A method for ensemble expansion and improved definition of forecast distributions from climate simulations

[2]*Hydrologic Research Center, San Diego, USA*

SUMMARY

Because of the inherently chaotic nature of the atmosphere, ensemble simulations are required to characterize a model's response to the prescribed boundary forcing in probabilistic terms, particularly if the focus is on the probabilities of extreme events. At the same time, substantial computer resources are needed to produce routinely ensemble seasonal climate forecasts of sufficient size to make suitably reproducible estimates of such probabilities. We describe a method for artificially expanding the effective number of members in ensemble climate simulations on a seasonal basis, thereby reducing uncertainty in estimated probability distributions. As described here, the method involves calculating seasonal statistics using monthly values from all possible combinations of ensemble members. Under certain assumptions, this method is able to increase the effective ensemble size of an N-member M-month seasonal forecast by a factor of (asymptotically) M. One key assumption in this regard is that, aside from the effects of prescribed boundary conditions, the month-to-month values from a particular ensemble member are linear independent. This paper describes the behaviour of the ensemble expansion technique using both idealized and actual ensemble forecast data under a variety of conditions, drawing comparisons with an alternative parametric approach for ensemble expansion. A method for testing the assumption of linear independence in model simulations is also presented.

KEYWORDS: Probability distributions　Re-sampling　Seasonal climate prediction

1. INTRODUCTION

Atmospheric general circulation models (AGCMs) driven with observed, or predicted, sea surface temperatures (SSTs) can reproduce aspects of the observed interannual variability in the large-scale circulation, and of such socio-economically important variables as temperature and precipitation (e.g. Graham 1994; Rowell, *et al.* 1995; Kumar *et al.* 1996; Davies *et al.* 1997; Shukla 1998; Mason *et al.* 1999; Branković and Palmer 2000; Frederiksen *et al.* 2001; Goddard *et al.* 2001). The sensitivity of simulated atmospheric circulations, and the resulting need for ensemble simulations, has spurred considerable innovative work concerning the interpretation and potential use of ensemble seasonal climate predictions. In this regard, ensemble climate simulations were used first to establish reproducible estimates for certain low-order statistics (e.g. the mean) of a model's response to altered boundary conditions (generally SST), and to explore 'potential predictability'—the degree to which individual ensemble members reproduce aggregate ensemble statistics (i.e. a model's ability to predict its own behaviour, e.g. Rowell *et al.* (1995)). More recently there has been increasing interest in using ensemble results to establish, compare, and potentially use the information contained in the probability distributions of the results (Harrison *et al.* 1995; Anderson 1996; Anderson and Stern 1996; Déqué 1997, 2001; Ward and Navarra 1997; Mason and Graham 1999, 2002; Palmer *et al.* 2000; Palmer 2002; Wilks 2002). Indeed, efforts to apply climate prediction information have shown that, because the predictability of the actual climate response is inherently limited, and because the cost-benefit functions for potential applications of climate forecast information are frequently complicated (or at least discontinuous), probabilistic climate forecasts are a necessity if appropriate

† Corresponding author: Hydrologic Research Center, 12780 High Bluff Drive, Suite 250, San Diego, CA 92130, USA. e-mail: ngraham@hrc-lab.org

responses (hedges) to those forecasts are to be developed (e.g. Georgakakos *et al.* 1998; Palmer *et al.* 2000).

Although the number of simulations necessary to obtain an adequately reproducible estimate of the mean seasonal climate response to particular boundary forcing is generally not prohibitively large (Déqué 1997; Kumar *et al.* 2001), and often practicable for both seasonal predictions and multi-decadal validation simulations, this is not the case for establishing more or less reproducible estimates of the true underlying probability-density or cumulative-density functions (PDFs or CDFs) from the discrete probability mass functions (PMFs) of a simulated variable for an individual season, especially if the concern is the likelihood of more extreme outcomes. This is true for actual seasonal climate predictions, as well as for the extended retrospective simulations required to relate model behaviour to real world outcomes in order to develop appropriate strategies for end users to apply such information, and to quantify potential long-term socio-economic benefits that could be accrued through the application of such strategies.

The potential practical benefits and computational costs of producing large ensembles motivate the development of methods for increasing the accuracy of ensemble-derived PMFs without actually performing additional model integrations (artificial ensemble expansion). Wilks (2002) has explored this idea in the context of ensemble numerical weather prediction using a parametric approach. In this methodology, the parameters of a selected statistical distribution are fitted to the ensemble output, and that distribution is then used to estimate exceedence probabilities. The results showed that parametric ensemble expansion has the potential to improve estimates of the probabilities of extreme events, subject to the character of model errors and appropriate choice of statistical distribution.

We present a re-sampling-based ensemble expansion method, and explore its effectiveness from the perspective of seasonal ensemble climate forecasting using both synthetic and actual AGCM seasonal forecasts. The basis of this method is different from the parametric approach explored by Wilks (2002) and, as shown in our results, has both comparative strong points and weaknesses in various situations.

In the presentation that follows section 2 describes the models and model datasets used in the analyses; section 3 describes the methodology, provides some examples with idealized and actual data, and outlines a method for testing for linear independence. Section 4 gives a brief summary and discussion.

2.   MODEL DATA

The AGCM data used in this study come from the ECHAM3 climate model developed at the Max Planck Institute for Meteorology at the University of Hamburg. For the experiments described here the model was configured using 19 vertical layers and triangular-42 (T42) spectral truncation, giving a physical grid resolution of approximately 2.8° latitude and longitude. The model includes interactive prognostic clouds, a five-layer soil temperature model, spatially varying vegetation, and an explicit representation of the planetary boundary layer. The basic formulation of the model is described in DKRZ (1992), and some baseline characteristics are given in Roeckner *et al.* (1992). The model (superseded by newer versions) has been used in a wide variety of studies of climate variability and predictability (e.g. Bengtsson *et al.* 1993; Barnett 1995; Graham and Tyree 1998; Goddard and Graham 1999; Goddard and Mason 2002).

Results from two ensemble simulations with the ECHAM3 model are used in this paper. One of these is a ten-member ensemble covering the period 1950–98. The SST boundary conditions for this simulation (designated ECHAM-AMIP here) come from

the 'reconstructed' dataset described by Smith *et al.* (1996). The second ensemble simulation is a 100-member ensemble (designated ECHAM-100), using September 1997 SST anomalies (US National Centers for Environmental Prediction (NCEP) optimal interpolation analysis (Reynolds and Smith 1994) from the last week in September 1997) added to the long-term climatologies for September through December (thus these simulations are a retrospective persisted SST forecast for October–December 1997).

Initial conditions for these simulations were prepared as follows. A set of ten 'seed' initial conditions were taken from the ECHAM-AMIP simulations described above—these initial conditions come from multi-decade ensemble simulations and, aside from the effects of prescribed SSTs, are completely independent. Simulations were run out from each of these for 1 September to 31 December 1997, with restart files written out for 2–10 September. These restart files, re-dated to 1 September, provided restart conditions for an additional 90 (9 days × 10 ensembles) simulations. The entire month of September (using the SST data for 1–30 September) was then used for model equilibration, and to allow the 90 sets of ensemble 'siblings' to diverge (thus, the ensemble members diverge because of different initial conditions rather than different SSTs). The results from the resulting set of 100 ensembles (10 originals + 90 siblings) for October–December are used for the results presented here. Though the restart conditions provided by this technique are not completely independent, the 90 individual sibling members do have 30 days in which to diverge (well beyond the timescale of dynamical predictability), and it is unlikely that any residual interdependence has much effect on the results presented here.

3. METHODOLOGY AND RESULTS

(a) Theory

Consider an N-member ensemble of monthly forecasts covering a season composed of M months, with each member denoted as $f_{N,M}$. Ordinarily N seasonal totals F_N are formed as:

$$F_J = \sum_{I=1}^{M} f_{J,I} \quad J = 1, N, \tag{1a}$$

or equivalently as seasonal averages:

$$F_J = \frac{1}{M} \sum_{I=1}^{M} f_{J,I} \quad J = 1, N. \tag{1b}$$

Such sums will be referred to as 'ordinary'. Assume for the moment that sequential monthly values of a given variable within a particular ensemble member are linearly independent, beyond the effects of prescribed lateral boundary conditions (e.g. SSTs) common to all ensemble members, so that:

$$\langle f_{J,I} f_{J,I+1} \rangle = \langle f_{K,I} f_{L,I+1} \rangle \quad K \in 1, N, \quad L \in 1, N, \tag{2}$$

where the angle brackets indicate the expectation of the enclosed expression. This assertion (denoted here as 'internal linear independence') strictly implies that linear correlations introduced by dynamical memory, or by such non-prescribed boundary conditions, are negligible (and, more weakly, that knowledge of $f_{J,I}$ contains little information concerning $f_{J,I+1}$). (Of course, this is not necessarily the case—such

correlations might arise from: (i) synoptic fluctuations that have sufficiently long time-scales, and overlap the end of one month and the beginning of the next; (ii) natural low-frequency atmospheric variability unrelated to horizontal boundary conditions; (iii) the effects of non-prescribed boundary conditions such as soil moisture and snow cover; or (iv) interactions between factors (i) to (iii).) Under this assumption of internal linear independence, a larger and equally probable set of seasonal forecasts can be generated as:

$$F_L^* = f_{I,1} + f_{J,2} + \cdots + f_{K,M} \quad I \in 1, N, \quad J \in 1, N, \quad \ldots,$$
$$K \in 1, N, \quad L = 1, N^M. \tag{3}$$

Sums formed in this manner will be denoted 'permuted'. This larger set of forecasts has the potential advantage of allowing better characterization of the spread of the actual PMFs (or CDFs) of forecast values when an ensemble of simulations are conducted with a particular set of prescribed lateral-boundary conditions (note that the estimated ensemble mean, and its associated uncertainty, are the same as for the ordinary sums). To the degree that the forecast model results provide information about the future state of the real world, this advantage translates into an improved ability to estimate the probability of actual events. We refer to this technique as ELVIS (EnsembLe Variability from Inferred Statistics).

