Tellus (2005), 57A, 398–408
Printed in UK. All rights reserved

TELLUS

A method for statistical downscaling of seasonal ensemble predictions

By HENRIK FEDDERSEN* and UFFE ANDERSEN, *Danish Meteorological Institute, Lyngbyvej 100, DK-2100 Copenhagen, Denmark*

(Manuscript received 30 March 2004; in final form 26 July 2004)

ABSTRACT

A model output statistics based method for downscaling seasonal ensemble predictions is outlined, and examples of ensemble predictions of precipitation and 2-m temperature are verified against observing stations in Scandinavia, Europe, north-western America, the contiguous United States and Australia. The downscaling from seasonal ensemble predictions from coupled ocean/atmosphere general circulation models to daily precipitation time series for individual observing stations is performed in three steps: (i) a spatial downscaling of ensemble mean seasonal means from dynamical model output to station level by means of patterns derived from a singular value decomposition analysis of model output and observations; (ii) application of the downscaling transformation to the model output ensemble and subsequent calibration of the downscaled ensemble; (iii) a stochastic generation of daily precipitation conditioned on predictions of the probability of a wet day in the season and daily persistence. In the majority of the examples, the downscaling is found to provide more skilful predictions than the raw dynamical model output.

1. Introduction

Several potential applications exist for seasonal to interannual climate prediction, including crop yield prediction and prediction of tropical disease, which are both treated elsewhere in this issue (Cantelaube and Terres, 2005; Marletto et al., 2005; Morse et al., 2005). Most existing application models require seasonal climate input on a spatial scale much smaller than that of present-day dynamical seasonal climate prediction models. In addition to inaccuracy associated with lack of horizontal resolution, coupled ocean/atmosphere models suffer from a substantial drift away from the observed climate (Stockdale, 1997), a drift that also needs to be corrected. A variable that is often poorly predicted on local scale is precipitation, which is also one of the most important variables for many applications. One approach to improve poor predictions of precipitation is that of statistical downscaling. Statistical downscaling aims at specifying the local field (the predictand, e.g. precipitation) from a large-scale field (the predictor), which is accurately predicted by the dynamical model.

The choice of predictor depends on the predictand. Two conditions should be satisfied: (i) it must be possible to specify the predictand accurately from the predictor, and (ii) the predictor should be well predicted by the dynamical model. For precipitation, large-scale fields, such as mean sea level pressure, geopotential height, relative humidity, vorticity and divergence, are all predictor candidates (Wilby and Wigley, 2000), but also precipitation predicted by the dynamical model can—and will in this paper—be used as predictor.

A natural first approach to the specification of the predictand is to use linear regression, but more advanced methods (e.g. artificial neural networks) can also be applied. An important choice regards the choice of 'training' data set. Traditionally, statistical downscaling makes use of relationships that are derived between observed (or analysed) predictor and predictand fields, such as, for example, the 700-hPa height field and precipitation. However, predictions based on this perfect prognosis approach (Wilks, 1995) are sensitive to model errors in the predictor field. If, instead, the training predictor is chosen as a model field, model systematic errors will automatically be accounted for in the predictions. This approach is known as model output statistics (MOS; Wilks, 1995) and is widely applied to numerical weather predictions to obtain point predictions. MOS predictions are not normally referred to as downscaling, as the predictor is normally obtained from immediately surrounding model grid points, rather than from large-scale fields. The major drawback of MOS is the need for a long series of hindcasts using an unaltered model. That is, every time the dynamical model undergoes a major upgrade, a long series of hindcasts must be recomputed in order to derive new MOS relations that take into account possibly altered systematic errors of the dynamical model.

*Corresponding author.
e-mail: hf@dmi.dk

The DEMETER (Development of a European Multi-model Ensemble system for seasonal to inTERannual prediction) set of model hindcasts (Palmer et al., 2004) provide a data set that is well suited for a MOS-based downscaling. To date, there is only limited experience with downscaling of seasonal predictions, whereas in climate change modelling statistical downscaling has been applied extensively, but using the perfect prog approach (e.g. von Storch et al., 1993). As a result of the different nature of the problem, climate change modelling (and subsequent downscaling) cannot be verified in the same manner as seasonal predictions.

In addition to spatial downscaling, some applications may also require a downscaling in time, i.e. time series in a higher temporal resolution specified from seasonal or monthly time series (Goddard et al., 2001). Most crop yield models, for example, require daily weather input. The raw output from dynamical models is available as the required daily values, but as surface variables typically are available as grid cell averages and the horizontal resolution is relatively coarse in seasonal prediction models, the resulting time series are smoother than time series that are observed at single stations, particularly for precipitation. Alternatively, synthetic daily time series can be generated using a stochastic weather generator, which is conditioned on seasonal output of the dynamical model. Stochastic weather generators have been widely used for simulating weather (precipitation, temperature, global radiation, etc.) for use in crop yield models (Richardson, 1981; Parlange and Katz, 2000). They have also been used in connection with climate change studies (Wilks, 1992; Katz, 1996; Mearns et al., 1997; Semenov and Barrow, 1997; Palutikof et al., 2002), but little has been done in relation to seasonal prediction.

It has been documented that dynamical models can provide skilful seasonal forecasts, i.e. forecasts that are better than climatology, particularly in the tropics (Stockdale et al., 1998), but it has also been demonstrated that the prediction skill, notably for precipitation, can be improved using statistical techniques to correct the raw model output. Feddersen et al. (1999) showed how statistical correction using the leading modes of a singular value decomposition analysis (SVDA), also known as a maximum covariance analysis (MCA), could improve the skill of seasonal precipitation simulations made with an atmospheric general circulation model forced by observed sea surface temperature. Similar results using related statistical correction methods applied to a number of different atmospheric model simulations have been reported for a number of later studies (Gershunov et al., 2000; Kharin and Zwiers, 2001; Tippett et al., 2003; Widmann et al., 2003; Kang et al., 2004).

