Improving the Analog Ensemble Wind Speed Forecasts for Rare Events

STEFANO ALESSANDRINI

National Center for Atmospheric Research, Boulder, Colorado

SIMONE SPERATI

Ricerca sul Sistema Energetico, Milano, Italy

LUCA DELLE MONACHE

Center For Western Weather and Water Extremes, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California

(Manuscript received 7 January 2019, in final form 26 April 2019)

ABSTRACT

An analog-based ensemble technique, the analog ensemble (AnEn), has been applied successfully to generate probabilistic predictions of meteorological variables, wind and solar power, energy demand, and the optimal bidding in the day-ahead energy market. The AnEn method uses a historical time series of past forecasts from a meteorological model or other prediction systems and observations of the quantity to be predicted. For each forecast lead time, the ensemble set of predictions is a set of observations from the past. These observations are those concurrent with the past forecasts at the same lead time, chosen across the past runs most similar to the current forecast. Recent applications have demonstrated that the AnEn introduces a conditional negative bias when predicting events in the right tail of the forecast distribution of wind speed, particularly when the training dataset is short. This underestimation increases when the predicted event occurs less frequently in the available historical data. A new bias correction for the AnEn using wind observations from more than 500 U.S. stations is tested to reduce the AnEn's underestimation of rare events. It is shown that the conditional negative bias introduced by the AnEn in its standard application is significantly reduced by our novel approach. Also, the overall probabilistic AnEn performances improve when predicting wind speed higher than 10 m s^{-1} as demonstrated by lower values of the continuous ranked probability score. These improvements can be attributed to an increased reliability achieved by introducing the proposed bias correction algorithm.

1. Introduction

The analog ensemble (AnEn) technique has been recently used to generate probabilistic predictions of 10-m wind speed and 2-m temperature (Delle Monache et al. 2013, hereafter DM13) starting from a deterministic meteorological forecast. The theoretical basis for the analog approach was provided by Hamill and Whitaker (2006) who used it to calibrate probabilistic predictions of 24-h accumulated precipitation from a numerical weather prediction (NWP) ensemble.

The AnEn uses a conditional sample of past observations (or analysis values) of the quantity to be predicted, appropriately chosen from a historical dataset, to build an ensemble forecast. Given a forecast at any location, the most similar forecasts are selected from a historical dataset, and the concurrent past observations are used to build the ensemble. An archive of NWP deterministic forecasts is used for that purpose, but other deterministic prediction systems (e.g., from statistical methods) can be used as well. The AnEn aims at identifying past meteorological conditions or weather regimes in which the error (defined as the difference between forecast and observation) probability density function (PDF) was similar. If those past analog conditions are adequately identified, then the past observations' errors can be used to infer the future error PDF, as they are samples from the same distribution.

Unlike NWP dynamical ensembles, there is no need for an initial perturbation strategy or stochastic physics

Corresponding author: Stefano Alessandrini, alessand@ucar.edu

DOI: 10.1175/MWR-D-19-0006.1

to initialize and run an NWP model multiple times. On the other hand, a historical dataset of NWP runs with the same configuration is necessary for the error PDF to remain similar. This requirement very often limits the possibility of building an archive longer than a few years. Despite using a reforecast dataset from 12 to 15 months long, DM13 demonstrated that the AnEn mean could outperform the ensemble mean of Environment and Climate Change Canada Regional Ensemble Prediction System (REPS) in terms of accuracy up to 48 h for 10-m wind speed and 2-m temperature. Also, the AnEn exhibited better probabilistic skills, such as statistical consistency and reliability than a calibrated version of REPS while using fewer computational resources.

Several studies have demonstrated the AnEn's adaptability to a wide range of applications. Here, of particular interest is its application to renewable energies. The AnEn predictions are usually naturally calibrated and reliable in the first 3 days, which is often the time horizon of interest for end users in the energy sector. Also, the AnEn can generate power predictions without using a power curve, which is a function used to compute solar or wind power from solar radiation or wind speed values. In fact, once the past analog dates are identified, the generated power at those dates can be used to build the ensemble forecast. Alessandrini et al. (2014, 2015a) and Junk et al. (2015) used the AnEn for generating hourly wind speed and wind power probabilistic predictions for up to 72 h ahead. Similar applications for solar power can be found in Alessandrini et al. (2015b), Cervone et al. (2017), and with an extension to regional power production in Davò et al. (2016). A more recent contribution (Sperati et al. 2017) used the AnEn to generate probabilistic predictions of 10-m wind speed over a two-dimensional grid using values from the analysis field as ground truth instead of measurements. They showed that up to 72 h the AnEn based on the European Centre for Medium-Range Weather Forecasts (ECMWF) deterministic model outperforms the ECMWF Ensemble Prediction System (ECMWF-EPS) over a domain slightly larger than the Italian peninsula, using a fraction of the computational resources necessary to generate ECMWF-EPS. Other AnEn applications include Alessandrini et al. (2018) for predictions of maximum intensity of tropical cyclones, Keller et al. (2017) for downscaling a reanalysis dataset of precipitation, Nagarajan et al. (2015) for AnEn applications using several variables and models, and Djalalova et al. (2015) for air quality predictions.

It is evident that the likelihood of finding good analogs increases with longer training datasets. As a matter of fact, the requirement of generating a reforecast dataset by an NWP with the same configuration often limits the available training length and some issues arise as explained hereafter. The AnEn analogy is assessed through ranking past forecasts by their L2 norm distance to the current forecast computed in $\mathbb{R}^N$ space, with N being the number of predictors. If infinite training were available in a stationary climate, an infinite number of perfect analogs (with a distance equal to zero) would be available (Hamill and Whitaker 2006). In real cases, the shorter the training, the more likely the points in the $\mathbb{R}^N$ space representing the selected analog forecast will be distributed farther away from those corresponding to the current forecast. Also, the finite number of analogs introduces a sampling error in representing the predicted PDF which can affect the quality of the predictions. In practice, if the distance between the current and the analog forecast is high enough, the verifying past observations might not be samples from the same error PDF. The current study demonstrates that this issue causes a negative bias in the AnEn wind speed predictions when the deterministic NWP wind speed forecast lies in the right tail of its climatological distribution based on the available historical dataset. This negative bias has also been observed by Plenković et al. (2018). Also, Hamill et al. (2015) found similar biases when the precipitation forecast was unusual as measured in terms of its percentile relative to the climatological distribution of forecasts (q_f). To alleviate this bias, their approach focused on both selecting a smaller number of ensemble members for increasing q_f, and on extending the training by searching for analog forecasts at near locations. In Hamill et al. (2015), the number of analogs ranges from 100 ($q_f < 0.75$) to 20 ($q_f > 0.95$) with a training's length of 12 years. For most previous AnEn applications the training dataset is shorter than 12 years (we use 12–15 months in the current study), and an optimal performance in terms of RMSE is often obtained with about 20 members even for $q_f < 0.75$. Also, when working with station networks instead of gridded data, extending the training by searching for analogs at neighboring stations is generally not as straightforward, depending on the observation density and location. This holds particularly for wind speed predictions, for which the error distribution is often determined by small-scale topographic and roughness features that are hardly similar in different stations.

Alternatively, the method proposed herein is based on linear regression analysis and aims at reducing the conditional bias without changing the number of analog members depending on q_f. A linear regression is performed between observed and predicted wind speed values at each lead time and location, and the resulting slope is used to adjust the AnEn members when the wind speed forecast exceeds a certain quantile of the historical forecast wind speed distribution. Hence, the proposed model is a combination of the original AnEn

algorithm and linear regression. As a term of reference, an ensemble is also generated by using only a multiple linear regression analysis and the performances compared with the new version of the AnEn.

This paper is organized as follows. In section 2, we discuss the datasets used in the study. Section 3 describes the AnEn method we use. Section 4 introduces our new bias correction approach for rare events. Also, alternative approaches for improving the analog forecast for rare events proposed in the literature are described. Section 5 describes the multiple linear regression ensemble. The performance verification of the AnEn and the bias correction are discussed in section 6. A summary and discussion are provided in section 7.

2. Datasets

The same dataset and experiment setup as DM13 are used to show the improvement achieved by the AnEn with the proposed bias correction method when compared to its standard version. Observations and raw model predictions used to generate the AnEn forecasts are available over a 457-day period (1 May 2010–31 July 2011) with the last 100 days used as the verification period. The AnEn requires a training dataset including past forecasts and observations. As in DM13, to mimic a real-time operational condition, the training period consists of 12 months for the first forecast of the verification period (initialized 23 April 2011) and increases to 15 months for the last one (initialized 31 July 2011).

The observational dataset comprises hourly 10-m AGL wind speed observations from 550 routine aviation weather reporting stations (METAR, surface). The stations are located within CONUS (see Fig. 1), and are characterized by different topographic, land-use types, and climate conditions. The regional version of the Environment Canada (EC) deterministic (15 km) Global Environmental Multiscale (GEM) model is used to generate the AnEn predictions. For each day in the dataset, 0–48-h forecasts initialized at 0000 UTC of 10-m wind speed, 10-m wind direction, 2-m temperature, and surface pressure at the station locations are available at 3-h intervals.

