Tellus (2001), 53A, 500–512
Printed in UK. All rights reserved

TELLUS
ISSN 0280–6495

Seasonal predictability of tropical rainfall: probabilistic formulation and validation

By MICHEL DÉQUÉ*, *Météo-France, Centre National de Recherches Météorologiques, 42 av. Coriolis, F-31057 Toulouse Cedex 1, France*

(Manuscript received 6 October 2000; in final form 5 March 2001)

ABSTRACT

Two idealized seasonal forecast experiments are performed by prescribing monthly observed SSTs to atmospheric GCMs. The first one uses 3 different models, each with 9 individual forecasts (PROVOST experiment). The second one uses an improved version of one of the 3 models and larger ensembles consisting of 120 members. Both experiments show that forecast scores are maximum in the tropics during winter and during summer. The relatively high correlations in the tropics (0.4 to 0.7) imply, however, that the forecasts explain less than 50% of the variance of the observations. The raw probabilistic forecasts obtained by the empirical probability distribution of the forecast members exhibit very little skill, when evaluated by a euclidian distance versus the climatological forecast. The lack of reliability can be partly corrected by a simple statistical adaptation. Moreover, when the skill is evaluated by an economical value in a cost/loss approach, the model forecasts are more efficient than the climatological forecast. A more realistic evaluation of the probabilistic skill is obtained by replacing observed by statistically predicted SSTs. A simple but efficient method is used, which lets each member of the ensemble develop its own SST anomalies. Although lower, skill is significant in the tropics.

1. Introduction

The development of coupled ocean–atmosphere models, and progress in ocean data assimilation have recently led to substantial interest in numerical seasonal prediction (Palmer and Anderson, 1994; Shukla, 1998, and many other papers which have been devoted to the subject but cannot be quoted here). Earlier studies using specified observed sea surface temperatures (Blackmon et al., 1983; Shukla and Wallace, 1983; Lau, 1985; Owen and Palmer, 1987) have demonstrated potential predictability during strong ENSO years. The European Union PROVOST (PRediction Of climate Variations On Seasonal-to-interannual Timescales) project aimed at generalizing these results by using systematic hindcasts for the 1979–93 period (Brankovic and Palmer, 2000). The seasonal prediction results presented here involve three different models, which makes the conclusions little model-dependent. Predictive skill is found in both the tropics and extratropics, but is maximum in the former region. This is easy to understand as the tropical SST shows large and persistent anomalies which are able to generate a coherent response in the atmospheric model.

We concentrate in this paper on the precipitation predictive skill. Indeed, seasonal precipitation forecasts are more useful than temperature forecasts in the tropics through their impact on agriculture and hydrology. Section 2 presents the experimental setup, and Section 3 the mean scores. Even though the scores are relatively high compared to other regions or other parameters, they are too low to allow a fully confident use of the forecasts. The probabilistic approach is studied in Section 4 and a proper way of scoring is examined.

* email: deque@meteo.fr

All the previous results refer to potential predictability, as the SST is assumed to be perfectly predicted. The European Union DEMETER (Development of a European Multi-model Ensemble system for seasonal to inTERannual prediction) project is a new attempt to strengthen or weaken the PROVOST results by using coupled ocean–atmosphere models. In Section 5, we evaluate the forecast skill when the SST is predicted using a simple statistical scheme based on persistence. Conclusions are given in Section 6.

2. Experiments

2.1. PROVOST

The PROVOST project is based on the ECMWF reanalysis (Gibson et al., 1997) which provides intial atmospheric conditions, SST boundary conditions, and verification data for some fields. This reanalysis covers 1979 through 1993 and is referred to as ERA15 hereafter. Four centers participated in the hindcast exercise: the European Centre for Medium-range Weather Forecasts as a coordinator, the UK Meteorological Office, Météo-France (the French meteorological service), and Electricité De France (the French electricity company). The four partners and their experiments will be referred to as ECMWF, UKMO, MF and EDF respectively. The last two partners used the same spectral model, ARPEGE1, but with two horizontal resolutions, namely T42 and T63. ECMWF used a T63 horizontal resolution of its spectral model IFS (cycle13), and the horizontal resolution used by the UKMO grid point model corresponds rather to a T42 resolution. The ERA15 period (1979–1993) allows production of 15 years of forecasts (only 14 winters are available in the case of ECMWF). Forecasts for 4 seasons were produced (only winter in the case of EDF forecasts), but we limit the present study to winter and summer. Each forecast extends up to 4 months: we restrict the study to the average over the last three months: JFM, corresponding to winter, and JAS, corresponding to summer. Each forecast consists of 9 individual integrations starting from ERA15 conditions lagged by 24 h.

We consider the 36 (or 27 when only 3 models are available) individual integrations as a multi-model. Anomalies are calculated by subtracting the climatology of the corresponding model based on the 14 or 13 other years. This is necessary because the models do not have the same systematic errors (except MF and EDF), see Doblas-Reyes et al. (2000) or Palmer et al. (2000).

2.2. PROVOST revisited

A second series of seasonal hindcasts has been recently produced with a new version of the ARPEGE model. This version (numbered 3 at Météo-France) includes 5 years of model development since the version of ARPEGE, used in PROVOST (numbered 1). The models share the same name and parts of code, but the scientific calculations are very different. A description of ARPEGE1 can be found in Déqué and Piedelievre (1995). ARPEGE3 uses a semi-lagrangian advection (ARPEGE1 uses an eulerian one), a two time-level discretization (ARPEGE1 uses a leap frog scheme). The spectral truncation is T63 (T42 in ARPEGE1), the 31 vertical levels are those of ERA15 (in ARPEGE1, 20 levels out of the 31 were in the stratosphere) and the time step is 30 min (15 min in ARPEGE1). If we except the convection scheme (Bougeault, 1985), which has undergone only minor changes, all other physical parameterizations have been modified or replaced. The radiation is now the Morcrette (1990) scheme. The cloud-precipitation-vertical diffusion scheme uses the statistical approach of Ricard and Royer (1993). The soil scheme no longer has a deep relaxation towards climatology, but a 4-layer diffusion scheme; a few other improvements have been also brought to the ISBA soil vegetation scheme (Douville et al., 2000). The orographic gravity wave drag has been improved by the addition of mountain blocking and lift effect (Lott and Miller, 1997; Lott, 1999).

