Tellus (1996), 48A, 518–537
Printed in UK – all rights reserved

TELLUS
ISSN 0280–6495

Statistical and dynamical long-range atmospheric forecasts: experimental comparison and hybridization

By JÉRÔME SARDA and GUY PLAUT, *Institut Non-Linéaire de Nice, 1361 Route des Lucioles, 06560 Valbonne, France;* CARLOS PIRES and ROBERT VAUTARD* *Laboratoire de Météorologie Dynamique, Université Pierre & Marie Curie, Boite 99, 4 Place Jussieu, Paris, France*

(Manuscript received 4 August 1995; in final form 4 March 1996)

ABSTRACT

We perform a direct comparison between a statistical forecast model using space-time principal components as predictors and a series of experimental long-range (up to 44 days) dynamical forecasts performed at Météo-France using a simplified-physics version of the former french operational forecast model "Emeraude". The comparison is made possible by forecasting the same upper air quantity, the monthly averaged 50 kPa geopotential height, at the same forecast dates and lead times, using the same skill measure. A disappointing result is that most of the skill differences are nonsignificant, due to the small number of forecast cases used (40). Moreover, the skill comparisons are obscured by inherent biases due to the combination of trends, interdecadal variability and systematic errors. We use a very conservative significance testing procedure taking into account these problems. A careful examination of the skill leads to the conclusion that the statistical model performs better in the long run than the dynamical one. Of particular relevance is the question whether the valuable information contained in the empirical and the dynamical forecasts differ. If such is the case an appropriate combination of both forecasts, and/or forecasting algorithms could lead to an improvement of the skill. This is our second purpose: we propose here two hybrid procedures which combine objectively the dynamical and statistical predictive information contents. The skill of these hybrid models is compared to that of the two original models. One of the two hybrid schemes is shown to be significantly more skillful, at long lead times, than the dynamical model.

1. Introduction

The development of computer resources has, during the last decade, raised the possibility of using large numerical models for monthly-to-seasonal atmospheric predictions. These models are generally derived from the currently operational short-range forecast models, by reducing horizontal resolution and simplifying the various physical parametrizations. This dynamical approach of long-range prediction relies on the hypothesis that, given the computer time and memory constraints, the most skillful model must be the most physically accurate one, in its representations of the complex processes of the earth-ocean-atmosphere system. Numerous experiments have been conducted to assess the validity of dynamical models for long-lead forecasts (Tracton et al., 1989; Brankovic et al., 1990; Milton and Richardson, 1991; Déqué and Royer, 1992; Palmer and Anderson, 1994 (for a thorough review)).

However, despite the fact that dynamical forecasts are based on the most scientific grounds, they are still at an experimental stage and apparently fail to exhibit significantly higher skill than empirical forecasts (Van den Dool, 1994), such as provided in an operational fashion by various meteorological and climatic centers (Folland

* Corresponding author.

and Woodcock, 1986; Livezey and Barnston, 1988; Barnston, 1994). Such empirical forecasts are mostly based on linear techniques such as canonical correlation analysis (Barnett and Preisendorfer, 1987; Barnston, 1994), or on analogue techniques (Bergen and Harnack, 1982; Folland and Woodcock, 1986; Livezey and Barnston, 1988). Admittedly, the reason for the apparent failure of dynamical models in the long range rests upon its inherent numerical constraints: the large computing time required inhibits the possibility of carrying out long validation experiments necessary to measure skill with any statistical reliability. This difficulty is increased by the fact that the skill is anyway expected to be low. Moreover, development, tuning and validation of new physical parametrizations require many of these long validation experiments, and it is plausible that the full validation of a long-range forecast model takes longer than its own obsolescence time.

At present, a major part of the long-range forecast error of dynamical models is the systematic error (Tracton et al., 1989; Brankovic et al., 1990; Déqué, 1991). Conversely, empirical forecast schemes are likely to mimic the climate of the past, but are less able to forecast its drift, which is not a serious limitation for time scales shorter than a season. However, if the climate trends (or its ultra-low frequency fluctuations) are due to modifications in the occurrence frequencies of particular high-amplitude events, models trained from the past climate hold the knowledge of these events, and therefore should be able to adapt themselves to the modified environment. For the above reasons, one is lead to doubt that, for several years or perhaps decades, dynamical models will undeniably beat empirical models, and will be used as such for operational long-range forecasting.

There is, to our knowledge, no quantitative investigation of the difference between statistical and dynamical models' skill for real-atmosphere long range forecasts. This is mostly due to the quoted difficulties in assessing dynamical forecasts skill. The skill of statistical models, by contrast, has been addressed in many articles (Folland et al., 1986; Barnett and Preisendorfer, 1987; Livezey and Barnston, 1988; Barnston, 1994). Another difficulty comes from the nature of the output forecasts. Statistical forecasts are often put into discrete, categorical, probabilistic forecasts, while dynamical forecasts are generally continuous and deterministic. Therefore, an important problem is to use the same skill index for both.

The first goal of the present study is to provide a direct comparison between a statistical forecast model, using *space-time principal components* (ST-PCs) as predictors, as developed by Vautard et al. (1996; hereafter VPP, see also Vautard, 1995), and a general circulation forecast model, namely the former french numerical weather prediction model EMERAUDE. The long-lead dynamical forecast experiments with EMERAUDE have already been described in Déqué and Royer (1992), and Déqué et al. (1994) with skill estimation procedures somewhat different to that we shall use. Here, the direct comparison is made possible by forecasting the same upper-air quantity, the 50 kPa geopotential height, at the same forecast dates and lead times, using the same measure of skill.

An interesting point is to examine whether there are certain circumstances where our simple statistical model formulation can perform better, in the long-range forecasting sense, than a general circulation dynamical model. Above all, we address here the question whether the valuable information contained within the statistical and the dynamical forecasts differ. If such is the case, an appropriate combination of both strategies could lead to an increase of the prediction skill. This is our second goal. We propose here two hybrid forecast schemes where the statistical and the dynamical predictive knowledges are combined. The idea of combining statistical knowledge (obtained from data processing) with dynamical knowledge (obtained from the laws of physics) is not new: A simple hybrid formulation has been proposed, for instance, by Madden (1981), Déqué et al. (1994), who used either a combination of damped persistence and simplified GCMs or the dynamical model's output statistics in order to improve the forecast skill. In these studies, however, the statistical formulation was not as sophisticated as up-to-date statistical models. In a recent study, Jianping et al. (1993) proposed a more elaborate analogue-dynamical forecast combination, where only perturbations from the best analogue are integrated from the dynamical model. The skill was indeed improved relative to the analogue model itself, but the validation procedure

was performed over a very limited number of cases. Meteorological centers producing on an experimental or even on a routine basis long-range forecasts generally use some blending procedure: the strategy is most often to perform, on the one hand, *ensemble average forecasts*, by perturbing the initial condition with perturbations representative of the initial error, and on the other hand, to use sophisticated statistical models. The final forecast provided to the public is generally a blend of the two outcomes, involving a large degree of subjectivity. We propose here objective hybridization methods for this blending procedure.

In Section 2, we describe the data set, the validation procedure, and the forecast experiments performed with the dynamical and statistical models. Section 3 is devoted to a comparison of the skill of the models in extrapolating the space-time principal components. In Section 4, the skills of the models in forecasting grid-point values of the 50 kPa 30-day averages are compared. Section 5 describes the formulation of two hybrid models and their skill is compared with those of the original models. Section 6 contains a summary and a short discussion.

2. Data, models and methodology

2.1. Basic data set

Throughout this article, we use as predictor (for the statistical model) the geopotential height at 50 kPa over the extratropical northern hemisphere (NH: 25°N–86°N), whereas the predictand is restricted to 30-day averages of 50 kPa heights over the midlatitude Atlantic area (ATL: 30°N–70°N, 80°W–40°E). The original datasets are the daily analyses of the 50 kPa geopotential height from NMC and ECMWF. NMC data are used for the period 1 April 1950–30 April 1983, and are initially given on a regular 5×5 degree grid. The ECMWF analyses, covering the whole globe, are used for the period 1 May 1983–31 January 1994. Both NMC and ECMWF data are interpolated onto a T21 gaussian grid.

2.2. Validation procedure and predictands

For the dynamical forecasts, a set of 40 experiments, performed at METEO-FRANCE, is available at dates specified in Table 1. These dates are all contained within the cold seasons 1983–1984 to 1992–1993. Hence, for the construction of the statistical model, we use a single learning period (1 April 1950–31 March 1983); all verified dates belong to the complementary period, the verification period. Regardless of the existence of any trend in the data, annual cycle removal, principal component analysis, tercile separator estimation (see below), are carried out using data belonging to the learning period only. Such a procedure differs from the classical cross-validation (Tukey, 1958), in which data are successively used for learning model coefficients and for forecast verification. However, the hybrid models construction, as well as the removal of systematic errors, requires the knowledge of some dynamical forecasts, for which we shall use a cross-validation technique, by omitting in these calculations the cold season over which the

Tables 1. *Starting dates of the 40 Emeraude ensemble forecasts (Déqué et al., 1994)*

Nov.	Dec.	Jan.	Feb.
16 Oct. 1983	16 Nov. 1983	15 Dec. 1983	16 Jan. 1984
16 Oct. 1984	16 Nov. 1984	15 Dec. 1984	16 Jan. 1985
16 Oct. 1985	16 Nov. 1985	15 Dec. 1985	19 Jan. 1986
19 Oct. 1986	16 Nov. 1986	14 Dec. 1986	18 Jan. 1987
18 Oct. 1987	15 Nov. 1987	13 Dec. 1987	17 Jan. 1988
16 Oct. 1988	20 Nov. 1988	18 Dec. 1988	22 Jan. 1989
16 Oct. 1989	16 Nov. 1989	16 Dec. 1989	16 Jan. 1990
16 Oct. 1990	16 Nov. 1990	16 Dec. 1990	16 Jan. 1991
16 Oct. 1991	16 Nov. 1991	16 Dec. 1991	16 Jan. 1992
16 Oct. 1992	16 Nov. 1992	16 Dec. 1992	16 Jan. 1993

forecast is verified against the analysis, as in Déqué and Royer (1992).

