Tellus (2003), 55A, 16–30
Printed in UK. All rights reserved

TELLUS
ISSN 0280–6495

Combining dynamical and statistical ensembles

By M. S. ROULSTON[1,2]* and L. A. SMITH[1,2], [1]*Pembroke College, Oxford University, Oxford, UK;* [2]*Centre for the Analysis of Time Series, London School of Economics, London, UK*

(Manuscript received 10 September 2001; in final form 1 July 2002)

ABSTRACT

A prediction accompanied by quantitative estimates of the likely forecast accuracy is inherently superior to a single "best guess" forecast. Such estimates can be obtained by "dressing" a single forecast using historical error statistics. Dressing ensemble forecasts is more complicated, as one wishes to avoid double counting forecast errors due, for example, to uncertainty in the initial condition when that uncertainty is explicitly accounted for by the ensemble (which has been generated with multiple initial conditions). The economic value of dressed forecasts has been demonstrated by previous studies. This paper presents a method for dressing ensembles of any size, thus enabling valid comparisons to be made between them. The method involves identifying the "best member" of an ensemble in a multi-dimensional forecast space. The statistics of the errors of these best members are used to dress individual forecasts in an ensemble. The method is demonstrated using ECMWF ensemble forecasts, which are compared with the ECMWF high-resolution best guess forecasts. It is shown that the dressed ECMWF ensembles have skill relative to the dressed ECMWF best guess, even at the maximum lead time of the ECMWF forecasts (10 days). The approach should be applicable to general ensemble forecasts (initial condition, multi-model, stochastic model etc.), allowing better informed decisions on forecast aquisition and forecast system development.

1. Introduction

Estimating the uncertainty associated with a forecast is crucial if the forecast is to be used in a decision making situation. Traditional "deterministic" dynamical forecasts are generated by evolving the best estimate of the initial state of the system forward in time, under the dynamics of a deterministic forecast model. In this paper, any single forecast will be referred to as a *best guess* forecast[1]. The uncertainty of such a forecast can be estimated from the statistical properties of the forecast errors, obtained by analysing historical forecast–verification pairs. A forecast–verification pair is an archived forecast and its subsequent verification. Whether the appropriate verification is a direct observation or an analysis is an important issue, but in this paper direct observations will be used. Common implementations of this method usually assume that the error statistics are stationary, and do not exploit any state dependence of the predictability of the system. In principle, statistical inverse models of forecast error, as a function of atmospheric state, can be constructed using archived forecast–verification pairs; however, in practice the complexity of such models is severely constrained by data availability.

Dynamical ensembles are constructed using a forward modelling approach to quantify the state dependence of predictability[2]. The members of a dynamical ensemble have different initial conditions, often constructed by perturbing the analysis. This method

*Corresponding author. Address: Pembroke College, Oxford, OX1 1DW, UK.
e-mail: roulston@maths.ox.ac.uk

[1]The last thing the authors wish to do is to needlessly introduce jargon into the field, but since the essence of this paper lies in the distinction between the different forecasts, they have introduced some terminology that clarifies these differences. To aid the reader the terminology is summarised in Table 1.

[2]In meteorology, the word "ensemble" alone is used often when discussing dynamical ensembles under a single model.

Table 1. *Glossary of different ensemble forecast types*

Best guess forecast	The forecast obtained by evolving the analysis.
Dynamical ensemble	An ensemble obtained by evolving an ensemble of diffferent initial conditions. This is often referred to as just an "ensemble" in meteorology.
Daughter ensemble	An ensemble generated around an individual member of a dynamical ensemble using forecast error statistics.
Hybrid ensemble	The ensemble obtained by combining all the daughter ensembles of all the members of a dynamical ensemble.
Poor man's ensemble	An ensemble that consists of forecasts generated by different models.

is used operationally by the ECMWF in Europe and NCEP in the U.S. (Molteni et al., 1996; Toth and Kalnay, 1997 and references therein). As well as incorporating uncertainty in the initial condition, dynamical ensembles can also contain stochastic parameterisations or use multiple models to assess the impact of model inadequacy (Houtekamer et al., 1996; Buizza et al., 1999; Stensrud et al., 1999; Evans et al., 2000; Palmer, 2001). Multi-model ensembles that combine the operational forecasts produced by different forecasting centres are commonly referred to as "poor man's ensembles" (Mylne et al., 2002). The computational demand of integrating numerical weather prediction models places a practical constraint on the size of dynamical ensembles. The ECMWF generates the largest operational ensembles, with 51 members. The size of ensembles generated using error statistics is not so limited: ensembles with over 100 000 members are operationally feasible.

The different sizes of ensembles complicates efforts to compare approaches. Furthermore, while a dynamical ensemble can attempt to quantify initial condition uncertainty, there will, inevitably, be residual errors in the forecast due to the finite size of the ensemble. Before a meaningful comparison of different forecasts can be made, the impact of these residual errors should be accounted for. This can be done by "dressing" each member of the dynamical ensemble with its own statistical error ensemble. The ensemble constructed by dressing an individual member of a dynamical ensemble will be called a *daughter ensemble*. Constructing a daughter ensemble requires an estimate of the error statistics associated with *individual* members. The combined daughter ensembles of each member of a dynamical ensemble will be referred to as a *hybrid ensemble*. In this framework, the best guess forecast can be considered to be a single member ensemble. The upper limit on the size of a hybrid ensemble set by the size of the forecast–verification pair archive. The ability to correct for identifiable shortcomings of operational forecasts, and hence their economic value, will be diminished if this archive is too small.

A method for estimating the individual ensemble member error statistics is presented in this paper. The method relies upon identifying the "best member" of the ensemble. Once the best member error statistics have been estimated they can be used to generate daughter ensembles around individidual members of a dynamical ensemble. It will be shown that the resulting hybrid ensembles give more skillful probabilistic forecasts than ensembles generated statistically around the best guess forecast. Essentially, hybrid ensembles allow a user to have "the best of both worlds": ensembles as large as any statistically generated ensemble and which reflect all sources of error, while at the same time containing the information on state-dependent predictability that is contained in a dynamical ensemble.

2. The "best member" concept

Given an ensemble forecast and a verification, it is unlikely that any member of the ensemble will match the verification exactly. This fact can be accounted for by adding an uncertainty distribution to *each member* of the ensemble. To do this, one must know the appropriate uncertainty associated with an *individual* ensemble member. It is proposed that the appropriate uncertainty is the uncertainty associated with the *best member* of the ensemble, where the best member is defined as the member that is closest to the verification in the full state space of the model. Figure 1 illustrates how the best member might be misidentified due to projection effects. In Fig. 1 ensemble member H is closest to the verification (X) in the full model space of variable 1 and variable 2. If, however, only variable 1 is

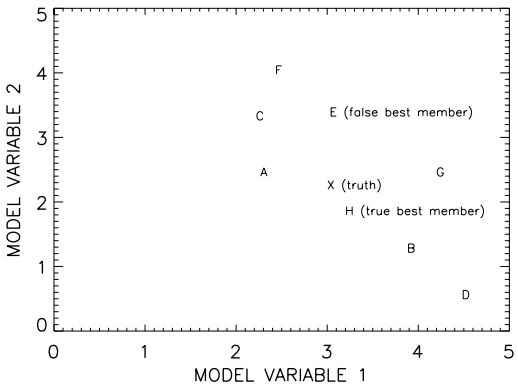

Fig. 1. An illustration of the concept of best members and false best members of an ensemble. Suppose that the model state space is two-dimensional. The symbol X represents the verification. The symbols A to H represent an eight member ensemble. If both the verification and the ensemble members were projected onto model variable 1 then ensemble member E would be identified as the best member (closest to the verification). However, in the full, two-dimensional, model state space it can be seen that H is the true best member. E was falsely identified as a best member due to projection effects.

used to identify the best member, E would be identified as such (closest to the verification in variable 1). This type of misidentification, due to projection, would lead to an underestimate of the error associated with each member of the ensemble. The key point illustrated by Fig. 1 is that, even if one is only interested in forecasting variable 1, both variables 1 and 2 are relevant to identify the best member, and thus estimate the error distribution of variable 1 appropriate for individual ensemble members. The impact of higher-dimensional dynamics on the observed variables can be modelled statistically. (The Appendix gives a formal justification for the best member approach.) The error associated with individual ensemble members depends upon the size of the ensemble. For a one-member ensemble the appropriate error statistics are just the error statistics of that forecast; however, as the number of ensemble members increases, the magnitude of the best member error will decrease because the expected distance between the best member and the verification will decrease.