3. The analog ensemble

In this section, we briefly describe the AnEn algorithm introduced by DM13. The basic idea behind the AnEn is to exploit a dataset of past forecasts over a specific location generated by an NWP model and a time series of past observations at the same location. In this application, 10-m wind speed is the predictand but the algorithm can be applied to predict other variables as

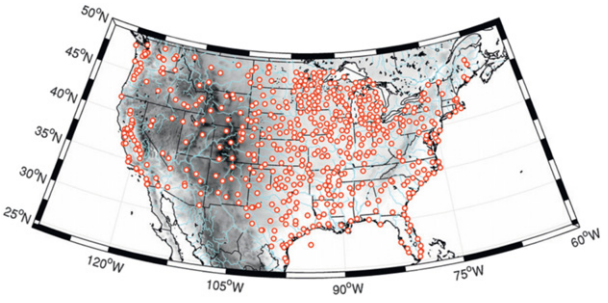

FIG. 1. Spatial distribution of the 550 stations from the routine aviation weather reports (METAR, surface), providing the observations of 10-m wind speed and 2-m temperature used in this study. Darker shading corresponds to higher terrain elevation, rivers are indicated in light blue, and the U.S. state and international borders in black (adapted from DM13).

demonstrated by several previously mentioned studies. The dataset of past forecasts must contain a set of meteorological variables used as predictors, which for the current study are derived from the GEM model and include 10-m wind speed, 10-m wind direction, 2-m temperature, and surface pressure. In the AnEn construction, the predictors are used to detect a given number of past forecasts similar to a future forecast, and the corresponding past concurrent verifying observations form the future ensemble forecast. The basic idea is to find past situations when the model error PDF (a distribution of the differences between predicted and observed 10-m wind speed in this case) is similar to the PDF of the future forecast. If such situations are found then the past and the future verifying observation are sampled from the same PDF. The degree of similarity of the future forecast at a given lead time t and location (hereafter referred to as target forecast) to past potential analog forecasts at the same lead time (ta) and location is assessed by computing the distance ($D_{t,\mathrm{ta}}$), which is

$$D_{t,\mathrm{ta}} = \sum_{P} w_P D_{P,t,\mathrm{ta}}, \tag{1}$$

where

$$D_{P,t,\mathrm{ta}} = \sqrt{(P_t - P_{\mathrm{ta}})^2 + (P_{t-3\mathrm{h}} - P_{\mathrm{ta}-3\mathrm{h}})^2 + (P_{t+3\mathrm{h}} - P_{\mathrm{ta}+3\mathrm{h}})^2}, \tag{2}$$

and subscripts t and ta represent the lead time of a forecast in the future and in the past, respectively. In Eqs. (1) and (2), P is the value of the predictor normalized by its standard deviation computed over the historical dataset at the lead time t and w_P is the weight assigned to each predictor. The summation in Eq. (1) is

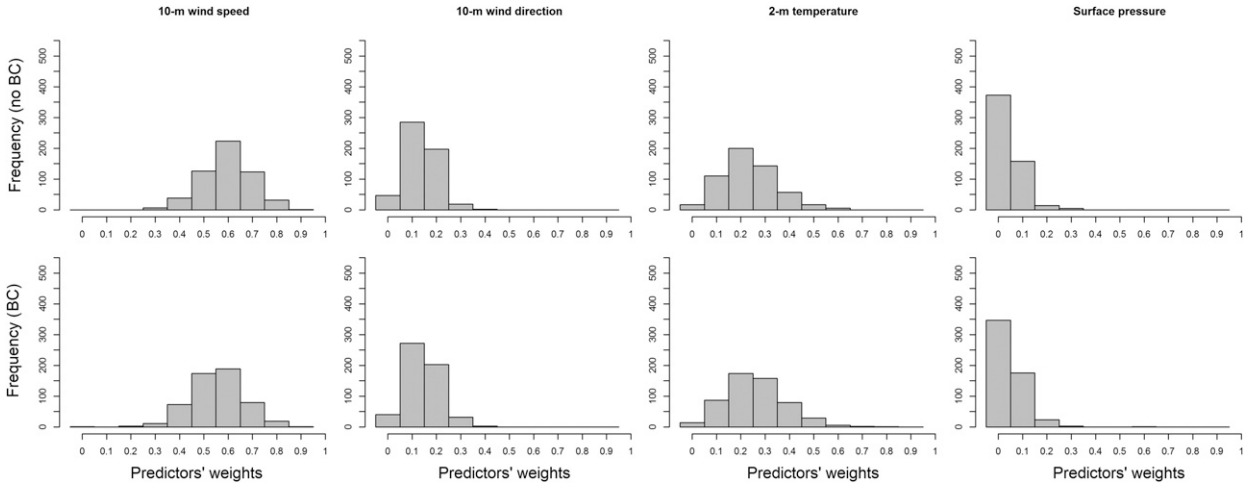

FIG. 2. The distribution of the weights across the stations received by each predictor as a result of the brute force optimization using (top) the standard analog technique (AnEn) and (bottom) the modified version with the bias correction (AnEnBc) as described in section 4a.

over the available predictors (P) which are four in this application as mentioned earlier. This distance is computed over a time interval of ±3 h to also consider similar trends in the past predictions (the 3-h window is motivated by the 3-h time increment availability of the forecasts in this dataset). For each target forecast in the verification dataset, all the distances ($D_{t,\mathrm{ta}}$) with the past forecasts in the training dataset are computed. As in DM13, the past forecasts with the 21 smallest distances are selected and the corresponding 21 wind speed observations used as ensemble members.

In DM13, the weights w_P were each specified as equal to 1, thus assigning the same importance to each predictor. Several subsequent applications (Junk et al. 2015; Alessandrini et al. 2015b) have demonstrated that a brute-force weight optimization (which is computationally feasible with a limited number of predictors, as in the current study) can increase the AnEn performance and has also been carried out in the current study. The weights' optimization is performed independently at each location by choosing the combination that minimizes the continuous ranked probability score (CRPS) over the training dataset. Since only four predictors are used, four corresponding weights can be set. All the possible combinations defined with the constraint $\sum_{P=1}^{4} w_P = 1$, where $w_i \in [0, 0.1, 0.2, \ldots, 1]$, are tested for the AnEn prediction over the training dataset using a leave-one-out approach over the training period. Specifically, for each forecast the AnEn predictions are issued for all possible combinations of weights using all the remaining runs in the training for the analog search.

The weight distributions across the stations received by each predictor after the brute-force optimization are presented in Fig. 2. It is worth noting that 10-m wind speed generally gets the highest weight (about 0.6) as expected, followed by 2-m temperature (about 0.3), by 10-m wind direction (about 0.2) and surface pressure which receives most of the time zero. Having the temperature as the second most important predictor allows the AnEn to select the past analog dates most likely in the same season which indicates that there is a seasonal component in the GEM 10-m wind speed forecast error. Similar, positive weights received by wind direction suggest that the errors are often determined by local topographic features not adequately represented in the GEM 10-m wind speed forecast.

4. AnEn bias correction for rare events

a. *A combination of AnEn and linear regression (AnEnBc)*

To simplify the description of the bias correction (BC) method, the assumptions of having 10-m wind speed as the predictand and only one predictor (10-m wind speed) for the analog selection are made. Also, the ±3 h trend is neglected. Equations (1) and (2) simplify as follows:

$$D_{t,\mathrm{ta}} = D_{P,t,\mathrm{ta}}, \tag{3}$$

$$D_{P,t,\mathrm{ta}} = \sqrt{(P_t - P_{\mathrm{ta}})^2}, \tag{4}$$

where P represents 10-m predicted wind speed.

Figure 3 presents a scatter diagram for station 107 of 9-h wind speed predictions compared with corresponding observations. In this example, the target forecast is equal to 17.9 m s^{-1}, indicated by a red × in the plot, and for simplicity only 10 analog members are generated; the selected closest analog forecasts are the red circles. The binned distribution of past wind speed predictions

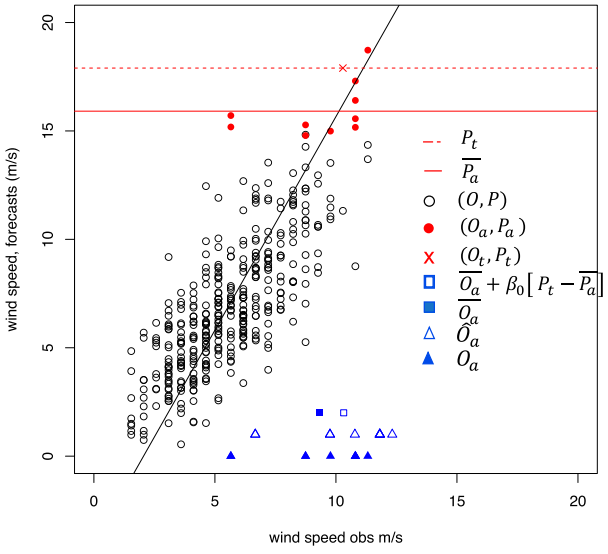

FIG. 3. Scatter diagram of wind speed 9-h forecasts plotted against observations at station 107. The red × symbol indicates the target forecast of 17.9 m s^{-1} used as an example. The expression $P_t - \overline{P_a}$ is the distance between the dashed and the solid red lines. The sloping black linear regression line is obtained by using the observed wind speed as the response variable and the predicted wind speed as the explanatory variable. On the x axis, the filled and empty blue triangles represent the AnEn members before (O_a) and after ($\hat{O}_a$) the bias correction. The blue squares indicate the AnEn mean $\overline{O_a}$ before (filled) and after the correction (empty). It is worth noting that all the blue symbols have no reference to the y axis.

