Probabilistic Weather Prediction with an Analog Ensemble

LUCA DELLE MONACHE

National Center for Atmospheric Research, Boulder, Colorado

F. ANTHONY ECKEL

National Weather Service Office of Science and Technology, Silver Spring, Maryland

DARAN L. RIFE

GL Garrad Hassan, San Diego, California

BADRINATH NAGARAJAN AND KEITH SEARIGHT

National Center for Atmospheric Research, Boulder, Colorado

(Manuscript received 28 September 2012, in final form 14 April 2013)

ABSTRACT

This study explores an analog-based method to generate an ensemble [analog ensemble (AnEn)] in which the probability distribution of the future state of the atmosphere is estimated with a set of past observations that correspond to the best analogs of a deterministic numerical weather prediction (NWP). An analog for a given location and forecast lead time is defined as a past prediction, from the same model, that has similar values for selected features of the current model forecast. The AnEn is evaluated for 0–48-h probabilistic predictions of 10-m wind speed and 2-m temperature over the contiguous United States and against observations provided by 550 surface stations, over the 23 April–31 July 2011 period. The AnEn is generated from the Environment Canada (EC) deterministic Global Environmental Multiscale (GEM) model and a 12–15-month-long training period of forecasts and observations. The skill and value of AnEn predictions are compared with forecasts from a state-of-the-science NWP ensemble system, the 21-member Regional Ensemble Prediction System (REPS). The AnEn exhibits high statistical consistency and reliability and the ability to capture the flow-dependent behavior of errors, and it has equal or superior skill and value compared to forecasts generated via logistic regression (LR) applied to both the deterministic GEM (as in AnEn) and REPS [ensemble model output statistics (EMOS)]. The real-time computational cost of AnEn and LR is lower than EMOS.

1. Introduction

A deterministic numerical weather prediction (NWP) model forecast can provide useful information for decision-making. Its utility, however, is fundamentally limited as it represents only a single plausible future state of the atmosphere from a continuum of possible states, resulting from imperfect initial conditions and model deficiencies that lead to nonlinear error growth during model integration (Lorenz 1963). Accurate knowledge of that continuum, the forecast probability density function (PDF), provides considerably more utility to decision-making (NRC 2006; AMS 2008; Gill et al. 2008; Hirschberg et al. 2011).

Epstein (1969) proposed to generate a forecast PDF via stochastic dynamic NWP, where the range of possible solutions is integrated forward by incorporating uncertainty into the model's prognostic equations. That approach requires computing power not currently feasible for operations. Leith (1974) proposed a Monte Carlo approximation to stochastic dynamic forecasting, referred to here as an NWP ensemble, where the deterministic NWP model is run multiple times (called

Corresponding author address: Luca Delle Monache, National Center for Atmospheric Research, Research Applications Laboratory, P.O. Box 3000, Boulder, CO 80307-3000.
E-mail: lucadm@ucar.edu

DOI: 10.1175/MWR-D-12-00281.1

ensemble members) over the valid period with plausible variations to each separate run. The NWP ensembles have been created using different model initial conditions (e.g., Toth and Kalnay 1993, 1997; Molteni et al. 1996; Bishop and Toth 1999; Houtekamer et al. 2005, 2009; Kuhl et al. 2007), parameterizations within a single model (e.g., Stensrud et al. 2000; Hacker et al. 2011), approaches (e.g., Buizza et al. 1999; Eckel and Mass 2005; Teixeira and Reynolds 2008; Bowler et al. 2008; Berner et al. 2009), numerical schemes (e.g., Thomas et al. 2002), and models (e.g., Houtekamer et al. 1996; Krishnamurthi et al. 2000; Hou et al. 2001; Wandishin et al. 2001), and coupled to ocean and land surface ensembles (e.g., Holt et al. 2009).

In this study the forecast PDF is estimated using a set of n past verifying observations corresponding to the n best analogs (past model predictions) to a current deterministic model forecast. An analog for a given location and forecast lead time is defined as a past prediction from the same model that has similar values for selected features of the current model forecast as proposed by Delle Monache et al. (2011). The verifying observation for each analog is thus a member of the *analog ensemble* (AnEn).

Several past studies explored use of analog-based methods for producing both deterministic and probabilistic weather predictions. Van den Dool (1989) generated 12-h 500-hPa height forecasts from a 15-yr dataset by finding analogs of the current analysis with past analyses over a localized area with a radius of about 900 km, and then used the 12-h subsequent analysis to each analog as a plausible forecast. His results revealed the ability of this approach to predict the forecast skill of an NWP model, as indicated by a strong spread-skill relationship in a 10-member, analog-based ensemble (see Fig. 9 of van den Dool 1989). Zorita and von Storch (1999) tested a relatively simple downscaling technique based on analogs for daily and monthly winter rainfall over the Iberian Peninsula. They found that their analog method performs similar to more complex downscaling techniques, and that it can be applied to both normally and nonnormally distributed variables because it is fully nonparametric. Hamill and Whitaker (2006, hereinafter HW06), who provide a theoretical basis of the analog approach (see section 2), found analogs for the mean of an NWP ensemble over a 25-yr reforecast dataset for probabilistic prediction of 24-h precipitation. They tested several analog-based methods and found dramatic improvement over the raw NWP ensemble, as well as skill competitive with a logistic regression (LR) technique, which is the same baseline method analyzed in this study (explained below). Messner and Mayr (2011) assessed the skill of different configurations of the analog methods proposed by HW06 in an idealized model setting and found similarly promising results, particularly for longer lead times. Klausner et al. (2009) proposed a computationally efficient ''similar day method,'' based on a historical dataset of observations for the 0–6-h prediction of near-surface wind; their approach exhibited skill superior to both a climatological and persistence forecast. Panziera et al. (2011) tested an analog approach based on radar observations for very short-term orographic precipitation predictions and found that their method performed better than persistence for lead times beyond 1 h, and better than a limited area NWP prediction for lead times up to 4 h. Delle Monache et al. (2011) proposed two postprocessing analog-based methods to improve 1–24-h NWP predictions of 10-m wind speed, which proved to drastically reduce random and systematic errors of the raw NWP prediction and considerably improve correlation between forecasts and observations. The k-nearest neighbors approach (KNN or k-NN; Fukunaga 1990), based on the concept of analogy, has been explored extensively in hydrology (Gangopadhyay et al. 2009; Hopson and Webster 2010, and references therein) and more recently to downscale seasonal weather predictions (Wu et al. 2012).

The pioneering contribution of van den Dool (1989) blazed a path for others to follow. In this paper we recognize that the analog approach is not merely useful as a calibration technique for an NWP ensemble, as performed in HW06, but also as a means to generate uncertainty (i.e., probabilistic) information from a purely deterministic forecast. Our focus then is to compare and contrast the NWP ensemble and AnEn approaches, along with LR to produce probability from a deterministic forecast, to explore the approaches' relative benefits. The AnEn has potential advantages and disadvantages relative to an NWP ensemble. One advantage may be to significantly lower the computational expense of generating an ensemble as AnEn requires only a single model forecast, as opposed to the multiple model runs of an NWP ensemble. Another advantage is that forecast uncertainty is based solely upon past observations, thereby eliminating the need to simulate all sources of NWP forecast uncertainty via sophisticated and computationally intensive techniques, and perhaps also avoiding the need for postprocessing calibration. The AnEn attempts to capture flow-dependent error growth by assigning the observed errors from similar past flows, described by the high-resolution deterministic model, to the current model forecast. A disadvantage may be the additional cost of generating a long history of model forecasts from a frozen modeling system needed for finding good analogs, even though these reforecast datasets are already produced in operational

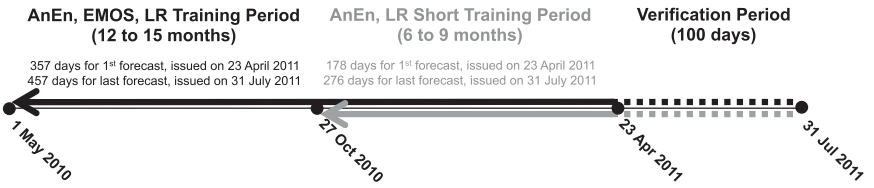

FIG. 1. Timeline of the available datasets. The arrows show the training periods over which observations and model predictions are available to produce the ensemble model output statistics (EMOS), the analog ensemble (AnEn), and logistic regression (LR). The dotted lines indicate the verification period.

centers to support successful forecast calibration (Hamill et al. 2004, 2006, 2008, 2013; HW06; Hamill and Whitaker 2007; Wilks and Hamill 2007; Hagedorn et al. 2008; Wilks 2009). Of particular interest concerning the reforecast requirement is the relative sensitivity of AnEn versus calibrated ensemble forecasts to the reforecasts' length. These are some of the issues explored in this paper.

In this work AnEn is evaluated for 0–48-h probabilistic predictions of 10-m wind speed and 2-m temperature at 550 surface stations over the contiguous United States (CONUS), over the 23 April–31 July 2011 period. Analogs for AnEn are found over the previous 12–15 months using the regional version of the Environment Canada (EC) deterministic (15 km) Global Environmental Multiscale (GEM) model. The skill of AnEn is compared to the skill of a state-of-the-science NWP ensemble, the 21-member EC Regional (33 km) Ensemble Prediction System (REPS), and LR applied to both the deterministic 15-km GEM and REPS [ensemble model output statistics (EMOS)].

In section 2 the datasets and the prediction systems used in the experiments are described followed by an analysis of the results in section 3, including statistical consistency, reliability, sharpness, resolution, and value of the probabilistic predictions. Section 4 presents a sensitivity analysis of the AnEn algorithm; section 5 discusses the results and presents conclusions.

2. Research datasets

This section describes the surface observations and the four different systems that produce 0–48-h 3-hourly predictions (initialized at 1200 UTC) utilized for the analysis presented in section 3. Figure 1 shows the timeline of the available datasets. Observations and raw model predictions are available over the 457-day (15 months) period of 1 May 2010–31 July 2011, with the verification period consisting of the last 100 days of the available data. Several of the forecast systems described below require a training dataset of past forecasts and observations (black arrow). To take full advantage of the available data and mimic real-world forecast operations, the training period increases from 12 months for the first verified forecast (initialized 23 April 2011) to 15 months for the last (initialized 31 July 2011). It is worth noting that in operational settings much longer (i.e., multiyear reforecasts; Hamill et al. 2006, 2013) training datasets may be available to further improve these forecast systems. Sensitivity of the forecast systems' performance to a shorter training dataset (6–9 months, gray arrow) is also performed and presented in section 4.

a. Observations

The observational dataset includes hourly 10-m AGL wind speed and 2-m AGL temperature measurements from 550 aviation routine weather-reporting stations (METAR, surface) collected over the 457-day period. The 550 stations are distributed throughout CONUS (Fig. 2) spanning a wide range of topographic complexity, land-use types, and weather regimes, thus allowing for a robust analysis of forecast skill.

