Calibrated Probabilistic Mesoscale Weather Field Forecasting: The Geostatistical Output Perturbation Method

Yulia GEL, Adrian E. RAFTERY, and Tilmann GNEITING

Probabilistic weather forecasting consists of finding a joint probability distribution for future weather quantities or events. It is typically done by using a numerical weather prediction model, perturbing the inputs to the model in various ways, and running the model for each perturbed set of inputs. The result is then viewed as an ensemble of forecasts, taken to be a sample from the joint probability distribution of the future weather quantities of interest. This is typically not feasible for mesoscale weather prediction carried out locally by organizations without the vast data and computing resources of national weather centers. Instead, we propose a simpler method that breaks with much previous practice by perturbing the *outputs*, or deterministic forecasts, from the model. Forecast errors are modeled using a geostatistical model, and ensemble members are generated by simulating realizations of the geostatistical model. The method is applied to 48-hour mesoscale forecasts of temperature in the North American Pacific Northwest between 2000 and 2002. The resulting forecast intervals turn out to be empirically well calibrated for individual meteorological quantities, to be sharper than those obtained from approximate climatology, and to be consistent with aspects of the spatial correlation structure of the observations.

KEY WORDS: Empirical calibration; Ensemble forecast; Geostatistical simulation; Probabilistic weather prediction.

1. INTRODUCTION

In this article, we propose a way to obtain probabilistic mesoscale weather forecasts that are empirically calibrated, sharp, and applicable to whole weather fields simultaneously, rather than just to individual weather events. A probabilistic weather forecast is a (joint) probability distribution of a set of future weather quantities, to be distinguished from a point or deterministic forecast, which is just a single forecast of the quantities. Mesoscale weather forecasts are local forecasts that have resolutions on the order of 1–12 km and typically cover areas on the order of 500–1,000 kilometers square, compared with global and synoptic forecasts that have resolutions typically on the order of 30–100 km and much larger, sometimes planetary, areas of coverage. We say that a probabilistic forecast is calibrated if events declared to have probability p occur a proportion p of the time on average, and we say that it is sharp if prediction intervals are shorter on average than intervals with the same probability content derived from the long run marginal distribution (sometimes called climatology).

Up to about 1955, all practical weather forecasting was done by humans who integrated the available information subjectively, using their professional experience. Bjerknes (1904) proposed that weather forecasting be done by dynamically solving a system of seven partial differential equations in seven unknowns that represent the state of the atmosphere. To do this requires the specification of initial conditions and lateral boundary conditions. Richardson (1922) described a vision of doing this numerically, but it was not until 1955 that numerical solution of the systems of differential equations began to become possible thanks to the advent of computers. The quality of numerical weather predictions improved steadily, and by about 1995 synoptic models consistently provided good point forecasts up to about 3 days ahead.

Up to about 1995, numerical weather forecasting was mostly done in practice on global and synoptic scales and required vast amounts of computing resources. As a result, it was done mostly in a small number of national weather centers that had considerable data and computing resources, including supercomputers. The resulting forecasts were then released for public use. Local forecasters, such as those working for the media, aviation, shipping, and the military, typically produced forecasts for their areas of interest essentially by subjectively adjusting the synoptic forecasts and interpolating between the grid points, using knowledge of local terrain and weather patterns.

The past 10 years have seen a revolution in the practice of numerical weather prediction. Increased model resolution and improved model physics have made mesoscale numerical weather prediction possible. The MM5 (National Center for Atmospheric Research–Penn State Mesoscale Model Generation 5) is the most used mesoscale model. The advent of MM5 and fast desktop computers have made local numerical weather prediction possible, and now thousands of organizations are doing it, instead of the handful of weather organizations worldwide a decade ago. Typically, these organizations obtain the initial conditions for MM5 from global or synoptic forecasts provided by the large weather forecasting organizations.

Probabilistic numerical weather prediction has been much slower to develop than point forecasts. Epstein (1969) proposed that it be carried out by specifying uncertainty in the initial and lateral boundary conditions, and propagating these through to the quantities being forecast. Leith (1974) proposed doing this in practice by Monte Carlo simulation, generating an *ensemble* of different initial conditions, running each of them forward using the model to obtain forecasts, and using the resulting set of forecasts as a predictive probability distribution of the future weather quantities being forecast. Murphy and Winkler (1979) called for operational probabilistic temperature forecasts. By the 1990s, three viable methods had been developed:

Yulia Gel is Visiting Assistant Professor, Department of Statistics, George Washington University, Washington, DC 20052 (E-mail: *ygl@gwu.edu*). Adrian E. Raftery is Professor of Statistics and Sociology (E-mail: *raftery@stat.washington.edu*) and Tilmann Gneiting is Associate Professor (E-mail: *tilmann@stat.washington.edu*), Department of Statistics, University of Washington, Seattle, WA 98195-4322 The authors are grateful to Mark Albright, Eric Grimit, and Clifford Mass for helpful discussions and for providing data, and to the editor, the associate editor, and the referees for constructive comments. This research was supported by the Department of Defense Multidisciplinary University Research Initiative (MURI) program administered by the Office of Naval Research under grant N00014-01-10745.

Journal of the American Statistical Association
September 2004, Vol. 99, No. 467, Applications and Case Studies
DOI 10.1198/016214504000000872

the breeding–growing modes method used by the U.S. National Centers for Environmental Prediction (Toth and Kalnay 1993), the singular vector method used by the European Centre for Medium-Range Weather Forecasts (Molteni, Buizza, Palmer, and Petroliagis 1996), and the perturbed observations method used by the Meteorological Service of Canada (Houtekamer, Lefaivre, Derome, Ritchie, and Mitchell 1996). Hamill, Snyder, and Morss (2000) compared these methods in an ideal model context and concluded that the perturbed observations method works best.

Ehrendorfer (1997) and Palmer (2000) reviewed techniques of probabilistic weather prediction that were in operational use by the mid and late 1990s. However, these methods do not apply directly to probabilistic mesoscale forecasting. The initial conditions being perturbed are typically specified by on the order of 10 million numbers. The perturbed observations method, for example, perturbs the observations on which the estimate of the initial conditions is based and then runs a cycle of data assimilation to turn them into initial conditions for the model. An organization that is running MM5 locally typically does not have access either to the observations used to generate the initial conditions or to the computing resources needed to perform the data assimilation. Also, errors in model physics are particularly important for mesoscale forecasts (Stensrud and Fritsch 1994a,b). Methods that perturb the initial conditions directly in a simple way are questionable, because the resulting sets of initial conditions may violate the equilibrium conditions of atmospheric physics, and so may give unstable results and hence not be usable.

There have been several mesoscale probabilistic forecasting methods developed using a range of initial conditions from different global models, including the ETA-Regional Spectral Model ensemble (Wandishin, Mullen, Stensrud, and Brooks 2001), the 1998 Storm and Mesoscale Ensemble Experiment (SAMEX; Hou, Kalnay, and Droegemeier 2001), and the University of Washington MM5 ensemble (Grimit and Mass 2002). Neither of the first two ensembles showed an ability to predict forecast reliability well. The third ensemble did, but the prediction intervals produced were far too narrow (Raftery, Balabdaoui, Gneiting, and Polakowski 2003), thereby suggesting unwarranted confidence in the forecasts. Specifically, observed temperatures fell much more often outside the ensemble range than would be expected from calibrated forecasts.

We propose to develop an easy to use mesoscale probabilistic forecasting method by directly perturbing the model *output*, or point forecasts, in contrast to the traditional approach of perturbing model inputs. If outputs (forecasts) are perturbed independently, one meteorological quantity at a time, the properties of overall fields will not be well forecast, because, for example, there will be no spatial correlation, while actual error fields show substantial spatial correlation. To avoid this, we model the errors using a geostatistical model that preserves the field's spatial correlation structure. We generate our ensembles by simulating realizations from the resulting spatial random field model. The result is a simple method that uses only the point forecasts, does not use simulated or perturbed observations or initial conditions, and implicitly incorporates uncertainty due to errors in model physics. In our numerical experiments, it turns out to be both empirically calibrated and sharp, and also to reproduce spatial properties of the observed field.

In Section 2, we describe the geostatistical output perturbation (GOP) method, including the basic statistical model, parameter estimation, geostatistical simulation method, and ways to verify the resulting model and forecasts. In Section 3, we apply the method to forecasting temperatures in the North American Pacific Northwest and show the results, including in-sample verification statistics and an out-of-sample predictive check. Finally, in Section 4, we discuss possible improvements to the methodology.

2. THE GEOSTATISTICAL OUTPUT PERTURBATION METHOD

We now describe the geostatistical output perturbation method. First we outline the underlying statistical model. Then we describe how it can be estimated from data and how realizations can be simulated from it efficiently. Finally, we explain how we go about verifying probabilistic forecasts.

2.1 Statistical Model

Let $\tilde{Y}(s,t)$ be the MM5 forecast value of a meteorological variable, $Y(s,t)$, at the spatial point $s \in \Re^2$, verifying at time t, at a given forecast lag. We focus on forecasting $\{Y(s,t) : s \in S\}$, simultaneously for all s on a grid of points S in the forecast region, but where t and the forecast lag are fixed. Our goal is to produce calibrated probabilistic forecasts of the kind of two-dimensional images that operational forecasters look at, where the whole image is calibrated, rather than just the individual forecasts of which it is composed.

Let $X(s,t)$ be a finite set of variables that correspond to location s and time t, and that are thought to be related to forecast bias. These might include functions of time of year or time of day, and functions of space such as latitude, longitude, altitude, distance from the ocean, and land use. Then our model is

$$Y(s,t) = \mathbf{a}^T X(s,t) + \big(\mathbf{b}^T X(s,t)\big)\tilde{Y}(s,t) + w(s,t). \quad (1)$$

Here $\mathbf{a}$ and $\mathbf{b}$ are parameter vectors, and $w(s,t)$ is a mean-zero stationary Gaussian space–time stochastic process model. Thus $\mathbf{a}^T X(s,t)$ models the additive bias of the forecasts from the numerical weather prediction model and $\mathbf{b}^T X(s,t)$ models the multiplicative bias.

