Calibrated probabilistic hub-height wind forecasts in complex terrain

David Siuta*, Gregory West, and Roland Stull

The University of British Columbia, Vancouver, BC Canada

Thomas Nipen

Norwegian Meteorological Institute, Oslo, Norway

Corresponding author address: David Siuta, The University of British Columbia, 2020 - 2207 Main Mall, Vancouver, BC Canada V6T 1Z4.

E-mail: dsiuta@eos.ubc.ca

ABSTRACT

This work evaluates the use of a WRF ensemble for short-term, probabilistic, hub-height wind-speed forecasts in complex terrain. We test for probabilistic-forecast improvements by increasing the number of planetary-boundary-layer schemes used in the ensemble. Additionally, we evaluate several prescribed uncertainty models used to derive forecast probabilities based on knowledge of the error within a past training period. A Gaussian uncertainty model provided calibrated wind-speed forecasts at all wind farms tested. Attempts to scale the Gaussian distribution based on the ensemble mean or variance values did not result in further improvement of the probabilistic-forecast performance. When using the Gaussian uncertainty model, a small-sized six-member ensemble showed equal skill to that of the full 48-member ensemble. We introduce a new uncertainty model called the pq distribution that better fits the ensemble wind-forecast-error distribution. Results indicate that gross attributes (central tendency, spread, and symmetry) of the prescribed uncertainty model are more important than its exact shape.

1. Introduction

The renewable-energy sector has grown rapidly as global economies aim to derive a larger portion of their electric generation from renewable sources such as wind and solar. One of the primary challenges with using renewable-energy sources is integrating these sources into existing electric grids. Power from these sources is generated only when the wind blows or sun shines, and cannot be controlled to the same extent as traditional sources (e.g., coal, natural gas, nuclear, hydro).

Wind forecasts are used to ensure that electric supply and demand always remain in sync through the scheduling of energy-reserve resources and market trading (Monteiro et al. 2009). Because winds are highly variable and forecasts imperfect, planners must schedule *spinning* and *stand-by* energy reserves to add flexibility to the electric grid. These reserves are traditional energy sources (e.g., natural gas and coal) that are scheduled ahead of time to dispatch additional power to the grid with as little as five minutes notice (Monteiro et al. 2009). However, reserve scheduling is complicated by the spin-up time required before power can be dispatched to the grid from these sources. Energy-reserve spin-up times can range from several hours to over a day (Ahlstrom et al. 2013). Thus, accurate forecasts of wind-energy generation are required at short-term horizons (0-48 hours) to determine how many reserves are needed, to schedule these reserves ahead of time, and to adjust already committed reserves in the event of forecast changes (Monteiro et al. 2009).

For market trading, energy traders bid excess resources to the market, with some deadlines as little as 30 minutes prior to scheduled delivery (Monteiro et al. 2009; Draxl et al. 2014). Penalties can exist in the event of an imbalance between the bid power and that actually delivered (Pinson et al. 2006).

Planners and market traders use wind forecasts based on numerical weather prediction (NWP) models to gauge future wind-energy production. Such NWP models are the largest source of un-

certainty in wind-energy generation estimates (Monteiro et al. 2009). Forecast uncertainty arises from incorrect initial-condition (IC) sources, incomplete knowledge of atmospheric physics (e.g., turbulence), and improper assimilation of observational data (Buizza et al. 2005). Better estimation of forecast uncertainty, and higher forecast accuracy lead to more efficient energy planning, optimized market trading opportunities, and significant monetary savings (Pinson et al. 2006; Marquis et al. 2011; Mahoney et al. 2012; Wilczak et al. 2015).

Probabilistic weather forecasts are the primary tools used to gain insight on forecast uncertainty. Earlier forecasts used for wind-energy planning were deterministic, meaning they did not provide information on forecast uncertainty (Pinson et al. 2006). However, in the last 10 years, probabilistic forecasts have begun to be adopted throughout the industry as planners started to recognize the value of quantifying forecast uncertainty. Monteiro et al. (2009), Giebel et al. (2011), and Zhang et al. (2014) provide extensive reviews of the history of probabilistic forecasts for wind energy.

Zhang et al. (2014) detail the two standard approaches for creating probabilistic forecasts for wind energy: parametric and non-parametric methods. Parametric methods prescribe a probability distribution of a particular shape that is typically representative of the past forecast-error distribution. Such distributions are described by location and scale parameters and have the benefit of being computationally cheap (Zhang et al. 2014). Early methods dressed distributions around single deterministic forecast models, with several distributions being tested. Lange (2005) found that the forecast-error distribution calculated from the German Weather Service (DWD) numerical forecasts are well-dressed by a Gaussian distribution. They provided the forecast-error-distribution results for 10-m wind speeds, not hub-height winds, and likely only at a location in flat terrain (the example location is confidential and a large quantity of the stations tested were in flat terrain). When converted to wind power, forecast-error distributions have been fit by Gaussian (Lange 2005), Beta (Bludszuweit et al. 2008), and generalized logit-normal distributions (Pinson 2012).

Non-parametric approaches do not make a shape assumption. One example is an ensemble of NWP models. Ensemble forecast systems provide a distribution of possible forecast outcomes (Fig. 1a), and can be produced by NWP using a variety of methods: 1) using multiple IC sources, 2) stochastically perturbing a single IC, 3) perturbing observations used during data assimilation for observation error, 4) varying model physical parameterization configurations, 5) varying NWP dynamical cores, 6) varying model grid lengths, or 7) using stochastic kinetic energy backscatter (SKEBS) (Stensrud et al. 2000; Grimit and Mass 2002; Buizza et al. 2005; Eckel and Mass 2005; McCollor and Stull 2008a; Candille 2009; Berner et al. 2009).

An ensemble forecast distribution can be converted into a probability distribution (Fig. 1b). This is commonly done by using the empirical distribution of ensemble members (Anderson 1996). Bayesian Model Averaging has also been successfully applied to wind-forecast systems (Hoeting et al. 1999; Raftery et al. 2005; Sloughter et al. 2010; Courtney et al. 2013). Kernel density estimation is another non-parametric method that links one or more explanatory variables to forecast probability density functions (Juban et al. 2007). While such methods have been shown to work well, the latter two require large datasets and can become computationally expensive (Juban et al. 2007; Zhang et al. 2014).

A hybrid approach is to parameterize the moments of a prescribed probability distribution (such as Gaussian) as a function of statistical properties of an ensemble (e.g., the ensemble mean, ensemble variance, and ensemble forecast-error distribution; Gneiting et al. (2005); Nipen and Stull (2011)). This method requires running computationally expensive ensembles, but does not require an extensive training dataset. Gneiting et al. (2005) refer to this method as ensemble model output statistics (EMOS).

Probabilistic forecasts are most useful when the forecast probability of an event matches its observed frequency of occurrence. Distributions from raw ensembles are often under-dispersive

(Stensrud et al. 2000; Eckel and Mass 2005; McCollor and Stull 2008c), and must undergo bias correction and probabilistic calibration (Buizza et al. 2005; Nipen and Stull 2011). When the forecast probabilities match the relative frequency of occurrence, a probabilistic forecast is *reliable* or *calibrated*. Calibrated probabilistic forecasts can be trusted as accurate estimates of forecast uncertainty. However, as Nipen and Stull (2011) discuss, even forecasts based on climatology can be probabilistically calibrated, so assessing calibration is not enough. One method that can be used to differentiate probabilistically-calibrated forecasts is through evaluating forecast sharpness (Gneiting et al. 2005; Pinson et al. 2006; Juban et al. 2007). Forecast sharpness is a measure of the width of probabilistic spread, with a deterministic forecast being perfectly sharp (Juban et al. 2007). Pinson et al. (2006) suggest forecast calibration is the primary concern of probabilistic forecasts for wind energy, while any improvements in sharpness represent added value.

The use of ensembles in probabilistic forecasts for hub-height wind speeds is growing, but studies evaluating probabilistic skill are still relatively limited (Junk et al. 2015). Deppe et al. (2012) performed an ensemble study over the flat terrain of Iowa using the Weather Research and Forecasting (WRF, Skamarock et al. 2008) model version 3.1.1., but did not assess probabilistic-forecast calibration. Deppe et al. (2012) used an ensemble of different planetary-boundary-layer (PBL) schemes and IC sources. They found a multi-PBL ensemble had a more accurate ensemble-mean forecast when compared to an ensemble formed from perturbed initial conditions, although the latter had larger ensemble variance. Therefore, a multi-PBL ensemble might be a suitable choice when using the EMOS method of generating probabilistic forecasts.

Since none of the existing PBL schemes available in WRF were designed for complex terrain, and because many wind features in complex terrain are dependent on PBL evolution, we posit that an ensemble containing several PBL schemes may be beneficial. The main differences between PBL schemes are the treatment of vertical mixing and the statistical order of turbulence closure

(Stull 1988; Stensrud 2007; Deppe et al. 2012). Differences in PBL schemes could be expected to be larger in complex terrain than was found by Deppe et al. (2012) over flat terrain.

For complex terrain, better methods to produce calibrated, sharp probabilistic hub-height wind-speed forecasts are needed. This study evaluates the performance of multi-PBL, multi-IC, multi-grid-length WRF ensembles at four wind farms in the mountainous terrain of British Columbia using the empirical ensemble distribution and EMOS methods. The eight PBL schemes used here are the Yonsei University Scheme (YSU, Hong et al. 2006; Hu et al. 2013), Asymmetric Convective Model version 2 (ACM2, Pleim 2007), Medium Range Forecast (MRF, Hong and Pan 1996), Mellor-Yamada-Janjic (MYJ, Janjić 1994), Mellor-Yamada-Ninno-Nakanishi and Niino Level 2.5 (MYNN, Nakanishi and Niino 2006), Quasi-Normal Scale Elimination (QNSE, Sukoriansky et al. 2005), University of Washington (UW, Bretherton and Park 2009), and Grenier-Bretherton-McCaa (GBM, Grenier and Bretherton 2001) schemes.

Further, this study provides insight on probabilistic forecasting techniques for hub-height wind-speed forecasts using the WRF model. It also adds knowledge about the shape of the hub-height wind-speed forecast-error distribution in complex terrain. This work is a follow-up to Siuta et al. (2016), who evaluated the deterministic-forecast performance of this dataset and the deterministic-forecast sensitivity to the choice of PBL-scheme, grid length, and IC source. The rest of this paper is organized as follows: Section 2 describes the research methodology used. Section 3 provides a discussion of the results. Finally, Section 4 discusses general conclusions and suggestions for future work.

2. Methodology

A year-long (June 2014 - May 2015) study using the WRF model (version 3.5.1) was done to evaluate the effectiveness of a multi-PBL, multi-IC, multi-grid-length ensemble forecast in com-

plex terrain. Each day, four, two-way nested meshes with 36-, 12-, and 4-km grid lengths (Fig. 2) produced 48 individual forecasts with hourly output covering a forecast horizon of one day at four wind farms in British Columbia. The 48 forecasts were created by running the domain setup in Fig. 2 with initial and boundary conditions from the U.S. National Centers for Environmental Prediction 0000 UTC 0.5 degree Global Forecast System (GFS) and 0000 UTC 32-km North American Mesoscale Model (NAM, grid 221), and with eight PBL schemes in the WRF-ARW dynamical core (Fig. 3). For the NAM, the 32-km grid is the only output available that extends far enough north for our WRF domains. Table 1 provides a summary of the WRF-model settings used.

Forecasts for each domain were created by using the nearest-neighbor grid cell in the horizontal, and the closest model level to wind-turbine hub height in the vertical. Vertical resolution in the boundary layer was increased with six levels in approximately the lowest 100 m. The authors recognize that horizontal and vertical interpolation methods may improve deterministic hub-height (roughly 50- to 150-m AGL) wind forecasts, but we do not apply them here to better isolate uncertainties in the boundary-layer wind forecast due to PBL scheme choice.

Forecasts from each ensemble member, as well as hub-height wind-speed observations were used in the Component-Based Post-Processing System (COMPS, Nipen 2012) to post-process and evaluate the forecast. Post-processing is anything applied to the raw forecast for improvement, including bias removal, choice of uncertainty model, and any calibration done to adjust the uncertainty model. COMPS is a modular post-processing and verification system, analogous to how WRF is modular for NWP. COMPS uses a series of namelists containing user-defined post-processing options that are applied to the raw forecast input.

The four wind farms used in this study are located on mountain ridge tops, with elevations between 500 and 1000 m MSL. We use wind-farm-averaged nacelle (hub-height) wind-speed ob-

servations for post-processing and model evaluation. These observations from the independent power producers are confidential and were provided to us by the local utility company, BC Hydro. BC Hydro, who purchases the wind power, performed quality control on the data. Wind-farm averaged observations have been shown to better match power generation and output from NWP models over that of individual turbines or meteorological towers. Farm-averaged values remove sub-grid-scale (intra-wind-farm) variability not currently resolved by NWP models (Cutler et al. 2012). Wind-farm names and locations remain anonymous in the figures shown here. However, statistics and aggregated results based on these observations are provided.

We evaluate several factors affecting the probabilistic hub-height wind-speed forecast in complex terrain, including the effect of using multiple PBL schemes, applying bias correction, and the choice of uncertainty model. These tests are summarized by Table 2 and are described in detail in the following subsections.

a. Effect of number of PBL schemes in ensemble

This study is a follow-up to that by Siuta et al. (2016), who used the same data set to examine deterministic hub-height wind-speed forecast sensitivity to the choice of PBL scheme, grid length, and IC source. The outcome of that study showed that the grid length and PBL scheme had the most influence on deterministic-forecast accuracy, but the most influential factor varied by location, season, and time of day. Appendix A provides the Mean Absolute Error (MAE) scores for the bias-corrected forecasts used in this prior study. Siuta et al. (2016) found the ACM2 PBL scheme to be the best-performing of all the PBL schemes, and the 12 km to be the best performing grid, when averaged over all four wind farms over the entire year. However, the best PBL scheme (and grid length) differed for individual wind-farms and seasons (see appendix A).

In this study we start with only the ACM2 PBL scheme (six total ensemble members), and then test for short-term probabilistic-forecast improvements by adding PBL schemes one-by-one into the ensemble. We start with the ACM2 PBL scheme because of the prior results mentioned above.

b. Bias correction technique

We bias correct each individual ensemble member prior to forming the ensemble mean. Wind speeds can never be negative. To satisfy this condition, we use a degree-of-mass-balance (DMB) multiplicative bias-correction technique (Grubišić et al. 2005; McCollor and Stull 2008c; Bourdin et al. 2014) applied to each ensemble member. The current bias correction factor, DMB_{t}, is calculated from the ratio of past forecast and observation pairs, weighted by an e-folding time τ:

$$DMB_t = \frac{\tau - 1}{\tau} DMB_{t-1} + \frac{1}{\tau} \frac{F_{t-1}}{O_{t-1}}. \tag{1}$$

Here, $DMB_{\mathrm{t\text{-}1}}$ is the bias correction factor from the previous day weighted by $(\tau - 1)/\tau$. We used a τ of 30 days because it was found to be the optimum training period when averaged over all locations (Appendix A). The ratio of the mean of the previous day's forecasts ($F_{\mathrm{t\text{-}1}}$) and observations ($O_{\mathrm{t\text{-}1}}$) is weighted by $1/\tau$. The result is that the influence of the most recent forecast-observation pairs decreases with an e-folding time of τ. This recursive bias correction minimizes the need to store an ever-growing data set for training.