The present work is an extension to the above studies in that we apply statistical downscaling/correction to predictions from more realistic coupled ocean/atmosphere forecasting systems, i.e. sea surface temperature is not prescribed, but part of the prediction; we deal with full ensembles, not only the ensemble mean; we look into the benefits of multi-model forecasting after statistical downscaling has been applied individually to each ensemble from three different models; we demonstrate how to generate daily time series of precipitation using a stochastic weather generator conditioned on downscaled seasonal mean predictions.

The paper is organized as follows. Following this introduction, in Section 2 we describe the downscaling methodology, comprising spatial downscaling of seasonal mean precipitation, downscaling and calibration of seasonal ensemble predictions and stochastic generation of daily precipitation conditioned on downscaled seasonal mean predictions. In Section 3, the statistical downscaling method is illustrated by several examples, and conclusions are presented in Section 4.

2. Methodology

Statistical downscaling allows endless possibilities for the choice of predictor fields. As we are mainly interested here in outlining a method for statistical downscaling, we intend to keep things simple and let fine-tuning of the method depend on the particular application.

2.1. Downscaling seasonal ensemble mean predictions

As systematic errors in the DEMETER models cannot be ignored, a MOS-based downscaling is preferred, following the approach in Feddersen et al. (1999) and Feddersen (2003). The predictor field is chosen to be the model output equivalent of the predictand, i.e. if the predictand is precipitation in a region, then the predictor is chosen as precipitation predicted by the dynamical model in a region encompassing the predictand region. An SVD analysis (Bretherton et al., 1992) is applied to a training data set of predictor and predictand, and observed precipitation is linearly regressed on the leading SVD modes. The regression equation is derived using the ensemble mean of the model predictions, as ensemble averaging reduces climatic noise and so is an estimate of the predictable part of the ensemble.

Following Bretherton et al. (1992), the standardized predictor (ensemble mean model output), $\mathbf{x}_i$, and standardized predictand, $\mathbf{y}_i$, time series (where index i denotes time) are expanded in terms of patterns ($\mathbf{g}_m$, $\mathbf{h}_m$) and time series ($u_{m,i}$, $v_{m,i}$),

$$\mathbf{x}_i \approx \sum_m u_{m,i}\mathbf{g}_m, \tag{1}$$

$$\mathbf{y}_i \approx \sum_m v_{m,i}\mathbf{h}_m. \tag{2}$$

The time series are given by projection of the standardized fields on the respective patterns, i.e.

$$u_{m,i} = \mathbf{x}_i \cdot \mathbf{g}_m, \tag{3}$$

and similarly for $v_{m,i}$. By choosing the patterns $\mathbf{g}_m$ and $\mathbf{h}_m$ as singular vectors of an SVD decomposition of the cross-covariance matrix of the $\mathbf{x}$ and $\mathbf{y}$ fields, the covariance between $u_{1,i}$ and $v_{1,i}$ is maximized, and the covariances between $u_{m,i}$ and $v_{m,i}$

are maximized subject to the condition that $\mathbf{g}_m$ is orthogonal to $\mathbf{g}_1, \ldots, \mathbf{g}_{m-1}$, and $\mathbf{h}_m$ is orthogonal to $\mathbf{h}_1, \ldots, \mathbf{h}_{m-1}$.

An estimate for the predictand is obtained by multiple linear regression on the $u_{m,i}$ time series, i.e.

$$\hat{\mathbf{y}}_i = \sum_m \mathbf{A}_m u_{m,i}, \tag{4}$$

where $\mathbf{A}_m$ is calculated so as to minimize the expected root-mean-square (rms) difference between $\hat{\mathbf{y}}_i$ and $\mathbf{y}_i$.

The number of modes m (pairs of patterns and time series) that are included in the sums above should ideally explain the fraction of the covariance between GCM output and observations that is not due to climatic noise. In practice, this is accomplished by sorting the singular vectors of the SVD analysis according to the corresponding singular values (that are proportional to the covariance fraction explained by the singular vectors) and including singular vectors corresponding to the singular values in descending order until the singular values, or close pairs of singular values, are small compared to the first singular values and are no longer significantly different from the subsequent singular values (see North et al., 1982, for a more detailed discussion of the related problem of the spectrum of eigenvalues of empirical orthogonal functions). For the DEMETER predictions we typically include between two and six singular vectors.

The predictand comprises precipitation from a set of observing stations. A choice has to be made regarding the dynamical model output predictor region: it should encompass the corresponding observations and be large enough to resolve the relevant large-scale patterns. For example, for downscaling over Scandinavia the model output region is chosen to include a good part of the North Atlantic in order to resolve the North Atlantic Oscillation pattern (Feddersen, 2003).

2.2. *Downscaling and calibrating seasonal ensemble predictions*

Having thus obtained a procedure (which is nothing but a linear transformation of the model output) for downscaling the ensemble mean, the natural extension to downscaling of the full ensemble is to apply the same linear transformation to the individual ensemble members, i.e. to project the individual standardized ensemble members, $\mathbf{x}_{i,j}$ on to the ensemble mean SVD modes, $\mathbf{g}_m$ as in eq. (3) and apply the regression equation (eq. 4) to each ensemble member. However, the linear regression does not, in general, 'explain' all variance. Consequently, we expect the ensemble spread to be too small to capture the verifying observations, so that a calibration of the ensembles is necessary.

