Using Evolutionary Programs to Maximize Minimum Temperature Forecast Skill

PAUL J. ROEBBER

University of Wisconsin–Milwaukee, Milwaukee, Wisconsin

(Manuscript received 22 March 2014, in final form 4 February 2015)

ABSTRACT

Evolutionary program ensembles are developed and tested for minimum temperature forecasts at Chicago, Illinois, at forecast ranges of 36, 60, 84, 108, 132, and 156 h. For all forecast ranges examined, the evolutionary program ensemble outperforms the 21-member GFS model output statistics (MOS) ensemble when considering root-mean-square error and Brier skill score. The relative advantage in root-mean-square error widens with forecast range, from 0.18°F at 36 h to 1.53°F at 156 h while the probabilistic skill remains positive throughout. At all forecast ranges, probabilistic forecasts of abnormal conditions are particularly skillful compared to the raw GFS guidance.

The evolutionary program reliance on particular forecast inputs is distinct from that obtained from considering multiple linear regression models, with less reliance on the GFS MOS temperature and more on alternative data such as upstream temperatures at the time of forecast issuance, time of year, and forecasts of wind speed, precipitation, and cloud cover. This weighting trends away from current observations and toward seasonal (climatological) measures as forecast range increases.

Using two different forms of ensemble member subselection, a Bayesian model combination calibration is tested on both ensembles. This calibration had limited effect on evolutionary program ensemble skill but was able to improve MOS ensemble performance, reducing but not eliminating the skill gap between them. The largest skill differentials occurred at the longest forecast ranges, beginning at 132 h. A hybrid, calibrated ensemble was able to provide some further increase in skill.

1. Introduction

The utility of deterministic weather forecasts is undermined by uncertainties arising from sensitive dependence on initial conditions (Lorenz 1963 and many others) and model error (e.g., Harrison et al. 1999; Stensrud et al. 2000; Orrell 2005). Probabilistic forecasts are the logical alternative, since they can be used to quantify the degree of this uncertainty. In meteorology, probabilistic forecasts are most often generated using ensembles of numerical weather prediction (NWP) models. Such ensembles use perturbed initial conditions (via breeding, singular vectors, or Monte Carlo methods), sometimes accompanied by variable model physics (e.g., Buizza et al. 1999; Stensrud et al. 2000; Shutts and Palmer 2007). As forecast range increases, the forecasts will diverge (at rates dependent upon the characteristics of the flow) until such time as the separation between forecasts is that of randomly chosen atmospheric states (i.e., the forecast is no better than climatology).

Despite this strong conceptual basis, NWP model ensemble forecasts have been hampered by the problem of underdispersion [e.g., Buizza (1997); see Novak et al. (2008) for a discussion of the effect on operations]. The use of perturbed model physics and statistical postprocessing has improved (e.g., Harrison et al. 1999; Stensrud et al. 2000; Hamill and Whitaker 2007 and references therein) but not resolved this performance problem. Further, owing to computational constraints, NWP model ensembles are typically restricted to a relatively small number (order 20–50 members), and thus the estimate of the probability distribution function and accompanying spread is less optimal than might be the case with a larger sample, particularly where more unusual event detection is concerned.

Evolutionary programming (EP) is a process by which the principles of evolution are used to map a particular set of inputs to a desired output. Work in EP has been ongoing since the 1960s (e.g., Fogel 1999), although the application of the idea in meteorological fields has been relatively limited (Yang et al. 1996; Lakshmanan 2000;

Corresponding author address: Paul J. Roebber, Atmospheric Science Group, Department of Mathematical Sciences and School of Freshwater Sciences, University of Wisconsin–Milwaukee, 3200 North Cramer Ave., Milwaukee, WI 53211.
E-mail: roebber@uwm.edu

