

# Bayesian Model Averaging for Wind Speed Ensemble Forecasts Using Wind Speed and Direction

SIRI SOFIE EIDE AND JOHN BJØRNAR BREMNES

*Norwegian Meteorological Institute, Oslo, Norway*

INGELIN STEINSLAND

*Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway*

(Manuscript received 5 July 2017, in final form 19 October 2017)

## ABSTRACT

In this paper, probabilistic wind speed forecasts are constructed based on ensemble numerical weather prediction (NWP) forecasts for both wind speed and wind direction. Including other NWP variables in addition to the one subject to forecasting is common for statistical calibration of deterministic forecasts. However, this practice is rarely seen for ensemble forecasts, probably because of a lack of methods. A Bayesian modeling approach (BMA) is adopted, and a flexible model class based on splines is introduced for the mean model. The spline model allows both wind speed and wind direction to be included nonlinearly. The proposed methodology is tested for forecasting hourly maximum 10-min wind speeds based on ensemble forecasts from the European Centre for Medium-Range Weather Forecasts at 204 locations in Norway for lead times from +12 to +108 h. An improvement in the continuous ranked probability score is seen for approximately 85% of the locations using the proposed method compared to standard BMA based on only wind speed forecasts. For moderate-to-strong wind the improvement is substantial, while for low wind speeds there is generally less or no improvement. On average, the improvement is 5%. The proposed methodology can be extended to include more NWP variables in the calibration and can also be applied to other variables.


## 1. Introduction

Ensemble forecasts have become an essential tool in making optimal decisions in weather-dependent settings. However, because of deficiencies in the models and the perturbation methods, raw ensemble forecasts are often not reliable in a probabilistic sense (e.g., Hamill and Colucci 1997). To better calibrate ensembles to local conditions, a variety of statistical methods have been proposed during the last decade. The overall aim is to provide probabilistic forecasts that are calibrated and sharp, that is, that can be trusted and are as certain as possible (Gneiting et al. 2007). Many methods assume a parametric probability distribution for the variable of

interest and let its parameters depend on the ensemble forecasts in various ways. Most often the mean of the distribution is characterized by each ensemble member, while the spread is related to the ensemble variability (e.g., Gneiting et al. 2005; Thorarinsdottir and Gneiting 2010). Mixtures of parametric distributions provide more flexibility and more closely resemble the weather-dependent variations in the ensembles. The basic idea is that each ensemble member is dressed with a probability distribution and that the sum of these constitutes the probabilistic forecast. The most widely used mixture approach is Bayesian model averaging (BMA; Raftery et al. 2005; Slougher et al. 2010). Other methods make no distributional assumptions and focus on quantiles. In these, quantiles are modeled separately as functions of ensemble members or ensemble statistics to better account for variations seen in the ensemble distributions and data (e.g., Bremnes 2004; Taillardat et al. 2016).

In contrast to statistical postprocessing of deterministic forecasts, most of the methods for ensembles only use the ensemble forecast of the variable under

---

 Denotes content that is immediately available upon publication as open access.

---

*Corresponding author:* John Bjørnar Bremnes j.b.bremnes@met.no

study as a predictor variable. There may be several reasons for this practice. First, the total number of predictors is increased by a factor equal to the number of ensemble members, and dealing with high-dimensional grouped data is challenging. Second, more variables often call for more complex nonlinear relationships, which are not straightforward to introduce. However, there are a few recent studies that demonstrate that including additional variables in the ensemble can improve the postprocessed forecast. [Scheuerer and Hamill \(2015\)](#) used precipitable water in precipitation forecasting with a shifted, censored gamma distribution and concluded it had a positive impact during the warm season. [Taillardat et al. \(2016\)](#) added ensemble statistics of several variables in quantile regression forests and demonstrated improvements of about 3% and 6% for wind speed forecasts at two sites. [Messner et al. \(2017\)](#) included variable selection directly in the parameter estimation using boosting and showed improvements of about 4%–12% depending on lead time for minimum and maximum temperature forecasting at five stations. Common for all of these methods is the use of summarizing statistics like the ensemble mean and standard deviation.

The objective of this article is to propose and test a flexible methodology that allows for including the complete ensemble forecast of several variables within a BMA framework. The basic idea is to allow for additional variables in the conditional mean of each member and to use flexible regression methods for describing possible nonlinearities. The approach is demonstrated for wind speed forecasting using ensemble forecasts of wind speed and wind direction for Norwegian synoptic stations. A large number of these stations are located in mountains, narrow fjords, valleys, and along the rugged coastline, which may strongly affect local wind conditions. These small-scale features are not well represented in current ensemble prediction systems, and statistical calibration methods can often improve local wind forecasts considerably. Further, the relation between model and locally observed wind speeds may vary with flow direction. Hence, it is reasonable to include wind direction as additional predictive information. One practical use of wind forecasts is for operations that have a threshold for maximum wind speed (e.g., construction work). Therefore, the probabilistic forecasts are evaluated for their probability of exceeding thresholds as well as their overall performance.

The remainder of this article is structured as follows. The data are presented in [section 2](#). [Section 3](#) describes the methods and models we propose as well as the evaluation scheme used. The results are presented in [section 4](#), and some final reflections and conclusions are given in [section 5](#).

## 2. Data

This article considers forecasts of wind speed and wind direction at 10-m height at 204 synoptic stations in Norway from the European Centre for Medium-Range Weather Forecasts (ECMWF) Ensemble Prediction System and measurements of hourly maximum 10-min-average wind speed. The measurement stations were selected from the Norwegian Meteorological Institute's network of automatic synoptic stations with the requirement of at least 70% data availability. Most, if not all, of the measurement data have been processed by a simple automatic quality control system, and some are also manually assessed. For the period under study, the ECMWF Ensemble Prediction System had a horizontal resolution of about 32 km and consisted of 51 members; one unperturbed (control) and 50 exchangeable pairwise symmetrically perturbed members. All forecasts were generated at 0000 UTC with lead times of +12, +36, +60, +84, and +108 h and were bilinearly interpolated to the locations of the stations. The observations and corresponding forecasts used are from the period from 1 January 2014 to 31 December 2015, yielding two years' worth of data.

[Figure 1](#) shows the locations, their mean observed wind speeds, and their multiplicative biases. The multiplicative bias is defined as the mean wind speed of the ensemble control forecasts divided by the mean observation. Mean observed wind speeds range from 1.7 to 11.0  $\text{ms}^{-1}$ . Norway has a varied topography, with mountain ranges with peaks and plateaus, and valleys and fjords, as well as lowlands. A considerable number of the stations are therefore located in complex terrain. The dark blue dots in [Fig. 1](#), representing high average wind speeds, mainly correspond to sites along the coast and in mountain areas. The sites with the largest discrepancy between the mean forecast and mean observed wind speeds are generally located in inland mountain regions.

