Using Evolutionary Programming to Add Deterministic and Probabilistic Skill to Spatial Model Forecasts

PAUL J. ROEBBER

University of Wisconsin–Milwaukee, Milwaukee, Wisconsin

(Manuscript received 26 September 2017, in final form 21 May 2018)

ABSTRACT

Evolutionary programming is applied to the postpocessing of ensemble forecasts of temperature on a spatial domain. These forecasts are obtained from the 11-member Reforecast V2 ensemble over the region from 24°–53°N to 125°–66°W for the period 1 January 1985–14 May 2011. The evolution is based upon a static ecosystem model that holds constant the number of individuals (algorithms), using a fixed rate of introduction of new algorithms and removal of existing algorithms. Each algorithm adheres to a specific underlying genetic architecture, and the selection pressure on the ''species'' is according to deterministic performance (root-mean-square error) on a training dataset. On a 2325-case, independent test dataset, the method improved root-mean-square error and ranked probability score relative to the Reforecast ensemble by 0.31°F (8.7%) and 3.3%, respectively, across the domain, with 96% of the grid points showing simultaneous improvements in both measures. The use of input information by the evolutionary programming algorithms varied by region; while the algorithm forecasts at all locations are fundamentally tied to the Reforecast ensemble forecast, northeastern locations found snow cover to be the next most useful input, whereas southwestern locations preferentially employed precipitable water. An adaptive form of the approach, developed to be readily implemented into operations, is tested in the absence of improving inputs but is found to only slightly degrade performance (1.2% in root-mean-square error and 0.6% in ranked probability skill score). A number of future extensions are discussed.

1. Introduction

Postprocessing of numerical weather prediction (NWP) model forecasts provides an opportunity to extract more forecast skill from these data, and, relative to the computational requirements of producing the model data itself, it has a relatively low cost. A number of postprocessing approaches have been examined in the meteorological literature, including multiple linear regression (i.e., MOS; Glahn and Lowry 1972), artificial neural networks (e.g., Koizumi 1999; Kuligowski and Barros 2001; Hennon et al. 2005; Roebber et al. 2007), evolutionary programming (EP; Yang et al. 1996; Bakhshaii and Stull 2009; Roebber 2010, 2015a), quantile mapping (e.g., Scheuerer and Hamill 2015), ensemble Kalman filtering (e.g., Houtekamer and Mitchell 1998), Bayesian model averaging (Raftery et al. 2005), and Bayesian model combination (BMC; Roebber 2015a), either alone or along with bias correction (e.g., Cui et al. 2012).

Roebber (2015a) used bias correction in combination with BMC in the context of single-site EP ensembles, where the inputs were a combination of observed and NWP data, and showed that approach to be effective. Here, we will modify this approach, extending it to spatial data involving only NWP ensemble model inputs, and document the relative gains in both deterministic and probabilistic skill that are obtainable over the raw NWP model forecasts, multiple regression models using the same inputs, and the bias-corrected NWP ensemble. Finally, we will apply the spatial EP methodology using an adaptive form.

In section 2, we describe the dataset used in the analysis and detail the postprocessing methodology. In section 3, we apply the method to the problem of 72-h temperature forecasts on a North American domain and discuss future extensions of the method. In section 4, we present a summary of the paper.

2. Methodology

a. Meteorological data

The data used in this study are analysis and 72-h, 2-m temperature forecasts valid at 0000 UTC, obtained

Corresponding author: Paul J. Roebber, roebber@uwm.edu

DOI: 10.1175/MWR-D-17-0272.1

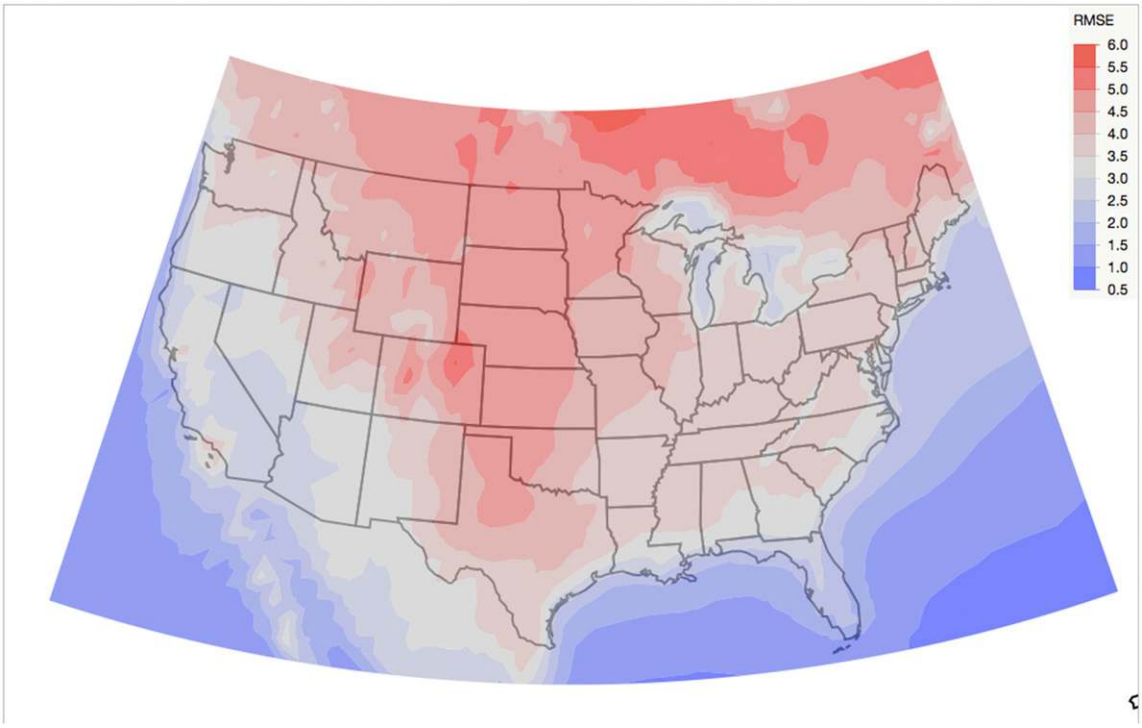

FIG. 1. Root-mean-square error (°F) of the bias-corrected, control 2-m temperature forecast from the 11-member 72-h Reforecast version 2 ensemble, valid at 0000 UTC, for the period 1 Jan 2005–14 May 2011. Verification is based upon the 0-h Reforecast V2 analysis.

from the Reforecast V2 ensemble (RFv2; 11 members; Hamill et al. 2013) 1° latitude–longitude data for the region from 24°–53°N and 125°–66°W (Fig. 1). Forecasts are verified using the corresponding 0-h analysis for the period 1 January 1985–14 May 2011.

Additionally, we obtain 72-h RFv2 forecasts for 850-hPa temperature, cloud cover, precipitable water, 10-m wind speed, and snow cover in excess of 1 in., all from the control member of the ensemble only. We compute the cosine of the solar zenith angle at local noon for each date at each grid point following:

$$\cos Z = \sin\left(\frac{\pi}{180}\varphi\right) \times \sin\left(\frac{\pi}{180}D\right) + \cos\left(\frac{\pi}{180}\varphi\right) \times \cos\left(\frac{\pi}{180}D\right), \tag{1}$$

where Z is the solar zenith angle, φ is the latitude (°), and D is the declination angle (°), given by

$$D = 0.396\,372 - 22.913\,27 \times \cos\left(\frac{\pi}{180}g\right) + 4.025\,43 \times \sin\left(\frac{\pi}{180}g\right) - 0.387\,205 \times \cos\left(2\frac{\pi}{180}g\right) + 0.051\,967 \times \sin\left(2\frac{\pi}{180}g\right) - 0.154\,527 \times \cos\left(3\frac{\pi}{180}g\right) + 0.084\,798 \times \sin\left(3\frac{\pi}{180}g\right) \tag{2}$$

and g (degrees) is

$$g = \frac{360}{365.25} \times (\text{Day} + 0.5), \tag{3}$$

where Day is the day of the year. We compute the climatological 2-m temperature (based on the analysis data), and extract the minimum, 20th percentile, median, 80th percentile, and maximum members of the 2-m temperature forecasts from the 11-member RFv2 ensemble. Finally, we compute an analog forecast for 2-m temperature, using as the analog search database all cases from the beginning of the available data (1 January 1985) up to 10 days prior to the forecast in question. The analog is defined based upon the mean absolute difference in the forecast minimum, 20th percentile, median, 80th percentile, and maximum value obtained from the 11-member RFv2 ensemble. The forecast is then formed

using the observed 2-m temperature of the analog case as defined in Delle Monache et al. (2011).

The sample is split with 4383 days (1 January 1985–31 December 1996) for training, 2922 days (1 January 1997–31 December 2004) for cross validation, and 2325 days (1 January 2005–14 May 2011) for testing. In the EP procedure (see sections 2b–d below), the training data are used to develop the solutions, while the cross-validation data are used to determine which of these solutions to retain. Both training and cross-validation data are then used again to assign weights for BMC. Finally, the test data are used to evaluate performance—all results reported in this paper are based on this independent test dataset of 2325 days.

