Received: 19 August 2023 | Revised: 19 December 2023 | Accepted: 24 February 2024

DOI: 10.1002/qj.4701

Quarterly Journal of the Royal Meteorological Society RMetS

RESEARCH ARTICLE

Lead-time-continuous statistical postprocessing of ensemble weather forecasts

Jakob Benjamin Wessel | **Christopher A. T. Ferro** | **Frank Kwasniok**

Department of Mathematics and Statistics, University of Exeter, Exeter, UK

Correspondence
Jakob Benjamin Wessel, Department of Mathematics and Statistics, University of Exeter, Exeter, UK.
Email: jw1301@exeter.ac.uk

Funding information
Engineering and Physical Sciences Research Council, Grant/Award Number: 2696930

Abstract

Numerical weather prediction (NWP) ensembles often exhibit biases and errors in dispersion, so they need some form of postprocessing to yield sharp and well-calibrated probabilistic predictions. The output of NWP models is usually at a multiplicity of different lead times and, even though information is often required on this range of lead times, many postprocessing methods in the literature are applied either at a fixed lead time or by fitting individual models for each lead time. However, this is (1) computationally expensive because it requires the training of multiple models if users are interested in information at multiple lead times and (2) prohibitive because it restricts the data used for training postprocessing models and the usability of fitted models. This article investigates the lead-time dependence of postprocessing methods in the idealized Lorenz'96 system as well as temperature and wind-speed forecast data from the Met Office Global and Regional Ensemble Prediction System (MOGREPS-G). The results indicate that there is substantial regularity between the models fitted for different lead times and that one can fit models that are *lead-time-continuous* that work for multiple lead times simultaneously by including lead time as a covariate. These models achieve similar and, in small data situations, even improved performance compared with the classical *lead-time-separated* models, whilst saving substantial computation time.

KEYWORDS

ensemble prediction, probabilistic weather forecasting, recalibration, statistical postprocessing, temperature, wind speed

1 | INTRODUCTION

Ensemble forecasts have become an essential tool in meteorology and climate science, providing valuable insights into the range of possible future weather and climate conditions. These forecasts are generated from multiple runs of numerical weather prediction (NWP) models, each with slightly different initial conditions, to quantify the uncertainty inherent in predicting complex atmospheric phenomena. However, ensemble forecasts are often biased and have errors in dispersion, which necessitates the application of statistical postprocessing techniques to improve their reliability and accuracy (Vannitsem et al., 2021).

Even though postprocessed forecasts are usually required by end users at a range of different lead times, many postprocessing methods in the literature are applied either at a single lead time (Gneiting et al., 2005; Raftery et al., 2005; Baran and Lerch, 2018) or by fitting a separate statistical model for each lead time (Gebetsberger et al., 2018; Allen et al., 2020; Roberts et al., 2023). This, however, is (1) computationally expensive, because if users are interested in multiple lead times it requires the training of multiple models, and (2) prohibitive, restricting the data used for training postprocessing models and the usability of fitted models.

In this work, we will first investigate how the parameters of a typical postprocessing method — Ensemble Model Output Statistics (EMOS), also known as Nonhomogeneous (Gaussian) Regression (NGR) — vary over the lead-time range to see whether there are any regularities. These results will then be used to see whether it is possible to fit models that work continuously over lead times, achieving equal performance to models fitted separately for each lead time whilst saving computational costs. The focus of this work is on EMOS/NGR-type techniques. Much more sophisticated postprocessing methods exist, but EMOS or EMOS extensions are still used operationally by different forecasting centres (Hess, 2020; Roberts et al., 2023) and, for the purposes of studying lead-time dependence of postprocessing methods, EMOS provides a well-performing and interpretable baseline. As part of future work, it might be possible to extend this study to include various other postprocessing methods.

The problem of lead-time dependence has not been considered much in the postprocessing literature. Pinson and Girard (2012); Hemri et al. (2013, 2015), and Engeland and Steinsland (2014) are interested in the temporal dependence structure over lead times for applications to wind speed and hydrological predictions. Hemri et al. (2015) fit separate models for different lead times and then smooth postprocessing parameters using cyclic splines. Here, however, the focus is different, as we are less interested in multivariate calibration across lead times, but rather in building computationally cheaper models that account for the lead-time character within the model itself rather than requiring subsequent adjustments. For long-range forecasts ranging from monthly to decadal timescales, a lead-time-dependent drift (or even trend) correction is often performed (Schaeybroeck and Vannitsem, 2018), and here we argue that such a correction is also sensible for shorter-range forecasts, together with adjustments for the diurnal cycle. Recently Mlakar et al. (2023) described neural networks jointly postprocessing forecasts at all lead times and exploring the importance of covariates at different lead times. The authors also comment that most approaches in the literature only consider fixed lead times. Their approach is different from ours, as they do not systematically consider the effects of lead time on the postprocessing and how to account easily for these effects. Rather, this is hidden within the neural network taking lead time as input.

Most closely related to the present study is the work by Dabernig et al. (2017a), who calculate standardized anomalies accounting for seasonality and diurnal cycle via bivariate splines and define EMOS models between the ensemble and observational anomalies. However, we believe that this approach is rather restrictive and that lead-time dependence as well as the diurnal cycle can be accounted for directly in the postprocessing model itself. This is conceptually simpler and computationally cheaper, providing directly interpretable model parameters and only requiring one model fit instead of multiple ones. Furthermore, defining standardized anomalies via splines requires rather large data sets, and spline estimation, especially of bivariate splines, can be unstable—particularly in small training data situations. This makes the method from Dabernig et al. (2017a) unusable in running-window training schemes or in the presence of frequent model updates, where it might be especially interesting to increase training data by merging across lead times. Furthermore, Dabernig et al. (2017a) argue that, after removing diurnal and seasonal cycle(s) from the values, the lead-time character disappears. However, this might not be the case, especially at longer lead times, due to the existence of model drift. The approach presented here integrates smoothly into existing postprocessing frameworks, requiring only small modifications, and is less sensitive to training data limitations or model upgrades.

This article is organized as follows. Firstly, we will describe the data and methods used for postprocessing both in the idealized Lorenz'96 system and also for 2-m temperature and 10-m wind-speed forecasts from the Met Office Global and Regional Ensemble Prediction System (MOGREPS-G). Secondly, we will look at parameter development over lead times for postprocessing models in the Lorenz'96 system and see whether it is possible to fit models accounting for the lead-time character within the model. Using an idealized system is advantageous, as it allows us to isolate a lead-time effect without data limitations. This analysis will then be extended in a second step by investigating data in a real forecasting system. As seasonality plays an important role in postprocessing models using real data, this latter section will first look at models accounting for seasonality within the model and then at models trained in a running window, subsequently to compare both approaches. A discussion and some outlooks for future work are given at the end.

2 | DATA

2.1 | Lorenz'96 system

The Lorenz'96 system (Lorenz, 1996) is a highly idealized model of the atmosphere. It is often used for the trialling of statistical postprocessing methods (Roulston and Smith, 2003; Wilks, 2006; Williams et al., 2014; Allen et al., 2019) as it includes chaotic dynamics but has no restrictions on data availability for training and evaluation. It is a coupled system of larger-scale variables X_k and subgrid-scale variables $Y_{j,k}$:

$$\frac{dX_k}{dt} = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F - \frac{hc}{b}\sum_{j=1}^{J} Y_{j,k}, \quad (1)$$

$$\frac{dY_{j,k}}{dt} = -cbY_{j+1,k}\left(Y_{j+2,k} - Y_{j-1,k}\right) - cY_{j,k} + \frac{hc}{b}X_k, \quad (2)$$

for $k = 1, \dots, K$ and $j = 1, \dots, J$. The system has cyclic boundary conditions: $X_{k-K} = X_{k+K} = X_k$, $Y_{j,k-K} = Y_{j,k+K} = Y_{j,k}$, $Y_{j-J,k} = Y_{j,k-1}$, and $Y_{j+J,k} = Y_{j,k+1}$. Here parameters $K = 8$, $J = 32$, $F = 20$, $h = 1$, $b = 10$, and $c = 10$ are used to ensure comparability with Wilks (2005, 2006); Williams et al. (2014); Allen et al. (2019). The system is integrated in time using a Runge–Kutta 4(5) scheme for 100,000 model time units (MTUs) and the output is sampled on a time grid with steps of $\Delta t = 0.1$ MTU. This is taken as ground truth for the atmospheric conditions. In the Lorenz'96 system, 1 MTU corresponds to 5 days.

A truncated version of the Lorenz'96 system only resolving the larger scales, together with a simple parametrization of the subgrid terms, is used to mimic an imperfect NWP model:

$$\begin{aligned}\frac{dX_k}{dt} &= -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F \\ &\quad - \left(\beta_0 + \beta_1 X_k + \beta_2 X_k^2 + \beta_3 X_k^3 + \beta_4 X_k^4\right).\end{aligned} \quad (3)$$

The parameters $\beta_0, \dots, \beta_4$ are fitted by regressing unresolved tendencies onto powers of X_k as in Wilks (2005); Kwasniok (2012); Christensen et al. (2015); Allen et al. (2019) and very similar parameter values are obtained. Based on this NWP model, an ensemble with 20 members is generated by perturbing the atmospheric ground truth with $N(0, 0.1^2)$ noise independently for each X_k and integrating the NWP system using a Runge–Kutta 4(5) scheme up to 3 MTU (the equivalent of 15 days). The output is again sampled on a time grid with steps of $\Delta t = 0.1$ MTU (the equivalent of 0.5 days) and, as the ensemble is exchangeable, it is summarised by the mean and standard deviation. A training data set is generated using the first 10,000 MTUs of the ground truth, by initializing a forecast every 0.2 MTU ($\sim$ 1 day), and a test data set using the latter 90,000 MTUs by initializing forecasts every 50 MTUs to ensure approximate independence of test samples. This leads to a data set containing 50,000 forecast–observation pairs in the training set (only the data for variable X_1 are taken, as the X_k are exchangeable) and 12,792 in the test set (1599 forecast observation pairs for each of the eight X_k variables, merged).

2.2 | MOGREPS-G forecasts of temperature and wind speed

To analyse lead-time dependence in an operational forecast system, we use 2-m temperature and 10-m wind-speed forecasts from MOGREPS-G (Walters et al., 2017; Porson et al., 2020), issued between April 1, 2019 and April 1, 2022. MOGREPS-G forecasts consist of 18 ensemble members (including the unperturbed control run), and are initialized four times a day at 0000, 0600, 1200, and 1800 UTC. MOGREPS-G has an approximately 20-km horizontal resolution in the midlatitudes. We consider the forecast initialized at 0000 UTC and lead times from $T + 06$ to $T + 198$ hours (8.25 days) at 6-hr intervals. The forecasts are regridded bilinearly to the locations of 30 surface synoptic observation (SYNOP) stations in the United Kingdom (see Figure 1 for the locations) and verified against temperature and wind-speed observations (observations of 2-m temperature and 10-min average of 10-m wind speed). The forecast ensembles are assumed to be exchangeable and are summarised by the ensemble mean and ensemble standard deviation.

3 | STATISTICAL POSTPROCESSING METHODS

3.1 | Ensemble model output statistics

EMOS, also often known as NGR (Jewson et al., 2004; Gneiting et al., 2005) is one of the most frequently used statistical postprocessing methods. It models the variable to forecast with a parametric distribution $\mathcal{D}(\boldsymbol{\theta})$, the parameters $\boldsymbol{\theta}$ of which depend on the ensemble prediction, which is often summarised by the ensemble mean and ensemble standard deviation.

For forecasts in the Lorenz'96 system, as well as 2-m temperature forecasts from MOGREPS-G, a normal distribution is used throughout this study, although it has been argued that a distribution with heavier tails might be more

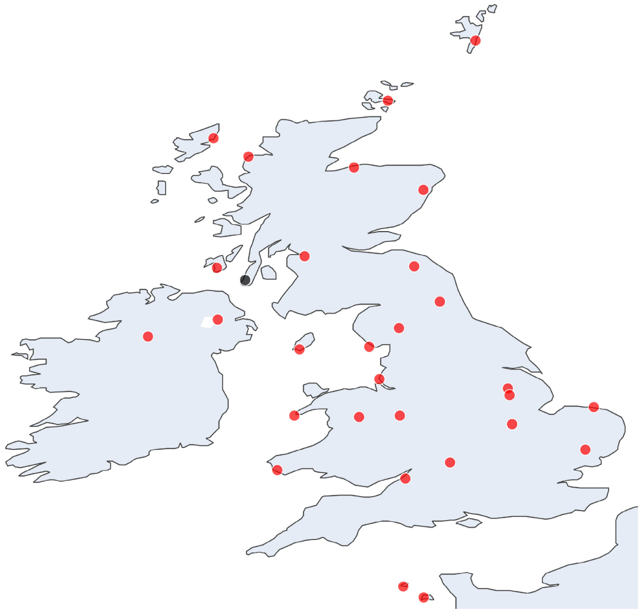


FIGURE 1 Locations of 30 SYNOP stations used as observational reference for MOGREPS-G forecasts. Station 20 considered below is indicated by a black dot.

sensible for temperature (Gebetsberger et al., 2018; Allen et al., 2021). Let T_t denote temperature or the Lorenz'96 variable X_k at a lead time t and $x_t^{(1)}, \dots, x_t^{(L)}$ the ensemble member forecasts, which we summarise by ensemble mean m_t and ensemble standard deviation s_t. The base EMOS then models the location parameter μ of a normal distribution as a linear function of the ensemble mean and the scale parameter σ as a linear function of the ensemble standard deviation:

$$T_t | x_t^{(1)}, \dots, x_t^{(L)} \sim \mathcal{N}(\mu_t, \sigma_t^2), \tag{4}$$

$$\mu_t = \alpha_t + \beta_t m_t, \tag{5}$$

$$\log \sigma_t = \gamma_t + \delta_t \log s_t. \tag{6}$$

Here the scale σ_t is log-transformed to ensure $\sigma_t > 0$ and, to make sure that parameters are on the same scale, the predictive ensemble standard deviation is log-transformed as well. This is nonstandard, as most of the time an identity link, together with optimization constraints on the parameters, is used in EMOS. However, a log link has been used in multiple publications (Gebetsberger et al., 2018; Lang et al., 2020) and comparisons for the Lorenz'96 system and MOGREPS-G forecasts found no performance differences between identity, quadratic, and log link. Therefore, the latter is used in this study because the model is simpler to handle, especially when adding covariates, whilst naturally ensuring positivity.