## 3. Models and methods

### a. BMA model for wind speed forecasts

BMA provides probabilistic forecasts as predictive probability density functions (PDFs) of the weather quantity  $Y$  based on an ensemble forecast with  $M$  members ([Raftery et al. 2005](#); [Sloughter et al. 2010](#)). Let  $f_m$  be the forecast of ensemble member  $m$ . Assuming that each forecast  $f_m$  corresponds to a component PDF  $g_m(y|f_m; \theta_m)$ , where  $\theta_m$  are parameters to be estimated, the predictive PDF of the weather quantity  $Y$  can be expressed as a weighted sum of the component PDFs associated with each ensemble member

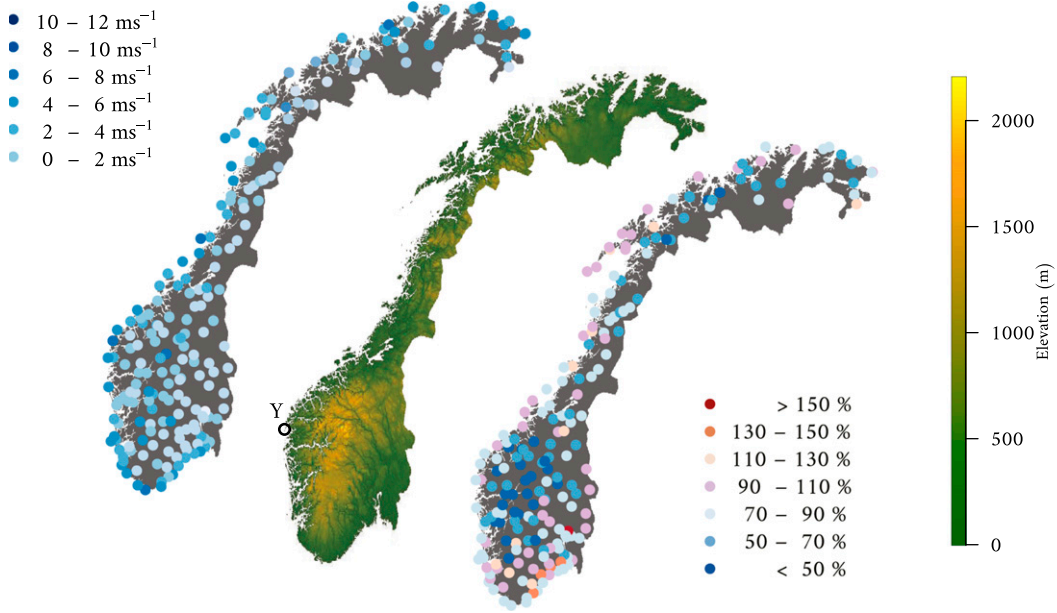


FIG. 1. (left) Mean observed wind speed at each location. (middle) Topography, with the location of Ytterøyane fyr (Y) marked. (right) Multiplicative bias of the wind speed forecast of the control member.

$$p(y|f_1, \dots, f_M; \theta_1, \dots, \theta_M) = \sum_{m=1}^M w_m g_m(y|f_m; \theta_m), \quad (1)$$

where  $w_m$  are weights based on the predictive performance of forecast  $f_m$  and  $\sum w_m = 1$ . Exchangeable ensemble members are given equal weights and a single set of corresponding parameters. Exchangeability is commonly assumed for ensemble members that use small, stochastic perturbations of the initial conditions in the model such as the ECMWF ensemble (Fraleay et al. 2010). However, the control member uses the best estimate of the initial conditions. Therefore, it is treated as a separate forecast, with its own weight and parameters. The remaining members are considered exchangeable and are given equal weights and parameters. We denote the weight and parameters of the control member  $w_c$  and  $\theta_c$ , and the weight and parameters of the perturbed members  $w_p$  and  $\theta_p$ . Following Eq. (1), our model can be specified as

$$p(y|f_0, \dots, f_{50}; \theta_c, \theta_p) = w_c g(y|f_0; \theta_c) + \sum_{m=1}^{50} w_p g(y|f_m; \theta_p), \quad (2)$$

where  $f_0$  is the control member and  $f_1, \dots, f_{50}$  are the perturbed ensemble members.

For component PDFs, we follow the suggestion of Sloughter et al. (2010) and use a gamma distribution, but other distributions could be used as well (e.g., Baran 2014). A gamma PDF can be specified by its mean  $\mu$  and

standard deviation  $\sigma$ . Sloughter et al. (2010) modeled these parameters as linear in the ensemble member forecast for each component PDF. In the linear model for the mean, coefficients are fitted separately for the control member and the perturbed members, while in the model for the standard deviation, the same coefficients are used for all ensemble members:

$$\mu_0 = b_{0c} + b_{1c} f_0, \quad (3)$$

$$\mu_m = b_{0p} + b_{1p} f_m, \quad m = 1, 2, \dots, 50, \quad \text{and} \quad (4)$$

$$\sigma_m = c_0 + c_1 f_m, \quad m = 0, 1, \dots, 50. \quad (5)$$

Assuming equal coefficients of the linear model for the standard deviation for all members simplifies the inference. This model will be referred to as BMA in the following.

*b. BMA utilizing wind speed and wind direction forecasts*

The ECMWF ensemble forecasts include both wind speed and wind direction. As a result of the coarseness of the forecasts compared to the local topography, it is reasonable that, for example, bias can change with wind direction. A way of utilizing the extra information in the wind direction forecasts is to also include the wind direction in the modeling of the mean of each component PDF. As wind direction is circular, a linear model is unreasonable. Hence, we want to introduce a nonlinear

model including both forecast wind speed and forecast wind direction. An established and flexible framework for nonlinear regression is thin-plate regression splines (see the [appendix](#) for a brief introduction to this concept). We denote the spline model  $s(f, d)$ , where  $s(\cdot)$  is the spline function, and  $f$  and  $d$  are the forecast wind speed and forecast wind direction, respectively. We follow the same approach as for the BMA model above and use gamma components that have different component mean models for the control and perturbed members, but the same linear model for the component standard deviation. Hence, we use Eq. (2), with gamma-distributed components with mean and standard deviation models for the ensemble members given by

$$\mu_0 = s_c(f_0, d_0), \quad (6)$$

$$\mu_m = s_p(f_m, d_m), \quad m = 1, 2, \dots, 50, \quad \text{and} \quad (7)$$

$$\sigma_m = c_0 + c_1 f_m, \quad m = 0, 1, \dots, 50, \quad (8)$$

where  $s_c(\cdot)$  and  $s_p(\cdot)$  are thin-plate regression spline functions for control members and perturbed members, respectively. We will use BMA-D to refer to this model.

### c. Model fitting and training scheme

We follow the approach for fitting BMA models introduced by [Raftery et al. \(2005\)](#) and, based on a training set, fit the model in two stages. First, the means of the component PDFs are fitted using linear regression for BMA, or regression splines for BMA-D (see the [appendix](#)). Next, the coefficients for the standard deviation and the model weights are fitted using a variant of the expectation–maximization (EM) algorithm.

