmospheric Sciences*, 50, 2674–2696.

Dabernig, M., Mayr, G.J., Messner, J.W. and Zeileis, A. (2017) Spatial ensemble post-processing with standardized anomalies. *Quarterly Journal of the Royal Meteorological Society*, 143, 909–916.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39, 1–22.

Eide, S.S., Bremnes, J.B. and Steinsland, I. (2017) Bayesian model averaging for wind-speed ensemble forecasts using wind speed and direction. *Weather and Forecasting*, 32, 2217–2227.

Epstein, E.S. (1969) Stochastic dynamic prediction. *Tellus*, 21, 739–759.

Ferranti, L., Corti, S. and Janousek, M. (2015) Flow-dependent verification of the ECMWF ensemble over the Euro-Atlantic sector. *Quarterly Journal of the Royal Meteorological Society*, 141, 916–924.

Franzke, C., Crommelin, D., Fischer, A. and Majda, A.J. (2008) A hidden Markov model perspective on regimes and metastability in atmospheric flows. *Journal of Climate*, 21, 1740–1757.

Gebetsberger, M., Messner, J.W., Mayr, G.J. and Zeileis, A. (2018) Estimation methods for nonhomogeneous regression models: minimum continuous ranked probability score versus maximum likelihood. *Monthly Weather Review*, 146(12), 4323–4338.

Gneiting, T., Balabdaoui, F. and Raftery, A.E. (2007) Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B*, 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. (2013) Combining predictive distributions. *Electronic Journal of Statistics*, 7, 1747–1782.

Greybush, S.J., Haupt, S.E. and Young, G.S. (2008) The regime dependence of optimally weighted ensemble model consensus forecasts of surface temperature. *Weather and Forecasting*, 23, 1146–1161.

Hamill, T.M., Bates, G.T., Whitaker, J.S., Murray, D.R., Fiorino, M., Galarneau Jr, T.J., Zhu, Y. and Lapenta, W. (2013) NOAA's second-generation global medium-range ensemble reforecast

dataset. *Bulletin of the American Meteorological Society*, 94, 1553–1565.

Hamill, T.M. and Colucci, S.J. (1997) Verification of ETA–RSM short-range ensemble forecasts. *Monthly Weather Review*, 125, 1312–1327.

Hamill, T.M., Whitaker, J.S. and Wei, X. (2004) Ensemble reforecasting: improving medium-range forecast skill using retrospective forecasts. *Monthly Weather Review*, 132, 1434–1447.

Hannachi, A., Straus, D.M., Franzke, C.L., Corti, S. and Woollings, T. (2017) Low-frequency nonlinearity and regime behavior in the northern hemisphere extratropical atmosphere. *Reviews of Geophysics*, 55, 199–234.

Horel, J.D. (1985) Persistence of the 500 mb height field during northern hemisphere winter. *Monthly Weather Review*, 113, 2030–2042.

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.

Junk, C., Delle Monache, L. and Alessandrini, S. (2015) Analog-based ensemble model output statistics. *Monthly Weather Review*, 143, 2909–2917.

Kimoto, M. and Ghil, M. (1993) Multiple flow regimes in the northern hemisphere winter. Part I: Methodology and hemispheric regimes. *Journal of the Atmospheric Sciences*, 50, 2625–2644.

Kober, K., Craig, G. and Keil, C. (2014) Aspects of short-term probabilistic blending in different weather regimes. *Quarterly Journal of the Royal Meteorological Society*, 140, 1179–1188.

Koch, S.E., Skillman, W.C., Kocin, P.J., Wetzel, P.J., Brill, K.F., Keyser, D.A. and McCumber, M.C. (1985) Synoptic scale forecast skill and systematic errors in the MASS 2.0 model. *Monthly Weather Review*, 113, 1714–1737.

Kondrashov, D., Ide, K. and Ghil, M. (2004) Weather regimes and preferred transition paths in a three-level quasigeostrophic model. *Journal of the Atmospheric Sciences*, 61, 568–587.

Kwasniok, F. (2007) Reduced atmospheric models using dynamically motivated basis functions. *Journal of the Atmospheric Sciences*, 64, 3452–3474.

