Seeking Consensus: A New Approach

PAUL J. ROEBBER

Atmospheric Science Group, University of Wisconsin—Milwaukee, Milwaukee, Wisconsin

(Manuscript received 20 May 2010, in final form 25 July 2010)

ABSTRACT

Simulated evolution is used to generate consensus forecasts of next-day minimum temperature for a site in Ohio. The evolved forecast algorithm logic is interpretable in terms of physics that might be accounted for by experienced forecasters, but the logic of the individual algorithms that form the consensus is unique. As a result, evolved program consensus forecasts produce substantial increases in forecast accuracy relative to forecast benchmarks such as model output statistics (MOS) and those from the National Weather Service (NWS). The best consensus produces a mean absolute error (MAE) of 2.98°F on an independent test dataset, representing a 27% improvement relative to MOS. These results translate to potential annual cost savings for electricity production in the state of Ohio of the order of $2 million relative to the NWS forecasts. Perfect forecasts provide nearly $6 million in additional annual electricity production cost savings relative to the evolved program consensus.

The frequency of outlier events (forecast busts) falls from 24% using NWS to 16% using the evolved program consensus. Information on when busts are most likely can be provided through a logistic regression equation with two variables: forecast wind speed and the deviation of the NWS minimum temperature forecast from persistence. A forecast of a bust is 4 times more likely to be correct than wrong, suggesting some utility in anticipating the most egregious forecast errors. Discussion concerning the probabilistic applications of evolved programs, the application of this technique to other forecast problems, and the relevance of these findings to the future role of human forecasting is provided.

1. Introduction

Margaret Thatcher, Prime Minister of the United Kingdom during 1979–90, famously stated in a 1981 speech: ''To me, consensus seems to be the process of abandoning all beliefs, principles, values and policies. So it is something in which no one believes and to which no one objects.'' Thatcher's comment emphasizes the primacy of informed, individual action, and certainly a highly skilled choice is demonstrably preferable to group think. In the latter case, there is a rush to conformity and the independence upon which a useful consensus depends is lost.

Despite Thatcher's assertion, a recent (July 2010) search of the American Meteorological Society's Web site reveals 1432 papers containing the words consensus in the text. Why this interest in consensus by the meteorological community? In the context of weather forecasting, Thompson (1977) provided a mathematical proof that an optimal linear combination of independent, imperfect forecasts is superior to that of individual predictions. The averaging process effectively acts as a nonlinear filter to focus on the component of the forecast that can be considered the most reliable (Toth et al. 1998). This theoretical principle has been shown to hold in real-world forecasting contexts (Sanders 1973, 1979; Bosart 1975; Gyakum 1986) where student-based consensus forecasts are competitive with, and sometimes superior to, those of highly experienced individuals.

Vislocky and Fritsch (1995, 1997) demonstrate that a multiple linear regression/model output statistics (MOS) consensus is likewise competitive with the best human forecasters. Given that consensus provides this standard of forecast excellence, it was used as a benchmark to intercompare scores from different forecast periods and varying sites in the National Collegiate Weather Forecasting Contest (e.g., Roebber et al. 1996). As well as the simple average or some other optimized combination (e.g., Krishnamurti et al. 1999; Vukicevic et al. 2008), another advantage of an array of skilled forecasts is that

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

DOI: 10.1175/2010MWR3508.1

it can be used to generate probabilities, which in some contexts can provide greater utility and/or value than a deterministic forecast (e.g., Richardson 2000; Zhu et al. 2002). In meteorology, where even the best forecast may not be readily apparent beforehand, probabilistic forecasts have come to be understood as a preferred means of managing this inherent uncertainty.

Despite these advantages, consensus forecasts (and their conceptual relation, ensemble averages) can fail, sometimes spectacularly (e.g., Toth et al. 1997; Bosart 2003). Since numerical weather prediction (NWP) is at the core of modern weather forecasting, issues related to the interplay between analysis error, model error, and their nonlinear growth are key. Ensemble forecast systems can be designed to manage these uncertainties by varying analyses [see Kalnay (2003) for a review of the extensive literature] and model physics/numerics (e.g., Brooks et al. 1995; Hamill and Colucci 1997; Tracton et al. 1998; Krishnamurti et al. 1999; Stensrud et al. 1999; Fritsch et al. 2000; Weigel et al. 2008; and others), and through calibration of resulting probability density functions (e.g., Hamill and Whitaker 2007). These improvements, however, do not completely resolve problems with reliability (the degree to which a set of forecasts of an event of certain probability are observed at a similar frequency) and resolution (the ability to distinguish nonevents from events). Zhang et al. (2002) show how key observations can substantially alter a model forecast of an important event and Bosart (2003) shows how in that instance, the operational model consensus can lead forecasters astray, despite the existence of contrary observations.

The generic problem of extracting useful information from imperfect forecast data can be understood qualitatively using a simple argument. Suppose that we have an observable quantity F that is of forecast interest and is itself a function of a state defined by two variables X and Y. To forecast F, we need to know what X and Y will be, but our estimates of these quantities contain some error (in actuality, our measurements of F will also contain error, but this is neglected here for simplicity). The forecast error for F will be a function of the unknown true state, and how much the error (randomly) scales according to that true state. Consequently, there will be a higher probability of larger errors for some states than others. This suggests that if we can identify these controlling states, and if we can test the sensitivity of those forecasts to errors in our estimate of these states, we can at least assign some level of confidence to the forecasts. Ensemble modeling methods are an attempt to do this. But there may well be other approaches that might be used to provide this information.

The focus of this paper is on the production of a skillful and valuable consensus forecast for surface temperature, using a different approach—a method known generally as evolutionary programming [see Bakhshaii and Stull (2009) and section 2b below for more details]. The focus is on generating deterministic rather than probabilistic forecasts because in many situations, forecast users must make a binary (yes–no) decision. We will develop a measure of forecast confidence, however, so that where an ability to manage risk exists, this information can be used in addition to the consensus forecast. We will show that the gains in skill provided by the forecast consensus derived from evolutionary programming translate into a considerable additional value relative to several forecast benchmarks, in the context of a simplified form of day-ahead electricity demand forecasting.

The paper is organized as follows. Section 2 provides an explanation of the forecast problem and construction of the dataset for model development and testing (section 2a), a general description of evolutionary programming and the specific method used here (section 2b), and details concerning electricity demand forecasts (section 2c). Section 3 presents the results of the experiments, while conclusions are presented in section 4.

2. Data and methods

a. Observed and forecast data

Hourly surface observations from the National Climate Data Center (NCDC) Global Surface Hourly database were collected for Columbus, Ohio (CMH), for the period 1995–2002. Matching upper-air data were obtained from the closest sounding site, located at Wilmington, Ohio (ILM), and were retrieved from the NCDC Integrated Global Radiosonde Archive (IGRA). Snowcover data were obtained from the NCDC Climate Visualization (CLIMVIS) database. MOS data were obtained from the National Oceanic and Atmospheric Administration (NOAA) Meteorological Development Laboratory (MDL) archive. Since nested grid model (NGM) MOS is the only guidance source available for the entire period of record, these data are used. The technique, described below, is not dependent on a specific form of guidance, however, and it is to be expected that improvements in MOS would lead to further improvements from the technique as well. Note, however, that this is not guaranteed, since the closer forecasts approach inherent limits to predictability the less opportunity exists for postprocessing methods to improve them. After quality control and accounting for missing data (only full data for all dates are considered), a total of 852 days were available for analysis.

TABLE 1. Potential predictors obtained or derived from basic data (see text for details).

