

Assessment of Probabilistic Wind Forecasts at 100 m Above Ground Level Using Doppler Lidar and Weather Radar Wind Profiles

KAROLIINA HÄMÄLÄINEN, ELENA SALTIKOFF, AND OTTO HYVÄRINEN

Finnish Meteorological Institute, Helsinki, Finland

VILLE VAKKARI

Finnish Meteorological Institute, Helsinki, Finland, and Unit for Environmental Sciences and Management, North-West University, Potchefstroom, South Africa

SAMI NIEMELÄ

Finnish Meteorological Institute, Helsinki, Finland

(Manuscript received 7 June 2019, in final form 16 December 2019)

ABSTRACT

Modern society is very dependent on electricity. In the energy sector, the amount of renewable energy is growing, especially wind energy. To keep the electricity network in balance we need to know how much, when, and where electricity is produced. To support this goal, the need for proper wind forecasts has grown. Compared to traditional deterministic forecasts, ensemble models can better provide the range of variability and uncertainty. However, probabilistic forecasts are often either under- or overdispersive and biased, thus not covering the true and full distribution of probabilities. Hence, statistical postprocessing is needed to increase the value of forecasts. However, traditional closer-to-surface wind observations do not support the verification of wind higher above the surface that is more relevant for wind energy production. Thus, the goal of this study was to test whether new types of observations like radar and lidar winds could be used for verification and statistical calibration of 100-m winds. According to our results, the calibration improved the forecast skill compared to a raw ensemble. The results are better for low and moderate winds, but for higher wind speeds more training data would be needed, either from a larger number of stations or using a longer training period.

1. Introduction

Probabilistic weather forecasts can provide valuable support for decision-making in many fields (Mylne 2002; Buizza 2019), such as energy production, aviation and agriculture. For instance, a farmer may be interested in the risk of strong winds or precipitation. Wind forecasts are also essential for aviation safety during landing and take off. Recently, increasing interest in probabilistic weather forecasting has come from the energy sector. The growing installed capacity of renewable energy has

created a need for new types of weather services. While hydropower needs long-term precipitation data and can be used as energy reserves, solar and wind power are more dependent on prevailing weather conditions. Currently, the wind energy sector is growing rapidly. During the year 2017, the amount of new wind power installed globally was 52.5 GW, raising the total installed capacity to 539 GW (Global Wind Energy Council 2017).

Wind energy companies use wind forecasts to maximize their income in energy markets (Pinson et al. 2007). In deregulated energy markets they can suffer financial penalties for both underprediction and overprediction of wind power (Thorarinsdottir and Gneiting 2010). Traditional deterministic wind forecasts based on a single model output are useful, but with a probabilistic wind forecast, the user gets a better overall picture of the uncertainty related to the forecast. The need for proper

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

Corresponding author: Karoliina Hämmäläinen, karoliina.hamalainen@fmi.fi

wind forecasts is growing because wind parks are getting bigger and their influence on the power network is increasing. To keep the power network in balance, it is important to know when, where and how much electricity will be produced. Also, the price of electricity depends on how it has been produced (U.S. Energy Information Administration 2019; Weron 2014). Wind energy is cheap compared to fossil fuels, but the wind energy producer needs to make a bid to the energy market and provide an as good as possible estimate of how much electricity it can provide the next day. If that goal is not achieved, there is a penalty of paying the price between the estimate and the actual cost, as described in Bremnes (2004). Jónsson et al. (2010) have written a very thorough and detailed review of how wind forecasts affect energy markets.

Due to energy markets, wind power producers are usually interested in the 48-h prediction horizon. For the spot markets, even shorter nowcasting or observation-based statistical methods are needed (Giebel et al. 2011; Simon et al. 2018; Messner and Pinson 2019; Roohollah 2019). However, for more extensive planning of energy reservoirs, electricity network balancing, etc., longer-range forecasts are needed (Miettinen and Holttinen 2017; Zhu and Genton 2012).

Even though probabilistic wind forecasts bring additional value for decision-making compared to deterministic forecasts (Pinson et al. 2006; Gneiting and Katzfuss 2014; Buizza 2019), they are not flawless. Increasing the ensemble size is one way to cover more probabilities. However, in many cases the forecasts are biased and thus the probabilities are biased too. This leads to either overdispersive or underdispersive probabilistic forecasts. Thus, statistical calibration methods are needed to minimize the bias, yet keep the ensemble spread realistic. The spread is considered realistic when it is equal to RMSE, and hence can cover the average mean error of the ensemble.

Many publications such as Gneiting et al. (2007, 2005), Hamill et al. (2008), Thorarinsdottir and Gneiting (2010), and Hagedorn et al. (2008) present methods on how to do the calibration. Furthermore, Junk et al. (2014) compared different calibration methods for 100-m wind forecasts using measurements from high observation masts.

The availability of operational mast measurements from the height of wind turbine nacelles is low. Therefore, in this paper we have investigated how we could improve the wind forecasts at 100-m height by using statistical calibration methods combined with a new type of ground-based remote sensing observations. The calibration method (BCT) was chosen based on our previous experiences in the GLAMEPS consortium (GLAMEPS 2015), in which the chosen method has been used for the

calibration of the operational forecasting model. The method used in this paper has been found to be efficient in correcting the 10-m winds. The lidar network of the Finnish Meteorological Institute (FMI) includes four observation locations operating in research mode, meaning that the measurement data are not collected frequently enough to support an operational service. The radar network, on the other hand, is operational and its spatial coverage is good, but many of the observations are made higher above the ground than the level of interest, 100 m. Several months of data are needed for calibration, and then at least one month for verification. Due to the research status of the lidar network, and to the fact that the weather radars provide good wind data only when it rains, both observation time series were patchy. For this study, lidar measurements for the last three weeks of February 2016 were used for training the statistical models and the calibration and verification was done for the following month, March 2016. However, March 2016 was a dry month, so for radars, we choose the last three weeks of June 2016 for the training period and the calibration and verification were done for July 2016. This paper describes the promising results obtained from the comparison of hub height wind speed measurements from lidar and radar to model output for the calibration of ensemble predictions. The goal of this study is to test if these new types of observations could bring value when forecasting the 100-m winds essential for wind energy production. Due to a lack of high measurement masts and the good coverage of the radar network, these types of observations could be found to be useful in the future if the height of the wind towers continue to increase.

In this paper, in section 2 we present the forecasts used, the new types of observations and the calibration methods. Section 3 is dedicated to the results. Section 4 presents the conclusions based on the results and the final section 5, is dedicated to wider discussion and future plans.

2. Materials and methods

a. The NWP model

The Integrated Forecast System–Ensemble System (IFS-ENS) is an operational global ensemble prediction system of the European Centre for Medium-Range Weather Forecasts (ECMWF). The ensemble data with 50 perturbed members, together with a High-Resolution deterministic run (HRES), were extracted from the ECMWF archive. The spatial resolution for the ensemble was 0.25° , and 0.125° for HRES. The ensemble members have 91 levels, and HRES has 137 levels in the vertical (ECMWF 2017;

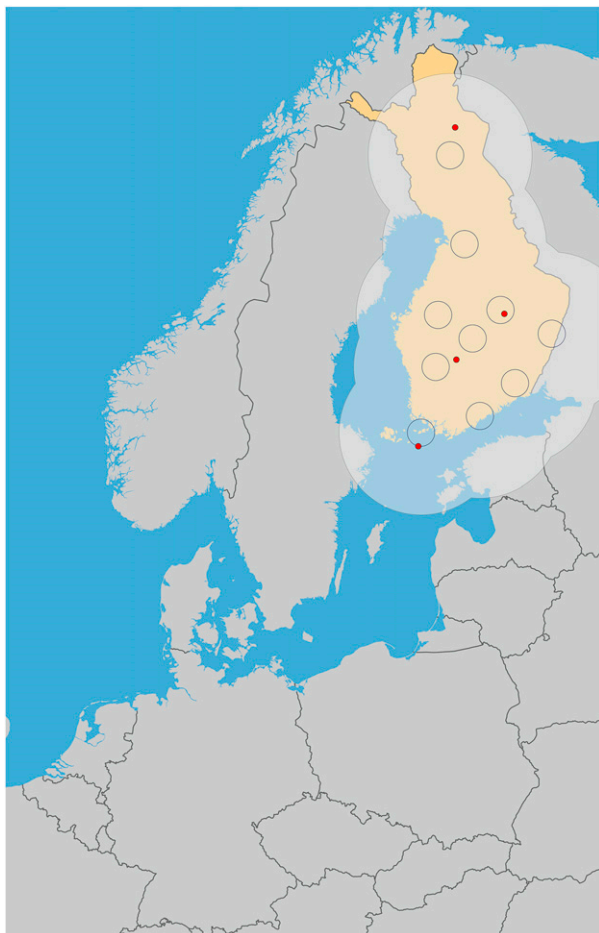


FIG. 1. The domain where IFS-ENS data were retrieved and calibrated. Red dots represent lidar stations. The locations of the 10 Finnish radars are marked with circles (radii 40 km). The shaded area describes the entire radar network coverage.

