Analysis of the Spread–Skill Relations Using the ECMWF Ensemble Prediction System over Europe

SIMON C. SCHERRER AND CHRISTOF APPENZELLER

Swiss Federal Office of Meteorology and Climatology (MeteoSwiss), Zurich, Switzerland

PIERRE ECKERT AND DANIEL CATTANI

Swiss Federal Office of Meteorology and Climatology (MeteoSwiss), Geneva, Switzerland

(Manuscript received 21 January 2003, in final form 19 December 2003)

ABSTRACT

The Ensemble Prediction System (EPS) of the European Centre for Medium-Range Weather Forecasts (ECMWF) was used to analyze various aspects of the ensemble-spread forecast-skill relation. It was shown that synoptic-scale upper-air spread measures can be used as first estimators of local forecast skill, although the relation was weaker than expected. The synoptic-scale spread measures were calculated based on upper-air fields ($Z500$ and $T850$) over western Europe for the period June 1997 to December 2000. The spread–skill relations for the operational ECMWF EPS were tested using several different spread definitions including a neural network-based measure. It was shown that spreads based on upper-air root-mean-square (rms) measures showed a strong seasonal cycle unlike anomaly correlation (AC)-based measures. The deseasonalized spread–skill correlations for the upper-air fields were found to be useful even for longer lead times (168–240 h). Roughly 68%–83% of small or large spread was linked to the corresponding high or low skill. A comparison with a perfect model approach showed the potential for improving the ECMWF EPS spread–skill relations by up to 25–30 correlation percentage points for long lead times.

Local forecasts issued by operational forecasters for the Swiss Alpine region, as well as station precipitation forecasts for Geneva were used to test the limits of the synoptic-scale upper-air spread as an estimator of local surface skill. A weak relation was found for all upper-air spread measures used. Although the probabilistic EPS direct model precipitation forecast for Geneva exhibited a considerable bias, the spread–skill relation was recovered at least up to 144 h. A neural network downscaling technique was able to correct the precipitation forecast bias, but did not increase the synoptic-scale spread surface-skill relation.

1. Introduction

The loss of forecast skill of synoptic-scale motions is caused by both the growth of small errors in the initial conditions (Lorenz 1963) and imperfect numerical models that cannot describe the laws of physics exactly (Palmer 2000). The relative importance of the model error is a subject of some discussion in the literature (Roulston and Smith 2003; Orrell et al. 2001). In order to tackle the chaotic behavior in medium-range weather forecasts, a Monte Carlo approach was introduced that extended the deterministic forecast system to a probabilistic forecast system (Palmer et al. 1990). Several runs are computed with different initial conditions spanning their ''space of uncertainty.'' The resulting distribution of forecasts gives the possibility of estimating and sampling the probability density function (PDF) space of possible weather parameters. The idea of creating predictions based upon Monte Carlo experiments was already being discussed in the late 1960s (Epstein 1969; Leith 1974). But since these experiments require huge amounts of computer power, operational probabilistic prediction was not feasible until the 1990s. Today the Monte Carlo system is called the Ensemble Prediction System (EPS) and was introduced at the U.S. National Meteorological Center, now known as the National Centers for Environmental Prediction (NCEP), in 1992 and in the same year at the European Centre for Medium-Range Weather Forecasts (ECMWF) (Toth and Kalnay 1993; Molteni et al. 1996).

The EPS paved the way for many new tools in medium-range weather prediction. First of all it allows an estimation of the entire PDF of forecast variables, which defines the basis for skill prediction (Palmer et al. 1990). It also provides information on possible alternative flow patterns (Chessa and Lalaurette 2001; Eckert et al. 1996). Another application is the prediction of extreme weather events (Buizza and Hollingsworth 2000).

Since the EPS is still under development, a major

Corresponding author address: Simon C. Scherrer, MeteoSwiss, Kraehbuehlstrasse 58, Postfach 514, CH-8044 Zurich, Switzerland.
E-mail: simon.scherrer@meteoswiss.ch

TABLE 1. Major model changes in the ECMWF EPS for the time period considered in this paper (indicated by gray lines in Figs. 4 and 5).

Date	Abbreviation (e.g., cycle)	Model change description
25 Nov 1997	4DVAR	New surface exchange treatment; new snow, albedo, and ice climate
16 Dec 1997	Pack_F	Modified deep convection, radiation, vertical diffusion
31 Mar 1998	cy18_r5	New orography and sea ice albedo
29 Jun 1998	cy18_r6	Use of temperature and surface pressure instead of geopotential heights
9 Mar 1999	cy19_r2	Introduction of 50 model levels
5 May 1999	cy21_r1	New SST analysis; ATOVS active, RTOVS passive*
12 Oct 1999	cy21_r4	New orography and climatology; 40-level EPS; 60-level OPR
27 Jun 2000	cy22_r3	New surface exchange treatment; new snow, albedo, and ice climate
12 Sep 2000	cy23_r1	12-h 4D-variational assimilation system
21 Nov 2000	cy23_r3	T511 for OPR; T255 resolution for EPS

* ATOVS stands for Advanced Television Infrared Observation Satellite (TIROS) Operational Vertical Sounder; RTOVS for Revised TOVS.

task is its evaluation and verification. This represents a major challenge since probabilistic forecast verification is ambiguous, and it is therefore difficult to define what kind of verification measure is best (Atger 1999; Molteni et al. 1996; Strauss and Lanzinger 1995; Wilson 1995).

In this study we attempt to investigate the quality of the ECMWF EPS spread–skill relations from a synoptic perspective, assess how reliable the EPS system is, and determine to what extent upper-air synoptic-scale spread can be used as a first guess for local forecast skill. It is obvious that the enormous amount of data provided has to be condensed not just for meteorologists but especially for communication to the public (e.g., in a newspaper). A simple way to tackle this problem is to define a single confidence index that describes the synoptic-scale forecast uncertainty for a particular target region. A straightforward index is the use of an averaged EPS spread.

Besides a general view on the longer-term variations of the skill and spread measures, the paper investigates spread–skill relations of an EPS toy model. A perfect model situation and an imperfect situation are used to discuss the limits of the spread–skill relations. The same is done with the ECMWF EPS to estimate possible improvements.

In reality the typical end user is interested in the multivariable medium-range forecast confidence for a given city or place on the surface and not in the estimation of skill of the upper-air variable. Hence, in the second part of this paper, the limits of synoptic-scale upper-air spread as an estimator of realistic operational local weather forecast uncertainty in Switzerland is addressed. Finally upper-air spread is used to estimate skill on a station basis.

In the data section 2, a short description of the data, the region of interest, and the time frame considered is given. In the methods section 3, the different spread and skill measures used are introduced. In section 4 longer-term variations of the synoptic-scale skill measures are presented. The same is done with respect to EPS spread in section 5. Climatological aspects of the toy model spread–skill relations as well as the operational ECMWF EPS spread–skill relations are given in section 6. Section 7 provides information on the application of synoptic-scale upper-air spread measures on local and station forecast skill measures. Finally the results are discussed and some conclusions are drawn in section 8.

2. Data

The operational EPS of the ECMWF was used in this study. It is based on 50 perturbations that are computed over the Northern Hemisphere using the singular vector technique; see Buizza and Palmer (1995) for details. The set of perturbations of the initial state allows 50 alternative forecasts to be run with a horizontal resolution of T_L159 and 31 vertical levels prior to November 2000. Thereafter a horizontal resolution of T_L255 and 50 vertical levels are used. Prediction 51 is the unperturbed control (CTR) run with the same horizontal and vertical resolution. Changes in the model configuration are made almost monthly. A compilation of major changes is given in Table 1. These changes have important impacts on the performance of the system (Simmons et al. 2000), but the impacts on the spread–skill relations are comparatively small, as will be shown later on.

Our analysis is based on a dataset covering the period from 1 June 1997 to 31 December 2000, a total of 1310 days. The 51 ensemble members of the 1200UTC run are used on a daily basis. The variables taken into account were the 500-hPa geopotential height surface, subsequently called $Z500$ and the 850-hPa temperature, or $T850$. The data grid was composed of 10 by 10 points, with a gridpoint distance of 2° in latitude and 3° in longitude ranging from 36° to 54°N in latitude and from 9°W eastward to 18°E in longitude (cf. Fig. 1). The area considered was chosen because it corresponds to the region covered by the neural network classification run at MeteoSwiss (Eckert et al. 1996).