DOI: 10.1175/MWR-D-14-00096.1

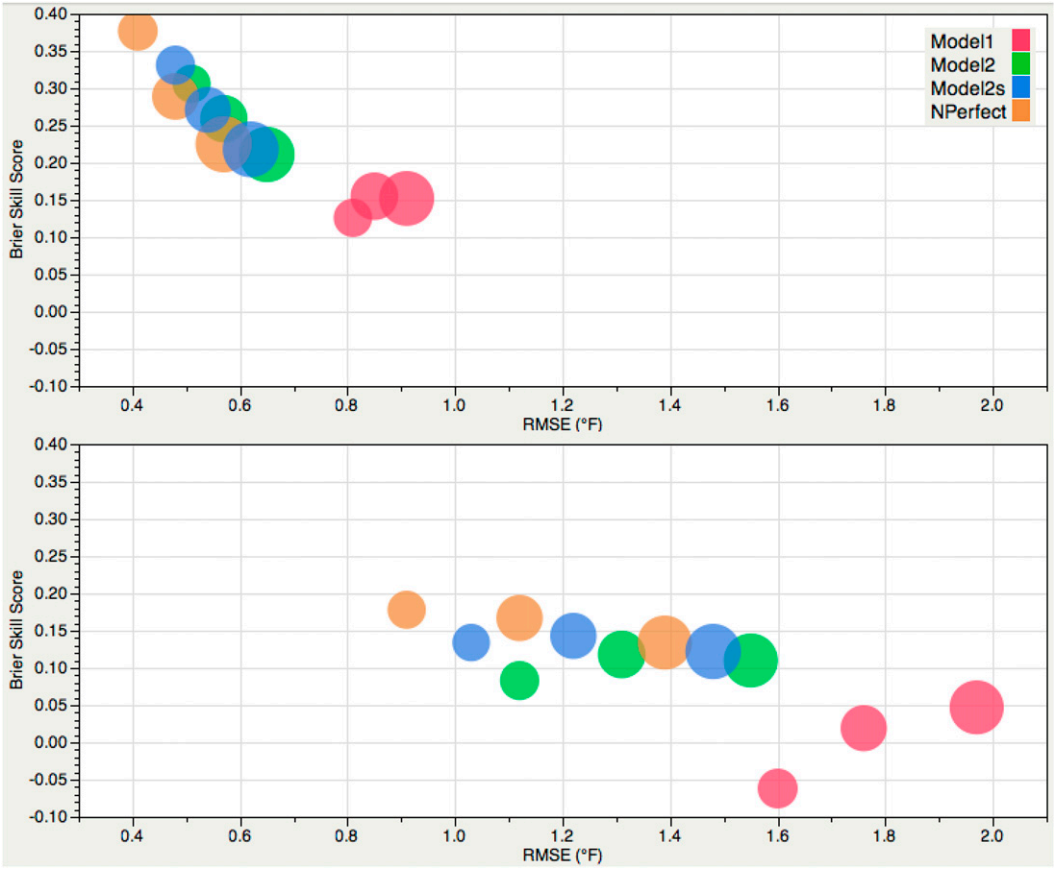

FIG. 1. Brier skill score vs root-mean-square error for four forecast versions of the Lorenz (1963) system. "Truth" is constructed using $\sigma = 10$, $\rho = 28$, and $\beta = 10/3$; model 1 uses $\sigma = 10$, $\rho = 26$, and $\beta = 9/3$; and model 2 uses $\sigma = 10$, $\rho = 27$, and $\beta = 9/3$. A stochastic physics version of model 2 (model 2 s) is also used, in which $\sigma = 10$, $\rho = 27 \pm 0.25$, and $\beta = 9/3 \pm 0.2$. Finally, "near perfect" is a model using $\sigma = 10$, $\rho = 27.25$, and $\beta = 9.25/3$. The results are shown for three values of analysis uncertainty: 0.2, 0.3, and 0.4 standard deviations of X, Y, and Z (small, medium, and large circles, respectively). Two forecast ranges are shown: (top) 5 steps and (bottom) 10 steps.

Coulibaly 2004; Haupt et al. 2006; O'Steen and Werth 2009; Bakhshaii and Stull 2009; Roebber 2010). Roebber (2013, 2015) developed a modified form of EP to generate large (order 1000–3000 member) ensembles. Unlike a NWP model postprocessor such as model output statistics (MOS), however, the method is not restricted solely to NWP inputs and in fact typically performs better when observations are included (e.g., Roebber 2010, 2015). Thus, it might more correctly be considered as a form of nonlinear statistical forecast model. Roebber (2010) was primarily focused on the best deterministic forecast produced through EP, whereas Roebber (2013, 2015) extends that work to investigate probabilistic as well as deterministic forecast performance.

Attempts to minimize ensemble mean forecast error, however, may work at cross purposes to maximizing the skill of probabilistic forecasts. Roebber (2015), for example, showed that seeking to minimize EP ensemble mean square error also tends to reduce the diversity of the solutions. A general example of this contention between deterministic and probabilistic skill is provided in Fig. 1, in the context of a series of simulations using versions of the Lorenz (1963) chaotic model (sample size of 40 000 points for each simulation). Three behaviors are apparent. First, for any amount of analysis error, as models improve, both the ensemble mean forecasts and the probabilistic forecasts improve (i.e., the points shift toward the upper left of the figure). Second, for models that contain relatively substantial "physics" errors, improvements in analyses will lead to better ensemble mean forecasts but not necessarily better probabilistic forecasts [as measured by the Brier skill score (BSS); Brier (1950)]. As is observed with more complex ensemble models, there is also the characteristic underdispersion. This is overcome only when the model physics improve past some unknown critical threshold. Third, this nonuniform behavior with respect to the ensemble mean and the probabilistic forecasts is

dependent upon forecast range. In the example, ''model 1'' exhibits this behavior for both forecast ranges examined, but it emerges in the regular and stochastic physics versions of ''model 2'' only at longer forecast range.

Using minimum temperature data for Chicago, Illinois, Roebber (2015) demonstrated that EP ensembles can produce forecasts superior to those of the 21-member GFS ensemble MOS system (see www.nws.noaa.gov/mdl/synop/enstxt.php), both in terms of the ensemble mean root-mean-square error (RMSE) and the BSS of the resulting forecasts. The degree of performance improvement provided by the EP ensemble was found to be dependent on the details of the calibration method employed, particularly with respect to probabilistic skill. For RMSE, the advantage ranged up to several tenths of a degree Fahrenheit, whereas for BSS this advantage decreased from +4.7% for the simplest calibration to +1.2% using Bayesian model combination [BMC; see Monteith et al. (2011)]. Further improvement was found by using BMC on a pooled set of ensemble members derived from the EP ensemble and the GFS MOS ensemble. While these improvements over a properly calibrated NWP forecast are incremental, in specific user contexts, they can provide considerable value [e.g., Teisberg et al. (2005) for electricity generation].

In this study, we will consider the GFS ensemble MOS as one model system and the EP ensemble as another model system. The results of Roebber (2015) suggest the possibility of ensemble behavior conceptually akin to model 2 (for EP) compared to model 1 (for GFS MOS) for a given analysis error and forecast range as in Fig. 1. We do not know how this characteristic behavior might change as forecast range changes, however. Second, we do not know how the mapping of inputs to the desired output (minimum temperature) will change as forecast range changes. Finally, we do not know how these characteristics might be altered using different calibration methods.

To explore these aspects, we will extend the 60-h forecast analysis for Chicago (ORD) of Roebber (2015) to include minimum temperature forecasts at 36, 84, 108, 132, and 156 h. For this purpose, observed and forecast data were collected for April 2007 through August 2012. We also produce a new 60-h EP ensemble (see section 2) in order to use the same approach to ensemble generation for all the forecasts examined in this study.

The structure of this paper is as follows. In section 2, the data and the EP method are summarized [readers interested in complete details should consult Roebber (2015)]. In section 3, the results of the analysis of ensemble forecasts of minimum temperature are presented for forecast ranges extending from 36 out to 156 h. Finally, in section 4, the conclusions are provided.

