Tellus (2000), 52A, 280–299
Printed in UK. All rights reserved

TELLUS
ISSN 0280–6495

Comparison of a single-model EPS with a multi-model ensemble consisting of a few operational models

By CHRISTINE ZIEHMANN, *Universität Potsdam, Institut für Physik, Nichtlineare Dynamik, Am Neuen Palais 10, D-14415 Potsdam, Germany*

(Manuscript received 23 December 1998; in final form 16 December 1999)

ABSTRACT

Since the introduction of operational ensemble forecasts in Numerical Weather Prediction (NWP) more than 5 years ago, the dispute on how to best determine the initial perturbations has largely dominated the direction of research in the field of ensemble prediction. While it is important to consider uncertainties in the initial condition, errors due to model physics or the model numerics and truncation provide another source of forecast errors and might also be considered in ensemble prediction. In this study, we compare the performance of 2 fundamentally different ensemble schemes. First, the ensemble prediction system (EPS) of the European Centre for Medium Range Forecasts is taken as a representative of the single-model approach based on the perfect model assumption and thus taking only the uncertainty in the observations into account. Second, a virtual ensemble comprised of the operational forecasts of 4 NWP centers as a “gratis” candidate of the multi-model approach which, in addition, takes model errors into account. The comparison is based on forecasts of 500 hPa fields over Europe for a summer and a winter period in 1997 and on diagnostics ranging from various measures for the performance of the ensemble means to the statistical consistency and discrimination properties of the ensembles. The different sizes of both ensembles poses the main difficulty for the interpretation of the results. If the ensemble size is not considered as a criterion for the evaluation, the results lead to controversial conclusions; but when penalizing for an overly large and inefficient ensemble the results are for the most part consistent, and one has to conclude that the multi-model ensemble performs better in most forecast aspects.

1. Introduction

The European Centre for Medium-Range Weather Forecasts (ECMWF) and the National Centers for Environmental Prediction (NCEP) have performed operational ensemble forecasts since 1992; at present several other numerical weather prediction centers run ensembles on an experimental or semi-operational level (Ehrendorfer, 1997; Sivillo et al., 1997; Toth et al., 1997). The common idea of ECMWF and NCEP is to take into consideration the uncertainties in the description of the initial state, a view which is justified by indications that the largest forecast errors often (but not always) arise from errors in the initial analysis (Rabier et al., 1996). Details of the ensemble schemes are described in Palmer et al. (1992) and Toth and Kalnay (1993) and references thereof. The primary difference between the schemes is the method used to construct the initial ensemble. This difference has been the subject of dispute (ECMWF, 1996) and shall not be discussed in this paper. Here it is important to note, that the ensemble scheme designs and evaluation methods of both centers imply 2 assumptions. The first implicit assumption is that the forecast model is perfect, since uncertainties in the model are not considered. When probabilities of

e-mail: chriss@agnld.uni-potsdam.de

events are derived from the ensemble the members are usually weighted equally, i.e., it is also assumed that each ensemble member occurs with the same likelihood.

The interpretation of forecast probabilities based upon a perfect ensemble and a perfect model is straightforward. In this case the occurrence of each member of the ensemble is equally likely, and the time evolution of the ensemble reflects the predictability of the system since model and system are equivalent. But even if this perfect model were known, it is not simple to construct a perfect ensemble for a given reference initial condition. If the dynamics is restricted to an attractor with a dimension less than the state space an ensemble distributed via the covariance matrix will include initial conditions not on this attractor. Such an initial ensemble will not reflect a probability distribution consistent with the long-term behavior of the system, which is also called the natural measure on this attractor (Eckmann and Ruelle, 1985). Consequently, at forecast time the ensemble will not reflect the predictability originating from uncertain knowledge in the initial condition, but may be dominated by transient effects. One method suggested by Smith (1996) (see also Smith et al., 1999) to determine a perfect ensemble for a given initial condition is to integrate the model forward in time and pick up those points on the trajectory which fall within a distance smaller than (an estimate of) the typical observational error. These analogs are "identical" with the initial condition within the observational uncertainty and they lie on the attractor. The resulting ensemble forecast distribution has been denoted "accountable" by Smith (1996), because its only shortcoming is due to sampling uncertainty. As the ensemble size increases, the distribution of the ensemble members will converge to the true system probability distribution arising from only observational uncertainty. For operational weather forecasting models, however, such an approach seems impossible.

Of course, models are not perfect. No numerical model simulates the physics and dynamics of the atmosphere perfectly. Madigan — in the discussion of the paper by Draper (1995) — puts this fact into rather pictorial words: "Model uncertainty is the Achilles heel of statistics;" and concludes: "to ignore it is to overstate your certainty and risk making poor predictions". The consequences of model imperfections for ensemble prediction were pointed out by Leith (1983): "A flaw in stochastic dynamic prediction methods and in Monte Carlo approximations to them is that model imperfections are not taken into account and thus sample clusters grow too slowly". The probability distribution derived from ensemble forecasts of imperfect models cannot be improved accountably through an increase of ensemble size. One may sample the forecast distribution better, but this sample distribution may not reflect the system distribution. In the worst case, this forecast distribution would yield unreliable probabilistic forecasts, a useless ensemble mean and a spread with no information on the observed forecast error. Pitcher (1977) recommended 20 years ago, that this could be partly remedied by the introduction of random forcing terms to simulate the effects of model imperfections. The idea of taking model imperfections into account has been realized by the Canadian "system simulation approach" to ensemble prediction which allows uncertainties in the observations as well as the model through different options for parameterizations of processes and uncertainties in surface parameter fields as for example the roughness length and albedo (Houtekamer et al., 1996). The combination of 2 ensembles forecasts from different centers can be also viewed as a system simulation approach and leads to a bigger ensemble of higher quality than the original ensembles (Harrison and Richardson, 1997).

A group of forecasts from different weather prediction centers may be considered a "gratis" multi-model ensemble (Balzer and Emmrich, 1997). This approach uses different models with different physics, numerics and truncations. Each model starts from its own analysis, thus differences in the initial conditions are also present. The disadvantage of the multi-model approach is the limitation to only a few members, however this may be compensated through the fact that the operational models are usually more highly developed (higher resolution, better physics) than the model versions used for the ensemble integration. Naturally, such gratis ensembles have been studied by operational forecasters. In the German Weather Service (DWD) the operational global models of the service and of the ECMWF are operationally interpreted in terms of local weather by means of a statistical interpretation scheme of the perfect prog type (Klein et al., 1959; Wilks, 1995) which uses the 1000 hPa and 500 hPa topography from a few grid-points around Germany as basic predictors and local weather parameters at German stations as predictands (Balzer, 1995). Balzer and Emmrich have demonstrated repeatedly that the average of the statistical forecasts from the 2 models performs better in terms of root mean square error than the statistical forecast from the ECMWF ensemble mean. The advantage over the ECMWF ensemble increases if the arithmetic average of 4 operational models from the ECMWF, NCEP, UKMO, DWD is used (Balzer and Emmrich, 1997). Even the major ensemble system upgrade at the ECMWF in December 1996 with a larger ensemble size and a higher resolution model to run the ensemble did not change the situation.

Stimulated by the above findings we carried out further investigations of these 2 ensembles, the ECMWF ensemble as a representative for a perfect model approach and the ensemble consisting of the ECMWF, NCEP, UKMO, and DWD forecasts as an example for a multi-model scheme. While the previous results of the comparison between these 2 ensembles were mainly based on the verification of local weather parameters at German stations, we will investigate here field forecasts for a larger region. In Section 2 details about both ensembles and the analyzed data sets are given. In Section 3 we introduce the verification strategy and show verification measures for several aspects of the performance of the 2 different ensembles. The results are discussed in Section 4 together with suggestions for the design of new ensemble schemes.

2. Description of the project and analyzed data

The design of this comparison experiment was influenced by 2 practical questions: first, the wish of the German Weather Service to conduct an investigation for a region which is relevant for German weather forecasters and second, the availability of the data on the ECMWF MARS system.

The ECMWF ensemble forecasts used in this study are based on the upgraded ensemble prediction system (EPS) started on 11 December 1996. In this configuration, the single deterministic forecast is integrated using a T213L31 model. The EPS is performed with a model of lower resolution (T159L31) in which a control forecast and an ensemble of 50 forecasts each starting from a slightly perturbed control analysis are integrated for 10 days. The initial perturbations are determined with a linearized version of a model of even lower resolution (T42L31) using the Singular Vector (SV) method (Palmer et al., 1992; Molteni and Palmer, 1993; Buizza and Palmer, 1995; Molteni et al., 1996). Thus the orientations of these SVs are defined by the dynamics of infinitesimal uncertainties which have grown the most with respect to an energy-based metric at 48 h. They are then scaled to be somewhat smaller than the size of the usual analysis error and integrated forward with the nonlinear model (T159L31) used to perform the ensemble forecasts. Obviously, this is a constrained ensemble with perturbations that are not randomly drawn from the distribution of possible analysis errors and it is doubtful that these ensemble members should be equally likely.

The ensemble members comprising the multi-model ensemble are the operational forecasts from the European and 3 national weather forecast centers, namely the ECMWF, UKMO, NCEP, and DWD. For details see "ECMWF (1995)" for the ECMWF operational model, Cullen (1993) for the UK model, Derber et al. (1998) for the NCEP model, and "Deutscher Wetterdienst (1995)" for the DWD model. In case of a multi-model ensemble it seems at first natural to weight the forecasts of these models equally, because it is unknown a priori which model will simulate the atmosphere the best. Optimal a posteriori weights used for the statistical interpretation scheme AFREG were determined for 2 seasons and showed a reasonable agreement with equal a priori weights of 1/4 for each model (Balzer and Emmrich, 1997).

