Received: 19 June 2025 | Revised: 19 January 2026 | Accepted: 1 April 2026

DOI: 10.1002/qj.70201

Quarterly Journal of the Royal Meteorological Society RMetS

RESEARCH ARTICLE

Statistical post-processing of operational dual-resolution wind-speed ensemble forecasts

Sándor Baran | **Mária Lakatos**

Faculty of Informatics, University of Debrecen, Debrecen, Hungary

Correspondence
Sándor Baran, Faculty of Informatics, University of Debrecen, Kassai út 26, H-4028 Debrecen, Hungary.
Email: baran.sandor@inf.unideb.hu

Funding information
National Research, Development and Innovation Office, Grant/Award Number: K142849; National Research, Development and Innovation Fund, Grant/Award Number: EKÖP-24-3-II

Abstract

Weather forecasting presents several challenges, including the chaotic nature of the atmosphere and the high computational demands of numerical weather prediction models. To achieve the most accurate predictions, the ideal scenario involves the lowest possible horizontal resolution and the largest ensemble size. This study provides a detailed comparative analysis of the forecast skill of the raw and post-processed medium- and extended-range wind-speed ensemble forecasts of the European Centre for Medium-Range Weather Forecasts issued at 9 km and 36 km horizontal resolutions respectively, and their various mixtures. We utilize the ensemble model output statistic approach for forecast calibration with three different spatial training data selection techniques. First, we investigate the performance of the 50-member medium-range and 100-member extended-range predictions (referred to as high and low resolution respectively) and their 150-member dual-resolution combination. Further, we examine whether the performance of raw and post-processed low-resolution forecasts can be improved by incorporating high-resolution ensemble members. Specifically, we extend a 50-member low-resolution extended-range forecast with 1, 2, 4, 8, 16, and 32 members from the high-resolution medium-range ensemble and compare the performance of these mixtures both before and after calibration. Our results confirm that, in general, all post-processed forecasts outperform the raw ensemble predictions in terms of probabilistic calibration and point forecast accuracy and that post-processing considerably reduces the differences between the various configurations. We also show that spatial resolution is superior to the ensemble size; augmenting a sufficiently large ensemble of high-resolution forecasts with low-resolution predictions does not necessarily result in a gain in forecast skill. However, our study also highlights the clear benefit of the other direction; namely, incorporating high-resolution members into low-resolution ensemble forecasts, where the most significant gains are observed in configurations with the highest number of high-resolution members.

KEYWORDS

dual-resolution forecasts, ensemble calibration, ensemble model output statistics, truncated normal distribution, wind speed

1 | INTRODUCTION

The advent of ensemble forecasting systems represented a major breakthrough in meteorology, fundamentally changing the way weather predictions are generated and interpreted. Unlike traditional single-run deterministic models, which produced a single outcome based on fixed initial conditions and often overlooked atmospheric uncertainties, ensemble systems introduced a transformative approach. Ensemble forecasts are the outputs of multiple runs of a numerical weather prediction model, each starting from slightly different initial conditions or parametrizations. Their purpose is to account for uncertainties in the atmospheric system and to provide a broader, probabilistic view of expected weather conditions rather than point forecasts. The European Centre for Medium-Range Weather Forecasts (ECMWF) is one of the most advanced weather centers, offering highly accurate models for various time ranges, produced with the Integrated Forecasting System (IFS; ECMWF, 2024). Their medium-range forecasts are generated four times a day, covering forecast horizons of 1–15 days, whereas longer-term forecasts extend up to 46 days. The accuracy of ensemble forecasts can be influenced by various factors, one of the key ones being the resolution of the weather prediction model's grid, particularly its horizontal resolution. The upgrade of ECMWF's IFS to Cycle 48r1 brought substantial improvements to medium-range forecasts (ENS), including an increase in horizontal resolution from 18 km ($T_{CO}639$) to 9 km ($T_{CO}1279$). For long-term forecasts (ENS extended), the resolution improved to 36 km ($T_{CO}319$), and the number of runs was increased from 2 days per week to a daily schedule.[1] Models with finer grids, such as those used for medium-range forecasts, are better equipped to handle smaller-scale weather phenomena, such as localized precipitation or strong wind gusts. In contrast, long-term forecasts focus on larger-scale atmospheric and oceanic processes.

However, ensemble forecasts come with challenges, such as high computational costs and the difficulty of balancing resolution and ensemble size. At a given spatial resolution, the computational cost is proportional to the ensemble size. Nonetheless, as discussed by Leutbecher and Ben Bouallègue (2020), to guarantee numerical stability and accuracy, the larger the resolution, the more integration time steps are required, and "the computational cost for a forecast day is approximately proportional to the number of grid points times the number of integration time steps needed". Thus, since the ratio between the grid resolutions of $T_{CO}1279$ and $T_{CO}319$ forecasts is 4:1, and the integration time steps are 450 s and 1200 s respectively, one medium-range forecast can be roughly traded against 40 extended-range predictions. Moreover, ensemble forecasts can suffer from biases and underestimation of uncertainties (Buizza, 2018). A widely used technique to address the latter issues is ensemble model output statistics (EMOS; Gneiting *et al.*, 2005), a statistical method designed to improve the accuracy of ensemble forecasts by calibrating the raw output. EMOS estimates the relationship between the raw ensemble predictions and the actual outcomes. It involves fitting a parametric model to the forecast ensemble, such as a normal distribution, and using it to adjust the ensemble mean and spread to better match observed reality. By doing so, it generates more reliable probabilistic forecasts that more accurately represent the uncertainty in the atmospheric system.

By integrating forecasts at distinct spatial resolutions (9 km and 36 km in this study), the dual-resolution approach capitalizes on the strengths of both high- and low-resolution ensembles. This method aims to combine the finer, smaller-scale details offered by medium-range forecasts with the broader, large-scale insights provided by extended-range forecasts, thereby improving the overall accuracy and reliability of predictions. Furthermore, owing to the computational constraints of the IFS, achieving an optimal balance between different forecast combinations is crucial. Leutbecher and Ben Bouallègue (2020) combined lower- and higher-resolution ensemble members to improve medium-range weather forecasts while staying within computational constraints. Their findings indicated that dual-resolution ensembles optimize 2-m temperature predictions, whereas single-resolution ensembles are more effective for 850 hPa temperature forecasts. Baran *et al.* (2019) subsequently examined whether the dual-resolution ensembles studied by Leutbecher and Ben Bouallègue (2020) remained superior to single-resolution ensembles after statistical post-processing. The results showed that statistical post-processing significantly reduced performance differences between various single- and dual-resolution configurations. In these studies, the researchers combined 50 forecast members at $T_{CO}639$ (18 km), 200 members at $T_{CO}399$ (29 km), and 254 members (45 km) at $T_{CO}255$ resolution. Later, Gascón *et al.* (2019) evaluated the predictive performance of raw and post-processed dual-resolution precipitation accumulation forecasts, utilizing non-parametric calibration methods, and Szabó *et al.* (2023) assessed the censored shifted gamma EMOS approach for the statistical post-processing of $T_{CO}639$–$T_{CO}399$ dual-resolution ensemble forecasts of the same variable, which were derived from experimental extended-range predictions. The aim of this study is to compare raw and post-processed ECMWF operational dual-resolution 10-m wind-speed forecasts and investigate the impact of incorporating varying numbers of higher-resolution members into the calibration process.

First, we study the skill of raw and post-processed 50-member high-resolution and 100-member low-resolution wind-speed ensemble predictions together with their 150-member combination, where, for forecast calibration, we utilize the EMOS approach. Then, we fix the number of lower-resolution members at 50, augment it with 1, 2, 4, 8, 16, and 32 higher-resolution members, and compare the predictive performance of these combinations before and after post-processing.

The structure of the article is as follows: Section 2 provides a detailed description of the wind-speed dataset under study. In Section 3, we review the calibration approaches used, along with the methods for parameter estimation and model verification. Section 4 presents the results, and Section 5 offers the conclusions.

2 | DATA

As mentioned in the Section 1, our aim is to compare the predictive performance of ECMWF dual-resolution wind-speed forecasts and to examine whether the ranking of competing predictions would differ after calibration.

The dataset at hand comprises 50-member medium-range forecasts of 10-m wind speed at $T_{CO}1279$ resolution and 100-member extended-range forecasts at $T_{CO}319$ resolution, obtained from perturbed initial conditions and/or parametrizations for 8726 synoptic observation (SYNOP) stations (see Figure 1) from July 1, 2023, to May 31, 2024, together with the corresponding validating observations. These station observations are 10-min averages of the observed 10-m wind speed at the reported observation time, whereas forecasts at station locations are ensemble predictions for the corresponding nearest grid points. All forecasts are initialized at 0000 UTC; and since the medium-range forecasts are limited to a 15-day horizon, the lead time of the extended-range predictions is also limited to 15 days with a time step of 24 hr.

3 | STATISTICAL POST-PROCESSING

In the last two decades, a large variety of post-processing approaches have been developed both in parametric- and nonparametric set-ups; for a systematic overview, see Vannitsem *et al.* (2021). Parametric methods provide full predictive distribution of the weather quantity at hand, whereas non-parametric approaches usually provide quantiles of the predictive law. Here, we focus on parametric post-processing, and since the aim is to explore the general tendencies in the calibration of dual-resolution forecasts, we consider the simple but still powerful EMOS approach. As mentioned, the EMOS method addresses the shortcomings of the raw predictions by applying a single parametric distribution to the ensemble outputs, where the parameters are related to the ensemble members or their descriptive statistics via appropriate (usually affine) link functions. Naturally, different meteorological variables require distinct probability distributions to best capture their unique properties. For instance, temperature is often represented by a normal distribution and its generalizations (Gneiting *et al.*, 2005; Taillardat, 2021), as its values generally exhibit a symmetric spread around the mean. In contrast, wind speed is always non-negative and skewed, making a truncated normal (Thorarinsdottir & Gneiting, 2010) or a log-normal (Baran & Lerch, 2015) distribution a more suitable choice. Note that, recently, owing to its flexibility in incorporating additional input covariates, the machine-learning-based counterpart of the EMOS approach, the distributional regression network (DRN; Rasp & Lerch, 2018), has gained more and more popularity. In the DRN approach, the link functions that connect the input features to the predictive distribution parameters are replaced by a neural network so that DRN models can better capture more general relationships between these quantities, which usually results in better forecast skill. However, one should also note

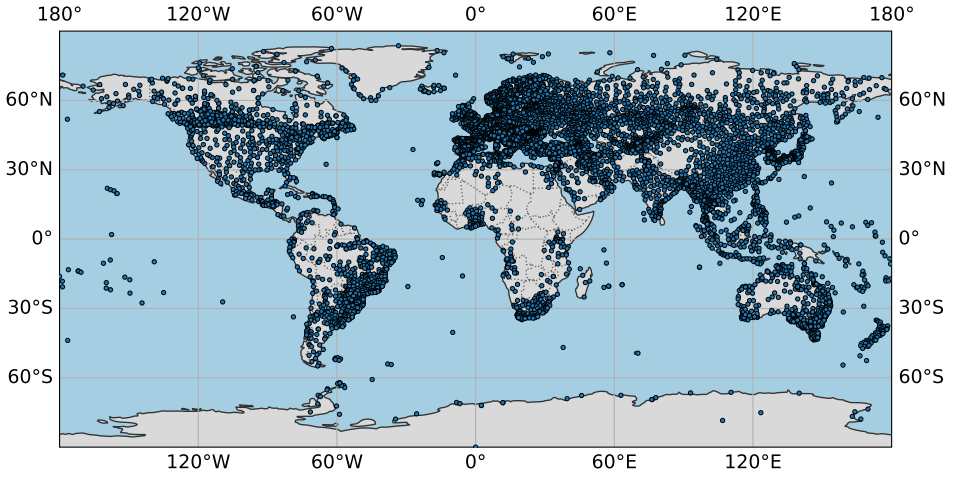

FIGURE 1 Location of SYNOP stations. [Colour figure can be viewed at wileyonlinelibrary.com]

that EMOS provides a well-defined statistical framework, making it more transparent and easier to interpret than machine-learning-based approaches, especially when deep neural networks are involved.