For wind speed, a truncated normal distribution is used throughout this study, which has been shown to work well in Scheuerer and Möller (2015) and earlier in Thorarinsdottir and Gneiting (2010) and Lerch and Thorarinsdottir (2013), although the variable of interest is maximum daily rather than average wind speed in the latter two. Let W_t denote wind speed at a lead time t and $x_t^{(1)}, \dots, x_t^{(L)}$, m_t, s_t, as above, the corresponding forecasts and summary statistics. The EMOS model for wind speed is

$$W_t | x_t^{(1)}, \dots, x_t^{(L)} \sim \mathcal{N}_0(\mu_t, \sigma_t^2), \tag{7}$$

$$\mu_t = \alpha_t + \beta_t m_t, \tag{8}$$

$$\log \sigma_t = \gamma_t + \delta_t \log s_t. \tag{9}$$

These models are fitted by minimizing the continuous ranked probability score (CRPS, see Equation 21), using the R package `crch` (Messner et al., 2022). Maximum-likelihood estimation was also explored and was found not to lead to substantial differences in performance. EMOS models are most of the time fitted separately for each station and lead time t. Within the Lorenz'96 system, the K variables are exchangeable, and thus a model fitted to any of them (or the merged data set) can be used for the other X_k.

3.2 | Lead-time-continuous postprocessing

In the following, models will be developed that only require a single fit to a data set of merged forecast–observation pairs of all lead times and are able to perform postprocessing for all of these. Such models will be referred to as *lead-time-continuous*. The base EMOS fitted separately for all lead times will be referred to as *lead-time-separated*.

Figure 2 shows the development of the parameters $\alpha_t, \beta_t, \gamma_t, \delta_t$ of lead-time-separated EMOS models fitted to X_1 in the Lorenz'96 system. As one can see in this situation without substantial data limitations the parameters vary fairly smoothly as a function of lead time. For the location parameter, the intercept α_t starts at zero and the multiplicative parameter β_t at one for small lead times, meaning that the ensemble mean captures the true dynamics fairly well. β_t then even rises a bit around 2 days, before falling linearly over lead time. This reflects the decreasing skill of the ensemble mean in predicting the observations, whilst the constant correction provided by the intercept α_t rises. Due to the log link, the scale parameters are less interpretable. For small lead times, the intercept γ_t is strongly negative and the multiplicative parameter δ_t near zero, so σ_t is very small, as the ensemble mean presents a very good prediction with little uncertainty. δ_t then rises very fast, peaking at around a lead time of 3 days before approaching a seemingly constant value. The intercept γ_t seems fairly

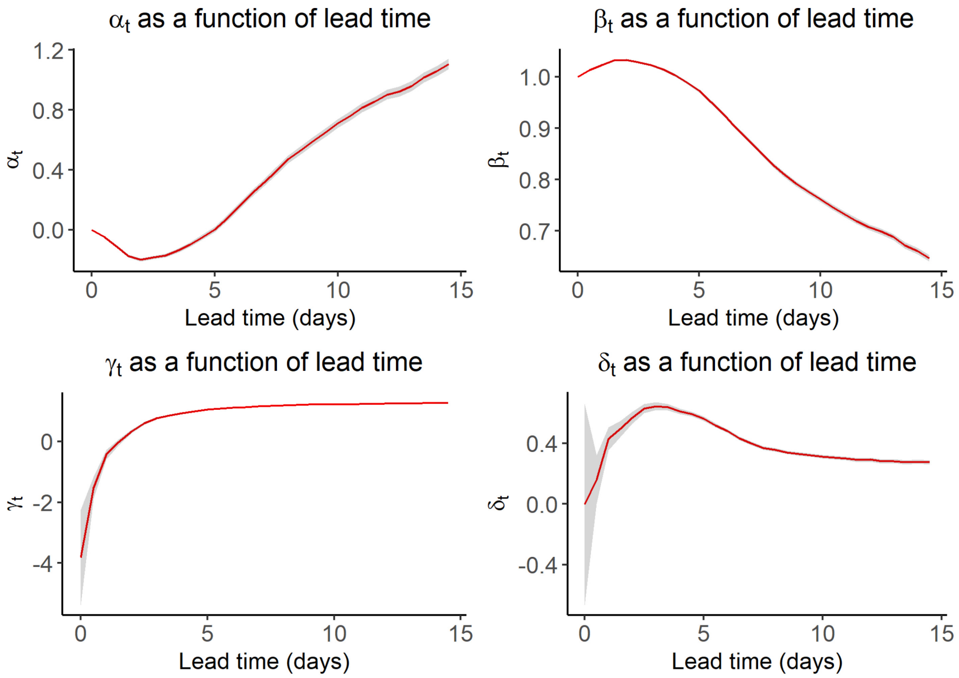


FIGURE 2 Dependence on lead time of parameters of EMOS postprocessing models fitted separately for each lead time in the Lorenz'96 system. The shaded areas indicate pointwise 95% confidence intervals.

stable after 5 days, slightly larger than one, indicating that the log ensemble spread captures a big part of the true uncertainty, but a constant correction is needed.

Given the smooth variation of the EMOS parameters as a function of lead time, one can try to account for this character using splines. The following lead-time-continuous EMOS models, including lead time t as a covariate for both μ_t and σ_t, will be analysed:

$$\text{Model 1:}\quad \mu_t = \alpha + \beta m_t + p_1(t), \tag{10}$$

$$\log \sigma_t = \gamma + \delta \log s_t + p_2(t). \tag{11}$$

$$\text{Model 2:}\quad \mu_t = \alpha + p_1(t) m_t + p_2(t), \tag{12}$$

$$\log \sigma_t = \gamma + p_3(t) \log s_t + p_4(t). \tag{13}$$

Here the functions $p_1, \dots, p_4$ are thin plate splines (Wood, 2003). These models are fitted with penalized maximum likelihood using iteratively weighted least squares (IWLS) and the backfitting algorithm. The amount of smoothing is chosen in a data-driven way by optimizing an adjusted AIC whilst fitting the model (see Rigby & Stasinopoulos, 2005 for an overview of fitting and smoothing estimation). The `bamlss` package (Umlauf et al., 2018) is used. Model 1 tries to account for the effect of lead time by using an additive correction over the level of the intercept α, thus representing lead-time-dependent change in the latter (the splines themselves have no constant term). Model 2 extends model 1 by also modelling the multiplicative EMOS parameters, β and δ, as splines of lead time t. We use thin plate splines, as they allow us to account for flexible nonlinear lead-time-dependent effects. These models will be evaluated in Section 5.1.

Seasonality plays an important role in real data and so lead-time-separated and lead-time-continuous models with seasonal components will be presented for the MOGREPS-G forecasts in the next section.

3.3 | Seasonality and diurnal cycle

Parameters of postprocessing methods but also NWP forecasting skill tend to vary depending on the time of year, which means that methods like EMOS need to account for seasonality. This is usually done by fitting models in a running window—as originally proposed by Gneiting et al. (2005)—or by including terms in the model that adjust for seasonal cycles as in Gebetsberger et al. (2018); Lang et al. (2020), and Chen et al. (2022). Dabernig et al. (2017b) and Messner et al. (2017) instead work on standardized anomalies to which they fit EMOS models. Such a strategy is also employed by Dabernig et al. (2017a)—one of the only publications looking at simultaneous postprocessing for multiple lead times—but the authors believe that using a running window or including adjustments in the model has advantages, as it is more interpretable and requires fewer model fits (see Section 1).

When including adjustments within the model, EMOS is trained on a fixed training period and applied to a testing data set: here we used data from April 1, 2019–December 31, 2020 for training and from January 1, 2021–April 1, 2022 for testing. Applying EMOS in a running window requires continuous retraining without a fixed training and testing period. Window sizes between 30 and 45 days have often been found to yield optimal performance for temperature postprocessing (Gneiting et al., 2005) and here we use 40 days, which is a common choice in the literature (Lang et al., 2020). Wind speed has been found in some cases to have a less pronounced seasonal cycle (Allen et al., 2022) and might be fitted in a fixed window. However, Scheuerer and Möller (2015) find a running window of 60 days to be optimal and in operational systems even shorter training window lengths are sometimes used, such as 15 days in the UK Met Office IMPROVER system (Roberts et al., 2023). As a compromise, we also work with a training-window size of 40 days for wind speed.

In the following, we will first outline EMOS models accounting for seasonality within the model together with lead-time-continuous models used in that case, to then show models using a running window.

3.3.1 | Seasonality within the model

To account for seasonal variation, sine and cosine transformations of the day of year (doy) are included in both the location and scale parameters of the EMOS models. This is similar to Gebetsberger et al. (2018), who, however, only allow seasonal variation in the location parameter. Cyclic splines as in Lang et al. (2020) were also tested, but, most likely due to the limited training data (1 year and 8 months), their estimates proved not fully reliable and using them did not improve performance. Thus the conceptually easier parametric model is used. For each lead time t and for both temperature and wind speed, the following model is fitted for location and scale parameters:

$$\mu_t = \alpha_t + \beta_t m_t + \lambda_{\mu,\mathrm{c},t} \cos\left(2\pi \frac{\mathrm{doy}}{366}\right) + \lambda_{\mu,\mathrm{s},t} \sin\left(2\pi \frac{\mathrm{doy}}{366}\right), \tag{14}$$

$$\log \sigma_t = \gamma_t + \delta_t \log s_t + \lambda_{\sigma,\mathrm{c},t} \cos\left(2\pi \frac{\mathrm{doy}}{366}\right) + \lambda_{\sigma,\mathrm{s},t} \sin\left(2\pi \frac{\mathrm{doy}}{366}\right). \tag{15}$$

As more flexible seasonal effects using cyclic splines did not lead to increased performance over simple sine–cosine transformations, nor give evidence for the role of higher harmonics in the seasonal cycle, only first-order ones are used. These models are fitted separately for each station and lead time on the training data set of MOGREPS-G forecasts between April 1, 2019 and December 31, 2020. Parameters are estimated using minimum CRPS estimation using the R package `crch` (Messner et al., 2022). This base model will be referred to as Separated–Seasonality Within the Model (S-SWM).