ECMWF 2016a,b). Figure 1 shows the area of interest for which the IFS-ENS data were extracted.

For each ensemble member, the fields of 100-m winds (U and V component) were used. Two years of ensemble data (2016–17) were extracted for the initial assessment. Only part of that data are used for the results presented in this paper, due to issues with observational data. One forecast run per day (started from 0000 UTC analysis) was extracted with 6-hourly intervals up to 15 days (+360 h). Higher temporal resolution would be preferred, but the ensemble data were only available with 6-h intervals in the ECMWF data archive. The lead times for 1–2 days and even up to 5 days are used for wind forecasting as well as for solar power. The longest lead times (up to 15 days) are important for hydropower and also to maintain the energy balance in the entire power network (consumption versus production and storage). However, this paper focuses only on wind power.

TABLE 1. Specifications for Halo Doppler lidars utilized in this study.

Wavelength	1.5 μm
Pulse repetition rate	15 kHz
Nyquist velocity	20 m s^{-1}
Sampling frequency	50 MHz
Velocity resolution	0.038 m s^{-1}
Points per range gate	10
Range resolution	30 m
Maximum range	9600 m
Pulse duration	0.2 μs
Lens diameter	8 cm
Lens divergence	33 μrad
Telescope	Monostatic optic-fiber coupled

b. Lidar wind profiles

Halo Photonics Streamline scanning Doppler lidars belonging to the Finnish Doppler lidar network (see Hirsikko et al. 2014), were used to measure wind profiles at four locations in Finland (Fig. 1): Utö (59.782°N, 21.372°E, 8 m MSL), Hyytiälä (61.847°N, 24.295°E, 179 m MSL), Vehmasmäki (62.738°N, 27.543°E, 190 m MSL), and Sodankylä (67.370°N, 26.630°E, 181 m MSL). Halo Photonics Streamline is a 1.5- μm pulsed Doppler lidar with a heterodyne detector (Pearson et al. 2009). At Utö and Hyytiälä we operated a full hemispherical scanning version of the lidar. At Vehmasmäki and Sodankylä we operated the Streamline Pro version, which has no moving parts on the outside and consequently the scanning is limited to a cone of 70°–90° elevation angle. All Doppler lidars were operated with the same specifications, which are given in Table 1.

The settings for vertical profiles of the velocity–azimuth display (VAD) scan were chosen for each site considering differences between the sites and requirements for other types of scanning in addition to horizontal wind retrievals. The VAD scan settings for each Doppler lidar site are summarized in Table 2. Of the sites, Utö is a rural marine location, Hyytiälä and Vehmasmäki are rural boreal forest sites, and Sodankylä is a subarctic forest site (Hirsikko et al. 2014).

All lidar measurements were postprocessed according to the Vakkari et al. (2019) algorithm to correct for uncertainty in signal-to-noise ratio (SNR). After the postprocessing, a SNR threshold correction of 0.005 (−23 dB) was applied to the data. Horizontal wind speed and direction were then obtained from VAD scans (i.e., conical scans at a fixed elevation angle). Wind retrievals were calculated from a sinusoidal fit (see Browning and Wexler 1968). In this study, we only utilize the wind speed at 100 m above ground level (AGL). Previously, VAD-based horizontal wind measurements with Doppler lidar have been found to compare well

TABLE 2. VAD scan settings for Doppler lidar sites.

Site	Elevation angle	N0. of azimuthal angles	Integration time per beam (s)	VAD schedule
Utö	15°	24	7	Every 15 min
Hyytiälä	30°	23	12	Every 30 min
Vehmasmäki	75°	24	5	Every 15 min
Sodankylä	75°	12	5	Every 10 min

with radar wind profiler and mast-based measurements (Päsche et al. 2015; Newsom et al. 2017).

The lidar data were available for 4 stations from 1 January to 10 October 2016. However, during a preliminary study of the observation quality, it was discovered that a large amount of data was missing from June and August. This resulted in poor quality training data that were not usable for the calibration of July and September. Due to good data quality during February and March 2016, this period was chosen to be tested. The three last weeks of February data were used for the training and to create calibration functions. These functions were then used for statistical correction of the forecast during March. The results of the calibrated ensemble forecast from March 2016 are hereby to represent the results of the lidar study.

c. Radar wind profiles

Doppler radars measure the component of wind parallel to the radar beam (known as “radial wind”) with high temporal and spatial accuracy. The resolution is even too fine for NWP data assimilation because the observations can reflect phenomena that are too small to be described in the model. Several methods have been developed to balance the disparate resolutions. Regional models typically use data thinning by pixel hopping or superobservations (averaging several neighboring pixels) (Salonen et al. 2009), while global models use data processed to vertical profiles of horizontal wind, by techniques called VAD (velocity–azimuth display) or VVP (velocity volume processing).

The VAD (Browning and Wexler 1968) and VVP methods (Waldteufel and Corbin 1979) were developed during the second half of the twentieth century. In both methods, radial wind measurements are collected from around the radar, organized as a function of azimuth and then a Fourier analysis is performed to find the wind speed (from the amplitude of the first Fourier component) and direction (from the phase of the first Fourier component), and some quality measures (from the higher Fourier components). The VVP method, which uses data from several elevation scans in a given volume, is considered more robust and stable than the VAD

method, although VVP basis functions are not inherently orthogonal (Boccippio 1995). VVP winds represent a potentially valuable source of lower-air data suitable for data analysis and other mesoscale meteorological applications (Bhowmik et al. 2011).

Holleman (2005) has assessed the suitability of VVP winds in the High Resolution Limited Area Model (HIRLAM) NWP model. The statistics of the difference between observations and model background of the VVP wind profiles were at least as good as those of the radiosonde profiles. The version of the model used in that study was hydrostatic, and had 31 vertical levels and a horizontal resolution of 55 km.

During the period of this study, FMI had 10 C-band weather radars (Fig. 1). The VVP profiles were generated using data from 8 elevation angles (10 elevations are measured), nominally 2–40 km from the radar. The two lowest measured elevation angles are excluded from the VVP calculations, as they have been noted to contain residual ground clutter, typically from objects that are moving, but not at wind speed, such as trees and masts swaying in the wind or rotating wind turbine blades. The actual height interval above sea level varies from radar to radar depending on the local topography and the height of the radar tower. It should also be noted that a weather radar measurement is not a point measurement, but includes data from an approximately cylinder-shaped slice in the atmosphere. The diameter of this cylinder depends on antenna properties and distance from the radar, and it is about 400 m at a range of 40 km. The nominal measurement height is the central point of this cylinder, and the efficiency decreases gradually to the edges following a Gaussian curve, indicated with blue shading in Fig. 2. The height intervals (from mean sea level) used are listed in Table 3.

The VVP products used in this study are created and calculated using the 3 lowest elevations with a radius smaller than 20 km around each radar, in slices of 200 m MSL (pink shading in Fig. 2). Thus, the “100 m data” actually include measurements from a 200-m-thick layer. The center of the beam at the lowest elevation angle (1.5°) used in this study reaches 200-m elevation at a range of 7.5 km and 400 m at 15 km, so the range of the used data is much smaller than the nominal value of 40 km, which is normally used to describe the metadata of the entire VVP profile.