At this stage, we are modeling only the spatial correlation in $w(s,t)$ and ignoring the temporal correlation, because the spatial correlation is what counts for getting calibrated images. For simplicity, and because it works well in the cases we have studied, we use the exponential spatial variogram model

$$\tfrac{1}{2}\operatorname{Var}\big(w(s_1,t) - w(s_2,t)\big) = \rho + \sigma^2\big(1 - \exp(-\|s_1 - s_2\|/r)\big) \quad (2)$$

whenever $s_1 \neq s_2$, where $\|\cdot\|$ is the Euclidean norm. This is a geostatistical model, and in geostatistical terminology, ρ is called the *nugget effect* and is usually thought of as the measurement error variance of observations, $\rho + \sigma^2$ is the marginal variance of $w(s,t)$ and is called the *sill*, and r is a *range* parameter and is measured in kilometers (Cressie 1993; Chilès and Delfiner 1999). The range parameter r is interpreted as follows. The error process $w(s,t)$ can be viewed as a sum of two

component processes: measurement error (viewed as spatially uncorrelated) and continuous spatial variation. The spatial correlation of the continuous spatial variation component process at distance d is $e^{-d/r}$. This spatial correlation declines from 1 at distance zero and reaches .05 at distance $3r$.

2.2 Parameter Estimation

We estimate the parameters of the model given by (1) and (2) using historical data on forecasts and observations. Typically we use data for a relatively homogeneous region over a recent time interval of length on the order of 3 months to a year, so as to avoid difficulties due to different patterns of model bias, changes in the numerical weather prediction model, and so on. It is possible to estimate the model using maximum likelihood or a fully Bayesian approach. Forecasts are on a grid and may correspond to a grid cell, whereas observations correspond to irregularly spaced locations, the so-called change of support problem, and a fully Bayesian approach has the advantage of being able to deal with this explicitly in a coherent way.

However, the datasets used for parameter estimation are typically very large, so full maximum likelihood estimation or fully Bayesian estimation tends to be prohibitively time-consuming. We, therefore, use a simpler and much faster three-stage estimation method that approximates maximum likelihood and works well in our implementation. First, we interpolate the forecasts, which are on a grid, to the observation locations, which are irregularly spaced, using bilinear interpolation. Then we estimate the coefficients $\mathbf{a}$ and $\mathbf{b}$ by linear regression, and compute the residuals $\hat{w}(s,t)$. Finally, we estimate the variogram parameters ρ, σ^2, and r by binning the residuals $\hat{w}(s,t)$ and using weighted nonlinear least squares with weights equal to the numbers of observations in the bins (Cressie 1993), as implemented in the R package geoR (Ribeiro and Diggle 2001).

2.3 Generating the Ensemble Members

The ensemble members are spatial forecasts specified on the model grid. They are generated simply by simulating realizations of the stochastic process given by (1) and (2), given the current forecast $\tilde{Y}$, and using the parameters estimated from the historical data. However, this is not as simple as it sounds, because it involves simulating a large number of correlated values simultaneously. For example, in the Pacific Northwest region that we consider, it would typically involve simulating 10,000 values or more. Because direct simulation from the very high-dimensional multivariate normal distribution is not feasible by standard techniques such as Cholesky decomposition of the covariance matrix, we must seek a more efficient method.

This is essentially the problem of generating the realizations of a stationary Gaussian random field, which has traditionally been solved by spectral methods, the turning bands method, moving average techniques, and a number of other approximative algorithms (Chilès and Delfiner 1999). We used the circulant embedding method of Wood and Chan (1994) and Dietrich and Newsam (1997) as implemented in the R package RandomFields (Schlather 2001). Contrary to the aforementioned techniques, the circulant embedding method of generating stationary Gaussian random fields is both fast and exact. Being exact means that the realizations have exactly the required multivariate normal distribution. The method is fast because it exploits the speed and efficiency of the fast Fourier transform. For simulations on a regular grid in $\Re^2$ and appropriate orderings of the grid points, the covariance matrix of the associated Gaussian random vector is a block Toeplitz matrix, with each block being Toeplitz itself. It can be embedded into a block circulant matrix, where all the blocks are circulant themselves and admit an eigenvalue decomposition in terms of a standard fast Fourier transform matrix. If all the eigenvalues of the block circulant matrix are positive, which is true for a large class of covariance structures, a random vector with the required multivariate distribution can be generated by the fast Fourier transform. The computational effort for a Gaussian random vector of size n is proportional to $n \log n$, which makes the exact simulation of grids with 10,000 or more correlated Gaussian values feasible.

2.4 Verifying and Assessing the Probabilistic Forecasts

We use two criteria to verify and assess our probabilistic forecasts: calibration of prediction intervals and sharpness of prediction intervals. The geostatistical approach ensures that the forecasts are consistent with key aspects of the spatial correlation structure of the observations.

To form, for example, 90% prediction intervals for individual future weather quantities, we simulate 19 realizations from the predictive random field, and take the minimum and the maximum as the endpoints of the 90% prediction interval. Alternatively, we could simulate 99 realizations, and take the 5th and 95th order statistics as the endpoints, or we could use another number of simulations, depending on our computational resources. We are interested in the average coverage and the average length of such intervals.

We also employ the rank histogram, a verification tool for probabilistic forecasts developed by atmospheric scientists (Anderson 1996; Hamill and Colucci 1997; Talagrand, Vautard, and Strauss 1997). Consider, for instance, an ensemble of 99 forecasts. If the ensemble is calibrated, the 99 forecasts and the observation will be exchangeable. Hence, if we order the combined set of 99 forecasts and 1 observed value, the rank of the observation will be equally likely to be any number between 1 and 100. The verification rank histogram, or simply rank histogram, is a histogram of these ranks. A uniform rank histogram indicates proper coverage of the prediction intervals at all levels. Deviations from uniformity imply a lack of calibration.

An important consideration here is that a single probabilistic forecast, that is, an ensemble on any given day, typically cannot be verified. The aforementioned quantities need to be computed as averages over many ensembles, and we do so below. Section 3.4 is concerned with the in-sample verification of the GOP ensemble. Section 3.5 provides an out-of-sample verification and predictive check, using new data on forecasts and observations made in 2001.

3. RESULTS

We now apply the GOP method to some data on temperature in the North American Pacific Northwest and show the results. We describe the data, the model estimation process, and the probability ensemble forecasts, and finally we give some results on the verification of the forecasting model.

3.1 Data

To estimate the model, we use forecast and observed temperatures during the period January–June 2000 in the North American Pacific Northwest. Temperature was measured at 0 hours Greenwich mean time (GMT; 00Z) on each of 102 days during this period at different observation locations. The number of observation locations varied by day, but was typically between 500 and 600; in all, there were 56,488 observations of temperature. Data on forecasts for about 80 days are missing, as a result of the numerical weather prediction model being run in research mode. The observation locations were of different types: for example, some were regular meteorological stations, some were snow monitoring stations, some were ships, and so on. The data were measured in degrees kelvin, where $x°$K is equal to $(x - 273.15)°$C. The observed temperatures ranged from 250.9 to 313.2, with mean 286.1, median 284.8, and standard deviation 8.5.

Forty-eight-hour forecasts verifying at each of the 102 times for which they are available were obtained. These forecasts were obtained from the MM5 model, initialized using the Aviation model of the National Weather Service's National Center for Environmental Prediction (NCEP) and run by researchers in Professor Clifford Mass's group at the University of Washington Department of Atmospheric Sciences. The forecasts were on a 12-km grid.

3.2 Parameter Estimation

To estimate the model, we first converted the forecasts from the model grid to the observation locations using bilinear interpolation. Because the grid is regular and fine relative to the observations (10,300 grid points compared with on the order of 500–600 observations on a typical day), it is unlikely that more complicated interpolation methods would lead to much better results.

For simplicity, we considered only a simple additive bias, a, and a simple multiplicative bias, b, in our model. The simplified model is

$$Y(s,t) = a + b\tilde{Y}(s,t) + w(s,t). \tag{3}$$

The regression estimate of the additive bias is 1.6 (standard error .4) and that of the multiplicative bias parameter is .995 (.002). Thus the additive bias is significant, but there is almost no multiplicative bias. The residuals from these regressions were computed and their variogram is shown in Figure 1, together with the fitted exponential variogram function (2). The estimated nugget is $\rho = .51$, the estimated variance of the continuous spatial variation component is $\sigma^2 = 7.2$, and the estimated range parameter is $r = 114$ km. The fit of the parametric variogram function is better than that often observed in geostatistical applications.

3.3 Ensembles of Forecasts: An Example

We now illustrate the use of our method to produce ensemble forecasts. We apply it to produce a probabilistic 48-hour ahead forecast of temperature in the Pacific Northwest, verifying on January 12, 2002 at 0 hours GMT (00Z); this forecast is based on information available on January 10, 2002 at 00Z. The probabilistic forecast applies to a time point $1\frac{1}{2}$ years after the period to which the data used to estimate the model pertain, so it is truly an out-of-sample forecast.

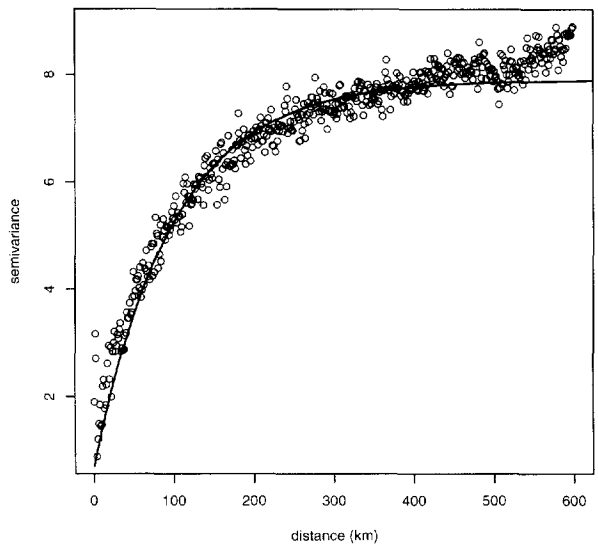

Figure 1. Empirical Spatial Variogram of $\hat{w}(s,t)$ With Fitted Exponential Variogram Function for Temperature in the North American Pacific Northwest, January–June 2000.

Figure 2 shows the gridded MM5 forecast, $\tilde{Y}$, produced by running MM5 initialized with the output from the synoptic NCEP Aviation model. This shows output on a 12-km grid. This figure also shows the bias-adjusted predictive mean, $a + b\tilde{Y}$.

Figure 3 shows four members of the forecast ensemble plotted on the 12-km grid. It is interesting to note that these plots are somewhat rougher than the point forecasts in Figure 2, reflecting the spatial roughness observed in actual data. The point forecasts are smoother because they represent the evolution of a system of partial differential equations that, over time, smooth out roughness to some extent, and also because, at least implicitly, they represent a kind of mean of a forecast distribution, which typically is smoother than an individual realization.