There are several training schemes one can use. For BMA it is common to use a sliding window scheme where the postprocessing models are fitted using the previous  $k$  days (e.g., [Raftery et al. 2005](#); [Erickson et al. 2012](#)). To avoid the computational cost of fitting a new model every day, a recomputation frequency is introduced, so that a model that is fitted using the previous  $k$  days for training is used for the following  $r$  days. This paradigm means we must choose the length of the sliding window  $k$  and the recomputation frequency  $r$ . To carry out this approach, models are fitted for each site and lead time with training periods of lengths 30, 60, 90, 120, 150, 180, 210, 240, 270, 300, 330, and 365 days and recomputation frequencies of 30 and 365 days.

### d. Validation

The validation approach includes the evaluation of reliability and sharpness as well as summarizing measures for probabilistic forecasts. The overall forecast quality is quantified using the continuous ranked

probability score (CRPS), which measures the difference between the forecast CDF and the observation ([Murphy 1988](#)). [Gneiting and Raftery \(2007\)](#) showed that the CRPS can be formulated in terms of two expectations and the computations here are based on this approach. For the continuous distributions generated by the BMA models, the expectations in the CRPS formula are estimated by simulating 10 000 samples, while for the raw ensemble the formula is used directly (i.e., with the ensemble treated as a discrete distribution with equal probability for each member). Further, the Brier score (BS; [Brier 1950](#)) is applied to measure the quality of the forecast probabilities of the wind speed exceeding thresholds of 5, 10, 15, and 20 m s<sup>-1</sup>. Since this article aims to compare forecast approaches, CRPS and BS are reported as skill scores ([Murphy 1988](#)) and denoted by CRPSS and BSS, respectively. The scores can then be interpreted as improvements relative to the score of a reference forecast system, which here is chosen to be the standard BMA. The CRPSS and BSS results of standard BMA are thus zero by definition. A positive skill score therefore indicates improvements over standard BMA and a negative skill score the opposite.

Reliability quantifies the degree to which forecast probabilities and probability distributions can be trusted and is often assessed using rank or probability integral transform histograms (e.g., [Gneiting et al. 2007](#)), but when considering a large number of different sites, lead times and models, this approach becomes infeasible. Instead, we focus on the reliability of selected quantiles. For quantiles at level  $\alpha$ , the proportion of measurements less than or equal to the quantile should be  $\alpha$  by definition. The proportions are used as indicators of reliability and are calculated separately for each site, lead time, and model in order to examine reliability in detail; for example, a rank histogram with data pooled over sites is not sufficient to evaluate reliability at the site level. To ease the assessment of the degree of reliability, 95% reference intervals are added by computing the 2.5nd and 97.5th percentiles of the binomial distribution with the quantile level and the number of validation cases as parameters. Sharpness shows the degree of concentration of the probability mass and is defined, here, by the average widths of the central 50% and 90% forecast intervals derived from quantiles (e.g., [Gneiting et al. 2007](#)). The shorter the interval, the better the sharpness.

All validation measures are computed for the whole test period. To further analyze the resulting postprocessing models, the scores are also computed for high, medium, and low forecast wind speeds. This computation is done by splitting the test data into three groups of equal size according to the raw ensemble median. The grouping is done separately for each site and lead time, which is especially important for the reliability since all forecasts

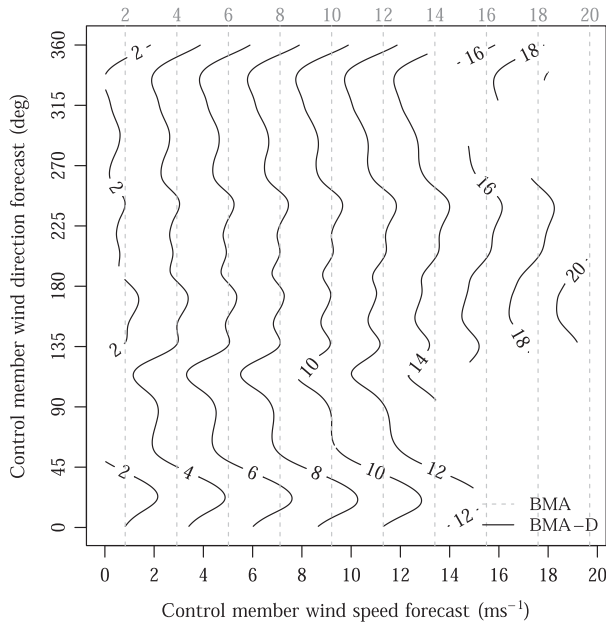


FIG. 2. Contour plot (solid lines) of the fitted thin-plate regression spline for the control member forecast at lead time +36 h from the site Ytterøyane fyr. The dashed lines show the fitted linear relation based on the control member forecast of wind speed only (BMA) at the same site and lead time.

should be reliable, and too much averaging may conceal systematic deviations from reliability.

### 4. Results

#### a. Choosing the training scheme

Comparing the CRPS for different training schemes for all lead times revealed that the length of the training period was more important than the recalculation frequency.

Increasing the number of training days led to a decrease in the CRPS, until about 200 days, at which time it stabilized. Refitting a new model every month had no noticeable advantage over fitting the model once a year. Therefore, the postprocessing models analyzed below are all only fitted once for each site and lead time, using 200 training days, and the results are evaluated for the last 365 days of the dataset.

#### b. BMA-D for Ytterøyane fyr

We first demonstrate BMA-D by studying the fitted model for Ytterøyane fyr, a lighthouse located on the west coast of Norway that often experiences strong wind. Figure 2 shows the thin-plate regression spline surface used for the control member in BMA-D for lead time +36 h for Ytterøyane fyr. Figure 2 illustrates how the resulting predicted wind speed in the BMA-D depends on both forecast wind speed and forecast wind direction. A forecast of 10 m s<sup>-1</sup> results in component means of 8 m s<sup>-1</sup> for a wind direction forecast of 15°, 10 m s<sup>-1</sup> for a wind direction of 50°, and 12 m s<sup>-1</sup> for a wind direction of 115°. For the BMA model the contour lines are linear (vertical dashed lines in Fig. 2), and an ensemble forecast of 10 m s<sup>-1</sup> would give a component mean of approximately 10.8 m s<sup>-1</sup>. Tracing the contour lines of the spline in Fig. 2, we observe that many of them have a marked peak at approximately 30°, suggesting that for a given forecast wind speed the winds from the north-northeast are generally weaker than for other forecast wind directions. In terms of CRPSS, using the forecast wind direction increased the score by about 12%. Similar and even stronger wind direction dependencies are seen for other sites.