The ATL area includes 176 grid-points. As in VPP and many other long-range predictability studies (Folland and Woodcock, 1986, Livezey and Barnston, 1988), the 176 predictands are classified into three equally-probable terciles (whose separators depend on the grid point). The categories are called A (above normal), N (near normal) and B (below normal). Calendar-time dependent tercile separators are calculated, as in VPP, by sorting, over the learning period, the monthly mean values of months covering at least 15 days of the current calendar month.

Each issued monthly-mean 50 kPa height tercile forecast (from any model) is verified against the actual tercile, and a 3×3 contingency table $n_{ij}(x)$ is built at each grid point x. The measure of skill used here is the categorical LEPS (linear error in probability space) score (Ward and Folland, 1991; Barnston, 1992). In practice, each entry of the contingency table $n_{ij}(x)$ (the number of cases with the i verified tercile and the j forecast tercile) is multiplied by a constant scoring weight W_{ij}, the same as used for terciles by Ward and Folland (1991) and VPP. The final score is obtained, at each grid point x, by summing up all the products,

$$S(x) = \frac{\sum_{i,j=1}^{3} n_{ij}(x) W_{ij}}{\sum_{i=1}^{3} n_i(x) W_{ii}}, \qquad (1)$$

where $n_i(x)$ is the number of verified occurences in category i.

This scoring system is equitable (Gandin and Murphy, 1992), that is, the weigths W_{ij} are such that the score $S(x)$ is 0 for random or a constant categorical forecasts and 1 for perfect forecasts. It also has correspondence with the standard correlation measure (Barnston, 1992). The values of W_{ij} are obtained from information theory and represent the probabilistic *value* of a forecast of tercile j when tercile i actually occurred (Ward and Folland, 1991). The weight matrix is $W_{11} = W_{33} = 1.35$, $W_{12} = W_{21} = W_{32} = W_{23} = -0.15$, $W_{31} = W_{13} = -1.30$ and $W_{22} = 0.30$. The global skill score S is estimated by simply averaging, over the geographical domain of the predictands, the values of the grid-point LEPS scores.

2.3. *Trends, skill biases, and statistical significance of the LEPS scores*

Here we have to face one major difficulty in assessing the confidence of the skill scores: Real-atmosphere data are not stationary in the statistical sense. These nonstationarities can be due, for instance, to the considerable changes in observation and assimilation procedures since the early 50s. However, the climate itself has its own natural trends that we are not able to separate from the analysis biases. It may be desirable to substract trends from a data set (Barnston, 1994) since it can inflate artificially the skill, especially when a cross-validation procedure is used. However, we choose here not to do so, mostly because the extrapolation of a linear trend calculated from the 33-year learning period is a disastrous estimate of it over the verification period. Indeed, trends in the 50 kPa are highly nonlinear, and probably largely dominated by interdecadal variability. Instead, our estimation of the statistical significance of the skill scores takes this trend into account.

Owing to the above choice, three difficulties arise in assessing skill significance. The first one is due to the bias of the LEPS skill score when verification terciles are not balanced, i.e., do not occur exactly one third of the time. In this case, a random model having the "right climatology", i.e., a number of forecast occurrences equal to the number of observed ones over the verification period, has a strictly positive skill score. The second difficulty arises from the combination of trend and systematic errors: Assume that the forecast model has a systematic error that is so large that it always forecasts a given extreme tercile; When this tercile turns out to have also the largest verified occurrence probability, the LEPS score is positively biased. In the opposite case, it is negatively biased. Finally, when the variance of the forecasts is much smaller than the observed variance, the N category is overforecast and the skill is negatively biased (Ward and Folland, 1991). In this case, one still has the possibility of inflating artificially the variance, but the inflation factor can only be estimated using cross-validation. We are not aware of any universal scoring system which avoid these peculiarities. We choose, here, to assess the statistical signific-

ance by a simple non-parametric test that takes into account the above problems.

The statistical significance test compares the measured LEPS score with that of a random forecaster predicting the same tercile frequencies as the model to be tested. Tests are done in a non-parametric way, almost like in Déqué and Royer (1992), but they are inherently model — and grid point — dependent: The 40 forecast tercile maps are shuffled 1000 times, but not the verified tercile maps, yielding 1000 contingency tables from which LEPS scores are calculated. The 900th value, as obtained by sorting them into ascending order, provides a one-sided 90% significance test above which the model score is expected to lie. In this way, all skill bias problems mentioned above also occur for the random forecaster. By substracting the average over the 1000 LEPS score values from the model score, one also obtains an *unbiased estimate* of the score.

This method of estimating score significance, and removing its bias, is very conservative: A simple persistence model, which is, by definition, able to take trends into account, has a vanishing unbiased score. Note also that, in the presence of trends, the unbiased score of perfect forecasts is lower than 1. In the results presented below, both unbiased and standard LEPS scores will be shown in order to allow discussion about the ability of the models to accomodate to the existence of trends.

2.4. *The dynamical model*

The EMERAUDE model (Déqué and Royer, 1992) is adapted from the previously-operational short-range forecast model developed at Météo-France between 1984 and 1992. It is a T42 spectral model, with 20 levels in hybrid vertical coordinates. The simplified physical parametrizations include radiation, convection, hydrological cycle, interactive cloudiness. The integration scheme is semi-implicit with a time-step of 20 mn. Boundary conditions are monthly climatological sea surface temperatures (SST), to which is added a persistent anomaly, computed from days -11 to -2 before the starting analysis. For further details about the model itself, the reader is referred to Déqué and Royer (1992).

The model is run up to 44 days ($D+44$) starting from the 40 dates listed in Table 1. Each forecast consists of a set of 5 lagged integrations starting from ECMWF analyses of days D-2 to D with a 12-h step. In this work, and unless specified otherwise, we always use the average of the ensemble forecast, which provides slightly improved skill, as mentioned by Déqué and Royer (1992). Forecasts of the 50 kPa height are therefore averaged over the 5 individual members, and finally placed into their respective terciles at each grid point.

Fig. 1 shows the large EMERAUDE systematic error in forecasting 30-day averages, over the ATL sector, calculated from the 40 cases, at a 14-day lead time, i.e., for periods ($D+15$, $D+44$). There is a large underestimation of the height in the mid-Atlantic, and a smaller overestimation over Europe, at all lead times. The systematic error is found to be very similar if calculated from only the first or second half of the experiments. In the following, we denote by DYN the categorical dynamical forecasts and DYN-S the forecasts obtained by removing the cross-validated systematic error before tercile calculation. As mentioned above, the cross-validated systematic error is obtained, for a given winter, from the 36 forecast cases issued during the 9 other winters.

2.5. *The statistical model*

The statistical model is that described by VPP (see also Vautard, 1995). It is based on a two-step procedure. First, "predictable components" are calculated, and are extrapolated, using a linear autoregressive model. Second, these extrapolated components are used as predictors for a specification stage using a simple analogue method. We recall here briefly its main characteristics, and the reader is referred to VPP for further technical details.

The first step consists in calculating 5-day averages (pentads) of the 50 kPa geopotential height field anomalies over the whole Northern Hemisphere on the T21 gaussian grid, and performing a principal component analysis (PCA) of these data (Preisendorfer, 1988). Then, space-time principal components (ST-PCs) are obtained from a multichannel singular spectrum analysis (MSSA: Broomhead and King, 1986a,b; Plaut and Vautard, 1994), which is a PCA in the *delay-coordinate phase space*, that is, the space of T-long sequences of consecutive state vectors defined by the first 10 S-PCs, as in VPP. These filtered

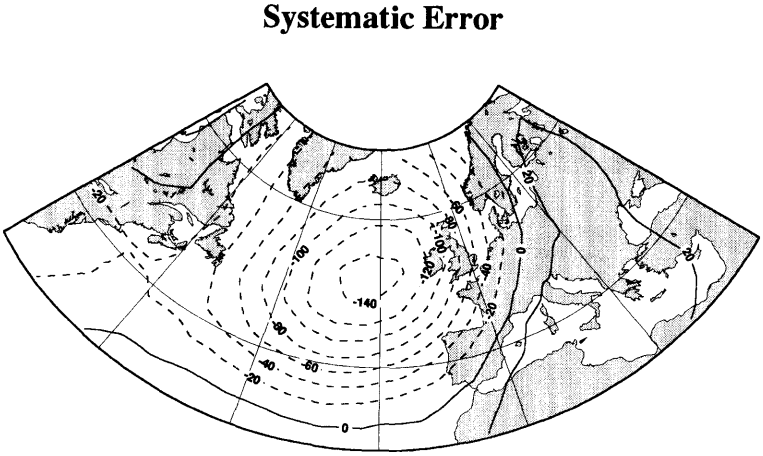

Fig. 1. Systematic error of the 50 kPa height monthly averages (days 15 to 44) of the EMERAUDE dynamical model.

