Calibrated Probabilistic Mesoscale Weather Field Forecasting: The Geostatistical Output Perturbation (GOP) Method [1]

Yulia Gel, Adrian E. Raftery and Tilmann Gneiting
University of Washington

Technical Report no. 427
Department of Statistics
University of Washington.

March 12, 2003

[1] Yulia Gel is Research Associate, Adrian E. Raftery is Professor of Statistics and Sociology, and Tilmann Gneiting is Assistant Professor, all at the Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195-4322. The authors are grateful to Mark Albright, Eric Grimit and Clifford Mass for helpful discussions and providing data. This research was supported by the DoD Multidisciplinary University Research Initiative (MURI) program administered by the Office of Naval Research under Grant N00014-01-10745.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE
12 MAR 2003

2. REPORT TYPE

3. DATES COVERED
00-03-2003 to 00-03-2003

4. TITLE AND SUBTITLE
Calibrated Probabilistic Mesoscale Weather Field Forecasting: The Geostatistical Output Perturbation (GOP) Method

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Washington,Department of Statistics,Box 354322,Seattle,WA,98195-4322

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**19**	

Abstract

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, often depending on data assimilation, 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 which 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 US Pacific Northwest in 2000 and 2002. The resulting forecast intervals turn out to be 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.

Contents

List of Figures

1 Introduction

In this paper, we propose a way of obtaining probabilistic mesoscale weather forecasts that are calibrated, sharp, and apply to whole weather fields simultaneously, rather than just 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 rather than a probability distribution. Mesoscale weather forecasts are local forecasts with resolutions on the order of 1–12 km, and typically cover areas on the order of 500–1000 kilometers square, compared with global and synoptic forecasts with 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 integrating the available information subjectively, using their professional experience. Bjerknes (1904) had 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 the first computers. The quality of numerical weather predictions improved steadily, and by about 1995 synoptic models consistently provided good point forecasts up to about three days ahead.

Up to about 1995, numerical weather forecasting was mostly done in practice on the 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 with considerable data and computing resources, including supercomputers. They then released their forecasts for public use. Local forecasters, such as those working for the media, aviation, shipping, and the military, would typically produce 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 ten 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, with the MM5 (NCAR–Penn State Mesoscale Model Generation 5) being 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 a handful of weather organizations worldwide a decade ago. Typically, they 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 solved 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, 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. By the 1990s three viable methods had been developed: the breeding growing modes method used by the US National Centers for Environmental Prediction (NCEP) (Toth and Kalnay 1993), the singular vector method used by the European Centre for Medium-Range Weather Forecasts (ECMRWF) (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) review 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 ten 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 these into initial conditions for the model. An organization running MM5 locally will typically not have access to either 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; Stensrud and Fritsch 1994b). Methods that perturb the initial conditions directly in a simple way are questionable, because the resulting sets of initial conditions will usually not be in thermal balance, 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 one did, but the prediction intervals produced were far too narrow.

We propose to develop an easy to use mesoscale probabilistic forecasting method by directly perturbing the model *output*, or point forecasts, in contrast with 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 which 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 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 of verifying the resulting model and forecasts. In Section 3 we apply the method to forecasting temperatures in the US Pacific Northwest and show the results. Finally, in Section 4 we discuss possible improvements to the methodology.

2 The Geostatistical Output Perturbation (GOP) 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 say 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 will 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, with the whole image being calibrated, rather than just the individual forecasts that make it up.

Let $X(s,t)$ be a finite set of variables corresponding to location s and time t, 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

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

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,

$$\frac{1}{2}\,\mathrm{Var}\left(w(s_1,t) - w(s_2,t)\right) = \rho + \sigma^2\left(1 - e^{-||s_1 - s_2||/r}\right) \tag{2}$$

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 0.05 at distance $3r$.

2.2 Parameter Estimation

We estimate the parameters of the model given by equations (1) and (2) using historical data on forecasts and observations. Typically we would use data for a relatively homogeneous region over a recent time interval of length on the order of three 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 would be possible to estimate the model using maximum likelihood, or a fully Bayesian approach, for example along the lines of Fuentes and Raftery (2002). Forecasts are on a grid and may correspond to a grid cell, while observations correspond to irregularly spaced locations, the so-called change of support problem, and the fully Bayesian approach has the advantage of being able to deal with with this explicitly in a coherent way.