The investigated variable is the geopotential height of the 500 hPa surface at 45 points on a 5° × 5° grid over Europe (10°W–30°E, 40°–60°N, see Fig. 1). Ensemble forecasts, f_i with $i = 1, \ldots, M$, for 24, 48, ..., 144 h are verified using the ECMWF analysis denoted by o. Tests using the UKMO analysis resulted in only small differences; therefore all results shown below are based on the ECMWF analysis. The ECMWF ensemble size is 50 for all forecast times, while the multi-model ensemble consists of 4 members (ECMWF, DWD, NCEP, UKMO) for the first 3 days and of only 3 members (ECMWF, DWD, UKMO) for the

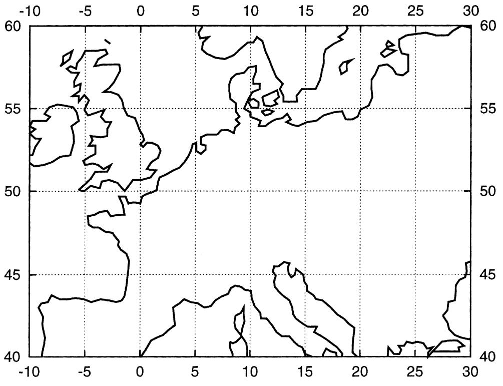

Fig. 1. Investigated region and the grid of 45 points.

longer forecast ranges, for which forecast data from NCEP were not available on the MARS system. These are the 2 fundamentally different ensembles contrasted in this study, and examples for a specific date and grid point are shown in the top 2 panels of Fig. 2 together with the verifying ECMWF analysis.

The lower 2 panels in Fig. 2 represent 2 reference ensemble configurations. They have been introduced to ease the interpretation of differences in the performance of the upper 2 practical ensembles. In the "perfect configuration" (lower left panel) one EPS member is drawn randomly as the verification. The member corresponding to the opposite singular vector perturbation is removed from the ensemble to avoid a biased initial ensemble. Consequently, there remain only 48 members for the perfect ensemble, but this size is still comparable with the original ECMWF ensemble size of 50. Systematic differences between the results obtained for the ECMWF ensemble and the perfect configuration reflect model error, as there is no model error in the perfect ensemble configuration. The "ECMWF sub-ensemble configuration" (lower right panel) consists of 2 pairs of ensemble members, each pair consisting of symmetric perturbations about the control. The 2 pairs are drawn at random from the total of 25 ECMWF pairs without replacement. This 4-member-subensemble is verified against the ECMWF analysis; the results of this reference configuration are helpful in interpreting the effects of ensemble size. Note that if the 4-member-subensemble is built from 4 random and independent ECMWF EPS members its performance is slightly worse than of the one built from 2 pairs. Choosing pairs instead of random members guarantees the coincidence of the ensemble mean with the ensemble mean of the total 50 members (as well as the unperturbed control run) for those time scales in which the dynamics can be considered as linear. This may be the reason for the better performance of the ensemble mean of pairs. Note that after the third forecast day 3-member-subensembles are chosen (one random pair and a single third random member) for a consistent comparison with the multi-model ensemble, which, after 3 days, also consists of only 3 members. In the example presented in the lower right panel of Fig. 2 it seems as if only 3 members are shown even between day 1 and day 3, but this is not the case; one member is not visible in this plot because one pair of EPS members falls on top of each other until day 3. This is certainly not a desirable feature of the EPS since it conflicts with the linearity assumption until day 2 under which anti-parallel vectors should remain anti-parallel (Smith and Gilmour, 1998).

Two periods of about 3 months length within the year 1997 were analyzed. The first data set considers the 1 to 6 day forecasts issued between 1 January and 31 March verifying on 2 January to 6 April ("winter"), the second data set consists of forecasts issued between 1 May and 31 July verifying on 2 May to 6 August ("summer"). Thus, this verification study is based on 2 independent samples each consisting of about 4000 forecasts (90 days × 45 grid-points) for each forecast projection time. Both reference ensemble configurations are repeated 11 × for each forecast with random selections of the 2 pairs of ensemble members (in case of the 4-member-subensembles) or the "verification" (in case of the perfect configuration). Note that these random selections are independent in each case, for each forecast projection time, and at each grid point. This procedure is justified by the order of the evaluation. As explained in the next section, the evaluation is performed first grid-point wise and then completed by averaging over the grid. Throughout this article the verification measures for these 4 ensemble configurations are denoted by the line types introduced in Figs. 2, 4.

3. Verification strategy, methods, and results

The assessment of the quality of an ensemble prediction system is a more difficult task than

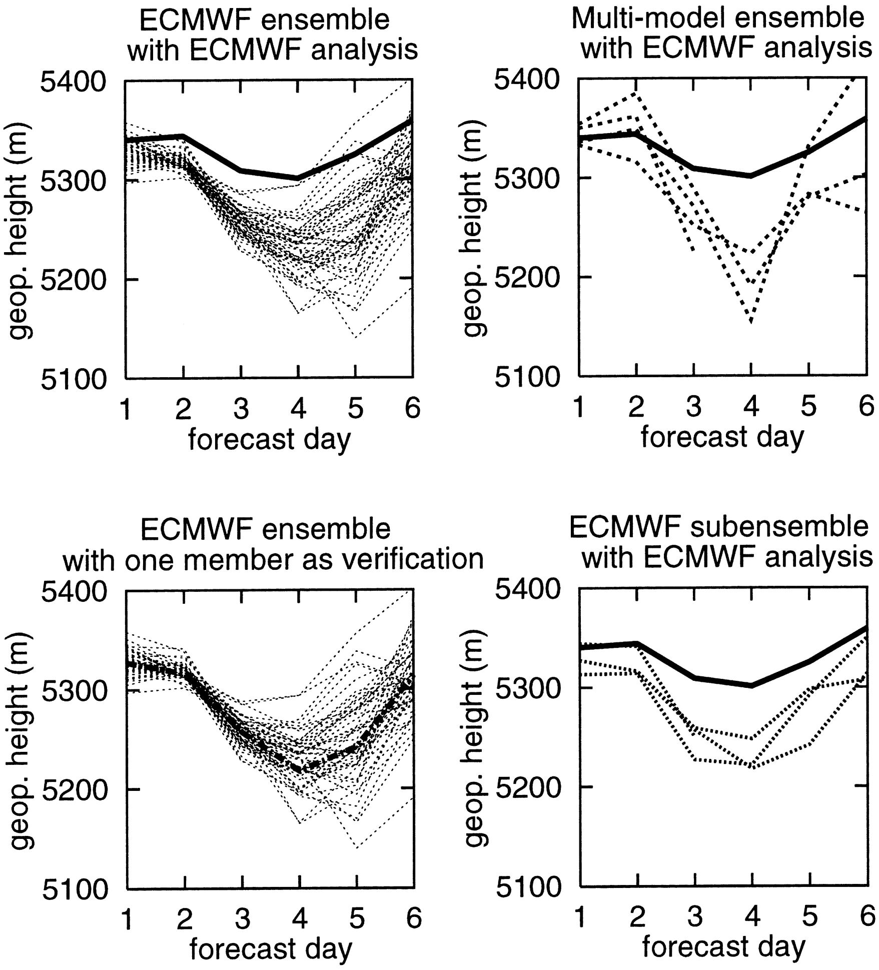

Fig. 2. The different ensemble configurations for grid-point 10°E–50°N and the 1 to 6 day forecasts issued on 1 January 1997. The top left panel shows the ECMWF ensemble (thin dotted) with the ECMWF verification analysis (solid). The right top panel shows the multi-model ensemble (short-dashed) again with the ECMWF analysis (solid). The lower left panel shows the "perfect configuration", where one ensemble member is picked at random and serves as verification analysis (dot-dashed) and the lower right panel shows an ECMWF sub-ensemble (dotted) with the ECMWF analysis (solid).

contrasting single deterministic forecast systems. The basic difficulty is that the predicted object (an approximation of the probability density function of possible states of the system) and the verifying object (the observed state) are of different nature. Several approaches for the evaluation of probabilistic forecasts are given in Murphy and Katz (1985). Recently, special strategies for EPS verification have been developed by Wilson et al. (1998) and new evaluation methods of probabilistic prediction systems by Talagrand et al. (1998). The 3 goals of ensemble prediction:

- to improve the forecast skill by the ensemble mean
- to predict the forecast skill using the dispersion of the ensemble
- to provide reliable and useful probabilistic forecasts

as originally formulated by Leith (1974) provide the guideline for the comparison of these 2 different ensembles. Several verification measures used to quantify the probabilistic aspects of the ensembles can be only calculated for scalar quantities, i.e., grid-point wise. Therefore, all verification measures are defined as sample averages at each grid point denoted by $\langle \quad \rangle$ and then averaged over the grid. There is one exception; the performance of the best ensemble member which will be defined in the next section. The effect of exchanging the order of averaging over grid and sample was tested for the root mean square error (rmse) of the ensemble means and lead to negligible differences.