3.1 | Truncated-normal EMOS model

To calibrate the single- and dual-resolution wind-speed forecasts, we apply the EMOS model using a normal distribution left truncated at zero $\mathcal{N}_0^\infty(\mu, \sigma^2)$ with location μ and scale $\sigma > 0$ (Thorarinsdottir & Gneiting, 2010). In the dual-resolution case, according to the horizontal resolution, we form two groups of statistically indistinguishable predictions, with M_L members for the low-resolution "ENS extended" forecasts and M_H for the high-resolution "ENS" predictions. Based on this grouping, the location and scale parameters of the EMOS predictive distribution are

$$\mu = a + b_\mathrm{H}^2 \bar{f}_\mathrm{H} + b_\mathrm{L}^2 \bar{f}_\mathrm{L} \quad \text{and} \quad \sigma^2 = c^2 + d^2 S^2, \qquad (1)$$

where a, b_H, b_L, c, and $d \in \mathbb{R}$ are the parameters to be estimated, $\bar{f}_\mathrm{H}$ and $\bar{f}_\mathrm{L}$ are the means of the high- and low-resolution forecasts respectively, and S^2 denotes the variance of the $(M_\mathrm{H} + M_\mathrm{L})$-member combined ensemble. In the following sections, such a configuration of M_L low- and M_H high-resolution ensemble members will be referred to as a combination $(M_\mathrm{L}, M_\mathrm{H})$. Note that one can incorporate the groups in the scale parameter as well and consider the variances of the high- and low-resolution components separately. However, our previous experience with dual-resolution (Baran *et al.*, 2019) and multimodel (Lerch & Baran, 2017) ensemble forecasts had shown that the use of more complex models for the scale does not result in a significant gain in forecast skill.

Alternatively, when the calibration is based solely on one resolution, the expression for location in Equation (1) has to be modified accordingly by fixing $b_\mathrm{L} = 0$ for the pure high-resolution ($M_\mathrm{L} = 0$) and $b_\mathrm{H} = 0$ for the pure low-resolution ($M_\mathrm{H} = 0$) case.

Drawing on the optimal score estimation principle proposed by Gneiting and Raftery (2007), the estimation of the parameters of EMOS predictive distributions involves minimizing the mean value of a proper scoring rule over a carefully chosen training dataset comprising past forecast–observation pairs. In most cases, the continuous ranked probability score (CRPS) defined by Equation (2) in Section 3.3 is favored, as it simultaneously evaluates the magnitude of forecast errors and the overall distributional performance, yielding a comprehensive assessment of predictive skill.

3.2 | Training data selection

The number of training days is a critical factor affecting the stability and adaptability of the EMOS model. Following Gneiting *et al.* (2005) and Thorarinsdottir and Gneiting (2010), we consider rolling training windows. In this approach, a shorter training period allows a quick adaptation, for instance, to changing seasonal biases but might result in noisy parameter estimates. On the other hand, a longer training period can lead to more stable parameter estimates but might come with challenges such as data availability. For a detailed comparison of time-adaptive training schemes in the context of EMOS modeling, we refer to Lang *et al.* (2020). The choice of a spatial selection strategy also plays a vital role in estimating the parameters of the EMOS model. This estimation can be based on data from a broader region, data from individual stations, or a third approach that incorporates information from similar stations. Each of these methods (regional, local, and semi-local) offers distinct benefits and drawbacks in terms of forecast accuracy and stability. In the regional method (Thorarinsdottir & Gneiting, 2010), the parameters of the EMOS model are estimated using data from all available stations, hence resulting in a single set of parameters for the whole ensemble domain. By using data from all stations within a larger geographic area, it could enable more stable and reliable parameter estimation. This approach is particularly beneficial in homogeneous regions where weather conditions are similar, as patterns learned from a larger dataset can be effectively applied to local forecasts. Additionally, it is computationally more efficient, as a single model is fitted for a larger region, reducing resource demand. However, its drawback is that it is less capable of capturing location-specific characteristics. In heterogeneous regions where weather conditions vary considerably, the regional model may not always provide accurate forecasts, as individual stations may have unique climatic features that differ from the model fitted to the entire area. In contrast, local learning relies solely on the data of individual stations, allowing it to better capture the special weather characteristics of a given location. Since a separate model is fitted for each station, this approach can provide more accurate forecasts in areas where the generalization of a regional model is not suitable. This is particularly useful in mountainous or coastal regions, where topographical conditions cause significant local variations in weather parameters. Although, in general, local modeling overperforms the regional approach, one of its drawbacks is the limited amount of data available. For the optimal training window lengths for various weather quantities, see Hemri *et al.* (2014). If a station has little historical data,

the estimated parameters may become more uncertain, and the model might overfit the training data, reducing its generalizability. Additionally, this approach is more computationally demanding, as a separate EMOS model must be fitted for each station, increasing computational costs. By partitioning the ensemble domain into smaller, similar areas, the clustering-based semi-local approach of Lerch and Baran (2017) capitalizes on the strengths of both regional and local techniques. In each of these homogeneous areas, regional modeling is applied, allowing forecasts to be fine-tuned to local conditions while still reflecting overall regional trends. To dynamically form these areas for each verification date, we employ k-means clustering using feature vectors proposed by Lerch and Baran (2017) that combine climatological data with the forecast errors from the raw ensemble during the training period. Ultimately, this strategy enhances the accuracy of modeling local variations and contributes to improved overall forecast performance.

3.3 | Verification metrics

We evaluate the predictive performance of both probabilistic and deterministic forecasts using a blend of traditional error metrics and proper scoring rules. Since the post-processing methods considered are applied independently to each lead time and location, their evaluation is naturally aligned with univariate scoring rules that assess the quality of one-dimensional predictive distributions.

To assess the forecast skill of deterministic forecasts, such as the ensemble/EMOS means and medians, we utilize the root-mean-squared error (RMSE) and the mean absolute error (MAE), respectively. In scenarios where a single point forecast is employed to represent the future outcome, the MAE is minimized by the median of the corresponding probabilistic prediction, whereas the RMSE is minimized by the mean (Gneiting, 2011).

The CRPS (Wilks, 2019, section 9.5.1) is a proper scoring rule used to evaluate the quality of probabilistic forecasts. It measures the difference between the predicted cumulative distribution function (CDF) and the empirical CDF of the observed value. For a forecast distribution F and an observation y, the CRPS is defined as

$$\mathrm{CRPS}(F,y) := \int_{-\infty}^{\infty} (F(z) - \mathbb{1}\{z \geq y\})^2 \, dz, \qquad (2)$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function. Note that for the truncated normal distribution, the CRPS has a closed form, allowing for an efficient parameter estimation, whereas for the raw ensemble, in Equation (2), one should consider the empirical CDF of the ensemble members (see Jordan *et al.*, 2019).

In addition, we include the Brier score (BS; Wilks, 2019, section 9.4.2) to evaluate forecast skill for binary events derived from continuous variables. Specifically, we assess the case in which the event of interest is whether the observed value y exceeds a given threshold z. The BS compares the forecast probability $1 - F(z)$ of this event with its actual outcome and is defined as

$$\mathrm{BS}(F,y;z) := (F(z) - \mathbb{1}\{z \geq y\})^2.$$

Moreover, the CRPS can be interpreted as the integral of the BS over all possible threshold values. In Section 4, following Veldkamp *et al.* (2021) for example, we consider thresholds $5\ \mathrm{m \cdot s^{-1}}, 10\ \mathrm{m \cdot s^{-1}}$, and $15\ \mathrm{m \cdot s^{-1}}$, corresponding to low, moderate, and high wind speeds respectively.

Furthermore, let $q_\tau(F)$ denote the τ-quantile ($0 \leq \tau \leq 1$) of a CDF $F(y)$; that is,

$$q_\tau(F) := F^{-1}(\tau) := \inf\{y : F(y) \geq \tau\}.$$

Consider the loss function

$$\rho_\tau(x) := \begin{cases} \tau|x|, & \text{if } x \geq 0, \\ (1-\tau)|x|, & \text{if } x < 0. \end{cases}$$

Then, for an observed value y, the quantile score (QS; e.g., Bentzien & Friederichs, 2014) is defined as

$$\mathrm{QS}_\tau(F,y) := \rho_\tau(y - q_\tau(F)).$$

In this study, we evaluate the QS at the 5th, 10th, 20th, 80th, 90th, and 95th percentiles of the predictive distributions.

The improvement of a forecast with respect to a reference predictive distribution F_{ref} can be quantified using corresponding skill scores. For a generic scoring rule S, the skill score comparing forecasts F and F_{ref} is defined as

$$SS(F,F_{\mathrm{ref}}) := 1 - \frac{\overline{S_F}}{\overline{S_{F_{\mathrm{ref}}}}},$$

where $\overline{S_F}$ and $\overline{S}_{F_{\mathrm{ref}}}$ denote the mean score values over the verification data corresponding to F and F_{ref} respectively (Murphy, 1973). Naturally, skill scores differ from the original metrics by being positively oriented, so higher values indicate better forecast quality. In Section 4, summarizing our results, for probabilistic forecasts, we consider the continuous ranked probability skill score (CRPSS), the quantile skill score (QSS), and the Brier skill score (BSS),

whereas for point forecasts we investigate skill scores corresponding to the MAE of the median (MAES) and the RMSE of the mean (RMSES). All skill scores presented are based on the mean score values over all verification dates and stations.

To gain insight into the uncertainty in the score values and the significance of the score differences, in Section 4 the reported skill scores are equipped with 95% confidence intervals. The required confidence bounds are derived from 2000 block bootstrap samples calculated with the help of the stationary bootstrap scheme with random block lengths following a geometric distribution with a mean proportional to the cube root of the length of the time series of skill scores (Politis & Romano, 1994). We first sample the time series of the spatially averaged scores for the actual and reference forecast using identical blocks and then calculate the 2000 skill-score values from means of the matching samples.

4 | RESULTS

As mentioned in Section1, we investigate the forecast skill of various combinations of raw high- and low-resolution wind-speed ensemble forecasts and their post-processed counterparts, where we utilize the truncated normal EMOS model introduced in Section 3.1. For the EMOS calibration, each lead time was modeled independently using the three distinct spatial selection strategies (regional, local, and semi-local) described in Section 3.2. However, to keep the focus on the comparison of the performance of the various combinations of high- and low-resolution predictions, in the following analysis we concentrate on the best-performing local EMOS approach alongside the raw forecasts. Results on regional and semi-local EMOS modeling can be found in Appendix A. To identify the optimal length of the rolling training period, training periods of 30, 60, and 90 days were considered, and the corresponding EMOS models were evaluated over a dedicated verification period from October 13, 2023, to May 31, 2024, spanning 232 calendar days, to ensure comparability between different model set-ups. The finally chosen 60-day training period was selected by minimizing the mean CRPS over this validation period while also considering additional verification metrics and balancing computational efficiency with predictive performance.

Based on this set-up, the full verification period was defined as September 13, 2023, to May 31, 2024, totaling 262 calendar days, and was applied consistently across all model configurations during the final evaluation phase.

4.1 | Performance of operational dual-resolution forecasts

In the following, we present a comparative analysis of the ECMWF operational forecasts at both high- and low-resolution and explore their connection to dual-resolution ensemble predictions. All three configurations, referred to as combinations $(0, 50)$, $(100, 0)$, and $(100, 50)$, are post-processed using the truncated normal EMOS model, and the skill of the resulting forecasts is subsequently assessed.

Figure 2a displays the mean CRPS values for raw pure low-, high-, and (combined) dual-resolution forecasts, alongside their post-processed counterparts. Note that the non-monotonic shape of the mean CRPS curves of the raw configurations is a result of the representativeness error in the verification (see also Baran *et al.*, 2021; Baran & Lakatos, 2024), which can be reduced, for instance, by perturbing the ensemble members (Ben Bouallègue *et al.*, 2020). Among the raw configurations, the $(100, 0)$ pure low-resolution forecast exhibits by far the lowest predictive skill, whereas the $(0, 50)$ pure high- and $(100, 50)$ dual-resolution forecasts perform comparably. Figure 2b, where the corresponding pure high-resolution raw/post-processed forecast is used as a reference (raw high-resolution ensemble for the ensemble forecasts and high-resolution forecast-based local EMOS for the local EMOS models), indicates a modest but significant advantage of this configuration over the dual-resolution raw forecasts between days 2 and 8. Post-processing substantially reduces the mean CRPS for all three combinations and shrinks the differences among them. According to Figure 2b, the mean CRPS values of the $(0, 50)$ and $(100, 50)$ combinations are very close to each other, although the deviation is significant for some lead times, and the EMOS model based on the pure low-resolution ensemble gradually catches up with them as the forecast horizon increases and, moreover, at days 14 and 15 outperforms the calibrated dual-resolution prediction.

