Extended-Range Probability Forecasts Based on Dynamical Model Output

JIANFU PAN AND HUUG VAN DEN DOOL

Climate Prediction Center, NCEP/NWS/NOAA, Camp Springs, Maryland

(Manuscript received 9 June 1997, in final form 10 April 1998)

ABSTRACT

A probability forecast has advantages over a deterministic forecast as the former offers information about the probabilities of various possible future states of the atmosphere. As physics-based numerical models find their success in modern weather forecasting, an important task is to convert a model forecast, usually deterministic, into a probability forecast. This study explores methods to do such a conversion for NCEP's operational 500-mb-height forecast and the discussion is extended to ensemble forecasting. Compared with traditional model-based statistical forecast methods such as Model Output Statistics, in which a probability forecast is made from statistical relationships derived from single model-predicted fields and observations, probability forecasts discussed in this study are focused on probability information directly provided by multiple runs of a dynamical model—eleven 0000 UTC runs at T62 resolution.

To convert a single model forecast into a strawman probability forecast (single forecast probability or SFP), a contingency table is derived from historical forecast–verification data. Given a forecast for one of three classes (below, normal, and above the climatological mean), the SFP probabilities are simply the conditional (or relative) frequencies at which each of three categories are observed over a period of time. These probabilities have good reliability (perfect for dependent data) as long as the model is not changed and maintains the same performance level as before. SFP, however, does not discriminate individual cases and cannot make use of information particular to individual cases. For ensemble forecasts, ensemble probabilities (EP) are calculated as the percentages of the number of members in each category based on the given ensemble samples. This probability specification method fully uses probability information provided by the ensemble. Because of the limited ensemble size, model deficiencies, and because the samples may be unrepresentative, EP probabilities are not reliable and appear to be too confident, particularly at forecast leads beyond day 6. The authors have attempted to combine EP with SFP to improve the EP probability (referred to as modified forecast probability). Results show that a simple combination (plain average) can considerably improve upon both the EP and SFP.

1. Introduction

A traditional dynamic forecast provides a definite future state of the atmosphere and does not, as such, offer information about how much the forecast should be trusted. A probability forecast, on the other hand, not only predicts the possible future states but also estimates the chances that these states will occur. A forecast of a system like the atmosphere should be looked upon as probabilistic rather than deterministic. The stochastic nature of the system requires forecasting probabilities of a number of probable future states instead of a definite single state. Also, the forecast is not perfect even if the system were deterministic, because of errors in data and methods used in making the forecast. The probability forecast can and should convey information about the accuracy of the forecast. For example, if 80% of the forecasts made by a model are verified to be categorically correct over a period of time when the model predicts warmer than normal temperatures, it is reasonable to make a forecast for warmth with a probability of 80% the next time the model predicts warm.

Traditionally, probability forecasts have been made based on statistical relationships derived from historical single forecast–verification data, and the physics of the system is not explicitly involved in converting a single forecast into a probability forecast. In contrast, a dynamic forecast takes a snapshot of the system and moves the system forward based on its knowledge of the physics. Among many other problems, such as the complexity of models and the difficulties in solving nonlinear equations, one disadvantage of the dynamic forecast is that probabilistic features of the forecast are difficult to infer from a single model run. Ensemble forecasts developed in recent years are designed to capture nondeterministic aspects of the atmosphere, but the interpretation of the probabilities provided by the ensemble may not be straightforward. The purpose of this study is twofold: to explore methods translating a physics-based dynamic forecast into a probability forecast and to examine probabilities provided by ensembles. Spe-

Corresponding author address: Dr. Huug van den Dool, Climate Prediction Center, NCEP/NWS/NOAA, Camp Springs, MD 20746.
E-mail: wd51hd@sgi84.wwb.noaa.gov

cifically, the question being investigated is how to formulate probabilities based on probability information in dynamic forecasts, especially with ensemble forecasting becoming operational in several weather centers around the world.

Traditional practice of model based probability forecasts includes perfect-prog, Model Output Statistics (MOS; Glahn and Lowry 1972), and other postprocessors (see Wilks 1995 for a review of these methods). A common feature of these traditional methods is that the probability forecast is made from statistical relationships developed from a series of paired observation and model forecast data, whereby model forecasts are treated as the predictors. A modern example of a physics-based probability forecast is the ensemble forecast, in which a number of forecasts are made from slightly different initial conditions (Tracton and Kalnay 1993). The differences among the members, or spread, are small initially, simulating the uncertainties in initial condition. The spread grows as the model integrations proceed, reflecting the growth of uncertainty in the forecast, and eventually become saturated at a level that is equivalent to the spread of randomly chosen states of the atmosphere. Ideally, the ensemble spread before the saturation level is expected to reflect the evolution of the atmosphere into a number of possible states and the ensemble will provide direct probabilistic information about future states. While it may be generally true that the state that most members agree on is more likely to occur, the use of the ensemble probability may not be straightforward as the information carried by the ensemble is not completely understood. For instance, probabilities from the ensemble are found to be unrealistically confident. Proper probability specification for the ensemble forecast is thus an outstanding problem.

In this study, we shall examine probability information in dynamical forecasts and use them to make a probability forecast. Two situations are considered in this study: a single model forecast and an ensemble of forecasts. Methods are explored for both situations. The single model case serves as a control for the ensemble forecast. The methods are applied to yield 500-mb heights (Z500) probability forecast for grid points over the Northern Hemisphere.

In section 2, we describe the data and methodologies used in the formulation and verification of probability forecasts. Section 3 discusses strategies of specifying probabilities given a model forecast, while section 4 presents verifications with different formulation strategies. Finally, in section 5, we present a summary and a brief discussion about our findings.

2. Method and data

a. Data

The primary data used in this study are daily Z500 at 0000 UTC from National Centers for Environmental Prediction–National Center for Atmospheric Research (NCEP–NCAR) reanalyses from five winter seasons (1 December to 28 February) from January 1985 to February 1989 (Kalnay et al. 1996), and the Climate Prediction Center's (CPC) reforecast data for the same 5-yr period (Schemm et al. 1996). The reforecasts are made with NCEP's reanalysis model starting from the daily reanalyses at 0000 UTC as initial conditions. The model resolution used is T62L28 (triangular truncation up to the maximum wavenumber of 62 and 28 levels in the vertical) and the forecasts are made out to 50 days from 0000 UTC initial conditions from 1 January 1985 to 28 February 1989. Forecasts up to day 15 are used in this study. These data are used to calculate climatologies and standard deviations of analyses and forecasts, and also to derive probabilities (see section 3a). To study the probability forecast based on the ensemble forecast, NCEP's operational ensemble forecasts from the 1994/95 to 1996/97 winters are used, along with the analysis data for the same period. NCEP's operational ensemble forecast consists of 17 individual runs per day, each with forecasts out to 16 days. Five of the 17 members are forecasts from yesterday at 1200 UTC, while the other 12 are initialized at 0000 UTC. One 0000 UTC member is an extension of a high-resolution run (T126L28, or triangular truncation up to wavenumber 126 and 28 levels in the vertical) beyond day 7, when the resolution is reduced to T62L28. The remaining 11 members all have resolutions of T62L28 and include a control run and five pairs of perturbation runs from the control. In this study, we choose to use only the 11 T62L28 0000 UTC members as they represent a ''consistent ensemble.'' All data used in this study are gridded fields with resolution of 2.5° lat $\times$ 2.5° long.

b. Standardization of anomalies

The forecast variable chosen is the 500-mb height anomaly, with ''anomaly'' defined as the departure from sample climatology. For dynamic forecasts, a common problem is systematic error or climate drift, which is defined as the algebraic mean difference between forecast and analysis (forecast–analysis) at individual grid points (Tracton et al. 1989; Klinker and Capaldo 1986). Countless previous investigations have shown that there were negative systematic height errors over most of the Northern Hemisphere particularly over the lower latitudes, and the NCEP model in recent years is no exception (Chen and van den Dool 1995a). Because of the negative bias, the distribution of the forecast anomalies tends to be shifted toward the negative side of the anomaly probability curve. That is, more negative anomalies are observed in forecasts than in analyses over the same period of time. Furthermore, climate simulations show discrepancies in variability between the model forecasts and the corresponding analyses (Chen and van den Dool 1995b). These errors in variability also affect the distribution of anomalies. Figure 1 shows the 5-yr refore-

