Predicting uncertainty in forecasts of weather and climate

(Also published as ECMWF Technical Memorandum No. 294)

By **T.N. Palmer**

Research Department

November 1999

Abstract

The predictability of weather and climate forecasts is determined by the projection of uncertainties in both initial conditions and model formulation onto flow-dependent instabilities of the chaotic climate attractor. Since it is essential to be able to estimate the impact of such uncertainties on forecast accuracy, no weather or climate prediction can be considered complete without a forecast of the associated flow-dependent predictability. The problem of predicting uncertainty can be posed in terms of the Liouville equation for the growth of initial uncertainty, or a form of Fokker-Planck equation if model uncertainties are also taken into account. However, in practice, the problem is approached using ensembles of integrations of comprehensive weather and climate prediction models, with explicit perturbations to both initial conditions and model formulation; the resulting ensemble of forecasts can be interpreted as a probabilistic prediction.

Many of the difficulties in forecasting predictability arise from the large dimensionality of the climate system, and special techniques to generate ensemble perturbations have been developed. Special emphasis is placed on the use of singular-vector methods to determine the linearly unstable component of the initial probability density function. Methods to sample uncertainties in model formulation are also described. Practical ensemble prediction systems for prediction on timescales of days (weather forecasts), seasons (including predictions of El Niño) and decades (including climate change projections) are described, and examples of resulting probabilistic forecast products shown. Methods to evaluate the skill of these probabilistic forecasts are outlined. By using ensemble forecasts as input to a simple decision-model analysis, it is shown that probability forecasts of weather and climate have greater potential economic value than corresponding single deterministic forecasts with uncertain accuracy.

Table of contents

8. VERIFYING FORECASTS OF UNCERTAINTY

As discussed, the output from an ensemble forecast can be used to construct a probabilistic prediction. In this section, we discuss two basic measures of skill for assessing a probability forecast: the Brier Score and the Relative Operating Characteristic. Both of these measures are based on the skill of probabilistic forecasts of a binary event E, as discussed in Section 7 above. For example E could be: temperatures will fall below 0°C in three days time; average rainfall for the next three months will be at least one standard deviation below normal; seasonal-mean rainfall will be below average and temperature above average, and so on.

8.1 The Brier score and its decomposition

Consider an event E which, for a particular ensemble forecast, occurs a fraction p of times within the ensemble. If E actually occurred then let $v = 1$, otherwise $v = 0$. Repeat this over a sample of N different ensemble forecasts, so that p_i is the probability of E in the i th ensemble forecast and $v_i = 1$ or $v_i = 0$, depending on whether E occurred or not in the i th verification ($i = 1, 2 \ldots N$).

The Brier score (*Wilks*, 1995) is defined by

$$b = \frac{1}{N}\sum_{i=1}^{N}(p_i - v_i)^2, \qquad 0 \le p_i \le 1, v_i \in \{0, 1\} \tag{47}$$

From its definition $0 \le b \le 1$, equalling zero only in the ideal limit of a perfect deterministic forecast. For a large enough sample, the Brier score can be written as

$$b = \int_0^1 [p-1]^2 o(p)\rho_{ens}(p)\mathrm{d}p + \int_0^1 p^2[1-o(p)]\rho_{ens}(p)\mathrm{d}p \tag{48}$$

where $\rho_{ens}(p)\mathrm{d}p$ is the relative frequency that E was forecast with probability between p and $p + \mathrm{d}p$, and $o(p)$ gives the proportion of such cases when E actually occurred. To see the relationship between (47) and (48) note that $\int_0^1 [p-1]^2 o(p)\rho_{ens}(p)\mathrm{d}p$ is the Brier score for ensembles where E actually occurred, and $\int_0^1 [p-0]^2 o(p)\rho_{ens}(p)\mathrm{d}p$ is the Brier score for ensembles where E did not occur.

Simple algebra on (48) gives *Murphy*'s (1973) decomposition

$$b = b_{rel} - b_{res} + b_{unc} \tag{49}$$

of the Brier score, where

$$b_{rel} = \int_0^1 [p - o(p)]^2 \rho_{ens}(p)\mathrm{d}p \tag{50}$$

is the reliability component,

$$b_{res} = \int_0^1 [\bar{o} - o(p)]^2 \rho_{ens}(p)\mathrm{d}p \tag{51}$$

is the resolution component

$$b_{unc} = \bar{o}[1-\bar{o}] \tag{52}$$

is the uncertainty component, and

$$\bar{o} = \int_0^1 o(p)\rho_{ens}(p)\mathrm{d}p \tag{53}$$

is the (sample) climatological frequency of *E*.