One might wonder whether a limiting factor in methods such as this, which map inputs to outputs (EP, multiple linear regression, artificial neural networks, etc.), and which rely on older training data, is the relatively rapid pace of climate change? First, if the training data contain sufficient exemplars of what will be represented in the future climate, this should not restrict in any way its successful future application. Second, where this is not the case, if a proportional response to a given predictor continues to hold in future climate states, then the method will still hold. On the other hand, if these response functions do not hold and there are insufficient exemplars of such future conditions in the training data, then one should not necessarily expect the method to perform well in the future. Ultimately, this argues for adaptive forms of such systems, a concept that has been explored by Roebber (2015c) and that is further discussed in section 3b below.

Although the 1° grid spacing of the RFv2 is somewhat coarse, the long forecast history of these data, which provide for a wide variety of conditions in all seasons, makes it an excellent resource for the training and testing of postprocessors. In section 3c, we provide a discussion of how these ideas might benefit from application to a higher-resolution dataset, such as the High Resolution Rapid Refresh (HRRR; Benjamin et al. 2016).

b. Bias correction

Each forecast input F_t at time t is bias corrected following Cui et al. (2012), such that

$$B_t = 0.95 \times B_{t-1} + 0.05 \times (F_t - O_t), \tag{4}$$

where B_t is the accumulated bias, F_t is the uncorrected forecast, and O_t is the verification. The corrected forecast at time t (Fc_t) is then formed by

$$\mathrm{Fc}_t = F_t - B_t. \tag{5}$$

c. Bayesian model combination postprocessing for spatial evolutionary programs

For the spatial training, we take some advantage of the autocorrelation of meteorological fields. We accomplish this by applying the concept of von Neumann neighborhoods, defined by the grid point in question and the four immediately adjacent points, using BMC (Monteith et al. 2011) to weight the resultant set of five forecast inputs. Note that the number of weight comparisons required by BMC scales as W^N, where W is the number of raw weight levels and N is the number of forecasts for a given time step. Larger neighborhoods than the five-point one used here could be envisioned, but come at computational expense. Here, we use only 6 levels of raw weights (0–5), which for all possible weights then require 6^5 comparisons, a number that is quite tractable with modest computing power even over a grid of 1800 points. For obvious reasons, we require that at least 1 of the 5 inputs have nonzero weight and thus, the possible normalized weights (where the set of 5 weights sum to 1) range from 0 to 1, with the smallest possible nonzero weight being 0.048 and the largest possible less than unity being 0.833.

For each set of 5 forecasts, we cycle through the 6^5 possible combinations, producing a weighted forecast. At each grid point, for each set of 5 forecasts, we compute the posterior probability for the particular model weight combination (e) given the training data (D) according to the following formula:

$$p(e|D) \propto \frac{1}{4^{10}}(1-\varepsilon)^r \varepsilon^{n-r}, \tag{6}$$

where ε is estimated using the average error rate of the model combination on the training data, r is the number of correct predictions, and n is the total number of training cases. Readers interested in the rationale for this formulation should consult Monteith et al. (2011).

Here, we define a model combination as correct for a given date if the weighted forecast has lower squared error than that of the median forecast obtained from the set of 11 RFv2 forecasts at the point and time in question. The final weight combination that is selected is the one that maximizes the logarithm of Eq. (6), taking into account that ε must be less than 0.5.

Under the assumption that each individual member k of the weighted forecast is normally distributed, one can produce a wide variety of probability distribution shapes, where for each individual member, the mean is the forecast value, and the variance o^2 is estimated as follows:

$$o^2 = \frac{\sum_n \sum_k w_k (F_k - O)^2}{n}, \tag{7}$$

where w_k is the normalized weight of the kth forecast, F_k is the kth forecast, O is the observed value, and the summation is over the five forecasts and the n training cases. The overall probability distribution function is then simply the sum of these five individual distributions, weighted by w_k (i.e., a normal mixtures distribution). Thus, this system can be used for both deterministic [e.g., root-mean-square error (RMSE)] and probabilistic [e.g., ranked probability skill score (RPSS)] verification and will be assessed accordingly in section 3.

Given the relatively coarse 1° latitude–longitude grid, one might expect some difficulties to emerge along orographic boundaries (complex terrain, coastlines), since the neighborhood sampling in those areas may include substantially unalike points. One could envision a procedure where such points are excluded from the neighborhood, but for simplicity here, we will simply allow the BMC weighting process to empirically discount these locations through the above process. The general issue will be addressed in the discussion.

d. Evolutionary programming

The basic architecture of EP algorithms as developed by Roebber (2015a) consists of a sum of 10 IF-THEN equations involving linear and nonlinear combinations of predictors. Such architecture is quite flexible, yet because of the structural logic, allows for interpretation of the forecasts so-produced. The genome is composed of the variables, mathematical operators, and coefficients contained within each of the 10 IF-THEN lines that make up the sum. The number of algorithm lines is subjective and is chosen to insure sufficient algorithm complexity to handle the variety of situations under consideration.

Initially, a population of algorithms is created with this underlying structure, but in which all the variables, operators, and coefficients are chosen at random. In each iterative step (hereafter, generation), the fitness of individual algorithms are evaluated and better performing members are preferentially allowed to pass their genetic material to the next generation (in some cases, with mutations that provide an important source of innovation). This process allows improvements in performance from one generation to the next until some measure of convergence is achieved. Here, we follow these basic principles but with some differences in detail with respect to Roebber (2015a).

Let F be the EP algorithm, such that

$$F = \langle \mathrm{RFv2}' \rangle + \sum_{j=1}^{5} \partial_j, \tag{8}$$

TABLE 1. Variables used as inputs to the evolutionary programs. Bias-corrected inputs are bold. Note that this list is not inclusive of all variables that could be included, but represent a selection of variables that can be reasonably defended on the basis of both meteorological and postprocessing considerations.

72-h forecast 850-hPa temperature
2-m climatological temperature (monthly basis)
72-h forecast cloud cover
72-h forecast precipitable water
72-h forecast 10-m wind speed
72-h forecast snow cover in excess of 1 in.
Cosine of the solar zenith angle
72-h analog forecast 2-m temperature
Min of the 72-h forecast 2-m temperatures from the 11 RFv2 members
20th percentile of the 72-h forecast 2-m temperatures from the 11 RFv2 members
Median of the 72-h forecast 2-m temperatures from the 11 RFv2 members
80th percentile of the 72-h forecast 2-m temperatures from the 11 RFv2 members
Max of the 72-h forecast 2-m temperatures from the 11 RFv2 members
Unity

where j refers to the algorithm line, $\langle \mathrm{RFv2}' \rangle$ is the normalized RFv2 ensemble mean forecast (where the sum of the five lines represents the adjustment to the normalized RFv2 ensemble mean temperature), and each line j is expressed as follows:

$$\mathrm{IF}\left(V_{1,j} O_{R,j} V_{2,j}\right) \mathrm{THEN}\, \delta_j = \left(C_{1,j} V_{3,j}\right) O_{1,j} \left(C_{2,j} V_{4,j}\right) \times O_{2,j} \left(C_{3,j} V_{5,j}\right), \tag{9}$$

where $V_{1,j}$, $V_{2,j}$, $V_{3,j}$, $V_{4,j}$, and $V_{5,j}$ can be any of the input variables from Table 1; $C_{1,j}$, $C_{2,j}$, and $C_{3,j}$ are real-valued multiplicative constants in the interval $[-1, 1]$; $O_{R,j}$ is one of two relational operators ($\leq$, $>$); and $O_{1,j}$, $O_{2,j}$ can be either the addition or multiplication operators. This notation is such that $V_{1,j}$, $V_{2,j}$, etc. indicate the first variable in line j, the second variable in line j, and so on. The input variables are normalized to [0, 1] based on the minimum and maximum of the training data, and the final forecast is then reconstituted to the proper non-normalized value based upon the minimum and maximum of the training data verification.

This structure, unlike that of Roebber (2015a), specifies a baseline forecast formed by the RFv2 ensemble mean and the succeeding five lines then each represent an adjustment to that baseline forecast. For example, weather forecasters know that under strongly radiative conditions (clear skies, light winds), the minimum temperature will be lower if there is snow cover, so here one of the five IF-THEN conditional lines might test for the presence of

snow and calculate a downward adjustment to the ensemble mean forecast in that case. Additionally, the astute reader will note that the number of IF-THEN conditional lines has been reduced from the 10 lines of the prior studies (e.g., Roebber 2015a,b,c) to the 5 in this study. This has been done to reduce computational complexity—examination of algorithms trained on other datasets using the 10 line form revealed that most of the lines were not used in the final, trained forms.

The EP algorithms are trained using the following steps:

1) Bias correct all input data that are 2-m temperature forecasts (e.g., the bold items of Table 1).
2) Normalize all the input data in the range 0–1.
3) Randomly initialize a population of 50 algorithms at each of the 60 × 30 grid points.
4) Evaluate the RMSE of each algorithm on the training and cross-validation datasets.
5) At each grid point, sort the 50 algorithms in order from lowest to highest RMSE on the training data.
6) Eliminate the 10 worst performers based on the training RMSE.
7) Select the top-ranked performer based on the training RMSE that is within some distance (see below) of the current grid point and fill 1 of the available 10 slots with its clone.
8) Clone the 9 best performers based on the training RMSE at the current grid location to fill the remaining 9 slots.
9) For each of the 10 clones, randomly select one of the 5 lines and one of the 11 genetic components of the line (5 variables, 3 operators, and 3 coefficients) and randomly replace it (sampling with replacement, so the clone may or may not mutate).