Variable name	Interpretation	Variable name	Interpretation
PERS	Persistence	ΔTD12	Change in TD12
T00	MOS 0000 UTC temperature	Δ*V*	Change in MOS wind
TD00	MOS 0000 UTC dewpoint	ΔUshr	Change in Ushr
TD12	MOS 1200 UTC dewpoint	ΔVshr	Change in Vshr
T850	850-hPa temperature	ΔUad	Change in Uad
V00	MOS 0000 UTC wind speed	ΔVad	Change in Vad
V06	MOS 0600 UTC wind speed	ΔSNOW	Change in SNOW
V12	MOS 1200 UTC wind speed	ΔCLOUD	Change in CLOUD
Ushr	Zonal wind shear	ΔSOUTH	Change in SOUTH
Vshr	Meridional wind shear	ΔNORTH	Change in NORTH
Uad	Zonal temperature advection	ΔT850	Change in T850
Vad	Meridional temperature advection	F1	Wind speed factor
SNOW	Snow cover (yes–no)	F2	Temperature, dewpoint factor
CLOUD	Cloud cover (yes–no)	F3	Temperature, dewpoint trend factor
SOUTH	Southerly wind (yes–no)	F4	Meridional temperature advection factor
NORTH	Northerly wind (yes–no)	F5	Cloud-cover factor
ΔMOS	Change in MOS temperature	F6	Zonal temperature advection factor
ΔT00	Change in T00	F7	Snow-cover factor
ΔTD00	Change in TD00	F8	Southerly wind direction factor

These data are roughly evenly distributed through the four seasons: 24% [December–February (DJF)], 36% [March–May (MAM)], 23% [June–August (JJA)], and 17% [September–November (SON)], respectively. Initial interest in the area of central and southwest Ohio was spurred by the prevalence of large minimum temperature forecast errors in forecasts issued by the National Weather Service (NWS). Differences between the observed and NWS 24-h forecast minimum temperature of at least 6°F (hereafter, a ''forecast bust'') occurred on 17.6% of the days. Examination of these forecast days shows a mix of synoptic situations, with advective conditions being the most frequent [67%; as defined by approximate 0000 UTC 850-hPa temperature advection of either sign of at least 0.5°C $(6\ h)^{-1}$], followed by radiative (15%; as defined by 24-h MOS 1200 UTC winds of less than or equal to 3 kt). Frontal and miscellaneous other types account for the remainder.

Based on forecast experience and data availability, the following surface variables and/or their combinations were initially selected for consideration as inputs to model development (Table 1): persistence (the next day's minimum temperature will be the same as today's); the MOS minimum temperature forecast; the previous day's MOS minimum temperature forecast validating at the same time as the current forecast (i.e., the day-2 MOS forecast from the previous day); the MOS surface temperature at 0000 UTC; the MOS surface wind speeds and directions at 0000, 0600, and 1200 UTC; the MOS surface dewpoint temperatures at 0000, 0600, and 1200 UTC; the average MOS overnight cloud coverage (SCT or less versus BKN or more); and snow on the ground (greater than or equal to 1 in.).

Some observed upper-air variables at 0000 UTC were also examined given their availability relative to similar measures from a model: the 850-hPa wind speed and direction, the 850-hPa temperature, the east and north components of surface-to-850-hPa wind shear, the east and north components of temperature advection [based on the vector product of the surface-to-850-hPa wind shear (an approximate measure of the horizontal temperature gradient) and the 850-hPa wind speed]. These are considered to be perfect prognostic measures [i.e., the observations are assumed to be equivalent to forecasts for the purposes of model development; see Klein et al. (1959)], since forecasts from 1200 UTC should not deviate substantially at this short range. Snow cover is also treated as a perfect prognostic measure, since it is defined as a binary variable, and in many instances the snow cover will be known ahead of time. In some instances, where rapid melting might occur, or where new snow will be falling on bare or nearly bare ground, an actual forecast of whether or not the snow the next morning will exceed one inch would be needed. This, in most cases, will not be a demanding forecast and imperfections in such a forecast are not considered in the assessment that follows.

It is desirable to reduce dimensionality, a topic that is addressed in more detail in section 2b. Usually this is accomplished through domain knowledge (here, physical understanding of the minimum temperature forecast issue). One method for doing so is to limit the number of independent variables (inputs) though factor analysis.

The objective of factor analysis (Gorsuch 1983) is to reduce a set of correlated variables to a smaller number of factors, which are linear combinations of the correlated variables, and are associated with a particular process or input. In this study, principal component analysis (Hotelling 1933) is used to extract a set of eight factors (Table 1) from the original set of inputs. The extracted factors are constrained to be orthogonal (i.e., the factors are not correlated). Because some information is lost, there is a cost associated with this data extraction. In the course of the analysis, it is necessary to determine whether the loss of information is compensated by the reduction of dimensionality (see section 2c). Thus, for the analyses, both the raw data and the extracted factors are provided as inputs. The data are randomly split into three sections: model development (512 cases), model cross validation (170 cases), and independent test dataset (170 cases) for model evaluation. Although this is a standard procedure for statistical model development, because of temporal correlation in meteorological data, it is possible that this selection process reduces independence between the development data and the test data. Since only 29% of all possible days from 1995 to 2002 are used in this dataset, the likelihood of such correlation is substantially reduced. Nonetheless, this possibility will be considered in section 3.

The 12-h reforecast (Hamill et al. 2006) 2-m temperature ensemble mean forecasts valid at 1200 UTC are used to provide an ensemble benchmark for forecast comparison. The reforecast dataset is produced from a 1998 version of the National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS). The 15-member ensembles are constructed from runs initialized at 0000 UTC and the forecast data are archived on a 2.5° grid. In this study, the values at 40°N, 82.5°W are used to represent CMH. Bias correction is accomplished through regression between the observed and reforecast temperatures for the 512 member development dataset.

b. Evolutionary programming

Bakhshaii and Stull (2009) provide a review of evolutionary programming and apply it to the specific problem of developing a deterministic ensemble forecast of precipitation in complex terrain using the individual members. Additional information on evolutionary programming can be found in Ferreira (2006). The conceptual basis is as follows. Suppose that we have a well-defined problem with a clear measure of success. Suppose that we can construct solutions to the problem by performing various mathematical operations on a set of inputs. For example, a multiple linear regression equation is constructed from the sum of a series of coefficients multiplied by input variables, where the values of the coefficients and the set of variables to be used have been determined through some optimization process (e.g., least squares, stepwise regression, etc.).

In this case, it is possible to develop a computer program that generates algorithms that solve the defined problem by applying various operators and coefficients to the inputs. The level of success or ''fitness'' of the particular solution is then measured. The idea of fitness invokes evolutionary principles and suggests that if one starts from a very large set of random initial algorithms and allows ''fit'' algorithms to propagate some portion of their components to the next generation, then it may be possible to produce improved algorithms over time. This culling of the population in favor of stronger individuals through fitness pressure (i.e., maximizing fitness) and the exchange of ''genetic material'' between fit algorithms is at the heart of the process. The genetic material of the algorithm consists of the particular inputs to be used, the specific operators that are applied, and the value of the coefficients. To insure genetic diversity, a small percentage of poorly performing algorithms are also allowed to propagate their information to the next generation. Hereafter, we shall refer to all these algorithms as evolved programs (EPs).

