An Ensemble Forecasting Primer

JOEL K. SIVILLO AND JON E. AHLQUIST

Department of Meteorology and NOAA/FSU Cooperative Institute for Tropical Meteorology, The Florida State University, Tallahassee, Florida

ZOLTAN TOTH

GSC, Environmental Modeling Center, NCEP, NWS/NOAA, Washington, D.C.

(Manuscript received 18 January 1996, in final form 26 June 1997)

ABSTRACT

An ensemble forecast is a collection (an ensemble) of forecasts that all verify at the same time. These forecasts are regarded as possible scenarios given the uncertainty associated with forecasting. With such an ensemble, one can address issues that go beyond simply estimating the best forecast. These include estimation of the probability of various events and estimation of the confidence that can be associated with a forecast.

Global ensemble forecasts out to 10 days have been computed at both the U.S. and European central forecasting centers since December 1992. Since 1995, the United States has computed experimental regional ensemble forecasts focusing on smaller-scale forecast uncertainties out to 2 days.

The authors address challenges associated with ensemble forecasting such as 1) formulating an ensemble, 2) choosing the number of forecasts in an ensemble, 3) extracting information from an ensemble of forecasts, 4) displaying information from an ensemble of forecasts, and 5) interpreting ensemble forecasts. Two synoptic-scale examples of ensemble forecasting from the winter of 1995/96 are also shown.

1. Introduction

An ensemble forecast is a collection of two or more forecasts that verify at the same time. These forecasts start from different initial conditions and/or are based on different forecasting procedures. The various forecasts all represent possibilities given the uncertainties associated with forecasting. From these possibilities, one can estimate probabilities of various events as well as an average (''consensus'') forecast.

This primer is targeted at the operational forecaster. Its intent is to discuss in a nonmathematical way ensemble forecasting's philosophy, implementation, and challenges. General principles are stressed more than operational details, which are subject to frequent change. Also included are synoptic-scale examples of operational ensemble forecasts from the winter of 1995/96 carried out at the U.S. National Centers for Environmental Prediction (NCEP).

The Internet, particularly through the World Wide Web, is a conduit for real-time operational ensemble forecasts and an increasing amount of information about ensemble forecasting. A Web search for ''ensemble forecasting'' using searching facilities such as Alta Vista, Yahoo, etc., will reveal various sites.[1]

Several other introductions to ensemble forecasting have been published. From the U.S. side, these include Toth and Kalnay (1993) and Toth et al. (1997) and, from Europe, Palmer et al. (1990) and Palmer (1993). The best recent reference, containing many reports on a wide variety of ensemble forecasting investigations, is the (unrefereed) preprints volume from the American Meteorological Society's 11th Conference on Numerical Weather Prediction, held in August 1996.

An ensemble of forecasts can be composited into a single forecast by means of a weighted average (Van den Dool and Rukhovets 1994), but an ensemble also contains additional information. The forecasts in an ensemble suggest possibilities whose probabilities can be estimated (Anderson 1996). These probabilities currently require a correction because the ensemble often underestimates the range of possibilities (Hamill and Colucci 1997a; Zhu et al. 1996; Buizza 1997). An ensemble of forecasts can also be used to estimate the reliability of the composite forecast (Wobus and Kalnay

Corresponding author address: Dr. Jon E. Ahlquist, Department of Meteorology/CITM, The Florida State University, Tallahassee, FL 32306-3034.
E-mail: ahlquist@met.fsu.edu

[1] At the time of writing, NCEP's ensemble forecasting activities were documented at http://nic.fb4.noaa.gov:8000, this being the Web page for NCEP's Environmental Modeling Center. This Web address is subject to change, however.

1995; Buizza 1997). Further, a forecast ensemble can suggest where additional special observations might be targeted to improve forecast accuracy (Bishop and Toth 1996; Emanuel et al. 1996).

Besides using a single model to generate an ensemble of forecasts, an ensemble can arise from the use of two or more different forecast techniques or numerical models. For example, Vislocky and Fritsch (1995) demonstrated that the average of Model Output Statistics (MOS) from the Limited Fine Mesh and the Nested Grid Model (NGM) is more accurate than individual MOS from either model. As another example, hurricane forecasting in the U.S. is carried out by generating various forecasts using different numerical and climatological methods. Another ensemble is the collection of global forecasts issued by forecasting services around the world (Tracton and Kalnay 1993, 380; Richardson et al. 1996).

An earlier example of ensemble forecasting is the World War II D Day forecast, which considered predictions from three independent forecasting teams using different techniques (Shaw and Innes 1984; Fuller 1990, 85–100). As the D Day forecast example shows, ensemble forecasting does not have to involve computers. If two or more human forecasters make separate forecasts and then intercompare them, they are working with an ensemble of forecasts.

2. Background

On 7 December 1992, NCEP (then called the National Meteorological Center) began computing 10-day ensemble forecasts on an operational basis (Toth and Kalnay 1993). Later that month, experimental ensemble forecasts began at the European Centre for Medium-Range Weather Forecasts (ECMWF) and were issued on Saturday, Sunday, and Monday (Molteni et al. 1996, 73, 75–76). In May 1994, ensemble forecasts became part of ECMWF's daily operational routine. Canada's and several European countries' operational centers are also concerned with ensemble forecasting (see, e.g., Dubreuil 1996; Houtekamer et al. 1996; Åkesson 1996; Harrison 1996; Richardson et al. 1996). The U.S. Navy and the meteorological services of Japan and South Africa have begun ensemble forecasting. India and Australia are planning ensemble forecasting operations. Operational ensemble forecasting follows years of background work by researchers around the world. We begin by reviewing some of this work.

By the early 1950s, some meteorologists considered applying statistical methods to weather prediction to cope with the uncertainties inherent in forecasting. See Gleeson (1961) for a review of this earlier work and for an outline of how one can view the forecasting problem in terms of evolving probabilities.

