STATISTICAL, DYNAMICAL AND HYBRID LONG-RANGE FORECAST APPROACHES USING PREDICTABLE COMPONENTS

R Vautard, C Pires
Laboratoire de Météorologie Dynamique
Paris, France

G Plaut and J Sarda
Institut Non-linéaire de Nice
Nice, France

Summary: We raise several theoretical and technical issues about long-range forecasting. The first issue is the comparison between statistical models and dynamical models. We demonstrate, using a very simple chaotic dynamical system, that even with a perfect model, dynamical ensemble prediction may require a very accurate knowledge of the initial error statistics in order to beat a simple linear model. We also propose several methodologies in order to combine dynamical and statistical forecasts. These methodologies are tested in a perfect model environment using a simple general circulation model (GCM), and in a real-atmosphere environment using a realistic GCM. These methodologies are based on the identification and the prediction of "predictable components" such as provided by the multichannel singular spectrum analysis.

1. INTRODUCTION

Research on long-range forecasting, on the time scale of several weeks to a season has become abundant within the previous years. This is mainly because computer resources made possible the numerical integration, over long periods, of physical and dynamical models derived from the physics laws (Tracton et al., 1989; Brankovic et al., 1990; Milton and Richardson, 1991; Déqué and Royer, 1992; Palmer and Anderson, 1994, for a thorough review). Well before this possibility was offered to scientists, however, long-range forecasts were carried out by meteorological centers on a routine basis. The forecasting techniques were far different since they were based on empirical grounds. Their success, which was difficult to assess, was questionable. Today, the long records of observed and analysed meteorological variables at least allows the validation of long-range forecast models, but despite the introduction of many degrees of realism in the state-of-the-art numerical models, the long-range predictability is still very poor, with correlations of the order of 0.2-0.3 on average. For lead times exceeding, say, 15 days, empirical models still beat dynamical models (Van den Dool, 1994).

Admittedly, the reasons for the apparent failure of dynamical models are 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 (i) the skill is anyway expected to be low, and (ii) the extraction of a predictable signal usually requires ensemble forecasting, that is, many simulations of the same forecast period. Finally, 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.

The first purpose of this article is to compare the skill of dynamical and statistical (empirical) model approaches, using various examples. In particular, we examine a theoretical aspect of the problem: we argue that not only model errors can explain why a statistical model beats a dynamical model. The lack of knowledge of the initial error statistics can also produce the same effects; We show, by using a simple nonlinear deterministic dynamical system with 5 variables that even within a perfect model framework, a simplified linear (and therefore containing model errors) model can provide more skillful forecasts than dynamical ensemble forecasts using the perfect model and a "bred growing modes" technique for generating perturbations.

Our second purpose is to examine the possibility of hybrid forecasts combining in an "optimal way" a dynamical forecasting system with a statistical forecasting system. Section 2 is devoted to a theoretical examination and discussion of the relative skill of statistical and dynamical models, and of the various technique for building hybrid models. Section 3 contains experimental forecast results obtained within a perfect model framework, the model being a simple quasi-geostrophic one. Section 4 contains results performed with the former french operational general circulation model EMERAUDE for real-atmosphere forecasts. Most of the results presented in this article (especially those of Sections 3 and 4) can be found in Vautard et al. (1995), Pires et al. (1995) and Sarda et al. (1995) to which the reader is referred for technical details.

2. THEORETICAL ASPECTS ABOUT LONG-RANGE FORECASTING

2.1 **Statistical and dynamical prediction as an initial-value problem**

The hope to forecast qualitative characters of the future weather for lead times exceeding the so-called deterministic period (15 days, say), is primarily based on the existence of slowly-evolving climatic modes such as the El Nino phenomenon or regular low-frequency phenomena such as oscillations or slow Rossby waves. It is expected that a significant part of the atmospheric predictable signal is due to slow, predictable variations of the boundary forcing. In this respect, long-range forecasting should not be considered as an initial-value problem. When long-range forecasts are issued from an atmospheric general circulation model, three sources of forecast errors are therefore expected: Errors in the initial conditions, model errors and errors in the

boundary conditions. Mathematically speaking, the third source is not different from the second, unless the forecasts are issued from a coupled atmosphere-ocean model, in which case interface flux errors can also arise from bad initial values.

Since our purpose here is mostly theoretical, we shall only consider the first two above error sources. Let us assume, in a more general way, that the problem is to forecast the future evolution of a (nonlinear) deterministic dynamical system of the form

$$\frac{dX(t)}{dt} = F(X(t)) \ , \tag{1}$$

where F is a deterministic function, and $X(t)$ is the vector representing the state of the system at time t. Both statistical and dynamical approaches are, in a way or another, based on the extrapolation of an initial condition $X(0)$ using a model function $G(X)$ instead of $F(X)$. We shall denote $G_S(X)$ the statistical function and $G_d(X)$ the dynamical function. When the dynamical model is perfect, $G_d=F$. It would be rather astonishing that this occurred for the statistical model!

The fact that the statistical model beat the dynamical model can obviously arise from the fact G_S is a better approximation than G_d, especially since the statistical model is built from real data, while many physical parametrizations involve more or less arbitrary coefficients. The purpose of the next section is more interesting. We want to show that even for a perfect dynamical model, the errors in the initial conditions can lead to the same conclusion. Hence we consider here that our dynamical model is perfect, i.e. $G_d=F$.

Under this assumption, the forecast error only comes from errors in the initial conditions X(0). The initial condition one has at hand is, in fact

$$X' = X(0) + e \ , \tag{2}$$

where e is the error. Of course, e is unknown, but its distribution can be approximated (Lönnberg and Hollingsworth, 1986; Houtekamer and Derome, 1995). Let us denote by ρ the probability density function (PDF) of the error e. Since the system is fully deterministic, if the initial error PDF are known, the PDF of the true future value at time t is also known, and can be obtained by integration of the perfect model from a large sample of random initial error pullings. This latter PDF is

$$\rho_t(X(t)) = p(X(t)|X') \ , \tag{3}$$

where p(A|B) denotes the conditional probability of event A knowing event B. The best single prediction, in the least square sense, that can be given is therefore the average of the random variable X(t) whose distribution is given by Eq. (3). Thus, in a perfect-model environment, the

best forecast is obtained by the classical ensemble average forecast strategy. Another possibility is to use a Bayesian strategy by choosing the peak of ρ_t. In any case, the key problem is to approximate as closely as possible the initial error PDF. By the way, this PDF is not unique since it may result also from various assumptions. For instance, it may depend on the current flow or on the flow over the past few days, i.e.

$$\rho(e) = p(e|X'(t)) \tag{4a}$$

or

$$\rho(e) = p(e|X'(t), X'(t-\alpha), X'(t-2\alpha), \ldots) \ . \tag{4b}$$

The dependence of initial error statistics on the past flow is obvious since classical assimilation processes use a "first guess" resulting obtained as a forecast from the previous analysis cycle. The first conclusion is that there is not a unique ensemble average forecast, and that sets of initial perturbation pullings can be obtained in various ways. Which initial perturbation generation strategy to use in order to have an optimal forecast is a difficult question. Probably, the quality of the forecast in a perfect model depends mostly on the entropy H(ρ) of the chosen PDF ρ,

$$H(\rho) = -\int_e \rho(e) \log(\rho(e)) de \ , \tag{5}$$

which measures the information content of the PDF; If H is low, the PDF is more "peaked" and the forecast cloud is more concentrated around the true forecast. Unfortunately, there is, to our knowledge, no general way of estimating the "minimum-entropy" initial error PDF, which would give the best forecast, and hence reach the predictability limit of the system.

Unfortunately also, not all initial-error PDFs are compatible with the true error. If a given strategy, such as the "bred growing modes" (Toth and Kalnay, 1993), or the "optimal perturbations" (Molteni and Palmer, 1993; Palmer et a., 1993) is used systematically, the resulting PDF may be inconsistent with the true error statistics and generate erroneous ensemble average forecasts. Some of these aspects are discussed in Houtekamer and Derome (1995). We now turn to an example.

2.2 Experiments with a simple dynamical system

In this section, we demonstrate that a perfect ensemble-average forecast model can provide less skillful long-range predictions than a simple linear (thereby imperfect) model, and that the defficiency is due to bad estimations of the initial error statistics. The dynamical system under consideration is the 5-variable model proposed by Molteni et. al (1993), and was designed in the first place to illustrate the modulation by tropical forcing of the mid-latitude regimes frequencies.

This model consists in the classical 3-variable model of Lorenz (1963), coupled to an harmonic oscillator. Its evolution equations are:

$$\begin{aligned}
\frac{dx}{dt} &= -\sigma x + \sigma y + \alpha v \\
\frac{dy}{dt} &= -zx + \rho x - y + \alpha u \\
\frac{dz}{dt} &= xy - bz \\
\frac{du}{dt} &= -\Omega v - ku - \alpha y \\
\frac{dv}{dt} &= \Omega u - kv - \alpha x
\end{aligned} \tag{6}$$

The coupling is conservative since the sum of all cubic terms vanish in the energy equation. We set the parameter values to b=8/3, ρ=30 and σ=10 (the same values as used by Molteni et al. (1993)). The oscillating variables are *u* and *v*. They are much more predictable and slow than the "Lorenz" variables *x*, *y* and *z*. The doubling time of small errors of the Lorenz variables is about 0.7 units, hence a proper units correspondence with the real atmosphere would be that 1 Lorenz unit correspond to about 5 days. The oscillatory variables, which mimic some predictable intraseasonal oscillation (namely the Madden and Jullian one), must therefore have a period of about 50 days. Hence we chose Ω=0.7. The dissipation parameter *k* is set to 0.1, as in Molteni et al. (1993). The coupling parameter α, which introduces nonlinearity in the oscillation's behaviour, is set to α=0.4 (quasi-linear case), or α=0.8 (nonlinear case). Figure 1 shows a time sequence of the two variables *x* and *u* in the two cases. Even in the quasi-linear case, the oscillation undergoes chaotic fluctuations. We shall focus only on the prediction of the predictable component *u*.

A long run of 100000 units is carried out and is output every 0.1 units. In order to simulate the atmospheric analysis errors, a random normal perturbation is added to each variable, with a variance equal to 9% of each variable's variance. The system is next integrated 0.1 time units, and the resulting perturbation simulates the analysis errors. Relative to the initial random error introduced, the "analysis" error has grown along the unstable modes and decayed along the stable modes, which is also the case of errors which undergone the assimilation process (Toth and Kalnay, 1993; Pires et al., 1995). The analysis error has a total variance equal to 10% of the total variance of the system.