If it is assumed that the monthly values ($f_{N,M}$; and hence the ordinary and permuted sums or averages, F_J and F_L^*) are normally distributed, then the behaviour of the permuted and ordinary sums can be conveniently compared in terms of variance using the χ^2 distribution and its associated degrees of freedom. If it is found that for an N-member ensemble and M-month season the central tendency of $\langle F^{*2} \rangle$ is unbiased, and the spread of the distribution of $\langle F^{*2} \rangle$ is smaller for permuted sums than for ordinary sums (i.e. the degrees of freedom, df, is larger), then the permuted sums will produce more accurate estimates of the true model probabilities (taken from a hypothetical very large ensemble).

Consider a simple idealized case of a two-month seasonal total with linearly independent, normally distributed, monthly forecast values of equal variance ($\sigma_1^2 = \sigma_2^2$), where:

$$\sigma_I^2 = \langle (f_{J,I} - \overline{f_I})^2 \rangle, \quad J \in 1, N, \quad I \in 1, 2 \tag{4}$$

(here the overbar represents the ensemble mean). Clearly there are the N ordinary sums, and given that:

$$\langle F^2 \rangle = \frac{1}{N-1} \sum_{J=1}^{N} (F_J - \overline{F})^2, \tag{5}$$

then $(N-1) \times \langle F^2 \rangle$ is distributed as χ^2_{N-1}, where $N-1$ indicates df. In contrast, the set of permuted sums is made up of N^M members, each constructed from M monthly forecast values. If one such set of N is selected that is constructed without repetition of a given monthly value (the ordinary sums form one such set), then all the $N^M - N$ others (e.g. all of the non-ordinary sums) are constructed with at least one repeated monthly term. The appearance of these repeated values greatly reduces the df associated with estimates of variance of the permuted sums. The result is that for permuted sums the sum of squares $(N^M - 1) \times \langle F^{*2} \rangle$, where:

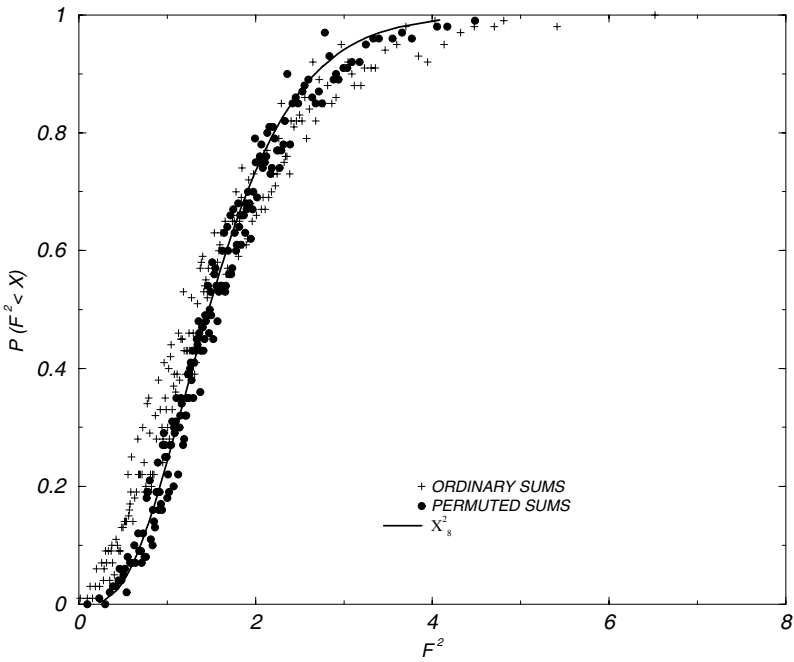$$\langle F^{*2} \rangle = \frac{1}{(N^M - 1)} \sum_{L=1}^{N^M} (F_L^* - \overline{F}^*)^2, \tag{6}$$

Figure 1. Cumulative frequency distributions of the variance of the ordinary (crosses) and permuted (circles) sums (see text) for simulated ensemble forecasts constructed from linearly independent, normally distributed (0, 1) random variables, with the number of 'months' comprising the 'season', $M = 2$, and the number of ensemble members, $N = 5$, from 10 000 Monte Carlo experiments. The solid line shown is for a χ^2 distribution with degrees of freedom, df, given by $M \times (N - 1) = 8$. F^2 is the estimated ensemble variance (as in the discussions of Eqs. (5) and (6)).

is distributed as χ^2 with df $= M \times (N - 1)$ rather than df $= N^M - 1$. Nevertheless, $M \times (N - 1)$ is larger than $N - 1$, and this potential increase in effective sample size can be realized at modest computational cost.

In the discussion that follows, the inverse χ^2 distribution is used to demonstrate that the formulation of df for $\langle F^{*2} \rangle$ is correct, and to compare its behaviour under different circumstances with that of $\langle F^2 \rangle$. Figure 1 shows distributions of $\langle F^2 \rangle$ and $\langle F^{*2} \rangle$ for a case with $M = 2$, $N = 5$ and $\sigma_1 = \sigma_2 = 1$, derived from a hypothetical very large ($n = 10\,000$) Monte Carlo population of linearly independent, normally distributed $(0, \sigma_I^2)$ monthly forecasts; the χ^2 distribution for $\langle F^{*2} \rangle$ is also shown and the fit is clear. In this example, the permutation procedure has raised the effective sample size from 5 to 9 (df from 4 to 8).

The second case shown (Fig. 2) is for an ensemble size $N = 10$, $M = 3$ and $\sigma_1 = \sigma_2 = \sigma_3 = 1$. The large reduction in uncertainty achieved by using the permuted sums in estimating the ensemble variance is apparent, in this case yielding a nearly three-fold increase (from 10 to 28) in effective ensemble size (df increases from 9 to 27). Again, the fit of the χ^2 distribution is excellent.

Next, consider the case in which the monthly variances are not equal. Let $f_{i,1} = g_{i,1} + ag_{i,1}$ (where a is a scalar constant), $f_{i,2} = h_{i,2}$, and assume that both $h_{i,j}$ and $g_{i,j}$ are linearly independent $N \sim (0, \sigma_I^2)$ random variables. Assuming a is positive, the variance of $f_{i,1}$ will be larger than that of $f_{i,2}$. In such a case with $M = 2$ the ensemble
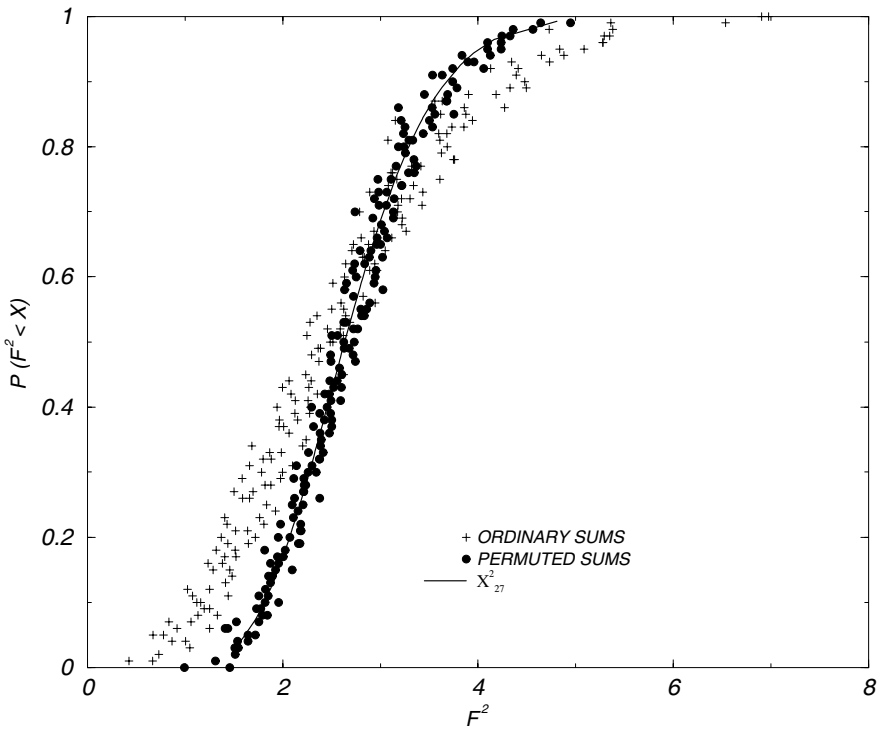
Figure 2.   As Fig. 1 but with $M = 3$, $N = 10$ and χ^2 distribution with df $= 27$.

sums could be written:

$$
\begin{aligned}
(g_{1,1} + ag_{1,1}) + h_{1,2} &= F_1 \\
(g_{2,1} + ag_{2,1}) + h_{2,2} &= F_2 \\
&\vdots \\
(g_{N,1} + ag_{N,1}) + h_{N,2} &= F_N
\end{aligned}
\tag{7}
$$

Because the terms inside the brackets in Eq. (7) remain together, the central tendencies of $\langle F^2 \rangle$ and $\langle F^{*2} \rangle$ are equal (i.e. the latter are unbiased). However, as seen in Fig. 3, the distribution of $\langle F^{*2} \rangle$ does not follow $\chi^2_{(M \times (N-1))}$. This is because the values from month 1 ($f_{i1} = g_{i,1} + ag_{i1}$) dominate the variance of the permuted sums, thus reducing the df. The effect is to push the distribution of the permuted sums (6) towards the distribution of the ordinary sums (5). As noted above, the expectation of $\langle F^{*2} \rangle$ is unaffected, but the benefits of permutation (as expressed by reduction in the scatter of the estimates of $\langle F^{*2} \rangle$ in comparison with $\langle F^2 \rangle$) are diminished. If the monthly variances (σ_i^2) are known, an approximate (and ad hoc) correction for the degrees of freedom in the χ^2 test is:

$$
R = \left(\frac{T_1}{T_2} \right)^4,
\tag{8a}
$$

where,

$$
T_2 = \left(M \sum_{i=1}^{M} \sigma_i^2 \right)^{1/2},
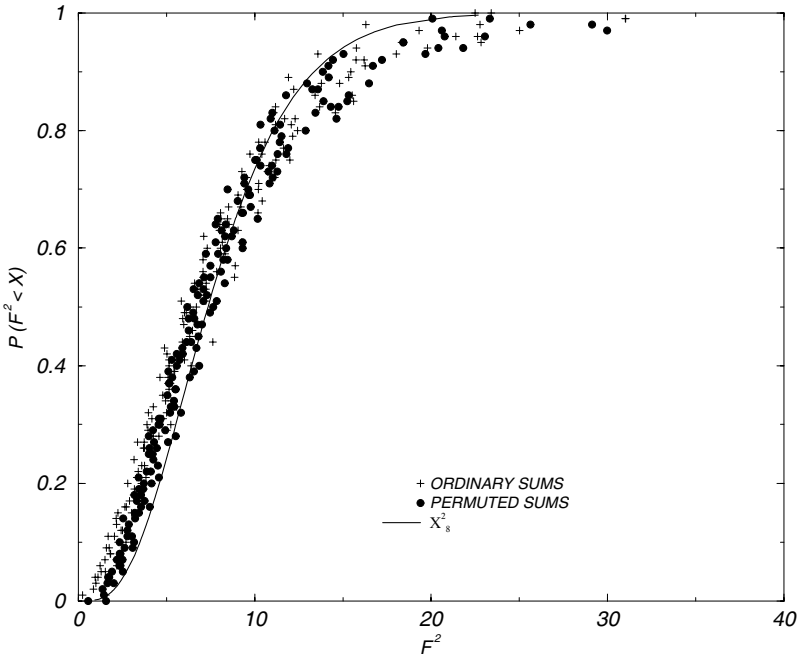\tag{8b}
$$