This second series is also a re-forecasting study of the ERA15 period with 15 winters and 15 summers. The difference with PROVOST is the use of 120 members. In this case, it is not possible to use the traditional lagged average (Hoffman and Kalnay, 1983) approach to generate the ensembles. The method used to produce ensembles is important for short- and medium-range forecasting (Houtekamer and Derome, 1995; Molteni et al., 1996; Anderson, 1997). In the case of seasonal forecasts, especially if the first month is discarded, the way to produce the members is

less important, since the growth of initial errors saturates after a few weeks. A simple Monte Carlo approach has been used here. The perturbation is a linear combination of the atmospheric variables for the 9 initial states of PROVOST for a given year. Nine random weights are generated and normalized so that their sum is 0 and the sum of their squares is 0.2. No perturbation is calculated for the surface and soil variables (except for surface pressure). The perturbation is added to the ERA15 initial condition of the 5th day in the 9 lagged situations (i.e., 26 November for winter cases and 27 May for summer cases).

The mean spread of tropical precipitation (averaged in the 30°N–30°S belt) for JFM is 1.4, 1.2, 1.3 and 1.1 mm/day in the 4 PROVOST experiments (ECMWF, UKMO, MF and EDF, respectively). The multimodel produces a higher spread of 1.5 mm/day, as expected. In the new experiment, the spread of the 120 members is 1.2 mm/day. This is comparable with the spread of the individual models in PROVOST, and shows that the method used to produce the perturbation has a similar effect as lagged averaging. The new forecast experiment will be hereafter referred to as PROVOST2.

3. Forecast skill

The deterministic skill in PROVOST and PROVOST2 has been measured by the mean anomaly correlation (Déqué, 1997). This score corresponds to the anomaly correlation coefficient when applied to a map, and to the time correlation when applied to a time series. Analyses used for the verification come from the monthly database produced by Xie and Arkin (1996). Fig. 1 shows the correlation as a function of latitude for JFM and JAS. Values greater than 0.5 are obtained for both seasons and both experiments in the tropics only. The scores are not symmetric with respect to the equator, but are higher in the winter rather than in the summer hemisphere. One can see that PROVOST has slightly better scores than PROVOST2. Table 1 shows different scores for the tropical belt (30°N–30°S). In the case of PROVOST2, sub-ensembles of 36 members (27 members in summer) have been considered, drawn at random among the 120 members available, in order to get a forecast configuration, suitable for comparison with the PROVOST multimodel. The sub-sampling procedure is repeated 200 times, so that a 95% interval is calculated by sorting the 200 scores. This interval is [0.48, 0.50] in JFM, and [0.40, 0.42] in JAS. In both seasons, the PROVOST multimodel is superior to PROVOST2 when using the same ensemble size. Even with 120 members, skill in PROVOST2 predictions is less than in PROVOST (correlations of 0.49 in JFM and 0.42 in JAS). This result shows that it is preferable to have several models with smaller ensembles each than to have one model with a very large ensemble. In Déqué (1997), it is shown that an ensemble size greater than 3 does not widely increase the mean anomaly correlation for tropical precipitation (with a single model). This result is confirmed with PROVOST and PROVOST2. The main reason for producing large ensembles in our case is the generation of probability forecasts (see next sections).

To ensure that these results are not just artifacts, a pseudo-model forecast is considered by using a scrambling procedure. A permutation P of the 15 years (1979 through 1993) is produced and the forecast associated with the observation of year Y is the PROVOST multimodel forecast of year $P(Y)$. The score of this series of forecasts is calculated. This procedure is repeated 200 times. Then, the mean and the 95% interval are calculated for the anomaly correlation. For both seasons, the mean score is 0. Interestingly, the distribution is skewed in JFM with an interval of $[-0.17, 0.30]$, whereas it is $[-0.13, 0.14]$ in JAS. This phenomenon is also observed with PROVOST2 data, and could be due to the presence of 2 El Niño winters

Table 1. *Anomaly correlation for tropical precipitation with the 4 models in PROVOST, the PROVOST multimodel and PROVOST2 model with 36-member or 27-member ensembles*

	ECMWF	UKMO	MF	EDF	Multi	PROVOST2
JFM	0.51	0.45	0.41	0.48	0.55	0.49
JAS	0.31	0.35	0.38		0.44	0.41

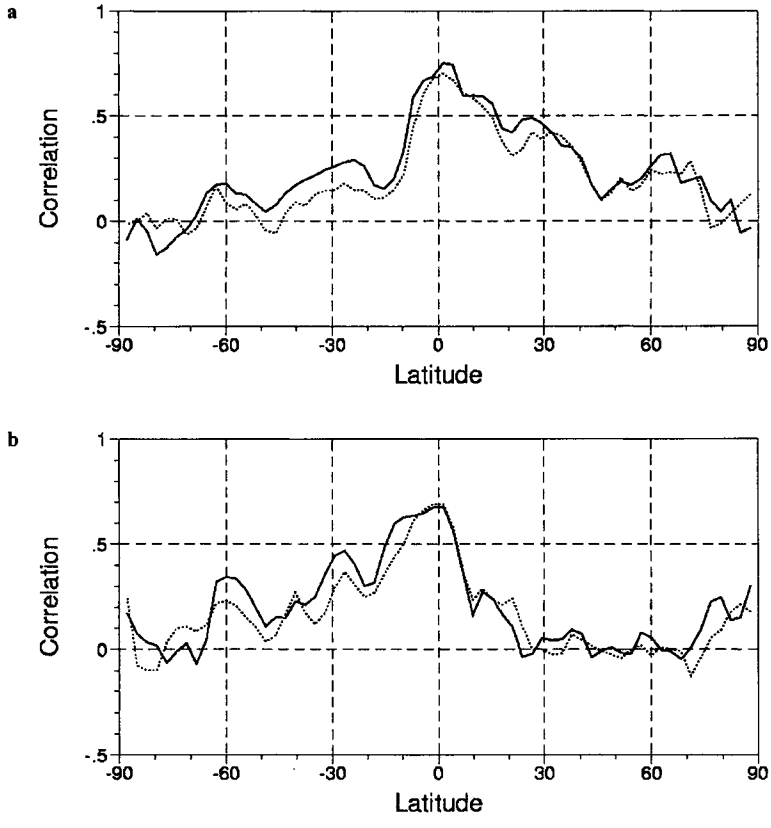

Fig. 1. Mean anomaly correlation as a function of latitude for JFM (a) and JAS (b) precipitation; PROVOST (solid line) and PROVOST2 (dot line).