Each wind value in our radar data was based on 3500–5000 measurement pixels, within the 200-m-thick volume around the radar. The pixel data were processed into single wind speed value and direction to represent a point measurement at the location of the radar. This was done by using Fourier analysis, as mentioned earlier in this chapter. In dry situations, many pixels contained no

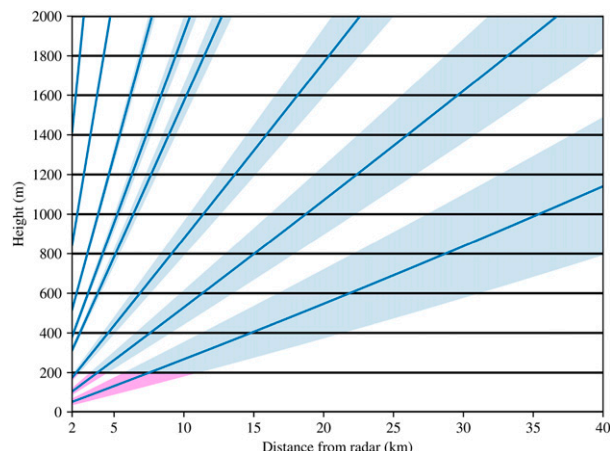


FIG. 2. Geometry of the FMI VVP profiles. Blue line indicates center and shading nominal 1° beamwidth of radar beam at 8 different elevations. The black horizontal lines are the border boxes, in which the data are processed. Only the data within the lowest box is used in this study (marked with pink shading).

data. We removed all wind values which were based on less than 100 valid data pixels. The pixel values were also used to calculate a standard deviation within the calculation volume. If the value was larger than 0.9 times the wind value, the value was disregarded. Both of these quality control measures were made to remove outliers. The purpose was to exclude those cases where the measured Doppler speed was not from particles moving with the wind (raindrops, snowflakes, insects, seeds) but from scatterers with their own velocity (migrating birds, airplanes or residual ground clutter). This method will also reject some cases dominated by mesoscale phenomena smaller than 80 km (thunderstorms, sea breeze).

Initial analysis was performed for data from 10 weather radar stations from 1 January 2016 to 31 December 2017. It showed that the quality and availability of the data are not homogenous. C-band weather radar cannot measure the movement of air, but the movement of some particles moving in the air. In the case of raindrops, snowflakes or even small insects, the speed of these particles represents the speed of the wind fairly well. In the case of birds or airplanes, the observed values are not the same as the wind speed, and in case of dry weather, there may be no data at all. For example, 16% of the Kuopio weather radar wind observations during the year 2016 were not available, either because of dry weather or because quality control had removed the speed estimates. Month-to-month differences were large, in March 45% of data were missing. March 2016 was a very dry month in the Kuopio area: monthly rainfall for the Kuopio Karttula weather station was 15 mm (while it was 64.1 mm in February). To have an equally long time series with

TABLE 3. Height (in m MSL) of the layer used to define the VVP winds from weather radars.

Ikaalinen	47–247
Anjalankoski	61–261
Kuopio	132–332
Korpo	139–339
Luosto	67–267
Utajärvi	82–282
Vantaa	118–318
Vimpeli	200–400
Kesälahti	26–226
Petäjävesi	129–329

lidars, July 2016 was selected as the verification period, and all 10 radars were used.

d. Verification metrics

The reliability of probabilistic forecasts reflects the average agreement between the forecast values and the measurements. For example, if the forecasted probability of wind speed exceeding a certain threshold is 20%, the observed wind speed should exceed the threshold in 20% of those cases. The reliability of the probabilistic wind forecast was investigated with the rank histogram (Wilks 2006). The rank histogram presents how well the ensemble spread can represent the actual variability of the observations. This is done by determining which bin of the sorted ensemble members of the forecast the observations fall into. If the distribution of binned observations in the histogram is uniform, the ensemble forecast is considered reliable. However, a *U*-shaped histogram indicates that the ensemble forecast is underdispersive, which means that the ensemble distribution is too narrow and many observations lay outside of the forecasted distribution. Furthermore, if the *U* shape is also asymmetric it indicates that the ensemble is biased. A bell-shaped histogram, on the other hand, indicates that the ensemble forecast is overdispersive, meaning that the distribution is too wide.

The quality of the probabilistic wind forecast was estimated using the Brier skill score (BSS), which measures the magnitude of the forecast error over the reference forecast (sample climatology), as described in Wilks (2006). The Briers skill score varies between $-\infty$ and 1. A positive result means that the model is more accurate than the reference. A 0 value means that the model has no skill, while negative values imply that the reference model is more accurate. The Brier skill score is calculated with the following formula:

$$\text{BSS} = 1 - \frac{\text{BS}}{\text{BS}_{\text{ref}}}, \quad (1)$$

$$\text{BS} = \frac{1}{N} \sum_{i=1}^N (P_i - O_i)^2. \quad (2)$$

The Brier score (BS) is defined in Eq. (2) (Brier 1950). It is the error of the probabilistic forecast computed over the verification sample. N is the number of forecasted cases. The variable P represents the probability of the forecasted event and varies between 0 and 1. The variable O represents the observations, and is given a value of either 1 (occurs) or 0 (does not occur). So, overall the Brier score has a value between 0 and 1, where 0 means perfect forecast.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^N (f_i - o_i)^2}{N}} \quad (3)$$

In addition, the accuracy of the forecast was evaluated by using root-mean-square error (RMSE) as given by Eq. (3). The bias (e.g., mean error) defined as the difference between the actual forecasted value f and the actual observation o . The main goal of ensemble calibration is to identify systematic errors in the ensemble forecast and try to reduce biases while keeping the ensemble spread within realistic limits, by widening or narrowing the ensemble spread. In an ideal probabilistic forecast, the ensemble spread should be as large as the RMSE, to be able to cover all the uncertainties.

e. Statistical calibration of IFS-ENS

A good ensemble prediction system should be able to cover all climatological possibilities, from normal weather situations to weather extremes, with minimal bias. In reality, this goal is not very often reached. Typically ensemble forecasts are underdispersive as our wind verification results show (see section 3). Hence, statistical calibration is needed to expand the ensemble spread and to minimize the mean error of the ensemble.

In this study, the Box–Cox t distribution (BCT) within the General Additive Model for Location, Scale and Shape (GAMLSS) was used (GLAMEPS 2015). The GAMLSS method is described in more detail in Rigby and Stasinopoulos (2005, 2006). The wind distribution is not Gaussian. It can never have negative values and the distribution is normally tilted to the left, favoring low and moderate winds. Thus the BCT better describes the wind distribution and its error distribution (Ioannidis et al. 2018).

The BCT is defined by four parameters representing the median (μ), variance (σ), skewness (ν), and kurtosis (τ) together with regression coefficients a – i :

$$\mu = a + b \times \text{MEAN} + c \times \text{ELEV}, \quad (4)$$

$$\sigma = \exp\{d + e \times \log(\text{STD}) + f \times \log[\max(1, \text{ELEV})]\}, \quad (5)$$

$$\nu = g + h \times \text{MEAN}, \quad (6)$$

$$\tau = \exp(i), \quad (7)$$

where MEAN stands for the ensemble mean, STD stands for the ensemble standard deviation, and ELEV refers to the model elevation above mean sea level. ELEV was chosen as an explanatory predictor to obtain a better variance in more heterogeneous terrain.

The statistical training of the model was done by comparing historical point observations against historical forecasts, at the observation stations points. Based on model wind and error distribution the calibration functions were adjusted by iterating the parameters (μ , σ , ν , and τ) to produce more reliable probabilities compared to the raw ensemble. The iteration was done by the GAMLSS-package in *R*. After the parameters had been estimated to generate the best correction functions the calibration took place. During the calibration these four functions [Eqs. (4)–(7)] were used to correct the forecasted 100-m wind distribution. Once the total distribution was set, the conversion back to members took place. The conversion was done by extracting evenly spaced quantiles from the distribution. The quantiles were then reordered to resemble the raw ensemble, resulting in coherent calibrated model fields.

For the calibration, training datasets (correction functions) were generated for each month separately. The training period was defined to be three weeks before the change of the month. Each training set was used during the calibration for the following month (please, see recommendations for operational usage in section 5). The training was done for each lead time, forecast cycle (0000 UTC), and parameter, independently.