In order to perform ensemble average forecasts, we use the "bred growing modes" (BGM) technique designed by Toth and Kalnay (1993), with two sets of parameters; For the first one, the breeding cycle is taken as 0.1 units and the breeding starts 0.8 units before the analysis time, that is 8 breeding cycles are completed, starting from small initial random perturbations with a

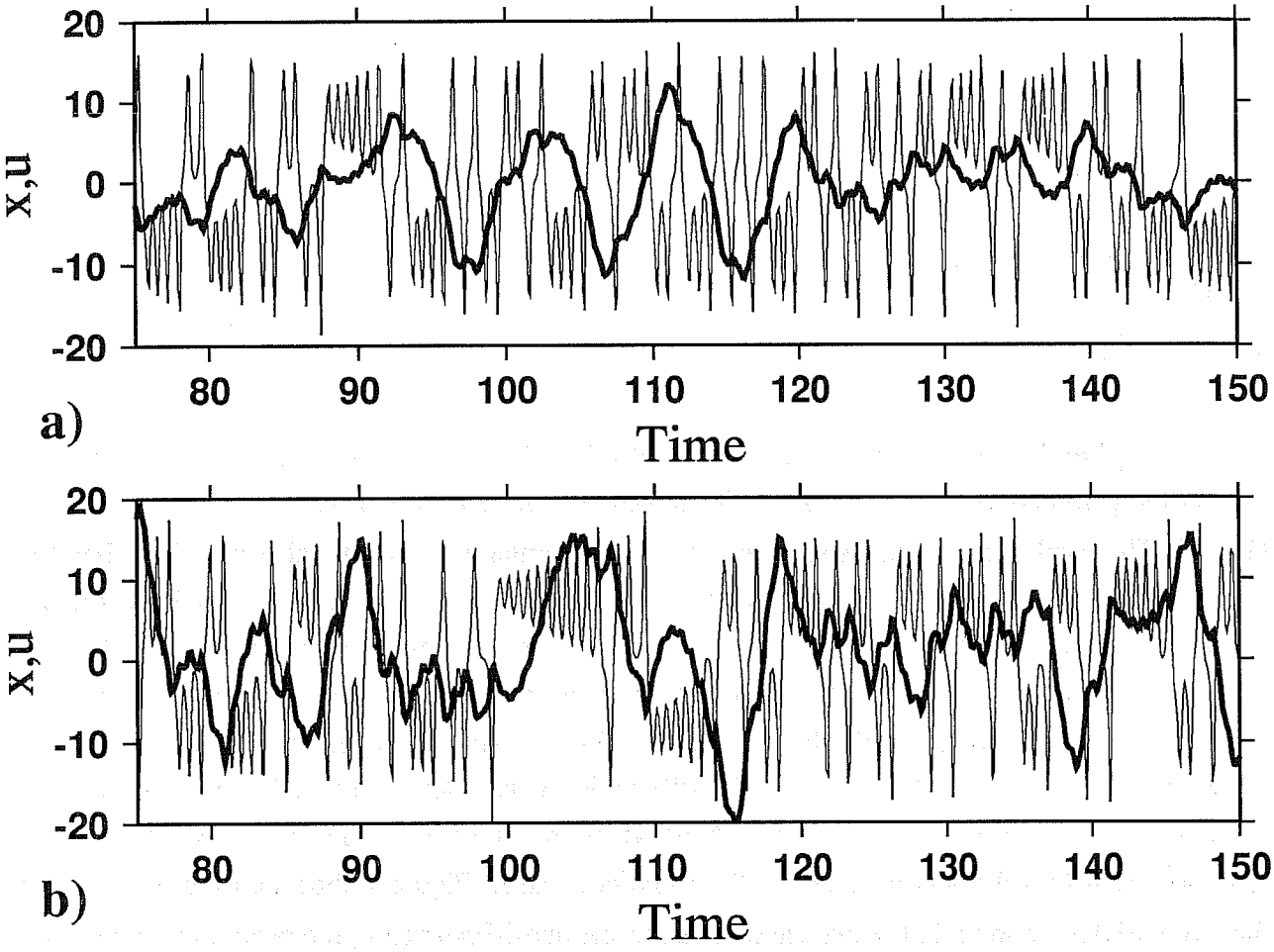

Figure 1: Typical time sequence of the two variables x (light curves) and u (heavy curves) for α=0.4 (panel a) and α=0.8 (panel b).

variance equal to 0.01% of each variable's variance. At the end of the breeding cycle, the perturbation is rescaled to 10% of the total variance (of all variables). In this manner, the initial perturbation PDF is erroneous: it has a too large variance along unstable directions and a too small variance along stable directions: the resulting forecasts will exhibit too much spread. In order to generate perturbations with statistics closer to the analysis error statistics, we generate sets of perturbations along only one breeding cycle, with an initial perturbation size of 9% and the same final rescaling. In this case, the major difference between the perturbation PDF and the analysis error PDF is due to rescaling. Indeed, analysis errors do not have a fixed size. We also finally generate sets of perturbations having a more correct PDF, by simply omitting the rescaling stage of the last experiment. This case will be denoted the "correct error statistics" case (CES).

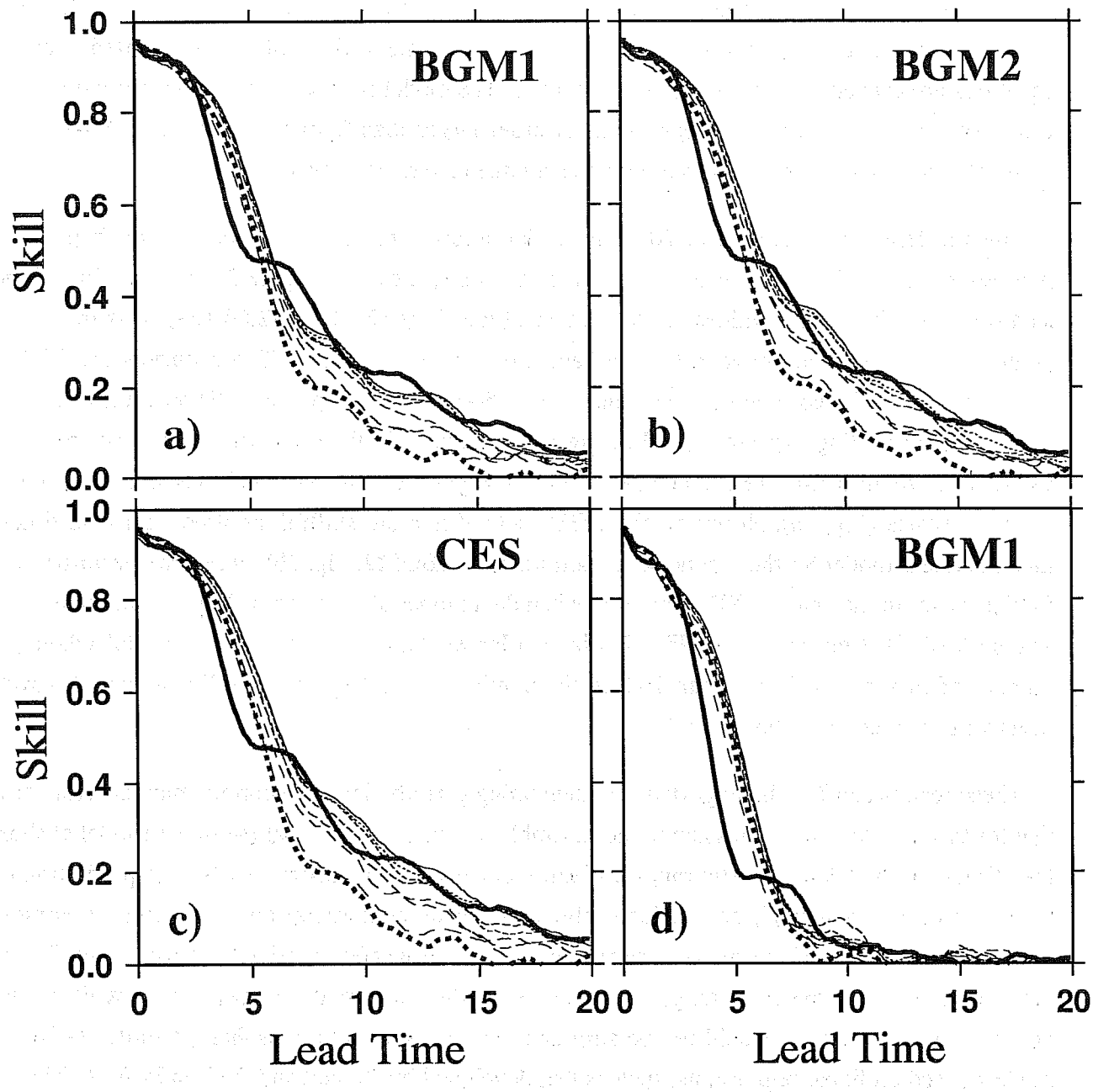

Figure 2: Correlation scores of the forecast of the u variable using the imperfect linear model (heavy curves), the perfect dynamical control forecasts (heavy dashed curves) and various ensemble average forecast techniques with 2, 4, 8, 16, 32, 64 and 128 number of members. The dashes decrease as the number of member increases. a) For the bred growing modes perturbations with 8 breeding cycles and α=0.4; b) same as a) but with 1 breeding cycle; c) CES perturbations (see text); d) Same as a) but for α=0.8.

The simple, imperfect, linear forecast model of the u variable simply consists in omitting the coupling term in the last two equations of (6). Single integrations of the uncoupled u and v equations are carried out from the analysed values. The model is close to the statistical model that would obtain by a simple auto regression, of order larger than 2, of the u variable. Finally the "control experiment" denotes the single forecast issued from the analysis.

The forecasts are extended to 20 units, which allows statistics to be computed from 5000 independent cases. The skill of these various forecasts is shown in Figure 2, as a function of the lead time, and for various values of the number of members (2,4,8,16,32,64,128). It is measured as the time correlation between the true and the forecast u values. These curves are highly significant since the correlation is calculated over 5000 independent cases. Their bumps are not due to undersampling, but are typical of low-order models. In the quasi-linear case, for the first BGM set of perturbations (BGM1) and the control experiment, the skill is lower than that of the linear forecasts (Fig. 2a). However, the BGM method is more skillful, at almost all lead times, than the linear model for the second set of perturbations (BGM2; Fig. 2b), where the perturbations PDF is closer to the analysis PDF, but only when the number of members is larger than about 100. The skill of CES perturbations (Fig. 2c) is also larger than that of the linear model when the number of members is larger than 100. In the nonlinear case (Fig. 2d), the linear model hardly beats all ensemble prediction models.

