Model Consensus

J. M. FRITSCH, J. HILLIKER, AND J. ROSS

Department of Meteorology, The Pennsylvania State University, University Park, Pennsylvania

R. L. VISLOCKY

The Forecast Institute, State College, Pennsylvania

(Manuscript received 13 December 1999, in final form 30 May 2000)

ABSTRACT

Consensus forecasts from the control runs of several operational numerical models are compared to 1) the control-run forecasts of the individual models that compose the consensus and to 2) other consensus forecasts generated by varying the initial conditions of the various individual models. It is found that the multimodel consensus is superior to the individual control runs and to the consensus forecasts constructed from ensembles of runs generated by varying model initial conditions. The source of the forecast improvement by model consensus is not the result of a simple cancellation of errors as a result of an overall positive bias in one model and an overall negative bias in another. Rather the main improvement stems from overlapping differences in the sign of the errors associated with forecasts of individual traveling disturbances. The results suggest that variations in model physics and numerics play a substantial role in generating the full spectrum of possible solutions that can arise in a given numerical forecast.

1. Introduction

Using ''consensus'' as the basis for a decision emanates from the notion that the cumulative knowledge of all concerned individuals is greater than that of a single individual. The more complex the issue and the more factors involved in the decision, the more likely that a single individual cannot fully comprehend nor successfully integrate the relevant factors to arrive at the ''correct'' decision. Thus, it follows that, in dealing with highly complex situations, a single individual's decision will be prone to error and bias. Clearly, the atmosphere, with its multitude of processes interacting nonlinearly, and with our very limited ability to observe and define its three-dimensional state, constitutes an exceedingly complex system of which each of us, as individuals, have limited knowledge. Therefore, forecasting the state of the atmosphere would appear to be a situation wherein consensus would be of great value. Experiments to test this notion have confirmed that a blended or average forecast is superior to that of an individual or of a single forecast guidance product (Sanders 1973; Bosart 1975; Winkler et al. 1977; Gyakum 1986). Mathematically, the strength of consensus forecasting rests on the foundation that if two or more sources of imperfect forecasts have some independent and nonzero forecast skill, then the performance of an optimal blend[1] of the predictions will be superior to any of the individual sets of forecasts (Thompson 1977).

In addition to the improvements that result from applying consensus to forecasts made by humans, improvements have also been documented when consensus is applied to objective techniques such as statistical forecasting guidance (Fraedrich and Leslie 1987; Verret and Yacowar 1989; Vislocky and Young 1989; Vislocky and Fritsch 1995) and to numerical model output. In particular, it has been shown that a simple mean of an ensemble of forecasts generated by varying model initial conditions is as good or better than control-run simulations run with resolutions higher than the ensemble members (Toth and Kalnay 1993; Tracton and Kalnay 1993; Molteni et al. 1996; Hamill and Colucci 1997; Tracton et al. 1998; Stensrud et al. 1999). Although the mean of the ensemble is, in general, not the optimal blend, Toth et al. (1998) have pointed out that it provides

Corresponding author address: J. M. Fritsch, Department of Meteorology, 503 Walker Building, The Pennsylvania State University, University Park, PA 16802.
E-mail: fritsch@ems.psu.edu

[1] As considered here, the optimal blend is constructed by weighting each component of the consensus in a manner that minimizes the differences between *past* forecasts and observed conditions. Typically, this is accomplished using statistical techniques such as multiple regression wherein least squares minimization is applied to determine the appropriate weights.

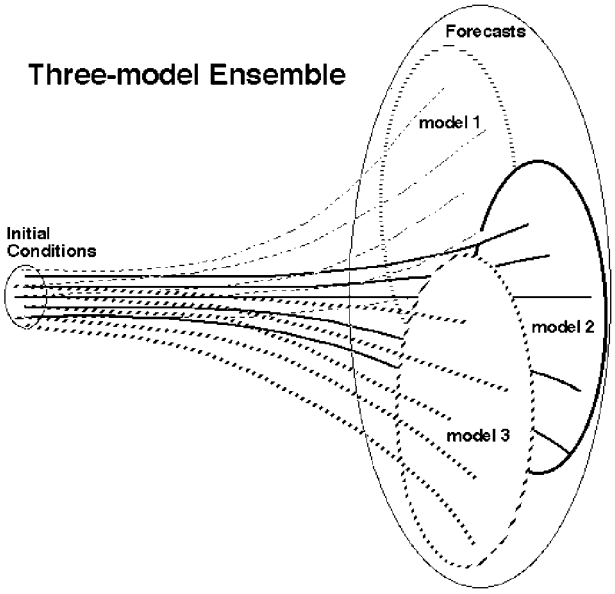

FIG. 1. Three-model ensemble. Large ellipse (thin solid line) represents the total set of possible solutions (forecasts) based upon variations in initial conditions, model physics, and model numerics. Small ellipses (labeled 1, 2, and 3) within the large ellipse indicate the range of solutions for each of the three models.

nonlinear filtering that damps the unpredictable components of the forecast and thereby provides an overall improvement in the forecast.

Recently, Brooks et al. (1995), Hamill and Colucci (1997), Stensrud et al. (1999), and Tracton et al. (1998) have recognized that *variations in model physics and numerics* may provide independent information to an ensemble and, therefore, based upon Thompson (1977), have the potential to further improve model consensus forecasts. In other words, the spectrum of possible independent solutions may be far greater than that generated simply by varying the initial conditions. A schematic depicting the concept of a broader envelope of solutions based upon variations in model physics and numerics is presented in Fig. 1. The figure suggests that without allowing for variations in model physics and numerics, a set of solutions derived from variations in the initial conditions alone may not be capable of generating the correct solution. Indirect evidence that this may be true is suggested by the fact that the observed 500-mb height contour in spaghetti diagrams occasionally lies outside the envelope of forecast solutions generated by varying the model's initial state (Toth et al. 1997). Moreover, as pointed out by Stensrud et al. (1999), there are many explicit examples in the literature that directly infer how a difference in model physics can produce an entirely different family of solutions. Perhaps one of the most intuitive examples of this is the difference in solutions that would occur between two models where one uses a convective parameterization that includes moist downdrafts and the other does not. It is well known that the lifting that initiates and organizes convection in some mesoscale convective systems is a consequence of moist-downdraft-generated surface-based cold pools. In fact, in certain instances, mesoscale convective systems will actually propagate upstream (i.e., with a component against the environmental flow at all levels) as a consequence of the lifting of the environment by convectively generated cold pools (Stensrud and Fritsch 1994a,b; Zhang and Fritsch 1988). Thus, without the moist downdrafts in the convective parameterization, an entire set of solutions, including the *correct* solution, may never materialize in the ensemble. Based upon numerical sensitivity experiments conducted by Klemp and Wilhelmson (1978), Benjamin and Carlson (1986), Mullen and Baumhefner (1988), Zhang et al. (1994), Zhang and Harvey (1995), Alapaty et al. (1998), Clappier (1998), and many others, similar arguments can be made for forecasts of phenomena ranging in scale from extratropical cyclones down to individual convective clouds.