in the sample which might produce high scores in a few scrambled combinations. In the case of La Niña, the tropical precipitation response is not the opposite anomaly, so that it is not possible to get highly negative scores by scrambling. This skewed behavior can be reproduced by a simple exercise: consider a series with two "1" and thirteen "0", and generate a random correlation between the original series and a scrambled one. The population of scores you obtain contains only 3 values: 1 (frequency 1/105), 11/26 (frequency 26/105) and −4/26 (frequency 78/105). For other variables like temperature or for other regions like midlatitudes, the distribution is not skewed. Nevertheless, the scores obtained in PROVOST and PROVOST2 are above the 95% level and cannot be attributed to an artifact of the verification procedure.

At this stage, we have obtained forecasts with a skill, which is both statistically significant and non negligible, as far as the tropical region is concerned. However, a correlation of 0.50 is a relatively small one, and any reasonable use of this kind of forecast must be a probabilistic approach.

4. Probability forecasts

4.1. Ranked probability score

The simplest way to produce a probability forecast using an ensemble of integrations is to consider each member as an equiprobable realization of the truth. A probability is obtained by counting the number of members which predict a particular event (e.g., a positive anomaly). The implicit assumption is that the model is perfect, since the truth is assumed to follow the same

distribution as the members. Once a series of forecasts has been produced, a score can be calculated, to evaluate the skill of the forecasting system. A simple and robust way is to use Euclidian metrics between the forecast and verification data. The Brier Score (BS) is the squared difference between the forecast probability and the verification, averaged for the different categories (Brier, 1950). The ranked probability score (RPS) has been introduced by Epstein (1969) for categories derived from a continuous variable, which is mostly the case in meteorology. However, in the case of two categories, the RPS and the BS are proportional. For example, let us assume that we define a precipitation threshold r. If p is the forecast probability that the dry category occurs ($R < r$) and v the corresponding verification (i.e., $v = 1$ if the dry category is observed, $v = 0$ otherwise), the RPS reads:

$$\mathrm{RPS} = (p - v)^2. \tag{1}$$

It can be remarked that when the forecast probability is obtained by counting the members which belong to the category, the RPS is simply the mean square error (MSE) applied to the transformed data: the precipitation R in the forecast members or in the verification is replaced by 1 if $R < r$, and by 0 otherwise.

The RPS is a distance, not a skill score. It must be compared to a reference value. It can be demonstrated that the skill-less forecast which minimizes the RPS consists of using the interannual observed frequency of the category as the prediction. This poor man's forecast is called a climatological probability forecast and its RPS is noted RPSc. So, a probability skill score can be introduced as:

$$\mathrm{PSS} = 100(1 - \mathrm{RPS}/\mathrm{RPSc}). \tag{2}$$

Fig. 2 shows the RPS in PROVOST and PROVOST2 and the RPSc as a function of threshold r. In fact, the thresholds depend on the season, the geographical location, and are different for the model and verification data. They are calculated as:

$$r = m(R) + \alpha\sigma(R), \tag{3}$$

where $m(R)$ is the local averaged precipitation, $\sigma(R)$ the interannual standard deviation, and α a dimensionless parameter. Eq. (3) means that a normalized precipitation index is compared to the threshold α. 9 different values for α have been taken, to examine different configurations of users of the forecasts. We have chosen for α, the deciles (10-quantiles) of a Gaussian distribution. This does not mean that the probability of the lower category is 1/10, 2/10, ..., 9/10, i.e., a Gaussian hypothesis is not made for the interannual distribution. This is just a convenient way to remove the systematic error and have reasonable dry and wet categories whatever the region in the tropical belt. In particular $\alpha = 0$ (5th decile) corresponds to a forecast of a positive versus a negative anomaly. One could estimate the actual deciles by counting, but with only 15 years this non-parametric method is unsuitable. The vertical bars in Fig. 2 indicate the average observed frequency (noted f_0 in the following) for each category: they match reasonably the diagonal (0,0 to 10,1) and this shows the suitability of the Gaussian deciles. This fit is also found for the model frequency (not shown). The mean and standard deviations are calculated with 14 or 13 years (i.e., excluding the target year) to avoid introducing verification values in the forecast process. For the same reason, the climatology probability forecast is calculated with the other 14 or 13 years.

As can be seen in Fig. 2, the scores of model forecasts for both PROVOST and PROVOST2 are very close to the climatology. For any threshold, the PSS is less than 10%: the maximum of 7% is obtained with PROVOST and wet thresholds. These poor scores can be partly explained by the fact that our method is biased. Murphy (1973) decomposed the BS or RPS into three terms (see also Palmer et al. (2000)):

- The reliability term is due to the fact that an event predicted with a probability of 0.2 does not occur, on the average in 20% of the cases. This term is zero for a climatological forecast.
- The resolution term measures the intrinsic capability of the forecast system to discriminate the year-to-year events. This term contributes negatively and is zero for a climatologcal forecast.
- The uncertainty term does not depend on the forecast system, but only on the mean observed frequency of the category f_0. It corresponds to the RPS of a climatological forecast $f_0(1 - f_0)$.