These results confirm the importance of generating perturbations in a correct manner. Note also that for this simple dynamical system, the ensemble forecasts always beat the linear model at short lead times, in the "deterministic range". Finally, our results emphasize that for the prediction of "predictable components", one major problem is the computational cost; For this dynamical system, the number of members required to beat the imperfect model by only about 5% of correlation skill in the long range is above 100, when the model is perfect as well as the perturbation set. Thus, it would not be surprising that for the real atmosphere, simple statistical models based on linear regressions, such as that developed by Barnett and Preisendorfer (1987) or Barnston (1994), still beat ensemble forecasts for a long time...

2.3 **Hybrid approaches**

The problem is now to take advantage of both forecasting systems. General circulation models have an enormous advantage with respect to statistical models: they can predict the weather where data records are absent or not long enough. Hence the question is whether one can improve the prediction of the future climate by combining both approaches. On areas where data coverage is large, one would expect improvements only if the two model predictions bear different information contents. That is, if statistical and dynamical models have equivalent skill and error statistics, one cannot expect an hybrid approach to be more skillful than the best of the two.

However, taking into account that dynamical models are *de facto* not perfect, some improvements by the combination with statistical models are expected. Roughly speaking, hybrid models would correct the dynamical model errors. Alternatively, statistical models are expected to represent well the past climate, but generally fail to forecast nonstationarities, such as those due to anthropic effects. In this case, dynamical models would help to correct the statistical model defficiencies.

One very simple hybridization technique is borrowed from estimation theory: the *best linear unbiased estimator* (the BLUE; see, e.g. Jazwinski, 1970; Gelb, 1986); It simply consists in calculating the optimal (in the least square sense) linear combination between two estimates of the same quantity. Therefore, if f_1 and f_2 are forecasts of the same quantity g, the BLUE is

$$f = af_1 + (1-a)f_2 \quad , \tag{7}$$

where

$$a = \frac{v_1 - c}{v_1 + v_2 - 2c} \quad , \tag{8}$$

v_1 and v_2 are the respective variances of f_1-g and f_2-g and c is their covariance. The biggest limitation to this approach (as well as other hybrid approaches) is that these latter coefficients must be calculated from a long training series, and therefore one must have at hand many independent dynamical forecast cases. This approach will be tested in Section 3 and 4.

Other hybridization techniques can be developed. For instance, the BLUE technique can be applied at each step of the dynamical forecast, yielding a particular objective *nudging* technique. Nonlinear hybridization methods, borrowed from data assimilation fields are presently tested at LMD. For instance, the statistical model forecasts can be considered as "future observations", and incorporated within a 4-D variational assimilation scheme.

3. PERFECT MODEL EXPERIMENTS

3.1 **The dynamical model**

In this Section, we perform an array of long-range forecast experiments using the 3-level quasi-geostrophic model developed by Marshall and Molteni (1993), to which the reader is referred for technical details. The horizontal resolution is a triangular truncation at total wavenumber 21 (T21). The constant forcing of the model is obtained as the residual of the potential vorticity time derivative equations over the winter months (December through March) using the ECMWF analysis for the period 1983-93. The model is therefore ran under perpectual winter conditions. The model has a fairly realistic climate and variability given its simplicity. Recent analysis of long

simulations revealed the existence of intraseasonal oscillations in the angular momentum (Strong and Vautard, 1995, manuscript in preparation), with realistic maintenance mechanisms. It has also been shown to be relatively successful in simulating extratropical low-frequency variability and phenomena like planetary flow regimes (Marshall and Molteni, 1993; Michelangeli, personnal communication).

A long control simulation of 25000 days has been carried out, and output data are sampled on a daily basis. These data are considered throughout the Section as the "true data". Artificially erroneous analyses were generated by adding a perturbation to the true geopotential height field. This perturbation was obtained by adding a small random error at the beginning of the control run, running the model from the perturbed state and rescaling the perturbation every five days. Once these small perturbations are constructed, two set of erroneous analyses are generated, with fixed geopotential height rms error amplitudes of 10m and 30m (when averaged over the three levels), which are close to the order of magnitude of actual observation errors, by simply rescaling the current perturbation on a daily basis.

Two series of 60-day forecasts are produced by integrating the model from the two sets of erroneous observations. In order to have at hand independent experiments, the successive forecasts are separated by 80 days. Therefore, 310 independent long-range forecasts are produced for the two sets of initial errors. In the following, the experiments conducted with an initial error of 10m will be referred to as the "small error experiments", and the other set as the "large error experiments". Note that we are not considering here ensemble forecasts, but only the equivalent of the "control" experiments of Fig. 2.

The quantity to be forecast here is the the 30-day average of the 50 kPa geopotential height over the Euro-Atlantic sector (80W-40E; 30N-70N: the ATL domain). Therefore, at the end of the forecast stage, 30-day averages are calculated. Moreover, on this time scale, it would be illusory to expect interesting skill from continuous forecasts, and therefore the predictand values are classified into three equally-probable catgeories, at each grid point: Above, below, and near normal. The category separator values are calculated from the "true data". The continuous forecasts are thus transformed into categorical forecasts.

Monthly-mean 50 kPa height tercile forecasts (from any model) are verified against the true tercile, and a 3x3 contingency table is built at each grid point. The measure of skill used here is the categorical LEPS (linear error in probability space) score (Ward and Folland, 1991): Each entry of the contingency table is multiplied by a constant scoring weight, the same as used for terciles by Ward and Folland (1991). The final skill score is obtained, at each grid point x, by summing up all the products and dividing by the score of perfect forecasts. This scoring system is equitable in the sense of Gandin and Murphy (1992), that is, the weigths are such the score is 0 for random or a constant categorical forecasts and 1 for perfect forecasts. 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. The statistical significance of the skill scores is estimated, in a nonparametric way, as in Vautard et al. (1995).

Both for the construction of the statistical model (see Section 3.2) and for the estimation of the dynamical model error statistics, data must be separated into a learning period and a verification period, the first being used for data processing and statistical coefficients tuning, and the second for the verification of models' skill. As we have at our disposal long data series, we use a simplified version of the cross-validation procedure (Tukey, 1958). The control run is divided into two parts, (containing 15000 and 10000 days respectively). The statistical model building is performed on each of the two parts (the learning period) and verified on the other part, which allows noninflated estimation of skill. The global skill is calculated by averaging scores over the two verification periods (with weights according to their respective lengths).

3.2 **The statistical model**

The statistical model is that developed by Vautard et al. (1995) (see also Vautard, 1995). It is based on a two-step procedure. First, "predictable components" are identified, 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.

The first step consists in performing a principal component analysis (PCA) of the 50 kPa geopotential heights in order to obtain *spatial principal components* (S-PCs). 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 10 S-PCs. These filtered ST-PCs achieve a good compromise between predictability and explained variance (Vautard et al., 1995). T is called the window length, and is fixed to 90 days in the present study.

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. They also behave similarly as the "u variable" of the Molteni et al. (1993) model (see Section 2). The extrapolation of the ST-PCs is performed by a linear autoregressive model whose coefficients are tuned from the learning period. These extrapolations run up to 60 days ahead.

The second step is the *specification* (or "downscaling") step. The problem is to estimate, from the forecast ST-PCs, the value at each ATL grid point of the 30-day 50 kPa geopotential field average. Instead of estimating directly the values of the predictand, one estimates crudely its conditional probability distribution (CPD), discretized into the three terciles. The CPD is

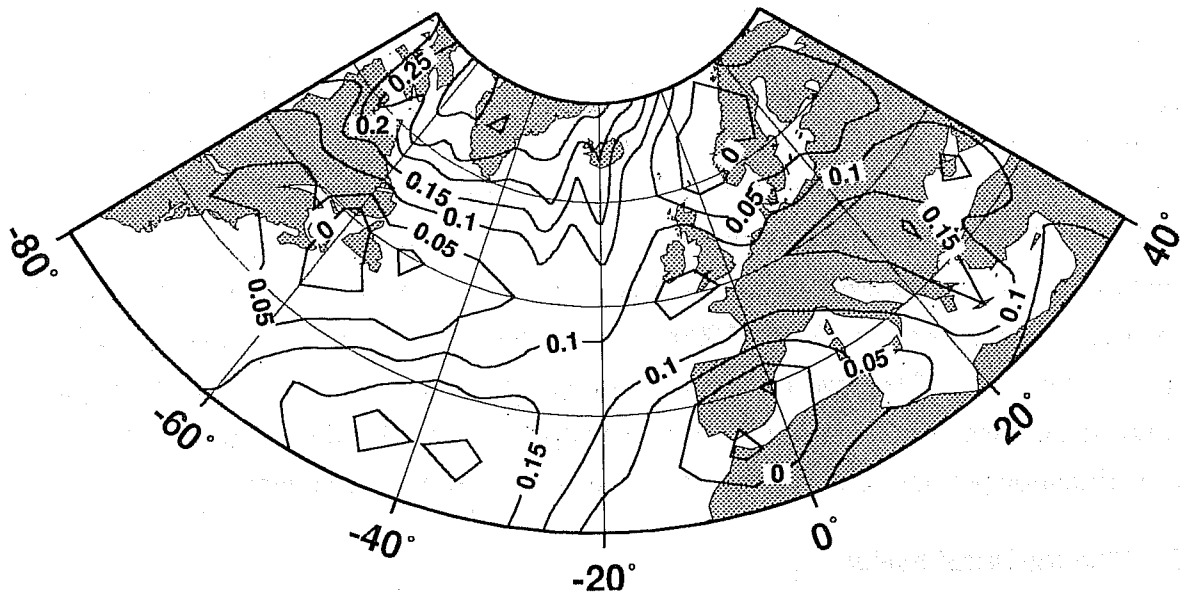

Figure 3: Cross-validated distribution of the LEPS Score of the statistical model applied to the forecast of 30-day 70 kPa averages with a lead time of 15 days. The figure is borrowed from Vautard et al. (1995). The validation is performed over 40 winters (december through march).

calculated by seeking, within the learning period, the 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 to find analogues is the pattern correlation in the vector space spanned by the first 10 ST-PCs. This statistical model is called STA in the following.

This statistical model has been tested by Vautard et al. (1995). They applied it to the forecast of monthly mean 70 kPa heights, and showed that the most predictable feature of this field over the North Atlantic was the so-called North-Atlantic Oscillation (NAO) for winter forecasts. Figure 3 shows the skill score pattern obtained by a cross-validation of this statistical model over 40 years, at a lead time of 15 days, that is for the prediction of the DAY15-DAY45 monthly average. There are three centers of significant predictability: Near greenland, In the Southern part of the domain and over Europe. These centers reflect the predictability of the NAO.