The factor DMB_{t} is calculated for each individual ensemble member and then applied to the raw ensemble-member forecast F_t to produce a bias-corrected ensemble-member forecast $\hat{F}_t$:

$$\hat{F}_t = \frac{F_t}{DMB_t}. \tag{2}$$

Eqs. (1) and (2) utilize data such as would be available in a true operational setting—current and future observations are not available when the forecast is made.

c. Choice of uncertainty model

The method used to convert the distribution of raw or bias-corrected ensemble members to a probability distribution is called the uncertainty model. The uncertainty model must not assign any probabilities to wind speeds below zero.

We evaluate five uncertainty models, four of which are shown in Table 2. The first method, referred to as Raw (Table 2), uses the empirical distribution of ensemble members to assign probability. This is a basic method that could be used by those producing probabilistic forecasts over areas where observations do not exist.

The second uncertainty model, referred to as GNS (Table 2) assumes the form of a Gaussian distribution dressed about the ensemble mean. This Gaussian forecast probability distribution ($\mathcal{N}_t$) is described at any time t by

$$\mathcal{N}_t(\bar{F}_t, \sigma_t^2). \tag{3}$$

Here, $\overline{F_t}$ is the bias-corrected ensemble mean calculated from (1) and (2), where each member is equally weighted. The variance, σ_t^2, is adaptively estimated from the square of the past forecast errors calculated from the ensemble mean, weighted by the same 30-day e-folding time used in (1).

The third uncertainty model, GSEV (Table 2), scales the variance of the Gaussian distribution based on the ensemble variance. For this method, variance is calculated using a linear regression of the form:

$$\sigma_t^2 = mx + b \tag{4}$$

In (4), x is the raw or bias-corrected ensemble variance, and constants m and b are determined by regression against the square of the past errors (calculated from the bias-corrected ensemble mean). When used in this manner, the GSEV method could capitalize on theorized spread-skill relationships (Wilks 2011; Grimit and Mass 2002). Namely, when ensemble variance is smaller (larger), the spread of the Gaussian distribution (4) could be smaller (larger), assuming a relationship exists. Generally, linear correlation is used to quantify spread-skill relationships, with values over 0.6 considered strong relationships (Grimit and Mass 2007).

The fourth uncertainty model, GSEM (Table 2), scales the Gaussian distribution by the ensemble mean. For this method, x in (4) represents ensemble mean. This method could allow the uncertainty model to have larger (smaller) spread when the ensemble-mean-forecast wind speed is high (low). For the GSEM and GSEV uncertainty models, the regressed variables in (4) (e.g., the ensemble mean or the ensemble variance, and the past squared errors) are also adaptively updated with an e-folding time of 30 days.

Fig. 4 illustrates how these three Gaussian-based methods could differ for an idealized three-member ensemble. We posit that the GSEV and GSEM methods could allow for sharper probabilistic forecasts than the GNS method. For each of these distributions, we also test for probabilistic forecast improvements resulting from adding more PBL schemes to the ensemble, and through ensemble-member bias correction. An overview of the tests performed is given in Table 2.

A fifth uncertainty model, based on a new distribution (which we call the pq distribution), is also tested. It allows for optimization of kurtosis to better match the distribution of past forecast errors (shown later). Probabilistic-forecast results using this model did not improve on those from the

Gaussian-based models. The authors felt it important to show the results of this experiment, but details will be relegated to Appendix B.

d. Verification metrics

As our focus in this paper is probabilistic forecasting, we concentrate our forecast evaluation on distributions-oriented verification metrics, which use the joint distribution of forecasts and observations (Wilks 2011).

These types of metrics provide advantages over deterministic measures when assessing probabilistic forecasts. Because they evaluate the full probabilistic-forecast distribution, they provide insight into how well forecast uncertainty is represented by the forecast probability distribution. Accurate representation of forecast uncertainty allows for optimum decision making and cost minimization (Pinson et al. 2006; Juban et al. 2007; McCollor and Stull 2008b).

As with deterministic verification, the choice of verification metrics should be decided based on the purpose of the forecast and the needs of the end user. Gneiting et al. (2005, 2007) and Juban et al. (2007) propose that ideal probabilistic forecasts achieve calibration while maximizing forecast sharpness (both are described in more detail later). For this we use the Probability Integral Transform (PIT) histogram, reliability diagram, and continuously ranked probability score (CRPS). Skill scores are also used to quantify improvements in verification metrics of a test forecast configuration over that of a reference forecast (Wilks 2011).

Calibration is a measure of how well forecast probability represents the actual frequency of event occurrence (Wilks 2011; Nipen and Stull 2011; Gneiting et al. 2007, 2005). The PIT histogram provides a concise method of evaluating ensemble probabilistic calibration. PIT values are calculated by finding which forecast percentile matches the associated observation for individual forecast-observation pairs. A histogram is generated by counting the frequency of occurrence of

observations in each forecast percentile bin. This histogram should be flat for a calibrated forecast, and the deviation from flatness can be used to gauge calibration as was shown in Nipen and Stull (2011).

When the PIT histogram is not flat, it can indicate two things: 1) the forecast is under- or over-dispersive (needs calibration), or 2) the forecast has bias. Both of these situations can be fixed by calibration methods that adjust the probability distribution, resulting in flatter PIT histograms.

Gneiting et al. (2007) highlight that while the use of PIT histograms is prudent to address forecast calibration, it is hardly the only important aspect of a probabilistic forecast. They mention that forecasts that maximize sharpness, while maintaining calibration, are best. In addition, forecast calibration may not hold true over a subset of events. For wind-energy applications, sharper probabilistic forecasts (when also probabilistically calibrated) result in more efficient energy reserve resource planning and market trading opportunities because uncertainty (typical error from the ensemble mean) in the forecast is reduced (Juban et al. 2007). Because even forecasts based on climatology can be probabilistically calibrated, forecasts must be sharper than climatology to be considered skillful.

We use the observations from a previous 15-day window to define climatology at each hour of the day. Our choice of a 15-day window is two-fold: (1) we have limited observations available with only a year-long dataset, and (2) a shorter definition of climate (over that which averages all days of the year) could better represent seasonality. We do not use any observations in the future to define climate as this may cause the climate-forecast skill to falsely appear better.

Reliability diagrams allow us to assess forecast calibration for a subset of events given a threshold (Wilks 2011). For this study, we use wind-speed thresholds of 5, 15, and 20 m s^{-1} to represent low winds, turbine-rated winds, and high winds, respectively. Turbine-rated speeds are those at

which wind turbines start to generate maximum power output (i.e., increasing winds will no longer increase power output). This speed can differ between turbine models.

We insert a sharpness histogram within the reliability diagram. Sharpness is a function of only the forecast, and is a measure of the width of a forecast probability distribution (Gneiting et al. 2007). Given a forecast threshold, it is a count of how many times this threshold value lands in each probability bin. Forecasts that most often place this threshold near either the 0^{th} or 100^{th} percentiles are said to be sharp, while forecasts that frequently place the forecast near the 50^{th} percentile are not sharp. Narrower forecast probability distributions are sharper. While a sharper forecast is useful because it is more decisive, caution must be used to make sure that sharp forecasts are also probabilistically calibrated (i.e., a falsely confident forecast is not useful).

The CRPS is analogous to the MAE, but for probability. It is a measure of how well probability is assigned around each observed value (Hersbach 2000). A lower CRPS indicates a forecast that better assigns probability with respect to actual event occurrence (i.e., it assigns higher probability for the event). A perfect forecast will have a CRPS of 0. The CRPS reduces to the MAE for deterministic forecasts (the cumulative probability distribution takes the form of a step function), which makes it possible to compare ensemble and deterministic forecasts in a probabilistic sense. In addition, the CRPS provides a strictly proper scoring metric of assessing both forecast calibration and sharpness in a single score (Gneiting et al. 2005). Strictly proper scores promote honesty in the sense that forecasters can not hedge their forecasts to achieve a better verification score (Wilks 2011). The CRPS skill score is calculated by comparing the CRPS score of any probabilistic forecast to that of a reference probabilistic forecast.

Lastly, we will refer to root mean square error (RMSE) as a deterministic measure of forecast accuracy and also relate it to probabilistic-forecast sharpness, since RMSE is directly related to the spread of the Gaussian uncertainty model.

3. Results and Discussion

a. Raw ensemble distribution as a probability distribution

Tests R1-R8 (Table 2) are designed to quantify improvements in probabilistic calibration due to adding more PBL schemes to a short-range ensemble forecast in complex terrain. Starting with the best PBL scheme (ACM2, test R1), additional PBL schemes were added to the ensemble in tests R2-R8. Each additional PBL scheme added six total members to the ensemble, comprised of three grid sizes and two ICs. R1-R8 used the binned, raw-ensemble distribution as the forecast-probability distribution. Tests RB1-RB8 followed the same method, but used ensemble-member bias correction.

Fig. 5 shows the improvement in probabilistic calibration resulting from adding up to eight PBL schemes (i.e., improvement over R1). For the raw-ensemble distribution, forming a multi-PBL scheme, multi-IC, and multi-grid ensemble leads to large improvements in probabilistic calibration over an ensemble consisting of just a single PBL scheme (six members). For example, using 2 PBL schemes (12 members; R2, Table 2) results in 9% calibration improvement over 1 PBL scheme (R1, Table 2). A statistically significant 19% improvement in calibration (relative to R1, Table 2) results from using all eight PBL schemes (R8, Table 2). Throughout this manuscript we refer to statistical significance as exceeding the $p = 0.05$ significance level in a Student's t-test.

Once bias correction is applied, using all eight PBL schemes (RB8, Table 2) provides a 27% improvement in calibration (relative to RB1, Table 2). A 14% improvement is found when using only two PBL schemes (RB2, Table 2). Differences in forecast calibration between RB1 and RB 8 are statistically significant. Additionally, improvements in calibration through the application of bias-correction also pass statistical significance testing (e.g., comparing tests R1-R8 to RB1-RB8).

Even though R8 and RB8 were the best within their respective group of experiments, their PIT histograms (Fig. 6) show that the binned, raw-ensemble distribution is under-dispersive, even with a 48-member ensemble. The probability forecast is too confident and events too-often fall at the extremes of the distribution. To fix this, a better uncertainty model is needed, and we address this through our next set of tests.

b. Prescribed probability distributions dressed on the ensemble mean

Tests G1-G8 (Table 2) use the GNS uncertainty model dressed about the raw ensemble mean with spread (variance) based on the past squared errors of the ensemble mean. Using this method with the best overall PBL scheme (G1), or all eight PBL schemes (G8), results in a biased forecast (Fig 7). To remove the bias, we performed bias correction on each ensemble member in tests GB1-GB8. These also used the GNS uncertainty model, but centered the forecast probability distribution on the bias-corrected ensemble mean. The removal of the bias resulted in a nearly flat PIT histogram, indicating a calibrated forecast (GB1 and GB8 in Fig. 7). A separate calibration step is therefore not needed.

Next, we tested two additional Gaussian-based uncertainty models that allow the distribution to scale by either the ensemble variance or the ensemble mean. First, we scaled the distribution by the ensemble variance in tests GSV1-GSV8 (no ensemble-member bias-correction, Table 2) and GSVB1-GSVB8 (with ensemble-member bias correction, Table 2). Tests GSV1-GSV8 and GSVB1-GSVB8 show nearly identical results to using a non-scaling Gaussian uncertainty model (GNS). PIT histograms for GSV1 and GSV8 show a biased probabilistic forecast (GSV1 and GSV8; Fig. 7), that result in a well-calibrated forecast once ensemble-member bias correction is applied (GSVB1 and GSVB8; Fig. 7). Tests GSM1-GSM8 and GSMB1-GSMB8 yielded similar results (Fig. 7) for scaling the Gaussian distribution by the ensemble mean. When using any of

the Gaussian-based uncertainty models, improvements in calibration through the use of the bias-corrected ensemble members passed significance testing.

Fig. 8 shows the actual distributions of wind-speed forecast errors with respect to the bias-corrected ensemble mean. Each histogram is for a different wind farm. By inspection, these distributions do not have a Gaussian shape. These distributions have more probability near the center (kurtosis = 3.5 to 6.6) than a Gaussian (kurtosis = 3), but they are nearly symmetric (skewness of less than 0.5, close to the Gaussian zero skewness). This motivates two questions: a) what aspects of the Gaussian distribution allow it to work so well to produce calibrated probabilistic forecasts, and b) can we find a different theoretical distribution that is a better fit to the observed error distributions? We address question one next, and address question two in section 3f and in the appendix.

Perhaps all that is needed to produce calibrated forecasts is to have almost any distribution shape that has central tendency, spread about the mean, and is symmetric. The fact that the Gaussian distribution has tails that unphysically extend to $\pm\infty$ appears to be irrelevant to its ability to satisfactorily dress the ensemble mean. So the next question is how much spread in the distribution is needed?

First, we investigate whether larger forecast errors correspond to a) greater spread among the raw-ensemble members, or b) larger ensemble-mean-forecast winds. To do this, we calculated the coefficient of determination (r^2) for the linear relationship between the squared error of the bias-corrected ensemble mean, and either the bias-corrected ensemble variance or bias-corrected ensemble mean. We found the relationship between the bias-corrected ensemble mean and the squared error had r^2 values of 0.027 or less. This means that only 2.7% of the error variance can be explained by the linear relationship between the ensemble variance and squared-error magnitude (e.g., the spread-skill relationship does not apply at these four locations in complex terrain).

The corresponding analysis comparing the bias-corrected ensemble mean with the squared error magnitude resulted in r^2 values no larger than 0.053. The implications of this are that the Gaussian distributions used in GB1-GB8, GSVB1-GSVB8, and GSMB1-GSMB8 are similar, since the scaling relationships are weak. At some of the four locations, the differences between the GNS, GSEV, and GSEM distributions are statistically insignificant. Because of this, our next analyses will focus on the GNS distribution used in tests GB1-GB8.

An r^2 value of 1 is not expected or feasible, as several studies have noted (Whitaker and Loughe 1998; Grimit and Mass 2007; Hopson 2014). Some of those studies, and Wang and Bishop (2003), propose potential methods for extracting a stronger spread-skill relationship. Exploring those methods is beyond the scope of this study. Hopson (2014), however, concludes that in the case of a weak-spread relationship, an invariant spread, derived from ensemble-mean error, may be used instead. We reach the same conclusion here.

Having addressed calibration, we now look at sharpness. For the GNS uncertainty model, spread is only a function of the past error of the bias-corrected ensemble mean (3). The smaller the error (which can be measured by RMSE), the narrower the prescribed forecast probability distribution, and the sharper the forecast. We address this in the next section.

c. Use of ensembles to improve forecast sharpness

The benefit in using ensembles in the short-term forecast horizons has been debated, especially for near surface variables like hub-height winds (Mass et al. 2002; Eckel and Mass 2005; Warner 2011). We approach this question by looking at the effects of an ensemble on forecast sharpness. When using the GNS uncertainty model, forecast systems that have lower RMSE must have smaller probabilistic spread (improved sharpness).

Fig. 9 shows the average RMSE for each ensemble member, as well as the six-member bias-corrected ensemble from the best PBL scheme (GB1, Table 2) and the full 48-member bias-corrected ensemble (GB8, Table 2). While probabilistic calibration could be achieved by dressing a Gaussian distribution around any of these *individual* ensemble members, the sharpest forecasts are produced by dressing the distribution on the ensemble means from GB1 or GB8, which have the lowest RMSE. While we show that the best-PBL ensemble (GB1) and full-PBL ensemble (GB8) have equivalent sharpness, it is not possible to know which PBL scheme is best apriori, unless it has been shown to consistently perform best over an extended period of time. Nonetheless, these results indicate that larger ensembles do not necessarily perform better in short-term forecasts than smaller, selective ensembles. This is an important find for short-term wind planning as it reduces computational cost, while still providing sharp, probabilistically calibrated forecasts.

d. Probabilistic forecasts for wind events

While the bias-corrected ensemble forecasts show statistical calibration for the data used in the training period, they are not necessarily calibrated over subsets of this training data. Forecasts of rare events, which in our case are high-wind events, can be uncalibrated. To assess calibration over different wind-speed subsets, we provide the reliability diagrams using event thresholds of 5, 15, and 20 m s^{-1} for tests RB1, RB8, GB1, GB8, and also for a GNS distribution dressed on the single-best bias-corrected ensemble member (ACM2 12-km GNS; not in Table 2) (Fig. 10).