3.1. Performance of the ensemble mean and the best ensemble member

The average performance of the ensemble mean,

$$\bar{f} = \frac{1}{M} \sum_{i=1}^{M} f_i,$$

relative to the verification analysis, *o*, will be investigated first. Often it is argued that one should not compare the performance of the ensemble mean of a variable with an individual observed value of the same variable. There are 2 reasons for this. First, an ensemble mean can be unphysical and it may not be realizable by the system. This is visualized in Smith et al. (1999) where the time evolution of a perfect initial ensemble in the Lorenz (1963) system is shown. After some evolution time the ensemble "splits" with one part of the ensemble members visiting one wing of the attractor while the other ensemble members chose the other butterfly wing. The ensemble forecast distribution becomes bimodal and the ensemble mean at values close to zero is not an observable state of the system. Second, a mean value is statistically a different quantity with a smaller variance than an individual realization of the same variable. If the ensemble mean is built by M independent realizations of a variable with variance σ^2, then the variance of the mean is only σ^2/M. This has the effect that the rmse of an ensemble mean forecast is in general smaller than that of a single forecast, because the variance of the forecasts with respect to the forecast mean contributes to the rmse. It is well known that the error variance of a single forecast which is drawn at random from the climate distribution is twice as large as the error variance of the climate mean forecast, i.e., the error based on the mean of this distribution (see, e.g., the appendix in Hayashi (1986)).

While we keep the above problems in mind, we nevertheless consider the ensemble mean. First, because we compare the performance of 2 ensemble means with respect to the same observation or analysis. Given the argument of forecast variance reduction, one should then expect that the ECMWF ensemble will show a smaller rmse, simply because its ensemble size is larger. As we shall see below, this is not the case for all forecast times. A second reason to consider the ensemble mean is that it is widely used by forecasters. In addition, we shall also compare the statistics of the best forecast member of each ensemble which does not share the inherent problems of a mean value as discussed in the above paragraph.

First, maps of ensemble mean performance for both the ECMWF and the multi-model ensemble are considered. Fig. 3 shows rmse and bias of both ensemble means for the third forecast day of the winter data set. While the geographical patterns of rmse and bias are similar for both ensembles the magnitudes of the errors are larger for the ECMWF ensemble mean. Next, the distribution of the distances of the ensemble members from the ensemble mean as "indicator for predictability" is investigated. Here, the average squared distance of the ensemble members from their means (spread) is calculated and correlated with the square error of the ensemble mean (skill) at each grid point. The geographical agreement in the rmse and bias fields between both ensembles is not observed for the spread-skill-correlation (Fig. 3, lower panels). The largest correlations are slightly above 0.6 for the ECMWF ensemble but only 0.4 for the multi-model ensemble. Note that the spread-skill-correlation is determined with respect to the ensemble mean for both ensembles, because in the case of the multi-model ensemble no "control" forecast is available. Since the tangent model is centered about the control and not the mean, the spread-skill-correlations for the ECMWF ensemble shown here may be smaller than correlations calculated with respect to the control forecast (Molteni et al., 1996).

Figs. 4a, 5a summarize the results for both the winter and the summer sample as a function of the forecast time. The rmse of the multi-model

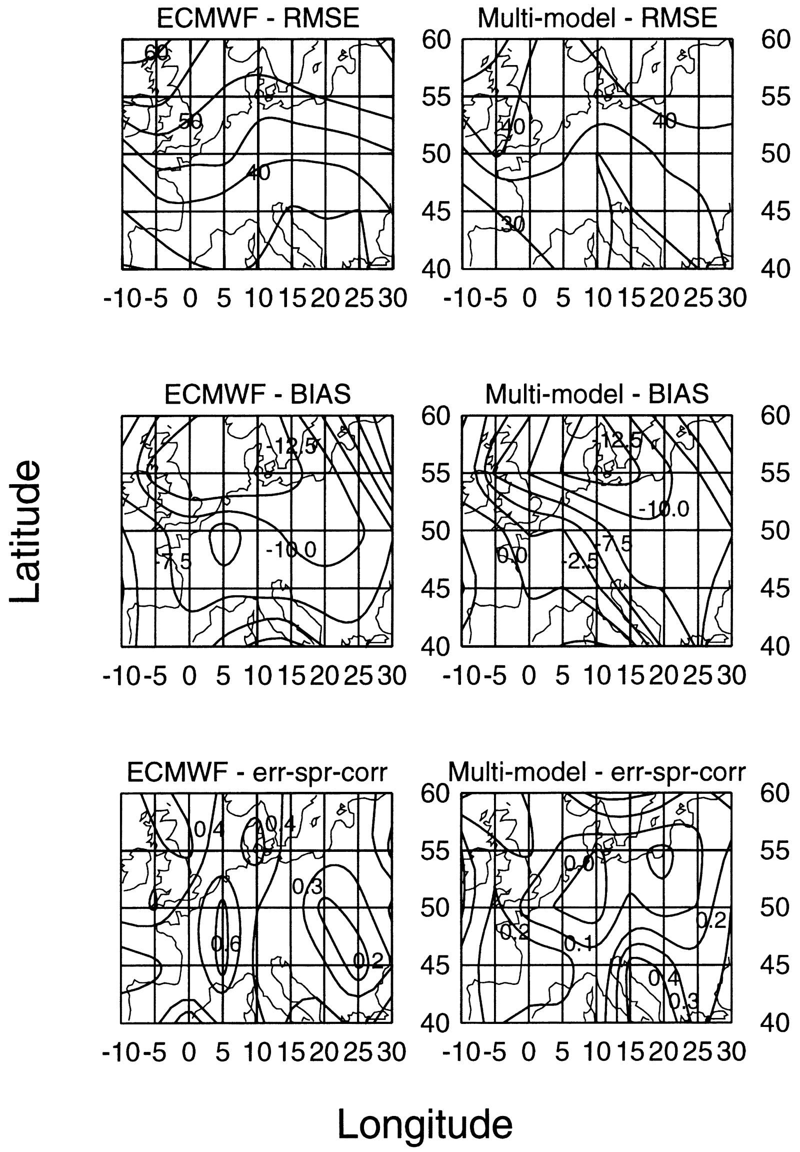

Fig. 3. Root mean square error, bias and spread-skill-correlation of the ensemble mean forecast for the ECMWF (left) and the multi-model ensemble (right) at day 3 in winter 97. Error units in the upper 4 panels are meters.

ensemble mean is somewhat smaller than that of the ECMWF ensemble mean until day 4 in the summer and until day 5 in the winter sample. This result cannot be explained merely due to the larger biases of the ECMWF ensemble mean in both samples: the error variance, $\mathrm{Var}(E) = \mathrm{mse} - \mathrm{bias}^2$, i.e., the mean square error of the forecasted anomalies, $\bar{f} - \langle \bar{f} \rangle$ with respect to the observed anomalies, $o - \langle o \rangle$, shows in winter and summer qualitatively the same behavior as the uncorrected mse. The standard linear correlation coefficient (not the anomaly correlation coefficient) is larger for the multi-model ensemble for all forecast times in the winter sample and up to day 5 in the summer sample (figures not shown).