It is of interest to compare the forecast ensemble with the observed values. This is not straightforward, because the forecasts are on a relatively fine grid, whereas the observation locations are irregularly spaced and much sparser. We display the observations by interpolating the values to a fine grid using kriging (Cressie 1993; Chilès and Delfiner 1999), as implemented in the R package fields (Nychka 2003), and plotting the result, as shown in the top row of Figure 4. The gridded ensemble members in Figure 3 are not directly comparable to the plot of observations in Figure 4 and a visual comparison is misleading.

To make a valid visual comparison, we plot the ensemble members in a different way. We first estimate the ensemble forecast values at the observation locations, using simple bilinear interpolation; the grid is fine enough that other interpolation methods would yield similar results. We then interpolate the resulting forecasts by kriging to a fine grid, exactly as was done for the observations. The results are shown in Figure 4, with the station locations overlaid. The lower two rows show four ensemble plots, which can be compared to the actual observations on January 12, 2002 at 0 hours GMT (00Z) shown in the upper row.

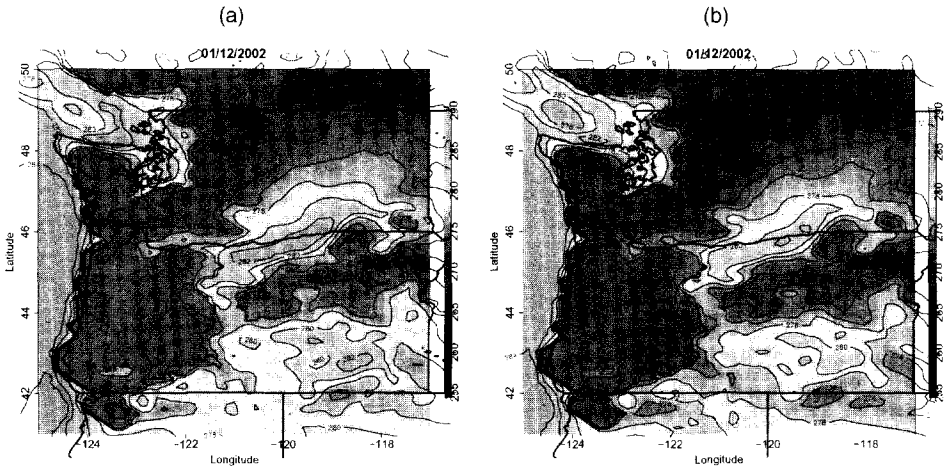

Figure 2. (a) The Gridded MM5 48-Hour Ahead Forecast, $\tilde{Y}$, of Temperature in the Pacific Northwest Verifying on January 12, 2002 at 0 Hours GMT; (b) The Bias-Adjusted Predictive Mean, $a + b\tilde{Y}$.

The forecasts seem to capture many aspects of the observations fairly well, and the observation and ensemble forecast plots look similar in the sense that the plots seem compatible with their having been generated by the same process. We simulated 99 realizations and formed 90% prediction intervals by taking the 5th and 95th order statistics as endpoints. The actual coverage on this specific date was 90.8%, and the correlation between point forecasts and observations was .66.

3.4 In-Sample Verification of the Forecasts

We consider verification of the forecasts in two ways, as described in Section 2.4: coverage of prediction intervals and sharpness of prediction intervals. The quantities reported are averages over 56,488 observations of temperature during 102 days in the period January–June 2000.

For coverage of prediction intervals, we considered two intervals: 66.7% and 90%. The 66.7% interval contained the

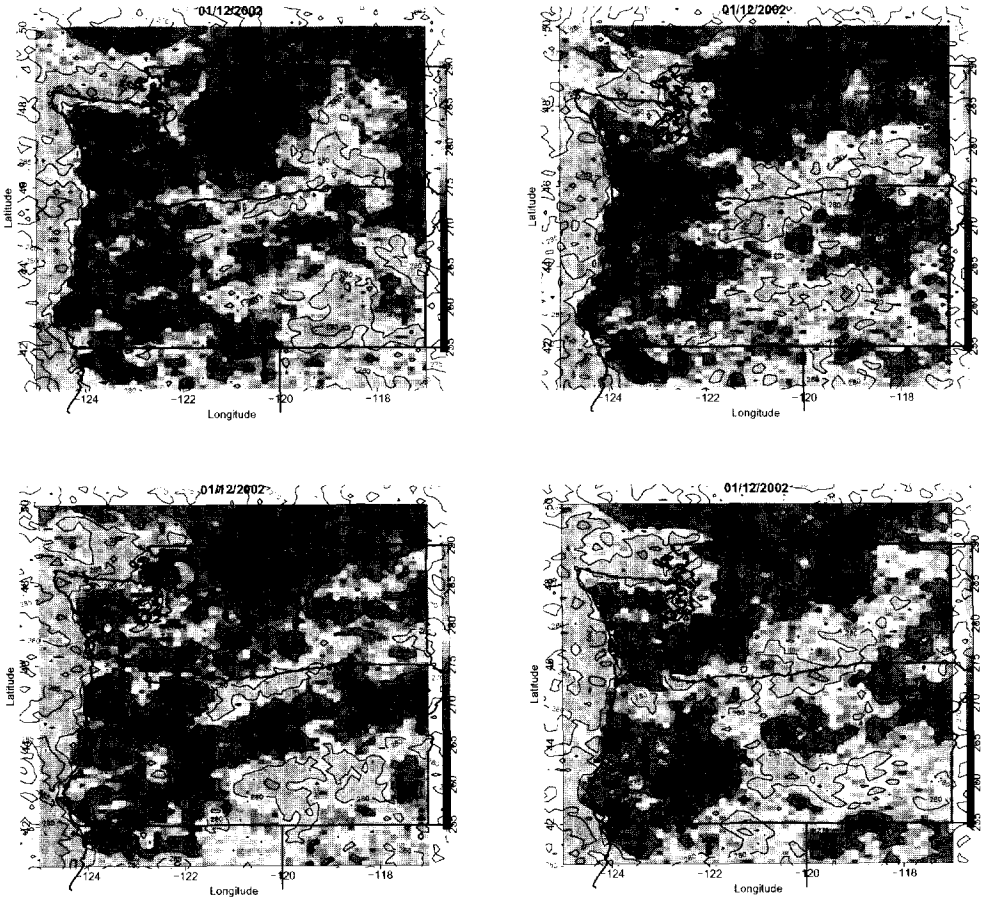

Figure 3. Ensemble of Forecasts for Temperature on January 12, 2002 Using Gridded MM5 Output.

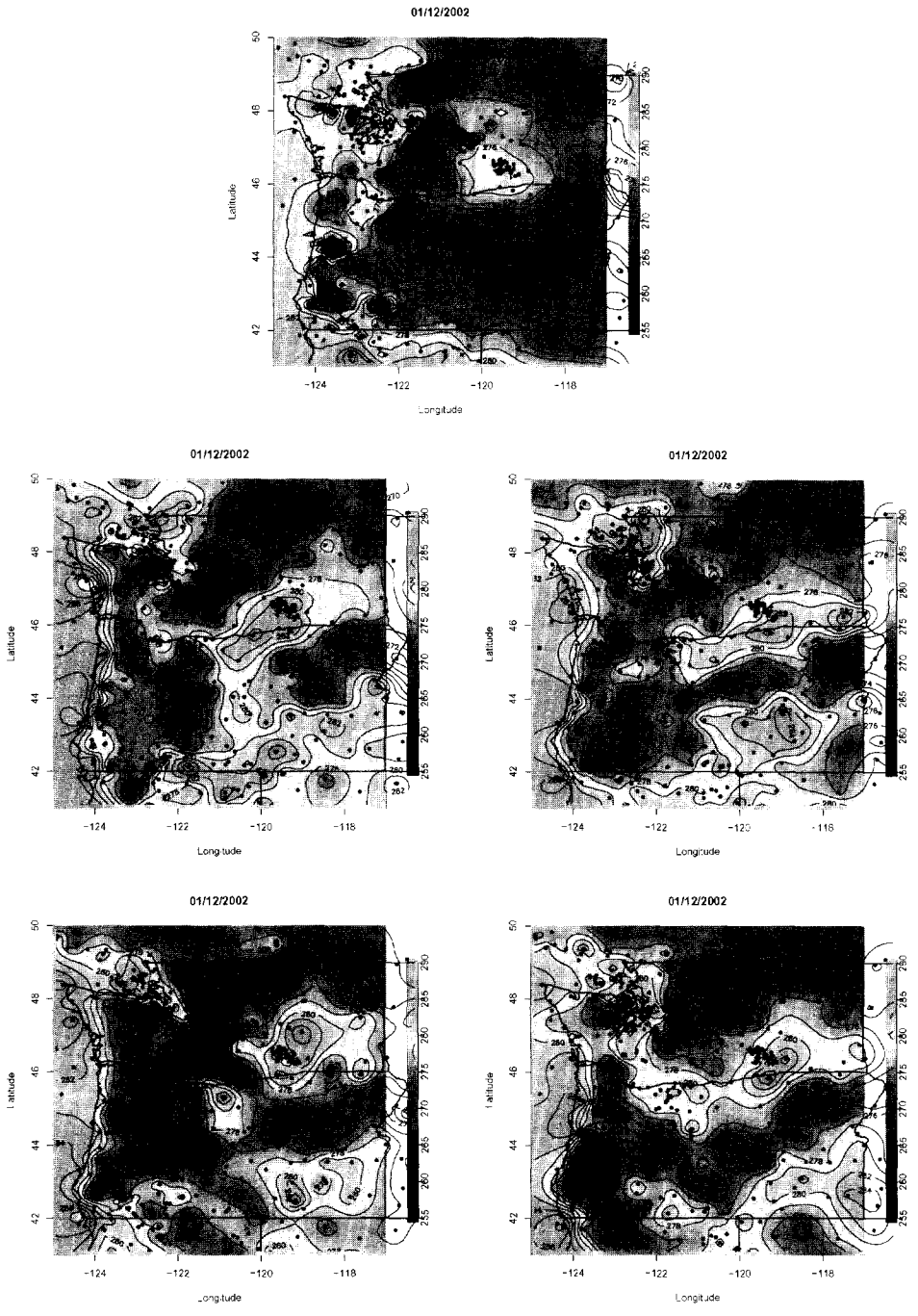

Figure 4. Observations and Ensemble Plots Compared for Temperature on January 12, 2002. The ensemble plots (lower two rows) were interpolated to the observation locations and then interpolated again to a fine grid by kriging. The observation locations are overlaid.

true value 68.1% of the time; the 90% interval contained the true value 90.8% of the time. The verification rank histogram, shown in Figure 5(a), is relatively close to being uniform, thereby indicating proper coverage of the prediction intervals at all levels.