Quantitative evidence that a multimodel consensus provides a superior forecast was first presented by Rousseau and Chapelet (1986). They averaged the output from five global models (France, Germany, Japan, the United Kingdom, and the United States). Further evidence of the superiority of model consensus forecasts was presented by McCalla and Kalnay (1988) when they combined the operational model outputs from the Medium Range Forecast model (MRF), the European Centre for Medium-Range Weather Forecasts, and the U.K. Met. Office. Krishnamurti et al. (1999) have likewise demonstrated the superiority of a multimodel consensus when they combined output from eight research and operational models from around the globe. Moreover, Wobus and Kalnay (1995) have shown that the correlations among multiple models are, by far, the most important predictor of forecast skill.

The primary goal of this paper is to document how a multimodel consensus, constructed from an ensemble of the control runs of several different operational numerical models, compares to (a) the performance of the individual model control runs, and (b) to the consensus constructed from ensembles based strictly upon variations in model initial conditions. A second goal is to determine *how* a multimodel consensus produces improvements over the forecasts of the individual members.

Section 2 provides a description of the components of several consensus (ensemble average) forecasts and experiments. Results are presented in section 3. The final section provides a brief summary and a few concluding remarks.

2. Data and methodology

Two basic experiments were conducted. The first experiment was designed to examine how the consensus of an ensemble of different operational models compares to the consensus of forecasts generated by varying

the initial conditions of a single model. The second experiment was designed to test how a multimodel consensus compares to the performance of the individual models that make up the consensus.

For any given initial time, the control-run initial conditions for each of the various operational models are typically very similar to each other. This similarity prevails even though factors such as differences in data cutoff time and/or differences in analysis techniques can introduce slight variations in the initial states. On the other hand, each operational model contains significant differences in model physics and numerics, and therefore it is likely that an ensemble composed of several different operational models can provide a broader representation of the full suite of possible solutions than that which could be generated by varying the initial state of a single model.

For the first experiment, 17 complete ensemble cases from the period 25 July–26 December 1995 were available from the output of the National Centers for Environmental Prediction (NCEP) pilot program on short-range ensemble forecasting (SREF) (Brooks et al. 1995). Cases before this period were excluded due to missing Aviation Model (AVN) runs. Fifteen ensemble members were specially created for each experimental day: 10 were integrations of the Eta Model (Black 1994; Rogers et al. 1995) and 5 were generated using the regional spectral model (RSM) (Juang and Kanamitsu 1994). The Eta Model was configured with 80-km horizontal resolution and 38 levels; the RSM was run with 80-km horizontal resolution and 28 sigma levels. The five RSM initial conditions consisted of the MRF model (Kanamitsu et al. 1991) control-run analysis and four ''breeding of growing modes'' (Toth and Kalnay 1993) analyses. The breeding technique was also used to generate four of the Eta Model initial conditions, while the remaining six Eta Model initial states were constructed from different analyses interpolated to the Eta Model grid. Specifically, these analyses are 1) the Nested Grid Model (NGM) regional analysis (Dimego et al. 1992), 2) the aviation run of the MRF (Parrish and Derber 1992), 3) a static Eta model optimum interpolation (Rogers et al. 1995), 4) the Eta model data assimilation system (Rogers et al. 1996), 5) the MRF control run (Parrish and Derber 1992), and 6) the three-dimensional variational analysis (Parrish et al. 1996). Further information about the model runs and each of the ensemble members is provided by Hamill and Colucci (1997) and Stensrud et al. (1999).

In addition to the 15 model runs described above, output from the Limited-Area Fine-Mesh model (LFM) (Gerrity 1977) and the NGM (Hoke et al. 1989) was also archived, thus making a total of 17 runs available for the construction of various model consensus forecasts. All models were run to 48-h lead times. Forecasts (48 h) of height, temperature, and humidity were obtained at the 1000-, 850-, 700-, and 500-mb layers for each model run. Upper-air observations at sounding stations within the domain bounded by 25°–50°N and 65°–125°W (roughly the 4000 km by 3000 km contiguous portion of the United States) were selected as the verification. Gridded model output for the nearest four grid points surrounding each upper-air sounding site was linearly interpolated to the site and then the absolute error was calculated. Below ground level values were not included in the calculations. In all, there were roughly 4000 observations (17 cases, four levels, and over 60 stations) in the verification. However, considering that there was *vertical* dependence among the errors, the dataset should be considered to contain independent information for only about 1000 records. Moreover, allowing for the *horizontal* dependence among the observations reduces the independence of the dataset still further. In particular, according to Schlatter (1975), observations from rawinsondes must be separated by approximately 2000 km or more to be considered fully independent. Therefore, the reader is cautioned that only a fraction of the 4000 observations can be considered as truly independent from one another for any given case.

From the 17 model runs, the following consensus forecasts were constructed:

1) *Eta multibred*—the Eta control run plus the four Eta breeding runs,
2) *RSM multibred*—the RSM control run plus the four RSM breeding runs,
3) *multianalysis*—the Eta control run plus the five Eta runs generated from different analyses,
4) *total initial condition*—the Eta control run plus the four Eta breeding runs plus the five Eta runs generated from the different analyses,
5) *multimodel*—the Eta control plus the LFM plus the NGM plus the RSM control runs, and
6) *superblend*—a combination of 3 and 5 above.

In addition to these consensus forecasts, a ''smoothed Eta'' product was computed. This was generated by taking the Eta control run and smoothing the grid field by using an 81-point spatial smoother. Mathematically, this means that the Eta control forecast for a specific grid point equals the average of the 81 grid values nearest to the grid point of interest.

For the second experiment, a model consensus composed of the 32-km Eta, the 126-wave AVN, and the 80-km NGM was constructed for each of the 0000 and 1200 UTC 48-h forecasts generated in the two periods 1 December 1998–31 January 1999 and 1 June–31 July 1999. Similar to the first experiment, 48-h forecasts of height, temperature, and humidity were obtained at the 850-, 700-, and 500-mb levels for each model run. The forecasts were verified against the upper-air observations at sounding stations within the same domain as described for the first experiment. Gridded model values for the nearest (in the horizontal) four grid points surrounding each upper-air sounding site were linearly interpolated to the site and then the actual (positive or

negative) error and the absolute error were calculated. For the 4-month period, there were over 35 000 records with complete data available for the verification (≈190 cases for three levels at 60–70 sounding stations). For each of the three forecast parameters (height, temperature, and relative humidity), the mean absolute error was computed at each of the three pressure levels for all sounding sites. The errors were computed for all 48-h forecasts, for each model, and for the model consensus. The models and the multimodel consensus were then ranked from 1 (best) to 4 (worst) based upon their performance. Analyses of the error fields were constructed for each of the selected pressure levels using the objective analysis system described by Cahir et al. (1981).

