A Strategy for Verification of Weather Element Forecasts from an Ensemble Prediction System

LAURENCE J. WILSON

Recherche en Prévision Numérique, Atmospheric Environment Service, Dorval, Quebec, Canada

WILLIAM R. BURROWS

Recherche en Prévision Numérique, Atmospheric Environment Service, Toronto, Ontario, Canada

ANDREAS LANZINGER*

European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom

(Manuscript received 11 October 1997, in final form 8 April 1998)

ABSTRACT

Using a Bayesian context, new measures of accuracy and skill are proposed to verify weather element forecasts from ensemble prediction systems (EPSs) with respect to individual observations. The new scores are in the form of probabilities of occurrence of the observation given the EPS distribution and can be applied to individual point forecasts or summarized over a sample of forecasts. It is suggested that theoretical distributions be fit to the ensemble, assuming a shape similar to the shape of the climatological distribution of the forecast weather element. The suggested accuracy score is simply the probability of occurrence of the observation given the fitted distribution, and the skill score follows the standard format for comparison of the accuracy of the ensemble forecast with the accuracy of an unskilled forecast such as climatology. These two scores are sensitive to the location and spread of the ensemble distribution with respect to the verifying observation.

The new scores are illustrated using the output of the European Centre for Medium-Range Weather Forecasts EPS. Tests were carried out on 108 ensemble forecasts of 2-m temperature, precipitation amount, and windspeed, interpolated to 23 Canadian stations. Results indicate that the scores are especially sensitive to location of the ensemble distribution with respect to the observation; even relatively modest errors cause a score value significantly below the maximum possible score of 1.0. Nevertheless, forecasts were found that achieved the perfect score. The results of a single application of the scoring system to verification of ensembles of 500-mb heights suggests considerable potential of the score for assessment of the synoptic behavior of upper-air ensemble forecasts.

The paper concludes with a discussion of the new scoring method in the more general context of verification of probability distributions.

1. Introduction

Until the advent of ensemble prediction systems (EPSs), verification of numerical weather prediction results usually consisted either of comparisons of gridded model output with analyses, or of interpolated model output with point observations. Many different measures have been defined and used, such as mean absolute and root-mean-square error, anomaly correlations, and the S1 score, to name a few. These scores are computationally different and express different aspects of the model performance, but in all cases the elements of the verification sample are produced by space and time matching of single point deterministic forecasts from the model with the corresponding observation. The ''forecast error'' is determined by a simple difference between the forecast value and the observed value of the weather element. The forecast value is presented as a ''best estimate'' and any associated uncertainty is not estimated. In terms of the verification, uncertainty is simply an unspecified component of the error.

A complete approach to forecasting the uncertainty in the future state of the atmosphere would mean solving the Liouville equation. However, this is impractical in realistic situations (Ehrendorfer 1994). An EPS attempts

* Current affiliation: ZAMG, Wetterdienststelle Salzburg, Salzburg, Austria.

Corresponding author address: Recherche en Prévision Numérique, Atmospheric Environment Service, 2121 route Transcanadienne, Suite 500, Dorval, PQ, H9P 1J3, Canada.
E-mail: lawrence.wilson@ec.gc.ca

to quantify this uncertainty using a set of perturbed initial conditions and/or perturbed model formulations, depending on the system. The ensemble represents an attempt to estimate the full range of possible outcomes given what is hoped is a realistic range of possible initial conditions (and model formulations). Whereas each member of the ensemble is a deterministic forecast, a specific model trajectory, the ensemble of deterministic forecasts is considered to be an estimate of the distribution of all the model's predicted variables at each grid point and at each projection time. This is the basic output of an EPS.

Verification methods applied to ensemble forecasts have so far been directed toward two main purposes: diagnostic assessment of the characteristics of the ensemble distribution and verification of probability forecasts derived from the ensemble. Examples of the former include Buizza (1997) and Molteni et al. (1996) for the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble system, and Hamill and Colucci (1997) for National Centers for Environmental Prediction (NCEP) ensembles. Buizza (1997) searches for a relationship between ensemble accuracy and the accuracy of the unperturbed control forecast using the anomaly correlation coefficient and root-mean-square (rms) differences. These are applied to individual ensemble members with respect to the verifying analysis. Molteni et al. (1996) also looked for a spread–accuracy relationship, used the Brier score (Brier 1950) to evaluate the probability of occurrence of synoptic flow patterns, and calculated the accuracy of the ensemble mean using the root-mean-square error (rmse) with respect to the analysis. Hamill and Colucci (1997) measured the performance of short-range NCEP ensemble forecasts using histograms of verification rank. These are produced by ordering the N ensemble forecast values from lowest to highest, numbering (ranking) the $N + 1$ intervals in order, and tallying over a sample the ranks of the intervals in which the observations fall. Rank histograms (sometimes called Talagrand diagrams) give information on the distribution of the ensemble forecasts relative to the distribution of the observations in the verification sample. However, they do not give information on the accuracy of the ensemble forecast since one could randomly select ranks from the ensemble distributions as ''observations'' and get a better result.

EPSs have also been assessed in terms of probability forecasts derived from the ensemble distribution (Hamill and Colucci 1997; Akesson 1996; Buizza and Palmer 1998). To accomplish this, the predicted variable to be verified is first divided into two or more categories separated by threshold values, then the probabilities of occurrence of the events defined by the categories are estimated from the ensemble, and standard methods of verification for probability forecasts are applied. For example, one might be interested in the probability of occurrence of extreme temperatures as defined by an anomaly of at least 8°C. Verification methods for probability forecasts include reliability diagrams, signal detection theory [Mason 1982; Stanski et al. 1990.; see Buizza and Palmer (1998) for an example], along with summary scores such as the Brier score (Brier 1950), the rank probability score (Epstein 1969), and related skill scores. These methods all measure one or more of the attributes of a probability forecast, as identified by Murphy (1993).

In general, one must collect a sufficiently large sample of ensemble forecasts to permit the estimation of their accuracy level with sufficient confidence. However, information of a more diagnostic nature can be obtained through the synoptic assessment of individual cases (Sivillo et al. 1997; Toth et al. 1997). One of the tools for this kind of assessment has been the ''spaghetti'' plot, where a single contour, usually from 500-mb charts, is plotted for all the members of the ensemble, the control, and the verifying analysis. The amount of scatter among the contours is an indication of the ensemble spread, and a subjective estimate of the accuracy can be obtained by comparing the ensemble of contours with the analysis. One must be careful interpreting spaghetti plots because a large scatter of the plotted contours signifies less spread in the ensemble in areas of flat gradient than it does in areas of steep gradient, and it has been suggested (Toth et al. 1997) that a full field map of the control be used alongside a spaghetti plot. Other diagnostic tools that have been used include maps of statistics of the ensemble such as the spread and maps of probabilities estimated from the ensemble, all displayed for individual valid times.

While all of the methods mentioned above clearly indicate that the choice of measure depends on the purpose of the verification, none addresses directly the issue of verifying the basic output of an ensemble forecast, that is, an estimated probability distribution, against a single observation. To apply them, one must resort to verifying the individual observed values (from observations or analyses) against a statistic of the ensemble (e.g., the ensemble mean; Toth and Kalnay 1997), or compile a distribution of verification results on individual ensemble members (e.g., the best member; Buizza and Palmer 1998), or estimate probabilities from the ensemble. In this paper, a new verification method is proposed that is designed for the quantitative evaluation of ensembles of deterministic weather element forecasts. The method takes into account the probabilistic nature of the ensemble output and, as shown below, can be applied to a single forecast or summarized over a sample of ensemble forecasts. The two new scores are introduced in section 2, and their characteristics discussed. To illustrate the performance of the scores, results of tests on ensemble forecasts from ECMWF are shown in section 3. Finally, in section 4 we offer some discussion of the new scores in the more general context of assessment of ensemble forecasts.

2. A method for the verification of ensembles of deterministic weather element forecasts

a. Measurement of the ensemble distribution accuracy

Bayes's theorem concerns the relationship among three distributions of a random variable X. These are the prior distribution, the likelihood function, and the posterior distribution. The prior distribution incorporates what is known about the distribution of X prior to a stochastic experiment, the gathering of a sample, for example. The likelihood function is the sampling distribution of X, consisting of the full set of probabilities of obtaining each particular value of the variable under specific sampling conditions. The posterior distribution represents the knowledge of the distribution of X following the stochastic experiment. The posterior distribution is in general changed from the prior distribution, in light of information based on the sample. A full description of Bayes's theorem and Bayesian inference is given in Hays and Winkler (1971).