A reliability diagram (*Wilks*, 1995) is one in which $o(p)$ is plotted against p for some finite binning of width δp. In a perfectly reliable system $o(p) = p$ and the graph is a straight line oriented at 45° to the axes, and $b_{rel} = 0$. Reliability measures the mean square distance of the graph of $o(p)$ to the diagonal line.

Resolution measures the mean square distance of the graph of $o(p)$ to the sample climate horizontal line. A system with relatively high b_{res} is one where the dispersion of $o(p)$ about is as large as possible. Conversely, a forecast system has no resolution when, for all forecast probabilities, the event verifies a fraction $o(p) = \bar{o}$ times.

The term b_{unc} on the right-hand side of (49) ranges from 0 to 0.25. If *E* was either so common, or so rare, that it either always occurred or never occurred within the sample of years studied, then $b_{unc} = 0$; conversely if *E* occurred 50% of the time within the sample, then $b_{unc} = 0.25$. Uncertainty is a function of the climatological frequency of E, and is not dependent on the forecasting system itself. It can be shown that the resolution of a perfect deterministic system is equal to the uncertainty.

When assessing the skill of a forecast system, it is often desirable to compare it with the skill of a forecast where the climatological probability $\bar{o}$ is always predicted (so $\rho_{ens}(p) = \delta(p - o)$). The Brier score of such a climatological forecast is $b_{cli} = b_{unc}$ (using the sample climate), since, for such a climatological forecast $b_{rel} = b_{res} = 0$. In terms of this, the Brier skill score, B, of a given forecast system is defined by

$$B = 1 - \frac{b}{b_{cli}} \tag{54}$$

$B \leq 0$ for a forecast no better than climatology, and $B = 1$ for a perfect deterministic forecast.

Skill-score definitions can similarly be given for reliability and resolution, i.e.

$$\left[B_{rel} = 1 - b_{rel}/b_{cli}\right] \tag{55}$$

$$\left[B_{res} = b_{res}/b_{cli}\right] \tag{56}$$

For a perfect deterministic forecast system, $B_{rel} = B_{res} = 1$. Hence, from Eqs. (49) and (54)

$$B = B_{res} + B_{rel} - 1 \tag{57}$$

Fig. 13 shows two examples of reliability diagrams for the ECMWF EPS taken over all day-6 forecasts from December 1998 - February 1999 over Europe (cf. Fig. 8). The events are $E_{>4}$, $E_{>8}$:- lower tropospheric temperature being at least 4°C, 8°C greater than normal. The Brier score, Brier skill score, and Murphy decomposition are shown on the figure.