3. Methods

a. Verification measures

The verification of probabilistic forecasts is complex. Generally, the verification is a measure of the relation-

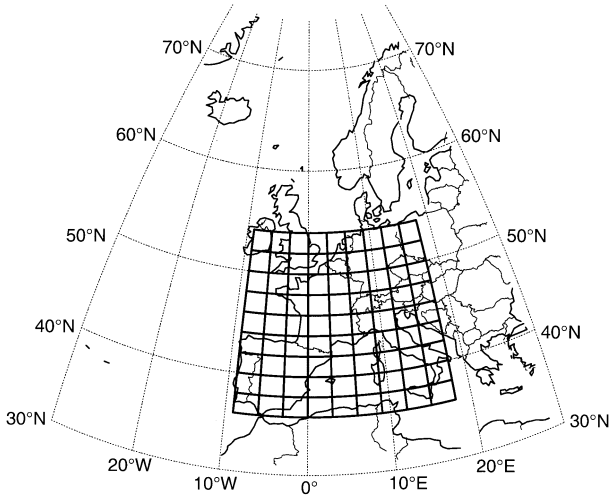


FIG. 1. Geographical domain and location of the 10 × 10 grid points used in this study (mesh grid). Longitudinal resolution is 3° ranging from 9°W eastward to 18°E; meridional resolution is 2°, ranging from 36° to 54°N.

ship between a set of forecasted values and a set of corresponding observations (Wilks 1995). The most basic quantity used to verify an EPS is the skill of the ensemble mean (EM) forecast (Buizza and Palmer 1998). Simple skill measures considered in this study are the ensemble root-mean-square error (rmse), the ensemble anomaly correlation (AC), and a local surface score introduced below.

1) ENSEMBLE ROOT-MEAN-SQUARE ERROR

The rmse operates on the gridded forecast and observed fields by simply averaging the individual squared differences between the two at each of the M grid points (Brankovic et al. 1990):

$$\text{rmse} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(y_m - o_m)^2}, \qquad (1)$$

where y_m is the predicted value of the corresponding variable and o_m the analyzed value for the considered time range. As written above, the formula is valid for one forecast–observation pair. To verify the EPS with 50 members we define the ensemble rmse (rms_{ENS}) as the arithmetic mean of all ensemble members rmse's:

$$\text{rms}_{\text{ENS}} = \frac{1}{N}\sum_{i=1}^{N}\text{rmse}_i, \qquad (2)$$

where N stands for the number of ensembles. Note that the rms_{ENS} is a function of the ensemble spread and the EM error and that it is never zero for an EPS, since the ensemble members are slightly different from each other. This will have an impact on the spread skill relations discussed later.

2) ENSEMBLE ANOMALY CORRELATION

The AC is another measure of association. It operates on pairs of gridpoint values in the forecast and analyzed fields, respectively. The forecast and observation values are first converted to anomalies. Therefore the climatological average of the analyzed field is subtracted at each grid point. The AC is computed according to Wilks (1995):

$$\text{AC} = \frac{\sum_{m=1}^{M}[(y_m - C_m)(o_m - C_m)]}{\sqrt{\sum_{m=1}^{M}(y_m - C_m)^2\sum_{m=1}^{M}(o_m - C_m)^2}}, \qquad (3)$$

where y_m is the predicted value of the corresponding variable, o_m the analyzed value at grid point m, and C_m is the climatological value of the analyzed variable o_m given by the daily climatological value of the field at each grid point separately.

The values are bounded to ± 1 and they are not sensitive to any bias in the forecast. The better the patterns coincide, the more the AC approaches $+1$. In analogy to the rms_{ENS}, an AC_{ENS} that is the arithmetic mean of all ensemble members ACs is defined. A physical complication appears in this simple averaging due to the fact that the AC is a highly nonnormally distributed variable. The average is thus computed after transforming the AC values into an approximately normal distributed variable by applying a classical Fisher z transformation (Ledermann 1984). The z-transformed average is then retransformed into the original space:

$$\text{Fisher } z \text{ transformation,} \quad z_j = 1.151 \log_{10}\left(\frac{1 + \text{AC}_j}{1 - \text{AC}_j}\right);$$

$$\text{averaging,} \quad \bar{z} = \frac{1}{N}\sum_{j=1}^{N} z_j; \quad \text{and}$$

$$\text{retransformation,} \quad \text{AC}_{\text{ENS}} = \frac{10^{\bar{z}/1.151} - 1}{10^{\bar{z}/1.151} + 1}, \qquad (4)$$

where N stands for the number of ensembles again. If all members are identical and the analysis is equal the predicted values, then AC_{ENS} is 1. Note that AC_{ENS} is never 1 for an EPS since the ensemble members are invariably slightly different from each other.

b. Definitions of synoptic-scale ensemble spread

There is no unambiguous definition of the ensemble spread. Many definitions are in use. Perhaps the most obvious one is the standard deviation with respect to the EM or the CTR forecast. Other possibilities in defining spread are, for example, entropy (Eckert et al. 1996; Ott 1993), bin spread (Ziehmann 2001), or the newly introduced mode population (Toth et al. 2001). All classical ensemble spread measures have their drawbacks. They depend on geographic location, annual cy-

cle, and lead time. For this study, the following spread measures were considered (being aware of possible other reasonable spread measures).

1) ROOT-MEAN-SQUARE ERROR SPREAD

An rms spread (SP_{rmse}) is defined, which is constructed as the average of the sum of the rmse's between every EPS member and all other members. It has to be determined for each physical field separately:

$$T^{(i)(j)} = \mathrm{rmse}[W^{(i)}, W^{(j)}]; \qquad \Delta = \underbrace{\sum_i \sum_j T^{(i)(j)}}_{i<j}$$

$$\mathrm{SP}_{\mathrm{rmse}} = \frac{N(N-1)}{2}\Delta, \tag{5}$$

where W^i and W^j represent the N 10×10 gridpoint physical fields (e.g., $T850$, $Z500$) and N is the number of ensembles considered (in the case of the ECMWF EPS, N is 51, including the CTR run). There is no ensemble member reference (like CTR or EM) in this spread measure. Each member is the center once, which implies that the extreme members are assumed to be equally important as the members around the EM.

2) ANOMALY CORRELATION SPREAD

The AC between all ensemble members is computed applying a Fisher z transform again, including the control run. To determine the AC spread (SP_{AC}), a reverse Fisher z transform is performed:

$$T^{(i)(j)} = \underbrace{\mathrm{AC}[W^{(i)}, W^{(j)}]}_{i<j};$$

$$\underbrace{z^{(i)(j)} \equiv 1.151 \log_{10}\left[\frac{1 + T^{(i)(j)}}{1 - T^{(i)(j)}}\right]}_{\text{Fisher } z \text{ transform}}$$

$$\Delta = \underbrace{\sum_i \sum_j z^{(i)(j)}}_{i<j} \qquad \overline{z} \equiv \frac{N(N-1)}{2}\Delta;$$

$$\underbrace{\mathrm{SP}_{\mathrm{AC}} \equiv \frac{10^{\overline{z}/1.151} - 1}{10^{\overline{z}/1.151} + 1}}_{\text{reverse Fisher } z \text{ transform}}, \tag{6}$$

where again $W^{(i)}$ and $W^{(j)}$ represent the 10×10 gridpoint physical fields (e.g., $T850$ or $Z500$) and N is the number of ensembles. Again, no ensemble member reference is considered in this spread measure (like CTR or EM). As for the SP_{rmse}, each member is the center once, which implies that the extreme members are equally important as the members around the EM.

3) GEOMETRIC ENTROPY S

This measure is used to determine the spread of the EPS following a classification of all ensemble members based on a self-organizing Kohonen artificial neural network (NN) used at MeteoSwiss with 12×12 units (neurons) (Eckert et al. 1996). The NN was originally developed to classify weather patterns and applied to find clusters in the EPS forecasts. The 144 units of the neural network were trained with a 20-yr climatology. The classification was done on two fields together: $Z500$ and $T850$. The seasonal mean climatological fields were subtracted before presentation to the NN. The functionality of the resulting network is to map topologically the meteorological fields onto the two-dimensional map of units representing typical weather situations. The topological mapping can be understood in the following sense: two similar forecasts will be associated with two close units on the chart; the two units associated with the forecasts will be distant on the chart if the two forecasts are different. Once the EPS members are classified on the NN, a measure of the order of the elected units distribution on the map is possible using the classical entropy E (Ott 1993), defined by the population frequency for each unit f_{ij}:

$$E = \sum_{f_{ij} \neq 0} f_{ij} \log \frac{1}{f_{ij}}. \tag{7}$$

But the classical entropy is independent of geometrical distance (in our case the physical distance defined by $Z500$ fields between the units on the neural network). Therefore, two neighboring units contribute the same to the entropy as two very distant units. In order to circumvent this fact and incorporate in the spread measure the differences of the forecasted solutions, a new form of entropy, a geometric entropy S, is introduced (Eckert and Cattani 1997, 70–72). This measure takes full advantage of the topology of the neural map, because it includes the distances on the map of the physical fields $Z500$ and $T850$ of the units in an MSE fashion:

$$T^{(ij)(kl)} = \mathrm{MSE}[W^{(ij)}, W^{(kl)}]; \qquad \Delta = \sum_{ij} \sum_{kl} T^{(ij)(kl)}$$

$$Q^{(ij)(kl)} = \frac{1}{N^2} f_{ij} f_{kl} \frac{PQ(PQ-1)}{\Delta} T^{(ij)(kl)}$$

$$S = \underbrace{\sum_{ij} \sum_{kl} Q^{(ij)(kl)}}_{ij \neq kl} \ln\left(\frac{1}{Q^{(ij)(kl)}}\right), \tag{8}$$

where $W^{(ij)}$ and $W^{(kl)}$ represent the weighted physical anomaly fields $T850$ (units in 10^{-1} K) or $Z500$ (units in m) corresponding to the unit $[i, j]$ and $[k, l]$; that is,

$$W^{(ij)} = \alpha W^{(ij)}_{Z500} + (1-\alpha) W^{(ij)}_{T850}; \qquad \alpha = 0.1, \tag{9}$$

where α is empirically determined in order to equally weight the parameters $Z500$ and $T850$. This scheme was applied because the values of the fields are neither nor-

malized (standardized) nor in the same units. Here, N is the number of ensemble members, P and Q are the dimensions of the neural network (in our case, $P = Q = 12$), and f_{ij} and f_{kl} stand for the number of forecasts classified to the neural units $[i, j]$ and $[k, l]$.

c. Skill measure for a regional surface forecast—The KOMIFRI score

In order to verify the ''written'' medium-range weather forecast issued by a forecaster, the so-called KOMIFRI score used at MeteoSwiss is introduced here. This score is semiquantitative and gives an averaged measure for the surface forecast confidence for meteorologically different parts of a small country like Switzerland (area of ~41 300 km^2). Three weather elements are taken into account: precipitation, sunshine, and the 2-m temperature.

The forecast and the verifying observations are first categorized into four precipitation classes, five sunshine classes, and seven 2-m temperature classes. This is done for three meteorologically different regions in Switzerland separately, that is, western, eastern, and southern Switzerland. The KOMIFRI score is constructed by applying an evaluation matrix to the forecasted and the observed classes. The three elements—precipitation, sunshine, and temperature—receive a weight of 40%, 40%, and 20%, respectively. The score is constructed in such a way that the maximum possible score (very good forecast) is 100 and the minimum (very bad forecast) is 0. It is important to note that the weather forecast is accomplished by a human forecaster using a variety of models as well as the ECMWF EPS predictions.

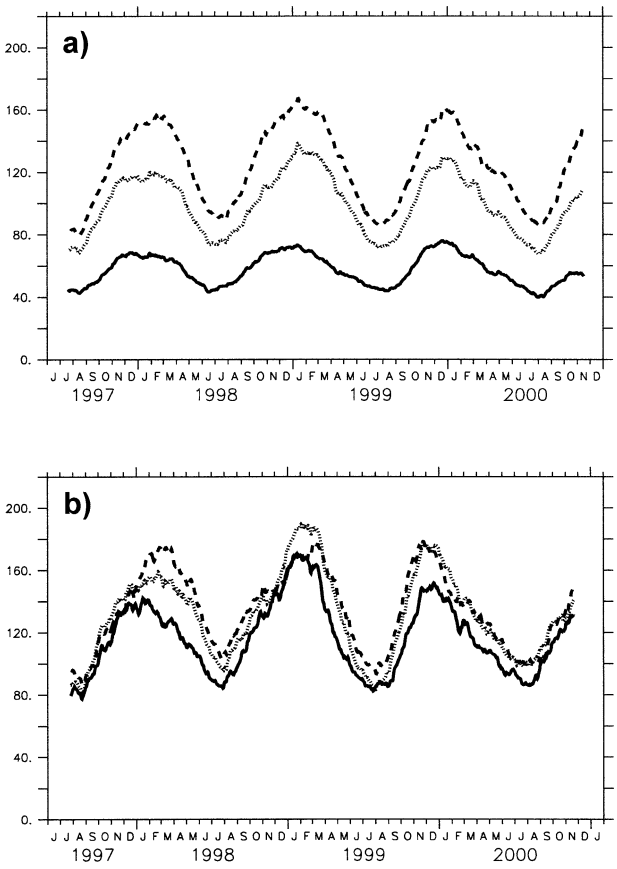


FIG. 2. (a) Rmse time series from 1 Jun 1997 to 31 Dec 2000 of the $Z500$ EPS EM forecast for the 96- (solid), 168- (stippled), and 240-h (dashed) lead time forecasts. (b) As in (a) but for the corresponding 96-, 168-, and 240-h persistence forecasts. All values are 90-day boxcar smoothed. Units are in m in both panels.

d. Skill measure for station-based forecasts

For station-based forecast verifications the Brier skill score (BSS) is used, a skill score measure of the Brier score (BS), which is a measure for dichotomous events, with $o_k = 1$ if the event occurs and $o_k = 0$ if the event does not occur (Wilks 1995):

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{clim}}},$$

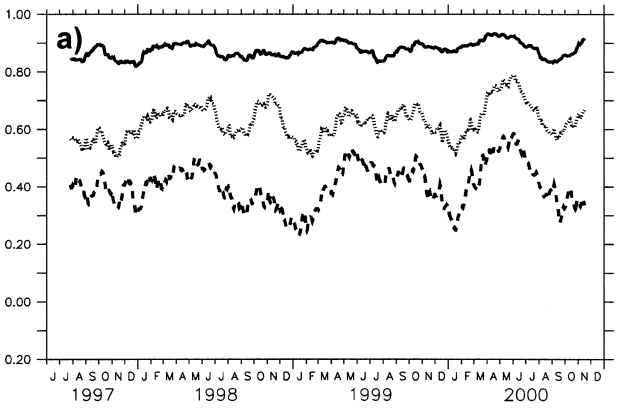
where

$$\mathrm{BS} = \frac{1}{n}\sum_{k=1}^{n}(y_k - o_k)^2, \qquad (10)$$

where y_k stands for the predicted probability of occurrence and n for the number of forecast–event pairs. The $\mathrm{BS}_{\mathrm{clim}}$ is the Brier score of a climatological forecast. A perfect deterministic forecast has a BSS of 1.

4. Long-term variations of skill measures

In order to assess the properties and the value of the various skill measures, their seasonal and year to year variabilities are analyzed. First, the skill of the ECMWF EPS over western Europe for a forecast range of 4–10-day lead time is considered. Figure 2a shows the time series of the $Z500$ rmse with respect to the EM as a domain average. Displayed are 90-day boxcar-smoothed values. There is an obvious seasonal cycle of the rmse values. Low values are observed in summer and high ones in winter. The rmse values and the seasonal cycle increase with increasing lead time (96–168–240-h predictions). No significant trends are found for the period of investigation. In order to understand the seasonal cycle in the rmse, the rmse of a $Z500$ persistence forecast was computed (Fig. 2b). Here an x-h persistence forecast of a field is defined as the particular field x h before the target time of the forecast. The persistence forecasts also show a strong seasonal cycle, but there is only a minor difference between the different forecast lead times. This first of all indicates that the persistence forecast does not have any skill in the medium range and second that the seasonal cycle is an inherent atmospheric property. Note further the strong interannual variability in the persistence forecasts. There are time periods where



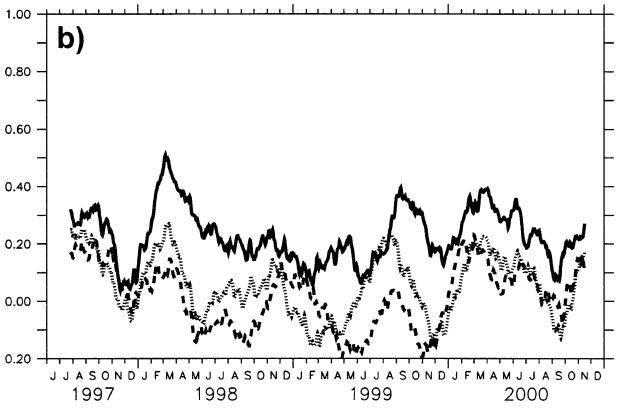


FIG. 3. As in Fig. 2 but for the *Z*500 AC. The AC values are dimensionless.

the rmse of the 96- and 240-h persistence forecasts have almost the same value (e.g., November and December 1997), whereas in others, the 96-h rmse is clearly lower (e.g., March and April 1998). More variability is observed in winter than in summer. Very similar results are found for the *T*850 variable (not shown).