During the 1960s, Lorenz (1963, 1965, 1969) investigated fundamental aspects of atmospheric predictability. He demonstrated that weather, even when viewed as a deterministic system, may have a finite prediction time (1963). Further, predictability varies with different weather situations in a way not easily discernible by naked eye examination of weather maps (1965). He also calculated that the average limit to atmospheric predictability at planetary scales is on the order of 10 days (1969). More recent estimates (Simmons et al. 1995) suggest that a 10-day average forecasting limit may be a little too conservative, but probably not by more than a few days. See also Reynolds et al. (1994, 1281–1282) for an analysis of forecast error and Stoss and Mullen (1995) for a discussion of how the skill of 48-h NGM 500-hPa height forecasts varied with the initial flow regime. Below, we discuss Lorenz's work in more detail.

Even if forecasting models included accurate representations of all physical processes, a fundamental limit to atmospheric predictability arises due to imprecise initial conditions (Lorenz 1963). We can never expect to measure all variables at all levels with absolute precision. There will always be some space between observing stations, and conditions there will not be known exactly.

Errors of different spatial scales grow at different rates (Lorenz 1969). On average, the fastest error growth occurs at small scales. For example, an airmass thunderstorm may be unpredictable 2 h into the future. Smaller scales have even shorter prediction timescales. Thus, the loss of predictability occurs first at the smallest scales and ''propagates'' to larger scales (Lorenz 1969; Shaw 1981). This flow of uncertainty is in the opposite direction to the turbulent cascade of energy described in L. F. Richardson's poem:

> Big whirls have little whirls
> which feed on their velocity.
> Little whirls have lesser whirls,
> and so on to viscosity.

To describe the flow of information from small to large scales, the authors offer an inversion of the 1733 quatrain by Jonathan Swift that inspired Richardson[2]:

> So, naturalists observe, a flea
> will bite its dog most readily.
> The dog, surprised, will bite its master,
> who changes course with actions rasher.

This poem illustrates the so-called butterfly effect, named after a 1972 talk by Lorenz entitled, ''Does the flap of a butterfly's wings in Brazil set off a tornado in Texas?'' (Lorenz 1993, 14–15, 181–184). Lorenz ''avoided answering the question, but noted that if a

[2] Here is the original of Swift's metaphorical comment on authors feeding off the work of others.

> So, naturalists observe, a flea,
> Hath smaller fleas that on him prey.
> And these have smaller still to bite 'em;
> And so proceed ad infinitum.

single flap could lead to a tornado that would not otherwise have formed, it could equally well prevent a tornado that would otherwise have formed'' (op cit., 14).

By 1970, Tatarskiy (1969), Epstein (1969), and Gleeson (1970) had proposed ways to forecast probabilities. These procedures do not involve an ensemble but rather forecast statistical quantities directly. The difficulty with these approaches is that they require an enormous amount of calculation, even for computer resources available in the foreseeable future (Leith 1974, 410). As an alternative, Leith (1974) demonstrated that an ensemble of roughly 10 forecasts seemed to be large enough to make real improvements in 6–10-day forecasts. With the advances in computing that were taking place at that time, ensemble forecasting, or ''Monte Carlo'' forecasting as Leith called it, became a distinct possibility.

During the 1980s, ensemble forecasts were computed in a research mode to establish procedures and assess utility. One of the simplest approaches to ensemble forecasting is to consider a collection of forecasts issued at different times but that all verify at the same time. This technique, known as the Lagged Average Forecast (LAF) method, was discussed by Hoffman and Kalnay (1983). For example, a 24-h forecast made this morning could be considered along with a 36-h forecast made last night, a 48-h forecast from yesterday morning, etc. The advantage of the LAF method is that it uses forecasts that already exist. Its chief disadvantage is that the forecasts in an LAF ensemble are not even close to being equal contenders, since the newest (hence, shortest range) forecast will almost always be considerably more accurate than the oldest (hence, longest range) forecast.

As computer power increased, it became possible to compute multiple forecasts that all start at the same time. In December 1992, the NCEP operational forecast ensembles consisted of 14 forecasts, 4 of which were computed at 0000 UTC and the remaining 10 of which were computed 12–48 h earlier (Tracton and Kalnay 1993, Fig. 3b). In 1996, the NCEP ensemble contained 17 forecasts of which 12 were computed at 0000 UTC and 5 were computed 12 h earlier (Kalnay and Toth 1996). As computer power continues to increase, the number of forecasts in an ensemble, the complexity of the model, the model resolution, and the length of the forecast are all expected to increase. For example, in December 1996, ECMWF expanded its ensemble to 51 forecasts in which the dynamics were computed at T159 resolution and the model's physical processes (radiation, precipitation, etc.) were computed at T106.

As of 1997, both NCEP and ECMWF used ensemble forecasts primarily for synoptic and planetary scales in the multiday forecast range, but ensembles are potentially useful at all space and time scales. Even at forecast lead times of a few hours, mesoscale features in an ensemble of forecasts will differ. After a few days, synoptic-scale forecasts will exhibit noticeable differences, with planetary-scale forecasts diverging after that. This is the cascade of uncertainty mentioned earlier (Lorenz 1969).

The potential utility of short-range ensemble forecasting was discussed at a workshop in Washington, D.C., in July 1994 (Brooks et al. 1995). The principal recommendation of that workshop was to perform a pilot study in which an ensemble of regional 48-h forecasts would be computed weekly. Preliminary results from these regional ensemble experiments involving the NCEP Eta and Regional Spectral Model were just becoming available at the time this review was written (Tracton et al. 1997). For example, Hamill and Colucci (1997b) reported that an adjusted ensemble of short-range precipitation forecasts is more skillful than the NGM MOS for all categories of precipitation amount, although probability of precipitation was more skillful from NGM MOS than from the ensemble estimate. The ensemble required an adjustment because the range of values in the ensemble underestimated the range of values in the verifications (Hamill and Colucci 1997a).

3. Ensemble forecasting philosophy

The purpose of ensemble forecasting is to recognize the inherent uncertainty of weather forecasting. It asks, what forecasts are possible? Its goals are to increase average forecast accuracy, to estimate the likelihood of various events, and to estimate the decay of forecast skill with increasing lead time.

Computer weather forecasts are uncertain for a variety of reasons: incomplete (and inaccurate) observations of initial conditions, incomplete knowledge of the dynamical and physical equations that govern the atmosphere, and the further approximations associated with converting differential and integral equations into forms that can be solved by computer within the available time. These factors are compounded by Lorenz's (1963, 1969) finding that errors in initial conditions, no matter how small, impose a limit on how far into the future a skillful weather forecast is possible, even if the governing equations were known exactly, which they are not.