Measurements of 2-min average 10-m wind speed and 5-min average 2-m temperature have observational error (95% confidence interval) of ±2.0 kt (or 5% if wind speed is greater than 40 kt) and ±0.1°C, respectively (NOAA 1998). To account for those errors in the verification process (e.g., Anderson 1996), verification of raw REPS output is performed only after first adding random white noise (scaled by the observation error) to each REPS member. Such a procedure is not performed for the other three prediction systems described below since the observational error is actually incorporated into their forecast process.

b. Prediction systems

1) ANALOG ENSEMBLE

The AnEn seeks to estimate the probability distribution $[f(\cdot)]$ of the observed value of the predictand variable given a model prediction, which can be represented as

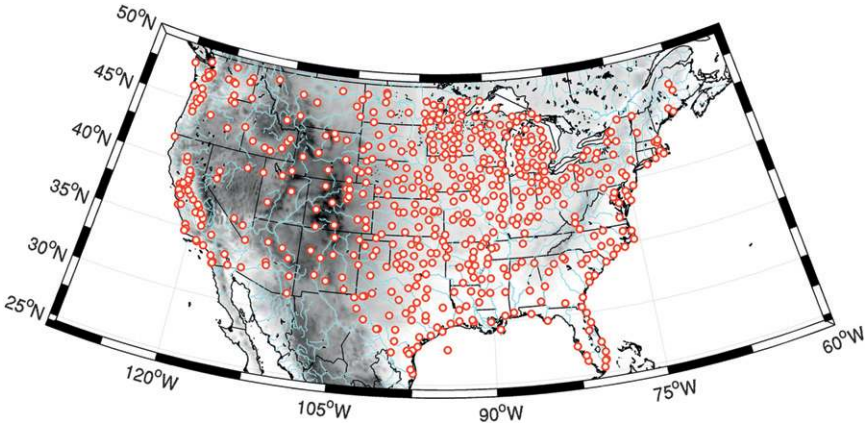

FIG. 2. Spatial distribution of the 550 stations from the aviation routine weather reports (METAR, surface), providing the observations of 10-m wind speed and 2-m temperature used in this study. Darker shading corresponds to higher terrain elevation, rivers are indicated in light blue, and the U.S. state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

$$f(y \mid \mathbf{x}^f), \tag{1}$$

where, at a given time and location, y is the observed future value of the predictand variable, and $\mathbf{x}^f = (x_1^f, x_2^f, \ldots, x_k^f)$ contains the values of k predictors from the deterministic model prediction at the same location and over a time window centered over the same time.

As shown in Fig. 3, the AnEn method generates samples of y given $\mathbf{x}^f$ via three main steps using a history of cases, called the analog training period, in which both the NWP deterministic prediction and the verifying observation are available (a minimum of 6 months in this study). Analogs are sought independently at each location and for each lead time (black square in step 1), and thus also for each time of day since only 1200 UTC forecasts are used. The best-matching historical forecasts for the current prediction are selected as the analogs (blue boxes in step 1). An analog may come from any past date within the training period (i.e., a day, week, or several months ago). Next, each analog's verifying observation is selected as a member of AnEn (green boxes in step 2). Taken all together, these observations constitute the ensemble prediction for the current forecast (orange circles in step 3).

For step 1 above, the quality of an analog (i.e., closeness of the match) is determined by the following metric, as proposed by Delle Monache et al. (2011):

$$\|F_t, A_{t'}\| = \sum_{i=1}^{N_v} \frac{w_i}{\sigma_{f_i}} \sqrt{\sum_{j=-\tilde{t}}^{\tilde{t}} (F_{i,t+j} - A_{i,t'+j})^2}, \tag{2}$$

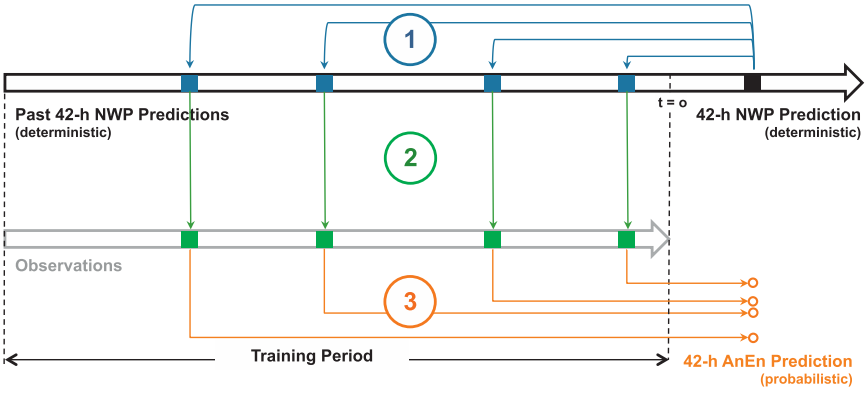

FIG. 3. Schematic representation of the process for finding four members of the analog ensemble (AnEn) at one forecast lead time. A detailed description of the three main steps is found in section 2b(1).

where F_t is the current NWP deterministic forecast valid at the future time t at a station location; $A_{t'}$ is an analog at the same location and with the same forecast lead time but valid at a past time t'; N_v and w_i are the number of physical variables used in the analogs search and their weights, respectively; σ_{fi} is the standard deviation of the time series of past forecasts of a given variable at the same location and forecast lead time; $\tilde{t}$ is equal to half the number of additional times over which the metric is computed; and $F_{i,t+j}$ and $A_{i,t'+j}$ are the values of the forecast and the analog in a time window for a given variable. Section 4 explores the sensitivity to key choices of the analog ensemble algorithm proposed below.

Similarly to Delle Monache et al. (2011), $\tilde{t}$ is set to 1 so that the time window is ±3 h given the 3-h forecast interval. The weights w_v are also set to one. In searching for analogs for 10-m wind speed predictions, N_v is equal to 4 and includes 10-m wind speed and direction, 2-m temperature, and surface pressure (which were chosen as a reasonable set of predictors out of the 15-km GEM output, as confirmed by the high-quality probabilistic predictions presented in section 3). For 2-m temperature analogs, N_v is equal to 3 and includes 2-m temperature and 10-m wind speed and direction. For the latter, the appropriate definition of difference for a circular variable has been taken into account.

The analog searching algorithm is highly flexible and allows the search to occur over a time window of any specified width; however, the current 3-hourly output for GEM dictates that, with the choice of $\tilde{t}$ equal to 1, the window's width is 6 h. The idea is to find past forecasts (at a specific location) that predicted similar temporal trends and not simply values of the forecasted physical variables (i.e., the predictors) at one lead time. Including multiple predictor variables that exhibit correlations to the predictand further helps distinguish the analogs by perhaps identifying specific weather regimes. To rule out possible differences in skill related to sampling error when comparing AnEn performance to that of the 21-member REPS, AnEn uses only the 21 best analogs. No calibration is performed on AnEn.

The deterministic NWP prediction used to generate AnEn is the EC Regional 15-km GEM, which used 58 vertical levels up to 20 September 2010 and 80 vertical levels thereafter. The implications of this choice in terms of computational costs when comparing REPS with AnEn are discussed in the concluding section.

While the basic steps of AnEn have parallels with the analog ensemble approach in HW06, there are many distinct differences to note, as follows:

- AnEn is generated using a deterministic dynamical model prediction, rather than being based on the mean of an NWP ensemble as in HW06. So instead of being seen as a postprocessing method to calibrate an existing ensemble, in this study the AnEn is a procedure to generate an ensemble.
- Here analogs are searched independently for every location and over a 3-point time window, whereas in HW06 the analog matching is performed over a limited-sized 16-point region independently for every forecast lead time.
- While in this study the analog metric [Eq. (2)] is multivariate, HW06 found the best performance of the analog ensemble with a metric computation based only on the variable of interest (i.e., precipitation in their case).
- The AnEn searches for analogs throughout the available historical dataset, whereas HW06 limited their search to a ±45-day window around the date of the forecast in previous years.

2) LOGISTIC REGRESSION

Forecasts are formulated using logistic regression, which is a model output statistics (MOS) technique specifically designed to produce probabilistic forecasts (Wilks 2006). While the mechanics of LR are quite different from AnEn, both approaches consider the past relationship between predictor variable(s) and the predictand to produce a forecast of the predictand given the predictors' values in the current forecast cycle. One difference with LR is that the predictand is the probability of an event, such as the probability of 10-m wind speed greater than $5\,\mathrm{m\,s^{-1}}$, rather than the value (or PDF) of 10-m wind speed itself. A nonlinear function is fit to past pairs of the predictor(s) and the predictand, which as an observed value takes on a probability of either 1.0 (event occurred) or 0.0 (event did not occur) (Wilks 2006):

$$p = \frac{e^{(b_0 + b_1 x_1 + \cdots + b_K x_K)}}{1 + e^{(b_0 + b_1 x_1 + \cdots + b_K x_K)}}, \tag{3}$$

where p is the probability of the event, x_K are the K predictor variables, and b_K are the regression coefficients.

To make a fair comparison with AnEn, LR uses the same training dataset (i.e., deterministic 15-km GEM forecasts and METAR station observations). Also like AnEn, training (and application) for LR is performed separately for each location, each forecast lead time, and for each forecast initialization within the verification period using all historical data available at forecast initialization time. A difference from AnEn is that in deriving the probability of an event, LR takes into account the complete history of forecasts and observations whereas

AnEn is based only on a small subset of forecasts within the training dataset that closely match current forecast conditions. Given a much longer training period than available in this study, LR may be trained over seasonally relevant subsets of the historical data, but would still in general include much more data than AnEn. Furthermore, the two methods may have different sensitivities to training data length, as discussed in section 4.

Recall that the four available 15-km GEM predictors are 10-m wind speed, 10-m wind direction, 2-m temperature, and surface pressure. For 10-m wind speed probabilistic predictions, LR was found to perform best using all four predictors, but with a power transformation (square root) of 10-m wind speed to make its distribution more normal (Hamill et al. 2008). For 2-m temperature probabilistic predictions, only the 2-m temperature predictor is used since including any (or any combination) of the other predictors produced inferior results.

When an event occurs infrequently within the training data, LR is prone to producing poorly skilled predictions. Predictand values in the training data are predominantly 0.0, with only a very few values of 1.0, making it difficult to fit a dependable curve. The same issue exists for the inverse situation of an event that occurs frequently (i.e., training data have mostly predictand values of 1). This issue, of course, is the general challenge of probabilistic forecasting of rare events, which also exists for AnEn and is discussed further in section 3. For LR, one approach that can alleviate this problem is weighting rare or extreme forecast events within the training data, as described in Hamill et al. (2008). This type of weighting, as well as other related techniques, was thoroughly tested and provided no significant improvement for the variables of interest in this study, so only the standard logistic regression technique was employed.

