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

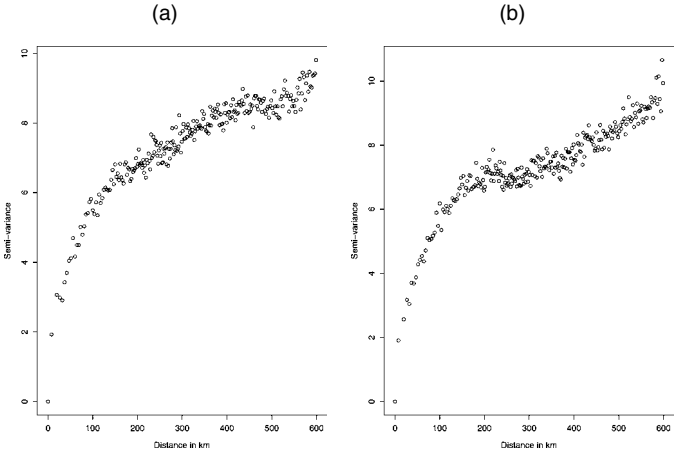

Figure 1. Directional Variograms for Temperature Forecast Errors in the North American Pacific Northwest, January–June, 2000: (a) North–South; (b) East–West.

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.