An early application of this idea to the sorting of a random sequence of numbers showed that it was possible to develop unique solutions that approached the efficiency of algorithms devised by experienced programmers (e.g., Hillis 1990). Since the weather forecast problem is nonlinear and there is no unique solution, multiple approaches are possible, and evolutionary programming is a potential means for generating a set of skillful, yet independent, solutions. In this context, it is similar to each member of a group of experienced forecasters, who when exposed to the same information, develops a set of independent forecasts based on his or her individual reasoning, training, and experience [see section 5a of Roebber et al. (2002) and references therein for a discussion of scientific forecasting]. As discussed previously, the independent information contained in these forecasts allows for a more skillful consensus to be constructed. Here, we use evolutionary programming to construct the independent forecasts and then combine these to form a consensus forecast.

An obstacle common to complex, high-dimensional problems is the so-called curse of dimensionality. Essentially there is an exponential increase in the volume of the solution space associated with adding dimensions to the problem. Practically, this means that describing the true state of a high-dimensional problem will require very large amounts of data, otherwise large portions of the volume are never explored. Since the observational

data are inevitably finite, however, it is often better to sacrifice dimensionality to better fit a lower-dimensional solution. The success of this approach rests on the assumption that the solution function is quasi-constant for much of the space, and this sacrifice is less costly than might initially appear to be the case. Bakhshaii and Stull (2009) recognize this point in stating that ''a scientific relationship should not be more complex than needed.'' Of course, what is sufficient is not known beforehand and must be determined empirically based upon the problem at hand.

These considerations are of central concern when considering the operators to allow in EP development. In the case of multiple linear regression, for example, the operators are addition and multiplication. But many more operators are possible, including trigonometric, hyperbolic, relational, and other functions. To constrain dimensionality, and as a second but important consideration, to increase the likelihood that we may be able to develop EP whose forecast logic is understandable (noting that there is no guarantee that the logic *will be* understandable; see section 3 for the results), we restrict the operators here to addition, multiplication, and relational functions (i.e., greater than or equal to, less than or equal to).

Commercial software is available to do evolutionary programming (e.g., Ferreira 2010). There are many possible approaches to the method, however, and to limit the problem, the evolutionary programming code for this specific problem was locally developed (in FORTRAN). An algorithm structure (i.e., the EP that produces a minimum temperature forecast) consists of up to 10 lines of the following form:

$$\text{if} \quad (\text{var1}\ \mathrm{O_r}\ \text{var2}) \quad \text{then}$$
$$\partial_i = (\text{c1} \times \text{var3})\mathrm{O}(\text{c2} \times \text{var4})\mathrm{O}(\text{c3} \times \text{var5}). \qquad (1)$$

Here var1, var2, ... var5 can be any of the list of inputs (Table 1) normalized to the range from 0 to 1 by the range of the variable; c1, c2, and c3 are constants; $\mathrm{O_r}$ is a relational operator ($\geq$, $\leq$); and O is addition or multiplication. The minimum temperature is then recovered by

$$T_{\min} = \left(\sum_{i=1}^{10} \partial_i\right)(\text{Max} - \text{Min}) + \text{Min}, \qquad (2)$$

where Max and Min are the largest and smallest values of minimum temperature in the development dataset. Here $T_{\min}$ is rounded to the nearest integer (°F) to match the observed temperature data and other temperature elements (e.g., MOS, NWS, etc.).

This structure, which represents a genetic code of up to 110 elements (10 lines times 11 alterable components per line), is flexible and always produces functional algorithms. The form of the solutions can be categorized as piecewise multiple linear or nonlinear regression equations. The possible combinations are enormous, however, providing very specialized fits to the data (and also emphasizing the need to control overfitting). Two of the best solutions are shown and discussed in section 3.

Given this structure, how do the algorithms evolve toward improved solutions? The measure of fitness used, which was arrived at through trials with various criteria, is the mean absolute error (MAE). The procedure is as follows. Every algorithm is checked to determine if it satisfies the MAE fitness criterion. This fitness threshold is slowly tightened over time, so that in any training iteration it is 3% higher than the current best solution or 7.5% lower than the fitness threshold in the last iteration, whichever is the lesser change. This insures that fitness pressure continues to increase and the population continues to evolve toward improved solutions. If a given algorithm does not satisfy the required fitness, it may still be preserved for possible reproduction in order to enhance population diversity (initially, there is a 25% chance of this, which reduces monotonically with training iterations to a minimum of 5%).

Once the ''parent pool'' has been determined, the next generation is produced. Reproduction occurs in three ways: cloning (10% chance), crossover (40% chance), and mutation (50% chance). Cloning is simply taking an exact copy of a parent algorithm (in this implementation, the ''mother''; note though that this terminology is purely for exposition as any algorithm can serve as mother or father and may in fact do both in a given iteration). Crossover is a process where portions of the two parent algorithms are combined to form a new algorithm, while mutation means randomly altering one element of one algorithm.

In each iteration, the set of algorithms designated as ineligible for reproduction (the vast majority) are replaced. Parents are selected at random from the set of algorithms previously identified as eligible for reproduction (including those randomly selected from the unfit population). The current best solution (based on MAE for the training dataset) is always cloned but is also available for other forms of reproduction. In crossover mode, any of the 10 lines indicated by (1) may be exchanged at 50% probability, so that the most probable result is that five lines of the new generation algorithm will be provided by the mother and five by the father, but this proportion can vary (up to and including a nearly 0.2% chance that the next generation would be a perfect clone of either the mother or of the father). In mutation,

the next generation will be the same as that of a clone, except that one of the 11 elements of one of the 10 lines of (1) would be randomly altered.

Since parental sampling is without replacement, this means that there are multiple reproductive opportunities for fit individuals, but also no guarantee that a specific individual will pass any of its code to the next generation. The relative probabilities of cloning, crossover, and mutation cannot be theoretically specified and are arrived at in this instance through trial and error. There is a plausible argument that mutation is the primary driving force for innovative solutions, since in some sense crossover only acts to recombine existing strategies while mutation can develop new ones. For this problem, a relatively large mutation rate improved algorithm convergence. Initial population size is also an important, theoretically unspecified, but tunable parameter. Here, an initial algorithm population size of 20 000 is used. Training is stopped when a minimum number of training iterations have been conducted, a best solution has been found that satisfies a minimum MAE criterion for the training data, and the cross-validation error is no longer decreasing. For this dataset, the minimum number of training iterations was set to 100, and the minimum MAE was 3.20.

Training data are provided as bootstrap samples, that is, in a given training iteration, 512 cases are provided based on sampling from the development dataset with replacement. Bootstrap sampling and cross validation are both used to increase the generality of the EP. By repeating the process with different random initial populations, multiple independent solutions are developed. A drawback of iterative approaches of this kind is the time-consuming nature of the training process. Once the algorithms are trained, however, these algorithms are no more costly than simple regression equations.

c. *Electricity demand forecasts*

Utilities use forecasts of power demand (the so-called load forecast) to determine their daily generation needs. The load forecast is used by the power system operating center to plan the most economical mix of generation resources that will fill the expected demand. Errors in the demand forecast lead to costs to utility operations associated with the startup and shutdown of generation units and the costs of those units once in service [see Teisberg et al. (2005) for an overview]. In this paper, the temperature forecasts are placed in the context of load forecasting as a means of measuring the value of improvements. Utility data are proprietary and difficult to obtain, but there is enough information available from various sources (e.g., Valor et al. 2001; Rosenzweig and Solecki 2001; Cheng et al. 2001; Teisberg et al. 2005) to develop an approximate method for valuing the temperature forecasts in this study.