2. Data and methods

Roebber (2015) describes a method for generating EP ensembles. Each algorithm is constructed as a series of up to 10 IF-THEN equations involving multiplicative and additive combinations of predictors, up to cubic powers, with each IF-THEN statement treated conceptually as a gene (hereafter EP-gene), such that the ith algorithm has the following form:

$$F_i = \sum_{j=1}^{10} \partial_{ij}, \tag{1}$$

where

$$\text{IF}(V_{1ij} O_{Rij} V_{2ij})$$
$$\text{THEN} \quad \partial_{ij} = (C_{1ij} V_{3ij}) O_{1ij} (C_{2ij} V_{4ij}) O_{2ij} (C_{3ij} V_{5ij}),$$

and V_{1ij}, V_{2ij}, V_{3ij}, V_{4ij}, and V_{5ij} can be any of the input variables; C_{1ij}, C_{2ij}, and C_{3ij} are real-valued multiplicative constants in the range ± 1; O_{Rij} is a relational operator ($\leq$, $>$); and O_{1ij} and O_{2ij} are addition or multiplication operators. The input variables and the forecast variable are normalized to (0, 1) based on the minimum and maximum of the training data.

Successive generations are produced through mating (i.e., combination) between qualified algorithms. What follows is a brief overview but interested readers should consult Roebber (2015) for full details. As shown in that paper, simulated sexual selection (preferential combination of the most successful algorithms) with disease [to limit saturation of the ''gene pool'' by a small number of individuals; see, e.g., Hillis (1990)] accelerates improved algorithm performance. There is a 50% probability of a gene experiencing a mutation, which can arise through a random change to a new variable, operator, or coefficient, or through a transposition (copy error). Although the innovation introduced by mutation is mostly counterproductive, the occurrence of occasional favorable combinations drives progress toward better solutions (Roebber 2015).

A total of 20 developmental groupings are established, in analogy with ecological niches, with each grouping allowing different combinations of inputs, disease susceptibility, and mating practices. Algorithms can migrate between niches when populations dwindle; genetic exchange between niches can also occur through partner selection, depending on the ''cultural practice'' of that niche. These rules seek to balance algorithmic independence (promoted by genetic isolation) with broad sharing of successful approaches.

Natural selection pressure is provided through a gradually tightening mean square error (MSE) threshold

during training. Thus, the most successful algorithms as defined by this moving threshold propagate their approaches through subsequent generations. A secondary selection pressure is defined based upon the correspondence between an algorithm's cumulative distribution function of forecast temperature and that of the observations for the training data. The first generation of algorithms is initialized with random combinations of variables, operators, and coefficients. These are evaluated and subsequent generations are subjected to the above evolutionary forces until convergence criteria based upon performance on the training data is satisfied.

To limit the search volume of possible solutions, the set of forecast inputs are drawn from a master set of potential predictors (Table 1) for each forecast range which have been further limited based upon consideration of colinearity (Table 2). This was accomplished simply by sequentially eliminating one of any pair of model-based variables for which the squared correlation coefficient exceeded 0.80. The multiple linear regression (MLR) for each forecast range is produced from the full potential set of predictors, through a stepwise regression procedure (Table 3). Most of the predictors are derived from NWP forecasts, based on the GFS ensemble MOS or the 0000 UTC cycle of the extended-range, deterministic GFS-based MOS (MEX) guidance. Additionally, a predictor ($[T]_{+24h}$) is formed from the GFS ensemble MOS mean temperature forecast for the same valid time but issued 24 h earlier (e.g., for a 60-h minimum temperature forecast issued at 0000 UTC 2 April and validating for the period 0000–1200 UTC 4 April, $[T]_{+24h}$ would be defined by the 21-member GFS ensemble mean MOS minimum temperature 84-h forecast issued at 0000 UTC 1 April).

Model data were supplemented with observations, including the last observed maximum temperature prior to forecast issuance (e.g., for the above forecast issued at 0000 UTC 2 April, this persistence temperature would be formed by the maximum temperature at ORD between 1200 UTC 1 April and 0000 UTC 2 April), and the average temperature at three upstream sites [Des Moines, Iowa (DSM); Minneapolis, Minnesota (MSP); and St. Louis, Missouri (STL)] for the 24 h leading up to the time of forecast issuance (e.g., for a forecast issued at 0000 UTC 2 April, the 24-h temperature would be the average temperature from 0000 UTC 1 April to 0000 UTC 2 April). An additional measure of seasonality is provided by the sine and the cosine of the Julian day. More information on this set of predictors can be found in Roebber (2015) (note, however, that here, factor analysis is not used to generate any predictors).

The input data are split into two samples: a training sample, consisting of all available dates for the period 1 April 2007–12 March 2010, and an independent test sample, consisting of the data for the period 13 March 2010–31 August 2012. Although cross validation may well be useful in this context to limit overfitting and thus provide superior performance on independent data, owing to the need to maximize the available training data, it was not used in this study.

TABLE 1. List of predictors considered in the analysis. Cues based on observations are shown in bold. Brackets indicate that the cue is derived from the GFS ensemble.

Predictor	Description
PRS_{Tmax}	Observed max temperature between 1200 and 0000 UTC for forecast issuance at 0000 UTC
PRS_{SNW}	Existence of at least 1 in. of snow on the ground at forecast issuance
$[T]$	Ensemble average forecast min temperature from the 21-member GFS ensemble MOS
$[CC]_{min}$	Min forecast cloud cover from the 21-member GFS ensemble MOS (0 = clear, 1 = partly cloudy, 2 = overcast)
$[CC]_{max}$	Max forecast cloud cover from the 21-member GFS ensemble MOS (0 = clear, 1 = partly cloudy, 2 = overcast)
$[CC]$	Ensemble average forecast cloud cover from the 21-member GFS ensemble MOS (0 = clear, 1 = partly cloudy, 2 = overcast)
$[PP]_{min}$	Min forecast precipitation probability from the 21-member GFS ensemble MOS
$[PP]_{max}$	Max forecast precipitation probability from the 21-member GFS ensemble MOS
$[PP]$	Ensemble average forecast precipitation probability from the 21-member GFS ensemble MOS
$[T]_{+24h}$	Ensemble average forecast minimum temperature 24 h after verification time from the 21-member GFS ensemble MOS
Sin(JD)	Sine of the Julian day
Cos(JD)	Cosine of the Julian day
WS	Forecast wind speed from the GFS-based MOS (MEX)
WS_{GRB}	Forecast wind speed from the GFS-based MOS (MEX) for Green Bay, WI
CC_{GRB}	Forecast cloud cover from the GFS-based MOS (MEX) for Green Bay, WI (0 = clear, 1 = partly cloudy, 2 = overcast)
PP_{GRB}	Forecast precipitation probability from the GFS-based MOS (MEX) for Green Bay, WI
WS_{DVN}	Forecast wind speed from the GFS-based MOS (MEX) for Davenport, IA
CC_{DVN}	Forecast cloud cover from the GFS-based MOS (MEX) for Davenport IA (0 = clear, 1 = partly cloudy, 2 = overcast)
PP_{DVN}	Forecast precipitation probability from the GFS-based MOS (MEX) for Davenport, IA
DSM	**24-h average temperature at Des Moines, IA**
MSP	**24-h average temperature at Minneapolis, MN**
STL	**24-h average temperature at Saint Louis, MO**
MLR	Multiple linear regression forecast from model data