In practice, it is not necessary to identify the best member of an ensemble in the full model space. It is only necessary that the subspace in which the identification is made is high enough so that projection effects are unlikely to lead to misidentification. The minimum number of variables that must be used to satisfy this condition must be determined empirically. To do this the idea of a *false best member* is introduced. This is reminiscent of the idea of a *false nearest neighbour* in nonlinear dynamics (Kennel et al., 1992; Abarbanel, 1995). Let the N ensemble members be described by d-dimensional vectors, $\mathbf{x}_i$ $(i = 1, \ldots, N)$. Let the verification be described by the d-dimensional vector $\mathbf{y}$. d is the number of forecast variables being considered. Let the normalised distance between the ith ensemble member and verification, in the space of d variables, be written as $R_{i,d}$, where

$$R_{i,d}^2 = \sum_{k=1}^{d} \frac{(x_{i,k} - y_k)^2}{\Omega_k^2}. \tag{1}$$

Here Ω_k is the standard deviation of the kth components of the forecast vectors.

The best member in this d-dimensional space is the one which has the minimum $R_{i,d}^2$. If this member is the true best member, then it should remain the best member when additional variables are included. Or, in symbols, if $\min R_{i,d}^2 = R_{j,d}^2$ then $\min R_{i,d+1}^2 = R_{j,d+1}^2$. If this condition does not hold then the best member can be classed as a *false* best member (FBM). The fraction of FBMs, averaged over past forecasts, gives an indication of an operational lower bound on the minimum number of variables required. The additional variables that are included can either be the same quantity at different spatial locations or they can be at the same location, but for different forecast lead times. Different forecast quantities might also be used. Once a choice of forecast variables has been made that gives a low fraction of FBMs, the best member of each ensemble in the forecast–verification archive can be identified. After identifying the best members, the errors of these particular ensemble members can be calculated. There are several approaches to generating errors with which to dress the current forecast. The simplest approach is to sample from the archive of best member errors. This will ensure that the dressing errors have the same distribution as the historical best member errors, and also that correlations between the errors on different variables are reproduced. This is the approach that will be used in all the examples in this paper, although it can be refined say, by only sampling from the same season as the current forecast. More complicated stochastic error models can also be constructed, with the best member errors used to fit their parameters. The limited number of forecast–verification pairs in the archive places constraints on the complexity of such models.

3. Example: Lorenz-95

The Lorenz-95 system was used to illustrate the procedure for combining statistically generated ensembles with dynamical ensembles outlined above. This spatially distributed system was introduced by Lorenz (1995) as a relatively simple analogue of the atmosphere. It contains external forcing, dissipation and convection. The system is described by

$$\frac{\mathrm{d}\tilde{X}_i}{\mathrm{d}t} = \tilde{X}_{i-1}(\tilde{X}_{i+1} - \tilde{X}_{i-2}) - \tilde{X}_i + F - \frac{c}{b}\sum_{j=1}^{n}\tilde{Y}_{i,j} \tag{2}$$

$$\frac{\mathrm{d}\tilde{Y}_{i,j}}{\mathrm{d}t} = cb\tilde{Y}_{i,j+1}(\tilde{Y}_{i,j-1} - \tilde{Y}_{i,j+2}) - c\tilde{Y}_{i,j} + \frac{c}{b}\tilde{X}_i \tag{3}$$

where the $\tilde{X}_i$s and $\tilde{Y}_{i,j}$s have cyclic boundary conditions and are loosely analogous to meridional averages. The values of the parameters used were $i = 1, \ldots, 8$, $j = 1, \ldots, 4$, $F = 8$ and $b = c = 10$. With these parameters the $\tilde{Y}_{i,j}$s are "fast" variables. They have a timescale 10 times faster than the $\tilde{X}_i$ variables, but their amplitudes are 10 times smaller (Hansen, 1998). Equations (2) and (3) were used to provide "truth". The forecast model used was an imperfect representation of the Lorenz-95 system. The imperfection was introduced by replacing eq. (3) with a parameterisation of the impact of the $\tilde{Y}$ variables on the tendency of the $\tilde{X}$ variables. Figure 2 shows the last term on the RHS of eq. (2) plotted against the value of $\tilde{X}$ for a sample of points on the attractor of the Lorenz-95 system. This term (containing the $\tilde{Y}$s) in eq. (2) was replaced with a linear function of X to give the following forecast model,

$$\frac{\mathrm{d}X_i}{\mathrm{d}t} = X_{i-1}(X_{i+1} - X_{i-2}) - X_i + F + \alpha X_i + \beta \tag{4}$$

where the closure scheme coefficients, α and β, were estimated from Fig. 2 to be $\alpha = -0.1543$ and $\beta = -0.1039$. Figure 3 shows four forecasts of the Lorenz-95 system, made using eq. (4). In the forecasts in Fig. 3 the initial condition is the true initial condition of the full system projected into the model space ($\tilde{X}_i = X_i$ for all i). The state-space of the system described by eqs. (2) and (3) is 40-dimensional (eight $\tilde{X}$ variables and 32 $\tilde{Y}$ variables). The state-space of the model described by eq. (4) is only eight-dimensional (eight X variables).

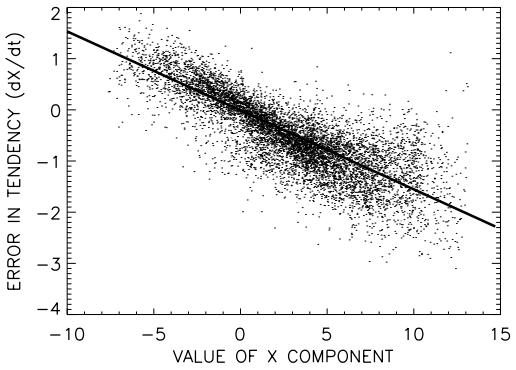

Fig. 2. The error in the tendency of the $\tilde{X}$ variables, given by eq. (2), that is introduced when the the impact of the $\tilde{Y}$ variables is ignored, as a function of the value of $\tilde{X}$. The line is the fit that was used to find the values of α and β in the closure scheme used in eq. (4).

To simulate the effect of observational uncertainty, normally distributed errors, with standard deviations of 0.2 ($\approx$1% of the range of X_i), were added on to the initial values of the X_i. Ensembles were constructed by adding perturbations, with the same distributions as the errors, onto the erroneous initial conditions. Ensembles with 2, 4, 8, 16 and 32 members were generated. The ensembles were *imperfect* ensembles since the initial states were not constrained to lie on the system attractor, hence the forecast PDF is not expected to be accountable even if the model is perfect (Smith, 1997; Smith et al., 1999). The equations were integrated using a fourth-order Runge–Kutta scheme. The trajectories were sampled at intervals of 0.1 dimensionless time units.