Figure 3.   As Fig. 1, but for monthly values with unequal variances ($\sigma_1 = 1$, $\sigma_2 = 3$). Note that the χ^2 distribution (for df = 8) overestimates the decrease in the spread of the variance of the permuted sums.

and

$$T_1 = \sum_{i=1}^{M} \sigma_i, \tag{8c}$$

so that

$$R \times N \times M \times \langle F^{*2} \rangle \tag{9}$$

is distributed approximately as $\chi^2_{(R \times N \times (M-1))}$.

Note that T_2 gives the sum of the monthly standard deviations if the monthly variances had been equal, while T_1 gives the actual value. R has a maximum value of 1.0 when $\sigma_1 = \sigma_2 = \cdots = \sigma_M$, and a minimum value of M^{-2} when all of the variance of the sums is accounted for by a single month (for example $\sigma_1 = 1$ and $\sigma_2 = \cdots = \sigma_M = 0$). Figure 4 gives an example for a case where $\sigma_1 = 1$, $\sigma_2 = 3$ and $\sigma_3 = 6$ and $N = 10$. The adjusted distribution has a relatively good fit to the data while the unadjusted curve fits poorly.

Next consider the case where the stipulation of internal linear independence does not hold, and some internal (non-prescribed) memory in the model climate system produces some non-zero correlation between sequential monthly values within a given ensemble member. To explore this case, Eq. (7) is rewritten so that $f_{i,2} = ag_{i,1} + h_{i,2}$:

$$\begin{aligned}
g_{1,1} + (ag_{1,1} + h_{1,2}) &= F_1 \\
g_{2,1} + (ag_{2,1} + h_{2,2}) &= F_2 \\
&\vdots \\
g_{N,1} + (ag_{N,1} + h_{N,2}) &= F_N
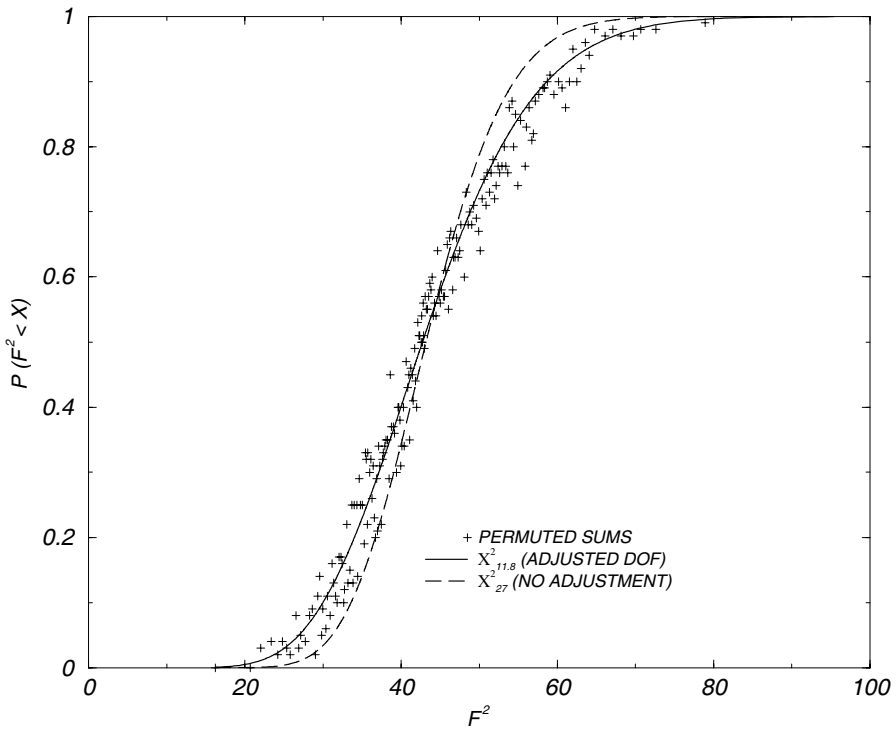\end{aligned} \tag{10}$$

Figure 4.   As in Fig. 2, but for monthly values with unequal variances ($\sigma_1 = 1$, $\sigma_2 = 3$, $\sigma_3 = 6$), and in this case the χ^2 distribution for the ordinary sums is not shown. χ^2 distributions are shown for df $= 27$ (not accounting for the differences in variance) and for df $= 11.8$ (from Eq. (7)).

where a now produces the correlation between $f_{i,1} = g_{i,1}$ and $f_{i,2} = ag_{i,1} + h_{i,2}$. It is clear that, for ordinary sums, Eqs. (7) and (10) are equivalent and give the familiar expression for the variance of the sums:

$$\langle F^2 \rangle = \langle f_1^2 \rangle + 2 \langle f_1 f_2 \rangle + \langle f_2^2 \rangle = \sigma^2 \{(1+a)^2 + 1)\}, \tag{11}$$

with correlation (for Eq. (11)):

$$R_{1,2} = \frac{a}{(a^2 + 1)^{1/2}}. \tag{12}$$

Note that the equivalence between Eqs. (7) and (10) for ordinary sums does not hold for permuted sums, because the expressions inside parentheses in Eq. (10) are bound together so that the covariance term in Eq. (11) frequently does not contribute to the total variance (contrast this with the situation for unequal variance in Eq. (7)). For example, in the $M = 2$ case, the N^M permuted sums are comprised of the N ordinary sums (where $ag_{i,1}$ remains with $g_{i,1}$, and the covariance term in Eq. (11) is non-zero), and a group of $N^M - N$ sums where $ag_{i,1}$ is with $g_{j,1}$ ($i \neq j$) and the expected value of the covariance term is zero. For this latter group (and under the stipulation that $\langle g_{i,1}^2 \rangle = \langle h_{i,2}^2 \rangle = \sigma^2$), $\langle F^{*2} \rangle = \sigma^2 \times (2 + a^2)$ is smaller than Eq. (9) by $2a\sigma^2$, and its distribution behaves as that of the sums of three random variates (with standard deviations σ, σ, and $a\sigma$). Thus for correlated monthly values (i.e. with the correlations deriving from internal processes) in the $N = 2$ case, the population of $\langle F^{*2} \rangle$ is comprised of members from two distributions with different central tendencies (and likely different shapes).
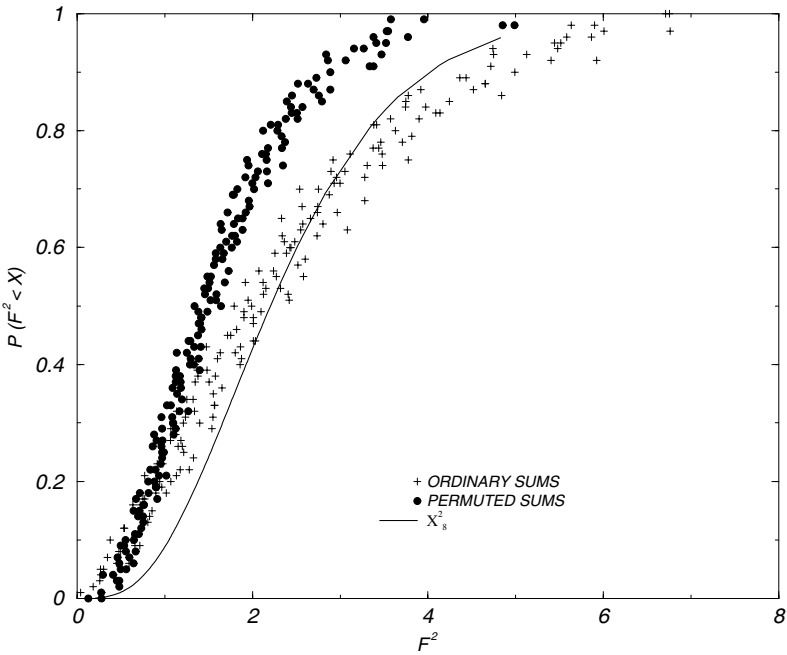
Figure 5. As Fig. 1, but with the nonlinearly independent variables ($R_{1,2} = 0.5$). Note that the permuted sums are biased low, and that the χ^2 distribution (for the expected sum of squares ($\langle F^2 \rangle = 2$) and df = 8) fits very poorly.

Figure 5 shows the distributions of $\langle F^2 \rangle$ and $\langle F^{*2} \rangle$ for the $M = 2$, $N = 5$ case with $\sigma_1 = \sigma_2 = 1$ and $R_{1,2} = 0.5$. The bias between the central tendencies of the distributions is obvious, as is the poor fit to $\langle F^{*2} \rangle$ of the χ^2 distribution (calculated using the expected variance from Eq. (8)). For $N > 2$ the situation becomes more complicated with varying numbers of correlated and partially correlated terms involved in the permuted sums, but the problems of bias and mixed distributions remain.

To summarize: the discussion above, concerning the distributions of the variances of ordinary and permuted sums constructed from normally distributed monthly values, highlighted the three main points listed below:

(i) For the case of no internal linear independence and equal monthly variances, the behaviour of the permuted sums is straightforward, and provides an increase in the effective ensemble size by a factor of M as measured by the df for the χ^2 distribution.