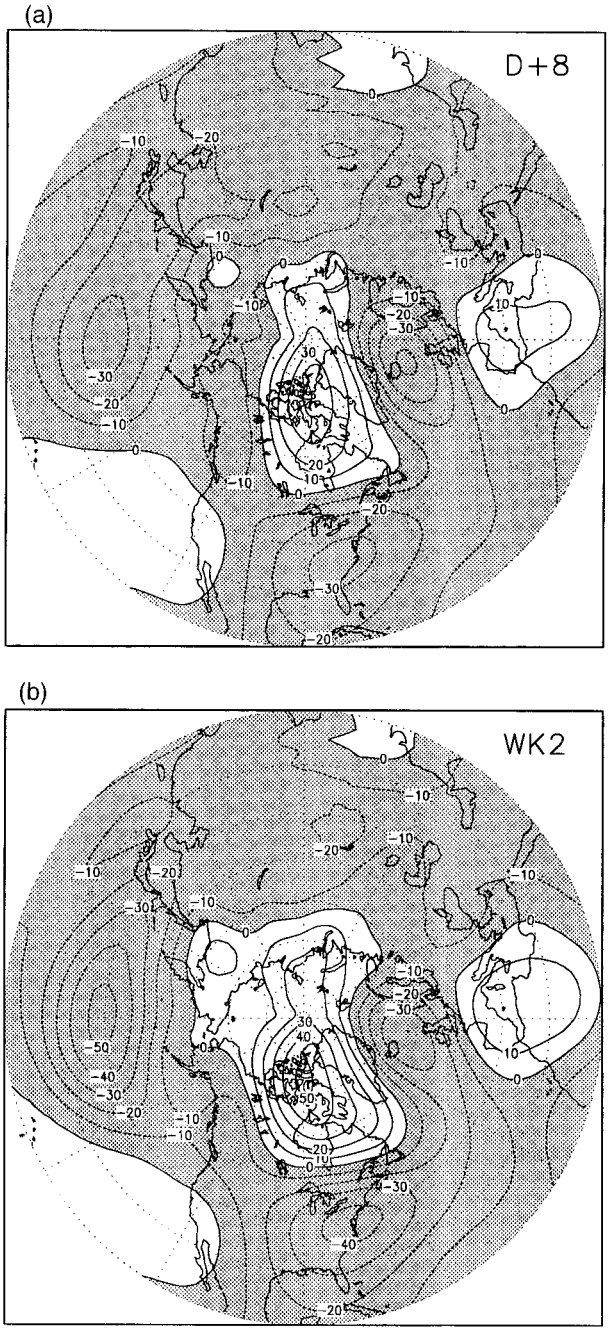

FIG. 1. Reforecast systematic error derived from the five winters (Dec–Feb, Jan 1985–Feb 1989). (a) D + 8 forecasts and (b) WK2. Units are meters. Contour interval is 10 m. Areas with negative errors are shaded.

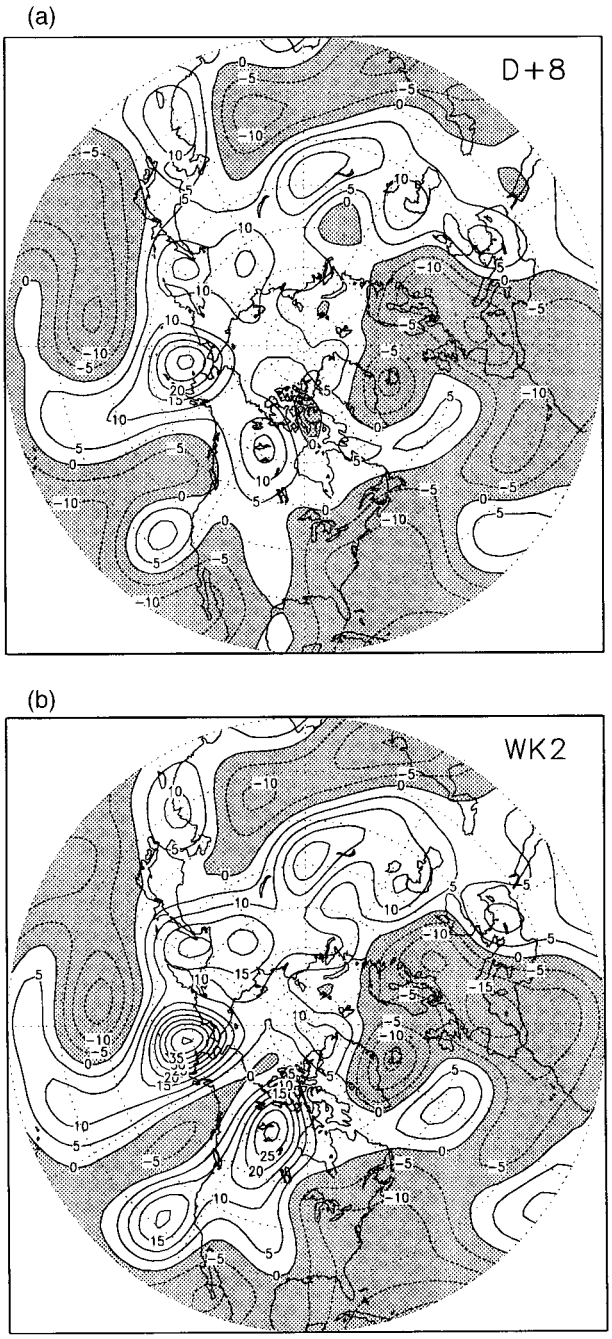

FIG. 2. Reforecast standard deviation percentage error derived from the five winters (Dec–Feb, Jan 1985–Feb 1989). (a) D + 8 forecast, and (b) WK2. Contour interval is 5%. Areas with negative errors are shaded.

cast systematic errors at the ranges of D + 8 (average of day 6 to day 10) and WK2 (average of day 8 to day 14). The values are averages of forecast minus analysis over the three winter months averaged over 5 yr. Negative errors cover most of the Northern Hemisphere except the polar region, the west coast of North America, and the southwestern part of Europe. Large negative errors are observed over the central North Pacific and southeastern part of North America. Figure 2 presents percentage errors in standard deviations for the same two forecast ranges. In general, forecasts have higher variability over the higher latitudes of continents except

in western Europe. Mostly the errors are less than 20% except in the far northern Pacific Ocean in both D + 8 and WK2, and over Canada in WK2.

To alleviate these problems to a certain extent, forecast anomalies are standardized with lead-time-dependent *forecast climatology* and *forecast standard deviation*, while the standardization of analysis anomaly is made from *observed climatology* and *observed standard deviation*. Standardized anomalies can be expressed as the following:

$$\text{for analysis,} \qquad z'_a = (z_a - \text{zc}_a)/\text{sd}_a;$$

$$\text{for forecast,} \quad z'_f(\tau) = [z_f(\tau) - \text{zc}_f(\tau)]/\text{sd}_f(\tau); \quad (1)$$

where z'_a and z'_f are standardized anomalies of analysis heights (z_a) and forecast heights (z_f), respectively; zc_a and sd_a are observed climatology and standard deviation; and $\text{zc}_f(\tau)$ and $\text{sd}_f(\tau)$ are forecast climatology and standard deviation at lead time τ. The climatology data calculated for each day of the year are derived from harmonically smoothed 5-yr daily averages. The standardization with (1) can only be done if we have multiyear forecast data from the same model.

c. Definition of categories

In this study, we consider a three-class categorical forecast of 500-mb-height anomalies. The three equal classes, below normal (B), normal (N), and above normal (A), are derived based on the normal distribution. The three classes are ''equal'' in terms of their occurrence likelihood in a normal distribution. For a standardized normal distribution with zero mean and unitary standard deviation, the boundaries for the three equal classes are ± 0.4308. The definitions of the three classes are

$$\begin{aligned} &\text{class B} && z' \le -0.4308, \\ &\text{class N} && -0.4308 < z' < 0.4308, \\ &\text{class A} && z' \ge 0.4308, \end{aligned} \quad (2)$$

where z' is standardized anomaly and the definition applies to both forecasts and analyses. The authors are aware that the 500-mb distribution is not precisely normal in certain areas, but do not think it is worth the effort to be more precise with only 5 yr of data.

d. Time-mean forecast

Current operational medium-range efforts at NCEP's CPC is directed toward D + 8, but D + 8 will probably be replaced by WK2 in the near future (O'Lenic et al. 1996). To make a time-mean forecast such as D + 8 or WK2, the climatology and standard deviation need to be available for the time-mean fields. We calculated D + 8 and WK2 model climatologies, as well as the 5- and 7-day mean analysis climatology and standard deviation.

TABLE 1. Single forecast probability table. A probability is calculated as the percentage of verified cases over all cases of the forecast category.