Figure 3 shows time series of forecasts for lead time +36 h and corresponding observations for July 2015. The forecasts

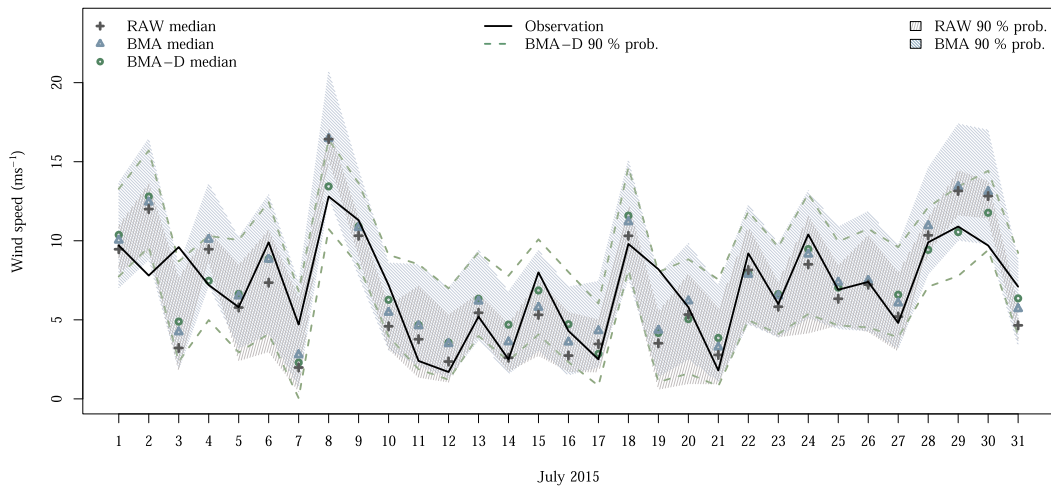


FIG. 3. Time series plots of the 5%, 50%, and 95% forecast quantiles for RAW, BMA, and BMA-D and corresponding observations at Ytterøyane fyr for July 2015 and at lead time +36 h.



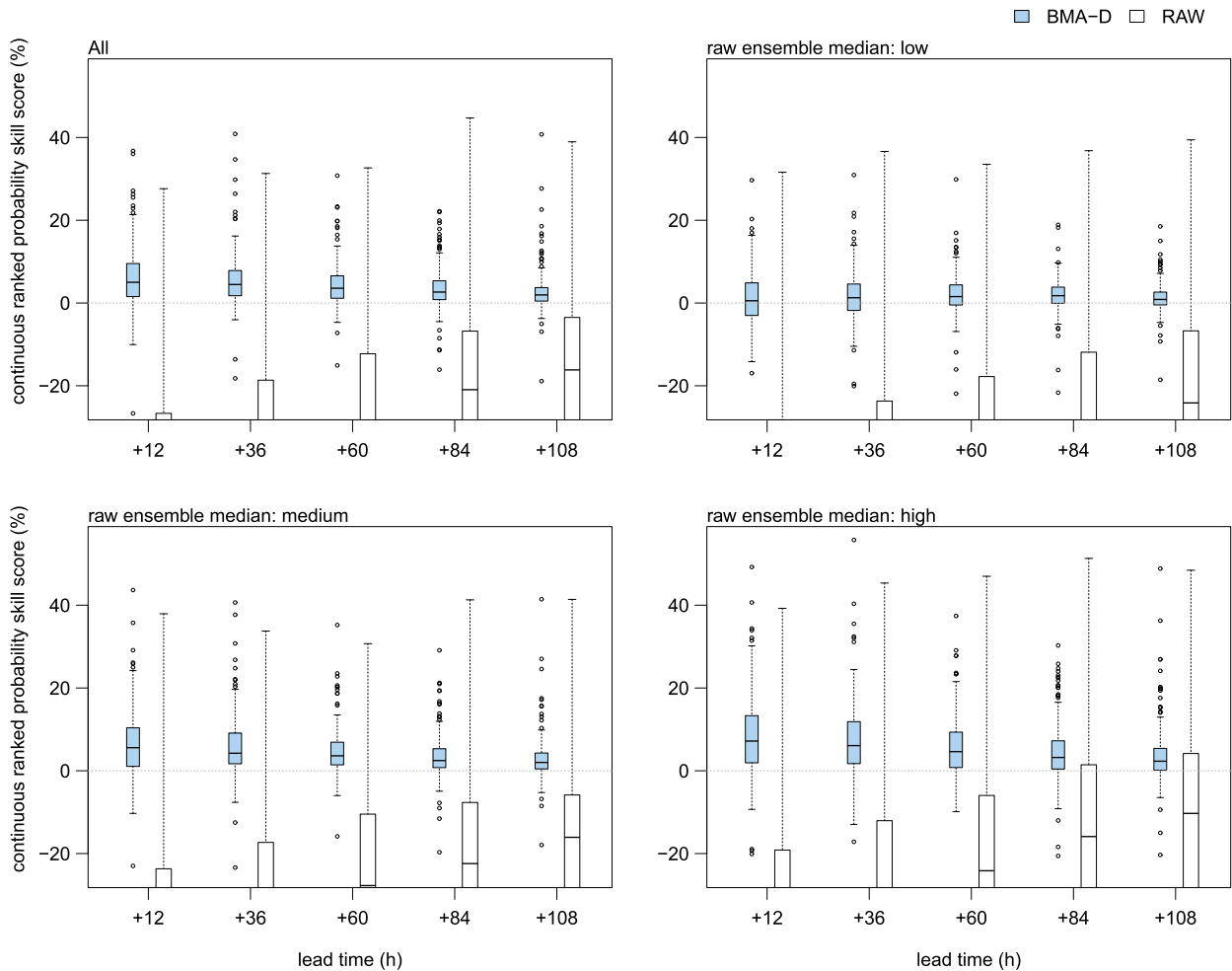


FIG. 4. Boxplots of the CRPSSs (%) of BMA-D and RAW relative to BMA over all locations for lead times of +12, +36, . . . , and +108 h. (top left) The results for all data, and the results for data where the raw ensemble median is categorized as (top right) low, (bottom left) medium, or (bottom right) high. The dashed gray lines at zero indicate the CRPSSs for BMA. The vertical axes are cut to highlight the relation between BMA-D and BMA.

of the raw ensemble (RAW), BMA, and BMA-D are displayed in terms of the 5%, 50%, and 95% quantiles. The most noticeable feature is that too many observations (about 30%) are outside the raw ensemble 90% forecast interval, indicating that the raw ensemble is too sharp. The BMA and BMA-D intervals are similar, but where they differ most (4, 8, and 29 July), the BMA-D intervals tend to be more centered around the measurements.

### c. Overall performance

Figure 4 shows boxplots of the CRPSSs for BMA-D and the raw ensemble relative to BMA for all sites, with the distributions being over the 204 sites. In the top right and bottom panels in Fig. 4, the boxplots show the same score for forecasts grouped according to the raw ensemble median. The top left panel in Fig. 4 clearly shows that BMA-D does better in terms of CRPS than BMA for the majority of the sites used, especially for shorter lead times. The

proportions of sites with positive CRPSS for BMA-D for lead times of +12, +36, +60, +84, and +108 h are 84.8%, 87.7%, 87.7%, 83.3%, and 81.9%, respectively. The average CRPSSs for BMA-D for the lead times are 6.1%, 5.5%, 4.4%, 3.5%, and 2.9% respectively. The raw ensemble performs poorly as a result of model weaknesses, such as bias, not being properly calibrated, and the NWP model wind not quite being representative of the local wind conditions. The remaining panels in Fig. 4 show that the improvement is especially pronounced when the raw ensemble median is medium or high. A possible explanation may be that in situations with weak wind, the wind is more often locally driven, and since the ECMWF model is not able to model these processes well, the large-scale model wind direction is of less importance in these cases.