Kwasniok, F. (2019) Fluctuations of finite-time Lyapunov exponents in an intermediate-complexity atmospheric model: a multivariate and large-deviation perspective. *Nonlinear Processes in Geophysics*, 26, 195–209.

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 nonhomogeneous regression. *Nonlinear Processes in Geophysics*, 27, 23–34.

Leith, C. (1974) Theoretical skill of Monte Carlo forecasts. *Monthly Weather Review*, 102, 409–418.

Lerch, S. and Thorarinsdottir, T.L. (2013) Comparison of nonhomogeneous regression models for probabilistic wind-speed forecasting. *Tellus A*, 65, 21206.

Lorenz, E.N. (1969) The predictability of a flow which possesses many scales of motion. *Tellus*, 21, 289–307.

Majda, A.J., Franzke, C.L., Fischer, A. and Crommelin, D.T. (2006) Distinct metastable atmospheric regimes despite nearly Gaussian statistics: a paradigm model. *Proceedings of the National Academy of Sciences of the USA*, 103, 8309–8314.

Marshall, J. and Molteni, F. (1993) Toward a dynamical understanding of planetary-scale flow regimes. *Journal of the Atmospheric Sciences*, 50, 1792–1818.

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(1), 137–147.

Michelangeli, P.-A., Vautard, R. and Legras, B. (1995) Weather regimes: recurrence and quasi stationarity. *Journal of the Atmospheric Sciences*, 52, 1237–1256.

O'Lenic, E.A. and Livezey, R.E. (1989) Relationships between systematic errors in medium range numerical forecasts and some of the principal modes of low-frequency variability of the northern hemisphere 700 mb circulation. *Monthly Weather Review*, 117, 1262–1280.

Raftery, A.E., Gneiting, T., Balabdaoui, 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.

Roebber, P.J. (1998) The regime dependence of degree day forecast technique, skill, and value. *Weather and Forecasting*, 13, 783–794.

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 Büermann, L. (2014) Spatially adaptive post-processing of ensemble forecasts for temperature. *Journal of the Royal Statistical Society: Series C*, 63, 405–422.

Scheuerer, M. and Hamill, T.M. (2015) Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Monthly Weather Review*, 143, 4578–4596.

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.

Serreze, M.C., Carse, F., Barry, R.G. and Rogers, J.C. (1997) Icelandic low cyclone activity: climatological features, linkages with the NAO, and relationships with recent changes in the northern hemisphere circulation. *Journal of Climate*, 10, 453–464.

Siegert, S., Stephenson, D.B., Sansom, P.G., Scaife, A.A., Eade, R. and Arribas, A. (2016) A Bayesian framework for verification and recalibration of ensemble forecasts: how uncertain is NAO predictability?. *Journal of Climate*, 29, 995–1012.

Sloughter, 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.

Smyth, P., Ide, K. and Ghil, M. (1999) Multiple regimes in northern hemisphere height fields via mixture model clustering. *Journal of the Atmospheric Sciences*, 56, 3704–3723.

Stoss, L.A. and Mullen, S.L. (1995) The dependence of short-range 500-mb height forecasts on the initial flow regime. *Weather and Forecasting*, 10, 353–368.

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.

Talagrand, O. (1997). Evaluation of probabilistic prediction systems. In:, *Proceeding ECMWF Workshop on predictability*. Reading, UK: ECMWF, 20-22 October 1997, pp. 1-18.

Thompson, D.W. and Wallace, J.M. (1998) The Arctic oscillation signature in the wintertime geopotential height and temperature fields. *Geophysical Research Letters*, 25, 1297–1300.

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*, 173, 371–388.

Tibaldi, S. and Molteni, F. (1990) On the operational predictability of blocking. *Tellus A*, 42, 343–365.

Vannitsem, S., Wilks, D.S. and Messner, J. (2018) *Statistical Postprocessing of Ensemble Forecasts*. Amsterdam: Elsevier.

Wallace, J.M. and Gutzler, D.S. (1981) Teleconnections in the geopotential height field during the northern hemisphere winter. *Monthly Weather Review*, 109, 784–812.

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

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

How to cite this article: Allen S, Ferro CAT, Kwasniok F. Recalibrating wind-speed forecasts using regime-dependent ensemble model output statistics. *Q J R Meteorol Soc*. 2020;146:2576–2596. https://doi.org/10.1002/qj.3806