RB1 and RB8 are under-dispersive for an event threshold of 5 m s^{-1}, while GB1, GB8, and the ACM2 12-km GNS are calibrated. However, for the 15 and 20 m s^{-1} thresholds, all forecasts exhibit poor calibration, typically overforecasting probabilities. The horizontal dashed line in Fig. 10 represents the climatological frequency of events above the given threshold. For the 15 and 20 m s^{-1} thresholds, these events represent less than 5% of the total observed distribution. For the 15

m s^{-1} threshold, forecasts have near-zero skill (close to red dashed line, Fig. 10). For the 20 m s^{-1} threshold, the probabilistic forecast has no skill.

Insets within the panels of Fig. 10 are sharpness diagrams for the event thresholds. These diagrams indicate how often the event threshold falls in each forecast probability bin. Sharper forecasts will more frequently give event probabilities closer to 0% or 100%, while less sharp forecasts give event probabilities closer to 50%. For all thresholds, we see that the raw bias-corrected-ensemble distribution (RB1 and RB8) produces the sharpest forecasts. However, sharpness without calibration is worthless. The high sharpness is a result of the under-dispersive, over-confident nature of raw-ensemble distributions. Each of the GNS distributions (GB1, GB8, ACM2 12-km GNS in Fig. 10) produce sharp forecasts for each of these event thresholds.

e. Summary of skill vs. computation cost

As a summary of the forecast improvements obtained by using probabilistic-based forecasts, we plot the CRPS skill score, which allows us to compare deterministic and ensemble forecasts in a probabilistic sense (Fig. 11). We use the worst-performing deterministic forecast (the bias-corrected 36-km QNSE forecast) as the reference forecast to gauge short-term probabilistic-forecast improvements. Using the best deterministic forecast (12-km ACM2 forecast) provides an 11% improvement in the CRPS, averaged over all stations. A six-member ensemble forecast based on the best PBL scheme using the raw-ensemble distribution as the uncertainty model (RB1, Table 2) improves the CRPS by an additional 17% (or 28% over the worst deterministic forecast). Using the raw distribution with the full 48-member ensemble (RB8, Table 2), provides an additional improvement in CRPS of 3%, or a 31% improvement over the worst deterministic forecast.

Fig. 11 shows the large improvement that is gained by dressing the ensemble mean with an uncertainty model. The bias-corrected ACM2 12-km GNS provides a better probabilistic forecast

than that of RB8 (on average), at a fraction of the computational cost. CRPS is improved over RB8 by an additional 5%. Test GB1, which uses all the ACM2-based forecasts, provides the most CRPS skill, with a 41% improvement over the worst deterministic forecast, and an additional 5% improvement over the best deterministic forecast dressed by the GNS distribution. No further improvement is gained with a larger ensemble for this short-term forecast (GB8, Fig. 11), while computational cost increases. All ensemble forecasts in Fig. 11 (e.g., tests RB1, RB8, ACM2 12-km GNS, GB1, and GB8) have probabilistic skill (increased sharpness) over a climatology forecast while the two deterministic forecasts do not (the ACM2 12-km deterministic and the 36-km QNSE deterministic reference forecasts; not explicitly shown).

f. Shape of the prescribed distribution

In section 3b, we suggested that the exact shape of the prescribed distribution used to dress the ensemble mean is less important than the distribution's gross attributes: central tendency, spread, and symmetry. As a preliminary test of this hypothesis, we devise a new distribution, called the pq distribution. This distribution can better fit the shape of the observed wind-forecast-error distribution (Fig. 12). The pq distribution is symmetric and bounded; see Appendix B for details.

When Fig. 11 was re-calculated using CRPS from the best-fitting pq distributions, the results (not shown) were nearly identical to those from the GNS distribution. The lack of improvement in CRPS through the use of a better-fitted distribution (pq) is preliminary evidence that the distribution's gross attributes are more important, rather than the exact shape. Although the wind-forecast errors were nearly symmetric for the four wind farms studied here, there might be other situations where the forecast-error distributions may be asymmetric or even multi-modal (i.e., not Gaussian; not pq), and other distributions would better describe the errors.

4. Conclusions and Future Work

We detailed methods to produce calibrated hub-height wind-speed forecasts over complex terrain from a multi-PBL-scheme, multi-IC, multi-grid-length WRF ensemble. Tests evaluated the effects of having multiple PBL schemes in the ensemble, bias correction, and uncertainty model choice. Probabilistic forecast performance was evaluated based on improvements in the PIT histogram, reliability diagram, sharpness, CRPS, and RMSE.

For the binned raw-ensemble distribution, increasing the number of PBL schemes available improved forecast calibration. However, this distribution was under-dispersive, and remained probabilistically uncalibrated even after removing the bias from each individual ensemble member.

To improve probabilistic calibration, we tested if a prescribed uncertainty model could be used to better represent forecast uncertainty. Three Gaussian-based uncertainty models and one pq-model were evaluated. The first had no scaling; the distribution was based on the past error. The second and third Gaussian-based uncertainty models attempted to scale the Gaussian distribution based on a linear regression of either the ensemble variance or the ensemble mean against the square of the past errors from the ensemble mean. Using any Gaussian uncertainty model without individual member bias correction yielded biased PIT histograms. However, bias correcting each individual ensemble member resulted in probabilistically-calibrated forecasts for all three Gaussian-based uncertainty models.

When using the prescribed uncertainty models, using additional PBL schemes did not result in improved probabilistic calibration. Therefore, large ensembles are not necessary to produce probabilistically calibrated forecasts. Instead, the addition of multiple PBL schemes improved sharpness. Although the RMSE (and thus sharpness) was similar when using the full 48-member ensemble and the reduced 6-member ensemble using only the ACM2 PBL scheme, the single best

PBL scheme may not be known apriori. If the best PBL scheme is known, a selective ensemble can be used to save computation costs.

Additionally, the ensemble-mean RMSE, and thus sharpness, may be improved through the use of ICs produced by other agencies. Examples include the Canadian Global Deterministic Prediction System, the Fleet Numerical Meteorology and Oceanography Center Navy Global Environmental Model, the UK Metoffice Unified Model, and the European Centre for Medium Range Weather Forecasts Integrated Forecast System. The reason we suggest testing the use of these other sources is because the underlying data assimilation system used for both the GFS and the NAM is the same: the Gridpoint Statistical Interpolation (Shao et al. 2016). Differences between ICs may be larger when the sources come from distinctly different agencies (those using different data-assimilation techniques).

Our error analysis shows that the Gaussian distribution describes the forecast error well, but that forecast error has little relationship to either the ensemble mean or the ensemble variance. We also found that gross attributes (central tendency, spread, symmetry) of a prescribed distribution are more important than its exact shape.

5. Acknowledgments

The authors acknowledge BC Hydro, Mitacs, and the Canadian Natural Science and Engineering Research Council for providing the funding to make this extensive project possible. We also thank GDF SUEZ Energy North America, Altagas Ltd, Alterra Power Corp, and Capital Power for allowing us access to their data. In addition, we thank Magdalena Rucker and Doug McCollor for their contributions to this effort. Finally, we would like to acknowledge the three anonymous reviewers for their input which greatly improved this manuscript.

6. Appendix A

Table A1 provides the deterministic MAE scores for the bias-corrected forecasts used to identify the best-performing PBL scheme in a separate study (Siuta et al. 2016). Fig. A1 quantifies the annual accuracy skill gain or loss (MAESS) at each individual wind-farm site by using the DMB bias-correction scheme and a training period of 30 days. Averaged over all locations, bias correction always resulted in accuracy improvements (purple bars, Fig. A1). For sites 1-3, improvements in accuracy over the raw forecasts ranged from 0-12%. At site 4, MAE was improved by up to 56% (the 4-km grid using the MRF PBL scheme).

Fig. A2 shows the variation in the annual MAESS at all four wind-farm sites averaged over all 48 bias-corrected forecasts. While we do not show training periods less than 5 days in length, we tested a training period of only 1 day and found MAESS to be significantly worse (-40% or worse), thus it is not included in Fig. A2. We find that the benefits of a longer training period level out between 20 and 30 days.

7. Appendix B

We introduce the meteorological community to a new frequency distribution that we devised for section 3. The distribution has a simple un-normalized one-sided form:

$$f_s(x) = (1 - x^p)^q \tag{B1}$$

where p and q are the shape parameters and the distribution is defined between 0 and 1. It can be made into a normalized (unit area) symmetric probability density (f) with arbitrary scaling

parameter S and center-location parameter x_o:

$$f(x) = \frac{1}{A}\left[1 - \left|\frac{x - x_o}{S}\right|^p\right]^q \quad \text{for} \quad (x_o - S) \leq x \leq (x_o + S) \tag{B2}$$

where the area A under the un-normalized symmetric curve is

$$A = \frac{2S\Gamma(q+1)\Gamma(1/p)}{p\Gamma[q+1+(1/p)]} \tag{B3}$$

and where Γ is the gamma function. The pq distribution is similar, but not identical, to the Kumaraswamy (1980) distribution and the beta distribution (Jambunathan 1954).

The pq distribution has a theoretical mean absolute deviation of

$$MAD = S\frac{\Gamma[q+1+(1/p)]\Gamma(2/p)}{\Gamma[q+1+(2/p)]\Gamma(1/p)} \tag{B4}$$

and variance of

$$\sigma_x^2 = S^2\frac{\Gamma[q+1+(1/p)]\Gamma(3/p)}{\Gamma[q+1+(3/p)]\Gamma(1/p)}. \tag{B5}$$

The fourth statistical moment is

$$\mu_4 = S^4\frac{\Gamma[q+1+(1/p)]\Gamma(5/p)}{\Gamma[q+1+(5/p)]\Gamma(1/p)} \tag{B6}$$

from which the kurtosis can be found as μ_4/σ_x^4. Skewness is zero for this symmetric distribution. Sometimes a situation may exist where part of the distribution tail on one or both sides must be cut off. For this situation, numerically integrate to get the area under the curve and to get the higher

moments, using evenly-spaced samples from (B2) instead of using the theoretical expressions (B3) - (B6).

If you are given an observed distribution with scatter, then you can find the best-fit parameters (p, q, S, x_o) for the pq distribution as follows. If you know apriori (or want to fix) one or more of the parameters (typically S and/or x_o), then do so and find the remaining parameters by one of two methods. You can use the method of moments, where you first calculate the statistical moments of the observation data, and then use the the Newton-Raphson or other root-finding algorithm to find the pq-distribution parameters that give the same statistical moments. You will need as many statistical moments as unknown parameters.

A second approach is brute force, which we used here to help us learn more about the sensitivity of the distribution to the choice of parameters. Namely, loop through all reasonable values of p and q, and for each (p, q) combination find the mean absolute error between the pq distribution and the observed frequency distribution (after normalizing the observations to have unit area under its histogram). The best (p, q) combination is the one with minimum error. Fig. B1 shows an example for one of the wind-farm forecast-error distributions.

While stepping through each (p, q) combination, we also produced a plot of the pq-distribution MAD, variance and kurtosis (Fig. B2). You can use this as a poor-mans method of moments. Namely, if you know the variance and kurtosis (or the MAD and kurtosis) from your observation data, then you can use those values in Fig. B2 to find the approximate p and q values.

Although the pq distribution cannot fit skewed distributions, its versatility for symmetric distributions is remarkable. Fig. B3 shows extremes that approximate a delta function, a linear ramp shape, and a near-uniform distribution. Fig. B4 shows that q controls the slope of the tails, while p controls the shape of center of the distribution. The last frame in Fig. B4 shows how the pq distribution can approximate Gaussian distributions, particularly for values of $p \approx 1.85$.

References

Ahlstrom, M., and Coauthors, 2013: Knowledge is power: Efficiently integrating wind energy and wind forecasts. 45–52 pp., doi:10.1109/MPE.2013.2277999.

Anderson, J. L., 1996: A method for producing and evaluating probabilistic precipitation forecasts from ensemble model integrations. *Journal of Climate*, **9 (7)**, 1518–1530, doi:10.1175/1520-0442(1996)009⟨1518:AMFPAE⟩2.0.CO;2.

Berner, J., G. J. Shutts, M. Leutbecher, and T. N. Palmer, 2009: A spectral stochastic kinetic energy backscatter scheme and its impact on flow-dependent predictability in the ECMWF ensemble prediction system. *Journal of the Atmospheric Sciences*, **66 (3)**, 603–626, doi:10.1175/2008JAS2677.1.

Bludszuweit, H., J. A. Dominguez-Navarro, and A. Llombart, 2008: Statistical analysis of wind power forecast error. 983–991 pp., doi:10.1109/TPWRS.2008.922526.

Bourdin, D. R., T. N. Nipen, and R. B. Stull, 2014: Reliable probabilistic forecasts from an ensemble reservoir inflow forecasting system. *Water Resources Research*, **50 (4)**, 3108–3130, doi:10.1002/2014WR015462.

Bretherton, C. S., and S. Park, 2009: A new moist turbulence parameterization in the Community Atmosphere Model. *Journal of Climate*, **22 (12)**, 3422–3448, doi:10.1175/2008JCLI2556.1.

Buizza, R., P. L. Houtekamer, G. Pellerin, Z. Toth, Y. Zhu, and M. Wei, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Monthly Weather Review*, **133 (5)**, 1076–1097, doi:10.1175/MWR2905.1.

Candille, G., 2009: The multiensemble approach: The NAEFS example. *Monthly Weather Review*, **137 (5)**, 1655–1665, doi:10.1175/2008MWR2682.1.

Courtney, J., P. Lynch, and C. Sweeney, 2013: High resolution forecasting for wind energy applications using bayesian model averaging. *Tellus A*, **65 (0)**, doi:10.3402/tellusa.v65i0.19669.

Cutler, N. J., H. R. Outhred, and I. F. MacGill, 2012: Using nacelle-based wind speed observations to improve power curve modeling for wind power forecasting. *Wind Energy*, **15 (2)**, 245–258, doi:10.1002/we.465.

Deppe, A. J., W. A. Gallus, and E. S. Takle, 2012: A WRF ensemble for improved wind speed forecasts at turbine height. *Weather and Forecasting*, **28 (1)**, 212–228, doi:10.1175/WAF-D-11-00112.1.

Draxl, C., A. N. Hahmann, A. Peña, and G. Giebel, 2014: Evaluating winds and vertical wind shear from weather research and forecasting model forecasts using seven planetary boundary layer schemes. *Wind Energy*, **17 (1)**, 39–55, doi:10.1002/we.1555.

Eckel, F. A., and C. F. Mass, 2005: Aspects of effective mesoscale, short-range ensemble forecasting. *Weather and Forecasting*, **20 (3)**, 328–350, doi:10.1175/WAF843.1.

Giebel, G., R. Brownsword, G. Kariniotakis, M. Denhard, and C. Draxl, 2011: The state-of-the-art in short-term prediction of wind power - a literature overview, 2nd edition. Tech. rep., Riso DTU, 109 pp.

Gneiting, T., F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69 (2)**, 243–268, doi:10.1111/j.1467-9868.2007.00587.x.

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, **133 (5)**, 1098–1118, doi:10.1175/MWR2904.1.

Grenier, H., and C. S. Bretherton, 2001: A moist PBL parameterization for large-scale models and its application to subtropical cloud-topped marine boundary layers. *Monthly Weather Review*, **129 (3)**, 357–377, doi:10.1175/1520-0493(2001)129⟨0357:AMPPFL⟩2.0.CO;2.