To assess sharpness of prediction intervals, we computed the average length of the 66.7% (90%) prediction interval, which was 5.5 (9.4) degrees. For comparison, we computed the difference between the 16.7th (5th) and and 83.3rd (95th) percentiles of the marginal distribution of the observations, corresponding to the lengths of 66.7% (90%) prediction intervals based on sample climatology, which were 17.2 (28.3) degrees. Hence, the prediction intervals from the GOP method were considerably shorter, while remaining empirically calibrated.

3.5 Predictive Check

To assess the predictive performance of the GOP method, we applied it to gridded MM5 temperature forecasts in the period January–June 2001. In doing so, we retained the parameter estimates obtained by using data for forecasts and observations in the period from January to June 2000, as described in Sections 2.2 and 2.3. We discuss the results subsequently.

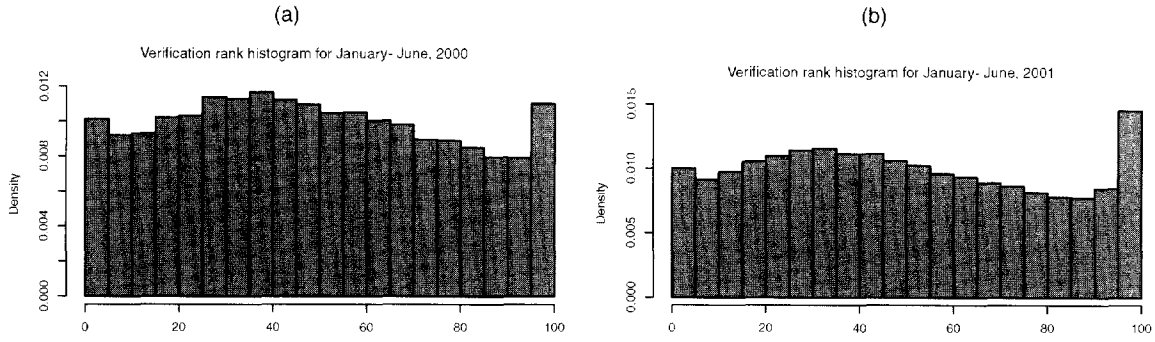

Figure 5. Verification Rank Histograms for GOP Ensemble Predictions. Rank histogram for (a) in-sample predictions and (b) out-of-sample predictions (June–July 2001).

The quantities reported are averages over 85,091 observations of temperature in the period January–June 2001. The 66.7% (90%) GOP prediction interval contained the true value 67.2% (88.0%) of the time. The verification rank histogram, shown in Figure 5(b), indicates a larger number of particularly high observations than anticipated by the ensemble. However, the rank histogram is much more uniform than that typically observed in conventional ensembles (Hamill and Colucci 1997) and the deviation is moderate. The average length of the 66.7% (90%) prediction interval was 5.5 (9.1) degrees.

4. DISCUSSION

We have proposed a method for mesoscale probabilistic forecasting that is feasible for local users of MM5 and other mesoscale models. It breaks with many previous approaches by not perturbing the inputs to the model, but perturbing the outputs from the model—the forecasts. The spatially correlated behavior of observed weather is reproduced using a geostatistical model and the ensembles are generated by simulating realizations from this model. In our numerical experiments, the resulting method turns out to be empirically well calibrated for individual forecast quantities, to be sharper than climatology, and to reproduce the spatial correlation behavior of observations.

One interesting aspect of our results is that the forecast ensemble members look rougher than the point forecasts. Meteorologists often look at point forecasts like Figure 2 and at plots, not of actual observations, but of an "analysis," which is an estimate of the current state of the atmosphere using the numerical weather prediction model. The analysis is made by combining the model's prediction with data, so it is smoother than data. This suggests that the adoption of calibrated probabilistic mesoscale forecasting may involve something of a culture change: forecasters must get used to looking at images like Figure 3, as well as the smoother point forecasts such as Figure 2 that they are used to. Houtekamer and Mitchell (1998, 2001) proposed generating ensembles with different initial values by taking an analysis, or "background field" as it is sometimes called, perturbing it by adding errors drawn from a random field model, and then running the numerical weather prediction model forward from each resulting initialization to obtain an ensemble of forecasts. This differs from our approach in that we perturb the forecast field and use the resulting simulations directly as an ensemble of forecasts; we run the numerical weather prediction model only once.

There are many ways in which our method, as currently implemented, could be improved. The most obvious is bias correction. In our implementation, we used a very simple bias correction method, although in principle the method allows for the use of many independent variables for this purpose, such as time of year or time of day, and functions of space such as latitude, longitude, altitude, distance from the ocean, and land use. Linear regression methods for correcting the biases of deterministic meteorological prediction models are known as *model output statistics* (Wilks 1995). A modern hierarchical Bayes approach was proposed by Nott, Dunsmuir, Kohn, and Woodcock (2001).

Our method is designed to be based on a relatively limited space–time window, given the effects of model changes over time and spatial inhomogeneity. In our experiments, we used a 6-month period in the Pacific Northwest to fit the model. More research is needed to determine the best temporal window and to develop more systematic ways to decide what it should be.

The statistical model underlying our work is quite simple and surprisingly effective given its simplicity. Nevertheless, various elaborations might improve its performance. Allowing a more general spatial covariance class, such as the Matérn class or related models (Gneiting 1999), taking account explicitly of the different spatial scales of the observations and the forecasts, allowing for non-Gaussian forecast error distributions by using a transformation, and taking account of temporal autocorrelation might all improve the results. These enhancements could all be done by taking a fully Bayesian approach using Markov chain Monte Carlo methods. The Bayesian model calibration technique of Kennedy and O'Hagan (2001), for example, used observational data to learn about model inadequacy. Such an approach is quite expensive computationally, however, and given the vast amounts of data involved in weather forecasting and the need for real-time forecasts, it may not be feasible for a while yet. The gains from such elaborations seem likely to be incremental rather than transformative.

Many applications of probabilistic weather forecasting focus on a single weather variable. Farmers and transportation managers, for instance, are often concerned about the probability of temperatures below the freezing point. The weather risk insurance and weather derivatives industries typically focus on

heating or cooling degree days, which are derived from temperature. The organizers of an outdoor event might be interested primarily in the probability of precipitation. Still, multivariable probabilistic weather field forecasts have important applications, such as ship and aircraft routing. Fitting vector-valued random function models to forecast error fields and simulating realizations thereof is methodologically challenging, but possible in principle (Daley 1991; Chan and Wood 1999). Such an approach would generate joint probability distributions of temperature and wind components, or other weather variables, that would be a useful extension of the GOP technique.

Finally, it might be feasible to improve the GOP method by using information from conventional mesoscale ensembles. Typically, the individual members of a conventional mesoscale ensemble differ from each other with respect to the source of the initial and boundary conditions used. Grimit and Mass (2002) showed a clear relationship between the variation among forecasts based on initial conditions supplied by different weather centers and the mean absolute forecast error, the so-called spread-skill relationship. This relationship could be exploited to obtain a better assessment of the spread of the predictive distribution in the present context. Specifically, regression models for the predictive mean and predictive variance in terms of the individual predictions from the conventional ensemble could be developed. Fitting a stationary geostatistical model to historical error fields, simulating standardized realizations from this model, and transforming locally by taking account of the site-specific predictive mean and predictive variance might be a way to form a statistical ensemble. Alternatively, the ideas of Bayesian model averaging (Hoeting, Madigan, Raftery, and Volinsky 1999; Raftery et al. 2003) could be combined with the present framework. The resulting method would, essentially, make the marginal variance of the space–time stochastic process model $w(s,t)$ in (1) vary temporally and spatially, with variance depending on the ensemble spread between forecasts based on initial conditions provided by different weather centers. Another approach to combining conventional and statistical ensembles, and thereby exploiting the spread-skill relationship, was suggested by Roulston and Smith (2002).

Hybrid methods of this type hold the promise to retain the advantages of either kind of ensemble, while avoiding their disadvantages. Conventional ensembles yield a dynamic assessment of uncertainty, and the associated weather field plots provide more realistic synoptic-scale features that are difficult to achieve with statistical approaches. On the other hand, the GOP method was empirically calibrated in our experiments, provides a more realistic reproduction of the mesoscale variability of weather parameters, and generates hundreds of ensemble members in real time, all major challenges for conventional ensembles. A decade or two from now, optimal operational ensembles might well combine conventional and statistical ensembles in sophisticated ways.

[Received March 2003. Revised January 2004.]

REFERENCES

Anderson, J. L. (1996), "A Method for Producing and Evaluating Probabilistic Forecasts From Ensemble Model Integrations," *Journal of Climate*, 9, 1518–1530.

Bjerknes, V. (1904), "Das Problem der Wettervorhersage, Betrachtet vom Standpunkte der Mechanik und der Physik," *Meteorologische Zeitschrift*, 21, 1–7.

Chan, G., and Wood, A. T. A. (1999), "Simulation of Stationary Gaussian Vector Fields," *Statistics and Computing*, 9, 265–268.

Chilès, J.-P., and Delfiner, P. (1999), *Geostatistics: Modeling Spatial Uncertainty*, New York: Wiley.

Cressie, N. A. C. (1993), *Statistics for Spatial Data*, New York: Wiley.

Daley, R. (1991), *Atmospheric Data Analysis*, Cambridge, U.K.: Cambridge University Press.

Dietrich, C. R., and Newsam, G. N. (1997), "Fast and Exact Simulation of Stationary Gaussian Processes Through Circulant Embedding of the Covariance Matrix," *SIAM Journal on Scientific Computing*, 18, 1088–1107.

Ehrendorfer, M. (1997), "Predicting the Uncertainty of Numerical Weather Forecasts: A Review," *Meteorologische Zeitschrift N. F.*, 6, 147–183.

Epstein, E. S. (1969), "Stochastic Dynamic Prediction," *Tellus*, 21, 739–759.

Gneiting, T. (1999), "Correlation Functions for Atmospheric Data Analysis," *Quarterly Journal of the Royal Meteorological Society*, 125, 2449–2464.

Grimit, E. P., and Mass, C. F. (2002), "Initial Results of a Mesoscale Short-Range Ensemble Forecasting System Over the Pacific Northwest," *Weather and Forecasting*, 17, 192–205.