Although load depends on several variables, including day of week and season, temperature is the primary driver (e.g., Valor et al. 2001 and references therein; Teisberg et al. 2005). Load projections use hourly forecast data for the day-ahead market, so working from a 1200 UTC forecast cycle and a midnight-to-midnight period requires forecasts in the 18–42-h range. Valor et al. (2001) and others show that it is possible to construct demand curves from degree-day data with reasonable accuracy. In this study, we focus on next-day minimum temperatures and exploit the high correlation between minimum and maximum temperature ($r = 0.924$ for CMH for 1 full year of observed minimum and maximum temperature data) to estimate a daily average temperature from the minimum temperatures. Real-world demand forecasting requires accounting for additional details (including hourly frequency), but this approximate method is sufficient to give useful value estimates as a proof of concept.

Based on the data of Valor et al. (2001), Rosenzweig and Solecki (2001), and Cheng et al. (2001), the estimate of the daily electricity demand (D, in MWh) is

$$
\begin{aligned}
D(T \le 5^\circ\mathrm{F}) &= 157\,250,\\
D(5 < T \le 68) &= 112\,680 + 11.3(68 - T)\\
&\quad + 33.1(68 - T)^2 - 0.35(68 - T)^3,\\
D(87 \le T < 68) &= 112\,680 + 34.8(T - 68)\\
&\quad + 315.9(T - 68)^2 - 10.3(T - 68)^3,
\end{aligned}
$$

and

$$
D(87 < T) = 157\,250, \tag{3}
$$

where T is the daily average temperature (°F). This curve (Fig. 1) is intended to represent the demand characteristics of the state of Ohio, as represented by limited public data provided by American Electric Power (AEP) Ohio. Data for New York State (Rosenzweig and Solecki 2001) show higher demand during the summer season (from air conditioning) than winter (for heating), while data primarily from southern Ontario (Cheng et al. 2001) show comparable demand in both seasons, and others (not shown) indicate higher demand in the cold season. For simplicity, we assume that demand is similar and saturates as in Valor et al. (2001) at both seasonal extremes (Fig. 1).

Based on Hobbs et al. (1999) as reproduced in Teisberg et al. (2005), the increased production cost resulting from forecast errors is approximated by

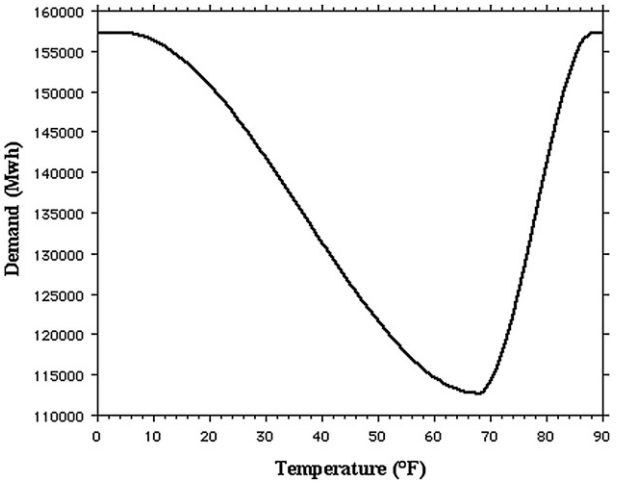

FIG. 1. Estimated daily electricity demand (MWh) as a function of daily average temperature (°F) for Ohio.

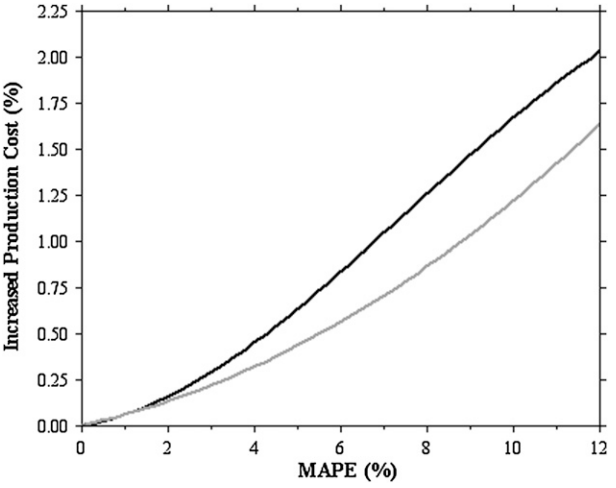

FIG. 2. Increased electricity production cost (%) as a function of mean absolute percent error in forecast demand. The black (gray) curve represents the production cost for the southern (northern) region. See text for details.

$$\mathrm{IP}_S = 0.000\,37\ \mathrm{MAPE} + 0.000\,23\ \mathrm{MAPE}^2 - 0.000\,01\ \mathrm{MAPE}^3, \quad (4)$$

where MAPE is the mean absolute percent error in the demand forecast obtained from (3). The values obtained from (4) (Fig. 2) are intentionally toward the higher end of the curves provided by Teisberg et al. (2005), similar to the south values [note that the values from (4) must be multiplied by 100 to recover percentages]. Since we lack precise information concerning production costs and CMH may have some characteristics from both ''north'' and ''south'' regions, we produce a second curve:

$$\mathrm{IP}_N = 0.000\,52\ \mathrm{MAPE} + 0.000\,07\ \mathrm{MAPE}^2, \quad (5)$$

which is similar to the less expensive north curves of Teisberg et al. (2005). The production cost savings S for that day (in thousands of dollars), resulting from a better forecast, is then

$$S = (\$45)D(\mathrm{IP}_{\mathrm{ref}} - \mathrm{IP}), \quad (6)$$

where an electricity rate of \$45 per MWh is assumed (based on continued fuel cost increases since Hobbs et al. 1999), D is the demand obtained from (3) using the observed mean temperature, and $\mathrm{IP}_{\mathrm{ref}}$ is obtained from (4) or (5) using the reference forecast (this reference will be provided in turn by persistence, NWS, MOS, a regression equation and in the case of perfect forecasts, also an EP consensus). Both (4) and (5) will be used to represent the range of cost savings that might be attained relative to the reference. A constant electricity rate of \$45 per MWh is an additional simplification as this rate will adjust with demand, typically rising as demand rises, but should be sufficient to provide approximate estimates of value in current dollar terms.

3. Results

Thirteen individual EPs are developed for forecasting minimum temperature, following the procedure outlined in section 2b. There is no particular reason to expect that the choice of 13 EPs is optimal. Rather, this choice was dictated by the need to have sufficient EPs to evaluate varying sizes of consensus, understanding that training an individual EP is computationally intensive. Results are presented with respect to performance on the test dataset.

As expected, EP consensus performance increases with increasing number of members (Fig. 3). The average performance, however, asymptotes within just a few members and the value of additional members appears to be primarily in reducing the interensemble volatility, depending on the specific set of individual algorithms selected to form the consensus. Despite this, even the worst performing algorithm is more than 12% better than MOS by this measure, while the best consensus (consisting of three members) improves relative to MOS by nearly 27%. All of these EP forecasts also exceed the NWS forecasts, although these are themselves about 8% better than MOS. Although NGM MOS is not the most recent or accurate form, the gap between more modern forms of MOS and NGM MOS is not this large. For example, comparing the north-central U.S. region for two summer and two winter months in 2006–07, representing several thousand forecasts, we find that the

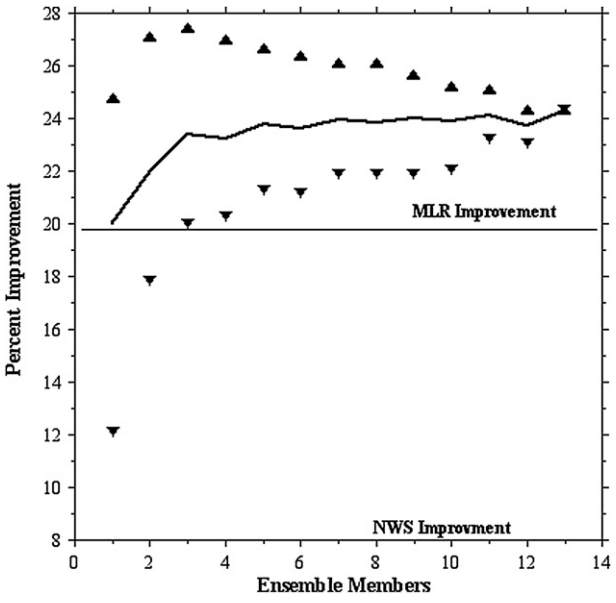

FIG. 3. Percent improvement of MAE for individual and EP consensus forecasts relative to MOS, as a function of the number of individual members. The solid line indicates the average of all possible forecasts for that number of members, and the points indicate the best and worst performing individual or consensus. NWS and MLR improvements relative to MOS are shown as the labeled horizontal lines.