The fractions of false best members, as a function of trajectory length for a single X variable, were estimated. The result is shown in Fig. 4. The fraction of FBM falls rapidly as the lead time is increased, until a lead time of about 1.2 time units. Beyond this lead time the fraction of FBM remains approximately constant, at around 7%. Figure 4 indicates that best members should be identified using forecasts out to at least 1.2 time units. The best members of each ensemble in the historical forecasts were identified using a trajectory of 4.0 time units. The best member error statistics were then calculated. A further 1000 forecast–verification pairs were used for forecast validation. Hybrid ensembles of 512 members were generated. In the case of the best guess forecasts, the single forecast was dressed with a statistically generated ensemble of 512 members. The dynamical ensembles with 2, 4, 8, 16 and

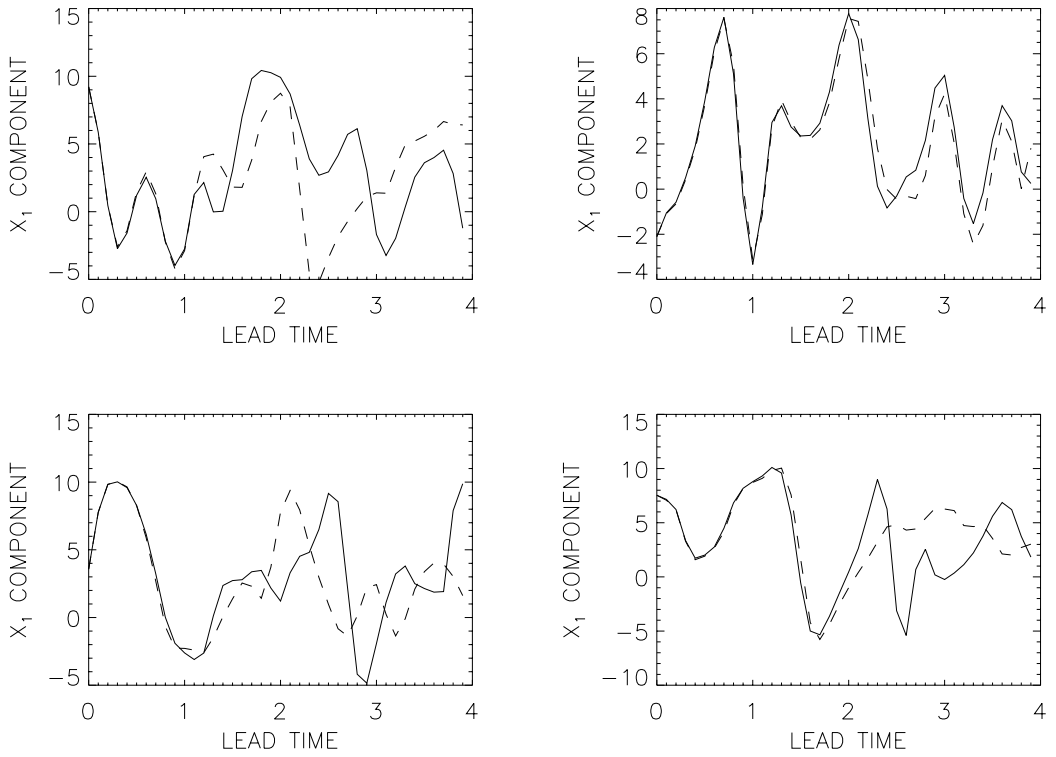

Fig. 3. Forecasts of the Lorenz-95 system, given by eqs. (2) and (3), made using eq. (4). For these forecasts there was no error in the initial condition. The deviation of the forecasts (dashed line) from the verification (solid line) is the result of replacing the impact of the $\tilde{Y}$ variables with a closure approximation in eq. (4).

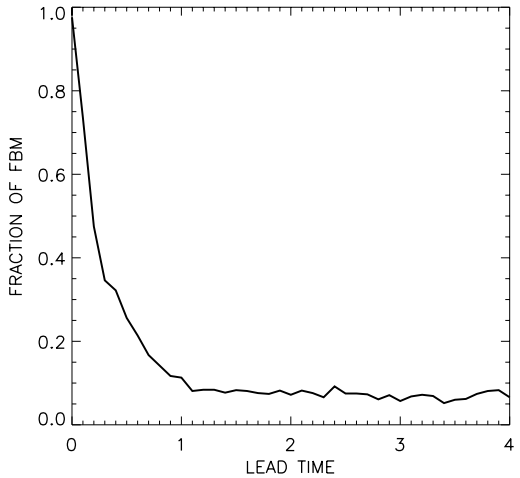

Fig. 4. The fraction of false best members identified in the Lorenz-95 ensemble forecasts, as a function of the length of the forecast trajectory used. A set of 1000 forecasts was used.

32 members were dressed with daughter ensembles of 256, 128, 64, 32 and 16 members, respectively. Therefore, all the hybrid ensembles had 512 members. Making a *direct* comparison of the ensembles of different sizes would introduce a bias against the smaller ensembles.

The best member error statistics are shown in Fig. 5. The thick line is the standard deviation of the errors for the best guess forecasts. The standard deviation of the errors of the best members of the dynamical ensembles is substantially smaller than the error of the best guess forecasts at each lead time. The total error is essentially partitioned into two components; the spread of the dynamical ensemble and the residual error.

Figure 6 compares the dressed best guess and the hybrid, dynamical-statistical ensembles for two particular forecasts. The left panels show the dressed best guess and the right panels show the hybrid ensembles. In the top panels, the spread of the dynamical ensemble indicates that the system is in a particularly

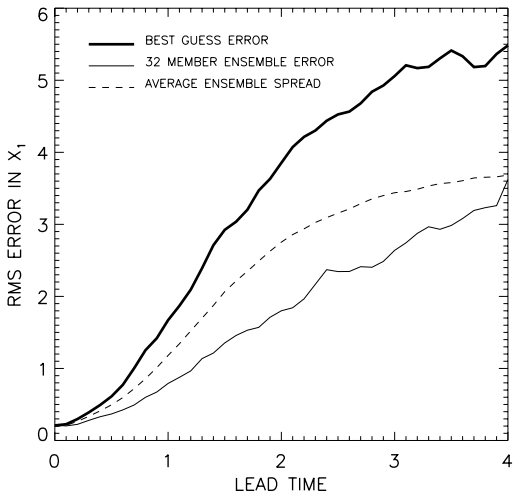

Fig. 5. A comparison of the dynamical ensemble spreads, the RMS error of the best guess forecast and the standard deviation of the best member error, for the 32 member ensemble forecasts of the Lorenz-95 system.

unpredictable regime. In the bottom panels, the dynamical ensemble spread is low, indicating higher predictability. The dressed best guesses, however, have static error statistics. The result is that the dressed best guess ensembles are substantially wider than is needed to capture truth. The hybrid ensemble is constructed by dressing each of the 32 dynamical ensemble members with its own, tighter, daughter ensemble. The result is a tighter hybrid ensemble. Dressing the best guess using static error statistics leads to ensembles which are too narrow in low predictability regimes, and ensembles which are unneccesarily wide in high predictability regimes. Both types of error reduce the value of the forecast.

The forecasts were evaluated using two different scoring rules for probabilistic forecasts. The value of X_1 was discretised into M bins of width 2.0 units for calculating both scoring rules. Let $p_{i,t}$ be the forecast probability of the verification falling in the ith bin for forecast t, where $t = 1, \ldots, T$. Let $j(t)$ be the index of the bin in which the verification was observed to fall, for forecast t.