Grimit, E. P., and C. F. Mass, 2002: Initial results of a mesoscale short-range ensemble forecasting system over the Pacific Northwest. *Weather and Forecasting*, **17 (2)**, 192–205, doi:10.1175/1520-0434(2002)017⟨0192:IROAMS⟩2.0.CO;2.

Grimit, E. P., and C. F. Mass, 2007: Measuring the ensemble spread-error relationship with a probabilistic approach: Stochastic ensemble results. *Monthly Weather Review*, **135 (1)**, 203–221, doi:10.1175/MWR3262.1.

Grubišić, V., R. K. Vellore, and A. W. Huggins, 2005: Quantitative precipitation forecasting of wintertime storms in the sierra nevada: Sensitivity to the microphysical parameterization and horizontal resolution. *Monthly Weather Review*, **133 (10)**, 2834–2859, doi:10.1175/MWR3004.1.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, **15 (5)**, 559–570, doi:10.1175/1520-0434(2000)015⟨0559:DOTCRP⟩2.0.CO;2.

Hoeting, J. A., M. Madigan, A. E. Raftery, and C. T. Volinsky, 1999: Bayesian model averaging: A tutorial. *Statistical Science*, **14 (4)**, 382–417, doi:10.1214/ss/1009212519.

Hong, S.-Y., Y. Noh, and J. Dudhia, 2006: A new vertical diffusion package with an explicit treatment of entrainment processes. *Monthly Weather Review*, **134 (9)**, 2318–2341, doi:10.1175/MWR3199.1.

Hong, S.-Y., and H.-L. Pan, 1996: Nonlocal boundary layer vertical diffusion in a medium-range forecast model. *Monthly Weather Review*, **124 (10)**, 2322–2339, doi:10.1175/1520-0493(1996)124⟨2322:NBLVDI⟩2.0.CO;2.

Hopson, T. M., 2014: Assessing the ensemble spreaderror relationship. *Monthly Weather Review*, **142 (3)**, 1125–1142, doi:10.1175/MWR-D-12-00111.1.

Hu, X.-M., P. M. Klein, and M. Xue, 2013: Evaluation of the updated YSU planetary boundary layer scheme within WRF for wind resource and air quality assessments. *Journal of Geophysical Research: Atmospheres*, **118 (18)**, 10,410–490,505, doi:10.1002/jgrd.50823.

Jambunathan, M. V., 1954: Some properties of beta and gamma distributions. *The Annals of Mathematical Statistics*, **25 (2)**, 401–405, doi:10.1214/aoms/1177728800.

Janjić, Z. I., 1994: The step-mountain Eta coordinate model: Further developments of the convection, viscous sublayer, and turbulence closure schemes. *Monthly Weather Review*, **122 (5)**, 927–945, doi:10.1175/1520-0493(1994)122⟨0927:TSMECM⟩2.0.CO;2.

Juban, J., N. Siebert, and G. N. Kariniotakis, 2007: Probabilistic short-term wind power forecasting for the optimal management of wind generation. *2007 IEEE Lausanne Power Tech*, IEEE, 683–688, doi:10.1109/PCT.2007.4538398.

Junk, C., S. Späth, L. von Bremen, and L. Delle Monache, 2015: Comparison and combination of regional and global ensemble prediction systems for probabilistic predictions of hub-height wind speed. *Weather and Forecasting*, **30 (5)**, 1234–1253, doi:10.1175/WAF-D-15-0021.1.

Kumaraswamy, P., 1980: A generalized probability density function for double-bounded random processes. *Journal of Hydrology*, **46 (1-2)**, 79–88, doi:10.1016/0022-1694(80)90036-0.

Lange, M., 2005: On the uncertainty of wind power predictionsanalysis of the forecast accuracy and statistical distribution of errors. *Journal of Solar Energy Engineering*, **127 (2)**, 177–184.

Mahoney, W. P., and Coauthors, 2012: A wind power forecasting system to optimize grid integration. *IEEE Transactions on Sustainable Energy*, **3 (4)**, 670–682, doi:10.1109/TSTE.2012.2201758.

Marquis, M., J. Wilczak, M. Ahlstrom, J. Sharp, A. Stern, J. C. Smith, and S. Calvert, 2011: Forecasting the wind to reach significant penetration levels of wind energy. *Bulletin of the American Meteorological Society*, **92 (9)**, 1159–1171, doi:10.1175/2011BAMS3033.1.

Mass, C. F., D. Ovens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bulletin of the American Meteorological Society*, **83 (3)**, 407–430, doi:10.1175/1520-0477(2002)083⟨0407:DIHRPM⟩2.3.CO;2.

McCollor, D., and R. Stull, 2008a: Hydrometeorological short-range ensemble forecasts in complex terrain. Part I: Meteorological evaluation. *Weather and Forecasting*, **23 (4)**, 533–556, doi:10.1175/2008WAF2007063.1.

McCollor, D., and R. Stull, 2008b: Hydrometeorological short-range ensemble forecasts in complex terrain. Part II: Economic evaluation. *Weather and Forecasting*, **23 (4)**, 557–574, doi:10.1175/2007WAF2007064.1.

McCollor, D., and R. Stull, 2008c: Hydrometeorological accuracy enhancement via postprocessing of numerical weather forecasts in complex terrain. *Weather and Forecasting*, **23**, 131–144.

Monteiro, C., R. Bessa, V. Miranda, A. Botterud, J. Wang, and G. Conzelmann, 2009: Wind power forecasting: State-of-the-art 2009. *Argonne National Laboratory*, 1–216, doi:10.2172/968212.

Nakanishi, M., and H. Niino, 2006: An improved MellorYamada Level-3 model: Its numerical stability and application to a regional prediction of advection fog. *Boundary-Layer Meteorology*, **119 (2)**, 397–407, doi:10.1007/s10546-005-9030-8.

Nipen, T., and R. Stull, 2011: Calibrating probabilistic forecasts from an NWP ensemble. *Tellus A*, **63 (5)**, 858–875, doi:10.1111/j.1600-0870.2011.00535.x.

Nipen, T. N., 2012: A component-based probabilistic weather forecasting system for operational usage. PhD. dissertation, Dept. of Earth, Ocean, and Atmospheric Sciences, University of British Columbia. [Available online at https://open.library.ubc.ca/collections/24/items/1.0053570].

Pinson, P., 2012: Very-short-term probabilistic forecasting of wind power with generalized logit-normal distributions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **61 (4)**, 555–576, doi:10.1111/j.1467-9876.2011.01026.x.

Pinson, P., J. Juban, and G. N. Kariniotakis, 2006: On the quality and value of probabilistic forecasts of wind generation. *2006 International Conference on Probabilistic Methods Applied to Power Systems*, IEEE, 1–7, doi:10.1109/PMAPS.2006.360290.

Pleim, J. E., 2007: A combined local and nonlocal closure model for the atmospheric boundary layer. Part I: Model description and testing. *Journal of Applied Meteorology and Climatology*, **46 (9)**, 1383–1395, doi:10.1175/JAM2539.1.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, **133 (5)**, 1155–1174, doi:10.1175/MWR2906.1.

Shao, H., and Coauthors, 2016: Bridging research to operations transitions: Status and plans of community GSI. *Bulletin of the American Meteorological Society*, **97 (8)**, 1427–1440, doi: 10.1175/BAMS-D-13-00245.1.

Siuta, D., G. West, and R. Stull, 2016: WRF hub-height wind forecast sensitivity to PBL scheme, grid length, and initial-condition choice in complex terrain. *Weather and Forecasting*, submitted.

Skamarock, W., J. Klemp, J. Dudhia, D. Gill, D. Barker, W. Wang, X.-y. Huang, and M. Duda, 2008: A description of the Advanced Research WRF version 3. NCAR Technical Note NCAR/TN-475+STR, doi:10.5065/D68S4MVH.

Sloughter, J. M., T. Gneiting, and A. E. Raftery, 2010: Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *Journal of the American Statistical Association*, **105 (489)**, 25–35, doi:10.1198/jasa.2009.ap08615.

Stensrud, D. J., 2007: *Parameterization Schemes: Keys to Understanding Numerical Weather Prediction Models*. Cambridge University Press, 459 pp.

Stensrud, D. J., J.-W. Bao, and T. T. Warner, 2000: Using initial condition and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. *Monthly Weather Review*, **128 (7)**, 2077–2107, doi:10.1175/1520-0493(2000)128⟨2077:UICAMP⟩2.0.CO;2.

Stull, R. B., 1988: *An Introduction to Boundary Layer Meteorology*. Kluwer Academic Publishers, 666 pp.

Sukoriansky, S., B. Galperin, and V. Perov, 2005: Application of a new spectral theory of stably stratified turbulence to the atmospheric boundary layer over sea ice. *Boundary-Layer Meteorology*, **117 (2)**, 231–257, doi:10.1007/s10546-004-6848-4.

Wang, X., and C. H. Bishop, 2003: A comparison of breeding and Ensemble Transform Kalman Filter ensemble forecast schemes. *Journal of the Atmospheric Sciences*, **60 (9)**, 1140–1158, doi:10.1175/1520-0469(2003)060⟨1140:ACOBAE⟩2.0.CO;2.

Warner, T. T., 2011: *Numerical weather and climate prediction*. Cambridge University Press, 526 pp.

Whitaker, J. S., and A. F. Loughe, 1998: The relationship between ensemble spread and ensemble mean skill. *Monthly Weather Review*, **126 (12)**, 3292–3302, doi:10.1175/1520-0493(1998)126⟨3292:TRBESA⟩2.0.CO;2.

Wilczak, J., and Coauthors, 2015: The Wind Forecast Improvement Project (WFIP): A public-private partnership addressing wind energy forecast needs. *Bulletin of the American Meteorological Society*, **96 (10)**, 1699–1718, doi:10.1175/BAMS-D-14-00107.1.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed., Academic Press, 676 pp.

Zhang, Y., J. Wang, and X. Wang, 2014: Review on probabilistic forecasting of wind power generation. *Renewable and Sustainable Energy Reviews*, **32**, 255–270, doi:10.1016/j.rser.2014.01.033.

LIST OF TABLES

Table 1. WRF-model configurations used for this study. 37

Table 2. Summary of the tests performed. PBL schemes are detailed in section 2. Bias correction refers to the degree-of-mass-balance (DMB) multiplicative bias correction applied to each individual ensemble member before ensemble statistics are computed. A description of the uncertainty models is provided in section 2c. . . . 38

Table A1. Mean Absolute Error (MAE) for each bias-corrected forecast initialized off the GFS. Sites 1-4 are the anonymous wind farm locations. Statistics are divided into five blocks, Annual, Summer (June - August), Fall (September - November), Winter (December - February), and Spring (March - May). MAE is in m s^{-1}. 39

Table A2. Mean Absolute Error (MAE) for each bias-corrected forecast initialized off the NAM. Sites 1-4 are the anonymous wind farm locations. Statistics are divided into five blocks, Annual, Summer (June - August), Fall (September - November), Winter (December - February), and Spring (March - May). MAE is in m s^{-1}. 40

TABLE 1. WRF-model configurations used for this study.

Model Detail	Setting
WRF Core	ARW Version 3.5.1
Grid Dimensions	36 km: 100x76, 12 km: 136x103, 4km: 103x103
Land Surface	Noah
Microphysics	WRF Single-Moment 5-class
Shortwave Radiation	Dudhia
Longwave Radiation	Rapid Radiative Transfer Model
Surface Layer	MM5 similarity (used with YSU, ACM2, MRF, MYNN, GBM, and UW PBL schemes) or Eta similarity (MYJ PBL scheme only) or QNSE surface layer (QNSE PBL scheme only)

TABLE 2. Summary of the tests performed. PBL schemes are detailed in section 2. Bias correction refers to the degree-of-mass-balance (DMB) multiplicative bias correction applied to each individual ensemble member before ensemble statistics are computed. A description of the uncertainty models is provided in section 2c.

Test ID	Uncertainty Model	Bias Correction		PBL schemes used								Count of
		Yes	No	ACM2	MYJ	YSU	GBM	UW	MRF	QNSE	MYNN	Ensemble Members
R1	Raw		X	X								6
R2	Raw		X	X	X							12
R3	Raw		X	X	X	X						18
R4	Raw		X	X	X	X	X					24
R5	Raw		X	X	X	X	X	X				30
R6	Raw		X	X	X	X	X	X	X			36
R7	Raw		X	X	X	X	X	X	X	X		42
R8	Raw		X	X	X	X	X	X	X	X	X	48
RB1	Raw	X		X								6
RB2	Raw	X		X	X							12
RB3	Raw	X		X	X	X						18
RB4	Raw	X		X	X	X	X					24
RB5	Raw	X		X	X	X	X	X				30
RB6	Raw	X		X	X	X	X	X	X			36
RB7	Raw	X		X	X	X	X	X	X	X		42
RB8	Raw	X		X	X	X	X	X	X	X	X	48
G1	GNS		X	X								6
G2	GNS		X	X	X							12
G3	GNS		X	X	X	X						18
G4	GNS		X	X	X	X	X					24
G5	GNS		X	X	X	X	X	X				30
G6	GNS		X	X	X	X	X	X	X			36
G7	GNS		X	X	X	X	X	X	X	X		42
G8	GNS		X	X	X	X	X	X	X	X	X	48
GB1	GNS	X		X								6
GB2	GNS	X		X	X							12
GB3	GNS	X		X	X	X						18
GB4	GNS	X		X	X	X	X					24
GB5	GNS	X		X	X	X	X	X				30
GB6	GNS	X		X	X	X	X	X	X			36
GB7	GNS	X		X	X	X	X	X	X	X		42
GB8	GNS	X		X	X	X	X	X	X	X	X	48
GSV1	GSEV		X	X								6
GSV2	GSEV		X	X	X							12
GSV3	GSEV		X	X	X	X						18
GSV4	GSEV		X	X	X	X	X					24
GSV5	GSEV		X	X	X	X	X	X				30
GSV6	GSEV		X	X	X	X	X	X	X			36
GSV7	GSEV		X	X	X	X	X	X	X	X		42
GSV8	GSEV		X	X	X	X	X	X	X	X	X	48
GSVB1	GSEV	X		X								6
GSVB2	GSEV	X		X	X							12
GSVB3	GSEV	X		X	X	X						18
GSVB4	GSEV	X		X	X	X	X					24
GSVB5	GSEV	X		X	X	X	X	X				30
GSVB6	GSEV	X		X	X	X	X	X	X			36
GSVB7	GSEV	X		X	X	X	X	X	X	X		42
GSVB8	GSEV	X		X	X	X	X	X	X	X	X	48
GSM1	GSEM		X	X								6
GSM2	GSEM		X	X	X							12
GSM3	GSEM		X	X	X	X						18
GSM4	GSEM		X	X	X	X	X					24
GSM5	GSEM		X	X	X	X	X	X				30
GSM6	GSEM		X	X	X	X	X	X	X			36
GSM7	GSEM		X	X	X	X	X	X	X	X		42
GSM8	GSEM		X	X	X	X	X	X	X	X	X	48
GSMB1	GSEM	X		X								6
GSMB2	GSEM	X		X	X							12
GSMB3	GSEM	X		X	X	X						18
GSMB4	GSEM	X		X	X	X	X					24
GSMB5	GSEM	X		X	X	X	X	X				30
GSMB6	GSEM	X		X	X	X	X	X	X			36
GSMB7	GSEM	X		X	X	X	X	X	X	X		42
GSMB8	GSEM	X		X	X	X	X	X	X	X	X	48

Table A1. Mean Absolute Error (MAE) for each bias-corrected forecast initialized off the GFS. Sites 1-4 are the anonymous wind farm locations. Statistics are divided into five blocks, Annual, Summer (June - August), Fall (September - November), Winter (December - February), and Spring (March - May). MAE is in m s^{-1}.