Steps 4–9 are repeated through 50 generations, saving the best individual algorithm at each grid point, based on the cross-validation RMSE from any time during the training. Note that while all the algorithms share the same underlying structure, the variables, operators, and coefficients are populated randomly and then allowed to evolve—thus, in general, the algorithms will be structurally distinct both at and across grid points.

Eliminating the 10 worst performers (step 6) and not allowing them to mutate to the next generation can result in the loss of genetic diversity. This is a trade-off that is made in order to create ecosystem ''space'' for new forms to develop (the 10 new algorithms replace the 10 that have been eliminated). The assumption is made that 50 algorithms at a grid point are able to provide sufficient diversity for innovative solutions to be discovered. One could proceed without step 6, but this would require continuously increasing the carrying capacity and the overhead associated with carrying large numbers of poorly performing algorithms is undesirable. This is particularly the case since, as is shown in section 3a, large increases in carrying capacity do not translate to large improvements in algorithm performance. More sophisticated ecosystem dynamics can be employed that would provide a natural counter to unrestrained population growth while still promoting both innovation and diversity and is currently being explored (see section 3c for further discussion).

The first clone is selected from some normally distributed location away from the grid point (nx, ny), following Basak et al. (2010), where

$$\sigma_M = 1 + \left(1 - \frac{i}{100}\right)^3 \left|\cos\frac{\pi}{10} i\right| \left(\frac{M}{10} - 1\right) \qquad (10)$$

and i is the generation number, and M is the number of grid points in the x direction (60) or the y direction (30). Clones 2–10, however, are always produced locally (i.e., at nx, ny). It should be noted that the ''next generation'' is not introduced until the current generation has completed the procedure at all grid points, so the cloning process is independent of the ordering of those grid points (i.e., if a point at X, Y produces a clone at $X + 1$, Y and X, Y is processed before $X + 1$, Y, when the procedure reaches point $X + 1$, Y, it does not yet know about the clone).

This distant clone selection idea borrows from the technique of invasive weed optimization (IWO; e.g., Mehrabian and Lucas 2006) such that as generations increase, the probability of selecting a clone from farther away decreases (i.e., falling from 98% at the first generation to 45% at the 50th generation). The rationale for this scheme is that as generations increase, algorithms are expected to have already achieved substantial local optimization, yet we do not wish to entirely eliminate the possibility that a useful innovation could be introduced from more distant locations. Basak et al. (2010) showed that the addition of the cosine function allows for more rapid detection of optimal solutions, since the probability of selecting a clone locally increases to 100% every 10th iteration (from 5, 15, 25 etc.).

In the above procedures, we do not use the following features that Roebber (2015a) showed to be beneficial: genetic cross-over (simulated sexual selection), training niches (here, we rely on the neighborhoods to perform this function), the transposition form of mutation, fatal disease, and an external performance criterion (instead, we use the ranking process to provide the necessary selective pressure). As will be shown below, application of this somewhat simplified training process on the spatial domain produces good performance both for the individual algorithm at a grid and when these are combined using BMC (section 3).

TABLE 2. The 72-h, 2-m temperature forecasts across the latitude–longitude domain extending from 24°–53°N to 125°–66°W. Shown are RMSE (°F) for the RFv2 ensemble mean without and with bias correction (BC), a multiple linear regression (MLR) trained at each grid point using the 13 predictors (Table 1) and the RFv2 ensemble mean without and with bias correction (BC), the analog forecast, the EP without and with BC and with the combination of bias correction and BMC (BC, BMC), and the adaptive EP without bias correction and with the combination of bias correction and BMC (BC, BMC). Note that all computations are based on the test data (2325 days from 1 Jan 2005 to 14 May 2011). Also shown is the RPSS, using the ensemble-size correction methodology of Weigel et al. (2007).

	RFv2 (BC)	MLR (BC)	Analog	EP (BC)	EP (BC, BMC)	Adaptive (BC, BMC)
RMSE	3.86 (3.55)	3.97[a] (3.59)	5.00	3.45[a] (3.27)	3.45[a] (3.24)	3.54[a] (3.28)
RPSS	0.631	—	—	—	0.664	0.658

[a] The forecasts are not bias corrected but the inputs have been, as per Table 1.

3. Results

There are substantial spatial forecast patterns in RMSE, dictated in large part by latitude and physiography and likely also by the density of the verifying observations. In the North American domain, the continental climate of the interior leads to larger temperature variability and larger forecast RMSE [Fig. 1; note that the correlation between the standard deviation in the analysis temperature at each grid point (based on monthly data) and the RFv2 72h forecast RMSE across this domain is 0.93]. Thus, interpretation of performance should be accounted for in relation to this ''background'' level of error.

The procedures described in section 2 were applied to the RFv2 dataset, and the results averaged over the domain excluding the 4 boundaries [i.e. we verify on a 58×28 grid rather than 60×30 (Table 2), and also at individual grid points]. By way of comparison, we also include verification for forecasts produced using an ordinary least squares multiple linear regression (MLR) with the 13 predictors (Table 1) along with the RFv2 ensemble mean, where the MLR equations are developed independently at each grid point using the training and cross-validation data from before.

No attempt was made to control for overfitting of the MLR; however, the number of predictors is of the same order as operational versions of the National Oceanic and Atmospheric Administration (NOAA) model output statistics (MOS) and the RMSE of the MLR is within 3% between the training and independent test data, suggesting little generalization error. Other potentially useful techniques for this application, which have not been evaluated here, include nonhomogeneous Gaussian regression (Gneiting et al. 2005; Hemri et al. 2014), quantile regression forests (Taillardat et al. 2016) and nonhomogeneous boosting (Messner et al. 2017). No claim is made here of the relative superiority with respect to any of these techniques; such evaluations will need to be conducted in the context of the specific forecast context of concern. Here, we provide the MLR and the RFv2 as standard baselines, along with bias corrections and BMC as further improvements.

For the test data, the RMSE, relative to the 11-member RFv2 ensemble mean, is 0.31°, 0.27°, 0.41°, and 0.62°F lower for the bias-corrected RFv2 ensemble, the bias-corrected multiple linear regression, the EP, and the EP using bias correction and BMC, respectively (Table 2; note that in this paper, we use the Fahrenheit temperature scale to be consistent with the available input and verification temperature data). Thus, while substantial improvement in the RFv2 ensemble mean forecast is obtained simply through the Cui et al. (2012) bias correction, these gains are doubled when using this correction on EP along with BMC. Although ordinary least squares regression is a form of bias correction, the Cui et al. (2012) correction to the MLR reduces RMSE across the domain considerably but notably the bias-corrected MLR does not improve upon the bias-corrected RFv2 ensemble mean forecast. While the analog forecast is not competitive as a stand-alone, as will be shown in the relative weights analysis below, it does provide some useful information in the form of another input to the EP algorithms.

While national forecast services work toward continuous skill improvements, an incremental advance of this scale (0.31°F) might not appear to have much value in the public forecast context. However, as shown in Roebber (2010), using electricity demand information provided by Teisberg et al. (2005), an improvement of this order can result in annual cost savings in this sector for an Ohio-sized utility of $1.5 million (U.S. dollars) or more. In recent years, natural gas has become an increasingly important component of energy production in the United States, and spot market prices are highly sensitive to temperature. Although a comprehensive analysis of the relationship between forecast error and cost is highly complex, using data provided by the U.S. Energy Information Administration (2014) for temperature sensitivity and natural gas consumption, an approximate estimate is that a forecast improvement of this scale

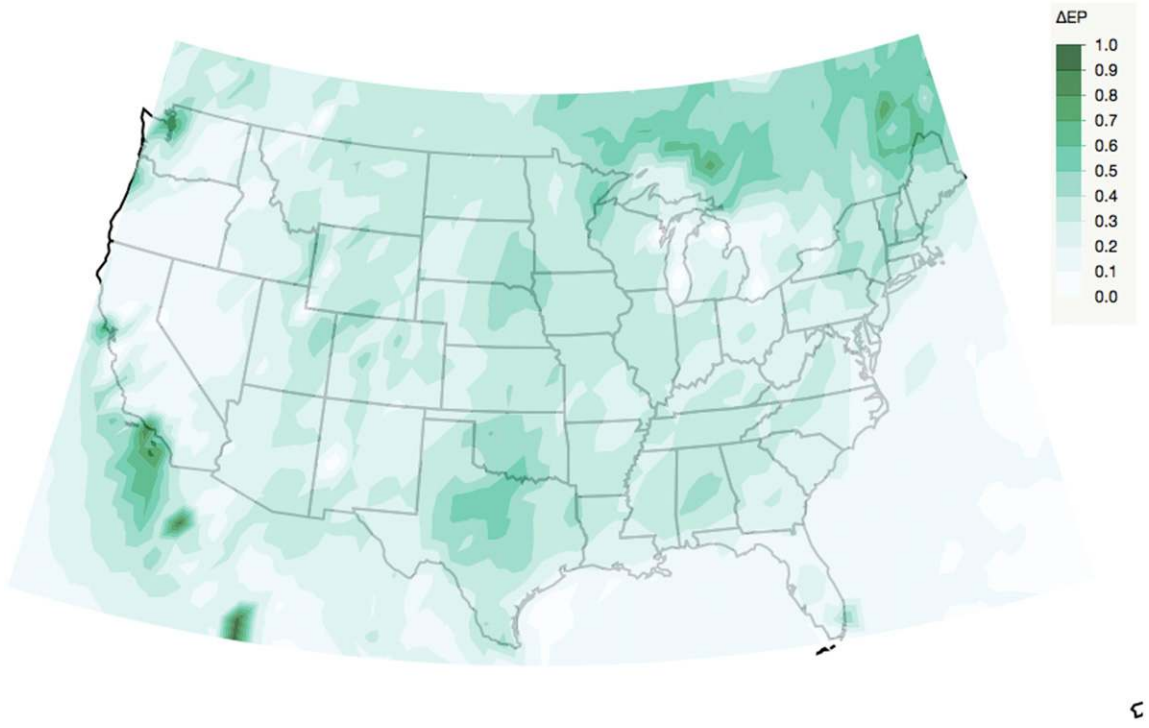

FIG. 2. Root-mean-square error (°F) improvement of the evolutionary program using Bayesian model combination relative to the bias-corrected RFv2 ensemble mean.

during cold outbreaks in the New York energy market might lead to savings on the order of \$250,000 day^{-1} (U.S. Energy Information Administration 2014).