3. Results

a. Comparison of initialization consensus to multimodel consensus

The mean absolute errors for the Eta control run, the smoothed Eta product, and for each of the consensus forecasts for the first experiment are presented in Fig. 2. The errors are averaged over all stations and pressure levels. It is immediately clear from Fig. 2 that the multimodel consensus forecast is superior to the other consensus forecasts, being, on average, over 12% more accurate than the Eta control run. A statistical *t* test indicates that the differences between the multimodel consensus and the Eta control run are significant at the 98% confidence level for all three parameters (height, temperature, and humidity). The pronounced superiority of the multimodel consensus is particularly interesting since (a) the other consensus formulations had more members (5–10) than the multimodel consensus (4) and (b) the multimodel consensus used forecasts from the LFM and NGM, models that are inferior to the Eta. Certain other comparisons are also worth noting. In particular, simply smoothing the model output reduced errors, on average, over 3%—nearly as great as the 4% improvement produced by the multianalysis ensemble. Toth and Kalnay (1997), in an analysis of the performance of global model ensembles, also found that smoothing reduced errors. Moreover, they demonstrated that over 60% of the gain that the ensemble has over the same resolution control forecast (in terms of root-mean-square error) is retained after optimally smoothing both the control and the ensemble mean. Nevertheless, in the present case, were it not for the many other benefits derived from running ensembles, it would be questionable whether or not it would be worth the time and expense of running a multianalysis ensemble just to gain less than a 1% improvement over simply smoothing the control run. Clearly, it could be argued that the greatest value of single-model ensembles does not stem from the fact that the ensemble mean is slightly more accurate than the control run or a spatially smoothed forecast. Rather, ensembles with many members provide quantitative guidance about the distribution of possible events and, most importantly, an indication of the possibility of extreme events.

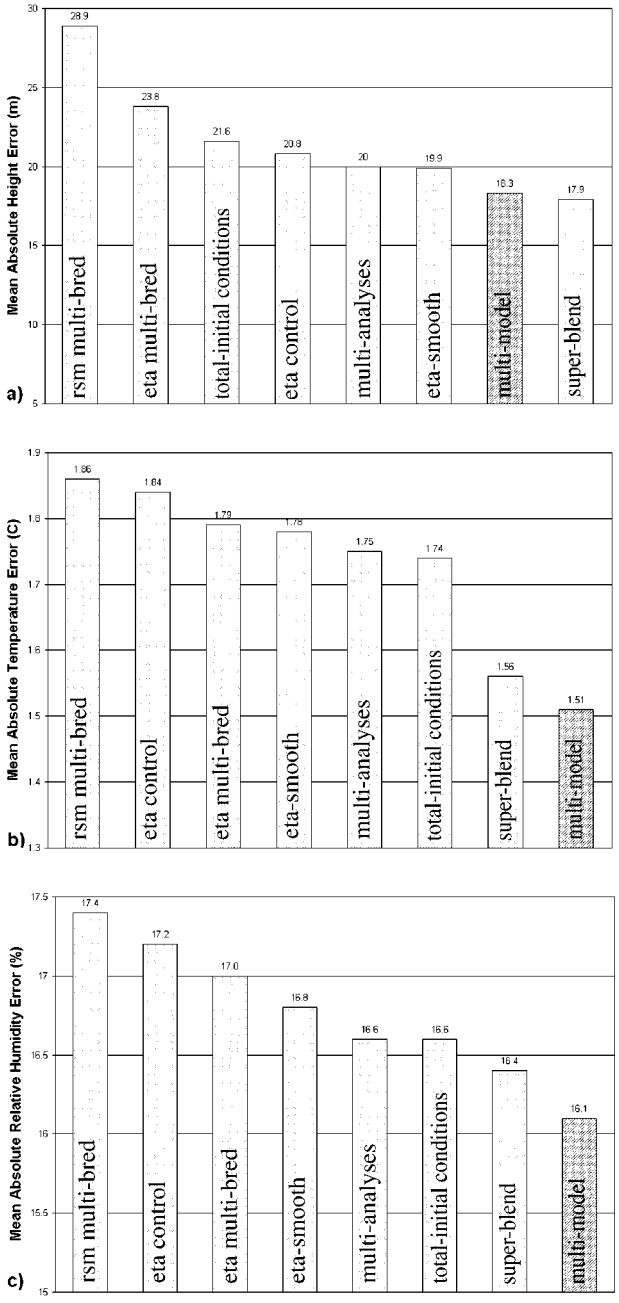

FIG. 2. Mean absolute errors for (a) height (m), (b) temperature (°C), and (c) relative humidity (%). Errors are averaged for all sounding stations and for all levels (1000, 850, 700, and 500 mb).

It is also of interest that inclusion of the Eta analysis runs with the multimodel consensus to form the ''superensemble'' did not, in general, improve the forecasts over what the multimodel ensemble produced, suggesting that *most of the important differences in the solutions were captured by the control runs of the four different models.*

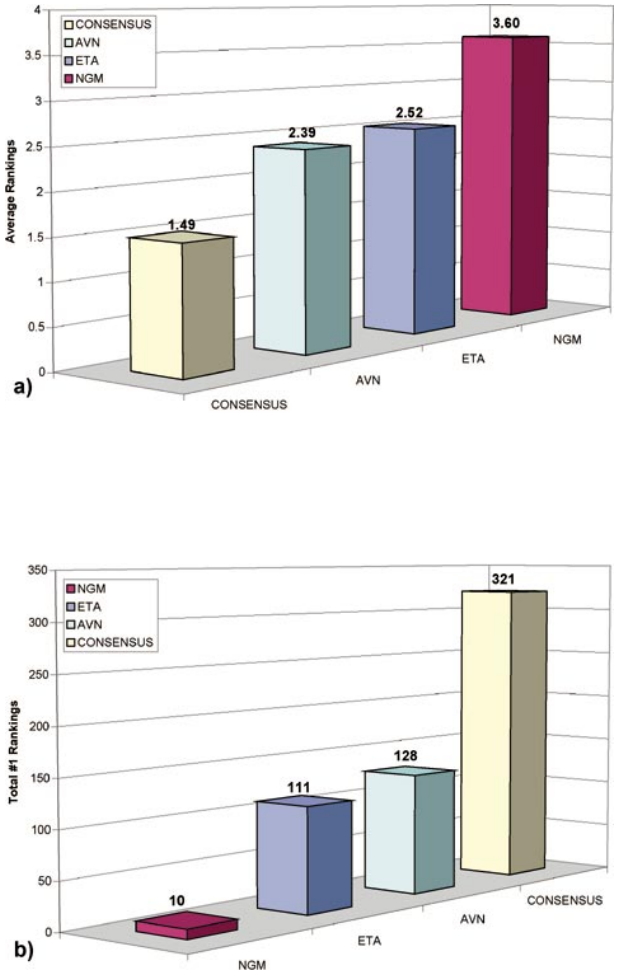

FIG. 3. (a) Average ranking and (b) the number of 1 rankings for each of the three models (Eta, NGM, and AVN) and for the multimodel consensus. The rankings are for all parameters at all three model levels (850, 700, and 500 mb) for the Dec–Jan/Jun–Jul 1998–99 database.

b. Comparison of individual operational models to multimodel consensus

Figures 3a,b show the mean ranking and the number of 1 rankings for each of the three models (Eta, NGM, and AVN) and for the multimodel consensus. The rankings are for all parameters at all three model levels for all of the cases in the database. It is clear that the multimodel consensus far outperforms the individual models.