Initial Condition	GFS																							
PBL Scheme	YSU			ACM2			MRF			MYJ			MYNN			QNSE			UW			GBM		
Grid Length (km)	36	12	4	36	12	4	36	12	4	36	12	4	36	12	4	36	12	4	36	12	4	36	12	4
Annual																								
Site 1	2.30	2.16	2.45	2.28	2.12	2.41	2.27	2.20	2.46	2.34	2.15	2.50	2.35	2.26	2.52	2.41	2.24	2.62	2.40	2.19	2.53	2.35	2.15	2.47
Site 2	1.83	1.82	1.79	1.76	1.77	1.68	1.81	1.86	1.82	1.83	1.85	1.77	1.92	1.92	1.86	1.91	1.93	1.82	1.85	1.88	1.85	1.78	1.80	1.79
Site 3	1.81	1.77	1.77	1.78	1.72	1.73	1.77	1.73	1.69	1.78	1.71	1.79	1.87	1.81	1.80	1.89	1.81	1.87	1.86	1.77	1.83	1.75	1.72	1.76
Site 4	1.87	1.61	1.62	1.85	1.61	1.60	1.86	1.66	1.66	1.82	1.60	1.62	1.83	1.65	1.64	1.88	1.62	1.63	1.85	1.64	1.67	1.80	1.61	1.61
Summer																								
Site 1	2.04	1.99	2.14	1.95	1.90	2.08	1.97	1.91	2.03	2.10	2.01	2.14	2.10	2.03	2.13	2.22	2.14	2.28	2.14	2.03	2.21	2.09	2.00	2.15
Site 2	1.70	1.70	1.64	1.58	1.59	1.53	1.69	1.65	1.58	1.72	1.72	1.62	1.80	1.78	1.68	1.80	1.81	1.71	1.75	1.77	1.69	1.72	1.71	1.64
Site 3	1.64	1.61	1.60	1.56	1.53	1.52	1.66	1.62	1.55	1.61	1.57	1.56	1.75	1.70	1.64	1.72	1.67	1.68	1.71	1.65	1.60	1.62	1.61	1.58
Site 4	1.58	1.21	1.32	1.57	1.21	1.28	1.64	1.25	1.34	1.64	1.21	1.37	1.64	1.26	1.36	1.62	1.22	1.36	1.59	1.23	1.38	1.57	1.17	1.29
Fall																								
Site 1	2.32	2.13	2.36	2.32	2.12	2.34	2.31	2.19	2.43	2.34	2.10	2.41	2.36	2.30	2.51	2.42	2.21	2.52	2.42	2.22	2.51	2.38	2.14	2.42
Site 2	1.79	1.83	1.82	1.72	1.77	1.71	1.76	1.92	1.87	1.77	1.87	1.78	1.93	2.00	1.94	1.86	1.93	1.81	1.82	1.92	1.86	1.76	1.84	1.81
Site 3	1.75	1.70	1.72	1.70	1.62	1.66	1.73	1.68	1.68	1.68	1.63	1.76	1.81	1.73	1.77	1.80	1.72	1.87	1.78	1.71	1.82	1.68	1.62	1.71
Site 4	1.96	1.84	1.74	1.97	1.86	1.73	1.95	1.90	1.79	1.86	1.82	1.72	1.89	1.88	1.74	1.97	1.88	1.76	1.99	1.91	1.81	1.85	1.82	1.73
Winter																								
Site 1	2.67	2.45	2.98	2.76	2.45	3.01	2.76	2.62	3.14	2.75	2.44	3.10	2.78	2.62	3.10	2.71	2.47	3.14	2.84	2.44	3.06	2.80	2.40	2.99
Site 2	2.03	1.90	1.92	2.03	1.93	1.80	2.05	2.05	2.05	2.01	1.97	1.87	2.11	2.01	1.94	2.04	1.99	1.87	2.01	1.98	1.95	1.90	1.87	1.88
Site 3	1.96	1.85	1.79	2.05	1.89	1.87	1.92	1.85	1.79	1.96	1.81	1.91	2.02	1.90	1.84	2.03	1.85	1.91	1.98	1.81	1.91	1.86	1.78	1.83
Site 4	2.16	1.97	1.98	2.10	1.93	1.94	2.06	1.99	2.02	2.08	1.98	1.97	2.07	1.98	1.98	2.15	1.97	1.95	2.10	1.95	1.99	2.03	1.96	1.96
Spring																								
Site 1	2.16	2.06	2.32	2.11	2.00	2.21	2.04	2.07	2.25	2.17	2.03	2.36	2.17	2.09	2.33	2.29	2.16	2.54	2.20	2.09	2.36	2.16	2.06	2.33
Site 2	1.82	1.84	1.78	1.74	1.77	1.69	1.76	1.81	1.80	1.80	1.85	1.82	1.85	1.88	1.86	1.94	1.99	1.89	1.82	1.87	1.92	1.76	1.79	1.82
Site 3	1.90	1.93	1.95	1.83	1.83	1.87	1.77	1.75	1.75	1.88	1.83	1.91	1.89	1.90	1.93	2.01	1.97	2.03	1.95	1.92	2.00	1.85	1.85	1.91
Site 4	1.68	1.33	1.35	1.67	1.35	1.34	1.72	1.41	1.39	1.64	1.29	1.35	1.68	1.38	1.38	1.71	1.31	1.37	1.67	1.35	1.42	1.67	1.35	1.38

Table A2. Mean Absolute Error (MAE) for each bias-corrected forecast initialized off the NAM. Sites 1-4 are the anonymous wind farm locations. Statistics are divided into five blocks, Annual, Summer (June - August), Fall (September - November), Winter (December - February), and Spring (March - May). MAE is in m s^{-1}.

Initial Condition	NAM																							
PBL Scheme	YSU			ACM2			MRF			MYJ			MYNN			QNSE			UW			GBM		
Grid Length (km)	36	12	4	36	12	4	36	12	4	36	12	4	36	12	4	36	12	4	36	12	4	36	12	4
Annual																								
Site 1	2.39	2.17	2.48	2.37	2.14	2.47	2.32	2.19	2.49	2.40	2.15	2.53	2.39	2.23	2.52	2.47	2.27	2.65	2.49	2.23	2.56	2.42	2.16	2.47
Site 2	1.81	1.83	1.76	1.75	1.80	1.70	1.81	1.86	1.79	1.81	1.87	1.78	1.88	1.89	1.83	1.90	1.95	1.84	1.84	1.88	1.85	1.74	1.81	1.81
Site 3	1.87	1.81	1.80	1.83	1.75	1.76	1.81	1.77	1.72	1.85	1.76	1.84	1.90	1.84	1.82	1.92	1.82	1.89	1.88	1.80	1.86	1.82	1.77	1.79
Site 4	1.88	1.63	1.64	1.85	1.63	1.63	1.88	1.66	1.67	1.82	1.62	1.67	1.85	1.67	1.67	1.88	1.63	1.66	1.86	1.68	1.71	1.82	1.64	1.65
Summer																								
Site 1	2.07	1.95	2.13	1.99	1.86	2.08	2.01	1.90	2.09	2.09	1.94	2.16	2.12	1.99	2.17	2.18	2.06	2.29	2.16	2.02	2.20	2.10	1.95	2.11
Site 2	1.73	1.72	1.70	1.63	1.66	1.62	1.73	1.74	1.67	1.80	1.72	1.70	1.84	1.76	1.74	1.86	1.84	1.77	1.84	1.80	1.76	1.74	1.72	1.71
Site 3	1.80	1.80	1.71	1.70	1.69	1.63	1.75	1.75	1.66	1.77	1.74	1.69	1.88	1.84	1.76	1.81	1.76	1.75	1.84	1.78	1.69	1.78	1.75	1.68
Site 4	1.60	1.26	1.36	1.58	1.28	1.35	1.64	1.27	1.36	1.66	1.27	1.43	1.66	1.31	1.41	1.63	1.27	1.41	1.61	1.30	1.43	1.58	1.23	1.34
Fall																								
Site 1	2.44	2.19	2.51	2.45	2.15	2.48	2.37	2.19	2.52	2.46	2.16	2.52	2.42	2.29	2.59	2.52	2.28	2.66	2.53	2.27	2.62	2.50	2.20	2.52
Site 2	1.75	1.83	1.74	1.71	1.82	1.69	1.74	1.89	1.81	1.74	1.88	1.75	1.88	1.95	1.82	1.84	1.94	1.84	1.78	1.93	1.86	1.71	1.88	1.79
Site 3	1.80	1.73	1.77	1.75	1.67	1.70	1.80	1.75	1.75	1.74	1.64	1.77	1.88	1.82	1.85	1.85	1.73	1.87	1.84	1.74	1.87	1.74	1.68	1.77
Site 4	2.00	1.86	1.77	2.01	1.86	1.74	1.98	1.87	1.80	1.87	1.80	1.74	1.95	1.91	1.78	2.00	1.85	1.76	1.99	1.92	1.80	1.91	1.84	1.75
Winter																								
Site 1	2.73	2.39	2.90	2.81	2.45	3.01	2.74	2.56	3.06	2.77	2.42	3.06	2.80	2.52	2.99	2.75	2.48	3.12	2.92	2.45	2.98	2.81	2.39	2.94
Site 2	1.95	1.92	1.82	1.91	1.92	1.75	1.98	2.00	1.92	1.93	1.98	1.83	2.02	1.99	1.89	1.98	2.02	1.82	1.93	1.96	1.91	1.81	1.84	1.86
Site 3	1.98	1.82	1.76	2.05	1.84	1.84	1.94	1.83	1.74	1.99	1.81	1.92	2.03	1.89	1.80	2.04	1.84	1.91	1.96	1.80	1.91	1.91	1.76	1.77
Site 4	2.11	1.92	1.95	2.05	1.88	1.90	2.06	1.96	1.97	2.02	1.94	1.96	2.04	1.96	1.95	2.07	1.95	1.97	2.04	1.96	2.00	2.01	1.96	1.98
Spring																								
Site 1	2.32	2.17	2.40	2.25	2.09	2.32	2.16	2.11	2.27	2.26	2.1	2.37	2.22	2.14	2.33	2.41	2.27	2.55	2.34	2.18	2.44	2.25	2.11	2.33
Site 2	1.83	1.85	1.80	1.76	1.82	1.72	1.78	1.81	1.77	1.77	1.89	1.86	1.79	1.84	1.85	1.93	2.00	1.95	1.80	1.85	1.88	1.70	1.80	1.86
Site 3	1.89	1.90	1.95	1.80	1.81	1.88	1.75	1.73	1.74	1.88	1.86	1.98	1.83	1.82	1.88	1.99	1.94	2.03	1.88	1.89	1.96	1.86	1.87	1.95
Site 4	1.73	1.37	1.42	1.70	1.42	1.43	1.78	1.45	1.47	1.71	1.36	1.46	1.71	1.42	1.46	1.75	1.35	1.42	1.72	1.43	1.52	1.71	1.43	1.47

LIST OF FIGURES

Fig. 1. Example ensemble forecast (A) and corresponding probabilistic forecast (B). 43

Fig. 2. WRF domains used in this study. The 36-, 12-, and 4-km grids are bound by the red, blue, and black boxes, respectively. 44

Fig. 3. Flow chart describing the formation of the 48-member WRF ensemble. 45

Fig. 4. Illustration of ensemble-forecast-meteogram and Gaussian-based uncertainty models (above the meteogram). Individual ensemble members are indicated by the colored, dashed curves and the ensemble mean by the black curve. Each forecast time has a probability density function based on the Gaussian distribution. The grey curve indicates a Gaussian distribution that does not scale (GNS) during the forecast. The pink curve is a Gaussian distribution that scales with ensemble variance (GSEV), while the red curve scales with the ensemble mean (GSEM). 46

Fig. 5. Improvement in PIT histogram calibration for one year of hourly wind forecasts by using more PBL schemes in the ensemble for the raw ensemble distribution (R2-R8, Table 2) and bias-corrected ensemble distribution (RB2-RB8) uncertainty models. Improvement is based on reduction in deviation between bins of the PIT histogram (such as for the PIT histogram in Fig. 6). Larger improvement is better. 47

Fig. 6. PIT histograms indicating an under-dispersive ensemble for tests R8 and RB8. Under-dispersion occurs when observed events fall too often at or outside the extremes of the ensemble forecast distribution. Flatter PIT histograms are better. 48

Fig. 7. PIT histograms comparing the results of the three Gaussian-based distributions for the six- and 48-member ensembles, prior to and after bias correction. Labels are the test name (Table 2). Flatter PIT histograms are better (closer to the horizontal dashed line). 49

Fig. 8. Actual wind-speed forecast-error distribution about the ensemble mean in test GB8 for each of the four wind farms. 50

Fig. 9. Annual RMSE for each bias-corrected ensemble member, the six-member ensemble in test GB1, and the 48-member ensemble in test GB8, averaged over all four wind sites. Ensemble members are named under the convention [PBL][IC][Grid (km)]. Smaller RMSE is better. . . 51

Fig. 10. Reliability diagrams for tests RB1, RB8, GB1, GB8, and the GNS distribution dressed over the bias-corrected 12-km ACM2 deterministic forecast initialized off the GFS (ACM2 12km GNS) for event thresholds of 5, 15, and 20 m s^{-1}. Reliability curves closer to the diagonal 1:1 line (thick grey) are better (i.e., are more calibrated). The blue dashed line represents the climatological event threshold frequency. The red dashed line represents the no-skill line. The inset figure is the sharpness histogram and is a count of the number of occurrences the event threshold is forecast in each probability bin (from 0^{th} to 100^{th} percentile in 10% increments.) Sharper forecasts assign mostly 0^{th} and 100^{th} percentiles while non-sharp forecasts issue mainly 50^{th} percentile forecasts. 52

Fig. 11. CRPS skill scores for test RB1, RB8, GB1, GB8, the ACM2 12-km GNS, and the bias-corrected deterministic 12-km ACM2 forecast initialized off the GFS. Skill is calculated relative to the worst performing deterministic forecast, the 36-km QNSE scheme initialized with the NAM. Larger CRPS Skill Score is better. 53

Fig. 12. Typical 30-day training period past error distribution (vertical bars) fit by the pq probability distribution (smooth curve). Bins are every 0.2 m s^{-1}. 54

Fig. A1. Annual MAE skill score (MAESS), showing improvement in MAE resulting from bias correction. Skill is relative to the equivalent raw wind forecast. Colors represent individual wind farm sites, with purple bars indicating the site-averaged performance. Larger positive values are better. 55

Fig. A2. Effect of the training period length on forecast accuracy (MAESS), averaged over all 48 bias-corrected deterministic wind forecasts at each wind-farm site. 56

Fig. B1. (a) Relative error (contoured, arbitrary units) between the observed and pq distributions as a function of parameters p and q. X marks the minimum error; namely, the best p and q values. S was fixed apriori at 10 m s^{-1} wind-forecast error, and p, q, and x_o were varied to find the distribution errors. (b) Resulting best-fit pq distribution (curve) to the observed histogram of wind-speed forecast errors (vertical bars). Note that this fitting method of minimizing the mean absolute error does not overly weight the tails of the distribution. 57

Fig. B2. Contour plots of (a) MAD (dashed line) and kurtosis (solid line), and (b) variance (dashed line) and kurtosis (solid line) for the pq distribution as a function of the p and q shape parameters. 58

Fig. B3. Some extreme examples of the pq distribution. 59

Fig. B4. Sample of some of the shapes produced by the pq distribution. The last shape is approximately Gaussian. 60