ST-PCs achieve a good compromise between predictability and explained variance (VPP). T is called the window length, and is fixed to 90 days in the present study. Since S-PCs represent pentads, the numerical window length is 18, and the lag-covariance matrix diagonalized by MSSA is 180×180.

The behaviour of the ST-PCs bears similarity with that of forced-damped stochastic oscillators, which justifies the use of autoregressive models for their time extrapolation. The leading 10 ST-PCs are extrapolated using a linear autoregressive model with, as regressors, the leading 100 ST-PCs taken τ days before. Assume, for instance, that present is time t. In order to forecast 50 kPa averaged over the period $(t+s,\ t+s+30)$ (s is called, as in Barnston (1994), the *lead time*), the current ST-PCs, constructed with the latest T-long sequence of observed data, are extrapolated up to the latest time of the average $t' = t+s+30$.

The second step is the *specification* (or "downscaling") step. The problem is to estimate, from the leading 10 extrapolated ST-PCs, the value at each ATL grid point of the 30-day 50 kPa geopotential field average. VPP proposed a probabilistic approach, which has also been used for various specification problems (Klein, 1983). Instead of estimating the values of the predictand, one estimates crudely its conditional probability distribution (CPD), discretized into the three terciles. The CPD is calculated by seeking, within the learning period, the 500 nearest neighbours (analogues) of the forecast 10-dimensional ST-PC vector, and counting the number of occurrences of predictand values falling within each tercile. Then, a deterministic decision is made by forecasting the tercile having the largest conditional probability. The similarity measure used for determining the best analogues is the pattern correlation in the 10-dimensional space spanned by the leading 10 ST-PCs. The use of pattern correlation was also made in long-range forecasting by Bergen and Harnack (1982).

The above forecast scheme was shown by VPP to provide forecast skill larger than the direct forecast by the analogues, skipping the first extrapolation stage. The technical differences between the scheme used by VPP and the present one are as follows.

(a) The window length is $T=90$ days, instead of 180 days in VPP. This choice increases slightly the skill because it reduces the number of parameters to be estimated.

(b) Predictands and predictors are the 50 kPa heights instead of 70 kPa heights in VPP. This is not a major change for monthly means, owing to their intrinsic barotropic nature. Also, in VPP, the prior PCA was performed over the ATL domain only. Here PCA is performed over the global Northern Hemisphere, which adds information to the predictor and increases slightly the skill.

(c) The pool of analogues is restricted to the winter season November through March.

Learning ST-PCs are sampled every 5 days, yielding a pool of about 1000 possible analogues. Our results are fairly insensitive to the choice of these statistical model's parameters. The statistical forecasts will be denoted STA hereafter. Since there are climate differences between the learning and the verification periods, the statistical model has "systematic errors" due to the fact that it is built only from the 1950–1983 analyses. In order to make a fair comparison with DYN-S forecasts, the systematic errors of the linear ST-PC extrapolation can be corrected in the same manner as for the dynamical model. The resulting forecasts are called STA-S.

3. Forecast of "predictable modes"

3.1. Dynamical versus empirical extrapolations of the ST-PCs

In order to assess the skill of each model in extrapolating the low-frequency modes of the atmosphere, we first examine the average pattern correlation between extrapolated and actual ST-PCs. The pattern correlation is defined as the cosine of the angle between two ST-PC vectors (of dimension 10). The dynamically-extrapolated ST-PCs are calculated after the forecast, by projecting the average forecast onto the corresponding ST-EOFs. Since ST-PCs are 90-day averages of the 50 kPa anomalies, weighted by the ST-EOF coefficients, their extrapolation by the dynamical model uses ECMWF analyses in the early part of the 90-day window, and model forecasts in the later part.

Fig. 2 shows the average pattern correlations using the two extrapolation models, calculated from the 40 forecast cases, as a function of the extrapolation time. Also shown is the average pattern correlation obtained from individual members of the dynamical ensemble forecast, and the average pattern correlation obtained from the linear regression model, but calculated over all dates belonging to the winter season (instead of from the 40 cases only). As mentioned above, winter is defined as the period november-march. Fig. 2 shows basically that all extrapolation models have almost the same extrapolation skill. The ensemble-averaging operation increases slightly the skill. Dynamical extrapolations, when averaged over their five members, beat the linear-regression extrapolations at all lead times considered here.

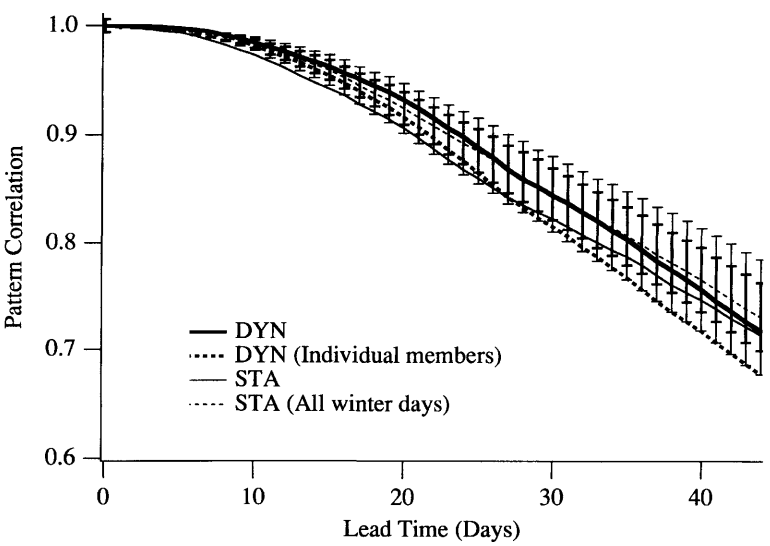

Fig. 2. Average pattern correlations between extrapolated ST-PC vectors, using the dynamical (DYN) and the statistical (STA) models, and actual ST-PC vectors (as obtained from ECMWF analysis after projection onto the ST-EOFs). Note that lead time is defined here as the lag between extrapolation time and the time of the forecast. Light vertical bars: standard deviation for 40 starting dates randomly choosed (see text). Heavy error bars: the same for 100 random starting dates.

In order to check to what extent does the restriction to 40 cases limit the statistical significance of our comparison, we randomly generate 1000 sets of 40 starting analysis dates (4 within each of the 10 winters 1983–1992) and forecast the corresponding ST-PCs with our linear statistical model. In this way we have at our disposal 1000 values of the pattern-correlation at any lead-time and are able to compute a mean and a standard-deviation (light vertical bars of Fig. 2). Notice however that these cases are not independent. Since we cannot check in a similar way the pattern correlations of the dynamically forecast ST-PCs, we can do no more than conjecture the following (see Fig. 2): (i) The 40 cases under study are probably not favourable to the STA model, since the corresponding skills lie in the lower part of significance interval; (ii) No definite conclusion may be drawn from the 40 cases in assessing the difference between dynamical and statistical models skill; (iii) Even if one used 100 cases (10 within each of the 10 winters), instead of 40 (heavy bars on Fig. 2), and provided the dynamical model pattern correlations do not change much, more definite conclusions could hardly be drawn.

3.2. Spread of dynamical forecasts and skill

Many meteorological centers now carry out dynamical extended-range forecasts using several realizations of the initial conditions in order to provide an ensemble forecast. The actual atmosphere's state is thus expected to lie within the cloud of forecasts, which in principle is rigorously true when (i) the cloud of initial conditions encompasses the actual one (Epstein, 1969), and (ii) when the model is perfect. In practice, condition (i) is hard to verify while condition (ii) is definitely not satisfied. Fig. 3 shows, for instance, the evolution of each of the five individual dynamical forecasts projected onto the plane spanned by ST-EOFs 3 and 4, for the forecast case starting on 18 january 1987. Also shown in Fig. 3 are the verifying ST-PCs obtained from ECMWF data and the extrapolation obtained by the linear regression model. For this particular case, the verification ST-PCs do not lie within the cloud of forecasts, even when the systematic error of the model is removed (Fig. 3b). This indicates that systematic errors may not be the only source of the displacement of the forecast cloud with respect to reality. However, for this example, we do not have any mean of determining whether condition (i) or (ii) are violated. The perturbation generation procedure is very simple here, and the simulation of analysis errors can certainly be improved by the use of more sophisticated schemes such as the "bred growing modes" (Toth and Kalnay, 1993), or the "optimal perturbations" (Mureau et al., 1993).

4. Predictability of monthly-mean geopotential heights

In this section, we compare the LEPS skill scores of the 30-day average tercile forecasts, for the DYN, DYN-S, STA and STA-S models.

4.1. Global scores

The global skill scores of the models DYN and STA are displayed in Fig. 4a as a function of lead time. Lead time is defined, as in Barnston (1994), as the time interval between the latest known data and the first day of the average to be forecast. Since DYN forecasts are carried out for periods of 44 days, the lead times run here from 0 to 14 days. The DYN forecasts have a clearly higher skill at shorter lead times, but for lead times exceeding 10 days, the score difference with STA forecasts becomes nonsignificant. Note that, due to the small number of cases, the skill score of the STA model turns out to be higher at long lead times than at short lead times. When the score of STA is calculated from all winter days (november through march), it decreases monotonically with lead time, as expected. Once again, the 40 selected cases turn out to have below-average score. Also shown on Fig. 4a is the average score of the single-member DYN forecasts, which is slightly lower than that of the ensemble average forecasts. Finally, these scores are clearly higher than those obtained from a simple persistence model, where the forecast tercile is the latest-known tercile, i.e., that of the 50 kPa 30-day average ending at the day of the forecast.