The well-known problem with regression-based statistical predictions having less variance than observations is frequently accounted for by inflating the statistical predictions, i.e. the predicted anomalies are rescaled or inflated so that the variance of the inflated predictions is increased so as to match the variance of the corresponding observations (Klein et al., 1959). An unfortunate side effect of this approach is that the predictions after the inflation are no longer optimal in a least-squares sense. However, for ensemble predictions we can 'add' extra variance to the predictions and still keep the ensemble mean fixed by increasing the ensemble dispersion. Using the method of analysis of variance (ANOVA) it is found that the total variance of the predictions for each station can be estimated as the sum of the variance of the ensemble mean and the internal variability reflected in the ensemble dispersion of the individual ensembles (Rowell et al., 1995; Rowell, 1998). Thus, for the predictand $y_{i,j}$, where indices i and j here, and in the following, denote time and ensemble member, respectively, we have that the total variance for each station is the sum of the ensemble mean variance and internal variance (the average ensemble dispersion).

With N being the number of years, n the number of ensemble members, $\hat{\mu}_i$ the estimated ensemble mean of the downscaled variable,

$$\hat{\mu}_i = \frac{1}{n}\sum_{j=1}^{n} \hat{y}_{i,j}, \tag{5}$$

and $\bar{y}$ the observed climatology,

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i, \tag{6}$$

an estimate for the ensemble mean variance, $\hat{\sigma}^2_{\mathrm{em}}$, is given by

$$\hat{\sigma}^2_{\mathrm{em}} = \frac{1}{N-1}\sum_{i=1}^{N}(\hat{\mu}_i - \bar{y})^2, \tag{7}$$

and an estimate for the internal variability, $\hat{\sigma}^2_{\mathrm{int}}$, by

$$\hat{\sigma}^2_{\mathrm{int}} = \frac{1}{N(n-1)}\sum_{i=1}^{N}\sum_{j=1}^{n}(\hat{y}_{i,j} - \hat{\mu}_i)^2. \tag{8}$$

An estimate for the total variance for each station is given by

$$\hat{\sigma}^2_{\mathrm{tot}} = \hat{\sigma}^2_{\mathrm{em}} + \frac{n-1}{n}\hat{\sigma}^2_{\mathrm{int}}, \tag{9}$$

where the factor $(n-1)/n$ is included in order to obtain an unbiased estimate (Rowell et al., 1995; Rowell, 1998).

When $\sigma^2_{\mathrm{tot}} < \sigma^2_{\mathrm{obs}}$, i.e. when the downscaled variable does not account for all the observed variance, then a stochastic 'noise' term, η is added to each ensemble member of the downscaled variable in order to introduce the variance that is not accounted for (von Storch, 1999), i.e.

$$y^*_{i,j} = \hat{y}_{i,j} + \eta. \tag{10}$$

If η is statistically independent of $\hat{y}_{i,j}$, then the total variance of $y^*_{i,j}$ is given by eq. (9) with addition of the variance of the noise, σ^2_η, to the internal variability, i.e.

$$\hat{\sigma}^2_{\mathrm{tot}^*} = \hat{\sigma}^2_{\mathrm{em}} + \frac{n-1}{n}\left(\hat{\sigma}^2_{\mathrm{int}} + \sigma^2_\eta\right). \tag{11}$$

If the total variance of $y^*_{i,j}$ is to match the observed variance then

$$\hat{\sigma}^2_\eta = \frac{n}{n-1}\left(\hat{\sigma}^2_{\text{obs}} - \hat{\sigma}^2_{\text{tot}}\right). \tag{12}$$

If identical levels of noise are added to each of the N ensembles, then the ensemble dispersion of the 'noisy' ensembles, y^*_i, is given by

$$D^2_{y^*_i} = D^2_{y_i} + \sigma^2_\eta, \tag{13}$$

where the dispersion of the ith downscaled ensemble is estimated as

$$\hat{D}^2_{y_i} = \frac{1}{n-1}\sum_{j=1}^{n}(\hat{y}_{i,j} - \hat{\mu}_i)^2. \tag{14}$$

It follows from eq. (13) that addition of noise to the ith ensemble leads to an increase of the ensemble dispersion by a factor

$$a_i^2 = 1 + \frac{\sigma^2_\eta}{D^2_{y_i}}. \tag{15}$$

Thus, the desired average ensemble dispersion can be obtained by rescaling or inflating the individual ensemble members relative to the ensemble mean:

$$\tilde{y}_{i,j} = \hat{\mu}_i + a_i(\hat{y}_{i,j} - \hat{\mu}_i). \tag{16}$$

Alternatively, the variance that is not accounted for by the ensemble mean can be added to the forecast, using eq. (10) directly. The probability distribution of the stochastic term, η is in general not known, except in the limit where the ensemble mean explains none of the observed variance. In this case where there is no predictive skill, it makes sense to assume that the variance is described by the climatological distribution, so that a probability forecast is always given by the climatological distribution. In the other limit where all the observed variance is accounted for by the linearly transformed ensemble members, nothing should be added. In between the two limiting cases, one approach is to use a distribution which is obtained by scaling the climatological distribution such that the total variance matches the observed variance. A first approach would be to assume white noise, but results of Huth et al. (2001) suggest that this is not sufficient, and so modelling the noise by an autoregressive process may yield better results.

Additional complications can be expected for a multivariate noise model where the different parameters cannot be assumed independent (temperature and precipitation are, for example, not independent in winter in Northern Europe where a wet and mild winter is more likely than a wet and cold winter). The ensemble inflation method (16) preserves both autocorrelations and multivariate relations and so is preferred throughout this paper.