Analysis / Forecast	B	N	A
B	p_{11}	p_{12}	p_{13}
N	p_{21}	p_{22}	p_{23}
A	p_{31}	p_{32}	p_{33}

3. Probability formulation

Probability formulation is a process that determines probabilities of future states given one or more forecasts, made by a model or otherwise. The future states are divided here into the three predefined categories. Given a model forecast, the forecast probabilities are to be assigned to each of the three forecast categories (the three probabilities add up to 1), usually with the biggest probability value assigned to the category into which the model forecast falls.

a. Single forecast probability (SFP)

A single model forecast provides a single value that falls into one of our three categories. A simple but laborious prescription of probabilities for a single model forecast is to construct a contingency table based on historical forecast and analysis data. A probability table is derived by calculating the frequency distribution for each forecast category from the historical data. For example, in all cases where the model predicts category B, 60% of the cases are categorically verified as B, 20% are N, and 20% are A (numbers here are fictitious). Probability specification when the model forecasts B on a future occasion is then simply (B, N, A) = (0.6, 0.2, 0.2). Likewise, probabilities can be derived for forecast categories N and A. The probability table for a single forecast is shown in Table 1, in which the following is satisfied:

$$p_{i1} + p_{i2} + p_{i3} = 1, \qquad i = \{1, 2, 3\}.$$

Ideally, each column in Table 1 should also add up to 1. The discrepancy in the distributions between analyses and forecasts causes the sums of the columns to not be exactly 1. The values on the diagonal in the table (p_{ii}) are probabilities of correct forecasts in each category in historical cases and are a measure of the model skill, albeit on dependent data. The familiar Heidke score is based on the diagonal p_{ii}. In a typical table with forecast skill, the values p_{11}, p_{22}, and p_{33} are the largest compared to the other two values in their respective rows. The terms p_{13} and p_{31} should be the smallest. In a situation of negligible forecast skill all p_{ij}'s approach ⅓, except for (minor) distributional discrepancies and sampling.

In making an SFP, the specification for a single model forecast is made by first determining its forecast category. Probabilities are then obtained from the corresponding row in the probability table. For instance, if

the model predicts category A, the probability forecast will be (B, N, A) = (p_{31}, p_{32}, p_{33}).

In the past, a similar probability table was derived from CPC's historical operational long-range forecasts, which were subjective forecasts made from a variety of tools including statistical methods, model forecasts, and forecasters' experience, and was used to guide probability specification for the forecast (Gilman 1986). In this study, we try to establish an objective probability table for dynamical forecasts.

One noticeable drawback of SFP is that the method does not discriminate one case from another once a model forecast category is determined, since the probability table represents ''climatological'' results for the model. If some cases are more predictable than others, SFP will not reveal that information. Medium-range dynamical forecast practice has demonstrated (or at least strongly suggested) that there is a large variability in forecast skill from case to case, because of the variability in predictability of different regimes (Molteni and Tibaldi 1990; Tracton 1990). It is desirable that cases with higher predictability be identified ahead of time and predicted accordingly with higher confidence. This is particularly important at longer forecast lead times like WK2, at which average forecast skills become marginal by any standard. Since the introduction of the ensemble forecasting strategy at NCEP in late 1992, additional information has become available about individual cases.

SFP can be viewed as a simple MOS approach. However, it is not our intention to construct an MOS formulation for its own sake. Instead, SFP is used to provide a realistic and objective control measure of the probability forecasts.

b. Ensemble probability (EP)

Unlike the single forecast where the solution of the future state is unitary, an ensemble forecast provides a probabilistic solution of the future. These forecasts fall into the different categories and reflect different probable states of the atmosphere. A probability forecast from an ensemble is constructed by calculating the percentage of ensemble members in each category. These percentages, used as ensemble probabilities and referred to as EP = (p_B, p_N, p_A), exclusively use information provided by the ensemble and do not use any historical data as in SFP except for correction of mean and standard deviation. They are virtually identical in information to the spread of the ensemble.

While qualitatively EP reflects the probabilities that the categories will happen in the future, the reliability of these probability values are questionable. In most cases, the ensemble probabilities are too confident and the ensemble spread is too small to catch future reality. Techniques have been developed to generate dynamically efficient initial perturbations. These techniques are apparently somewhat effective as ensemble mean forecasts are found to be better than the control forecast (van den Dool and Rukhovets 1994). Surprisingly few studies have been made concerning the probability features of the ensemble. A binned probability ensemble (BPE) technique was proposed by Anderson (1996b). In BPE, ensemble forecasts are used to partition the forecast into a number of bins, each of which has an equal probability of containing the ''true'' forecast. The method evaluates the consistency of ensemble predictions and observations. Ensemble probability forecast for short ranges was discussed by Hamill and Colucci (1996). Precipitation probabilities of ensemble forecast were studied by Akesson (1996). To estimate the probabilities from the ensemble, a certain minimum sample size is required so that the estimation is statistically meaningful. The choice of initial perturbations is currently constrained by dynamical and computational constraints. The ensemble size is limited in forecast operation because making many dynamical forecasts is still expensive in terms of computer resources even at low resolution.

c. Modified forecast probability (MFP)

The previous two methods are totally different approaches to constructing a probability forecast from model outputs in terms of how the current forecast and historical data are used. SFP bases its probability forecast on the model forecast history once a model forecast is given. The probability is reliable and appropriate for a single model run. On the other hand, EP exclusively relies on the ensemble spread specific to the case. It is conceivable that the two types of probabilities should be combined to provide an improved probability forecast.

We believe that both SFP and EP provide valuable information on the probability on the future states. However, we struggled to find a way to combine the two with a solid theoretical basis. As one such effort, an averaged specification is made by first applying the SFP table to individual ensemble members, assuming the ensemble members have compatible performance as individual forecasts. This is the main reason we took only 0000 UTC T62L28 members. The probabilities from individual members are then averaged according to the EP probabilities. The procedure can be expressed as

$$\begin{aligned} p'_B &= p_B p_{11} + p_N p_{21} + p_A p_{31}, \\ p'_N &= p_B p_{12} + p_N p_{22} + p_A p_{32}, \\ p'_A &= p_B p_{13} + p_N p_{23} + p_A p_{33}, \end{aligned} \tag{3}$$

where the symbols on the right-hand side are the same as in SFP and EP, and (p'_B, p'_N, p'_A) is the averaged probabilities that satisfies the following relation:

$$p'_B + p'_N + p'_A = 1.$$

This approach is somewhat similar to a consensus

TABLE 2. SFP table over North America (30°–60°N, 80°–125°W) derived from the five-winter reforecast data (Dec–Feb, Jan 1985–Feb 1989).

Observation / Forecast	D + 8			WK2		
	B	N	A	B	N	A
B	0.576	0.288	0.130	0.460	0.329	0.204
N	0.294	0.385	0.314	0.320	0.334	0.339
A	0.120	0.284	0.590	0.190	0.315	0.489

forecast (Vislocky and Fritsch 1995). It turns out that the averaged specification is more conservative than SFP as it has probabilities lower than or equal to p_{11}, p_{22}, or p_{33} in the SFP table. The conservativeness of the averaged specification will cause underforecasting of classes A and B and is not desired. For the sake of argument, we will refer to SFP as a lower bound of probability estimation. Ideally, a probability table like the one in SFP should be derived as a function of ensemble probabilities for ensemble forecast. This requires a very large sample of ensemble forecasts and is practically impossible, at least with the resources available right now. We propose that a new specification can be derived from the averaged probability (p'_B, p'_N, p'_A), which represents an extremely conservative estimation, and the EP probability. If all the members agree on a category, a 100% EP forecast will be made for that category. In general, we believe that the EP represents an approximate upper limit of the probability one can assign to the favored class, the class that the majority of the ensemble members agree on. We further assume that a reasonable probability lies between EP and the averaged probability described in (3). As an experiment, we construct a modified forecast probability (MFP) as the plain mean of the probability in (3) and EP. The new MFP specification, (p''_B, p''_N, p''_A), is simply expressed as

$$p''_B = (p'_B + p_B)/2,$$
$$p''_N = (p'_N + p_N)/2,$$
$$p''_A = (p'_A + p_A)/2. \quad (4)$$

By no means can this be argued to be the optimal combination of the SFP table and the EP probability. However, we hope that the average of the two extreme estimates used in the MFP could provide a first guess of the elusive optimal specification.