However, the data sets used for parameter estimation are typically very large, and so full maximum likelihood estimation or fully Bayesian estimation tend 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 **a** and **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. Direct simulation from the very high-dimensional multivariate normal distribution is not feasible by standard techniques such as Cholesky decomposition of the covariance matrix, and 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, or 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, and 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, with all blocks being circulant themselves, and admitting 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 could 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 fifth and 95th order statistics as the endpoints, or we could use another number of simulations, depending on our computational resources. We will be interested in the average coverage and the average length of such intervals.

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 will do so in Section 3.4.

3 Results

We now apply the GOP method to some data on temperature in the US 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 US Pacific Northwest. Temperature was measured at 0 hours 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. 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 degrees Kelvin is equal to $(x - 273.15)$ degrees Centigrade. 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 we had observations 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 (NWS/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.

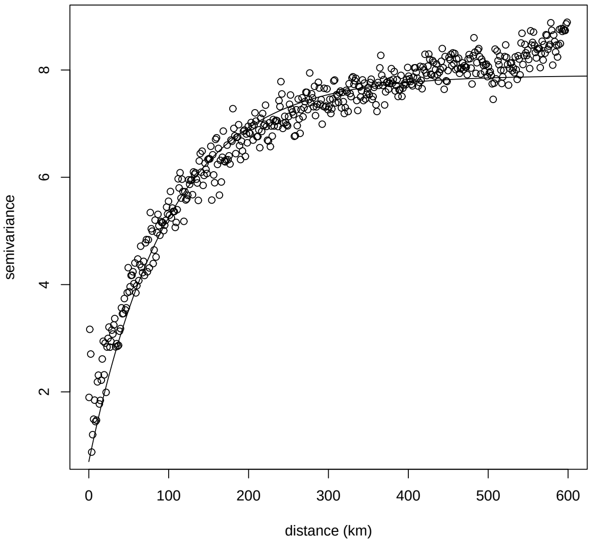
3.2 Parameter Estimation

In order 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 to 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 reads

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

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



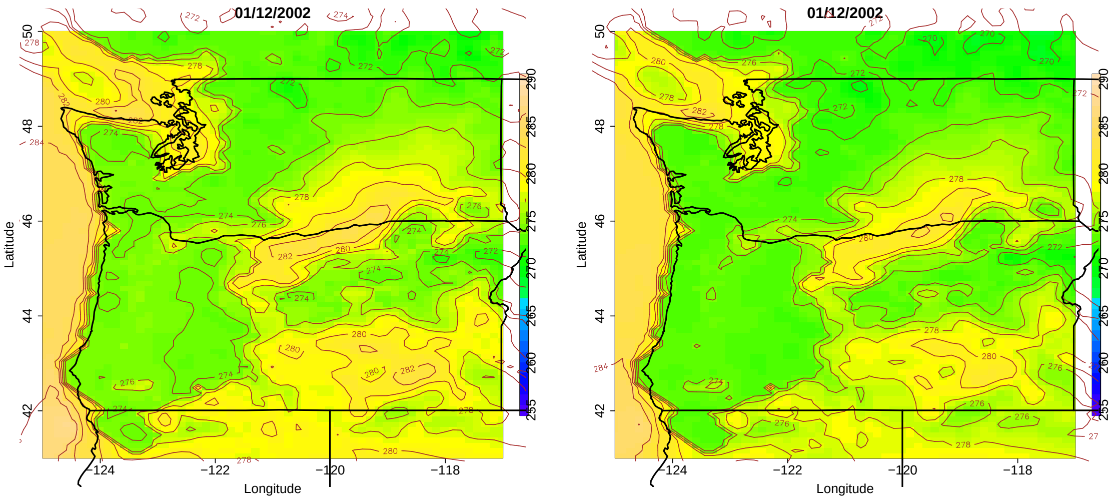
and the regression estimate of the additive bias was 1.6 (standard error 0.4), and that of the multiplicative bias parameter was 0.995 (0.002). Thus the additive bias was significant, but there was almost no multiplicative bias. The residuals from these regressions were computed, and their variogram is shown in Figure 1, together with the fitted exponential variogram function (2). The estimated nugget was $\rho = 0.51$, the estimated variance of the continuous spatial variation component was $\sigma^2 = 7.2$, and the estimated range parameter was $r = 114$km. The fit of the parametric variogram function is better than what is 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 US 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 a year and a half after the period to which the data used to estimate the model pertain, so it is a truly an out of sample forecast.