Hamill, T. M., and Colucci, S. J. (1997), "Verification of Eta-RSM Short-Range Ensemble Forecasts," *Monthly Weather Review*, 125, 1312–1327.

Hamill, T. M., Snyder, C., and Morss, R. E. (2000), " A Comparison of Probabilistic Forecasts From Bred, Singular-Vector, and Perturbed Observation Ensembles," *Monthly Weather Review*, 128, 1835–1851.

Hoeting, J. A., Madigan, D. M., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial" (with discussion), *Statistical Science*, 14, 382–401.

Hou, D., Kalnay, E., and Droegemeier, K. K. (2001), "Objective Verification of the SAMEX'98 Ensemble Forecast," *Monthly Weather Review*, 129, 73–91.

Houtekamer, P. L., and Mitchell, H. L. (1998), "Data Assimilation Using an Ensemble Kalman Filter Technique," *Monthly Weather Review*, 126, 796–811.

——— (2001), "A Sequential Ensemble Kalman Filter for Atmospheric Data Assimilation," *Monthly Weather Review*, 129, 123–137.

Houtekamer, P. L., Lefaivre, L., Derome, J., Ritchie, H., and Mitchell, H. L. (1996), "A System Simulation Approach to Ensemble Prediction," *Monthly Weather Review*, 124, 1225–1242.

Kennedy, M. C., and O'Hagan, A. (2001), "Bayesian Calibration of Computer Models" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 63, 450–464.

Leith, C. E. (1974), "Theoretical Skill of Monte-Carlo Forecasts," *Monthly Weather Review*, 102, 409–418.

Molteni, F., Buizza, R., Palmer, T. N., and Petroliagis, T. (1996), "The ECMWF Ensemble System: Methodology and Validation," *Quarterly Journal of the Royal Meteorological Society*, 122, 73–119.

Murphy, A. H., and Winkler, R. L. (1979), "Probabilistic Temperature Forecasts: The Case for an Operational Program," *Bulletin of the American Meteorological Society*, 60, 12–19.

Nott, D. J., Dunsmuir, W. T. M., Kohn, R., and Woodcock, F. (2001), "Statistical Correction of a Deterministic Numerical Weather Prediction Model," *Journal of the American Statistical Association*, 96, 794–804.

Nychka, D. (2003), "The Fields Package," reference manual, available at *http://www.cran.r-project.org/src/contrib/PACKAGES.html*.

Palmer, T. N. (2000), "Predicting Uncertainty in Forecasts of Weather and Climate," *Reports on Progress in Physics*, 63, 71–116.

Raftery, A. E., Balabdaoui, F., Gneiting, T., and Polakowski, M. (2003), "Using Bayesian Model Averaging to Calibrate Forecast Ensembles," Technical Report 440, University of Washington, Dept. of Statistics.

Ribeiro, P. J., and Diggle, P. J. (2001), "geoR: A Package for Geostatistical Analysis," *R News*, 1(2), 14–18.

Richardson, L. F. (1922), *Weather Prediction by Numerical Process*, Cambridge, U.K.: Cambridge University Press.

Roulston, M. S., and Smith, L. A. (2002), "Combining Dynamical and Statistical Ensembles," *Tellus*, 55A, 16–30.

Schlather, M. (2001), "Simulation and Analysis of Random Fields," *R News*, 1(2), 18–20.

Stensrud, D. J., and Fritsch, J. M. (1994a), "Mesoscale Convective Systems in Weakly Forced Large-Scale Environments. Part II: Generation of a Mesoscale Initial Condition," *Monthly Weather Review*, 122, 2068–2083.

——— (1994b), "Mesoscale Convective Systems in Weakly Forced Large-Scale Environments. Part III: Numerical Simulations and Implications for Operational Forecasting," *Monthly Weather Review*, 122, 2084–2104.

Talagrand, O., Vautard, R., and Strauss, B. (1997), "Evaluation of Probabilistic Prediction Systems," in *Proceedings of the Workshop on Predictability*, Reading, U.K.: European Centre for Medium-Range Weather Forecasts, pp. 1–25.

Toth, Z., and Kalnay, E. (1993), "Ensemble Forecasting at the NMC: The Generation of Perturbations," *Bulletin of the American Meteorological Society*, 74, 2317–2330.

Wandishin, M. S., Mullen, S. L., Stensrud, D. J., and Brooks, H. E. (2001), "Evaluation of a Short Range Multimodel Ensemble System," *Monthly Weather Review*, 129, 729–747.

Wilks, D. S. (1995), *Statistical Methods in the Atmospheric Sciences*, San Diego: Academic Press.

Wood, A. T. A., and Chan, G. (1994), "Simulation of Stationary Gaussian Processes in $[0, 1]^d$," *Journal of Computational and Graphical Statistics*, 3, 409–432.

Comment

Claudia TEBALDI and Doug NYCHKA

Forecasting the weather presents a unique context for statistics, blending physical modeling with complicated observational data to produce information that is used at many different levels of sophistication. We are pleased that Gel, Raftery, and Gneiting (GRG) have brought this area to the attention of a statistical audience. In this discussion we give the reader a broader view of the use of *ensemble* techniques in numerical weather prediction (NWP). We have some comments about the use of ensembles idea presented by GRG and also present some of our recent analysis of the value of ensemble forecasts.

1. THE VALUE OF A FORECAST AND QUANTIFYING FORECAST SKILL

Weather forecasts have many users and, of course, the value and form of a forecast may depend on its intended purpose. Perhaps the most common use of a forecast is the estimate, say maximum surface temperature for a point location and a companion measure of uncertainty (e.g., Nychka's daughter asks him each morning what the temperature will be in Boulder so she can choose her outfit for school; she then asks him if he is "sure" about the forecast). In contrast, the geostatistical output perturbation (GOP) method goes beyond point forecasts using representations of the spatial covariance of the forecast accuracy to yield an ensemble of meteorological fields. The variability about the mean surface quantifies the uncertainty. Although this gives a significantly richer inference concerning the forecast, we also contend that it targets a sophisticated consumer.

To illustrate the distinction between point forecasts versus ensembles of fields, consider the following example. The Colorado Department of Transportation must make a decision whether to salt a highway to prevent icing. This decision is based on whether at any point along the highway the temperatures will dip below freezing. Thus, in statistical language the inference is whether the minimum of the field over a particular domain (the highway) has a high probability of being below freezing. To our mind, GRG give an elegant solution to this problem. For each ensemble field, the minimum temperature along the route of the highway must be found. The result is an empirical distribution of minimum temperatures that attempts to incorporate the spatial dependence of errors in the field and so may be more accurate in assessing the potential for icing. We are not sure how a correct inference would be drawn from just point forecasts of temperature with accompanying standard errors, so GRG's approach seems particularly useful in this context.

It is not clear that the man on the street or the forecaster on the evening news can interpret ensembles of fields and draw straightforward conclusions on the confidence he or she has in the forecast. In this respect, we question the need for a cultural change in forecaster attitude toward realizations of the GOP method. Based on the preceding example, it may be that specific applications of the forecast will benefit from ensemble fields, but in many cases a pointwise assessment of a best guess plus or minus a range of uncertainty, a simple probability density function, or a number between 0 and 1 that characterizes the degree of confidence in the forecast will do. Accordingly, in the last part of this discussion we focus more on the problem of obtaining more accurate inferences for point forecasts.

1.1 Ensemble Forecasting

A statistician can think of an ensemble as a discrete sample whose empirical distribution approximates a continuous distribution of interest. An idealized ensemble is a random sample from the posterior distribution for the state of the atmosphere given all past data and incorporating all known physical models of the flow.

Let $\mathbf{x}_t$ denote the vector of meteorological variables on a spatial grid that describe the state of the atmosphere at time t. The entire physical and geographic knowledge of the atmosphere's dynamical behavior can be subsumed by a function g, the NWP model, such that

$$\mathbf{x}_{t+1} = g(\mathbf{x}_t).$$

One way to make a forecast is to take the best estimate of the atmosphere's state, say $\hat{\mathbf{x}}_t$, and apply g. In atmospheric science, significant intellectual and computing resources have been aimed at constructing the closest approximation to the actual trajectory of the atmospheric state vector (in terms of an NWP model g) and generating the most realistic spread around it (in terms of ensemble members). Two factors contribute to the difficulty of this enterprise: uncertainty in initial conditions and model error. Referring to the notation, these two factors are

Claudia Tebaldi is Project Scientist (E-mail: *tebaldi@ucar.edu*) and Doug Nychka is Senior Scientist (E-mail: *nychka@ucar.edu*), National Center for Atmospheric Research, Box 3000, Boulder, CO80307.

Journal of the American Statistical Association
September 2004, Vol. 99, No. 467, Applications and Case Studies
DOI 10.1198/016214504000000935

errors in $\hat{\mathbf{x}}_t$ and errors in g. Our view from a statistical perspective is that the atmospheric sciences community has devoted the best of their statistical and numerical analyses to the problem of characterizing the uncertainty in the initial conditions (Toth and Kalnay 1993, 1997; Molteni, Buizza, Palmer, and Petroliagis 1996; Buizza, Miller, and Palmer 1996; Mitchell and Houtekamer 2000). This research has produced sophisticated approaches for initializing the ensemble members by "interesting" or "effective" perturbations of the best guess $\hat{\mathbf{x}}_t$. On the other hand, the characterization of model error seldom has been undertaken (Orrell 2002; Smith 2000; Smith, Ziehmann, and Fraedrich 1999).

In view of these daunting gaps between our understanding of the atmospheric processes and their approximation through NWP models, it is perhaps reasonable to target ensemble forecasts to a lesser, but still valuable goal: providing a rough idea of the uncertainty around NWP's best guess. To this end, GRG offer an interesting perspective, which is a natural extension of previous activity to calibrate forecasts with observations. In the atmospheric sciences, this has had success for many years in the form of the model output statistics (MOS) technique (Glahn and Lowry 1972). However, to our knowledge, MOS has been carried out only location by location, that is to say, independently at each observing station, and separately for each weather variable of interest.