4. Evaluation of the methods

In this section, we shall present results of the probability tables derived from the 5 yr of analysis and reforecast data. The evaluation of the three methods discussed in the previous section are made with data from three independent winter seasons from 1994/95 to 1996/97, during which NCEP's operational ensemble forecasts are available. The results are assessed with both a reliability test and the mean-squared-error (mse) measure. We are particularly interested in comparing EP with SFP. Originally, all statistics, including the probability table and probability forecast, are derived at grid points with the resolution of 2.5° lat × 2.5° long. Except when we discuss geographical distributions, however, most results are presented as statistics accumulated over a region to avoid having to arbitrarily pick points from the gridded fields. Another approach is that when we discuss the probability scores for teleconnection indices (see section 4d), a single index stands for a large area (like that covered by PNA), as well, but only the low-frequency variations are represented.

a. The probability table

Given the 5-yr reanalysis and reforecast data, we have derived the SFP tables for daily forecasts of projections from day 1 to 15, as well as for time-mean forecasts of D + 8 and WK2. As an example, Table 2 is the SFP probability table for D + 8 and WK2 forecasts over the North America region from 30° to 60°N and from 80° to 125°W. The D + 8 forecast shows that probabilities of correct forecasts are about 0.25 over the random forecast (⅓) for categories B and A. Probabilities of category N, given a forecast for N, are only marginally higher than that of the random forecast. Probabilities of forecasts missed by one category are slightly below that of the random forecast, while the probabilities are much smaller for forecasts missed by two categories. It is also interesting to note that the four probabilities for forecasts missed by one category have similar magnitudes, and the same is true for forecasts missed by two categories. The rows for extreme categories, B and A, depart significantly more from (⅓, ⅓, ⅓) than the N row, suggesting that N is by far the most difficult category of the three to forecast (for an explanation see van den Dool and Toth 1991). The sums of columns in Table 2 are not exactly 1's because we assumed a normal distribution, but the deviations from 1 are small.

The WK2 forecast shows generally smaller departures from 33.33% than does D + 8 and the differences between the three categories are smaller, as the model loses skill with forecast lead time, and p_{ij} is driven toward ⅓ for all i and j. Nevertheless, the probabilities of the two extreme categories are still impressive, on dependent data, that is, and the probabilities of forecasts missed by two categories remain small compared to the other elements.

It is interesting to study the spatial distribution of the

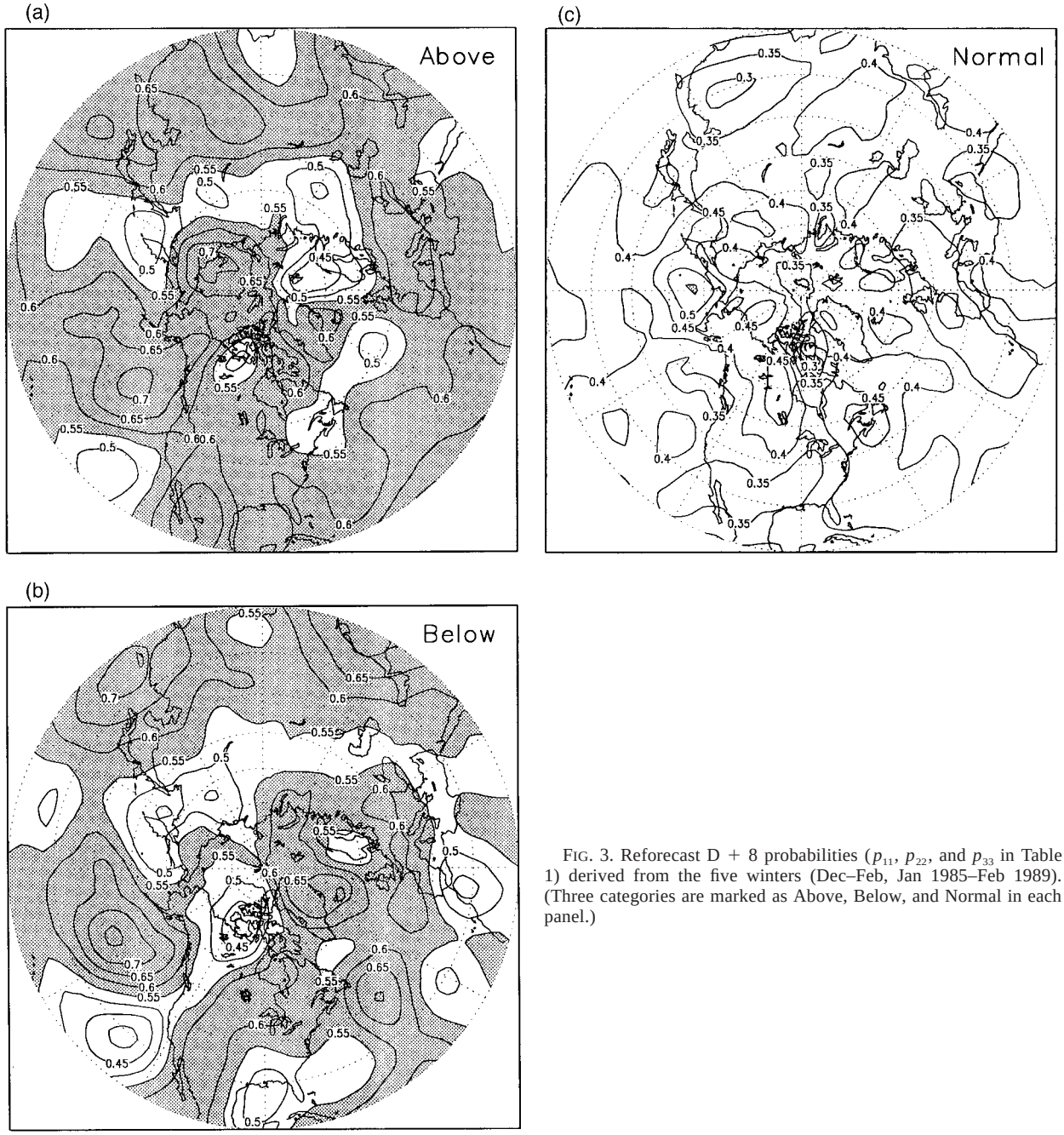

FIG. 3. Reforecast D + 8 probabilities (p_{11}, p_{22}, and p_{33} in Table 1) derived from the five winters (Dec–Feb, Jan 1985–Feb 1989). (Three categories are marked as Above, Below, and Normal in each panel.)

elements of the probability table. As expected, the probabilities show nonuniform distributions in space, as the model has a nonuniform skill distribution over the globe. Figure 3 presents three terms, p_{11}, p_{22}, and p_{33}, of the D + 8 probability table derived from the reforecast and reanalysis data over the five winters. The three terms are, respectively, the probabilities of a correct forecast of category B, N, and A, given a model forecast in the same categories. Since probabilities of a three-category random forecast are ⅓, areas with probabilities larger than this threshold value imply skill over these regions. For category A (marked as ''Above'' in the figure), probabilities over 0.6 are found over the northeastern Pacific, southwest Mexico, western-central Europe around 45°N, and most of southeastern Asia. The category B has a somewhat similar pattern over the Pacific–North America region. However, the center over Europe is not as clear and the high probability region over southeastern Asia is pushed to the coastal areas. Meanwhile, higher probabilities are noted over the At-

lantic. The asymmetry between B and A could be caused by the model's systematic error. For example, it is found that there is large negative bias in heights over western Europe (Pan and van den Dool 1995), which may affect the model's performance of forecasting the category B even after systematic error correction as in section 2b. As is well known, category N seems to be the most difficult class to forecast as the probabilities (Fig. 3c) are much lower than those of the other two categories. The distribution is relatively uniform over the continents.

Figure 4 shows the same probability distributions for the WK2 forecast with the same shading as Fig. 3. Overall, the probability values are expectedly dampened toward 33.33% compared with those of the D + 8 forecast. The regions with relatively high probability, however, are similar to those for D + 8. These regions include the northeastern Pacific, western North Atlantic, and southeastern Asia for categories A and B. Since the model has higher skills over these regions, probabilities above the random value can be assigned at least to the extreme categories (A and B) at the WK2 range, particularly over the oceanic areas. With regard to class N (not shown), there is really no skill for the normal category in WK2 forecast, especially over the continents.

Because of the similarity between B and A, and their distinction from N, we shall combine the cases of B and A into an extreme category (A&B) for verification purposes, and N as a separate category for verification purposes. The combined results are simply the average of B and A. Most of the discussions will be focused on the extreme categories as they are more interesting and rewarding to forecast. It should also be mentioned that forecasting regime transition is a very important part of the extended-range forecast. This is a more difficult subject and is beyond this study.