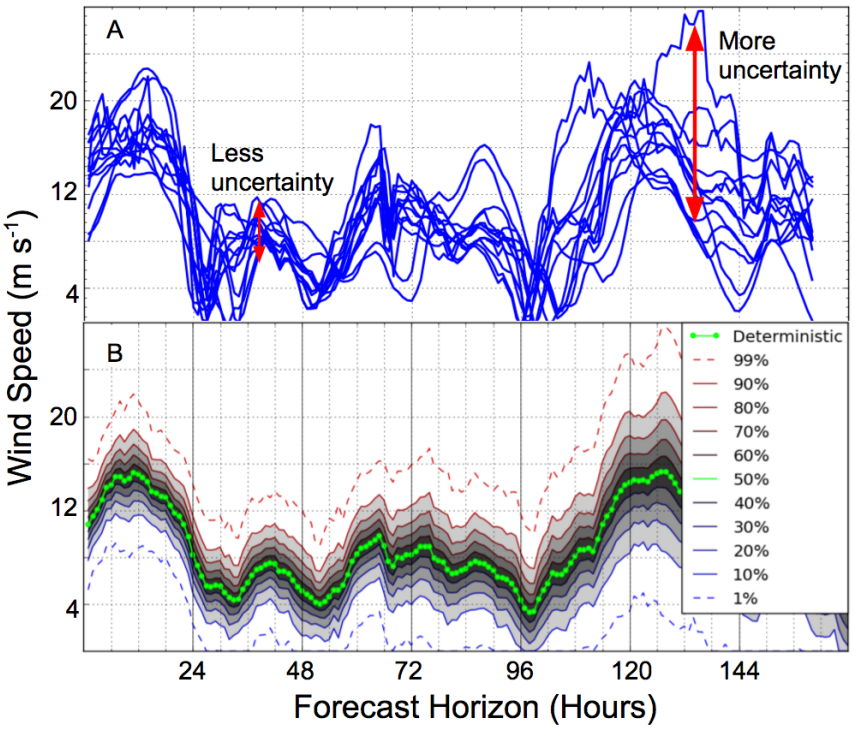
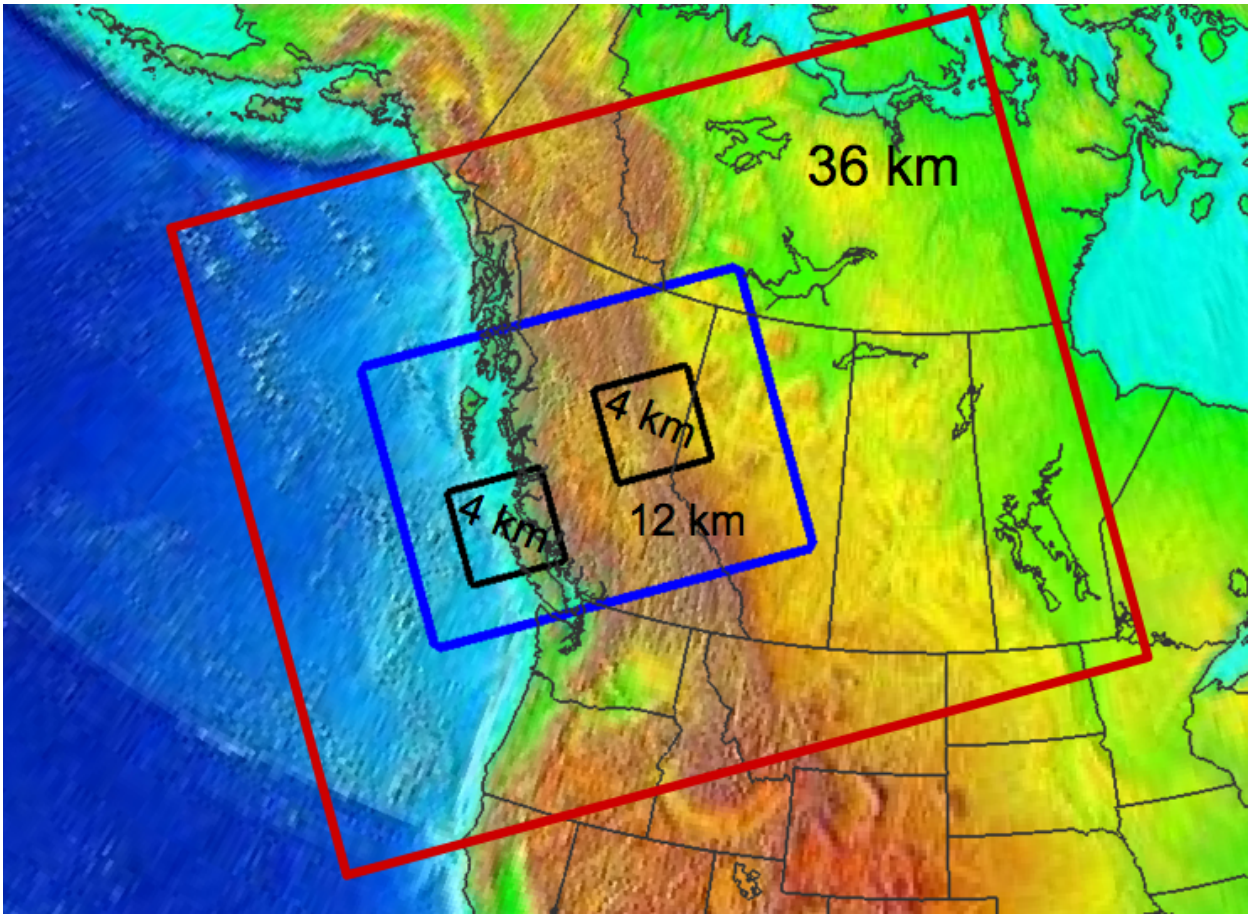


FIG. 1. Example ensemble forecast (A) and corresponding probabilistic forecast (B).



826 FIG. 2. WRF domains used in this study. The 36-, 12-, and 4-km grids are bound by the red, blue, and black
827 boxes, respectively.

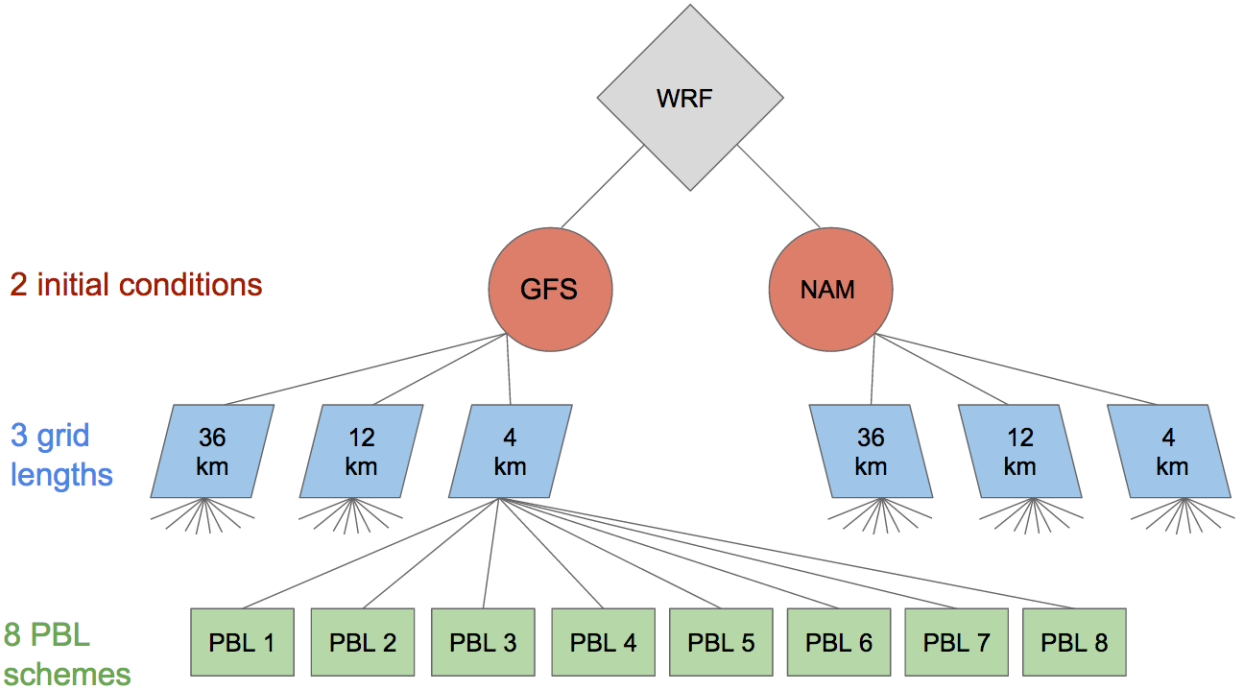
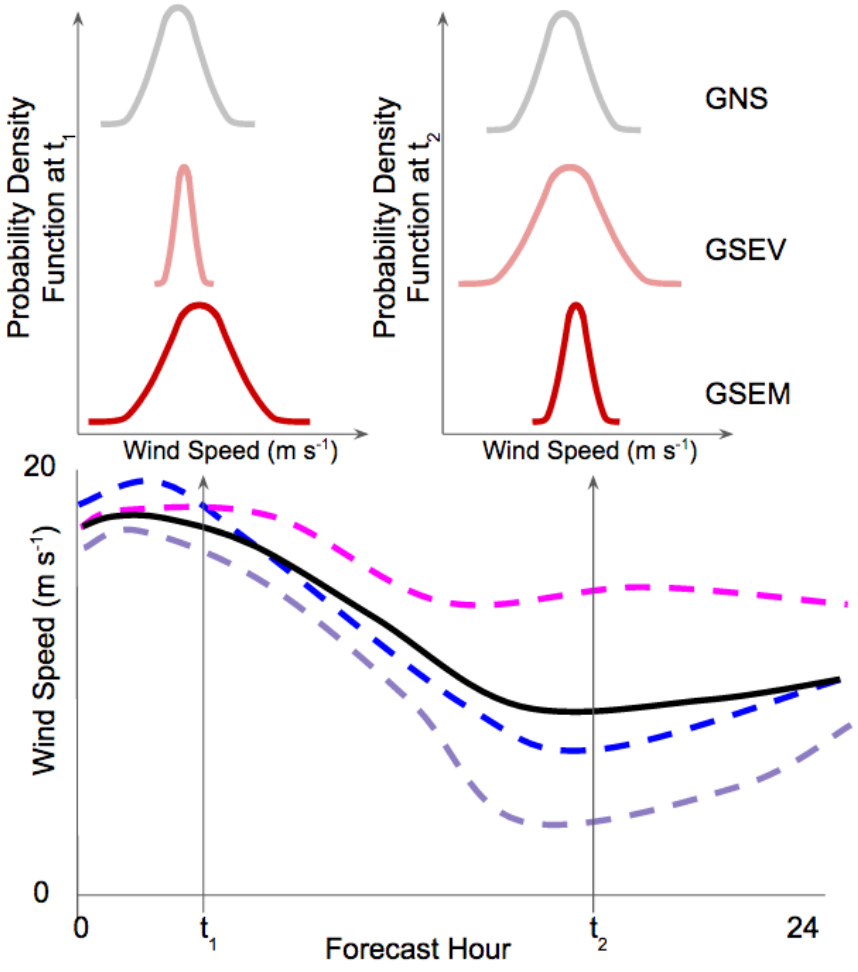
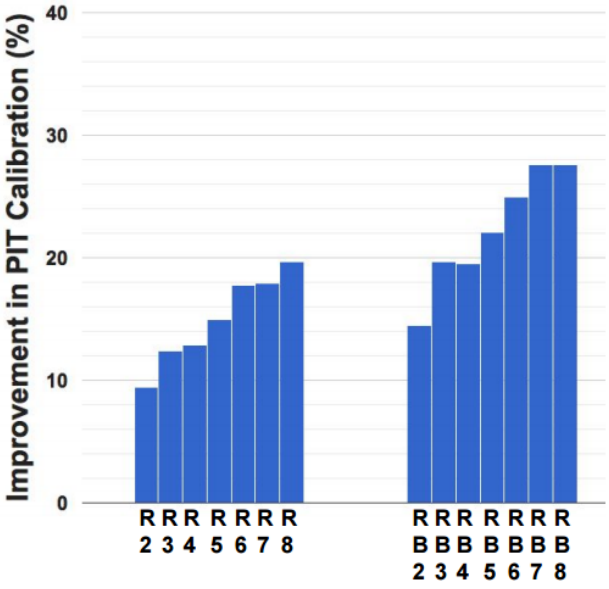


FIG. 3. Flow chart describing the formation of the 48-member WRF ensemble.



828 FIG. 4. Illustration of ensemble-forecast-meteogram and Gaussian-based uncertainty models (above the me-
829 teogram). Individual ensemble members are indicated by the colored, dashed curves and the ensemble mean by
830 the black curve. Each forecast time has a probability density function based on the Gaussian distribution. The
831 grey curve indicates a Gaussian distribution that does not scale (GNS) during the forecast. The pink curve is a
832 Gaussian distribution that scales with ensemble variance (GSEV), while the red curve scales with the ensemble
833 mean (GSEM).



834 FIG. 5. Improvement in PIT histogram calibration for one year of hourly wind forecasts by using more
835 PBL schemes in the ensemble for the raw ensemble distribution (R2-R8, Table 2) and bias-corrected ensemble
836 distribution (RB2-RB8) uncertainty models. Improvement is based on reduction in deviation between bins of
837 the PIT histogram (such as for the PIT histogram in Fig. 6). Larger improvement is better.