2.3. *Stochastically generated daily precipitation conditioned on seasonal predictions*

In order to generate daily time series of precipitation (downscaling in time) we apply a stochastic weather generator which is a two-state (precipitation/no-precipitation) first-order Markov chain, where the probability of precipitation on a day is conditioned on occurrence of precipitation on the previous day (Katz, 1977). This model is characterized by the transition probabilities:

P_{01} = probability of precipitation, given no precipitation the previous day;
P_{11} = probability of precipitation, given precipitation the previous day.

Alternatively, the probabilities can be expressed as the probability of a wet day, π, and the first-order autocorrelation or persistence, d, of the daily precipitation occurrence series. π and d are related to the transition probabilities by (Katz, 1996)

$$\pi = \frac{P_{01}}{1 + P_{01} - P_{11}}, \tag{17}$$

$$d = P_{11} - P_{01}. \tag{18}$$

As π is confined to the interval [0, 1], and d is confined to the interval $[-1, 1]$ (although in practice $d > 0$) we transform π and d, using the log-odd transform and the Fisher Z-transform, respectively, to obtain (Wilks, 1999)

$$\pi' = \ln\left(\frac{\pi}{1-\pi}\right), \tag{19}$$

$$d' = \frac{1}{2}\ln\left(\frac{1+d}{1-d}\right). \tag{20}$$

π' and d' can be specified linearly from the $u_{m,i}$ time series similarly to the way in which the seasonal mean precipitation, and y_i is specified from $u_{m,i}$ in eq. (4). Having predictions of π' and d', it is straightforward to transform back to predictions of P_{01} and P_{11}, and then stochastically simulate daily sequences of wet and dry days.

For wet days, the precipitation amount is specified by randomly sampling the probability distribution for precipitation. Daily precipitation is generally well described by a gamma distribution (Stephenson et al., 1999), but downscaling experiments where ERA-15 data were used as predictor, showed that only the mean (and not the variance) could be skilfully predicted, so instead we sample daily precipitation from an exponential distribution (which is a special case of the gamma distribution) that is completely characterized by the daily mean precipitation on a wet day. The latter is given by the downscaled seasonal mean precipitation divided by π times the number of days in the season.

Additional weather parameters, such as minimum and maximum 2-m temperature, potential evaporation and total solar radiation, are normally assumed to follow normal distributions

with mean and possibly variance conditioned on the occurrence of precipitation and downscaled from the dynamical seasonal forecast model. Time series of the additional weather parameters are modelled by a first-order autoregressive process, but it is beyond the scope of this paper to demonstrate all aspects of stochastic weather generators. Details can be found, for example, in Katz (1996) and Parlange and Katz (2000).

3. Downscaling examples

In order for the statistical downscaling to produce robust results, a relatively long period of model hindcasts and corresponding observations is required. In the following, the 40-yr period 1961–2000 is chosen. Model hindcasts are available for this period from three of the modelling groups in DEMETER: Météo-France, the European Centre for Medium-Range Weather Forecasts (ECMWF) and the UK Meteorological Office (Palmer et al., 2004).

The downscaling examples are limited by limited availability of long records of observational data. In the following, we consider monthly means of observed precipitation and 2-m temperature from two data sets. (i) Nordklim (data available from http://www.smhi.se/hfa_coord/nordklim/; Tuomenvirta et al. 2001), which covers the Nordic region (Denmark, Finland, Iceland, Norway and Sweden). Figure 1 shows the location of Nordklim stations used in the following; only stations in Scandinavia and Finland are used. The five Danish stations for which daily data are available (Laursen, 2004) are indicated with '+' signs. (ii) The World Meteorological Organization (WMO) Global Climate Observing System (GCOS; data available from http://www.wmo.ch/web/gcos/gcoshome.html).

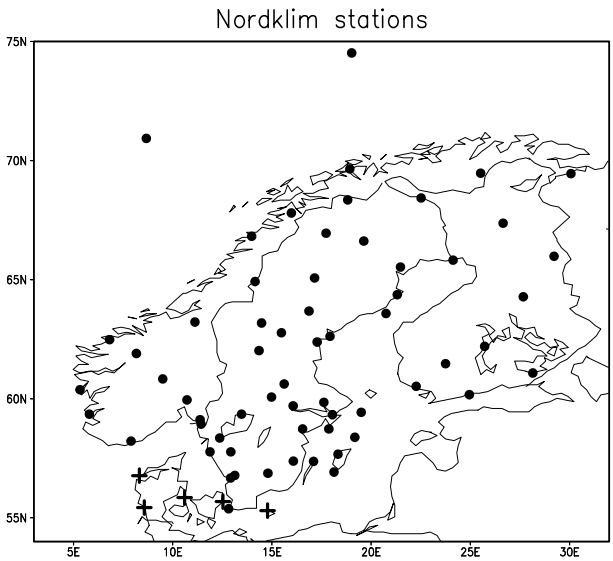

Fig 1. Location of stations used for downscaling in Scandinavia. The five Danish stations for which daily data are available are indicated with '+' signs.

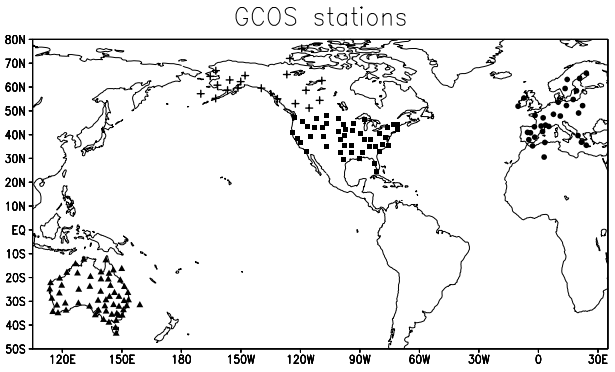

Fig 2. Location of stations used from the GCOS data set. The symbols indicate four different downscaling regions.