Once the reliability diagram (Wilks, 1995) is known, it is possible to modify the probabilities to make the system reliable. For example, if the

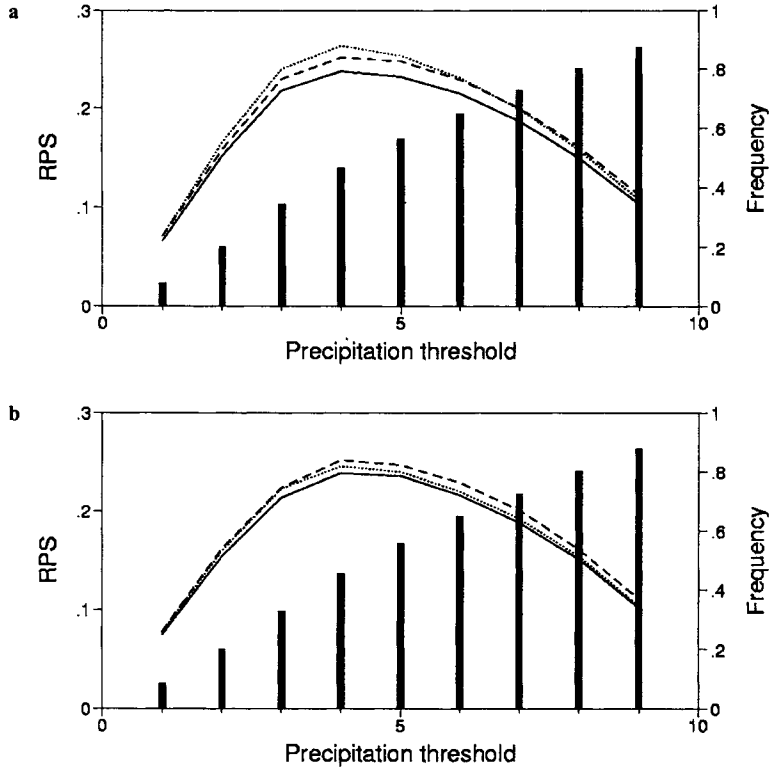

Fig. 2. Mean RPS for PROVOST (solid line), PROVOST2 (dot line) and climatological (dash line) probability forecasts of precipitation in two categories as a function of precipitation threshold (given by eq. (3)); the leftmost categories are the driest. The RPS is averaged over the tropical belt for JFM (a) and JAS (b). The bars correspond to the observed frequency of each category.

forecasts of 0.2 correspond on average to an observed frequency of 30%, then a forecast of 0.2 can be modified to 0.3. This method is similar to the systematic error removal in deterministic prediction. The problem is that we have only 15 years, and our estimated reliability diagram cannot involve the target year (otherwise we are cheating). When several reliability diagrams are plotted they appear to have the form:

$$f(p) = p - \beta p(p-1)(p-f_0), \tag{4}$$

where p is the probability, $f(p)$ the frequency of occurrence of the phenomenon when this probability is predicted and β a coefficient to be adjusted. Eq. (4) is the simplest formula which ensures that the fitted probability is equal to the probability for a probability 0, f_0 and 1. The coefficient β has to be adjusted by least squares. When β varies with location, PSSs between 10% and 20% are obtained if we do not exclude the target year, but smaller PSSs than without correction (i.e., $\beta = 0$) are obtained when we exclude it. So, for the sake of robustness, we chose the value for β (one per year and per threshold) which minimizes the squared difference, averaged over the tropical belt and for the forecasts of the 14 other years, between the actual and the predicted frequencies. For year i, β is given by:

$$\beta = \frac{\left\langle \sum_{j \neq i} (p_j - v_j) p_j (p_j - 1)(p_j - f_0) \right\rangle}{\left\langle \sum_{j \neq i} p_j^2 (p_j - 1)^2 (p_j - f_0)^2 \right\rangle}, \tag{5}$$

where p_j and v_j are the probability forecast and the verification respectively for year j and $\langle\ \rangle$ is the spatial average. The value for β is about -3 for the medium thresholds and -1.5 for the

extreme thresholds in the case of PROVOST. In the case of PROVOST2, absolute values are slightly larger (about -4 and -2 respectively). Fig. 3 shows the reliability diagram and its polynomial fit for the JFM forecasts of PROVOST with the 2nd and the 5th thresholds. In both cases, the model tends to exaggerate the extremes: small probabilities are predicted too often whereas high probabilities are predicted too scarcely. Table 2 shows the extent to which this correction improves the PSS for threshold 5 (corresponding to a climatological distribution of about 50% in each category).

4.2. *Cost-loss model*

One should not be surprised by the fact that the PSS is low, even after correction. This measure index is very severe because we attempt to reproduce a Dirac function (the observation) with an ensemble with large spread. This can be compared with the use of RMS to measure deterministic skill: ensemble averaging reduces signal amplitude, and the ensemble mean is closer to climatology than to the observation, when using a quadratic distance. This is why the anomaly correlation is preferred as a deterministic score: it does not depend on amplitude and measures an angle instead of a distance in phase space.

We need such a score to verify probabilistic forecasts, and the cost-loss model (see Palmer et al., 2000 for details) is a good approach. A grid point (in the tropics), a season (JFM or JAS) and a threshold for precipitation can be chosen. If a user can take an action which costs C but which