Figure 3 shows the evolution of parameters $\alpha_t, \beta_t, \gamma_t$, and δ_t as a function of lead time t at station 20, which seems typical within the data set for both temperature and wind speed. Two effects can be seen. Firstly there seems to be a substantial diurnal cycle in the parameters—especially for the location parameters α_t and β_t, but also for the scale parameters γ_t and δ_t, even though it is less pronounced there. This cycle can be interpreted directly in terms of forecasting skill: during the day (1200, 1800 UTC) ensemble forecasts are less skilful, leading to larger intercepts α_t and smaller multiplicative parameters β_t, whilst the

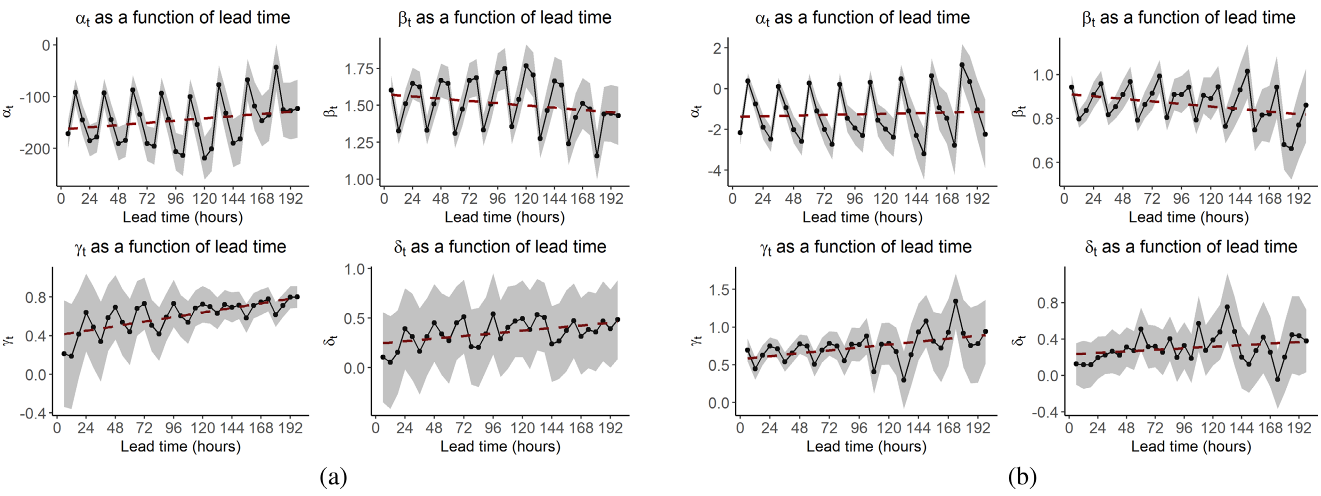


FIGURE 3 Development of EMOS parameters as a function of lead time for models fitted separately for different lead times for (a) temperature and (b) wind speed (right panel). The dashed line represents the best linear fit. The shaded areas indicate pointwise 95% confidence intervals.

behaviour is reversed during the night. Secondly, over the lead-time range there seem to be general trends in the parameters. In the location, the intercept α_t goes down and then up for temperature, whilst staying flat for wind speed. Similarly, the multiplicative parameter β_t goes down for wind speed and up and then down for temperature. In the scale, the intercept parameter γ_t rises for both temperature and wind speed. This indicates decreasing forecasting skill and increasing uncertainty as the lead time grows. The δ_t parameter has more uncertainty than the other parameters. It is quite small for small lead times, particularly for temperature, indicating that at these lead times the ensemble standard deviation is not indicative of the true uncertainty. There is an indication of an upwards trend over the lead time; however, given the large confidence bands this might not be robust.

Given this parameter evolution, it seems that when building lead-time-continuous models it is important to account for the effects of the diurnal cycle and general drift. During model building it was found to be beneficial to account for the two effects separately. There are different ways of doing so, but, given that this data set is at six-hourly resolution, we decided to include a factor or categorical variable, ψ_{tod}, accounting for the time of day tod, and to model the effect of lead time, t, with a linear approximation. If more lead times were available, one could use factors with more levels, random effects to regularise the factor levels, sine–cosine transformations of the time of day, or spline terms. However, even with a MOGREPS-G data set at one-hourly resolution, accounting for the diurnal cycle using a factor for the hour of the day was found to perform reasonably well.

The following model is fitted as lead-time-continuous for both temperature and wind speed separately to each station, again using minimum CRPS estimation:

$$\begin{aligned}\mu_t = \alpha + \beta m_t + \psi_{\mu,\text{tod}} + \phi_\mu t \\ + \lambda_{\mu,\text{c}} \cos\left(2\pi \frac{\text{doy}}{366}\right) + \lambda_{\mu,\text{s}} \sin\left(2\pi \frac{\text{doy}}{366}\right) \\ + \psi_{\mu,\text{c,tod}} \times \lambda_{\mu,\text{c}} \times \cos\left(2\pi \frac{\text{doy}}{366}\right) \\ + \psi_{\mu,\text{s,tod}} \times \lambda_{\mu,\text{s}} \times \sin\left(2\pi \frac{\text{doy}}{366}\right),\end{aligned} \tag{16}$$

$$\begin{aligned}\log \sigma_t = \gamma + \delta \log s_t + \psi_{\sigma,\text{tod}} + \phi_\sigma t \\ + \lambda_{\sigma,\text{c}} \cos\left(2\pi \frac{\text{doy}}{366}\right) + \lambda_{\sigma,\text{s}} \sin\left(2\pi \frac{\text{doy}}{366}\right) \\ + \psi_{\sigma,\text{c,tod}} \times \lambda_{\sigma,\text{c}} \times \cos\left(2\pi \frac{\text{doy}}{366}\right) \\ + \psi_{\sigma,\text{s,tod}} \times \lambda_{\sigma,\text{s}} \times \sin\left(2\pi \frac{\text{doy}}{366}\right).\end{aligned} \tag{17}$$

This model will in the following be referred to as Continuous–Seasonality Within the Model (C-SWM). It corresponds to a standard EMOS with non-lead-time-dependent intercepts (α, γ) and multiplicative parameters (β, δ), which is, however, fitted on a data set of merged lead times. Similarly to the lead-time-separated model, sine–cosine transformations of the day of year are included to account for seasonality, with non-lead-time-dependent parameters. To account for the effects of the diurnal cycle, the time of day tod is treated as a factor variable (ψ_{tod}) with four levels for the four forecast times each day (0000, 0600, 1200, 1800 UTC) and included as a main effect and in interaction with each of the seasonal terms to adjust for seasonal variations in the diurnal cycle. This leads to time-of-day-dependent intercepts and seasonality parameters for both location and scale. To account for lead-time-dependent drift, a linear term $(\phi_{\mu,\sigma} t)$ is included. Except for the time-of-day–seasonality interaction, only additive lead time adjustments are used and especially the multiplicative parameters β and δ do not vary with lead time, seasonal cycle, or diurnal cycle. This will be discussed later.

3.3.2 | Seasonality in a running window

In a running window, the base EMOS (Section 3.1) is fitted separately for each lead time. This model will be abbreviated as Separated–Running Window (S-RWIN). For lead-time-continuous postprocessing, the following model is fitted for both temperature and wind speed, abbreviated Continuous–Running Window (C-RWIN):

$$\mu_t = \alpha + \beta m_t + \psi_{\mu,\text{tod}} + \phi_\mu t, \tag{18}$$

$$\log \sigma_t = \gamma + \delta \log s_t + \psi_{\sigma,\text{tod}} + \phi_\sigma t. \tag{19}$$

Here the time of day is included again as a factor or categorical variable (ψ_{tod}) to account for the diurnal cycle and $\phi_\mu t$ and $\phi_\sigma t$ are terms accounting for forecast drift or trends. Again, all adjustments for lead-time effects are additive and the multiplicative parameters β and δ are constant across lead time. Both lead-time-separated and lead-time-continuous models are fitted using minimum CRPS estimation on a running window of length 40 days.

An overview of all lead-time-separated and lead-time-continuous models considered, together with their abbreviations, is provided in Table 1.

TABLE 1 Overview of postprocessing models fitted to MOGREPS-G data.

Abbreviation	Model	Training period	Stations	Lead time
S-SWM	EMOS including seasonality terms	Fixed	Separate for each station	Separated
C-SWM	EMOS including seasonality terms, time of day, and lead time	Fixed	Separate for each station	Continuous
S-RWIN	EMOS	Running window	Separate for each station	Separated
C-RWIN	EMOS including time of day and lead time	Running window	Separate for each station	Continuous

4 | VERIFICATION

The goal of statistical postprocessing has been phrased by Gneiting et al. (2005) as maximizing *sharpness* subject to *calibration*. Calibration is usually assessed by investigating probability integral transform (PIT) histograms (Gneiting et al., 2007; Thorarinsdottir and Schuhen, 2018), which are histograms of the predictive cumulative distribution function (CDF) evaluated at the observation. These histograms should be uniform if the forecast is calibrated, whilst a U-shape indicates underdispersion and an inverse U-shape overdispersion. In the case of ensemble forecasts, verification rank histograms can be used instead, which show the distribution of the rank of the observation within the ensemble (Anderson, 1996; Hamill and Colucci, 1997). These should again be uniform under the assumption of calibration. The amount of (mis)calibration shown by PIT histograms can also be summarised numerically by the reliability index (RI, Thorarinsdottir & Schuhen, 2018; Delle Monache et al., 2006), which is

$$\mathrm{RI} = \sum_{i=1}^{I} \left| \zeta_i - \frac{1}{I} \right|. \qquad (20)$$

Here I is the number of bins in the histogram and ζ_i the observed proportion of PIT values in bin i.

Proper scoring rules can be used to assess calibration and sharpness together. One of the most common proper scores is the CRPS (Hersbach, 2000):

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} [F(z) - 1\{y \leq z\}]^2 dz. \qquad (21)$$

Here F represents the issued predictive CDF and y the corresponding observation. The CRPS is negatively oriented, meaning that a smaller score implies better performance. Closed-form expressions exist for the CRPS for forecasts issued as normal (Gneiting et al., 2005) and truncated normal (Thorarinsdottir and Gneiting, 2010) distributions. When comparing a model F_1 with a baseline model F_0 we can also define a skill score:

$$\mathrm{CRPSS} = 1 - \frac{\sum_{i=1}^{n} \mathrm{CRPS}(F_1^{(i)}, y^{(i)})}{\sum_{i=1}^{n} \mathrm{CRPS}(F_0^{(i)}, y^{(i)})}. \qquad (22)$$

Here $F_0^{(1)}, F_1^{(1)}, \ldots, F_0^{(n)}, F_1^{(n)}$ are predictive CDFs for observations $y^{(1)}, \ldots, y^{(n)}$ given by model F_1 and F_0. The continuous ranked probability skill score (CRPSS) can be interpreted directly as the relative improvement over the baseline, with 100 × CRPSS representing the percentage improvement or decrease.

For evaluating multivariate forecasts across all lead times (the forecasts issued at a common issue date), the energy score (Gneiting et al., 2008) and the p-Variogram score (Scheuerer and Hamill, 2015) are used:

$$\mathrm{ES}(F, y) = \mathbb{E}_{\mathbf{X} \sim F} \|\mathbf{X} - \mathbf{y}\| - \frac{1}{2} \mathbb{E}_{\mathbf{X}, \mathbf{X}' \sim F} \|\mathbf{X} - \mathbf{X}'\|, \qquad (23)$$

$$Var_p(F, y) = \sum_{i,j=1}^{d} w_{ij} \left(|y_i - y_j|^p - \mathbb{E}_{\mathbf{X} \sim F} |X_i - X_j|^p \right)^2, \qquad (24)$$

with non-negative weights w_{ij}. These are all taken as $w_{ij} = 1$ here and $p = 0.5$. The X_i and y_i correspond to the ith elements of the d-dimensional vectors. In this work, evaluation of the energy and p-Variogram scores is done via their formulation for an ensemble of L multivariate samples $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(L)}$ from F:

$$\mathrm{ES}(F, y) = \frac{1}{L} \sum_{i=1}^{L} \|\mathbf{X}^{(i)} - \mathbf{y}\| - \frac{1}{2L^2} \sum_{i,j=1}^{L} \|\mathbf{X}^{(i)} - \mathbf{X}^{(j)}\|, \qquad (25)$$

$$\mathrm{Var}_p(F, y) = \sum_{i,j=1}^{d} w_{ij} \left(|y_i - y_j|^p - \frac{1}{L} \sum_{k=1}^{L} \left| X_i^{(k)} - X_j^{(k)} \right|^p \right)^2. \qquad (26)$$

Here we use an ensemble of $L = 500$ draws from the predictive distributions issued. The R package `scoringRules` (Jordan et al., 2019) is used for evaluating the CRPS, energy, and p-Variogram scores.

To assess forecasts of certain values, the threshold-weighted CRPS (Gneiting and Ranjan, 2011),

$$\mathrm{twCRPS}(F,y) = \int_{-\infty}^{\infty} [F(z) - 1\{y \le z\}]^2 \omega(z) dz, \qquad (27)$$

can be used for some non-negative weight function $\omega(z)$. Here we use an indicator function as weight function $\omega(z) = \mathbb{1}(z \ge \tau)$ with some threshold τ, to evaluate behaviour in the upper tail only. To our surprise, no closed-form expressions of the threshold-weighted CRPS for predictive normal and truncated normal distributions could be found in the literature. Thus we derived expressions in these cases and verified them against numerical integration. These expressions are given in the Appendix for reference.

5 | RESULTS AND DISCUSSION

5.1 | Lorenz'96 system

After introducing both the data as well as methods used in this study, we now come to results in the Lorenz'96 system. EMOS models 1 and 2 (lead-time-continuous models, Equations 10–13) are fitted and compared with lead-time-separated EMOS models (Equations 4–6). Figure 4 shows the CRPS as a function of lead time for the different models.

The performance of the lead-time-continuous model 2 is virtually identical to that of the lead-time-separated model. The performance of model 1 (which excludes interactions between lead time and ensemble covariates in both location and scale parameters) is slightly worse. Models without an effect for lead time for one of μ_t or σ_t were also tested, but yielded strongly worse performance and calibration (not shown).