Although the GCOS coverage is global, sufficiently long time series are only available in Europe, North America and Australia, and even here the network of GCOS stations is relatively sparse. Figure 2 shows the location of GCOS stations used in the following. The stations are divided into four different downscaling regions (indicated by different symbols in Fig. 2): Europe, north-west America (Alaska and western Canada), the contiguous United States and Australia.

The Nordklim data set is a quality controlled data set where both monthly temperature and precipitation data have been tested for homogeneity for many stations, and possible inhomogeneities (e.g. caused by instrument changes) have been adjusted. The GCOS data are used 'as is', i.e. the quality control, if any, that has been applied varies from country to country. Stations have only been included in the following if the data series are complete or nearly complete. Missing data have been replaced by the 1961–2000 seasonal average in the training of the downscaling regression equations, whereas missing data are ignored in the validation.

3.1. Downscaling seasonal ensemble mean predictions

The statistical downscaling has been applied to model hindcasts for Europe, north-west America, the contiguous United States, Australia and Scandinavia for a lead time of two months for four different seasons. Tables 1 and 2 show cross-validated mean anomaly correlations (Déqué and Royer, 1992; Déqué, 1997) for the multi-model ensemble mean. The skill scores are mostly positive, indicating modest predictive skill beyond that of climatology. In the cross-validation, one year is withheld from the predictand data set and a prediction is made for the withheld year (Michaelsen, 1987). This is repeated for all years yielding 40 yr of predictions for validation. The statistical significance of the skill scores has been tested using a Monte Carlo type resampling approach, where the time series of observed precipitation and temperature have been randomly resampled 500 times in order to estimate a distribution for 'random' skill scores, and the actual

Table 1. Multi-model (Météo-France, ECMWF and UK Meteorological Office models) mean anomaly correlation for cross-validated ensemble mean predictions of seasonal precipitation statistically downscaled to stations in five selected regions for seasons JFM, AMJ, JAS and OND. Period is 1961–2000, and lead time is two months. Statistical significance at the 5% level is indicated by bold numbers

	JFM	AMJ	JAS	OND
Europe	0.22	0.15	0.12	0.17
NW America	−0.11	−0.17	0.12	−0.23
Contiguous US	**0.23**	0.05	−0.07	0.02
Australia	0.12	0.07	**0.22**	**0.28**
Scandinavia	**0.27**	0.12	0.08	0.06

Table 2. As Table 1, but for seasonal mean 2-m temperature

	JFM	AMJ	JAS	OND
Europe	0.22	**0.23**	**0.35**	0.19
NW America	**0.29**	0.13	**0.24**	−0.07
Contiguous US	0.23	0.10	**0.36**	0.05
Australia	**0.19**	**0.22**	**0.32**	0.23
Scandinavia	0.13	**0.28**	0.14	0.07

Table 3. Comparison of mean anomaly correlation for individual models and multi-model for selected seasonal ensemble mean predictions

	Precipitation		2-m temperature	
	Europe JFM	Scand. JFM	Europe JAS	Scand. AMJ
Météo-France	0.07	0.11	0.16	0.33
ECMWF	0.30	0.09	0.35	0.14
UK Met Office	0.03	0.28	0.25	0.15
Multi-model	0.22	0.27	0.35	0.28

skill score has been compared to this distribution. The significance at the 5% level is indicated in Tables 1 and 2.

The multi-model approach has a positive impact on the mean anomaly correlation skill score. In cases where the skill varies considerably between the individual models, the skill of the multi-model is comparable to the skill of the best of the individual models. This is illustrated in Table 3, showing mean anomaly correlations for the most skilful cases in Europe and Scandinavia. Note that no single model is consistently better or worse than the other models.

The predictive skill varies geographically within the downscaling region as well as from year to year. As an example, Fig. 3 shows time series of anomaly correlations for predictions of 2-m temperature for Europe in the JAS season, and Fig. 4

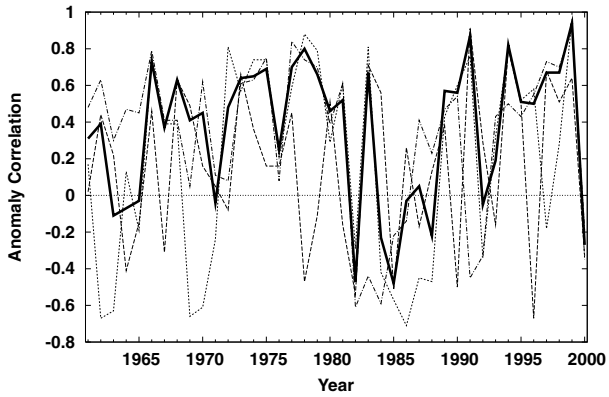

Fig 3. Cross-validated anomaly correlation for 2-m temperature predictions for Europe in the JAS season. The dashed curves represent individual models, and the solid curve represents multi-model.