Figure 3 shows a similar overall pattern in performance differences, this time based on MAE values. Among the raw forecasts, the pure high- and dual-resolution set-ups again perform best, though the distinction between them is now more noticeable, and according to the skill scores of Figure 3b the advantage of the former is significant for all forecast horizons studied. Similar to the CRPS, local EMOS post-processing results in a solid decrease in the MAE for all three combinations, with the pure low resolution significantly lagging behind up to day 10. Note that the only case where the incorporation of low-resolution terms is significantly beneficial is the local EMOS approach; however, even in this case, only up to day 2.

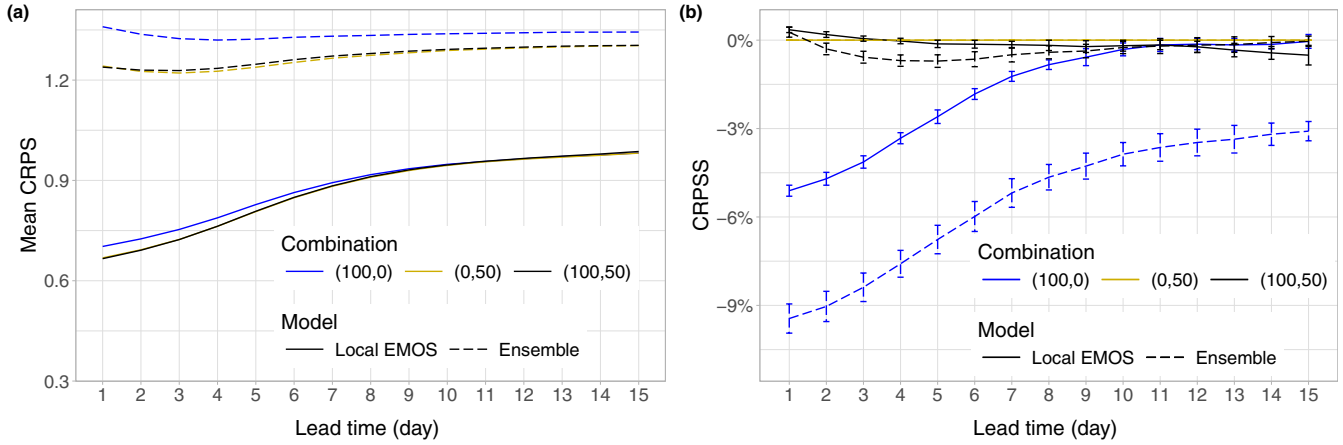

FIGURE 2 (a) Mean continuous ranked probability score (CRPS) of pure low-resolution (100, 0), pure high-resolution (0, 50), and combined dual-resolution (100, 50) raw and post-processed wind-speed forecasts and (b) continuous ranked probability skill score (CRPSS) with respect to the corresponding pure high-resolution prediction with 95% confidence intervals as functions of the lead time. EMOS: ensemble model output statistics. [Colour figure can be viewed at wileyonlinelibrary.com]

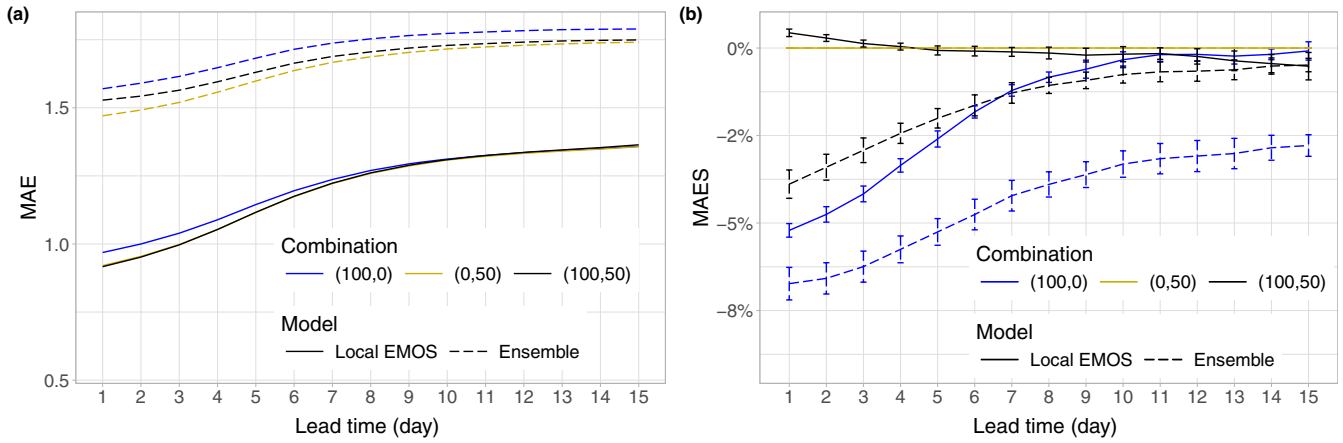

FIGURE 3 (a) Mean absolute error (MAE) of the median of pure low-resolution (100, 0), pure high-resolution (0, 50), and combined dual-resolution (100, 50) raw and post-processed wind-speed forecasts and (b) MAE skill score (MAES) with respect to the corresponding pure high-resolution prediction with 95% confidence intervals as functions of the lead time. EMOS: ensemble model output statistics. [Colour figure can be viewed at wileyonlinelibrary.com]

An identical conclusion can be drawn from studying the RMSE values of the mean forecasts and the matching skill scores; therefore, we refrain from presenting the corresponding figures.

Figure 4 shows the QSS of raw and post-processed wind-speed forecasts from pure low-resolution (100, 0) and combined dual-resolution (100, 50) ensembles, relative to the corresponding pure high-resolution (0, 50) forecasts for selected percentiles, with 95% confidence intervals as functions of lead time. The pure low-resolution ensemble forecast exhibits the same behavior as in Figures 2b and 3b; it significantly underperforms its pure high-resolution counterpart for all lead times and all quantiles studied (note that the QSS for the 50th percentile coincides with the MAES). In contrast, the raw dual-resolution ensemble forecast provides a significant advantage over the (0, 50) combination for all lead times in most cases. The improvement is highest at the most extreme quantiles and gradually decreases towards the center, completely fading for the 80th percentile (and for the 50th percentile as well; see Figure 3b) and after day 2 for the 90th percentile. A different situation can be observed in the case of EMOS post-processed predictions; however, the shapes of the corresponding QSS curves are mostly in line with the matching graphs of Figures 2b and 3b. For short lead times, the pure low-resolution EMOS forecast is significantly behind the reference pure high-resolution-based EMOS model but gradually catches up with the increase of the forecast horizon. In general, the deviations are the largest in the central percentiles, where the corresponding

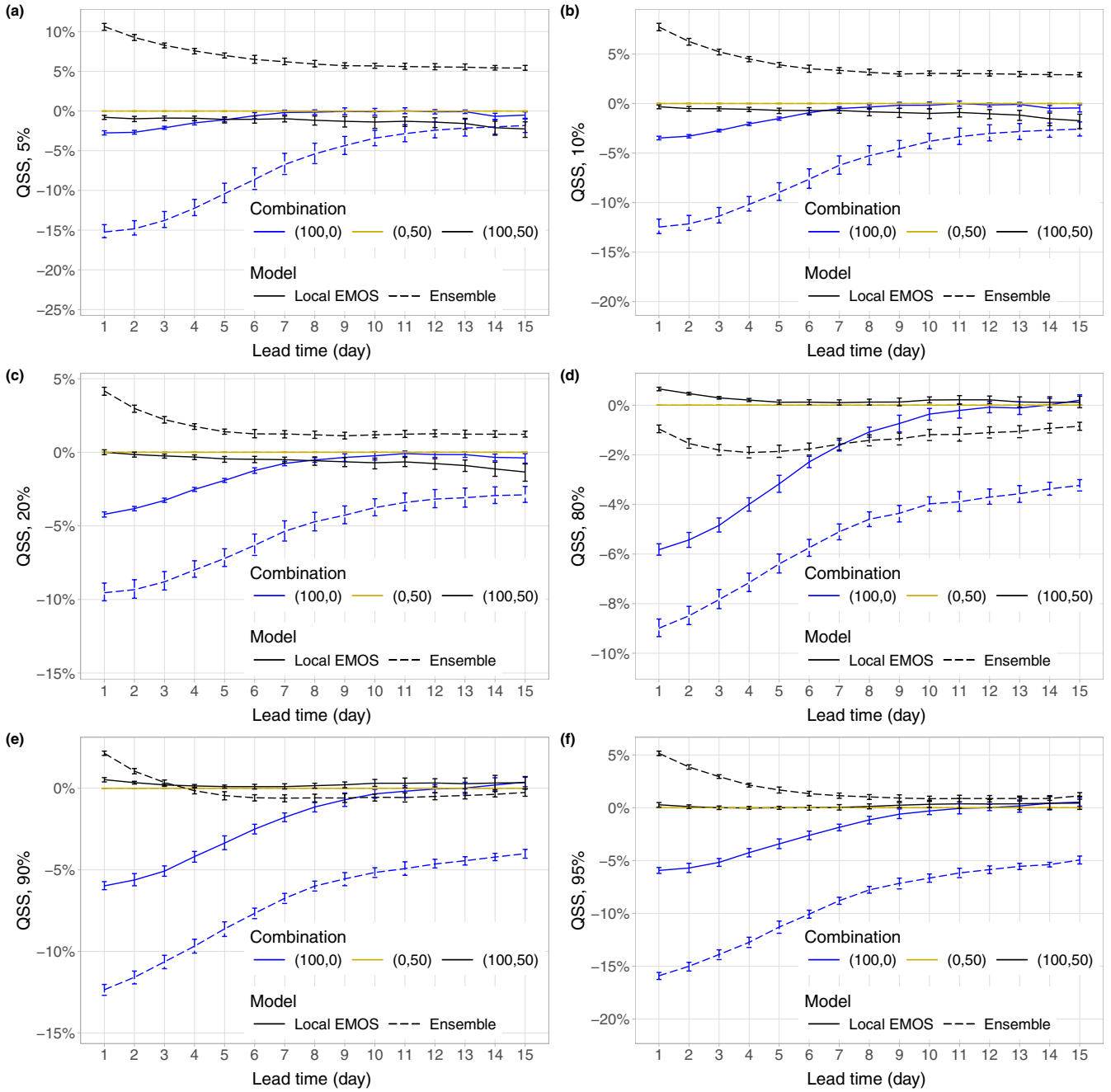

FIGURE 4 Quantile skill score (QSS) of pure low-resolution (100, 0) and combined dual-resolution (100, 50) raw and post-processed wind-speed forecasts with respect to the corresponding pure high-resolution (0, 50) predictions for percentiles (a) 5, (b) 10, (c) 20, (d) 80, (e) 90, and (f) 95 with 95% confidence intervals as functions of the lead time. EMOS: ensemble model output statistics. [Colour figure can be viewed at wileyonlinelibrary.com]

QSS fails to be significantly negative only at the longest lead times. Finally, dual-resolution post-processed forecasts result in significantly positive QSS only for the first 2–3 days and only for the larger percentiles.