Naturally, it is of interest to understand why this is so. Analysis of the performance of the individual models and the multimodel consensus on a daily basis provides some insight. For example, Fig. 4 shows a 26-day wintertime period of twice-daily mean *absolute* temperature errors (at the 850-, 700-, and 500-mb levels) for the 48-h forecasts from each of the models and for the multimodel consensus. Since the absolute errors of each of the three models typically are greater than the multimodel consensus errors, it is clear that there must be instances where the errors from one model are opposite in *sign* from one of the other models. If this were to happen on a regular basis, for example, if one model had a cold bias and another had a warm bias, then the errors would tend to cancel each other and the advantage of model consensus could be easily understood. Whether or not this is true should be evident from an analysis of the bias of each model. Unfortunately, accurately identifying bias is an extremely difficult task. This is because bias varies as a function of location, model level, model parameter, time of day, season, synoptic pattern, etc. Nevertheless, in order to explore the possibility that offsetting biases are contributing to the superior performance of model consensus, a crude measure of bias for one of the forecast parameters (temperature) was constructed for each of the models. Specifically, Fig. 5 shows the *overall* mean temperature error (bias) and the distribution of the *individual* mean errors as a function of their sign and magnitude. The overall mean is the average of the 850-, 700-, and 500-mb temperature errors at all sounding sites and for all of the 48-h runs in the wintertime sample. Individual means are the average of the errors at all sounding sites for a single 48-h forecast. It is readily evident that all of the models, and therefore the model consensus as well, have a slight cold bias.[2] This suggests that the improvement of model consensus over the individual models does not result from a simple cancellation of error as a result of an overall warm bias in one model and an overall cold bias in another. Note, however, that the mean error (Fig. 5) is typically much smaller than the mean absolute error (Fig. 4), again indicating that there must be large error cancellations as a result of the averaging in constructing the model consensus.

In order to explore this issue further, the 700-mb temperature errors at three sounding sites (Salem, OR; Omaha, NE; Pittsburgh, PA) were plotted (Fig. 6) for the same 26-day period shown in Fig 4. Based upon the sequence of errors at these three sites, it is clear that much of the improvement of multimodel consensus over the individual models stems from the fact that the individual model errors often have opposite signs. In fact, on approximately one-third to one-half of the days, there are significant (>1°C) error cancellations—and some events exhibit cancellations of 4°C or more.

To understand better how/when the offsetting errors are occurring, the 700-mb error fields for each model for each 48-h forecast were plotted for the North American region and compared to the error fields from the multimodel consensus. On most days, the error patterns for each model are very similar to each other and therefore, of course, the consensus is similar as well. However, in other instances, the error patterns are substan-

[2] The NGM bias is minimal since a statistical correction for a pronounced cold bias was introduced starting with the 1200 UTC 21 October 1987 run (National Weather Service 1986, 1987).

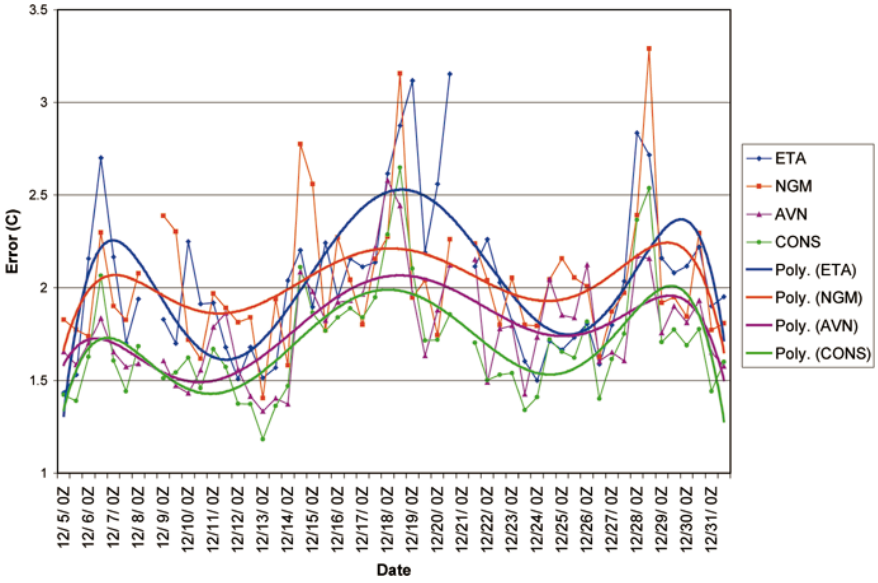

FIG. 4. Mean absolute temperature errors (°C) at 850, 700, and 500 mb for the twice-daily 48-h forecasts from the Eta, NGM, AVN, and model consensus for 5–31 Dec 1998. Gaps in the plots indicate missing data. Heavy solid lines show the polynomial best fit to the data from each model and to the model consensus.

tially different. For example, Fig. 7 shows the patterns for a day when the model errors differed in sign over much of the eastern half of the nation, especially over the Great Lakes and New England. As a result, the errors of the multimodel consensus were, in general, considerably less in those regions compared to the errors associated with each of the component models.

Although Figs. 6 and 7 show how model consensus benefits from cancellation of the errors from the individual models, it is also evident from Fig. 6 that, in some situations, the error of one of the models is so large that the model consensus becomes worse than the forecasts from the remaining two models (e.g., 1200 UTC 18 Dec; Fig. 6b). Therefore, if one of the model consensus members is consistently much less accurate than the other members, the benefits of computing the consensus may be insufficient to overcome the negative effects of the poorly performing model.

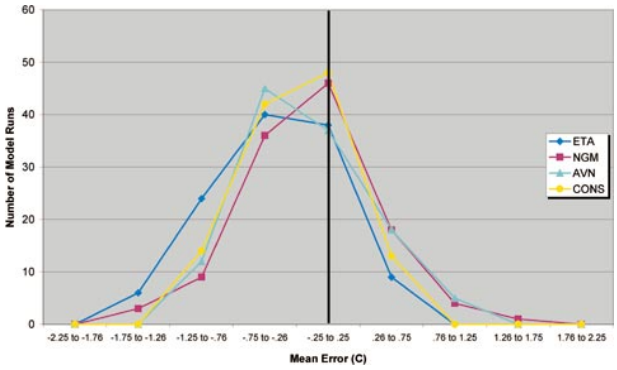

FIG. 5. Frequency distribution of average temperature error (°C) at the 850-, 700-, and 500-mb levels for the Eta, NGM, AVN, and model consensus for all soundings in the Dec–Jan/Jun–Jul 1998–99 database.