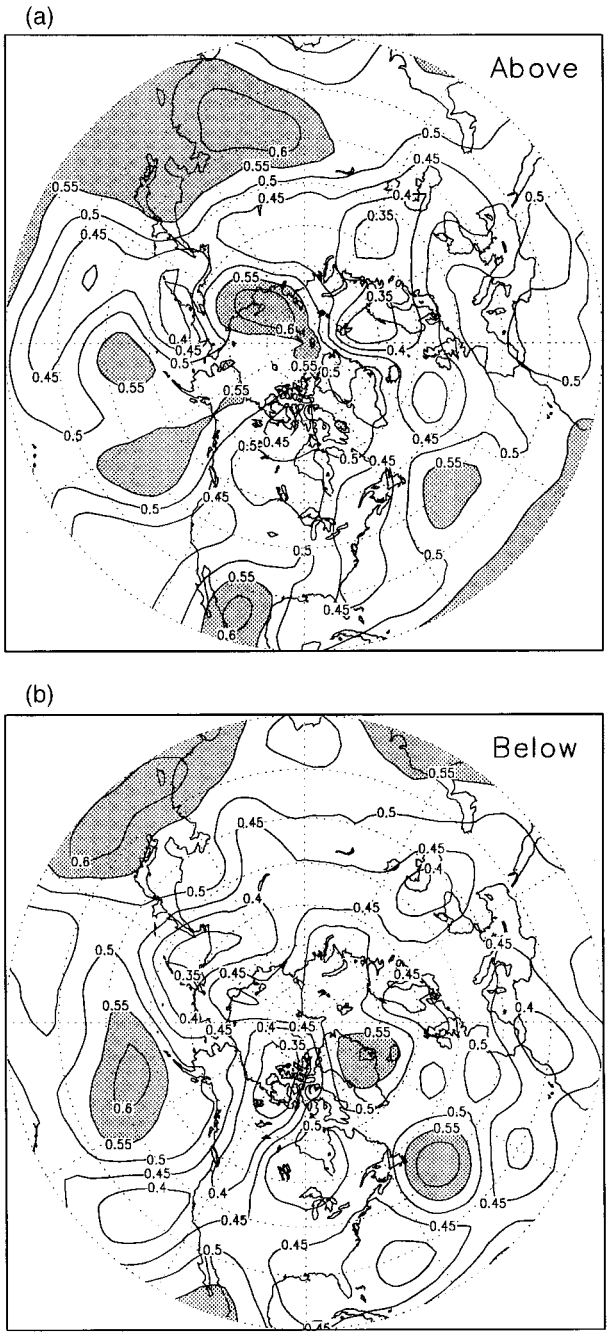

FIG. 4. Reforecast WK2 probabilities (p_{11} and p_{33} in Table 1) derived from the five winters (Dec–Feb, Jan 1985–Feb 1989).

b. Reliability test

A forecast probability is said to be reliable if the probability value is comparable to the observed conditional (or relative) frequency of the forecasted event (Wilks 1995). When the probability values are continuous, as is the case in this study, probability intervals (bins) are used to tally the cases and to calculate the frequency. Here we use bins of width 0.2, that is, 0–0.2 (denoted as 0.1), . . . , 0.8–1.0 (denoted as 0.9). The observed relative frequency is then compared with the forecast probability, and a reliability diagram can be constructed to show the relationship between the observed relative frequencies and corresponding probabilities. The reliability is perfect if the observed relative frequency is the same as the forecast probability. If the observed relative frequency is larger than the probability, the probability is an underspecification or underforecast. In this case there is potential to specify higher probabilities in the forecast. In a probability forecast, the probability should be specified as high or as low as possible as long as it remains reliable. If the observed frequency is lower than the probability, on the other hand, the probability is an overspecification or overforecast. An overspecified probability forecast exaggerates the chance that a category would occur, misleads users, and damages credibility.

Applied to this study, a reliability diagram can be constructed for any of the three forecast categories sep-

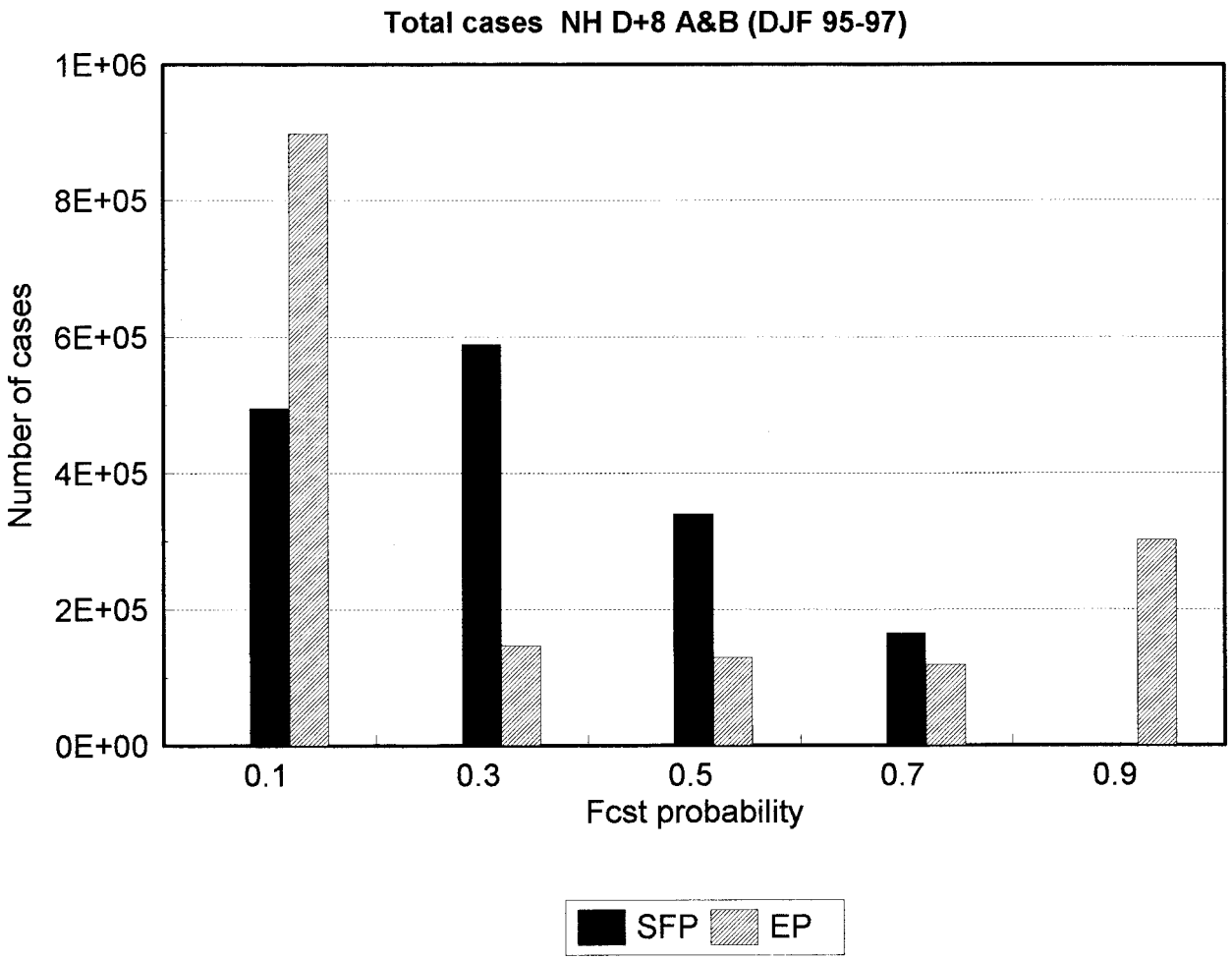

FIG. 5. Histogram of the number of binned cases of 0500 UTC anomaly D + 8 probabilities for categories A and B over the Northern Hemisphere (20°–80°N). Period: Dec–Feb, 1994/95–1996/97.

arately or combined. The table-based SFP has perfect reliability on the dependent dataset from which the table is derived. Except for sampling error, the specification remains reliable for future forecasts as long as the model does not change and maintains the same level of performance as before. In reality, model performance varies according to forecastability of circulation regimes. Also, models never stay quite the same. Improvement of the model often, but not always, leads to an underconfident forecast; however, the changes of the T62L28 operational model since the reanalysis 1994 version have been minimal.

Figure 5 shows the distribution of total number of cases for each forecast probability interval (A&B combined) for three consecutive winters (1995–97). The SFP method shows its conservativeness relative to the EP method as most cases are forecasts with probabilities less than 0.6 and have a maximum near the climatological probability of 0.33. On the other hand, EP is an aggressive specification as it forecasts a considerable number of cases with probabilities larger than 0.8, which will be shown to be overconfident (see Fig. 6 below). A large number of cases in the 0.0–0.2 range and in the 0.8–1.0 range are each other's complement since A and B are combined into one graph here. The remarkable minimum of EP for the 0.3–0.7 range shows that the model suggests much higher skill than can be realized that far ahead.