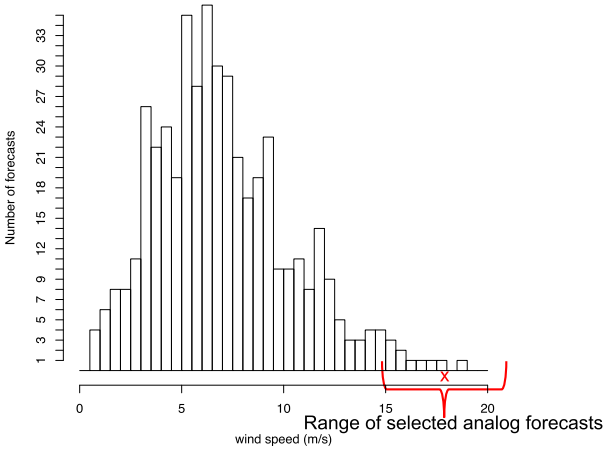

FIG. 4. Binned distribution of the number of wind speed 9-h ahead forecasts at station 107. The red × symbol indicates the target forecast of 17.9 m s^{-1} used as an example for describing the bias correction method. The red parentheses indicate the range of wind speeds for the analog forecasts if 10 ensemble members are chosen.

at the same lead time and station is plotted in Fig. 4. The wind speed range including the selected 10 analog predictions is indicated in red. Here, 9 of the 10 analog wind speed predictions are lower than the target forecast. In fact, if the target prediction is located in the right tail of the forecast distribution, given its decreasing trend, it is more likely to have analog wind speed predictions lower than the target one. The likelihood of finding analog forecasts lower than the target one is more enhanced by a shorter training which limits the possibility of finding similar predictions (with an infinite training dataset, an infinite number of equal analog predictions would be available). Being the mean of the analog forecasts' distribution lower than the target forecast, a negative bias is introduced when comparing the mean of the observed wind speed corresponding to the analog forecasts and the observed wind speed corresponding to the target prediction.

For a wind speed prediction (P), the corresponding observed wind speed (O) may be expressed as

$$O = \beta_0 P + \beta_1 + \varepsilon, \tag{5}$$

where ε is a random error and β_0 and β_1 are the coefficients (slope and intercept) resulting from a linear regression with O as a response variable and P as an explanatory variable. In general, there are some assumptions about the errors ε that should be independent, normally distributed, and should have zero mean for all P. Similarly, the following relationships of the analog and the target forecasts to the corresponding observations are obtained with

$$O_a = \beta_0 P_a + \beta_1 + \varepsilon_a \tag{6}$$

and

$$O_t = \beta_0 P_t + \beta_1 + \varepsilon_t, \tag{7}$$

where the subscripts a and t indicate the analog and the target forecasts and the corresponding observations, respectively. By first taking the mean over the analog forecast events a in Eq. (6), then the mean over all the forecast events of both Eqs. (6) and (7) and then subtracting Eq. (6) from Eq. (7), the following can be obtained:

$$\langle O_t \rangle - \langle \overline{O_a} \rangle = \beta_0[\langle P_t \rangle - \langle \overline{P_a} \rangle] + \langle \varepsilon_t \rangle - \langle \overline{\varepsilon_a} \rangle, \tag{8}$$

where the overbar represents the mean over the selected members and $\langle\ \rangle$ over different forecast events. Given that the bias of the AnEn mean ($\overline{O_a}$) forecasts is equal to $\langle O_t \rangle - \langle \overline{O_a} \rangle$, Eq. (8) shows that the bias has a systematic component equal to $-\beta_0[\langle P_t \rangle - \langle \overline{P_a} \rangle]$ and a random component equal to $\langle \varepsilon_t \rangle - \langle \overline{\varepsilon_a} \rangle$. The systematic component of the bias becomes significantly positive when the

target forecast P_t is within the right tail of the forecast distribution, where it is more likely to be greater than the mean of the analog forecast ($\overline{P_a}$). In general, the random component might not be equal to zero, but in this study we aim at correcting only the systematic component. Also, if $\langle \varepsilon_t \rangle = 0$ can be considered a good approximation over a large number of forecast events, from the central limit theorem, $\langle \overline{\varepsilon_a} \rangle = 0$. In that case, the random component of the AnEn bias tends to zero. The proposed bias correction consists of adjusting the AnEn members O_a for each forecast event by removing the systematic component as follows,

$$\hat{O}_a = O_a + \beta_0[P_t - \overline{P_a}], \tag{9}$$

where $\hat{O}_a$ represents the bias-corrected members.

For the case of 9-h wind speed predictions at station 107 depicted in Fig. 3, $P_t - \overline{P_a}$ is indicated by the distance between the dashed and the solid red line. The black line is plotted according to the coefficients resulting from the linear regression with the observed wind speed as the response variable and the predicted wind speed as the explanatory variable. On the x axis, the filled and empty blue triangles represent the AnEn members before (O_a) and after ($\hat{O}_a$) the bias correction. The blue squares indicate the AnEn mean $\overline{O_a}$ before (filled) and after (empty) the correction ($\overline{\hat{O}_a} = \overline{O_a} + \beta_0[P_t - \overline{P_a}]$). It can be seen that after the adjustment, the AnEn mean ($\overline{\hat{O}_a}$) underestimation with respect to the verifying observation O_t is corrected.

The proposed method works if the assumption of a linear relationship between predicted and observed values is acceptable, which may not be true in general. For variables other than wind speed, Eqs. (6) and (7) should be adapted to account for the nonlinearity. For example, in applications related to wind power prediction (e.g., Junk et al. 2015; Alessandrini et al. 2015a) using observed power and wind speed predictions, the terms P_a and P_t on the right-hand side of Eqs. (6) and (7) should be replaced by $F(P_a)$ and $F(P_t)$ with F being the power curve function.

Also, Eq. (9) can be applied for any prediction P_t and may be suitable if $P_t < \overline{P_a}$, which usually occurs when P_t is in the left tail of the forecast distribution. For wind speed, several empirical tests carried out over the training dataset adopted in this work have shown that a conditional application of Eq. (9) avoids a performance degradation as measured by the root-mean-square error (RMSE). These tests (not shown) have demonstrated that an optimal RMSE is obtained if Eq. (9) is applied when $P_t > Q_{90}(P_t)$, with Q_{90} being the 90th quantile of the climatological forecast distribution independently computed at any station and lead time. As a matter of fact, without using the threshold the target forecast is often very close to the analog forecasts mean, which makes the BC adjustments small and noisy leading to a degradation of the root-mean-squared error. This threshold should be considered as a tuning parameter for the AnEn application and might require a specific optimization for variables other than wind speed or different datasets. Also, for variables with a two-tail distribution (e.g., 2-m temperature) adopting an additional condition for the left tail of the PDF [e.g., $P_t < Q_{10}(P_t)$] may also be beneficial.

b. *Extending the training with observations from neighboring stations*

As already mentioned, Hamill et al. (2015) proposed to extend the training dataset from which selecting the analogs by including observations from supplemental neighboring grid points. These additional grid points were selected based upon the similarity of the observed climatology of rainfall, the similarity of terrain characteristics and some constraints on their distance. Clearly, extending the training increases the likelihood that the analog forecasts are very similar to the target forecast even when the target forecast is rare, which should alleviate the conditional negative bias for wind speed predictions. In this work, three selection criteria to add supplemental locations are tested. They are much simpler than what proposed in Hamill et al. (2015) and are meant to explore the potential feasibility for wind speed.

The first two criteria consist of extending the training dataset of each station by including the training dataset of supplemental stations only if their distance is lower than 20 km (neigh_20 km) and 50 km (neigh_50 km). Respectively, only 31 and 173 stations had at least one supplemental location within a range of 20 and 50 km. The underlying assumption of this approach is that two nearby locations might share similar error structure determined by similar topography and land use. The third selection method is to use the 10 supplemental stations with the most similar bias (neigh_bias), defined as the difference between the mean GEM wind speed predictions and the observations.

5. Multiple linear regression ensemble

An ensemble based on a multiple linear regression (MLR_En) is described in this section and used as a base of reference for the AnEnBc in the subsequent analyses.

The use of linear regression–based techniques to correct errors in NWP outputs are commonly referred to as model output statistic (MOS). The earliest application goes back to the work of Glahn and Lowry (1972) in which predicted wind speeds and wind components at

different pressure levels were used as predictors in a multiple linear regression with surface wind speed observations as the predictand.

Since the novel version of the AnEn, as described in section 4a, is a combination of the AnEn and a linear regression analysis, we aim at comparing it with a simpler model based only on the latter technique. Similar to Glahn and Lowry (1972), a multiple linear regression is used here to correct the GEM wind speed forecast.

At each station and for each forecast lead time an independent multiple regression analysis is carried out on the training dataset (see section 2) by computing the regression parameters (b) in the equation:

$$\mathrm{ws}_{\mathrm{obs}} = b_0 + b_1\mathrm{ws}_f + b_2 v_f + b_3 u_f + b_4 P_f + b_5 T_f, \qquad (10)$$

where ws_f, v_f, u_f, P_f, and T_f are, respectively, 10-m wind speed, 10-m v and u components, the surface pressure, and the 2-m temperature from GEM while $\mathrm{ws}_{\mathrm{obs}}$ is the observed 10-m wind speed. To check whether adding predictors other than ws_f improves the regression, we compared the adjusted R^2 of the Eq. (10) with the one from the regression using ws_f as the only predictor. For each station, Eq. (10) is used only if the adjusted R^2 is higher, which happens in about 95% of the cases. In the remaining 5% of the cases, we used ws_f as the only predictor. The ensemble members are then generated by taking the values corresponding to the quantiles of the Gaussian distribution of the regression's residuals and adding them to the wind speed predicted by Eq. (10). Also, for consistency with AnEn, 21 members are used to build the MLR_En. It is worth noting that for each location the MLR_En spread, defined as the standard deviation of the members about their mean, depends only on the forecast lead time but not on the predicted wind speed. Hence, differently from AnEn, MLR_En's ensemble spread cannot be considered flow dependent.