3) REPS

The direct output of REPS is used in this study to represent forecasts from a state-of-the-science, short-range NWP ensemble. First described in Li et al. (2008), the system was upgraded for its 2011 operational implementation, which is the version used here. The REPS consists of 21 72-h forecasts that use a North American domain of the GEM model, with grid spacing of 0.3° (~33 km) and 28 vertical levels. All REPS members have the same model configuration but apply perturbations to physical tendencies (Buizza et al. 1999). Each REPS member gets initial conditions (used for cold starts) and 3-hourly boundary condition updates from the direct model output of a different member of the 21-member Global Ensemble Prediction System (GEPS), which is run using a grid spacing of ~66 km and 40 vertical levels. The GEPS initial conditions are generated with the ensemble Kalman filter technique using a 12-h update cycle and 96 ensemble members (Houtekamer et al. 2009). In addition to perturbations to physical tendencies, GEPS simulates model uncertainty using multiple physics as well as the kinetic energy backscatter approach (Shutts 2005).

4) ENSEMBLE MODEL OUTPUT STATISTICS

Calibrated REPS forecasts are formulated via a form of the ensemble model output statistics (EMOS) technique, originally proposed by Gneiting et al. (2005) using multiple linear regression with predictors being the ensemble members' forecast values as well as the ensemble spread. This study follows Hamill et al. (2008) in performing EMOS with LR (as described above) using two predictors—the REPS ensemble mean forecast and the square root of ensemble spread. The EMOS training data periods are broken up by location and forecast lead time to match both AnEn and LR as described above. The 10-m wind speed ensemble mean predictor is also transformed by taking its square root, as performed for LR.

3. Results

The performance of the probabilistic forecast systems described in section 2b is compared by examining key attributes of probabilistic predictions, namely statistical consistency, reliability, sharpness, resolution, and value. While these attributes and their associated metrics are briefly reviewed below, thorough descriptions can be found in Jolliffe and Stephenson (2003) and Wilks (2006). A detailed analysis with a different dataset of the analog ensemble mean results can be found in Delle Monache et al. (2011).

Forecasts for 10-m wind speed and 2-m temperature from 0 to 48 h (at 3-h increments) initialized daily at 1200 UTC are verified from 23 April to 31 July 2011 (roughly the last 3 months of the 15-month research dataset; see Fig. 1 for details) against the observations described in section 2a. The forecast probability for an event threshold is computed from an ensemble's members using the rank method with uniform probability in each rank (Hamill and Colucci 1997).

a. Statistical consistency

An ensemble is statistically consistent when its members are indistinguishable from the truth (i.e., the PDF from which the members are drawn is *consistent* with the PDF from which the truth is drawn; Anderson 1996). If so, an observation ranked among the corresponding

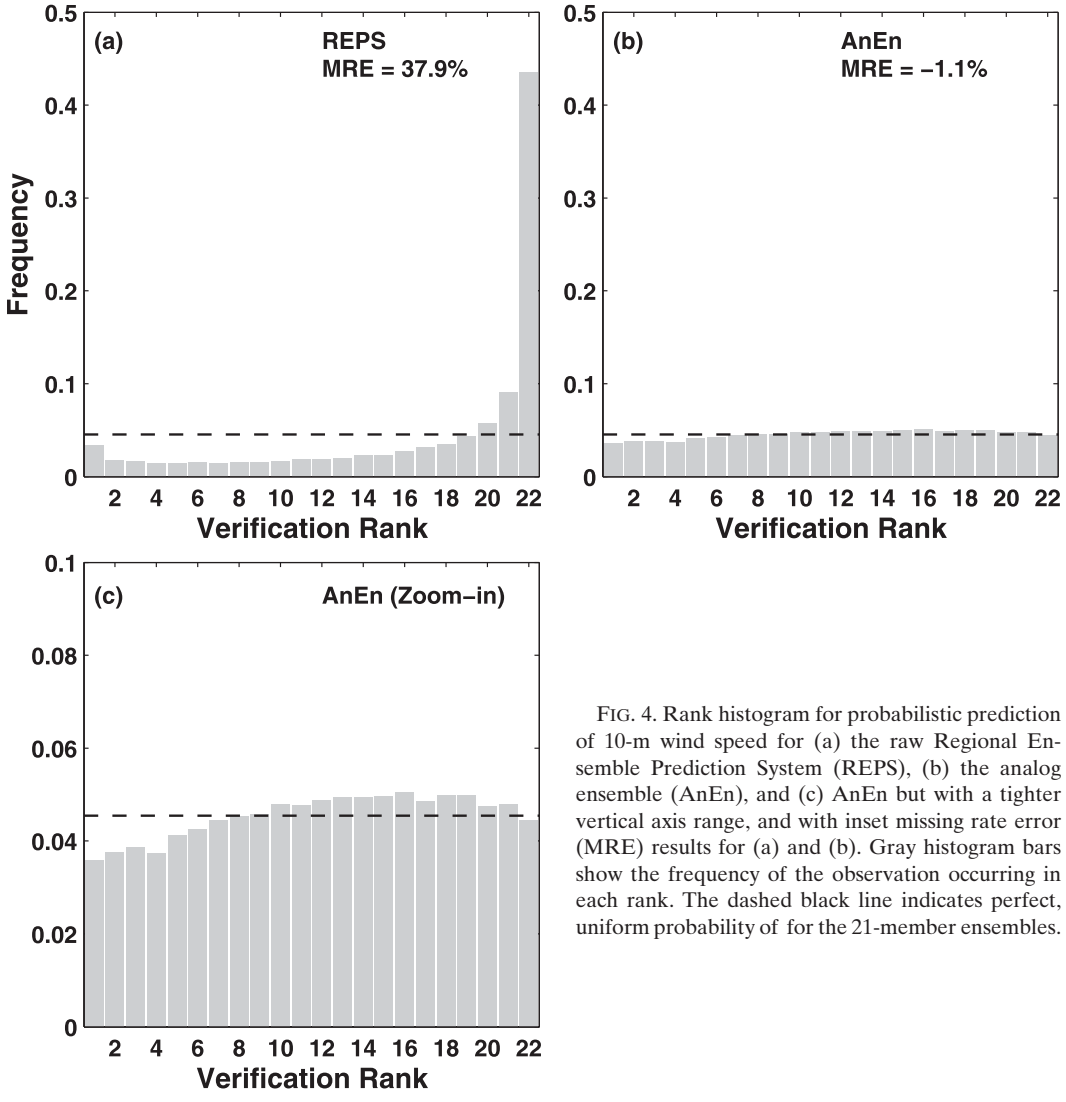

FIG. 4. Rank histogram for probabilistic prediction of 10-m wind speed for (a) the raw Regional Ensemble Prediction System (REPS), (b) the analog ensemble (AnEn), and (c) AnEn but with a tighter vertical axis range, and with inset missing rate error (MRE) results for (a) and (b). Gray histogram bars show the frequency of the observation occurring in each rank. The dashed black line indicates perfect, uniform probability of for the 21-member ensembles.

ordered ensemble members is equally likely to take any rank i in the range $i = 1, 2, \ldots, n + 1$, where n is the number of ensemble members. Collecting the rank of the observation over a number of cases and plotting the results generates a rank histogram, which tests as flat [i.e., uniform rank probability of $1/(n + 1)$] for a statistically consistent ensemble (Anderson 1996; Hamill and Colucci 1997; Talagrand et al. 1997). However, that test is a necessary but not sufficient condition for statistical consistency (Hamill 2001).

Figure 4 shows the rank histograms for 9-h forecast 10-m wind speed for REPS and AnEn as well as 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, $2/(n + 1)$. Note that rank histograms, as well as the other statistical consistency plots below, are designed for ensemble forecasts so only REPS and AnEn results are displayed in this subsection. The REPS rank histogram (Fig. 4a) reveals a severe lack of statistical consistency, with a notable negative forecast bias (observed wind speed often greater than all REPS members) as well as an underspread condition (highest probabilities in the two outer ranks). The AnEn has much better statistical consistency, displayed by a nearly uniform rank histogram (Fig. 4b) with a slightly overspread condition (MRE equal to −1.1%), as is more evident in Fig. 4c where a tighter vertical axis range is used. Note that 2-m temperature and other forecast lead times (not shown) indicated similar results.

Examination of statistical consistency over all forecast lead times is accomplished following the general definition by Talagrand et al. (1997) that the mean square error of the ensemble mean should match the average

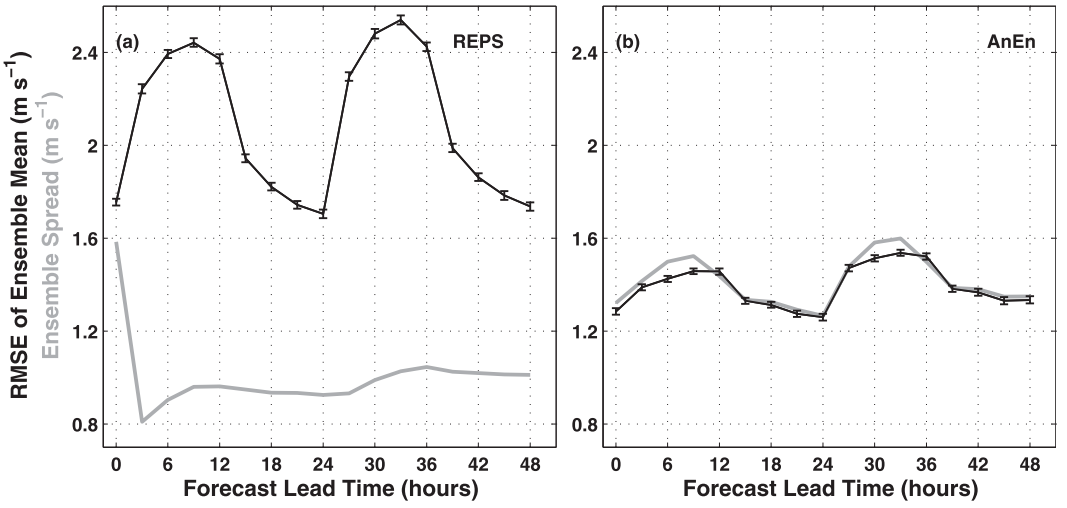

FIG. 5. Dispersion diagram for probabilistic prediction of 10-m wind speed (a) for REPS and (b) AnEn. The black line is the root-mean-square error (RMSE) of the ensemble mean, while the gray line is the average ensemble spread (m s^{-1}). The 95% bootstrap confidence intervals for RMSE are indicated by the error bars.

ensemble variance over a large number of verifications. Comparing the square root of those two statistics (to display results with the predicted variables' natural unit) over all forecast lead times produces a dispersion diagram that reveals whether an ensemble is properly dispersive (i.e., able to simulate average forecast error growth).