Figure 6 compares reliability between SFP and EP when applied to D + 8 forecasts for the three winters, and over the Northern Hemisphere from 20° to 70°N. The results combine the two categories B and A. The thick line is where the reliability is perfect. It is noted that, for the three winters, the table-based SFP considerably underforecasts at the probability range of 0.4–0.6, slightly overforecasts at 0.6–0.8, and has no cases for 0.8–1.0 at all. An underforecast in SFP is possible when the model has better skills during the particular forecast period than in the historical period from which the table is derived, or the model was biased in the case of model change. On the other hand, the EP method consistently overforecasts probabilities when the fore-

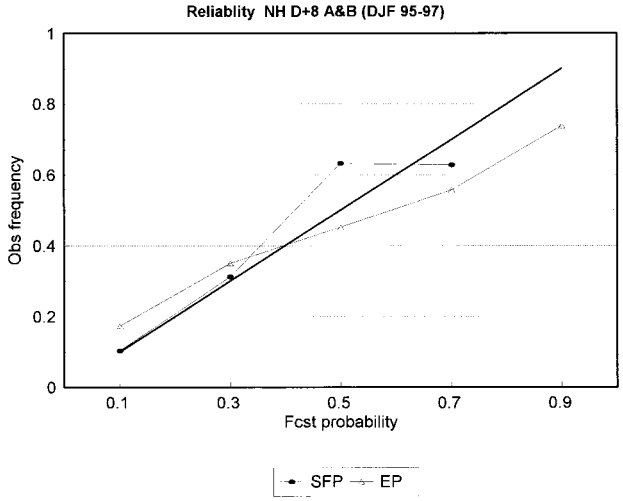

FIG. 6. Reliability diagram of Z500 anomaly D + 8 probability forecast of categories A and B over the Northern Hemisphere (20°–80°N) grids and from Dec–Feb, 1994/95–1996/97.

cast probability is 0.4 or higher, and underforecasts probabilities when the probability is 0.4 or lower. In other words, EP probabilities are too confident and, as there are so many cases (Fig. 5), this is a serious deficiency.

A simple way to adjust the EP probabilities is systematic error correction on the probabilities. For example, Zhu et al. (1997) adjusted the probabilities based on the reliability from a previous period. This virtually requires a probability table for all categories and all probability intervals. For a three-category forecast and the probability interval of 0.2, the matrix size is $3 \times 3 \times 5$. A large sample size is necessary to derive this matrix with enough accuracy.

The corresponding results for the WK2 forecast are shown in Figs. 7 and 8. SFP forecasts show good reliability throughout the probability intervals. The overforecasting by EP is clearly seen at probabilities higher than 0.4. The EP outrageously forecasts a large number of cases with probability of 0.8–1.0, while the reliabilities at WK2 are marginal.

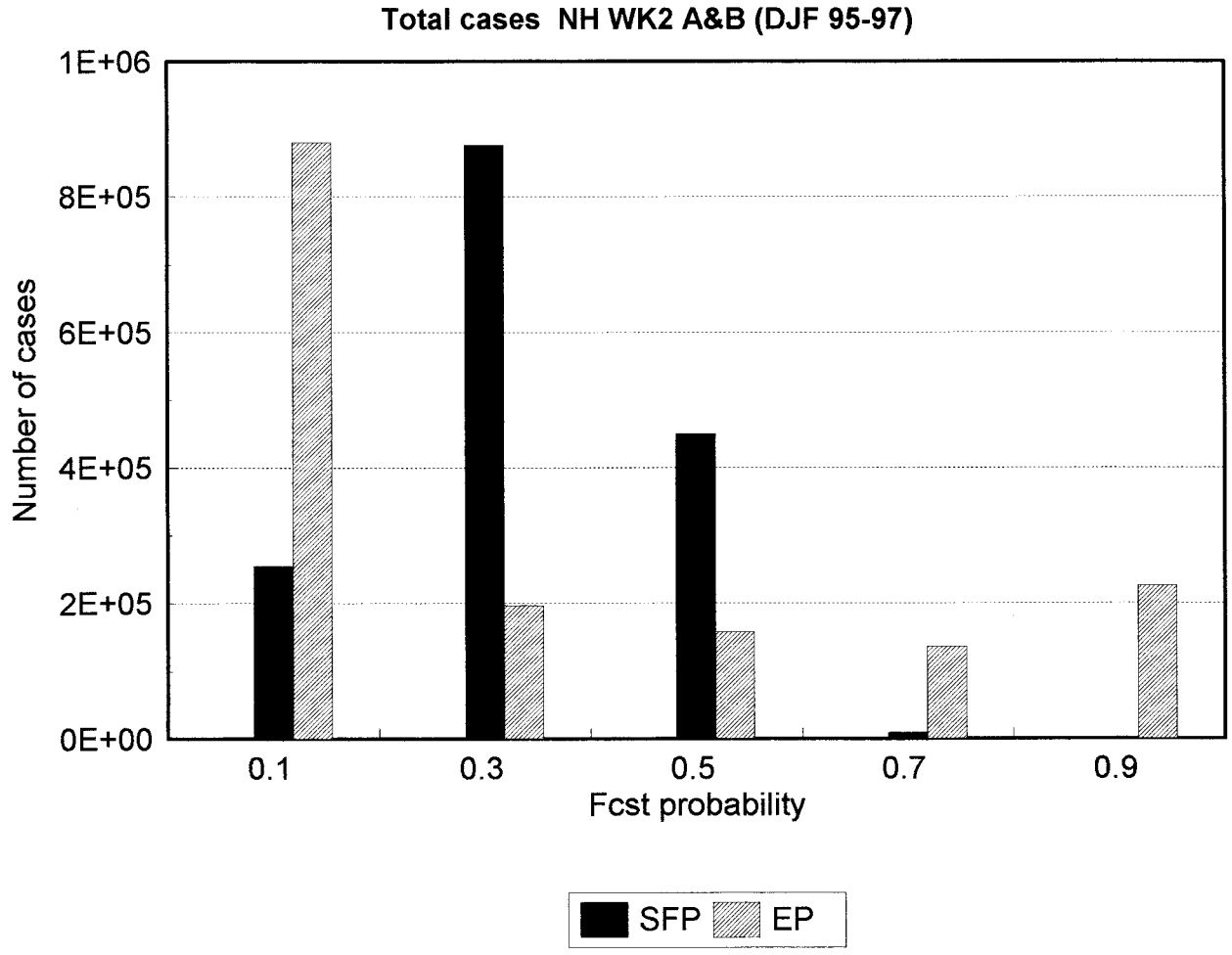

FIG. 7. Same as Fig. 5 except for WK2 forecast.

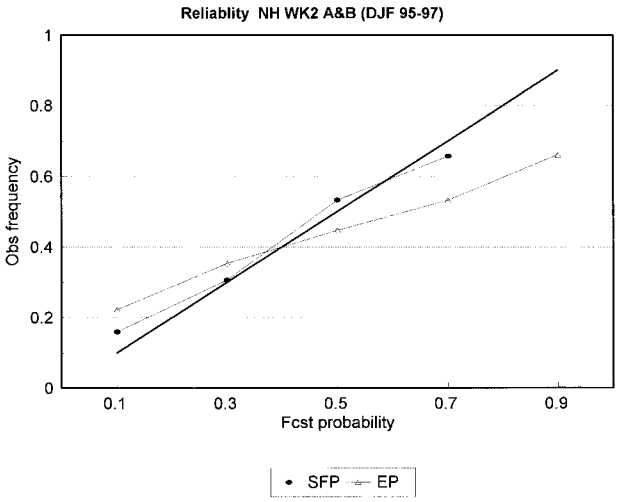

FIG. 8. Same as Fig. 6 except for WK2 forecast.