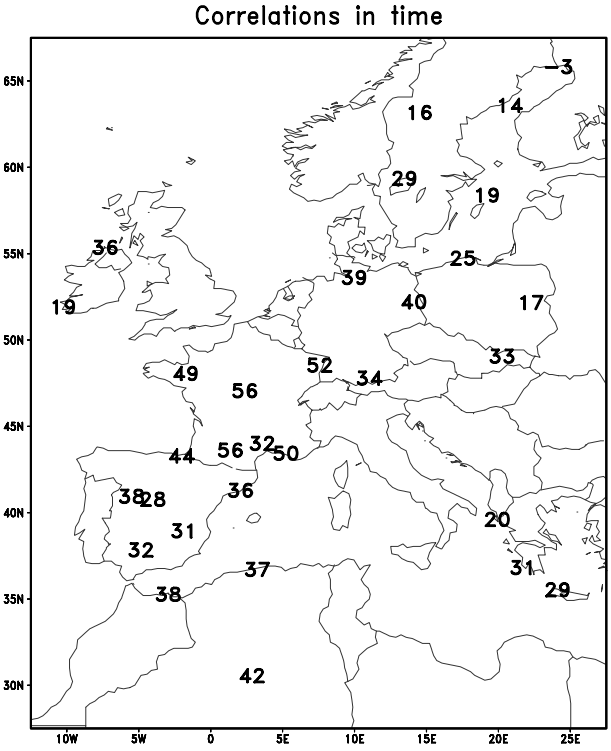

Fig 4. Cross-validated anomaly correlations (in %) in time for downscaled 2-m temperature predictions for Europe in the JAS season, 1961–2000.

shows correlations in time for the multi-model version of the same predictions. Figure 3 suggests that it takes at least two bad model predictions (anomaly correlation near zero or negative) before the multi-model (three-model) prediction fails. Correlations in time less than 0.6, as in Fig. 4, are representative for correlations also for other seasons and regions. However, the geographical distribution of the correlations depends strongly on the season. Point correlations in time based on only 40 values

Table 4. Comparison between downscaled and raw multi-model output skill for precipitation in seasons JFM, AMJ, JAS and OND. '+' ('−') indicates that downscaled predictions for most years are more (less) skilful than raw model output in terms of anomaly correlations. Statistical significance at the 5% level is indicated by '++' or '−−'

	JFM	AMJ	JAS	OND
Europe	+	++	+	+
NW America	−	+	+	−
Contiguous US	+	−	+	−−
Australia	0	+	+	++
Scandinavia	++	++	+	+

Table 5. As Table 4, but for 2-m temperature

	JFM	AMJ	JAS	OND
Europe	+	+	+	−
NW America	+	+	+	−
Contiguous US	+	−	++	+
Australia	+	−	+	0
Scandinavia	+	+	−	++

or less can be sensitive to missing data, while the mean anomaly correlation is fairly robust. Both the geographical and seasonal variations are in qualitative agreement with correlations between direct model output and ERA-40 reanalyses (see the DEMETER verification web page, http://www.ecmwf.int/ research/demeter/d/charts/verification/).

The predictive skill for 2-m temperature is generally higher than for precipitation. However, if the raw multi-model ensemble mean predictions are validated against observations (using model output from the grid cell that includes the station location and subtracting constant model bias), we find that the downscaling generally adds more skill for precipitation than for temperature, particularly in Europe and Scandinavia. Results of comparisons between downscaled and raw model output are shown in Tables 4 and 5 where '+' ('−') denotes that the downscaled predictions for a majority of the years are more (less) skilful than the raw model output. The comparison is based on time series of cross-validated anomaly correlations (as in Fig. 3), which allows for a simple test for statistical significance: with the null hypothesis that the downscaled and raw model predictions 'win' the same number of times in terms of anomaly correlation, a test in the binomial distribution gives that the null hypothesis can be rejected at a 5% level when one or the other type of prediction wins in 26 or more years out of 40 yr (corresponding to 65% of the years). If we only include years for which the 'winning' anomaly correlation is positive, then slightly more than 65% of the years are required for statistical significance. Statistical significance is indicated by '++' and '−−' in Tables 4 and 5.

We note that while this test for statistical significance is simple to apply, it does not provide the definitive answer as to whether the downscaled predictions are more skilful than the raw model output; anomaly correlations indicate only one aspect of skill—other skill scores may lead to different results. Moreover, the stations are unevenly distributed, so if there is a high concentration of stations in an area where the downscaled predictions are, e.g. more skilful than the raw model output, then the test for statistical significance is biased. Also, a user of the downscaled predictions may not be interested in the whole downscaling region, so if the test for statistical significance is applied only to the stations that are relevant for him, the result may be different.

3.2. Downscaling seasonal ensemble predictions

The full ensemble is downscaled by applying the ensemble mean based downscaling transformation to each member of the ensemble, i.e. the individual standardized ensemble members are downscaled by projecting them on to the model singular vectors from the SVD analysis, and temperature or precipitation is specified by use of regression equations such as eq. (4). Subsequently, the downscaled ensemble is inflated by rescaling the ensemble about the ensemble mean by an amount that would make the total variance of the ensemble inflated predictions for the training period match the observed variance, as described in Section 2.2.

In order to validate whether the downscaled ensemble predictions are statistically consistent with the verifying observations and capable of resolving different events, we consider rank histograms (Anderson, 1996; Hamill and Colucci, 1997) and relative operating characteristics (ROC; Stanski et al., 1989).

The rank histogram shows the distribution of the verifying observations in terms of ranked forecast ensemble members. If all ensemble members a priori are equally likely, as is implicitly assumed when probability forecasts are derived directly from the ensemble, simply by determining the fraction of the members that forecast a certain event, then the verifying observation is equally likely to fall in any interval between two neighbouring, ranked ensemble members, including the open ended intervals 'less than the smallest ensemble member' and 'greater than the largest ensemble member', and the rank histogram will be approximately flat. A flat rank histogram indicates statistical consistency between forecasts and observations (but is not a sufficient condition; Hamill, 2001). Conversely, if the rank histogram is not flat, then the intervals between the ensemble members are not equally likely. In particular, a U-shaped histogram, which is often encountered in medium-range forecasts, is indicative of underdispersive ensembles that too often do not capture the verifying observations.