For a closer interpretation of these results, one

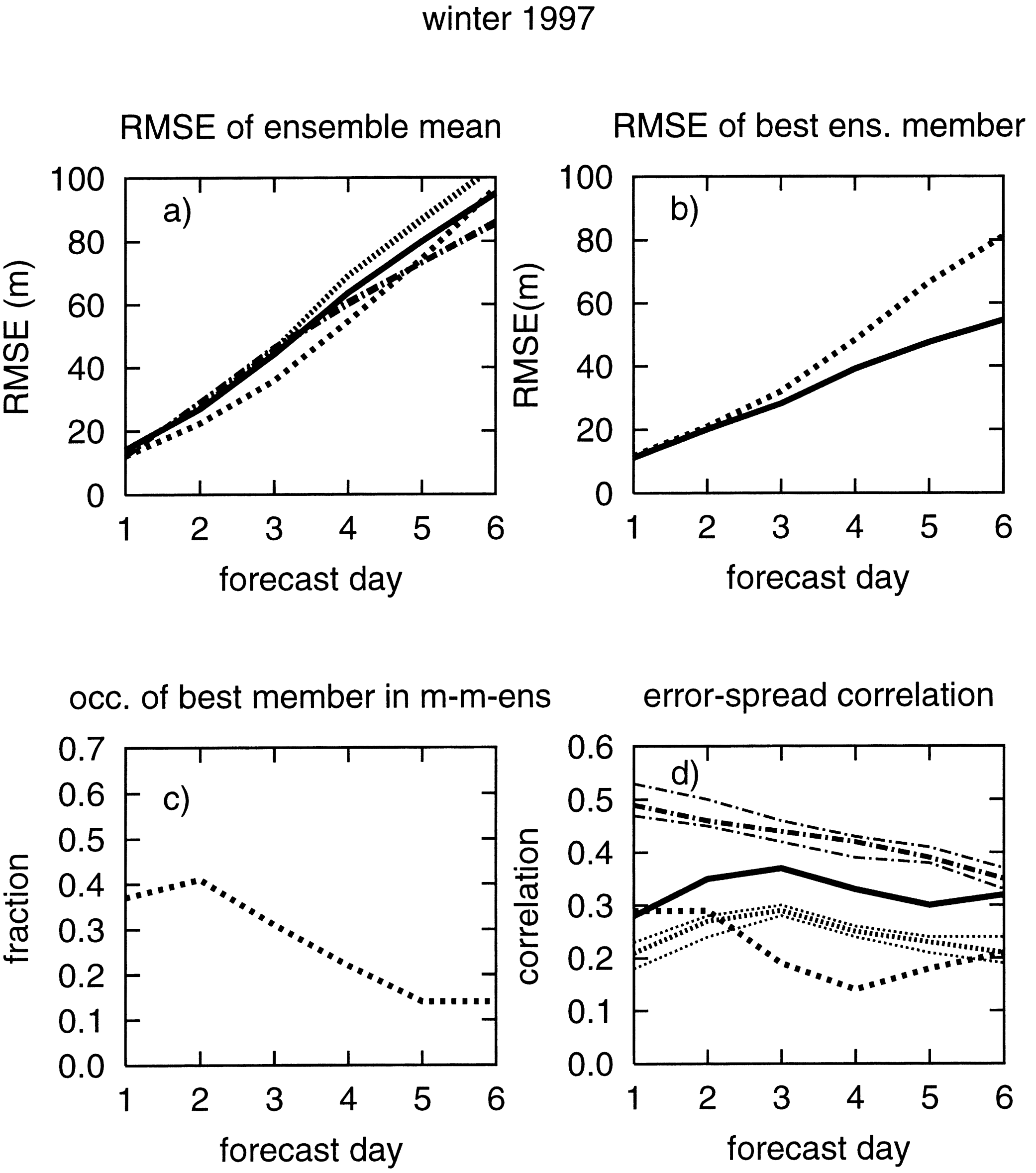

Fig. 4. Summary of some results for the winter sample showing the average rmse of the ensemble mean forecast (a) and of the best ensemble member (b). Panel (c) shows the fraction of cases in which the best member of the *combined* ensemble is found in the multi-model ensemble. Panel (d) shows the spread-skill-correlation. In all following figures the same line styles for the 4 ensemble configurations will be used as here. The ECMWF ensemble with thick solid lines, the multi-model ensemble with thick dashed lines, the perfect configuration dot-dashed, and the 4-member ECMWF subensembles dotted. In case of the 2 reference configurations the median is shown with thick lines and the 10 and 90% quantiles with thinner lines of the same line type. Note that in panel (a) the distributions are very narrow, in panel (d) the 10, 50, and 90% quantiles are better visible.

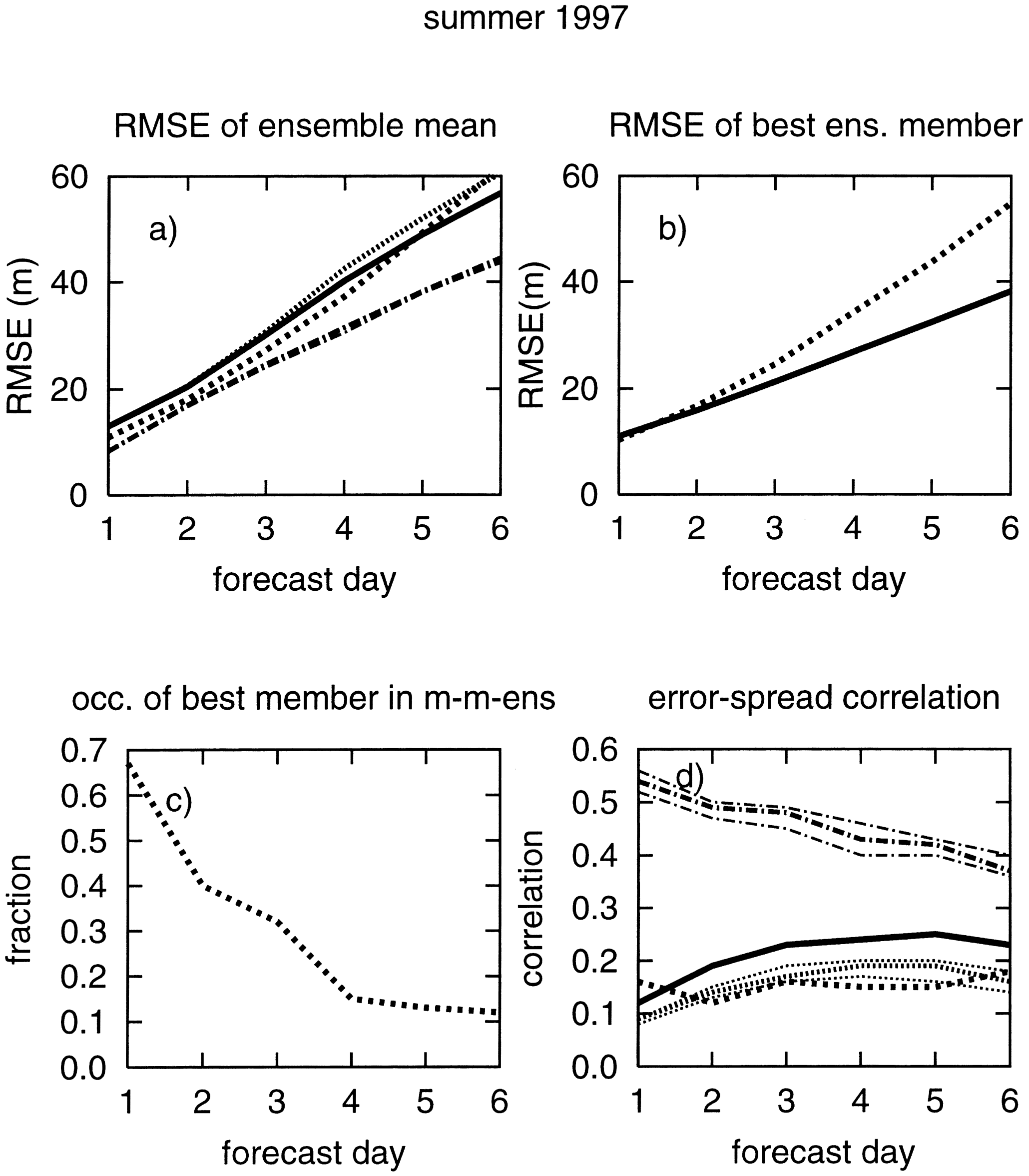

Fig. 5. Same as Fig. 4 but for the summer sample.

has to keep in mind that both rmse and correlation coefficient favor forecasts with small forecast variances Var$\{F\}$, which can be easily seen when the error variance is split up into its components Var(E) = Var$(F-O)$ = Var$\{F\}$ + Var$\{O\}$ − 2Cov$\{F, O\}$. If one is not willing to accept the reduction of forecast variance as a "predictional success" then the correlation weighted by the ratio of forecast and observed standard deviations, Cov$\{F, O\}$/Var$\{O\}$, may be a more suitable verification measure than the correlation itself, because it penalizes forecasts with a too small ratio between the forecasted and observed variances Var$\{F\}$/Var$\{O\}$. In terms of this weighted correlation the multi-model ensemble mean outperforms the ECMWF ensemble mean for all forecast times and in both samples (figures not shown). With increasing forecast time the variance of the ensemble mean forecast Var$\{F\}$ decreases from about 100% of the variance of the observations Var$\{O\}$ at day 1 to about 70% [93%] at day 6 for the ECMWF [multi-model] ensemble in the summer sample and 65% [94%] at day 6 in the winter sample. As expected, the ratio between ensemble mean forecast variance and observed variance is smaller for the ECMWF ensemble because of its larger ensemble size. This ratio would become $(1/M)$% if both ensembles were built from M members randomly sampled from a constant population distribution, which would be the case if the model had forgotten its initial conditions. Until day 6, however, this ratio is still much larger than $(1/M)$% for both ensembles.

The perfect ensemble configuration leads in the summer sample as expected to the smallest rmse (Fig. 5a). It is interesting to note that in the winter sample (Fig. 4a) this is not the case. Since the rmse in the perfect configuration is closely related with the average spread in the sample, one may conclude that in the winter sample the spread is slightly too large at day 2 and 3 compared with realistic forecast errors. The 4-member-subensembles from the ECMWF EPS show consistently larger errors than the multi-model ensemble in both the winter and the summer sample.

Next we present results for the "best ensemble members". The best ensemble member within each ensemble is defined as that member for which the rmse is the smallest over the grid for the forecast time in question. Figs. 4b, 5b show the rmse of the best ensemble member of the ECMWF and the multi-model ensemble in the winter and the summer sample. In both samples the rmse of the best ensemble members do not differ much until the third forecast day, but the rmse of the best ECMWF ensemble member remains considerably smaller than that of the multi-model ensemble for the longer forecast ranges. This demonstrates the greater sampling potential of larger ensembles, however, the error of the best ensemble member is probably also strongly related to differences in the resolution of the models used. Higher resolution models would be expected to show relatively larger errors at longer ranges than lower resolution models because of differences in predictability of small scales compared to large scales (assuming all models maintain an activity level corresponding to the atmosphere's activity). Figs. 4c, 5c show the fractions of cases in which the best member of the combined ensemble (i.e., all members of both the ECMWF and the multi-model ensemble) is a member of the multi-model-ensemble. In agreement with the above rmse results, these figures show a fraction which is much larger at short forecast ranges as would be expected if all members in the combined ensemble were equally likely. This fraction would be $\frac{4}{54}$ (and after the third forecast day $\frac{3}{53} \simeq 0.06$). It should be also noted that the definition of the best member used here is a purely static one, because the best member can be a different one for each forecast time. The best dynamical ensemble member, which is the best but for the total forecast range may be also an interesting verification aspect.