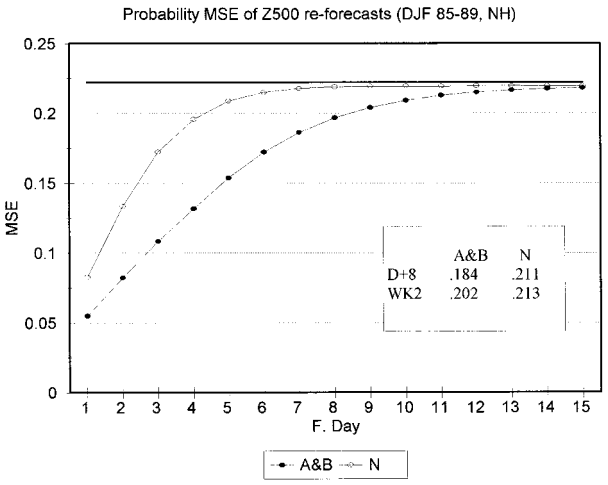

FIG. 9. Mse verification of Z500 anomaly probability forecast for the five-winter reforecasts (20°–80°N).

c. Mean squared error

To measure the skill of a probability forecast, the mse of the probabilities over an area S is defined as

$$\text{mse} = \frac{1}{n}\frac{1}{S}\sum_{s}\sum_{i=1}^{n}(p_{i,s} - \lambda_{i,s})^2, \qquad (5)$$

where s is grid index, i is the category index, and n is total number of categories; $p_{i,s}$ is forecast probability and $\lambda_{i,s}$ is the verification. In our three-category forecast, n is 3. For the three-category forecast, (p_1, p_2, p_3) is also symbolized by (p_B, p_N, p_A) in later discussion, representing the probability forecast of the three categories, B, N, and A. The verification datum $\lambda_{i,s}$ for (B, N, A) has one of three forms: (0, 0, 1), (0, 1, 0), or (1, 0, 0). Lower mse values indicate more accurate forecasts as the larger forecast probabilities are assigned to categories that verify. In a no-skill random forecast (one $p_i = 1$, the other two zero), mse has a value of 0.4444. However, as the no-skill level for the forecast error, we take mse = 0.2222, which is attained by climatological probabilities (all $p_i = ⅓$). The mse defined here is analogous to the ranked probability score (RPS), as both skill measures are essentially an extension of the Brier score to the multicategory probability forecast verification. A more complete discussion on different types of scoring methods can be found in Wilks (1995). Both MSE and RPS are constructed based on the squared differences between forecast probabilities and the verifying binary observations. The difference is that RPS uses cumulative probabilities so as to better reward–punish for absence–occurrence of two-class errors. Mse, however, has the advantage of easy decomposition. That is, partial verification can be made on a subset of the categories. In this study, we shall separate the extreme categories (B and A, $i = 1$ and 3) from the normal category N ($i = 2$). Use of the mse definition in (5) turns out to be convenient to compare different forecast categories. Mse also has commonality with traditional categorical scoring.

In our verification, the mse of probabilities is calculated. As a reference point, the mse for a no-skill three-category forecast, (⅓, ⅓, ⅓), is ⅔ = 0.2222. Forecasts with smaller mse are more accurate. Figure 9 shows the mse of the reforecast (SFP only) over the five winters in the developmental dataset, with forecast lead times from 1 to 15 days, and also for D + 8 and WK2 (the table in the figure). The five winters of reforecasts serve as dependent data for the probability table. Again the two extreme categories are combined and separated from category N. Clearly category N is more difficult to forecast, as its mse approaches that of a no-skill forecast by day 6. The difference is already very large at day 1, indicating the difficulty in describing weak anomalies in the model even at the beginning of the forecast. For the extreme categories, on the other hand, skill is prevalent until about day 12. The D + 8 and WK2 forecasts show some skill for categories B and A, while category N has almost no skill.

Figure 10 shows mse averaged over all three classes for the three independent winters of ensemble forecast, in which the three methods of probability forecast are compared. It is indicated that the forecast is improved by using ensembles at the early stages of the forecast although the gain of EP over SFP is quite modest. As the forecast goes on, the EP method loses control and reaches the no-skill level by day 8 or 9. Clearly, the probability forecast is not realistic by using ensemble probability alone. SFP is much better behaved at larger lead times. It is interesting to notice that the modified method MFP, which is only a simple use of ensemble information, shows improvement over the other two methods in terms of mse. This suggests that SFP is too conservative and EP does provide useful information after some postprocessing. The gains of MFP are the largest at day 6 to 7.

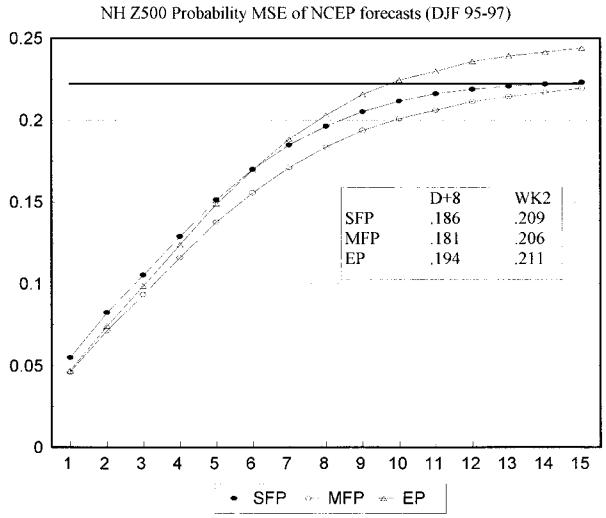

FIG. 10. Mse verification of Z500 anomaly probability forecast for the three-winter NCEP forecasts (20°–80°N).

One interesting feature in Fig. 10 is that EP is better than its control forecast, SFP, at the early stages of the forecast before it becomes worse than the control after about day 6. This is the reverse of what is normally found in terms of anomaly correlation skill, when the ensemble mean is used. The mse skill here and the traditional anomaly correlation skill measure different aspects of the forecast.

d. Application to teleconnection index forecast

Since the probabilities show a nonuniform distribution over the Northern Hemisphere (see Figs. 3 and 4), forecast ability is higher in some places than in others. Forecasts should be focused on areas or flow aspects where the probabilities are higher than average (Branstator et al. 1993). Not surprisingly, these areas correspond to the ones where low frequency patterns are climatologically active. We are therefore motivated to test our probability forecast method in predicting teleconnection patterns. The indices for the PNA pattern, North Atlantic oscillation (NAO), and Eurasian (EU) pattern are defined with the standardized anomalies:

TABLE 4. Mse verification of index forecast using MFP (results are average for the winters of 1954/95–1996/97).

Lead Index	D + 8		WK2	
	B & A	N	B & A	N
PNA	0.160	0.237	0.167	0.242
NAO	0.148	0.214	0.169	0.210
EU	0.135	0.215	0.136	0.242

$$\mathrm{PNA} = (1/4)[z(20\mathrm{N}, 160\mathrm{W}) - z(45\mathrm{N}, 165\mathrm{W}) + z(55\mathrm{N}, 115\mathrm{W}) - z(30\mathrm{N}, 85\mathrm{W})]$$

$$\mathrm{EU} = -(1/4)z(55\mathrm{N}, 20\mathrm{E}) + (1/2)z(55\mathrm{N}, 75\mathrm{E}) - (1/4)z(40\mathrm{N}, 145\mathrm{E})$$

$$\mathrm{NAO} = (1/2)[z(37.5\mathrm{N}, 25\mathrm{W}) - z(65\mathrm{N}, 22.5\mathrm{W})],$$

where PNA and EU have the same definition as in Wallace and Gutzler (1981). The indices are defined with the anomalies at a number of critical places. Most of these places happen to have high probability as shown in Figs. 3 and 4.

Similar to the SFP probability table, Table 3 presents probabilities of the three teleconnection indices derived from the five winters of data. Only probabilities of correct forecast categories (equivalent to p_{11}, p_{22}, and p_{33} in Table 1) are presented at this time. The probabilities in Table 3 have similar features to those of grid points over the North America region (main diagonal in Table 2). For example, categories B and A have much higher probabilities than N. Meanwhile, probabilities of the PNA index are higher than for grid points over North America in general. Table 4 shows the mse verification of the index probability forecast for the three winters. The scores are considerably better than those of grid points (see Fig. 8).

5. Summary and discussion

In this study, we have explored three methods to convert a deterministic forecast into a probability forecast. The first method, SFP, is based on a single model forecast and is made from a probability table derived from the model's historical forecasts and analyses. The table-based probabilities have good reliability. One drawback of SFP is that the method cannot take advantage of information particular to a case, such as ensemble spread in an ensemble forecast. SFP is a relatively conservative

TABLE 3. SFP probabilities of teleconnection indices derived from the five-winter reforecast data (Dec–Feb, Jan 1985–Feb 1989). Only diagonal elements of SFP table are shown.