In carrying out a MOS analysis, there is an important benefit of ensemble methods that is not exploited in the GRG study, but should be mentioned. The map g taking the (discretized) atmosphere from time t to $t + 1$ is nonlinear and often can amplify small features. By applying g to *each* member of the ensemble, a new ensemble for time $t + 1$ is obained and the resulting spread includes the nonlinear amplification and distortion that are well known for geophysical fluids. These features are termed flow dependent because the particular transformations of g depend partly on the state $\mathbf{x}$. For some states, g is nearly linear, whereas for others it can be sensitive to small perturbations of $\mathbf{x}$. By initializing, at time t, the ensemble members in a way that accounts for the uncertainty in initial conditions, the resulting spread among members at time $t + 1$ includes the flow-dependent nature of the uncertainty and is a function of the large scale weather patterns at the time of initialization. Part of our work with the NEXTCAST system described briefly in the next section makes use of this uncertainty that is tied to the current state and the dynamical properties of the atmosphere.

2. ENSEMBLE SPREAD AND CONFIDENCE IN THE FORECAST

Here we present some current work on relating the ensemble spread to the actual error in a forecast, but for point locations. In the past, the failure of accounting for model errors besides those in the initial condition has hampered the production of ensembles whose spread is representative of the actual error. Our work is based on the observation that measures of spread of the ensembles are more useful when the ensemble is built by forecasts from *different* NWP models. This so-called poor-man's ensemble is readily available and is less costly than one derived by multiple runs of the same NWP model run under perturbed initial conditions.

A poor-man ensemble in concert with extensive statistical postprocessing is the heart of the NEXTCAST forecasting system, under development by the Research Applications Program division of the National Center for Atmospheric Research (Mahoney 2001a,b). This system provides automatic, continuously updated, timely forecasts of many weather parameters (e.g., temperature; probability, phase and amount of precipitation; fraction of cloud cover; wind speed and direction; dewpoint temperature) at thousands of sites over the coterminous United States at lead times out to a few days.

NEXTCAST is a modular system, every module representing a—more or less—independent forecast, thus having the characteristics of a poor-man's ensemble. Different NWP models, statistical forecasts, climatology, and persistence are combined to produce the NEXTCAST ensemble. The final product is a weighted average of the single forecasts, whose weights depend on the recent relative performances of the single modules. Although spatial coherence of the final station forecasts is not enforced directly, the derivation from spatially coherent single forecasts suggests that some degree of spatial cohesion will be observed. Forecasts at points in between stations are inferred by simple bilinear interpolation of the anomalies with respect to a 30-year climatology. This part of NEXCAST can be interpreted as a fairly sophisticated MOS exercise, but is nonlinear and based on a relatively short and continuously updated time window. To this extent, it is more complicated than the linear bias adjustment made by GRG.

The users of these forecasts (such as engineers at the Department of Transportation in several states that are testing a version of NEXTCAST for road weather applications) expressed interest in a measure of confidence to be attached to each forecast at the time of issue. We exploited the property that the NEXTCAST ensemble spread exhibits a robust relationship with the size and distributional properties of the actual forecast error.

We considered pairs of spread measure (mean standard deviation among the NEXTCAST modules) and forecast error, collected over a thinned network of sites, over many days sparsely sampled between September 2001 and May 2002, and different lead times ranging from 12 to 84 hours. Figure 1 shows the quantiles of the error distribution as a function of ensemble spread values for some of the meteorological variables forecasted. Larger values of spread, signaling disagreement among models and usually associated with synoptic conditions harder to forecast, are associated with error distributions that are more diffuse and shifted toward larger values. Conversely, smaller values of spread, indicative of agreement among models, are usually associated with easier to forecast synoptic conditions, thus with tighter error distributions concentrated on smaller values. It is possible to fit parametric distributions to the errors stratified by spread values, and the gamma family gives a good approximation when fitting errors in absolute value.

This solution is tailored to specific locations, parameters, and seasons. It provides an answer to the question of forecast accuracy, in a way that does preserve the information of flow dependency, under the assumption that the ensemble

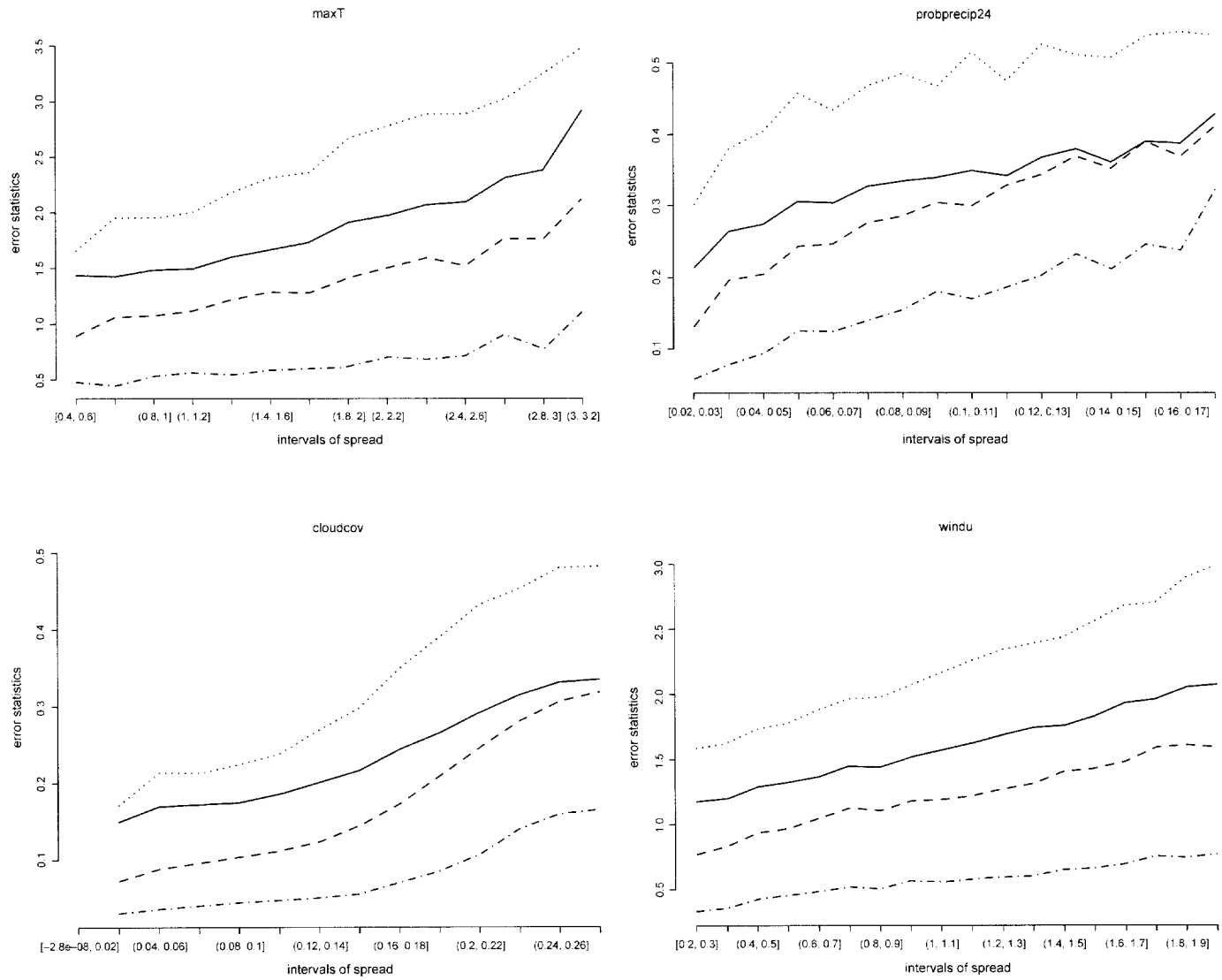

Figure 1. Relationship Between Spread of the Ensemble and Quantiles of the Forecast Error Distribution for Maximum Temperature, Probability of Precipitation, Cloud Cover, and u-Component of the Wind. The four lines correspond to first quantile (- - -), median (— —), mean (——), and third quantile (·····) of the error distribution. Along the x axis are binned values of the ensemble spread.

spread is a surrogate of such information. At least in the case of multimodel ensembles, the evidence is in favor of this claim (Ziehmann 2000). However, compared to GRG's GOP method, it cannot provide a spatially coherent picture of error covariance.

3. SUMMARY

The authors have applied elegant statistical methods to the area of ensemble forecasting and, thus, have brought it into the spotlight for the larger statistical community. Although the GOP method could be improved through the use of the richer physical content of real ensembles, we also believe that the information that the GOP ensemble delivers has a complementary value, taking the traditional MOS approach one large step further by providing a spatially coherent forecast calibration.

So we conclude with a suggestion; apply the GOP to the single members of a multimodel ensemble or estimate the GOP for a representative best member of a single model ensemble (as in Roulston and Smith 2002) and then perturb the whole set of members. This approach may embody the best of a dynamical *and* statistical treatment of the uncertainties at the roots of NWP's challenge.

ADDITIONAL REFERENCES

Buizza, R., Miller, M., and Palmer, T. N. (1999), "Stochastic Simulation of Model Uncertainty in the ECMWF Ensemble Prediction System," *Quarterly Journal of the Royal Meteorological Society*, 125, 2887–2908.

Glahn, H. R., and Lowry, D. A. (1972), "The Use of Model Output Statistics (MOS) in Objective Weather Forecasting," *Journal of Applied Meteorology*, 11, 1203–1211.

Mahoney, W. P. (2001a), "An Advanced Weather Information Decision Support System for Winter Road Maintenance," Proceedings of the Eight World Congress on Intelligent Transport Systems, 30 September–4 October 2001, Sydney, Australia.

——— (2001b), "An Advanced Winter Road Maintenance Decision Support System," Proceedings of the Intelligent Transportation Society of America Conference, 4–7 June 2001, Miami Beach, FL.

Mitchell, H. L., and Houtekamer, P. L. (2000), "An Adaptive Ensemble Kalman Filter," *Monthly Weather Review*, 128, 416–433.
Orrell, D. (2002), "Model Error in Weather Forecasting. Does Chaos Matter?" Available at *http://www.beatrizl.freeserve.co.uk/AGUposter.htm*.
Smith, L. A. (2000), "Disentangling Uncertainty and Error: On the Predictability of Nonlinear Systems," in *Nonlinear Dynamics and Statistics*, ed. A. Mees, Boston: Birkhauser.
Smith, L. A., Ziehmann, C., and Fraedrich, K. (1999), "Uncertainty Dynamics and Predictability in Chaotic Systems," *Quarterly Journal of the Royal Meteorological Society*, 125, 2855–2886.
Toth, Z., and Kalnay, E. (1997), "Ensemble Forecasting at NCEP and the Breeding Method," *Monthly Weather Review*, 125, 3297–3319.
Ziehmann C. (2000), "Comparison of a Single-Model EPS With a Multi-Model Ensemble Consisting of a Few Operational Models," *Tellus*, 52A, 280–299.