3.2. *Spread-skill relation*

The superiority of the multi-model ensemble is not observed in the spread-skill-relation (Figs. 4d, 5d). Here, the small multi-model ensemble is only advantageous for the first forecast day. This is probably due to the fact that the ECMWF ensemble is known to have much too small a spread at the early forecast times, because its initial spread (at zero forecast time) is chosen smaller than average analysis errors in order to represent realistic day 2 errors (Talagrand et al., 1998). For the larger forecast times the ECMWF ensemble shows the higher (but still not practically useful) correlations. The main reason for this is that the multi-model ensemble is simply too small to capture the expanding orientations in this very high dimensional phase space. This interpretation is supported by the results for the ECMWF 4-member-subensembles which show significantly smaller correlations than the original 50 member ECMWF ensemble. In the winter sample, however, the spread-skill-correlations of the multi-model ensemble are even significantly smaller than those of the ECMWF sub-ensembles. One might expect positive correlations between spread and skill due to the existence of regions in the state space in which uncertainties are more likely to grow or more likely to decay, i.e., the variability of predictability (Smith et al., 1999). This concept of stretching and shrinking regions in state space cannot be directly applied to the multi-model approach, because there are several model state spaces. In some models uncertainties may grow while they may shrink in another. If the majority of the different models in a multi-model ensemble represent reasonable models of the atmospheric dynamics, however, it is expected that such a set of models should also be able to indicate high predictability when the true atmosphere is in a state of extended predictability. To summarize, if the goal is to reach the largest spread-skill-correlations a single model EPS appears to be more appropriate, however, possibly conflicting with another goal, namely to improve the forecast skill by the ensemble mean. It seems, that the 3 goals of ensemble prediction as originally stated by Leith (1974) cannot be optimized simultaneously within a single ensemble scheme.

On the other hand, it is known that even in a perfect model/ensemble environment spread will not be perfectly correlated with the error of any individual forecast (Barker, 1991). In the perfect ensemble configuration the maximum correlations are around 0.5 (0.6) in winter (summer). When the spread is small, then the mean trajectory should be close to the verifying trajectory, however when spread is large this need not be the case. Therefore, a better indication whether the forecast dispersion captures the error behavior may be the fraction of cases, in which the ensemble mean forecast error is smaller than the maximum distance of any ensemble member from the ensemble mean. In a perfect model perfect ensemble environment this fraction should be very close to one; only sampling problems (too few ensemble members) can lead to smaller values. For the multi-model ensemble in the winter sample this fraction is slightly above 50% at day 1 and decreases to 35% at day 6. A prediction that the squared forecast error of the ensemble mean will be smaller than the maximum distance found in the ensemble would be incorrect in more than in half of the cases and is considered as useless. For the ECMWF ensemble the results are a bit better from day 2 on with fractions slightly above 60% decreasing to values around 55% at day 6. In the summer sample, both ensembles yield useless results with fractions not larger than 50%.

3.3. Statistical consistency of the ensemble

Statistical consistency of the ensemble is the next ensemble quality which will be investigated more closely. It is the agreement between the a priori predicted distributions with the a posteriori observations. Talagrand et al. (1998) define this agreement by the condition that for each possible probability distribution p, the a posteriori verifying observations are distributed according to p in those circumstances when the ensemble system predicts the distribution p. One approach to test for statistical consistency is to determine rank histograms (Hamill and Colucci, 1998), which are also called Talagrand diagrams. The same idea has been also published under the term “binned probability ensemble” technique by Anderson (1996). A rank histogram is an extension of the idea of the bias for a single deterministic forecast to the case of an ensemble of forecasts. One determines the frequency of occurrences of the observed values in the intervals defined by the single ensemble members. If the ensemble members are indistinguishable from each other in the sense that they possess the same probability of occurrence, the frequency should be the same for all intervals including the 2 extreme intervals outside the range of ensemble values. An example from the summer sample showing the rank histograms of both ensembles for the day 3 forecasts at grid point 10°E–50°N is given in Fig. 6. Besides the number of cases in all $M+1$ classes the expected number and confidence limits based on the equally likely occurrence of all intervals with $p=1/(M+1)$ are indicated. The probability of drawing randomly a certain interval k times in N trials is binomial distributed

$$P(k)=\frac{N!}{k!(N-k)!}\,p^k q^{N-k} \quad \text{with} \quad q=1-p.$$

The cumulative binomial probability $P_c(k)$ of drawing randomly k or more times a specific ensemble member ranges from 0 for $k=N+1$ to 1 for $k=0$ and provides the lower [upper] confidence bound N_l [N_u] when $1-P_c(N_l)$ [$P_c(N_u)$] becomes larger than 0.05. Numerically it is approximated by the incomplete beta function (Press et al., 1992) because N is too large to calculate the $P(k)$ directly. We see that at this grid point both ensembles significantly predict too high values. Note that the confidence bounds are based on the assumption of independent trials, while the daily data at this grid point are temporally correlated. Nevertheless, the results will remain noteworthy even under wider confidence bounds.

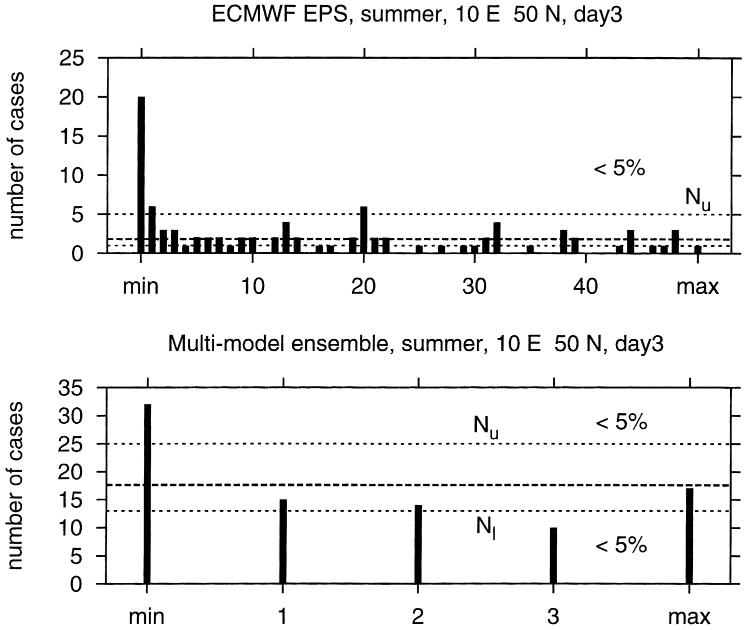

Fig. 6. Rank histograms for the ECMWF (top) and multi-model ensemble (bottom) at grid-point 10°E–50°N for day 3 forecasts of the summer sample. The bars indicate the number cases found in the $M+1$ intervals. The thick dashed lines show the expectation values, the thin dashed lines indicate confidence limits at N_u and N_l defined by the probabilities of finding counts above or below these lines by chance to be smaller than 5%, where independence of the data is assumed.

Another special aspect of statistical consistency is the frequency of "outliers", i.e., cases in which the observation falls outside the range of ensemble values. A certain proportion of observations will be expected by chance to be outliers. Obviously, the larger the ensemble size, the larger is the possibility that the ensemble will cover even extreme values towards the tails of the distributions. To be able to compare the outliers statistics of ensembles of different ensemble sizes, we introduce the "effectivity" r as the ratio of the observed frequency of outliers, f_{obs}, and the probability of outliers expected by chance, p_{theo}. If the ensemble members were all equally likely with the M ensemble members dividing the whole range into $M+1$ equally likely probability classes, then the probability for the outlier class is $p_{\mathrm{theo}}=2/(M+1)$ and consequently the ratio r is

$$r=\frac{(M+1)f_{\mathrm{obs}}}{2}. \tag{1}$$

The optimal case is when $r=1$. If r is larger than 1 the ensemble does not cover the observation, either because its spread is too small and/or the ensemble is biased. If r is smaller than 1, the ensemble does cover the observation but this could be due to worst and best case estimates which are too conservative. In this case the spread is too large. Fig. 7 shows the results for both the winter and the summer sample. As expected the large ensemble size of the ECMWF ensemble provides a smaller percentage of outliers in the winter and the summer sample (7a and c) than the multi-model ensemble. If we compare the effectivity r, however, the multi-model ensemble has almost optimal values around 1, while for the ECMWF ensemble the spread is too small especially on day 1 before it levels at $r\simeq 2$ in the winter and $r\simeq 3$–4 in the summer sample (7b and d). Note that the effectivity r has to be interpreted with care since its maximum value, $r_{\max}$, depends on the ensemble size. The maximum $r_{\max}$ is reached when in each case the verification falls outside the ensemble range, i.e., $f_{\mathrm{obs}}=1$, leading to $r_{\max}=(M+1)/2$. Therefore, $r_{\max}$ is 2.5 for the multi-model ensemble and a factor of 10 larger for the