6. Results

In this section, the AnEn's performance with the novel bias correction method (AnEnBc) is assessed and compared to the original version (AnEn) and to MLR_En using common verification metrics for evaluation of deterministic and probabilistic predictions. It is worth noting that both AnEn and AnEnBc methods used in this study employ optimized weights (section 3) while in DM13 equal weights were assigned to the predictors. This choice is motivated by previously documented superior performance when the weights are optimized, (e.g., Junk et al. 2015).

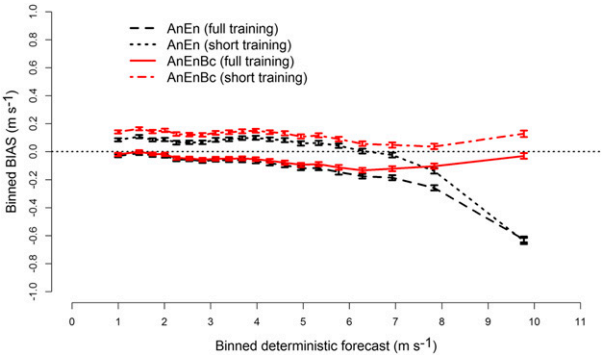

FIG. 5. AnEn and AnEnBc ensemble mean bias as a function of the wind speed from the Global Environmental Multiscale (GEM) model averaged over equally populated bins. The 1-yr training period and the shorter 9-month training periods are used for both AnEn and AnEnBc. The error bars indicate the 95% bootstrap confidence intervals.

a. Bias

A deterministic, single-valued forecast can be obtained from any ensemble prediction by taking the mean of the ensemble members at any forecast lead time. In this study, we analyze the performance of the AnEn members mean ($\bar{f}$). A total of 21 members are used to be consistent with DM13.

The bias, also known as the systematic error, is defined as

$$\mathrm{bias} = \langle \bar{f} \rangle - \langle o \rangle, \qquad (11)$$

where $\langle\ \rangle$ indicates the mean over all the available observation/forecast pairs. The bias measures the average under/overestimation of the forecasts compared to observations. In Fig. 5, the bias is computed for both the AnEn and AnEnBc ensemble means, as a function of the mean wind speed from the Global Environmental Multiscale (GEM) model computed for equally populated bins. To gain insight on the analog models' performances, tests have been carried out both with the whole 1-yr-long training (starting 1 May 2010) and with the shorter 9-month-long training (starting 1 August 2010).

Without the BC, the AnEn conditional negative bias gets larger (in absolute value) for higher values of the GEM wind speed deterministic predictions on which the AnEn is based. On the other hand, AnEnBc is not affected by the same conditional bias even though a slight positive bias is evident with the shorter training.

A similar analysis to evaluate the methods based on the additional locations (neigh_bias, neigh_50 km and neigh_20 km) as described in section 4b is presented in Fig. 6. All three methods exhibit a slight improvement over the standard AnEn in terms of reducing the

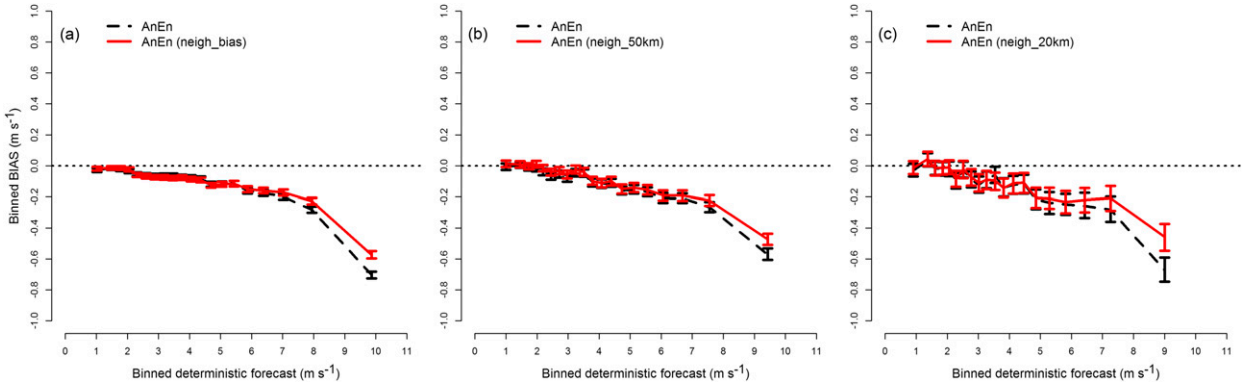

FIG. 6. AnEn (black), and (left) AnEn neigh_bias (red), (middle) AnEn neigh_50 km (red), and (right) AnEn neigh_20 km (red) ensemble mean bias as a function of the wind speed from the Global Environmental Multiscale (GEM) model averaged over equally populated bins. For the plots including the models neigh_50 km and neigh_20 km, the bias has been computed considering only the stations for which at least one supplemental location has been used. The error bars indicate the 95% bootstrap confidence intervals.

conditional negative bias for high wind speed. However, when looking at the root-mean-squared error, all three methods show significantly worse performance than AnEn (not shown). For this reason, they have not been included in the evaluation presented from now on.

In Fig. 7, bias as a function of the forecast lead time is shown for AnEn, AnEnBc, and MLR_En for the whole set of observations (case 0) and for two subsets restricted to observed wind speed greater than $5\,m\,s^{-1}$ (case 5) and $10\,m\,s^{-1}$ (case 10). According to the 5%–95% bootstrap confidence intervals, for all three cases, AnEnBc improves AnEn by reducing the negative bias a statistically significant amount at all the lead times. For observed wind speed greater than $10\,m\,s^{-1}$, the negative bias worsens in the standard version of the AnEn and is only partially alleviated by AnEnBc, with a reduction (in absolute value) of about $1\,m\,s^{-1}$ at all lead times. Regarding MLR_En, its bias usually ranges between AnEn and AnEnBc for case 0, getting very similar to AnEnBc for cases 5 and 10. It is worth noting the diurnal cycle in the bias trend for all the models, especially for cases 5 and 10. The absolute value of the bias is lower around 0900 UTC, which corresponds to early morning over CONUS and is also the time of the day with the highest observed mean wind speed over all the stations (not shown).

b. *Root-mean-squared error and centered root-mean-squared error*

The root-mean-square error (RMSE) is a common verification metric for deterministic predictions, and is a quadratic score index, that gives higher weights to larger forecast errors. The RMSE measures both systematic error (bias) and random errors and is defined as follows:

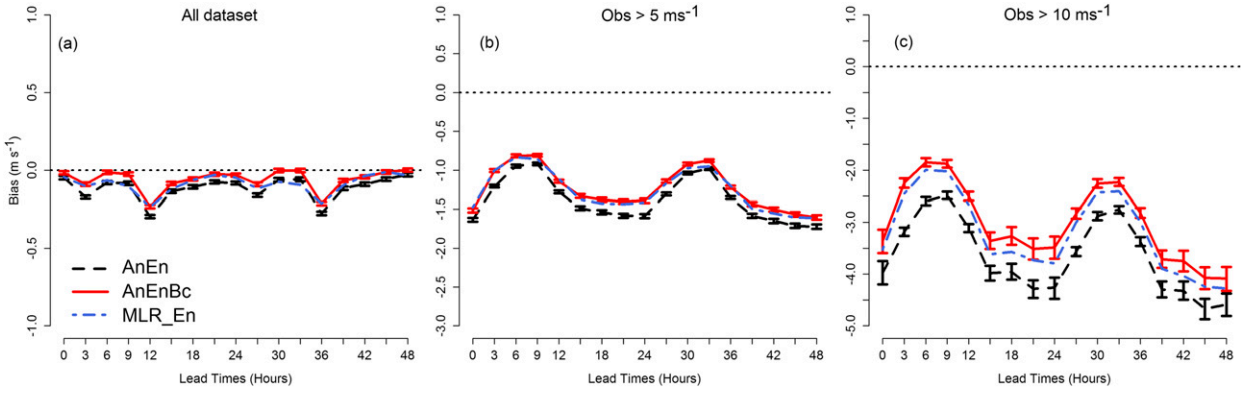

FIG. 7. Bias as a function of the forecast lead time for the AnEn, AnEnBc, and MLR_En ensemble means for 10-m wind speed using (a) all the observations, (b) wind speed greater $5\,m\,s^{-1}$, and (c) $10\,m\,s^{-1}$ using AnEn (dashed line, black), AnEnBc (solid line, red), and MLR_En (dot–dashed, blue). Note the different ranges of the vertical axis in (a)–(c). The error bars indicate the 5%–95% bootstrap confidence intervals. They are plotted only for AnEn and AnEnBc to reduce the clutter.