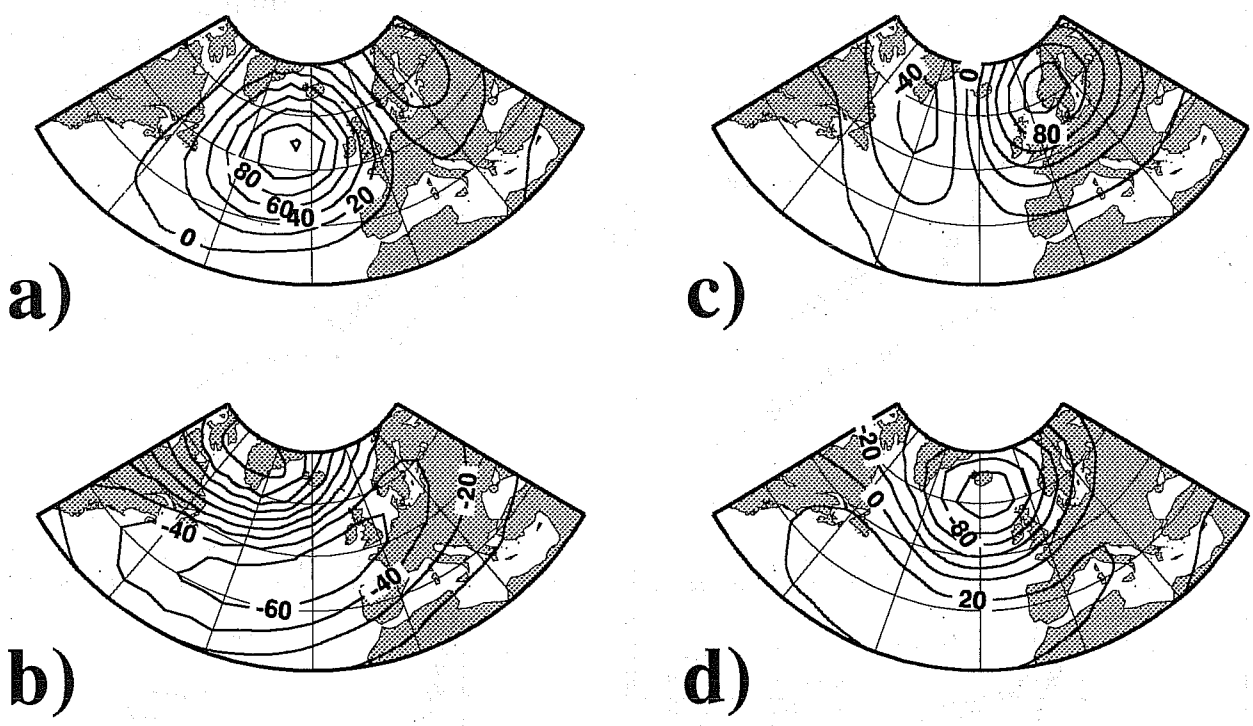

Figure 4: The four Atlantic weather regimes identified by Michelangeli et al. (1995). The patterns displayed are the anomalies (in m) of the 70 kPa geopotential heights associated to the cluster centroids. The figure is borrowed from Vautard et al. (1995).

The model was also tested for the prediction of the occurrences of the four atlantic weather regimes identified by Michelangeli et al. (1995). These latter were obtained from a cluster analysis technique. The centroids of these clusters are displayed in Fig. 4. In this application, the predictand was the the *number of days*, during a forthcoming period of 30 days, of each regime. These variables were also classified into three equally-probable terciles, and forecast in the same manner as the 50 kPa heights. The skill scores obtained for each regime are displayed in Fig. 5, as a function of the lead time, and for various values of the number of forecast ST-PCs used as predictors in the specification stage. Also shown on Fig. 5 are the scores of a persistence model, which simply consists in predicting the same tercile as the latest observed tercile, and another one-step model using only spatial principal components as direct predictors and the same analogue technique as for the specification stage. In all cases, the ST-PC model beats the persistence model, and most of the time (especially at long lead times), beats the S-PC model. The important feature of Fig. 5 is that the regimes of Fig. 4 are not equally predictable. The most predictable regime is the second one, which consists basically in a positive phase of the NAO. Forecasts of the "blocked" regime (regime 3) bear hardly any significance.

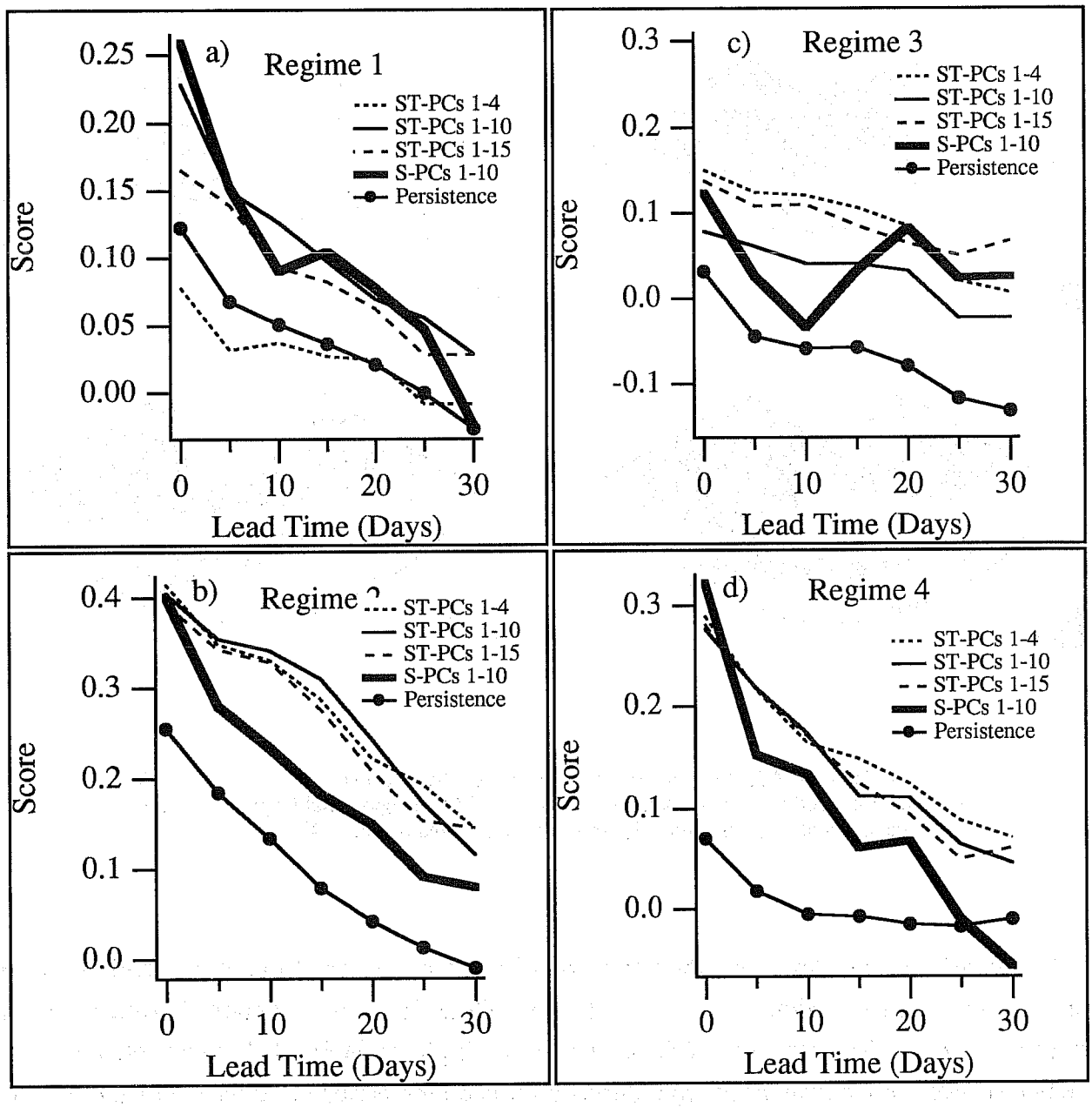

Figure 5: LEPS skill scores of various statistical models applied to the prediction of the forthcoming occurrences of the four weather regimes within time periods of 30 days, as a function of the lead time. The heavy curves correspond to a one-step direct analogue model using the leading 10 spatial PCs as predictors. The curves with circles correspond to the persistence model. The other curves correspond to the two-step ST-PC model described in the text, using various numbers of predictor ST-PCs (see legend on the graph). The figure is borrowed from Vautard et al. (1995).

3.3 Two 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 the calculation of the ST-PCs using the 50 kPa height field forecast by the dynamical model, at the same lead time as for the statistical model, and then to apply the specification stage exactly as before, using the analogue method. This is the first, simple hybrid statistical-dynamical model, called hereafter HYB1.

The second hybrid model uses the BLUE procedure applied onto extrapolated ST-PCs: Each couple of ST-PCs, extrapolated by the linear regression and by the dynamical model, are combined using Equations 7 and 8, yielding "hybrid ST-PCs". Once again, these are used as predictors for the specification stage which is the same as before. This second hybrid approach is called HYB2 in the following.

3.4 Skill of the models

Going back to the quasi-geostrophic model experiments, the global LEPS scores of all models are displayed on Fig. 6 as a function of the lead time, for the two initial error cases. The DYN model produces good forecasts at a short lead time, but its skill decreases dramatically faster than the skill of the other models. This strongly indicates that when only a few filtered predictable components are considered and then "downscaled" to the target predictand, there is a significant increase in skill relative to direct estimation of terciles from model integrations. In fact, for both initial error amplitude cases, the DYN model loses any skill once the celebrated "deterministic period" of about 10-15 days is not contained within the predictand averaging period. The amplitude of the initial error essentially affects the skill of the forecast for short lead times, where there is some skill left. Note that the skill of the dynamical model is perhaps underestimated in the long run, relative to its counterpart for the real atmosphere since the QG model does not take into account any possible sources of skill coming from persistence or slow variations of boundary conditions.

The STA model is fairly insensitive to the initial error amplitude, and beats marginally the HYB1 model for lead times greater than about 15 days, independently of the error size. As shown by the 10-90% confidence interval, this difference is however nonsignificant. The HYB2 model almost always performs better than the best of STA and HYB1. Once again, this improvement is marginal, but systematic. Note finally the important skill increase between the original DYN model and the most sophisticated HYB2 model.