Figure 2 shows the gridded MM5 forecast, $\tilde{Y}$, produced by running MM5 initialized with

Figure 2: The Gridded MM5 48-hour Ahead Forecast of Temperature in the US Pacific Northwest Verifying on January 12, 2002 at 0 hours GMT, and the Bias-Adjusted Predictive Mean.



(a) The gridded MM5 forecast, $\tilde{Y}$.

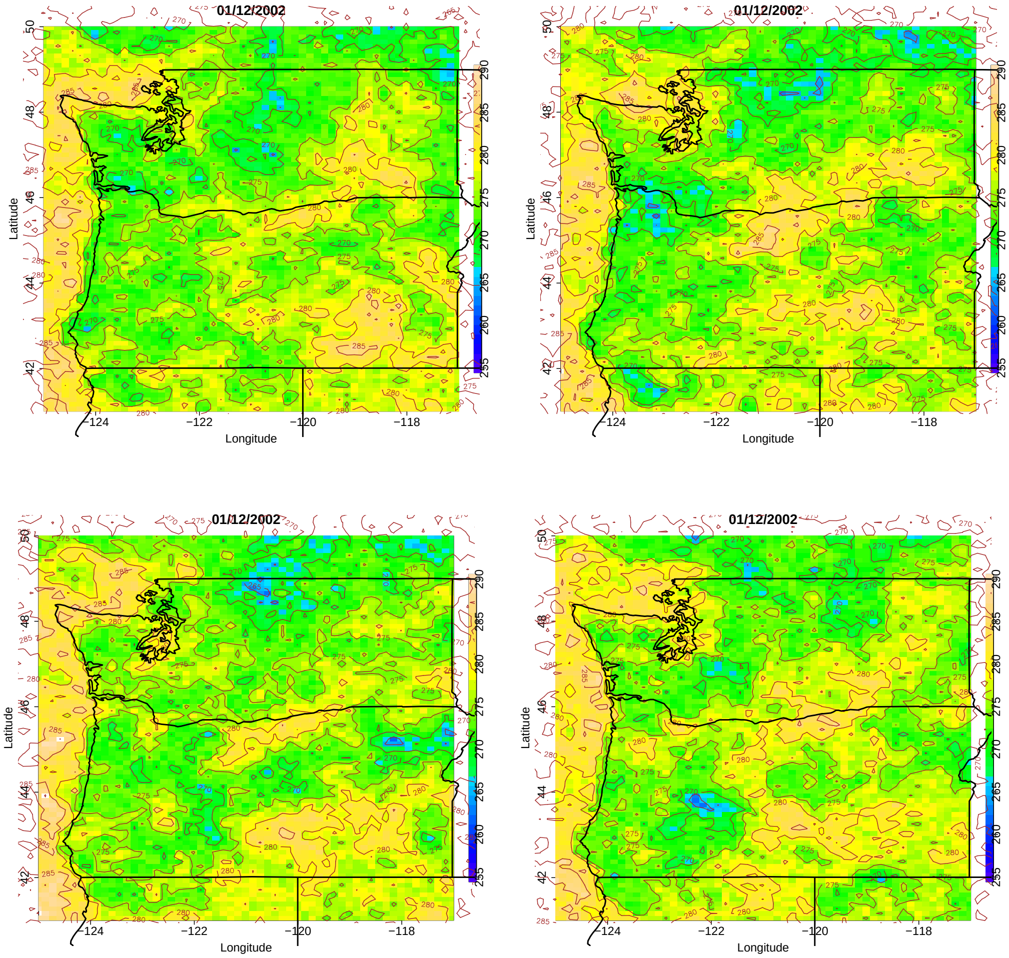
(b) The predictive mean, $a + b\tilde{Y}$.

the output from the synoptic NCEP Aviation model. This shows output on a 12km 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 12km grid. It is interesting to note that these 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 will typically be 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, while the observation locations are irregularly spaced, and much sparser. We show 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, and a visual comparison is misleading.