Figure 5 shows the mean BS along with the matching skill scores for the previously investigated raw and post-processed forecasts with respect to the corresponding high-resolution (0, 50) predictions. The mean BS values (Figure 5a,c,e) demonstrate that the advantage of post-processing diminishes progressively with increasing wind-speed thresholds. This indicates that though the truncated normal EMOS model generally performs well, for high wind-speed values the EMOS models utilizing distributions with heavier tails, such as log-normal (Baran & Lerch, 2015) or truncated generalized extreme value (Baran *et al.*, 2021), can be more suitable. At

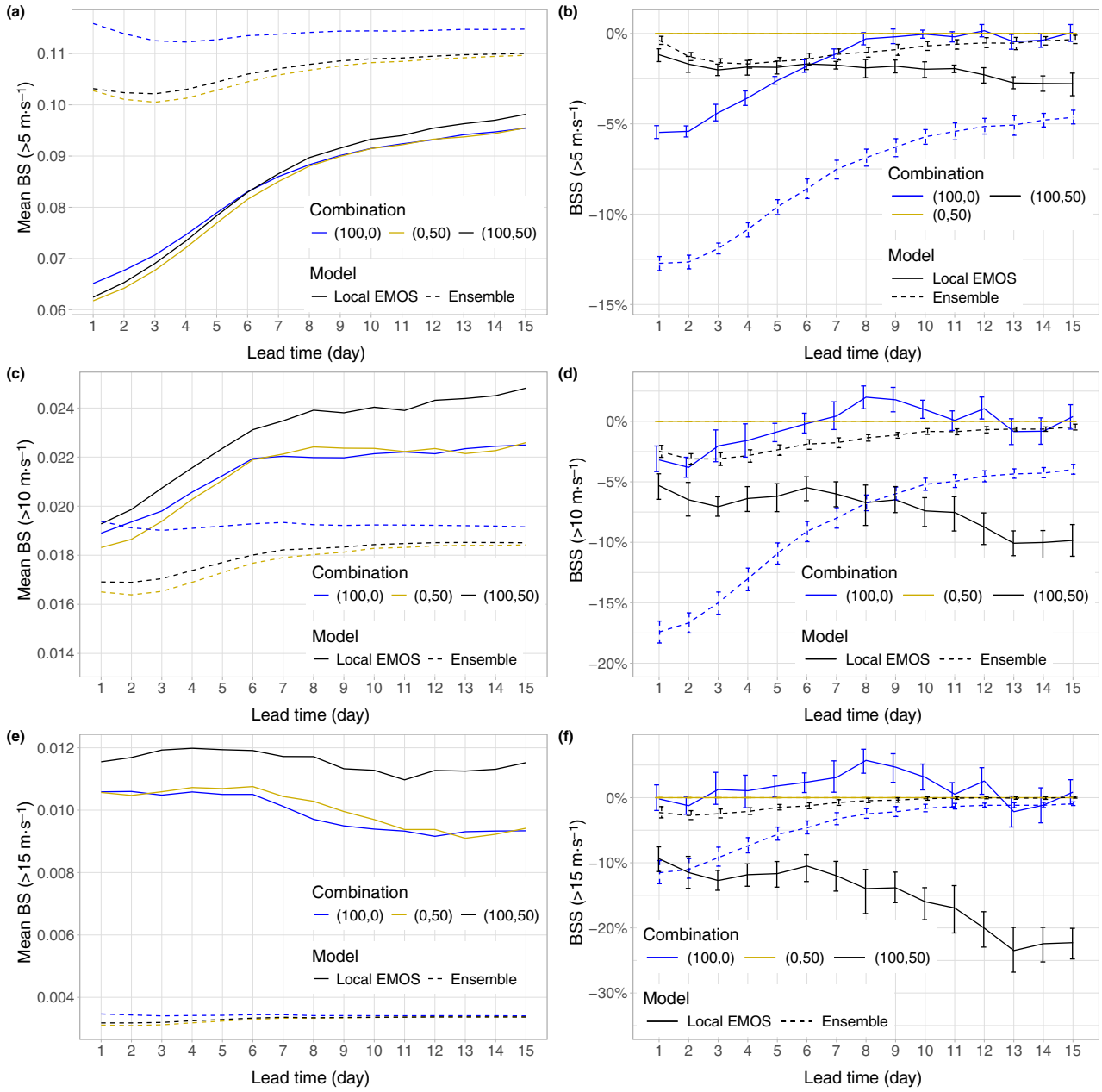

FIGURE 5 (a, c, e) Mean Brier score (BS) of pure low-resolution (100, 0), pure high-resolution (0, 50), and combined dual-resolution (100, 50) raw and post-processed wind-speed forecasts and (b, d, f) Brier skill score (BSS) of pure low-resolution (100, 0) and combined dual-resolution (100, 50) forecasts with respect to the corresponding pure high-resolution (0, 50) predictions with 95% confidence intervals for thresholds (a, b) 5 $m \cdot s^{-1}$, (c, d) 10 $m \cdot s^{-1}$, and (e, f) 15 $m \cdot s^{-1}$ as functions of the lead time. EMOS: ensemble model output statistics. [Colour figure can be viewed at wileyonlinelibrary.com]

the lowest threshold of 5 $m \cdot s^{-1}$, the advantage of the post-processed models is highly pronounced. For shorter lead times, EMOS models that also build on high-resolution forecasts are superior; however, as the forecast horizon increases, the purely low-resolution forecast begins to catch up and, after day 7, outperforms the (100, 50) mixture. According to Figure 5b,d,f, whereas for higher wind-speed thresholds the post-processed pure low-resolution (100, 0) forecasts show the highest average advantage, this benefit is only significant for certain forecast lead times (from days 8 to 11). In contrast, the dual-resolution (100, 50) forecasts exhibit significantly lower skill across all prediction horizons and thresholds examined, and their disadvantage becomes increasingly pronounced as the threshold increases. This behavior can be explained by the short supply of appropriate training data at the tails, meaning that EMOS models based on single-resolution forecasts having fewer parameters are preferred. A different pattern can be observed for the raw forecasts: here, the dual-resolution (100, 50)

configurations exhibit a performance more comparable to that of the high-resolution (0, 50) forecasts. However, neither configuration shows a significant advantage over pure high-resolution forecasts at any forecast lead time or threshold.

4.2 | Forecast mixtures

In the following, our goal is to investigate how the inclusion of high-resolution ($T_{CO}1279$) ensemble members affects the predictive performance of the forecasts when gradually added to a base of 50 low-resolution ($T_{CO}319$) predictions. We separately assess the forecast skill of the raw ensemble (Section 4.2.1) and the locally calibrated EMOS models (Section 4.2.2); the latter choice of the training data selection method is based on the findings of Section 4.1. Specifically, we assess the predictive performance of the raw and post-processed ECMWF wind-speed forecasts by gradually adding 1, 2, 4, 8, 16, and 32 high-resolution ensemble members to the original set of 50 low-resolution forecasts.

4.2.1 | Ensemble predictions

Figure 6 illustrates the predictive accuracy of these combined ensembles. Figure 6a presents the mean CRPS across lead times, and Figure 6b shows the CRPSS relative to the baseline (50, 0) low-resolution ensemble. The inclusion of high-resolution members results in consistent improvements in forecast skill, particularly at short to medium lead times. The greatest improvement (nearly 10% CRPSS) is achieved by the (50, 32) configuration at day

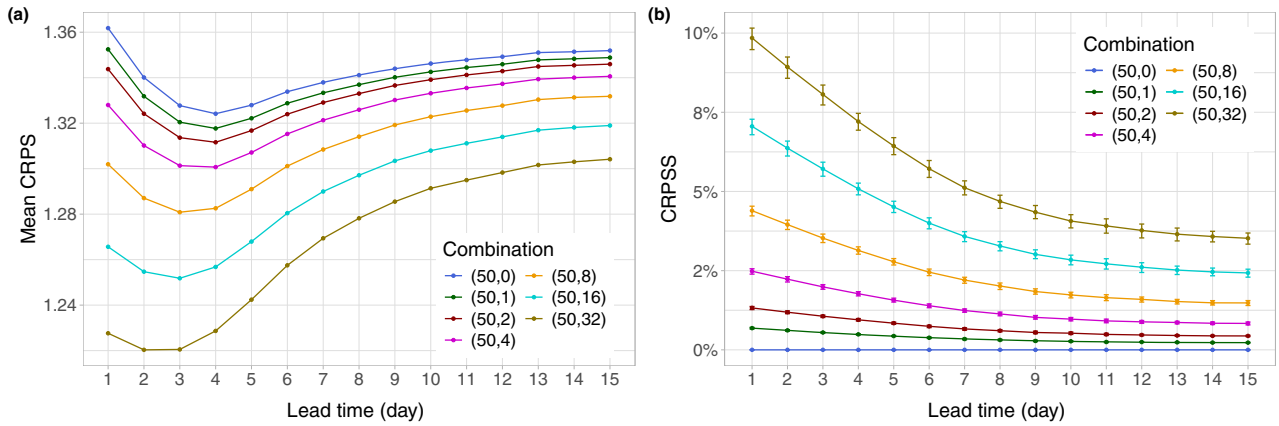

FIGURE 6 (a) Mean continuous ranked probability score (CRPS) of various combinations of low- and high-resolution raw wind-speed ensemble forecasts and (b) continuous ranked probability skill score (CRPSS) of mixtures containing high-resolution members with respect to the pure low-resolution (50, 0) prediction with 95% confidence intervals as functions of the lead time. [Colour figure can be viewed at wileyonlinelibrary.com]

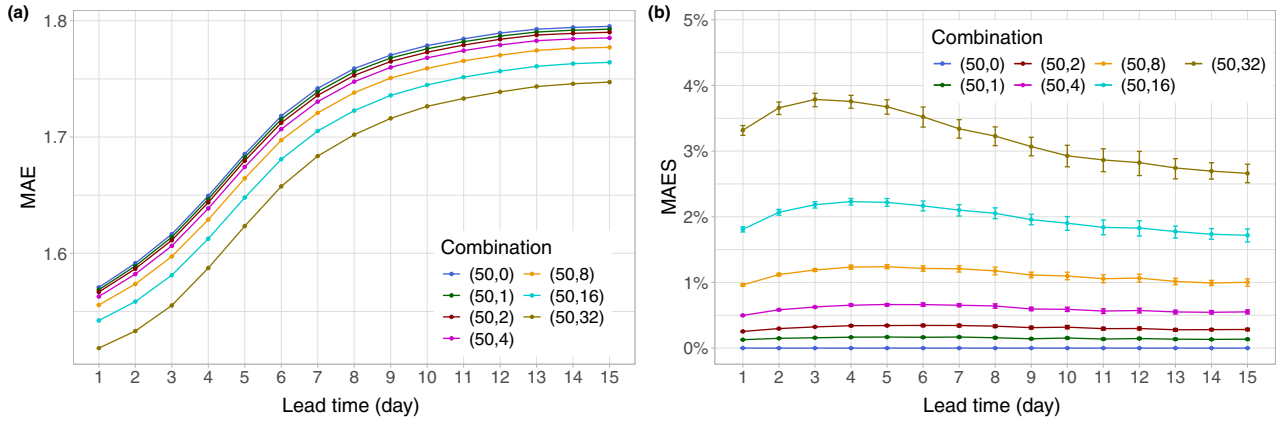

FIGURE 7 (a) Mean absolute error (MAE) of the medians of various combinations of low- and high-resolution raw wind-speed ensemble forecasts and (b) MAE skill score (MAES) of mixtures containing high-resolution members with respect to the pure low-resolution (50, 0) prediction with 95% confidence intervals as functions of the lead time. [Colour figure can be viewed at wileyonlinelibrary.com]

1, followed by (50, 16) with gains of up to 7% and (50, 8) with approximately 5%. Even configurations with only one or two additional high-resolution members, such as (50, 1) or (50, 2), yield slight improvements. Although the (50, 32) set-up provides the most substantial overall benefit relative to the reference, the magnitude of this advantage decreases with increasing forecast lead time across all model configurations; nevertheless, the improvement remains statistically significant for all lead times and ensemble configurations, as confirmed by 95% bootstrap confidence intervals.