(i) Ignorance. Ignorance is a logarithmic scoring rule which is equal to the information deficit (in bits) of someone in possession of the forecast (Roulston and Smith, 2002). The average ignorance is given by

$$IGN = -\frac{1}{T}\sum_{t=1}^{T} \log_2 p_{j(t),t}. \tag{5}$$

(ii) Ranked probability score (RPS). The ranked probability score is a quadratic scoring rule. It is the name given to the Brier score when applied to a situation of more than two possible outcomes (Brier, 1950). It is given by

$$RPS = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{M}\left[p_{i,t} - \delta_{ij(t)}\right]^2 \tag{6}$$

where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$.

The ignorance and RPS are both proper scoring rules that combine together the reliability, and the resolution, of a probabilistic forecast. Both scores are defined such that a lower score is better. The ignorance score can be nicely interpreted in terms of gambling returns. If fair odds are set on each outcome by a casino based on their probabilistic forecast, and a gambler bets according to her probabilistic forecast, so as to maximise her expected return, then her expected return per bet is given by

$$100 \times (2^{IGN_{casino} - IGN_{gambler}} - 1)\% \tag{7}$$

where IGN_{casino} and $IGN_{gambler}$ are the ignorances of the casino and gambler respectively. A reduction in ignorance of 1 bit is equivalent to halving the number of possible outcomes and doubling your expected rate of return (Kelly, 1956; Cover and Thomas, 1991).

To estimate the significance of improvements in either skill score, a bootstrap resampling technique was used (Efron and Tibshirani, 1986). The time series of scores for each forecast in the validation set was divided into blocks. The length of the blocks was chosen so that the block-averaged skill score was not strongly correlated between successive blocks. Fifty new time series of block-averaged skill scores were generated by resampling the blocks, with replacement. These 50 time series were used to estimate the average skill score, and the uncertainty on this average.

Figure 7 shows the average skill scores for the 512-member dressed best guess and for 512-member hybrid ensembles constructed by dressing each of 32 dynamical forecasts with its own 16-member daughter ensemble. The thickness of the curves corresponds to 1 standard deviation about the estimate of the mean obtained using bootstrap resampling. Both skill scores indicate a moderate, but statistically significant, improvement in forecast skill when the hybrid ensemble is used instead of the dressed best guess. The difference in ignorance at a lead time of 2 units is 0.6 bits. If this difference is interpreted in terms of gambling

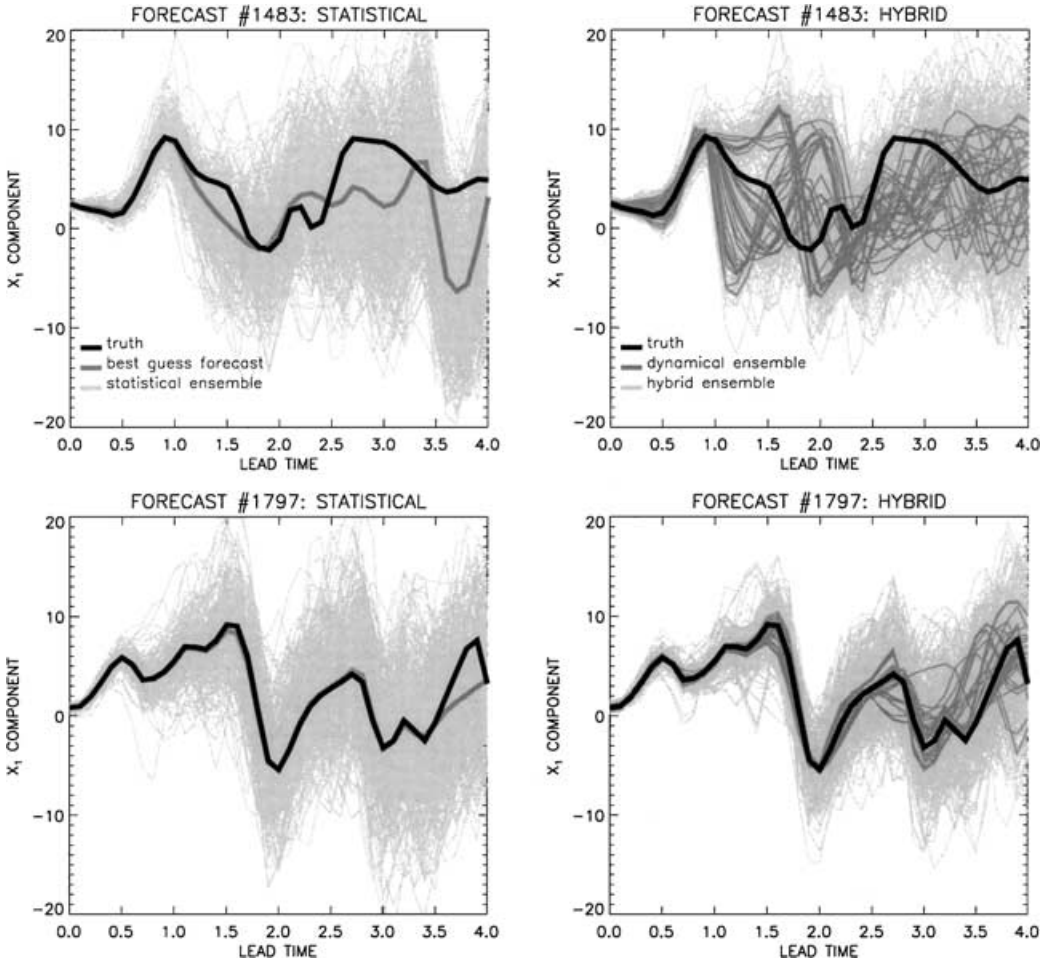

Fig. 6. Statistical ensemble (left panels) and hybrid ensemble forecasts (right panels) of the Lorenz-95 system in a low predictability state (top panels) and a high predictability state (bottom panels). The left panels are ensembles constructed by dressing a best guess forecast with 512 statistical trajectories constructed using historical error statistics. The right panels are hybrid, dynamical-statistical forecasts constructed by dressing each 32 member dynamical ensemble member with its own 16 member statistically generated ensemble constructed using the best member error statistics. In the high predictability state the dressed best guess forecast is unnecessarily wide, whereas the hybrid ensemble is tighter because its width varies with the width of the dynamical ensemble around which it is constructed.

returns it implies that a gambler who places optimal bets according to the dressed ensemble can expect an average return of 52% per bet if the fair odds are set based on the dressed best guess forecast. Figure 8 shows both skill scores as a function of dynamical ensemble size when the size of the hybrid ensemble is held constant at 512 members.

To illustrate the fair comparison of dynamical ensembles of different sizes, Fig. 8 contrasts dynamical ensembles of 1, 2, 4, 8, 16 and 32 members. In each case the dynamical ensembles were dressed with hybrid ensembles of 512 members. The forecast skill was evaluated as a function of dynamical ensemble size. Increasing the size of the dynamical ensemble leads to an improvement in both skill scores. Over the range of ensemble sizes evaluated, each doubling in the size of the dynamical ensemble leads to a reduction in ignorance of approximately 0.1 bit (equivalent to a 7% increase in gambling returns). In this case the model was held constant, but the same approach could