In the present context, the random variable X could be any of the surface or upper-air weather elements predicted by the EPS and the prior distribution might be the climatological distribution of weather element X at a particular location, for a particular day. The prior distribution might also be a persistence distribution, consisting of the probabilities of occurrence of various values of X under specific antecedent conditions. Either way, the prior distribution represents knowledge about X before the analysis is performed and before the ensemble forecast is run. The set of perturbed analyses and the ensemble forecast can be considered to be the (rather sophisticated) equivalent of the likelihood function, a ''sampling'' methodology by which a more refined estimate of the distribution can be determined, for each location and forecast time. Then, the ensemble distribution of the variable X at a point and time becomes an estimate of the posterior distribution.

Again in a stochastic context, the observation x of random variable X is a sample of size one, a single member extracted from the distribution of X. The ensemble forecast can be expected to have provided additional site- and time-specific information to refine the estimate of the probability distribution of X. Thus, if the model has skill, the probability of occurrence of a specific observation should be higher under the ensemble distribution than under the prior distribution. It is this probability, conditional on the posterior (ensemble) distribution that can be used to verify the EPS prediction. This can be written as $P(X = x_{obs} | X_{eps})$, for weather element X, where X_{eps} refers to the ensemble distribution.

Ideally, one might compute this probability directly from the ensemble distribution. However, a reasonably reliable delineation of the full probability density function (pdf), or alternatively the cumulative distribution function, would require an ensemble size of many hundreds or thousands of members. For example, one would need an ensemble of more than 100 members to expect to locate the first or 99th percentile of the distribution, assuming the members are randomly selected from the ensemble distribution. With ensemble sizes typical of current operational systems (now 50 members for the ECMWF EPS), it is necessary to use a parametric approach to describe the distribution, to fit the parameters of a chosen distribution, then estimate the probabilities from the fitted distribution. Fitting of a specific distribution also makes it possible to add any available knowledge about the expected shape of the ensemble distribution.

As a first approximation, it is reasonable to expect that the ensemble forecast would exhibit a distribution of the same family as the climatological distribution of the same weather element. For example, temperatures usually follow a normal distribution. Other relatively smooth elements such as upper-air temperatures and geopotential heights can also be expected to be normally distributed. Wind speeds have been shown to fit a Weibull distribution (Somerville and Bean 1979); precipitation amounts fit a gamma distribution (Hamill and Colucci 1998) or a kappa distribution (Mielke 1973). The gamma, kappa, and Weibull distributions are all suited to positive definite variables such as wind speed and precipitation amount. Variables such as cloud amount, which are defined in the range [0,1], can be fit with a beta distribution (Somerville and Bean 1979). The beta distribution includes not only the U shape typical of cloud amount, but also the uniform distribution and near-normal forms. Table 1 summarizes the distributions that have been found to fit observations of weather elements.

Once a distribution has been chosen and fitted, the probability $P(X = x_{obs} | X_{eps})$ can be computed. An appropriate window may be chosen to define the probability of a ''sufficiently correct'' forecast for a specific application. For example, we define a range ΔX from the observed value of X that constitutes a correct forecast. Thus, the window is defined by the limits $[x - \Delta X, x + \Delta X]$. The window can also be allowed to vary over the range of the weather element and be tuned for operational considerations. As an example, if winds less than 4 m s^{-1} are considered operationally unimportant, this can be used as a fixed window for assessment of wind forecasts. For precipitation, the width of the window can increase with increasing precipitation amounts, for example, as a geometric progression with limits $[X/c, cX]$ where c is a constant. A 5 mm window may not be too large for amounts over 25 mm, but it certainly is too large for amounts in the range of 1 mm. The choice of window is to some extent subjective and based on operational considerations. In any case, the window width should not span more than one climatological standard deviation on either side of the observed value. The window can always be chosen to reflect the specific accuracy requirements of the forecast, and their variation over the range of the variable.

TABLE 1. Summary of theoretical distributions that have been found to fit observations of weather.

Weather element	Distribution	Characteristics
Temperature, geopotential height, upper-air temperatures	Normal	Two-parameter, mean and standard deviation Symmetric, bell-shaped
Precipitation (QPF)	Gamma, kappa, cube-root normal	Gamma: Two-parameter—"shape and spread;" positively skewed. Applies to variables bounded below; approaches normal when well away from lower bound. Kappa: similar to gamma in form, but not as well known. Cube-root normal: The cube root of precipitation amount has been found to be approximately normally distributed.
Wind speed	Weibull	Two-parameter; negatively skewed; applies to variables bounded below.
Cloud amount	Beta	Two-parameter; a family of distributions including the uniform and U shaped as special cases. Intended for variables that are bounded above and below, such as probability estimates and cloud amount. Negatively or positively skewed, depending on parameters.
Visibility	Lognormal	Normal distribution with logarithmic x axis; applies to positive-definite variables.

Since the results and interpretation of the score also depend on the window, the chosen window size(s) should be reported along with the results. It should be noted that the score can be defined without the window, by using the pdf evaluated at the observation; however, this was considered to give less meaningful results.

Figure 1 is a schematic of the proposed scoring system for temperature as an example. Two hypothetical ensemble distributions are shown: a sharp distribution with small standard deviation such as might occur in a short-range forecast and a distribution with greater spread such as might occur in a medium-range forecast. A hypothetical climate distribution is also shown. The score is given by the probability within the window centered on the verifying temperature, that is,

$$\mathrm{P}(T_{\mathrm{obs}} \,|\, T_{\mathrm{eps}}) = \int_{T-\Delta T}^{T+\Delta T} f(T_{\mathrm{eps}})\, dT, \tag{1}$$

where ΔT defines the limits of the correct range (1°C in this example), $f(T_{\mathrm{eps}})$ is the fitted pdf, and T_{obs} is the observed temperature.

Up until this point, it has been assumed that the ensemble distribution is unimodal. Unimodal distributions are not likely to fit well if the ensemble is trying to predict a multimodal distribution. If there is sufficient evidence that the ensemble distribution is multimodal, as revealed, for example, by a cluster analysis, then the distribution can be treated as separate distributions of the same form, and the parameters estimated separately for each distinct cluster. The score value is then given by the sum of the products of the prior probability of occurrence of each cluster and the likelihood of obtaining the observation given the occurrence of the cluster,

$$\begin{aligned}\mathrm{P}(x_{\mathrm{obs}} \,|\, X_{\mathrm{eps}}) = \mathrm{P}(x_{\mathrm{obs}} \,|\, X_{\mathrm{eps1}})\mathrm{P}(X_{\mathrm{eps1}}) + \mathrm{P}(x_{\mathrm{obs}} \,|\, X_{\mathrm{eps2}})\mathrm{P}(X_{\mathrm{eps2}}) \\ + \ldots \mathrm{P}(x_{\mathrm{obs}} \,|\, X_{\mathrm{eps}n})\mathrm{P}(X_{\mathrm{eps}n})\end{aligned} \tag{2}$$

for n clusters, where the conditional probabilities are computed following Eq. (1). With the small ensemble sizes in operational use, the distribution parameters cannot be reliably estimated for more than two or three clusters. For example, when fitting a unimodal normal distribution, a sample size of at least 4 is needed to define the density at the location of the mode (Silverman 1986) within a mean-square-error of 0.1. The probability of occurrence of the cluster is estimated by the percentage of ensemble members identified with each cluster.

b. Skill of the distribution of ensemble forecasts

The score defined above can easily be extended to a skill score in the usual format (Stanski et al. 1990), that is, skill equals the actual improvement over the standard forecast divided by the total possible improvement over the standard forecast. In the present case, this takes the form

$$\mathrm{Skill} = \frac{\mathrm{P}(x_{\mathrm{obs}} \,|\, X_{\mathrm{eps}}) - \mathrm{P}(x_{\mathrm{obs}} \,|\, X_{\mathrm{std}})}{1.0 - \mathrm{P}(x_{\mathrm{obs}} \,|\, X_{\mathrm{std}})}. \tag{3}$$

The variable X_{std} is usually the climatological distribution for the valid date of the forecast, but it could also be a persistence distribution. In the latter case, the appropriate distribution is the climatological distribution for the date conditioned on the initial value of the weather element. Since the score consists of probability estimates, a perfect score is always 1.0. In the case of multimodal distributions, the first term in the numerator of the skill score is estimated using Eq. (2).

Computation of the skill score requires not only the set of verification scores for the weather element in question, but also a set of climatological distributions for the weather element at the same locations as for the score. The climatological distributions need be computed only infrequently. In the results shown below, the climatological distributions are based on about 30 yr of observations, and the climatology is recomputed every 10 yr.