The DYN scores are significant at the 90% confidence level at all lead times. Note that our significance test leads to 10%–90% error bars lying entirely in the positive score domain. This results from the combination of trends and system-

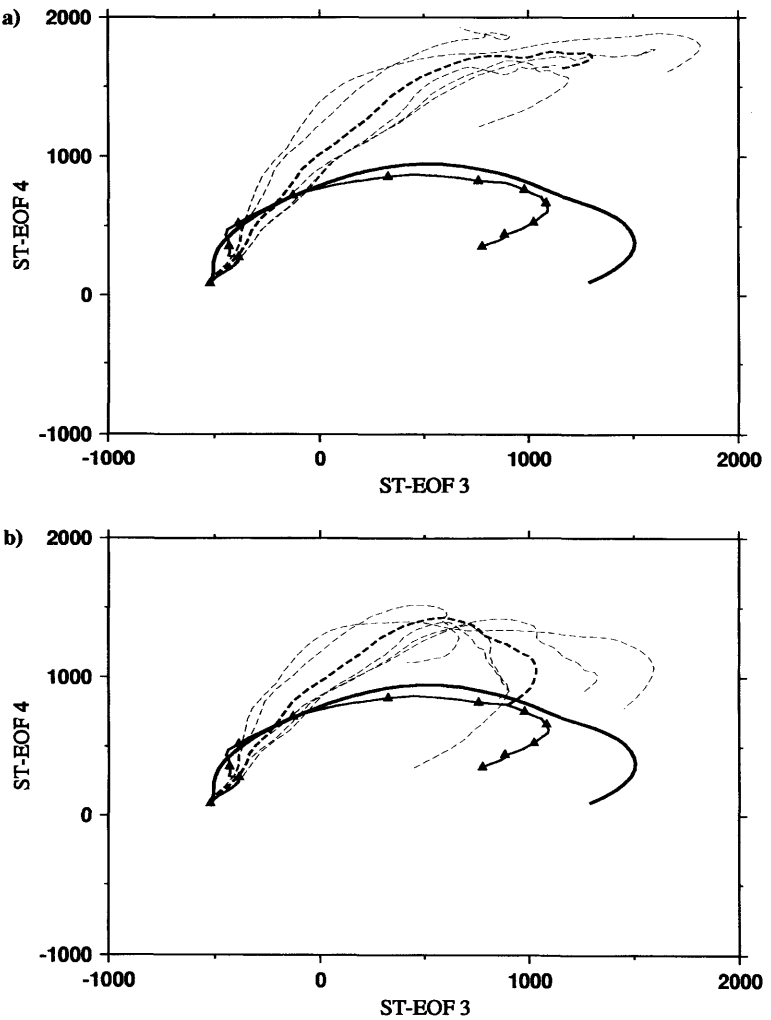

Fig. 3. Extrapolations of ST-PCs issued from various models, starting on January 18, 1987, projected onto the plane spanned by ST-EOFs 3 and 4: (a) Raw dynamical forecasts (DYN), and (b) corrected ones (DYN-S). Light dashed lines: individual members of the ensemble forecasts; heavy dashed lines: average DYN and DYN-S extrapolations; heavy solid line: extrapolation from the linear regression scheme; line with triangles: projected ECMWF verifying analyses. The curves all start from a single point representing the state of the ST-PCs at the time of the forecast.

atic errors mentioned in Section 2. The STA model score, when calculated from the 40 cases, is only marginally significant at some lead times; however, the STA score calculated from all winter days (not shown) is, in fact, significant at all lead times. Note finally that the confidence intervals are tighter for the DYN model than for the STA model. This is due to the combination of two factors: (i) STA tends to produce more extreme forecasts than DYN (see VPP for explanations), and (ii) LEPS scores rewards (penalizes, respectively) more correct (erroneous, respectively) extreme tercile forecasts than near-normal forecasts. Note finally that the persistence model has, as anticipated in Subsection 2.3, a nonsignificant score. This confirms the severity of our significance test.

Fig. 4b shows the unbiased scores (see

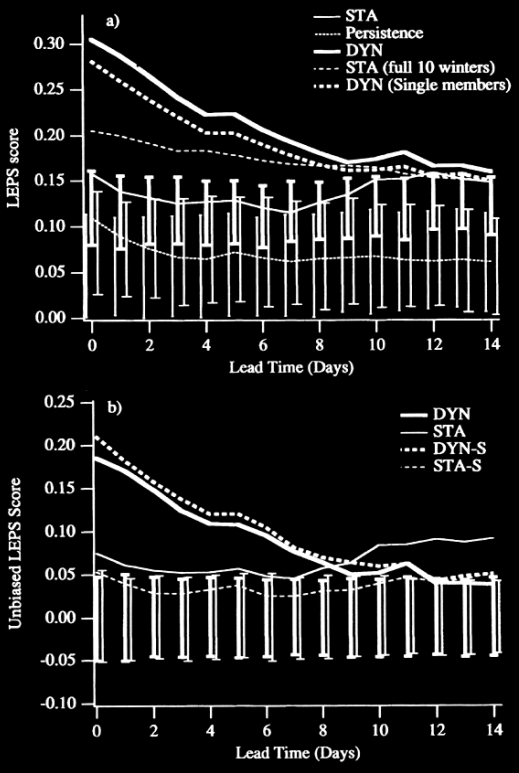

Fig. 4. Global LEPS scores of the DYN, STA, and persistence models, versus lead-time, for the forecasts of the 50 kPa heights. (a) Raw forecasts for the DYN, STA, persistence, and individual-members of the dynamical model. Heavy error bars represent the 10%–90% confidence interval for the random forecaster associated to the DYN model; Light error bars stand for the STA model, and light bars without caps stand for the persistence model. See also the legend on the graph. (b) Same as (a) for the *unbiased* LEPS score of the DYN, DYN-S, STA, STA-S models. Heavy (light, respectively) error bars are associated to the DYN-S (STA-S, respectively) model.

Subsection 2.3 for definition) of the DYN, STA, DYN-S and STA-S models. The STA and DYN unbiased score curves cross at a lead time of 8 days. A comparison of the difference between the two curves with the size of the confidence intervals shown in Fig. 4a indicates, however, that the significance of this difference cannot be fully established from our experiments. Figure 4b also shows that the removal of systematic errors does not clearly improve the unbiased skill of the models. The STA-S scores are even almost systematically non-significant at the 90% confidence level.

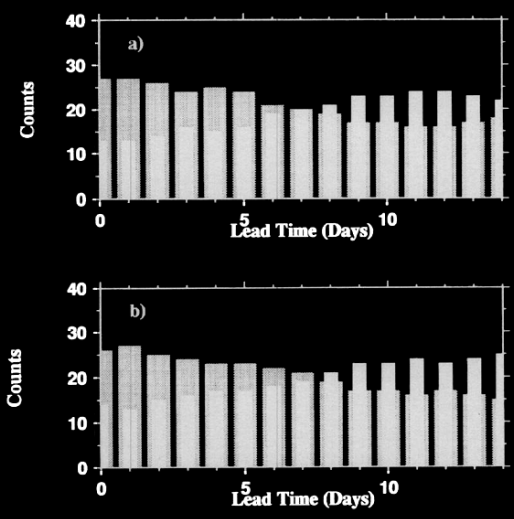

Fig. 5. (a) Number of cases for which DYN beats STA (shaded bars), and reverse (solid bars), versus lead-time. (b) The same for DYN-S and STA-S.

The variations of the respective skills of the DYN and STA models with lead time are also illustrated on Fig. 5a (Fig. 5b, respectively for the DYN-S and STA-S models) by calculating the number of cases, among the 40 considered, for which one model provides a better forecast than the other. The skill measure is again the LEPS score of the contingency matrix, but is cumulated over the 176 ATL grid points only, for a given case and lead time. For lead times longer than 8 days the STA (resp. STA-S) forecasts seem more accurate than the DYN (DYN-S, respectively) forecasts, in terms of number of successes.

A complementary way to characterize both model performance is to examine the cumulated contingency tables they produce. With 40 experiments and 176 grid-points, a random forecaster should provide roughly 780 events in each cell of the contingency table. The most outstanding features of DYN forecasts at a 14 day lead-time (Table 2) mainly consist in a 50% excess of right tercile B forecasts, together with an equivalent shortage of wrong tercile A forecasts. In fact, the DYN model forecasts almosts twice as many B terciles as A terciles. The STA model is more balanced, but with an excess of extreme tercile forecasts, which is inherent to the Bayesian formulation of the statistical model (Van den Dool and Toth (1991) and VPP for explanations).

Table 2. *Contingency tables of 30-day average Z500 grid-point forecast for a lead time of 14 days*

	Emeraude (DYN) (DYN experiments)			Statistical (STA) (STA experiments)			Statistical (STA) 10 full winters		
	B	N	A	B	N	A	B	N	A
tercile B	1249	885	805	1218	1056	769	1204	980	784
tercile N	725	798	842	283	391	466	335	355	436
tercile A	374	555	807	847	791	1219	804	909	1233

Rows correspond to forecast terciles, and columns to observed terciles. There are $40 \times 176 = 7004$ possibilities. The 10 full winter table has been renormalized to make comparison easier.