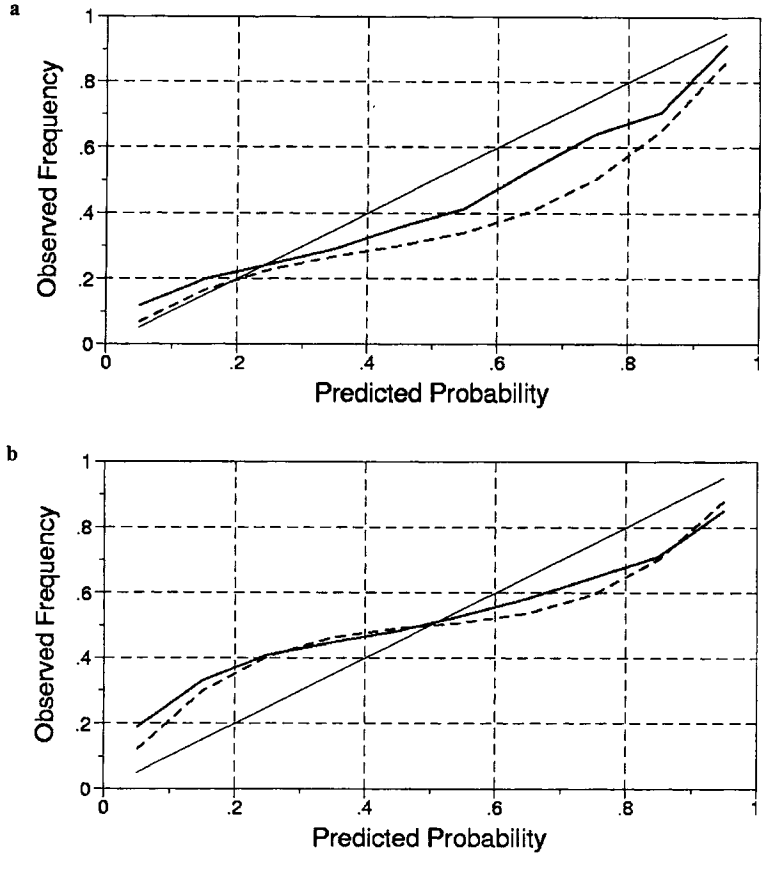

Fig. 3. Reliability diagram for the JFM PROVOST probability forecasts of tropical precipitation in two categories for the 2nd (a) and 5th (b) precipitation threshold. The solid curve is the empirical estimate and the dash curve the fitted polynomial.

Table 2. *Probability skill score for tropical precipitation with the PROVOST multimodel and PROVOST2 model (120 members) with and without reliability correction; binary forecast with threshold 5 (positive versus negative anomaly)*

	PROVOST	PROVOST2	PROVOST (corrected)	PROVOST2 (corrected)
JFM	6%	−2%	9%	5%
JAS	5%	3%	9%	8%

avoids a loss L when the first category occurs, a comparison of the average over 15 years of money spent according to three different strategies can be made:

- S1: always take action if C/L is less than f_0, never otherwise; mean expense is $E1$.
- S2: take action only when the probabilistic forecast of the first category is above a threshold t (to be determined by the user); this probability threshold should not be confused with precipitation threshold r; mean expense is $E2$.
- S3: read the future and take action only when the first category occurs; mean expense is $E3$.

Mean expense $E2$ is generally less than $E1$ (climatology forecast) and always greater than $E3$ (perfect forecast). The value of the forecast is defined as:

$$V = \frac{E1 - E2}{E1 - E3}. \tag{6}$$

V depends only on the forecasts and on C/L and can be plotted for C/L varying between 0 and 1 (beyond 1, $E1 = E3$ and V cannot be defined). The difference between $E1$ and $E3$ is maximum for $C/L = f_0$. When the cost is too low or too high with respect to the loss, the user does not need any forecast for taking his decision. So, one can simplify the evaluation by considering only the case of $C/L = f_0$ for which potential gain from the forecast is maximum, and thus a forecast may be useful for the decision. The denominator of eq. (6) simply becomes $f_0(1 - f_0)$. The choice of probability threshold t must be done without cheating, i.e., excluding the target year when optimizing the strategy, otherwise values obtained are about twice as large with the short 15-year sample.

Fig. 4 shows for the 9 precipitation thresholds, the forecast value, i.e., percentage of money saved with respect to a perfect forecast. Forecast value is mostly above 10%, which shows potential benefits of a probability forecast for a user with a cost/loss ratio in agreement with the frequency of the event to be predicted. The forecast years have been scrambled, and the empirical distribution of forecast value in the case of no skill has been estimated from 100 samples: the 95% confidence interval is [−7%, 2%] for the 5th threshold, and the upper boundary of the interval is below 6% whatever the threshold. When the empirical reliability correction described in the last section is applied, the value is generally smaller: the two-step optimization in cross-validation mode is not robust enough. The sensitivity to the choice of C/L ratio has been evaluated by increasing and decreasing this ratio by 10%: in both cases, the forecast value remains above 10%, so that our results are not restricted to a lucky user who has the optimal ratio.

Figs. 5 and 6 show the geographical distribution of the forecast value for both experiments and both seasons in the case of two equiprobable categories. As expected, regions with a value above 10% are generally located in the tropics and preferentially over the oceans. One can remark however that regions like India, Brazil or western Africa have positive values during the rainy season.

The source of the probabilistic skill comes from the year to year variability of the ensemble mean and from the statistical adjustments (correction of the reliability or optimization of the threshold t), but not from the year to year variability of ensemble spread. For each year, we have linearly modified the forecast members of the ensemble, so that the ensemble mean remains unchanged, but the ensemble standard deviation is the average of the 15 ensemble standard deviations. Thus, each ensemble has the same spread. The results for the PSS as well as for the forecast value are quasi identical to the results with the unmodified forecasts. In Déqué et al. (1994) a similar conclusion had been obtained, but an ensemble size of 5