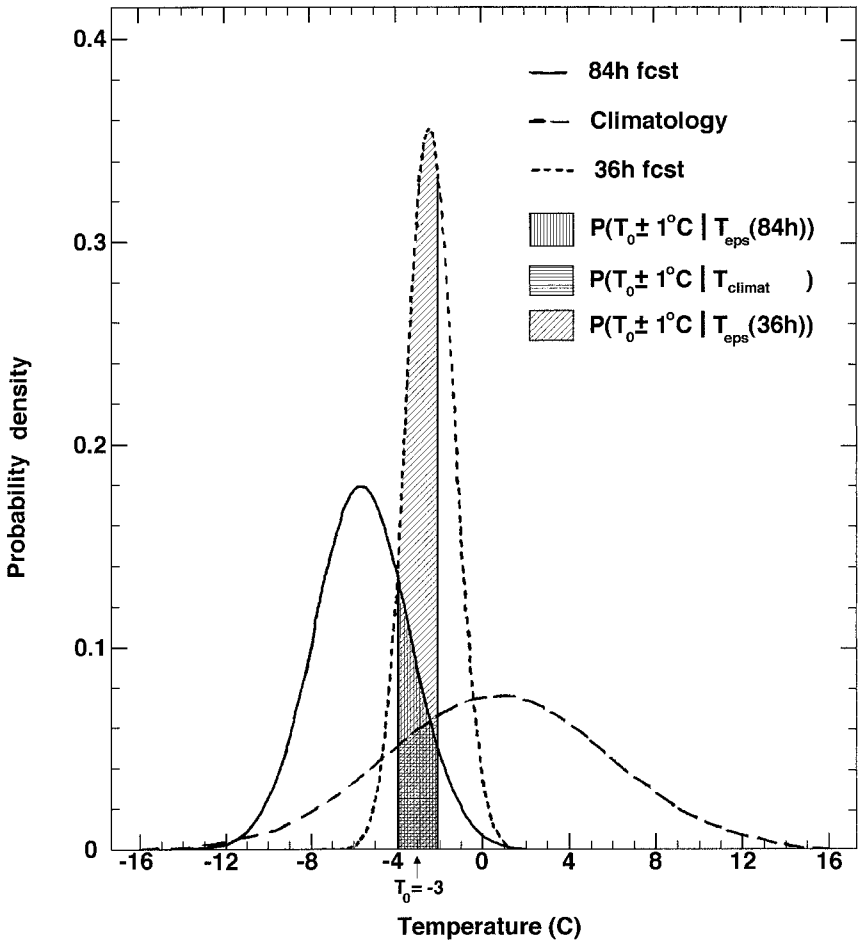

FIG. 1. Schematic illustration of the probability score for temperature within 1°C of the observed temperature, for a 36-h forecast (dashed), an 84-h forecast (solid), and climatology (long dashes). Hatched areas indicate the probability of a correct forecast given the two ensembles and the climatological distribution.

c. Characteristics of the scoring system

The probability score suggested in section 2a has some desirable characteristics as a verification measure. First, the score is simple by definition and, since it is a probability with a range of 0–1, it is as easy to understand as any probability. Whether the score is applied to a single forecast or averaged over a sample of ensembles, the numerical values that are produced can be understood as probabilities. Second, it is positively oriented. A perfect score is obtained if all the ensemble members predict the verifying observation. Third, the score is sensitive to both the spread (variance) of the distribution and to its location with respect to the verifying observation. Thus, it encourages precise forecasting (small ensemble spread) and accurate placing of the distribution. Conversely, dispersed forecasts (large spread) and sharp forecasts (small spread) that miss the event are both penalized with this score. The ideal forecast is both accurately positioned (reliable) and sharp.

The score bears some relationship to existing verification scores, most notably the Brier score (PS; Brier 1950) and the rank probability score (RPS; Epstein 1969). The Brier score is the mean square probability error. For a single forecast, the Brier score is related to the new score by

$$\mathrm{Sc} = 1 - \sqrt{\mathrm{PS}}, \tag{4}$$

where Sc is the new score, and PS is defined for a category that is always the same as the window for Sc in which the observation occurs. The Brier score is usually summed over a sample of forecasts; there is not a simple relationship between averages of the new score and averages of the Brier score, since the new score is linear and the Brier score is quadratic. It should also be noted that the Brier score normally applies to verification problems where the forecast is completely stated in advance. That is, it is applied to probability forecasts of the occurrence of specified fixed categories. To use

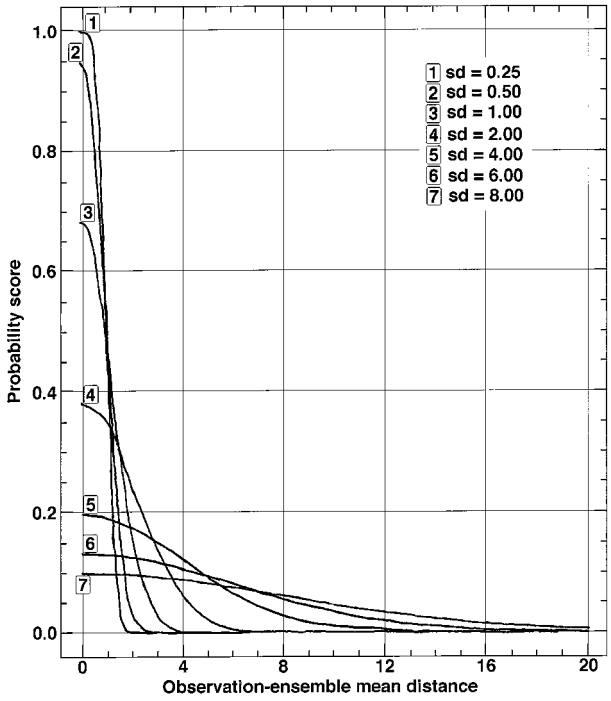

FIG. 2. Score values for temperature as a function of ensemble mean–observation difference assuming a normal distribution, for different ensemble standard deviations. A window of ±1°C is assumed.

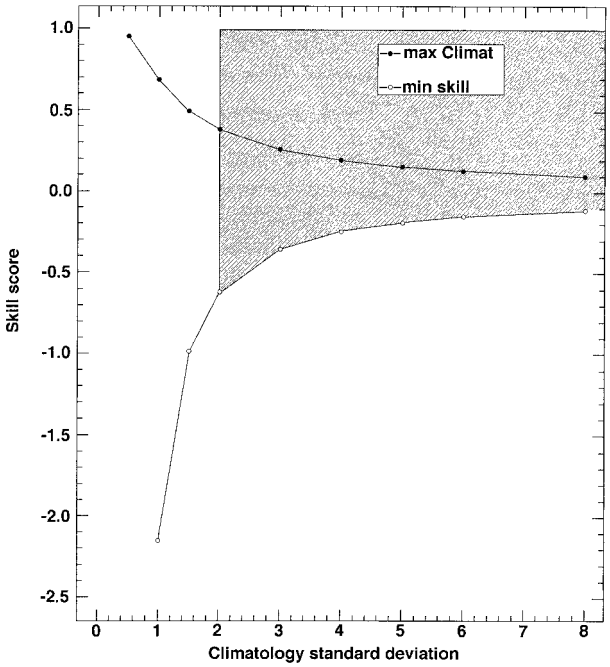

FIG. 3. Range of the skill score (hatched area), as a function of the standard deviation of climatology, assuming that the climatological standard deviation will be at least 2. A window of ±1°C is assumed.

the Brier score in the present application would mean verification of forecasts that can be specified only after the occurrence of the observation, because the category definition varies with the observation. This would at least confuse the interpretation of the score values and could be considered improper since the forecast that is being verified cannot be completely stated in advance.

The RPS is a quadratic scoring rule that takes into account the distribution of probabilities over prespecified categories. Like the new score, the RPS is positively oriented and is sensitive to the location and sharpness of the forecast distribution with respect to the verifying observation. The RPS is not computationally related to the new score, however, even for a single forecast. The RPS is also intended for verification of probabilities of fixed categories; its use for variable categories would be subject to the same difficulties as for the Brier score.

An example of the response of the score is shown in Fig. 2, for temperature, with a ±1°C window for a correct forecast. The greater the ensemble standard deviation, the lower the maximum possible score. For example, if the ensemble standard deviation is 2°C, the highest attainable score is less than 0.40. The figure also shows that the score is sensitive to the location of the ensemble (mean) with respect to the verifying observation. Ensemble mean–observation differences of more than 2.5°C limit the score value to less than 0.2, no matter what the ensemble spread. The set of curves in the figure forms an envelope for the score values within which the scores will lie. It is evident that the score is rather strict; high values occur only for sharp ensemble distributions and accurately placed ensembles. Low standard deviation in the ensemble allows high verification values, but they are achieved only if the whole ensemble is accurately placed.