Figure 3a shows the time series for the EPS EM *Z*500 forecast in terms of the AC. There is no clear seasonal cycle in the series as was already found in other studies, for example, in Buizza (1997). In terms of the AC, the *Z*500 EM forecast is useful up to a lead time of at least 168 h (7 days). Again an AC of the persistence forecast has been calculated. Figure 3b shows that these AC values are small and do not have a seasonal cycle but that they show a very irregular behavior. Even slightly lower values are found for *T*850 (not shown). It is important to note that these fundamental differences between the AC and rmse are inherent in the applied measures. This stresses once again that forecast verification depends on the verification measure chosen and on the season considered. For a detailed description of the advantages and drawbacks of AC and rms measures, the reader is referred to Persson (2001).

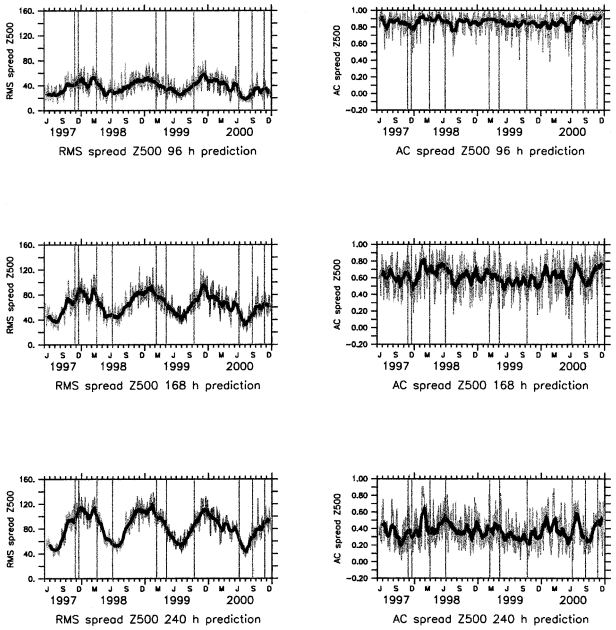


FIG. 4. The 1 Jun 1997–31 Dec 2000 time series of the (left) *Z*500 SP_{rmse} and the (right) *Z*500 SP_{AC} for the (top) 96-, (middle) 168-, and (bottom) 240-h lead time predictions. Daily values are displayed in grayscale. The thick black curve shows the 90-day boxcar-smoothed values. Vertical gray lines indicate important EPS model changes (cf. Table 1). For SP_{rmse} units are in m; SP_{AC} units are dimensionless.

5. Long-term variations of spread measures

Similar to the skill measures, the long-term variation of the spread measures are analyzed using the full time series. Figure 4 shows SP_{rmse} and SP_{AC} as defined in the methods section. Shown are the *Z*500 time series. Analogous to the verification measures, the SP_{rmse} increases with increasing lead time. The SP_{AC} decreases with increasing lead time. Again there is a strong seasonal cycle of the rmse values increasing with longer lead time, whereas no clear seasonal cycle can be identified in the AC time series.

A similar but less obvious picture of the seasonal variation of the SP_{rmse} measure is found for *T*850 (not shown). Clearly lower SP_{rmse} is found for *T*850 after the model change on 27 June 2000. Apparently these model changes led to smaller spreads. Other changes seem not to have a profound influence on the characteristics of the spread.

Figure 5 shows the time development of the geometric entropy, the rms-based spread measure of the neural network classified EPS forecasts. The behavior of the geometric entropy is very similar to the SP_{rmse} shown in Fig. 4. An increasing mean and variance as well as an amplification of the seasonal cycle are observed for longer lead times. Some influences of model changes (indicated with gray lines in Figs. 4 and 5) are seen for the second half of 2000 but otherwise they cannot be linked to strong changes in the spread measures. Hence they are assumed to be of minor importance for this

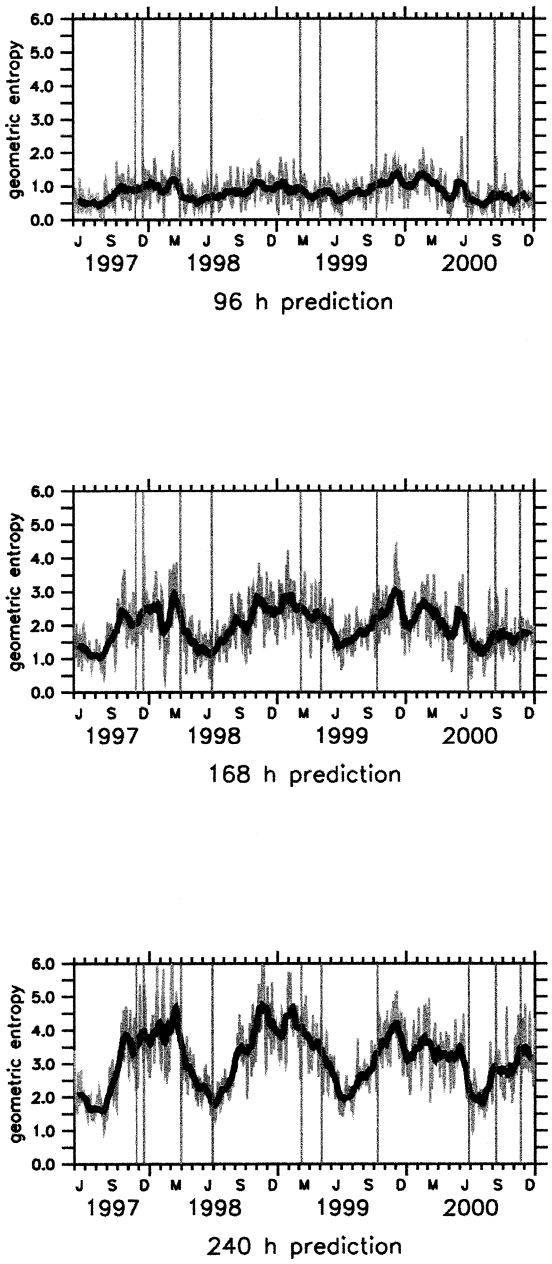


FIG. 5. As in Fig. 4 but for the geometric entropy on the neural network.

study. The neural network MSE-based spread is highly correlated with the direct model output (DMO) SP_{rmse}. The correlation coefficients between DMO SP_{rmse} and NN geometric entropy are 0.75 (96 h), 0.84 (168 h), and 0.88 (240 h). The slight increase in the correlation coefficients with increasing lead time is because of the increasing seasonal cycle. The NN geometric entropy measures how the EPS spans all possible weather scenarios defined by $Z500$ and $T850$. These scenarios are usually more manifold in winter than in summer, consistent with the seasonal cycle.

6. Climatological aspects of upper-air spread–skill relations

Before the upper-air spread–skill relation of the ECMWF system is explored in detail, the expected spread–skill relation is discussed based on a simplified so-called EPS toy model.

a. EPS toy model

For the real forecast system both the rms and the AC spreads are based on statistical relations between two-dimensional fields. Hence, a toy model that simulates synoptic-scale spreads cannot be just a simple known distribution like a one-dimensional normal distribution. In addition, $Z500$ fields are highly autocorrelated in space. This is a fundamental characteristic of the atmospheric flow that needs to be correctly modeled.

The EPS toy forecast is constructed as follows. First, a ''basic'' forecast is randomly chosen from the 21 yr of past analyzed fields. Subsequently the field is normalized; that is, the mean is subtracted and the residuals are divided by the standard deviation. The normalized field provides the base forecast for all 51 members of the EPS forecast to be constructed, but is itself not part of the ensemble. Second, 51 different ''noise'' terms are computed and added to the base forecast with a random weighting for each ensemble and ensemble member to construct the 51 member ensemble; that is,

$$z_i^k = \underbrace{\Gamma^k(z)}_{\text{''basic forecast''}} + \underbrace{\underbrace{N_i^k\langle 0,\, \sigma^k\rangle}_{\text{random factor}}\ \underbrace{\Gamma_i^k(z)}_{\text{arbitrary analysis}}}_{\text{''noise''}}, \qquad (11)$$

where k is the number of ensembles constructed, i is the index for each member, $\Gamma^k(z)[\Gamma_i^k(z)]$ is a random draw for each ensemble (member) from the normalized 21 yr of daily z analyses, $N_i^k\langle 0,\, \sigma^k\rangle$ is a random realization of a normal distribution with zero mean and a standard deviation σ^k that is different for each ensemble, k, and σ^k is a random realization of a uniform distribution of the interval $[\varepsilon = 0.2,\ \omega = 3]$. Because a medium-range forecast system has a spread larger than zero, ε is chosen to be greater than zero. The value 0.2 leads to reasonable results. The upper value of ω is chosen empirically to describe the observed tendency to produce an upper limit for medium-range spread. Temporal correlations are not considered in this simple toy system.

The second important point is how the ''observed'' analysis that is used to verify the toy model is constructed. Two possibilities are considered: a perfect model approach and an imperfect unreliable toy model.