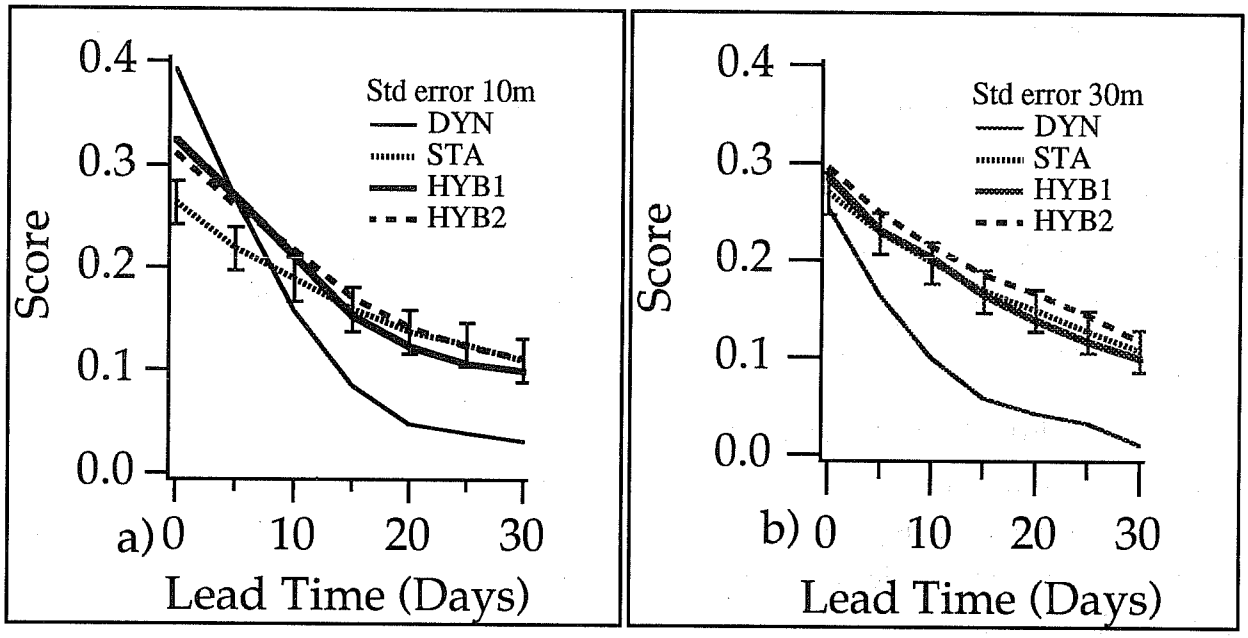

Figure 6: Global LEPS scores of the 50 kPa height tercile forecasts in the ATL region using the DYN, STA, HYB1 and HYB2 models. See legend on the graph. a) Small error case; b) Large error case. The error bars represent the 10-90% confidence interval associated to the STA model. The figure is borrowed from Pires et al. (1995).

The geographical distribution of LEPS score are presented in Fig. 7 for a lead time of 15 days, for the four models, and for the small error case. Regions of relatively high skill of the DYN forecasts are rare. The threshold value for which skill scores are significant at the 90% level is about 0.05. Regions of significant skill are therefore located over the Southwestern and Northeastern parts of the ATL domain. All the ST-PC-based forecasts display a similar skill pattern, with significant skill in the mid-Atlantic and near polar regions. This skill pattern is reminiscent of the skill pattern of Fig. 3 for the real atmosphere. The distortion observed in the skill pattern relative to that figure is consistent with the distortion between the NAO simulated by the QG model (as identified, for instance, by the first ATL EOF) and the actual NAO pattern. The fact that HYB1 and STA have almost similar score patterns confirms that their errors are highly correlated, if not almost equal, in which case no major skill increase can be expected from the BLUE procedure of the HYB2 model.

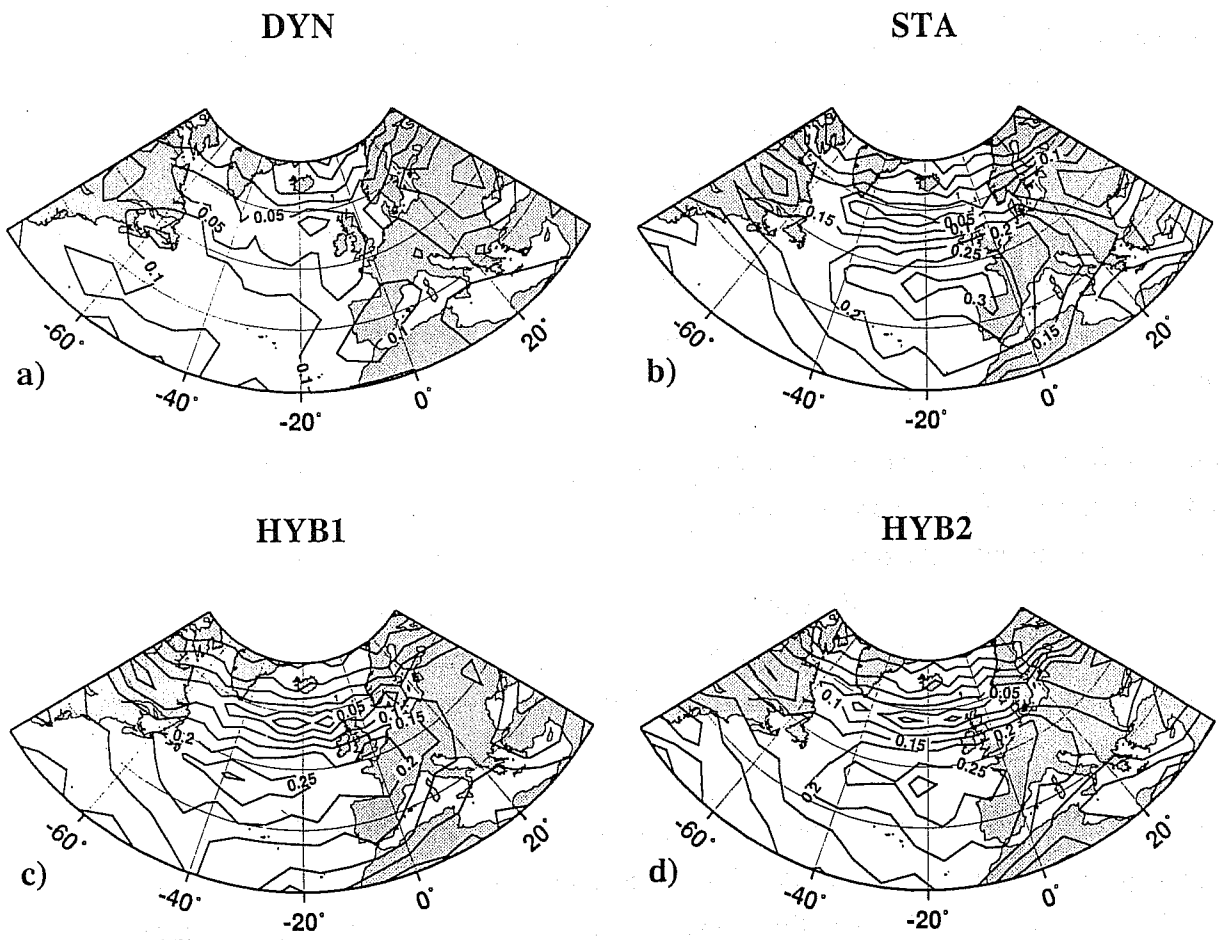

Figure 7: Geographical distribution of the LEPS score for the various model forecasts, with a lead-time of 15 days. a) DYN. b) STA. c) HYB1. d) HYB2. After Pires et al. (1995).

4. REAL ATMOSPHERE EXPERIMENTS

4.1 **Experimental design**

In this Section, we compare the skill of the statistical model described in Section 3.2 with the skill of 44-day forecasts issued at METEO-FRANCE (Déqué and Royer, 1992) from the EMERAUDE model (the "DYN" model), which was the former operational weather forecast model used in France. The experimental design is constrained by the EMERAUDE experiments which were carried out before 1992.

For this application, the EMERAUDE model is a spectral T42 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 ran up to 44 days (D+44) starting from 40 independent dates occurring in the winter season. The starting dates lie near 15 October, 15 November, 15 December and 15 January of each year for the cold seasons 1983-1984 to 1992-1993. Each forecast consists of a set of 5 lagged integrations starting with ECMWF analyses of days D-2 to D with a 12 hours time step. We use the average of the ensemble forecast, which provides slightly improved skill, as mentioned by Déqué and Royer (1992), than individual forecasts. The predictand is, as in Section 3, the 30-day average 50 kPa geopotential height over the ATL domain. The ensemble-average forecasts are also ouput into a categorical format, as above, by placing each 30-day grid point average into its respective tercile.

Since all forecasts are performed over a ten-year period, this period is taken as the verification period while the 1950-1983 period is used for training the statistical model. The tercile separators, for instance, are calculated from the learning period. However, the hybrid models construction, as well as the removal of systematic errors (see below), require 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 forecast is verified against the analysis, as in Déqué and Royer (1992).

The statistical model is the same as above, with differences due to the seasonal character of real-atmosphere data: Before any processing, the annual cycle is removed from the analyses of the 50 kPa height. As above, the statistical model is called STA for simplicity. The hybrid models HYB1 and HYB2 are also defined as in the previous Section, and tested.

4.2 Systematic errors, trends and skill biases

The systematic error of the dynamical model is large. Figure 8 shows the systematic error of 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. Therefore, one expects skill increases by removing the systematic error. This can be done by calculating it in cross-validation, for a given winter, from the 36 forecast cases issued during the 9 other winters. The resulting forecasts are denoted DYN-S, and will also be tested. The 50 kPa heights exhibit some climate differences between learning (1950-1983) and verification periods (1983-1993). Therefore, the statistical model also has a "systematic error". It is possible to remove this error in the same manner as for the dynamical model (STA-S).

In both cases, the use of data posterior to the forecast can inflate the skill, due to nonstationarities in the predictand. This is particularly true when using the LEPS skill score. In particular, three difficulties arise is assessing the significance of the LEPS skill score. The first one is due to the bias of the LEPS skill score when verification terciles are not balanced, i.e. do

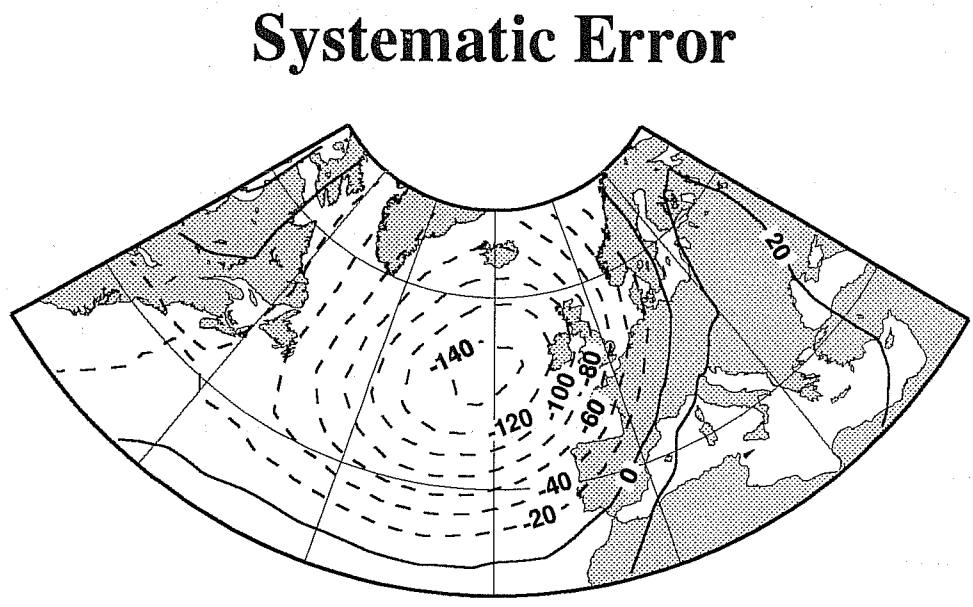

Figure 8: Systematic 50kPa error of the dynamical model for forecasts of 30-day averages at a lead time of 14 days. After Sarda et al. (1995).