TABLE 2. List of potential (all variables) and actual predictors (those marked with an ×) used in the EP training for each forecast range. Note that for each forecast range, the MLR involves different combinations of predictors, which are arrived at through stepwise multiple linear regression (see Table 3). Cues based on observations are shown in bold. Brackets indicate that the cue is derived from the GFS ensemble. Also shown in parentheses following each forecast range heading is the size of each EP ensemble after training.

	Forecast range					
Predictor	36 h (1666)	60 h (2713)	84 h (1860)	108 h (2163)	132 h (1680)	156 h (4749)
PRS_{Tmax}					×	×
PRS_{SNW}	×	×	×	×	×	×
$[T]$			×	×		
$[CC]_{min}$	×	×				×
$[CC]_{max}$					×	×
$[CC]$				×		×
$[PP]_{min}$		×			×	×
$[PP]_{max}$		×				×
$[PP]$			×		×	
$[T]_{+24h}$					×	
Sin(JD)	×	×	×	×	×	×
Cos(JD)	×	×	×	×	×	×
WS	×	×	×	×	×	×
WS_{GRB}			×	×	×	×
CC_{GRB}	×	×	×	×	×	×
PP_{GRB}	×	×	×	×	×	×
WS_{DVN}	×	×	×	×	×	×
CC_{DVN}					×	×
PP_{DVN}	×	×	×	×	×	×
DSM	×	×	×	×	×	×
MSP	×	×	×	×	×	×
STL	×	×	×	×	×	×
MLR	×	×	×	×	×	×

As in Roebber (2015), a simple bias correction and calibration procedure is applied to the ensemble forecasts. In particular, the ensemble forecast distributions are approximated as Gaussian (normal), with the mean fixed to the ensemble mean and the variance calibrated according to

$$\sigma^2 = I \times S^2, \tag{2}$$

where S^2 is the ensemble variance for a given forecast and I is an inflation factor, set to a fixed value for all cases. The inflation factor is fit over the entire training sample by requiring that 90% of the observations are contained within the 5th and 95th percentile temperatures implied by the corresponding normal distribution. To remove any bias from the ensemble mean, we also use a weighted correction following Cui et al. (2012):

$$B_{new} = (1 - w)B_{past} + wB_{current}, \tag{3}$$

where $B_{current}$ is the forecast error from the last day, B_{past} is the past accumulated bias, and B_{new} is the updated accumulated bias (which will become B_{past} for the next forecast). The next forecast is then adjusted by subtracting B_{new} from the raw forecast. Here, we set w = 0.15. More sophisticated calibration is possible and that provided by BMC, following up on Roebber (2015), is also explored in the discussion (section 3).

As in Fig. 1, verification will be conducted using a deterministic measure (RMSE) and a probabilistic measure (BSS). In the latter case, the score is computed based on standard deviations from the monthly climatic normal. First, each forecast F and matching observation O is standardized by

$$V' = \frac{V - \overline{C}}{\sigma_C}, \tag{4}$$

where V is the forecast or observed value, $\overline{C}$ is the climatological mean value for the month in which the observation occurs, and σ_C is the monthly standard deviation of the once-daily observation. A total of 33 bins of width 0.25 standard deviations are specified, covering ±4 standard deviations (note that values greater than 4σ or less than -4σ are placed in bins 33 and 1, respectively). The BSS is then computed as

$$\mathrm{BSS} = 1 - \frac{\sum_N \sum_j (pO_{Nj} - pF_{Nj})^2}{\sum_N \sum_j (pO_{Nj} - pC_j)^2}, \tag{5}$$

where the summation is over the j = 33 bins and the N cases; and pO_{Nj}, pF_{Nj}, and pC_j are the observed probability of bin j for case N, the forecast probability of bin

TABLE 3. List of predictors used in the stepwise MLR for each forecast range. Cues based on observations are shown in bold. Brackets indicate that the cue is derived from the GFS MOS ensemble. An evaluation of the relative importance of predictors is provided by the relative weights analysis (see text and Fig. 2).

	Forecast range					
Predictor	36 h	60 h	84 h	108 h	132 h	156 h
PRS_{Tmin}		×				
PRS_{PP}	×	×				
$[T]_{min}$	×	×	×			
$[T]$				×		×
$[TD]_{max}$				×	×	
[CC]		×				
$[PP]_{min}$		×				
$[PP]_{max}$						×
[PP]					×	
$[T]_{+24h}$				×	×	×
Sin(JD)	×					
Cos(JD)		×	×			
WS	×					×
T_{GRB}			×	×		×
TD_{GRB}					×	
WS_{GRB}					×	
PP_{GRB}			×			
T_{DVN}	×					
TD_{DVN}		×	×			
WS_{DVN}	×			×		
PP_{DVN}	×	×	×	×		

TABLE 4. GFS MOS and EP ensemble deterministic (RMSE) and probabilistic [Brier skill score (BSS)] performance on the independent test data. GFS and EP ensemble probabilities are calibrated based on an inflated variance procedure (see text for details).