Errors in midlatitude initial conditions are currently small enough and forecasting models are good enough that, for the first day or two, forecast errors at synoptic and larger scales are often governed by linear processes. That means two things:

1) The size of the forecast error is directly proportional to the size of the initial error. For example, if the initial error could be cut in half, then the forecast error would be cut in half.
2) It is meaningful to separate initial errors into categories (such as initial ''upper air'' errors versus initial surface observation errors, or initial errors over various geographic regions) and to consider independently their contribution to the forecast error.

By a couple of days into a forecast, though, synoptic-

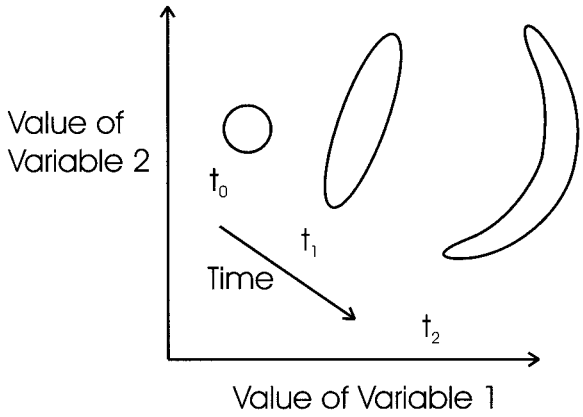

FIG. 1. Schematic demonstrating how small perturbations in initial conditions evolve into larger perturbations. An initially small circle of perturbations at time t_0 will evolve by linear processes into an ellipse, an example of which is shown at time t_1. When the magnitude of perturbations increases, their growth is governed by nonlinear effects, which limit the most extreme growth and cause the ellipse to fold, as shown at time t_2.

scale errors reach a magnitude where they are affected by nonlinear processes. Then, 1) the size of the forecast error is not proportional to the size of the initial error, and/or 2) it is not possible to separate the forecast error into independent components. When nonlinearities are involved, errors do not simply add together; they interact, often in complicated ways.

The growth of forecast errors occurs as sketched in Fig. 1. During the time when forecast errors are governed by linear processes, an initially circular range of possibilities for two variables will evolve into an elongated ellipse, an example of which is shown at time t_1. Physically, the elliptical shape means that different initial errors grow at different rates. Modest nonlinear error growth has occurred by time t_2, when the ellipse in Fig. 1 has begun to fold and deform into a more arbitrary shape. Physically, this means that the error in one variable begins to have a more complicated relationship to the error in another variable. If the forecast time would proceed even further, the loop in Fig. 1 would continue to stretch, fold, and spread until it approached the climatological scatter that would be expected between the two variables. At that point (not shown), climatology would be as good a forecast as anything.

Considerable mathematical theory exists to describe error growth by linear processes, but linear processes govern only the growth of small errors, and that applies only to the first day or two of synoptic-scale forecasts. Ensemble forecasting has its primary value for the more general case when forecast errors are influenced by nonlinear processes. Nonlinear effects typically become important first for small scales and last for planetary scales, with the details varying with weather conditions.

In order to carry out ensemble forecasting effectively, questions such as the following must be answered: 1) How should one construct an ensemble of forecasts; that is, how should the forecasts in an ensemble be different from each other? 2) How many forecasts should be in the ensemble? 3) How can forecast information be extracted from an ensemble of forecasts? 4) What are the best ways of presenting this information to forecasters? 5) How does one interpret ensemble forecast products? Only the last question is the direct responsibility of the field forecaster, but some understanding of the other points will help field forecasters address that last question better. Therefore, we devote a section to each of the questions above.

a. Constructing an ensemble of forecasts

For midlatitude synoptic scales, uncertainty with initial conditions seems to be a larger source of forecast error than are model deficiencies (Reynolds et al. 1994). As a particular example of this, Daley (1991, 1) noted that Hollingsworth et al. (1985) ''discovered that the predicted evolution of the Presidents' Day snowstorm (of 18–19 February 1979) was extremely sensitive to *small errors in the initial analysis in the northwestern Pacific four days earlier*. In other words, a small localized error in the initial analysis affected the forecast for locations far removed in space and time.''

Because current ensemble forecasting at both NCEP and ECMWF is focused on the consequences of initial value errors, both centers are generating an ensemble of forecasts by starting a forecast model from a variety of initial conditions. How one should choose an ensemble of initial states is not an easy matter to decide. Computer limitations restrict the number of forecasts that can compose an ensemble, so care must be taken to choose an ensemble of initial conditions that is small enough to fit within available computer resources but large and diverse enough to mimic statistics of the possibilities.

The number of initial conditions in an ensemble will always be small compared to the infinite number of possible atmospheric states. In order to cope with this situation, both NCEP in Washington and ECMWF in Europe have chosen to concentrate on those initial conditions that are likely to cause the largest forecast errors, although they do this in different ways (see Toth and Kalnay 1993).

Following Lorenz (1965), ECMWF computes those perturbations (i.e., possible errors in initial conditions) that, under linearized dynamics, would grow the most during the first 48 h of the forecast (Buizza 1997). An ensemble of initial conditions is then created by adding and subtracting these fastest-growing potential errors to and from a global analysis. These potential errors are called ''singular vectors.''

NCEP generates initial conditions by ''breeding growing modes'' (Toth and Kalnay 1993, 1997). In this procedure, perturbations from the most recent previous forecast ensemble are scaled and then added and sub-

tracted from the current analysis to form an ensemble of initial conditions for the next ensemble forecast.

Once NCEP and ECMWF create ensembles of initial conditions, they both use fully nonlinear forecasting models to compute ensembles of extended forecasts.

ECMWF's procedure is based on the assumption that all initial errors are equally likely, while NCEP's procedure emphasizes errors arising from the use of an earlier forecast to form the ''first guess'' for the global analysis. Mathematical theory exists for ECMWF's procedure, which requires extensive calculations using a separate linearized model. ''Breeding of growing modes'' allows for nonlinear effects by using the nonlinear forecasting model. It has the advantage of being computationally inexpensive because it exploits existing forecast products. Neither approach should be regarded as the final word. Both methods of picking an ensemble of initial conditions represent reasonable choices given the present state of the art. Research is under way to improve both approaches.