The perfect reliable toy model uses as the observed analysis another draw from the same distribution from which the ensembles members were constructed. The

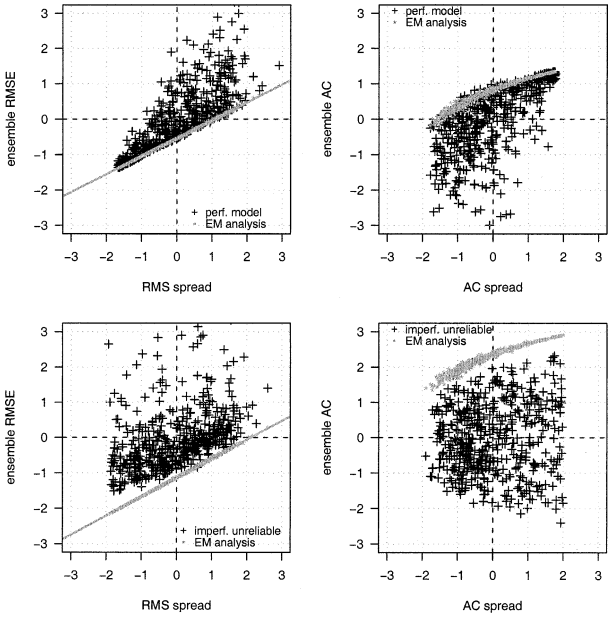


FIG. 6. Spread–skill scatterplots of an EPS toy model. Spread and skill measures are shown in normalized form (black crosses). The lower bound of the error is given by the cases, where the EM was taken as analysis (gray stars and best linear fit as gray line): (left) SP_{rmse} vs rms_{ENS} and (right) SP_{AC} vs AC_{EPS}, for (top) a perfect model approach and (bottom) for an imperfect and completely unreliable system. See text for definition of the different analyses used to determine forecast skill.

imperfect unreliable system uses an observed analysis that is constructed in the same way as an EPS member, with the exception that the ''basic forecast'' is an arbitrary analysis and hence different from the corresponding ensemble. In general this observed analysis has nothing in common with the ensemble members used. Thus it is a prototype of a very unreliable system.

To quantify how well different synoptic-scale spread measures estimate upper-air skill, a comparison of spread with the corresponding skill of the forecast is needed. The simplest and most obvious way to go is to plot spread versus skill. Figure 6 shows spread–skill scatterplots for 500 random realizations of the perfect and the imperfect unreliable EPS toy models described above. Shown are SP_{rmse} versus rms_{ENS} (top left) and SP_{AC} versus AC_{ENS} (top right) for the perfect toy ensemble. The imperfect unreliable toy ensemble results are depicted in the bottom panels. All measures are shown in normalized form. Thus small rms-based spreads appear with negative numbers. Also shown is the minimal rms_{ENS} (as stars and linear regression line) and the maximal AC_{ENS} (as stars) expected that arise when the EM is used as the verifying observed analysis.

The perfect EPS toy model spread–skill relation supports the rule saying that small spread within the EPS is often associated with a high forecast confidence. Earlier works found that large spread on the other hand is not necessarily bound to low forecast skill (Barker 1991; Molteni et al. 1996). There are several large spread cases where the skill is at its lowest possible value (high skill); that is, the analysis is near the EM forecast. This results in a typical fanning in the spread–skill plots and is an inherent property of a skillful probabilistic EPS system.

A different conclusion arises for an unreliable system. Here small spread is linked with low skill since the analysis often lies outside the ensemble. As a consequence the fan gets wider and the rms-based correlation coefficient between spread and skill decreases from 0.78 to 0.43 (Fig. 6, bottom panels). The decrease is even clearer for the AC-based measures, where the correlation coefficient decreases from 0.63 to 0.01. The linear correlation coefficient between rms- or AC-based spread measures is only a simple quantitative measure of association between spread and skill. It should be used with caution, especially since larger spreads are inherently linked to larger minimal ensemble errors as can clearly be seen in both of the rms panels in Fig. 6. This also results in a still positive rms correlation coefficient for the unreliable imperfect system. Alternative and more proper ways of measuring spread error relations that do separate the ensemble mean and spread component such as ignorance or likelihood could be used to circumvent this disadvantage (e.g., Roulston and Smith 2002; Jewson 2003).

b. ECMWF EPS

The quality of the ECMWF EPS spread–skill relation is assessed similarly to the above toy model, that is, (i) by plotting scatterplots of spread versus skill and computing simple linear correlation coefficients, (ii) by building contingency tables based on the cases lying in different quadrants, and (iii) by comparing the actual EPS performance with a perfect and totally unreliable imperfect model approach to assess potential improvements in today's system without considering model errors.

Before quantifying the spread–skill relations, the data have been preprocessed in the following way. Since the seasonal cycle is not a main interest in this study all spread and skill measures were deseasonalized. The low-frequency seasonal signal was removed by applying a 14-day boxcar smoothing and subtracting the smooth signal from the original time series. This makes the distributions and the spread–skill relations more Gaussian than the original ones. As an example for a deseasonalized series, the left panel of Fig. 7 shows the deseasonalized SP_{rmse} time series for 240-h lead times. The higher variance in winter is retained. The PDF is almost Gaussian (Fig. 7, right). For the highly non-Gaussian distributed AC and SP_{AC}, a Fisher z transform was applied to the data before the deseasonalization (Ledermann 1984). In order to analyze all spread and skill measures in a simple and comparable manner, all measures were again normalized.

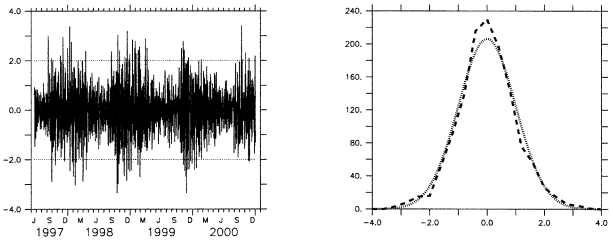


FIG. 7. (left) Deseasonalized SP_{rmse} for 240-h lead time predictions. The thin dotted lines indicate ± 2 standard deviations, respectively. (right) The PDF of the 240-h SP_{rmse} (thick, long dashed) and the fit of a normal distribution (stippled line).

1) SPREAD–SKILL RELATIONS: SCATTERPLOTS AND CORRELATION COEFFICIENTS

Figure 8 shows spread–skill relations using scatterplots of the normalized and deseasonalized SP_{rmse} versus rms_{ENS} for $Z500$. The black lines indicate linear fits to the data. A comparison with the toy EPS spread–skill scatterplots reveals several differences. First the reliable toy EPS shows a narrower spread–skill relation for small spreads than does the actual ECMWF EPS. Second the operational ECMWF relation lacks a clear fanning for high spreads. This may be partially explained by the deseasonalization of the operational EPS data, which made the distributions more normal. Also shown is the lower bound of the error (normalized and deseasonalized rms_{ENS}) given a certain SP_{rmse} (gray lines). The approximate intercept of the ''lowest possible rms_{ENS} given a certain SP_{rmse}'' line was constructed using the forecast case where the observed analysis was closest to the EM. Compared to the EPS toy model the gap between the lowest observed and the theoretically lowest possible rms_{ENS} becomes evident in the operational ECMWF EPS. This indicates that the system is not perfect.

In general the verifying analysis lies within the forecast ensemble, since small spread is predominantly linked with high skill (negative–negative quadrant). In addition the deseasonalized spread–skill relation seems to be approximately linear and symmetric with respect to the 45° line. It is therefore legitimate to use linear spread–skill correlation coefficients and contingency tables as a quality measure for the EPS performance concerning spread–skill relations.

The spread–skill correlation is on the order of 0.7 for the early medium range (96-h prediction). This correlation coefficient decreases with increasing prediction time (0.56 for the 168-h and 0.45 for the 240-h predictions). Simply due to the large number of degrees of freedom ($\gg 100$), a correlation of 0.45 for the 240-h lead time prediction would still be judged statistically significant at the 99% level. The toy EPS on the other hand showed that a completely unreliable system gets an rms-based spread–skill correlation coefficient of order 0.4, suggesting that the relation for the 240-h prediction might not be real.