Given the considerable spatial variability in error, it is of interest to consider the spatial pattern of the relative improvement of the EP using BMC compared to the bias-corrected RFv2 ensemble mean (Fig. 2). The largest improvements tend to be where the largest errors in the RFv2 occur (the correlation between the reduction in RMSE and the RMSE of the bias-corrected RFv2 ensemble mean is 0.54), the exception to this being in the extreme southwest of the domain (offshore of California and Mexico) where the RMSE is low.

An event that features both large temperature variability and large forecast errors is the Great Plains cold surge. An example of such an event, and the improvement shown by the EP, occurred during the period 27 November–1 December 2006 (Fig. 3). We will focus on the grid point nx = 27, ny = 10 (33°N, 99°W), which is located about 250 km west of Dallas, Texas. The analyzed 0000 UTC temperature at this grid point fell from 71° to 28°F from 30 November to 2 December, resulting in RMSE for the bias-corrected RFv2 ensemble and the EP 72-h forecasts of 5.58° and 2.88°F, respectively. The EP forecast improvement was primarily accomplished through its relative capture of the temperature drop between 30 November and 1 December, a deterministic advantage that is also reflected in the 72-h probabilistic forecast valid for 0000 UTC 1 December (Fig. 4).

It is of some interest to understand how this probabilistic shift depicted in Fig. 4 was accomplished. In this instance, perhaps not surprisingly, it appears to have been primarily the result of pulling in useful upstream information. Consider the five grid points making up the BMC calculation numbered as follows: to the west (grid point 1), to the south (grid point 2), at the point in question (grid point 3), to the north (grid point 4), and to the east (grid point 5), with their BMC weighting computed as 30%, 10%, 20%, 30%, and 10%, respectively. Clearly, upstream information at grid points 1 and 4 is being weighted heavily (60%), and in this situation we find that the forecasts (both the EP and RFv2 at grid point 4, and the EP at grid point 1) are indicating cold air in proximity to the grid point of interest (grid point 3). One can then ask what contributed to the colder temperatures in the individual EP forecast at grid point 1? The algorithm places considerable emphasis on the extremes of the RFv2 ensemble forecast (the minimum and maximum values), and in particular, where relatively large differences occur indicating the presence of a significant temperature gradient, the forecast is skewed colder. The overall effect of these weights is to shift the probability distribution toward colder temperatures. We note that a similar effect was

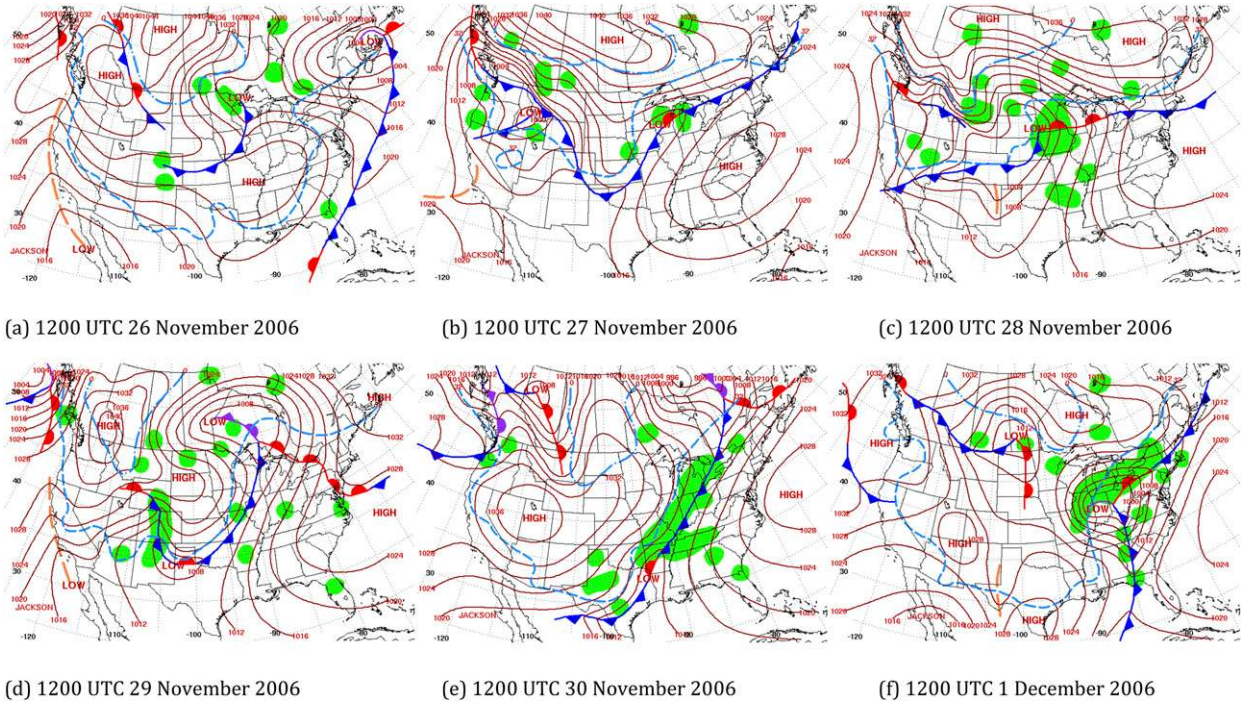

FIG. 3. (a)–(f) Surface analyses from 1200 UTC 26 Nov to 1200 UTC 1 Dec 2006, obtained from the NOAA Daily Weather Map series.

accomplished in the deterministic forecast at grid point 3 owing to the same sequence of genetic code as at grid point 1—large differences in the RFv2 minima and maxima skew the forecast colder.

To quantify the probabilistic forecasts overall, we compute the ranked probability score[1] (RPS; Epstein 1969; Murphy 1969, 1971) and the associated skill score (RPSS) referenced to climatology (Table 2). To correct the RPSS for the negative bias associated with ensemble size, we apply the correction methodology of Weigel et al. (2007). The RPSS calculation shows a small (3.3%) improvement for the EP relative to the bias-corrected RFv2 ensemble across the entire domain. There is consistency in this result; however, as EP improvement in both RMSE and RPSS exists at grid points simultaneously over 96% of the domain.

The longitudinal axis of lowest RFv2 predictability, based on combined ranks for RMSE and RPSS, is along 99°W, making it of interest to inspect relative performance there (Fig. 5). Although the latitudinal profiles are similar, the EP is consistently better than the bias-corrected RFv2 in both deterministic and probabilistic performance. Some insight can be gained into these characteristics by examination of forecast error distributions. When considering the spatial distribution of the difference in probabilities associated with specific forecast error increments between the EP and the bias-corrected RFv2 ensemble (Fig. 6), a shift from higher toward lower forecast errors is apparent in most regions.

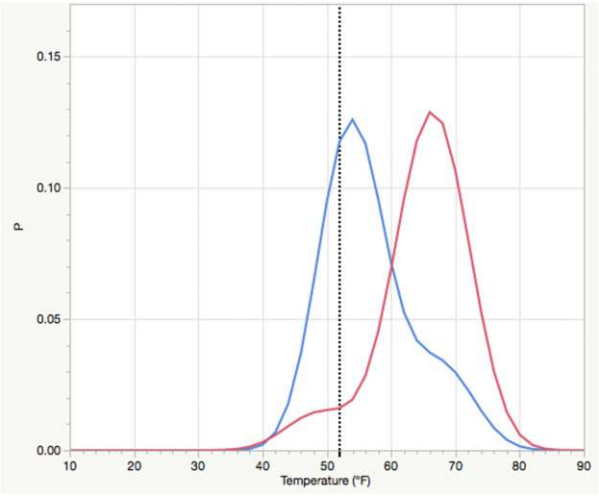

FIG. 4. Evolutionary program (blue) and RFv2 bias-corrected ensemble (red) probabilistic temperature forecasts for 0000 UTC 1 Dec 2006 for the grid point corresponding to 33°N, 99°W (located about 250 km west of Dallas, TX). The analyzed temperature is the vertical dashed line.

[1] Probabilities are computed within temperature intervals of width 2°F from −80°F to +120°F. Thus, each bin has a verification of 1 (for occurrence) and 0 for all other 2°F bins for that case, and the mean square error summed across all the bins becomes the multicategory equivalent of the Brier score, the ranked probability score.