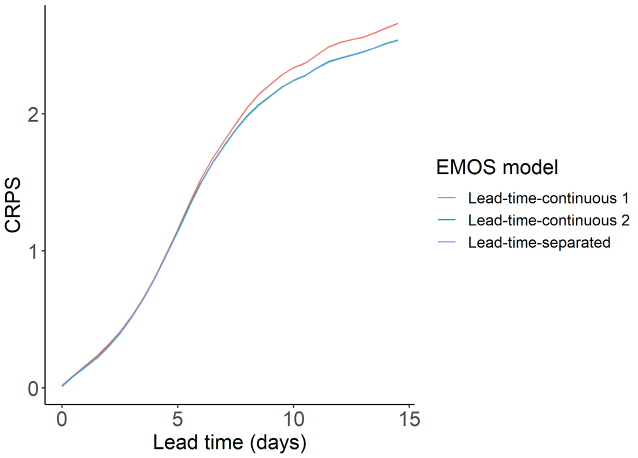


FIGURE 4 CRPS as a function of lead time for the lead-time-continuous models 1 and 2 compared with lead-time-separated EMOS in the Lorenz'96 system. The green line corresponding to the lead-time-continuous model 2 is hidden behind the blue line representing the lead-time-separated model, thus indicating equal performance in terms of CRPS.

To assess calibration, we investigate PIT histograms. Figure 5 shows the histograms for a lead time of 5 days (histograms for other lead times are similar). The PIT histogram for model 1 has an indication of overdispersion, some of which also seems to remain for the lead-time-continuous model 2, which, however, is better calibrated. It seems that the interactions included in model 2 are important to capture the main aspects of the distribution. The lead-time-separated EMOS seems to have good calibration (RI 0.061) and a slight edge over the lead-time-continuous model 2 (RI 0.072), however not strongly so.

5.2 | MOGREPS-G forecasts

5.2.1 | Seasonality within the model

Now we compare lead-time-continuous and lead-time-separated postprocessing for the MOGREPS-G forecasts. We consider the seasonal models (S-SWM and C-SWM) in this section and the running-window models (S-RWIN and C-RWIN) in the next section.

Figure 6 shows the CRPS averaged over all stations as a function of lead time, for lead-time-separated (S-SWM) and lead-time-continuous (C-SWM) models. Across all lead times, no substantial differences in performance between the lead-time-continuous and lead-time-separated models can be seen for both temperature and wind speed.

Figure 7 shows the PIT histograms (with 20 bins) merged across locations for 48-hr lead time for both lead-time-continuous (C-SWM) and separated (S-SWM) models for temperature and wind speed. In both models, some miscalibration remains for temperature and wind speed. Temperature forecasts seem to have a left tail that is too light and a right tail that is too heavy, which indicates that a skewed distribution might be more appropriate for the data, which has also been found in Allen et al. (2021). Wind-speed forecasts seem to have some amount of overdispersion after postprocessing, indicating overcorrection of the usually underdispersive raw ensemble. No substantial differences between the lead-time-continuous and separated model can be seen for temperature and wind speed. This remains similar over the lead times (not shown). The distribution across locations of differences in the reliability index RI between S-SWM and C-SWM

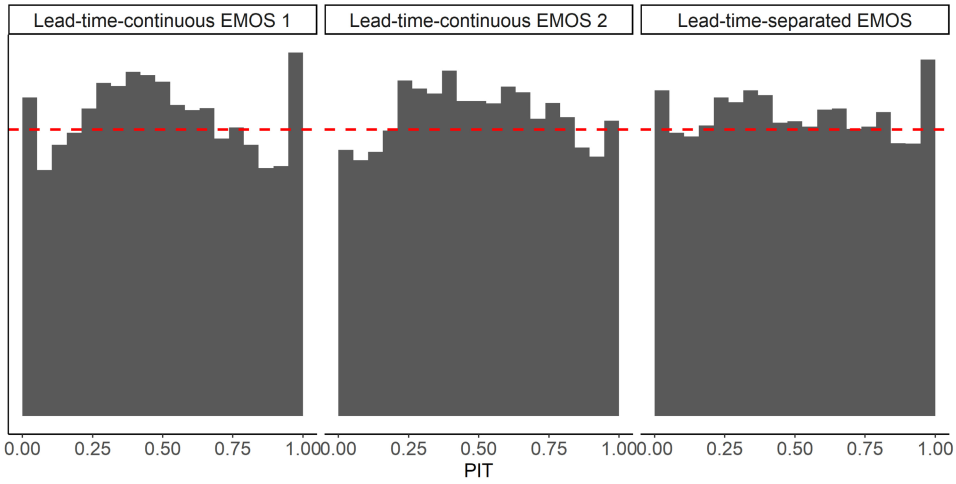
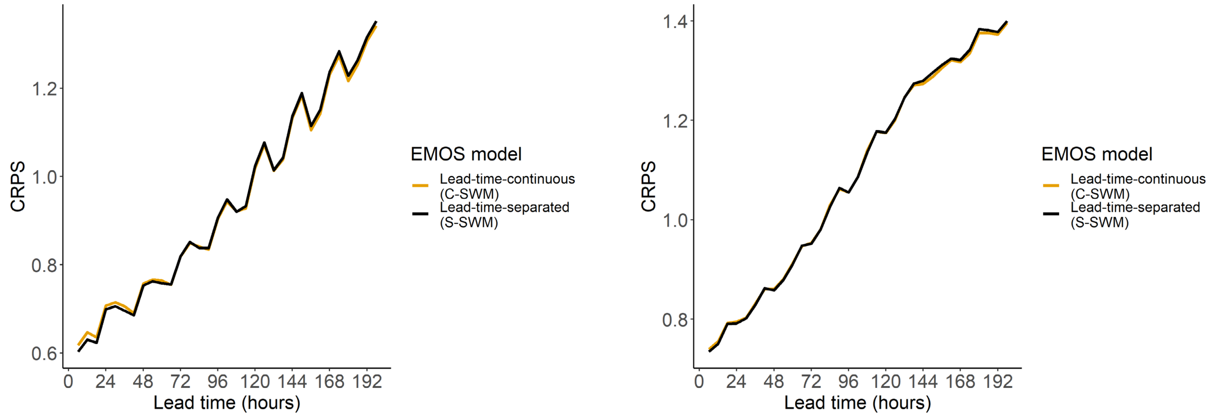


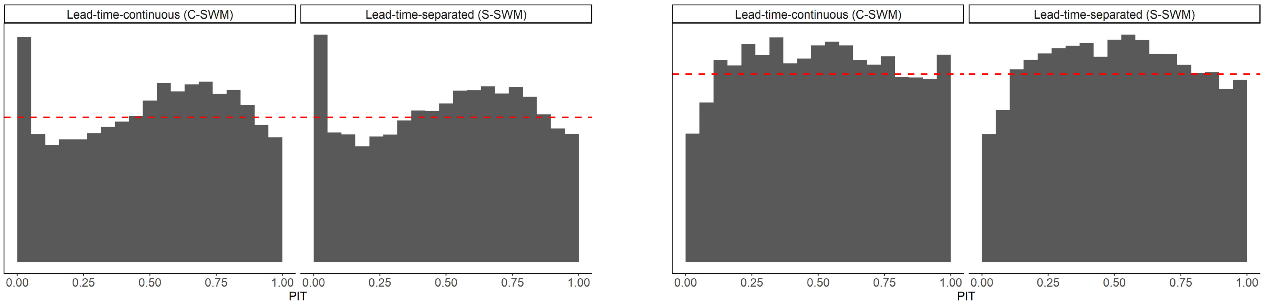
FIGURE 5 PIT histograms for a lead time of 5 days for the lead-time-continuous models 1 and 2 and the lead-time-separated EMOS in the Lorenz'96 system. The red horizontal line indicates perfect calibration.



(a)

(b)

FIGURE 6 CRPS as a function of lead time averaged over all stations for lead-time-continuous (C-SWM, orange) and lead-time-separated EMOS (S-SWM, black) that include seasonality adjustments within the model. (a) Temperature. (b) Wind speed.



(a)

(b)

FIGURE 7 PIT histograms for 48-hr predictions merged across stations for the lead-time-continuous (C-SWM, left) and lead-time-separated EMOS (S-SWM, right) that include seasonality adjustments within the model. The red horizontal lines indicate perfect calibration. (a) Temperature. (b) Wind speed.

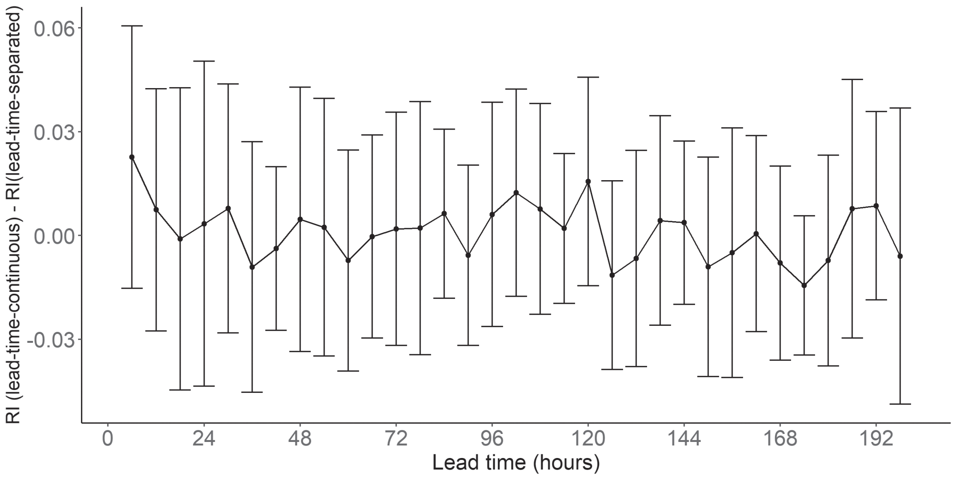


FIGURE 8 Distribution across locations of the difference in reliability index (RI) between lead-time-continuous (C-SWM) and lead-time-separated (S-SWM) temperature models including seasonality adjustments within the model, shown as a function of lead time. The solid line represents the mean, whilst the bars indicate one standard deviation.

is plotted against lead time in Figure 8 for the temperature forecasts. If this difference is negative, it means that the lead-time-continuous model has better calibration and vice versa if it is positive. As one can see, no substantial difference occurs at any lead time. This is the same for wind speed (not shown).

Training for the models was done on a portable computer under a 64-bit Windows 10 operating system (Intel Quad Core i5-1145G7 @ 2.60 GHz, 16 Gb RAM), without parallelisation. For temperature, the mean training time per location was 2.69 s (standard deviation 0.178 s) for the lead-time-continuous models (C-SWM) and 31.89 s (3.82 s) for the separated ones (S-SWM). For wind speed, training times were 7.45 s (1.79 s) for C-SWM and 99.22 s (7.61 s) for S-SWM models. This amounts to computational savings of around 90%, even though it is important to note that the times are generally quite small. Out-of-sample model application times are negligible compared with the model training times but were also improved using the lead-time-continuous models. In an operational context, where models might need to be loaded from disk into memory for application, this might be important.

Alternative models to the lead-time-continuous model C-SWM were also tested. Firstly, for temperature the interaction between the diurnal and seasonal cycles in Equations 16 and 17 seems to be crucial for the model performance. Models without this interaction have significantly worse performance, especially at earlier lead times—both overall, but even more strongly at stations with a considerable diurnal cycle. For wind speed, it seems that the interaction does not add much to the model performance and models removing it have similar (but not better) performance at all lead times. This is most likely due to the less pronounced diurnal and seasonal cycles that wind speed has. Secondly, it seems that the main lead-time/drift effect $\phi_{\mu,\sigma}t$ is not too important for model performance in terms of CRPS, even though parameter plots (see Figure 3) indicated such an effect. Models removing this effect did have similar (but not better) performance and $\phi_{\mu,\sigma}$ was generally estimated to be very small for both temperature and wind speed. This is different from models in a running window. Experiments using splines to capture an additive lead-time effect—albeit fitted using maximum likelihood—did also not indicate substantial nonlinear effects after removing the diurnal cycle and, when not accounting explicitly for the diurnal cycle (e.g., using the above tod factor), splines mainly reproduced it and had worse performance in terms of CRPS than C-SWM models in Equations 16 and 17. Models including interactions between the diurnal cycle or lead time and the ensemble mean and standard deviation were tested; however, no increase in performance was observed. This stands in contrast to the parameter plots in Figure 3 indicating a diurnal cycle and trend in β_t and δ_t. One explanation is that these multiplicative parameters might not be as crucial as one might expect and the adjustments provided by them can be partly obtained by additive corrections. Related to this is the fact that experiments fixing the multiplicative ensemble mean parameter β to different values and estimating all other parameters through CRPS minimization led to only slightly-worse performing models (in terms of CRPS). Especially for the C-SWM models, the CRPS seemed very robust to different specifications of β (not shown). This is somewhat surprising, as the EMOS model contains very few parameters in total, which, however, seem to take similar roles in correcting and recalibrating the forecast. Thus, even though in the lead-time-continuous model (C-SWM, Equations 16 and 17) the multiplicative correction is lead-time-independent, this aspect can be adjusted for using additive corrections. Simplified models without the main lead-time effect and without interactions between diurnal and seasonal cycle for wind speed have been tried; however, interestingly they did not improve in performance compared with the full

C-SWM model in Equations 16 and 17. It seems that, given the amount of training data available, this model adjusts well to the noninformativeness of some covariates.