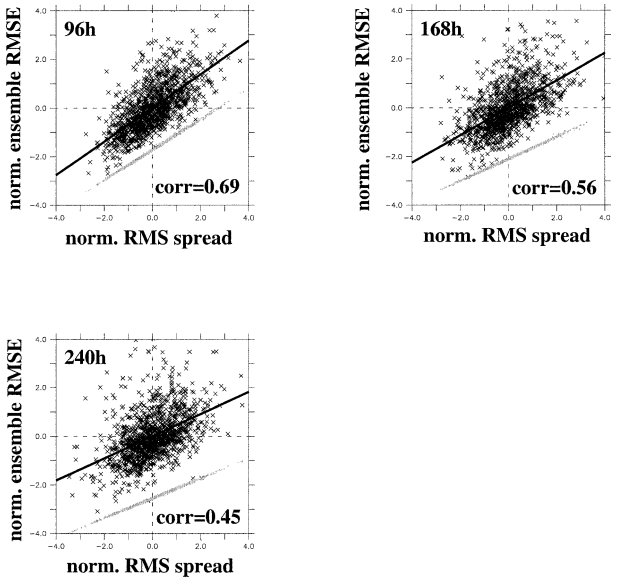


FIG. 8. Scatterplots of daily deseasonalized and normalized SP_{rmse} of the geopotential height at 500 hPa vs deseasonalized and normalized rms_{ENS} at 500 hPa (including a linear fit to the data shown as black line). The lower bound of the spread–skill relation is given by the cases where the EM was taken as analysis (gray plus signs); see text for more information. The time range considered is 1 Jun 1997–31 Dec 2000 for 96-, 168-, and 240-h lead time predictions. Corr values are linear correlation coefficients.

The $Z500$ SP_{AC} versus AC_{ENS} scatterplots are somehow less linear but show a similar picture with correlation coefficients decreasing from very high 0.74 for the 96-h prediction to 0.56 (0.42) for the 168-h (240 h) lead time prediction (Fig. 9). In contrast to the conclusion based on rms measures, a comparison with the toy

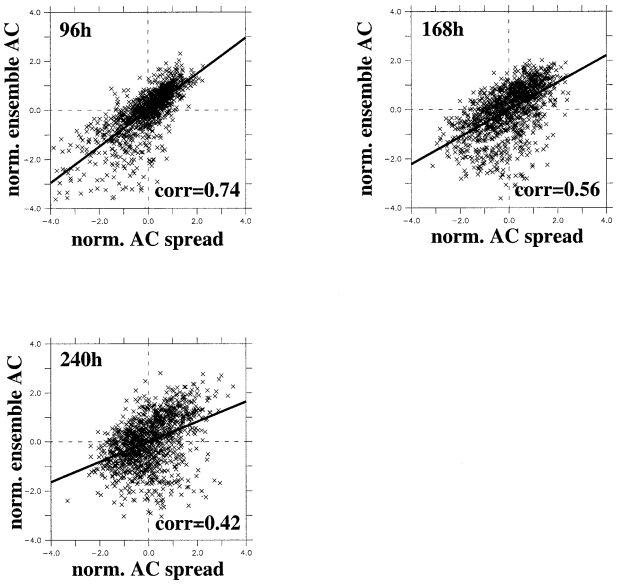


FIG. 9. As in Fig. 8 but for $Z500$ SP_{AC} and $Z500$ AC_{ENS} skill measures.

TABLE 2. Contingency table for $Z500$ SP_{rmse} spread and $Z500$ rms_{ENS} skill. The quadrants are defined by the means. Lead times considered are 96, 168, and 240 h. The numbers are based on all daily EPS predictions between 1 Jun 1997 and 31 Dec 2000. ''Small spread–high skill'' and ''large spread–low skill'' values are in boldface. Note that high skill denotes low rms_{ENS}.

	96 h		168 h		240 h	
*Z*500 rms	Small spread	Large spread	Small spread	Large spread	Small spread	Large spread
High skill (%)	**42.4**	11.7	**40.7**	13.6	**37.8**	16.5
Low skill (%)	8.6	**37.2**	11.8	**34.0**	14.5	**31.1**

EPS suggests that the spread–skill relation is useful up to 240-h lead time. Note that the Fisher *z*-transformed SP_{AC} values do not completely follow a Gaussian distribution. Thus the Fisher *z* transformation is probably not the most optimal transformation method to use in this case and the conclusions drawn from linear measures should be handled with care.

2) CONTINGENCY TABLES

Contingency tables are probably the simplest way of looking at a system's ability to estimate skill from spread. Tables for the rmse and AC spread–skill relations are shown in Tables 2 and 3. They list the percentage number of cases hitting quadrants of values below and above the means of the measure applied. For the *Z*500 rms spread–skill table, 69%–80% of the small (large) spread is linked to the corresponding high (low) skill. Small spread and low skill are less common than large spread and high skill. Large spread is not necessarily linked with low skill. The *Z*500 AC spread–skill contingency table shows that 68%–83% of large (small) SP_{AC}, that is, SP_{AC} values near 1 ($\ll$1), are linked to high (low) skill (Table 3). These results have to be handled with some care since for longer lead times the transformed SP_{AC} values in particular are not normally distributed.

3) COMPARISON WITH PERFECT MODEL APPROACHES AND POTENTIAL IMPROVEMENTS

Similar to the toy model approach discussed above, a perfect ECMWF EPS has been defined where the observed analysis field was one of the ensemble members. The associated spread–skill correlation coefficients are shown in Fig. 10. Overall the rms- and AC-based measures are similar. The correlation coefficients are highest for 96-h lead time and decrease with increasing lead time. The corresponding operational ECMWF EPS spread–skill correlations are lower but show a similar decline with longer lead times. The difference between the perfect model approach and the operational system increases from about 11% (7%) correlation for 96-h rmse (AC) based measures to about 26% (32%) correlation for 240-h lead time. A closer comparison shows that cases exist where the observed analysis lies outside the ensemble, especially for small to medium EPS spread. This is a feature that was also observed in the unreliable toy EPS. Hence, it seems that larger improvements can be expected for longer lead times and cases where the system produces too small spreads.

For completion a completely unreliable imperfect ECMWF EPS was defined where the observed analysis field was an arbitrarily selected analyzed field. No significant spread–skill correlations are found (dashed lines in Fig. 10). Thus the actual ECMWF EPS system is much closer to a perfect case than to the completely unreliable imperfect case.

7. Spread–skill relations for local surface forecasts

Typical end users of operational products need local surface forecasts. Therefore the bench forecaster performance is usually measured in terms of local surface parameters (the sensible weather). In the next two sections, the limits of the synoptic-scale upper-air spread measures as an estimator of skill are tested regarding (a) a local surface score taking into account several weather elements and (b) a skill measure for station-based precipitation forecasts. It tests to what extent synoptic-scale upper-air spread information of single variables such as *Z*500 can be used to determine local skill. It is clear that an alternative would be to use directly local spread to estimate local skill, but this is beyond the scope of this work.

a. Synoptic-scale upper-air spread—KOMIFRI score relations

In Fig. 11, the three standardized synoptic-scale upper-air spread measures are plotted against the KOMIFRI score, a semisubjective score used to verify the

TABLE 3. As in Table 2 but for *Z*500 AC measures. Note that large SP_{AC} means a narrow ensemble.

	96 h		168 h		240 h	
*Z*500 AC	Large spread	Small spread	Large spread	Small spread	Large spread	Small spread
High skill (%)	**49.1**	8.8	**38.8**	14.1	**32.0**	17.6
Low skill (%)	8.3	**33.7**	13.0	**34.1**	14.7	**35.7**

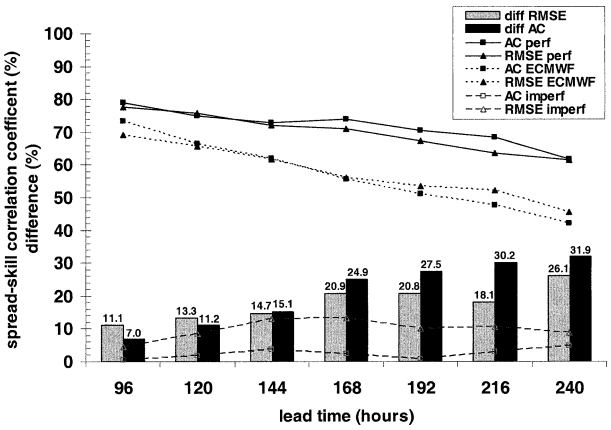


FIG. 10. Spread–skill correlation coefficients and their possible improvements for 4–10-day predictions. The solid lines show the spread–skill relations for the ECMWF perfect model EPS. The operational (imperfect unreliable) ECMWF EPS spread–skill correlations are dotted (dashed). Squared symbols denote AC relations; triangles denote rmse relations. The columns at the bottom show the possible improvement of the ECMWF EPS against the perfect model (%). The gray columns are rmse based; the black ones are AC based.

medium-range predictions at MeteoSwiss as described earlier in the methods section. The general decrease of the score with increasing lead time can be seen by the horizontal lines indicating the mean score for each forecast day. Similar to the upper-air spread–skill relations, all spread measures show a tendency to produce higher local forecast skill with small EPS spread. However the relation is ambiguous. There is also no spread measure that performs persistently better or worse than the others. In particular, the SP_{AC} and SP_{rmse} seem to show a certain unexpected increase in skill with large spread. This may be just an artifact of the analysis caused by the smaller numbers of cases of extreme spread. The number of cases considered in each spread interval is indicated with numbers in Fig. 11.