Regional mesoscale models add further considerations to the problem of assigning initial conditions. These include choice of boundary conditions as well as initial conditions. Also, ensembles can be generated by varying model physics; see, for example, Bresch and Bao (1996).

b. The number of ensemble members

A major concern in operational ensemble forecasting is how many forecasts to include in an ensemble. Because of computer limitations and time constraints, ensemble forecasting involves a trade-off among various factors: model initialization, model complexity, the number of forecasts in the ensemble, and the amount of time, both computer and human, needed to process the information in an ensemble of forecasts.

Traditionally, when a single forecast was computed, it used the highest-resolution, most sophisticated model that could be run in the available time. Tracton and Kalnay (1993, section 2) showed that for forecasts beyond about 5 days, the horizontal resolution of NCEP's global spectral model could be cut in half (from T126 to T62) with little loss in forecast skill. Further, even a model run at T62 for 10 days was almost as skillful for the 6–10-day forecast range as a T126 model used for the first 5 days, followed by T62 resolution for days 6–10. The reason for the small difference in forecast skill is that scales beyond T62 are forecasted with almost no accuracy beyond 5 days. These are important facts because the time needed to compute a T62 forecast is about a ninth of that needed to compute a T126 forecast. In fact, NCEP's first operational ensemble included three 10-day forecasts at T62 and a T126 7-day forecast extended to 10 days at T62, all in the same time previously used to compute a single 10-day forecast at T126.

In addition to synoptic- and planetary-scale ensemble forecasts, researchers are investigating the utility of 0–48-h mesoscale ensemble forecasts. Current experiments involve an ensemble of 10 Eta Model forecasts with 80-km resolution and five forecasts from the Regional Spectral Model. The 10 eta forecasts can be computed more quickly than 1 Eta Model forecast with 29-km resolution (H. Brooks 1995, personal communication). An even larger number of forecasts is desirable to estimate the probability of extreme mesoscale events, such as severe thunderstorms (Brooks et al. 1996).

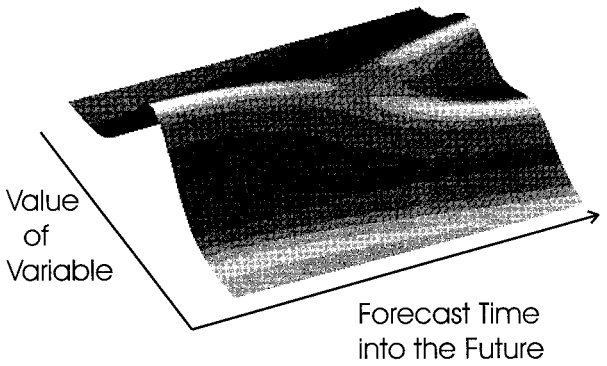

FIG. 2. Hypothetical probability distribution of the value an arbitrary variable, such as temperature at a point in space. Forecast time increases from left to right. Note that the width of the distribution generally increases as time increases. In this example, the initially Gaussian distribution ultimately becomes bimodal, which could occur due to uncertainty related to a frontal passage.

c. Extracting forecast information from an ensemble

From the perspective of ensemble forecasting, future weather is inherently uncertain; temperature, pressure, etc., cannot be forecasted with complete accuracy every time. One can, though, estimate forecast statistics such as the mean and standard deviation of the ensemble.

Figure 2 shows the hypothetical time evolution of the probability distribution for, say, temperature at a given point. Due to uncertainties in initial conditions, there is a range of possibilities even at the starting time of the forecast. Here, we have chosen to represent this initial uncertainty by a ''bell'' curve (a Gaussian distribution), although that distribution may not be appropriate for every case. As time proceeds into the forecast, the width of the distribution will generally increase, meaning that the range of likely values increases and the forecast is less certain. In Fig. 2, the hypothetical distribution even becomes bimodal (two peaked). A bimodal distribution could arise, for example, over uncertainty regarding a frontal passage. If a front has *not* passed at the forecast time, there is one set of possibilities, while another set of possibilities exists if a front *has* passed.

Since by definition it is impossible to forecast a random variable correctly every time, then how should one forecast? The usual starting point is to choose a way to measure forecast error. Once the rules for measuring forecast error have been set, one can then devise a procedure to minimize that error when averaged over many