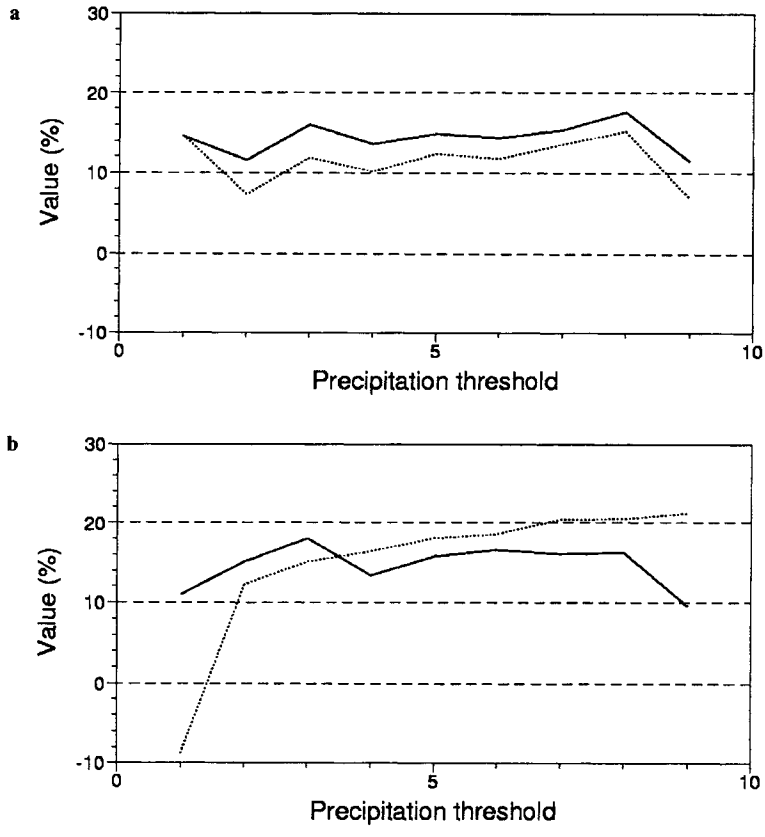

Fig. 4. Forecast value (defined by eq. (6)) of PROVOST (solid line) and PROVOST2 (dot line) probability forecasts of tropical precipitation in two categories as a function of the precipitation threshold for JFM (a) and JAS (b).

may have been too small to provide a good estimate of the standard deviation. Here with 120 members, we are sure that the estimate of spread for a given ensemble is well represented.

5. Less idealized forecasts

5.1. Experiments

The skill presented in the previous sections does not really correspond to the skill which can expected from an operational forecast. Indeed, the most important forcing, i.e., the tropical SST is imposed by observed data which are not available at the time of forecast production. Some meteorological centers like NCEP or ECMWF use coupled ocean–atmosphere models in operational mode. The European project DEMETER will involve carrying out hindcast predictions of the past 40 years with several coupled models (in the spirit of PROVOST) and is planned to be achieved in 2003.

If the forecast target is the next season, persistence of SST anomaly in the tropics is an unexpensive and bias-free way to get SST predictions. The UKMO and Météo-France are using this method for their real-time monthly to seasonal forecasts. Here, we propose a slightly more complex approach by a red noise process:

$$\mathrm{SST}(m) = A\,\mathrm{SST}(m-1) + B\,\mathrm{WN}(m), \qquad (7)$$

where $\mathrm{SST}(m)$ is the SST anomaly of a given month, $\mathrm{SST}(m-1)$ is the SST anomaly of the previous month, and $\mathrm{WN}(m)$ is a random component (white noise with mean 0 and variance 1). A and B are coefficients fitted by least squares (linear

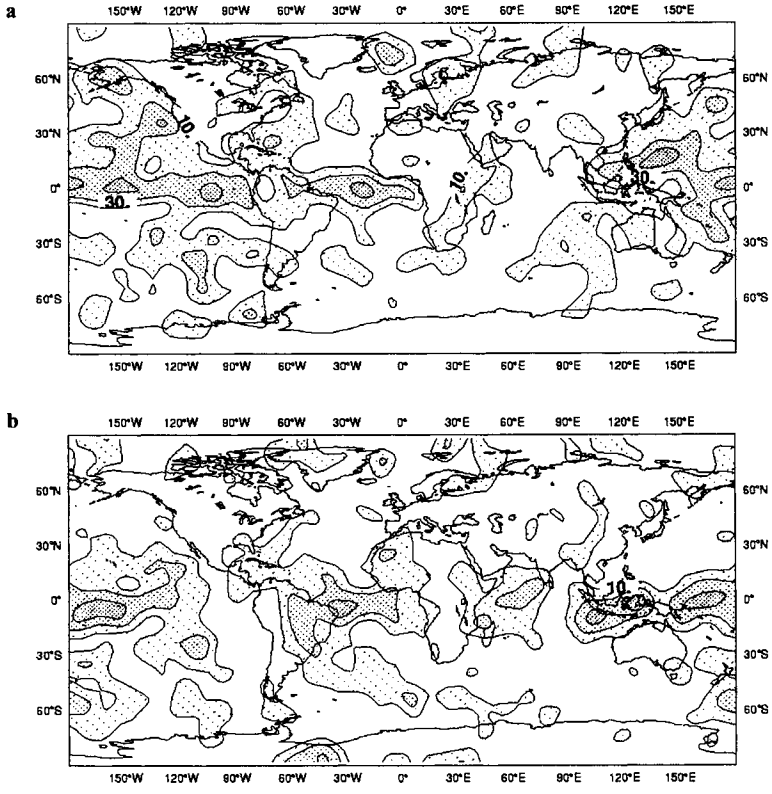

Fig. 5. Forecast value of precipitation probability forecasts for two equiprobable categories (5th threshold) and a cost/loss ratio of 0.5; PROVOST results for JFM (a) and JAS (b); contours 10%, 30% and 50% with shading above 10%.

auto-regression) during the learning period 1950–1978. Since the SST field has a spatial consistency, the auto-regression is calculated with the first 20 principal components instead of grid-point values. A and B depend on the EOF number and on the calendar month: the first EOF (ENSO pattern) is more persistent than the next ones, and is more persistent in JFM than in JAS. The SST anomaly field is subsequently reconstructed with the corresponding 20 EOFs.

This method allows generation of ensembles of SST fields which are close together in the tropics for the first month of the forecast, because SST is persistent, but which progressively diverge with the course of the forecast. When $A = 1$ and $B = 0$, we have pure persistence. $B = 0$ is better in terms of quadratic error, but leads to a climatological SST almost everywhere after 3 months, even with strong El Niño events like 1982–83. With $B \neq 0$, some ensemble members can exhibit El Niño like anomalies, whilst others exhibit a weak or an opposite pattern. The response of the GCM being non-linear, a better deterministic forecast can be expected. Moreover, this technique takes into account the uncertainty about SST evolution, so probabilistic forecasts should improve as well.