Examination of the daily mean absolute error and the error patterns for the 190 cases revealed several other interesting properties of the model forecasts. In particular, Fig. 4 shows that there are 1–2-week cycles when the magnitude of the mean absolute temperature errors varies by 30%–50%. These large changes in the size of the average errors suggests that, as might be expected, certain large-scale patterns are much more prone to producing errors than others. Moreover, in the periods with large errors, the model consensus does not perform as well, relative to the individual models, as it does in periods with relatively small errors (Fig. 4). This is because the periods with generally larger errors tend to produce instances of extremely large errors by one of the individual models and, as pointed out above, the errors can become so large that the consensus forecast becomes worse than that of the remaining two models (see Fig. 6). A cursory examination of the upper-air patterns that correspond to the alternating periods of large and small errors in wintertime indicates that the periods of small errors correspond to patterns wherein the jet stream passes over Alaska and western Canada before entering the United States. The periods of large errors generally correspond to patterns when a strong flow enters the United States from the Pacific. For example, Fig. 8a shows the observed 500-mb height field for 21 December 1998. This analysis corresponds to one of the initialization times that produced 48-h forecast errors that were relatively small (i.e., 23–24 Dec in Fig.

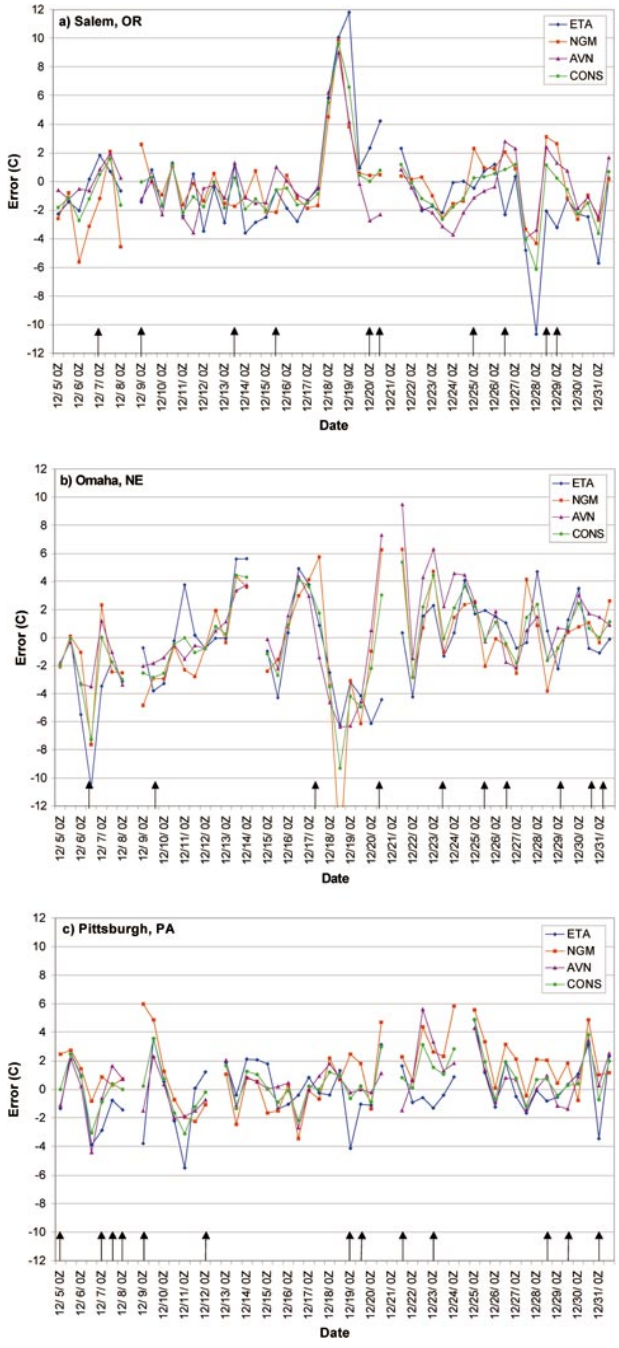

FIG. 6. Temporal sequence of the 700-mb 48-h temperature errors (°C) for the Eta, NGM, AVN, and model consensus for 5–31 Dec 1998 at (a) Salem, OR; (b) Omaha, NE; and (c) Pittsburgh, PA. Vertical arrows indicate temperature error cancellations ≥1°C.

4). Disturbances entering the United States were reasonably well sampled as they passed through the observing network over Alaska and western Canada. Conversely, during the large error period shown in the center of Fig. 4 (i.e., 17–18 Dec), a series of strong disturbances, one of which exhibited an 80 m s^{-1} jet at the 300-mb level, entered the United States from the poorly sampled eastern Pacific region (Fig. 8b). The disturbances propagated inland across the Rocky Mountains and then plunged southeastward through the Mississippi Valley and over the southeastern Appalachians. An analysis of the error patterns during this period (and several others) revealed that the models tended to err in unison for periods of several days. For example, notice in Fig. 9 how all of the models started producing a large positive temperature error along the Canadian border on 15–16 December. This area of positive errors was associated with the 80 m s^{-1} jet from the Pacific and can be readily followed in all of the model forecasts for eight successive model runs. As shown in Fig. 9, the contiguous area of positive temperature errors propagated southeastward across the Mississippi Valley, over the southern Appalachians, and then northeastward along the Atlantic coast.

The similarity of the error patterns among the various models, and the continuity of the error patterns from one model run to the next (in the Fig. 9 sequence and in many others not shown) suggests that 1) there is very little difference in the initial conditions of the various models, and 2) each subsequent initialization of the models with new observations is unable to ''purge'' the errors transmitted into each successive model cycle by the first guess field. This raises the issue of whether there is too much weighting of the model forecast (first guess field) in constructing the model initial conditions over well-sampled land areas. Nevertheless, there evidently are enough differences in the forecasts so that the model consensus can still produce significant improvements relative to the performance of the individual models.

4. Summary and concluding remarks

The consensus of forecasts from the control runs of several operational numerical models was compared to 1) the control-run forecasts of the individual models that compose the consensus and to 2) other consensus forecasts generated by varying the initial conditions of the various individual models. It was found that the multimodel consensus produced smaller mean absolute errors than those generated by individual operational-model control runs. The multimodel consensus also produced smaller errors than the consensus forecasts constructed from ensembles of runs generated by varying model initial conditions. The results indicate that variations in model physics and numerics play an important role in generating the full spectrum of possible solutions that can arise in a given numerical forecast.