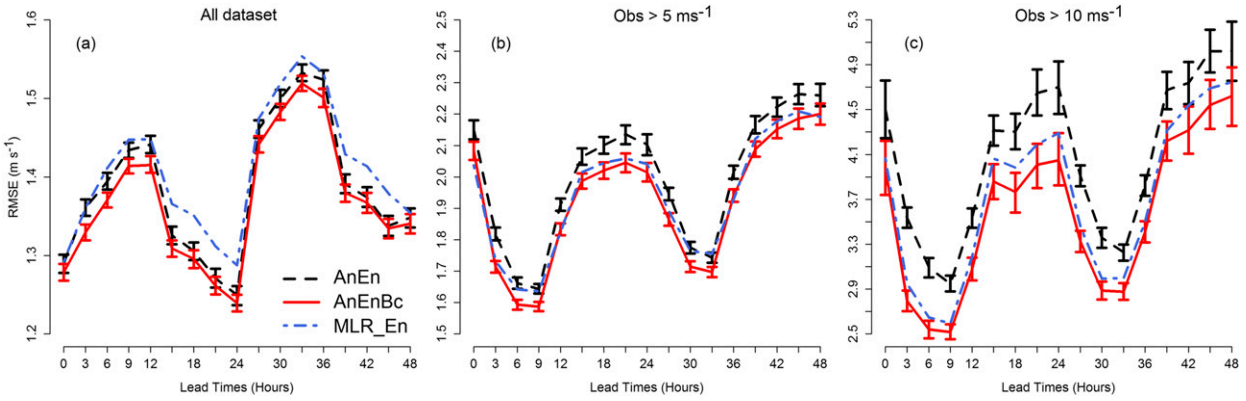

FIG. 8. Root-mean-square error (RMSE) as a function of the forecast lead time for the AnEn, AnEnBc, and MLR_En ensemble means for 10-m wind speed using (a) all the observations, (b) wind speed greater 5 m s^{-1}, and (c) 10 m s^{-1} using AnEn (dashed line, black) and AnEnBc (solid line, red). Note the different ranges of the vertical axis in (a)–(c). The error bars indicate the 5%–95% bootstrap confidence intervals.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\bar{f}_i - o_i)^2}, \tag{12}$$

where N is the total number of forecast events.

If the bias is removed from each forecast error, the centered root-mean-squared error (CRMSE) can be obtained, which includes random errors and residual conditional biases. The CRMSE is defined as follows:

$$\mathrm{CRMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[(\bar{f}_i - \langle\bar{f}\rangle) - (o_i - \langle o\rangle)]^2}. \tag{13}$$

In Figs. 8 and 9 , the RMSE and the CRMSE as a function of the forecast lead time for the AnEn and AnEnBc are plotted using the whole set of observations and the two subsets (cases 5 and 10).

For the RMSE, the AnEnBc outperforms AnEn at all lead times and in all the three cases. The improvements are statistically significant only up to 12 h ahead for case 0, at all the lead times except at 48 h for case 5, and at all the lead times for case 10. More significant improvements gained from using AnEnBc in case 10 are consistent with the AnEn members being corrected by the BC algorithm only if $P_t > Q_{90}(P_t)$. Thus, AnEn and AnEnBc predictions are different only for the 10% of the forecast events that are likely corresponding to higher wind speed observations. For MLR_En, RMSE is worse than both AnEn and AnEnBc for case 0, becoming more similar to AnEnBc for cases 5 and 10.

For CRMSE, AnEnBc improves over AnEn for case 0, but generally gets worse than AnEn for cases 5 and 10, which means that the BC algorithm introduces random errors. In all the three cases and at all the lead times, the

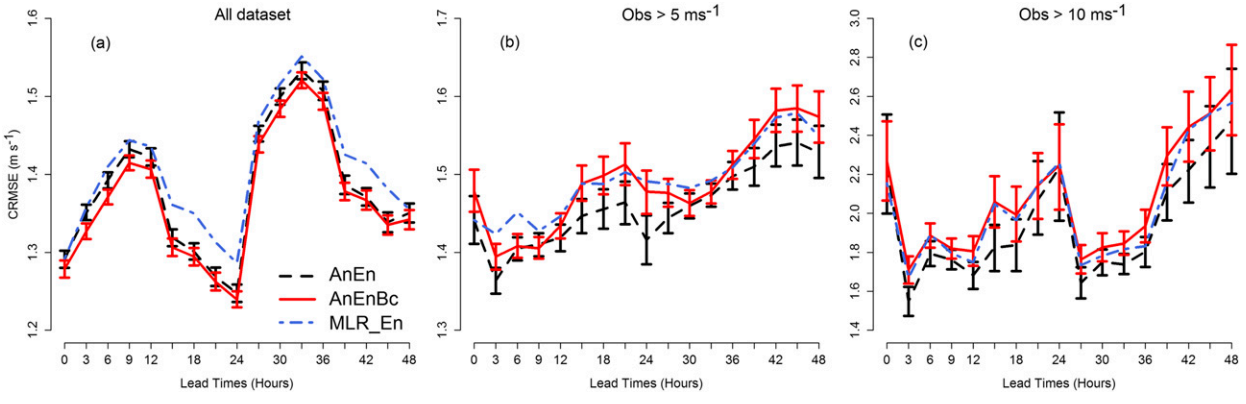

FIG. 9. Centered root-mean-square error (CRMSE) as a function of the forecast lead time for the AnEn, AnEnBc, and MLR_En ensemble means for 10-m wind speed using (a) all the observations, (b) wind speed greater 5 m s^{-1}, and (c) 10 m s^{-1} using AnEn (dashed line, black) and AnEnBc (solid line, red). Note the different range of the vertical axis in (a)–(c). The error bars indicate the 5%–95% bootstrap confidence intervals.

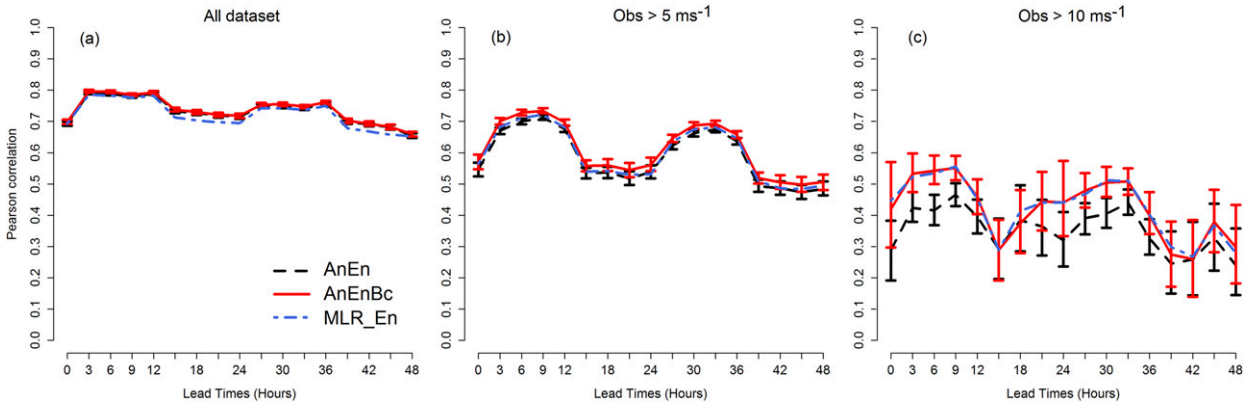

FIG. 10. Pearson correlation coefficient as a function of the forecast lead time for the AnEn, AnEnBc, and MLR_En ensemble means for 10-m wind speed using (a) all the observations, (b) wind speed greater 5 m s^{-1}, and (c) 10 m s^{-1} using AnEn (dashed line, black) and AnEnBc (solid line, red). The error bars indicate the 5%–95% bootstrap confidence intervals.

differences between the two models are not statistically significant. For MLR_En, CRMSE as RMSE is the worst for case 0 and very similar to AnEnBc for cases 5 and 10. Comparing the results in Figs. 7–9, we conclude that the RMSE improvements of AnEnBc over AnEn result from a significant reduction of the systematic component of the errors (bias), which can compensate the increased random component (CRMSE). A similar conclusion can be drawn when comparing MLR_En with AnEn on cases 5 and 10. Also, MLR_En's worst performance in terms of RMSE over the whole dataset (case 0) can be attributed to a higher random component of its errors.

c. *Pearson correlation coefficient (CC)*

The CC is defined as

$$\mathrm{CC} = \frac{1}{N}\frac{\sum_{i=1}^{N}(\bar{f}_i - \langle\bar{f}\rangle)(o_i - \langle o\rangle)}{\sigma_{\bar{f}}\sigma_o}, \qquad (14)$$

where $\sigma_{\bar{f}}$ and σ_o are the standard deviations of the forecasts (AnEn mean) and the observations, respectively. The CC measures the strength of the linear association between two variables. It ranges between $[-1, 1]$ with 1 being the best achievable correlation.

In Fig. 10, the CC as a function of the forecast lead time for the AnEn and AnEnBc are plotted using the whole set of observations and the two subsets (cases 5 and 10). The CC is very similar for AnEn and AnEnBc when calculated for the entire dataset. For cases 5 and 10 AnEnBc outperforms AnEn, but with the bootstrap intervals indicating statistically significant improvements only for few lead times. The conditional bias reduction in AnEnBc compared to AnEn compensates the introduction of some noise enough not to worsen the CC. MLR_En correlation values are very similar to AnEnBc except for case 0 where AnEnBc significantly outperforms MLR_En indeed because of a lower CRMSE.

d. *Rank histograms*

A rank histogram can be used to assess the statistical consistency of an ensemble, and indicates whether the members of an ensemble system are statistically indistinguishable from the observations. In a consistent ensemble, an observation ranked among the corresponding ordered ensemble members is equally likely to take any rank in the range of the whole forecast PDF (Anderson 1996). An ensemble is perfectly statistically consistent when its rank histogram is flat and has a uniform rank probability of $1/(n + 1)$ (Hamill 2001), with n equal to the number of ensemble members. A rank histogram can be presented together with the missing rate error (MRE), which is the fraction of observations lower (higher) than the lowest (highest) ranked prediction above or below the expected missing rate of $1/(n + 1)$. A larger positive (negative) MRE reveals a more under dispersive (over dispersive) ensemble.