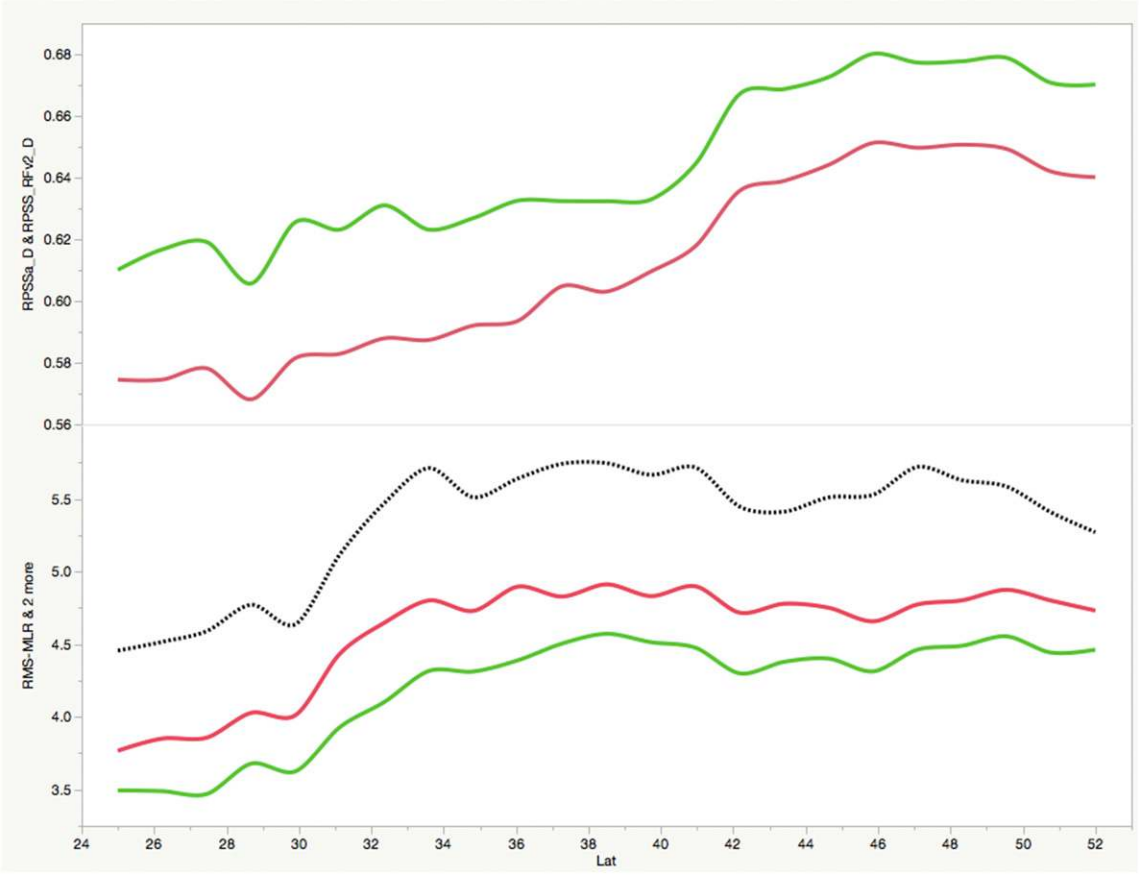

FIG. 5. (top) Ranked probability skill score (larger is higher skill) and (bottom) root-mean-square error (lower is better performance) for the evolutionary program (green) and bias-corrected RFv2 ensemble (red) at 99°W, extending from 25° to 52°N. Also shown is the root-mean-square error for the multiple linear regression (black dashed).

What information does the EP use in order to gain this forecast advantage? This can best be answered through application of relative weights (Cooksey 1996; Roebber 1998: Roebber 2015b). This method provides a scaled "weighting" from 0 to 100 of each forecast input, based on a regression analysis that assigns the amount of variance in the forecast that can be uniquely attributed to that input. This can be obtained from the squared semipartial correlation (sr^2), which is the difference between the squared multiple correlation (R^2) from a regression using all available forecast inputs and a second R^2 based upon a regression where the input of interest has been removed. The sr^2 of the ith input is then converted to a relative weight for M inputs as follows:

$$rw_i = 100 \frac{sr_i^2}{\sum_{i=1}^{M} sr_i^2}, \tag{11}$$

with the M weights summing to 100 units.

The highest weighted forecast input for the EP at most locations is the median, bias-corrected RFv2 forecast, an unsurprising result since the ensemble mean forms the baseline forecast and the ensemble median and the mean are highly correlated. Nonetheless, this weighting accounts for only about 50% of the variance and there is considerable variability in the next most important piece of forecast information (Fig. 7). For example, in the northeastern portion of the domain, the second-most important input is often snow cover, whereas in the desert southwest, it is precipitable water, and in the northern Great Plains, it is the cosine of the solar zenith angle. In a few locations, the analog forecast is found to be a useful secondary source of information for these forecasts. Finally, we note that the list of predictors (Table 1) is not comprehensive and the improvements reported here are not necessarily the limit of what is obtainable using these approaches.

Rank histograms (Anderson 1996; Hamill and Colucci 1997; Talagrand et al. 1997) are a means of

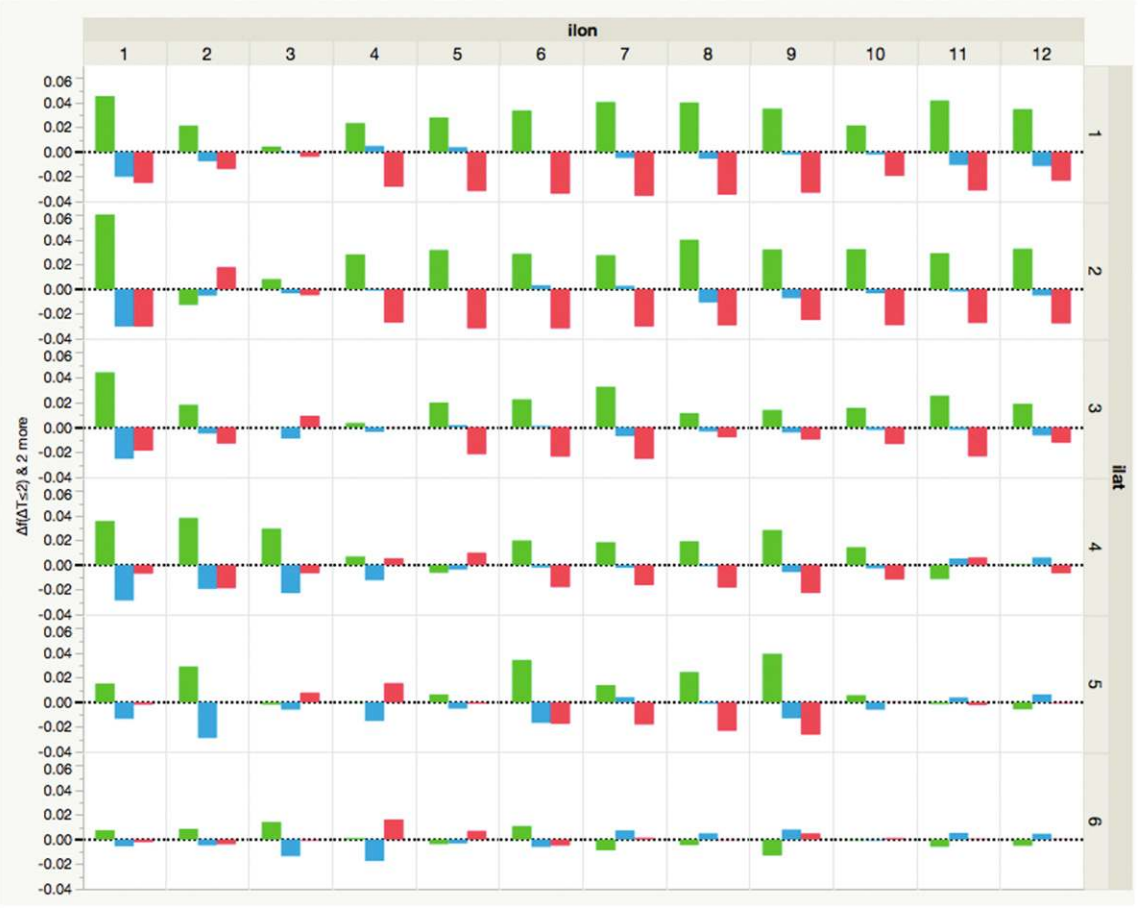

FIG. 6. Difference (evolutionary program − RFv2 ensemble) in cumulative probabilities of forecast errors ≤ 2° (green), > 2° and ≤ 5° (blue), and > 5° (red). Distributions are shown for latitudes of (bottom) 25°–30°; (top to next to bottom) 30°–35°, 35°–40°, 40°–45°, 45°–50°, and >50°N; and for longitudes of (left) 125°–120° and (next to left to right) 120°–115°, 115°–110°, 110°–105°, 105°–100°, 100°–95°, 95°–90°, 90°–85°, 85°–80°, 80°–75°, 75°–70°, <70°W.

quantifying the reliability of forecast probabilities, that is, the ensemble relative frequency should be a reliable indication of the observed occurrence. An extension of this is to consider the frequency of *outliers*, those events that verify outside of the range indicated by the ensemble (Siegert et al. 2011). Specifically, one expects for a consistent K-member ensemble that outliers will occur with a base rate of $2/(K + 1)$. For the 11-member, bias-corrected RFv2, the U-shaped rank histogram (Fig. 8, top) with a 20.9% excess outlier percentage indicates substantial underdispersion, a problem common to numerical weather prediction models. The rank histogram of the five-member, bias-corrected EP ensemble still shows underdispersion (Fig. 8, bottom), however the excess outlier percentage is reduced to 4.8%, indicating improved reliability. These EP results are comparable to those reported for the ECMWF and NCEP ensemble prediction systems by Buizza et al. (2005).

a. Carrying capacity

The EP algorithms in Roebber (2015a) were developed for a single location, and accordingly the carrying capacity of the population was set to a relatively large number (10 000 individuals). Here, we are training algorithms across an 1800 point grid, and have set the carrying capacity to 50 individuals per grid location, which keeps the computational demands of the training process relatively equivalent to that of Roebber (2015a). A question that naturally arises is whether this reduced carrying capacity constrains the ability of the method to develop the best solutions.