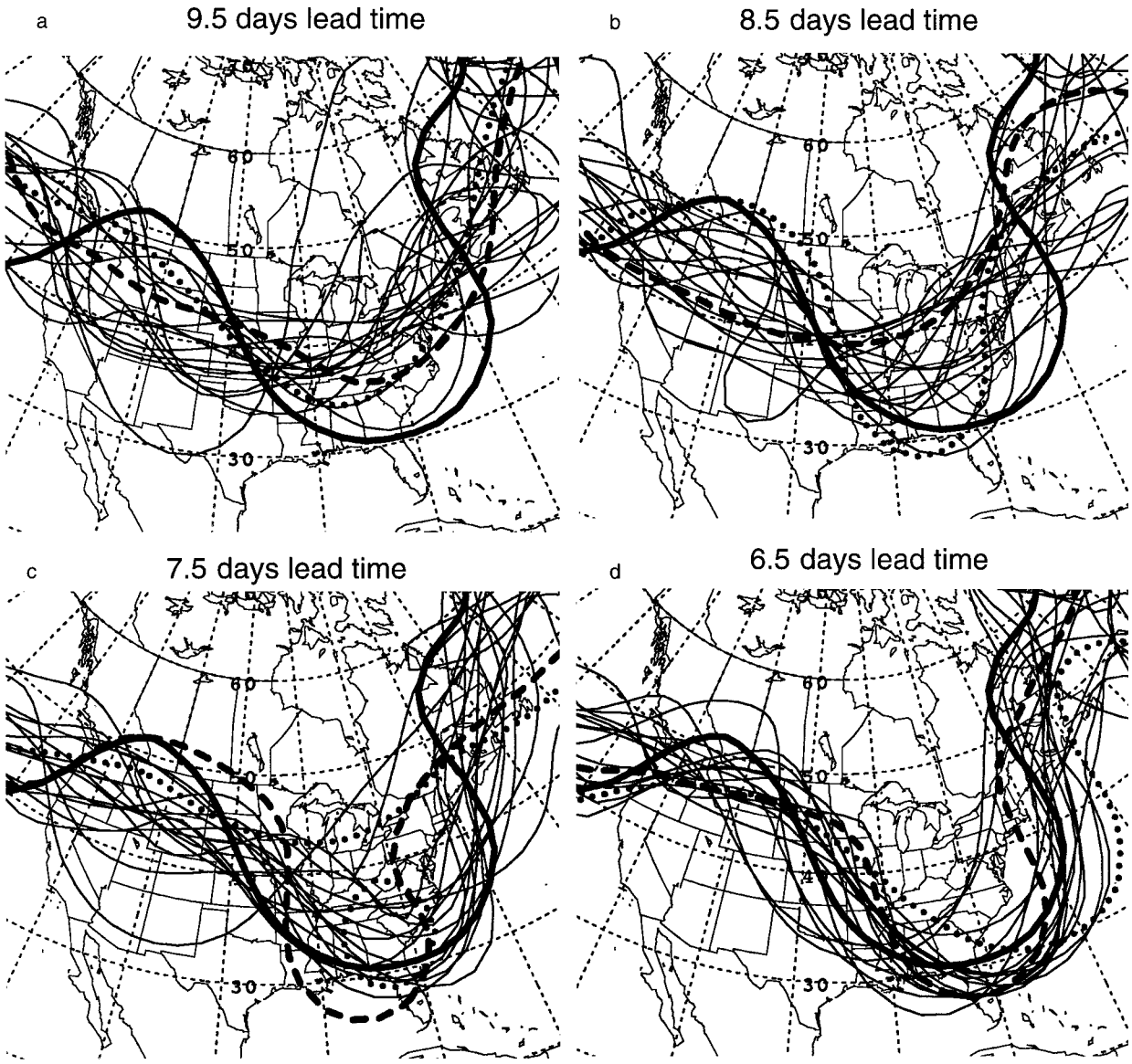

FIG. 3. The 5640-m contour line of 500-hPa height field from all 17 NCEP ensemble forecast members verifying at 1200 UTC 15 Nov 1995. The dotted line marks the 0000 UTC high-resolution control forecast (MRF) and the heavy solid line is the verifying analysis. (a)–(h) Ensembles with 9.5-, 8.5-, . . . , 2.5-day lead times (all valid at the same time). In (a)–(d), the 1200 UTC high-resolution control is highlighted with heavy dashed lines.

cases (Van den Dool and Rukhovets 1994). Indeed, anyone who has participated in a contest for human forecasters quickly learns to optimize (''hedge'') a forecast to minimize error. Exactly the same principle applies to computer forecasting: one ''tunes'' the procedure to minimize error. A number of different scoring schemes are in common use (Toth 1991). Optimizing forecast skill with respect to one scoring scheme does not in general, optimize forecast skill with respect to another scoring scheme. In addition, adjusting parameters to increase forecast skill for one part of the world might decrease forecast skill for another part of the world.

NCEP currently computes its best 6–10-day forecasts as a weighted average of ensemble forecasts. The weighting coefficients have been chosen to minimize the root-mean-square error (rms error) during an earlier evaluation period (Van den Dool and Rukhovets 1994). A weighted average is used because not all members of a forecast ensemble are equally skillful. (Some forecasts are older than others, some have lower resolution than others, etc.)

We emphasize that, because of the statistical nature of forecasting, the skill of ensemble forecasting compared to that of a more traditional single forecast can be judged only when a large number of cases are scored. A small number of case studies is not enough to compare