In Fig. 11, a rank histogram is compiled for AnEn, AnEnBc, and MLR_En together with their corresponding MRE values. The overall negative bias affecting AnEn is evident by the bins on the right side of the histogram being more populated than those on the left. The improvement in overall bias of AnEnBc compared to AnEn is also characterized by its flatter rank histogram. The MRE values are very low for both AnEn and AnEnBc, which indicates a good statistical consistency for the methods. For MLR_En, it is the worst model in terms MRE with an over dispersive behavior

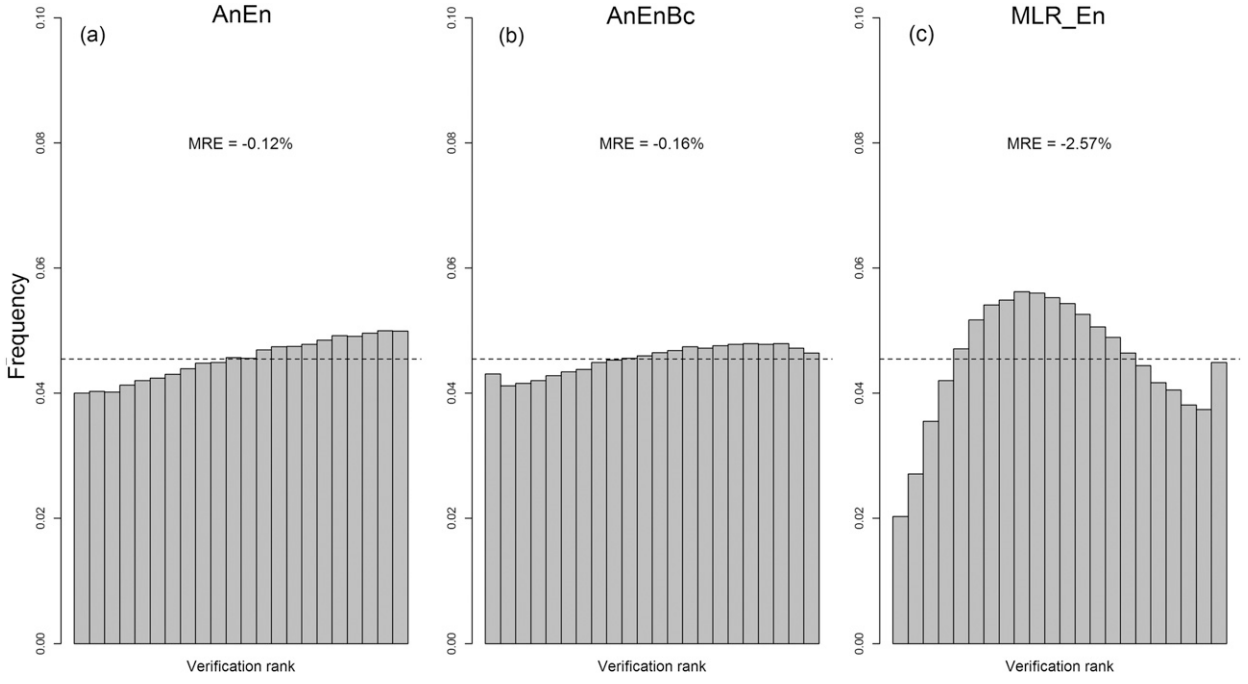

FIG. 11. Rank histograms for probabilistic prediction of 10-m wind speed for (a) AnEn, (b) AnEnBc, and (c) MLR_En with annotated missing rate error (MRE) results. The gray histogram bars show the frequency of occurrence of the observation in each rank. The dashed line indicates a perfect, uniform probability for a 21-member ensemble.

indicating that the assumption of a Gaussian residual distribution does not hold in general.

e. *Spread–skill*

Binned spread–skill diagrams can assess the ability of an ensemble system to quantify its own uncertainty (Fortin et al. 2014; Hopson 2014). In a spread–skill diagram, the ensemble spread is compared to the RMSE of the ensemble mean over small class intervals (i.e., bins) of spread, instead of considering its overall average (Van den Dool 1989; Wang and Bishop 2003). A good correlation in the spread–skill diagram is an indication that an ensemble system is able to forecast its own error (Hopson 2014). Binned spread–skill diagrams for both AnEn and AnEnBc are presented in Fig. 12. Each bin has the same number of forecast/observation pairs, which results in bins of different width.

Note that introducing the BC leads to a slight reduction of the RMSE for the bins on the right. This is consistent with the lower RMSE for AnEnBc compared to AnEn (Fig. 8). The overall spread–skill relationship looks very similar for the two models even though the reduction of the RMSE in the highest bins for AnEnBc is noticeable with a more significant departure from the diagonal. In fact, it is worth to notice that the BC technique does not modify the AnEn spread, but only the ensemble mean that in turn affects the RMSE.

MLR_En also exhibits an excellent spread–skill relationship. Considering that for MLR_En the spread does not change with respect to the predicted wind speed for a fixed station and lead time, it can be

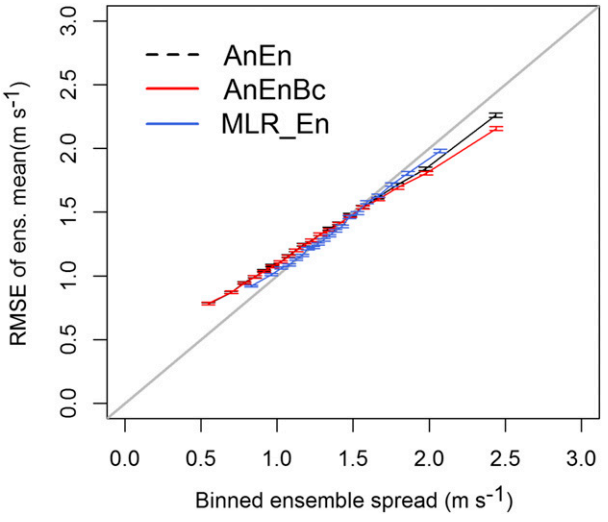

FIG. 12. Binned spread–skill diagram of 10-m wind speed for AnEn (black), AnEnBc (red), and MLR_En (blue) calculated over all forecast lead times. The error bars indicate the 95% bootstrap confidence interval. The diagonal 1:1 line represents a perfect spread–skill trend. For each diagram, the ensemble spread is binned into 20 equally populated class intervals.

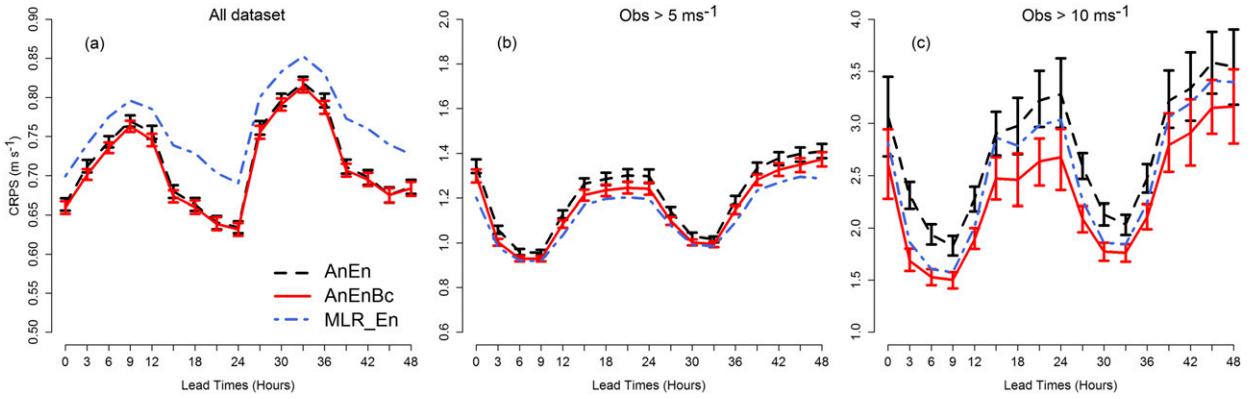

FIG. 13. Continuous ranked probability score (CRPS) as a function of the forecast lead time for the probabilistic prediction of 10-m wind speed using (a) all the observations, (b) wind speed greater 5 m s^{-1}, and (c) 10 m s^{-1} using AnEn (dashed line, black), AnEnBc (solid line, red), and MLR_En (dot–dashed line, blue). Note the different ranges of the vertical axis in (a)–(c). The error bars indicate the 95% bootstrap confidence intervals.

concluded that most of the RMSE variability depends on the lead time and location.

f. CRPS

The continuous ranked probability score (CRPS) is a metric used to assess the overall quality of an ensemble system (Carney and Cunningham 2006). It can generally be expressed as:

$$\mathrm{CRPS} = \frac{1}{N}\sum_{i=1}^{N}\int_{-\infty}^{\infty}\left[F_i^f(x) - F_i^0(x)\right]^2 dx, \qquad (15)$$

where $F_i^f(x)$ is the cumulative distribution function (CDF) of the probabilistic forecast and $F_i^0(x)$ is the CDF of the observation for the ith ensemble prediction/observation pair, and N is the number of available forecast events. It has been shown (Hersbach 2000) that the CRPS is equivalent to the mean absolute error if evaluating a deterministic (single valued) forecast. A lower value of CRPS indicates better performances, with 0 indicating a perfect score. CRPS has the same units as the forecasted variable. In Fig. 13, the CRPS as a function of forecast lead time is shown for both AnEn and AnEnBc for the three cases (cases 0, 5, and 10).