Class Index	D + 8			WK2		
	B	N	A	B	N	A
PNA	0.684	0.488	0.739	0.577	0.374	0.659
NAO	0.627	0.431	0.731	0.533	0.380	0.621
EU	0.573	0.346	0.643	0.423	0.267	0.528

method, particularly when applied to longer leads like D + 8 (6–10-day mean) and WK2 (8–14-day mean). In the second method, EP, the probabilities are simply derived from the ensemble members. The EP method fully makes use of the ensemble-provided probabilistic information, which is specific to the case under consideration. The EP probabilities are typically much too confident for large probability anomalies. In our experiment of D + 8 and WK2 forecasts with the three winters of NCEP's operational ensemble forecast data, there are a considerable number of occasions where EP predicts large height anomalies with probabilities over 0.6, but these probabilities are found not reliable and are too confident. The third method, MFP, is the average of a conservative estimate of SFP probabilities and the EP probabilities.

There is no question that, for a single model forecast, the table-based SFP provides a useful probability specification while keeping the probabilities formally reliable. There is also no doubt that an ensemble adds information. With the ensemble, however, questions remain as to how to precisely extract probabilistic information and how to use this information to improve the final probability forecast. The EP probabilities directly come out of the ensemble but are usually not reliable. We believe that a method should be developed to combine the SFP and EP probabilities, and MFP is a preliminary experiment. Reliability tests and verification with our data show that MFP is a much improved method, but by no means can it be argued as optimal. Further study is necessary to resolve this issue.

In recent years, many efforts, such as the ''breeding'' process or optimal singular vectors (Toth and Kalnay 1993; Mureau et al. 1993), have been devoted to constructing the ensemble perturbations. Initial perturbations should effectively represent the uncertainties in the initial conditions such that all possible future states are captured (or bracketed) as completely as possible. The perturbation sampling problem has also been studied with a perfect model (Anderson 1996a). Based on the results here, the ensemble forecast does not appear to have enough spread as the ensemble tends to cluster on a single state (B, N, or A). It will be interesting to consider this spread problem when constructing initial perturbations, model perturbations, or both. Another issue is the size of the ensemble (Deque 1997).

Although the initiative of the ensemble forecasting is to mimic samples of possible future states, operational practice using the ensemble appears based on a more traditional philosophy. That is, the ensemble spread is just ''dynamical noise'' and the ensemble mean is the best approximation of a single future state. This may be due to the convenience of using the mean and the difficulty of interpreting the spread. Improvement using the ensemble mean over any single forecast is widely agreed on in practice (van den Dool and Rukhovets 1994).

When the ensemble mean is used in our probability forecast, the ensemble mean is treated as a single forecast and therefore a probability table like Table 1 could be used for specification. We have applied the table derived from the reforecast data to the ensemble mean and found (not shown) that the ensemble mean verifies better than the single control forecast. This improvement presumably comes from the smoothing of the ''dynamical noise.'' However, the way it is used now, the ensemble mean does not beat MFP. It can be argued, though, that applying the SFP table to the ensemble mean is an underspecification. Instead, a probability table should be derived directly from ensemble mean forecasts. This again requires a large number of ensemble forecast samples, which are not available right now to our knowledge. The other end of the question is that the MFP is very likely not the optimal combination of the table and EP. A definitive comparison can be made only when both the probability table of ensemble mean is used and the MFP is better developed. We leave this topic for future study.

In this study, we used the 11-member ''consistent ensemble'' instead of a more mixed ensemble of forecasts that differ not only in perturbations but also in resolutions, initial condition time, and so on. In doing so, we could use the probability lookup table for all members as they should have the same skill level as individual forecasts, and the systematic error corrections was equally applicable to each member. The use of the consistent ensemble also provides an opportunity to examine the scheme creating the initial perturbations. The practical disadvantage is that the ensemble has probabilities that are too confident and much more so for the 11-member consistent ensemble than for the 17-member ensemble, which is a mix of resolutions and different initial condition times. Even when including model perturbations, the spread may be too small and the experience is that on some occasions all model solutions go confidently in one direction while reality goes in another.

Acknowledgments. The authors thank Dr. J. Schemm for providing the reforecasts data. We are also thankful to Ed O'Lenic, Dave Unger, and to three anonymous reviewers for their constructive comments and suggestions to improve this paper.

REFERENCES

Akesson, O., 1996: Comparative verification of precipitation probabilities from the ECMWF ensemble prediction system and from the operational T213 forecast. Preprints, *15th Conf. on Weather Analysis and Forecasting,* Norfolk, VA, Amer. Meteor. Soc., J31–J34.

Anderson, J. L., 1996a: Selection of initial conditions for ensemble forecasts in a simple perfect model framework. *J. Atmos. Sci.*, **53,** 22–36.

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

Branstator, G. A., W. Mai, and D. Baumhefner, 1993: Identification

of highly predictable flow elements for spatial filtering of medium and extended range numerical forecasts. *Mon. Wea. Rev.*, **121,** 1786–1802.

Chen, W. Y., and H. M. van den Dool, 1995a: Low-frequency anomalies in the NMC model and reality. *J. Climate,* **8,** 1369–1385.

——, and ——, 1995b: Forecast skill and low-frequency variability in the NMC DERF90 experiments. *Mon. Wea. Rev.*, **123,** 2491–2514.

Deque, M., 1997: Ensemble size for numerical seasonal forecasts. *Tellus,* **49A,** 74–86.

Gilman, D. L., 1986: Expressing uncertainty in long range forecasts. *Namias Symposium,* J. O. Roads, Ed., Scripps Institution of Oceanography Reference Series 86-17, 174–187.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics forecasting through model consensus. *Bull. Amer. Meteor. Soc.,* **76,** 1157–1164.

Hamill, T., and S. Colucci, 1996: Eta/RSM ensemble usefulness for short-range forecasting. Preprints, *15th Conf. on Weather Analysis and Forecasting,* Norfolk, VA, Amer. Meteor. Soc., J43–J45.

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis project. *Bull. Amer. Meteor. Soc.,* **77,** 437–471.

Klinker, E., and M. Capaldo, 1986: Systematic errors in the baroclinic waves of the ECMWF model. *Tellus,* **38A,** 215–235.

Molteni, F., and S. Tibaldi, 1990: Regimes in the wintertime circulation over northern extratropics. II: Consequences for dynamical predictability. *Quart. J. Roy. Meteor. Soc.,* **116,** 1263–1288.

Mureau, R., F. Molteni, and T. N. Palmer, 1993: Ensemble prediction using dynamically conditioned perturbations. *Quart. J. Roy. Meteor. Soc.,* **119,** 299–324.

O'Lenic, E., and Coauthors, 1996: Format, methods and estimated skill of proposed operational forecasts for week two (days 8–14). *Proc. 21st Annual Climate Diagnostics Workshop,* Huntsville, AL, NOAA/NWS/NCEP/CPC, 162–165.

Pan, J., and H. van den Dool, 1995: On the geographical distribution of forecast skill in 500 hPa height in the 6–10 day range for NH winter. *Proc. 20th Annual Climate Diagnostics Workshop,* Seattle, WA, NOAA/NWS/NCEP/CPC, 109–112.

Schemm, J. E., H. M. van den Dool, and S. Saha, 1996: Application of a multi-year DERF experiment towards week 2 and monthly climate prediction. *Proc. 21st Climate Diagnostics and Prediction Workshop,* Huntsville, AL, NOAA/NWS/NCEP/CPC, 166–169.

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

Tracton, M. S., 1990: Predictability and its relationship to scale interaction processes in blocking. *Mon. Wea. Rev.,* **118,** 1666–1695.

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

——, K. C. Mo, W. Chen, E. Kalnay, R. Kistler, and G. White, 1989: Dynamical extended range forecasting (DERF) at the National Meteorological Center. *Mon. Wea. Rev.,* **117,** 1604–1635.

van den Dool, H. M., and Z. Toth, 1991: Why do forecasts for near-normal fail to succeed? *Wea. Forecasting,* **6,** 76–85.

——, 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 concensus. *Bull. Amer. Meteor. Soc.,* **76,** 1157–1164.

Wallace, J. M., and D. S. Gutzler, 1981: Teleconnections in the geopotential height field during the Northern Hemisphere winter. *Mon. Wea. Rev.,* **109,** 784–812.

Wilks, D. S., 1995: Statistical methods in the atmospheric sciences. *International Geophysics Sciences,* R. Dmowska and J. R. Holton, Eds., Vol. 59, Academic Press, 199–283.

Zhu, Y., G. Lyengar, Z. Toth, S. Tracton, and T. Marchok, 1996: Objective evaluation of the NCEP gloabl ensemble forecasting system. Preprints, *15th Conf. on Weather Analysis and Forecasting,* Norfolk, VA, Amer. Meteor. Soc., J79–J82.