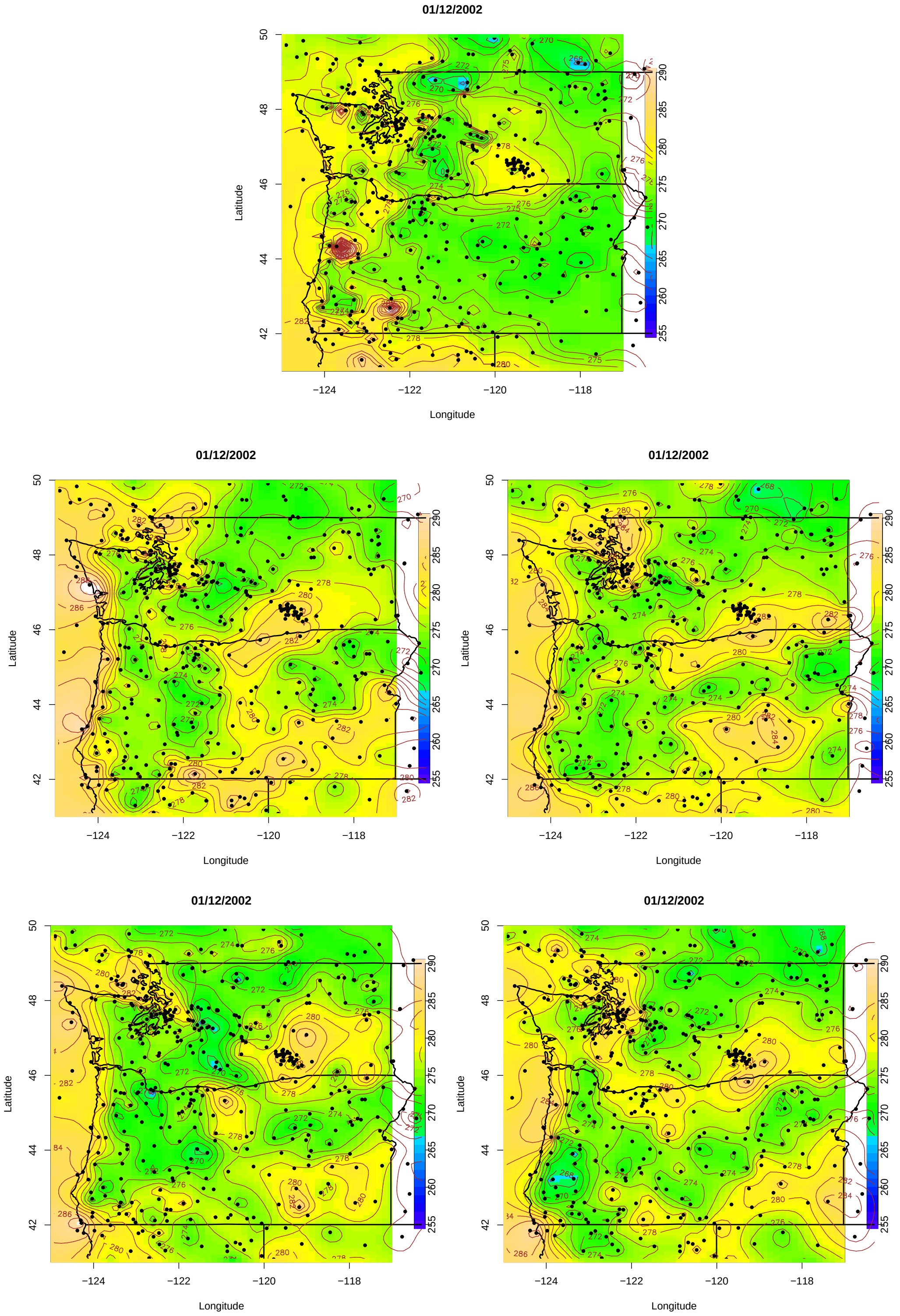
Figure 3: Ensemble of Forecasts for the Temperature on January 12, 2002 Using the Gridded MM5 Output.



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 bottom two rows show four ensemble plots, as compared to the actual observations on January 12, 2002 at 0 hours GMT (00Z), shown in the top row.

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

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



formed 90% prediction intervals by taking the fifth 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 0.66.

3.4 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: the 66.7% and 90% prediction intervals. The 66.7% interval contained the true value 68.1% of the time, and the 90% interval contained the true value 90.8% of the time. This indicates that the probability forecasts of individual temperatures were well calibrated.