AVN MOS minimum temperature forecasts show an approximate 8% reduction in MAE compared to NGM MOS for those same dates. Furthermore, some of these MOS advantages would be passed to the EP forecasts that use them or other components of improved forecasts.

To further examine performance, the development dataset (i.e., performance on the training data) is used to select the best individual algorithm and the best 3-member EP consensus, in addition to the 13-member EP consensus. The ranked performance in terms of MAE (Table 2) reveals a wide gap between MOS and the EP forecasts. The NWS forecasts show the usual improvement relative to MOS (e.g., Roebber and Bosart 1996), but also trail the EP forecasts. Finally, these results reveal that the bias-corrected reforecast ensemble temperature forecasts, although they are shorter forecast range than MOS (cf. 12 and 24 h, respectively), are comparable in skill to MOS and substantially inferior to the EP. This poorer performance likely arises from the coarse resolution of the reforecast model (see section 2a). For example, despite bias removal, for the independent *test* dataset, the reforecast temperatures exhibit a 0.5°F warm bias, some of which likely relates to the inability of the model grid to resolve local orographic details important to minimum temperature.

TABLE 2. MAE (°F) for all forecast systems, determined from the 170-case independent test dataset. See text for details concerning each forecast system.

Forecast	MAE (°F)
Persistence	5.40
Reforecast ensemble	4.09
NGM MOS	4.04
National Weather Service	3.72
MLR	3.24
13-member EP consensus	3.06
Best EP	3.04
3-member EP consensus	2.98

It seems likely that more state-of-the-art ensemble prediction systems would exhibit better performance than found here for the GFS reforecast ensemble (e.g., Hagedorn et al. 2008).

To what degree does temporal correlation compromise the independence of the test data? Suppose that the entire dataset contained no missing dates. In that instance, each test data point would have either a training or a cross-validation data point exactly one day prior to it. In that case, the standard of reference would be persistence. As noted, however, the MAE for persistence is 5.4°F (Table 2), while the MAE of the actual next closest development point (training or cross validation, when accounting for missing data) is 6.3°F. Given that the MAE for the EPs is considerably improved over either of these measures at 3.0°F, it is safe to conclude that the test data are sufficiently independent and that these results are representative of potential future performance.

Although the EP performance is promising, two other questions naturally arise. First, given the success of the EP forecasts, which as previously noted take the form of piecewise multiple (nonlinear) regression in the implementation here, one is prompted to ask whether a standard multiple linear regression equation would be competitive with EP. This may seem doubtful, given that MOS itself is a form of multiple linear regression, but it is of interest to assess how well the MOS equations have been optimized. Second, what do these forecast improvements mean, in practical terms? To answer this question, we translate the temperature forecasts to value in the electricity demand setting as detailed in section 2b, and examine the impact of these improvements on the occurrence and anticipation of forecast busts.

A stepwise, multiple linear regression using the inputs of Table 1 produces an equation of the following form:

$$T_{\min} = -0.768 + 0.233 \times \mathrm{PERS} + 0.174 \times \mathrm{T00} - 0.125 \times \mathrm{TD00} + 0.675 \times \mathrm{TD12} + 0.122 \times \mathrm{V00} - 2.578 \times \mathrm{SNOW}, \quad (7)$$

TABLE 3. Estimated annual electricity production cost savings relative to a specific reference forecast for three EP forecast combinations and for perfect forecasts. For example, the 3-member EP consensus shows cost savings of $4,551,048–$5,148,233 relative to persistence forecasts and $2,344,289–$3,060,658 relative to MOS. The electricity production is representative of demand for the state of Ohio. The range in cost savings is derived from northern and southern region production cost curves (see text for details).

Forecast	Reference forecast	Annual cost savings (relative to reference)
3-member EP consensus	Persistence	$4,551,048–$5,148,233
	NGM MOS	$2,344,289–$3,060,658
	National Weather Service	$1,912,717–$2,673,840
	MLR	$431,735–$617,368
13-member EP consensus	Persistence	$4,406,294–$4,971,263
	NGM MOS	$2,200,988–$2,883,157
	National Weather Service	$1,763,595–$2,488,594
	MLR	$288,187–$441,195
Best EP forecast	Persistence	$4,346,516–$4,919,913
	NGM MOS	$2,173,406–$2,861,197
	National Weather Service	$1,743,678–$2,474,432
	MLR	$267,359–$412,496
Perfect forecast	Persistence	$8,887,648–$10,858,674
	NGM MOS	$6,777,936–$8,867,976
	National Weather Service	$6,332,291–$8,430,765
	MLR	$4,889,634–$6,431,663
	3-member EP consensus	$4,352,400–$5,708,561

where the temperature (°F) is rounded to the nearest integer (hereafter MLR). The most influential variables in MLR, as indicated by the standardized regression coefficients, are in decreasing order: TD12, PERS, T00, and TD00. MLR is consistent with the *average* performance of a single EP but falls short of the best EP and the EP consensus forecasts (Fig. 3, Table 2). Given that MLR represents a substantial improvement over MOS, it is noteworthy that of its six variables, four are obtained from MOS, including three of the four most influential. This supports the notion that the MOS minimum temperature equation is not fully optimal, at least over the eight sampled years at this site (CMH) and further suggests that while multiple linear regression approaches will not provide the full capability of EP, there is room for improvement using this sort of ''nested'' regression technique.

Next, we turn to the value of these forecasts (Table 3). Even for electricity operations of the scale of a single state, the value of these forecast improvements potentially runs into the millions of dollars. Despite the relatively good performance of MLR, the additional development effort needed to generate EP consensus forecasts appears worthwhile, given that the potential annual cost savings run to more than $600,000. If this technique is scalable, these savings would increase considerably. For example, for the several upper midwestern U.S. states covered by the Midwest Independent Transmission Operator, where electricity demand is roughly an order of magnitude larger than that assumed to generate the figures in Table 3, then an annual savings of the order of $20 million could be expected relative to NWS forecasts. It is also apparent from these data that incremental improvements, such as changing from a 13-member to an optimized 3-member EP consensus, results in considerable value improvements.

Despite these encouraging results, there is room for further improvement, at least in principle. Perfect forecasts for the Ohio region provide potential cost savings of as much as $8 million relative to NWS forecasts and $5.7 million above the 3-member EP consensus. To answer why this might be, we next examine the issue of forecast busts. As noted in section 2a, busts occur in a wide variety of synoptic settings but most prominently under advective conditions. Presumably timing errors are a part of the problem. For the 170-event independent test dataset, 41 events or approximately 24% of the cases are forecast busts as defined by the NWS forecast error. How much can this record be improved by using EP? The 3-member EP consensus reduces the number of busts to 28 or 16% of the cases. This is indeed considerably better, but also suggests (unsurprisingly) that there may be limits to what can be done with current forecast capabilities.