Figure 5 shows the dispersion diagrams for 10-m wind speed. The underdispersion of REPS is evident (Fig. 5a), and the above conclusion of good statistical consistency by AnEn is now evident at all lead times, with slight excess spread by AnEn at hours 6, 9, 30, and 33 (Fig. 5b). The dramatic drop in REPS spread from analysis time (i.e., hour 0) to the 3-h lead time can be explained by the cold-start initialization of REPS. Each REPS member uses a GEPS member's initial condition, with no data assimilation or model spinup. Much of the diversity in the initial conditions likely collapses in the first few REPS time steps as the solutions adjust to the new model grid and converge to the new attractor (i.e., a set of states toward which the dynamical system asymptotically approach in the course of its evolution; Lorenz 1993). Figure 5b shows that the good statistical consistency of AnEn comes from not only higher spread than REPS but also from a lower RMSE of the ensemble mean—achieved from use of a higher-resolution NWP model as well as downscaling (i.e., adding information at smaller scales via the observations that compose AnEn). Note that the downscale benefit is realized also by EMOS, while both the downscaling and higher model resolution benefits are realized by LR, making them more fair comparisons to AnEn, as seen below.

A more in-depth assessment of statistical consistency at a particular forecast lead time is possible with a binned spread-skill plot (Fig. 6), which compares ensemble spread to RMSE of the ensemble mean over small class intervals of spread rather than just considering the overall average spread as in the dispersion diagram (e.g., van den Dool 1989; Wang and Bishop 2003). Good statistical consistency now requires the two metrics to match at all values of ensemble spread (i.e., results along the plot's 1:1 diagonal). For forecast hour 42 of 10-m wind speed, REPS forecasts are highly underspread for all spread values (Fig. 6a), while AnEn exhibits a much better spread-skill relationship with a slight conditional bias in the second moment of the forecast PDF—underspread at smaller values and overspread at higher spreads (Fig. 6b). This conditional bias is an effect of the limited sampling by AnEn (given the 21 members and the finite historical dataset available) that can only be seen when the analysis is stratified (i.e., when ensembles that satisfy a certain criteria are evaluated separately; Siegert et al. 2012). While this bias can be effectively corrected via postprocessing calibration, only unaltered AnEn results are presented since they are fairly well calibrated. This result (which is similar at other lead times) indicates that AnEn is indeed able to capture the flow-dependent forecast uncertainty since the AnEn spread dependably reflects the forecast error variance.

b. Reliability

Ideally, a large set of 30% probability forecasts will verify with a 30% occurrence rate of the event (called the observed relative frequency). In perfectly reliable

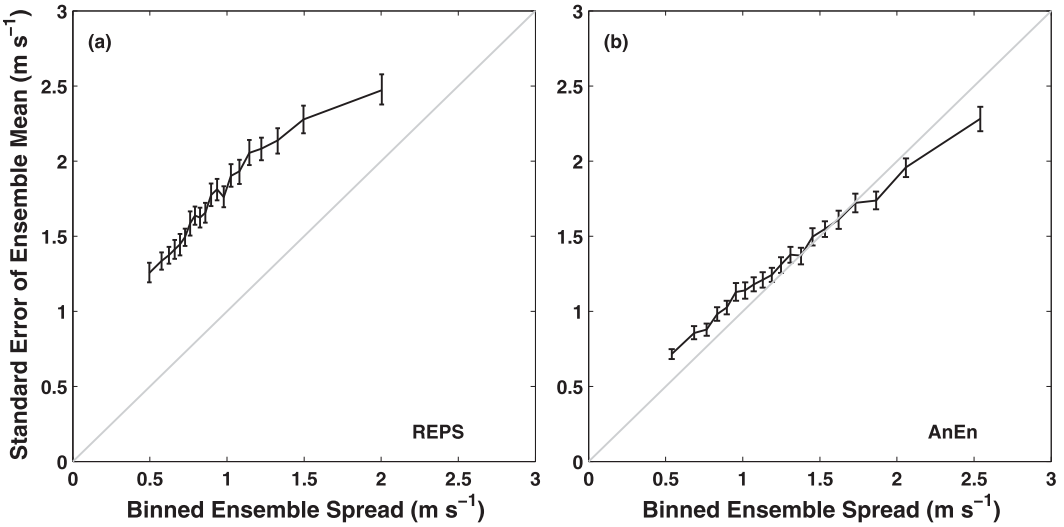

FIG. 6. Binned spread-skill plot for forecast hour 42 of 10-m wind speed for (a) REPS and (b) AnEn. The error bars indicate the 95% bootstrap confidence interval, while the diagonal 1:1 line represents the perfect spread-skill line. For each plot, ensemble spread is binned into 20 equally populated class intervals.

(or calibrated) forecasts, the observed relative frequency equals the forecasted probability for any given level of probability, resulting in the 1:1 diagonal line on a reliability diagram that plots class intervals of forecast probability against observed relative frequency (Jolliffe and Stephenson 2003; Wilks 2006).

Figure 7 shows reliability results (black lines with error bars) for the event of 10-m wind speed greater than 5 m s^{-1} at forecast hour 9. The REPS forecasts are the least reliable, with notable underestimation (overestimation) of the observed relative frequency below (above) approximately 0.7 (Fig. 7a). The EMOS forecasts have rather good reliability (Fig. 7b), with imperfections perhaps from limited training of the calibration routine and/or error variations between the dependent and independent data.

Both LR and AnEn (Figs. 7c,d, respectively) forecasts are also imperfect but exhibit roughly the same degree of good reliability as EMOS. Results were consistent also at other forecast lead times and thresholds. This result is important in this study's comparison of the NWP and analog ensemble approaches, and is the reason why EMOS was produced. A revealing comparison of any two ensembles is possible once their forecasts are similarly reliable and have been trained (and/or calibrated) using the same history of forecasts and observations. Poor reliability, associated with systematic errors, can mask the ability of an ensemble to predict the random forecast error, so that the intrinsic quality of an ensemble may only be evident after a simple calibration is applied.

Figure 8 is similar to Fig. 7, except for forecast hour 33 and for the prediction of 2-m temperature less than 15°C, chosen colloquially as a jacket versus no-jacket weather threshold. Conclusions concerning the forecast systems' reliability are about the same with the exception that REPS, while still exhibiting the lowest reliability, is far more reliable compared to the wind speed forecasts. This may be a result of the 33-km GEM having better skill at predicting 2-m temperature at the METAR stations compared to 10-m wind speed prediction (as confirmed below in Fig. 11) and/or REPS being better able to simulate the forecast uncertainty in the GEM forecast.

c. *Sharpness*

A sharper (more narrow) forecast PDF has a greater concentration of probability density and produces probability values more toward the extremes (i.e., close to 0% or 100%) for any given event threshold. Sharpness, which is a property of the forecasts only, is diagnosed in a reliability diagram by plotting how often (relative frequency) each class interval of probability is used (the gray lines with square markers in Fig. 7). A sharper forecast leads to better resolution (see next subsection) if the forecasts are reliable (Gneiting et al. 2004).

For forecast hour 9 and 10-m wind speed greater than 5 m s^{-1}, the REPS forecasts (Fig. 7a) are very sharp with the majority occurring in the 0%–10% range, but this is due to overconfidence as indicated by the poor reliability. The EMOS forecasts (Fig. 7b) have lower yet trustworthy sharpness resulting from the calibration's correction of REPS overconfidence. AnEn sharpness (Fig. 7d) is comparable to EMOS (Fig. 7b) and LR forecasts (Fig. 7c). In Fig. 8 (for forecast hour 33 and for

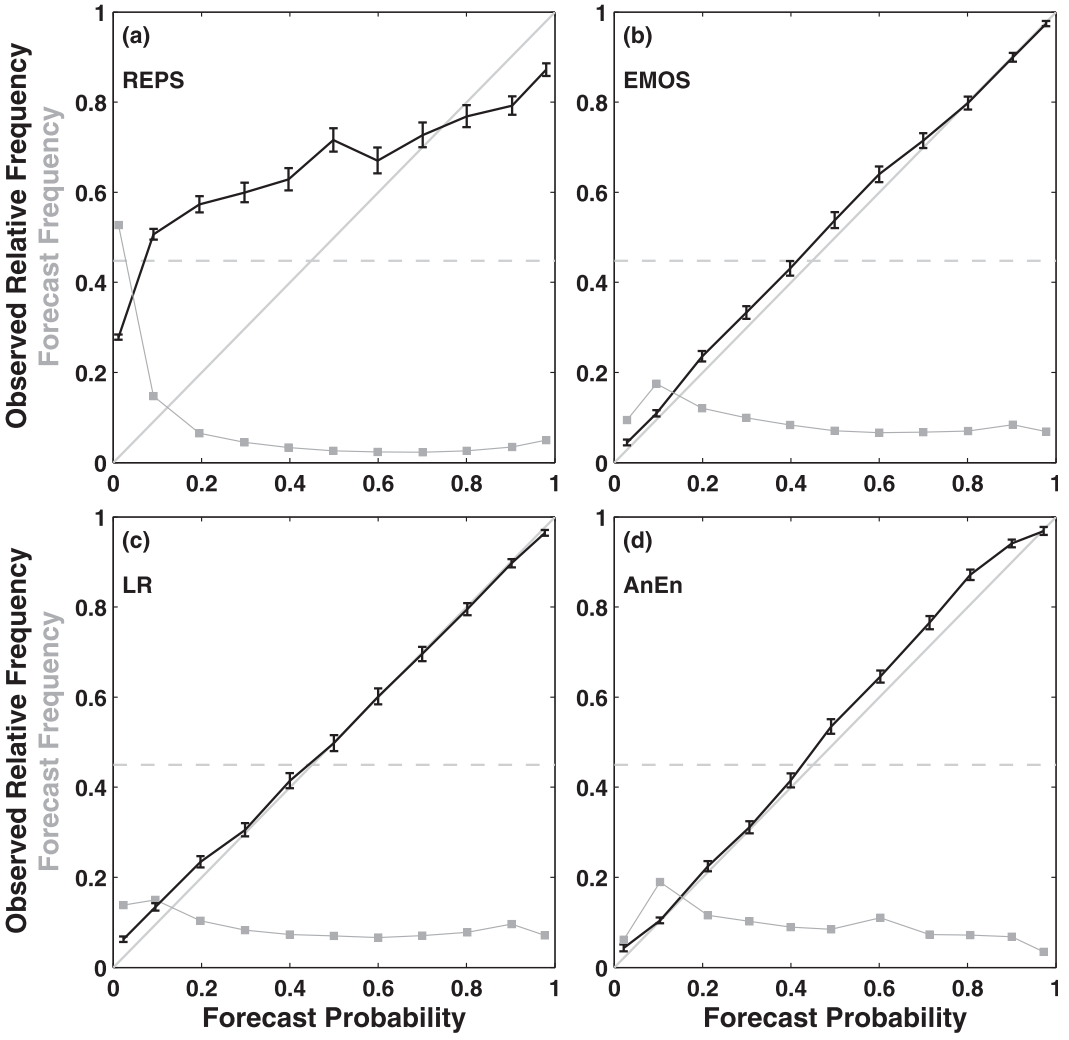

FIG. 7. Reliability (black lines with vertical error bars) and sharpness (gray lines with square marks) for (a) REPS, (b) ensemble model output statistics (EMOS), (c) logistic regression (LR), and (d) AnEn. Results are shown for forecast hour 9 and 10-m wind speed greater than 5 m s^{-1}. The horizontal dashed line represents the event's observed frequency over the verification period (i.e., sample climatology), while the diagonal 1:1 line represents the perfect reliability. The error bars indicate the 95% bootstrap confidence interval.