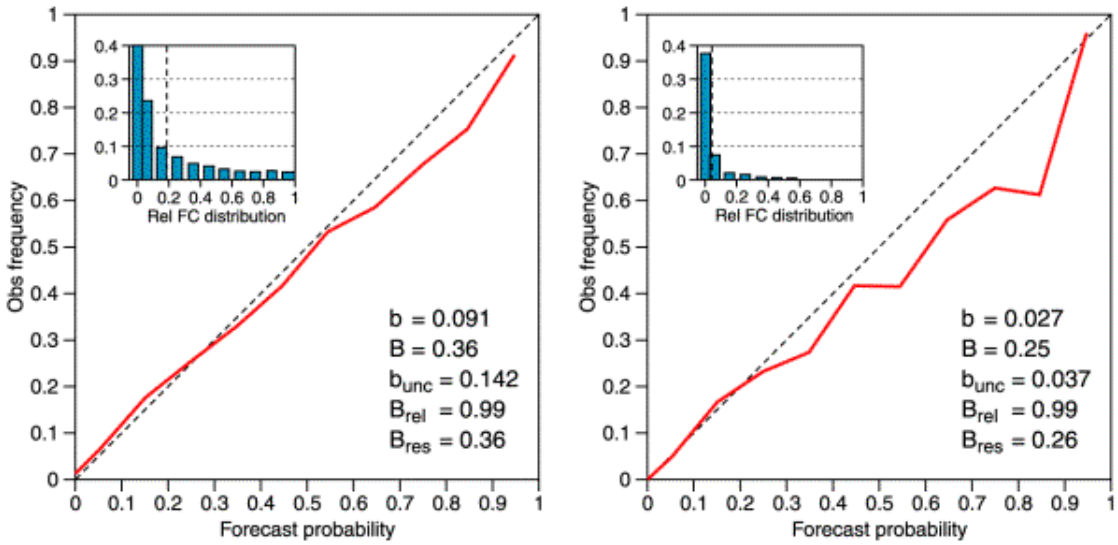

Figure 13. Reliability diagram and related Brier score skill score and Murphy decomposition for the events: (a) 850 hPa temperature is at least 4°C above normal and (b) at least 8°C above normal, based on 6-day forecasts over Europe from the 50-member ECMWF ensemble prediction system from December 1998 - February 1999. Also shown is the pdf $\rho_{ens}(p)$ for the event in question.

The reliability skill score B_{rel} is extremely high for both events. However, the reliability diagrams indicate some overconfidence in the forecasts. For example, on those occasions where $E_{>4}$ was forecast with a probability between 80% and 90% of occasions, the event only verified about 72% of the time. However, it should be remembered that the integrand in Eq. (50) is weighted by the pdf $\rho_{ens}(p)$, shown in each reliability diagram. In both cases, forecasts where $p > 0.4$ are relatively rare and hence contribute little to B_{rel}.

To see why probability forecasts of $E_{>4}$ have higher Brier skill scores than probability forecasts of $E_{>8}$, consider Eq. (57). From Fig. 13 , whilst B_{rel} is the same for both events, B_{res} is larger for $E_{>4}$ than for $E_{>8}$. This can be seen by comparing the histograms of $\rho_{ens}(p)$ in Fig. 13 which are more highly peaked for $E_{>8}$ than for $E_{>4}$; there is less dispersion of the probability forecasts of the more extreme event about its climatological frequency, than the equivalent probability forecasts of the more moderate event. This is hardly surprising; the more extreme event $E_{>8}$ is relatively rare (its climatological frequency is ~ 0.04) and most of the time is forecast with probabilities which almost always lie in the first probability category ($0 \le \delta p \le 0.1$). In order to increase the Brier score of this relatively extreme event, one would need to increase the ensemble size so that finer probability categories can be reliably defined. (For example, suppose an extreme event has a climatological probability of occurrence of p_{rare} Let us suppose that we want to be able to forecast probabilities of this event which can discriminate between probability categories with a band width comparable with this climatological frequency, then the ensemble size S_{ens} should be $>>(1/p_{rare})$.) With finer probability categories, the resolution component of the Brier score can be expected to increase. Providing reliability is not compromised, this will lead to higher overall skill scores.

However, this raises a fundamental dilemma in ensemble forecasting given current computer resources. It would be meaningless to increase ensemble size by degrading the model (e.g. in terms of "physical" resolution) making it cheaper to run, if by doing so it could no longer simulate extreme weather events. Optimising computer resources that on the one hand ensemble sizes are sufficiently large to give reliable probability forecasts of extreme but rare events, and that on the other hand the basic model has sufficient complexity to be able to simulate such events, is a very difficult balance to define.

The Brier score and its decomposition provide powerful objective tools for comparing the performance of different

probabilistic forecast systems. However, the Brier score itself does not address the issue of whether a useful level of skill has been achieved by the probabilistic forecast system. In order to prepare the ground for a diagnostic of probabilistic forecast performance which determines potential economic value, we first introduce skill score originally derived from signal detection theory.

8.2 Relative operating characteristic

The relative operating characteristic (ROC; *Swets*, 1973; *Mason* 1982; *Harvey et al.*, 1992) is based on the forecast assumption that E will occur, providing E is forecast by at least a fraction $p = p_t$ of ensemble members, where the threshold p_t is defined a priori by the user. As discussed below, optimal p_t can be determined by the parameters of a simple decision model.

Consider first a deterministic forecast system. Over a sufficiently large sample of independent forecasts, we can form the forecast- model contingency matrix giving the frequency that E occurred or did not occur, given it was forecast or not forecast, i.e.