Similar to RMSE, CRPS improvements for AnEnBc compared to AnEn are larger for case 10, which means that the BC improves the overall probabilistic skill of AnEn when predicting events with observed wind speed higher than 10 m s^{-1}. The improvements are statistically significant up to 36 h ahead except for lead time 0 and lead time 18 where the confidence intervals barely overlap.

When comparing MLR_En to AnEn and AnEnBc, it is evident that MLR_En is competitive only for case 5. For case 0, it is the worst model while for case 5 it slightly outperforms AnEn, but it is worse than AnEnBc.

The different components of the CRPS contain information about which attributes of a probabilistic prediction leads to the improvement of the AnEnBc over the AnEn. We have used the function ''crpsDecompostion'' provided by the R package ''verification'' (NCAR Research Applications Laboratory 2014) to compute the CRPS and its components that will be discussed hereafter. As demonstrated by Hersbach (2000), the CRPS can be decomposed into three components similarly to the Brier score (Brier 1950; Murphy 1973). These three components are the reliability (REL), resolution (RES), and uncertainty (UNC). The REL component measures how well the forecasted probabilities match the observed probabilities, with smaller REL indicating a better match. The REL attribute may also be assessed by compiling the previously discussed rank histograms. The RES component measures the system improvements compared to a climatological probabilistic forecast, which is a single probability value of an event observed in the dataset. In general, the resolution attribute of a system reflects how well the different forecast frequency classes can separate the different observed frequencies from the climatological mean. The UNC component measures how the situations in the dataset are predictable by a climatological forecast. UNC is related to the variability of the observations and therefore depends on the observations only. For a more detailed demonstration of how to derive the three components, we refer the reader to Hersbach (2000). For the purpose of this paper, we just note that CRPS can be expressed as CRPS = REL + CRPSPOT, where CRPSPOT = UNC − RES is the

TABLE 1. Decomposition of the continuous ranked probability score (CRPS) for AnEn, AnEnBc, and MLR_En predictions of 10-m wind speed using all the observations (case 0), wind speed greater 5 m s^{-1} (case 5), and 10 m s^{-1} (case 10). The best CRPS, CRPSPOT, and REL scores for each case are in bold font.

	Threshold (m s^{-1})	CRPS	CRPSPOT	REL
AnEn	0	0.723 < 0.726 < 0.729	0.722	0.004
AnEnBc	0	0.719 < **0.721** < 0.723	**0.718**	**0.003**
MLR_En	0	0.767 < 0.769 < 0.772	0.764	0.005
AnEn	5	1.140 < 1.146 < 1.153	0.849	0.297
AnEnBc	5	1.100 < 1.106 < 1.112	0.872	**0.234**
MLR_En	5	1.070 < **1.074** < 1.081	**0.826**	0.248
AnEn	10	2.319 < 2.371 < 2.421	1.040	1.331
AnEnBc	10	1.929 < **1.970** < 2.015	1.131	**0.839**
MLR_En	10	2.072 < 2.116 < 2.132	**0.984**	1.133

potential CRPS. The potential CRPS is the CRPS in a perfectly reliable (REL = 0) forecasting system.

The values of CRPS (with bootstrap confidence intervals), CRPSPOT, and REL computed over all the lead times are reported in Table 1. The confidence intervals show that AnEnBc significantly outperforms AnEn only for case 10 in terms of the overall CRPS. We note that CRPSPOT is very similar for the two models and the improvements of the CRPS can be mainly attributed to a reduction of REL (i.e., AnEnBc is better than AnEn in terms of reliability). AnEn, being affected by a negative bias, tends to be underconfident, which means that underestimates the probabilities of occurrences of high wind speed, a problem that is alleviated with the introduction of BC. Regarding MLR_En, its lowest resolution (highest CRPSPOT) leads to the worst CRPS for case 0. On the contrary, improvements for case 5 (where MLR_En exhibits the lowest CRPS) and 10 are achieved for a better resolution that cannot compensate for the degrading reliability to still outperform AnEnBc on case 10.

g. *ROCSS*

The receiver operating characteristic (ROC) is an approach commonly used to assess the ability of a probabilistic forecasting system to distinguish situations leading to the occurrence and nonoccurrence of an event which is also called discrimination (Mason 1982). For a probabilistic forecast, a ROC curve is obtained by plotting the false alarm rate (i.e., false alarms divided by total of nonoccurrence of the event) against the hit rate (i.e., the correct forecasts divided by the total occurrences of the event) for different probability thresholds. For each threshold, an event is considered a hit/false alarm if both the predicted probability is higher than the threshold and the event occurred/did not occur. The area delimited by the ROC curve is the ROC index, with a greater area corresponding to a higher number of hit rates, showing a better ability of the forecast system to discriminate. ROC index is widely used in decision making. For instance, the decision to undertake actions based on likelihood of occurrence of a particular meteorological event can depend on the forecasted probability of the same event exceeding a certain threshold. ROC skill score (ROCSS) is the translation of the ROC score into a standard skill score, with a value equal to 1 corresponding to a perfect forecast and values lower than 0 denoting a system performing worse than climatological forecasts.

In Fig. 14, ROCSS as a function of forecast lead times is shown for AnEn and AnEnBc with events having wind speeds exceeding 5 m s^{-1} (Fig. 14a) and 10 m s^{-1} (Fig. 14b). For the 5 m s^{-1} threshold, there is no difference evident between AnEn and AnEnBc. However, for the 10 m s^{-1} threshold, AnEnBc outperforms AnEn at all lead times even though the confidence intervals indicate that the improvements are not statistically significant. MLR_En exhibits the worst ROCSS at case 5 and it is still generally worse than AnEnBc at case 10.

7. Summary

A new bias correction (BC) method for improving the analog ensemble (AnEn) model for wind speed predictions of rare events has been presented. It has been demonstrated that the AnEn, in its earliest formulation (Delle Monache et al. 2013) is affected by a conditional negative bias when predicting events in the right tail of the forecast distribution. This conditional negative bias increases as the deterministic wind speed prediction is larger and as the training dataset gets shorter. The proposed method is based on a linear regression analysis between forecast and observations performed independently at each lead time and location. Each member is adjusted by adding a factor proportional to the difference between the target forecast and the mean of the past analog forecasts multiplied by the coefficient obtained after the linear regression analysis. In contrast

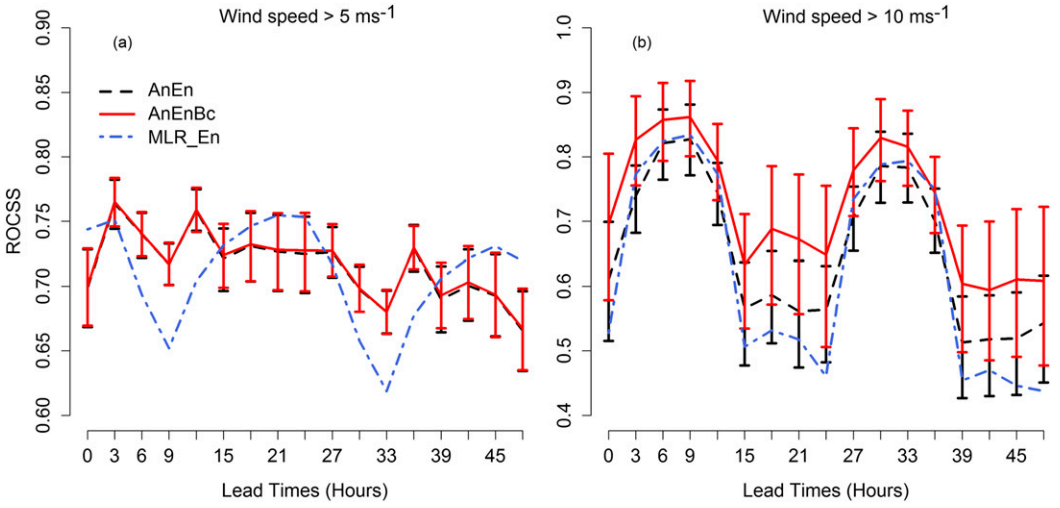

FIG. 14. Relative operating characteristic skill score (ROCSS) as a function of the forecast lead time for the probabilistic prediction of 10-m wind speed greater than (a) $5\,m\,s^{-1}$, and (b) $10\,m\,s^{-1}$ using AnEn (dashed line, black), AnEnBc (solid line, red), and MLR_En (dot–dashed line, blue). Note the different ranges of the vertical axis in (a) and (b). The error bars indicate the 95% bootstrap confidence intervals.

to previous approaches (Hamill et al. 2015), the proposed BC aims to keep the same number of members regardless of the magnitude of the target forecast. Also, other approaches similar to those previously explored by Hamill et al. (2015) for rainfall predictions have been tested. They consist of searching for analogs in neighboring stations or in stations where the GEM wind speed forecasts exhibit similar biases. In both cases, the novel BC approach was more effective for reducing the AnEn conditional bias as well as the RMSE. In fact, NWP prediction errors for wind speed are more likely dependent on topographic and roughness features at a lower spatial scale than those of rainfall. It is then harder to find stations with similar features unless a very dense observation network is available.