2-m temperature below 15°C) the relative differences in sharpness among the four prediction systems are minimal compared to the wind speed analysis.

d. Resolution and value

Resolution measures the forecasts' ability to a priori sort out when an event occurs or not (Murphy 1973). Probability forecasts with perfect resolution forecast 100% on occasions when the event occurs and forecast 0% when the event does not occur. The Brier skill score (BSS), which is the RMSE of probabilistic forecasts, can be broken up into reliability, resolution, and uncertainty (Wilks 2006). Figure 9c shows the resolution for forecasts of 10-m wind speed greater than 5 m s^{-1} by the four prediction systems. Note that the uncertainty, which depends solely on the sample climatology and is the highest possible value of resolution, is the orange line plotted in Fig. 9a. The AnEn, EMOS, and LR display similar and greatly superior ability to resolve this event at all lead times when compared to REPS, which, as discussed above, is due to higher model resolution (for AnEn and LR) and downscaling (for AnEn, EMOS, and LR). The AnEn, EMOS, and LR have very similar resolution at most lead times, with the differences being not statistically significant. The BSS results (Fig. 9a) yield similar conclusions concerning the relative skill (as well as value as explained below) of the forecast systems.

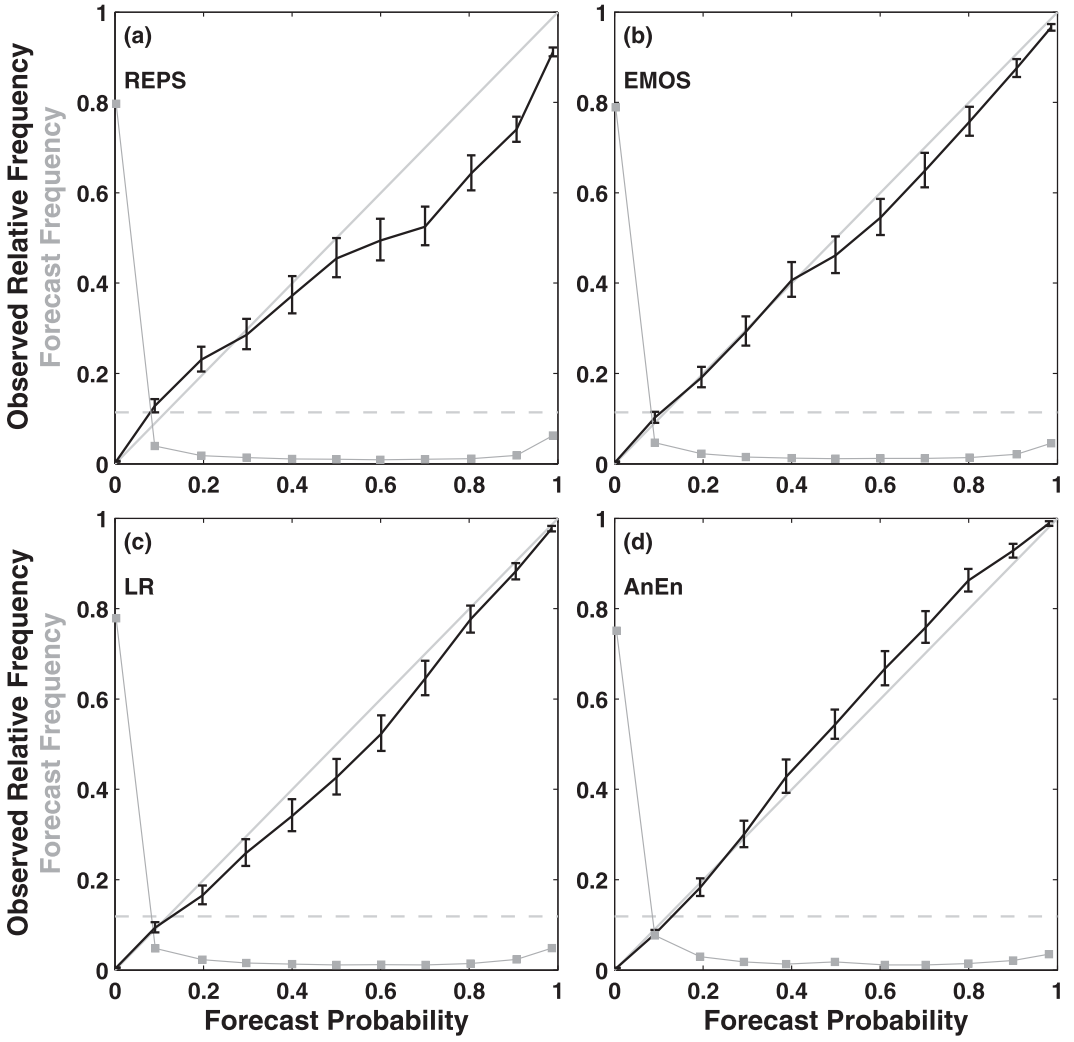

FIG. 8. As in Fig. 7, but for forecast hour 33 and for 2-m temperature below 15°C.

Signal detection theory (Mason 1980) can analyze whether probabilistic forecasts, when translated into binary decisions, may be useful to the end user, and is employed here to examine the relative value of the forecasts to individual users. As explained by Jolliffe and Stephenson (2003), the BSS presents the lower bound of overall value while the upper bound is revealed by the relative operating characteristic (ROC) skill score (ROCSS). The ROCSS is based on the ROC curve (Mason 1982), which plots the false alarm rate (false alarms divided by total nonoccurrences of the event) against the hit rate (correct forecasts divided by total occurrences of the event) to show the forecast's ability to discriminate. The ROC curve (as well as the ROCSS) thus depends upon resolution and not reliability, and the area under the ROC curve, known as the ROC score, conveys overall forecast value (Mason and Graham 1999). The ROCSS translates the ROC score into a standard skill score so that a ROCSS equal to 1 comes from perfect forecasts and a ROCSS lower than 0 indicates lower performance than climatological forecasts.

Figure 10 shows the ROCSS results over all lead times for forecasts of 10-m wind speed greater than $5\,\mathrm{m\,s^{-1}}$ (a common event with sample climatology varying from 18.1% at 1200 UTC to 44.9% at 2100 UTC) as well as $10\,\mathrm{m\,s^{-1}}$ (a rare event with sample climatology varying from 0.6% at 1200 UTC to 3.4% at 2100 UTC). For the $5\,\mathrm{m\,s^{-1}}$ event threshold (Fig. 10a), AnEn, EMOS, and LR exhibit a very similar ROCSS, which agrees with the BSS resolution term (Fig. 9c). For the $10\,\mathrm{m\,s^{-1}}$ event threshold (Fig. 10b), AnEn has higher ROCSS than both EMOS and LR, although only the differences from LR are statistically significant. The AnEn superiority over EMOS likely stems from AnEn's use of the 15-km GEM

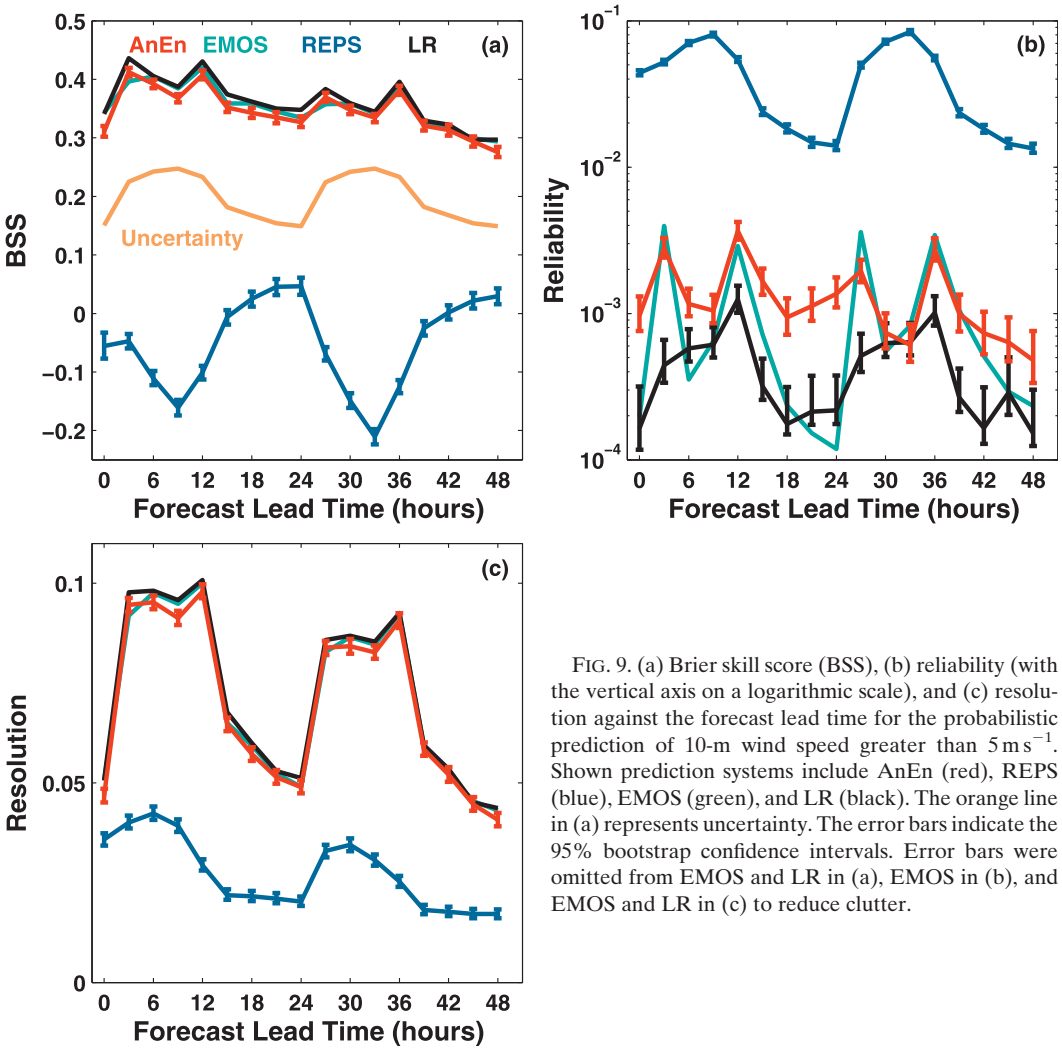

FIG. 9. (a) Brier skill score (BSS), (b) reliability (with the vertical axis on a logarithmic scale), and (c) resolution against the forecast lead time for the probabilistic prediction of 10-m wind speed greater than $5\,m\,s^{-1}$. Shown prediction systems include AnEn (red), REPS (blue), EMOS (green), and LR (black). The orange line in (a) represents uncertainty. The error bars indicate the 95% bootstrap confidence intervals. Error bars were omitted from EMOS and LR in (a), EMOS in (b), and EMOS and LR in (c) to reduce clutter.

versus EMOS's use of the 33-km GEM. AnEn is able to identify smaller-scale flows, as well as their uncertainty, that EMOS cannot resolve as well. AnEn's superiority to LR for the $10\,m\,s^{-1}$ event threshold is likely due primarily to LR's difficulty is fitting a dependable regression line in the case when the event is rarely observed within the training dataset. AnEn apparently may not suffer as severely in forecasting rare events, perhaps because it avoids some loss of information that LR may experience when creating a binary observation for the event.