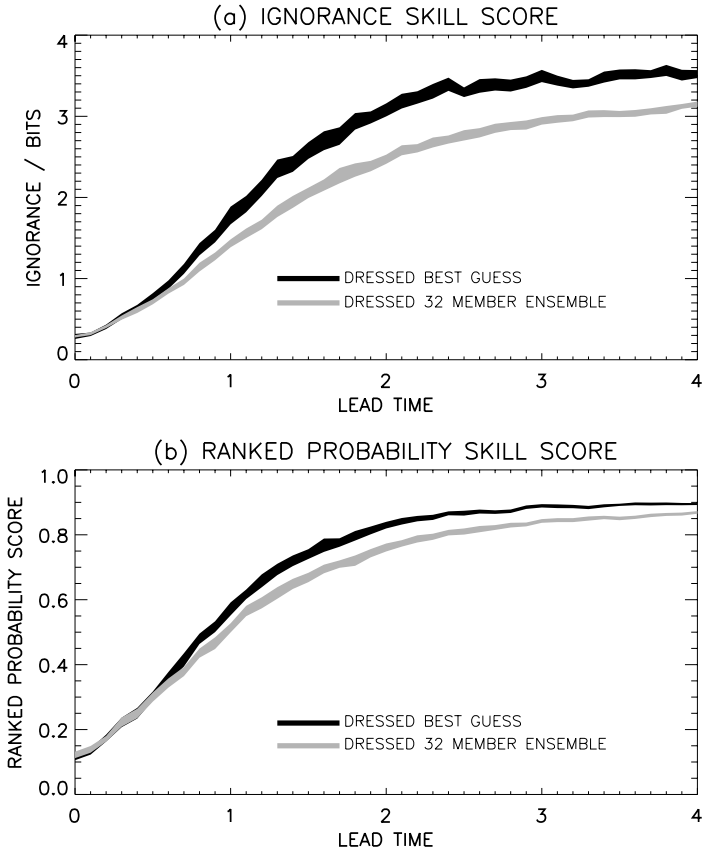

Fig. 7. Average skill scores of ensemble forecasts of the Lorenz-95 system. The curves are for the case of best guess, deterministic forecasts and 32 member dynamical ensembles. The best guess forecast was dressed with a statistically generated ensemble of 512 members, while each member of the dynamical forecast was dressed with its own 16 member statistically generated ensemble, thus giving a hybrid ensemble of 512 members. (a) Ignorance (logarithmic skill score), (b) ranked probability score (quadratic). The thickness of the lines indicates the 1σ uncertainty in the average skill score estimated using bootstrap resampling (Efron and Tibshirani, 1986).

be used if, say, computational cost was held constant. Therefore the advantages of increasing ensemble size over increasing model resolution can be determined without sampling errors due to finite ensemble size confusing the issue.

4. Example: ECMWF ensembles

To illustrate the best member method with real ensemble weather forecasts, ECMWF temperature forecasts were used. The ECMWF issues a 10 d forecast initialised with the best estimate of the state of the atmosphere. In addition, they issue a 51 member ensemble. The ensemble members are integrated with a lower-resolution version of the ECMWF global model. The leading ensemble member is initialised with the best estimate of the initial condition, while the other 50 are initialised with initial conditions constructed by perturbing this initial condition in the space of the leading 25 singular vectors of the linearised model (Palmer, 2000).

For this study, four European stations were used: Tromsö, Heathrow, Frankfurt and Warsaw. The archive for determining error statistics consisted of forecasts and observations from February to December 1999, while the verification was done with forecasts and observations from January to November 2000.

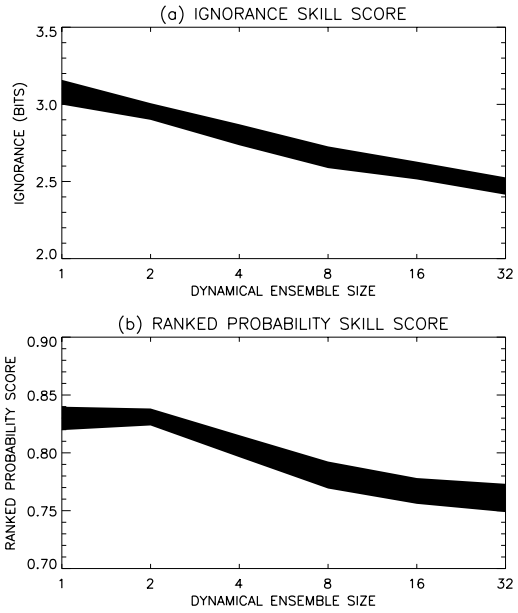

Fig. 8. Average skill scores of ensemble forecasts of the Lorenz-95 system as a function of the size of the dynamical ensemble for a fixed lead time of 2.0 time units. In all cases the number of members in the hybrid ensemble was 512. (a) Ignorance (logarithmic skill score), (b) ranked probability score (quadratic). The thickness of the lines indicates the 1σ uncertainty in the average skill score estimated using bootstrap resampling (Efron and Tibshirani, 1986).

The fraction of FBM for the forecasts, as a function of the length of the forecast trajectory, was determined. Figure 9 shows the estimated fraction of FBM for the case of London's Heathrow airport. The fraction of FBM decreases rapidly as the forecast trajectory is extended from 1 to 4 d and then appears to stabilise, at just over 20%, when trajectories longer than about 120 h are used. The full 10 d forecast trajectories were used to determine the best members of the ensembles in the archive. Figure 10 compares the growth of the error of the best guess forecast with that of the error of the best ensemble member for the Heathrow forecasts. The spread of the dynamical ensemble is also shown. At short lead times, the best member error is greater than the ensemble spread, indicating that model inadequacy and residual initial condition error dominate. Note, however, the best member error grows more slowly than the ensemble spread. At a lead time of 8 d the contributions of ensemble spread and best member error to overall error are comparable.

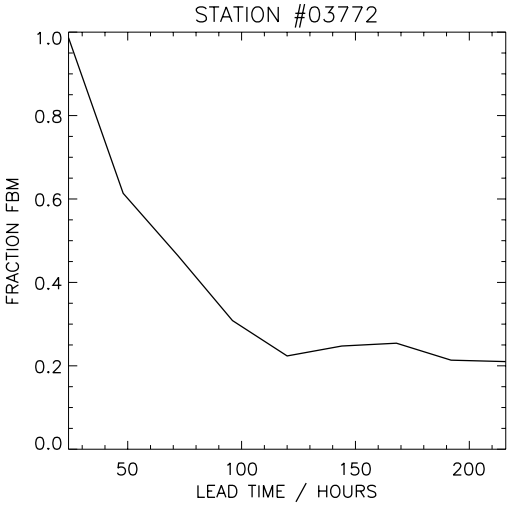

Fig. 9. The fraction of false best members identified in the Heathrow temperature forecasts, as a function of the length of the forecast trajectory used to identify the best member of the ensemble.

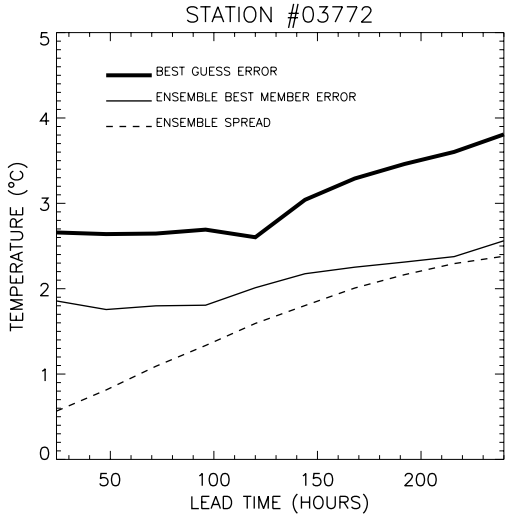

Fig. 10. A comparison of the average standard deviations of the dynamical ensembles, the error associated with the best member of the 51 member ensemble and the best member errors for the ECMWF forecasts of temperature at London's Heathrow airport.