The improvement of model consensus over the individual models did not result from a simple cancellation of errors as a result of an overall positive bias in one model and an overall negative bias in another; for the sample parameter that was investigated (upper-air temperature), the sign of the overall bias was the same for each model. Rather, much of the improvement of

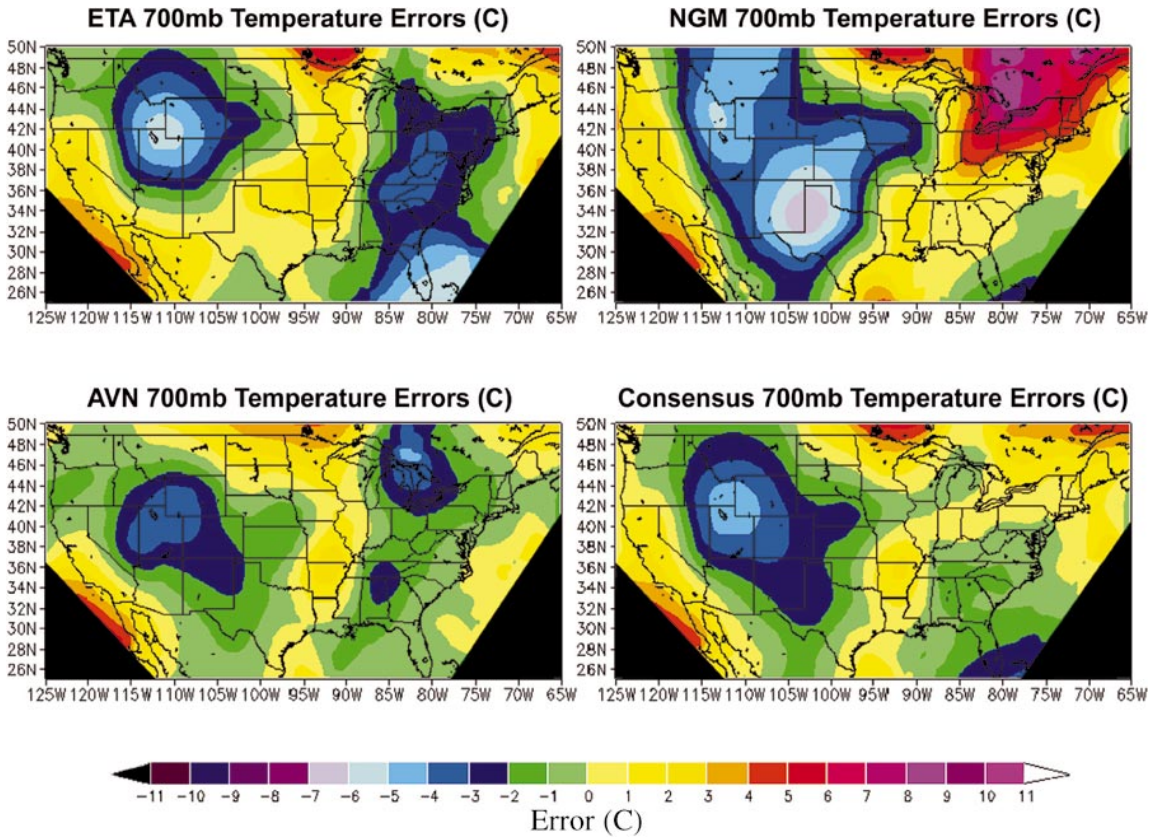

FIG. 7. Horizontal distribution of the 700-mb 48-h temperature errors (°C) for the Eta, NGM, AVN, and model consensus for 0000 UTC 9 Dec 1998.

the multimodel consensus over the individual models stemmed from the fact that the individual model errors often had opposite signs over mesoscale regions and would therefore cancel each other in the multimodel consensus. In fact, on approximately one-third to one-half of the days investigated, there were significant (>1°C) error cancellations at individual sounding locations where the model forecasts were verified. For some events, cancellations were as large as 4°C or more.

It is important to note that, while the model consensus examined here appears to offer more accurate forecasts than the individual members of the consensus, this would not be true if one of the consensus members were consistently much less accurate than the other members. In this situation, the benefits of computing the consensus are insufficient to overcome the negative effects of the poorly performing model. Therefore, it appears that model consensus will provide a more accurate forecast only if the individual components of the consensus have similar levels of skill.

Since models will never be perfect, there will always be the potential to improve forecasts through model consensus. Each member of the consensus could evolve as parameterizations, data assimilation techniques, and initialization procedures are improved. The various parameterizations and numerical procedures could be mixed and matched among the different models. And, since the physics and numerics of the different combinations will interact differently and nonlinearly, a broad ensemble of reasonable solutions should result. To achieve further improvement, this approach could be coupled with the variations in model initial conditions. Equivalently, as suggested by Z. Toth (1999, personal communication), one could use the same model infrastructure and simply vary the formulations for various physical and numerical processes.

It is well known that statistical postprocessing of model output greatly improves the accuracy of the forecast over that of the raw model output (e.g., Dallavalle 1996). Therefore, it is likely that the most accurate forecast guidance can be obtained from statistically postprocessing and then optimally combining the output from several different models or from statistically postprocessing multimodel output products. Vislocky and Fritsch (1995) tested this concept by averaging the model output statistics (MOS) guidance for an ensemble of two members: the NGM and the LFM. Their results demonstrated conclusively that the average of the statistically postprocessed output from the two models was far superior to the performance of the statistically postprocessed output from either of the individual models. Optimally combining