In terms of model ranking, a similar pattern can be observed in Figure 7 when analyzing the median of the forecasts: as more members are added to the ensemble consisting solely of low-resolution forecasts, the skill consistently improves. Moreover, it appears that, for all model configurations, the positive impact of high-resolution members is most pronounced at days 3 and 4, and this advantage becomes increasingly evident as more high-resolution members are included. However, unlike with the CRPS, increasing lead time does not necessarily correspond to a decreasing advantage relative to the

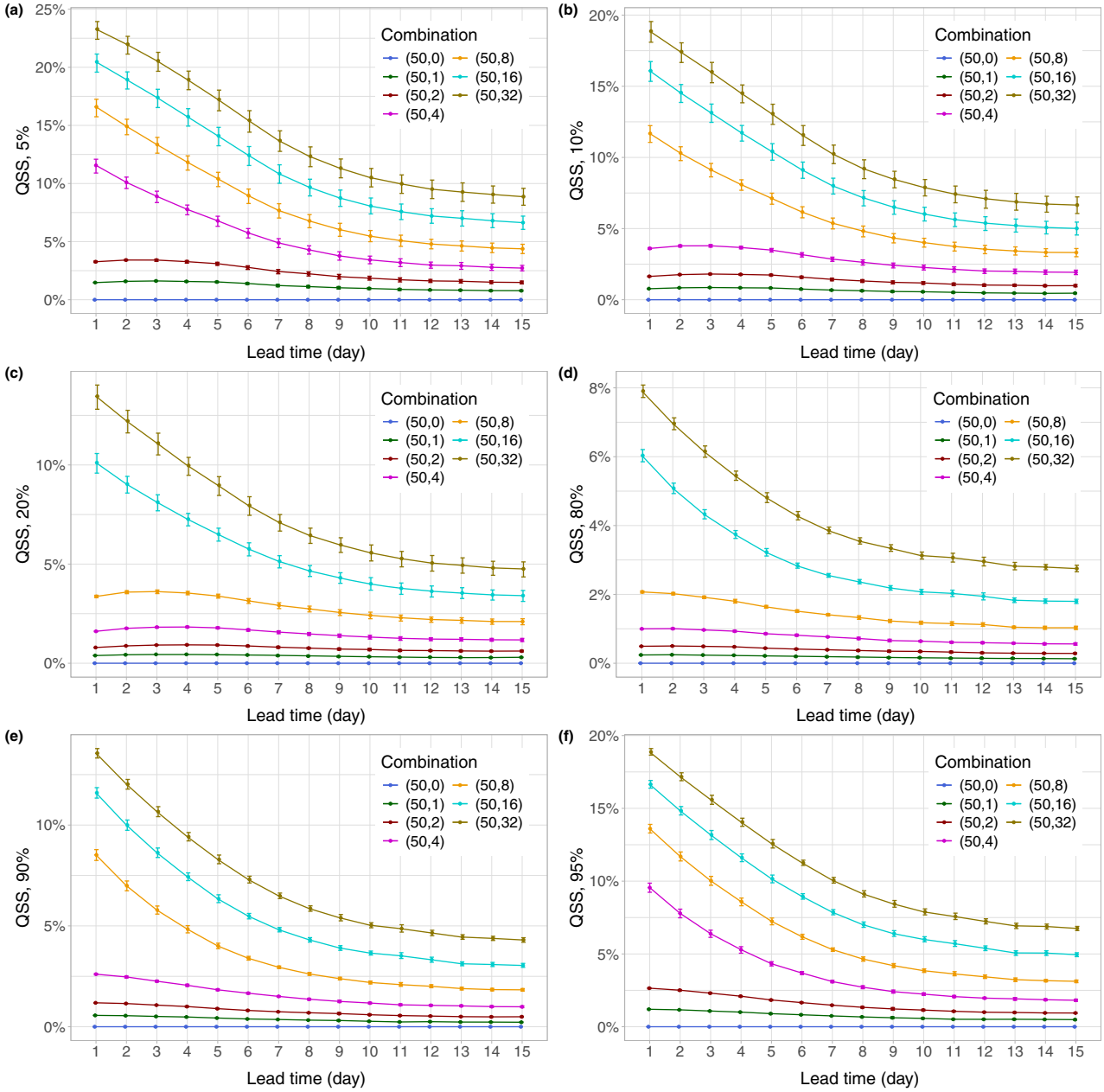

FIGURE 8 Quantile skill score (QSS) with respect to the pure low-resolution (50, 0) ensemble prediction of wind speed for percentiles (a) 5, (b) 10, (c) 20, (d) 80, (e) 90, and (f) 95 of mixtures containing high-resolution members with 95% confidence intervals as functions of the lead time. [Colour figure can be viewed at wileyonlinelibrary.com]

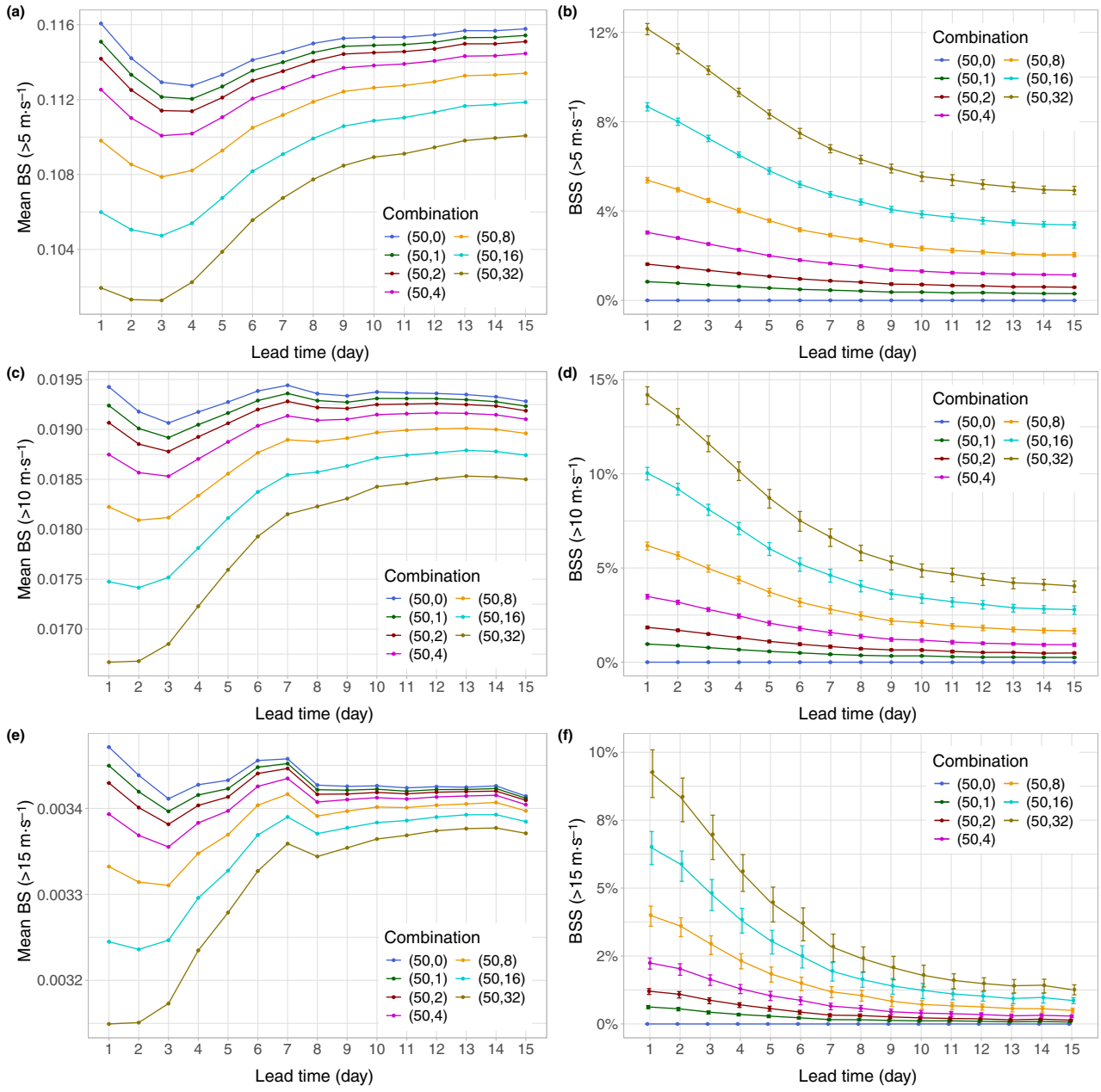

FIGURE 9 (a, c, e) Mean Brier score (BS) of various combinations of raw low- and high-resolution wind-speed forecasts and (b, d, f) Brier skill score (BSS) of mixtures containing high-resolution members with respect to the pure low-resolution (50, 0) prediction with 95% confidence intervals for thresholds (a, b) 5 $m \cdot s^{-1}$, (c, d) 10 $m \cdot s^{-1}$, and (e, f) 15 $m \cdot s^{-1}$ as functions of the lead time. [Colour figure can be viewed at wileyonlinelibrary.com]

reference forecast here. This highlights that the benefit of adding high-resolution ensemble members is particularly clear in the median forecasts, further underscoring their value in enhancing predictive accuracy, especially at these intermediate lead times.

The aforementioned overall forecast ranking is also confirmed by the RMSE values of the forecast means. However, similar to the CRPS, the benefit of adding high-resolution members gradually diminishes with increasing forecast lead time, a pattern that is most pronounced for forecasts using a larger number of high-resolution members. The corresponding figure can be found in Appendix B (Figure B.1).

Figure 8 shows various QSSs for all previously considered ensemble configurations containing high-resolution members with respect to the pure low-resolution forecast. The overall ranking of the models remains consistent with that observed in Figures 6 and 7. However, the magnitude of the improvement varies across different percentiles. For the more extreme percentiles (5% and 95%), forecasts using

only four high-resolution members show a notably greater advantage compared with other percentiles, closely followed by those incorporating 8, 16, and 32 high-resolution members. More generally, we observe that symmetrically positioned percentiles (such as 20%–80% and 10%–90%) exhibit similar trends in model performance as more high-resolution members are added, though the magnitude of the improvements differs. For instance, at the 20% and 80% percentiles, the largest gains are achieved by configurations including 16 or 32 high-resolution members, whereas at the 10% and 90% percentiles the model using eight high-resolution members also ranks among the top performers.

Figure 9 reporting the BSs for the events of wind speed exceeding thresholds 5, 10, and 15 $m \cdot s^{-1}$ for different raw forecast mixtures does not change the general picture. The overall trend is similar across all three thresholds: the more high-resolution members are added to the 50 low-resolution members, the better the predictive performance becomes. However, for all thresholds, these improvements become less pronounced as the forecast lead time increases. Based on Figure 9b,d,f, significant differences in forecast performance can be observed between the ensemble configurations; however, for higher wind-speed thresholds, the confidence intervals become increasingly wide, indicating greater uncertainty.

4.2.2 | Post-processed forecasts

In the following, we evaluate the predictive performance of the forecast mixtures introduced in the previous section, post-processed using the local EMOS method, as this model performed best for the operational forecasts evaluated in Section 4.1.

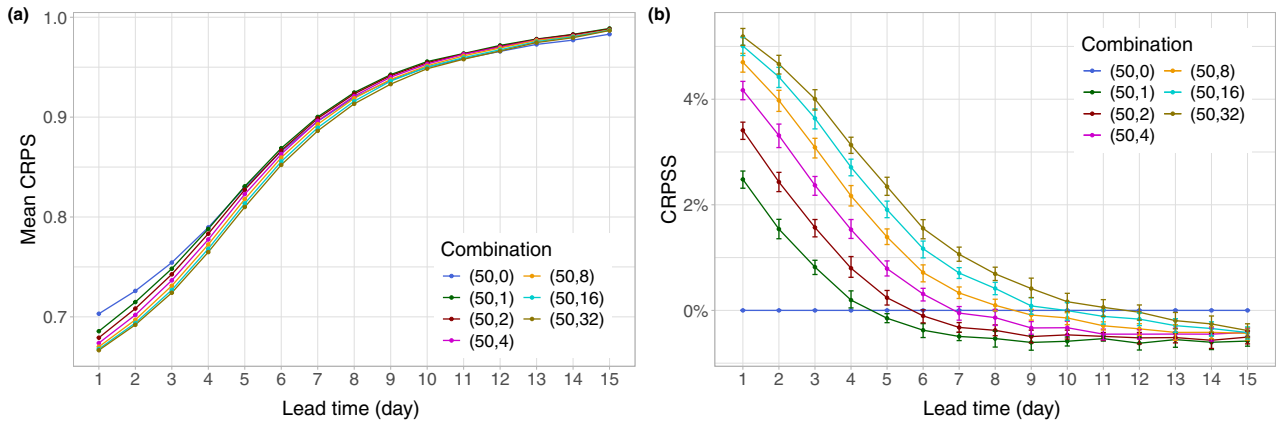

FIGURE 10 (a) Mean continuous ranked probability score (CRPS) of various combinations of locally post-processed low- and high-resolution wind-speed forecasts and (b) continuous ranked probability skill score (CRPSS) of mixtures containing high-resolution members with respect to the pure low-resolution (50, 0) prediction with 95% confidence intervals as functions of the lead time. [Colour figure can be viewed at wileyonlinelibrary.com]