Figure 11 compares the dressed best guess forecast with the dressed dynamical ensemble forecast for two different days. The dressed best guess ensembles (left panels) are both constructed using the same error

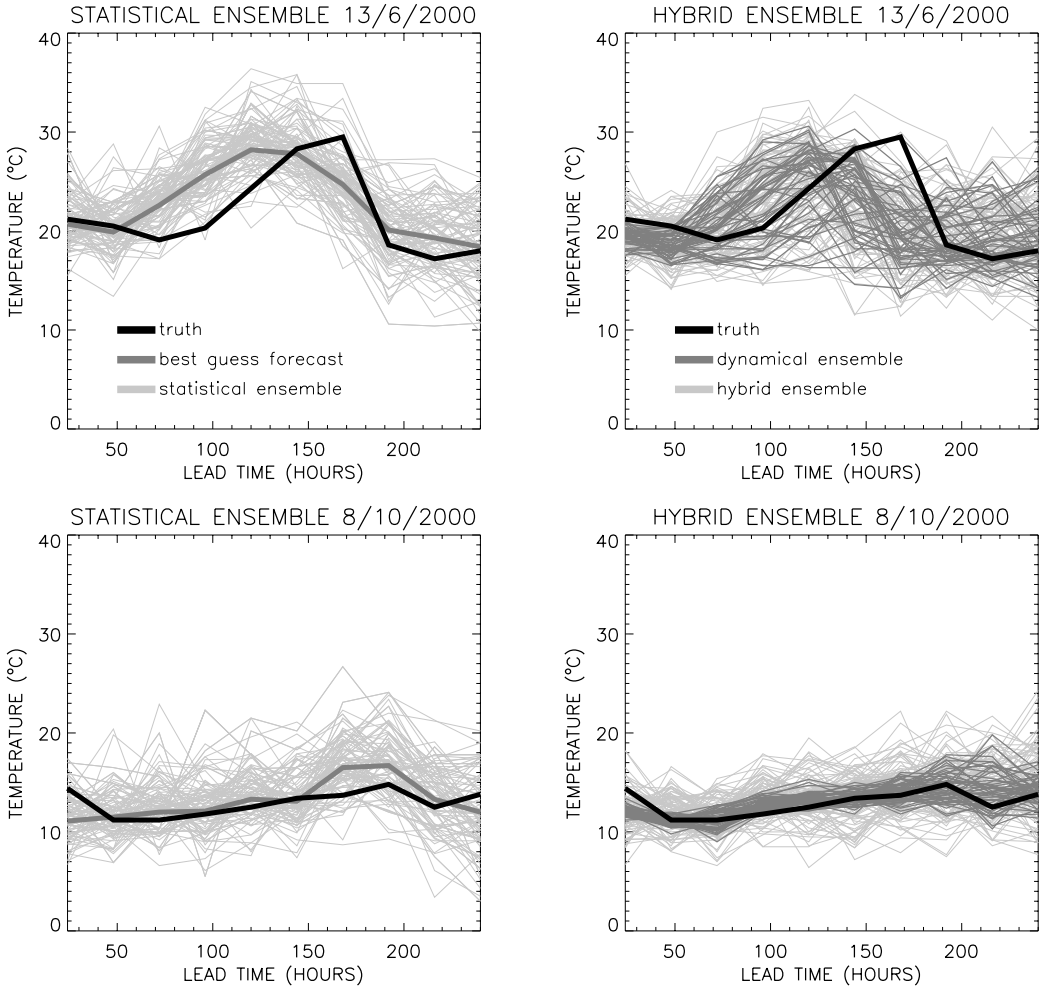

Fig. 11. A comparison of 10 d forecasts for temperature at London's Heathrow airport for two different days. The dressed best guess forecasts (left panels) were constructed by dressing the ECMWF best guess forecasts with 102 statistical error trajectories. The hybrid ensembles (right panels) were made by dressing each of the 51 members of the ECMWF dynamical forecasts with a two-member daughter ensemble.

statistics, and thus have the same spread for a given lead time. The hybrid ensembles (right panels) are constructed around the ECMWF dynamical ensembles, and their width can vary with the width of the dynamical ensembles. The top panels show a forecast when the dynamical had a large spread, but at a lead time of 170 h the verification fell outside the range of the dynamical ensemble. It did, however, fall within the hybrid ensemble. The bottom panels show a day when the ECMWF dynamical ensemble was tight. Under such conditions the dressed best guess ensemble was unnecessarily wide.

To evaluate the forecasts in the test period (January–November 2000) the ignorance and ranked probability skill scores were used. Temperature was quantised into 2 °C bins to calculate both the skill scores. The uncertainty in the estimates of the mean skill scores were again estimated using bootstrap resampling. The average skill score for each successive 10 d period was calculated. These averages were then resampled, with replacement, to obtain an estimate of the mean and also the uncertainty on this estimate. Figure 12 shows the results for the four stations. The hybrid ensembles, constructed by dressing the dynamical ensemble, yield

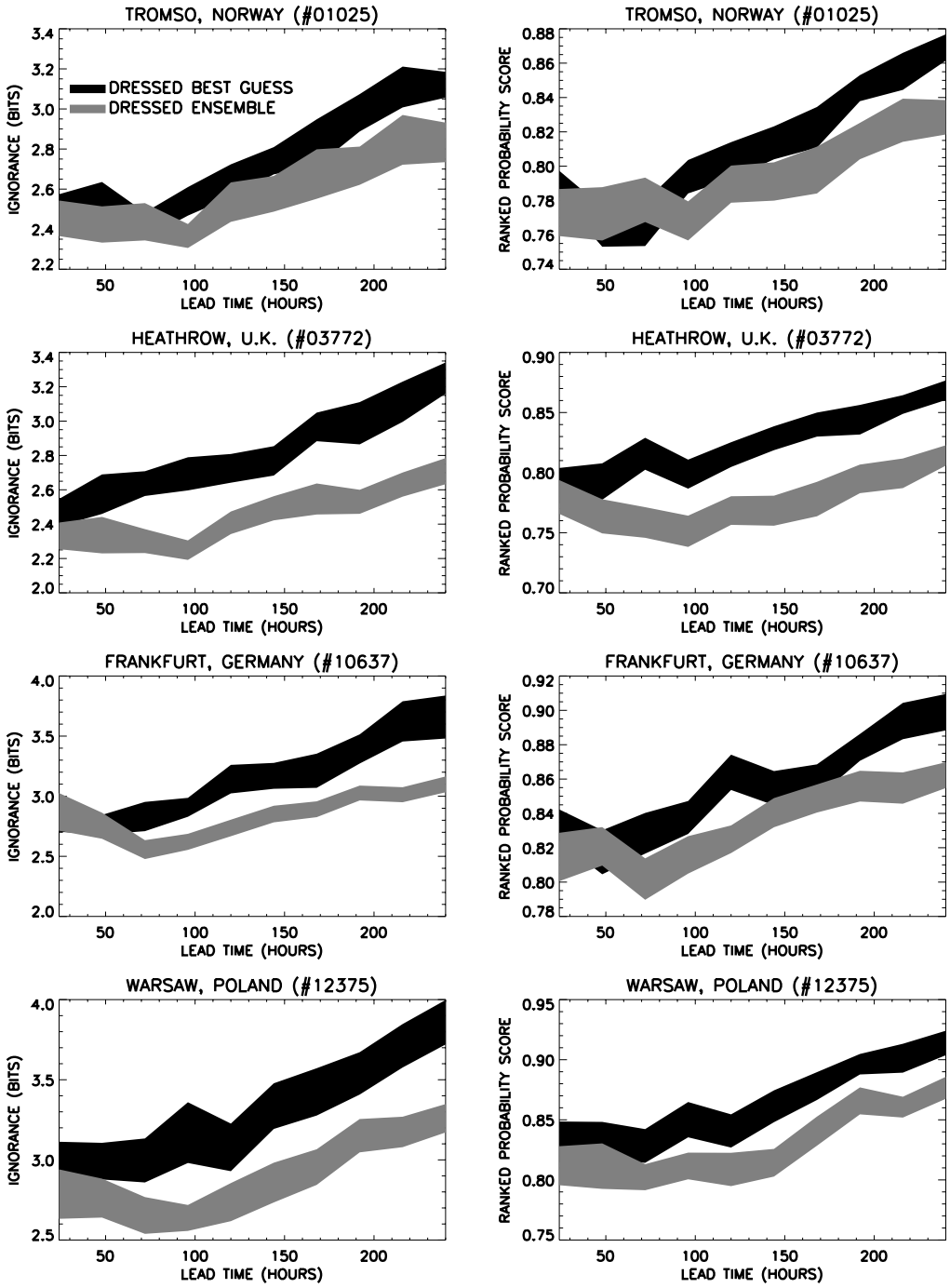

Fig. 12. Average skill scores for the ECMWF forecasts at four European stations. The left panels are ignorance and the right panels are ranked probability score. The temperature was quantised in 2 °C bins to calculate both skill scores. The thickness of the lines corresponds to the 1σ uncertainty in the estimate of the mean obtained using bootstrap resampling (Efron and Tibshirani, 1986).