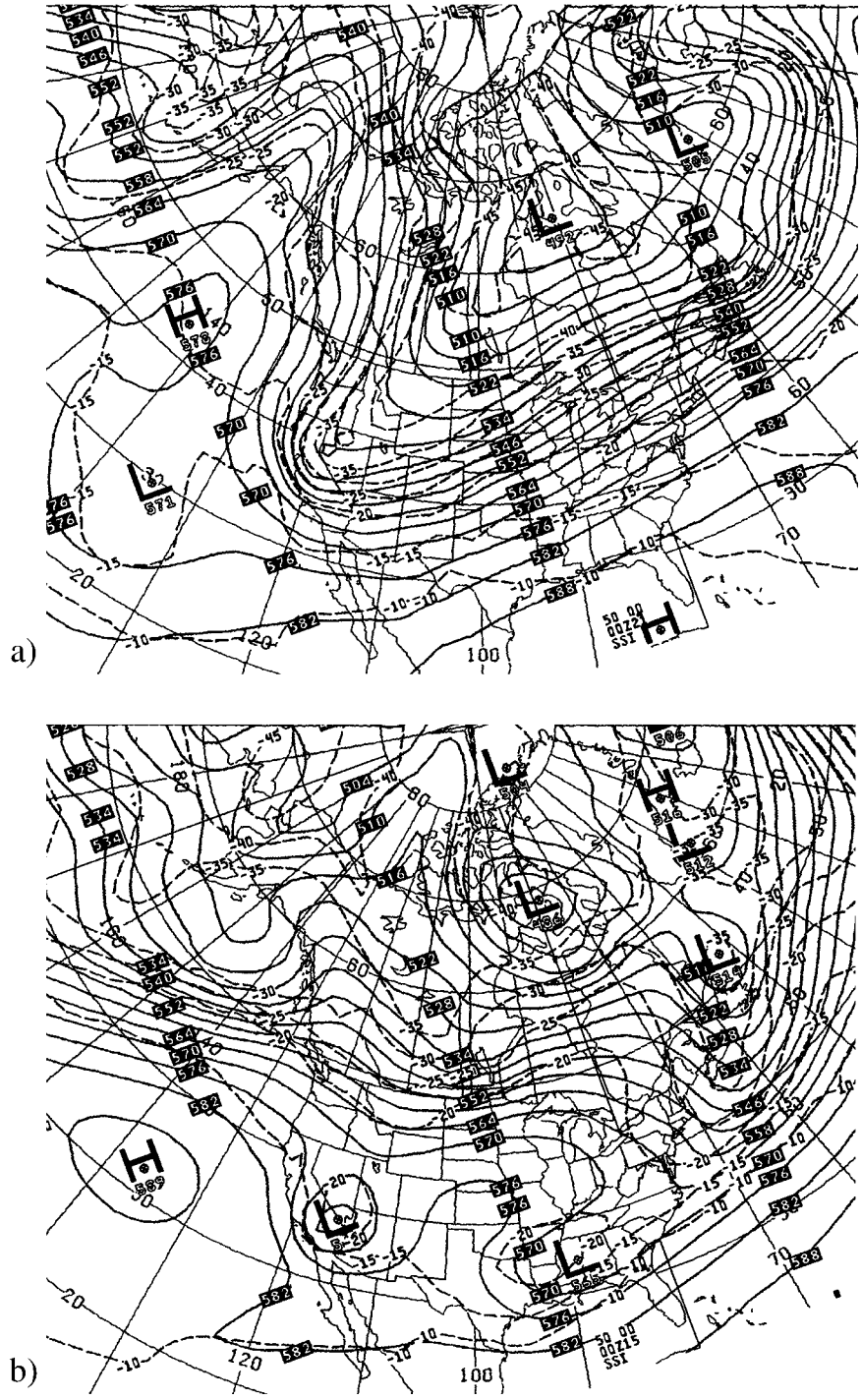

FIG. 8. Observed 500-mb height field (dm) for (a) 0000 UTC 21 Dec and (b) 0000 UTC 15 Dec 1998.

the model output further improved the superiority of the multimodel ensemble over the performance of the individual members. Recently, Krishnamurti et al. (1999) confirmed the Vislocky and Fritsch (1995) experiment by optimally combining the output from eight global models and demonstrating that the skill of the optimal blend far surpassed that from any of the individual members of the multimodel ensemble.

The model consensus constructed in the present study might have been further improved if the bias of each model were removed *before* constructing the consensus or if, as demonstrated by Krishnamurti et

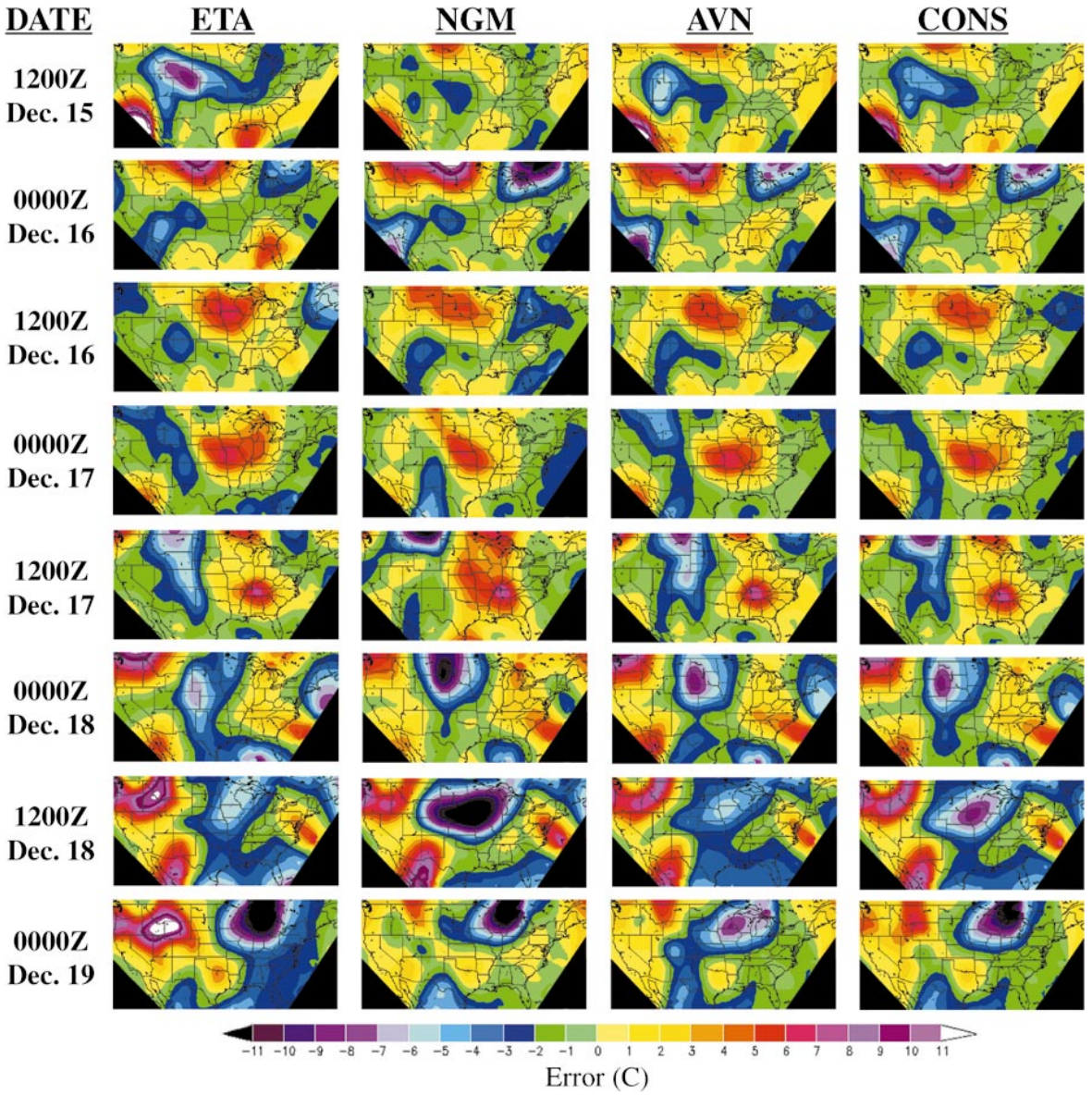

FIG. 9. Temporal sequence of the horizontal distribution of the 700-mb 48-h temperature errors (°C) for the Eta, NGM, AVN, and model consensus for 1200 UTC 15 Dec–0000 UTC 19 Dec 1998.

al. (1999), an optimal blend of the output from the various multimodel ensemble members was constructed. However, for an unbiased consensus to have operational utility, the models would either have to be ''frozen'' (as is done in the traditional MOS approach and in Krishnamurti et al.'s study) or the biases would have to be recomputed every time a change is made to one of the models—a daunting task. On the other hand, drawing again upon the analogy to human consensus, it is clear that the biases of human forecasters change with time (and probably with each forecast event), yet consensus continues to do exceedingly well in forecast competitions. Thus, while statistical postprocessing does produce a huge improvement in forecast skill and should be routinely invoked in generating the best weather forecasting guidance, it appears that significant improvements can be obtained by simply generating a multimodel consensus of the raw model output.

Acknowledgments. The authors appreciate the helpful comments and suggestions from Steven Tracton, Zoltan Toth, Eugenia Kalnay, and Craig Bishop. We are also grateful to David Stensrud for facilitating acquisition of the SREF data. This work was supported by USWRP/National Science Foundation Grant ATM-9714154.