4.2. Do models forecast trends?

The title of the present subsection may appear somewhat provocative since trends involve at least multi-annual observations, while the present study deals with intraseasonal forecasting. It should be more exact to ask whether models trained with a given "climate" are able to take into account the climate differences between learning and verification periods. Particularly, the statistical model should eventually (for infinite lead times) provide a forecast that converges toward the climate of the learning period. The key question is whether this convergence time exceeds the lead times under consideration. We demonstrate here that such is the case.

We display in Fig. 6 the difference between the verified probabilities of tercile A and B at a 14-day lead time, and the same difference but for STA and DYN forecast terciles. Relative to the 1950–1983 climatology, the verified dates display a clear lack of A terciles, in the area South of Greenland, and an excess of A tercile over Europe. The systematic drift of the DYN model towards its own climate leads to a large excess of B terciles in the mid-Atlantic area, together with an excess of A terciles in the Mediterranean area. Although the pattern showed in Fig. 6b is correlated with that of Fig. 6a, one cannot conclude that the DYN model is able to correctly take the trend into account, since the systematic error of the model (see Fig. 1) has significant correlation with the pattern of Fig. 6a: it is so large that in the mid-Atlantic area the tercile B is forecast for all the 40 cases. The correlation between systematic error and trend is likely to be due to pure chance.

The "systematic error" of the STA model can only stem from the trend itself and only results in *underestimation* of the trend, as revealed by comparison of Fig. 6a and 6c. However, the trend pattern itself is largely recovered. For this model, skill inflation due to trends is due to true ability of the model to adapt to the trend, and is thus not artificial. Therefore, one must be very carefull in interpreting scores in the presence of trends. Actually, going back to Fig. 4, the low unbiased score of the STA model indicates that over these 40 cases, the greatest part of the STA model's skill comes from adaptation to the trend itself, while for short leads, DYN forecasts tend to have a true skill in forecasting the variability about this trend.

4.3. Local scores

As already mentioned by Déqué and Royer (1992), local scores may be hardly seen as significant in a 40-case study. The results presented below confirm this point. Figs. 7a-d show the grid-point LEPS scores of the DYN, DYN-S and STA models, at a 14-day lead time. The DYN model achieves the highest (lowest, respectively) scores South of Greenland and over the Mediterranean area (resp. over the Eastern Atlantic). However, it is likely that these peak scores are mostly due to the combination of trends and systematic errors, as mentioned above, since the peak values occur precisely where the trends are highest (compare with Fig. 6a). When the systematic error of DYN is removed, the score pattern still exhibits a large area of negative values over the Eastern Atlantic. The STA model (Fig. 7c) has a more homogeneous score pattern, with negative values only in scattered areas, over the very Eastern part of the domain and North Africa. Its score also exhibits larger values over Europe and South of Greenland.

In order to have a better estimate of the STA score pattern, Fig. 7d shows the score map obtained

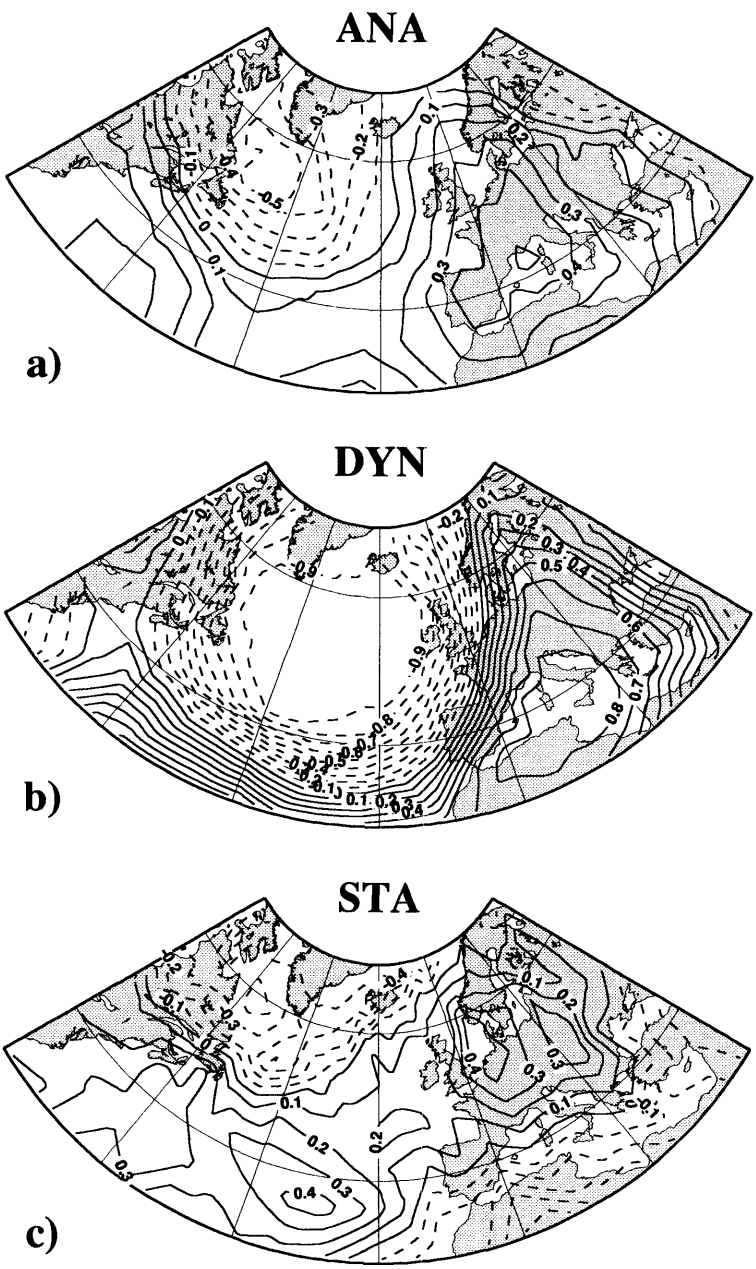

Fig. 6. Difference between the occurrence probabilities of the A tercile and the B tercile, calculated from the 40 forecast cases at lead time 14 days, for (a) the verified ECMWF analysis (ANA), (b) terciles forecast by the DYN model and (c) terciles forecast by the STA model. Contour interval is 0.1.

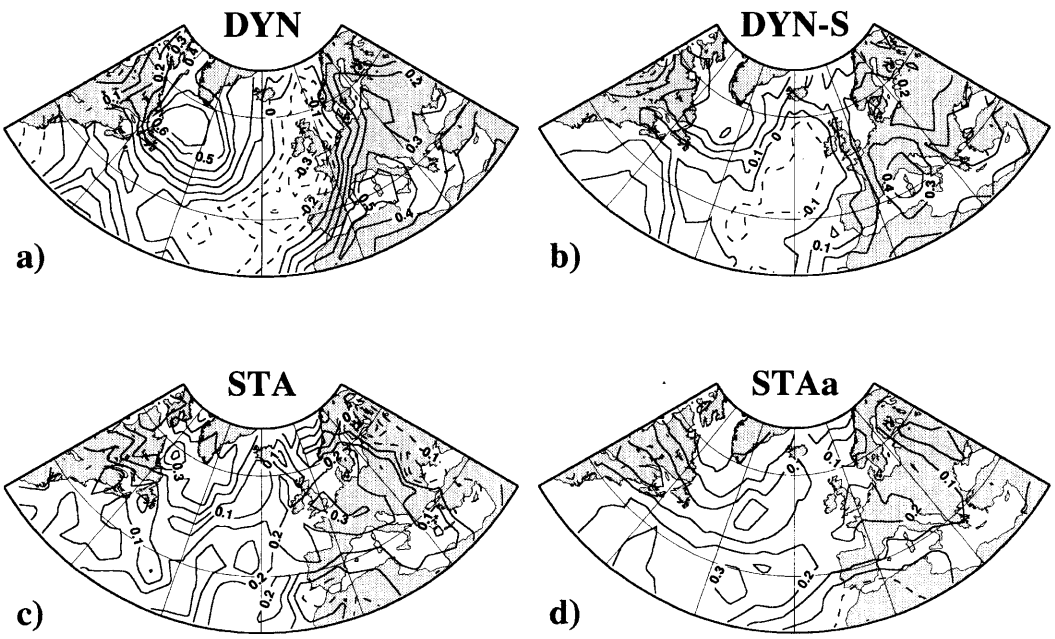

Fig. 7. Geographical distribution of the (biased) LEPS score for the DYN (a), DYN-S (b), and STA (c) models. In panel (d), the score of the STA model, calculated from all possible predictand days within the cold season along the verification period (10 winters) is represented (STAa). Contour interval is 0.1.

from forecasts over all winter days of the verification period. The resulting map strongly resembles the score map of VPP, who also found significant skill in 3 major areas of the same domain: near Greenland, over the Southern Atlantic and over Europe. These significant skill areas were shown to correspond to an enhanced predictability of the NAO seasaw (Déqué, 1988; VPP).