The list of variables in Table 1 is not exhaustive; other variables, perhaps some relating to upwind trends, might in some circumstances provide useful information sufficient to further reduce events with the largest errors. Nonetheless, limits to predictability are intrinsic to weather forecasting. If we cannot always predict the temperature very well, perhaps we can predict those occasions when we are unlikely to predict the temperature well. This may have some utility and value by providing the opportunity for better planning and scheduling in those situations.

To consider this possibility, we employ logistic regression using the same variables as before and develop a probabilistic equation for forecast busts from the development dataset, where random case selection is employed to insure comparable numbers of busts and nonbusts. The equation uses two variables: the MOS wind speed forecast at 1200 UTC (V12) and the size of the deviation of the NWS forecast from persistence, where the stronger the winds and the larger the forecast deviation, the more likely there will be a forecast bust. The coefficients for

TABLE 4. Contingency table for observed and forecast busts, where a bust is defined as a NWS minimum temperature forecast error for CMH greater than 6°F. Forecast busts are predicted based upon a logistic regression equation (see text for details). POD is 0.61, FAR is 0.59, bias is 1.49, critical success index (CSI) is 0.33, and the odds ratio (OR) is 4.02.

		Observed		
		Bust	No bust	Total
Forecast	Bust	25	36	61
	No bust	16	93	109
Total		41	129	170

these variables are quite similar (0.116 and 0.101, respectively), meaning that all else being equal, a 1-kt increase in 1200 UTC forecast wind speed will increase the probability of a forecast bust by approximately 12%, while a 1° increase in temperature deviation from persistence will increase that probability by 11%. These two variables indicate the familiar result that as synoptic activity increases and/or conditions are changing, the forecast challenge increases.

The performance of this equation in classifying cases from the independent test dataset is summarized in Table 4. Detection of forecast busts is high [probability of detection (POD) = 0.61], but an overforecast bias is problematic [false-alarm ratio (FAR) = 0.59, Bias = 1.49]. Still, the odds ratio shows that it is 4 times more likely that a forecast of a bust will be correct than wrong, again suggesting that this equation might provide some utility in anticipating the more extreme forecast errors.

Finally, by way of example, we turn to interpretation of two of the EP. The best EP, as measured by MAE in the training dataset, can be reduced to the following logic:

if (V12 $\geq$ ΔMOS) and (F2 $\geq$ ΔMOS) then

$$\delta = -1.056 + 0.727 \times \mathrm{TD12} + 0.182 \times \mathrm{PERS} + 0.019 \times \mathrm{V00} - 0.034 \times \mathrm{SNOW} + 0.067 \times \mathrm{Vad} \times \mathrm{CLOUD} + 0.061 \times \mathrm{Uad};$$

else if (V12 $<$ ΔMOS) and (F2 $<$ ΔMOS) then

$$\delta = 0.145 + 0.727 \times \mathrm{TD12} + 0.182 \times \mathrm{PERS} + 0.019 \times \mathrm{V00} - 0.034 \times \mathrm{SNOW} + 0.067 \times \mathrm{Vad} \times \mathrm{CLOUD};$$

else if (V12 $<$ ΔMOS) and (F2 $\geq$ ΔMOS) then

$$\delta = -0.213 + 0.727 \times \mathrm{TD12} + 0.182 \times \mathrm{PERS} + 0.019 \times \mathrm{V00} - 0.034 \times \mathrm{SNOW} + 0.067 \times \mathrm{Vad} \times \mathrm{CLOUD} + 0.061 \times \mathrm{Uad};$$

else if (V12 $\geq$ ΔMOS) and (F2 $<$ ΔMOS) then

$$\delta = -0.698 + 0.727 \times \mathrm{TD12} + 0.182 \times \mathrm{PERS} + 0.019 \times \mathrm{V00} - 0.034 \times \mathrm{SNOW} + 0.067 \times \mathrm{Vad} \times \mathrm{CLOUD};$$

end if; (8)

where δ is the normalized temperature forecast sum that is then scaled up as in (2).

The basis is the 1200 UTC dewpoint temperature forecast, modified by persistence and with adjustments for wind speed (windy is warmer) and snow cover (snow is colder). A further adjustment occurs for meridional temperature advection under cloudy conditions, specifically warmer (colder) for warm (cold) advection under cloudy skies, but with no adjustment if skies are clear. Beyond this baseline, further adjustments occur depending on two additional conditions related to prevailing temperatures, windiness, and the variability in successive MOS forecasts. For warmer conditions, the forecast temperature further adjusts according to the strength of the zonal temperature advection, while for colder conditions, the baseline holds subject to a constant offset. For windy conditions or depending on the relative consistency of the successive model forecasts, further adjustments to the offset ensue.

The logic of this forecast algorithm is understandable. The overnight minimum temperature closely tracks the 1200 UTC dewpoint temperature, while persistence forms a reasonable baseline temperature forecast. High winds lead to increased mixing and reduced radiative cooling, while forecasters have long understood that snow cover results in colder overnight temperatures. The conditional basis of the adjustment to meridional temperature advection is interesting. Further examination of the data reveals that the MOS forecasts are biased toward underestimating the temperature response to meridional advection under cloudy conditions (not shown).

Note that it would be possible to produce such equations using a standard technique such as multiple linear regression (the nonlinear terms involving the product Vad $\times$ CLOUD can be reduced to linear by noting that CLOUD is 0 or 1 and transforming the equations with additional if statements). Exploring all the intricacies of the data needed to accomplish this, however, would be at best cumbersome and most likely infeasible. Allowing the solution to evolve through fitness pressure provides a means to reveal these connections readily and produces the small but valuable gains shown relative to MLR (Fig. 3, Tables 2 and 3).

A second EP is provided for comparison, with logic as follows:

if (Vad $<$ T00) and (SNOW) then

$$\delta = 0.062 + 0.676 \times \text{TD12} + 0.105 \times \text{PERS} - 0.018 \times \text{Vshr} \times \Delta\text{Ushr} + 0.053 \times \Delta V \times \text{T00} - 0.103 \times \text{Vad};$$

else if (Vad $<$ T00) and not (SNOW) then

$$\delta = 0.057 + 0.676 \times \text{TD12} + 0.105 \times \text{PERS} + 0.053 \times \Delta V \times \text{T00};$$

else if (Vad $\geq$ T00) and (SNOW) then

$$\delta = 0.062 + 0.707 \times \text{TD12} + 0.105 \times \text{PERS} - 0.018 \times \text{Vshr} \times \Delta\text{Ushr} + 0.053 \times \Delta V \times \text{T00} - 0.103 \times \text{Vad} - 0.094 \times \Delta\text{SNOW} \times \Delta\text{SOUTH};$$

else if (Vad $\geq$ T00) and not (SNOW) then

$$\delta = 0.057 + 0.707 \times \text{TD12} + 0.105 \times \text{PERS} + 0.053 \times \Delta V \times \text{T00} - 0.094 \times \Delta\text{SNOW} \times \Delta\text{SOUTH};$$

end if; (9)

where δ is as in (8).