We have performed a new series of 15 forecasts for DJFM and JJAS in the same way as in the previous sections. The forecast is based on 60-member ensembles and uses the SST prediction scheme given by eq. (7). The initial situations do not need to be perturbed since the SSTs have their own spread. This experiment will be referred to as PERSIST.

5.2. *Scores*

As far as deterministic scores are concerned, the anomaly correlation for tropical precipitation is

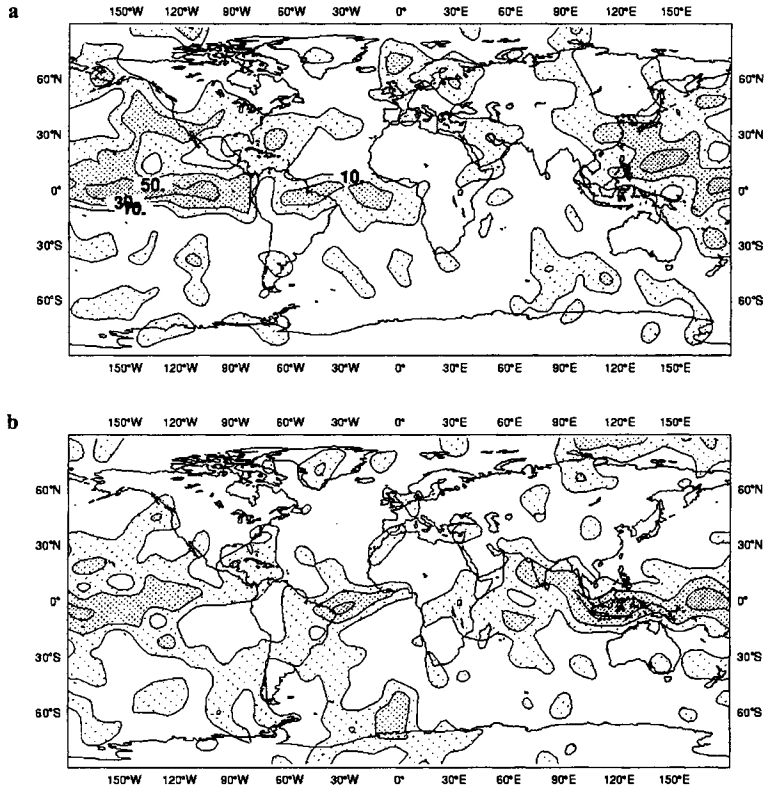

Fig. 6. As Fig. 5, but for PROVOST2 experiment.

0.42 in JFM and 0.37 in JAS. As expected, these scores are less than in PROVOST2, particularly in JFM. However they remain interesting, and a good coupled model would be expected to have a skill score part way between that of PERSIST and that of PROVOST2.

The probability forecasts, evaluated by forecast value, exhibit some skill. In JFM, the 5th threshold (positive versus negative precipitation anomaly) has a value of 8%. In JAS, forecast value is 11%. For the other precipitation thresholds, the value varies from 3% to 12%. Fig. 7 shows that values greater than 10% are obtained over a few continental areas in the tropics. This is an indication of the usefulness of such forecasts. When averaged over land points of the tropical belt, forecast value in JAS is 4%, which remains outside the 95% confidence interval of $[-5\%, 2\%]$ obtained by scrambling the 15 years of the forecast to estimate the forecast value of a skill-less forecast system.

6. Conclusion

Analyzing some results from the PROVOST experiment and from two additional experiments has shown that even though there is a clear possibility of a valuable use of numerical seasonal predictions, it is necessary to be modest when speaking of skill. The skill in probabilistic prediction comes essentially from the ability of the model to evaluate the mean response to SST anomalies. Unless another formulation of probability forecast is proposed (e.g., based on analogs), the role of ensembles or multimodel ensembles is to filter out noise or cancel individual model deviations. This explains why the PROVOST multimodel ensemble with 36 or 27 members is more successful than a "unimodel" ensemble of 120 members. In the tropics, large ensemble sizes are not necessary, even for a probabilistic forecast: for a "unimodel" ensemble, 10 members appears sufficient. However

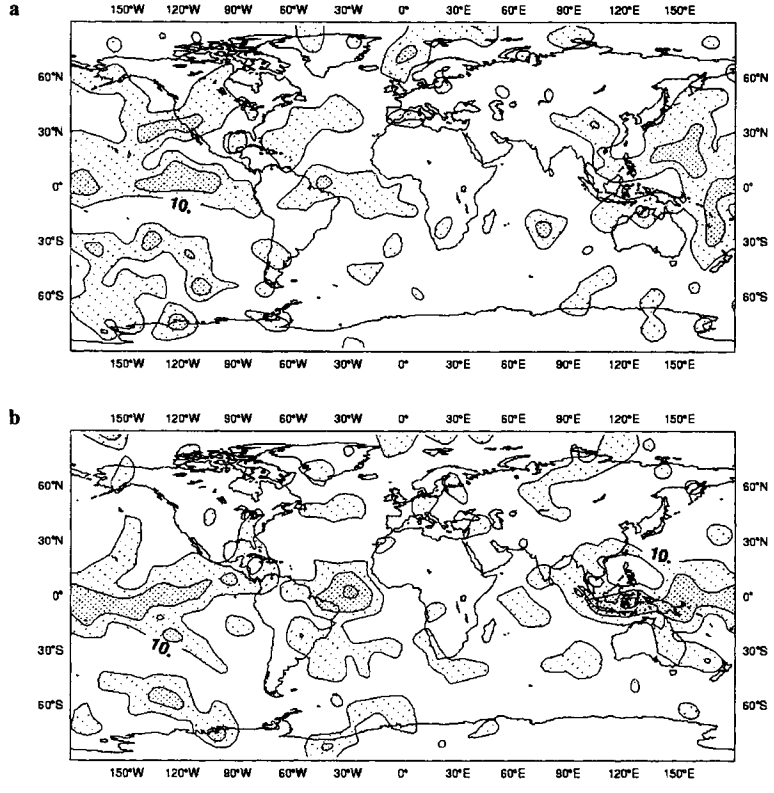

Fig. 7. As Fig. 5, but for PERSIST experiment.

in the case of spreading SSTs (PERSIST experiment) a perceptible skill increase from 10 to 60 members was found to occur. It appears far more important to have more years in the forecast sample. The reliability correction or the model value optimization require robust estimates, which is not the case with 15 years as shown by the difference between the training and the cross-validation modes. In the DEMETER project, it is planned to have 30 years of forecasts (possibly 40 years since it is based on the ERA40 reanalyses).