The *unbiased* score maps of the three models, at a lead time of 14 days, are shown in Figs. 8a–c. Both DYN and DYN-S still exhibit large areas of negative skill, with scattered peaks which are now located at different places relative to that of the biased skill maps. The significance (not shown) of these maps is, as anticipated, very poor. Only areas of score roughly higher than 0.2 are significant at the 90% confidence level. Some scattered grid points satisfy this condition, for both DYN and DYN-S, while the STA score map exhibits three areas of significant skill: over the Mediterranean area, along a ridge extending from the South Atlantic to the British Isles, and near the Labrador sea. It can be pointed out that, like for its systematical error, the pattern correlation of DYN appears rather robust against partitions of DYN experiments (not shown).

These maps also show that the distribution of skill of the STA and DYN models are qualitatively different. One should therefore expect to gain some skill from the combination of the two forecasts, which is the purpose of hybridization. The fact that the skill of the two models is very different not only results from the systematic errors of the dynamical model, but also from its systematical deficiencies, that is, systematic errors in higher-order moments.

5. Hybrid forecasts

5.1. The hybrid models

If indeed ST-PCs are predictable features, they should be well extrapolated by the dynamical model itself. One way to improve the statistical model is therefore to replace the linear extrapolation of the ST-PCs by a dynamical one and to apply the specification stage exactly as before, using the analogue method. This is the first, simple

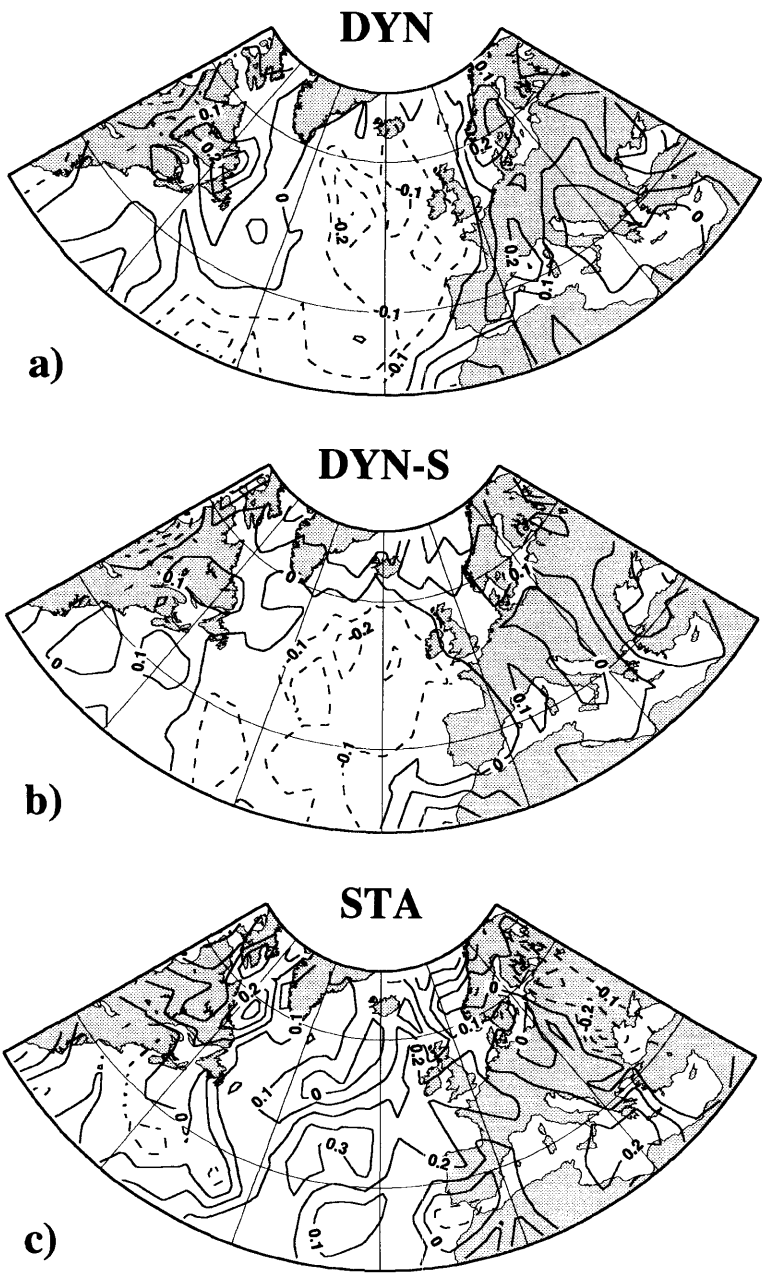

Fig. 8. Same as Fig. 7, but for the unbiased LEPS score of the DYN, DYN-S and STA models calculated from the 40 cases only.

hybrid statistical-dynamical model, called hereafter HYB1. This very simple method has the advantage of retaining, from the dynamical forecasts, only the most predictable modes. The systematic error of the dynamical model is not removed before the projection onto ST-EOFs, hence the HYB1 forecasts only use statistics from the learning period.

The STA forecasts and the HYB1 forecasts both use extrapolated ST-PCs as predictors for the specification stage. If one of the two extrapolation models turns out to be largely more skillful than the other, then one should use it for the specification. However, the extrapolation error can be of the same order of magnitude, but have different statistics. In this case, it should be useful to combine these extrapolations in order to reduce further the extrapolation error. The second hybrid model, denoted hereafter HYB2, consists in calculating a *best linear unbiased estimate* (BLUE: Jazwinski, 1970; Gelb, 1986) of the two extrapolated sets of ST-PCs (STA and HYB1), and applying again the same specification procedure as for the STA model.

Let us denote $a_k(t+\tau)$ the actual kth ST-PC at time $t+\tau$ (which uses observed data from the window $(t\text{-}T+\tau,\ t+\tau)$), $\bar{a}_k(t+\tau)$ the kth ST-PC extrapolated from the linear regression model, and $\tilde{a}_k(t+\tau)$ the kth ST-PC extrapolated by the dynamical model. The BLUE procedure provides another extrapolated estimate $\hat{a}_k(t+\tau)$ which is a linear combination of the two extrapolations,

$$\hat{a}_k(t+\tau)=h_k(\tau)\bar{a}_k(t+\tau)+(1-h_k(\tau))\tilde{a}_k(t+\tau). \qquad (3)$$

The weights $h_k(\tau)$ are those which minimize the rms error between $\hat{a}_k(t+\tau)$ and the observed ST-PC $a_k(t+\tau)$ over the learning period. In fact, the BLUE procedure results from a more general estimation problem, based on a maximum likelihood procedure which assumes that errors follow a gaussian distribution. Given this assumption, the new estimate (3) gives actually the peak of the conditional distribution of $a_k(t+\tau)$ given its two estimates $\bar{a}_k(t+\tau)$ and $\tilde{a}_k(t+\tau)$.

When one of the two extrapolations is more skillful on average, its weight is naturally found larger. In principle, an even better estimate can be provided from a multivariate estimate, by a global minimization of the rms error between the estimated ST-PC *vector* $\hat{A}(t+\tau)$ (of dimension 10), and the observed ST-PC vector $A(t+\tau)$, which would take into account the cross correlation between ST-PC errors. However, this approach fits weighting matrices instead of single weights. In our case, this would lead to fitting $10\times 10=100$ coefficients instead of 10 coefficients. Since the number of independent forecasts is small, overfitting problems make the multivariate approach useless. The lag-dependent weights $h_k(\tau)$ are given by

$$h_k(\tau)=\frac{\bar{\sigma}_k^2-c_k}{\bar{\sigma}_k^2+\tilde{\sigma}_k^2-2c_k}, \qquad (4)$$

where $\bar{\sigma}_k^2$ and $\tilde{\sigma}_k^2$ are the respective variances of the (lag-dependent) errors of the linear and the dynamical extrapolations of the kth ST-PC, and c_k is the covariance between their errors. The weights are calculated in cross-validation mode, as mentioned in Subsection 2.2, as for the removal of the systematic error: The BLUE coefficients are calculated, for a given verification winter, using the extrapolated ST-PCs from the 9 other winters. The number of extrapolations used to calculate these coefficients is thus 36, which is probably too small a number to substancially improve the forecast skill. Note that for the same reason, although it results in increased correlations in the learning period ST-PCs, the use of less-constrained parametrizations, such as a linear regression, instead of eq. (3) would lead to increased overfitting problems.

The pattern correlation between the three sets of extrapolated ST-PCs (remember the ST-PCs of HYB1 are by definition those DYN), and the actual ST-PCs, is shown in Fig. 9, in the same format as in Fig. 2. Although the difference between the three curves is nonsignificant in a statistical sense (compare with the error bars in Fig. 2), one notices that the linear combination between ST-PCs extrapolated from the dynamical model and the linear regression produces a marginal improvement, especially at long lead times. At shorter lead times, the skill of the BLUE

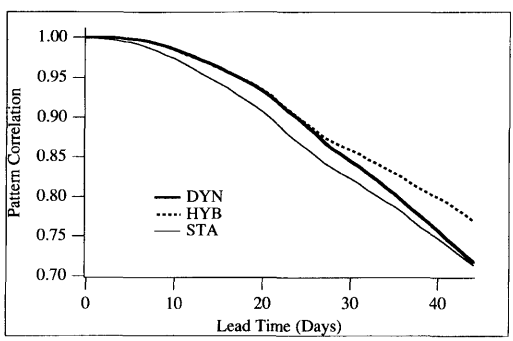

Fig. 9. Same as Fig. 2 for the dynamical (DYN), linear regression (STA), and the BLUE hybrid (HYB) extrapolation models.

extrapolation is the same as that of the dynamical extrapolation.