REFERENCES

Alapaty, K., R. Mathur, and T. Odman, 1998: Intercomparison of spatial interpolation schemes for use in nested grid models. *Mon. Wea. Rev.*, **126,** 243–249.

Benjamin, S. G., and T. N. Carlson, 1986: Some effects of surface heating and topography on the regional severe storm environment. Part I: Three-dimensional simulations. *Mon. Wea. Rev.*, **114,** 307–328.

Black, T., 1994: The new NMC mesoscale Eta model: Description and forecast examples. *Wea. Forecasting,* **9,** 265–278.

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

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

Cahir, J. J., J. M. Norman, and D. A. Lowry, 1981: Use of a real time computer graphics system in analysis and forecasting. *Mon. Wea. Rev.*, **109,** 485–500.

Clappier, A., 1998: A correction method for use in multidimensional time-splitting advection algorithms: Application to two- and three-dimensional transport. *Mon. Wea. Rev.*, **126,** 232–242.

Dallavalle, J. P., 1996: A perspective on the use of model output statistics in objective weather forecasting. Preprints, *15th Conf. on Weather Analysis and Forecasting,* Norfolk, VA, Amer. Meteor. Soc., 479–482.

Dimego, G. J., K. E. Mitchell, R. A. Petersen, J. E. Hoke, J. P. Gerrity, J. J. Tuccillo, R. L. Wobus, and H.-M. H. Juang, 1992: Changes to NMC's Regional Analysis and Forecast System. *Wea. Forecasting,* **7,** 185–198.

Fraedrich, K., and L. M. Leslie, 1987: Combining predictive schemes in short-term forecasting. *Mon. Wea. Rev.*, **115,** 1640–1644.

Gerrity, J. F., 1977: The LFM model—1976: A documentation. NOAA Tech. Memo. NWS NMC 60, U.S. Dept. of Commerce, Washington, DC, 68 pp.

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

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

Hoke, J. E., N. A. Phillips, G. J. DiMego, J. J. Tuccillo, and J. G. Sela, 1989: The Regional Analysis and Forecast System of the National Meteorological Center. *Wea. Forecasting,* **4,** 323–334.

Juang, H.-M., and M. Kanamitsu, 1994: The NMC nested regional spectral model. *Mon. Wea. Rev.*, **122,** 3–26.

Kanamitsu, M., and Coauthors, 1991: Recent changes implemented into the Global Forecast System at NMC. *Wea. Forecasting,* **6,** 425–435.

Klemp, J. B., and R. B. Wilhelmson, 1978: The simulation of three-dimensional convective storm dynamics. *J. Atmos. Sci.*, **35,** 1070–1095.

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

McCalla, C., and E. Kalnay, 1988: Short and medium range forecast skill and the agreement between operational models. Preprints, *Eighth Conf. on Numerical Weather Prediction,* Baltimore, MD, Amer. Meteor. Soc., 634–640.

Molteni, R., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122,** 73–119.

Mullen, S. L., and D. P. Baumhefner, 1988: Sensitivity to numerical simulations of explosive oceanic cyclogenesis to changes in physical parameterizations. *Mon. Wea. Rev.*, **116,** 2289–2329.

National Weather Service, 1986: Modeling of physical processes in the Nested Grid Model. NWS Tech. Procedures Bull. 363, National Oceanic and Atmospheric Administration, 5 pp.

——, 1987: Statistical correction for the NGM cold bias. Western Region Tech. Attachment No. 87-44, National Oceanic and Atmospheric Administration, 8 pp.

Parrish, D. F., and J. C. Derber, 1992: The National Meteorological Center's spectral statistical-interpolation system. *Mon. Wea. Rev.*, **120,** 1747–1763.

——, J. Purser, E. Rogers, and Y. Lin, 1996: The regional 3D-variational analysis for the Eta model. Preprints, *11th Conf. on Numerical Weather Prediction,* Norfolk, VA, Amer. Meteor. Soc., 454–455.

Rogers, E., D. G. Deaven, and G. J. DiMego, 1995: The regional analysis system for the operational ''early'' Eta model: Original 80-km configuration and recent changes. *Wea. Forecasting,* **10,** 810–825.

——, T. L. Black, D. G. Deaven, G. J. DiMego, Q. Zhao, M. Baldwin, N. W. Junker, and Y. Lin, 1996: Changes to the operational ''early'' Eta analysis/forecast system at the National Centers for Environmental Prediction. *Wea. Forecasting,* **11,** 391–413.

Rousseau, D., and P. Chapelet, 1986: A test of the Monte Carlo method using the WMO/CAS intercomparison project data. World Climate Research Programme Rep. 9, WMO/TD-No. 141, 6 pp.

Sanders, F., 1973: Skill in forecasting daily temperature and precipitation: Some experimental results. *Bull. Amer. Meteor. Soc.,* **54,** 1171–1179.

Schlatter, T. W., 1975: Some experiments with a multivariate statistical objective analysis scheme. *Mon. Wea. Rev.,* **103,** 246–257.

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. *Mon. Wea. Rev.,* **122,** 2068–2083.

——, and ——, 1994b: Mesoscale convective systems in weakly forced large-scale environments. Part III: Numerical simulations and implications for operational forecasting. *Mon. Wea. Rev.*, **122,** 2084–2104.

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

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

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74,** 2317–2330.

——, and ——, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **125,** 3297–3319.

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

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

Tracton, S., and E. Kalnay, 1993: Ensemble forecasting at NMC: Operational implementation. *Wea. Forecasting,* **8,** 379–398.

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

Verret, R., and N. Yacowar, 1989: Improvement of numerical weather element forecasts by combining forecasts from different procedures. Preprints, *11th Conf. on Probability and Statistics,* Monterey, CA, Amer. Meteor. Soc., 58–63.

Vislocky, R. L., and G. S. Young, 1989: The use of perfect prog forecasts to improve model output statistics forecasts of precipitation probability. *Wea. Forecasting,* **4,** 202–209.

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

Winkler, R. L., A. H. Murphy, and R. W. Katz, 1977: The consensus of subjective probability forecasts: Are two, three, . . . heads better than one? Preprints, *Fifth Conf. on Probability and Statistics*, Las Vegas, NV, Amer. Meteor. Soc., 57–62.

Wobus, R. L., and E. Kalnay, 1995: Three years of operational prediction of forecast skill at NMC. *Mon. Wea. Rev.,* **123,** 2132–2148.

Zhang, D.-L., and J. M. Fritsch, 1988: Numerical sensitivity experiments of varying model physics on the structure, evolution and dynamics of two mesoscale convective systems. *J. Atmos. Sci.,* **45,** 261–293.

——, and R. Harvey, 1995: Enhancement of extratropical cyclogenesis by a mesoscale convective system. *J. Atmos. Sci.,* **52,** 1107–1127.

——, J. S. Kain, J. M. Fritsch, and K. Gao, 1994: Comments on ''Parameterization of convective precipitation in mesoscale numerical models: A critical review.'' *Mon. Wea. Rev.,* **122,** 2222–2231.