To formally assess the significance of the CRPS results, the Diebold–Mariano (DM) hypothesis test (Diebold and Mariano 1995) was applied for two settings to compare

BMA-D and BMA. First, the test was performed separately for each lead time (over all sites and issue times). For all lead times the  $p$  values were extremely close to zero ( $<10^{-30}$ ), indicating significantly better scores for BMA-D compared to BMA. However, the DM test does not account for the complex spatial dependencies in the dataset (Hering and Genton 2011; Gilleland and Roux 2015), so the outcome should be interpreted with care. To avoid the spatial dimension, a second test was performed where CRPS was averaged over the sites for each issue time and lead time. The test was then carried out for each lead time and, again, the  $p$  values were extremely close to zero ( $<10^{-30}$ ). Thus, based on the hypothesis testing, the results are clearly in favor of BMA-D.

In Fig. 5 the CRPSs of BMA-D relative to BMA for each site is plotted on a map. As previously mentioned, BMA-D is slightly better at most sites. However, the greatest improvements occur in areas where the averaged observed wind speed is high (i.e., along the coast and in mountainous areas).

#### d. Exceedance performance

Figure 6 shows the Brier skill score of BMA-D and the raw ensemble relative to BMA for thresholds of 5, 10, 15, and 20  $\text{m s}^{-1}$ . The BSS for the BMA-D increases with the thresholds, and especially for the 15 and 20  $\text{m s}^{-1}$  thresholds, BMA-D is superior to BMA for the vast majority of the sites. This result supports the findings from Fig. 5—that it is particularly for higher wind speeds that the BMA performs poorly and that BMA-D solves this issue to a large extent.

The raw ensemble also appears to be equivalent to or even better than both BMA and BMA-D in terms of the Brier score for many sites at thresholds of 15 and 20  $\text{m s}^{-1}$ . However, this conclusion is mainly due to the fact that many sites very seldom experience strong winds. Consequently, the raw ensemble frequently achieves the perfect BS for high thresholds because high wind speeds are neither forecast by any ensemble member nor observed. At more than 50% of the sites, wind speeds greater than 15  $\text{m s}^{-1}$  are observed fewer than 10 times.

#### e. Reliability and sharpness

Figure 7 shows the reliability of the 5%, 50%, and 95% quantiles of the forecasts for lead times of +36 and +108 h. The raw ensemble's reliability is very poor, especially for the higher quantiles. It is generally slightly better for lead time +108 h than for +36 h, especially for the 5% and 95% quantiles, where the observed reliabilities are closer to the reference probabilities. The reliability for all quantiles is much better for BMA and BMA-D, although the reliability

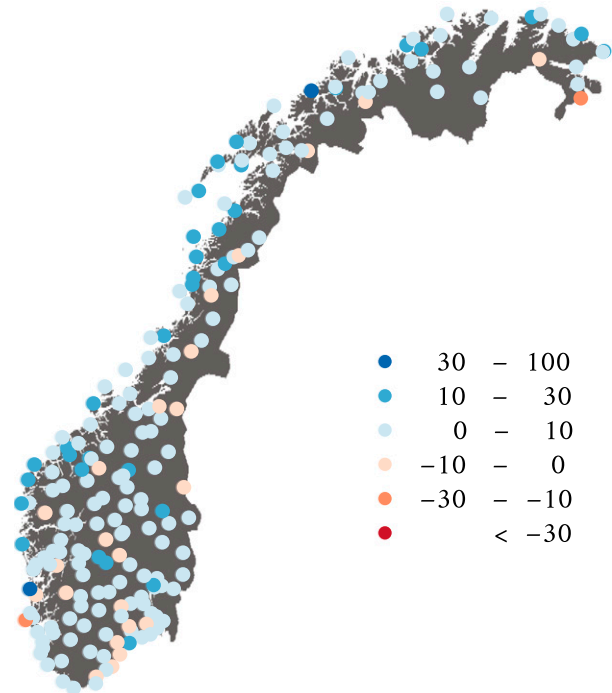


FIG. 5. CRPSs of BMA-D relative to BMA at each site for lead time +36 h.

of the 50% quantile is not very good. For the 95% quantile BMA is slightly better than BMA-D.

Figure 8 illustrates the sharpness of the BMA, BMA-D, and raw ensemble forecasts, with the two top panels showing the average lengths of the 50% and 90% prediction intervals. In general, the raw ensemble is considerably sharper than BMA and BMA-D, which is as expected. BMA-D also tends to give sharper predictions than BMA. For approximately 10 sites the average lengths of the 90% intervals are unrealistically high, indicating that for these sites the gamma distribution is not suitable. These stations are all among the windiest and for some of them, the forecast uncertainty does not differ much with lead time either. The bottom two panels in Fig. 8 show the average forecast interval lengths relative to BMA. From these findings, it can be concluded that for the majority of sites BMA-D produced sharper forecasts than BMA.

## 5. Discussion and conclusions

In this paper we have made use of the fact that the estimation of parameters in BMA models is split into two parts: regression for the mean parameters and expectation maximization for the weights and variance parameters. It is demonstrated that allowing the