		Occurs	
		No	Yes
Forecast	No	α	β
	Yes	γ	δ

Based on these values, the so-called "hit rat" (H) and "false-alarm rate" (F) for E are given by

$$\begin{aligned} H &= \delta/(\beta+\delta) \\ F &= \gamma/(\alpha+\gamma) \end{aligned} \tag{58}$$

Hit and false alarm rates for all ensemble forecast can be defined as follows. It is assumed that E will occur if $p \geq p_t$ (and will not occur if $p < p_t$). By varying p_t between 0 and 1 we can define $H = H(p_t)$, $F = F(p_t)$. In terms of the pdf $\rho_{ens}(p)$

$$\begin{aligned} H(p_t) &= \int o(p)\rho_{ens}(p)\mathrm{d}p/\bar{o} \\ F(p_t) &= \int \{1-o(p)\}\rho_{ens}(p)\mathrm{d}p/(1-\bar{o}) \end{aligned} \tag{59}$$

The ROC curve is a plot of $H(p_t)$ against $F(p_t)$ for $0 \leq p_t \leq 1$. A measure of skill is given by the area under the ROC curve (A). A perfect deterministic forecast has $A = 1$, whilst a no-skill forecast for which the hit and false alarm rates are equal, has $A = 0.5$.

Relative operating characteristic curve for seasonal timescale integrations (run over the years 1979-93, run with prescribed observed SST) for the event $E_{<0}$:- the seasonal-mean (December-February) 850 hPa temperature anomaly is below normal. Solid: based on a single model 9-member ensemble. bottom: based on a multi-model 36-member ensemble (see *Palmer et al.*, 2000) for more details. The area A under the two curves (a measure of skill) is shown.

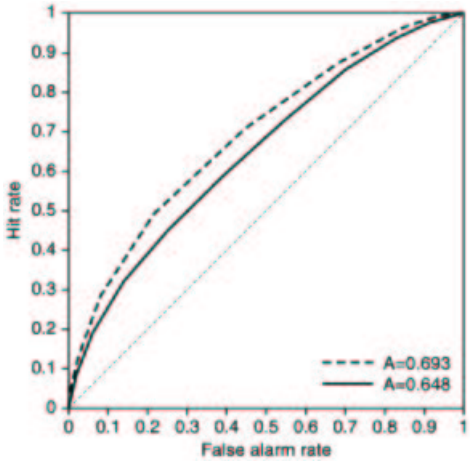

Figure 14. Relative operating characteristic curve for seasonal timescale integrations (run over the years 1979-93, run with prescribed observed SST) for the event $E_{<0}$:- the seasonal-mean (December-February) 850 hPa temperature anomaly is below normal. Solid: based on a single model 9-member ensemble. bottom: based on a multi-model 36-member ensemble (see *Palmer et al.*, 2000) for more details. The area *A* under the two curves (a measure of skill) is shown.

We illustrate in Fig. 14 the application of these measures of skill to a set of multi-model multi-initial condition ensemble integrations made over the seasonal timescale (*Palmer et al.*, 2000). The event being forecast is $E_{<0}$:- the seasonal-mean (December-February) 850 hPa temperature anomaly will be below normal. The global climate models used in the ensemble are the ECMWF model, the UK Meteorological Office Unified Model, and two versions of the French Arpège model; the integrations were made as part of the European Union "Prediction of Climate Variations on Seasonal to Interannual Timescales (PROVOST)". For each of these models, 9-member ensembles were run over the boreal winter season for the period 1979-1993 using observed specified SSTs. The values $H(p_t)$ and $F(p_t)$ have been estimated from probability bins of width 0.1. The ROC curve and corresponding *A* value is shown for the 9-member ECMWF model ensemble, and for the 36-member multi-model ensemble. It can be seen that in both cases, *A* is greater than the no-skill value of 0.5; however, the multi-model ensemble is more skilful than the ECMWF-model ensemble. Studies have shown that the higher skill of the multi-model ensemble arises mainly because of the larger ensemble size, but also because of a sampling of the pdf associated with model uncertainty.

9. The economic value of predicting uncertainty

Although *B* and *A* (defined in Section 8) provide objective measures of skill for ensemble forecasts, they do not determine measures of usefulness for seasonal forecasts. In an attempt to define this notion objectively, we consider here a simple decision model (*Murphy*, 1977; *Katz and Murphy*, 1997) whose inputs are probabilistic forecast information and whose output is potential economic value.

Consider a potential forecast user who can take some specific precautionary action depending on the likelihood that *E* will occur. Let us take some simple examples relevant to seasonal forecasting. If the coming winter is mild (*E*:- seasonal-mean temperature above normal), then overwintering crops may be destroyed by aphid growth. A farmer can take precautionary action by spraying. If the growing season is particularly dry (*E*:- seasonal-mean rainfall at least one standard deviation below normal), then crops may be destroyed by drought. A farmer can take precautionary action by planting drought-resistant crops. In both cases taking precautionary action incurs a cost *C* irre-

spective of whether or not E occurs (cost of spraying, or cost associated with reduced yield and possibly with more expensive seed). However, if E occurs and no precautionary action has been taken, then a loss L is incurred (crop failure).