To understand the performance of the lead-time-separated models (S-SWM) better, we analyse how lead-time-dependent their performance is. More concretely, we look at the lead times t' at which a lead-time-separated EMOS, trained on a data set consisting of forecast–observation pairs of lead time t, performs well for postprocessing. All forecasts considered here are initialised at the same time of day (0000 UTC), which means that each lead time corresponds to a single time of day. First experiments show that a model trained for a given time of day (e.g., 1200 UTC), but employed at another (e.g., 1800 UTC) have a strongly worse performance there compared with any model trained for this time of day (any model trained for 1800 UTC). Figure 9 therefore shows the CRPS skill score of models trained for a certain lead time (e.g., for 6 hr, x-axis) and employed at another lead time (e.g., 30 hr) relative to the appropriate model for this lead time (e.g., 30 hr), grouped by time of day (figure panels). The ith model (colour) here indicates the ith model for a certain time of day, so the ith model for time of day 0600 UTC is the model for lead time $i \times 24h + 6h$. Several effects can be observed. Firstly, it seems that models trained for later lead times have much worse performance at earlier lead times (up to 15% worse) and models trained for earlier lead times (especially category 0 and 1) have worse performance at longer lead times; however, this is not symmetric, as performance gets only up to 5% worse. Secondly, it seems for models in the middle category (model numbers 3–6) that before the lead time they are trained for they usually have strongly negative skill scores, but afterwards seem to give decent performance. This means, for example, that for 0600 UTC a model trained for lead time $2 * 24 + 6 = 54h$ can also be used at $j \times 24 + 6$ for $j \geq 2$. This might also add an explanation for the performance of lead-time-continuous EMOS without drift term $\phi_{\mu,\sigma} t$ because, even though this term is insinuated by the parameter plots (Figure 3), early to middle lead time models (necessarily without drift) still have strong performance at later lead times, thus the lead-time-continuous model (C-SWM) does not need to reproduce it.

5.2.2 | Seasonality in a running window

Now we compare lead-time-separated and lead-time-continuous postprocessing for the running-window models (S-RWIN and C-RWIN). Figure 10 shows the CRPS as a function of lead time for lead-time-continuous (C-RWIN) and separated (S-RWIN) model for temperature (Figure 10a) and wind speed (Figure 10b). Performance is similar at short lead times, but the lead-time-continuous models have considerably improved performance at longer lead times. At a lead time of 6 days, for example, the improvement corresponds to a reduction in CRPS by 4.4% for temperature and 5.6% for wind speed, fairly stable across all stations.

Figure 11 again shows the PIT histograms merged across locations for 48-hr predictions of temperature (Figure 11a) and wind speed (Figure 11b). There is again some slight indication of skew for temperature, and for

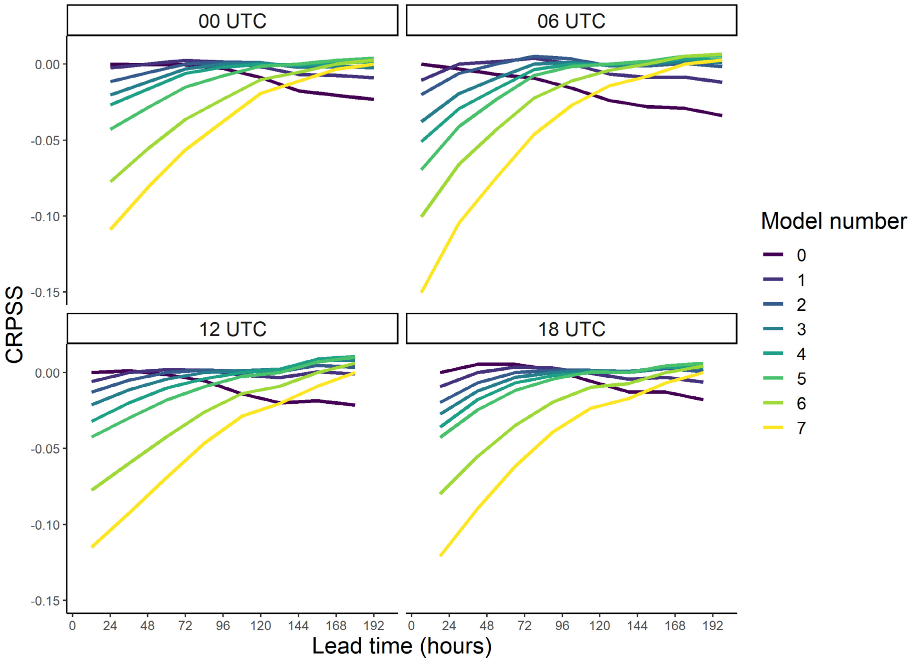


FIGURE 9 CRPSS of the ith model of each time-of-day category for predictions at another lead time t, respective to the lead-time-separated model trained for t. The ith model of each time-of-day category hereby refers to the model trained for lead time $24h \times i +$ tod: for example model 2 for 1200 UTC corresponds to the model trained for $2 \times 24h + 12h = 60h$.

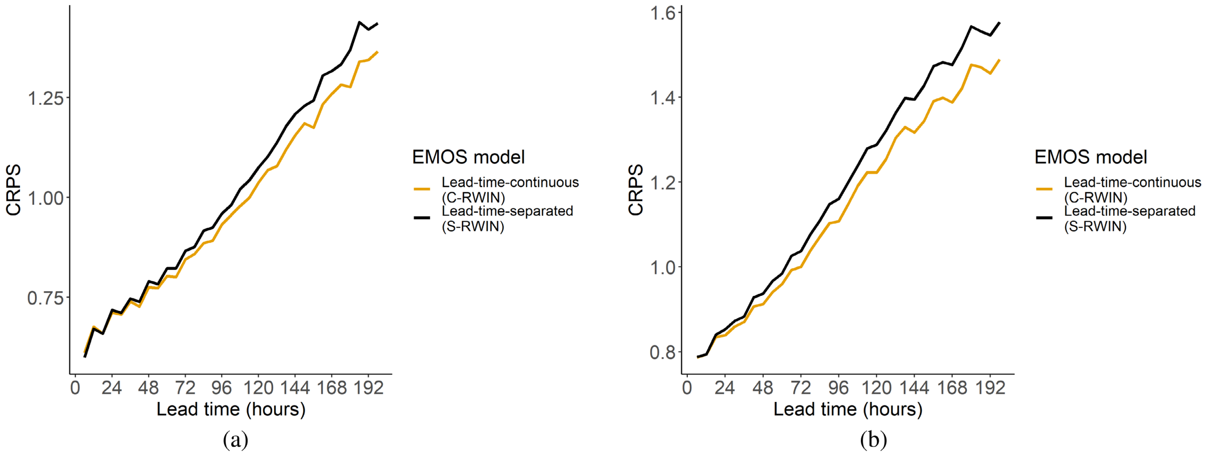


FIGURE 10 CRPS as a function of lead time averaged over all stations for lead-time-continuous (C-RWIN, orange) and lead-time-separated EMOS (S-RWIN, black), trained in a running window of 40 days. (a) Temperature. (b) Wind speed.

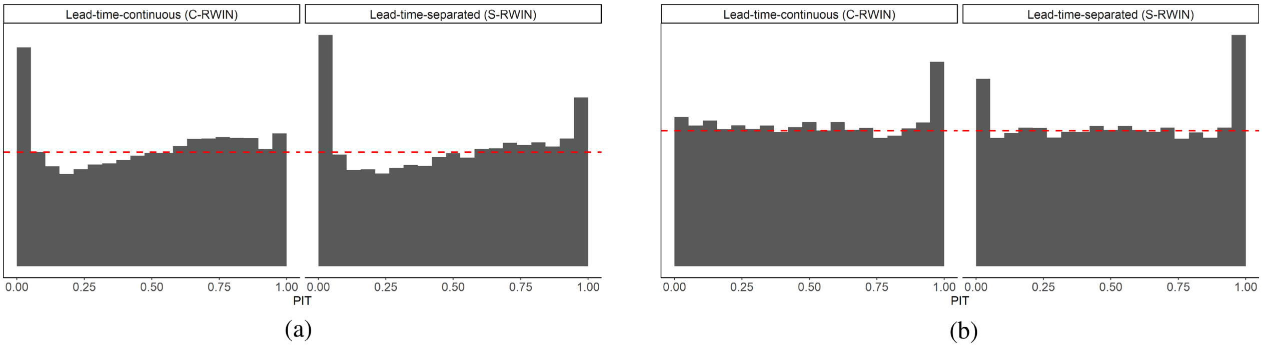


FIGURE 11 PIT histograms for 48-hr predictions merged across all stations for the lead-time-continuous (C-RWIN, left) and lead-time-separated EMOS (S-RWIN, right) trained in a running window of 40 days. The red horizontal lines indicate perfect calibration. (a) Temperature. (b) Wind speed.

wind speed it seems that the right tail might not be heavy enough. However, overall the calibration seems better than in the case where seasonality is included within the model. The lead-time-continuous models (left, C-RWIN) seem to have slightly better calibration than the lead-time-separated ones (right, S-RWIN), especially in the higher quantiles. This is similar across lead times, as can be seen in Figure 12, where the difference in reliability index seems to indicate slightly improved calibration by the lead-time-continuous models, with the trend in the difference in reliability index corresponding somewhat to the trend in CRPS in Figure 10. Only at early lead times of 12 hr does it seem that the calibration is actually slightly worse. This story is similar for wind speed (not shown).

In terms of computation time (not parallelized), the lead-time-continuous models represent a substantial improvement. For temperature, the mean time for training and prediction per location and running window was 0.152 s (0.012 s) for the lead-time-continuous models (C-RWIN) and 2.463 s (0.152 s) for the lead-time-separated ones (S-RWIN). This corresponds to an over 90% decrease in computation time, which might be helpful for operational use. Times and improvement are similar for wind speed.

The role of the drift terms $\phi_{\mu,\sigma}t$ in the C-RWIN models (see Equations 18,19) is interesting. For temperature, removing the drift term for the location parameter leads to worse performance at earlier lead times but improved performance for later lead times, which is especially pronounced for some stations with considerable diurnal cycle. This is difficult to account for using splines, the use of which did not lead to improved performance in terms of CRPS. It was possible to introduce a model including a factor accounting for early lead times, which did show improved performance at later lead times, whilst achieving the same at earlier ones, however, the gains were not considerable enough to investigate this model further. Removing the drift for the scale parameter did lead to

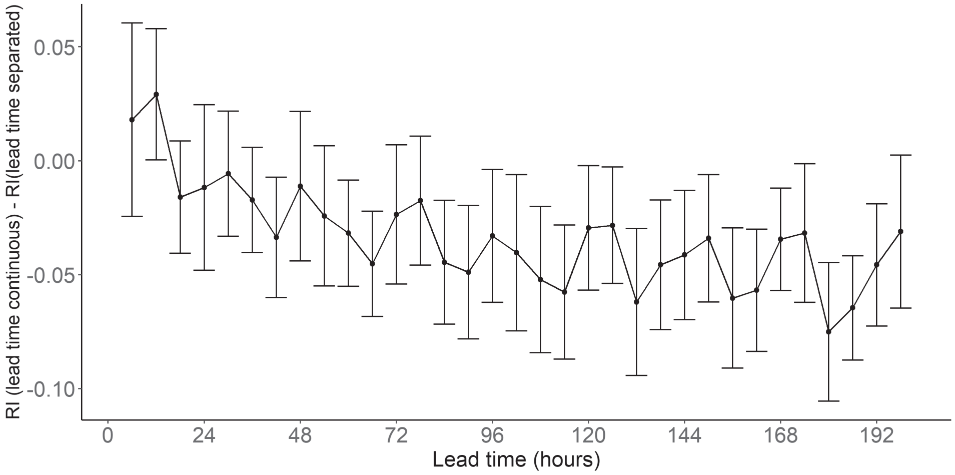


FIGURE 12 Distribution across locations of the difference in reliability index (RI) between lead-time-continuous (C-RWIN) and lead-time-separated (S-RWIN) temperature models trained in a running window, shown as a function of lead time. The solid line represents the mean whilst the bars indicate one standard deviation.

worse performance at all lead times. This is in contrast to the models including seasonality within the model (C- and S-SWM). One explanation for this phenomenon might be that drift is mostly relevant for small data situations and might be weather-pattern dependent. Furthermore, in small data situations the term might help to reduce the impact of later lead times—where errors are especially big — on the estimated EMOS parameters, whereas on larger data sets the effect of these errors is less strong. For wind speed, the main lead-time effect did not seem to add substantially to the model; however, removing it also did not lead to performance increases. Interestingly, base EMOS, without any adjustments for lead time nor diurnal cycle, trained simply on the data set of merged lead times in one running window, also performed fairly well for wind speed, with some deviations for earlier lead times during the day (1200 and 1800 UTC). There the tod adjustment is able to remove diurnal-cycle-dependent bias and miscalibration. For both temperature and wind-speed models, using different interactions between the time of day or drift and the ensemble mean and or standard deviation was tested, but this did not lead to improved performance in terms of CRPS.