3. Results

a. Model versus lidar

The lidar measurements from 4 stations were used in the calibration of the 100-m wind. The black lines in Fig. 3 show that the spread of the raw forecast is too narrow at the beginning of the forecast period. Also the RMSE increases as a function of the lead time. The positive outcomes of the calibration can be seen by comparing the black and yellow lines in Fig. 3. The calibration can widen the spread (line 1 in Fig. 3) and at the same time reduce the RMSE (line 2 in Fig. 3). Moreover, the wider spread stays in realistic limits by not exceeding the reduced RMSE. The overall

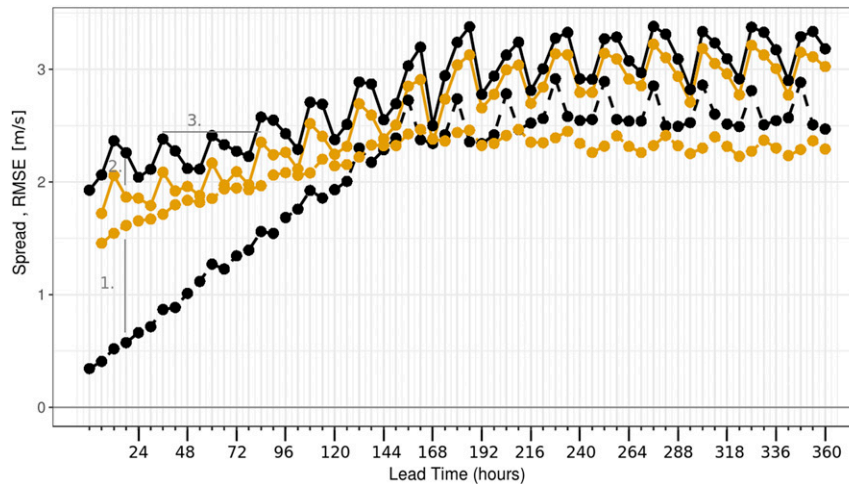
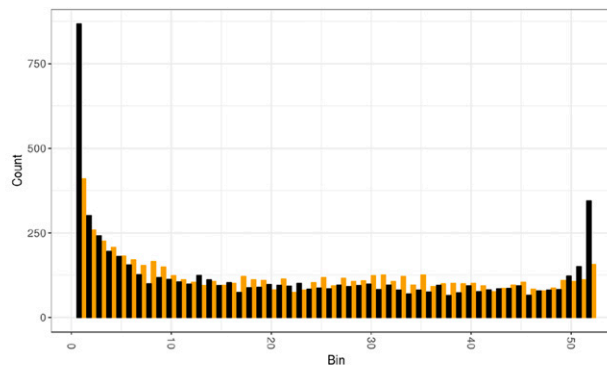


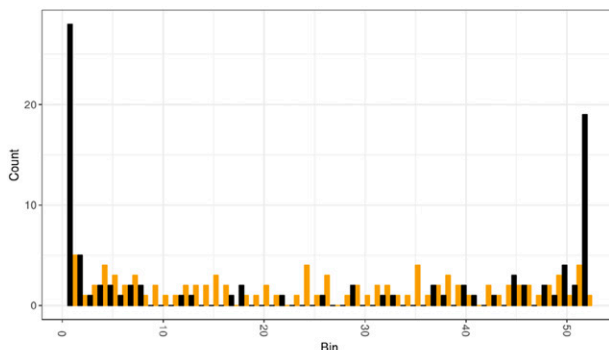
FIG. 3. Verification results for 100-m wind speed against lidar observations. Spread (dotted) and RMSE (solid) presented for raw ensemble model output (black) and calibrated dataset (yellow). The verification period is 1–31 Mar 2016. Lidar wind measurements from four Finnish stations are used for verification and calibration. Numbers 1–3 are explained in the text.

predictability is improved especially at the beginning of the forecast (line 3 in Fig. 3), which is also the most important part of the forecast from the wind energy point of view.

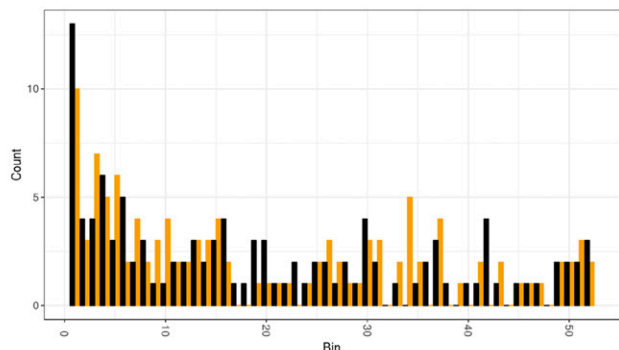
The fluctuation of RMSE between different lead times (seen in Fig. 3) is due to the local diurnal cycle. During the day time winds are stronger and the atmosphere is more convective, resulting in larger bias and



(a) All lead times



(b) Lead time +24h



(c) Lead time +132h

FIG. 4. Rank histogram of 100-m wind speed for the raw IFS-ENS (black) and ensemble calibrated with lidar observations (yellow). (a) All lead times, (b) +24 h, and (c) +132 h. Verification period is 1–31 Mar 2016.

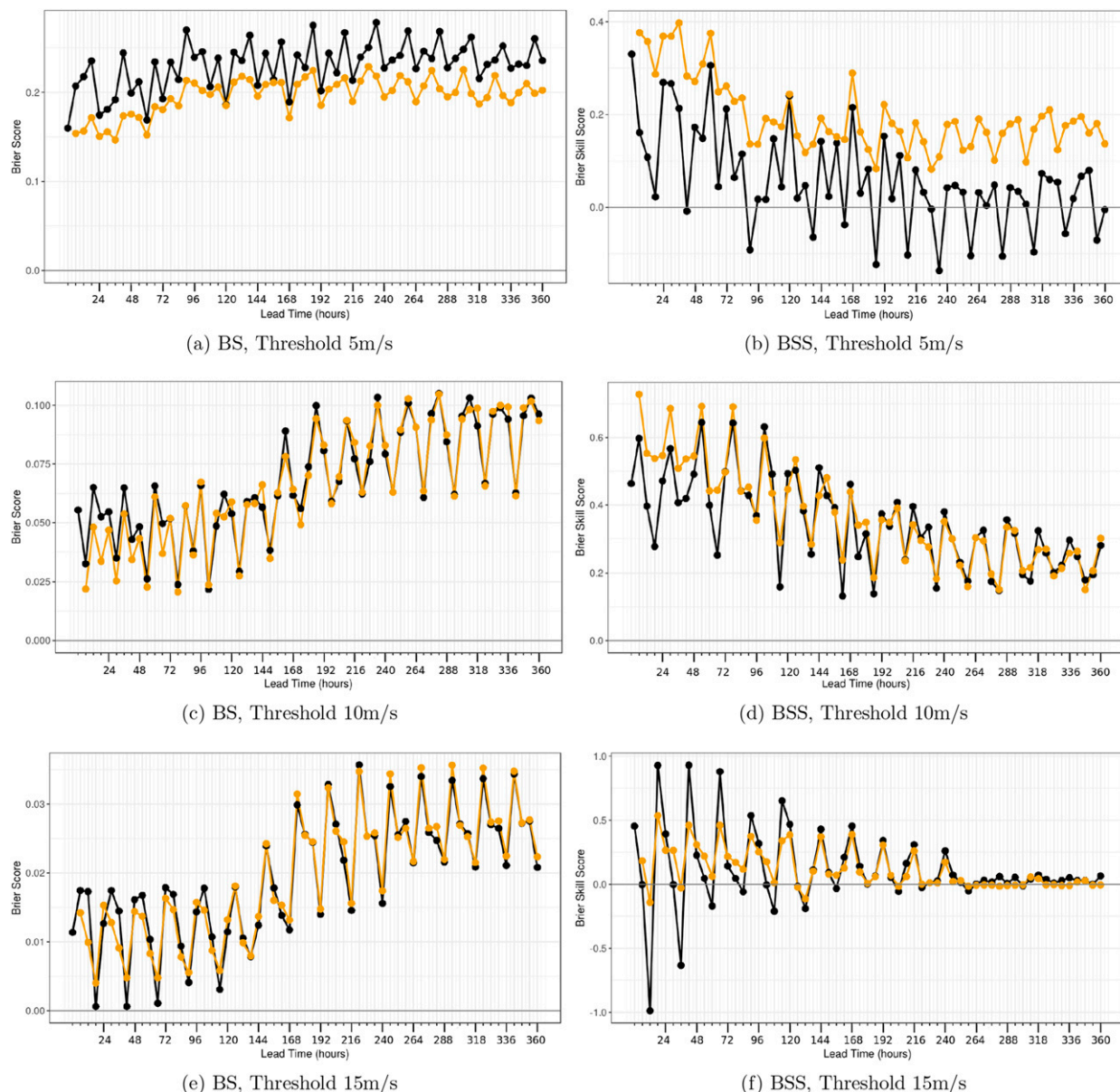


FIG. 5. (left) Brier scores and (right) Brier skill scores of 100-m wind speed for the threshold values (a),(b) 5 m s^{-1} , (c),(d) 10 m s^{-1} , and (e),(f) 15 m s^{-1} . The black lines represent raw IFS-ENS data, yellow lines the ensemble calibrated with lidars. The verification period is 1–31 Mar 2016.