statistically significant improvements in skill over the dressed best guess forecast at all four stations for lead times beyond 4 d. This result is true for both skill scores. The size of the improvement varies between stations, being relatively small for Tromsö but quite substantial for Heathrow. For short lead times there is much less improvement, which is to be expected since the ECMWF ensembles tend to be quite tight for the first 48 h of the forecast. This is also the reason why the skill of the dressed ensemble forecast sometimes improves with increasing lead time, for short lead times. As with the Lorenz-95 results, the improvement in the ignorance score can be interpreted in terms of gambling returns. For Tromsö the increase in gambling returns would be about 15% per bet at a lead time of 150 h, while at Heathrow it would be over 40%.

Finally, the importance of correctly identifying the best member will be demonstrated. Consider the situation where a user is only interested in the temperature forecast at a lead time of 4 d. If the best members of the ensembles are identified purely on how close they are to the verification at a lead time of 4 d the variance of the best member error, at this lead time, is much smaller than if the choice is based on the entire 10 d forecast trajectory. The resulting skill scores are shown in Fig. 13. The skill of the dressed ECMWF ensemble is diminished at all lead times, but the reduction is particularly large at the 4 d lead time used to identify the best members, in most cases leading to a less skillful forecast than the dressed best guess.

5. Discussion and conclusions

The concept of best member errors is a simple idea. Nevertheless, there are pitfalls to be wary of when attempting to identify the best member of an ensemble. Identification of the best member should be done using multivariate forecasts, even if only univariate forecast statistics are required. The number of forecast variables required can be estimated by looking at the fraction of *false best members*. Identifying the best members in too low a dimensional space can lead to underestimates of the error associated with the forecast variables included in the identification, which leads to a reduction in forecast skill.

Once the best member error statistics have been estimated, hybrid ensembles of an arbitrary size can be generated. This allows fair comparisons to be made between dynamical ensembles of different sizes, including single best guess forecasts. The choice of error model for dressing the dynamical forecasts is a separate issue. Attempting to condition error statistics on season and atmospheric state is a possibility. Due to the high-dimensionality of atmospheric models and their long recurrence times, any type of inverse modelling of state-dependent predictability faces severe limitations. The relatively small amounts of data that are often available strongly suggest that empirical error models should be parsimonious.

Using the method of false best members, it has been demonstrated that the 51 member ensemble forecasts, generated by the ECMWF, of temperature at four European locations contain information about state-dependent predictability that is *not* contained in single forecasts. It has also been shown, however, that model and residual initial condition error make a significant contribution to the total forecast error. This is especially true at shorter lead times. For this reason, the error statistics associated with individual ensemble members must be estimated to provide a more complete assessment of forecast uncertainty. The use of the best member method to obtain probabilistic forecasts at *specific locations* might be suboptimal. Forecasting models predict quantities which are defined at the resolution of the model. The forecast uncertainty of a physical observable at a single location will typically be higher than that of the model variable. Because of this, further forecast skill improvements may be possible by rescaling higher moments of the dynamical ensemble distribution when downscaling forecasts.

The best member method is applicable to all types of ensembles, not just those currently operational. If the ensemble members can be distinguished a priori (e.g. they are run under different models or initialised at different times), then each member will have different error statistics. In the ECMWF operational ensemble, for example, the leading ensemble member is distinguishable from the other 50 as it is the centre about which the others are generated[3]; hence it could be dressed differently. The value of the present paper lies in introducing the hybrid ensemble approach; this approach allows a fair comparison between very different forecast strategies, thereby easing their

[3]This is a topic for future discussion. In the results presented in this paper, all 51 members were treated as indistinguishable.

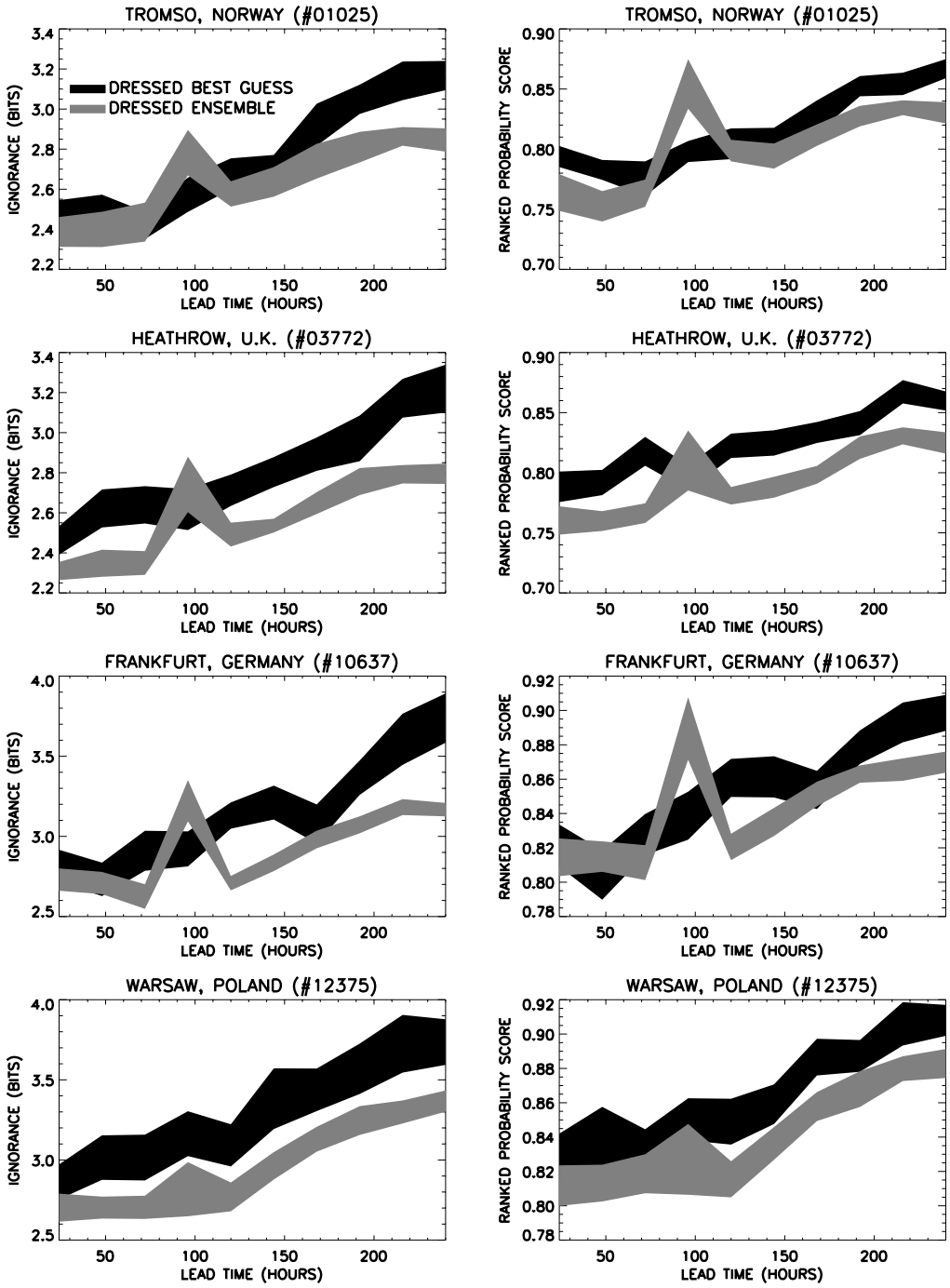

Fig. 13. As Fig. 12 but where the best members of the ensembles were identified using *only* the 4 d forecast. This leads to misidentification of the ensemble best members and subsequent underestimation of the appropriate errors.

improvement and evaluation of their relative economic value.