	GFS MOS		EP	
Range (h)	RMSE (°F)	BSS (%)	RMSE (°F)	BSS (%)
36	3.74	4.9	3.56	10.3
60	4.27	6.2	3.98	9.0
84	4.72	−10.7	4.31	7.3
108	5.53	2.1	4.88	5.7
132	5.86	1.5	4.33	3.3
156	6.58	−0.2	5.43	2.5

j for case N, and the climatological probability of bin j, respectively. Note that the sum over the j bins equals unity and that pO_{Nj} is 0 except in the one verifying bin, where it is unity.

3. Results and discussion

As noted in section 1, probabilistic and deterministic forecast skill need not move in concert. At all time ranges examined, however, the EP ensemble is superior to the GFS MOS ensemble in both the RMSE and the BSS (Table 4). Indeed, the gap in RMSE increases as forecast range increases from 36 h (0.18°F better) to 132/156 h (1.53°/1.15°F better), while the EP BSS remains positive through 156 h.

MLR using appropriate inputs can also improve forecasts beyond that obtained by the GFS ensemble MOS. For the test data, the MLR produces lower RMSE than that obtained from the GFS ensemble MOS but still 3%–7% higher than the EP ensemble mean. Thus, it is of interest to understand the weighting of forecast inputs for the MLR and the EP ensemble mean.

Using the method of relative weights (Cooksey 1996; Roebber 1998), the amount of variance in the forecast that can be uniquely attributed to a particular input is given by the squared semipartial correlation (sr^2), 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 with all forecast inputs save the one of interest. Under the assumption that there is no particular order for the inputs, and thus no clear partitioning of the shared variance, the overlapping variance among the inputs is removed. As a result, the sum of the sr^2 for the subset of inputs will always be less than the R^2 for the regression using all the inputs. The sr^2 of the ith input is then converted to a relative weight for a set of M inputs as

$$rw_i = 100 \frac{sr_i^2}{\sum_{i=1}^{M} sr_i^2} \tag{6}$$

with the M weights summing to 100 units.

For ease of interpretation, we have summed the rw_i into six representative categories: observations (persistence variables and the three upstream observations); NWP temperature (GFS MOS ensemble temperatures and dewpoints, $[T]_{+24h}$, and MEX temperatures and dewpoints at GRB and DVN), NWP cloud (GFS MOS ensemble cloud cover and MEX cloud cover at GRB and DVN), NWP precipitation (GFS MOS ensemble precipitation probabilities, and MEX precipitation probabilities at GRB and DVN), NWP wind (MEX wind speeds for ORD, GRB, and DVN), and seasonal (sine and cosine of the Julian day).

The results of this analysis (Fig. 2) demonstrate that the MLR forecast relies primarily on GFS MOS temperature data whereas the EP ensemble mean relies on a variety of forecast information. Interestingly, for the EP at shorter forecast ranges (e.g., 36–60 h), additional weight is placed on observations and forecasts of wind and precipitation, while at longer range, the primary weight is placed upon seasonality with some modification from a variety of forecasts and observations. This modification is important, however, since without it, the

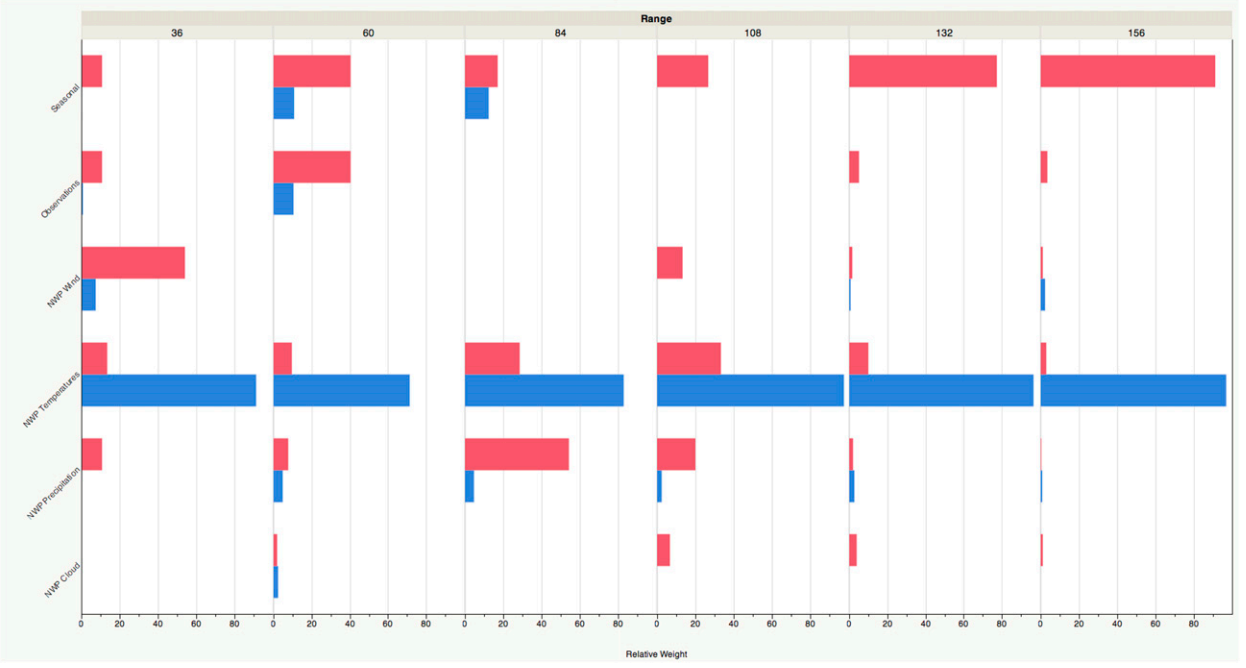

FIG. 2. Relative weights as a function of forecast range for the multiple linear regression model (blue) and the EP ensemble mean (red). Forecast ranges include 36, 60, 84, 108, 132, and 156 h.

EP ensemble would not be able to demonstrate skill with respect to climatology at those time ranges. This diversity in the cues (i.e., regressors), which drive the ensemble mean as well as in the individual ensemble members themselves, appears to be at the core of the method's ability to add skill relative to the GFS ensemble and regression.

