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

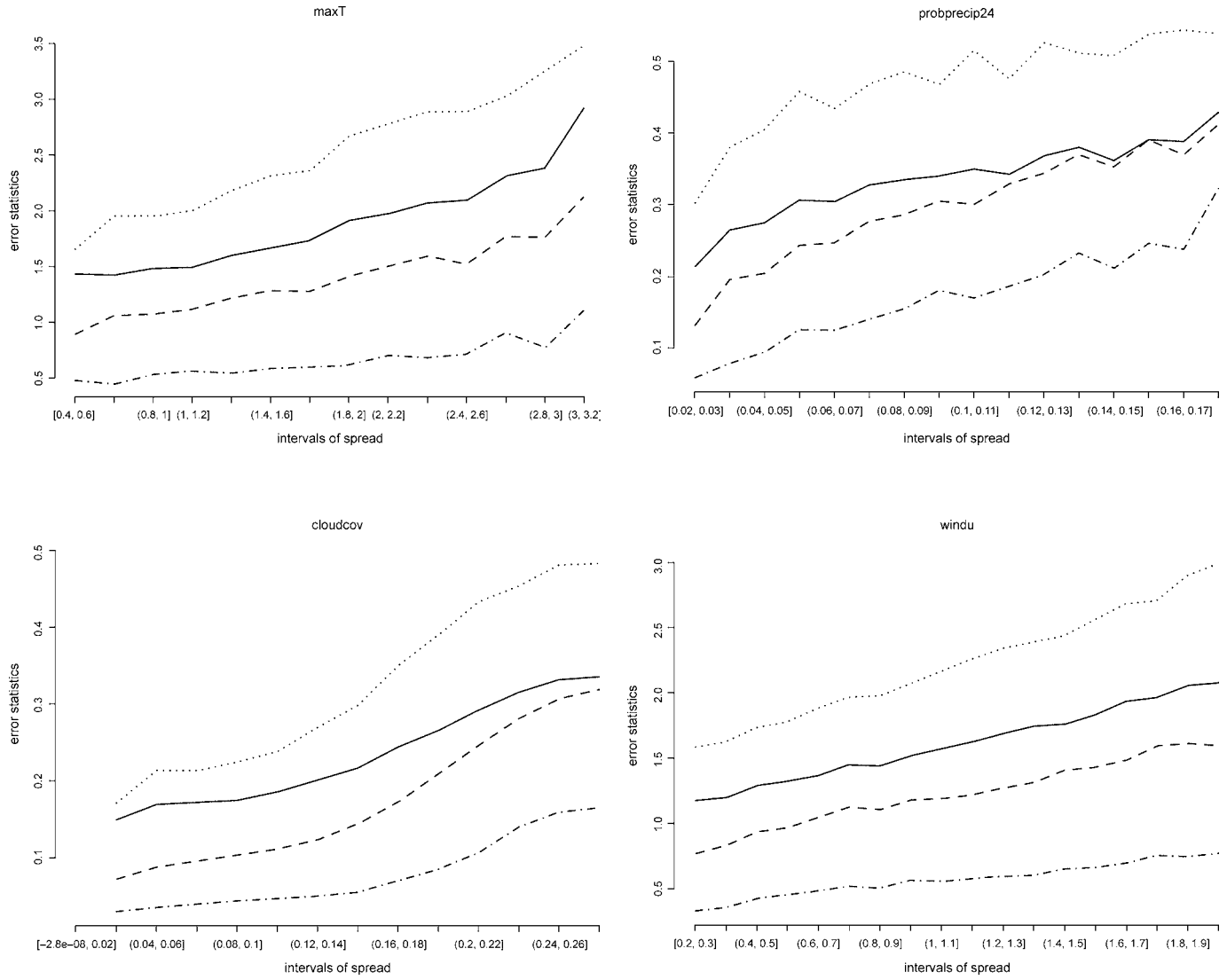

Figure 1. Relationship Between Spread of the Ensemble and Quantiles of the Forecast Error Distribution for Maximum Temperature, Probability of Precipitation, Cloud Cover, and u-Component of the Wind. The four lines correspond to first quantile (– · –), median (– –), mean (——), and third quantile (·····) 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