This simple "cost-loss" analysis is also applicable to much shorter range forecast problems (Richardson, 1999). For example, if the weather event was the occurrence of freezing conditions leading to ice on roads, and the precautionary action was to salt the roads, then C would correspond to the cost of salting, and L would be associated with the increase in road accidents, traffic delays etc.

In general, the expense associated with each combination of ac action /inaction and occurrence/non-occurrence of E is given in the decision-model contingency matrix

		Occurs	
		No	Yes
Take action	No	0	L
	Yes	C	C

It is presumed that the decision maker wishes wishes to maximise profits, or at least minimise overall expenses.

If only the climatological frequency $\bar{o}$ of E is known, there are two basic options: either always or never take precautionary action. Always taking action incurs a cost C on each occasion, whilst never taking action incurs a loss L only on the proportion $\bar{o}$ of occasions when E occurs, giving an expense $\bar{o}L$.

If the seasonal forecast data used in Section 8 above were used by a hypothetical decision maker, would his/her expenses would be reduced beyond what could be achieved using $\bar{o}$ alone? Consider first a deterministic forecast system with characteristics described by the forecast-model contingency matrix in Section 8.2. The user's expected mean expense M per unit loss is

$$M = \frac{\beta L + (\gamma + \beta)C}{L} \tag{60}$$

This can be written in terms of the hit-rate H and the false-alarm F using (58), so that

$$M = F\frac{C}{L}(1-\bar{o}) - H\bar{o}\left(1-\frac{C}{L}\right) + \bar{o} \tag{61}$$

For a perfect deterministic forecast $H = 1$ and $F = 0$, hence

$$M_{\text{per}} = \bar{o}\frac{C}{L} \tag{62}$$

To calculate the mean expense per unit loss knowing only $\bar{o}$, suppose first the decision maker always protects, then $M = C/L$. Conversely, if the decision maker never protects then $M = \bar{o}$. Hence if the decision maker knows only $\bar{o}$, M can be minimised by either always or never taking precautionary action, depending on whether $C/L < \bar{o}$, or $C/L > \bar{o}$, respectively. The mean expense per unit loss associated with a knowledge of climatology only is therefore

$$M_{\text{cli}} = \min\left(\frac{C}{L}, \bar{o}\right) \tag{63}$$

The value V of forecast information is defined as a measure of the reduction in M over M_{cli}, normalised by the maximum possible reduction associated with. a perfect deterministic forecast, i.e.

$$V = \frac{M_{\text{cli}} - M}{M_{\text{cli}} - M_{\text{per}}} \tag{64}$$

For a forecast system which is no better than climate, $V = 0$; for a perfect deterministic forecast system $V = 1$.

As discussed in Section 8, an ensemble forecast gives hit and false-alarm rates $H = H(p_t)$ and $F = F(p_t)$, as a functions of probability thresholds p_t (see (59)). Hence V is defined for each p_t, i.e. $V = V(p_t)$. Using (61), (62) and (63)

$$V(p_t) = \frac{\min\left(\frac{C}{L}, \bar{o}\right) - F(p_t)\frac{C}{L}(1-\bar{o}) + H(p_t)\bar{o}\left(1-\frac{C}{L}\right) - \bar{o}}{\min\left(\frac{C}{L}, \bar{o}\right) - \bar{o}\frac{C}{L}} \tag{65}$$

For given C/L and event E, the optimal value is

$$V_{\text{opt}} = \max_{p_t}[V(p_t)] \tag{66}$$

Figs. 15 -16 show examples of optimal value as a function of user cost/loss ratio. Fig. 15 is for the ECMWF day-6 ensemble weather prediction system and the event $E_{<4}$ (as in Fig. 13). The solid curve is the optimal value for the ensemble system, the dashed curve shows value for a single deterministic forecast (the unperturbed "control" integration in the ensemble). Peak value tends to occur for $C/L \square \bar{o}$; for such users, it makes little difference to the climatological expense M_{cli} whether they always protect, or never protect. The figure also illustrates the fact that the ensemble forecast has greater "value" than a single deterministic forecast. For some cost/loss ratios (e.g. $C/L > 0.6$), the deterministic forecast has no value, whereas the ensemble forecast does have value. The reason for this can be understood in terms of the fact that for a probabilistic forecast, different users (with different C/L) would take precautionary action for different forecast probability thresholds. A user who would suffer a catastrophic loss ($C/L \ll 1$) if E occurred, would take precautionary action even when a small probability of E was forecast. A user for whom precautionary action was expensive in comparison with any loss ($C/L \sim 1$) would take precautionary action only when a relatively large probability of E was forecast. The result demonstrates the value of a reliable probability forecast.