Figure 3 provides an example for a specific forecast. In this March case, the observed minimum temperature was almost three standard deviations above the climatological norm. The GFS ensemble MOS pdf is shifted correctly toward the warm extreme, and is also sharp. Unfortunately, the shift is not far enough and the sharpness of the pdf essentially represents a false confidence, since the forecast probability for the verified temperature is nil. In contrast, the EP ensemble is shifted farther to the warm extreme such that the pdf is nearly centered about the observed temperature. The pdf is not as sharp as that of the GFS ensemble, however, reflecting less certainty in the specific outcome, as is perhaps to be expected for an unusual case. This result is representative of the overall performance, which can be seen by comparing the BSS for the 60-h minimum temperature forecasts for only those cases in which either the forecast and/or the observation was at least two standard deviations from the climatological mean. For that set of "abnormal" events, the BSS for the EP ensemble was +10.5% compared to +6.4% for the GFS MOS ensemble, a 4.1% differential compared to 2.8% at this forecast range for all cases. This superior performance toward the tails of the distribution is consistent with Roebber (2013).

Since the EP ensemble uses the MLR (Table 2), it is reasonable to ask what relative benefit accrues to the EP given this lack of parity in inputs between the two approaches. The MLR, however, uses elements of the GFS MOS ensemble temperature (Table 3), and the resultant MLR is highly correlated with MOS ($R^2 = 0.988$), so it is likely that there would not be a substantial difference in the EP performance using either the MLR or the MOS as an input. To test this assertion, the 60-h minimum temperature EP was retrained, excluding the MLR but allowing the MOS ensemble mean temperature as an input instead. The result confirms the expectation—the RMSE over the test dataset for this latter EP was 3.99°F instead of 3.98°F.

Roebber (2015) used BMC to calibrate 60-h minimum temperature forecasts obtained from the GFS MOS ensemble and an EP ensemble. This approach was particularly effective in improving the performance of the MOS ensemble, but also resulted in some small further improvement in the EP ensemble RMSE. Here, we briefly describe the BMC procedure and the results obtained for the two ensembles.

BMC, rather than attempting to identify the "best model" as in Bayesian model averaging (BMA), instead

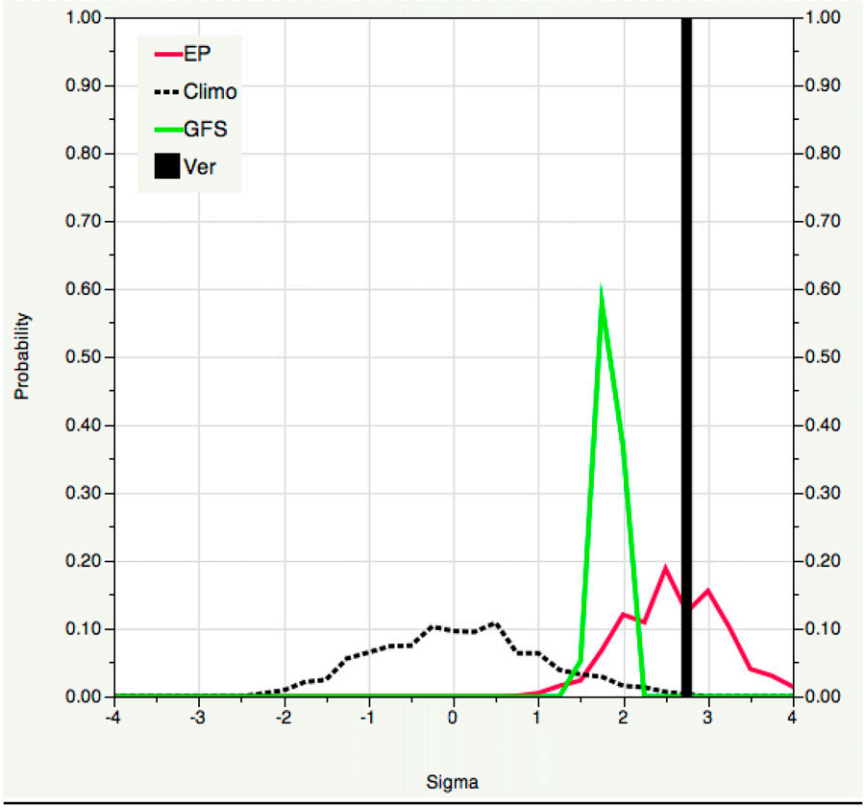

FIG. 3. Probability distributions for the 60-h forecasts for 20 Mar 2012, expressed as standard deviations from the climatological normal. Shown are the climatological distribution (dotted black), 21-member GFS MOS ensemble (green), EP ensemble (red), and verification (black bar).

seeks the ''best model combination.'' It has been shown to outperform BMA across a wide variety of datasets [see Monteith et al. (2011)]. Owing to computational considerations, however, both BMA and BMC are most suited to smaller ensembles. Accordingly, it is usual to perform a model selection preprocessing step (e.g., Hoeting et al. 1999). In the case of BMC with linear combination as used here, M^K possible combinations must be evaluated, where M is the number of raw weights and K is the number of ensemble members. Thus, for a target of $O(10^9)$ evaluations, the number of ensemble members falls from 30 to 15 as the number of raw weights increases from 2 to 4. In contrast, the smallest EP ensemble is the 36-h forecast, which consists of 1666 members.

To make the calculations tractable, we have opted to select 10 ensemble members with 4 possible weights, yielding approximately 10^6 possible combinations. We perform this selection by first ranking all ensemble members according to their MSE on the training data. Next, we compute the training data forecast variance between each possible ensemble pair, and in the case for which this variance is less than a critical threshold (set to $0.09°F^2$), the lower ranking member of the pair is eliminated from consideration. The 10 highest-ranking ensemble members by MSE become the model subset used in BMC. Each of these 10 ensemble members are bias corrected using (3), and then a weighted forecast is produced for each possible combination of the four weights (note that these weights are normalized to sum to unity). The posterior probability for the model combination e given the training data D is

$$p(e \mid D) \propto \frac{1}{4^{10}}(1 - \in)^r \in^{n-r}, \tag{7}$$

where $\in$ 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. Here, we consider a model combination to be correct for a given training case if the absolute error is less than or equal to 5°. After all combinations have been considered, the selected model is the one that maximizes the logarithm of (7).