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, 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 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 all these peculiarities. We choose, here, to assess the statistical significance by a simple test that takes into account the above problems.

The statistical significance test used here compares the measured LEPS score with that of a random forecaster predicting the same tercile frequencies as the model to be tested. These tests are therefore 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 of the 1000 LEPS score values to 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 forecast trends, has a vanishing unbiased score. Note also that, in the presence of trends, the unbiased score of perfect forecasts is different than 1. Finally, a non-random model able to forecast the trend, but not the variability around this trend has a vanishing unbiased score also. 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 forecast trends.

5.3 **Models skill**

The global (biased) skill scores of the models DYN and STA are displayed in Fig. 9a as a function of lead-time. The DYN forecasts have clearly higher skill at short 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 used, the skill score of the STA model turns out to be higher at long lead times. When the score of STA is calculated from all winter days (november through march), instead of only the 40 independent cases, it decreases monotonically with lead time, as expected (see Fig. 9a). The 40 cases selected turn out to have below-average score. Also shown on Fig. 9a is the score of the single-member DYN forecasts, which is slightly lower than that of the average DYN 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 null hypothesis of a random forecaster "knowing" the trends or the climate difference between learning and verification periods. The STA score, when calculated from the 40 cases, is only marginally significant at some lead times (actually at the longest lead times). The confidence limits of 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, and (ii) LEPS scores reward (resp. penalizes) more correct (resp. erroneous) extreme tercile forecasts than correct near-normal forecasts. Note finally that the persistence model has, as anticipated in Section 5.2, a nonsignificant score. This confirms the severity of our significance test.

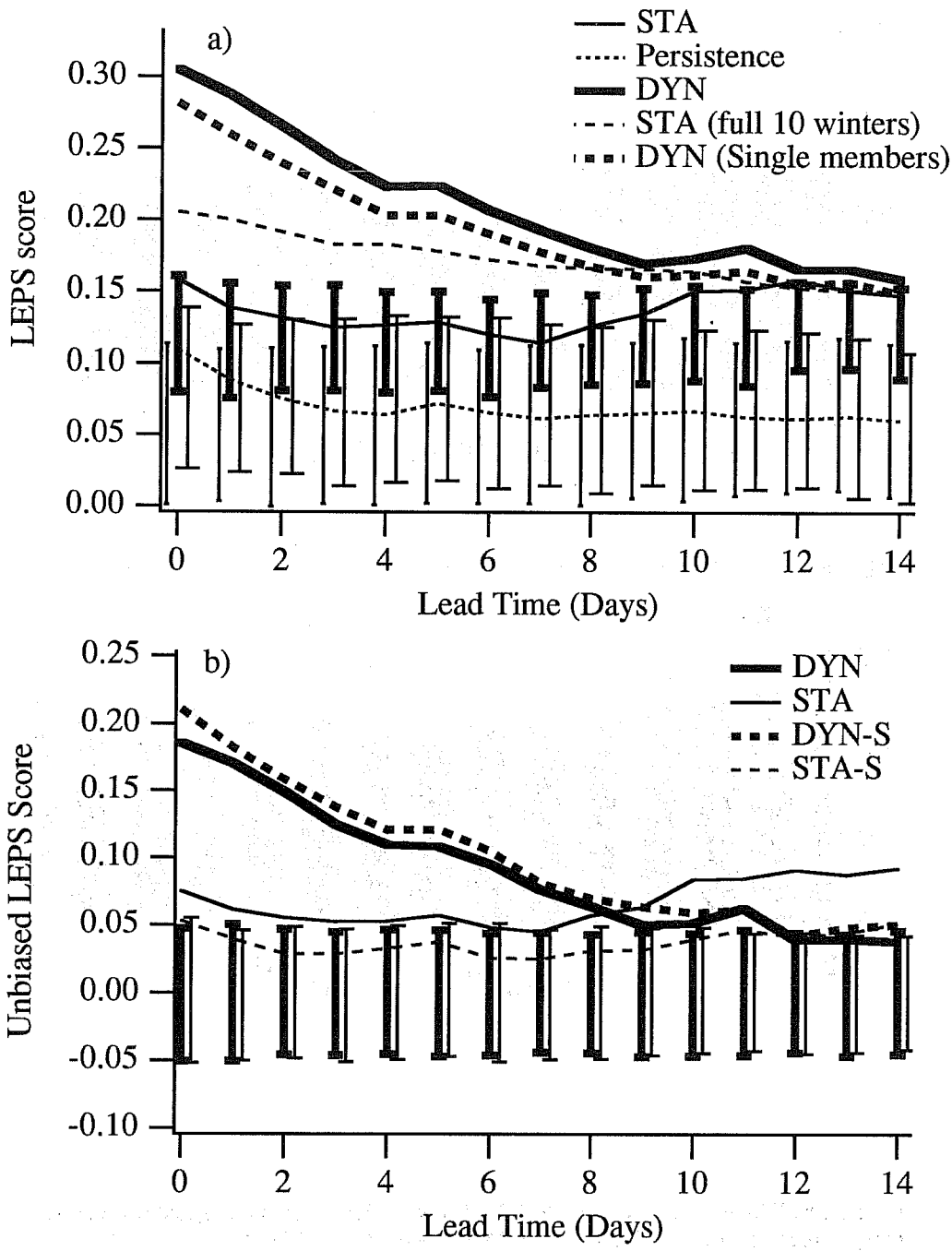

Figure 9: Global LEPS scores of the DYN, STA, and persistence models, versus lead-time, for the forecasts of the 50 kPa heights. a) Raw forecasts for the DYN, STA, persistence, and individual-members of the dynamical model. Heavy error bars represent the 10%-90% confidence interval for the random forecaster 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 (resp. light) error bars are associated to the DYN-S (resp. STA-S) model. After Sarda et al. (1995).

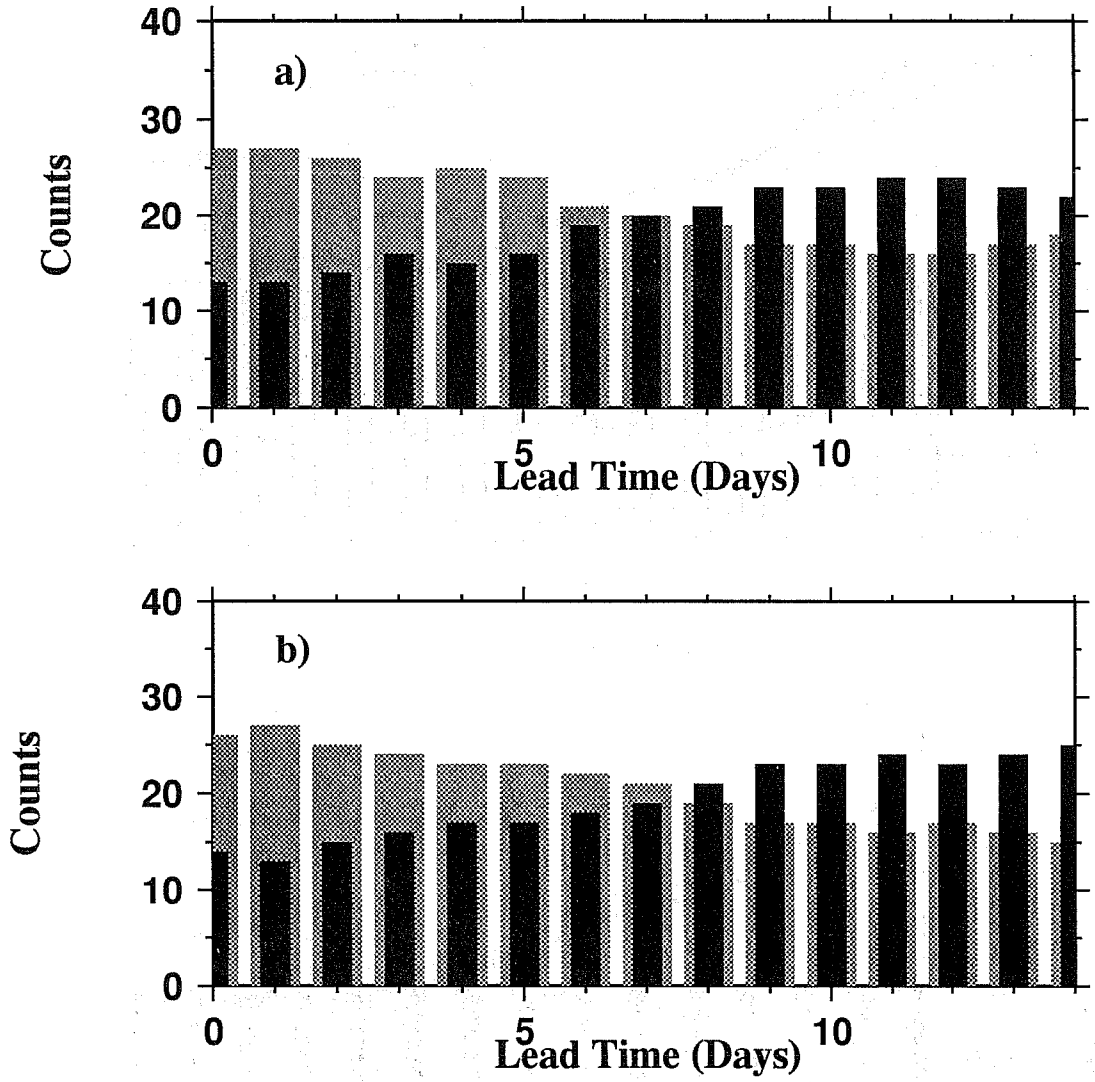

Figure 10: a)Number of cases for which DYN beats STA (shaded bars), and reverse (solid bars), versus lead-time. b)The same for DYN-S and STA-S. After Sarda et al. (1995).