RMSE. The other reason for fluctuation between the lead times is that we only used 0000 UTC (+2 h local time) in this study. More frequent model cycling (0000, 0600, 1200, and 1800 UTC) would smooth the diurnal cycle away from the verification results, due to different times of day for one lead time.

Figure 4 shows the rank histograms for the raw/uncalibrated 100-m wind forecast for IFS-ENS, compared against lidar observations. The shape of the rank histogram (Fig. 4a) indicates that the raw ensemble forecast is underdispersive and positively biased (leaning

more to the left side). Figure 4a shows that the larger spread of the calibrated dataset seen in Fig. 3 flattens the histogram, which is a desired feature of the ensemble forecast. Again, the improvement is larger at the beginning of the forecast. This is shown by comparing the outliers between the rank histograms for the lead times +24 h (Fig. 4b) and +132 h (Fig. 4c).

The Brier score and Brier skill score presented in Fig. 5 indicate that the calibration has the greatest effect on low wind speeds. With the low wind speed (threshold 5 m s^{-1}) the Brier score and Brier skill score are

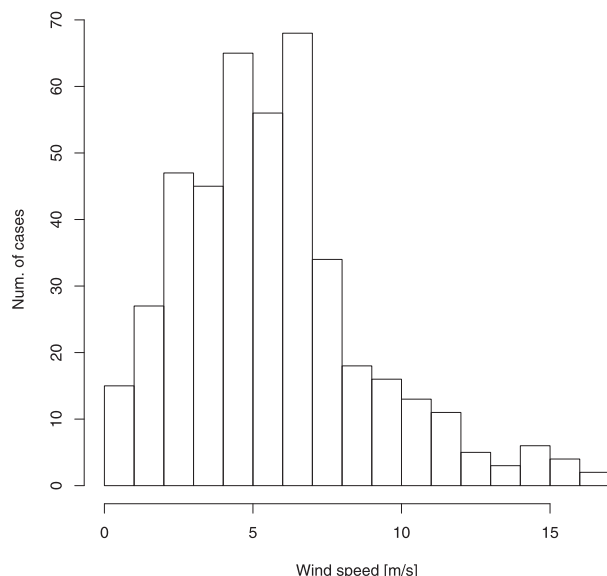


FIG. 6. Lidar wind speed (100-m height) categories for March 2016.

improved for all the lead times. However, with the higher wind speed threshold (10 m s^{-1}) the calibration seems to have a minor effect at the beginning of the model run. For the highest threshold (15 m s^{-1}) the model seems to have very low skill and especially after 10 days no skill at all. Figure 6 shows the number of wind speed observations from lidars as a function of wind speed. In this case, the reason for the poor performance of the calibration with a high wind speed threshold is the lack of observations. However, the results presented in Fig. 5f indicate (calibrated BSS mainly larger than 0, for the lead times up to 10 days) that the calibration could help to improve the results if more observations were available.

Similar fluctuations to those already seen with RMSE can also be identified from the Brier score and Brier skill score (see Fig. 5). This behavior is also due to the daily cycle. The Brier score is analog to bias in probability space. These probabilistic scores are usually interpreted with certain threshold values. Due to the diurnal cycle of wind it is more likely to exceed certain probability threshold values more often during different times of the day (in this study equal to lead time).

b. Model versus radar

Figure 7a shows the rank histograms for raw and radar-calibrated IFS-ENS datasets. These results indicate the same kind of underdispersive features already seen earlier in the lidar verification. However, when compared to radar observations the raw ensemble is negatively biased. The calibration can flatten the distribution and no outlier can be detected any longer. On

average this positive feature is seen until day 5 (Figs. 7b,c). This long-lasting skill to improve the forecast can also be seen in Fig. 8. The calibration method used is able to correct the spread and to reduce the RMSE. From Fig. 8, it can be seen that calibration can produce a spread that is the same magnitude as the RMSE. This is a feature of an optimal ensemble, in which the ensemble spread can cover all the uncertainties (Whitaker and Loughe 1998; Vannitsem et al. 2018).

As seen for the calibration with lidar observations, the lack of radar observations causes also challenges with the calibration of stronger wind speeds (Fig. 9). The Brier scores and Brier skill scores presented in Fig. 10 indicates that the calibration has a positive effect on weak and moderate wind speeds up to 10 m s^{-1} , but no skill for stronger winds exceeding 15 m s^{-1} . The improvement reaches until day 5 for the 5 and 10 m s^{-1} threshold values.

4. Conclusions

The results indicate that the calibration system described in this paper can improve the ensemble forecast when compared to the raw ensemble. The calibration results for weak and moderate wind speeds showed clear improvement. However, the results also state that the number of observations used is too small for stronger wind speeds for the calibration to further improve the scores.

The verification results indicate that the largest benefit is obtained within the first 3–5 days of the forecast. This finding is in line with previous studies such as Junk et al. (2014). In general, the 5-day weather forecasts have been found useful and representative as stated in Bauer et al. (2015).

The model bias compared to lidars is positive, but compared to radars negative. The average bias for the raw ensemble compared to lidar measurements was in the order of $0.5\text{--}1 \text{ m s}^{-1}$. The calibration was able to reduce the bias by 0.25 m s^{-1} (not shown in the figures). On the other hand, when comparing against radar winds the bias for the raw ensemble was in the order of -2 to -0.5 m s^{-1} . Within the first 5 days, the calibration was able to correct the bias to -0.5 to $+0.5 \text{ m s}^{-1}$. For the longer lead times (6–11 days) the calibration was overcorrecting and the bias was turned from negative to positive ($0\text{--}1 \text{ m s}^{-1}$). Overall the performance for the first 5 days was improved by using either of the observation types.

5. Discussion

The Box-Cox t distribution within the GAMLSS calibration system used in this study was built keeping