Comment

William BRIGGS

INTRODUCTION

Anything that encourages the use of probability forecasts in meteorology should be applauded. The authors' geostatistical output perturbation (GOP) method does this in a clever and computationally simple way that is somewhat similar in concept to model output statistics (MOS) for dynamical forecasts. The GOP method produces bias-corrected probability forecasts, not just bias-corrected point forecasts as MOS does, and so has the potential to be a superior approach. The GOP method also reinforces the notion that dynamical forecasts are not certain and that the variability in the output is important to understand.

ENSEMBLE FORECASTS

The GOP method takes the result of a dynamic field forecast, corrects its biases by formula, and then generates suites of forecast maps to form, in essence, a probability forecast of, say, a temperature field.

Ensemble forecasts work oppositely by perturbing the initial conditions or the parameterization of the dynamic model, running the model for each set of initial conditions or parameterizations, and gathering the end results to form a suite of dynamic field forecasts. This suite of individual forecasts must then be transformed to a single probability forecast by some other means. How to do this well is an open problem and is an area in which the authors of this article also work.

The true ensemble forecast should be a stronger forecast than that produced by the GOP method because, if done properly, the ensemble forecast samples from the whole of different possible future states of the atmosphere. Experience has shown that these future states can be dramatically different from one another and that the variability of the states is important for the forecast. The best method of perturbing the initial conditions is unsettled, and we have only begun to explore methods and the importance of what is called stochastic parameterization.

The GOP method takes only one possible future state from one dynamical model run and uses it to generate the forecast. The authors suggested some possibilities for combining ensemble forecasts and the GOP method, which is an area I hope they will pursue, because it is there that the GOP method will meet its greatest success.

USING THE OUTPUT

The authors rightly emphasize that standard weather maps of, say, surface temperature, are too smooth, which might lead meteorologists to subjectively underestimate the variability of the field forecast. This, in turn, might cause them to issue forecasts that are too certain. The GOP method rightly emphasizes this uncertainty, which is necessary because, unfortunately, uncertainty is only partially expressed in National Weather Service forecasts issued to the public; many meteorologists, for example, still give just one number for the high or low temperature forecasts. Forecasts like those produced by ensembles and statistical models like the GOP method will bring the realization that forecasts are certainly not certain and should be qualified with some kind of probability information.

The maps from the GOP emphasize the choppiness of the field, showing that spatial variability is far rougher than conventional maps. The authors now have to turn this idea into something that is useful operationally. So the big question is, "How many maps do you show the forecaster?" It is not clear that more is better, at least for human-issued forecasts. The amount of extra detail in the rough field is more accurate, but it may be more confusing and harder to assimilate, and could lead to worse forecasts. Some form of data compression will probably be needed.

Marginal density estimates or histograms of the variables of interest for specific locations culled from the GOP method members could easily be built. These would, of course, lose the spatial uncertainty inherent in the forecast, but would be easy to understand for the location at hand.

There are some ways to keep the spatial uncertainty while still reducing the overall complexity of the suite of maps. Spaghetti plots are one way. These are usually built from ensemble forecasts and are contour maps of, say, a temperature of 0°C of each ensemble member. The spread of the contours indicates the certainty of that contour level: tighter grouping implies greater confidence than looser grouping. These spreads could be used just as well with GOP method members. The same goes for mean maps (mean of the GOP method members) and variance maps, which are also standard ensemble picturing tools.

William Briggs is Assistant Professor, General Internal Medicine, Weill Medical College of Cornell University, New York, NY 10021 (E-mail: *wib2004@med.cornell.edu*).

Journal of the American Statistical Association
September 2004, Vol. 99, No. 467, Applications and Case Studies
DOI 10.1198/016214504000000944

VERIFICATION

It is easy to verify, that is, assess the accuracy of, a point forecast. If the forecast said that high would be 72 and it was actually 80, then it is simple to say "8 degrees off." However, how to verify field-point forecasts is still an open question. Field-point forecasts are the ordinary output from dynamical models (single numbers at each grid point). There are few quality methods or measures to do this; the anomaly correlation score is one measure, but it is poor in the sense that it boils down the entire model and observation field to one number, a number whose qualities are suspect and poorly understood.

Verifying ensemble or GOP suites is more difficult still. A tool that was developed, and used in this article is the rank-histogram plot, which is very nice, but, of course, does not allow the modeler to gain insight into where and why a particular set of forecasts succeeded or failed, especially spatially. Maybe individual rank histograms over subgrids could be displayed on a map to see if there are spatially varying differences in forecast quality. Alternatively, perhaps contour maps of a measure of departure from uniformity at each grid point could be displayed (rank histograms at each grid point would be impossible to see clearly).

Another curiosity is how to verify points that are forecasts for which there are no observations. The conventional approach is to convert the irregularly-spaced observations to the same grid as the forecast and then do verification.

I believe that models can only be ultimately evaluated at the points of the observation and nowhere else. Model grid points will not, of course, exactly correspond to observation location, and some interpolation from the model grid to the observations has to be done. This appears to be what the authors have done, but this results in an estimate of the model at those locations, not the actual model value. It is important to consider the uncertainty of these estimates in the eventual verification. It may be small, but even small errors can grow large when you add them across multiple ensemble members.

What should not be done, although it usually is, is to extrapolate the observations to the model grid and to compare the model and extrapolated observations on that entire grid, usually by producing an "analysis" of observations. This is somewhat fair in that we compare the forecast model with observations in a modellike sense because of the analysis. However, an analysis is not what the end user of a forecast sees. He or she gets the actual atmosphere, so the fairest verification method is to compare the actual forecast field with the actual observations.

There can be large gaps in observations, particularly for upper-level data, and extrapolating the observations to these gaps is not fair because the verification measure depends heavily on the extrapolation method used. The example I have in mind is the modeler who wants to know how far the model placement of an upper-level low pressure center is from the truth where there are no observations nearby the predicted center. Extrapolating or smoothing the observations to estimate where the low actually was is not wise because we never have actual data to see where the low actually was or even if it actually existed. In addition, we never know how well the extrapolation method truly performs. Our time would be better spent by trying to get more observations.

CONCLUSION

This is a fun and useful paper which was a pleasure to review.

Comment

Mark S. ROULSTON

The method proposed by Gel, Raftery, and Gneiting for generating probabilistic mesoscale forecasts is a relatively simple idea, but the authors have had to confront the "curse of dimensionality," which is a common feature of atmospheric modeling. It is this high dimensionality that often prevents statistical ideas from being applied to meteorological problems. Statistically modeling the forecast errors in two dimensions and then perturbing the output of numerical weather prediction (NWP) models is computationally far cheaper than the current methods of producing dynamical ensembles by running the NWP model multiple times. The authors suggest that the geostatistical output perturbation (GOP) method might be a viable alternative to dynamical ensemble forecasting for organizations that do not have the resources to initialize and run an NWP more than once per forecast cycle. I suggest, however, that GOP-type methods should be the benchmark against which dynamical ensemble prediction systems are evaluated. It would be interesting to see a comparison of the GOP method with dynamical ensembles at both the mesoscale and on the global scale where dynamical ensembles have become standard operational tools for forecasters. Questions such as whether a high resolution model combined with GOP is better than a dynamical ensemble of lower resolution model runs (requiring the same amount of computer time) could be addressed. Also, if the authors improve the GOP method by including information from conventional, dynamical ensembles, as they suggest might be possible, the trade-off between dynamical ensemble size and NWP model resolution could be investigated.

As the authors point out, the GOP method could be extended to include temporal correlations. A nonisotropic structure of spatial correlations is also worth investigating because the zonal (east–west) direction is the direction of prevailing flow and it seems quite likely that the error structure in this direction could differ from that in the meridional (north–south) direction. Researchers in the statistical community may suggest alternative

Mark S. Roulston is Assistant Professor, Department of Meteorology, The Pennsylvania State University, University Park, PA 16802 (E-mail: *roulston@met.psu.edu*).

Journal of the American Statistical Association
September 2004, Vol. 99, No. 467, Applications and Case Studies
DOI 10.1198/016214504000000926

ways to model the space–time correlation structure of NWP model errors. It is important to emphasize though, that in the meteorological community, the idea that *something* along the lines suggested by Gel. et al. should be done is not widely appreciated.

Finally, although the GOP method is conceptually straightforward, the high dimensionality of the problem has required the authors to use some advanced numerical techniques that will be unfamiliar to most meteorologists and other researchers who use mesoscale models. The authors should consider developing a software tool that implements the GOP method and that could be used in conjunction with the MM5 model and its successors.

Rejoinder

Yulia GEL, Adrian E. RAFTERY, Tilmann GNEITING, and Veronica J. BERROCAL

The past decade has seen a culture change in the practice of numerical weather prediction. Up to the early 1990s, numerical weather forecasting was an intrinsically deterministic endeavor. National and international weather centers used sophisticated computing resources to run carefully designed numerical weather prediction models. This is still the case today; however, as Hamill, Hansen, Mullen, and Snyder (2004) pointed out, "the most radical change to numerical weather prediction during the last decade has been the operational implementation of ensemble forecast methods." Ensemble forecasts seek to assess the uncertainty of the predictions, and methods of probabilistic numerical weather forecasting are now in vigorous development. Yet, probabilistic weather forecasting has largely bypassed the attention of the statistical community, with few, but notable exceptions, including Nychka (2000) and Gustafsson (2002). We thank the editor for bringing this exciting field to the attention of statisticians.

We are very grateful to the discussants for their insightful comments, which point to important future directions for research in this area. Key points raised are the connection between the geostatistical output perturbation (GOP) method and dynamical forecast ensembles, and the possibility of combining the two approaches; visualization, that is, what we should display and how we should do it; how to verify probabilistic forecasts of entire fields; and specification of the spatial correlation function.

1. THE GOP METHOD AND OTHER ENSEMBLE APPROACHES

All three discussions compared the GOP method with dynamical ensemble methods, and suggested that a combination of the two approaches would be fruitful. We strongly agree. Dynamical ensemble methods generate an ensemble of initial conditions and run the numerical weather prediction model forward from each of them in turn, whereas the GOP method instead perturbs the model output rather than its input. Dynamical ensembles have the advantage, pointed out by Tebaldi and Nychka, that they can capture nonlinear aspects of forecast uncertainty, but they typically require considerable resources in terms of data, data assimilation software, and computing power. The GOP method, on the other hand, is much faster and does not require any data beyond the deterministic forecast once it has been trained using historical data. We, therefore, endorse Roulston's suggestion that the GOP method be used as a benchmark for other ensemble methods; in this view, outperforming the relatively simple and cheap GOP method would be a minimal requirement for other more complex and costly ensemble methods.