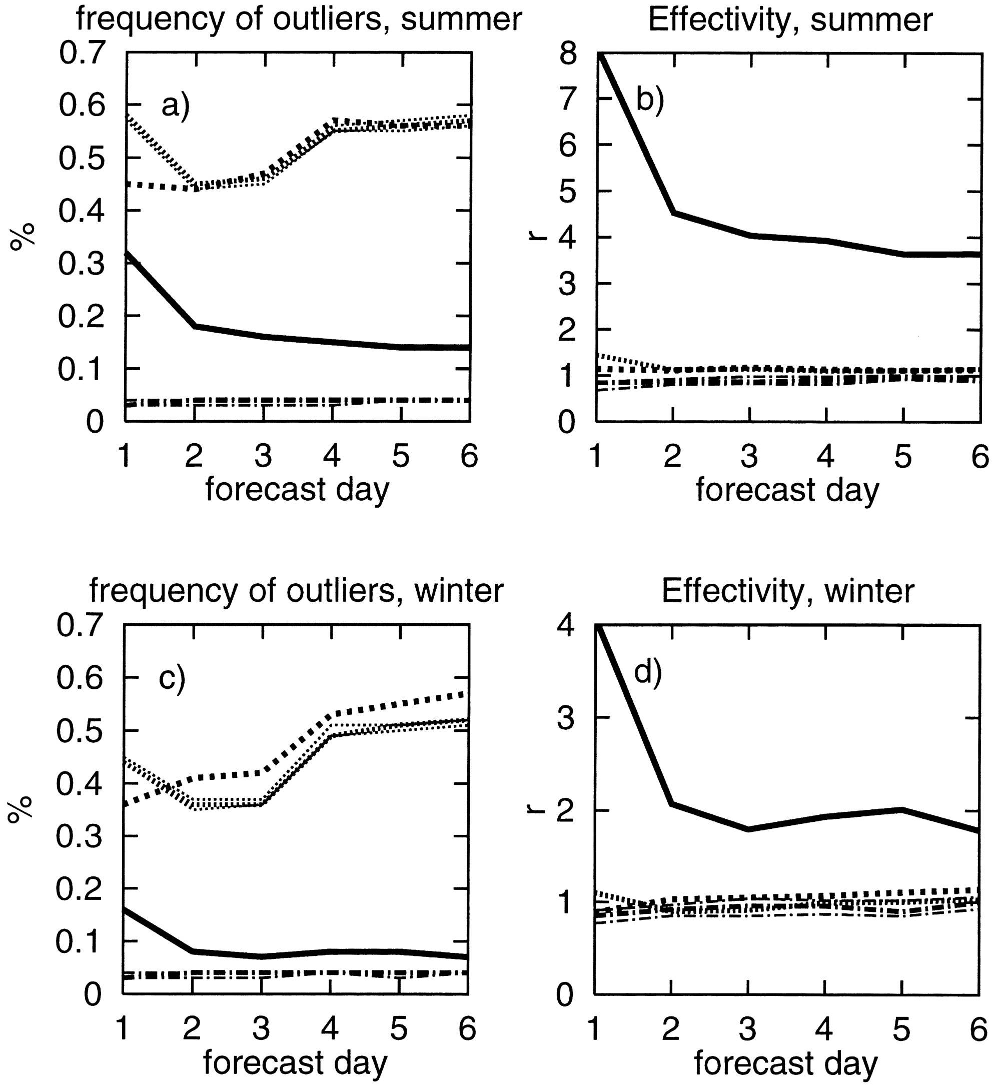

Fig. 7. Relative frequency of outliers, when the verification falls outside the interval spanned by the ensemble members, for the summer (a) and the winter sample (b) and the corresponding effectivities r in (c) and (d). Line types are explained in Fig. 4.

ECMWF ensemble. While the relative frequency favors large ensembles, the effectivity penalizes large ensembles if they do not reflect a wide enough and unbiased (i.e., appropriate) dispersion. In that case the ensemble size could be reduced without much loss of information about the possible forecast range, i.e., the ensemble is not effective. This becomes clear when the ECMWF subensembles are considered; the differences from the multi-model ensemble become very small. The multi-model ensemble shows the best r for the first forecast day. The results for the perfect configuration in Fig. 7b, d are also interesting. The effectivity appears to be only slightly but consistently too small during the first few days in both samples (i.e., even smaller than would be expected from the design of the perfect ensemble which is not exactly 1 but only $(\frac{2}{50})/(\frac{2}{49}) = 0.98$ because the "verification" is drawn from the total of the 50 EPS members but the SV perturbation opposite to it is then disregarded in the perfect ensemble). This behavior is not well understood, and may possibly be related to the results discussed next.

In rank histograms, the individual "names" of the ensemble members are ignored. It is irrelevant which one of the multi-model ensemble is the "DWD-model" or which of ECMWF ensemble is the ensemble member "1". Retaining this information is, of course, important when the ensemble is tested for "equal likelihood of the members". It is interesting to note the appearance of minima and maxima found in the total ensemble at certain ensemble members. This is shown in Fig. 8 for the day 1 and the day 6 forecasts in the winter sample. There appears to be a systematic preference of the minima and maxima to fall on the first few and last few EPS members on day 1, i.e., the ensemble members seem not to be equally likely in representing the maxima or minma values. This is not observed at day 6; now the histogram appears more flat. Fig. 8 is based on 4050 forecasts which are correlated in time and space, and hence might not provide a big enough sample to ensure the significance of this observation. The summer sample shows qualitatively the same behavior and supports the results, however, this systematic preference should be investigated in a further study based on a much larger data set. We also counted the number of cases in which minima and maxima appear as pairs, as would be expected under a perfect linear evolution of the pairs. On day 1, this is the case in about 60% of the pairs, indicating nonlinear effects already at the very early stage of uncertainty growth. A similar result was found

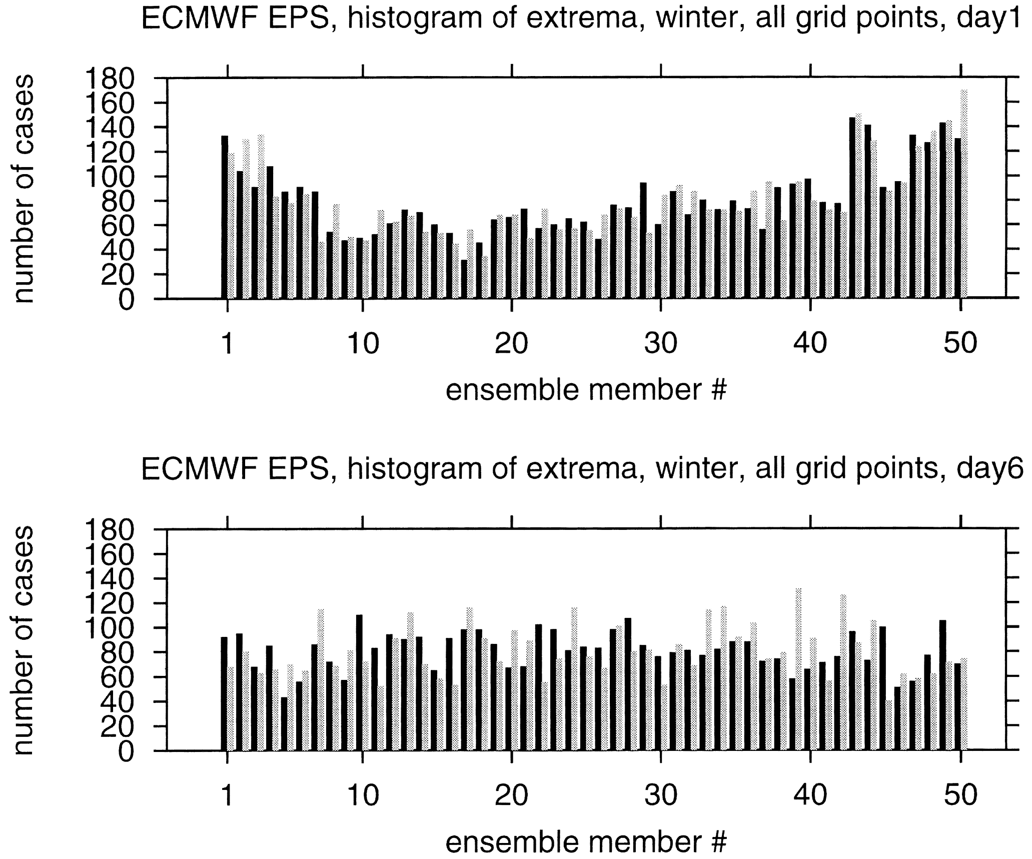

Fig. 8. Number of occurrence of minima (black) and maxima (gray) at certain (named) ECMWF EPS members for day 1 (top) and day 6 (bottom) in the winter sample at all grid points.

by Smith and Gilmour (1998). On day 6, minima and maxima appear as pairs in about 5% of the cases.

It would also be interesting to test how the verifying analysis falls onto the ensemble distribution using a standard (not ranked) histogram. If the ensemble members show different probabilities to be closest to the verification, one might use these a posteriori weights to improve the performance of the ensemble. Each time when an EPS system is changed, however, these a posteriori weights become invalid, therefore the real goal remains to produce ensembles with equally likely members.

Another integrated measure of statistical consistency of an ensemble proposed by Talagrand et al. (1998) is related to the average ensemble skill, i.e., the average squared forecast error of the ensemble mean, $\langle(o-\bar{f})^2\rangle$, and the average ensemble spread, i.e., the average variance of the ensemble forecasts about the ensemble mean $\langle(1/M)\sum_{i=1}^{M}(f_i-\bar{f})^2\rangle$. Statistical consistency requires the quantity

$$\mathrm{ENC}=\frac{\langle(o-\bar{f})^2\rangle}{\left\langle\frac{1}{M}\sum_{i=1}^{M}(f_i-\bar{f})^2\right\rangle}-1$$

to be $2/(M-1)$. Values larger than this indicate a too small spread (or a large bias). Both ensembles yield too large values of ENC in both samples; only days 2 and 3 in the winter sample yield negative values for the ECMWF ensemble thus a too large spread which had been already reported in Subsection 3.1. This result is surprising because often a generally too small spread of the ECMWF ensemble for the first forecast days is reported.

Note that a meaningless ensemble whose members were drawn at random from the climatological distribution would yield a perfect statistical consistency with $r=1$ and $\mathrm{ENC}=2/(M-1)$. The ability of the ensembles to provide not only statistically consistent but also useful probabilistic forecasts will be discussed in the next section.

3.4 *Verification of probabilistic forecasts derived from the ensembles*

For the evaluation of the ensemble in terms of probabilistic forecasts the continuous variable geopotential has to be transformed into a categorical variable by defining a threshold value below (above) which the event is considered to have (not) occurred. This threshold has been determined for all grid-points separately by the extreme values found in the analysis during the respective verification period by $x_{\mathrm{thresh}}=x_{\mathrm{min}}+(x_{\mathrm{max}}-x_{\mathrm{min}})/3$ in order to be sure that the event whose probability is forecasted has occurred at all.