For 2-m temperature less than 15°C (a common event with sample climatology varying from 42.3% at 1200 UTC to 12.0% at 2100 UTC), AnEn, EMOS, and LR forecasts again display similar skill (Fig. 11). The REPS results are included here since they are competitive at some forecast lead times, which supports the conclusion from the reliability diagrams (Fig. 8) that REPS performs relatively better for 2-m temperature forecasts.

Examining value in more detail at a specific forecast lead time can be done with an economic value diagram, which incorporates the cost/loss decision model (Richardson 2000; Jolliffe and Stephenson 2003). A value score (VS; Wilks 2001) is computed using ROC results combined with hypothetical costs C for a user to protect against a weather event and losses L incurred by failing to protect against an event occurrence. The VS, essentially a skill score for economic expense, is computed across all C/L ratios and thus applies generically to a variety of potential users and not to any specific level or type (monetary or other) of user expenses. However, a key assumption is that the users are normative—they consistently take action when the risk (i.e., forecast probability) exceeds the user's risk tolerance (i.e., the C/L ratio), thus minimizing expenses over many cases (Thompson 1950). That assumption means that similar to the ROCSS, the economic value diagram provides the upper bound of value.

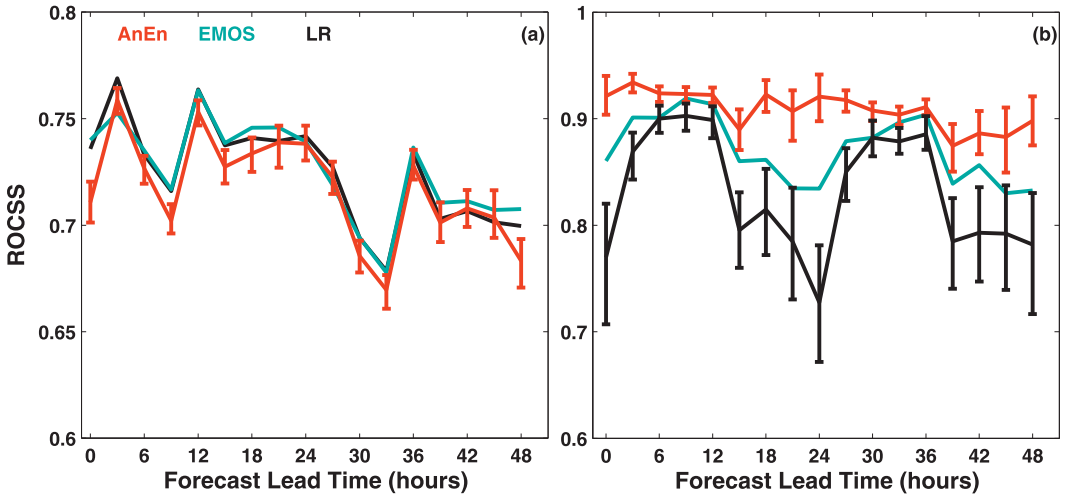

FIG. 10. Relative operating characteristic skill score (ROCSS) against the forecast lead time for the probabilistic prediction of 10-m wind speed greater than (a) $5\,m\,s^{-1}$ and (b) $10\,m\,s^{-1}$. Note the different range of the vertical axis. Shown prediction systems include AnEn (red), EMOS (green), and LR (black). The error bars indicate the 95% bootstrap confidence intervals. Error bars were omitted from EMOS and LR in (a) and from EMOS in (b) to reduce clutter.

Figure 12 shows two sample economic value diagrams for forecast hour 42 that typify results found at other lead times and event thresholds. Figure 12a supports the same general conclusions as above concerning relative value of the four predictive systems for forecasts of 10-m wind speed greater than $5\,m\,s^{-1}$. Figure 12b confirms that the result of Fig. 11 for 2-m temperature lower than 15°C (i.e., overall AnEn value equivalent to EMOS and LR) is true also for each value of the user C/L ratio.

4. Sensitivity analysis of the analog ensemble

The sensitivity to a number of parameters and implementation options in Eq. (2) and design choices for AnEn can be summarized as follows:

- Values for $\tilde{t}$ (the half-width time windows over which squared differences between analog and forecast values are computed for a given location) in the set $\{i\}_{i=0,1,\ldots,6}$, $i \in \mathbb{N}$, were tested, resulting in small differences between the different runs, with a value of $\tilde{t} = 1$ producing the best results (based on the highest correlation and lower RMSE of the ensemble mean across all the available observations and forecast lead times and the prediction of 10-m wind speed and 2-m temperature, and the fact the analog ensemble has the tendency of preserving statistical consistency regardless of the algorithmic options chosen, as explained at the end of section 5). For this reason, as in Delle Monache et al. (2011), $\tilde{t} = 1$ was used. Because the data used here have a 3-hourly frequency, this correspond to a comparison of the analog and prediction over a 6-h window that is able to capture the relevant information in terms of the predicted value and its trend, based on the results shown in section 3. Different datasets may have a different optimal value for $\tilde{t}$.
- As described in section 2, the analog predictors for 10-m wind speed predictions include 10-m wind speed and direction, 2-m temperature, and surface pressure, while for 2-m temperature the predictors are 2-m temperature and 10-m wind speed and direction. These choices were determined by selecting among the combinations of available variables the ones resulting in the

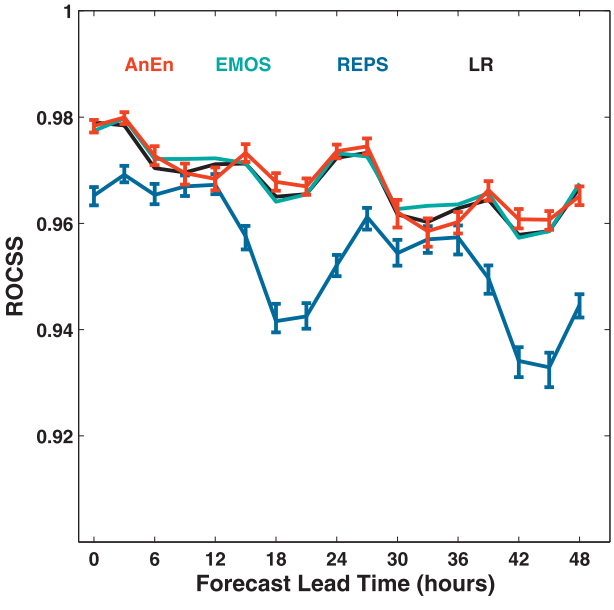

FIG. 11. As in Fig. 10, but for 2-m temperature lower than 15°C.

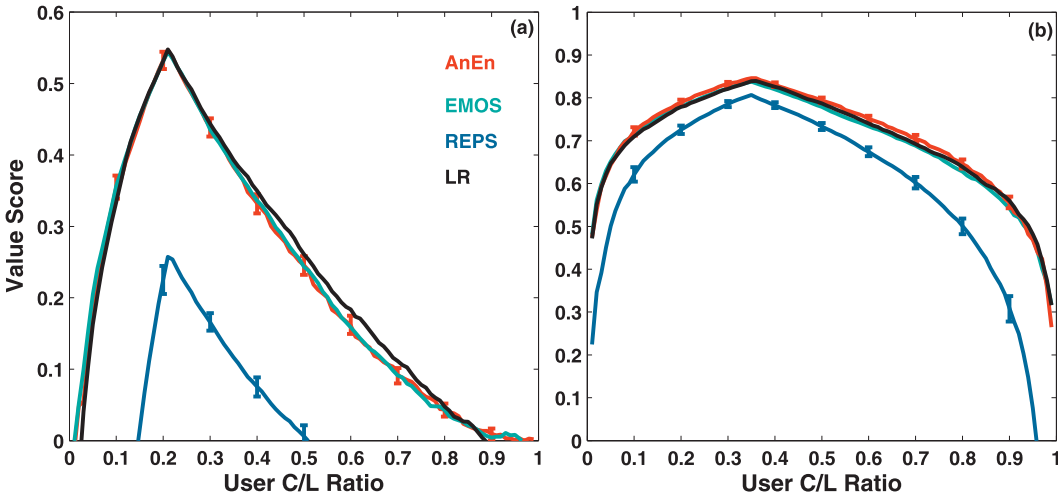

FIG. 12. Economic value score for forecast hour 42 for (a) 10-m wind speed greater than 5 m s^{-1} and (b) 2-m temperature lower than 15°C. Note the different range of the vertical axis. Shown prediction systems include AnEn (red), REPS (blue), EMOS (green), and LR (black). The 95% bootstrap confidence intervals are indicated by the error bars. Error bars were omitted from EMOS and LR to reduce clutter.

overall lowest RMSE and highest correlation of the ensemble mean prediction with the observations, as done for the choice of the width of the time window (see above).

- In this analysis we chose to run AnEn with 21 members to rule out possible difference in skill related to sampling error when comparing to REPS.
- No attempt was made to find optimal values for the weights w_i.

To explore the impact on AnEn from the NWP model resolution, AnEn is formulated and verified exactly as described above except using the REPS 20th ensemble member as the NWP model forecast from which analogs are found. These ensemble forecasts are called AnEn33 since they are generated with the 33-km GEM versus the 15-km GEM used by AnEn. Figure 13a not surprisingly shows that AnEn33 (dashed red line) performs worse than AnEn (solid red line). What is interesting is that AnEn33 results are not much worse than EMOS (see Fig. 10a, green line). Considering that AnEn33 real-time processing cost is 1/21 of EMOS (i.e., real-time run of 1 REPS member versus 21 members), AnEn33 is providing nearly all the value of EMOS (the well-calibrated NWP ensemble) at a much lower computational cost. As shown in Fig. 13b, these conclusions hold similarly for LR33 (black lines), which is LR run using the REPS 20th ensemble member.

AnEn (as well as other forecast techniques based on historical data) may be improved by using a larger training dataset constructed via reforecasting, which requires a large one-time expense but only a small increase in real-time processing from the extra searching for analogs. While the impact on AnEn of increased training could not be tested because of data limitations, the impact of decreased training is instead explored. The AnEnShort formulation is similar to AnEn except that it uses a shortened training data period that does not include the first 6 months of the original set (as shown in Fig. 1). Figure 14a shows that AnEnShort (red dashed line) performance falls below AnEn (solid red line), indicating that significant improvements in the analog ensemble may be achieved with even a modest increase in the available training data (Hamill et al. 2006). A portion of the improvement from AnEnShort to AnEn may be due to the increase in training data from 6+ months to 12+ months, which allows AnEn to train with more same-season data for each current forecast. Interestingly, as shown in Fig. 14b, LR results (black lines) show smaller improvements than AnEn when going from the short to the full training dataset, perhaps indication that a longer training dataset (e.g., multiyear) could be more beneficial to AnEn than to LR or EMOS.