The superior performance of C-RWIN over S-RWIN models (in terms of CRPS) is most likely due to better/stabler estimation of postprocessing parameters due to the availability of more data, without the need to estimate substantially more parameters. Different authors discuss the bias-variance trade-off that happens when choosing an optimal training window length (Gneiting et al., 2005; Scheuerer and Möller, 2015): when the training window becomes longer, the bias increases, as the model cannot adapt to seasonal/large-scale changes in external conditions; however, when the window becomes shorter the variance of parameter estimates increases, both leading to decreased performance. Merging data across lead times and reducing the variance in the parameter estimates might therefore allow users to reduce the size of a running-window/training data set. This can be important for operational use, as discussed by Hamill (2018) and others, because data sets available in practice can be of "less-than-ideal" quality and so developing techniques that utilize data more efficiently and work using limited training sample sizes can be important.

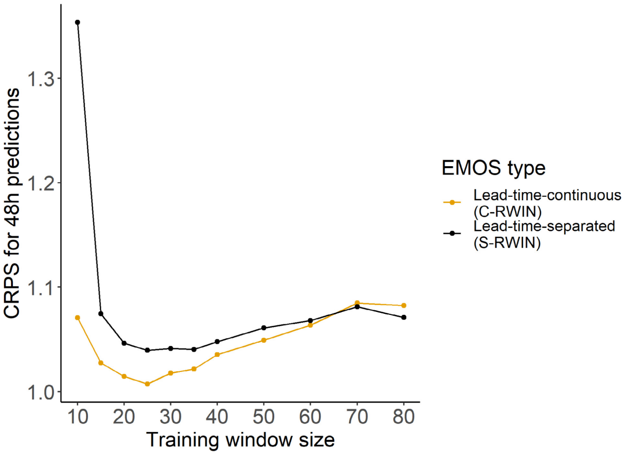


FIGURE 13 CRPS for 48-hr temperature predictions at station 20 as a function of the training-window size for lead-time-continuous (C-RWIN, orange) and lead-time-separated EMOS (S-RWIN, black) trained in a running window.

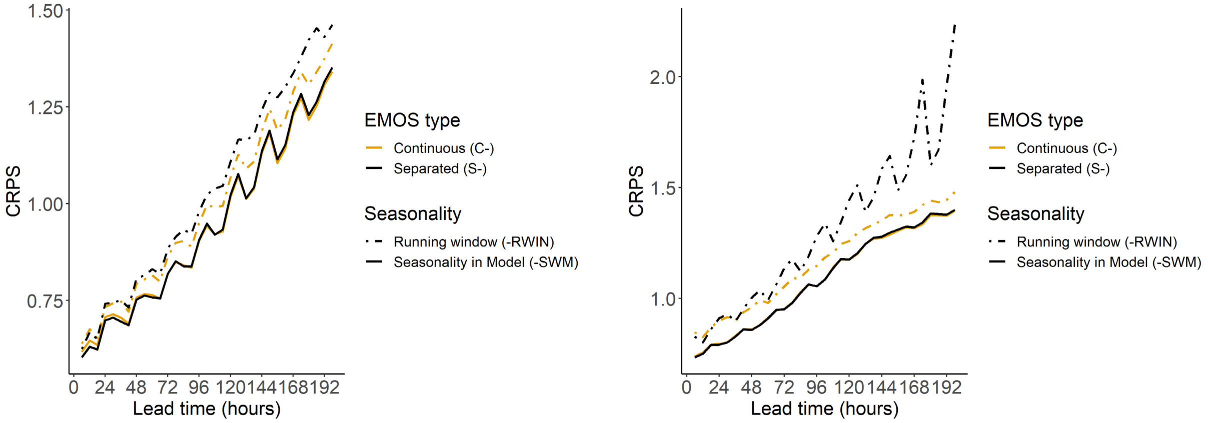
Figure 13 shows the performance for 48-hr temperature predictions at an example station (station 20) as a function of the training-window size for both lead-time-continuous (C-RWIN) and separated (S-RWIN) models. For the lead-time-separated model, there seems to be a relatively flat minimum around 35 days, with the typical behaviour of both shorter and longer windows leading to reduced performance. The lead-time-continuous model, however, profits from a shorter window of around

25 days, where a performance increase of around 3% can be seen compared with the optimal performance of the lead-time-separated model. This is most likely due to smaller bias in the estimated parameters because the data are more current, whilst the reduction in sample size from using more current data (which would usually lead to increased variance in the parameter estimates) is overcome by pooling over lead times. The optimal training window period of 25 days is also fairly stable across lead times (not shown), with some earlier lead times having even better performance at 20 days and some later ones at 30 days; however, these differences are not substantial. Training on a running window size of 25 days then leads to constant improvements across all lead times compared with the lead-time-separated model trained with a running window of 35 days. This improvement is even stronger than shown in Figure 10a, where both models were trained on 40 days running window. This performance increase is not only the effect of a longer training data set, as Figure 13 shows that increasing the training data set over 25 days leads to decreases in performance for the lead-time-continuous models. Rather it is the effect of a larger homogeneous training data set, where the shorter window means that the training data exhibit less variation in seasonal cycle and large-scale atmospheric conditions.

5.2.3 | Comparison

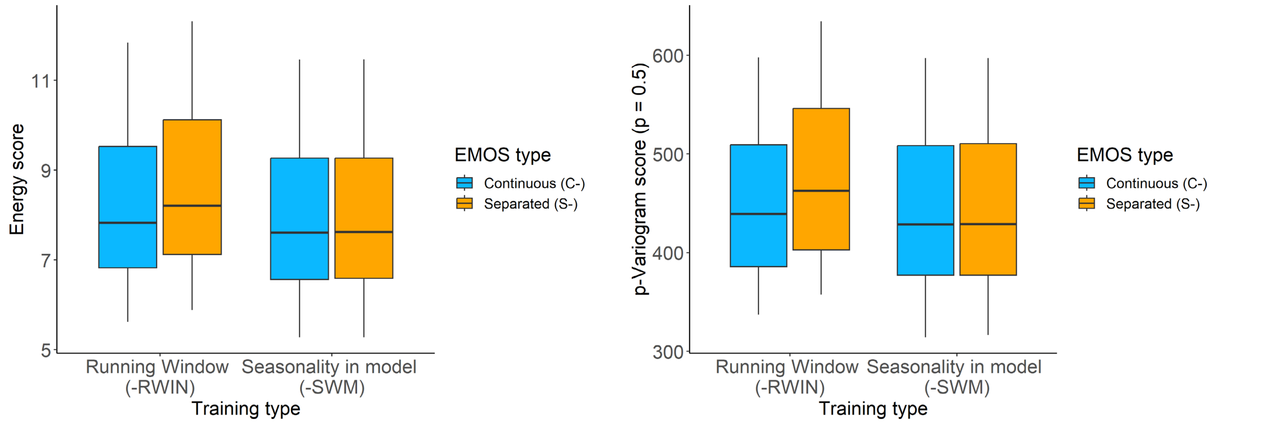
After looking at lead-time-continuous and separated models trained using both a running window and seasonality as terms in the model, we can compare both training routines overall. Figure 14 shows the CRPS as a function of lead time for lead-time-separated (S-models) and continuous models (C-models) trained in both a running window (-RWIN) and including seasonality within the model (-SWM) on the testing data set (December 1, 2021–April 1, 2022) for both temperature and wind speed. One sees that including seasonality within the model generally leads to improved performance compared with training in a running window. In these cases, the merging across lead times does not improve performance substantially in terms of CRPS, but it does cut down on computation time. This is consistent with Lang et al. (2020), who argue that as soon as EMOS training data from multiple years are available, approaches using the full data, rather than training in a running window, are superior. Unfortunately, however, due to NWP model updates and the computation cost associated with generating reforecasts, as well as possible issues attached to observational data, such data sets are often not available in practice (Hamill, 2018). Interestingly, whilst both lead-time-continuous models (C-models) and the S-SWM model for wind speed allow to alleviate time-of-day-dependent performance differences, these remain for all models for temperature. These diurnal-cycle-dependent differences in performance are also found in other studies (see e.g. Dabernig et al., 2017a) but are removed by the comparatively complex Atmosphere NETwork (ANET) model in Demaeyer et al. (2023); Mlakar et al. (2023). The reason for these variations in performance and which models are suitable to remove them is an interesting question for future research.

It is also possible to look at the relationship between forecast distributions at different lead times within one issued forecast, as this has been of interest in multiple studies (Pinson and Girard, 2012; Hemri et al., 2013,



(a) (b)

FIGURE 14 CRPS as a function of lead time averaged over all stations for lead-time-continuous (C-models, orange) and lead-time-separated EMOS (S-models, black) for both models trained in a running window (RWIN-models, dotted) and ones including seasonality adjustments within the model (SWM-models, solid). (a) Temperature. (b) Wind speed.



(a) (b)

FIGURE 15 Box plot of (a) energy and (b) p-Variogram score for multivariate wind-speed predictions (at all lead times) for lead-time-continuous (C-models, blue) and lead-time-separated EMOS (S-models, orange) for both models trained in a running window (RWIN-models, left) and ones including seasonality adjustments within the model (SWM-models, right).

2015; Engeland and Steinsland, 2014). Figure 15 shows the distribution across locations of the energy and p-Variogram scores for the multivariate vector of postprocessed wind-speed predictions issued at a common date. The scores were calculated from ensembles with L = 500 members drawn from the predictive distribution issued. Due to space issues, we omit the temperature results, which however are comparable. Within a running window, lead-time-continuous forecasts (C-RWIN compared with S-RWIN) help with the multivariate performance combined across lead times as measured by the energy and p-Variogram scores, most likely due to the reduced variance in the estimated parameters. However, for models not trained in a running window (-SWM models) no improvements can be seen.

We also evaluate the performance of different models in forecasting extremes. For wind-speed forecasts, Figure 16 shows the threshold-weighted CRPS skill score averaged over all stations and lead times, relative to the lead-time-separated model in a running window (S-RWIN), which we consider a baseline here. One can see that both models with seasonality within the model (-SWM) have strongly improved performance over the baseline model (S-RWIN), with even stronger relative improvement at extremes. The continuous model trained in a running window (C-RWIN) also improves upon the baseline S-RWIN model (albeit less), with a trend indicating that the relative improvement is again stronger at extremes than overall. No differences can be seen between the continuous and separated (C- and S-) models with seasonality within the model (-SWM). This behaviour, albeit aggregated here over all stations and lead times, can also be seen at each individual lead time and most stations (not shown). Temperature extremes were also evaluated similarly (not shown), with the SWM models again performing best, followed by the C-RWIN one, and with some indication of an increasing (with threshold) relative improvement of the C-RWIN model over the S-RWIN model, but not so for the others. These results should, however, be taken with care, as no large temperature extremes are present in the data set. These results provide another argument for increasing data size by merging across lead times and seem to indicate that larger data sizes are necessary for good forecasting of extremes.

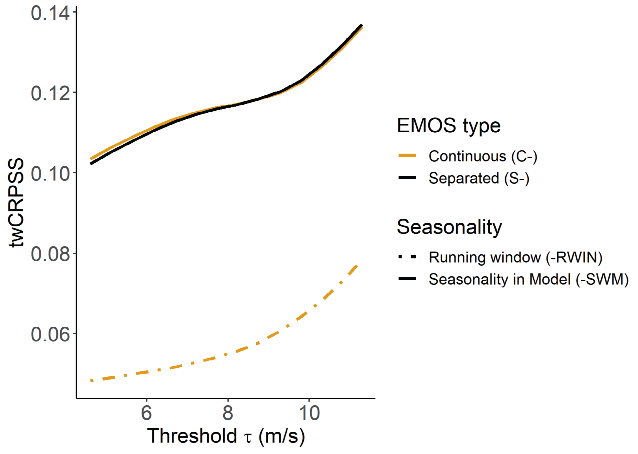


FIGURE 16 Threshold-weighted CRPS skill score for wind-speed forecasts, averaged over all stations and lead times, relative to the lead-time-separated model trained in a running window (S-RWIN). Thresholds here correspond to the 50th–95th percentile of wind-speed observations across all stations.

In terms of computation time, for both the running-window models (-RWIN) and the ones including seasonality within the model (-SWM), the lead-time-continuous (C-) models correspond to a substantial saving. This is more relevant for the models in a running window, as they require constant retraining. Especially in situations where more complex postprocessing models are used, the computational savings by merging across lead times can become relevant.

6 | CONCLUSIONS AND OUTLOOK

A lot of research on postprocessing of ensemble weather forecasts has focused on building models for separate lead times. However, as argued above, this can prove expensive and prohibitive for applications. In this work we have shown that for postprocessing there is a substantial degree of regularity between lead times with two combining effects: the diurnal cycle in interaction with seasonality and a possible lead-time-dependent drift. It proves beneficial to account for these effects separately, and by doing so it is possible to build models that work continuously over lead time. These models save substantially on computation time. For models trained on a fixed training data set and accounting for seasonality within the model, they have on-par performance and calibration compared with standard lead-time-separated models, whilst for models trained using a running window they improve (multivariate) performance across lead times and (temporal) calibration. This allows us to decrease training-window size with an increase in performance, due to faster adaptation to seasonality or changing large-scale conditions.