The AnEn with the BC method (AnEnBc) has been compared to the original AnEn using the same dataset as that used in Delle Monache et al. (2013), which is constructed from a 457-day period of hourly 10-m AGL wind speed observations from 550 aviation routine weather-reporting stations across the contiguous United States (CONUS) and meteorological predictions at these stations from the regional version of the Environment Canada (EC) deterministic (15 km) Global Environmental Multiscale (GEM) model. We have shown that AnEnBc improves AnEn by reducing the conditional negative bias for wind speed. The overall bias is more significantly improved when the verification is for higher wind speeds (cases 5 and 10 corresponding to wind speed greater than 5 and $10\,m\,s^{-1}$). The overall centered root-mean-squared error (CRMSE) is very similar for AnEn and AnEnBc but slightly deteriorates for cases 5 and 10. This means that the BC method introduces a random error, which because of the bias reduction, does not spoil the overall root-mean-square error (RMSE) reduction obtained with AnEnBc. In addition, the correlation coefficient is slightly improved with AnEnBc.

AnEn and AnEnBc have also been compared for several different attributes of a probabilistic prediction. In particular, AnEnBc has a better continuous ranked probability score than AnEn. The improvements are more significant when restricting the dataset to wind speed greater than $10\,m\,s^{-1}$ (case 10). By looking at the different components of the CRPS, it is possible to attribute the overall CRPS improvements to an increased reliability of AnEnBc. AnEnBc also outperforms AnEn in terms of discrimination even though the improvements are not statistically significant when looking at the singles lead times.

An ensemble generated by sampling the quantiles from the residuals' distribution of a multiple linear regression (MLR_En) has been used as a reference throughout the whole verification. As for the deterministic verification, MLR_En is, in general, competitive in terms of bias with AnEn, and slightly worse than AnEnBc. Over the whole dataset (case 0), the RMSE of MLR_Ens is the worst due to the highest random component of the errors (CRMSE). For larger wind speed (cases 5 and 10), AnEn can generally outperform MLR_En only when coupled with the BC technique. Similar conclusions can be drawn when looking at the probabilistic verification (CRPS).

MLR_En's CRPS is the worst for case 0, but competitive with AnEnBc for cases 5 and 10.

These conclusions suggest that the AnEn is more suitable than linear regression–based techniques in situations with a weak linear correlation between predictand and the predictor (for wind speed lower than $5\,m\,s^{-1}$). When a linear relationship holds better (for wind speed larger than $5\,m\,s^{-1}$), MLR_En becomes more competitive and could outperform AnEn if the BC technique was not introduced.

The proposed BC method has been tested for wind speed predictions but could be applied to any other variable. However, some adjustments might be necessary, since a linear relationship between model predictions and observations might not hold in general. Also, the choice of the quantile threshold to activate the BC should be optimized for different datasets. In terms of computational time, the introduction of BC did not require any significant additional costs.

Acknowledgments. The National Center for Atmospheric Research is sponsored by the National Science Foundation. We also wish to acknowledge the Kuwait Institute for Scientific Research (KISR) and Dr. Majed Al-Rasheedi for supporting part of this work. We also thank Stan Trier for his help with the NCAR internal review of the manuscript, Tom Hamill and an anonymous reviewer for reviewing the paper and providing useful feedback and suggestions.

REFERENCES

Alessandrini, S., F. Davò, S. Sperati, M. Benini, and L. Delle Monache, 2014: Comparison of the economic impact of different wind power forecast systems for producers. *Adv. Sci. Res.*, **11**, 49–53, https://doi.org/10.5194/asr-11-49-2014.

——, L. Delle Monache, S. Sperati, and J. N. Nissen, 2015a: A novel application of an analog ensemble for short-term wind power forecasting. *Renewable Energy*, **76**, 768–781, https://doi.org/10.1016/j.renene.2014.11.061.

——, ——, ——, and G. Cervone, 2015b: An analog ensemble for short-term probabilistic solar power forecast. *Appl. Energy*, **157**, 95–110, https://doi.org/10.1016/j.apenergy.2015.08.011.

——, ——, C. M. Rozoff, and W. E. Lewis, 2018: Probabilistic prediction of tropical cyclone intensity with an analog ensemble. *Mon. Wea. Rev.*, **146**, 1723–1744, https://doi.org/10.1175/MWR-D-17-0314.1.

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate*, **9**, 1518–1530, https://doi.org/10.1175/1520-0442(1996)009<1518:AMFPAE>2.0.CO;2.

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.

Carney, M., and P. Cunningham, 2006: Evaluating density forecasting models. Trinity College Dublin, Department of Computer Science, Tech. Rep. TCD-CS-2006-21, 12 pp.

Cervone, G., L. Clemente-Harding, S. Alessandrini, and L. Delle Monache, 2017: Short-term photovoltaic power forecasting using Artificial Neural Networks and an Analog Ensemble. *Renewable Energy*, **108**, 274–286, https://doi.org/10.1016/j.renene.2017.02.052.

Davò, F., S. Alessandrini, S. Sperati, L. Delle Monache, D. Airoldi, and M. T. Vespucci, 2016: Post-processing techniques and principal component analysis for regional wind power and solar irradiance forecasting. *Sol. Energy*, **134**, 327–338, https://doi.org/10.1016/j.solener.2016.04.049.

Delle Monache, L., F. A. Eckel, D. L. Rife, B. Nagarajan, and K. Searight, 2013: Probabilistic weather prediction with an analog ensemble. *Mon. Wea. Rev.*, **141**, 3498–3516, https://doi.org/10.1175/MWR-D-12-00281.1.

Djalalova, I., L. Delle Monache, and J. Wilczak, 2015: PM2.5 analog forecast and Kalman filtering post-processing for the Community Multiscale Air Quality (CMAQ) model. *Atmos. Environ.*, **119**, 431–442, https://doi.org/10.1016/j.atmosenv.2015.05.057.

Fortin, V., M. Abaza, F. Anctil, and R. Turcotte, 2014: Why should ensemble spread match the RMSE of the ensemble mean? *J. Hydrometeor.*, **15**, 1708–1713, https://doi.org/10.1175/JHM-D-14-0008.1.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203–1211, https://doi.org/10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560, https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

——, and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Wea. Rev.*, **134**, 3209–3229, https://doi.org/10.1175/MWR3237.1.

——, M. Scheuerer, and G. T. Bates, 2015: Analog probabilistic precipitation forecasts using GEFS reforecasts and climatology-calibrated precipitation analyses. *Mon. Wea. Rev.*, **143**, 3300–3309, https://doi.org/10.1175/MWR-D-15-0004.1.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570, https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Hopson, T. M., 2014: Assessing the ensemble spread–error relationship. *Mon. Wea. Rev.*, **142**, 1125–1142, https://doi.org/10.1175/MWR-D-12-00111.1.

Junk, C., L. Delle Monache, S. Alessandrini, G. Cervone, and L. von Bremen, 2015: Predictor-weighting strategies for probabilistic wind power forecasting with an analog ensemble. *Meteor. Z.*, **24**, 361–379, https://doi.org/10.1127/metz/2015/0659.

Keller, J. D., L. Delle Monache, and S. Alessandrini, 2017: Statistical downscaling of a high-resolution precipitation reanalysis using the analog ensemble method. *J. Appl. Meteor. Climatol.*, **56**, 2081–2095, https://doi.org/10.1175/JAMC-D-16-0380.1.

Mason, I. B., 1982: A model for the assessment of weather forecasts. *Aust. Meteor. Mag.*, **30** (740), 291–303.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–600, https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

Nagarajan, B., L. Delle Monache, J. Hacker, D. Rife, K. Searight, J. Knievel, and T. Nipen, 2015: An evaluation of analog-based post-processing methods across several variables and forecast models. *Wea. Forecasting*, **30**, 1623–1643, https://doi.org/10.1175/WAF-D-14-00081.1.

NCAR Research Applications Laboratory, 2014: Verification: Weather Forecast Verification Utilities. R package version 1.40, accessed 10 December 2018, https://CRAN.R-project.org/package=verification.

Plenković, I. O., L. Delle Monache, K. Horvath, and M. Hrastinski, 2018: Deterministic wind speed predictions with analog-based methods over complex topography. *J. Appl. Meteor. Climatol.*, **57**, 2047–2070, https://doi.org/10.1175/JAMC-D-17-0151.1.

Sperati, S., S. Alessandrini, and L. Delle Monache, 2017: Gridded probabilistic weather forecasts with an analog ensemble. *Quart. J. Roy. Meteor. Soc.*, **143**, 2874–2885, https://doi.org/10.1002/qj.3137.

Van den Dool, H. M., 1989: A new look at weather forecast through analogs. *Mon. Wea. Rev.*, **117**, 2230–2247, https://doi.org/10.1175/1520-0493(1989)117<2230:ANLAWF>2.0.CO;2.

Wang, X., and C. H. Bishop, 2003: A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci.*, **60**, 1140–1158, https://doi.org/10.1175/1520-0469(2003)060<1140:ACOBAE>2.0.CO;2.