To assess sharpness of prediction intervals, we computed the average length of the 90% prediction intervals. This was 9.4 degrees. For comparison, we computed the average difference between the fifth and 95th percentile of the marginal distribution of the values; this could be viewed as an approximation to the average length of a prediction interval based on a crude form of climatology. This was 28.3 degrees, so the prediction intervals from the GOP method were about 67% shorter, while remaining calibrated.

4 Discussion

We have proposed a method for mesoscale probabilistic forecasting, which is feasible for local users of MM5 and other mesoscale models. It breaks with many previous approaches by perturbing, not the inputs to the model, but 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 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, and so is smoother than data. This suggests that the adoption of calibrated probabilistic mesoscale forecasting may require something of a culture change, with forecasters getting used to looking at images that look like Figure 3, as well as the smoother point forecasts such as Figure 2 that they are used to looking at now.

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 have been known as "model output statistics" (Wilks 1995). A modern hierarchical Bayes approach is 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 six-month period in the US Pacific Northwest to fit the model. More research is needed to assess what the best temporal window would be, and to develop more systematic ways of deciding 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, and taking account of temporal autocorrelation, might all improve the results. These could all be done by taking a fully Bayesian approach using Markov chain Monte Carlo, and Fuentes and Raftery (2002) have described one way of doing this. 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.

One useful extension would be to take account of multiple models. Grimit and Mass (2002) have shown that there is 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 could be exploited to obtain a better assessment of the spread of the predictive distribution in the present context. This could be done, for example, by combining the ideas of Bayesian model averaging (Hoeting, Madigan, Raftery, and Volinsky 1999; Balabdaoui, Raftery, and Gneiting 2003) 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) temporally and spatially varying, depending on the spread between forecasts based on initial conditions provided by different weather centers. Another approach to exploiting the spread-skill relationship is suggested by Roulston and Smith (2002).

References

Balabdaoui, F., A. E. Raftery, and T. Gneiting (2003). Calibrated probabilistic weather forecasting using ensembles via Bayesian model averaging. *In preparation.*

Bjerknes, V. (1904). Das Problem von der Wettervorhersage, betrachtet vom Standpunkt der Mechanik und der Physik. *Meteorologische Zeitschrift 21*, 1–7.

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

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

Dietrich, C. R. and G. N. Newsam (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.

Fuentes, M. and A. E. Raftery (2002). Model validation and spatial interpolation by combining observations with outputs from numerical models via Bayesian melding. Technical Report 403, Department of Statistics, University of Washington.

Gneiting, T. (1999). Correlation functions for atmospheric data analysis. *Quarterly Journal of the Royal Meteorological Society 125*, 2449–2464.

Grimit, E. and C. Mass (2002). Initial results of a mesoscale short-range ensemble forecasting system over the Pacific Northwest. *Weather and Forecasting 17*, 192–205.

Hamill, T., C. Snyder, and R. E. Morss (2000). A comparison of probabilistic forecasts from bred, singular-vector, and perturbed observation ensembles. *Monthly Weather Review 128*, 1835–1851.

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

Hou, D., E. Kalnay, and K. K. Droegemeier (2001). Objective verification of the SAMEX'98 ensemble forecast. *Monthly Weather Review 129*, 73–91.

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell (1996). A system simulation approach to ensemble prediction. *Monthly Weather Review 124*, 1225–1242.

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

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

Nott, D. J., W. T. M. Dunsmuir, R. Kohn, and F. Woodcock (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, posted 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.

Ribeiro, P. J. and P. J. Diggle (2001). geoR: A package for geostatistical analysis. *R News 1(2)*, 14–18.

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

Roulston, M. S. and L. A. Smith (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 J. M. Fritsch (1994a). Mesoscale convective systems in weakly forced large-scale environments. Part II: Generation of a mesoscale initial condition. *Monthly Weather Review 122*, 2068–2083.

Stensrud, D. J. and J. M. Fritsch (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.

Toth, Z. and E. Kalnay (1993). Ensemble forecasting at the NMC: The generation of perturbations. *Bulletin of American Meteorological Society 74*, 2317–2330.

Wandishin, M. S., S. L. Mullen, D. J. Stensrud, and H. E. Brooks (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 G. Chan (1994). Simulation of stationary Gaussian processes in $[0, 1]^d$. *Journal of Computational and Graphical Statistics 3*, 409–432.