(ii) If the monthly variances are unequal, the effect is to reduce the df of the χ^2 distribution for the permuted sums towards that of the ordinary sums. An approximate ad hoc correction for df is given, for when the variances of the monthly values are known.

(iii) Where internal (and positive) autocorrelations are present:

- they act to inflate the expectation of the variance of the ordinary sums;
- they result in a low bias for $\langle F^{*2} \rangle$; and
- the distribution of $\langle F^{*2} \rangle$ is made up of an aggregate of χ^2 distributions with different expectations and degrees of freedom.

(b) Examples

In this section we present examples of the behaviour of the ELVIS procedure in idealized and more realistic cases. For the idealized cases, the following procedure is used:

(i) Generate large samples (typically 10 000) of M-month (typically $M = 3$) 'ensemble realizations' with month-to-month lag correlations $R_{N,N+1}$; these comprise the 'underlying population', and come either from a specified statistical distribution, or from AGCM results. Calculate the seasonal sum for each, sort these 'ordinary' sums into equally probable bins (typically 10) and record the breakpoints between bins.

(ii) (a) Randomly draw N of the M-month 'ensemble forecasts' from the population generated in (i) above (typically $N = 10$ and $M = 3$) and obtain the N ordinary sums and N^M permuted sums for each sample. Calculate the PMFs and CDFs from each population using the breakpoints from (i).

(ii) (b) In cases where the parametric expansion method is evaluated, the following approach is used. From the same N M-month forecasts described in (ii) (a), calculate the sample mean and standard deviation, and use a t-distribution (df $= N - 2$) to estimate the bin probabilities using the breakpoints from step (i).

(iii) Accumulate bin-wise total absolute probability errors (the differences between the sample probabilities and the reference values from (i)) for the PMFs and CDFs obtained in (ii) (a) and (b).

(iv) Repeat (ii) and (iii) many (typically 1000 times) to derive average absolute probability errors (AAPEs) for the PMFs and CDFs.

(Note that because the AAPE is uncorrected for the number of ensembles performed, its expectation will decrease as the number of ensembles increases (assuming the same underlying process). The measure is used here specifically because it has this property, i.e. to show how the magnitude of probability errors changes with ensemble size and expansion method.)

Figure 6(a) shows the AAPEs for the PMF for idealized linearly independent, normally distributed ensemble forecasts with $M = 3$ and $N = 10$. For the ordinary sums, the errors AAPEs are nearly constant at 0.07, or 70% of the underlying PMF values of 0.10. For permuted sums, the AAPE curve is U-shaped, with the smallest improvements at the extremes and the largest improvements near the middle of the distribution (about 70% and 15% of the AAPEs for the ordinary sums, respectively). For comparison, the AAPEs for the χ^2-equivalent ensemble size (i.e. $N = 28$, df $= 27$) are also shown. As with the ordinary sums for $N = 10$, the curve is flat, with the AAPEs at the tails of the distribution close to those from the permuted sums but considerably larger in the middle of the distribution. The distribution of AAPEs for the parametric method is also U-shaped, with errors generally slightly larger than for the permuted sums.

The U-shaped distribution of the AAPEs for the permuted and parametric sums is the result of improvements in the estimation of the spread of the forecast PDF (e.g. in the Gaussian case, better estimates of the standard deviation) in the absence of any reduction in uncertainty in the central tendency (e.g. the mean). For the permuted sums, this improvement arises from the larger sample size. For the parametric distribution, the reduction in uncertainty arises from the correct assumption that the underlying population is normally distributed. In either case, errors in the means of the ordinary sums are unchanged, resulting in increasing errors towards the tails of the distribution. Tests (not shown) reveal that if the permuted and parametric distributions are adjusted
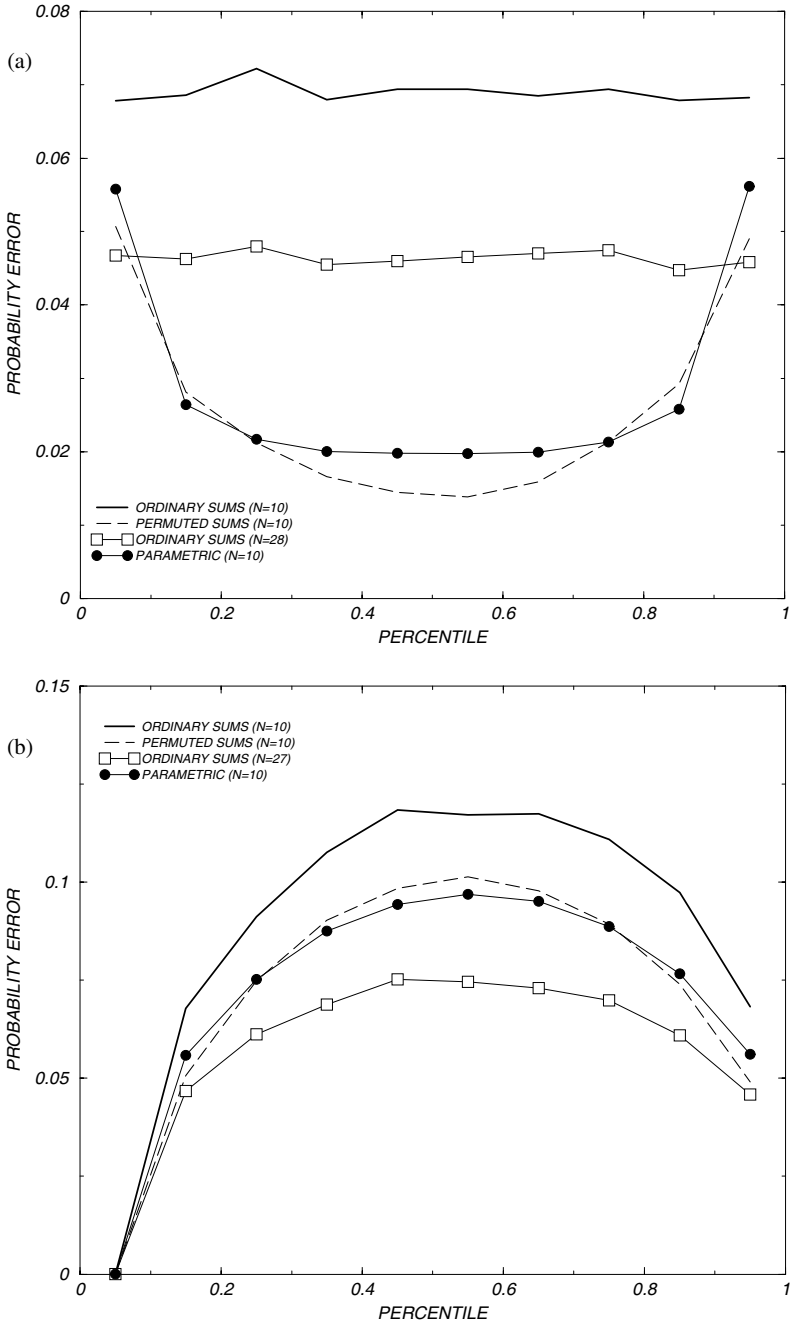
Figure 6. Average absolute probability errors (AAPEs) for a season with the number of months $M = 3$, having individual variances $\sigma_1 = 1$, $\sigma_2 = 1$, $\sigma_3 = 1$, and no serial correlation: (a) discrete probability mass functions (PMFs)) for ordinary and for permuted sums, with numbers of ensemble members $N = 10$ and $N = 28$; see the key for details. (b) As (a) but for cumulative-density function (CDFs)—accumulation begins at the left. All values are plotted at the mid-point of the respective bin.

to have the correct means, the AAPE curves become essentially flat except for the most extreme bins (where the effects of errors in the spread are concentrated).

Figure 6(b) uses the same data as Fig. 6(a), but shows the CDF's, thus giving an idea how the errors shown in Fig. 6(a) accumulate and affect estimates of exceedence probabilities. In this case, although the AAPEs obtained from the permuted sums are considerably smaller than those from the ordinary sums, the improvements are smaller than for the PMFs. Also, the AAPEs for the ordinary sums from the larger ensemble ($N = 28$) are now smaller than those for the permuted sums (tests (not shown) show that this difference is due primarily to errors in the mean of the permuted sums). Nevertheless, the ELVIS procedure gives 90th percentile AAPEs only slightly larger than those for ordinary sums from a 28-member ensemble. The AAPEs for the parametric distribution CDFs are very close to those for the permuted sums.

It is also of interest to examine the behaviour of the permuted sums in cases where the model climate variables have skewed distributions. In this instance, the underlying population was generated from random, linearly independent, log-normally distributed monthly values. The resulting distribution of the seasonal totals from the underlying population is mildly skewed (Fig. 7(a), and resembles that for historical annual precipitation in coastal central California). From the CDF of the AAPEs for this case, Fig. 7(b), it is apparent that although the skew negatively influences the performance of the ELVIS methodology for extreme positive values, the results are nevertheless an improvement over those obtained with ordinary sums. Also shown are the AAPEs obtained from the parametric expansion method, here using the incorrect assumption that the underlying population is normally distributed. The results are inferior to those both for ordinary and permuted sums for bins above the 50th percentile. The relatively poorer performance results from the difference between the actual and assumed background distributions, emphasizing the need for appropriate choice of parametric distribution when using the parametric approach.

The final example using idealized data, addresses the effects of violations of the assumption of linear independence (the presence of internally derived autocorrelations between the monthly values). Figure 8 compares the CDF AAPEs for ordinary and permuted sums, and as derived using the parametric method, for three cases (each with $M = 3$, $N = 10$, and equal monthly variances) in which the month-to-month correlations ($R_{1,2}$ and $R_{2,3}$) are set to 0.0, 0.2 and 0.4, respectively (the results for the first repeated from Fig. 6(b)). The AAPE curves for the permuted sums for these cases progressively approach that for the ordinary sums ($N = 10$) as the correlation increases. Interestingly, the effects at the high extreme (values exceeding the 90th percentile) show little deterioration in performance compared to the ordinary sums from 28-member ensembles. These results (and others not shown) suggest that for this case the ELVIS technique performs no worse than ordinary sums with internally derived autocorrelations up to 0.5, and continues to provide improved performance in the 90th percentile bin. This result is heartening for, as shown later, estimates derived from AGCM simulations suggest that internally derived month-to-month autocorrelations exceeding 0.4 are relatively rare for 2 m air temperature and even more so for precipitation.