5. Discussion and conclusions

This study compares an analog ensemble (AnEn), a 21-member system generated using the past forecasts and verifying observations of deterministic 15-km Global Environmental Multiscale (GEM) model runs, to a state-of-the-science numerical weather prediction (NWP) ensemble, the Environment Canada Regional Ensemble Prediction System (REPS) that consists of 21 runs of a 33-km version of GEM. For fair comparison, the direct output from REPS is calibrated to produce ensemble

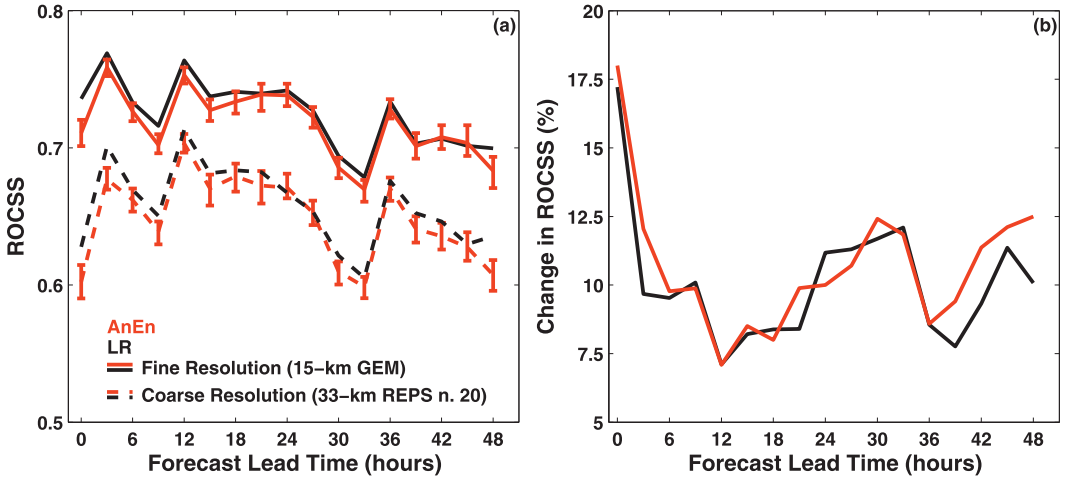

FIG. 13. (a) As in Fig. 10a, but without EMOS, and with the addition of AnEn and LR driven by a coarser numerical weather prediction model (dashed lines), and (b) the percentage change of ROCSS going from the coarse-resolution to the fine-resolution case. The 95% bootstrap confidence intervals are indicated by the error bars. Error bars are shown for AnEn only [in (a)] to reduce clutter.

model output statistics (EMOS) forecasts using the same historical data available to AnEn. A fourth system, logistic regression (LR), generates probabilistic forecasts from the 15-km GEM to provide another fair comparison with AnEn. An important difference between previous implementations of analog-based ensemble methods (e.g., HW06) and the AnEn method proposed here is that while the former are postprocessing procedures of a NWP ensemble, the latter produces an ensemble from a NWP deterministic run (as can be done with LR with its extended formulation; Wilks 2009). Another key difference is that in AnEn the analog search is fully localized in space.

All four forecast systems are tested for 0–48-h probabilistic predictions initialized at 1200 UTC of 10-m wind speed and 2-m temperature at 550 METAR stations over the contiguous United States for the 23 April–31 July 2011 period. The training set for AnEn, EMOS, and LR includes all data from 1 May 2010 up to the day the forecast would have been issued, as if the forecasts were produced in real time.

Analysis using a suite of ensemble and probabilistic forecast verification tools shows that REPS suffers from serious underdispersion and thus poor reliability. The EMOS exhibits far superior performance to REPS as the calibration procedure corrects for systematic error

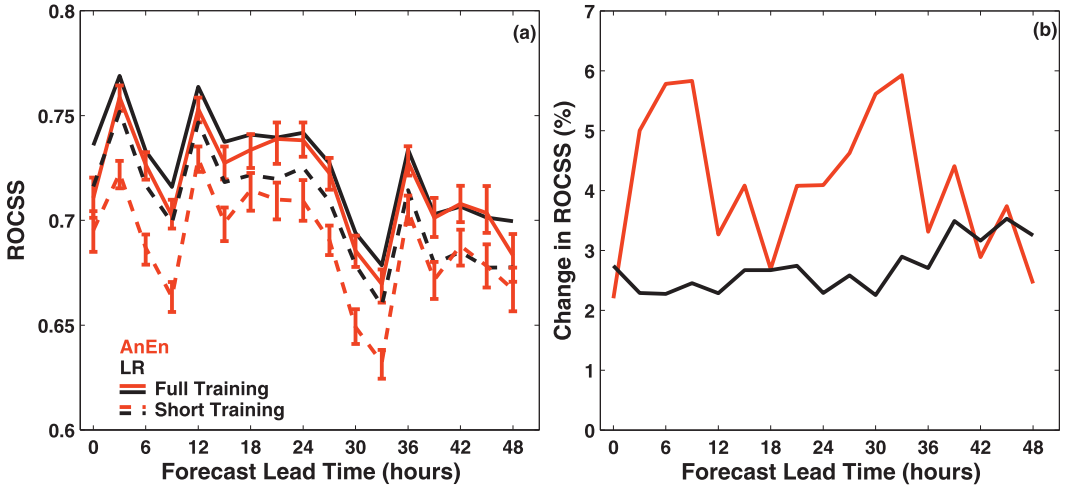

FIG. 14. As in Fig. 13, but with instead the addition of AnEn and LR with a shorter training period of 9 months (dashed lines).

and downscales the forecasts from the 33-km model grid to the station locations. Both EMOS and LR are generally competitive with AnEn, except for rare events (discussed below). An important finding for AnEn is that it is able to capture the situation- and flow-dependent behavior of the errors, as evident by its excellent spread-skill relationship (Fig. 7). This capability is from AnEn properly finding (i.e., filtering) observations from the past that are relevant to the atmospheric flow described in the current NWP forecast and thus represent valid samples of the forecast PDF (see below).

For the event of 10-m wind speed above $10\,\mathrm{m\,s^{-1}}$ (i.e., a rare event), AnEn exhibits superior skill to both LR and EMOS due to several factors. One obvious factor, for only EMOS, is the use of a higher model resolution of AnEn. A second factor may be that the logistic regression approach (used by both LR and EMOS) bases probabilistic forecasts on much more historical data (all available in this study), while AnEn uses only the best matching analogs of a given forecast, thus considering only the most appropriate information. A third factor for the AnEn superior performance of rare events is that logistic regression may be losing some information in the process of creating binary observations, whereas AnEn retains the original observations in constructing a forecast PDF.

The AnEn and LR seem to be more efficient than EMOS, as shown in Fig. 13 where AnEn and LR are generated using a single member of REPS (i.e., at $1/21$ of the computational cost of EMOS) and exhibit only a small decrease in performance with respect to EMOS. The general choice then is to run the members of a real-time NWP ensemble that also requires calibration using historical data, or to generate probabilistic forecasts from a single NWP forecast (allowing for smaller grid increments than any on the NWP ensemble members) using the same historical dataset that would be used to calibrate the NWP ensemble. The latter may be the preferred choice for applications where predictions are necessary at specific locations (e.g., renewable energy), but further studies (see below) are necessary to determine the relative benefit of the two options for applications where two- or three-dimensional fields are needed.

The greater efficiency of AnEn than EMOS can be explained by the following considerations. Grid spacing (i.e., model resolution) is an important factor in the quality of atmospheric prediction. In formulating an estimate of the forecast probability density function (PDF), an ensemble simulates uncertainty information only about atmospheric phenomena on scales resolved by the NWP model. Potential errors of unresolved scales must then be incorporated by widening the forecast PDF via calibration. Increasing model resolution allows for direct simulation of smaller scales and increased resolution (and value) of the probabilistic forecasts. Thus a key advantage of AnEn is the use of a 15-km model grid versus the EMOS use of a 33-km model grid. Comparing these two ensembles may seem unfair at first, but the point to consider is that the resources required to run any n-member NWP ensemble could be put toward producing a single NWP run at a much higher resolution, for which the analog method can then provide reliable forecast uncertainty information, perhaps resulting in more value for decision making by the end user.

A complex NWP model, even run at very fine resolution, is still a coarse approximation of the real atmosphere. An NWP ensemble, no matter how well designed, cannot sample from the true forecast PDF given the challenges of simulating both analysis and model uncertainties. Of the two sources of error (i.e., analysis and model), simulating model uncertainty is particularly daunting. Many techniques have been tried with varying degrees of success (see references in section 1) and typically yield an underdispersive NWP ensemble. This limitation can be compensated for by postprocessing calibration, as shown in this study.

Unlike an NWP ensemble, AnEn attempts to sample directly from the true forecast PDF [Eq. (1)], thus avoiding the challenges of simulating model uncertainty. If an infinite record of observations and predictions were available, it would be possible to find n analogs that are perfect matches to today's forecast, and the verifying observations would sample the true forecast PDF [as defined in Eckel et al. (2012)]. This process maps a point on the model attractor to a portion of the true attractor, which includes many possible true states that exist due to all uncertainties (from both analysis and model) in the forecast.[1] Using only a finite history of observations and model forecasts, AnEn approximates that process and introduces extra uncertainty. The n analogs are only similar (i.e., not perfect) matches so instead of mapping from a single point, AnEn effectively maps from n nearby points on the model attractor, each one associated with a different and likely overlapping portion of the true attractor. The result, relative to employing an infinite training period, is a wider spread by the analog ensemble members and decreased resolution of the forecast PDF.

[1] A similar concept has been explored in data assimilation with shadowing filters (e.g., Judd 2008).

Although resolution of AnEn forecasts is degraded by a finite training period, reliability is not. Similar to the way the spread of AnEn members increases as training length decreases, the root-mean-square error of the ensemble mean increases so that statistical consistency is maintained. Each imperfect analog contributes additional error to the AnEn mean, with worse matches creating larger error as well as creating larger spread among the members as a different portion of true states is sampled from. This effect may not hold true for a very small number of analog ensemble members, due to sampling errors, or for a very large number of members in which analogs become extremely unrepresentative.

This study's finding of greater efficiency by AnEn in producing skillful probabilistic forecasts is encouraging and motivates further investigations. Testing should be expanded from prediction of 10-m wind speed and 2-m temperature to other forecast variables (e.g., relative humidity, pressure, precipitation), from prediction at observation locations at the surface to upper-air forecasts over a three-dimensional grid, and also to include different and longer verification regions and periods. Research is also needed on the sensitivity of AnEn performance to key aspects of its formulation such as the number of members to use, aspects of the analog search (e.g., the set of predictors included, their weights, and the formulation of the analog-quality metric), and the length of the training dataset. As shown in Fig. 14, AnEn and LR greatly benefit from increased training, with AnEn perhaps benefiting the most from such extension, which may be due to the distinct differences in AnEn design discussed above. Testing could also be performed on a hybrid ensemble approach that combines both the analog and NWP ensemble by finding multiple analogs for each member of the NWP ensemble, which may calibrate the NWP ensemble members while generating a more thoroughly sampled forecast PDF.