As Baran and Lerch (2015) note, the computational cost for postprocessing is generally minor compared with the one for NWP model generation and model selection should therefore be based purely on performance. However, with the increasingly complex postprocessing methods and chains that are being built, this factor can become considerable and prove prohibitive. Taillardat and Mestre (2020) describe operational postprocessing at Météo France using quantile regression forests (Taillardat et al., 2016, 2019) and ensemble copula coupling (Schefzik et al., 2013). They show the benefits of switching to a high-performance computing (HPC) environment for postprocessing and note the extensive amounts of data being generated for model development and application. Furthermore, even though postprocessing runs on the Météo France supercomputer to generate results in time, operational demands come into place and can restrict methods chosen when they are too complex. Thus there is demand for methods that improve on computation time, whilst having equal performance. As shown in this study, merging across lead times can present a remedy.

Furthermore, with upcoming machine learning and artificial intelligence postprocessing methods (Rasp and Lerch, 2018; Kirkwood et al., 2021; Haupt et al., 2021; Chen et al., 2022) there is a need for larger data sets to train and evaluate models. As Hamill (2018) notes, due to the cost of generating reforecasts and practical constraints such as frequent model updates, data sets available in practice can often be of nonideal quality and may not have considerable size. The work presented here presents a remedy for small data sets when merging across lead times can improve performance. Furthermore, it gives a strategy to increase training data sizes for deep learning/artificial-intelligence-based methods, which are known often to be inefficient in their data usage (Marcus, 2018). Here, for the methods trained on a fixed training data set no increases in performance could be seen by merging across lead times; however, that might be due to the limits of what EMOS is able to do and the fact that variability in the parameters is already low for these models due to the larger data size. This might be different for deep-learning-based methods and future research into their data efficiency and the training data necessary is required. However, given the ease with which it is possible to incorporate lead-time information, future studies should aim to utilise this fully. Lead-time-continuous models could also easily be used in conjunction with other methods to enrich data size such as semilocal EMOS (Lerch and Baran, 2016), by using local time to account for the diurnal cycle as above.

Finally, this article has focused on developing methods for *lead-time-continuous postprocessing* at a weather timescale. However, similar methods might also be beneficial at longer timescales (intraseasonal/seasonal/decadal) and future work could aim to investigate lead-time dependence for long-range forecasts and try to extend the methods presented here, possibly addressing questions of *seamless prediction* across timescales.

AUTHOR CONTRIBUTIONS

Jakob Benjamin Wessel: conceptualization; formal analysis; investigation; methodology; software; visualization; writing – original draft. **Christopher A. T. Ferro:** conceptualization; methodology; supervision; writing – review and editing. **Frank Kwasniok:** conceptualization; funding acquisition; methodology; project administration; supervision; writing – review and editing.

ACKNOWLEDGEMENTS

Jakob Wessel was supported during this work by the Engineering and Physical Sciences Research Council (EPSRC) under grant award number 2696930. The authors thank

Gavin Evans and Jamie Kettleborough for discussions and comments, which have significantly improved the quality of this work, and Gavin Evans for providing the Met Office's MOGREPS-G forecasting and corresponding observational data. Jakob Wessel acknowledges the discussions with Fiona Spuler, especially regarding the models in Section 3.3.

DATA AVAILABILITY STATEMENT

The code for this study is available on GitHub at https://github.com/jakobwes/QJ-Lead-time-continuous-postprocessing. Unfortunately, the authors are unable to share the ensemble prediction and observational data; however, this can be requested from the UK Met Office.

ORCID

Jakob Benjamin Wessel https://orcid.org/0000-0003-2621-2477
Christopher A. T. Ferro https://orcid.org/0000-0002-9830-9270
Frank Kwasniok https://orcid.org/0000-0003-1421-4010

REFERENCES

Allen, S., Evans, G.R., Buchanan, P. & Kwasniok, F. (2021) Accounting for skew when postprocessing MOGREPS-UK temperature forecast fields. *Monthly Weather Review*, 149, 2835–2852 https://journals.ametsoc.org/view/journals/mwre/149/8/MWR-D-20-0422.1.xml

Allen, S., Ferro, C.A. & Kwasniok, F. (2019) Regime-dependent statistical post-processing of ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 145, 3535-3552.

Allen, S., Ferro, C.A. & Kwasniok, F. (2020) Recalibrating wind-speed forecasts using regime-dependent ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society*, 146, 2576-2596.

Allen, S., Ginsbourger, D. & Ziegel, J. (2022) Evaluating forecasts for high-impact events using transformed kernel scores. http://arxiv.org/abs/2202.12732. ArXiv:2202.12732 [stat]

Anderson, J.L. (1996) A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *Journal of Climate*, 9, 1518–1530 https://www.jstor.org/stable/26201352

Baran, S. & 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 https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.2521

Baran, S. & Lerch, S. (2018) Combining predictive distributions for the statistical post-processing of ensemble forecasts. *International Journal of Forecasting*, 34, 477–496 https://www.sciencedirect.com/science/article/pii/S016920701830030X

Chen, J., Janke, T., Steinke, F. & Lerch, S. (2022) Generative machine learning methods for multivariate ensemble post-processing. https://publikationen.bibliothek.kit.edu/1000151932

Christensen, H.M., Moroz, I.M. & Palmer, T.N. (2015) Simulating weather regimes: impact of stochastic and perturbed parameter schemes in a simple atmospheric model. *Climate Dynamics*, 44, 2195–2214. Available from: https://doi.org/10.1007/s00382-014-2239-9

Dabernig, M., Mayr, G.J., Messner, J.W. & Zeileis, A. (2017a) Simultaneous ensemble postprocessing for multiple lead times with standardized anomalies. *Monthly Weather Review*, 145, 2523–2531 https://journals.ametsoc.org/view/journals/mwre/145/7/mwr-d-16-0413.1.xml

Dabernig, M., Mayr, G.J., Messner, J.W. & Zeileis, A. (2017b) Spatial ensemble post-processing with standardized anomalies. *Quarterly Journal of the Royal Meteorological Society*, 143, 909–916 https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.2975

Delle Monache, L., Hacker, J.P., Zhou, Y., Deng, X. & Stull, R.B. (2006) Probabilistic aspects of meteorological and ozone regional ensemble forecasts. *Journal of Geophysical Research: Atmospheres*, 111 https://onlinelibrary.wiley.com/doi/abs/10.1029/2005JD006917

Demaeyer, J., Bhend, J., Lerch, S., Primo, C., Van Schaeybroeck, B., Atencia, A. et al. (2023) The EUPPBench postprocessing benchmark dataset v1.0. *Earth System Science Data*, 15, 2635–2635 https://doi.org/10.5194/essd-15-2635-2023

Engeland, K. & Steinsland, I. (2014) Probabilistic postprocessing models for flow forecasts for a system of catchments and several lead times. *Water Resources Research*, 50, 182–197 https://onlinelibrary.wiley.com/doi/abs/10.1002/2012WR012757

Gebetsberger, M., Messner, J.W., Mayr, G.J. & Zeileis, A. (2018) Estimation methods for nonhomogeneous regression models: minimum continuous ranked probability score versus maximum likelihood. *Monthly Weather Review*, 146, 4323–4338 https://journals.ametsoc.org/view/journals/mwre/146/12/mwr-d-17-0364.1.xml

Gneiting, T., Balabdaoui, F. & Raftery, A.E. (2007) Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 69, 243–268 https://www.jstor.org/stable/4623266

Gneiting, T., Raftery, A.E., Westveld, A.H. & Goldman, T. (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133, 1098–1118 https://journals.ametsoc.org/view/journals/mwre/133/5/mwr2904.1.xml

Gneiting, T. & Ranjan, R. (2011) Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, 29, 411–422. Available from: https://doi.org/10.1198/jbes.2010.08110

Gneiting, T., Stanberry, L.I., Grimit, E.P., Held, L. & Johnson, N.A. (2008) Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test*, 17, 211–235. Available from: https://doi.org/10.1007/s11749-008-0114-x

Hamill, T.M. (2018) Chapter 7 - Practical aspects of statistical postprocessing. In: Vannitsem, S., Wilks, D.S. & Messner, J.W. (Eds.) *Statistical Postprocessing of Ensemble Forecasts*. Amsterdam: Elsevier, pp. 187–217 https://www.sciencedirect.com/science/article/pii/B9780128123720000078

Hamill, T.M. & Colucci, S.J. (1997) Verification of eta–RSM short-range ensemble forecasts. *Monthly Weather Review*, 125, 1312–1327 https://journals.ametsoc.org/view/journals/mwre/125/6/1520-0493.1997.125.1312.2.0.co.2.xml

Haupt, S.E., Chapman, W., Adams, S.V., Kirkwood, C., Hosking, J.S., Robinson, N.H. et al. (2021) Towards implementing artificial intelligence post-processing in weather and climate: proposed actions from the Oxford 2019 workshop. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379, 20200091 https://royalsocietypublishing.org/doi/full/10.1098/rsta.2020.0091

Hemri, S., Fundel, F. & Zappa, M. (2013) Simultaneous calibration of ensemble river flow predictions over an entire range of lead times. *Water Resources Research*, 49, 6744–6755 https://onlinelibrary.wiley.com/doi/abs/10.1002/wrcr.20542

Hemri, S., Lisniak, D. & Klein, B. (2015) Multivariate postprocessing techniques for probabilistic hydrological forecasting. *Water Resources Research*, 51, 7436–7451 https://onlinelibrary.wiley.com/doi/abs/10.1002/2014WR016473

Hersbach, H. (2000) Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15, 559–570 https://journals.ametsoc.org/view/journals/wefo/15/5/1520-0434.2000.015.0559.dotcrp.2.0.co.2.xml

Hess, R. (2020) Statistical postprocessing of ensemble forecasts for severe weather at Deutscher Wetterdienst. *Nonlinear Processes in Geophysics*, 27, 473–487 https://npg.copernicus.org/articles/27/473/2020/

Jewson, S., Brix, A. & Ziehmann, C. (2004) A new parametric model for the assessment and calibration of medium-range ensemble temperature forecasts. *Atmospheric Science Letters*, 5, 96–102 https://onlinelibrary.wiley.com/doi/abs/10.1002/asl.69

Jordan, A., Krüger, F. & Lerch, S. (2019) Evaluating probabilistic forecasts with scoringRules. *Journal of Statistical Software*, 90, 1–37. Available from: https://doi.org/10.18637/jss.v090.i12

Kirkwood, C., Economou, T., Odbert, H. & Pugeault, N. (2021) A framework for probabilistic weather forecast post-processing across models and lead times using machine learning. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379, 20200099 https://royalsocietypublishing.org/doi/10.1098/rsta.2020.0099

Kwasniok, F. (2012) Data-based stochastic subgrid-scale parametrization: an approach using cluster-weighted modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 370, 1061–1086 https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2011.0384

Lang, M.N., Lerch, S., Mayr, G.J., Simon, T., Stauffer, R. & Zeileis, A. (2020) Remember the past: a comparison of time-adaptive training schemes for non-homogeneous regression. *Nonlinear Processes in Geophysics*, 27, 23–34 https://npg.copernicus.org/articles/27/23/2020/

Lerch, S. & Baran, S. (2016) Similarity-based semilocal estimation of post-processing models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66, 29–51 https://onlinelibrary.wiley.com/doi/abs/10.1111/rssc.12153

Lerch, S. & Thorarinsdottir, T.L. (2013) Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. *Tellus A: Dynamic Meteorology and Oceanography*, 65, 21206 http://a.tellusjournals.se/article/10.3402/tellusa.v65i0.21206/

Lorenz, E. (1996) Predictability: a problem partly solved. ECMWF. https://www.ecmwf.int/node/10829

Marcus, G. (2018) Deep learning: a critical appraisal. http://arxiv.org/abs/1801.00631. ArXiv:1801.00631 [cs, stat]

Messner, J., Zeileis, A. & Stauffer, R. (2022) Crch: censored regression with conditional heteroscedasticity. https://CRAN.R-project.org/package=crch

Messner, J.W., Mayr, G.J. & Zeileis, A. (2017) Nonhomogeneous Boosting for Predictor Selection in Ensemble Postprocessing. *Monthly Weather Review*, 145, 137–147 https://journals.ametsoc.org/view/journals/mwre/145/1/mwr-d-16-0088.1.xml

Mlakar, P., Merše, J. & Pucer, J.F. (2023) Ensemble weather forecast post-processing with a flexible probabilistic neural network approach. http://arxiv.org/abs/2303.17610. ArXiv:2303.17610 [physics]

Pinson, P. & Girard, R. (2012) Evaluating the quality of scenarios of short-term wind power generation. *Applied Energy*, 96, 12–20 https://www.sciencedirect.com/science/article/pii/S0306261911006994