The AAPE's for the parametrically fitted distributions are unaffected by the increasing autocorrelations, as both the underlying population and ordinary sums (from which the distribution parameters are estimated) are normally distributed. Thus, in cases where internally derived autocorrelations are high, and assuming the underlying distribution is well known, the parametric method has advantages over the resampling approach.

Analyses similar to the idealized cases described above were also conducted using the ECHAM-100 100-member AGCM ensemble data described in section 2.
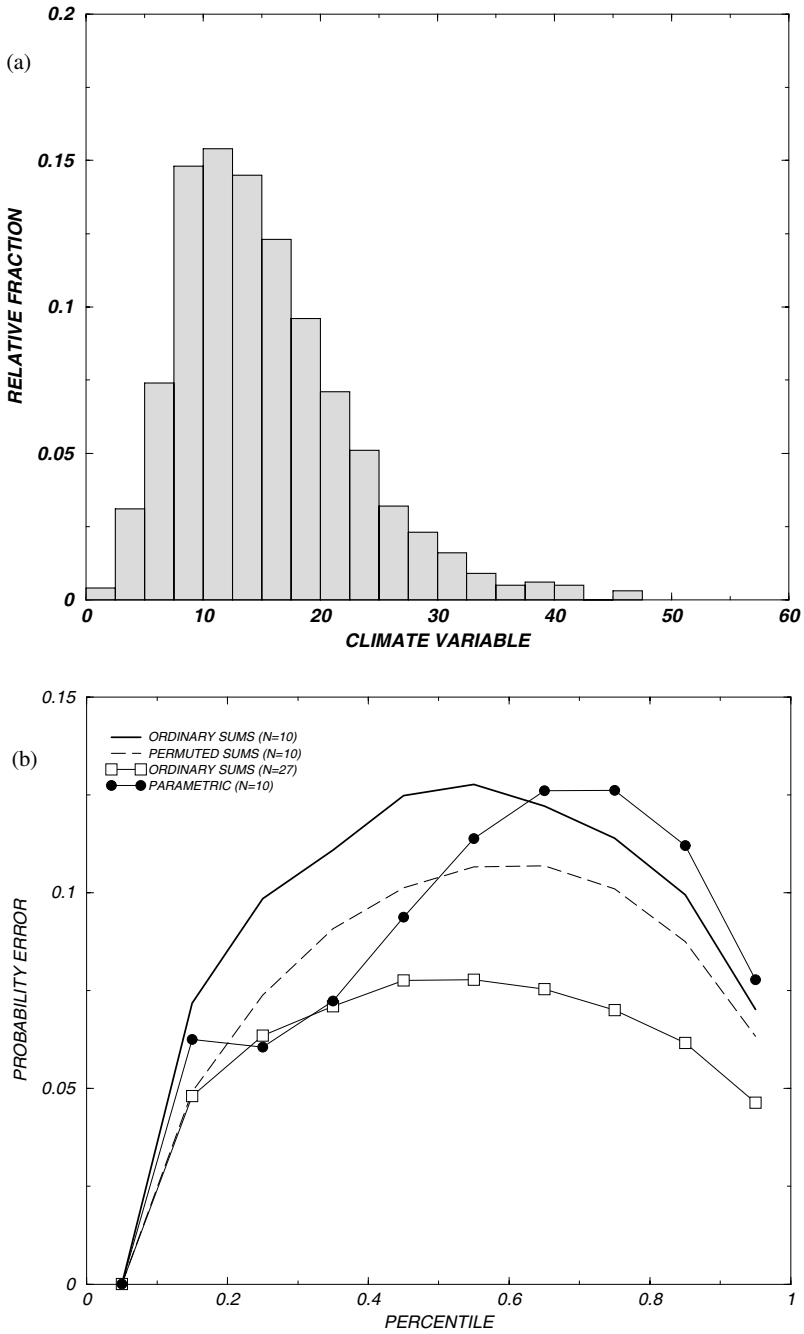
Figure 7. (a) Histogram of the frequency distribution of the underlying log-normal distribution used for the results shown in (b). (b) As Fig. 6(b), but for seasonal totals constructed from log-normally distributed monthly values (the seasonal distribution is shown in Fig. 7(a)). See text for details.
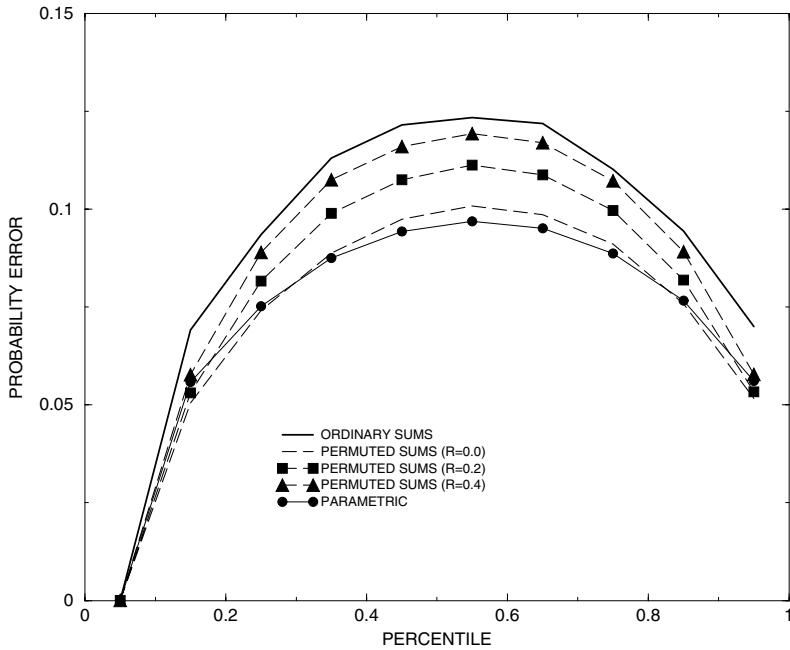
Figure 8.   As Fig. 6(b), but showing the effects of increasing lack of linear independence on the average absolute probability errors (AAPEs) of the permuted sums. Curves show the AAPEs for ordinary sums (solid), and permuted sums with month to month correlations: $R_{1,2} = R_{2,3} = 0.0$ (dashed), $R_{1,2} = R_{2,3} = 0.2$ (dashed with squares), and $R_{1,2} = R_{2,3} = 0.4$ (dashed with open triangles). The AAPE curve for the parametric approach is repeated from Fig. 6(b) for reference (correlations have no effect on the parametric estimates).

For these analyses, the statistically derived 'underlying' population of monthly values (see step (i) above) has been replaced with ensemble model values of central USA precipitation for October, November and December 1997 (a number of other variables and areas were also tested, with similar results). As with the idealized data, Monte Carlo tests were performed drawing random samples of 10 (and 28) 3-month ensembles and calculating ordinary and permuted seasonal sums†. In this example, the variances of the ensemble monthly means are not equal (the standard deviations for October–December are 0.99, 1.27 and 1.13, respectively), but the ad hoc correction given in Eq. (7) suggests that these minor departures from equality have negligible impact on the results. Intra-ensemble month-to-month correlations are also relatively low (0.15, 0.04, and −0.09, respectively), suggesting a minimal level of internally derived memory for regional precipitation in this instance. Figure 9(a) shows the PMF AAPEs for these data. As with the idealized data (Fig. 6(a)) the ELVIS permuted sums give better performance than the ordinary sums. The U-shaped curve seen in the idealized data (Fig. 6(a)) is not apparent (tests (not shown) with idealized data show this is due to the small size of the underlying population). The results for the CDF AAPEs using the AGCM data (Fig. 9(b)) are similar to those obtained from the idealized Gaussian data (integration smoothes the

† An obvious limitation to these trials is that the relatively small size (100) of the underlying population results in considerable repetition in the sampling. Tests with idealized data (not shown) show that this degree of over-sampling produces AAPE curves considerably rougher than those for the much larger populations used in the idealized cases.
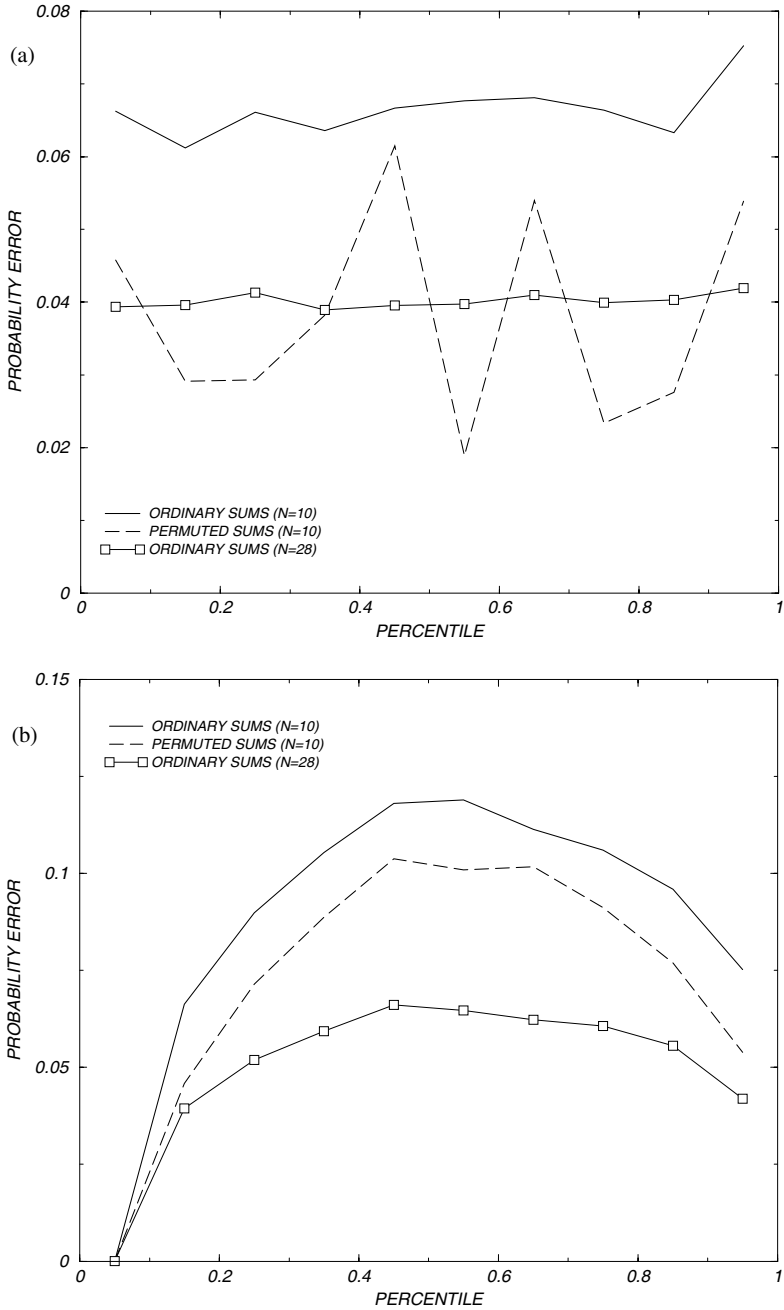
Figure 9. (a) As Fig. 6(a), but for atmospheric general circulation model ensemble forecasts of October–December 1997 precipitation in the central USA; for ordinary sums with $N = 10$ and $N = 28$, and permuted sums from the $N = 10$ population (see key). (b) As (a), but for cumulative-density function average absolute probability errors.