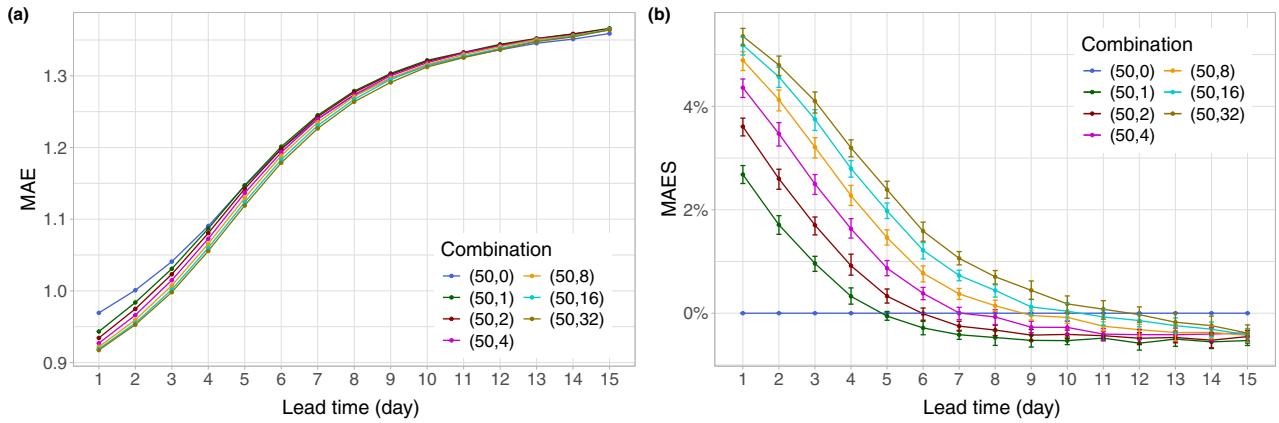

FIGURE 11 (a) Mean absolute error (MAE) of the medians of various combinations of locally post-processed low- and high-resolution wind-speed forecasts and (b) MAE skill score (MAES) of mixtures containing high-resolution members with respect to the pure low-resolution (50, 0) prediction with 95% confidence intervals as functions of the lead time. [Colour figure can be viewed at wileyonlinelibrary.com]

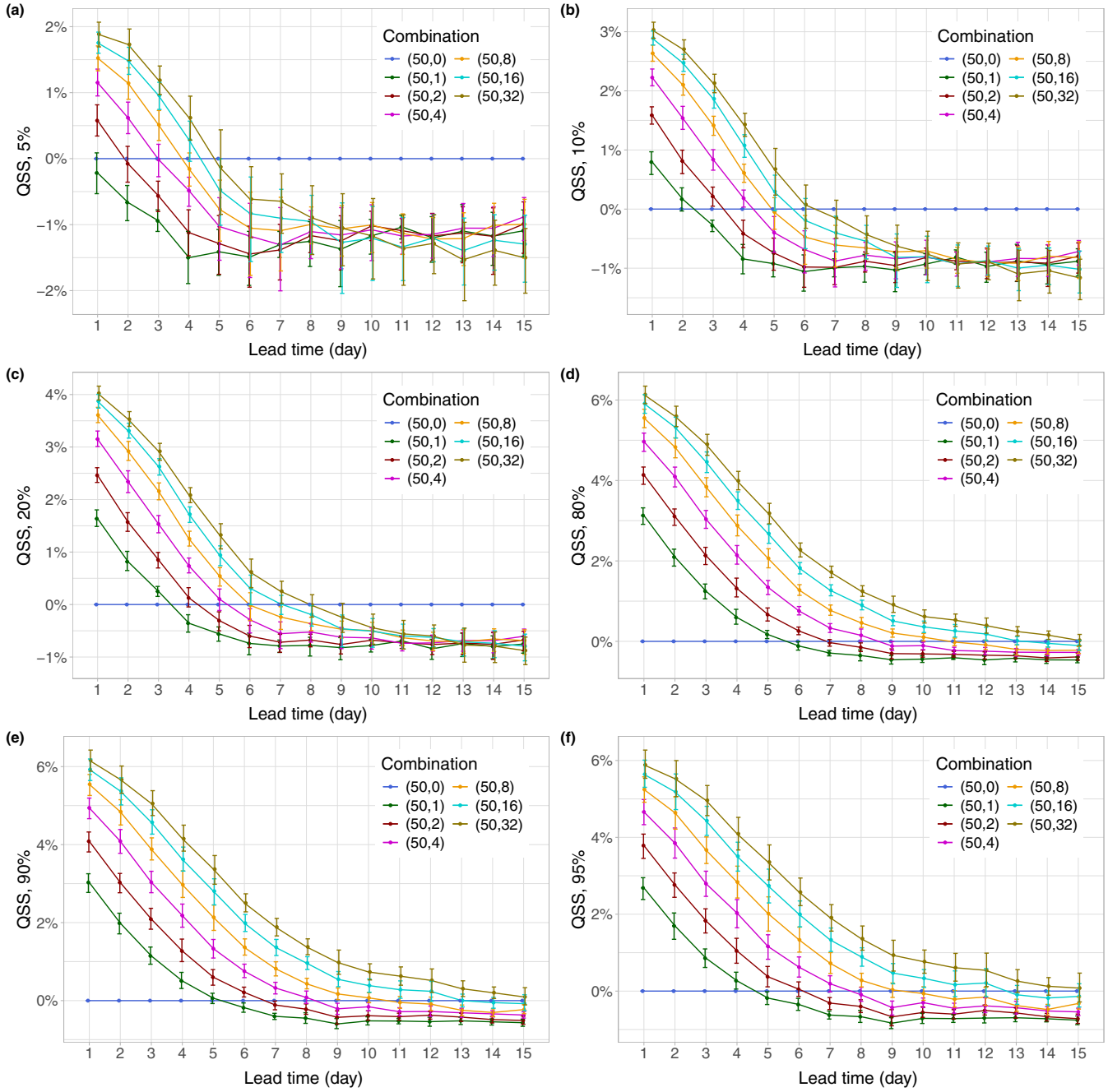

FIGURE 12 Quantile skill score (QSS) with respect to the locally post-processed pure low-resolution (50, 0) prediction of wind speed for percentiles (a) 5, (b) 10, (c) 20, (d) 80, (e) 90, and (f) 95 of post-processed mixtures containing high-resolution members with 95% confidence intervals as functions of the lead time. [Colour figure can be viewed at wileyonlinelibrary.com]

Figure 10a displays the mean CRPS values for the different post-processed forecast combinations. Similar to Figure 6a, the more high-resolution members that are included in the post-processing, the better the predictive performance. Furthermore, differences between the models gradually diminish as the forecast lead time increases, indicating that the benefit of including further high-resolution forecasts becomes negligible at longer lead times. One should also note that post-processing substantially reduces the deviations of the various combinations, which is completely in line with the findings of Baran *et al.* (2019) (see also Figure 2a). Finally, the skill scores of Figure 10b reveal that all models utilizing high-resolution predictions significantly outperform the reference forecast based solely on low-resolution members up to day 4. Beyond this point, the inclusion of additional high-resolution members helps maintain this performance advantage over longer lead times, with the

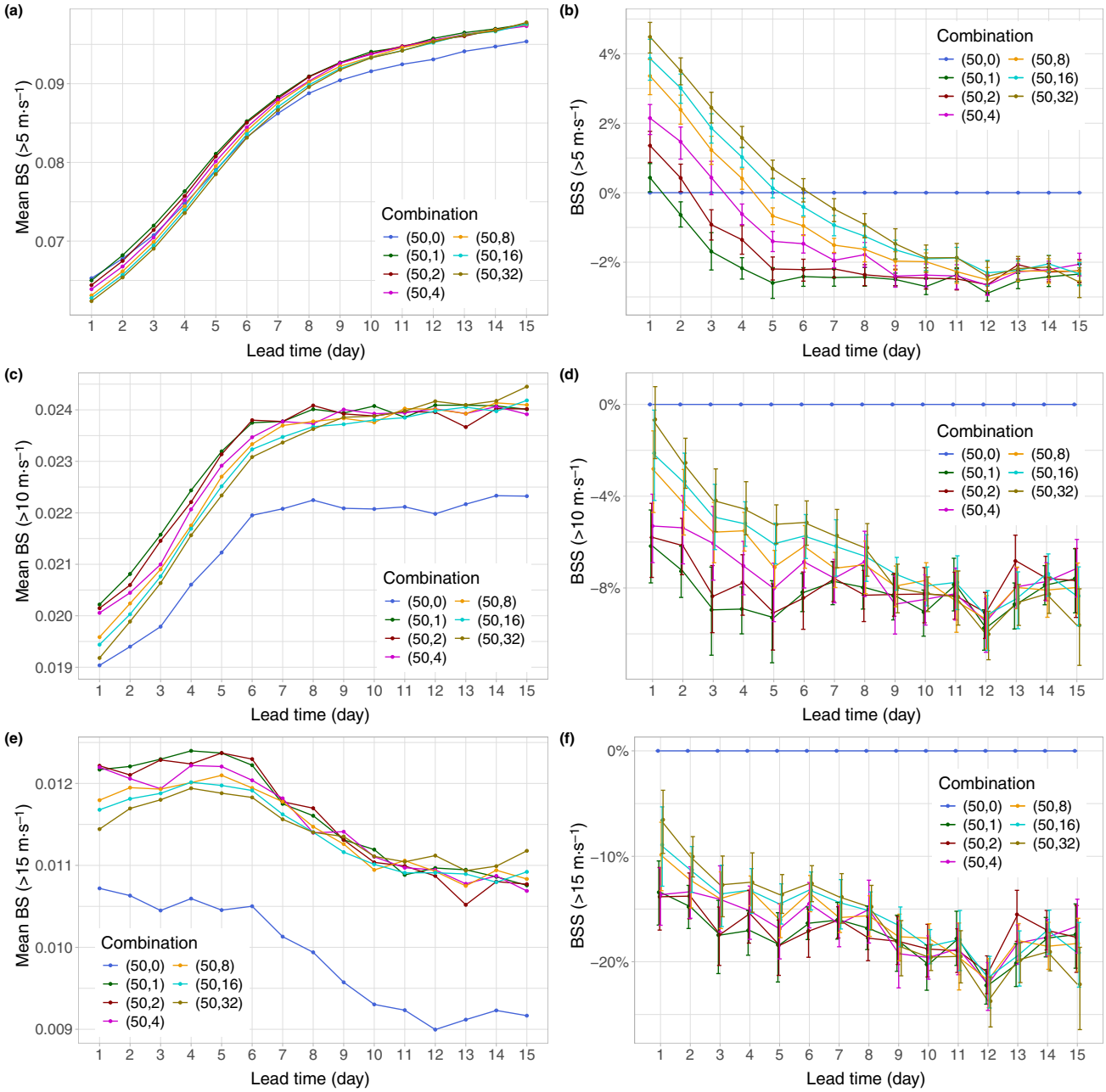

FIGURE 13 (a, c, e) Mean Brier score (BS) of various combinations of locally post-processed low- and high-resolution wind-speed forecasts and (b, d, f) Brier skill score (BSS) of mixtures containing high-resolution members with respect to the pure low-resolution (50, 0) prediction with 95% confidence intervals for thresholds (a, b) 5 m · s^{-1}, (c, d) 10 m · s^{-1}, and (e, f) 15 m · s^{-1} as functions of the lead time. [Colour figure can be viewed at wileyonlinelibrary.com]

(50, 32) combination showing the greatest overall benefit. These conclusions are further supported by Figure 11, which presents the MAE values of the EMOS medians and corresponding skill scores for the same post-processed forecast combinations, and also by the RMSE values of the EMOS means and related skill scores (see Figure B.2 in Appendix B).