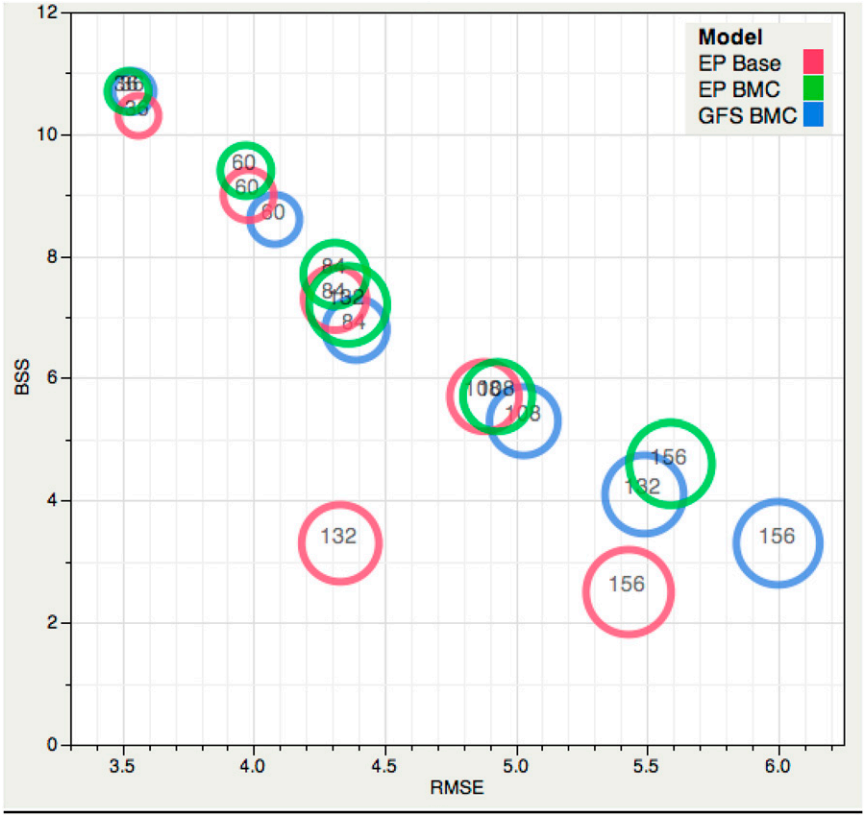

FIG. 4. Brier skill score vs root-mean-square error for minimum temperature forecasts at Chicago, IL, from the Bayesian model combination calibrated GFS MOS ensemble (blue) and the EP ensemble (green). For comparison, the EP ensemble using the simple calibration but retaining all members is also shown (red). The forecast ranges include 36, 60, 84, 108, 132, and 156 h and are indicated by the size of the circle and the label.

As in Roebber (2015), each of the 10 members of the weighted model are assumed to be normally distributed with mean according to the bias-corrected forecast and variance estimated as

$$o^2 = \frac{\sum_N \sum_k w_k (F_k - O)^2}{N}, \tag{8}$$

where w_k is the normalized weight of the ensemble member, F_k is the bias-corrected forecast of the ensemble member, O is the observed value, and the summation is over the 10 ensemble members and the N training cases (resulting in the possibility of multimodal or skewed probability forecasts).

Figure 4 shows the BMC for the GFS MOS and EP ensembles as compared to the base EP ensemble. Most apparent is the inevitable slide in skill toward climatology as forecast range increases (i.e., the shift from the top-left toward the bottom right of the figure). With respect to BMC calibration, the most notable result is that there is little improvement in the EP RMSE and in fact this measure increases slightly by 108 h relative to the base ensemble. Despite this, the BMC EP ensemble produces lower RMSE at all forecast ranges than that obtained by the BMC GFS MOS. The BMC calibration also improves probabilistic forecast skill for the EP ensemble, for all forecast ranges considered except 108 h, where the BSS remains at 5.7%. Thus, separation between ensembles in the BSS-RMSE space occurs at the longest forecast ranges, beginning at 132 h. Future research is needed to understand the origins of this result.

A question that naturally arises is how sensitive these results are to the model selection procedure? Certainly, our procedure is somewhat ad hoc. To evaluate this impact, we consider instead a Bayesian information criterion (BIC), which combines the error variance already used with a measure of model complexity (based on the number of EP parameters to be estimated). For model selection, we simply rank all individual ensemble members according to increasing BIC, and choose the 10 lowest such ensembles (i.e., the preferred model is that with the lowest BIC). This procedure, when applied to

the 108-h forecasts, resulted in a selection of ensemble members that was largely distinct from the set previously produced. Despite this, the BMC EP ensemble performance in RMSE and BSS was identical to that previously obtained with the base model selection procedure.

An intrinsic advantage of EP is the ability to generate large ensembles. This advantage is lost when applying the BMC methodology owing to the need for model selection. Thus, for the 108-h forecast, one instance in which the BMC approach reduces performance compared to the base EP ensemble, what would be the impact of increasing the ensemble member subset from 10 members to 21? This 21-member BMC was evaluated by cycling through two raw weight values for the 108-h forecast. Here, again, the ensemble performance in both RMSE and BSS was the same as that previously obtained with only 10 ensemble members. Owing to computational limitations, we cannot say what number of included members would be needed for performance to approach the levels obtained with the 2163-member base EP ensemble.

Finally, Roebber (2015) showed that a new BMC ensemble created by pooling the EP forecasts and the GFS forecasts (along with a multiple linear regression model) prior to model selection produced a further improvement in RMSE. This approach was repeated here, for all forecast ranges. Such a ''hybrid'' BMC, consisting of both EP and GFS ensemble members for 84-h forecasts, provided similar probabilistic forecast skill to the standard EP BMC, but reduced RMSE an additional 0.12°–0.53°F below the raw GFS MOS ensemble. However, for the other forecast ranges, the model selection procedure resulted only in sets of EP ensemble members and thus was unable to take advantage of this potential additional ''information.''