effects of over-sampling). The precipitation data used here are approximately normally distributed, so the parametric method gives results (not shown) much like those from the permuted sums.

(c)   A test for internal linear independence

The discussions earlier (for example, see Figs. 5 and 8) emphasize the need for estimates of the magnitude of internally derived month-to-month memory in model variables. One way to quantify the level of internal memory is to perform repeated ensemble control simulations with fixed prescribed boundary conditions (e.g. SST), and directly measure (for example) month-to-month lagged correlations (see discussions in Koster and Suarez 2001). In this section, we suggest a method to obtain (at least) first order expectations for internal autocorrelations if such control simulations are not available.

Consider an extended (multi-decade) ensemble of N simulations driven with prescribed observed SST. For a given variable, location and month a matrix of autocorrelations like that shown schematically below (for $N = 5$) can be constructed.

$$
\text{Month 1} \quad
\begin{bmatrix}
R_{1,1} & R^*_{1,2} & R^*_{1,3} & R^*_{1,4} & R^*_{1,5} \\
R^*_{2,1} & R_{2,2} & R^*_{2,3} & R^*_{2,4} & R^*_{2,5} \\
R^*_{3,1} & R^*_{3,2} & R_{3,3} & R^*_{3,4} & R^*_{3,5} \\
R^*_{4,1} & R^*_{4,2} & R^*_{4,3} & R_{4,4} & R^*_{2,5} \\
R^*_{5,1} & R^*_{5,2} & R^*_{5,3} & R^*_{5,4} & R_{5,5}
\end{bmatrix}
$$

$$\text{Month 2}$$

The rows and columns represent the ensemble members, the rows are for month 1 (e.g. June), and the columns for month 2 (e.g. July). Intra-ensemble lagged correlations (R) are along the diagonal, and inter-ensemble correlations (R^*) are the off-diagonal values. Because the prescribed boundary conditions affect all ensemble members, systematic differences between on- and off-diagonal correlations indicate the effects of 'memory' arising from sources other than the prescribed boundary conditions, for example from soil moisture or snow cover. In practice, the intra- and inter-ensemble correlations are calculated as single values. For example, if there are five ensemble members and a 20-year simulation, the covariances and variances required to calculate the intra-ensemble lagged correlation are formed through a series of 100 pairs (20 years and five inter-ensemble pairs), yielding a single value of R. Similarly, the inter-ensemble lagged correlation would be calculated from a series of 400 pairs (20 years and 20 intra-ensemble pairs), yielding a single value of R^*. The effect of internal memory is then estimated (assuming linearity) as a residual fraction of variance, then re-scaled as a correlation-like value: $R_{\text{MEM}} = (R^2 - R^{*2})^{1/2}$.

In the examples given here, R_{MEM} is defined only where the difference inside the parenthesis is positive; the few cases of inferred negative correlations are ignored.

As an example of application of this method, we show intra-ensemble and inter-ensemble 1-month lagged correlations (R and R^*, respectively), and their differences (R_{MEM}), for 2 m air temperature and precipitation from the 49-year, ten-member prescribed SST ensemble simulation (ECHAM-AMIP) described in section 2. Figure 10 shows the results for January and February 2 m temperature. The results suggest that, for this model, month-to-month correlations arising from internal processes are generally less than 0.3 over eastern Europe, the north-east of Africa, much of northern North America, and northern Asia. Correlations above 0.4 cover extensive portions

(a) 2 M TEMP: INTRA-RUN JAN-FEB CORRELATION
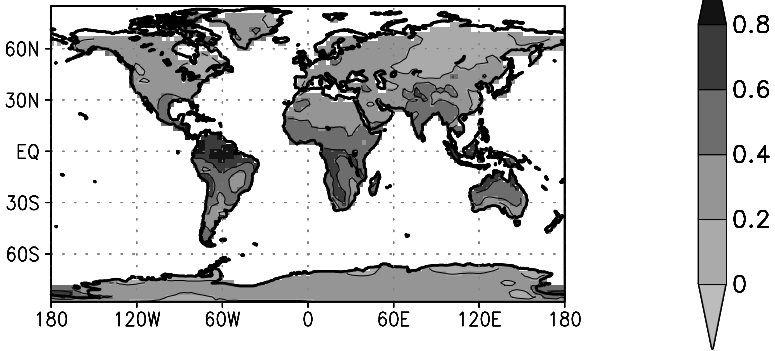
(b) 2 M TEMP: INTER-RUN JAN-FEB CORRELATION
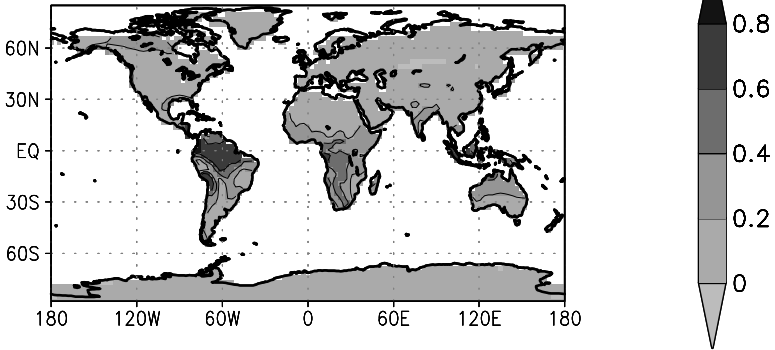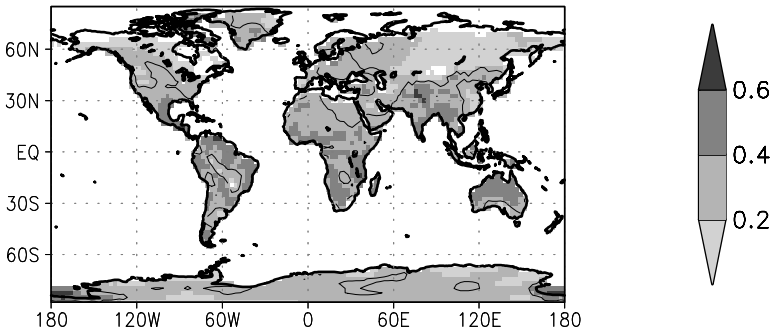
(c) 2 M TEMP: JAN-FEB R_{MEM}

Figure 10.   (a) Intra-ensemble correlations R, and (b) inter-ensemble pointwise correlations R^*, between 2 m air temperatures in January and February from a ten-member 50-year ensemble atmospheric general circulation model simulation; (c) the internal memory variance fraction R_{MEM}, scaled as a correlation-like variable (see text), for February arising from internal model processes. Areas shown as white are: oceanic regions, areas where the results did not exceed the 90% significance level from Monte Carlo tests, and the few areas of inferred negative internally derived autocorrelations.