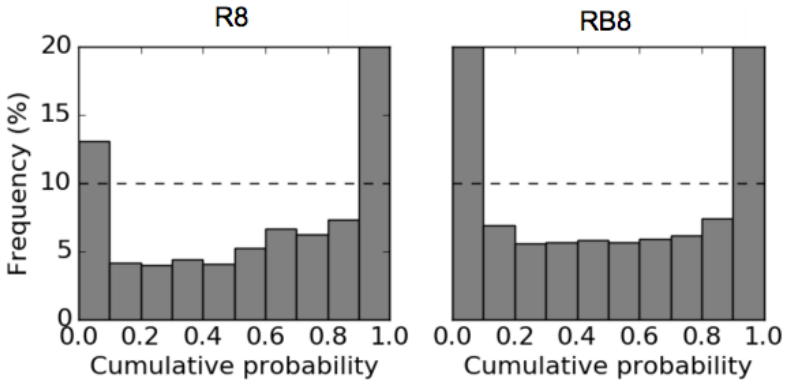
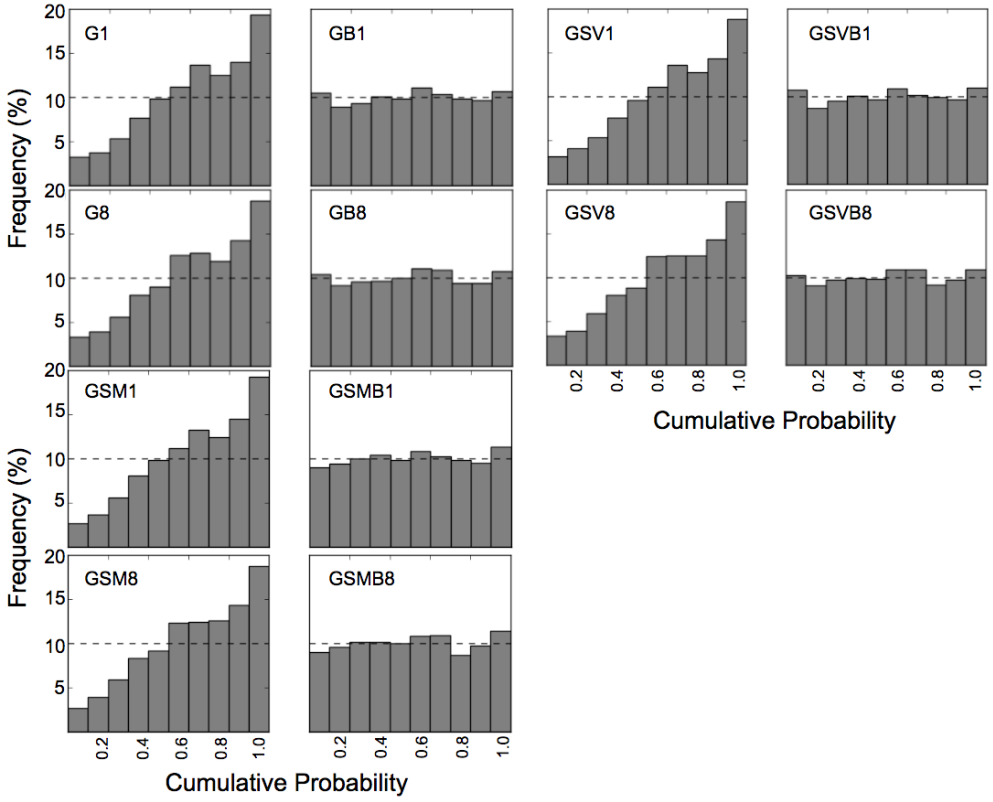
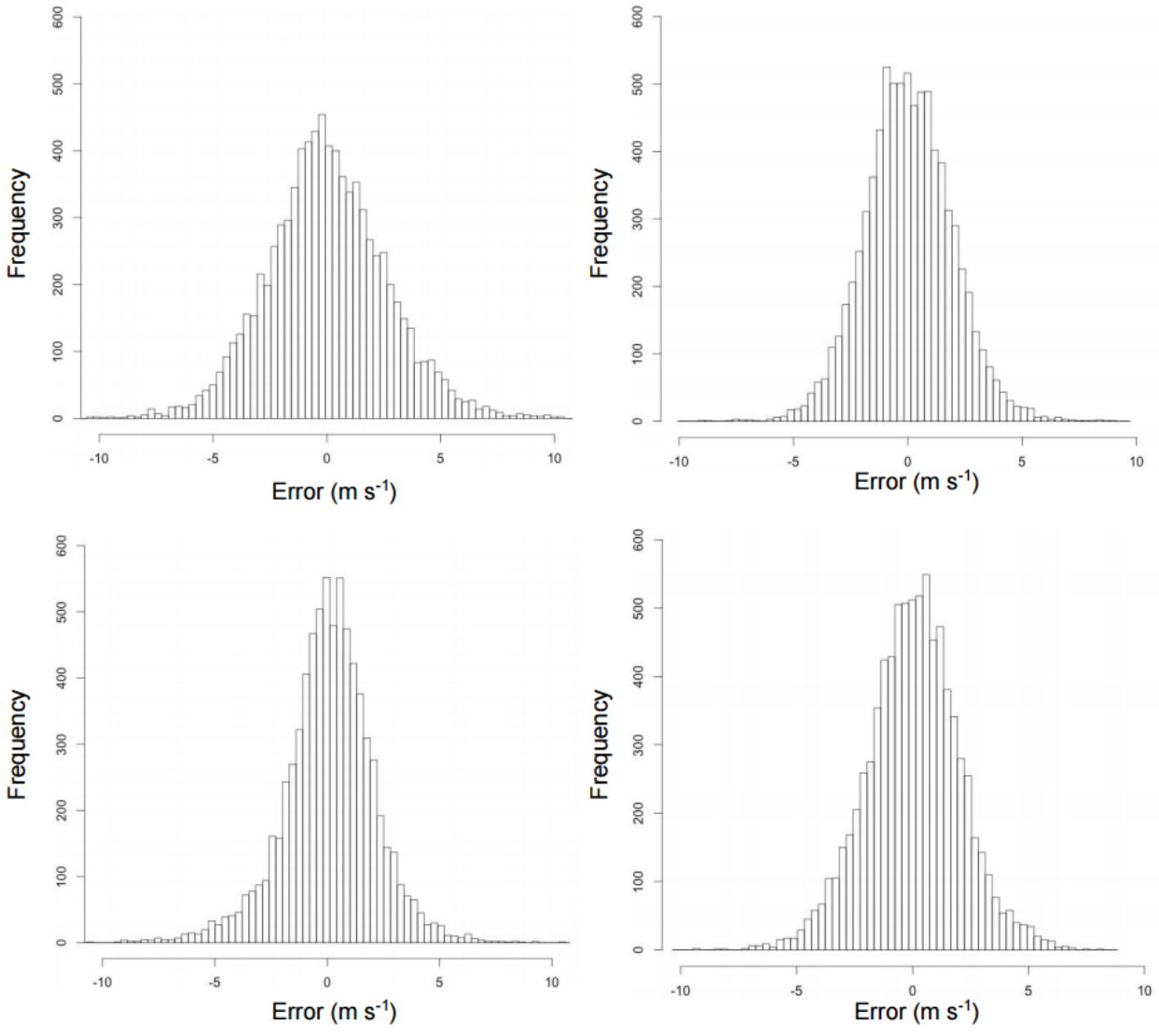


FIG. 6. PIT histograms indicating an under-dispersive ensemble for tests R8 and RB8. Under-dispersion occurs when observed events fall too often at or outside the extremes of the ensemble forecast distribution. Flatter PIT histograms are better.



841 FIG. 7. PIT histograms comparing the results of the three Gaussian-based distributions for the six- and 48-
842 member ensembles, prior to and after bias correction. Labels are the test name (Table 2). Flatter PIT histograms
843 are better (closer to the horizontal dashed line).



844 FIG. 8. Actual wind-speed forecast-error distribution about the ensemble mean in test GB8 for each of the
845 four wind farms.

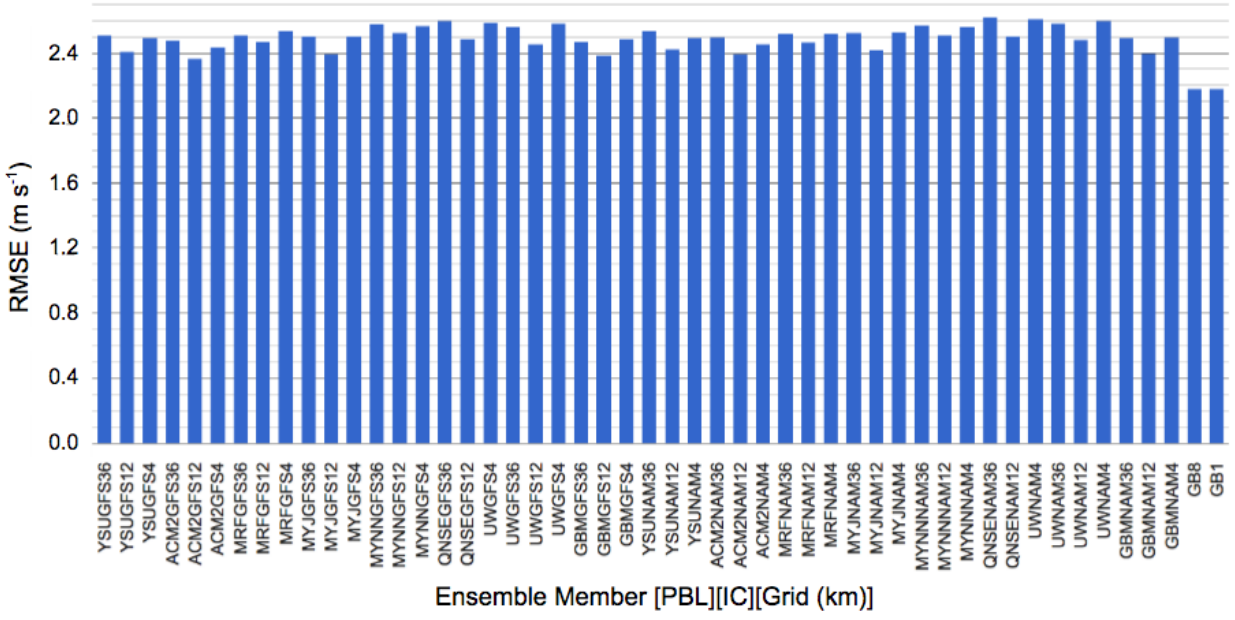


FIG. 9. Annual RMSE for each bias-corrected ensemble member, the six-member ensemble in test GB1, and the 48-member ensemble in test GB8, averaged over all four wind sites. Ensemble members are named under the convention [PBL][IC][Grid (km)]. Smaller RMSE is better.

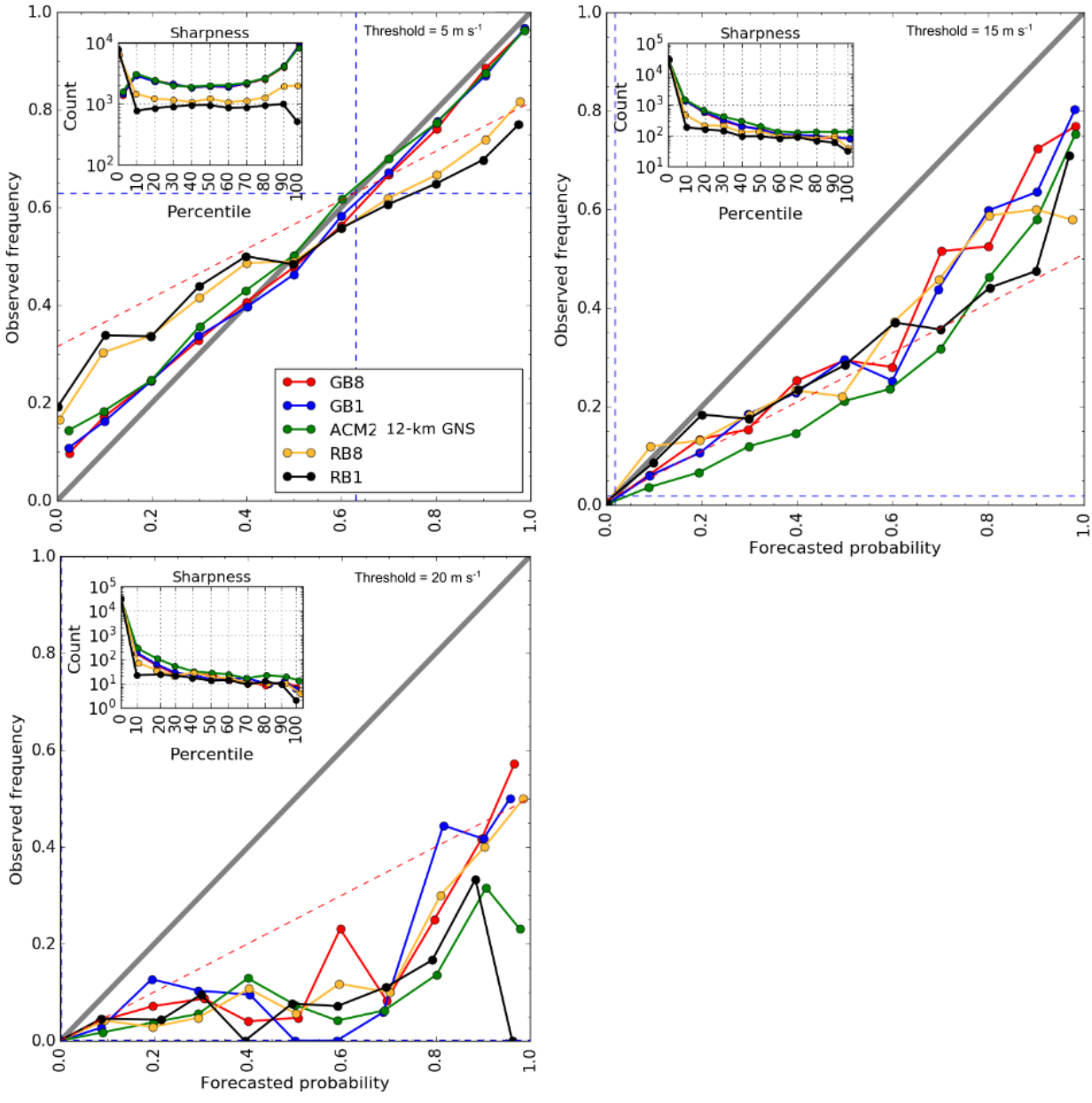


FIG. 10. Reliability diagrams for tests RB1, RB8, GB1, GB8, and the GNS distribution dressed over the bias-corrected 12-km ACM2 deterministic forecast initialized off the GFS (ACM2 12km GNS) for event thresholds of 5, 15, and 20 m s^{-1}. Reliability curves closer to the diagonal 1:1 line (thick grey) are better (i.e., are more calibrated). The blue dashed line represents the climatological event threshold frequency. The red dashed line represents the no-skill line. The inset figure is the sharpness histogram and is a count of the number of occurrences the event threshold is forecast in each probability bin (from 0^{th} to 100^{th} percentile in 10% increments.) Sharper forecasts assign mostly 0^{th} and 100^{th} percentiles while non-sharp forecasts issue mainly 50^{th} percentile forecasts.