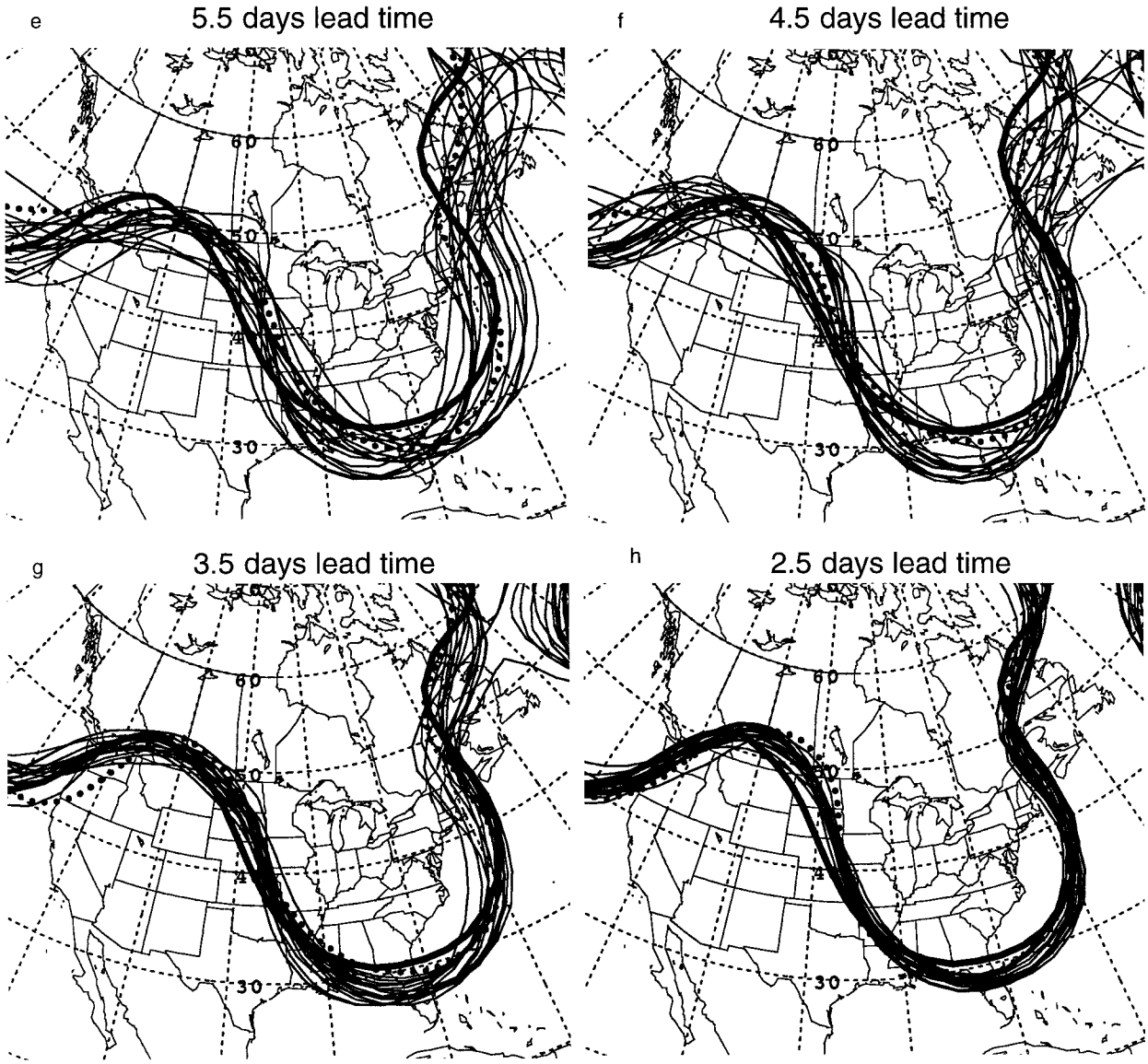

FIG. 3. (*Continued*)

competing procedures even for the traditional forecast approach (Daley and Chervin 1985).

d. Displaying information from ensemble forecasts

Even a single numerical forecast can provide more maps than any forecaster has time to use, while an ensemble can provide far more than that. How then can ensemble forecast information be conveyed without information overload?

A map of the weighted average of an ensemble forecast can easily replace a forecast map from a traditional single forecast. One can also supplement such an ensemble forecast with a map of the standard deviation among the forecasts. When the spread among forecasts is small, the average of the ensemble forecasts is probably accurate, at least in midlatitudes where model physics are good (Wobus et al. 1996; Buizza 1997). At the time this primer was written, maps of ensemble averages and standard deviations were issued daily on the Web.[3] A greater challenge is to display other aspects of ensemble information without overwhelming the user with maps.

One option is to plot more than one forecast on a single map. For example, consider the ensemble forecast of a rain/snow line. One can plot the 540-dam thickness line for the 1000–500-hPa layer for each of the forecasts in an ensemble, all on one map. Such a collection of

[3] The address at that time was http://nic.fb4.noaa.gov:8000 but is subject to change.

contour lines is called a spaghetti diagram. If the 540-dam isolines from the ensemble basically agree with one another, then the average forecast is likely to be skillful. If the isolines are widely dispersed over a region, then the forecast has greater uncertainty associated with it.

In the case of 540-dam isolines from the NCEP 17-member ensembles, the most equatorward line will mark the approximate boundary beyond which snow has low probability, and the most poleward 540-dam line will mark the approximate boundary beyond which rain has low probability. Precipitation in the intervening region could be liquid and/or frozen. Of course, one must also make the usual allowances for terrain height and any other pertinent factors in making rain versus snow decisions.

Another kind of information that can be gained from an ensemble of forecasts is the fraction of the ensemble that forecasts a certain kind of weather. This allows for an estimation of probabilities, although a correction is normally required because an ensemble tends to underestimate the range of possibilities (Hamill and Colucci 1997a; Zhu et al. 1996, Buizza 1997). For example, one can ask, what is the probability that city ''X'' will receive more than a certain amount of precipitation during a specified time interval? At the time of writing, examples of precipitation probabilities from ensemble forecasts could be found at the NCEP Web site.

Other ways also exist to display ensemble forecast information. For example, a page of small maps (called ''thumbnail sketches'' or ''postage stamps'') can show all the ensemble forecasts in miniature. Another option is to search the forecasts for possible clustering into a number of distinct patterns. Additional display methods can be expected in the future.

e. Interpreting ensemble forecasts

The biggest difference in appearance between a forecast map from a single numerical forecast and a map of the same field from an ensemble-average forecast is that the ensemble-average map will be smoother. The reason for this is that the ensemble-average emphasizes features that are similar from forecast to forecast and minimizes differences among forecasts. Since small scales are the least predictable, they will differ most from forecast to forecast, so small scales will be deemphasized in the average forecast. Therefore, spaghetti diagrams or thumbnail sketches showing information from all the forecasts are useful to suggest the range of possibilities.

When examining spaghetti diagrams, one should avoid putting too much trust in high-resolution members of an ensemble. In particular, while it is true that the T126 members of NCEP's ensemble are slightly more skillful than the T62 members, they are usually not as skillful as the ensemble average. In the operational example discussed in the next section, one of the lower-resolution ensemble members turned out to be more accurate than either of the two high-resolution members of the ensemble. There is no way of knowing which member of the ensemble will be most accurate until after the weather has occurred.

4. Ensemble forecasting example

In this section, we give examples of the kind of information an ensemble of forecasts can provide in excess of that available from a single control forecast. Our examples are spaghetti diagrams from the NCEP operational global ensemble forecast system, which is discussed in more detail in Toth et al. (1997). Here it suffices to say that the NCEP ensemble contains 17 forecasts each day, including two T126 high horizontal resolution control forecasts (one from 1200 UTC and another from 0000 UTC), and 15 forecasts at a lower T62 horizontal resolution, out of which 14 are from initial states that are perturbed by the breeding method (Toth and Kalnay 1993, 1997).

In our first example, we investigate how the NCEP ensemble performed in predicting an intense storm that hit the southeast United States on 15 November 1995. Figure 3 shows a sequence of spaghetti plots of the 5640-m contour for 500 hPa. All these maps are forecasts for 1200 UTC 15 November 1995 issued on consecutive days prior to the occurrence of the storm.