Fig. 16 shows optimal value from the ECMWF seasonal integrations described in section 7. Two different events are shown based on seasonal-mean rainfall forecasts over all regions in the tropics. They are: rainfall less than normal, and rainfall less than one standard deviation below normal (the latter being a possible objective definition of drought). It can be seen that the peak value for the two events occurs at different C/L, arising because the two events have different values of $\bar{o}$. The figure also shows that the value of the forecasts of the more extreme "drought" event is more valuable than the forecasts of rainfall simply below normal. It would appear that the reason for this is that such events tend to be forecast more during El Niño years, when seasonal predictability is high, than during other years, when predictability is lower.

An absolute "dollar" estimate of value can be attached to these curves, providing the user knows the "dollar" value

of their cost and loss. An attempt to do this for seasonal forecasts in southern Africa suggests, for some applications, a potential value of O(10^9 \$) (M. Harrison and N. Graham, personal communication).

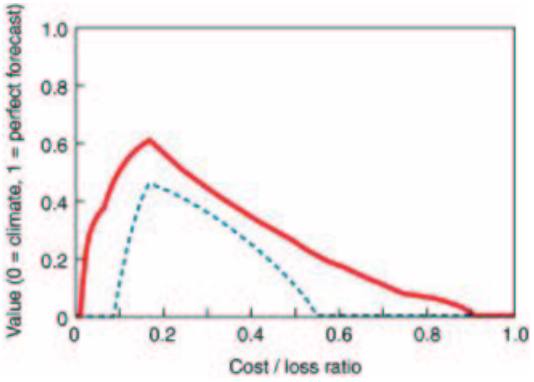

Figure 15: Potential economic value of the ECMWF ensemble prediction system as a function of user cost/loss ratio of day 6 weather forecasts over Europe (for the period December 1998 - February 1999) for the event $E_{<4}$: 850 hPa temperature at least 4°C below normal. Solid: value of the ECMWF ensemble prediction system. Dashed: value of a single deterministic forecast (the unperturbed "control" forecast of the EPS system). From *Richardson* (1999).

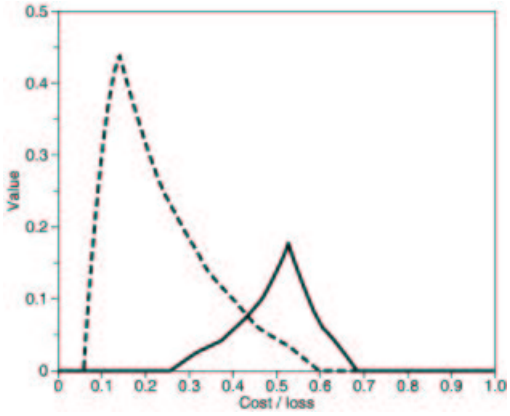

Figure 16: Potential economic value of seasonal ensembles as a function of user cost/loss ratio in the tropics, for the events $E_{<0}$: seasonal-mean rainfall below normal (solid), and $E_{<-\alpha}$: seasonal-mean rainfall at least one standard deviation below normal (dashed), based on 9-member ensembles of integrations with the ECMWF model over the years 1979-93, run with prescribed observed SST.

10. Concluding remarks

Our climate is a complex nonlinear dynamical system, with spatial variability on scales ranging from individual clouds to global circulations in the atmosphere and oceans, and temporal variability ranging from hours to millenia. Climate scientists interact with society through the latter's demands for accurate and detailed environmental forecasts: of weather, of El Niño and its impact on global rainfall patterns, and of man's effect on climate. The complexity of our climate system implies that quantitative predictions can only be made with comprehensive numerical models which encode the relevant laws of dynamics, thermodynamics and chemistry for a multi-constituent multiphase fluid. Typically such models comprise some millions of scalar equations, describing the interaction of circu-