The skill score has the theoretical range of $-\infty$ to 1.0. The lower bound is obtained only if the standard score is perfect, which would mean that all the relevant climatological values are the same as the observed value, that is, perfectly sharp and correct. In practice this does not occur for continuous variables such as temperature and wind, but is more likely in the case of episodic elements such as precipitation, where a dry climate can lead to near-perfect climate scores on dry days. The upper limit is obtained always with a perfect score for the ensemble forecast. Figure 3 shows schematically the possible range of values for the skill score, again using temperature as an example. The shaded area shows this range, assuming a ±1°C window and assuming that the climatological standard deviation will be at least 2°C. The greater the climatological standard deviation, the lower the maximum climatological score, and the greater the likelihood of a positive skill score. This means that the opportunity to improve on climatology is greater at locations with higher climatological variance (''difficult'' sites). Thus, the score takes into account differences in the predictability of the weather at the location. For the special case where the climatological mean value is observed, the ensemble still can

show skill by forecasting a sharper distribution than the climatological distribution.

3. Tests of the scoring method on ECMWF ensemble forecasts

The version of the ECMWF EPS used in tests of the new scoring method is described in Molteni et al. (1996) and Buizza and Palmer (1995). The EPS consists of a control run of a T63L19 version of the ECMWF model started from the operational analysis and 32 runs of the same model started from a perturbed analysis. The perturbed analyses are constructed by adding and subtracting 16 orthogonal linear combinations of singular vectors to the operational analysis. Operational products from the ensemble include ''postage stamp'' maps of all ensemble members, clusters of 500-mb height trajectories, probability ''plumes'' for specific locations, and maps of probabilities of precipitation and various extreme temperature and surface wind events. As noted above, these are produced by defining thresholds and estimating the probabilities directly from the ensemble under the assumption that each member is equally likely to occur (Molteni et al. 1996).

To assess the performance characteristics of the score, we used a sample of 108 days of weather element forecasts from the ECMWF ensemble system, for the period February to May 1996. Forecasts are valid every 12 h to 10 days, giving 20 forecast projections. We used the surface weather element forecasts, 2-m temperature, 10-m wind, and 12-h precipitation amount as interpolated from the model grid to 23 Canadian station locations. There are 32 members in the ensemble plus the unperturbed control forecast, for a total of 33 members that were used in the analysis. A single forecast of 500-mb heights was also available for testing the score as applied to upper-air variables, and to investigate the horizontal variability of the score.

The goal of the tests was to demonstrate whether and how easily the score can be used to diagnose the performance characteristics of the EPS. It was not the intention to verify the ECMWF EPS in a comprehensive way, but we have computed some summaries of the score values. A secondary goal of the tests was to evaluate the fit of the chosen distributions, at least in a qualitative way. The sample is probably not large enough to do a rigorous assessment of the assumption that the ensemble distribution is of the same family as the climatological distribution, but some qualitative evidence on that subject was revealed in the tests.

To compute the skill score, it is necessary to obtain climatological data for each location. For temperature and wind speed, we obtained the climatology from the archives for each station. The length of the archive depends on the station but is generally more than 30 yr. The climatological distributions were computed for each day of the year, separately for 0000 and 1200 UTC. Wind and temperature data were then time smoothed with a centered five-point binomial smoother, with weights of 1-4-6-4-1. Distributions (normal for temperature, Weibull and gamma for windspeed) were fit to the smoothed data. For precipitation, the climatology of 12-h amounts was computed from the more than 30-yr archives for each station, separately for 0000 and 1200 UTC. To stabilize the distribution estimates, we used a 31-day running average, centered on each day of the year. A gamma distribution was fit to the averaged data.

In the following sections, samples of the test results are described to illustrate the performance of the scoring system. For the surface weather elements, a single example of the application of the score and skill score is shown, then a summary of the score is shown over the 108-day test period. For 500-mb heights a spatial distribution of the score values is shown to illustrate possible diagnostic uses of the score when applied to upper-air weather elements.

a. 2-m temperature

Figure 4 shows the verification of two ensemble temperature forecasts, both valid at the same time. Windows for a correct forecast are ± 1°C. The plots show the actual ensemble distribution as a histogram, the fitted normal distribution, and the corresponding climatological distribution. The observed temperature of 11.3°C was near normal for the date. At both projections, the ensemble has indeed predicted a sharper distribution than the climatological distribution for the date. In the 3-day forecast, the ensemble distribution is accurately placed but the spread is relatively large (range from 8° to 19°C), resulting in a relatively low score value. At the longer range, 7 days, the ensemble spread is about the same, but the forecast has missed, the score drops to 0, and the skill score becomes negative. In this case, the climatological score is not large because the spread is large, even though a near-normal temperature was observed.

Figure 5 summarizes the score and the skill score for the same location, Pearson International Airport (PIA; Toronto, Ontario, Canada), for the entire 108-day sample of forecasts. The curves show that the forecasts valid at 0000 UTC remain skillful until about 8 days, but that there is no skill in the forecasts valid at 1200 UTC. Examination of some individual forecasts suggests that the ensemble forecast model is unable to resolve the thin surface-based nocturnal inversion layers that form especially in winter in continental climates. Thus, the 1200 UTC forecasts, near minimum temperature time in North America, contain many cases where the ensemble has underpredicted the intensity of the nocturnal surface inversion.

b. Precipitation

Precipitation ensemble forecasts were fit with a two-parameter gamma distribution. This distribution com-

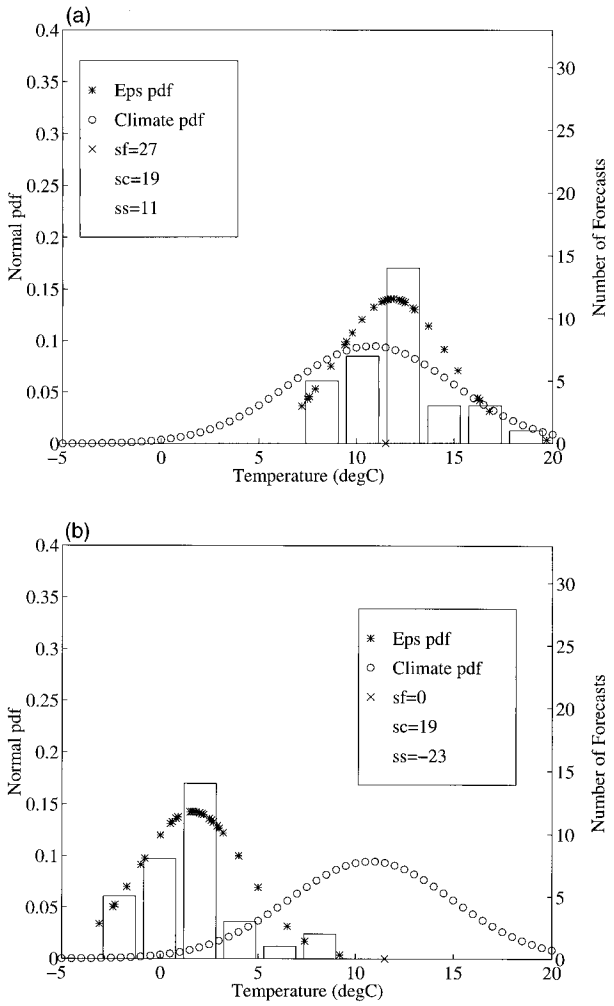

FIG. 4. Histogram of ECMWF ensemble temperature forecasts, fitted normal distribution (stars), and climatological distribution (circles) for (a) 72-h and (b) 168-h projections valid 17 May 1996 for Toronto, ON, Canada (PIA). The observed temperature is indicated by the cross on the abscissa. The probability score for the forecast is "sf," "sc" is the probability score for climatology, and "ss" is the skill score for this case. Score values are multiplied by 100.