We also examined the performance of the post-processed models based on the QSS, as shown in Figure 12. The overall behavior of the models is consistent with previous findings: the inclusion of more high-resolution members in the post-processing generally leads to greater improvements over the pure low-resolution reference forecasts. However, the magnitude of this improvement varies across different percentiles. For lower percentiles, the benefit of including high-resolution members is smaller. In an extreme case (at the 5th percentile) the (50, 1) combination even performs significantly worse

than the pure low-resolution reference. In contrast, the performance advantage of all models increases consistently with higher percentiles.

Figure 13 shows the mean BSs corresponding to the locally post-processed forecast combinations for the low ($5\ \mathrm{m \cdot s^{-1}}$), moderate ($10\ \mathrm{m \cdot s^{-1}}$), and high ($15\ \mathrm{m \cdot s^{-1}}$) wind-speed thresholds, and the matching skill scores with respect to the EMOS model based on the pure low-resolution $(50, 0)$ prediction. For low wind speeds, the patterns are similar to those observed in the previous CRPS and MAE figures. In general, including more ensemble members leads to better model performance, though, as before, this advantage holds only up to a certain forecast horizon. At low wind speeds, the benefit of the pure low-resolution forecast is now even more pronounced (Figure 13a); and according to Figure 13b, the advantage of models incorporating high-resolution members remains significant only up to day 5 at best. For the moderate and high wind-speed thresholds, the picture changes more drastically: none of the post-processed models show a statistically significant improvement over the pure low-resolution $(50, 0)$ reference. However, the superiority of the post-processed single-resolution forecasts in predicting tail events with low probability (the relative frequencies of wind-speed observations in the verification dataset exceeding $10\ \mathrm{m \cdot s^{-1}}$ and $15\ \mathrm{m \cdot s^{-1}}$ are 1.79% and 0.31% respectively) is in line with our findings based on Figure 5.

5 | CONCLUSIONS

We investigate the forecast skill of raw and post-processed 50-member medium-range and 100-member extended-range forecasts of 10-m wind speed up to day 15 at $\mathrm{T_{CO}1279}$ (high) and $\mathrm{T_{CO}319}$ (low) resolutions, respectively, and their various combinations. This study demonstrates that, among the raw ECMWF ensemble forecasts, the low-resolution $(100, 0)$ configuration consistently shows the weakest predictive performance for all verification measures investigated, whereas the high-resolution $(0, 50)$ and dual-resolution $(100, 50)$ set-ups perform similarly. Nevertheless, in most cases, the former maintains a slight but significant edge, especially for short lead times; the dual-resolution forecast is significantly superior to the single-resolution predictions only in terms of the mean QSs for the low and very high percentiles.

Post-processing, particularly with local and semi-local EMOS models, offers substantial improvements in forecast skill in terms of all evaluation metrics studied, except for the BSs for medium and high thresholds. The local method demonstrates the best overall performance, especially during the early forecast days. Moreover, when beneficial, statistical calibration considerably reduces the differences between the various forecast set-ups. In contrast to the raw ensemble predictions, the pure low-resolution EMOS forecast starts off lagging significantly behind the high-resolution EMOS benchmark but improves gradually over time, slowly closing the performance gap. Furthermore, for very short lead times, the use of EMOS post-processed dual-resolution forecasts seems to be advantageous; it is significantly superior to the single-resolution set-ups in terms of the mean CRPS and mean QS for percentiles not less than the median.

The aforementioned findings suggest that spatial resolution is superior to the ensemble size; augmenting a sufficiently large ensemble of high-resolution forecasts with low-resolution predictions does not necessarily result in a gain in forecast skill. Furthermore, for EMOS post-processed dual-resolution forecasts, the means/variances of the 100-member low-resolution forecasts are substantially less correlated with the corresponding EMOS location/squared scale parameters than their high-resolution counterparts based on 50 ensemble members. However, our study also highlights the clear benefit of the other direction; namely, incorporating high-resolution members into low-resolution ensemble forecasts. Again, based on multiple verification scores, the addition of high-resolution members consistently enhances the predictive performance of the raw ECMWF wind-speed forecasts for all forecast horizons studied. The most significant gains are observed in configurations with the highest number of high-resolution members, though even minimal inclusion yields slight improvements. After post-processing, the advantage of dual-resolution modeling depends on the number of high-resolution predictions incorporated. In general, the more high-resolution members that are involved, the longer the lead time until the superiority of the dual-resolution combination over the EMOS model relying merely on the low-resolution ensemble is significant.

The results of the current study suggest several avenues for further research. For post-processing, we utilized a very simple univariate approach incorporating neither spatial dependencies between the SYNOP stations nor temporal dependencies between the different forecast horizons. Thus, on the one hand, one can investigate the performance of more sophisticated state-of-the-art post-processing methods, such as the parametric machine-learning-based DRN (Rasp & Lerch, 2018) or versions of the non-parametric quantile regression (e.g., Bremnes, 2020; Song *et al.*, 2024) in the dual-resolution set-up. On the other hand, the study of multivariate post-processing approaches—classical two-step methods such as the ensemble copula coupling (Schefzik

et al., 2013) or the Schaake shuffle (Clark *et al.*, 2004) or data-driven techniques such as generative adversarial networks (Dai & Hemri, 2021) or scoring-rule-based generative models (Chen *et al.*, 2024)—can also lead us to a better understanding of the effects of mixing high- and low-resolution ensemble members. Finally, the increasing popularity of data-driven weather forecasts and the launch of the ensemble version of the ECMWF's Artificial Intelligence Forecasting System (AIFS-CRPS; Lang *et al.*, 2026), currently issued at a 28 km grid resolution, naturally induces the question of whether mixing IFS and AIFS-CRPS ensemble predictions results in improved predictive performance.

ACKNOWLEDGEMENTS

This work was supported by the EKÖP-24-3-II University Research Scholarship Program of the Ministry for Culture and Innovation, funded by the National Research, Development and Innovation Fund. We also gratefully acknowledge the support of the National Research, Development, and Innovation Office under grant no. K142849. Furthermore, we are indebted to Martin Leutbecher for his suggestions and for providing the ECMWF dual-resolution wind-speed data. Last, but not least, we thank the two anonymous reviewers, whose constructive comments helped to improve the manuscript.

DATA AVAILABILITY STATEMENT

The data used in this study consist of archived operational data from ECMWF. This is available under a CC BY 4.0 license and access can be requested via ECMWF's web archive (https://apps.ecmwf.int/archive-catalogue/?class=od).

The data that support the findings of this study are available from ECMWF. Restrictions apply to the availability of these data, which were used under license for this study. Data are available from https://apps.ecmwf.int/archive-catalogue/?class=od with the permission of ECMWF.

ENDNOTE

1ECMWF Newsletter No. 176, Summer 2023.

ORCID

Sándor Baran https://orcid.org/0000-0003-1035-004X
Mária Lakatos https://orcid.org/0009-0007-5574-0240

REFERENCES

Baran, S. & Lakatos, M. (2024) Clustering-based spatial interpolation of parametric postprocessing models. *Weather and Forecasting*, 39, 1591–1604.

Baran, S. & Lerch, S. (2015) Log-normal distribution based emos models for probabilistic wind speed forecasting. *Quarterly Journal of the Royal Meteorological Society*, 141, 2289–2299.

Baran, S., Leutbecher, M., Szabó, M. & Ben Bouallègue, Z. (2019) Statistical post-processing of dual-resolution ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 145, 1705–1720.

Baran, S., Szokol, P. & Szabó, M. (2021) Truncated generalized extreme value distribution-based ensemble model output statistics model for calibration of wind speed ensemble forecasts. *Environmetrics*, 32, e2678.

Ben Bouallègue, Z., Haiden, T., Weber, N.J., Hamill, T.M. & Richardson, D.S. (2020) Accounting for representativeness in the verification of ensemble precipitation forecasts. *Monthly Weather Review*, 148, 2049–2062.

Bentzien, S. & Friederichs, P. (2014) Decomposition and graphical portrayal of the quantile score. *Quarterly Journal of the Royal Meteorological Society*, 140, 1924–1934.

Bremnes, J.B. (2020) Ensemble postprocessing using quantile function regression based on neural networks and Bernstein polynomials. *Monthly Weather Review*, 148, 403–414.

Buizza, R. (2018) Ensemble forecasting and the need for calibration. In: Vannitsem, S., Wilks, D.S. & Messner, J.W. (Eds.) *Statistical postprocessing of ensemble forecasts*. Amsterdam: Elsevier, pp. 15–48.

Chen, J., Janke, T., Steinke, F. & Lerch, S. (2024) Generative machine learning methods for multivariate ensemble postprocessing. *The Annals of Applied Statistics*, 18, 159–189.

Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B. & Wilby, R. (2004) The schaake shuffle: a method for reconstructing space–time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology*, 5, 243–262.

Dai, Y. & Hemri, S. (2021) Spatially coherent postprocessing of cloud cover ensemble forecasts. *Monthly Weather Review*, 149, 3923–3937.

ECMWF. (2024) *IFS documentation CY49R1 – Part V: ensemble prediction system*. Reading: ECMWF.

Gascón, E., Lavers, D., Hamill, T.M., Richardson, D.S., Ben Bouallègue, Z., Leutbecher, M. et al. (2019) Statistical postprocessing of dual-resolution ensemble precipitation forecasts across Europe. *Quarterly Journal of the Royal Meteorological Society*, 145, 3218–3235.

Gneiting, T. (2011) Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106, 746–762.

Gneiting, T. & 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, A.H. & Goldman, T. (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133, 1098–1118.

Hemri, S., Scheuerer, M., Pappenberger, F., Bogner, K. & Haiden, T. (2014) Trends in the predictive performance of raw ensemble weather forecasts. *Geophysical Research Letters*, 41, 9197–9205.

Jordan, A., Krüger, F. & Lerch, S. (2019) Evaluating probabilistic forecasts with scoringRules. *Journal of Statistical Software*, 90, 1–37.

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.

Lang, S., Alexe, M., Clare, M.C.A., Roberts, C., Adewoyin, R., Ben Bouallègue, Z. et al. (2026) AIFS-CRPS: ensemble forecasting using a model trained with a loss function based on the continuous ranked probability score. *npj Artificial Intelligence*, 2, 18.

Lerch, S. & Baran, S. (2017) Similarity-based semilocal estimation of post-processing models. *Journal of the Royal Statistical Society*, 66C, 29–51.

Leutbecher, M. & Ben Bouallègue, Z. (2020) On the probabilistic skill of dual-resolution ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 146, 707–723.

Murphy, A.H. (1973) Hedging and skill scores for probability forecasts. *Journal of Applied Meteorology*, 12, 215–223.

Politis, D.N. & Romano, J.P. (1994) The stationary bootstrap. *Journal of the American Statistical Association*, 89, 1303–1313.

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

Schefzik, R., Thorarinsdottir, T.L. & Gneiting, T. (2013) Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, 28, 616–640.

Song, M., Yang, D., Lerch, S., Xia, X., Yagli, G.M., Bright, J.M. et al. (2024) Non-crossing quantile regression neural network as a calibration tool for ensemble weather forecasts. *Advances in Atmospheric Sciences*, 41, 1417–1437.

Szabó, M., Gascón, E. & Baran, S. (2023) Parametric postprocessing of dual-resolution precipitation forecasts. *Weather and Forecasting*, 38, 1313–1322.

Taillardat, M. (2021) Skewed and mixture of Gaussian distributions for ensemble postprocessing. *Atmosphere*, 12, 966.

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*, 173A, 371–388.

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.

Veldkamp, S., Whan, K., Dirksen, S. & Schmeits, M. (2021) Statistical postprocessing of wind speed forecasts using convolutional neural networks. *Monthly Weather Review*, 149, 1141–1152.

Wilks, D.S. (2019) *Statistical methods in the atmospheric sciences*, 4th edition. Amsterdam: Elsevier.

How to cite this article: Baran, S. & Lakatos, M. (2026) Statistical post-processing of operational dual-resolution wind-speed ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, e70201. Available from: https://doi.org/10.1002/qj.70201

APPENDIX A. PREDICTIVE PERFORMANCE OF LOCAL AND SEMI-LOCAL EMOS MODELS

Here, we extend the findings of Section 4.1 by comparing the CRPS and MAE values of the regional, local, and semi-local EMOS models based on the $(0, 50)$ pure high-, $(100, 0)$ pure low-, and $(100, 50)$ dual-resolution ensemble forecasts.