Acknowledgments. This work was made possible by support from the U.S. National Weather Service, the U.S. Defense Threat Reduction Agency, and the U.S. Army Test and Evaluation Command through an interagency agreement with the National Science Foundation. We are grateful to Martin Charron and Ronald Frenette (Environment Canada) for providing the REPS and GEM data, without which the presented analysis could not have been possible. This paper has been improved by thorough and insightful revisions provided by three anonymous reviewers, and the valuable comments and suggestions of Eric Grimit (3TIER); Sue Ellen Haupt, Tara Jensen, and Daniel Steinhoff (NCAR); Cliff Mass (University of Washington); and Thomas Nipen (University of British Columbia).

REFERENCES

AMS, 2008: Enhancing weather information with probability forecasts: An information statement of the American Meteorological Society. *Bull. Amer. Meteor. Soc.,* **89,** 1049–1053.

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate,* **9,** 1518–1530.

Berner, J., G. J. Shutts, M. Leutbecher, and T. N. Palmer, 2009: A spectral stochastic kinetic energy backscatter scheme and its impact on flow-dependent predictability in the ECMWF ensemble prediction system. *J. Atmos. Sci.,* **66,** 603–626.

Bishop, C. H., and Z. Toth, 1999: Ensemble transformation and adaptive observations. *J. Atmos. Sci.,* **56,** 1748–1765.

Bowler, N. E., A. Arribas, K. R. Mylne, K. B. Robertson, and S. E. Beare, 2008: The MOGREPS short-range ensemble prediction system. *Quart. J. Roy. Meteor. Soc.,* **134,** 703–722.

Buizza, R., M. J. Miller, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF Ensemble Prediction System. *Quart. J. Roy. Meteor. Soc.,* **125,** 2887–2908.

Delle Monache, L., T. Nipen, Y. Liu, G. Roux, and R. Stull, 2011: Kalman filter and analog schemes to postprocess numerical weather predictions. *Mon. Wea. Rev.,* **139,** 3554–3570.

Eckel, F. A., and C. F. Mass, 2005: Aspects of effective mesoscale, short-range ensemble forecasting. *Wea. Forecasting,* **20,** 328–350.

——, M. S. Allen, and M. C. Sittel, 2012: Estimation of ambiguity in ensemble forecasts. *Wea. Forecasting,* **27,** 50–69.

Epstein, E. S., 1969: Stochastic dynamic prediction. *Tellus,* **21,** 739–759.

Fukunaga, K., 1990: *Introduction to Statistical Pattern Recognition.* 2nd ed. Elsevier, 521 pp.

Gangopadhyay, S., B. L. Harding, B. Rajagopalan, J. J. Lukas, and T. J. Fulp, 2009: A nonparametric approach for paleohydrologic reconstruction of annual streamflow ensembles. *Water Resour. Res.,* **45,** W06417, doi:10.1029/2008WR007201.

Gill, J., and Coauthors, 2008: Guidelines on communicating forecast uncertainty. WMO TD 4122, 25 pp. [Available online at http://library.wmo.int/pmb_ged/wmo-td_1422_en.pdf.]

Gneiting, T., A. E. Raftery, F. Balabdaoui, and A. Westveld, 2004: Verifying probabilistic forecasts: Calibration and sharpness. Preprints, *17th Conf. on Probability and Statistics in the Atmospheric Sciences,* Seattle, WA, Amer. Meteor. Soc., 2.1. [Available online at https://ams.confex.com/ams/84Annual/techprogram/paper_68303.htm.]

——, ——, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.,* **133,** 1098–1118.

Hacker, J., and Coauthors, 2011: The U.S. Air Force Weather Agency's mesoscale ensemble: Scientific description and performance results. *Tellus,* **63A,** 625–641.

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.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.,* **129,** 550–560.

——, and S. J. Colucci, 1997: Verification of Eta-RSM short-range ensemble forecasts. *Mon. Wea. Rev.,* **125,** 1312–1327.

——, and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Wea. Rev.,* **134,** 3209–3229.

——, and ——, 2007: Ensemble calibration of 500-hPa geopotential height and 850-hPa and 2-m temperatures using reforecasts. *Mon. Wea. Rev.,* **135,** 3273–3280.

——, ——, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.,* **132,** 1434–1447.

——, ——, and S. L. Mullen, 2006: Reforecasts: An important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc.,* **87,** 33–46.

——, 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.

——, G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau Jr., Y. Zhu, and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble reforecast data set. *Bull. Amer. Meteor. Soc.,* in press.

Hirschberg, P. A., and Coauthors, 2011: A weather and climate enterprise strategic implementation plan for generating and communicating forecast uncertainty information. *Bull. Amer. Meteor. Soc.,* **92,** 1651–1666.

Holt, T., J. Pullen, and C. H. Bishop, 2009: Urban and ocean ensembles for improved meteorological and dispersion modelling of the coastal zone. *Tellus,* **61A,** 232–249.

Hopson, T. M., and P. J. Webster, 2010: A 1–10-day ensemble forecasting scheme for the major river basins of Bangladesh: Forecasting severe floods of 2003–07. *J. Hydrol.,* **11,** 618–641.

Hou, D., E. Kalnay, and K. K. Droegemeier, 2001: Objective verification of the SAMEX'98 ensemble forecasts. *Mon. Wea. Rev.,* **129,** 73–91.

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.,* **124,** 1225–1242.

——, H. L. Mitchell, G. Pellerin, M. Buehner, M. Charron, L. Spacek, and B. Hansen, 2005: Atmospheric data assimilation with an ensemble Kalman filter: Results with real observations. *Mon. Wea. Rev.,* **133,** 604–620.

——, H. L. Mitchel, and X. Deng, 2009: Model error representation in an operational ensemble Kalman filter. *Mon. Wea. Rev.,* **137,** 2126–2143.

Jolliffe, I., and D. Stephenson, 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science.* Wiley and Sons, 240 pp.

Judd, K., 2008: Forecasting with imperfect models, dynamically constrained inverse problems, and geometrically modified gradient descent. *Physica D,* **237,** 216–232.

Klausner, Z., H. Kaplan, and E. Fattal, 2009: The similar days method for predicting near surface wind vectors. *Meteor. Appl.,* **16,** 569–579.

Krishnamurti, T. N., C. M. Kishtawal, Z. Zhang, T. Larow, D. Bachiochi, E. Williford, S. Gadgil, and S. Surendran, 2000: Multimodel ensemble forecasts for weather and seasonal climate. *J. Climate,* **13,** 4196–4216.

Kuhl, D., and Coauthors, 2007: Assessing predictability with a local ensemble Kalman filter. *J. Atmos. Sci.,* **64,** 1116–1140.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.,* **102,** 409–418.

Li, X., M. Charron, L. Spacek, and G. Candille, 2008: A regional ensemble prediction system based on moist targeted singular vectors and stochastic parameter perturbations. *Mon. Wea. Rev.,* **136,** 443–462.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.,* **20,** 130–141.

——, 1993: *The Essence of Chaos*. University of Washington Press, 227 pp.

Mason, I. B., 1980: Decision-theoretic evaluation of probabilistic predictions. *Proc. WMO Symp. on Probabilistic and Statistical Methods in Weather Forecasting,* Nice, France, WMO, 219–228.

——, 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.,* **30,** 291–303.

Mason, S. J., and N. E. Graham, 1999: Conditional probabilities, relative operating characteristics, and relative operating levels. *Wea. Forecasting,* **14,** 713–725.

Messner, J. W., and G. J. Mayr, 2011: Probabilistic forecasts using analogs in the idealized Lorenz96 setting. *Mon. Wea. Rev.,* **139,** 1960–1971.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The new ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.,* **122,** 73–119.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.,* **12,** 595–600.

NOAA, 1998: Automated Surface Observing System (ASOS) user's guide. 74 pp. [Available online at http://www.nws.noaa.gov/asos/pdfs/aum-toc.pdf.]

NRC, 2006: *Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts*. National Research Council, National Academies Press, 124 pp.

Panziera, L., U. Germann, M. Gabella, and P. V. Mandapaka, 2011: NORA—Nowcasting of orographic rainfall by means of analogues. *Quart. J. Roy. Meteor. Soc.,* **137,** 2106–2123, doi:10.1002/qj.878.

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF Ensemble Prediction System. *Quart. J. Roy. Meteor. Soc.,* **126,** 649–667.

Shutts, G. J., 2005: A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Quart. J. Roy. Meteor. Soc.,* **131,** 3079–3102.

Siegert, S., J. Bröcker, and H. Kantz, 2012: Rank histograms of stratified Monte Carlo ensembles. *Mon. Wea. Rev.,* **140,** 1558–1571.

Stensrud, D. J., J. W. Bao, and T. T. Warner, 2000: Using initial condition and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. *Mon. Wea. Rev.,* **128,** 2077–2107.

Talagrand, O., R. Vautard, and B. Strauss, 1997: Evaluation of probabilistic prediction systems. *Proc. Workshop on Predictability,* Reading, United Kingdom, ECMWF, 1–25.

Teixeira, J., and C. A. Reynolds, 2008: Stochastic nature of physical parameterizations in ensemble prediction: A stochastic convection approach. *Mon. Wea. Rev.,* **136,** 483–496.

Thomas, S. J., J. P. Hacker, M. Desgagné, and R. Stull, 2002: An ensemble analysis of forecast errors related to floating point performance. *Wea. Forecasting,* **17,** 898–906.

Thompson, J. C., 1950: A numerical method for forecasting rainfall in the Los Angeles area. *Mon. Wea. Rev.,* **78,** 113–124.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.,* **74,** 2317–2330.

——, and ——, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.,* **125,** 3297–3319.

van den Dool, H. M., 1989: A new look at weather forecast through analogs. *Mon. Wea. Rev.,* **117,** 2230–2247.

Wandishin, M. S., S. L. Mullen, D. J. Stensrud, and H. E. Brooks, 2001: Evaluation of a short-range multimodel ensemble system. *Mon. Wea. Rev.,* **129,** 729–747.

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.

Wilks, D. S., 2001: A skill score based on economic value for probability forecasts. *Meteor. Appl.,* **8,** 209–219.

——, 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. Academic Press, 627 pp.

——, 2009: Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteor. Appl.,* **16,** 361–368.

——, and T. M. Hamill, 2007: Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Wea. Rev.,* **135,** 2379–2390.

Wu, W., and Coauthors, 2012: Statistical downscaling of climate forecast system seasonal predictions for the southeastern Mediterranean. *Atmos. Res.,* **118,** 346–356.

Zorita, E., and H. von Storch, 1999: The analog method as a simple statistical downscaling technique: Comparison with more complicated methods. *J. Climate,* **12,** 2474–2489.