The first panel of Fig. 3 contains ensemble forecasts with 9.5-day lead time. At this time range, both high-resolution control forecasts indicate a trough over the eastern third of North America, but the trough's intensity is much below that observed. The ensemble as a whole indicates a trough over the eastern two-thirds of the continent, with some members suggesting an intense storm, just as was later observed. In fact, one of the perturbed T62 forecasts was rather close to the verifying analysis. At day 8.5, the uncertainty in the position of the trough is much reduced. Only one member suggests a trough as far west as eastern New Mexico, while several members, including the 1200 UTC control, still indicate that the storm may not be intense. At day 7.5, many of the ensemble members indicate an intense storm over the southeast coast. It is not before 6.5-day lead time that all members agree that an intense storm will hit the southeast coast. There is still a lot of uncertainty, though, in the timing of this storm, as seen in the longitudinal position of the trough in the different ensemble members. This is also evident from comparing the two control solutions. If only two lagged control forecasts are available, a strong similarity between them, however, should not be taken as an indication of more reliable forecasts. In fact, these two solutions can be similar just by chance, as is the case at 9.5-day lead time in the first panel of Fig. 3 over the continental United States. In cases like that, only the inspection of a larger ensemble can reliably indicate the degree of uncertainty associated with the forecasts.

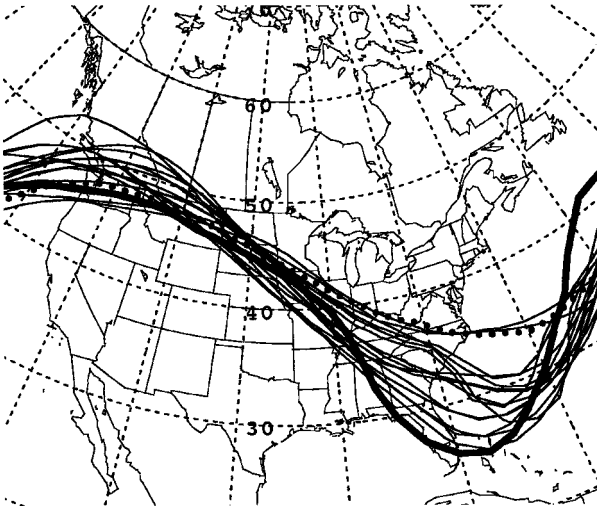

FIG. 4. Same as Fig. 3 except for 4.5-day lead time forecasts valid at 1200 UTC 13 Jan 1996.

At days 5.5 and 4.5, the uncertainty in the position of the wave is further reduced, and with day 3.5 or shorter lead times, the small ensemble spread indicates (and rightly so) high confidence in the forecasts. The degree of confidence will, of course, vary from day to day. One of the advantages of using an ensemble of forecasts is that we can learn about such changes in advance, before the forecasts actually verify. In contrast to the high confidence suggested by the 4.5-day forecast panel of Fig. 3, Fig. 4 shows an example where an East Coast storm at 4.5-day lead time was associated with large uncertainty in the ensemble. In this case (the second East Coast snow storm of 1996), the high-resolution control forecast indicated a much faster and less intense storm than what actually occurred, while some ensemble members were quite accurate. In situations like this, the forecaster has to realize that the situation is less certain, and the uncertainty also has to be conveyed in some form (alternate scenarios, probabilistic forecasts, etc.) to the users.

Concerning Fig. 3 again, at 3.5 days or shorter lead times, all forecasts are close together, indicating that initial errors are not likely to cause much error in the forecasts. The verification line is slightly north of the forecasts at the center of the trough, which may point to a possible case-dependent bias in the medium-range forecast (MRF) model, especially at lower resolution.

Experience with ensemble forecasts indicates that they can be effectively used to identify different forecast scenarios beyond that offered by the control forecast and to estimate the likelihood of these different solutions. More extensive subjective (Toth et al. 1997) and objective (Zhu et al. 1996) evaluations of ensemble forecasts confirm that when many ensemble members support a particular forecast pattern, then that pattern is likely to verify. In short, ensembles can be used to make reliable probabilistic forecasts, which also means that high and low confidence situations can be distinguished at the time forecasts are made. This can reduce forecast failures that would otherwise occur if only a single forecast were used.

5. Summary

An ensemble forecast is a collection of forecasts that all verify at the same time. Since December 1992, both NCEP and ECMWF have computed global ensemble forecasts where each forecast starts from different initial conditions. Experiments with subsynoptic-scale ensemble forecasting are also under way.

Ensemble forecasting is useful for improving average forecast accuracy, suggesting a range of possibilities and their probability, estimating the decay of forecast skill as a function of forecast lead time, and even suggesting where extra observations could be targeted.

One challenge with ensemble forecasting is to balance model sophistication and the number of forecasts in an ensemble, given that time and computer constraints will always exist. Another challenge is the selection and display of ensemble forecast information. Readers, especially operational forecasters, are invited to contribute comments and suggestions to the third author (e-mail: Zoltan.Toth@noaa.gov).

As the sophistication of numerical models improves along with the computer power needed to run them, forecasts should exhibit fewer and fewer biases. Ensemble forecasting is expected to become an increasingly important tool to characterize the nonsystematic errors that remain.

Acknowledgments. We thank (in alphabetical order) Jeffrey Anderson, Harold Brooks, Roberto Buizza, Brian Farrell, Tom Gleeson, Anthony Hollingsworth, Petros Ioannou, Eugenia Kalnay, Tim Palmer, Bernard Straus, Jeff Whitaker, and Richard Wobus for useful discussions. Richard Wobus prepared the final versions of Figs. 3 and 4. The first two authors acknowledge support from the NOAA/FSU Cooperative Institute for Tropical Meteorology, from NSF Grant ATM-9222595, and from The Florida State University.

REFERENCES

Åkesson, O., 1996: Comparative verification of precipitation probabilities from the ECMWF ensemble prediction system and from the operational T213 forecast. Preprints, *11th Conf. on Numerical Weather Prediction,* Norfolk, VA, Amer. Meteor. Soc., J31–J34.

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate,* **9,** 1518–1530.

Bishop, C. H., and Z. Toth, 1996: Using ensembles to identify observations likely to improve forecasts. Preprints, *11th Conf. on Numerical Weather Prediction,* Norfolk, VA, Amer. Meteor. Soc., 72–74.

Bresch, J. F., and J.-W. Bao, 1996: Generation of mesoscale ensemble members by varying model physics. Preprints, *11th Conf. on*

Numerical Weather Prediction Norfolk, VA, Amer. Meteor. Soc., 62–64.

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.