As a means of addressing this point, tests were run at a single grid point where the forecasts are relatively demanding (nx = 27, ny = 10; the same location discussed previously). For these tests, the training and testing is accomplished as previously except that all solutions are located only at the grid point in question, and

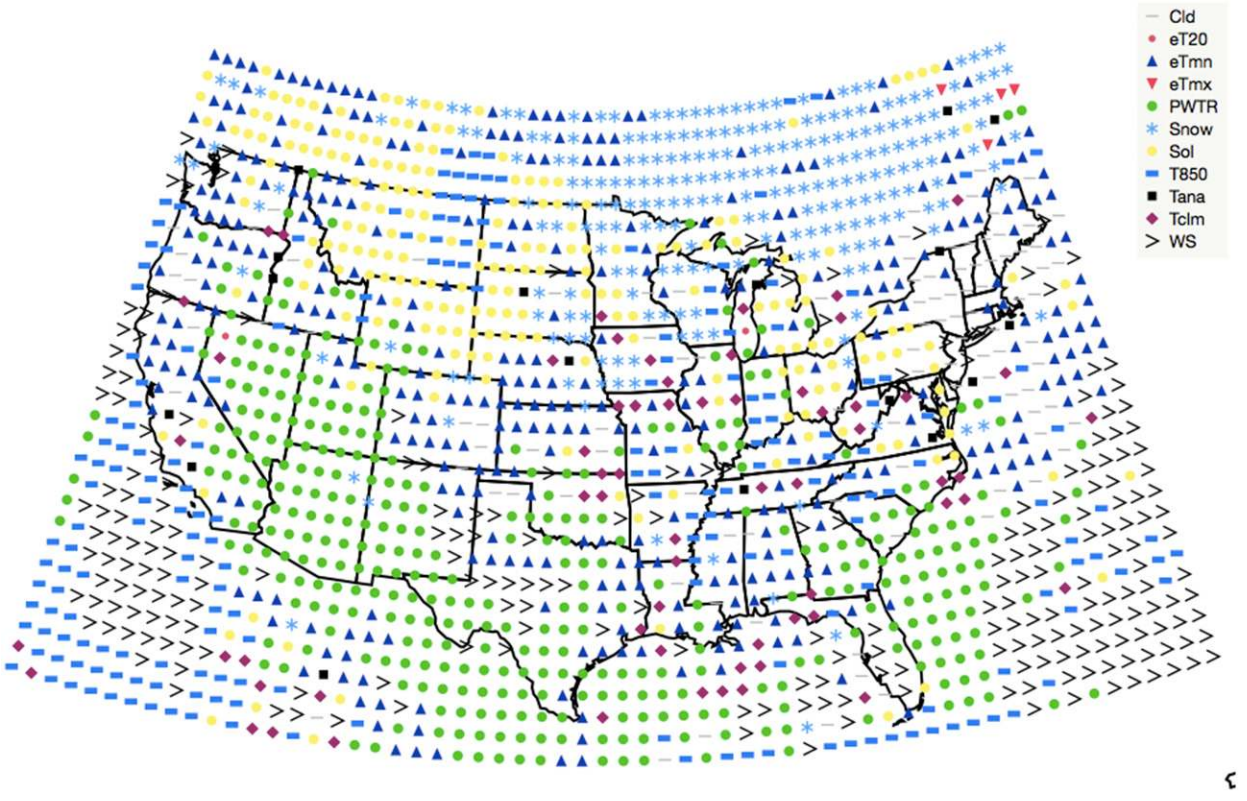

FIG. 7. Highest weighted forecast input (Table 1) for the evolutionary program solution at each grid, excluding the RFv2 median forecast member.

the carrying capacity is set to 50, 100, 500, 1000, and 10 000 individuals. We find that as carrying capacity increases from 50 to 10 000 individuals, the RMSE reduces by 2.7%. Given that larger RMSE reductions are obtainable through the much more efficient bias correction process [Eqs. (4)–(5)], using a smaller carrying capacity in exchange for faster processing seems worthwhile. Such trade-offs become more important when considering adaptive forms (section 3b) or higher-resolution domains (section 3c).

b. *Adaptive application*

Roebber (2015c) developed an adaptive evolutionary programming method suitable for single-site data. Here, we generalize that result to the spatially dependent approach explored in this article. As in Roebber (2015c), we employ a ''mixed-mode'' evolution wherein the overall IF-THEN genetic framework, which has evolved up to case M, uses the next generation to supply the forecast for case $M + 1$, but where the EP coefficients $C_{1,j}$, $C_{2,j}$, and $C_{3,j}$ are adjusted in an effort to optimize them for the case at hand.

The process begins exactly as in section 2d: we train an initial architecture using 50 generations and the 4383 training and 2922 cross-validation days, saving the best performing algorithm at each grid point at any point in this training, based on RMSE on the cross-validation data. We then evolve the IF-THEN architecture using a moving window of 2 yr of training data and 1 yr of cross-validation data, such that to produce a forecast for case 7306, the training interval is from case 6211–6940 inclusive, and the cross-validation interval is from case 6941–7305 inclusive, and these ranges are updated by 1 with each new forecast.

Here, the ''fast mode'' is that in which we adjust the coefficients $C_{1,j}$, $C_{2,j}$, and $C_{3,j}$ to minimize the RMSE on the cross-validation data for that particular IF-THEN architecture. If the performance of this newly evolved EP algorithm (including the fast mode coefficients), based on RMSE on the cross-validation data, is superior to that of the prior ''best'' algorithm at that grid point, then it becomes the new best performer and is the algorithm that is used to produce forecasts for that grid point until such time as it is itself replaced. The coefficients $C_{1,j}$, $C_{2,j}$, and $C_{3,j}$ are randomly selected from the interval $[-1, 1]$ and this selection is repeated 10 times, with each combination evaluated to determine performance. The choice of 10 selections is

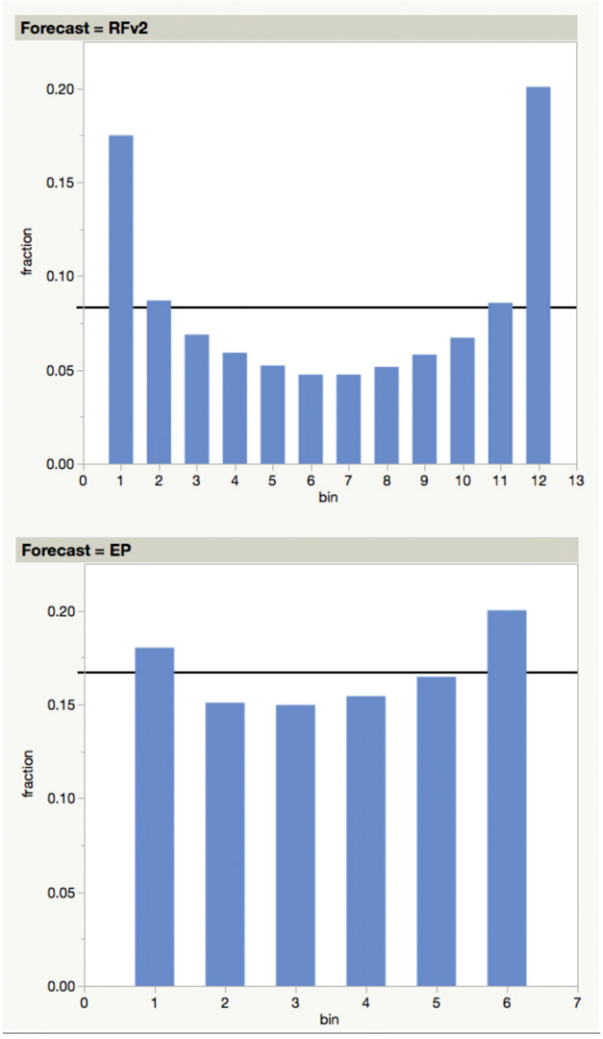

FIG. 8. Ranked histograms for (top) the 11-member, bias-corrected RFv2 and (bottom) the 5-member, bias-corrected EP ensemble. The base rate frequency for each bin is shown as the horizontal line.

arbitrary and is based on balancing the need to trial a variety of weights against computational considerations, as these calculations must be repeated at every grid point.