Next, the probability for the occurrence of the event needs to be computed. Wilson et al. (1999) state that one should not compute this probability directly from the ensemble distribution because of its too small size. Instead, they suggest a parametric approach to describe the distribution, first to fit parameters of an a priori chosen distribution and then to estimate the probabilities from the fitted distribution. Here, however, the simple approach of counting the number of ensemble members in which the event occurred and dividing by the ensemble size has been applied. Note that in both cases the underlying assumption is equal likelihood of the ensemble members. The multi-model ensemble is too small to allow probability classes. Therefore, discrete probabilities at $p_i=1/i$, $i=0, M$ are forecasted whenever none, one, two, three or all operational models predict the event. The ECMWF ensemble has been evaluated in the same way with $M+1$ discrete probabilities. In an early version of the paper the ECMWF ensemble was evaluated using the same number of probability bins as for the multi-model ensemble. Using finer bins leads to improved results for the ECMWF ensemble. Therefore we present all results in a consistent way, even when this systematically penalizes the ensemble with the smaller number of ensemble members.

Murphy (1993) suggested a set of 9 attributes to quantify the goodness of probability forecasts. Here we concentrate on reliability and discrimination and determine reliability and ROC diagrams (Stanski et al., 1989; Murphy and Katz, 1985; Mason, 1982). Two scalar quantities are derived from these graphical verification tools. The quantity “reliability” as the variance of the points from the diagonal weighted with the forecast frequency of the probability class, and the *area* above the curve in the ROC diagram. Often the area below the ROC curve is calculated (Stanski et al., 1989); here we choose the area above to orient the scoring in the same way as the reliability calculation, i.e., the smaller both values the better the probabilistic forecasts. Note that reliability is another measure to test for statistical consistency as the rank histograms, the effectivity *r*, and ENC in the previous section; the difference is that now the agreement between the consistency of the forecasted and the observed probabilities is tested. The ROC and its statistics consider the variability in the predicted probabilities and reflects the ability of the forecast to distinguish between conditions preceding occurrence and non-occurrence of the event.

Fig. 9 shows ROC diagrams for the day 6 forecasts in the winter and in the summer sample. The differences between the ECMWF and the multi-model ensemble are not very large, and both ensembles still provide useful information because the curves are still well above the diagonal, and so the area above the curves is still much smaller than 0.5. Fig. 9 also demonstrates nicely that the area above the ROC curve can be misleading if different numbers of probability bins are evaluated. Even if all points of the multi-model ensemble fall exactly on the curve for the ECMWF ensemble, the area will be larger for the multi-model ensemble merely because of the convex shape of ROC curves.

Fig. 10 summarizes the results of this Subsection. The scalar measure reliability is shown for both samples in Fig. 10a, c and indicates that the multi-model ensemble is more reliable in summer during the first 5 days, and in winter during the first 3 days. Since the ECMWF ensemble leads to a more complete estimate of the full ROC curves due to its larger ensemble size, the area above the curves needs to be interpreted with care. Mason (1982) suggested to fit curves assuming a normal-normal model for the hit and false alarm rates. Instead, we only contrast results of those ensembles shown in Figs. 10b, d which have about the same size. As expected, the perfect configuration leads to better results than the ECMWF ensemble. The ECMWF-subensembles yield consistently larger values than the multi-model ensemble, supporting the general result of this study that when taking the ensemble size into account, the multi-model ensemble appears to perform better than the ECMWF ensemble.

As for the ensemble mean results we investigated the effect of the bias of the ensemble also on the measures for effectivity, reliability and discrimination. While one can reduce the scalar reliability remarkably when the bias is removed, all other results remain qualitatively the same.

4. Discussion and conclusion

In this study, the performance of a modern EPS which explicitly takes uncertainties in the initial conditions into account has been compared with a multi-model ensemble which consists of the

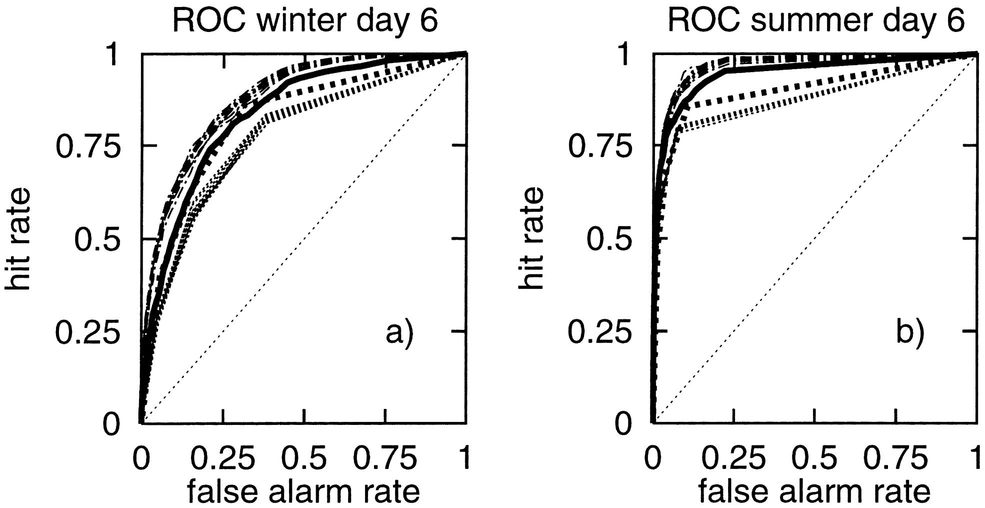

Fig. 9. ROC diagrams at forecast day 6 in the winter (a) and the summer sample (b). Line types are explained in Fig. 4. In this case not the percentiles but the results of *all* 11 repetitions for the 2 reference configurations are shown.

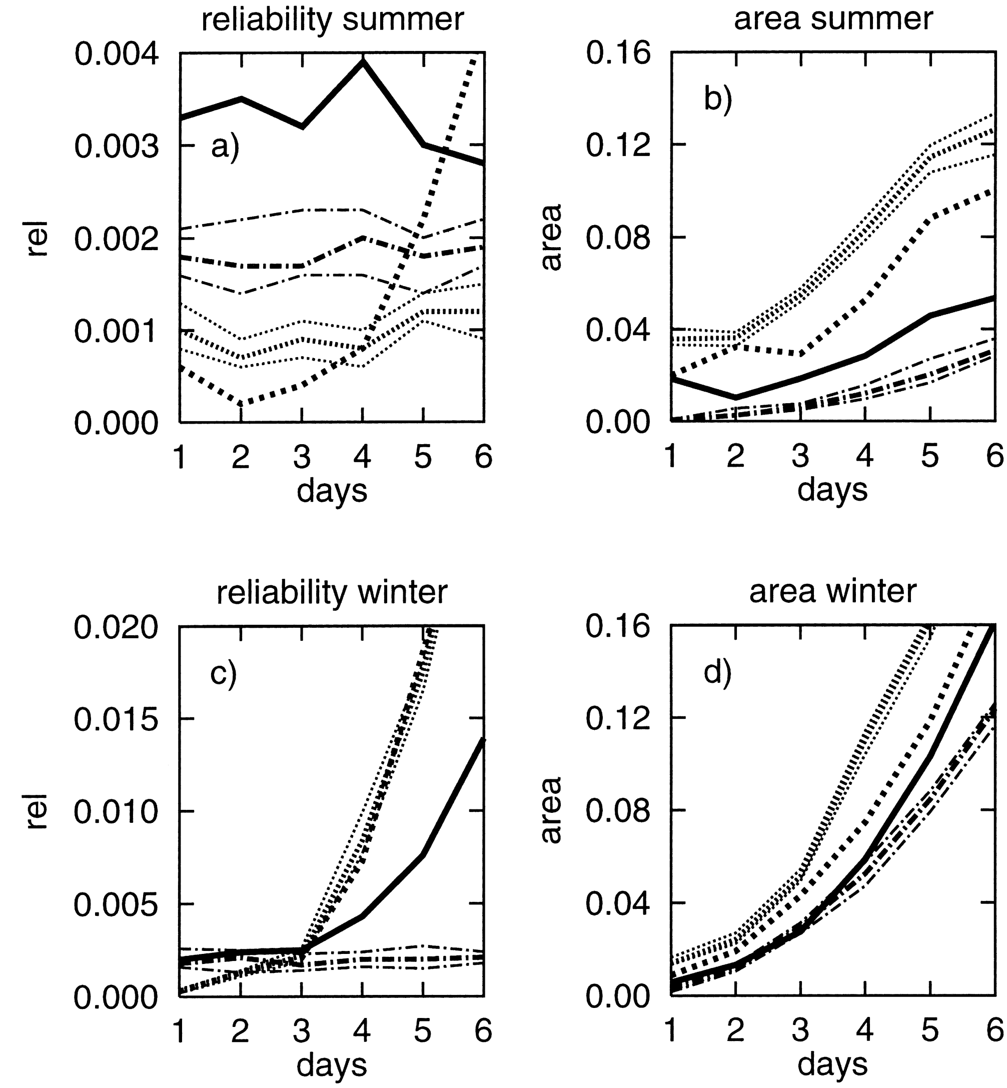

Fig. 10. Reliability and the area above the ROC curves in the summer (a) and (b) and the winter sample (c) and (d). Line types are explained in Fig. 4.

operational forecasts from 4 different national weather forecasting centers and thus implicitly considers model uncertainties.