——, D. J. Stensrud, and C. A. Doswell III, 1996: Application of short-range NWP model ensembles to severe storm forecasting. Preprints, *18th Conf. on Severe Local Storms,* San Francisco, CA, Amer. Meteor. Soc., 372–375.

Buizza, R., 1997: Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system. *Mon. Wea. Rev.,* **125,** 99–119.

Daley, R., 1991: *Atmospheric Data Analysis.* Cambridge University Press, 457 pp.

——, and R. M. Chervin, 1985: Statistical significance testing in numerical weather prediction. *Mon. Wea Rev.,* **113,** 814–826.

Dubreuil, P., 1996: Overview of NWP research directions at Environment Canada. Preprints, *11th Conf. on Numerical Weather Prediction,* Norfolk, VA, Amer. Meteor. Soc., 9J–11J.

Emanuel, K. A., E. N. Lorenz, and R. E. Morss, 1996: Adaptive observations. Preprints, *11th Conf. on Numerical Weather Prediction,* Norfolk, VA, Amer. Meteor. Soc., 67–69.

Epstein, E. S., 1969: Stochastic dynamic prediction. *Tellus,* **21,** 739–759.

Fuller, J. F., 1990: *Thor's Legions.* Amer. Meteor. Soc., 443 pp.

Gleeson, T. A., 1961: A statistical theory of meteorological measurements and predictions. *J. Meteor.,* **18,** 192–198.

——, 1970: Statistical-dynamic predictions. *J. Appl. Meteor.,* **9,** 333–344.

Hamill, T. M., and S. J. Colucci, 1997a: Evaluation of eta–RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.,* in press.

——, and ——, 1997b: Verification of eta–RSM short-range ensemble forecasts. *Mon. Wea. Rev.,* **125,** 1312–1327.

Harrison, M. S. J., 1996: Ensemble forecasting at the UK Meteorological Office. Preprints, *11th Conf. on Numerical Weather Prediction,* Norfolk, VA, Amer. Meteor. Soc., 21J–23J.

Hoffman, R. N., and E. Kalnay, 1983: Lagged average forecasting, an alternative to Monte Carlo forecasting. *Tellus,* **35A,** 100–118.

Hollingsworth, A., A. Lorenc, S. Tracton, K. Arpe, G. Cats, S. Uppala, and P. Kalberg, 1985: The response of numerical weather prediction systems to FGGE level IIb data. Part I: Analysis. *Quart. J. Roy. Meteor. Soc.,* **111,** 1–66.

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system approach to ensemble forecasting. *Mon. Wea. Rev.,* **124,** 1225–1242.

Kalnay, E., and Z. Toth, 1996: Ensemble prediction at NCEP. Preprints, *11th Conf. on Numerical Weather Prediction,* Norfolk, VA, Amer. Meteor. Soc., 19J–J20.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.,* **102,** 409–418.

Lorenz, E. N., 1963: Deterministic non-periodic flow. *J. Atmos. Sci.,* **20,** 130–141.

——, 1965: A study of the predictability of a 28-variable atmospheric model. *Tellus,* **17,** 321–333.

——, 1969: The predictability of a flow which possesses many scales of motion. *Tellus,* **21,** 289–307.

——, 1993: *The Essence of Chaos.* University of Washington Press, 227 pp.

Molteni, F., 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.

Palmer, T. N., 1993: Extended-range atmospheric prediction and the Lorenz model. *Bull. Amer. Meteor. Soc.,* **74,** 49–65.

——, R. Mureau, and F. Molteni, 1990: The Monte Carlo forecast. *Weather,* **45,** 198–207.

Reynolds, C., P. J. Webster, and E. Kalnay, 1994: Random error growth in NMC's global forecasts. *Mon. Wea Rev.,* **122,** 1281–1305.

Richardson, D. S., M. S. J. Harrison, K. B. Robertson, and A. P. Woodcock, 1996: Joint medium-range ensembles using UKMO, ECMWF and NCEP ensemble systems. Preprints, *11th Conf. on Numerical Weather Prediction,* Norfolk, VA, Amer. Meteor. Soc., 26J–27J.

Shaw, R. H., and W. Innes, Eds., 1984: *Some Meteorological Aspects of the D-Day Invasion of Europe: Proceedings of a Symposium.* Amer. Meteor. Soc., 170 pp.

Shaw, R. S., 1981: Strange attractors, chaotic behavior, and information flow. *Z. Naturforschung,* **36A,** 80–112.

Simmons, A. J., R. Mureau, and T. Petroliagis, 1995: Error growth and estimates of predictability from the ECMWF forecasting system. *Quart. J. Roy Meteor. Soc.,* **121,** 1739–1771.

Stoss, L. A., and S. L. Mullen, 1995: The dependence of short-range 500-mb height forecasts on the initial flow regime. *Wea. Forecasting,* **10,** 353–368.

Tatarskiy, V. I., 1969: The use of dynamic equations in the probability prediction of the pressure field. *Izv. Acad. Sci. USSR, Atmos. Oceanic, Phys.,* **5,** 293–297.

Toth, Z., 1991: Intercomparison of circulation similarity measures. *Mon. Wea Rev.,* **119,** 55–64.

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

Tracton, M. S., and E. Kalnay, 1993: Operational ensemble prediction at the National Meteorological Center: Practical aspects. *Wea. Forecasting,* **8,** 379–398.

——, J. Du, Z. Toth, and H. Juang, 1997: Short range ensemble forecasting (SREF) at NCEP/EMC. Preprints, *12th Conf. on Numerical Weather Prediction,* in press.

Van den Dool, H., and L. Rukhovets, 1994: On the weights for an ensemble-averaged 6–10 day forecast. *Wea. Forecasting,* **9,** 457–465.

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

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

——, Z. Toth, S. Tracton, and E. Kalnay, 1996: How the NCEP ensemble works: Synoptic examples. Preprints, *11th Conf. on Numerical Weather Prediction,* Norfolk, VA, Amer. Meteor. Soc., 29J–30J.

Zhu, Y., G. Iyengar, Z. Toth, M. S. Tracton, and T. Marchok, 1996: Objective evaluation of the NCEP global ensemble forecasting system. Preprints, *11th Conf. on Numerical Weather Prediction,* Norfolk, VA, Amer. Meteor. Soc., 79J–82J.