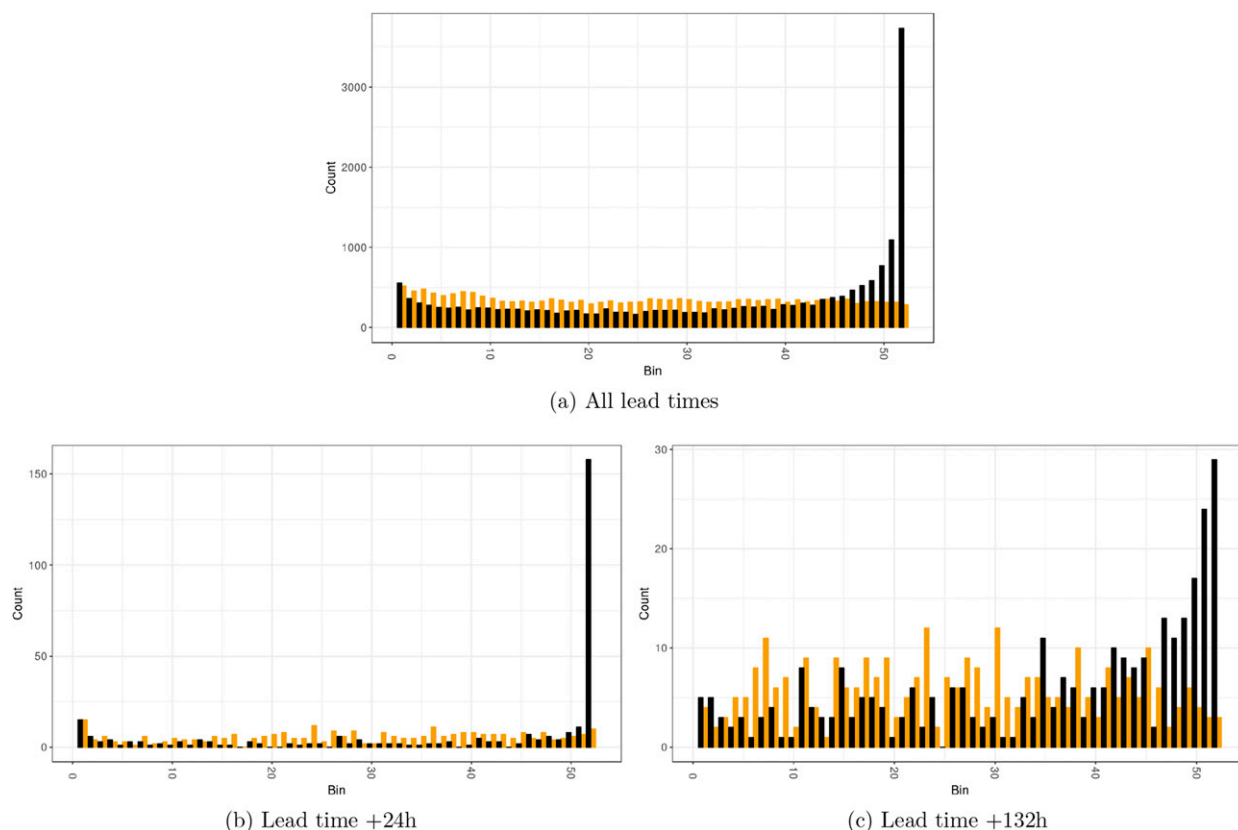


FIG. 7. Rank histogram for the raw IFS-ENS (black) and ensemble calibrated with radar observations (yellow). (a) All lead times, (b) +24 h, and (c) +132 h. The verification period is 1–31 Jul 2016.

possible operational usage in mind. Thus, the training period was chosen to be the past 21 days, instead of a huge amount of historical data being stored. The training set, including calibration coefficients, was created for each month separately by combining observation and forecast pairs. The coefficients were calculated for each lead time independently. The verification and calibration using lidar and radar wind data were done separately, because we wanted to test the suitability of each of these two observation types for such activities, before creating an operational flow of observational data.

Even though the calibration results using both lidar and radar data were good, the produced correction is in a different direction. Yet, both indicating that the raw ensemble was underdispersive, the bias was positive for the lidar and negative for the radar. This may be related to different measurement volumes of the instruments. A lidar measurement volume is in order of 100 m. When compared to model gridbox size they can be considered as point observations similar to a mast measurement (Tuononen et al. 2017). In our case there are two problems at the moment related to lidar measurements: 1) the number of stations for lidar measurements

is small and 2) the data are not collected and stored operationally/daily. On the other hand, the radar-based wind observations calculated with VVP represent the average wind over an observation volume with a radius of 10 km. Heterogeneous wind field inside of this volume, related to, for example, mesoscale weather phenomena, may add additional error sources. The positive aspects of the radar observations are (in our case): 1) the number of radar stations is larger, and 2) the data are available in real-time operationally. When comparing these two wind observation types against the model data, one can see that the RMSE is larger when using radar winds compared to results with lidar winds. This is natural due to the variability of wind at different scales reflecting the larger representative error in model-radar comparison. Furthermore, the quality of lidar wind observations can be more robust than radar-based wind measurements.

When comparing these two measurement types against the model the lidar verification was indicating overprediction, but the radar comparison was pointing underprediction. This may be related to the uneven distribution of the availability of radar data: a weather

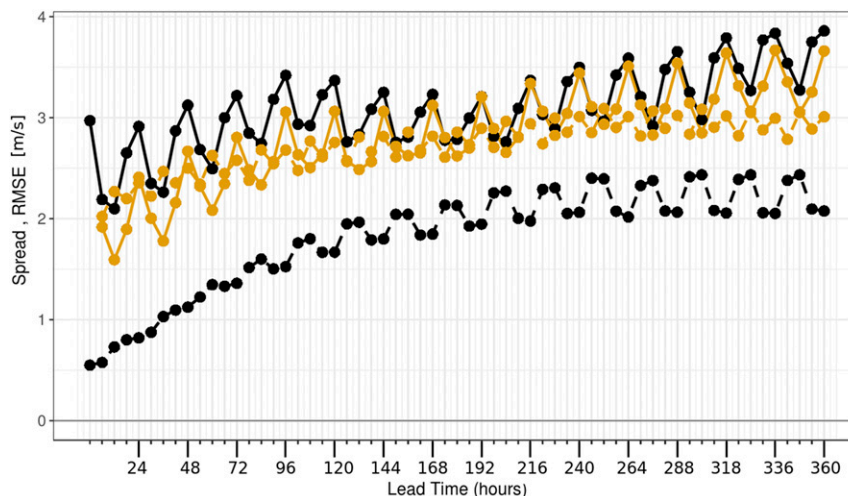


FIG. 8. Verification results for 100-m wind speed against radar observations. Spread (dotted) and RMSE (solid) presented for raw ensemble model output (black) and calibrated dataset (yellow). The verification period is 1–31 Jul 2016. Radar wind measurements from 10 Finnish stations are used for verification and calibration.

radar can only measure wind speed when there are particles in the air, not during calm, dry, high pressure situations. In the literature, the disparate volume of the model and observations has been discussed and studied mainly in the context of data assimilation as described in Salonen et al. (2009), Lindskog et al. (2004), and Gustafsson et al. (2018). The original resolution of radar data is of the order of 1 km or a few hundred meters, but the measurement of the radar is not wind that exists as a model forecast parameter, but just the component of it which is parallel to radar beam, known as radial wind. In variational data assimilation methods, the observation operator produces the model counterpart to the radial wind from the model horizontal wind U and V components at the observation location. Instead of a single observation, several pixels can be combined to a superobservation to better match model resolution. Lindskog et al. (2004) compared the approaches of assimilating Doppler weather radar data either as VAD or as superobservations in HIRLAM 3D-Var. The model used in that study had a horizontal resolution of 22 km and 31 vertical levels. The improvement of verification scores was similar for both approaches. In convective-scale models, assimilation of radial winds is used, but data thinning has replaced superobservations to avoid observation–error correlations (Gustafsson et al. 2018). It would be instructive to consider whether some of these methods could be useful for the calibration of probabilistic forecasts. The archived weather radar wind profiles used for this study were originally prepared for other users. In operational implementation, it would be worth testing different settings in the processing, such as

smaller integration volumes, perhaps even processing as sectors.

Before using this method as an operational forecast calibration system, the following practice should be adopted: 1) weekly update of the training coefficients, 2) a longer (1–3 months) training period, and 3) more frequent forecast cycling (0000, 0600, 1200, and 1800 UTC). The more frequent update ensures that the calibration adapts itself to changes in the model system (e.g., the upgrade of the operational model version). A longer training period window would increase the number of

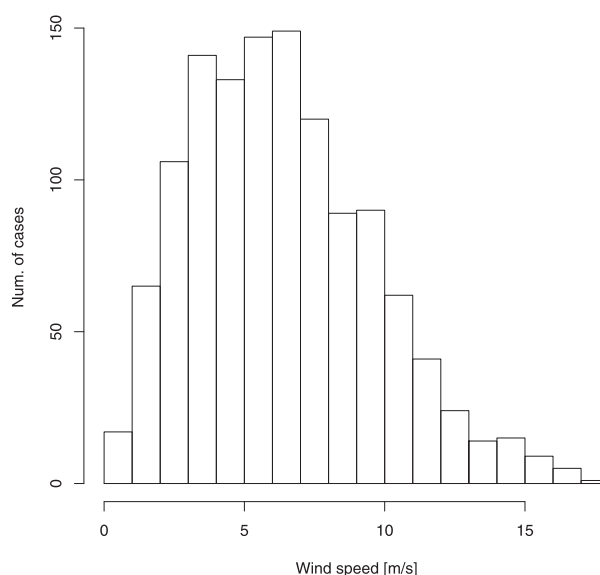


FIG. 9. Radar wind speed (100-m height) categories for July 2016.