4. Summary

Here, we have examined changes to deterministic and probabilistic forecast skill of the GFS ensemble MOS and evolutionary program ensembles for minimum temperature at Chicago, Illinois, for 36-, 60-, 84-, 108-, 132-, and 156-h forecast ranges. Likewise, we have considered how the reliance on particular forecast inputs change as forecast range changes.

We find that for all forecast ranges examined, the evolutionary program is superior to the operational GFS MOS ensemble for both RMSE and BSS. Further, the gap in RMSE increases with forecast range, from 0.18°F at 36 h to 1.53°F at 156 h. As with Roebber (2013, 2015), we find that the evolutionary program ensemble probabilistic forecasts are particularly effective compared to raw operational guidance when forecasting the likelihood of abnormal conditions for forecast ranges out to 156 h.

The weighting of the forecast inputs for the evolutionary program ensemble mean is distinct from that obtained for a multiple linear regression, with the primary difference being relatively less reliance on the GFS MOS temperature and relatively more weight being placed on alternative information (including observations such as the temperatures at upstream sites at the time of forecast issuance, time of year, and forecasts of wind speed, precipitation, and cloud cover). This weighting of evolutionary program inputs trends away from current observations and toward seasonal measures as forecast range increases.

Bayesian model combination requires that ensemble size be sufficiently small to render calibration feasible. Thus, evolutionary program ensemble size was reduced from $O(1000)$ to 10 members by eliminating ''similar'' algorithms (defined by their paired forecast variance) and selecting those remaining members with the lowest mean square errors. This calibration, when applied to both the evolutionary program and operational model ensembles, was able to improve the latter's performance and thus reduce but not eliminate the skill gap between the two ensembles. Notably, the calibration provided little to no improvement in evolutionary program root-mean-square error but did improve probabilistic forecast skill. These results were not sensitive to a second model selection procedure, or to extending the ensemble member subset from 10 to 21 members. The largest skill differentials occurred at the longest forecast ranges, beginning at 132 h. Further research is needed to understand the growth of the skill gap as forecast range increases. The ''hybrid'' procedure tested in Roebber (2015), in which both evolutionary program and operational ensemble members provide the sample from which the subset of members is selected for calibration, further increased deterministic forecast skill for the one forecast range in which both types were selected. While the overall skill differences between the calibrated ensembles are too small to make such efforts worthwhile in the context of public forecasts, they can provide value in particular applications such as in support of electricity generation. Since the skill differential increases as forecast range increases, it is of interest to determine whether this behavior extends to shorter-range, lower skill forecasts such as heavy precipitation.

Acknowledgments. This work was supported by UWM Research Foundation Catalyst Grant.

REFERENCES

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

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.,* **78,** 1–3, doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Buizza, R., 1997: Potential forecast skill of prediction and spread and skill distributions of the ECMWF prediction system. *Mon. Wea. Rev.,* **125,** 99–119, doi:10.1175/1520-0493(1997)125<0099:PFSOEP>2.0.CO;2.

——, M. Milleer, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.,* **125,** 2887–2908, doi:10.1002/qj.49712556006.

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

Coulibaly, P., 2004: Downscaling daily extreme temperatures with genetic programming. *Geophys. Res. Lett.,* **31,** L16203, doi:10.1029/2004GL020075.

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

Fogel, L. J., 1999: *Intelligence through Simulated Evolution: Forty Years of Evolutionary Programming.* John Wiley, 162 pp.

Hamill, T. M., and J. S. Whitaker, 2007: Ensemble calibration of 500-hPa geopotential height and 850-hPa and 2-m temperatures using reforecasts. *Mon. Wea. Rev.,* **135,** 3273–3280, doi:10.1175/MWR3468.1.

Harrison, M. S. J., T. N. Palmer, D. S. Richardson, and R. Buizza, 1999: Analysis and model dependencies in medium-range: Two transplant case studies. *Quart. J. Roy. Meteor. Soc.,* **125,** 2487–2516, doi:10.1002/qj.49712555908.

Haupt, S. E., G. S. Young, and C. T. Allen, 2006: Validation of a receptor–dispersion model coupled with a genetic algorithm using synthetic data. *J. Appl. Meteor. Climatol.,* **45,** 476–490, doi:10.1175/JAM2359.1.

Hillis, W. D., 1990: Co-evolving parasites improve simulated evolution as an optimization procedure. *Physica D,* **42,** 228–234, doi:10.1016/0167-2789(90)90076-2.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky, 1999: Bayesian model averaging: A tutorial (with comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors). *Stat. Sci.,* **14,** 382–417, doi:10.1214/ss/1009212519.

Lakshmanan, V., 2000: Using a genetic algorithm to tune a bounded weak echo region detection algorithm. *J. Appl. Meteor.,* **39,** 222–230, doi:10.1175/1520-0450(2000)039<0222:UAGATT>2.0.CO;2.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.,* **20,** 130–141, doi:10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

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, doi:10.1109/IJCNN.2011.6033566.

Novak, D. R., D. R. Bright, and M. J. Brennan, 2008: Operational forecaster uncertainty needs and future roles. *Wea. Forecasting,* **23,** 1069–1084, doi:10.1175/2008WAF2222142.1.

O'Steen, L., and D. Werth, 2009: The application of an evolutionary algorithm to the optimization of a mesoscale meteorological model. *J. Appl. Meteor. Climatol.,* **48,** 317–329, doi:10.1175/2008JAMC1967.1.

Orrell, D., 2005: Ensemble forecasting in a system with model error. *J. Atmos. Sci.,* **62,** 1652–1659, doi:10.1175/JAS3406.1.

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

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

——, 2013: Using evolutionary programming to generate skillful extreme value probabilistic forecasts. *Mon. Wea. Rev.,* **141,** 3170–3185, doi:10.1175/MWR-D-12-00285.1.

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

Shutts, G. J., and T. N. Palmer, 2007: Convective forcing fluctuations in a cloud-resolving model: Relevance to the stochastic parameterization problem. *J. Climate,* **20,** 187–202, doi:10.1175/JCLI3954.1.

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, doi:10.1175/1520-0493(2000)128<2077:UICAMP>2.0.CO;2.

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, doi:10.1175/BAMS-86-12-1765.

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, doi:10.1109/59.486125.