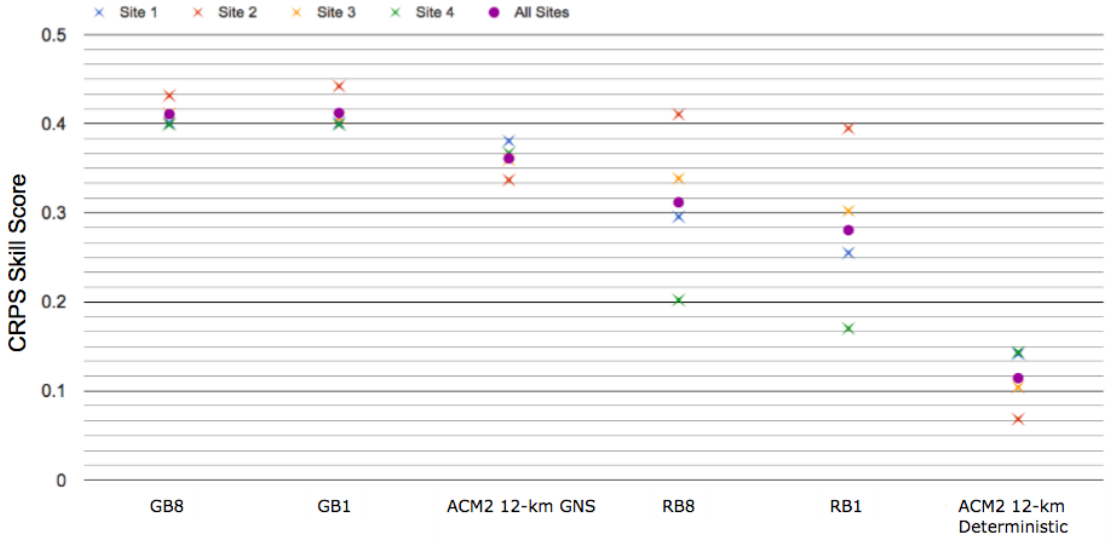
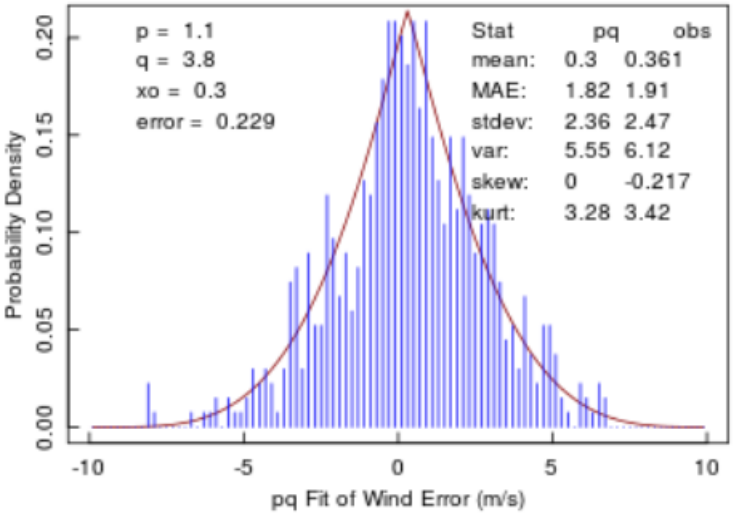
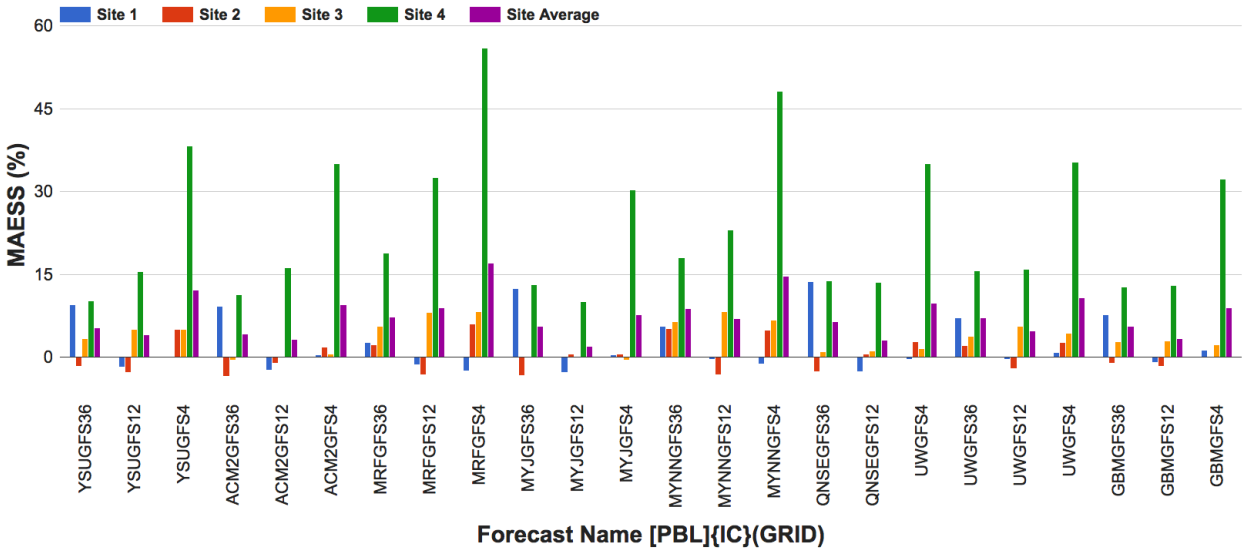


FIG. 11. CRPS skill scores for test RB1, RB8, GB1, GB8, the ACM2 12-km GNS, and the bias-corrected deterministic 12-km ACM2 forecast initialized off the GFS. Skill is calculated relative to the worst performing deterministic forecast, the 36-km QNSE scheme initialized with the NAM. Larger CRPS Skill Score is better.



860 FIG. 12. Typical 30-day training period past error distribution (vertical bars) fit by the pq probability distribu-
861 tion (smooth curve). Bins are every 0.2 m s^{-1}.



862 Fig. A1. Annual MAE skill score (MAESS), showing improvement in MAE resulting from bias correction.
863 Skill is relative to the equivalent raw wind forecast. Colors represent individual wind farm sites, with purple
864 bars indicating the site-averaged performance. Larger positive values are better.

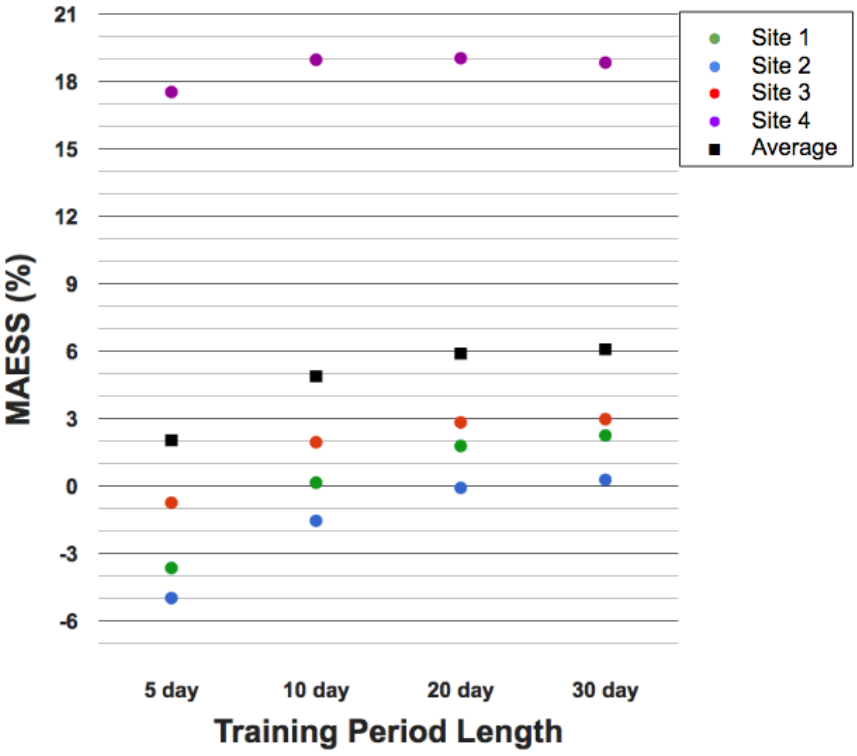


Fig. A2. Effect of the training period length on forecast accuracy (MAESS), averaged over all 48 bias-corrected deterministic wind forecasts at each wind-farm site.

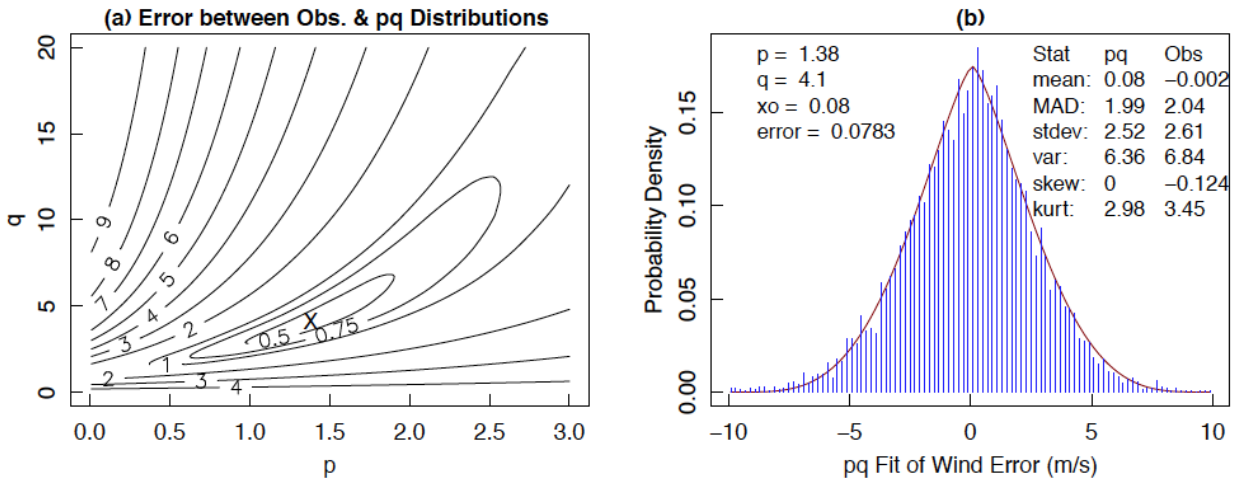
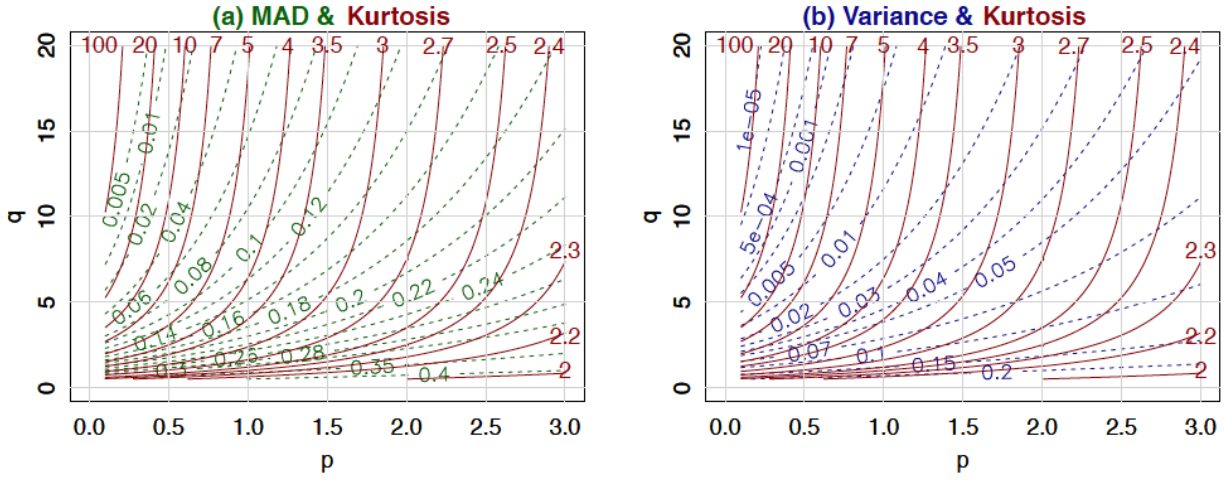


FIG. B1. (a) Relative error (contoured, arbitrary units) between the observed and pq distributions as a function of parameters p and q. X marks the minimum error; namely, the best p and q values. S was fixed apriori at 10 m s^{-1} wind-forecast error, and p, q, and x_o were varied to find the distribution errors. (b) Resulting best-fit pq distribution (curve) to the observed histogram of wind-speed forecast errors (vertical bars). Note that this fitting method of minimizing the mean absolute error does not overly weight the tails of the distribution.



872 FIG. B2. Contour plots of (a) MAD (dashed line) and kurtosis (solid line), and (b) variance (dashed line) and
873 kurtosis (solid line) for the pq distribution as a function of the p and q shape parameters.

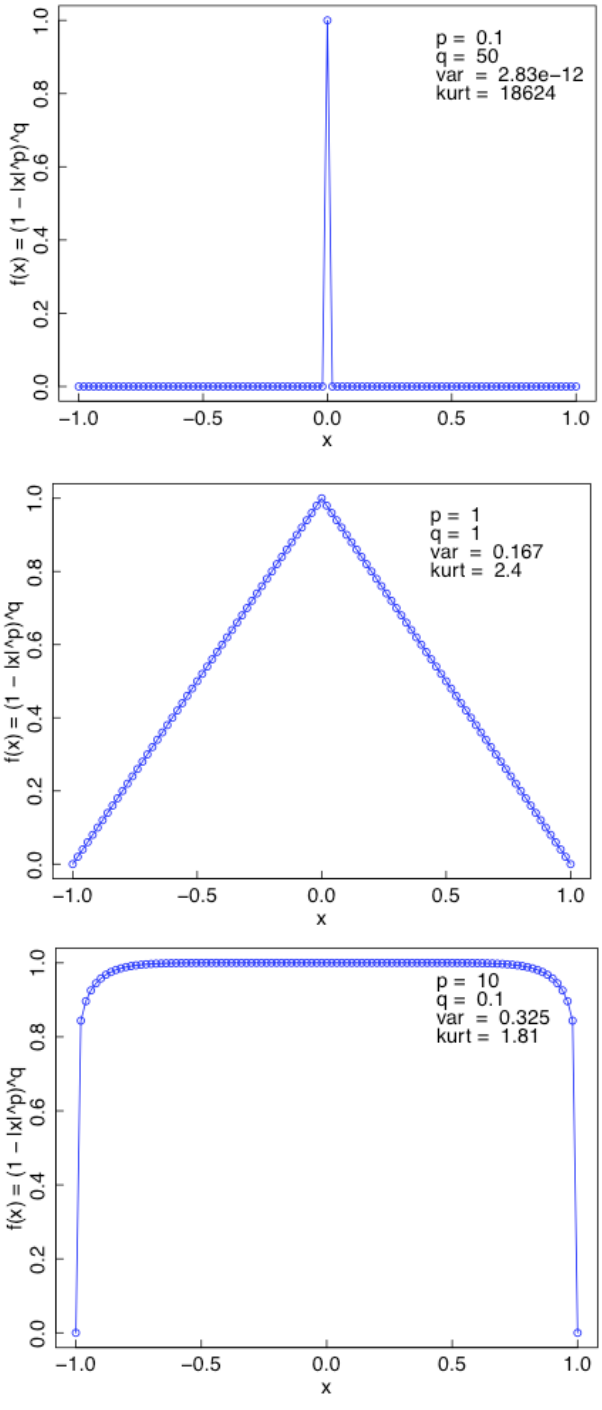
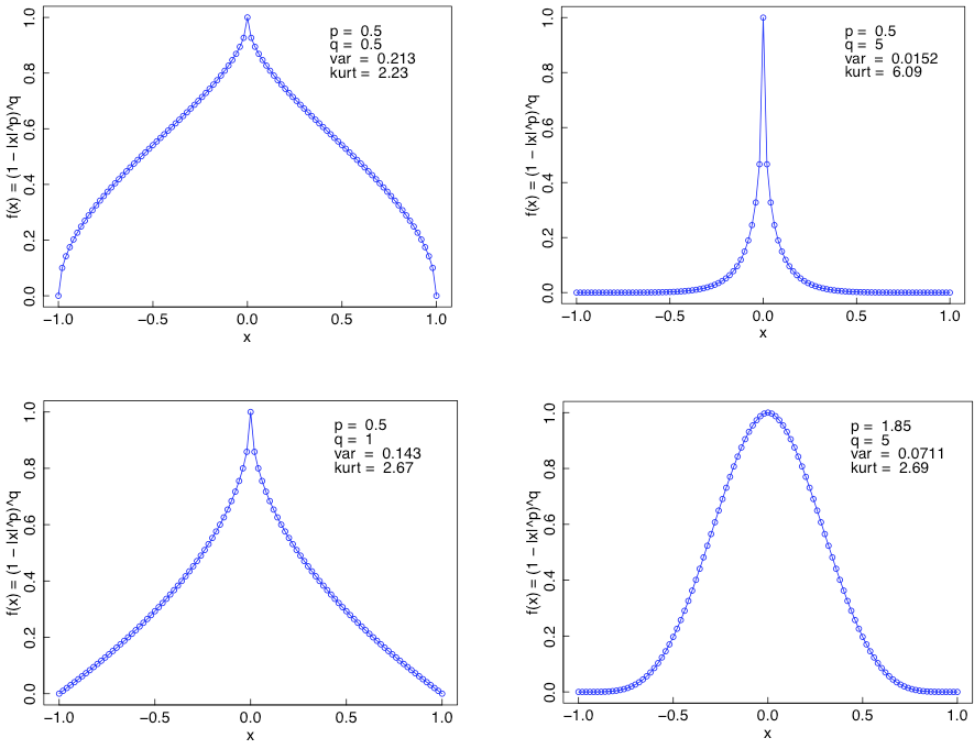


FIG. B3. Some extreme examples of the pq distribution.



874 FIG. B4. Sample of some of the shapes produced by the pq distribution. The last shape is approximately
875 Gaussian.