Another limitation of skill is geographical distribution. Most skill in the tropics comes from the Pacific ocean region in the PERSIST experiment. This arises from the fact that ENSO events are more persistent than events in other oceans. Here again, promising results are expected from the DEMETER experiment in which coupled ocean–atmosphere models will be used.

Nevertheless, our results prove that the choice by meteorological services to develop operational numerical seasonal forecasts is justified. As long as we are scientifically convinced that skill exhibited in long reforecasting exercises is not artificial, the fact that some (but not all) users can save 10% of what they could have saved with the exact vision of the future is a very positive result.

7. Acknowledgments

I acknowledge the help of Robin Clark and Jean François Guérémy for their valuable comments on the manuscript and their fruitful discussions. I would like also to thank ECMWF for providing the PROVOST and ERA15 results. This work was supported by the European Union Program Energy, Environment and Sustainable Development under contract EVK2-CT-1999-00024 (DEMETER).

REFERENCES

Anderson, J. L. 1997. The impact of dynamical constraints on the selection of initial conditions for ensemble predictions: low-order perfect model results. *Mon. Wea. Rev.* **11**, 2969–2983.

Brankovic, C. and Palmer, T. N. 2000. Seasonal skill and predictability of ECMWF PROVOST ensembles. *Q. J. R. Meteorol. Soc.* **126**, 2035–2068.

Blackmon, M. L., Geisler, J. E. and Pitcher E. J. 1983. A general circulation model study of January climate anomaly pattern associated with interannual variation of equatorial Pacific sea surface temperature. *J. Atmos. Sci.* **40**, 1410–1425.

Bougeault, P. 1985. A simple parameterization of the large-scale effects of cumulus convection. *Mon. Wea. Rev.* **113**, 2108–2121.

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

Déqué, 1997. Ensemble size for numerical seasonal forecasts. *Tellus* **49A**, 74–86.

Déqué, M. and Piedelievre, J. P. 1995. High resolution climate simulation over Europe. *Clim. Dyn.* **11**, 321–339.

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

Doblas-Reyes, F. J., Déqué, M. and Piedelievre, J. P. 2000. Multi-model spread and probabilistic seasonal forecasts in PROVOST. *Q. J. R. Meteorol. Soc.* **126**, 2069–2088.

Douville, H., Planton, S., Royer, J. F., Stephenson, D. B., Tyteca, S., Kergoat, L., Lafont S. and. Betts, R. A. 2000. The importance of vegetation feedbacks in doubled-CO_2 time-slice experiments. *J. Geophys. Res.* **105**, 14,841–14,861.

Epstein, E. S. 1969. A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.* **8**, 985–987.

Gibson, J. K., Kållberg, P., Uppala, S., Hernandez, A. and Serano, E. 1997. *ERA description.* ECMWF Re-analysis project report series. ECMWF, Shinfield Park, Reading, RG2 9AX, UK.

Hoffman, N. R. and Kalnay, E. 1983. Lagged average forecasting, an alternative to Monte Carlo forecasting. *Tellus* **35A**, 100–118.

Houtekamer, P. L. and Derome, J. 1995. Methods for ensemble prediction. *Mon. Wea. Rev.* **123**, 2181–2196.

Lau, N. C. 1985. Modelling the seasonal dependence of the atmospheric response to observed El Niños in 1962–76. *Mon. Wea. Rev.* **113**, 1970–1996.

Lott, F. 1999. Alleviation of stationary biases in a GCM through a mountain drag parameterization scheme and a simple representation of mountain lift forces. *Mon. Wea. Rev.* **125**, 788–801.

Lott, F. and Miller, M. J. 1997. A new subgrid-scale orographic drag parametrization: its formulation and testing. *Q. J. R. Meteorol. Soc.* **123**, 101–127.

Molteni, F., Buizza, R., Palmer, T. N. and Petroliagis, T. 1996. The ECMWF ensemble prediction system: Methodology and validation. *Q. J. R. Meteorol. Soc.* **122**, 73–199.

Morcrette, J. J. 1990. Impact of changes to the radiation transfer parameterizations plus cloud optical properties in the ECMWF model. *Mon. Wea. Rev.* **118**, 847–873.

Murphy, A. H. 1973. A new vector partition of the probability score. *J. Appl. Meteor.* **12**, 595–600.

Owen, J. A. and Palmer, T. N. 1987. The impact of El Niño on an ensemble of extended-range forecasts. *Mon. Wea. Rev.* **115**, 2103–2117.

Palmer, T. N. and Anderson, D. L.T. 1994. The prospects for seasonal forecasting — a review paper. *Q. J. R. Meteorol. Soc.* **120**, 755–793.

Palmer, T. N., Brankovic, C. and Richardson, D. S. 2000. A probability and decision-model analysis of PROVOST seasonal multi-model integrations. *Q. J. R. Meteorol. Soc.* **126**, 2013–2034.

Ricard, J. L. and Royer, J. F. 1993. A statistical cloud scheme for use in an AGCM. *Ann. Geophysicae* **11**, 1095–1115.

Shukla, J. 1998. Predictability in the midst of chaos: a scientific basis for climate forecasting. *Science* **282**, 728–731.

Shukla, J. and Wallace, J. M. 1983. Numerical simulation of the atmospheric response to equatorial Pacific sea surface temperature anomalies. *J. Atmos. Sci.* **40**, 1613–1630.

Wilks, D. S. 1995. *Statistical methods in the atmospheric sciences.* Academic Press. 467 pp.

Xie, P. and Arkin, P. A. 1996. Analyses of global monthly precipitation using gauge observations, satellite estimates, and numerical model predictions. *J. Climate* **9**, 840–858.