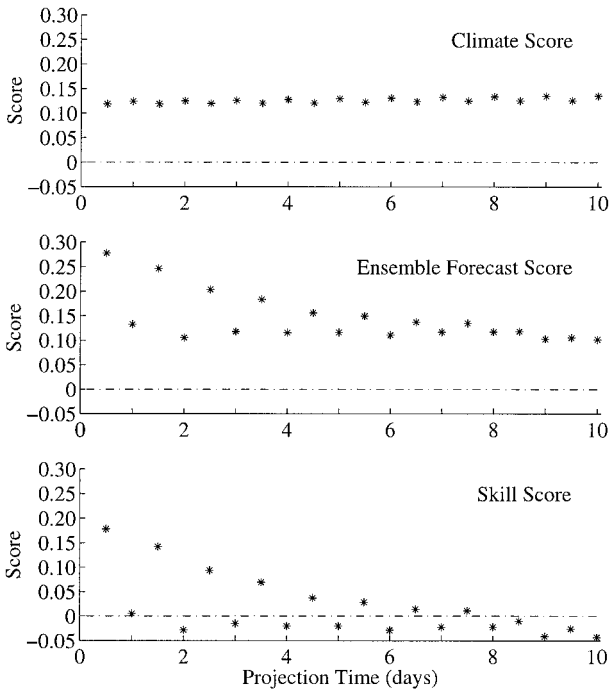

FIG. 5. Average climate score (top), ensemble forecast score (middle), and skill score (bottom) as a function of projection time for ensemble temperature forecasts for PIA, Toronto, ON, Canada, for 108 days between February and May, 1996.

prises both exponential shapes, with high probability density near zero, and skewed Gaussian shapes with modes greater than zero. We found the fit to be reliable for ensembles of 33 members; it did not fail on any of the cases. For cases where all the ensemble members predicted zero precipitation, the score value was set to 1.0 if precipitation did not occur and to 0.0 if it did. The distribution fit was attempted only if at least one of the ensemble members predicted some precipitation to occur. Windows for a correct forecast were chosen as follows. If less than 0.2 mm of precipitation occurred, a fixed window of 0.0–0.2 mm was used to define probabilities. If 0.2 mm or more was observed, the window was geometrically centered on the observed value, such that the upper limit was twice the lower limit. This means that the lower limit is defined as $P_{obs}/\sqrt{2}$ and the upper limit is $P_{obs}\sqrt{2}$. In practice, for example, the correct range for an observation of 2.0 mm is 1.41–2.83 mm, while the correct range for an observation of 10 mm is 7.1–14.1 mm. The expansion of the correct range with increasing precipitation accounts for the fact that small differences are more important for small amounts than with larger amounts. Such an exponential variation of the window also corresponds to typical probability density distributions of precipitation amount, which often are exponential. The choice of an exponent of 2 is somewhat arbitrary; other values could be chosen, for example, in response to particular operational requirements for accuracy and precision in the forecast.

Figures 6 and 7 show two examples of ensemble quantitative precipitation forecasts, for a case where none was observed and for a case where some precipitation was reported. Figure 6 demonstrates the high sensitivity of the skill score to relatively small errors in the ensemble placement in dry cases at a station where precipitation is relatively infrequent (Winnipeg, Manitoba, Canada). With only 3 of the 33 ensemble members forecasting precipitation at 3 days, the score remains positive. However, with six ensemble members forecasting precipitation at 7 days, the skill score has become strongly negative. In both cases, a relatively high absolute score value was obtained. We noticed a tendency toward too-frequent prediction of small amounts of precipitation among ensemble members in dry cases,

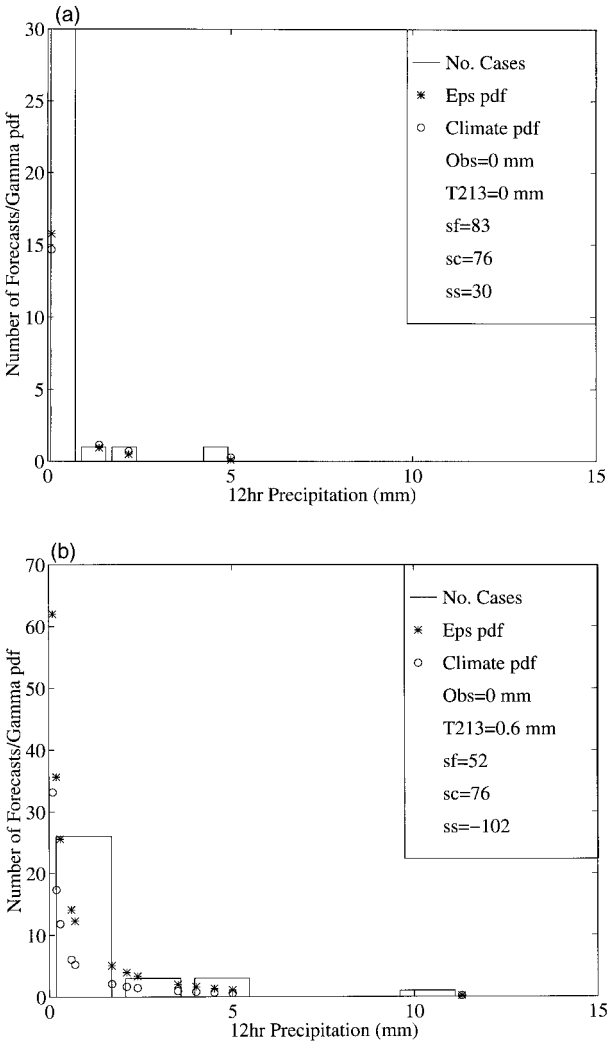

FIG. 6. Histogram of ECMWF ensemble 12-h quantitative precipitation forecasts, fitted gamma distribution (stars), and climatological distribution (circles) for (a) 72-h and (b) 168-h projections valid 5 May 1996 for Winnipeg, MB, Canada. The full resolution model prediction (''T213'') and three score values are shown in the box. No precipitation was observed.

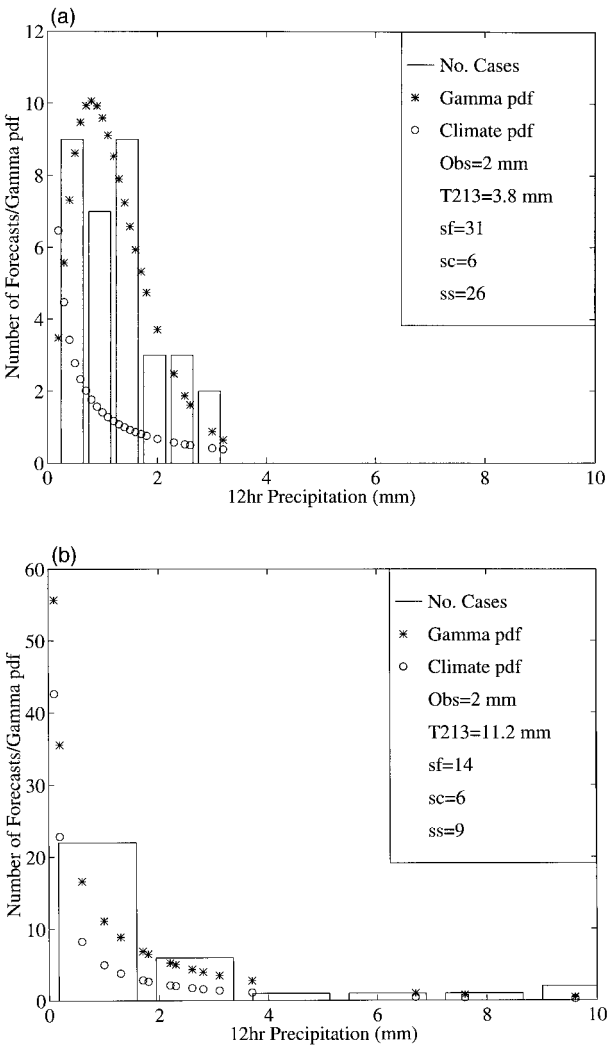

FIG. 7. Same as Fig. 6 but for Vancouver, BC, Canada, valid 23 May 1996. 2 mm of precipitation was observed.

especially for medium-range forecasts. Figure 7 documents by example the flexibility of the gamma distribution to fit both cases where the mode of the ensemble distribution is greater than zero (Fig. 7b) and cases where the lowest zero category has the greatest frequency (Fig. 7a). The climatological frequency of precipitation at this station is higher than at Winnipeg. Nevertheless, climatology does not score highly in this example because precipitation was observed, and the model shows skill at both 3 and 7 days.