Although (9) is fully nonlinear as it requires several variable product terms, the basis of the algorithm is similar to the first: use a combination of the 1200 UTC dewpoint temperature and persistence, modified by other factors. In this case, the other factors are distinct, however, leading to some independence in the forecasts. For example, this algorithm adjusts for snow cover by opposing the influence of warm advection, presumably a physical adjustment associated with cooling of warm air from below. In addition, for snow situations in which strong meridional cold air advection is occurring, temperatures are adjusted downward in proportion to the strengthening of the cold pool, perhaps reflecting a further diabatic adjustment. The maximum adjustment for this effect, however, is on the order of 2°F compared to 8°F for the cooling effects of the snowpack.

A third snow effect is included related to changes in snow cover when strong meridional warm air advection is occurring. If the winds have come up from the south in the past 24 h and the snow cover has disappeared during that time, then the temperatures are adjusted upward by about 7°F. If, on the other hand, the snow cover is unchanged, this adjustment is not made. Likewise, under a less usual synoptic scenario in which southerly winds have developed but there is a fresh snow cover, the temperature is dropped by the same amount. Finally, regardless of snow cover or meridional temperature advections, temperature adjustments are made based upon changes in overnight wind speed over the past 24 h. Specifically, for increasing winds, an upward adjustment of up to 4°F occurs, depending on how warm it is expected to be prior to the overnight period (cold conditions warm less). Conversely, if winds are decreasing, then this amount of cooling occurs, with the largest cooling occurring for warm conditions. This logic appears to pertain to the effectiveness of radiative cooling processes under wind mixing.

The forecast logic of these two algorithms is relatively sophisticated and is reminiscent of skilled, experienced forecasters examining the same set of information and coming to independent conclusions on the basis of their weighting of those elements. It is for this reason that the EP consensus forecasts are able to show additional improvement and value relative to the individual forecasts. The ability of these algorithms to reveal sophisticated reasoning suggests that they may also be useful as an aid in staff training; such an application bears further exploration.

4. Conclusions

A method for generating consensus forecasts invoking the principles of simulated evolution is explored for the application of minimum temperature forecasts at a site in Ohio (CMH). The resulting forecast algorithms [i.e., evolved programs (EPs)] take the form of piecewise multiple linear or nonlinear regression equations. The logic of these algorithms is interpretable in terms of physics that might be accounted for by sophisticated forecasters. Notably, the logic of individual algorithms is unique and thus the ability of consensus to act as a filter, focusing on those components of the forecast that can be considered the most reliable, acts to produce substantial increases in forecast accuracy relative to benchmarks such as MOS and forecasts issued by the NWS. In this regard, a three-member EP consensus, identified from a development dataset, is shown to provide the greatest accuracy for an independent test dataset (MAE of 2.98°F, representing a 27% improvement relative to MOS). These results are translated to approximate values in the context of electricity demand forecasting and show that potential annual cost savings for electricity production in the state of Ohio are of the order of $2 million relative to the NWS forecasts.

A nested multiple linear regression equation also exhibits marked improvement relative to MOS (MAE of 3.24°F), where the nesting occurs since MOS equations for temperature, dewpoint temperature, and wind speed are input to another regression equation intended to estimate the minimum temperature. This result suggests

that further optimization of multiple linear regression approaches are possible, given that the additional annual value obtained via the further improvements of the EP relative to MOS are on the order of $600,000 for Ohio. Room for improvement is considerable, however, with perfect forecasts providing nearly $6 million in annual electricity production cost savings relative to the best EP consensus.

Here, the value associated with forecast improvements using EP is shown only for electricity-demand forecasting. In that situation, value changes depending upon location on the temperature curve, but small forecast differences can produce large values. There are other contexts, however, where value is not continuous. Consider, for example, agricultural forecasts where the critical consideration is the freezing point. It is not necessarily the case that small forecast improvements will lead to increased values in that context, since what is important then is relative performance in detection of temperatures above or below the freezing point. These issues require further analysis.

The EP consensus is found to substantially reduce forecast busts (defined as minimum temperature forecast errors of at least 6°F), from 24% to 16% of the cases in the independent test dataset. It is shown that it is possible to provide guidance on those occasions when the forecasts are most likely to be poor through the development of a logistic regression equation with two variables: the MOS wind speed forecast at 1200 UTC and the size of the deviation of the NWS forecast from persistence. A 1-kt increase in 1200 UTC forecast wind speed will increase the probability of a forecast bust by approximately 12%, while a 1° increase in temperature deviation from persistence will increase that probability by 11%, indicating that as synoptic activity increases and/or conditions change, the forecast challenge increases. The equation shows some overforecast bias, but it is 4 times more likely that a forecast of a bust will be correct than wrong, suggesting that it might provide some utility in anticipating the most egregious forecast errors.

How might such information be used in decision making? One example is the electricity market. Electricity price in a deregulated market is set by supply and demand, with the supply based on submitted offers by the generators and demand based on bids from the companies that serve the electricity load (retailers). Since demand (and increasingly supply, as with wind or solar power) is partly a function of meteorology, there can be high price volatility such as when unexpected demand occurs that must be met. Both generators and retailers can enter into hedge contracts to protect themselves from this financial risk. Better information concerning uncertainties resulting from the weather will lead to fairer pricing in these contracts, since the pricing must in part reflect the anticipated volatility.

A search of the meteorological literature reveals only two applications of evolutionary programming to forecast problems to date: this study for minimum temperature and that of Bakhshaii and Stull (2009) for precipitation in complex terrain. In both applications, considerable promise is revealed; more research in this area, focused on optimizing this technique and applying it to a wider range of forecast problems, seems likely to be productive. In particular, the results presented here represent a proof of concept and motivate further exploration for a wider range of locations, longer periods of time, different forecast projections, improved inputs (such as more recent MOS products and/or state-of-the-art ensemble forecast products), and perhaps additional variables.

In this work, no attempt has been made to consider probabilistic applications of EP. Hamill and Whitaker (2007) and McCollor and Stull (2009) show that ensemble forecasts of 2-m temperature suffer from too little spread; that is, many observations fall outside of the extremes of the ensemble. Correction for model bias did not substantially alleviate this problem. The EP consensus forecasts suffer the same limitation for this dataset (not shown). Future work might be directed toward EP consensus calibration, which Hamill and Whitaker (2007) show is both effective and most necessary for surface data. This step would open useful probabilistic applications for EP consensus forecasts. Another useful direction would be to consider possible improvements to raw or calibrated ensemble predictions system output, using the ensemble predictions from state-of-the-art models (as opposed to the reforecast GFS used here) as input to EPs.

This begs an important question that has been posed in recent years: what is the future of human forecasting? Roebber and Bosart (1996) argue that the relative skill advantage of human forecasters reflects the ability to recognize instances of model forecast bias specific to certain synoptic situations. Bosart (2003) suggests that human forecasters are hamstrung by the lack of real-time data of sufficient resolution and quality, which renders it difficult to exercise their natural advantage relative to software in pattern recognition. In effect, forecasters run the risk of losing situational awareness by being relegated to the role of supervisor of a largely automated forecast system, that is, a system that is driven primarily by model guidance. The challenge is akin to that of a modern airline pilot, who must maintain alertness and skill over long inactive periods in preparation for those few instances where intervention is truly needed. In such a situation, it is not hard to imagine that a usual

lack of necessary observations will increase the likelihood of forecasters failing to recognize key observations when and where they do exist. This may particularly be the case in the context of forecast outliers, here, the 1-in-6 forecast situations that lead to large minimum temperature forecast errors.