The strengths of GOP and dynamical ensembles seem complementary, so combining them indeed seems like a good idea. We have been working on one approach to this. It starts with a Bayesian model averaging (BMA) approach to calibrating dynamical ensembles for forecasts at one point in space (Raftery et al. 2003). This generates a (univariate) predictive distribution that is a finite mixture of distributions, each one of which is centered around one of the (bias-corrected) forecasts in the ensemble. The mixture weights and distribution parameters are estimated from recent forecasts and observations. In experiments, it gave calibrated and sharp predictive distributions, and honored the observed correlation between absolute forecast errors and ensemble spread mentioned by Tebaldi and Nychka. Indeed it could be viewed as a different way to implement the idea mentioned by Tebaldi and Nychka in section 2 of their discussion. As Tebaldi and Nychka suggest, the BMA approach is related to the dressing method of Roulston and Smith (2002).

Our approach to combining the GOP method with dynamical ensembles starts by estimating a GOP model for each member of the ensemble. Weights and forecast variances for each ensemble member are then estimated using the BMA approach. The GOP spatial covariance function for each ensemble member is scaled using the estimated forecast variance for that member. Finally, several realizations are simulated from the GOP model that correspond to each of the ensemble members, with the number of realizations proportional to the weight for the ensemble member. We are currently implementing and evaluating this approach, which we call *Bayesian dressing*.

The ensemble model output statistics (EMOS) approach of Gneiting, Westveld, Raftery, and Goldman (2004) provides another option for ensemble postprocessing. The EMOS method fits a Gaussian predictive probability density function to ensemble output. The EMOS predictive mean is an optimal, bias-corrected weighted average of the ensemble member forecasts, with weights that are constrained to be nonnegative and associated with the skill of the ensemble member. The EMOS predictive variance is a linear function of the ensemble spread. In the EMOS–GOP approach, we perturb the EMOS predictive mean by simulated, spatially correlated error fields.

Journal of the American Statistical Association
September 2004, Vol. 99, No. 467, Applications and Case Studies
DOI 10.1198/016214504000000953

Bayesian dressing is a more principled and more elegant approach than EMOS–GOP. Indeed, EMOS can be viewed as a linear approximation to BMA, somewhat like Bayes linear methods provide an approximation to fully Bayesian procedures (Goldstein 1999). However, performance is an empirical question, and it remains to be seen which method performs best in the sense of maximizing the sharpness of the predictive distributions under the constraint of calibration.

2. VISUALIZATION

Tebaldi and Nychka give a wonderful example of a task for which the GOP method can be useful for sophisticated users, namely the decision whether to salt a highway to prevent freezing. They also suggest that GOP may not be so useful for less sophisticated users. We feel, however, that this is a matter of what summary of the forecast distribution to communicate and display, which should depend on the end use. If the right summary is chosen, it can be computed from the GOP output and provided to the user.

They give the example of Nychka's daughter asking him every day what the temperature will be in Boulder and then asking him if he is sure of his forecast. On a given day, he might say that it will be 68°F, but that he is not very sure. We suggest that a statement that it will probably be between 63 and 73°F could be at least as useful in helping Nychka's daughter choose her outfit. It could be understood between them, for example, that this means that there is 1 chance in 10 that the maximum temperature at her high school during the school day will be below 63°F and 1 chance in 10 that it will be above 73°F. This kind of statement is an immediate by-product of the GOP method and can easily be derived from its output and displayed. In this way, the GOP method can serve the needs of less sophisticated users too, provided that the right summaries of the realizations are displayed. Incidentally, there is some evidence that when such statements are given in terms of natural frequencies (1 chance in 10), users find them easier to interpret than when they are given in probabilities (10%; Hoffrage, Lindsey, Hertwig, and Gigerenzer 2000).

This leads to the more general question of what should be displayed and how. Briggs has provided a very insightful discussion. Nychka's daughter's hypothetical question and other similar ones can be answered by mapping summaries of univariate point probabilistic forecasts. For example, one might show maps of the median of the pointwise forecast distribution, and of the 10th and 90th percentiles; an example of this was given by Raftery et al. (2003). Nychka's daughter could read the answer to her question directly off such a map, as could other Colorado residents with less expert fathers! It is hard to see what direct use such users would make of statements of uncertainty as opposed to probabilities, but one could also show a map of a "margin of error," such as half the difference between the 10th and 90th percentiles, as a measure of uncertainty.

There are various ways to display and summarize the approximate posterior predictive distribution of the future weather field provided by the GOP realizations. Briggs has suggested spaghetti plots, and this is a very good idea, for which taking account of spatial correlation is vital. One could also show *stamp plots*, that is, simultaneous displays of several realizations arranged, for example, in a square 2×2, 3×3, or 4×4 array. The University of Washington MM5 ensemble (*http://www.atmos.washington.edu/~ens/view_uwme.cgi*) provides displays of this kind in a 3×3 format for a dynamical ensemble, and they have been found useful by forecasters in the Pacific Northwest region.

The question of which displays to provide is vital, as Briggs points out, and one to which statisticians have not yet given much attention. Such questions are essentially cognitive questions, and we are currently working with cognitive psychologists Earl Hunt and Susan Joslyn at the University of Washington to carry out experiments to assess the relative effectiveness of different ways to display this kind of probabilistic information.

3. FORECAST VERIFICATION

Briggs takes an exceptionally clear standpoint in the current debate on forecast verification in the atmospheric sciences. Should numerical weather prediction models be assessed by interpolating gridded model output to the observation locations or by interpolating the observations to the model grid? Briggs dismisses the latter approach and his argument is well taken. Interpolation from scattered observation locations to the model grid is frequently extrapolation; hence, the verification measure depends heavily on the extrapolation method used. We agree, and in ongoing joint work with Eric Grimit and Clifford Mass, we strive to quantify the effect. In contrast, interpolation from the model grid to scattered observation locations is straightforward. All but the most obscure interpolation techniques will yield similar results, thereby robustifying the verification approach.

4. SPATIAL CORRELATION

Roulston suggests that forecast error spatial correlations in the zonal (east–west) direction of prevailing atmospheric flow might differ from those in the meridional (north–south) direction. This is indeed a real possibility, and if it is the case, it should be taken into account in the modeling. The directional variograms in Figure 1 suggest, however, that for our data such differences are small, if indeed they exist at all. However, it is

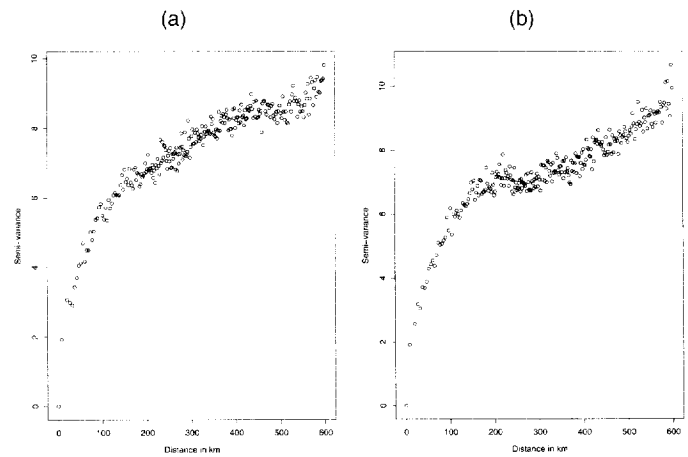

Figure 1. Directional Variograms for Temperature Forecast Errors in the North American Pacific Northwest, January–June, 2000: (a) North–South; (b) East–West.

quite possible that such differences do exist for other meteorological variables and regions, in which case a model that takes account of them should be considered.

It is not clear how much impact such differences would have on the performance of probabilistic forecasts. In a different meteorological context, Haslett and Raftery (1989) analyzed wind speed data where there was evidence of anisotropy (Guttorp and Sampson 1989). Nevertheless they used an istropic model, because it turned out that the anisotropic approach did not yield better performance in terms of the main goal of their study, namely the assessment of wind power at a new site. This suggests that it would be necessary to establish not only that such directional differences in spatial correlation exist, but that taking account of them is worth the increased effort and complication in terms of probabilistic weather forecasting performance.

5. SOFTWARE

Roulston suggests that we develop a software tool that implements the GOP method and could be used in conjunction with MM5. This is an excellent idea. We are currently developing an R package (tentatively named ProbForecastGOP) to do this, and we hope to make it publicly available soon at the Comprehensive R Archive Network at *http://lib.stat.cmu.edu/R/CRAN*.

ADDITIONAL REFERENCES

Gneiting, T., Westveld, A., Raftery, A. E., and Goldman, T. (2004), "Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation," Technical Report 449, University of Washington, Dept. Statistics. Available at *http://www.stat.washington.edu/www/research/reports*.

Goldstein, M. (1999), "Bayes Linear Analysis," in *Encyclopaedia of Statistical Sciences* (updated Vol. 3), eds. S. Kotz et al., New York: Wiley, pp. 29–34.

Gustafsson, N. (2002), "Statistical Issues in Weather Forecasting" (with discussion and reply), *Scandinavian Journal of Statistics*, 29, 219–243.

Guttorp, P., and Sampson, P. D. (1989), Discussion of "Space–Time Modelling With Long-Memory Dependence: Assessing Ireland's Wind Power Resource," by Haslett and Raftery, *Journal of the Royal Statistical Society*, Ser. C, 38, 32–33.

Hamill, T. M., Hansen, J. A., Mullen, S. L., and Snyder, C. (2004), "Meeting Summary: Workshop on Ensemble Forecasting in the Short to Medium Range," available at *http://www.cdc.noaa.gov/~hamill/EF_workshop_summary_25Jan.pdf*.

Haslett, J., and Raftery, A. E. (1989), "Space–Time Modelling With Long-Memory Dependence: Assessing Ireland's Wind Power Resource" (with discussion), *Journal of the Royal Statistical Society*, Ser. C, 38, 1–50.

Hoffrage, U., Lindsey, S., Hertwig, R., and Gigerenzer, G. (2000), "Communicating Statistical Information," *Science*, 290, 2261–2262.

Nychka, D. (2000), "Challenges in Understanding the Atmosphere," *Journal of the American Statistical Association*, 95, 972–975.