(a)   2 M TEMP: INTRA−RUN JUN−JUL CORRELATION
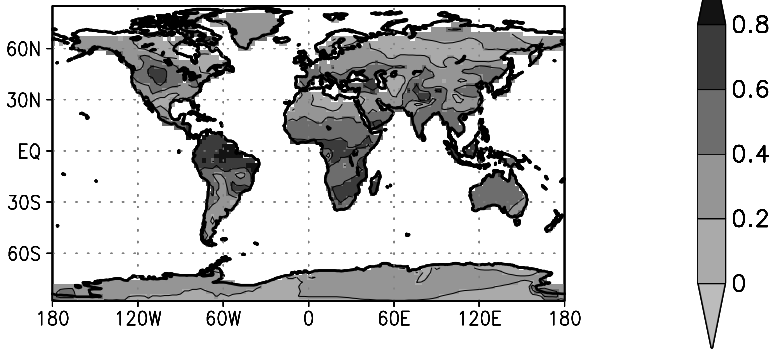
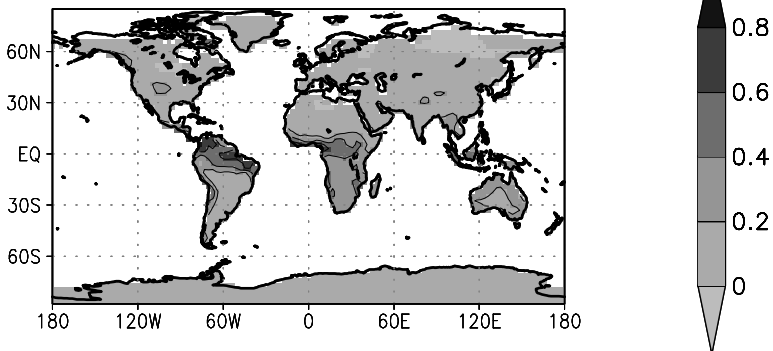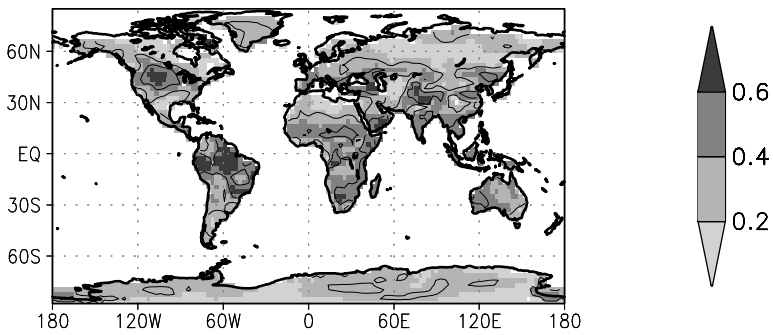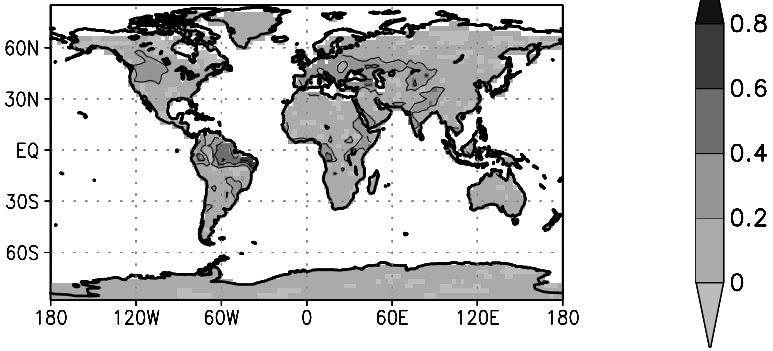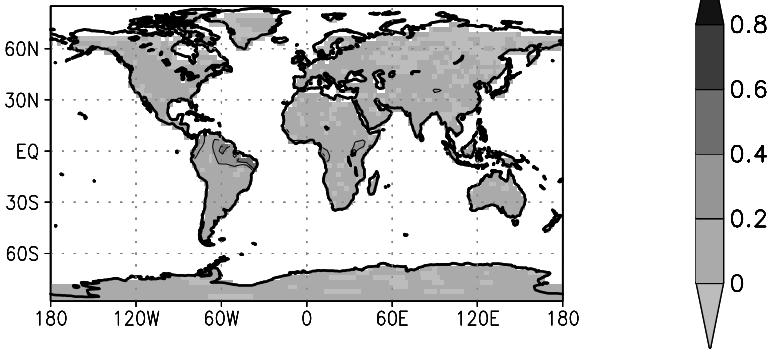(b)   2 M TEMP: INTER−RUN JUN−JUL CORRELATION

(c)   2 M TEMP: JUN−JUL R$_{\text{MEM}}$

Figure 11.   As Fig. 10 but for June and July 2 m temperatures.

(a) PRECIP: INTRA—RUN JUN—JUL CORRELATION

(b) PRECIP: INTER—RUN JUN—JUL CORRELATION
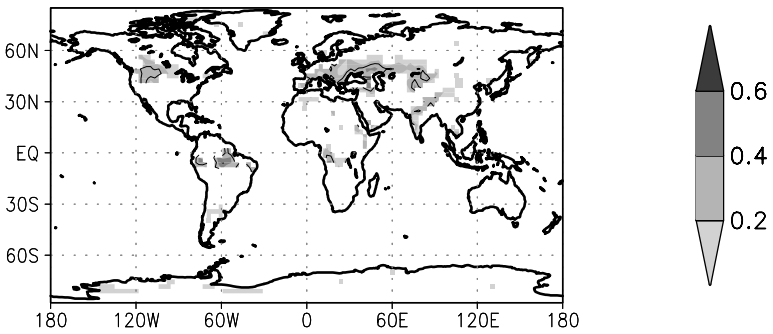
(c) PRECIP: JUN—JUL R$_{MEM}$

Figure 12. As Fig. 10 but for June and July precipitation.

of South America, southern Africa, southern Asia (particularly the Tibetan Plateau), and Australia. In these regions, the enhanced internal memory is likely to result from changes in snow cover and soil moisture. Interestingly, Monte Carlo tests indicate that the correlation differences are significant above the 90% level nearly everywhere, including areas where the autocorrelations are quite small (this is also true over large areas of the oceans (not shown)). Note that from the perspective of independence, it is the magnitude of the internal memory rather than the statistical significance that is of concern.

Figure 11 is like Fig. 10, but for June to July. For these months, the inter-ensemble correlations tend to be much higher than for January to February, with large areas of the northern USA and Canada, Europe, central and southern Asia, central South America, Australia, and Africa showing R_{MEM} values above 0.3 (reaching 0.6 in some areas). Again, changes in soil moisture (and snow cover in a few areas) probably produce these signals. Significance levels are above the 90% level over all continental regions.

Figure 12 is like Figs. 10 and 11, but for simulated June–July precipitation. For precipitation, the intra- and inter-ensemble correlations, and the values of R_{MEM} (Fig. 12(c)) are generally much lower than for 2 m air temperature. Correlations above 0.2 do appear in small regions of North and South America, and in a larger east–west oriented region extending across southern Europe and into west-central Asia, suggesting that the character of simulated precipitation variability in this region is modified somewhat by soil moisture. A similar analysis of January–February precipitation (not shown) shows values of R_{MEM} above 0.2 in only a few scattered areas.

From the perspective of the efficacy of the ELVIS technique, these results, together with those shown in Fig. 8, suggest that for this model ELVIS will provide improved estimates of probability distributions of simulated boreal winter season 2 m air temperature averages in many regions. For boreal summer, improvements can be realized in some areas, but these are less extensive than for winter. For simulated precipitation (subject to analysis of differences between monthly variances and distribution skewnesses) improvements can be realized in most areas for both seasons. Other tests with 500 hPa heights for January–February and June–July (not shown) indicate very low levels of internal memory. This finding is consistent with the idea that month-to-month internal memory in these simulations is principally controlled by soil moisture (snow cover in some areas) and has its most obvious (atmospheric) effects on near-surface variables (e.g. temperature), but little impact on simulated large-dynamics. Tests with another set of ensemble simulations performed with the NCEP MRF9 GCM (not shown), which is also T42 resolution, produced qualitatively similar results and suggests similar conclusions.

4. CONCLUSIONS

For seasonal climate forecasts (or retrospective simulations), ensembles are required to adequately resolve model response to a given evolution of boundary forcing. The potential utility of such forecasts requires that the following four conditions have been met.

(a) Suitably accurate forecasts of the boundary conditions are available.

(b) The AGCM response (described by statistics) to those boundary conditions carries some quantitative information about the response the of actual atmosphere, and thus the future state of the real world.

(c) Extended retrospective simulations are available to relate model behaviour to real world outcomes.

(d) The ensembles (both for retrospective and forecast simulations) are sufficiently large to allow the derivation of suitably reproducible statistics.

This paper deals with the last of these requirements (though it should be pointed out that satisfaction of (c) and (d) is required if (b) is to be satisfied). Many climate centres have the resources necessary to generate ensembles large enough to produce reasonably well constrained estimates of first order statistics such as the mean or the median. This is not generally the case if the desire is to make statements regarding the probabilities of more extreme model outcomes. We outline a resampling procedure, ELVIS, that allows the derivation of improved (more reproducible) estimates of simulated PMFs and CDFs of seasonal statistics (derived from monthly values) without increasing the number of ensemble members (see also Clark and Déqué 2003). The methodology uses exhaustive permutations of the individual monthly values from each ensemble member when deriving seasonal totals—essentially, for these monthly values the month index is retained but the ensemble index is not. The primary assumption underlying the utility of this approach is that, aside from the effects of prescribed boundary forcing, the month-to-month values of a given variable at a given location are more-or-less independent (in terms of internal memory). For some variables in some seasons which show relatively high internally derived autocorrelations (e.g. summer near-surface air temperature), this requirement represents a drawback of the re-sampling method. On the other hand, aside from improved estimates of ensemble spread, the method has the positive attributes that (i) derived values are consistent with model physical constraints, and (ii) cross-variable and cross-spatial relationships are conserved.

An alternative methodology for artificial ensemble expansion is to fit a statistical distribution to the ensemble results (e.g. the seasonal totals), and use that distribution to estimate event probabilities. This parametric approach has been explored by Wilks (2002), who documented the potential for reduced uncertainty in probability estimates in the context of ensemble numerical weather prediction. As it might be applied to seasonal simulations, the parametric technique has the positive attribute of being unaffected by serial independence, but has the following drawbacks.

- It requires choice of an appropriate statistical distribution.
- There is possible inconsistency with model physical constraints (events that are not possible within model physics may be given a finite probability).
- Cross-variable or cross-spatial relationships are not automatically retained.

The behaviour of normal and permuted sums from idealized ensembles (constructed from monthly values with normal or log-normal distributions) are investigated from several perspectives. Using the χ^2 distribution, it is shown that for an M-month season and an ensemble of size N, and under the assumptions of linear independence and equal monthly standard deviations, the permuted sums behave as if they come from an ensemble of $M \times (N-1) + 1\{\chi^2 \text{ df } M \times (N-1)\}$. Thus, a relative increase (of order M) in effective ensemble size can be realized with negligible computational effort.

In the case that monthly variances are not equal, the effective ensemble size (df) of the permuted sums is decreased (leading to incorrect assumptions about the uncertainty of probability estimates). An ad hoc correction factor giving the fractional reduction in df is suggested. The behaviour of the permuted sums in situations where the assumption of linear independence is violated is also examined. It is shown that the variance of the permuted sums is biased low, and that the (theoretical true) CDF of the variance of

those sums is composed of a multiple of cumulative χ^2 distributions. It seems likely that expressions for the bias and true CDF could be derived; on the other hand, in practice, estimated levels of linear dependence for a particular set of initial conditions would have large associated uncertainties (e.g. Koster and Suarez 2001). This uncertainty argues for a recognition of the potential effects arising from linear dependence, but against a formal correction scheme. This presentation has not discussed the impact of nonlinear dependence, and it is suggested that the ELVIS method be used with caution where antecedent conditions are known to have more than a modest impact on a particular variable and a particular point.