Since we are shortening the training intervals but not adding improved information to the system in this test, some degradation in performance is expected and does occur (Table 2). The relatively small increase in RMSE (0.09° and 0.04°F for the raw and bias-corrected forecasts, respectively) and decrease in RPSS (0.006), however, seems acceptable given the benefit of adding an adaptive capability that can effectively incorporate improved information without the need for retraining, as shown in Roebber (2015c).

c. *Future extensions*

The ''ecosystem model'' applied here is a simple, static one—fixed births and deaths at each iteration leading to a fixed population size for a single solution ''species.'' It is possible to implement more dynamic models, which could incorporate multiple species interactions over space and time, such as predation and/or competitive exclusion (e.g., Lugo and McKane 2008; Sprott 2008; Biswas et al. 2014). Preliminary experiments with such an approach suggest that it can produce additional, incremental gains in algorithm forecast skill, but that the most substantial contribution of such an approach may be in its potential to improve ensemble diversity and thereby increase probabilistic forecast skill (e.g., Carja et al. 2014). However, such improvements are not yet clearly established and remain a subject of current investigation. With regard to probabilistic improvements, application of a different cost metric for training may also be preferable. Here, we have employed RMSE and, as shown in Roebber (2015a,b), there is evidence that efforts to optimize with respect to deterministic skill can come at the expense of probabilistic skill. Experiments using RPS as a cost metric may shed some light on this question.

Although we have demonstrated relative success using this method for spatial data, extending the spatial ''footprint'' across three grid points of coarse resolution (i.e., ~300 km) likely reduces the effectiveness of the approach. In the current configuration, a ''nearby'' point 100 km distant, which may be at a much higher elevation, will be steeply discounted in the BMC process, but this effectively reduces the information available to the combination. A three-category stratification of grid points based upon terrain (water, sloped, not sloped) for the 1° × 1° data (Fig. 9) and for 3-km gridded data (Fig. 10) demonstrates the difference. A methodology that can account for ''likeness'' of spatially nearby points would potentially better utilize the ability of BMC to combine information. An example of a forecast dataset, well suited for the purpose of testing this procedure, would be the HRRR (Benjamin et al. 2016).

Using the 3-km HRRR, one might define neighborhoods as previously except in doing so, one would also specifically account for the likeness of grid points within an area. Consider a synoptic field ''overlaid'' on a region with complex orography (terrain, coastal zones). Substantial variations in temperature can arise from variable terrain, and within coastal zones (e.g., lake and sea- breeze boundaries along the Great Lakes and eastern, western, and southern shores, and back door cold fronts in New England). A considerable fraction of

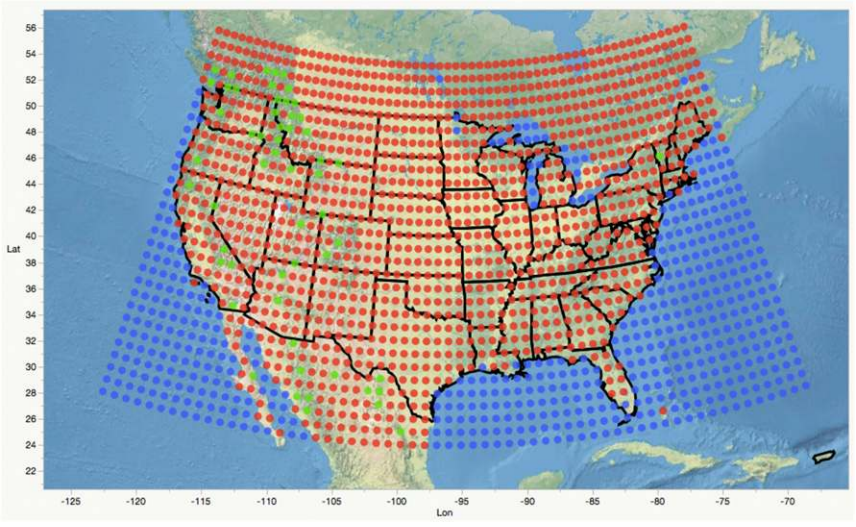

FIG. 9. A three-category stratification of grid points based upon terrain (water, sloped, not sloped) for the 1° × 1° data and for 3-km gridded data (se next, Fig. 10) demonstrates the difference. Shown are water (blue), sloped (green), and not sloped (red).

regional variability in such regions will necessarily reflect the interaction of the orography with the synoptic signal.

We can account for this by classifying each grid point in the domain according to its orographic similarity with surrounding points as in Fig. 10. One could then proceed radially outward from each point until some minimal set of *M*-like points are identified to form the collection for BMC. Thus, a point will be matched to all points of similar maximum slope classification by proceeding outward in all directions up to some distance *R* until the *M* points are obtained. This set of *M* points would constitute the neighborhood used in the BMC procedure.

Applying this type of classification to the map of Fig. 10, based upon a maximum slope of 300 m (3 km)$^{-1}$ (i.e., a grid point is considered sloped if the terrain increases or decreases at a rate greater than this amount), results in more than 96% of the domain points with at least 100 like points within a radial distance of 50 km or less. For the purposes of BMC, this is considerably more like points than are needed (or are feasible) to form ensembles and it is expected that some experimentation would be required to find the optimal combination of

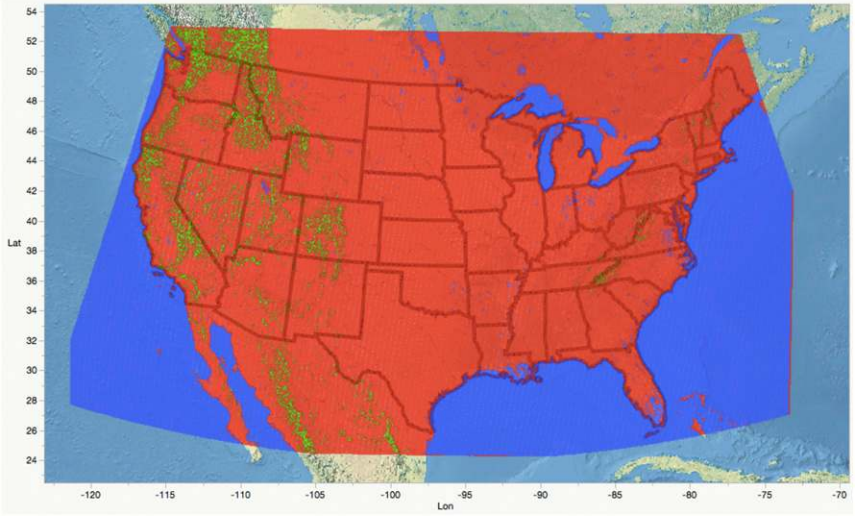

FIG. 10. As in Fig. 9, but for 3-km gridded data.

performance and computational efficiency. This latter would also be required given the $O(700)$ increase in grid points needed to be considered relative to the RFv2 for the otherwise similar spatial domain. Extension of these methods to other forecasts of interest (e.g., winter weather, severe weather, etc.) are likewise of interest. An alternative approach to identifying likeness of spatially nearby points is the ''mother–daughter'' methods of Deng and Stull (2005, 2007).

It is important to note that the present study has used *analysis data* rather than station observations for all forecast verifications (EP, MLR, RFv2, Analog). Prior work (e.g., Roebber 2015a) has shown that the EP method works well for station observations, so this is not a general limiting factor for the approach. Nonetheless, details related to translating forecasts from a grid point to a specific station location are nontrivial and might require some additional correction procedure as part of the interpolation. It might be expected that as higher-resolution analyses and forecasts become more ubiquitous, the translations will become less subject to error given the ability of the higher resolution to better capture physiographic variation with a region. In the present study, one should expect that the analysis is smoother than what one might see in the station data, and thus the reported errors for all the forecasts are likely smaller than that which would be obtained using station data; nonetheless, this bias applies to all the methods studied and we do not expect the relative performance of the methods reported here to change.

4. Summary

In this paper, we have developed a new approach to producing evolutionary program forecasts that accounts for the spatial dimension of the data. The approach is based upon a static ecosystem model in which the total number of individuals (forecast algorithms) at a grid location is held constant, for a single solution ''species.'' The evolution occurs based on selection pressure—the highest-performing individuals based on deterministic forecast skill are allowed to produce the next generation, and the poorest-performing individuals based upon that same measure are eliminated.

The approach was applied to 72-h forecast data for temperature, obtained from the 11-member Reforecast V2 ensemble, on a 1° latitude–longitude grid for the region from 24°–53°N to 125°–66°W. Training, cross validation, and testing were accomplished for the period 1 January 1985–14 May 2011. After bias correction, the method was shown to improve deterministic (root-mean-square error) and probabilistic (ranked probability skill score) forecasts compared to the Reforecast V2 ensemble across this domain by 0.31°F (8.7%) and 3.3%, respectively. This improvement was widespread, with 96% of the grid points showing improvements in both root-mean-square error and ranked probability skill score, although the key information used in these forecasts varied substantially with geography. An adaptive form of the evolutionary programming approach, which would be possible to implement operationally, was tested using fixed rather than improving inputs and found to lead to relatively little skill degradation (1.2% and 0.6% for deterministic and probabilistic, respectively). Future extensions to this work include the following:

- more comprehensive inputs;
- refinements to the adaptive approach, particularly with respect to the setting of algorithm coefficients under the ''fast evolutionary mode'';
- implementation of a dynamic rather than a static ecosystem model to enhance solution diversity, in combination with a success measure such as the ranked probability score; and
- application of the methodology to higher-resolution datasets where ensemble construction could account for local geographic variation.

Acknowledgments. This research was supported in part by the 2015 University Corporation for Atmospheric Research (UCAR) Developmental Testbed Center (DTC) Visitor program. I am grateful to UCAR and to the National Center for Atmospheric Research (NCAR) for their support, and to Louisa Nance (DTC), Isidora Jankov (NOAA-ESRL), and Trevor Alcott (NOAA-ESRL) for their interest in this work. This research was also supported in part by the Cooperative Institute for Research in the Atmosphere.