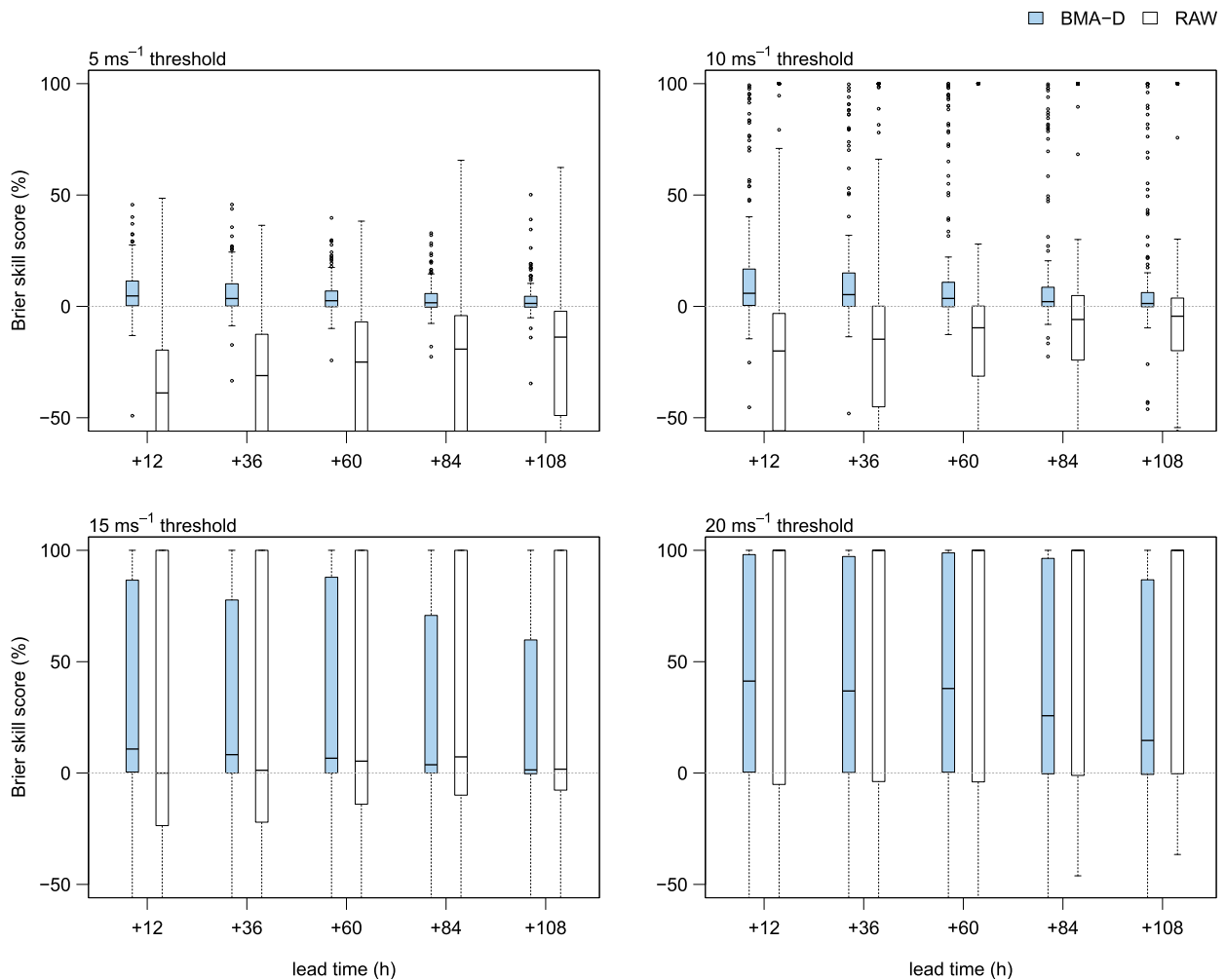


FIG. 6. Boxplots of the BSSs (%) of BMA-D and RAW relative to BMA for thresholds of 5, 10, 15, and 20 m s<sup>-1</sup>, over all locations for lead times of +12, +36, . . . , and +108 h. The dashed gray line at zero indicates the BSS for BMA. The vertical axes are cut to highlight the relation between BMA-D and BMA.

mean parameters for each ensemble member to be dependent on more than one predictor in a flexible manner leads to improvements over standard BMA modeling for wind speed forecasting at Norwegian stations. Nonlinear relationships were modeled by thin-plate splines, but any other regression method estimating conditional means could have been applied. In particular, most existing statistical postprocessing methods for deterministic forecasts can be used directly. Thus, valuable experience concerning relevant predictors and their functional relation to predictands in deterministic calibration is transferable to ensemble calibration. Further, flexible regression methods for the mean parameters can of course also be applied during the calibration of other variables, not only wind speed.

In our study, forecasts of wind speed and direction were used as predictors for all stations and lead times. On

average the improvements were about 5%, which is about the same as in Taillardat et al. (2016). However, the results were clearly site dependent; improvements up to about 40% were seen for some locations. At these sites, advanced users will no longer notice systematic weaknesses in the forecasts and thereby likely increase their confidence in the forecast system. Further advances could possibly be achieved by considering even more predictor variables. The scope for improvements, though, is likely more limited by the information in the data rather than the statistical methodology. Although the scores were generally better, degradation was noticed at about 15% of the stations. This suggests that wind direction was not a relevant predictor for these locations and that algorithms for choosing predictors would be useful. These algorithms could preferably be applied at the regression step or alternatively on the final mixture distribution. It should also