Porson, A.N., Carr, J.M., Hagelin, S., Darvell, R., North, R., Walters, D. et al. (2020) Recent upgrades to the Met Office convective-scale ensemble: an hourly time-lagged 5-day ensemble. *Quarterly Journal of the Royal Meteorological Society*, 146, 3245–3265 https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.3844

Raftery, A.E., Gneiting, T., Balabdaoui, F. & Polakowski, M. (2005) Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133, 1155–1174 https://journals.ametsoc.org/view/journals/mwre/133/5/mwr2906.1.xml

Rasp, S. & Lerch, S. (2018) Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, 146, 3885–3900 https://journals.ametsoc.org/view/journals/mwre/146/11/mwr-d-18-0187.1.xml

Rigby, R.A. & Stasinopoulos, D.M. (2005) Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54, 507–554 https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9876.2005.00510.x

Roberts, N., Ayliffe, B., Evans, G., Moseley, S., Rust, F., Sandford, C. et al. (2023) IMPROVER: the new probabilistic postprocessing system at the Met Office. *Bulletin of the American Meteorological Society*, 104, E680–E697 https://journals.ametsoc.org/view/journals/bams/104/3/BAMS-D-21-0273.1.xml

Roulston, M. & Smith, L. (2003) Combining dynamical and statistical ensembles. *Tellus A: Dynamic Meteorology and Oceanography*, 55, 16–30. Available from: https://doi.org/10.3402/tellusa.v55i1.12082

Schaeybroeck, B.V. & Vannitsem, S. (2018) Chapter 10 - Post-processing of long-range forecasts. In: Vannitsem, S., Wilks, D.S. & Messner, J.W. (Eds.) *Statistical Postprocessing of Ensemble Forecasts*. Amsterdam: Elsevier, pp. 267–290 https://www.sciencedirect.com/science/article/pii/B9780128123720000108

Schefzik, R., Thorarinsdottir, T.L. & Gneiting, T. (2013) Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, 28, 616–640 https://projecteuclid.org/journals/statistical-science/volume-28/issue-4/Uncertainty-Quantification-in-Complex-Simulation-Models-Using-Ensemble-Copula-Coupling/10.1214/13-STS443.full

Scheuerer, M. & Hamill, T.M. (2015) Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143, 1321–1334 https://journals.ametsoc.org/view/journals/mwre/143/4/mwr-d-14-00269.1.xml

Scheuerer, M. & Möller, D. (2015) Probabilistic wind speed forecasting on a grid based on ensemble model output statistics. *The Annals of Applied Statistics*, 9, 1328–1349 https://projecteuclid.org/journals/annals-of-applied-statistics/volume-9/issue-3/Probabilistic-wind-speed-forecasting-on-a-grid-based-on-ensemble/10.1214/15-AOAS843.full

Taillardat, M., Fougères, A.-L., Naveau, P. & Mestre, O. (2019) Forest-based and semiparametric methods for the postprocessing of rainfall ensemble forecasting. *Weather and Forecasting*, 34, 617–634 https://journals.ametsoc.org/view/journals/wefo/34/3/waf-d-18-0149.1.xml

Taillardat, M. & Mestre, O. (2020) From research to applications – examples of operational ensemble post-processing in France using machine learning. *Nonlinear Processes in Geophysics*, 27, 329–347 https://npg.copernicus.org/articles/27/329/2020/

Taillardat, M., Mestre, O., Zamo, M. & Naveau, P. (2016) Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Monthly Weather Review*, 144, 2375–2393 https://journals.ametsoc.org/view/journals/mwre/144/6/mwr-d-15-0260.1.xml

Thorarinsdottir, T.L. & Gneiting, T. (2010) Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173, 371–388 https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-985X.2009.00616.x

Thorarinsdottir, T.L. & Schuhen, N. (2018) Chapter 6 - Verification: assessment of calibration and accuracy. In: Vannitsem, S., Wilks, D.S. & Messner, J.W. (Eds.) *In Statistical Postprocessing of Ensemble Forecasts*. Amsterdam: Elsevier, pp. 155–186 https://www.sciencedirect.com/science/article/pii/B9780128123720000066

Umlauf, N., Klein, N. & Zeileis, A. (2018) BAMLSS: Bayesian additive models for location, scale, and shape (and beyond). *Journal of Computational and Graphical Statistics*, 27, 612–627. Available from: https://doi.org/10.1080/10618600.2017.1407325

Vannitsem, S., Bremnes, J.B., Demaeyer, J., Evans, G.R., Flowerdew, J., Hemri, S. et al. (2021) Statistical postprocessing for weather forecasts: review, challenges, and avenues in a big data world. *Bulletin of the American Meteorological Society*, 102, E681–E699 https://journals.ametsoc.org/view/journals/bams/102/3/BAMS-D-19-0308.1.xml

Walters, D., Boutle, I., Brooks, M., Melvin, T., Stratton, R., Vosper, S. et al. (2017) The Met Office unified model global atmosphere 6.0/6.1 and JULES global land 6.0/6.1 configurations. *Geoscientific Model Development*, 10, 1487–1520 https://gmd.copernicus.org/articles/10/1487/2017/

Wilks, D.S. (2005) Effects of stochastic parametrizations in the Lorenz'96 system. *Quarterly Journal of the Royal Meteorological Society*, 131, 389–407 https://onlinelibrary.wiley.com/doi/abs/10.1256/qj.04.03

Wilks, D.S. (2006) Comparison of ensemble-MOS methods in the Lorenz'96 setting. *Meteorological Applications*, 13, 243–256 https://onlinelibrary.wiley.com/doi/abs/10.1017/S1350482706002192

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

Wood, S.N. (2003) Thin plate regression splines. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 65, 95–114 https://www.jstor.org/stable/3088828

How to cite this article: Wessel, J.B., Ferro, C.A.T. & Kwasniok, F. (2024) Lead-time-continuous statistical postprocessing of ensemble weather forecasts. *Quarterly Journal of the Royal Meteorological Society*, 150(761), 2147–2167. Available from: https://doi.org/10.1002/qj.4701

APPENDIX. CLOSED-FORM EXPRESSIONS FOR THE THRESHOLD-WEIGHTED CRPS

The threshold-weighted CRPS of a forecast given as a normal distribution, with threshold τ and observation y, can be derived as

$$\mathrm{twCRPS}[N(\mu,\sigma^2),y] = \sigma\Bigg\{ -s[1-\Phi(s)]^2 + 2\varphi(s)[1-\Phi(s)] - \frac{1}{\sqrt{\pi}}[1-\Phi(\sqrt{2}s)]\Bigg\} \tag{A1}$$

if $y \le \tau$ and

$$\mathrm{twCRPS}[N(\mu,\sigma^2),y] = \sigma\Bigg\{ -s\Phi^2(s) + z[2\Phi(z)-1] + 2[\varphi(z)-\varphi(s)\Phi(s)] - \frac{1}{\sqrt{\pi}}[1-\Phi(\sqrt{2}s)]\Bigg\} \tag{A2}$$

if $y > \tau$, where φ and Φ are the standard normal density and cumulative distribution function, respectively, and

$$s = \frac{\tau-\mu}{\sigma}, \tag{A3}$$

$$z = \frac{y-\mu}{\sigma}. \tag{A4}$$

In the limit $\tau \to -\infty$, the usual CRPS for the normal distribution (Gneiting et al., 2005) is recovered:

$$\mathrm{CRPS}[N(\mu,\sigma^2),y] = \sigma\Bigg\{ z[2\Phi(z)-1] + 2\varphi(z) - \frac{1}{\sqrt{\pi}}\Bigg\}. \tag{A5}$$

The threshold-weighted CRPS of a forecast given as a (doubly) truncated normal distribution, truncated to the interval $[a,b]$, with threshold $\tau \in [a,b]$ and observation $y \in [a,b]$, can be derived as

$$\begin{aligned}\mathrm{twCRPS}[N_a^b(\mu,\sigma^2),y] = \sigma\Bigg\{ & -s\left[\frac{\Phi(\beta)-\Phi(s)}{\Phi(\beta)-\Phi(\alpha)}\right]^2 \\ & +2\varphi(s)\frac{\Phi(\beta)-\Phi(s)}{[\Phi(\beta)-\Phi(\alpha)]^2} \\ & -\frac{1}{\sqrt{\pi}}\frac{\Phi(\sqrt{2}\beta)-\Phi(\sqrt{2}s)}{[\Phi(\beta)-\Phi(\alpha)]^2}\Bigg\} \end{aligned} \tag{A6}$$

if $y \leq \tau$ and

$$\begin{aligned}\mathrm{twCRPS}[N_a^b(\mu,\sigma^2),y] = \sigma\Bigg\{ & -s\left[\frac{\Phi(s)-\Phi(\alpha)}{\Phi(\beta)-\Phi(\alpha)}\right]^2 \\ & +z\left[2\frac{\Phi(z)-\Phi(\alpha)}{\Phi(\beta)-\Phi(\alpha)}-1\right] \\ & +\frac{2}{\Phi(\beta)-\Phi(\alpha)}\left[\varphi(z)-\varphi(s)\frac{\Phi(s)-\Phi(\alpha)}{\Phi(\beta)-\Phi(\alpha)}\right] \\ & -\frac{1}{\sqrt{\pi}}\frac{\Phi(\sqrt{2}\beta)-\Phi(\sqrt{2}s)}{[\Phi(\beta)-\Phi(\alpha)]^2}\Bigg\} \end{aligned} \tag{A7}$$

if $y > \tau$, where φ and Φ are again the standard normal density and cumulative distribution function, respectively, and

$$\alpha = \frac{a-\mu}{\sigma}, \tag{A8}$$

$$\beta = \frac{b-\mu}{\sigma}, \tag{A9}$$

$$s = \frac{\tau-\mu}{\sigma}, \tag{A10}$$

$$z = \frac{y-\mu}{\sigma}. \tag{A11}$$

We remark for consistency that, in the limits $a \to -\infty$ and $b \to \infty$, we have $\Phi(\alpha) = 0$ and $\Phi(\beta) = \Phi(\sqrt{2}\beta) = 1$ and recover $\mathrm{twCRPS}[N(\mu,\sigma^2),y]$ as given above.

In the practically important case of left-truncation at zero, $a = 0$ and $b \to \infty$, we have $\Phi(\beta) = \Phi(\sqrt{2}\beta) = 1$, meaning that Equations A6 and A7 simplify to

$$\begin{aligned}\mathrm{twCRPS}[N_0(\mu,\sigma^2),y] = \sigma\Bigg\{ & -s\left[\frac{1-\Phi(s)}{1-\Phi(\alpha)}\right]^2 \\ & +2\varphi(s)\frac{1-\Phi(s)}{[1-\Phi(\alpha)]^2} \\ & -\frac{1}{\sqrt{\pi}}\frac{1-\Phi(\sqrt{2}s)}{[1-\Phi(\alpha)]^2}\Bigg\} \end{aligned} \tag{A12}$$

if $y \leq \tau$ and

$$\begin{aligned}\mathrm{twCRPS}[N_0(\mu,\sigma^2),y] = \sigma\Bigg\{ & -s\left[\frac{\Phi(s)-\Phi(\alpha)}{1-\Phi(\alpha)}\right]^2 \\ & +z\left[2\frac{\Phi(z)-\Phi(\alpha)}{1-\Phi(\alpha)}-1\right] \\ & +\frac{2}{1-\Phi(\alpha)}\left[\varphi(z)-\varphi(s)\frac{\Phi(s)-\Phi(\alpha)}{1-\Phi(\alpha)}\right] \\ & -\frac{1}{\sqrt{\pi}}\frac{1-\Phi(\sqrt{2}s)}{[1-\Phi(\alpha)]^2}\Bigg\} \end{aligned} \tag{A13}$$

if $y > \tau$. Setting $\tau = 0$, leading to $s = \alpha = -\mu/\sigma$, and using the symmetry $\Phi(x) + \Phi(-x) = 1$ allows us to recover the usual CRPS for a truncated normal distribution (compare Thorarinsdottir & Gneiting, 2010):

$$\begin{aligned}\mathrm{CRPS}[N_0(\mu,\sigma^2),y] = \sigma\Bigg\{ & z\left[2\frac{\Phi(z)-\Phi(\alpha)}{1-\Phi(\alpha)}-1\right] \\ & +\frac{2\varphi(z)}{1-\Phi(\alpha)}-\frac{1}{\sqrt{\pi}}\frac{1-\Phi(\sqrt{2}\alpha)}{[1-\Phi(\alpha)]^2}\Bigg\} \end{aligned} \tag{A14}$$

$$\begin{aligned} = \frac{\sigma}{\Phi\left(\frac{\mu}{\sigma}\right)}\Bigg\{ & \frac{y-\mu}{\sigma}\left[2\Phi\left(\frac{y-\mu}{\sigma}\right)+\Phi\left(\frac{\mu}{\sigma}\right)-2\right] \\ & +2\varphi\left(\frac{y-\mu}{\sigma}\right)-\frac{1}{\sqrt{\pi}}\frac{\Phi\left(\sqrt{2}\frac{\mu}{\sigma}\right)}{\Phi\left(\frac{\mu}{\sigma}\right)}\Bigg\}. \end{aligned} \tag{A15}$$