5.2. *Skill of the hybrid models*

Fig. 10a shows the (standard) global LEPS scores of the DYN, STA, HYB1 and HYB2 models. The global scores of HYB1 exceeds 0.24 at all lead times. Only at lead times smaller than 2 days does DYN perform better than HYB1. The score of HYB2 forecasts is smaller than that of HYB1, presumably due to sampling problems. The two hybrid models have scores significant at the 90% confidence level, using the same testing procedure as above. When considering unbiased scores (Fig. 10b), however, HYB1 does not beat STA for lead times longer than 10 days. By contrast, at long lead times, the BLUE procedure seems to provide the best skill (HYB2). By comparing the size of confidence intervals (Fig. 10a) with the difference between the DYN and HYB2 curves of Fig. 10b, one notices that the HYB2 procedure significantly improves the skill of the dynamical model in the long run, while the difference between HYB1, HYB2 and STA scores is nonsignificant.

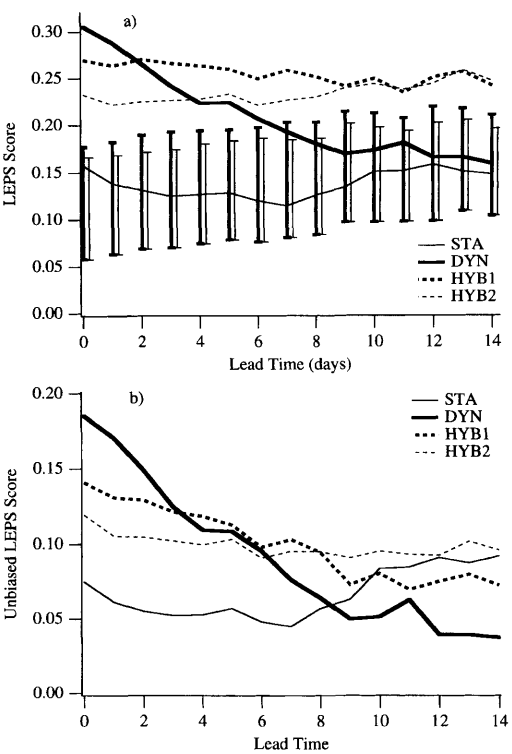

Fig. 10. (a) Same as in Fig. 4 (biased LEPS score), for the STA, DYN, HYB1 and HYB2 models (see legend on the graph). Heavy error bars stand for the HYB1 model and light error bars stand for the HYB2 model. (b) Same as Fig. 10a, but for the unbiased LEPS scores.

The biased/unbiased score maps of HYB1 and HYB2 are shown in Figs. 11a-d, which are to be compared with Figs. 7, 8. From a qualitative point-of-view, the skill of HYB1 and HYB2 do not differ greatly, with peaks South of Greenland and over Europe. The skill values are positive almost everywhere, and are much higher than those of STA and DYN. However, when skill bias is removed (Figs. 11c–d), there are still areas of negative skill, such as over Eastern Europe and the North Atlantic. Once again, grid-point skill scores are generally nonsignificant at the 90% confidence level, and it is hard to draw any definite conclusions from these figures. Note finally that HYB1, which does not use any data in the verification period, does correctly take into account the climate differences (or trends) between learning and verification (not shown).

The ideal behaviour of an hybrid model would be to perform better than the two original models. With the above formulation, this can only be true in a statistical sense. Moreover, due to the very different nature of the two models (DYN is continuous and deterministic while STA is discrete and probabilistic), the HYB1 and HYB2 models cannot be expected, in theory, to perform best even in a statistical sense. For instance, at short lead times, the skill of the hybrid models are lower than that of DYN. From the statistical point-of-view, the small size of the the dynamical forecasts set is a strong limitation for hybridization procedures. In the near future, long coupled ocean-atmosphere simulations should improve this shortcoming, and help us to identify what minimal number of cases is required in order to reach statistical convergence of the hybridization coefficients.

Finally, we emphasize that all these results are fairly insensitive to the statistical model's parameters. We have performed many experiments, by using the ATL sector only, as in VPP, for building the ST-PC sets, by changing the window length between 60 to 180 days, the number of predictors between 5 to 15, and the number of analogues between 200 to 700. All these experiments yielded

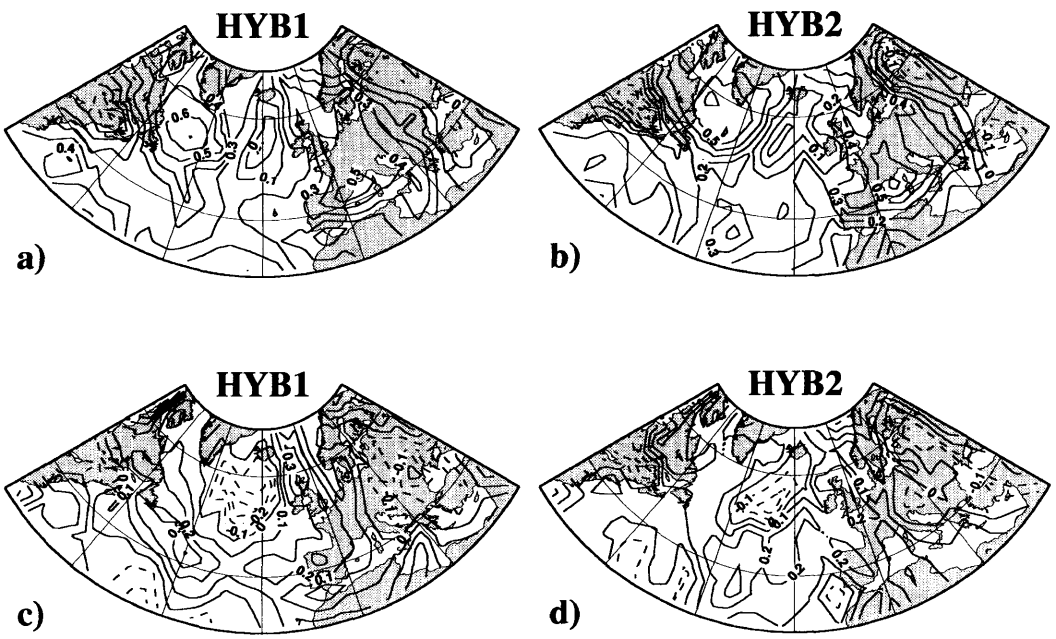

Fig. 11. Same as Fig. 7 and 8, but for the HYB1 (a) and HYB2 (b) biased scores, and the HYB1 (c) and HYB2 (d) unbiased scores.

the same qualitative conclusions. This allows us to place some confidence in our comparative study conclusions, despite the small number of forecast cases processed. However, these conclusions may a priori depend upon the choice of the 83–92 verification period.

6. Summary and conclusion

We performed a direct comparison between a statistical (empirical) two-step forecast model and a series of 40 experimental long range dynamical forecasts performed at Météo-France using a simplified-physics T42 version of the former French operational model Emeraude. For each forecast case, the dynamical model EMERAUDE was initialized with ECMWF analyses of particular fall or winter days between late 1983 and early 1993; anomaly of the boundary conditions was persisted along the forecast. The statistical model, designed by Vautard *et al.* (1996), consists in (i) extracting from the set of predictors its most "predictable components", the *space-time principal components* (ST-PCs), obtained through multi-channel singular spectrum analysis (MSSA), (ii) extrapolating these predictable components, and (iii) using these extrapolated ST-PCs in a specification procedure to forecast the predictand. The success of this methodology results from the filtering step, where unrelated noise produced by the forecast models is removed.

The predictands of our experiments are the 50 kPa height 30-day averages over the Atlantic-Europe sector. The comparison of the skill of both models is made possible by placing the deterministic dynamical forecasts at each grid point into terciles and using a common and objective procedure for assessing skill, the so-called LEPS scoring system (Ward and Folland, 1991), which quantifies actually the *value* of a categorical forecast. This score turns out to be biased when there are imbalances in the tercile frequencies between the learning period, from which the statistical model is built, and the verification period over which the true skill is estimated. These biases are enhanced in the particular case when the model has systematic errors. These problems are addressed carefully along the article.

The main outcome of the comparison is disap-

pointing: We are not able, with only 40 independent forecast cases, such as provided by the EMERAUDE experiments, to find any statistically significant skill difference between the two models, except at short lead times, where undoubtedly the dynamical model beats the statistical model. At longer lead times (typically for the forecast of the 50 kPa height of the average over days 15–44), the global skills are not distinguishable. However, a careful examination of biases and systematic errors leads us to argue that at long lead times, the statistical model beats the dynamical model, mostly because of the systematical deficiencies of the latter. An interesting result is that the statistical model turns out to be relatively skillful at adapting to the trends, or, more precisely, the climate differences between the learning period (1950–1983) and the verification period (1983–1993).