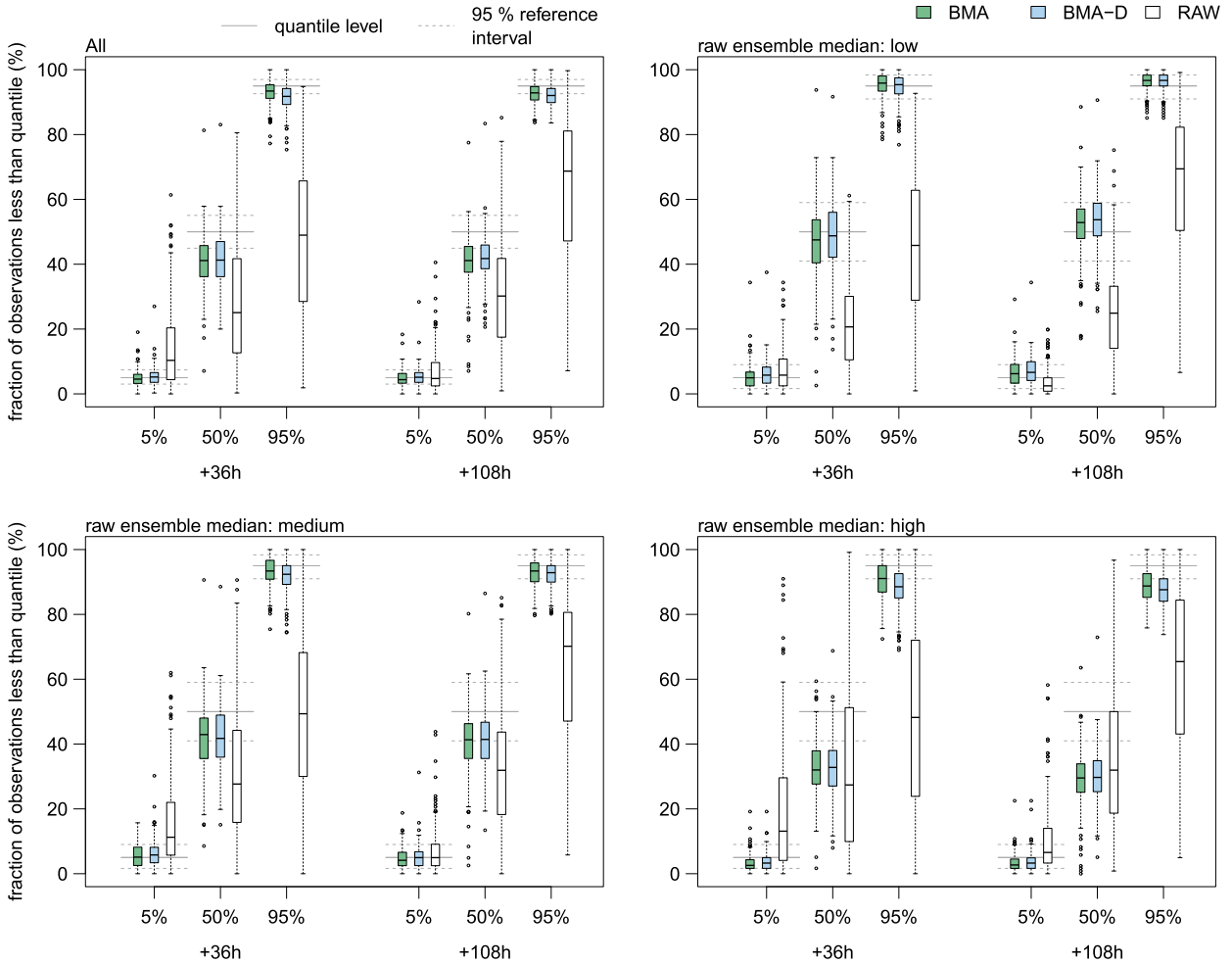


FIG. 7. Boxplots of the reliability of the 5%, 50%, and 95% forecast quantiles of BMA, BMA-D, and RAW over all sites for lead times of +36 and +108 h. (top left) The results for all data, and the results for data where the raw ensemble median is categorized as (top right) low, (bottom left) medium, or (bottom right) high. The solid gray lines indicate the quantile levels, while the dashed gray lines represent 95% reference intervals.

be added that for our specific problem and dataset, a bivariate wind vector approach could also be taken (Pinson 2012; Slughter et al. 2013; Schuhen et al. 2012). The main advantage of the wind vector approach is that the wind direction is adjusted or calibrated simultaneously.

A topic for future studies in ensemble postprocessing would be developing methods to assess how much is gained in forecast quality by using all ensemble members of several variables compared to only using summarizing statistics like the mean and standard deviations, as in Taillardat et al. (2016) and Messner et al. (2017). An advantage with our approach is that the relation between predictors is physically consistent since they are derived from the same scenario. By using summarizing quantities, the physical interpretation and consistency are lost.

*Acknowledgments.* The work of IS was funded by the Research Council of Norway, Project 250362.

## APPENDIX

### Thin-Plate Splines

Let  $f_m$  and  $d_m$  be forecasts of wind speed and wind direction from ensemble member  $m$ , and let  $y$  be the observed wind speed. Further, denote data of size  $n$  by  $\{(y_i, f_{mi}, d_{mi})\}, i = 1, 2, \dots, n$ . Thin-plate splines are then defined by the function  $s$  minimizing

$$\sum_{i=1}^n [y_i - s(f_{mi}, d_{mi})]^2 + \lambda \iint \left[ \left( \frac{\partial^2 s}{\partial f_m^2} \right)^2 + 2 \left( \frac{\partial^2 s}{\partial f_m \partial d_m} \right)^2 + \left( \frac{\partial^2 s}{\partial d_m^2} \right)^2 \right] df_m dd_m, \tag{A1}$$

where  $\lambda$  is a parameter that controls the degree of smoothness (Duchon 1977; Wahba 1990). Solving Eq. (A1)

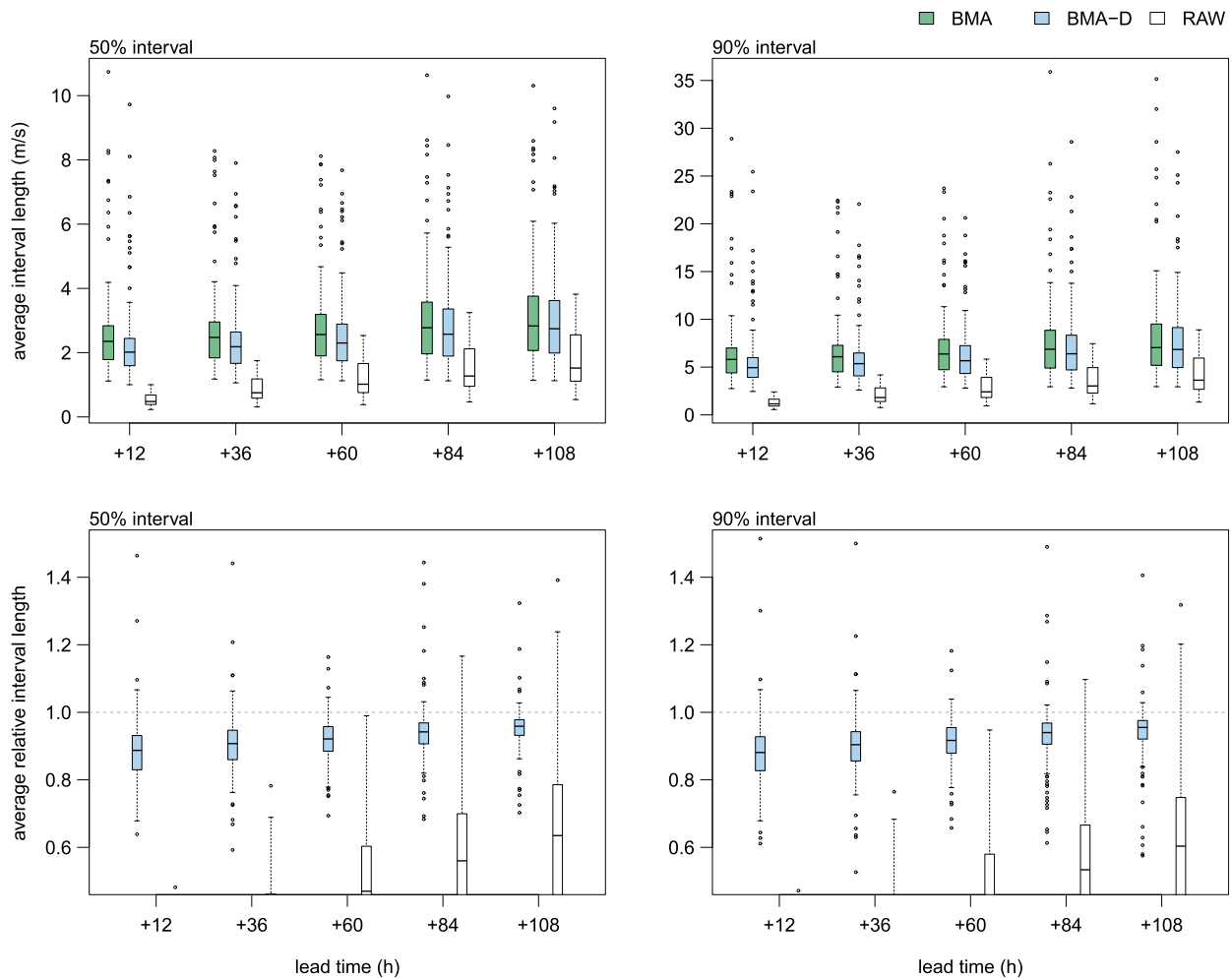


FIG. 8. Boxplots of sharpness for lead times of +12, +36, . . . , and +108 h. Results showing the average length of (top left) 50% and (top right) 90% forecast intervals for BMA, BMA-D, and RAW, and the average length of (bottom left) 50% and (bottom right) 90% forecast intervals of BMA-D and RAW relative to BMA (represented by the dashed gray lines at 1). The vertical axes are cut to highlight the relation between BMA-D and BMA.