Fig. 9b shows the unbiased scores 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, as observed for the quasi-geostrophic perfect model. A comparison of the difference between the two curves with the size of the confidence intervals shown in Fig. 9a indicates, however, that the significance of this difference cannot be established from our experiments. Figure 9b 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 nonsignificant at the 90% confidence level.

The variations of the respective skills of the DYN and STA models with the lead time are also illustrated on Fig.10a (resp. Fig. 10b 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 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 the the DYN (resp. DYN-S) forecasts, in terms of number of successes.

As explained in Section 5.2, the procedure used to remove score biases also removes the part of skill due to models' ability to forecast trends or climate differences between learning and verification periods. Actually, the STA model does exhibit some skill in forecasting the trend. This can be checked by looking at the frequency of STA forecasts in extreme terciles (not shown). There is in fact a clear tendency of the STA model to forecast more frequently extreme terciles that occur more frequently during the verification period. However, STA forecasts are carried out using only learning-period analogues for which these frequencies are balanced.

Local skill scores may be hardly seen as significant in a 40-case study. The results presented below confirm this point. Figs. 11a-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 (resp. lowest) 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. When the systematic error of DYN is removed, the score pattern still exhibits a large area of negative values over the Eastern Atlantic, but its peaks are now centered precisely where the trends are highest. The STA model (Fig. 11c) 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. 11d shows the score map obtained from forecasts over all winter days of the verification period. The resulting map strongly resemble the score map of Fig. 3, again corresponding to the higher predictability of the NAO. These maps also show that, unlike for the quasi geostrophic perfect model, the distribution of skill of the STA and DYN models are qualitatively different. One should therefore expect to gain some skill from the

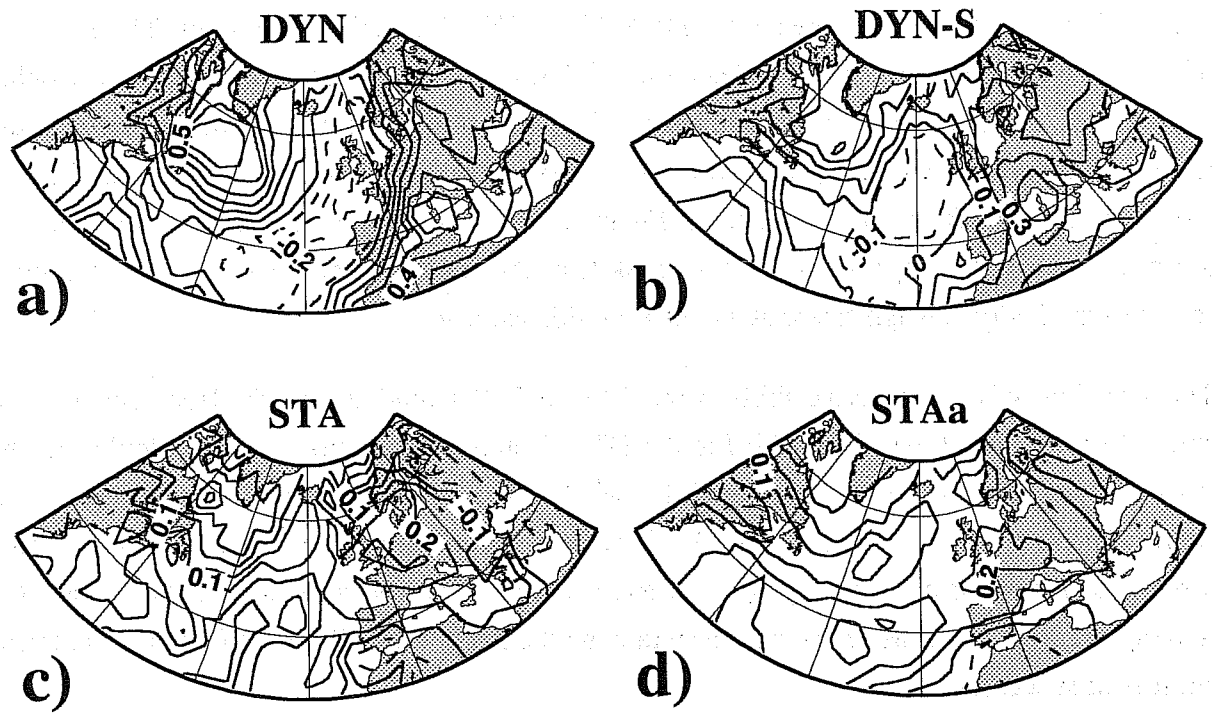

Figure 11: Geographical distribution of the (biased) LEPS score for the DYN (a), DYN-S (b), and STA (c) models. In panel d), the score of the STA model, calculated from all possible predictand days within the cold season along the verification period (10 winters) is represented (STAa). Contour interval is 0.1. After Sarda et al. (1995).

combination of the two forecasts, which is the goal of the hybrid models. 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, in a more general sense.

Figure 12a shows the (biased) 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 3 days does DYN perform better than HYB1. The score of HYB2 forecasts is smaller than that of HYB1, presumably due to undersampling problems. The two hybrid models have scores significant at the 90% confidence level, using the same testing procedure as above. When unbiased scores are considered (figure 12b), 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 unbiased skill (HYB2). By comparing the size of confidence intervals (Fig. 12a) with the difference between the DYN and HYB2 curves, one notices that the HYB2 procedure significantly improves the skill of the original dynamical model, while the difference between HYB1, HYB2 and STA scores is nonsignificant. All these curves show a qualitatively similar behaviour as for the perfect-model experiments presented in Section 3.

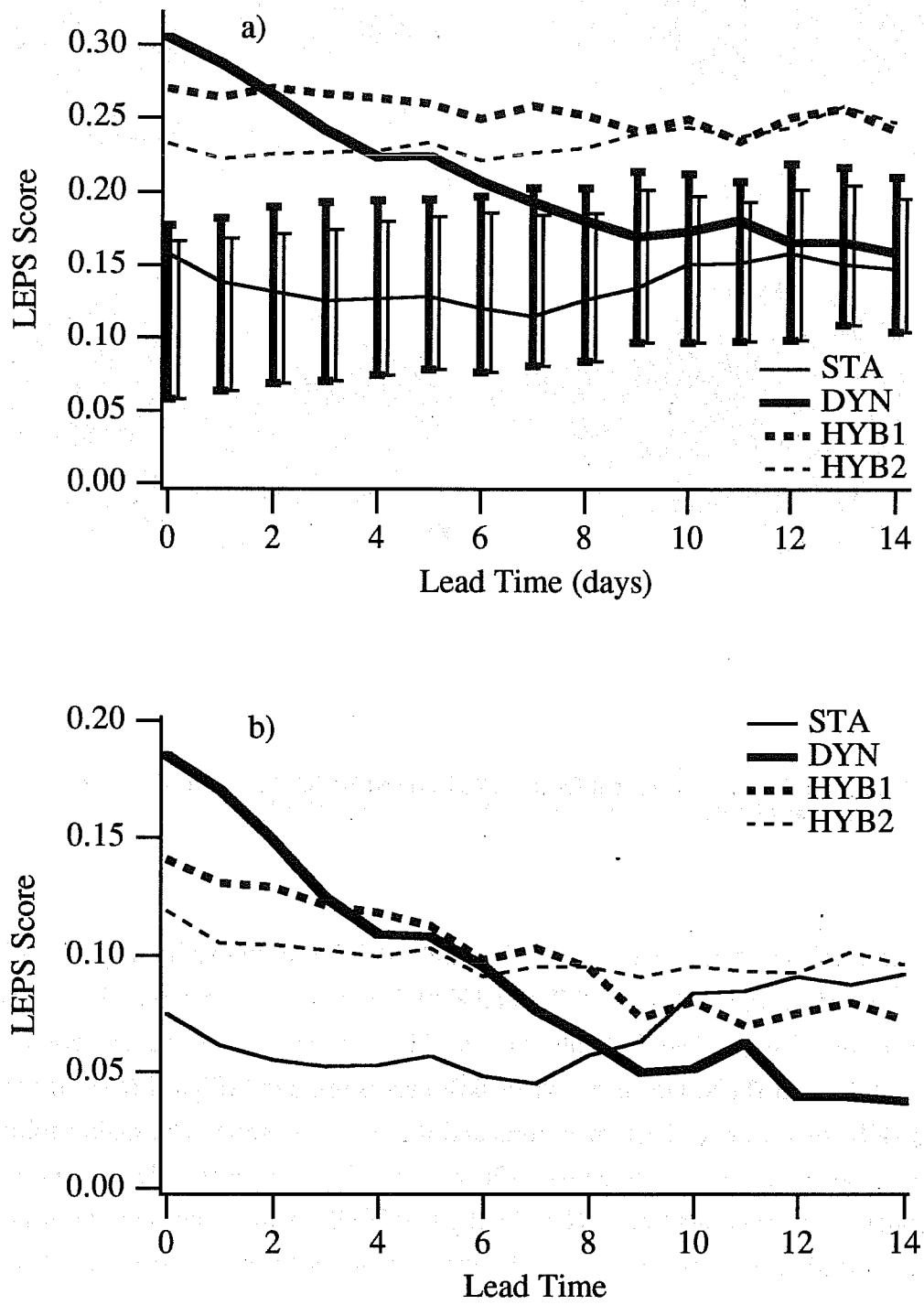

Figure 12: a) Same as in Fig. 9 (biased LEPS score), for the STA, DYN, HYB1 and HYB2 models (see legend on the graph). Heavy error bars stand for the HYB1 model and kight error bars stand for the HYB2 model. b) Same as Fig. 10a, but for the unbiased LEPS scores. After Sarda et al. (1995).