Average score and skill score values for Winnipeg are shown in Fig. 8. Skill remains positive for only 60 h because of the cumulative impact of too-frequent prediction of small amounts of precipitation on dry days.

c. Wind speed

Climatological wind speed observations have been found to fit a Weibull distribution (Somerville and Bean 1979). The Weibull distribution is similar to the gamma distribution, except that it has a somewhat steeper rise on the left-hand side. Our experience fitting the ECMWF ensembles indicated that there was some difficulty fitting the Weibull distribution, and the fit failed in many instances on the 108-day sample. This is most likely caused by the sample size not being large enough to define the parameters of the Weibull distribution with confidence. In those instances where the Weibull fit succeeded, the empirical ensemble distribution was distinctly Weibull-like, with the characteristic steep rise on the left-hand side. Further experimentation with combined ensembles confirmed that the sample (ensemble) size is a problem for fitting the Weibull distribution;

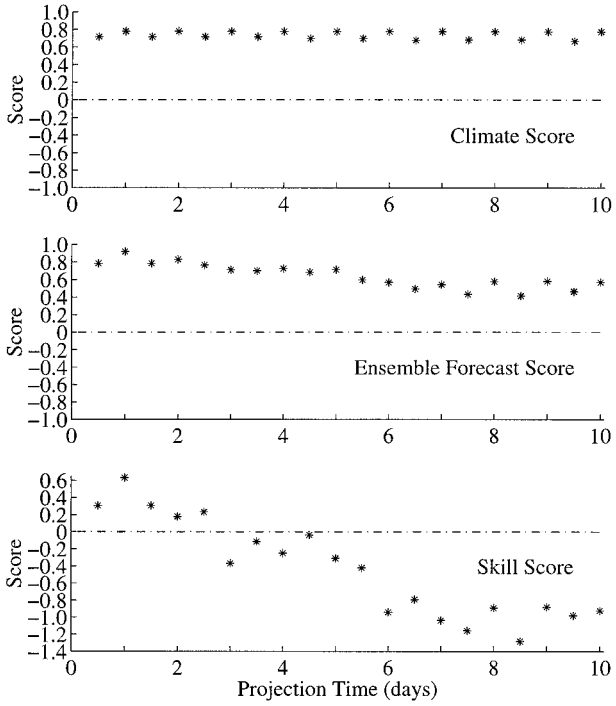

FIG. 8. Average climate, ensemble, and skill scores for Winnipeg ensemble precipitation forecasts for 108 days between February and May 1996.

once the sample size reached 120 or so, a fit was obtained in nearly all cases.

We found that the gamma distribution fit the wind speed data more reliably and gave probabilities that did not differ greatly from the Weibull probabilities. Figures 9b and 9c demonstrate this point. The histogram indicates a skewed distribution with a steep rise on the left side and tapered tail on the right. Scores for the 168-h forecast differed by only two points, 11 and 13 for the gamma and Weibull fits, respectively. The climatological score was nearly identical for the two distributions; however, the observation lies in the tail of both.

Figures 9a and 9b also show the comparison of wind speed verification using the gamma distribution for 3- and 7-day forecasts verifying at the same time. The ensemble is better placed with respect to the observation in the 3-day forecast, though the spread is nearly the same for the two forecasts. Once again, the score is sensitive to the placing of the ensemble distribution (bias), and the 72-h forecast achieves a much higher score than the 7-day forecast. The 3-day forecast has an ensemble distribution skewed the other way, with an extended tail on the left-hand side. The Weibull fit failed in this case.

d. Upper-air forecasts

In principle, the scoring system can be used for any weather element forecast by the model. So far, we have

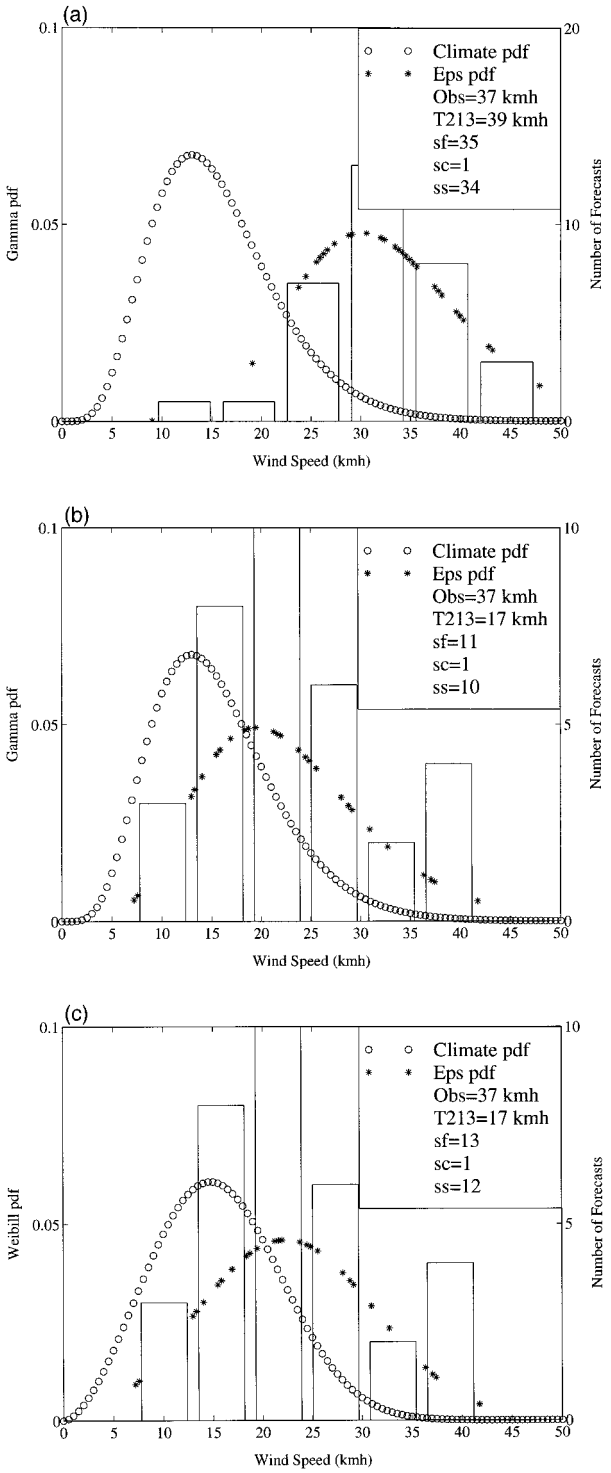

FIG. 9. Histogram of ECMWF ensemble wind speed forecasts, with fitted EPS and climate distributions and score values valid 22 March 1996. Score values are given in the box. (a) Gamma distribution, 72-h projection; (b) gamma distribution, 168-h projection; (c) Weibull distribution, 168-h projection.

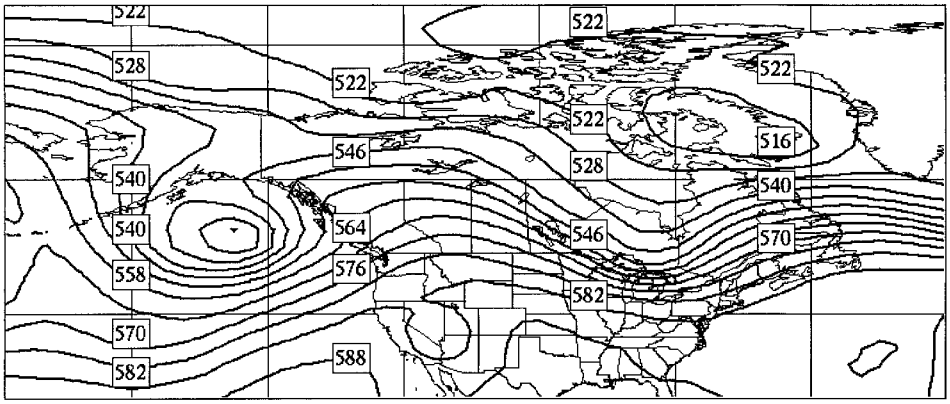

FIG. 10. Verifying analysis of 500-mb geopotential height, 0000 UTC 3 October 1996.

considered application to station-specific forecasts of surface-based weather elements. All ensemble forecasts were interpolated to the station location and matched with the station observation to carry out the verification. The scoring method can be applied equally well to gridded forecasts of upper-air variables such as 500-mb geopotential and 850-mb temperatures. To investigate this possibility, we considered a single ensemble forecast of 500-mb heights over a domain covering North America and adjacent oceans. The domain is large enough to permit examination of the large-scale flow pattern that is predicted by the ensemble and to assess the verification results in comparison with that pattern.

The verification method was applied at every grid point separately. The ensemble forecasts were available on a 2.5° × 2.5° lat–long grid. For verification, we used the Canadian analysis, normally available on a 1.5° × 1.5° global Gaussian grid, but interpolated by cubic splines to the same grid as the ECMWF forecasts. The analysis can therefore be considered to be independent of the model being verified since its background field comes from a different model.