is in general computationally challenging with respect to efficiency and numerical stability, and approximations are therefore common (Wood 2003). The smoothing parameter  $\lambda$  is chosen by generalized cross validation (Golub et al. 1979). In this work, the R software package *mgcv* (Wood 2006) is applied to fit the approximate thin-plate splines. Wind direction is a circular quantity and, ideally, fitted spline functions should be continuous at  $0^\circ/360^\circ$ . This constraint is not straightforward to include using the thin-plate spline implementation in the *mgcv* package. To make up for this shortcoming, some of the data were instead copied such that the data covered the interval  $[-90^\circ, 450^\circ]$  in the wind direction. An option would be to use wind vectors, but the results are less interpretable, and data would become sparse for both low and high values for each component.

## REFERENCES

- Baran, S., 2014: Probabilistic wind speed forecasting using Bayesian model averaging with truncated normal components. *Comput. Stat. Data Anal.*, **75**, 227–238, <https://doi.org/10.1016/j.csda.2014.02.013>.
- Bremnes, J. B., 2004: Probabilistic forecasts of precipitation in terms of quantiles using NWP model output. *Mon. Wea. Rev.*, **132**, 338–347, [https://doi.org/10.1175/1520-0493\(2004\)132<0338:PFOPIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(2004)132<0338:PFOPIT>2.0.CO;2).
- Brier, G. W., 1950: Verification of forecasts expressed in terms of probabilities. *Mon. Wea. Rev.*, **78**, 1–3, [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2).
- Diebold, F. X., and R. S. Mariano, 1995: Comparing predictive accuracy. *J. Bus. Econ. Stat.*, **13**, 253–263.
- Duchon, J., 1977: Splines minimizing rotation invariant semi-norms in Sobolev spaces. *Constructive Theory of Functions of Several Variables*, P. D. W. Schempp, and P. D. K. Zeller, Eds., Lecture Notes in Mathematics, Vol. 571, Springer, 85–100.

- Erickson, M. J., B. A. Colle, and J. J. Charney, 2012: Impact of bias-correction type and conditional training on Bayesian model averaging over the northeast United States. *Wea. Forecasting*, **27**, 1449–1469, <https://doi.org/10.1175/WAF-D-11-00149.1>.
- Fraley, C., A. E. Raftery, and T. Gneiting, 2010: Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Mon. Wea. Rev.*, **138**, 190–202, <https://doi.org/10.1175/2009MWR3046.1>.
- Gilleland, E., and G. Roux, 2015: A new approach to testing forecast predictive accuracy. *Meteor. Appl.*, **22**, 534–543, <https://doi.org/10.1002/met.1485>.
- Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.*, **102**, 359–378, <https://doi.org/10.1198/016214506000001437>.
- , —, A. H. Westveld III, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118, <https://doi.org/10.1175/MWR2904.1>.
- , F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc.*, **69B**, 243–268, <https://doi.org/10.1111/j.1467-9868.2007.00587.x>.
- Golub, G. H., M. Heath, and G. Wahba, 1979: Generalized cross-validation as a method for choosing a good ridge estimator. *Technometrics*, **21**, 215–223, <https://doi.org/10.1080/00401706.1979.10489751>.
- Hamill, T. M., and S. J. Colucci, 1997: Verification of Eta–RSM short-range ensemble forecasts. *Mon. Wea. Rev.*, **125**, 1312–1327, [https://doi.org/10.1175/1520-0493\(1997\)125<1312:VOERSR>2.0.CO;2](https://doi.org/10.1175/1520-0493(1997)125<1312:VOERSR>2.0.CO;2).
- Hering, A. S., and M. G. Genton, 2011: Comparing spatial predictions. *Technometrics*, **53**, 414–425, <https://doi.org/10.1198/TECH.2011.10136>.
- Messner, J. W., G. J. Mayr, and A. Zeileis, 2017: Nonhomogeneous boosting for predictor selection in ensemble post-processing. *Mon. Wea. Rev.*, **145**, 137–147, <https://doi.org/10.1175/MWR-D-16-0088.1>.
- Murphy, A. H., 1988: Skill scores based on the mean square error and their relationships to the correlation coefficient. *Mon. Wea. Rev.*, **116**, 2417–2424, [https://doi.org/10.1175/1520-0493\(1988\)116<2417:SSBOTM>2.0.CO;2](https://doi.org/10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2).
- Pinson, P., 2012: Adaptive calibration of  $(u, v)$ -wind ensemble forecasts. *Quart. J. Roy. Meteor. Soc.*, **138**, 1273–1284, <https://doi.org/10.1002/qj.1873>.
- Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, <https://doi.org/10.1175/MWR2906.1>.
- Scheuerer, M., and T. M. Hamill, 2015: Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted Gamma distributions. *Mon. Wea. Rev.*, **143**, 4578–4596, <https://doi.org/10.1175/MWR-D-15-0061.1>.
- Schuhen, N., T. L. Thorarinsdottir, and T. Gneiting, 2012: Ensemble model output statistics for wind vectors. *Mon. Wea. Rev.*, **140**, 3204–3219, <https://doi.org/10.1175/MWR-D-12-00028.1>.
- Sloughter, J. M., T. Gneiting, and A. E. Raftery, 2010: Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *J. Amer. Stat. Assoc.*, **105**, 25–35, <https://doi.org/10.1198/jasa.2009.ap08615>.
- , —, and —, 2013: Probabilistic wind vector forecasting using ensembles and Bayesian model averaging. *Mon. Wea. Rev.*, **141**, 2107–2119, <https://doi.org/10.1175/MWR-D-12-00002.1>.
- Taillardat, M., O. Mestre, M. Zamo, and P. Naveau, 2016: Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Mon. Wea. Rev.*, **144**, 2375–2393, <https://doi.org/10.1175/MWR-D-15-0260.1>.
- Thorarinsdottir, T. L., and T. Gneiting, 2010: Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroscedastic censored regression. *J. Roy. Stat. Soc.*, **173A**, 371–388, <https://doi.org/10.1111/j.1467-985X.2009.00616.x>.
- Wahba, G., 1990: *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59, Society for Industrial and Applied Mathematics, 161 pp., <https://doi.org/10.1137/1.9781611970128>.
- Wood, S. N., 2003: Thin plate regression splines. *J. Roy. Stat. Soc.*, **65B**, 95–114, <https://doi.org/10.1111/1467-9868.00374>.
- , 2006: *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC Press, 410 pp.