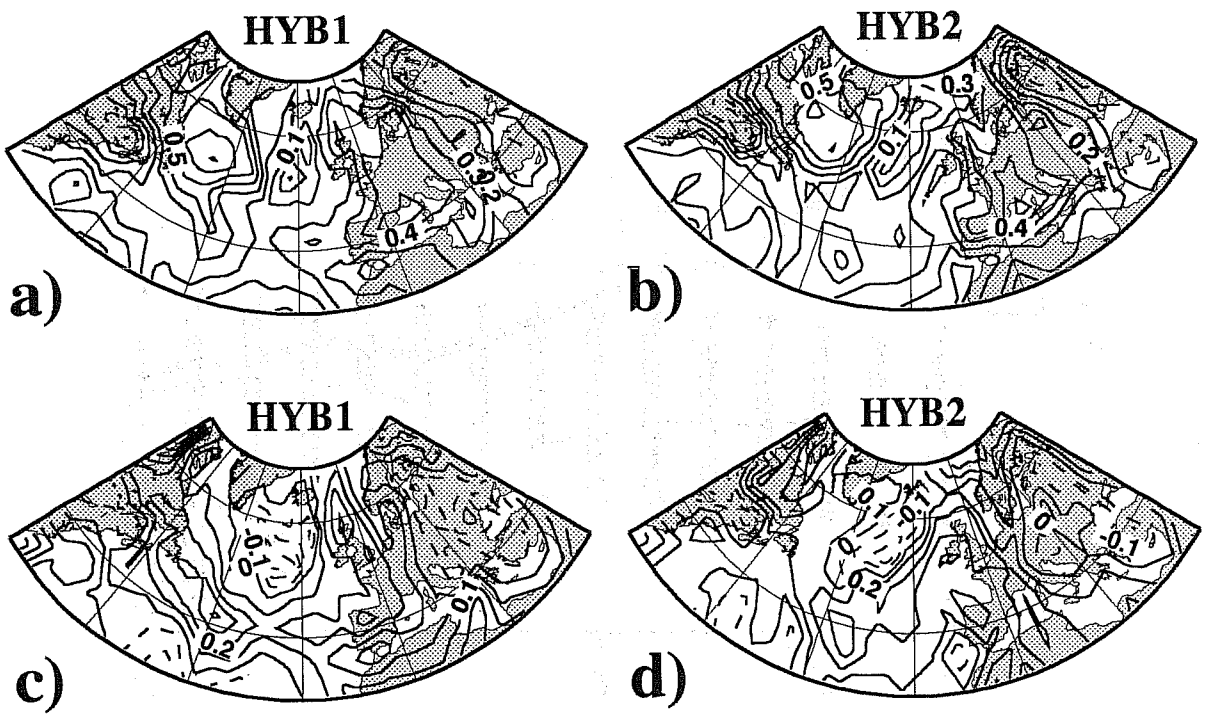

Figure 13: Same as Fig. 11, but for the HYB1 (a) and HYB2 (b) biased scores, and the HYB1 (c) and HYB2 (d) unbiased scores. After Sarda et al. (1995).

The biased/unbiased score maps of HYB1 and HYB2 are shown in Figs 13a-d. 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. 13c-d), there are areas of negative skill, such as over Eastern Europe and the North Atlantic. The main problem is that grid-point skill scores are generally nonsignificant at the 90% confidence level, and it is hard to draw any conclusions from this figure. Note finally that HYB1, which does not use any data in the verification period, does forecast correctly the climate differences between learning and verification (not shown).

Figure 14 recapitulates the skill of the various models, for a lead time of 14 days. Both in terms of score and significance, the HYB2 model performs best. All models but the persistence model have a significant score at the 90% confidence level, but only the statistical and hybrid models pass the test at the 95% confidence level.

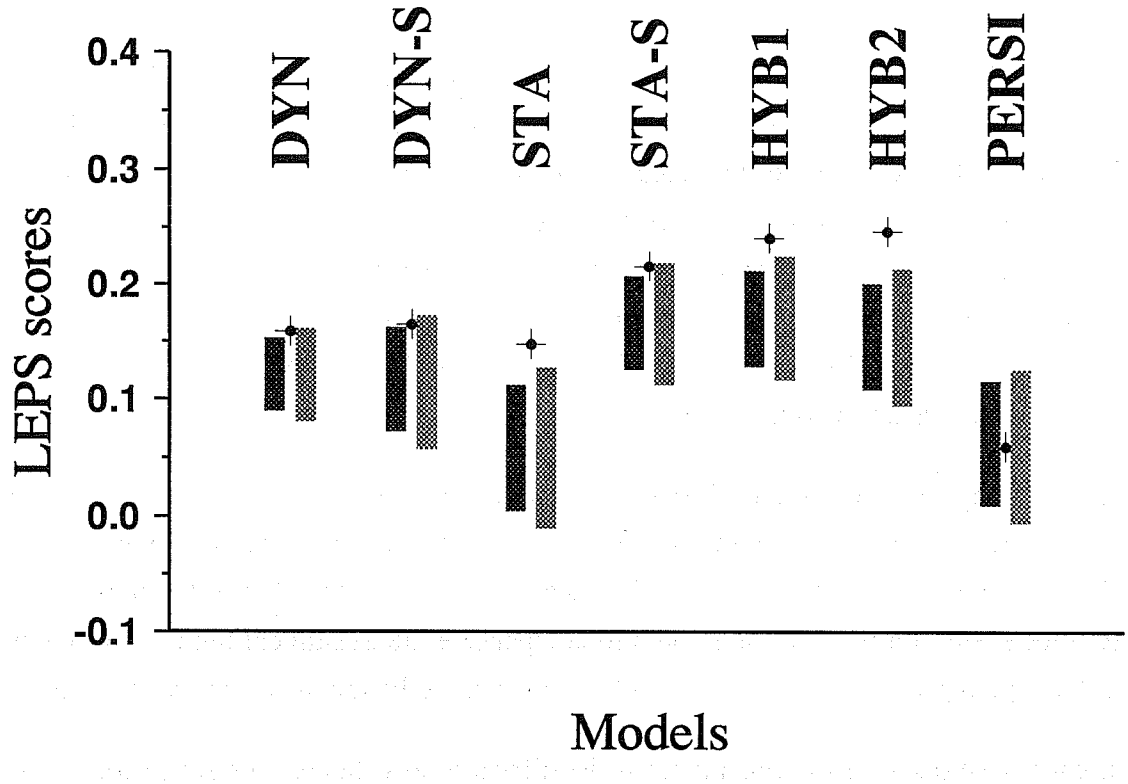

Figure 14: Global LEPS scores of the various models at a lead time of 14 days (Heavy dots with crosses), together with the 10%-90% (heavy shaded bars) and 5%-95% (light shaded bars) confidence limits of their associated testing random forecaster.

6. DISCUSSION

The whole content of this article demonstrates that the advantage of using large numerical models for long-range prediction, with respect to less computationally expensive statistical models is not obvious. The basic reasons are:

(i) Dynamical model errors still dominate the forecast errors for long lead times.

(ii) In some cases, even with a perfect model, one would have to know as accurately as possible the probability distribution of initial errors quite accurately in order to produce ensemble-average forecasts that would beat significantly simple, linear and therefore imperfect models.

(iii) Even when an accurate knowledge of the analysis error statistics is at hand, an important skill increase would probably require many perturbed integrations for the same forecast. We

exhibited a simple dynamical system for which at least 100 ensemble members are needed in order to beat a simpler imperfect model in that case.

However, dynamical models have important capabilities not shared by statistical models. These are the possibility of producing forecasts in areas of poor data coverage, and also the ability to predict climate changes. For the above reasons, it may be desirable to develop hybrid schemes that bear the positive aspects of the two approaches. We proposed here several ways to construct hybrid models. The one achieving the best ratio success/complexity consists in isolating predictable components (such as identified by the space-time principal components), extrapolating them with the dynamical model and finally "downscaling" them to the target predictand using simple schemes such as analogue schemes.

Another suggestion is to use classical estimation techniques as the "best linear unbiased estimator" (BLUE) in order to combine the extrapolations of the above predictable components. This last method does not appear to be significantly more successful than the previous one in a perfect model environment, but may help to compensate the dynamical model's deficiencies. In our real-world experiments, this second method turns out to be the most successful at long leads.

Another issue discussed in this article is the difficulty to validate long-range forecast models. Indeed, due to the low values of the scores, many independent forecast cases are required in order to exhibit any significant skill. We believe that our 40-case study (Section 5) is not sufficient. A relevant question is whether computational resources allow today any serious validation experiments.

Acknowledgements.

The permanence in France of Carlos Pires was supported in France by the grant BD/2023/92 of JNICT (Junta Nacional de Investigacao Cientifica e Tecnologica) - Portugal. This study was 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., 1994: Linear statistical short-term climate predictive skill in the Northern Hemisphere. *J. Climate,* 7, 1513-1564.

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, 113-144.

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

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

Gelb, A. (ed.), 1986: *Applied Optimal Estimation.* MIT Press, 374 pp.

Houtekamer, P., J. Derome, 1995: Methods for Ensemble Prediction. *Mon. Wea. Rev.*, **123**, 2181-2196.

Jazwinski, A. H., 1970: *Stochastic Processes and Filtering Theory.* Academic Press, 376 pp.

Lönnberg, P., and A. Hollingsworth, 1986: The statistical structure of short-range forecast errors as determined from radiosonde data. Part II: The covariance of height and wind errors. *Tellus*, **38A**, 137-161.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.*, **20**, 130-141.

Marshall, J., and F. Molteni, 1993: Toward a Dynamical Understanding of Planetary-Scale Flow Regimes. *J. Atmos. Sci.*, **50**, 1792-1818.

Michelangeli, P.-A., Vautard, R. and Legras B. 1995. Weather regimes : recurrence and quasi-stationarity. *J. Atmos. Sci.*, **52**, 1237-1256.

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, 205-245.

Molteni, F., L. Ferranti, T. N. Palmer, and P. Viterbo, 1993: A dynamical interpretation of the global response to equatorial Pacific SST anomalies. *J. Climate,* **6**, 777-795.

Molteni, F.,.and T. N. Palmer, 1993: Predictability and finite-time instability of the northern winter circulation. *Quart. J. Roy. Meteor. Soc.*, **119**, 269-298.

Palmer, T., F. Molteni, R. Mureau, R. Buizza, P. Chapelet, and J. Tribbia, 1993: Ensemble prediction. *Proc. ECMWF Seminar*, Vol. 1, Reading, U. K., ECMWF, 21-66.

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.

Pires, C., R. Vautard and O. Talagrand, 1995: On extending the limits of variational assimilation in nonlinear chaotic systems. *Tellus.* In press.

Pires, C., R. Vautard, J. Sarda and G. Plaut, 1995: Statistical, dynamical and hybrid long-range atmospheric forecasts. Part I: Perfect model experiments. *Tellus*, submitted.

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

Sarda, J., G.Plaut, C. Pires, R. Vautard, 1995: Statistical, dynamical and hybrid long-range atmospheric forecasts: Part II: Real-atmosphere experiments. *Tellus*, submitted.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 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., 1994: Long-range weather forecasts through numerical and empirical methods. *Dyn. Atmos. Oceans,* **20**, 247-270.

Vautard, R., C. Pires, G. Plaut, 1995: Long-Range atmospheric predictability using space-time principal components. *Mon. Wea. Rev.* In press.

Vautard, R., 1995: Patterns in time: SSA and MSSA. In *Analysis of Climate Variability*. Eds. H. von Storch and A. Navarra, Springer, Berlin, pp. 259-279.

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.