b. Synoptic-scale upper-air spread and station-based precipitation forecast skill

In this section the relation between synoptic-scale upper-air spread and station-based forecast skill is assessed. The local surface events considered are precipitation sums at Geneva that exceed 1 mm in 24 h. As EPS forecasts, the nearest gridpoint forecast and the NN-classified precipitation forecast are used (grid point La Dôle). The verification sums are station-based measurements from Geneva. The climatological values used are also based on station measurements. Brier skill scores for the >1 mm in 24 h events are computed for different lead times. The BSS values between small (smaller than average spread) and large (larger than average spread) synoptic-scale upper-air spread cases are compared in Fig. 12. Figure 12 shows the BSS for the DMO forecast (top) and for the NN classified forecasts

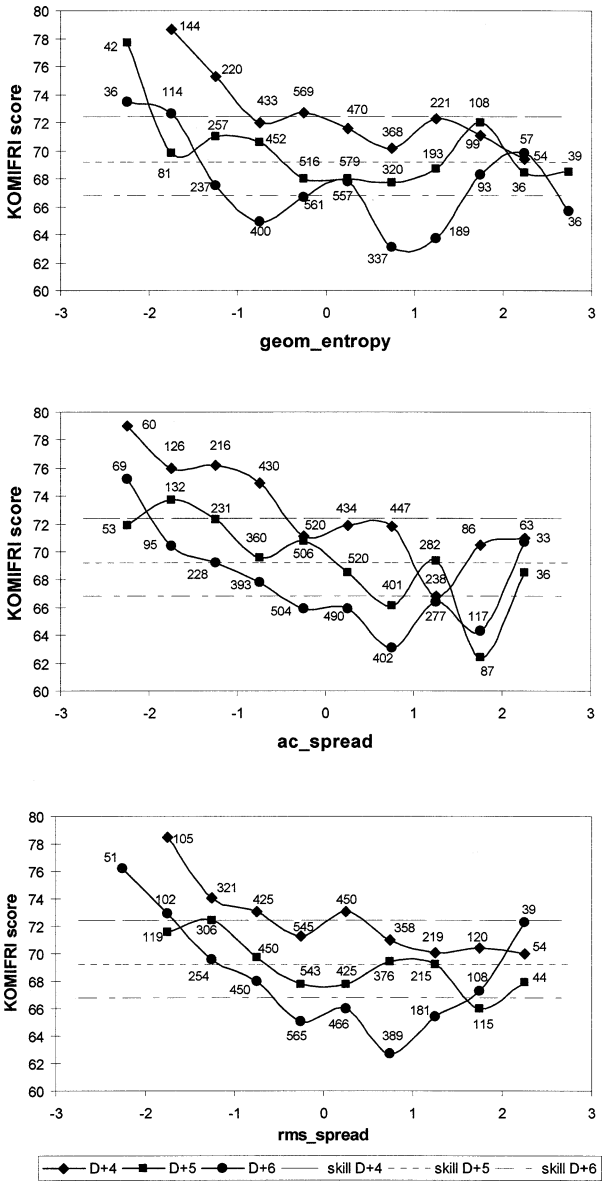


FIG. 11. Plots of standardized synoptic-scale upper-air spread measures against the KOMIFRI scores for the regions western, northern, and southern Switzerland taken together (top) for geometric entropy, (middle) for SP_{AC}, and (bottom) for SP_{rmse}. Shown are the spreads and scores for D+4 (96 h, diamonds), D+5 (120 h, squares), and D+6 (144 h, circles) lead time predictions. The time range considered is 5 Jan 1998–21 Jun 2000. The numbers near the symbols indicate how many cases are condensed in the symbol shown. Also given are the mean scores as horizontal lines for D+4 (96 h, long dashed), D+5 (120 h, short dashed), and D+6 (144 h, dashed–dotted). Extreme spread classes of less than 30 members are omitted.

(bottom). The lead times considered are D+4 (96 h), D+5 (120 h), and D+6 (144 h).

In general, the DMO BSS values are low or even negative for longer lead times. This suggests either very bad forecasts or a strong model bias, which decreases the skill score (Fig. 12a). The fact that the small spread cases show considerably higher skill scores than the

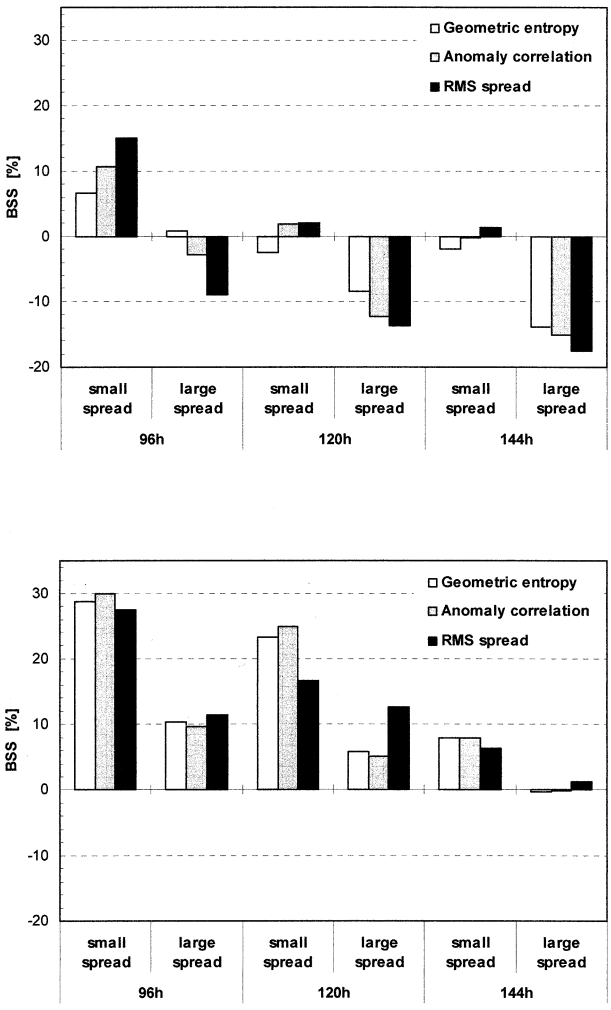


FIG. 12. BSSs for 96-, 120-, and 144-h precipitation forecast sums >1 mm (24 h)$^{-1}$ as a function of the categories 1) small spread, i.e., smaller than average spread of the whole sample, and 2) large spread, i.e., larger than average spread. The spread measures considered are geometric entropy, SP_{AC}, and SP_{rmse}. The station is Geneva. The BSS values are expressed in percent: (top) DMO and (bottom) NN classified.

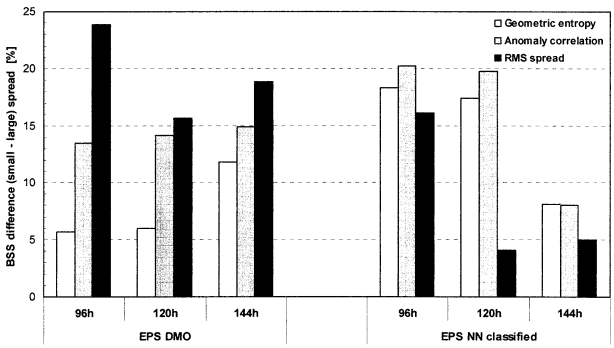


FIG. 13. BSS difference (%) between the small and large spread cases as defined in text for the spread measures geometric entropy, SP_{AC}, and SP_{rmse}. For (left) DMO station-based precipitation forecasts and (right) NN classified station-based precipitation forecasts.

large spread cases indicates that the EPS DMO precipitation forecasts have a strong model bias. This is found for all lead times and all spread measures.

Much better results in terms of absolute BSS values are found for the NN classified EPS forecasts, which indicates the ability to correct the forecast bias (Fig. 12b). The BSS values for the NN are positive or near zero for all three lead times. This may be no surprise, since the NN classification is a downscaling technique: All EPS forecast members were classified to one of the 144 NN units. A climatological precipitation value was associated with each NN unit. The precipitation value was computed as the average of the precipitation values associated with the NN units weighted by the numbers of forecasted hits per unit.

As expected, BSS decreases with longer lead time for both DMO and NN classified cases. No significant difference can be found between the different spread measures with the exception of the geometric entropy in the DMO case, which performs worse than the others.

A primary focus of this study is on skill estimation, that is, on the differences of the forecast performance for small and large synoptic-scale spread cases. Figure 13 shows the BSS differences between the small and large spread cases. This quantity is a direct measure of the ability to estimate forecast skill. For the DMO the difference values are generally between 13% and 24% for all lead times and spread measures (except for the NN-derived geometric entropy; Fig. 13, left). The differences are fairly constant over all lead times, indicating the potential for local skill estimation with synoptic-scale upper-air spread even for lead times of 144 h.