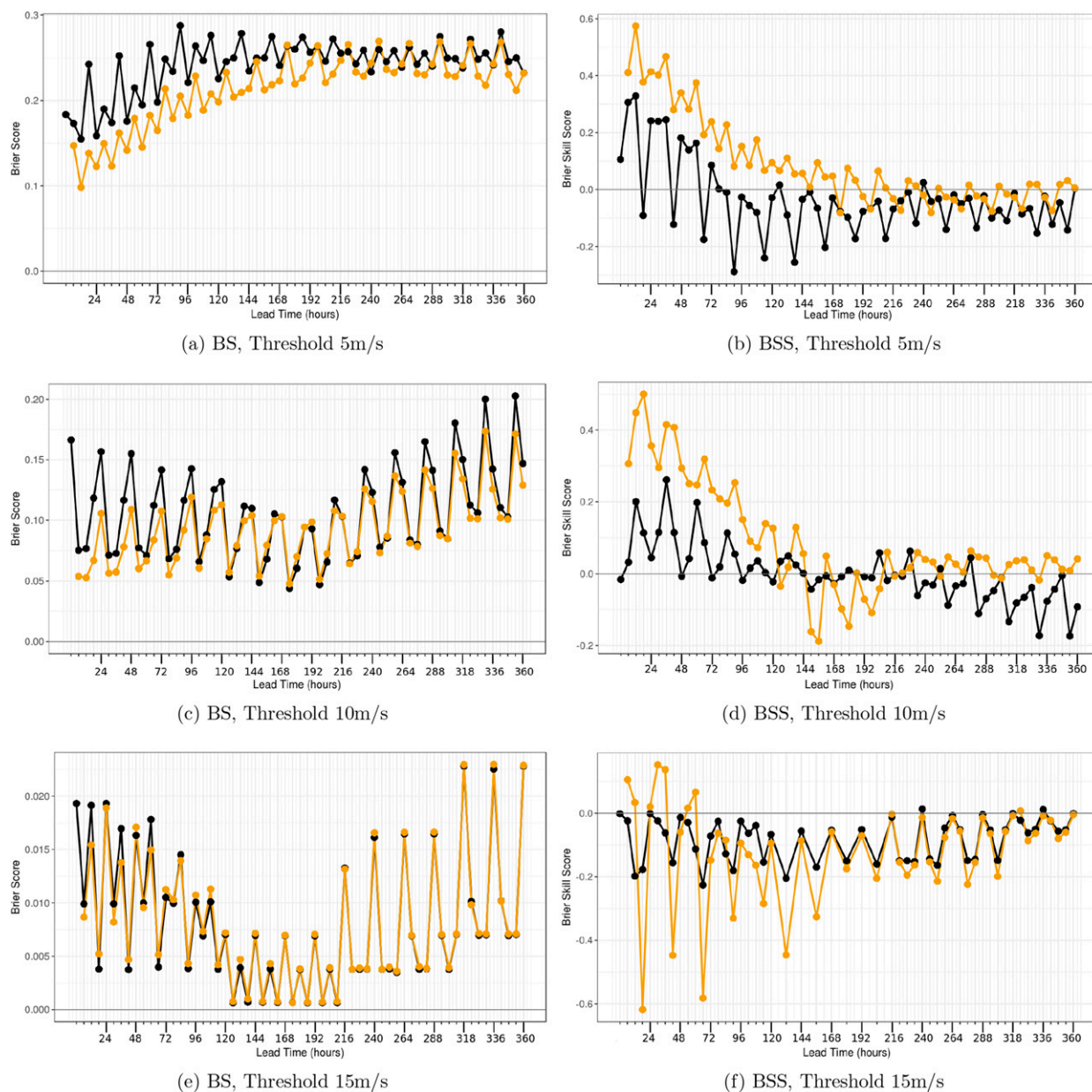


FIG. 10. (left) Brier scores and (right) Brier skill scores of 100-m wind speed for the threshold values (a),(b) 5 m s^{-1} , (c),(d) 10 m s^{-1} , and (e),(f) 15 m s^{-1} . The black lines represent raw IFS-ENS data, yellow lines the ensemble calibrated with radars. The verification period is 1–31 Jul 2016.

high wind speeds detected in the observations leading to better calibration results, including for strong wind speed cases. However, a longer training period requires larger data storage for the ensemble forecasts, which is expensive. An alternative approach to increase the volume of observations available for training to sufficiently cover the distribution of different weather situations would be to increase the number of observation stations.

FMI is a member of the MetCoOp collaboration together with SMHI (Sweden) and MET Norway (Müller et al. 2017). Together we run the operational

Harmonie-AROME EPS-system with 10 members (Frogner et al. 2019). In the future, we plan to use calibration methods for these high-resolution ensemble forecasts, to produce even more accurate probabilistic wind forecasts. We have already started testing some other parameters (temperature, precipitation, solar radiation) related to renewable energy production.

Acknowledgments. The study was mainly funded by Academy of Finland Project 284986 (VaGe) and partly by the Strategic Research Council, Finland, Project

292854 (BC-DC). We thank VaGe-project coordinators Hannele Holttinen and Juha Kiviluoma from VTT and other project partners for a fruitful conversation during the project. Related to international modeling consortium HIRLAM-C we want to send special thanks to Andrew Singleton and John Bjørnar Bremnes from the Norwegian Meteorological Institute for technical support and guidance, and a thank you to Jenna Ritvanen for drawing the VVP schematic.

REFERENCES

- Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, <https://doi.org/10.1038/nature14956>.
- Bhowmik, S. R., and Coauthors, 2011: Processing of Indian Doppler weather radar data for mesoscale applications. *Meteor. Atmos. Phys.*, **111**, 133–147, <https://doi.org/10.1007/s00703-010-0120-x>.
- Boccippio, D. J., 1995: A diagnostic analysis of the VVP single-Doppler retrieval technique. *J. Atmos. Oceanic Technol.*, **12**, 230–248, [https://doi.org/10.1175/1520-0426\(1995\)012<0230:ADAOV>2.0.CO;2](https://doi.org/10.1175/1520-0426(1995)012<0230:ADAOV>2.0.CO;2).
- Bremnes, J. B., 2004: Probabilistic wind power forecasts using local quantile regression. *Wind Energy*, **7**, 47–54, <https://doi.org/10.1002/we.107>.
- Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *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).
- Browning, K. A., and R. Wexler, 1968: The determination of kinematic properties of a wind field using Doppler radar. *J. Appl. Meteor.*, **7**, 105–113, [https://doi.org/10.1175/1520-0450\(1968\)007<0105:TDOKPO>2.0.CO;2](https://doi.org/10.1175/1520-0450(1968)007<0105:TDOKPO>2.0.CO;2).
- Buizza, R., 2019: Introduction to the special issue on “25 years of ensemble forecasting.” *Quart. J. Roy. Meteor. Soc.*, **145**, 1–11, <https://doi.org/10.1002/QJ.3370>.
- ECMWF, 2016a: Cycle 41r2 summary of changes. IFS documentation CY41r2, ECMWF, Reading, United Kingdom, accessed 24 February 2020, <https://www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model/cy41r2-summary-changes>.
- , 2016b: Cycle 43r1 summary of changes. IFS documentation CY43r1, ECMWF, Reading, United Kingdom, accessed 24 February 2020, <https://www.ecmwf.int/en/forecasts/documentation-and-support/evolution-ifs/cycles/cycle-43r1-summary-changes>.
- , 2017: IFS documentation. ECMWF, Reading, United Kingdom, accessed 24 February 2020, <https://www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model/ifs-documentation>.
- Frogner, I.-L., and Coauthors, 2019: HarmonEPS—The HARMONIE ensemble prediction system. *Wea. Forecasting*, **34**, 1909–1937, <https://doi.org/10.1175/WAF-D-19-0030.1>.
- 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 ed. ANEMOS.plus, 109 pp.
- GLAMEPS, 2015: GLAMEPS production system: Technical information for users. HIRLAM, NWP in Europe, accessed 24 February 2020, <http://hirlam.org/index.php/hirlam-programme-53/general-model-description/glameps>.
- Global Wind Energy Council, 2017: Annual market update 2017. Global Wind Report, Global Wind Energy Council, accessed 24 February 2020, <http://www.gwec.net>.
- Gneiting, T., and M. Katzfuss, 2014: Probabilistic forecasting. *Annu. Rev. Stat. Appl.*, **1**, 125–151, <https://doi.org/10.1146/annurev-statistics-062713-085831>.
- , A. E. Raftery, 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>.
- Gustafsson, N., and Coauthors, 2018: Survey of data assimilation methods for convective-scale numerical weather prediction at operational centres. *Quart. J. Roy. Meteor. Soc.*, **144**, 1218–1256, <https://doi.org/10.1002/qj.3179>.
- Hagedorn, R., T. M. Hamill, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Two-meter temperatures. *Mon. Wea. Rev.*, **136**, 2608–2619, <https://doi.org/10.1175/2007MWR2410.1>.
- Hamill, T. M., R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Mon. Wea. Rev.*, **136**, 2620–2632, <https://doi.org/10.1175/2007MWR2411.1>.
- Hirsikko, A., and Coauthors, 2014: Observing wind, aerosol particles, cloud and precipitation: Finland’s new ground-based remote-sensing network. *Atmos. Meas. Tech.*, **7**, 1351–1375, <https://doi.org/10.5194/amt-7-1351-2014>.
- Holleman, I., 2005: Quality control and verification of weather radar wind profiles. *J. Atmos. Oceanic Technol.*, **22**, 1541–1550, <https://doi.org/10.1175/JTECH1781.1>.
- Ioannidis, E., K. Whan, and M. Schmeits, 2018: Probabilistic wind speed forecasting using parametric and non-parametric statistical post-processing methods. KNMI Internal Rep. IR-2018-07, Royal Netherlands Meteorological Institute, Ministry of Infrastructure and Water Management, 78 pp., <http://bibliotheek.knmi.nl/knmipubIR/IR2018-07.pdf>.
- Jónsson, T., P. Pinson, and H. Madsen, 2010: On the market impact of wind energy forecasts. *Energy Econ.*, **32**, 313–320, <https://doi.org/10.1016/j.eneco.2009.10.018>.
- Junk, C., L. von Bremen, M. Kühn, S. Späth, and D. Heinemann, 2014: Comparison of postprocessing methods for the calibration of 100-m wind ensemble forecasts at off-and onshore sites. *J. Appl. Meteor. Climatol.*, **53**, 950–969, <https://doi.org/10.1175/JAMC-D-13-0162.1>.
- Lindskog, M., K. Salonen, H. Järvinen, and D. Michelson, 2004: Doppler radar wind data assimilation with HIRLAM 3DVAR. *Mon. Wea. Rev.*, **132**, 1081–1092, [https://doi.org/10.1175/1520-0493\(2004\)132<1081:DRWDAW>2.0.CO;2](https://doi.org/10.1175/1520-0493(2004)132<1081:DRWDAW>2.0.CO;2).
- Messner, J. W., and P. Pinson, 2019: Online adaptive lasso estimation in vector autoregressive models for high dimensional wind power forecasting. *Int. J. Forecasting*, **35**, 1485–1498, <https://doi.org/10.1016/j.ijforecast.2018.02.001>.
- Miettinen, J. J., and H. Holttinen, 2017: Characteristics of day-ahead wind power forecast errors in Nordic countries and benefits of aggregation. *Wind Energy*, **20**, 959–972, <https://doi.org/10.1002/we.2073>.
- Müller, M., and Coauthors, 2017: AROME-MetCoOp: A Nordic convective-scale operational weather prediction model. *Wea. Forecasting*, **32**, 609–627, <https://doi.org/10.1175/WAF-D-16-0099.1>.
- Mylne, K. R., 2002: Decision-making from probability forecasts based on forecast value. *Meteor. Appl.*, **9**, 307–315, <https://doi.org/10.1017/S1350482702003043>.