Guidelined by the originally formulated goals of ensemble prediction, several forecast attributes of both ensembles have been investigated and contrasted. The different sizes of both ensembles poses the main difficulty for the interpretation of the results. If the ensemble size is not considered as a criterion for the evaluation, the results lead to controversial conclusions. For example, the ensemble mean of the multi-model ensemble performs better for most forecast ranges, while the ECMWF EPS can provide a better "best" ensemble member for all forecast ranges because of its larger size. But, if one relates this result to the different ensemble sizes the best member is found disproportionately more often in the multi-model ensemble especially in the earlier forecast ranges (Figs. 4, 5). The larger sampling potential of the ECMWF ensemble also yields the better outlier statistics in terms of frequencies, yet the corresponding effectivities show that this sampling is not efficient (Fig. 7). Thus when taking the different ensemble sizes into account and penalizing for an overly large and inefficient ensemble, the results are for the most part consistent, and one has to conclude that the multi-model ensemble performs better in most forecast aspects. This result is supported by the 4-member-ECMWF-subensembles which also perform worse than the multi-model ensemble in most forecast aspects except the skill-predictability, yet at the cost of smaller skill. The discussion would become more complicated when not only the ensemble size but also costs were included into the evaluation of both types of ensembles which would, for example, arise by adding one new member to the ensemble.

Before concluding we would like to discuss some limitations in this analysis: (i) We do not present uncertainties for the results concerning the ECMWF and multi-model ensemble. Instead, we contrast the results with distributions of 2 reference ensemble configurations. The confidence limits with respect to these medians may provide an estimate for the uncertainties in the results for both the ECMWF and the multi-model ensemble. (ii) By verifying against the ECMWF analysis we may favor the ECMWF ensemble because all EPS members are integrated with a close relative of that model, while the ECMWF analysis is only akin to one of the members in the multi-model ensemble. It would be best to verify against observations; when these are not available either an independent analysis or the average of all used model analyses would be better. In this study, only one other analysis from the UK model was available and tests showed only very small differences in the results. (iii) The definition of a threshold value to transform the continuous variable into a categorical variable contains many more options than the one shown here. Despite of the shortcomings in this study our results are supported by similar findings of Talagrand et al. (1998) and Atger (1999) using independent data sets. Atger (1999) concludes that a multi-model ensemble performs better than the ECMWF and the NCEP ensembles up to +144 h.

Do our results indicate that model deficiencies need to be taken into account and that by tuning the initial conditions alone one cannot force the model close to a trajectory of the real atmosphere? Rabier et al. (1996) used the linear adjoint method and report about some cases when it is indeed not possible to reduce a large forecast error appreciably by a better initial condition. A nonlinear method to address this question is the method of "shadowing" as proposed by Smith and Gilmour (1998), in which one searches for a "dream perturbation" that would bring the model trajectory close to the observation within the observational uncertainty. Since the multi-model ensemble not only consists of different models but also uses different analysis we unfortunately cannot conclude finally from this study that it is the different models, that make the multi-model approach attractive especially in the early medium range. Nevertheless, the fact that a small ensemble consisting of forecasts from 4 different models with different physics, different analysis, different numerical schemes, and different resolutions outperforms a much larger ensemble in several key aspects is an interesting result per se and suggests that model uncertainty remains an important aspect to consider in ensemble prediction methods.

5. Acknowledgments

This study has been carried out under a contract with the German Weather Service (DWD) and special thanks are expressed to Dr. Volker Renner and Konrad Balzer. I received much support from the ECMWF: Tim Palmer helped to initiate this project, Roberto Buizza and Horst Boettger made the data available, and Norbert Kreitz patiently explained the ECMWF MARS system. I benefited from comments of Olivier Talagrand, Peter Houtekamer, and Leonard Smith on an early version of this paper and I am very grateful for the numerous helpful suggestions of Laurence Wilson.

REFERENCES

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

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

Balzer, K. 1995. Automatische Wettervorhersage mittels statistischer Interpretation. *Promet* **24**, 110–118.

Balzer, K. and Emmrich, P. 1997. Verification and intercomparison 1996. In: *Verification of ECMWF products in member states and co-operating states.* Report 1997, pp. 40–49. Available from ECMWF, Shinfield Park, Reading RG2 9AX, UK.

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

Buizza, R. and Palmer, T. 1995. The singular-vector structure of the Atmospheric Global Circulation. *J. Atmos. Sci.* **52**, 1434–1455.

Cullen, M. 1993. The unified forecast/climate model. *Meteorol. Mag.* **122**, 81–94.

Derber, J., Pan, H.-L., Alpert, J., Caplan, C., White, G., Iredell, M., Hou, Y.-T., Campana, K. and Moorthi, S. 1998. *Changes to the 1998 NCEP operational MRF model analysis-forecast system.* Technical Report 449, NOAA/NWS Tech. Procedure Bull. Available from Office of Meteorology, National Weather Service, 1325 East-West Highway, Silver Spring, MD 20910.

Deutscher Wetterdienst, 1995. *Quarterly reports.* Available from Deutscher Wetterdienst, Frankfurter Str. 135, D-63067 Offenbach, Germany.

Draper, D. 1995. Assessment and propagation of model uncertainty (with discussion). *J. R. Statist. Soc.* **B57**, 45–97.

Eckmann, J.-P. and Ruelle, D. 1985. Ergodic theory of chaos and strange attractors. *Rev. Mod. Phys.* **57**, 617–656.

ECMWF, 1995. *User guide to ECMWF products*, edition 2.1. Meteorological Bulletin M3.2. Available from ECMWF, Shinfield Park, Reading RG2 9AX, UK.

ECMWF, 1996. *Predictability (1995)*. Seminar Proceedings, 4–8 September 1995. Available from ECMWF, Shinfield Park, Reading RG2 9AX, UK.

ECMWF, 1998. *Predictability (1997)*. Seminar Proceedings, 20–22 October 1997. Available from ECMWF, Shinfield Park, Reading RG2 9AX, UK.

Ehrendorfer, M. 1997. Predicting the uncertainty of numerical weather forecasts: a review. *Meteorologische Zeitschrift N.F.* **6**, 147–183.

Hamill, T. M. and Colucci, S. J. 1998. Evaluation of Eta-RMS ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.* **126**, 711–724.

Harrison, M. and Richardson, D. 1997. *Developments in ensemble forecasting at UKMO.* Expert meeting on ensemble prediction system. Available from ECMWF, Shinfield Park, Reading RG2 9AX, UK, 78–80.

Hayashi, Y. 1986. Statistical interpretations of ensemble-time mean predictability. *J. Meteor. Soc. Japan* **64**, 167–181.

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

Klein, W. H., Lewis, B. M. and Enger, I. 1959. Objective prediction of five-day mean temperatures during winter. *J. Meteor.* **16**, 672–682.

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

Leith, C. 1983. *Large-scale dynamical processes in the atmosphere.* chapter: *Predictability in theory and practice.* Academic Press, N.Y.

Lorenz, E. 1963. Deterministic nonperiodic flow. *J. Atmos. Sci.* **20**, 130–141.

Mason, I. 1982. A model for assessment of weather forecasts. *Aus. Met. Mag.* **30**, 291–303.

Molteni, F., Buizza, R., Palmer, T. and Petroliagis, T. 1996. The ECMWF ensemble prediction system: methodology and validation. *Q. J. R. Meteorol. Soc.* **122**, 73–119.

Molteni, F. and Palmer, T. 1993. Predictability and finite-time instability of the northern winter circulation. *Q. J. R. Meteorol. Soc.* **119**, 269–298.

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

Murphy, A. and Katz, R. 1985. *Probability, statistics, and decision making in atmospheric sciences.* Westview, Boulder.

Palmer, T., Molteni, F., Mureau, R., Buizza, R., Chapelet, P. and Tribbia, J. 1992. Ensemble prediction. Technical report, Research Department Tech. Memo. No. 188. Available from ECMWF, Shinfield Park, Reading RG2 9AX, UK.

Pitcher, E. 1977. Application of stochastic dynamic prediction to real data. *J. Atmos. Sci.* **34**, 3–21.

Press, W., Teukolsky, S., Vetterling, W. and Flannery, B. 1992. *Numerical recipes* (2nd edition). Cambridge University Press.

Rabier, F., Klinker, E., Courtier, P. and Hollingsworth, A. 1996. Sensitivity of forecast errors to initial conditions. *Q. J. R. Meteorol. Soc.* **122**, 121–150.

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

Smith, L. 1996. *Accountability and error in ensemble forecasting.* In: ECMWF (1996).

Smith, L. and Gilmour, I. 1998. *Accountability and internal consistency in ensemble formation.* In: ECMWF (1998).

Smith, L., Ziehmann, C. and Fraedrich, K. 1999. Uncertainty dynamics and predictability in chaotic systems. *Q. J. R. Meteorol. Soc.* **125**, 2855–2886.

Stanski, H., Wilson, L. and Burrows, W. 1989. *Survey of* *common verification methods in meteorology*. World Meteorological Organization, World Weather Watch, Technical Report No. 8.

Talagrand, O., Vautard, R. and Strauss, B. 1998. *Evaluation of probabilistic prediction systems*. In: ECMWF (1998).

Toth, Z. and Kalnay, E. 1993. Ensemble forecasting at NMC: the generation of perturbations. *Bull. Am. Meteorol. Soc.* **74**, 2317–2330.

Toth, Z., Kalney, E., Tracton, M., Wobus, R. and Irving, J. 1997. A synoptic evaluation of the NCEP ensemble. *Wea. Forecasting* **12**, 140–153.

Wilks, D. S. 1995. *Statistical methods in atmospheric sciences*. Academic Press, San Diego.

Wilson, L. J., Burrows, W. and Lanzinger, A. 1999. A strategy for verification of weather element forecasts from ensemble prediction systems. *Mon. Wea. Rev.* **127**, 956–970.