Figures 5 and 6 show multi-model rank histogram examples for downscaled 2-m temperature and precipitation predictions. The examples are representative not only for the different seasons and regions, but also for the individual models. The rank histogram for 2-m temperature shows no systematic deviation from a flat rank histogram and indicates that the ensemble inflation provides well-calibrated predictions; the deviations that

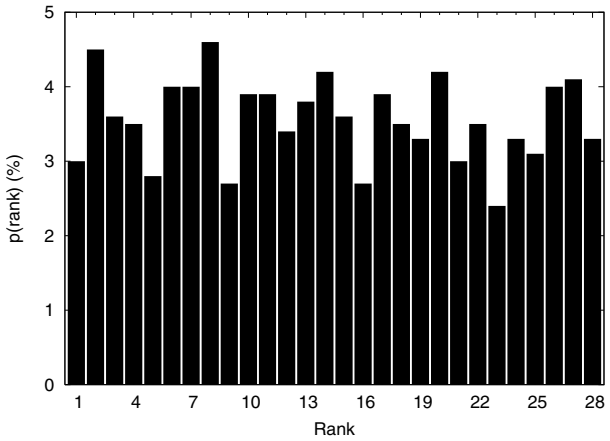

Fig 5. Rank histogram for downscaled 2-m temperature for Europe in JAS season for multi-model predictions in cross-validation mode.

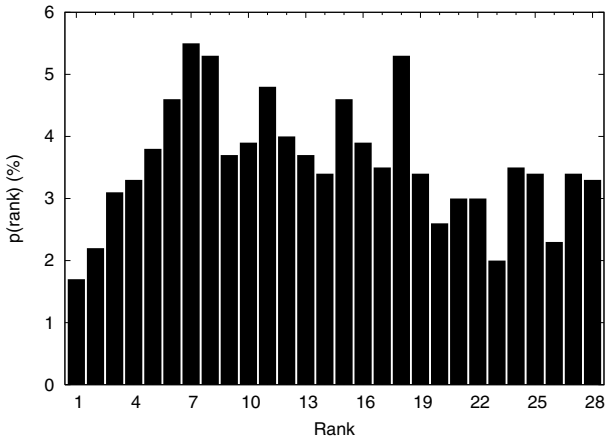

Fig 6. As Fig. 5, but for precipitation for Europe in JFM season.

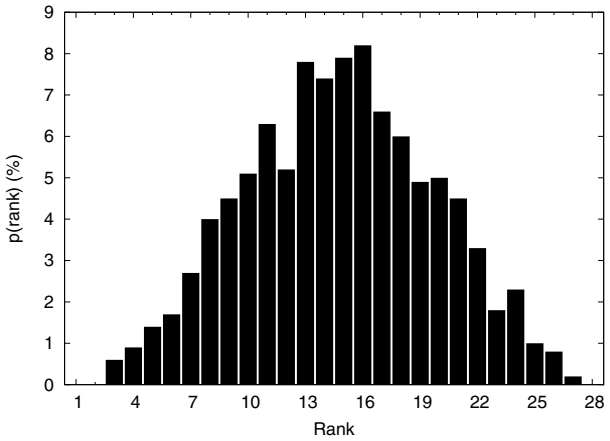

Fig 7. As Fig. 5, but for raw model output of 2-m temperature.

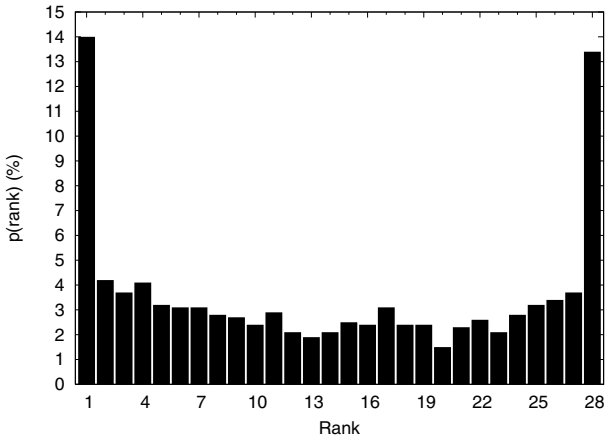

Fig 8. As Fig. 6, but for raw model output of precipitation.

are observed can be attributed to the limited sample size. For precipitation, there appears to be a tendency for the low end of the ensemble to be too small compared to observations, leading to a rank histogram where small ranks are underrepresented compared to the perfect, flat rank histogram. A possible cause for this problem is the application of the ensemble inflation to positively skewed precipitation distributions.

Similar rank histograms for raw model output (with constant model bias subtracted) are shown in Figs. 7 and 8. They indicate that the raw model temperature ensembles are overdispersive, while the raw model precipitation ensembles are underdispersive.

ROC curves are obtained by plotting hit rate versus false alarm rate for varying decision thresholds. The forecast is skilful if the hit rate exceeds the false alarm rate (i.e. if the area under the ROC curve exceeds 0.5). Figure 9 shows ROC curves for the same 2-m temperature and precipitation examples as in Figs. 5 and 6; the predicted events are temperature and precipitation in the upper tercile. The multi-model approach has a positive impact on the ROC score, which is partly caused simply by having an ensemble three times larger than for the individual models. Tables 6 and 7 list ROC areas for multi-model 2-m temperature and precipitation predictions. The ROC areas lie between 0.5 and 0.7, indicating modest skill with temperature ensemble predictions being, in general, slightly more skilful than precipitation ensemble predictions. Statistical significance at the 5% level, calculated as described in Section 3.1, is also indicated in Tables 6 and 7.