6. Acknowledgements

This work was supported by ONR DRI grant N00014-99-1-0056.

Appendix: The best member method

Let $\mathbf{y}$ be an m-component vector describing the verification. The verification can be decomposed into a *dynamical* component, $\mathbf{x}$, and a *statistical* component, ε:

$$\mathbf{y} = \mathbf{x} + \varepsilon. \tag{A1}$$

Note that the terms *dynamical* and *statistical* are being used in an operational sense. The dynamical component contains all processes contained in the forecasting model. The statistical component includes all contributions that are to be dealt with statistically, such as model inadequacy and residual initial condition error. The dynamical component need not be strictly deterministic if, for example, the model contains stochastic parameterisations. Conversely, the statistical component may include processes that are, in fact, deterministic, but which will be dealt with in a statistical manner. Let $\mathbf{x}_i$ be an ensemble of N dynamical forecasts ($i = 1, \ldots, N$). The best member of the ensemble is the member that has the "correct" dynamical component. If the verification is known, and the best member is identified, the contribution of ε can be determined. Over many forecasts, the statistical properties of the ε component can be estimated. For the following analysis, it will be assumed that ε has a multivariate normal distribution with uncorrelated components. The probability that the ith ensemble member has the correct dynamical component is

$$p_i = \frac{\exp\left[-\sum_{k=1}^{m}(y_k - x_{i,k})^2 \Big/ 2\sigma_k^2\right]}{\sum_{j=1}^{N}\exp\left[-\sum_{k=1}^{m}(y_k - x_{j,k})^2 \Big/ 2\sigma_k^2\right]} \tag{A2}$$

where y_k is the kth component of $\mathbf{y}$, and $x_{i,k}$ is the kth component of $\mathbf{x}_i$. If $\mathbf{x}_l$ is the correct dynamical component then $E[(y_k - x_{l,k})^2] = \sigma_k^2$. If Ω_k^2 is the variance of the kth component of all the ensemble members then an approximate expression for p_l, the probability assigned to the true best member, is

$$p_l \sim \frac{\exp(-m/2)}{\exp(-m/2) + (N-1)\exp\left[-\sum_{k=1}^{m}\left(\Omega_k^2 + \sigma_k^2\right) \Big/ 2\sigma_k^2\right]} \tag{A3}$$

To simplify eq. (A3) assume that all the σ_k are identical and all the Ω_k are also the same:

$$p_l \sim \frac{\exp(-m/2)}{\exp(-m/2) + (N-1)\exp(-m/2 - m\Omega^2/2\sigma^2)}. \tag{A4}$$

If $N \gg 1$ eq. (A3) becomes

$$p_l \sim \frac{1}{1 + N\exp(-m\Omega^2/2\sigma^2)}. \tag{A5}$$

An examination of eq. (A5) provides some insight into the conditions required for the correct identification of the best member. For a confident identification $p_l \approx 1$. Therefore, it is required that

$$N\exp(-m\Omega^2/2\sigma^2) \ll 1. \tag{A6}$$

From eq. (A6) it can be seen that if $\Omega \gg \sigma$ then the best member can be easily identified. That is, if the spread of the ensemble members is much greater than the uncertainty due to the statistical component, then the ensemble member that comes closest to the verification is highly likely to be the best member. If, however, $\Omega \approx \sigma$ and m is small then this is not the case. This is because if the size of ε is comparable to the spread of the $\mathbf{x}_i$, then the $\mathbf{x}_i$ closest to $\mathbf{y}$ is not necessarily the correct $\mathbf{x}$. In this situation, the chance of choosing the correct best member can be improved by increasing m. That is, the best member is identified as the closest $\mathbf{x}_i$ to $\mathbf{y}$ in a higher dimensional space. In practice, this can be done by comparing the forecast to the verification at multiple points in space and time. This means, that even if, one is only interested in forecasting a univariate quantity, the best member must be chosen on the basis of a multivariate forecast. From eq. (A6) it can also be seen that the probability of correctly identifying the best member falls as the ensemble size, N, increases. This makes sense; the more ensemble members there are, the higher the chance of misidentifying the best member. If N is increased then m should also be increased to ensure that the best member of the enlarged ensemble is correctly identified.

REFERENCES

Abarbanel, H. D. I. 1995. *Analysis of observed chaotic data*. Springer-Verlag, New York.

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

Buizza, R., Miller, M. and Palmer, T. N. 1999. Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.* **125**, 2887–2908.

Cover, T. M. and Thomas, J. A. 1991. *Elements of information theory*. John Wiley, New York, 542 pp.

Efron, B. and Tibshirani, R. 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat. Sci.* **1**, 54–77.

Evans, R. E., Harrison, M. S. J., Graham, R. J. and Mylne, K. R. 2000. Joint medium-range ensembles from the Met Office and ECMWF systems. *Mon. Wea. Rev.* **25**, 3104–3127.

Hansen, J. A. 1998. Adaptive observations in spatially-extended nonlinear dynamical systems, Ph.D. Thesis, Oxford University.

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

Kelly, J. 1956. A new interpretation of information rate. *Bell Sys. Tech. J.* **35**, 916–926.

Kennel, M. B., Brown, R. and Abarbanel, H. D. I. 1992. Determining minimum embedding dimension using a geometrical construction. *Phys. Rev. A* **45**, 3403–3411.

Lorenz, E. N. 1995. Predictability–a problem partly solved. In: *Predictability*. ECMWF, Seminar Proceedings, Shinfield Park, Reading, UK.

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

Mylne, K. R., Evans, R. E. and Clark, R. T. 2002. Multi-model multi-analysis ensembles in quasi-operational medium-range forecasting. *Q. J. R. Meteorol. Soc.* **128**, 361–384.

Palmer, T. N. 2000. Predicting uncertainty in forecasts of weather and climate. *Rep. Progr. Phys.* **63**, 71–116.

Palmer, T. N. 2001. A nonlinear dynamical perspective on model error: A proposal for non-local stochastic-dynamic parameterization in weather and climate prediction models. *Q. J. R. Meteorol. Soc.* **127**, 279–304.

Roulston, M. S. and Smith, L. A. 2002. Evaluating probabilistic forecasts using information theory. *Mon. Wea. Rev.* **130**, 1653–1660.

Smith, L. A. 1997. The maintenance of uncertainty. In: *Proc. Inte. School of Physics "Enrico Fermi"*, 177–246, Course CXXXIII, Società Italiana di Fisica, Bologna, Italy.

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

Stensrud, D. J., Brooks, H. E., Du, J., Tracton, M. S. and Rogers, E. 1999. Using ensembles for short-range forecasting. *Mon. Wea. Rev.* **127**, 433–446.

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