- Newsom, R. K., W. A. Brewer, J. M. Wilczak, D. E. Wolfe, S. P. Oncley, and J. K. Lundquist, 2017: Validating precision estimates in horizontal wind measurements from a Doppler lidar. *Atmos. Meas. Tech.*, **10**, 1229–1240, <https://doi.org/10.5194/amt-10-1229-2017>.
- Päschke, E., R. Leinweber, and V. Lehmann, 2015: An assessment of the performance of a 1.5 μm Doppler lidar for operational vertical wind profiling based on a 1-year trial. *Atmos. Meas. Tech.*, **8**, 2251–2266, <https://doi.org/10.5194/AMT-8-2251-2015>.
- Pearson, G., F. Davies, and C. Collier, 2009: An analysis of the performance of the UFAM pulsed Doppler lidar for observing the boundary layer. *J. Atmos. Oceanic Technol.*, **26**, 240–250, <https://doi.org/10.1175/2008JTECHA1128.1>.
- Pinson, P., J. Juban, and G. N. Kariniotakis, 2006: On the quality and value of probabilistic forecasts of wind generation. *2006 Int. Conf. on Probabilistic Methods Applied to Power Systems*, Stockholm, Sweden, IEEE, 1–7, <https://doi.org/10.1109/PMAPS.2006.360290>.
- , C. Chevallier, and G. N. Kariniotakis, 2007: Trading wind generation from short-term probabilistic forecasts of wind power. *IEEE Trans. Power Syst.*, **22**, 1148–1156, <https://doi.org/10.1109/TPWRS.2007.901117>.
- Rigby, R. A., and D. M. Stasinopoulos, 2005: Generalized additive models for location, scale and shape. *J. Roy. Stat. Soc.*, **54C**, 507–554, <https://doi.org/10.1111/j.1467-9876.2005.00510.x>.
- , and —, 2006: Using the Box-Cox t distribution in GAMLSS to model skewness and kurtosis. *Stat. Model.*, **6**, 209–229, <https://doi.org/10.1191/1471082X06st122oa>.
- Roohollah, A., 2019: Rapid Refresh Update Nowcasting with Harmonie-Arome. Norwegian Meteorological Institute MET Rep. 04/2019, 19 pp., https://www.met.no/publikasjoner/met-report/_attachment/download/0c336d8c-eff7-4915-9509-bbbfb1b5a198:682068250c751ca793587f7d94cc1d8208b9ed41/NOWWIND-H1-FINAL-REPORT.pdf.
- Salonen, K., H. Järvinen, G. Haase, S. Niemelä, and R. Eresmaa, 2009: Doppler radar radial winds in HIRLAM. Part II: Optimizing the super-observation processing. *Tellus*, **61**, 288–295, <https://doi.org/10.1111/j.1600-0870.2008.00381.x>.
- Simon, E., M. Courtney, and N. Vasiljevic, 2018: Minute-scale wind speed forecasting using scanning lidar inflow measurements. *Wind Energy Sci. Discuss.*, <https://doi.org/10.5194/wes-2018-71>.
- 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>.
- Tuononen, M., E. J. O'Connor, V. A. Sinclair, and V. Vakkari, 2017: Low-level jets over Utö, Finland, based on Doppler lidar observations. *J. Appl. Meteor. Climatol.*, **56**, 2577–2594, <https://doi.org/10.1175/JAMC-D-16-0411.1>.
- U.S. Energy Information Administration, 2019: Levelized cost and levelized avoided cost of new generation resources in the Annual Energy Outlook 2019. Independent Statistics and Analysis, U.S. Energy Information Administration, 25 pp., https://www.eia.gov/outlooks/aeo/pdf/electricity_generation.pdf.
- Vakkari, V., A. J. Manninen, E. J. O'Connor, J. H. Schween, P. G. van Zyl, and E. Marinou, 2019: A novel post-processing algorithm for Halo Doppler lidars. *Atmos. Meas. Tech.*, **12**, 839–852, <https://doi.org/10.5194/amt-12-839-2019>.
- Vannitsem, S., D. S. Wilks, and J. W. Messner, 2018: *Statistical Postprocessing of Ensemble Forecasts*. Elsevier, 362 pp.
- Waldteufel, P., and H. Corbin, 1979: On the analysis of single-Doppler radar data. *J. Appl. Meteor.*, **18**, 532–542, [https://doi.org/10.1175/1520-0450\(1979\)018<0532:OTAOSD>2.0.CO;2](https://doi.org/10.1175/1520-0450(1979)018<0532:OTAOSD>2.0.CO;2).
- Weron, R., 2014: Electricity price forecasting: A review of the state-of-the-art with a look into the future. *Int. J. Forecasting*, **30**, 1030–1081, <https://doi.org/10.1016/j.ijforecast.2014.08.008>.
- Whitaker, J. S., and A. F. Lough, 1998: The relationship between ensemble spread and ensemble mean skill. *Mon. Wea. Rev.*, **126**, 3292–3302, [https://doi.org/10.1175/1520-0493\(1998\)126<3292:TRBESA>2.0.CO;2](https://doi.org/10.1175/1520-0493(1998)126<3292:TRBESA>2.0.CO;2).
- Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. International Geophysics Series, Vol. 100, Academic Press, 648 pp.
- Zhu, X., and M. G. Genton, 2012: Short-term wind speed forecasting for power system operations. *Int. Stat. Rev.*, **80**, 2–23, <https://doi.org/10.1111/j.1751-5823.2011.00168.x>.