A second set of analyses examined the behaviour of ELVIS sums in terms of bin AAPEs for PMFs or CDFs of ordinary and permuted sums generated from specified statistical distributions. In general, the results show that permuted sums reduce the errors in estimating probabilities in comparison to ordinary sums. For normally distributed underlying populations, the reduction in PMF errors is largest near the centre of the distribution; for CDFs, the relative improvements (considering the upper part of the distribution) are largest for the uppermost (most extreme) bins. The parametric approach (applied under the correct assumption of an underlying Gaussian distribution) gives results close to those from the resampling method. For a mildly skewed (log-normal) distribution, the relative improvements in probability estimates from permuted sums (over ordinary sums) are somewhat reduced in comparison to a symmetrical distribution. An application of the parametric method to the log-normal case, and incorrectly assuming an underlying normal distribution, emphasized the need to select the correct form of underlying distribution when this approach is used.

Other tests examined the effects when the assumption of (internal) linear independence is violated. The results showed the expected increase in probability errors with increasing correlations, but revealed that for a 3-month season, ten-member ensemble and month-to-month internal autocorrelations of 0.4, probability errors for permuted sums remain smaller than those for ordinary sums. As noted, the parametric approach is insensitive to serial dependence.

The AAPE measure was also applied to a 100-member 3-month AGCM ensemble simulation. The results are basically consistent with those noted above, i.e. better results with both the resampling and permutation approach. The relatively small size of the underlying population and resulting over-sampling introduced 'noise' in the results, so that only qualitative conclusions are to be drawn from these tests. For this example, with approximately normally distributed monthly values and low (total) month-to-month autocorrelations, the performance of the parametric approach was essentially identical to the permutation method.

A final set of analyses examined a simple method for detecting internally derived memory from ensemble AGCM simulations with temporally varying prescribed boundary conditions (i.e. where extended multi-member control runs are not available). The technique appears to be effective in distinguishing cases where internal memory has a substantial effect on model behaviour. For the GCM used for these analyses (ECHAM3), the inferred internal memory is much larger for 2 m air temperature than for precipitation; this is consistent with the idea that the memory is driven largely by soil moisture and snow cover, and principally results from near-surface interactions rather than large-scale dynamics.

The improvements in effective ensemble size that can be realized through the ELVIS approach suggest extensions of the method. For example, in many applications of climate forecast information the interest is in the joint behaviour of two variables (e.g. temperature and precipitation). The larger sample afforded by permutation may

make it possible to better quantify the probabilities of particular outcomes in this regard. In a related manner, the method could be expanded to joint probabilities in the spatial sense—allowing, for example, better quantification of the simulated joint probabilities of low seasonal rainfall in one region and above normal rainfall in another.

A concluding point regarding the resampling approach concerns the time-scales of the aggregation and sub-aggregate sampling. In the work described here, seasons (aggregation) and months (sub-aggregate sampling) were used for convenience—monthly data were available—and because seasonal values are of some relevance to users. Probably somewhat shorter sub-aggregate scales could be used in some instances, thus allowing larger effective sample sizes. However, as sampling time-scales approach synoptic time-scales, the effects of dynamical internal memory will begin to have relatively larger—and likely less statistically benign—effects than those related to surface boundary conditions (e.g. see discussions in Wilks (2002)). Different aggregation times are also a possibility, with shorter periods allowing more precise forecast targets, and longer ones giving better sampling characteristics (but perhaps with less relevance).

A variety of studies make it clear that ensemble approaches to dynamical seasonal climate prediction allow the characteristics and performance of climate prediction systems to be described with much more refinement than deterministic measures (e.g. Barnett 1995; Anderson 1996; Branković and Palmer 1997; Mason and Graham 1999). Further, it is well established that climate forecasts framed in the form of probabilities are the most useful to end users (whose responses can be regarded as hedges, e.g. Palmer *et al.* (2000)). From either of these perspectives, more accurate estimates of the true PDF of model response to boundary forcing would be useful. In many cases, the ELVIS technique does allow improvements in the accuracy of such estimates through a simple, computationally efficient non-parametric approach, and it may prove a useful tool in a variety of applications.

REFERENCES

Anderson, J. L. | 1996 | A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate*, **9,** 1518–1530
Anderson, J. L. and Stern, W. | 1996 | Evaluating the potential predictive utility of ensemble forecasts. *J. Climate*, **9,** 260–269
Barnett, T. P. | 1995 | Monte Carlo climate forecasting. *J. Climate*, **8,** 1005–1022
Bengtsson, L., Schlese, U., Roeckner, E., Latif, M., Barnett, T. P. and Graham, N. E. | 1993 | A two-tiered approach to long-range climate forecasting. *Science*, **261,** 1026–1029

Branković, Č. and Palmer, T. N.   1997   Atmospheric seasonal predictability and estimates of ensemble size. *Mon. Weather Rev.*, **125,** 859–874

2000   Seasonal skill and predictability of ECMWF PROVOST ensembles. *Q. J. R. Meteorol. Soc.*, **126,** 2035–2067

Clark, R. and Déqué, M.   2003   Conditional probability seasonal predictions of precipitation. *Q. J. R. Meteorol. Soc.*, **129,** 179–193

Davies, J. R., Rowell, D. P. and Folland, C. K.   1997   North Atlantic and European seasonal predictability using ensembles of multidecadal atmospheric GCM simulations. *Int. J. Climatol.*, **17,** 1263–1284

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

2001   Seasonal predictability of tropical rainfall: Probabilistic formulation and validation. *Tellus*, **53,** 500–512

DKRZ   1992   'The ECHAM3 Atmospheric General Circulation Model'. Deutsches Klimarechenzentrum, Hamburg, Report No. 6. Bundesstrasse 55, D-2000 Hamburg 13 Germany

Frederiksen, C. S., Zhang, H. Q., Balgovind, R. C., Nicholls, N., Drosdowsky, W. and Chambers, L.   2001   Dynamical seasonal forecasts during the 1997/98 ENSO using persisted SST anomalies. *J. Climate*, **14,** 2675–2695

Georgakakos, K. P., Georgakakos, A. P. and Graham, N. E.   1998   Assessment of benefits of climate forecasts for reservoir management in the GCIP region. *GEWEX News*, **7,** 5–7

Goddard, L. and Graham, N. E.   1999   The importance of the Indian Ocean for GCM-based climate forecasts for eastern and southern Africa. *J. Geophys. Res.*, **104,** 19099–19116

Goddard, L. and Mason, S. J.   2002   Sensitivity of seasonal climate forecasts to persisted SST. *Clim. Dyn.*, **19,** 619–632

Goddard, L., Mason, S. J., Zebiak, S. E., Ropelewski, C. F., Basher, R. and Cane, M. A.   2001   Current approaches to seasonal-to-interannual climate predictions. *Int. J. Climatol.*, **21,** 1111–1152

Graham, N. E.   1994   'Experimental predictions of wet season precipitation in northeast Brazil'. Pp. 378–381 in Proceedings of the 18th climate diagnostics workshop, 1–5 November, 1993. Boulder, Colorado, USA

Graham, N. E. and Tyree, M.   1998   Simulation of 20th century temperature trends. *J. Geophys. Res.*, **103,** 28893–28908

Harrison, M. S. J., Richardson, D. S., Robertson, K. and Woodcock, A.   1995   'Medium-range ensembles using both the ECMWF T63 and unified models—An initial report'. UKMO Technical Report 153. Met Office, Exeter, UK

Koster, R. D. and Suarez, M. J.   2001   Soil moisture memory in models. *J. Hydrometeorol.*, **2,** 558–570

Kumar, A., Hoerling, M. P., Ji, M., Leetmaa, A. and Sardeshmukh, P.   1996   Assessing a GCM's suitability for making seasonal predictions. *J. Climate*, **9,** 115–129

Kumar, A., Barnston, A. G. and Hoerling, M. P.   2001   Seasonal predictions, probabilistic verifications and ensemble size. *J. Climate*, **14,** 1671–1676

Mason, S. J. and Graham, N. E.   1999   Conditional probabilities, relative operating characteristics and relative operating levels. *Weather and Forecasting*, **14,** 713–725

2002   Areas beneath the relative operating characteristics (ROC) and levels (ROL) curves: Statistical significance and interpretation. *Q. J. R. Meteorol. Soc.*, **128,** 2145–2166

Mason, S. J., Goddard, L., Graham, N. E., Yulaeva, E., Sun, L. and Arkin, P. A.   1999   The IRI seasonal climate prediction system and the 1997/98 El Niño event. *Bull. Am. Meteorol. Soc.*, **80,** 1853–1873

Palmer, T. N.   2002   The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Q. J. R. Meteorol. Soc.*, **128,** 747–774

Palmer, T. N., Branković, Č. and Richardson, D. S.   2000   A probability and decision-model analysis of PROVOST seasonal multi-model ensemble integrations. *Q. J. R. Meteorol. Soc.*, **126,** 2013–2033

Reynolds, R. W. and Smith, T. M.   1994   Improved global sea surface temperature analyses using optimal interpolation. *J. Climate*, **7,** 929–948

Roeckner, E. K., Arpe, K., Bengtsson, L., Brinkop, S., Dümenil, L., Esch, M., Lunkeit, F., Ponater, M., Rockel, B., Sausen, R., Schlese, U., Schubert, S. and Windelband, M.   1992   'Simulation of the present-day climate with the ECHAM Model: Impact of model physics and resolution'. Report No. 93. Max-Planck-Institut fur Meteorologie, Bundesstrasse 55, 20146 Hamburg, Germany

Rowell, D. P., Folland, C. K., Maskell, K. and Ward, M. N.   1995   Variability of summer rainfall over tropical north Africa (1906–92): Observations and modelling. *Q. J. R. Meteorol. Soc.*, **121,** 669–704

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

Smith, T. M., Reynolds, R. W., Livezey, R. E. and Stokes, D. C.   1996   Reconstructions of historical sea surface temperatures using empirical orthogonal functions. *J. Climate*, **8,** 1403–1420

Ward, M. N. and Navarra, A.   1997   Pattern analysis of SST-forced variability in ensemble GCM simulations: Examples over Europe and the tropical Pacific. *J. Climate*, **10,** 2210–2220

Wilks, D. S.   2002   Smoothing forecast ensembles with fitted probability distributions. *Q. J. R. Meteorol. Soc.*, **128,** 2821–2836