In the semi-local approach, clustering features were extracted from the training dataset, and k-means clustering was performed based on these features. In particular, following the approach of Lerch and Baran (2017), for a given location, we considered 12 equidistant quantiles of the climatological CDF and 12 equidistant quantiles of the CDF of the forecast error of the ensemble mean over the actual training period. The optimal number of clusters was determined alongside the length of the optimal training period by systematically testing a wide range of cluster numbers. Similar to the local EMOS, the corresponding semi-local models based on 30-, 60-, and 90-day rolling training windows were evaluated over the period from October 13, 2023, to May 31, 2024, using several different verification metrics. This analysis resulted in an optimal set-up of 90 clusters and a 60-day training period, the length of which is also used for the regional EMOS models. As mentioned in Section 4, the comparison of the various raw and post-processed forecasts is based on the 262-day verification period between September 13, 2023, and May 31, 2024.

Figure A.1 is an extension of Figure 2 with the CRPS values and corresponding skill scores of the regional and semi-local EMOS models; however, for the sake of a cleaner look, the error bars for the skill scores are now omitted. From the three EMOS approaches, the regional EMOS model results in the poorest mean CRPS values; and in terms of the forecast combinations, it resembles the performance of the raw ensemble. The $(100, 0)$ pure low-resolution forecast is far behind the $(0, 50)$ pure high-resolution one and the $(100, 50)$ mixture for all lead times, which are rather close to each other. However, according to Figure A.1b, the latter has a minor advantage. In terms of forecast combinations, local and semi-local EMOS behave similarly: mixing the forecasts barely results in an advantage over the high-resolution prediction, whereas models based on the pure low-resolution ensemble lag behind, but their skills are gradually improving as the forecast horizon increases.

According to Figure A.2, which is the counterpart of Figure 3, regional post-processing offers less of an advantage for median forecasts than it did for CRPS in Figure A.1. Meanwhile, the local and semi-local EMOS methods remain closely aligned, with the local approach

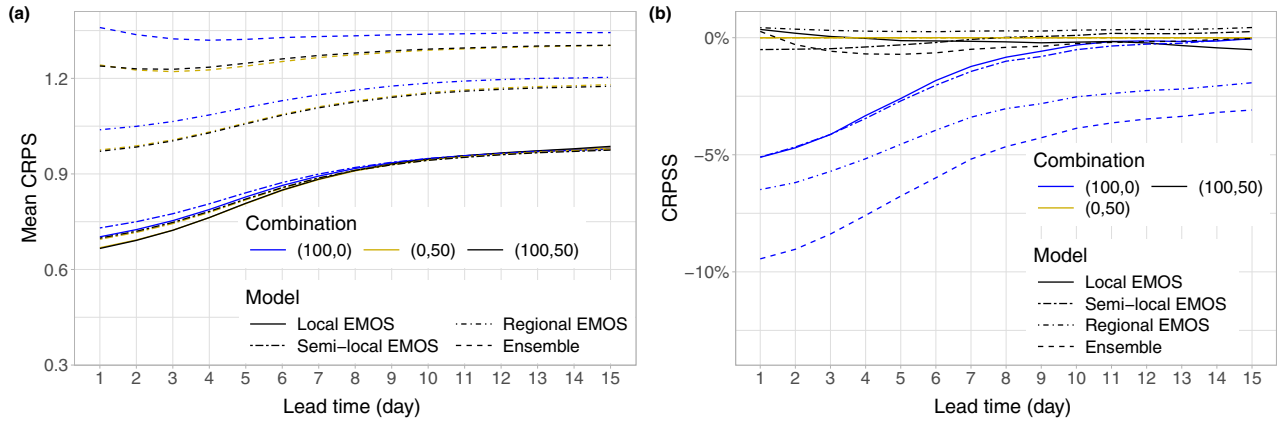

FIGURE A.1 (a) Mean continuous ranked probability score (CRPS) of pure low-resolution (100, 0), pure high-resolution (0, 50), and combined dual-resolution (100, 50) raw and post-processed wind-speed forecasts and (b) continuous ranked probability skill score (CRPSS) with respect to the corresponding pure high-resolution prediction as functions of the lead time. EMOS: ensemble model output statistics. [Colour figure can be viewed at wileyonlinelibrary.com]

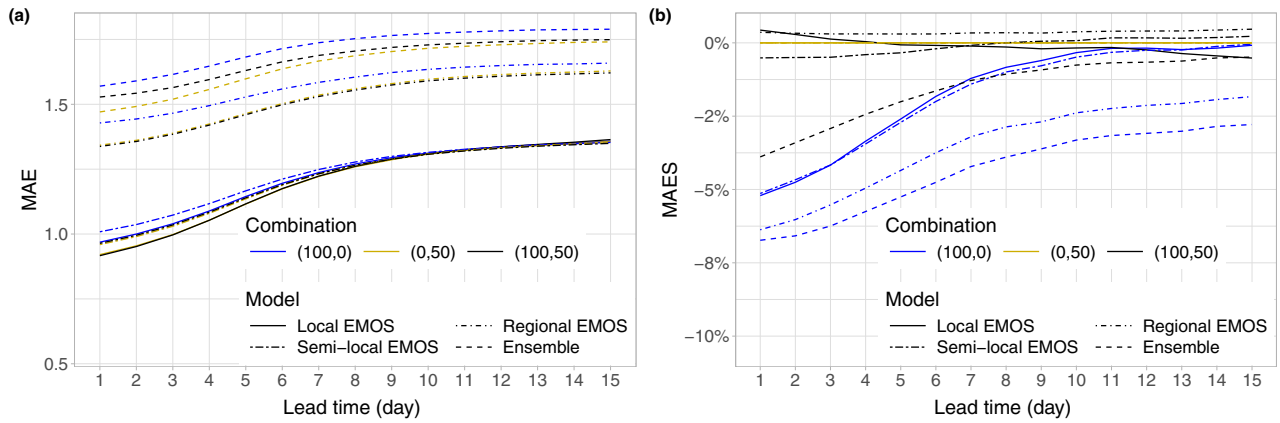

FIGURE A.2 (a) Mean absolute error (MAE) of the median of pure low-resolution (100, 0), pure high-resolution (0, 50), and combined dual-resolution (100, 50) raw and post-processed wind-speed forecasts and (b) mean absolute error skill score MAES) with respect to the corresponding pure high-resolution prediction with 95% confidence intervals as functions of the lead time. EMOS: ensemble model output statistics. [Colour figure can be viewed at wileyonlinelibrary.com]

providing the best performance across all resolution set-ups during the early forecast days.

APPENDIX B. ROOT-MEAN-SQUARE ERRORS FOR FORECAST MIXTURES

As mentioned in Section 4.2, in terms of the RMSE of the mean forecasts, the ranking of the various mixtures studied remains the same as in terms of the mean CRPS or MAE of the median, both for raw and post-processed predictions. However, there are some special characteristics in the behavior of this score that are discussed in the following.

Figure B.1 presents the mean RMSE and the corresponding skill score values for the raw ensemble combinations studied in Section 4.2.1. As expected, the longer the lead time, the larger the RMSE. The differences between the various combinations are significant for all forecast horizons; however, in contrast to the MAE skill scores in Figure 7b, the advantage of ensemble forecasts incorporating high-resolution members over the 50-member pure low-resolution ensemble prediction gradually fades with the increase of the forecast horizon.

As displayed in Figure B.2a, post-processing substantially improves the RMSE of the mean forecast for all combinations and greatly decreases the differences among them. The behavior of this score is highly in line with that of the mean CRPS and MAE of the EMOS median; however, as the comparison of Figure B.2b with Figures 10b and 11b reveals, all RMSE skill scores are slightly lower than the corresponding CRPSS and MAES values.

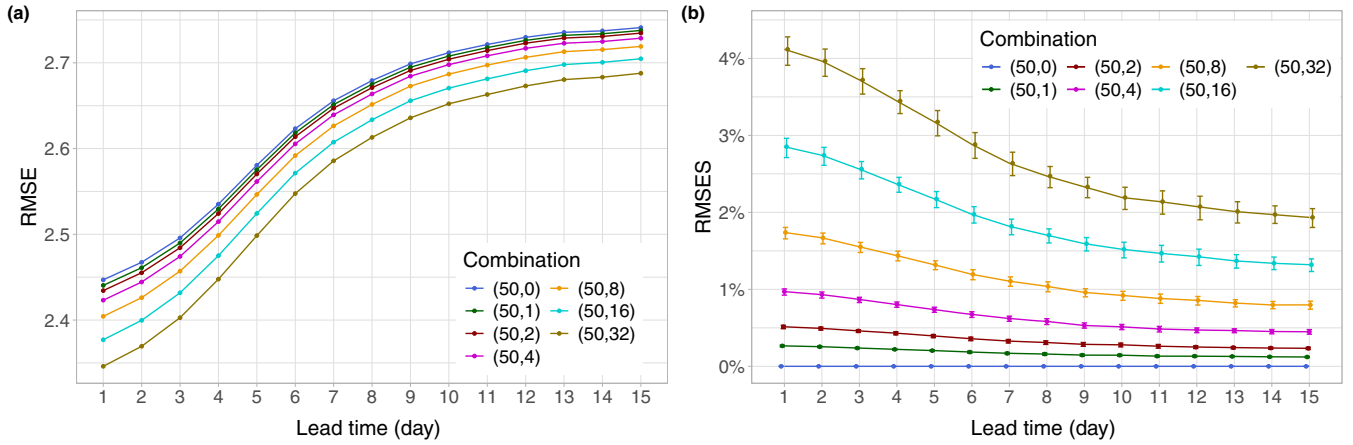

FIGURE B.1 (a) Root-mean-squared error (RMSE) of the means of various combinations of low- and high-resolution raw wind-speed ensemble forecasts and (b) RMSE skill score (RMSES) of mixtures containing high-resolution members with respect to the pure low-resolution $(50, 0)$ prediction with 95% confidence intervals as functions of the lead time. [Colour figure can be viewed at wileyonlinelibrary.com]

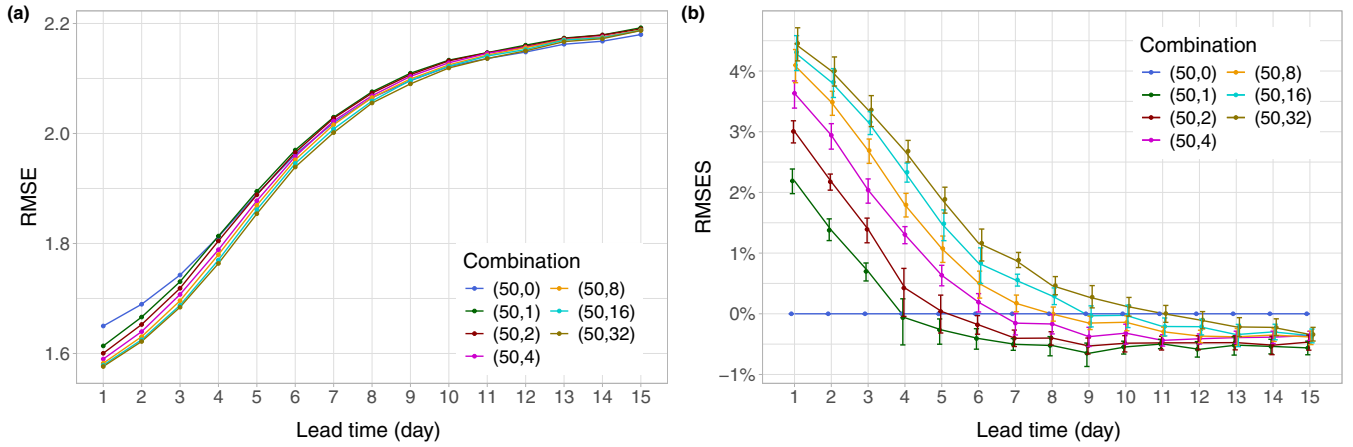

FIGURE B.2 (a) Root-mean-squared error (RMSE) of the means of various combinations of locally post-processed low- and high-resolution wind-speed forecasts and (b) RMSE skill score (RMSES) of mixtures containing high-resolution members with respect to the pure low-resolution $(50, 0)$ prediction with 95% confidence intervals as functions of the lead time. [Colour figure can be viewed at wileyonlinelibrary.com]