As for the surface-based observations, we had to decide on an appropriate window to define a sufficiently correct forecast and to choose a distribution to fit. Since upper-air fields are relatively spatially smooth and unbounded, we assumed that the normal distribution would be most appropriate to use. For window widths, we tried ±2, 4, and 6 dm. For this one case, we found that ±2 dm gave too much detail when the verification score was plotted; it seemed every little wiggle in the contour pattern was associated with its own pattern in the verification results, making the verification hard to interpret. On the other hand, a ±6 dm window gave too little detail, and the scores were near perfect well into the forecast. A window of ±4 dm seemed to give about the right amount of detail throughout the forecast period. This conclusion is based on one case; further experimentation would reveal whether ±4 dm could be used generally.

Figures 10 and 11 are the verifying analysis and the score values, respectively, for 500-mb geopotential, for 0000 UTC 3 October 1996. First, score values tend to be high over the Tropics and more variable outside the Tropics. This is because there is less variability in the 500-mb heights in the Tropics. Second, the score values

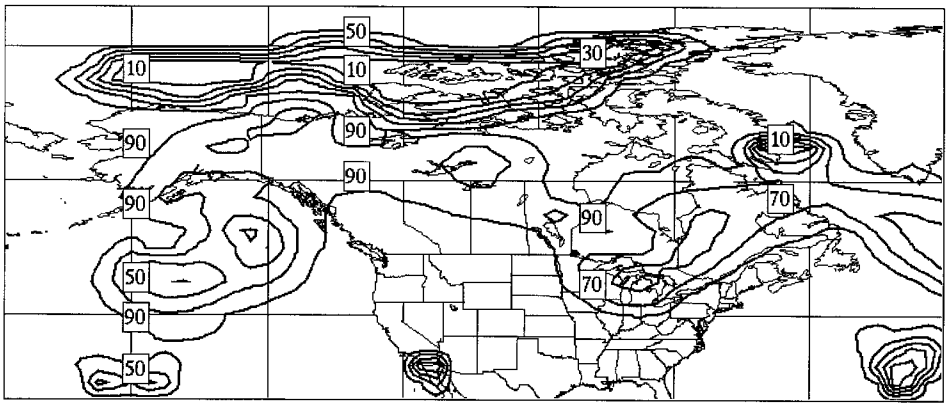

FIG. 11. Map of score values (%) for 36-h ECMWF 500-mb geopotential forecast valid 0000 UTC 3 October 1996. Score is computed assuming heights ±4 dam from the observed value are correct.

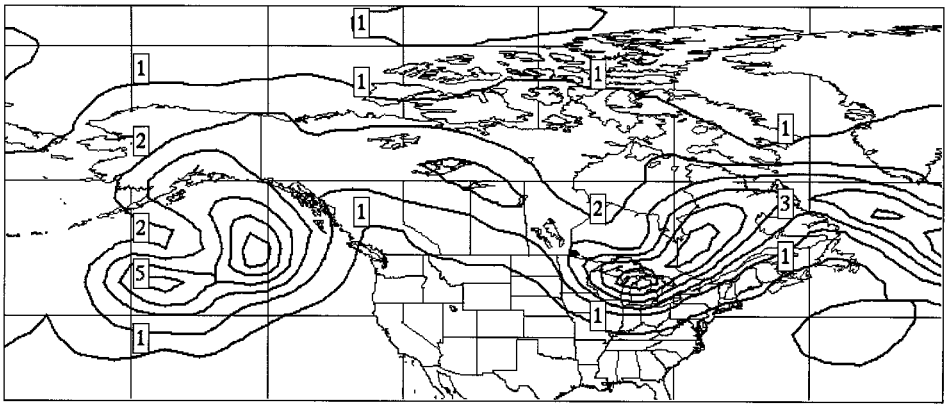

FIG. 12. Map of ensemble standard deviation (dam) for the case of Figs. 10 and 11.

dip modestly in the vicinity of the two main midlatitude lows, one in the Pacific and one over the Great Lakes. Third, there are several areas where the score values drop sharply to zero or near zero, most notably over the Arctic, but also near Labrador and at three locations in the Tropics. To explain these variations, it is useful to compare with Fig. 12, which shows the ensemble standard deviation (sd) in decameters for this case. The region of greatest standard deviation is along the northern side of the main jet at 500 mb. It stands to reason that the ensemble would be least confident in this area because it is the region of greatest baroclinity and potential for baroclinic instability. The pattern of the sd is similar to the pattern of the reduced score values in the midlatitude region, suggesting that the scores are reduced because of greater ensemble spread. The other regions of reduced score values do not correspond to areas of increased ensemble spread, suggesting that these regions are cases of positioning errors of the ensemble. Specifically, the reduced scores in the Arctic appear to be related to the low over Banks Island, which was apparently missed in the ECMWF model. The three small-scale drops in the score values in the Tropics are also related to weak troughs that appear in Fig. 10 and in the ECMWF analysis (not shown).

The score would appear to have considerable potential as a diagnostic tool in verification of upper-air ensemble forecasts. Using plots of both ensemble spread and the score, areas of bias error can be distinguished from areas where the error is more related to uncertainty in the ensemble forecast, especially in shorter-range forecasts. The score also gives a quantitative verification over the whole map. By contrast, spaghetti plots give verification information only along the particular contour that is plotted. Time series of score plots can also be used to diagnose source locations of error in the ensemble forecasts. For instance, an examination of the time series for the 1 October case indicated that the error over the Canadian Arctic originated north of Alaska in the 24-h forecast, spread east-southeastward, and expanded to cover the eastern Arctic by day 6. Errors in the main westerly belt tended to remain with the troughs with which they are associated, but in the longer ranges of the forecast there is an increasing tendency for a minimum in the score values to be located parallel to the western Cordillera. Whether this can be related to weaknesses in the topographical representation would require further study on many cases, but the scoring system could provide a valuable tool for such a study.

4. Discussion and conclusions

We have presented a strategy for the verification of weather element forecasts from an ensemble prediction system. In section 2 we presented and described the characteristics of two new verification measures designed for the assessment of the performance of the ensemble with respect to individual observations. Values obtained from the verification take the form of probabilities and can be easily understood whether they apply for a single point in space and time or whether they have been summarized over a large sample of forecasts. We have shown that values obtained from the new verification method are sensitive to both the location and spread of the ensemble distribution with respect to the verifying observation. The new measure comprises both a score and a skill score. The examples above used climatology as the standard forecast, but the skill score can be computed with reference to any available estimate of the probability distribution of the weather element.

The characteristics of the new scores were illustrated using 108 days of ensemble weather element forecasts from the ECMWF EPS. Results from the tests confirmed the sensitivity to the sharpness and location of the ensemble distribution with respect to the verifying observation. In fact, the score is quite severe; values above 0.25 are uncommon except in the shortest forecast ranges. The diagnostic potential of the score was illustrated both for point forecasts of surface weather elements and for ensembles of upper-air variables expressed on a grid. For surface variables, the score proved particularly sen-

sitive to errors in location of the ensemble with respect to the verifying observation, which permitted clear interpretation of the results in terms of model errors. For upper-air variables, experiments with one case indicated that, at least for short-range forecasts, it is possible to separate areas of large ensemble spread from areas where the ensemble has missed the forecast by using the new score in combination with a measure of ensemble spread.

Concerning the assumption about the distribution type, it is clearly valid if the ensemble members are randomly selected from the climatological distribution of each weather element for the verification date. The assumption might be refined by using the climatological distribution conditioned on the initial observation. Further study would be needed to determine whether these conditional distributions would differ significantly from the unconditional climatological distribution. Conditional climatology would also provide a more competitive standard for the skill score. The initial perturbations are not randomly selected; rather, the attempt is to identify perturbations that are most likely to grow and, therefore, have the greatest effect on the forecast. In terms of the forecast distribution, the intention is to find the extremes of the distribution with a relatively small number of ensemble members. In that sense, perhaps the ensemble is not a random selection at all. While the effect of the perturbations on the forecast is considered to be unpredictable beyond the shortest ranges, and therefore random, it does not follow that the underlying distribution is the same as the climatological, or even the conditional climatological, distribution. It is a subject of further work to try to determine on a very large sample of ensemble forecasts the underlying sampling distribution of the forecast weather elements. With respect to the present work, tests have shown that modest changes in assumed underlying distributions (Weibull to gamma) have a small impact on the probability estimates, and so the assumption may be a robust one.