The BSS differences for the NN classified EPS forecasts (Fig. 13, right) are comparable with the DMO values for short lead times (<120 h). Here the geometric entropy clearly performs better than for the DMO forecasts. The ability to estimate skill seems to decrease more rapidly with a lead time greater than 120 h. Overall the NN downscaling is good for bias corrections in precipitation forecasts. The skill estimation of NN classified forecasts based on NN and DMO spread measures is similar to that for the DMO forecasts.

8. Discussion and conclusions

As far back as 1987, Tennekes et al. stressed that ''no forecast is complete without a forecast of the forecast skill.'' In this paper we addressed several aspects of skill estimation using ensemble spread from a forecaster's perspective. The aims of this work were

- to provide insight into the properties of EPS spread and skill from a forecaster's perspective considering a multiyear time series over Europe,
- to test and analyze the ECMWF spread–skill relation for upper-air variables on synoptic scales and compare it with the expected relation from a EPS toy model,

- to point to limitations of the current operational ECMWF EPS, and
- to test an application where synoptic-scale upper-air spread is used as an estimator of local (multivariable) forecast skill.

The skill, spread, and the spread–skill relations of the ECMWF EPS *Z*500 and *T*850 fields have been investigated for the period 1 June 1997 to 31 December 2000. The measures considered were based on root-mean-square error and anomaly correlation. The region of interest was western Europe. It was shown that all rms measures (spread and skill) have a seasonal cycle, increasing with longer lead time. A 240-h SP_{rmse}-based confidence index in winter is up to a factor of 3 different from one applied in summer. The seasonal cycle was shown to be an inherent property of the atmosphere, at least for the variables *Z*500 and *T*850. Hence a skill estimation based solely on SP_{rmse} describes primarily the seasonal cycle of the rmse. This seasonality is not supported by the AC-based skill and spread measures. They do not suffer from a strong seasonal cycle.

A toy EPS was used to discuss basic spread–skill properties using a perfect and an imperfect unreliable system. It was confirmed that small spread is linked with small errors (high skill), whereas large spread can result in low or high skill.

ECMWF EPS spread–skill relations for upper-air fields were tested using spread–skill scatterplots and simple linear correlation coefficients, looking at contingency tables and comparing them with a perfect model approach. It was shown that the ECMWF EPS is in general a reliable system that is much nearer to the perfect model than to an imperfect unreliable system. Nevertheless there is some potential for improvement. The spread–skill correlation for the upper-air fields *Z*500 and *T*850 is high and useful for longer lead times (i.e., 168–240 h). Roughly 67%–83% of small or large spread was linked to the corresponding high or low skill.

A comparison with the toy EPS and the perfect model approach sheds some light on possible improvements. Both approaches showed that the system can be unreliable, especially when the spread is low. Spread–skill relation comparisons between the perfect model approach and the operational system revealed that potential improvements of about 11% (7%) correlation for 96-h rmse (AC) based measures to about 26% (32%) correlation for 240-h lead time would be possible if the observed analysis always lies in the ensemble. Note that this potential for improvement makes the assumption that the ECMWF model represents the true atmospheric system.

In general, the public is interested in skill prediction of local weather forecasts and the forecaster is rated by his or her ability to issue correct multivariable forecasts at a local or station scale. Thus a semiquantitative local skill measure used operationally at MeteoSwiss for the verification of medium-range predictions was used to test the synoptic-scale upper-air spread–local surface skill relation. A weak relation between local skill measures and the spread measures was found. No synoptic-scale spread measure was identified to perform clearly better than the others for surface predictions.

The absolute skill of the probabilistic DMO precipitation forecast on the local and station scale was low or even negative, indicating a potential forecast bias. This bias was successfully corrected by computing skills after a NN classification of the EPS forecasts. The spread–skill relation expressed as the BSS difference of the small and large spread cases was recovered for the DMO as well as the NN classified forecasts and all tested spread measures. The relation was found to be useful up to a lead time of 144 h even on the station scale.

The focus of this work was on the analysis of synoptic-scale upper-air EPS spread measures used to estimate upper-air as well as local surface forecast skill. An alternative approach would have been to use local instead of large-scale spread. Further research in this direction is needed.

Acknowledgments. This study was supported by the Swiss NSF through the National Centre for Competence in Research Climate (NCCR-Climate). Thanks are expressed to Christoph Schär for acting as supervisor of the thesis of Simon C. Scherrer on the subject and to Lionel Peyraud for carefully reading this paper. The comments of three anonymous reviewers lead to a significant improvement of the paper.

REFERENCES

Atger, F., 1999: The skill of ensemble prediction systems. *Mon. Wea. Rev.*, **127,** 1941–1953.

Barker, T. W., 1991: The relationship between spread and forecast error in extended range forecasts. *J. Climate*, **4,** 733–742.

Brankovic, C., T. N. Palmer, F. Molteni, S. Tibaldi, and U. Cubasch, 1990: Extended-range predictions with ECMWF models: Time-lagged ensemble forecasting. *Quart. J. Roy. Meteor. Soc.*, **116,** 867–912.

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

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

——, and ——, 1998: Impact of ensemble size on ensemble prediction. *Mon. Wea. Rev.*, **126,** 2502–2518.

——, and A. Hollingsworth, 2000: Severe weather prediction using the ECMWF EPS. *ECMWF Newsletter*, No. 89, 2–12.

Chessa, P. A., and F. Lalaurette, 2001: Verification of the ECMWF Ensemble Prediction System forecasts: A study of large-scale patterns. *Wea. Forecasting*, **16,** 611–619.

Eckert, P., and D. Cattani, 1997: Classification of ECMWF ensemble forecast members with the help of a neural network. Report on expert meeting on ensemble prediction system (17–18 June 1996), ECMWF, 75 pp. [Available from Library, European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, Berkshire RG2 9AX, United Kingdom.]

——, ——, and J. Ambühl, 1996: Classification of ensemble forecasts by means of an artificial neural network. *Meteor. Appl.*, **3,** 169–178.

Epstein, E. S., 1969: Stochastic dynamic prediction. *Tellus,* **21,** 739–759.
Jewson, S., cited 2003: Use of likelihood for measuring the skill of probabilistic forecasts. [Available online at http://arxiv.org/PS_cache/physics/pdf/0308/0308046.pdf.]
Ledermann, W., Ed., 1984: *Statistics.* Vol. 6. *Handbook of Applicable Mathematics,* J. Wiley and Sons, 1102 pp.
Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.,* **102,** 409–418.
Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.,* **20,** 130–141.
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.
Orrell, D., L. Smith, J. Barkmeijer, and T. N. Palmer, 2001: Model error in weather forecasting. *Nonlinear Processes Geophys.,* **8,** 357–371.
Ott, E., 1993: *Chaos in Dynamical Systems.* Cambridge University Press, 385 pp.
Palmer, T. N., 2000: Predicting uncertainty in forecasts of weather and climate. *Rep. Prog. Phys.,* **63,** 71–116.
——, R. Mureau, and F. Molteni, 1990: The Monte Carlo forecast. *Weather,* **45,** 198–207.
Persson, A., 2001: User guide to ECMWF forecast products. Meteorological Bulletin M3.2, ECMWF, 113 pp.
Roulston, M. S., and L. A. Smith, 2002: Evaluating probabilistic forecasts using information theory. *Mon. Wea. Rev.,* **130,** 1653–1660.
——, and ——, 2003: Combining dynamical and statistical ensembles. *Tellus,* **55A,** 16–30.
Simmons, A. J., and Coauthors, 2000: Forecasting system performance in summer 1999. ECMWF Tech. Memo. 322, ECMWF, 31 pp.
Strauss, B., and A. Lanzinger, 1995: Validation of the ECMWF Ensemble Prediction System. *Proc. Seminar on Predictability,* Vol. II, Reading, United Kingdom, ECMWF, 157–166.
Tennekes, H., A. P. M. Baede, and J. D. Opsteegh, 1987: Forecasting forecast skill. *Proc. Workshop on Predictability in the Medium and Extended Range,* Reading, United Kingdom, ECMWF, 277–302.
Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.,* **74,** 2317–2330.
——, Y. Zhu, and T. Marchok, 2001: The use of ensembles to identify forecasts with small and large uncertainty. *Wea. Forecasting,* **16,** 463–477.
Wilks, D. S., 1995: *Statistical Methods in the Atmosphere.* International Geophysics Series, Vol. 59, Academic Press, 467 pp.
Wilson, L. J., 1995: Verification of weather element forecasts from an ensemble prediction system. *Proc. Fifth Workshop on Meteorological Operational Systems,* Reading, United Kingdom, ECMWF, 114–126.
Ziehmann, C., 2001: Skill prediction of local weather forecasts based on the ECMWF ensemble. *Nonlinear Processes Geophys.,* **8,** 419–428.