Another important aspect is that the skill patterns of the two models are quite different, although, once again, poorly significant. Therefore, one expects that a proper combination of the two forecasting systems should improve their skill. The second major aim of this article is to develop and apply two hybrid statistical-dynamical models. The first one consists in the extrapolation of the ST-PCs by the dynamical model itself, instead of the statistical model, followed by the same specification procedure as for the statistical model. The second hybrid approach consists in combining the two ST-PCs extrapolations in an optimal manner, using a classical estimation technique, the *best linear unbiased estimate.*

The two hybrid schemes perform better than both the dynamical and statistical model, at long lead times, i.e. typically for the prediction of the 50 kPa average between days 15 and 44. The average skill of the second hybrid model, in particular, is shown to beat significantly the dynamical model. The skill values and their significance are recapitulated in Fig. 12, for the above lead time. Both in terms of score and significance, the HYB2 model performs best. All models but the persistence model have a score significant at the 90% confidence level, but only the statistical and hybrid models pass the test at the 95% confidence level. The fact that the confidence intervals lie in the positive score domain results from trends present in the data, which bias the

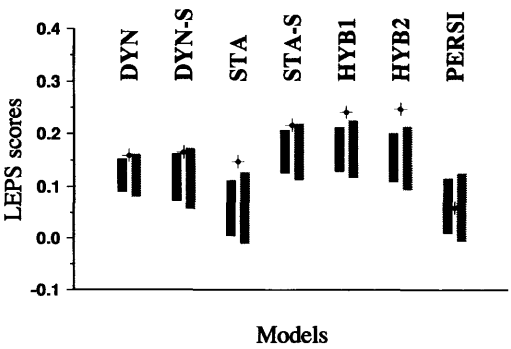

Fig. 12. Global LEPS scores of the various models at a lead time of 14 days (Heavy dots with crosses), together with the 10%–90% (heavy shaded bars) and 5%–95% (light shaded bars) confidence limits of their associated testing random forecaster.

score estimates. The conservative significance test performed here takes into account this bias. Fig. 12 demonstrates the advantage of the strategy consisting in extrapolating only predictable modes (this is the case for STA, HYB1, and HYB2) and specifying predictands at a second stage, regardless of the model used (empirical, dynamical or hybrid).

All the experiments performed throughout this article use upper-air predictands, but the specification (or downscaling) technique is fairly general, and has also been used in previous studies (Klein, 1983). Since models usually do not have an accurate representation of the surface weather, the application of our statistical or hybrid approach to surface weather prediction, which is straightforward, can be very useful. Such an application, for the seasonal prediction of Canadian surface-air temperature, is proposed by Vautard, Plaut and Brunet (1996; manuscript in preparation). We also verified that the statistical model could help curing the systematic deficiencies of the dynamical model, that is the model errors.

Many improvements could actually be brought to our empirical forecast scheme, such as integrating more fields in our predictor set, in particular SST anomalies. Further refinements of the specification technique can also be brought, such as the use of canonical correlation analysis (CCA: Hotelling, 1936; Barnett and Preisendorfer, 1987; see also Von Storch, 1995 for an application of

CCA to the downscaling problem). The hybrid schemes themselves can also be refined.

The lack of statistical significance in the results of this article is due to the small number of cases (40) used, and calls for a methodological study within a more idealized framework such as that provided by a simplified general circulation model, with a "perfect model" assumption. Using a simple 3-level quasi-geostrophic model as this ideal framework, Pires et al. (1996; manuscript in preparation) obtain the same qualitative conclusions as in the present article, but from many more forecast cases, which actually makes our results reliable and general.

7. Acknowledgements

It is a pleasure to thank M. Déqué, J.-F. Royer at the french meteorological service Météo-France for providing us with the outputs of the Emeraude experiments at our disposals. Fabio D'andrea helped us in extracting the NMC archive. We also benefitted from the useful comments from two anonymous referees). Carlos Pires was supported by the grant BD/2023/92 of JNICT (Junta Nacional de Investigacao Cientifica e Tecnologica) — Portugal. This work was also partly sponsored by the French power company Electricité de France.

REFERENCES

Barnett, T. P. and R. W. Preisendorfer, 1987. Origins and levels of monthly and seasonal forecast skill for United States surface air temperautres determined by canonical correlation analysis. *Mon. Wea. Rev.* **115**, 1825–1850.

Barnston, A. G. 1992. Correspondence among the correlation, RMSE, and Heidke forecast verification measures; refinement of the Heidke score. *Wea. Forecasting* **7**, 699–709.

Barnston, A. G., 1994. Linear statistical short-term climate predictive skill in the Northern Hemisphere. *J. Climate* **7**, 1513–1564.

Bergen, R. E. and R. P. Harnack, 1982. Long-range temperature prediction using a simple analog approach. *Mon. Wea. Rev.* **110**, 1083–1099.

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

Broomhead, D. S. and G. P. King, 1986a. Extracting qualitative dynamics from experimental data. *Physica D* **20**, 217–236.

Broomhead, D. S. and G. P. King, 1986b. On the qualitative analysis of experimental dynamical systems. *Nonlinear Phenomena and Chaos*, S. Sarkar, ed. Adam Hilger, pp. 113–144.

Déqué, M. 1988. 10-day predictability of the northern hemisphere winter 500-mb height by the ECMWF operational model. *Tellus* **40A**, 26–36.

Déqué, M. 1991. The probabilistic formulation. A way to deal with ensemble forecasts. *Annales Geophysicae* **6**, 217–224.

Déqué, M. and J.-F. Royer, 1992. The skill of extended-range extratropical winter dynamical forecasts. *J. Climate* **5**, 1346–1356.

Déqué, M., J.-F. Royer and R. Stroe, 1994. Formulation of gaussian probability forecasts based on model extended-range integrations. *Tellus* **46A**, 52–65.

Epstein, E. S., 1969. Stochastic dynamic prediction. *Tellus* **21**, 739–759.

Folland, C. K., and A. Woodcock, 1986. Experimental monthly long-range forecasts for the UK. Part I: Description of the forecasting system. *Meteor. Magazine* **115**, 301–318.

Folland, C. K., A. Woodcock and L. D. Varah, 1986. Experimental monthly long-range forecasts for the United Kingdom. Part III: Skill of monthly forecasts. *Meteor. Magazine* **115**, 377–393.

Gandin, L. S. and A. H. Murphy, 1992. Equitable skill scores for categorical forecasts. *Mon. Wea. Rev.* **120**, 361–370.

Gelb, A. 1986. *Applied optimal estimation.* MIT Press, 374 pp.

Hotelling, H. 1936. Relations between two sets of variants. *Biometrika* **28**, 321–377.

Jazwinski, A. H. 1970. *Stochastic processes and filtering theory.* Academic Press, 376 pp.

Jianping, H., Y. Yuhong, W. Shaowu, 1993. An analogue-dynamical long-range numerical weather prediction system incorporating historical evolution. *Quart. J. Meteor. Soc.* **119**, 547–565.

Klein, W. H. 1983. Objective specification of monthly mean surface temperature from mean 700-mb height in winter. *Mon. Wea. Rev.* **111**, 674–691.

Livezey, R. E. and A. G. Barnston, 1988. An operational multifield analog/antianalog prediction system for US seasonal temperatures. Part I: System design and winter experiments. *J. Geophys. Res.* **9D**, 10953–10974.

Madden, R. 1981. A quantitative approach to long-range prediction. *J. Geophys. Res.* **84**(C10), 9817–9825.

Milton, S. F. and D. F. Richarson, 1991. 30-day dynamical forecasts at the UK Meteorological Office. In: *Proceedings of ECMWF Workshop on New Developments in Predictability*, 13–15 November 1991, ECMWF, Shinfield Park, Reading, pp. 205–245.

Mureau, R., Molteni, F. and T. N. Palmer, 1993.

Ensemble prediction using dynamically conditioned perturbations. *Q. J. R. Meteorol. Soc.* **119**, 292–323.

Palmer, T. N. and D. L. T. Anderson, 1994. The prospects for seasonal forecasting — A review paper. *Quart. J. Roy. Meteor. Soc.* **120**, 755–794.

Plaut, G. and R. Vautard, 1994. Spells of low-frequency oscillations and weather regimes in the Northern Hemisphere. *J. Atmos. Sci.* **51**, 210–236.

Preisendorfer, R. W. 1988. *Principal component analysis in meteorology and oceanography*. C. D. Mobley, ed. Elsevier, Amsterdam, 425 pp.

Toth, Z. and E. Kalnay, 1993. Ensemble forecasting at NMC: The generation of perturbations. *Bull. Am. Meteorol. Soc.* **74**(12), 2317–2330.

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

Tukey, J. 1958. Bias and confidence in not-quite large samples. *Ann. Math. Stat.* **29**, 614.

Van den Dool, H. M. and Z. Toth, 1991. Why do forecasts "near normal" often fail? *Wea. Forecasting* **6**, 76–85.

Van den Dool, H. M., 1994. Long-range weather forecasts through numerical and empirical methods. *Dyn. Atmos. Oceans* **20**, 247–270.

Vautard, R., C. Pires and G. Plaut, 1996. Atmospheric long-range forecasts using space-time principal components. *Mon. Wea. Rev.* **124,** 288–307.

Vautard, R. 1995. Patterns in time: SSA and MSSA. In: *Analysis of climate variability*. von Storch, H. and Navarra, A. (eds.). Springer, Berlin, pp. 259–279.

Von Storch, H. 1995. Spatial patterns: EOFs and CCA. In: *Analysis of climate variability*, von Storch, H. and Navarra, A. (eds.). Springer, Berlin, pp. 227–257.

Ward, M. N. and C. K. Folland, 1991. Prediction of seasonal rainfall in the North Nordeste of Brazil using eigenvectors of sea-surface temperature. *Int. J. Climatol.* **11**, 711–743.