The requirement to fit distributions of an assumed type is both an advantage and a disadvantage. It is an advantage if one has prior knowledge of the family of distributions that the ensemble forecast should follow because it allows such information to be incorporated into the verification. The fitting of distributions can also lead to improved estimates of forecast probabilities from the ensemble (Hamill and Colucci 1998). Distribution fitting becomes a disadvantage when the ensemble is trying to predict more complicated forms such as multimodal distributions. One can allow for this by fitting a series of unimodal distributions as suggested above, provided a cluster analysis has been carried out to identify the different modes. However, this would require the use of a cluster analysis scheme that can determine whether there is sufficient evidence for multimodality as well as determining the number and location of the modes. The cluster analysis presently in use at ECMWF cannot do this.

There are at least two alternatives to the distribution assumption. One would be to estimate probabilities directly from the ensemble. Weights must be applied to the ensemble members, which can be equal if all members are assumed equally likely to occur, or unequal if not. If sufficient data exist, the probabilities can be calibrated using rank histograms (Hamill and Colucci 1997). This approach might be feasible on larger ensembles such as the 50-member ensembles now produced at ECMWF. One is still left with the problem of modeling the probability in the tails of the distribution, outside the ensemble. This could be handled by fitting an extreme value distribution such as the Gumbel (Hamill and Colucci 1998), but the total effort required might be as great as is necessary to use the method described here. Direct estimation of the probabilities would nevertheless lead to automatic accounting for multimodal distributions.

The second alternative to the distribution assumption would be to fit a pdf empirically using one of several distribution-fitting algorithms that are available. This would avoid the need to explicitly identify multimodality and would incorporate in a single procedure estimates of all probabilities, including those in the tails of the distribution. However, it is likely that ensembles would have to be considerably larger before empirical pdf's can be fit successfully. We plan to examine alternatives to the distribution assumptions in future work.

It is useful to consider the new verification measures in the context of an overall strategy for verification of weather element forecasts from the ensemble. If such a strategy were sought, it must not only provide for assessment of the forecasts for a variety of purposes, but also account for the different types of output from the ensemble. Existing verification measures are designed either for deterministic forecasts, involving the matching of a single forecast value with its verifying observation, or for probabilistic forecasts, involving the matching of the probability forecast of a specified event with the observation, normally in the form of a binary variable according to whether or not the specified event occurred. Following a framework for forecast verification (Murphy and Winkler 1987), each type of verification measure for deterministic forecasts can usually be identified with a corresponding measure for probability forecasts, in terms of the attributes of forecast quality that are measured. Thus, the mean square error of a deterministic forecast measures accuracy as does the Brier score or the RPS for a probability forecast (Murphy 1993). Reliability tables can be used to measure the reliability of probability forecasts, which is similar to the use of the mean error (bias) for deterministic forecasts and so on. Ensemble forecasts do not fit quite so neatly into this framework of scoring measures since the output is a collection of deterministic forecasts that are treated as a probability distribution. The new veri-

fication measures represent an attempt to account for this difference, while imposing a minimum of processing on the basic output of the ensemble. It is not necessary to turn the forecasts into deterministic forecasts by selecting a single member, which results in an incomplete verification, or by computing a statistic of the ensemble such as the ensemble mean. Nor is it necessary to impose specific categories on the output and calculate probabilities in order to use the score. We do not recommend verification of the ensemble mean against individual observations for two reasons. First, the ensemble mean has different statistical properties than the individual observation against which it is compared. Second, the ensemble mean is not itself a deterministic forecast because it is not a trajectory of the model. The ensemble mean tends to verify relatively well using quadratic scoring rules such as the rmse because it is a conservative forecast, rarely taking on extreme values. This is perhaps its appeal both as a forecast value and for verification of an EPS.

The relationship between the new score and existing scores for probability forecasts is discussed above. Since its properties are most closely related to the Brier and RPS scores and their respective skill scores, the new scores can be considered to measure the attributes of accuracy and skill in the context of an overall verification strategy. Effectively, the scores indicate the accuracy and skill of the ensemble in locating the observation value, while the Brier and RPS scores can be used directly to determine the accuracy and skill of probability forecasts of specific events generated from the ensemble.

It should be noted that the new verification system is not an attempt to verify the ensemble distribution itself. A probability distribution cannot be verified on the basis of a single outcome. One way to assess the distribution itself would be to assess the reliability (or ''statistical consistency'') of the distribution by comparing all forecast distributions of the same shape and location with the distribution of observations for those occasions (O. Talagrand 1996, personal communication). In practice, this would require a very large sample and is usually not feasible. However, the system described above could be extended to this application by fitting distributions to large numbers of ensembles, matching the parameters of the fitted distributions, and comparing the pooled forecast distribution of matched sets with the corresponding distribution of observations. A Kolmogorov–Smimov test will determine the degree of fit between the pooled forecast and observation distributions.

Finally, both the new scores seek to assess the ensemble output in terms of probabilities. At least, if probability forecasts from the ensemble are accurate, the forecast distribution is useful for estimating probabilities, whether or not the distribution function is of the correct shape and spread. This is an extension and quantification of the concept of ''usefulness'' of the ensemble forecast, as described by O. Talagrand (1997, personal communication). As such, the proposed verification strategy is oriented toward evaluation of the usefulness of the ensemble output to those who may use it in forecasting, rather than to an evaluation of the accuracy of the predicted distributions themselves as a consequence of the uncertainty in the initial conditions of the model. While aspects of the accuracy of the ensemble distribution itself can be determined using a very large sample as described above, or with a rank histogram, the interpretation of such results is not clear. Ensemble forecasts from an EPS such as the ECMWF EPS are interpreted as estimates of the distribution of forecasts arising from uncertainty in the initial conditions. However, the ensemble of errors (observation–forecast differences) also contains a component due to errors in the model simulation, which are not clearly separable from errors due to uncertainty in the initial conditions. One would not expect, therefore, that there should necessarily be a match with the distribution of observations in those circumstances where the model predicts a particular distribution. Furthermore, the perturbations themselves are far from randomly chosen, raising questions about whether the output ensemble distribution is really indicative of the true distribution of outcomes that would arise due to uncertainty in the initial conditions. Perhaps the pragmatic approach advocated here is the best that can be achieved, and most relevant to future use of EPS output.

Acknowledgments. The authors wish to express their appreciation to Horst Boettger for his assistance in obtaining the data used in the examples, and to Marcel Vallée for preparation of the data for use in the verification experiments. We are also indebted to Dr. O. Talagrand for stimulating discussions on ensemble verification strategies. Dr. Harold Brooks and two anonymous reviewers provided helpful suggestions on an earlier version of this manuscript.

As we were finalizing this paper for submission, we were saddened to learn of the passing of Dr. Allan Murphy. It is Allan's ideas that in large part inspired the work that is reported herein and we dedicate this paper to his memory.

REFERENCES

Akesson, 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.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probabilities. *Mon. Wea. Rev.,* **78,** 1–3.

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.

——, and T. N. Palmer, 1995: The singular-vector structure of the atmospheric global circulation. *J. Atmos. Sci.,* **52,** 1434–1456.

——, and ——, 1998: Impact of ensemble size on ensemble prediction. *Mon. Wea. Rev.,* **126,** 2503–2518.

Ehrendorfer, M., 1994: The Liouville equation and its potential usefulness for the prediction of forecast skill. Part II: Applications. *Mon. Wea. Rev.,* **122,** 714–728.

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.,* **8,** 985–987.

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

——, and ——, 1998: Evaluation of Eta–RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.,* **126,** 711–724.

Hays, W. L., and R. L. Winkler, 1971: *Statistics: Probability, Inference, and Decision.* Holt, Rinehart and Winston, 937 pp.

Mason, I., 1982: A model for the assessment of weather forecasts. *Aust. Meteor. Mag.,* **30,** 291–303.

Mielke, P., 1973: Another family of distributions for describing and analyzing precipitation data. *J. Appl. Meteor.,* **12,** 275–280.

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.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting,* **8,** 281–293.

——, and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.,* **115,** 1330–1338.

Silverman, B. W., 1986: *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, 175 pp.

Sivillo, J. K., J. E. Ahlquist, and Z. Toth, 1997: An ensemble forecasting primer. *Wea. Forecasting,* **12,** 809–818.

Somerville, P. N., and S. J. Bean, 1979: Probability modeling of weather elements. Preprints, *Sixth Conf. on Probability and Statistics in Atmospheric Sciences,* Banff, AB, Canada, Amer. Meteor. Soc., 173–175.

Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1990: A survey of common verification methods in meteorology. WMO World Weather Watch, Tech. Rep. 8, 115 pp.

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

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