REFERENCES

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.

Bakhshaii, A., and R. Stull, 2009: Deterministic ensemble forecasts using gene- expression programming. *Wea. Forecasting*, **24**, 1431–1451, https://doi.org/10.1175/2009WAF2222192.1.

Basak, A., S. Pal, S. Das, A. Abraham, and V. Snasel, 2010: A modified Invasive Weed Optimization algorithm for time-modulated linear antenna array synthesis. *Proc. 2010 IEEE Congress on Evolutionary Computation* (*CEC*), Barcelona, Spain, IEEE, https://doi.org/10.1109/CEC.2010.5586276.

Benjamin, S. G., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The rapid refresh. *Mon. Wea. Rev.*, **144**, 1669–1694, https://doi.org/10.1175/MWR-D-15-0242.1.

Biswas, R., C. Ofria, D. M. Bryson, and A. P. Wagner, 2014: Causes vs benefits in the evolution of prey grouping. *Proc. ALIFE 14: 14th Int. Conf. on the Synthesis and Simulation of Living Systems*, New York, NY, International Society for Artificial Life, 641–648.

Buizza, R., P. Houtekamer, G. Pellerin, Z. Toth, Y. Zhu, and M. Wei, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.*, **133**, 1076–1097, https://doi.org/10.1175/MWR2905.1.

Carja, O., U. Lieberman, and M. W. Feldman, 2014: Evolution in changing environments: Modifiers of mutation, recombination, and migration. *Proc. Natl. Acad. Sci. USA*, **111**, 17 935–17 940, https://doi.org/10.1073/pnas.1417664111.

Cooksey, R. W., 1996: *Judgment Analysis: Theory, Methods and Applications*. Academic Press, 407 pp.

Cui, B., Z. Toth, Y. Zhu, and D. Hou, 2012: Bias correction for global ensemble forecast. *Wea. Forecasting*, **27**, 396–410, https://doi.org/10.1175/WAF-D-11-00011.1.

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, https://doi.org/10.1175/2011MWR3653.1.

Deng, X., and R. Stull, 2005: A mesoscale analysis method for surface potential temperature in mountainous and coastal terrain. *Mon. Wea. Rev.*, **133**, 389–408, https://doi.org/10.1175/MWR-2859.1.

——, and ——, 2007: Assimilating surface weather observations from complex terrain into a high-resolution numerical weather prediction model. *Mon. Wea. Rev.*, **135**, 1037–1054, https://doi.org/10.1175/MWR3332.1.

Epstein, E., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.*, **8**, 985–987, https://doi.org/10.1175/1520-0450(1969)008<0985:ASSFPF>2.0.CO;2.

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.

Gneiting, T., A. E. Raftery, 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, https://doi.org/10.1175/MWR2904.1.

Hamill, T. M., 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 dataset. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565, https://doi.org/10.1175/BAMS-D-12-00014.1.

——, and S. J. Colucci, 1997: Verification of Eta-RSM short-range ensemble forecasts. *Mon. Wea. Rev.*, **125**, 1312–1327, https://doi.org/10.1175/1520-0493(1997)125<1312:VOERSR>2.0.CO;2.

Hemri, S., M. Scheuerer, F. Pappenberger, K. Bogner, and T. Haiden, 2014: Trends in the predictive performance of raw ensemble weather forecasts. *Geophys. Res. Lett.*, **41**, 9197–9205, https://doi.org/10.1002/2014GL062472.

Hennon, C. C., C. Marzban, and J. S. Hobgood, 2005: Improving tropical cyclogenesis statistical model forecasts through the application of a neural network classifier. *Wea. Forecasting*, **20**, 1073–1083, https://doi.org/10.1175/WAF890.1.

Houtekamer, P. L., and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.*, **126**, 796–811, https://doi.org/10.1175/1520-0493(1998)126<0796:DAUAEK>2.0.CO;2.

Koizumi, K., 1999: An objective method to modify numerical model forecasts with newly given weather data using an artificial neural network. *Wea. Forecasting*, **14**, 109–118, https://doi.org/10.1175/1520-0434(1999)014<0109:AOMTMN>2.0.CO;2.

Kuligowski, R. J., and A. P. Barros, 2001: Combined IR-microwave satellite retrieval of temperature and dewpoint profiles using artificial neural networks. *J. Appl. Meteor.*, **40**, 2051–2067, https://doi.org/10.1175/1520-0450(2001)040<2051:CIMSRO>2.0.CO;2.

Lugo, C. A., and A. J. McKane, 2008: Quasicycles in a spatial predator-prey model. *Phys. Rev. E*, **78**, 51 911–51 925, https://doi.org/10.1103/PhysRevE.78.051911.

Mehrabian, A. R., and C. Lucas, 2006: A novel numerical optimization algorithm inspired from weed colonization. *Ecol. Inform.*, **1**, 355–366, https://doi.org/10.1016/j.ecoinf.2006.07.003.

Messner, J. W., G. J. Mayr, and A. Zeileis, 2017: Nonhomogeneous boosting for predictor selection in ensemble postprocessing. *Mon. Wea. Rev.*, **145**, 137–147, https://doi.org/10.1175/MWR-D-16-0088.1.

Monteith, K., J. Carroll, K. Seppi, and T. Martinez, 2011: Turning Bayesian model averaging into Bayesian model combination. *Proc. Int. Joint Conf. on Neural Networks* (*IJCNN'11*), San Jose, CA, IEEE, 2657–2663, https://doi.org/10.1109/IJCNN.2011.6033566.

Murphy, A. H., 1969: On the "ranked probability score." *J. Appl. Meteor.*, **8**, 988–989, https://doi.org/10.1175/1520-0450(1969)008<0988:OTPS>2.0.CO;2.

——, 1971: A note on the ranked probability score. *J. Appl. Meteor.*, **10**, 155–156, https://doi.org/10.1175/1520-0450(1971)010<0155:ANOTRP>2.0.CO;2.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, https://doi.org/10.1175/MWR2906.1.

Roebber, P. J., 1998: The regime dependence of degree day forecast technique, skill and value. *Wea. Forecasting*, **13**, 783–794, https://doi.org/10.1175/1520-0434(1998)013<0783:TRDODD>2.0.CO;2.

——, 2010: Seeking consensus: A new approach. *Mon. Wea. Rev.*, **138**, 4402–4415, https://doi.org/10.1175/2010MWR3508.1.

——, 2015a: Evolving ensembles. *Mon. Wea. Rev.*, **143**, 471–490, https://doi.org/10.1175/MWR-D-14-00058.1.

——, 2015b: Ensemble MOS and evolutionary program minimum temperature forecast skill. *Mon. Wea. Rev.*, **143**, 1506–1516, https://doi.org/10.1175/MWR-D-14-00096.1.

——, 2015c: Adaptive evolutionary programming. *Mon. Wea. Rev.*, **143**, 1497–1505, https://doi.org/10.1175/MWR-D-14-00095.1.

——, M. R. Butt, S. J. Reinke, and T. J. Grafenauer, 2007: Real-time forecasting of snowfall using a neural network. *Wea. Forecasting*, **22**, 676–684, https://doi.org/10.1175/WAF1000.1.

Scheuerer, M., and T. M. Hamill, 2015: Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted Gamma distributions. *Mon. Wea. Rev.*, **143**, 4578–4596, https://doi.org/10.1175/MWR-D-15-0061.1.

Siegert, S., J. Bröcker, and H. Kantz, 2011: Predicting outliers in ensemble forecasts. *Quart. J. Roy. Meteor. Soc.*, **137**, 1887–1897, https://doi.org/10.1002/qj.868.

Sprott, J. C., 2008: Predator-prey dynamics for rabbits, trees and romance. *Unifying Themes in Complex Systems IV*, A. A. Minai and Y. Bar-Yam, Eds., Springer, 231–238, https://doi.org/10.1007/978-3-540-73849-7_26.

Taillardat, M., O. Mestre, M. Zamo, and P. Naveau, 2016: Calibrated ensemble forecasts using quantile regression forests

and ensemble model output statistics. *Mon. Wea. Rev.*, **144**, 2375–2393, https://doi.org/10.1175/MWR-D-15-0260.1.

Talagrand, O., R. Vautard, and B. Strauss, 1997: Evaluation of probabilistic prediction systems. *Proc. ECMWF Workshop on Predictability,* Reading, United Kingdom, ECMWF, 1–25. [Available from ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, United Kingdom.]

Teisberg, T. J., R. F. Weiher, and A. Khotanzad, 2005: The economic value of temperature forecasts in electricity generation. *Bull. Amer. Meteor. Soc.*, **86**, 1765–1771, https://doi.org/10.1175/BAMS-86-12-1765.

U.S. Energy Information Administration, 2014: Northeast natural gas spot prices particularly sensitive to temperature swings. *Today in Energy*, 11 August 2014, https://www.eia.gov/todayinenergy/detail.php?id=17491.

Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2007: The discrete Brier and ranked probability skill scores. *Mon. Wea. Rev.*, **135**, 118–124, https://doi.org/10.1175/MWR3280.1.

Yang, H. T., C. M. Huang, and C. L. Huang, 1996: Identification of ARMAX model for short-term load forecasting: An evolutionary programming approach. *IEEE Trans. Power Syst.*, **11**, 403–408, https://doi.org/10.1109/59.486125.