The development of EP will not alleviate this challenge, but rather exacerbate it. Evolutionary programming and other advanced methods that can provide competitive and reliable forecasts seem likely to reduce further the list of weather forecast tasks most suitable for humans. The question is how best to allocate available forecaster time; that is, to what tasks should forecasters be attached? Mass (2003) argues that forecasters should focus on those activities where their chances of adding skill and value are the greatest: nowcasting and weather threats to public safety. As this paper shows, it is already possible to replicate sophisticated human forecast strategies through computer methods. This argues even more strongly for the need to fully develop a scientific forecast community, informed by the necessary observations, and dedicated to focusing on high-impact, public forecasts. To do otherwise risks redundancy and quickly thereafter, irrelevance.

Acknowledgments. This research developed as an outgrowth of a research project concerned with minimum temperature forecast busts in southwestern Ohio (UCAR COMET S05-53805). I would like to add my heartfelt thanks to the late John Distefano, former NWS Science and Operations Officer at Wilmington, Ohio, whose ongoing interest in that project through his devastating illness was inspiring. Lisa Bengtsson-Sedlar collected much of the data for that project, while Melissa Schumann and Sarah Reinke assisted in that earlier analysis.

REFERENCES

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

Bosart, L. F., 1975: SUNYA experimental results in forecasting daily temperature and precipitation. *Mon. Wea. Rev.,* **103,** 1013–1020.

——, 2003: Whither the weather analysis and forecasting process? *Wea. Forecasting,* **18,** 520–529.

Brooks, H. E., M. S. Tracton, D. J. Stensrud, G. Dimego, and Z. Toth, 1995: Short-range ensemble forecasting: Report from a workshop, 25–27 July 1994. *Bull. Amer. Meteor. Soc.,* **76,** 1617–1624.

Cheng, S., J. Klaassen, N. Comer, H. Auld, D. MacIver, and A. Liu, 2001: The impacts of summer heat and air quality on energy and health in Ontario. Presentation to *Sustainable Energy Futures for Central Ontario: The Impacts of Extreme Weather, Climate Change and a Changing Regulatory Environment*; March 22, 2004, Toronto, ON, Canada, 9 pp. [Available online at http://www.pollutionprobe.org/Happening/pdfs/march22sustenergy/auld.pdf.]

Ferreira, C., 2006: *Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence*. 2nd ed. Springer, 478 pp.

——, cited 2010: GeneXproTools4.0: Modeling made easy. [Available online at http://www.gepsoft.com.]

Fritsch, J. M., J. Hilliker, J. Ross, and R. L. Vislocky, 2000: Model consensus. *Wea. Forecasting,* **15,** 571–582.

Gorsuch, R. L., 1983: *Factor Analysis*. L. Erlbaum Associates, 425 pp.

Gyakum, J. R., 1986: Experiments in temperature and precipitation forecasting for Illinois. *Wea. Forecasting,* **1,** 77–88.

Hagedorn, R., T. M. Hamill, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Two-meter temperatures. *Mon. Wea. Rev.,* **136,** 2608–2619.

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

——, 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.

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

Hillis, W. D., 1990: Co-evolving parasites improve simulated evolution as an optimization procedure. *Physica D,* **42,** 228–234.

Hobbs, B. F., S. Jitprapaikulsarn, S. Konda, V. Chankong, K. A. Loparo, and D. J. Maratukulam, 1999: Analysis of the value for unit commitment decisions of improved load forecasts. *IEEE Trans. Power Syst.,* **14,** 1342–1348.

Hotelling, H., 1933: Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.,* **24,** 417–498.

Kalnay, E., 2003: *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, 341 pp.

Klein, W. H., B. M. Lewis, and I. Enger, 1959: Objective prediction of five-day mean temperature during winter. *J. Meteor.,* **16,** 672–682.

Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z. Zhan, C. E. Williford, S. Gadgil, and S. Surendran, 1999: Improved weather and seasonal climate forecasts from a multimodel superensemble. *Science,* **285,** 1548–1550.

Mass, C. F., 2003: IFPS and the future of the National Weather Service. *Wea. Forecasting,* **18,** 75–79.

McCollor, D., and R. Stull, 2009: Evaluation of probabilistic medium-range temperature forecasts from the North American ensemble forecast system. *Wea. Forecasting,* **24,** 3–17.

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

Roebber, P. J., and L. F. Bosart, 1996: The complex relationship between forecast skill and forecast value: A real world analysis. *Wea. Forecasting,* **11,** 544–559.

——, ——, and G. S. Forbes, 1996: Does distance from the forecast site affect skill? *Wea. Forecasting,* **11,** 582–589.

——, D. M. Schultz, and R. Romero, 2002: Synoptic regulation of the 3 May 1999 tornado outbreak. *Wea. Forecasting,* **17,** 399–429.

Rosenzweig, C., and W. Solecki, 2001: Climate change and a global city: The potential consequences of climate variability and change—Metro East Coast. *Report for the U.S. Global Change Research Program,* National Assessment of the Potential Consequences of Climate Variability and Change for the United States, Columbia Earth Institute, Columbia University.

Sanders, F., 1973: Skill in forecasting daily temperature and precipitation: Some experimental results. *Bull. Amer. Meteor. Soc.,* **54,** 1171–1178.

——, 1979: Trends in skill of daily forecasts of temperature and precipitation, 1966–78. *Bull. Amer. Meteor. Soc.,* **60,** 763–769.

Stensrud, D. J., H. E. Brooks, J. Du, M. S. Tracton, and E. Rogers, 1999: Using ensembles for short-range forecasting. *Mon. Wea. Rev.,* **127,** 433–446.

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.

Thompson, P. D., 1977: How to improve accuracy by combining independent forecasts. *Mon. Wea. Rev.,* **105,** 228–229.

Toth, Z., E. Kalnay, S. M. Tracton, R. Wobus, and J. Irwin, 1997: A synoptic evaluation of the NCEP ensemble. *Wea. Forecasting,* **12,** 140–153.

——, Y. Zhu, T. Marchok, M. S. Tracton, and E. Kalnay, 1998: Verification of the NCEP global ensemble forecasts. Preprints, *12th Conf. on Numerical Weather Prediction,* Phoenix, AZ, Amer. Meteor. Soc., 286–289.

Tracton, S., J. Du, Z. Toth, and H. Juang, 1998: Short-range ensemble forecasting (SREF) at NCEP/EMC. Preprints, *12th Conf. on Numerical Weather Prediction,* Phoenix, AZ, Amer. Meteor. Soc., 269–272.

Valor, E., V. Meneu, and V. Caselles, 2001: Daily air temperature and electricity load in Spain. *J. Appl. Meteor.,* **40,** 1413–1421.

Vislocky, R. L., and J. M. Fritsch, 1995: Improved model output statistics forecasts through model consensus. *Bull. Amer. Meteor. Soc.,* **76,** 1157–1164.

——, and ——, 1997: Performance of an advanced MOS system in the 1996–97 National Collegiate Weather Forecasting Contest. *Bull. Amer. Meteor. Soc.,* **78,** 2851–2857.

Vukicevic, T., I. Jankov, and J. McGinley, 2008: Diagnosis and optimization of ensemble forecasts. *Mon. Wea. Rev.,* **136,** 1054–1074.

Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2008: Can multimodel combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quart. J. Roy. Meteor. Soc.,* **134,** 241–260.

Zhang, F., C. Snyder, and R. Rotunno, 2002: Mesoscale predictability of the ''surprise'' snowstorm of 24–25 January 2000. *Mon. Wea. Rev.,* **130,** 1617–1632.

Zhu, Y., Z. Toth, R. Wobus, D. Richardson, and K. Mylne, 2002: The economic value of ensemble-based weather forecasts. *Bull. Amer. Meteor. Soc.,* **83,** 73–83.