3.3. Stochastically generated daily precipitation conditioned on seasonal predictions

Observed daily precipitation time series were only available for the five Danish stations (indicated by '+' signs in Fig. 1), so the following is only an outline of the use of a stochastic weather generator to 'downscale in time'. Only precipitation is considered here, whereas elsewhere in this issue (Cantelaube and Terres, 2005; Marletto et al., 2005) also stochastically generated daily time series of minimum and maximum 2-m temperature,

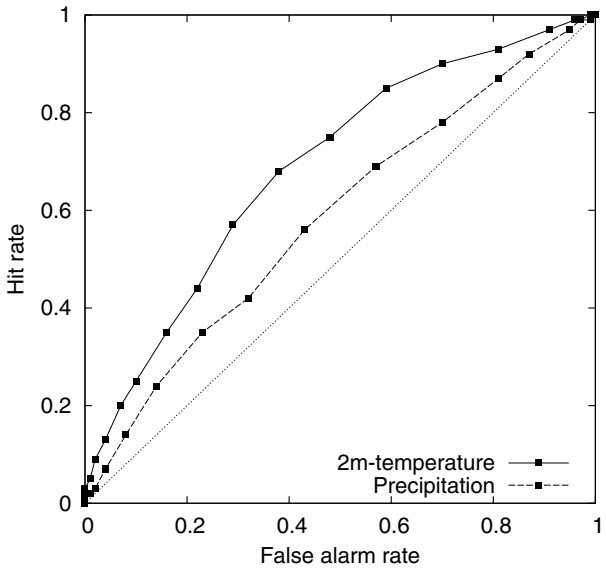

Fig 9. ROC curves for downscaled multi-model predictions for 2-m temperature (solid line) and precipitation (dashed line) for Europe in JAS season and JFM season, respectively.

Table 6. Multi-model (Météo-France, ECMWF and UK Meteorological Office models) ROC area for cross-validated ensemble predictions of upper tercile seasonal precipitation statistically downscaled to stations in five selected regions for seasons JFM, AMJ, JAS and OND. Period is 1961–2000, and lead time is two months. Statistical significance at the 5% level is indicated by bold numbers

	JFM	AMJ	JAS	OND
Europe	**0.58**	0.53	**0.54**	0.53
NW America	0.52	0.51	0.54	**0.57**
Contiguous US	**0.59**	0.53	0.52	0.53
Australia	**0.55**	**0.55**	**0.59**	**0.60**
Scandinavia	**0.59**	**0.59**	**0.55**	0.52

Table 7. As Table 6, but for 2-m temperature

	JFM	AMJ	JAS	OND
Europe	**0.67**	**0.65**	**0.69**	0.55
NW America	**0.62**	**0.61**	**0.65**	0.49
Contiguous US	**0.66**	**0.58**	**0.68**	0.52
Australia	**0.63**	**0.65**	**0.64**	**0.64**
Scandinavia	0.59	**0.67**	**0.60**	0.56

potential evaporation and total solar radiation are used in crop yield prediction.

As an example, we consider precipitation in the AMJ season for the station Nordby, which is the south-westernmost Danish station in Fig. 1. The skill for this station in terms of cross-validated correlations is shown in Table 8 for downscaled seasonal ensemble mean predictions, for predictions of the (log-odd transformed) probability of precipitation and for predictions of the (Fisher Z-transformed) daily persistence parameter. We note that the level of skill for prediction of the latter two is comparable to that of the conventional seasonal mean prediction.

Table 8. Cross-validated correlations for downscaled predictions of precipitation for station Nordby, Denmark in AMJ season for the three models

	Seasonal mean	Prob. wet day	Persistence
Météo-France	0.17	0.28	0.14
ECMWF	0.23	0.36	0.41
UK Met Office	0.04	−0.06	−0.05

The predicted probability of precipitation and daily persistence are used to generate ensembles of long daily sequences of occurrence of precipitation, using the method outlined in Section 2.3. On wet days, the precipitation amount is sampled from an exponential distribution whose mean is derived from the downscaled seasonal precipitation ensemble.

The distribution of daily precipitation in the AMJ season for Nordby for the 1961–2000 period is shown in Fig. 10 (note the logarithmic scale on the ordinate). The figure shows the observed distribution, the distribution of downscaled predictions, based on predictions for each year in the 1961–2000 period and stochastic generation of daily precipitation, and the distribution of raw model output of daily precipitation. We note that the precipitation in the dynamical models (raw model output) tends to be less intense than the observed precipitation, which is to be expected as the model precipitation approximately represents an average over a model grid cell. Also, the number of days with less than 1 mm precipitation is underestimated by the dynamical models. The stochastically generated daily precipitation suffers from the same deficiencies as the raw model output, but to a much lesser extent, and so agrees better with observations than the raw model output.

4. Concluding remarks

We have outlined a method for downscaling seasonal ensemble predictions. The main motivation for this work has been to enable existing crop yield models to take advantage of seasonal climate predictions.

The downscaling is performed in three steps: (i) a spatial downscaling of ensemble mean seasonal mean precipitation and 2-m temperature from dynamical model output to station level; (ii) application of the downscaling transformation to the model output ensemble and subsequent inflation (calibration) of the downscaled ensemble; (iii) a stochastic generation of daily precipitation conditioned on predictions of the probability of a wet day in the season and daily persistence.

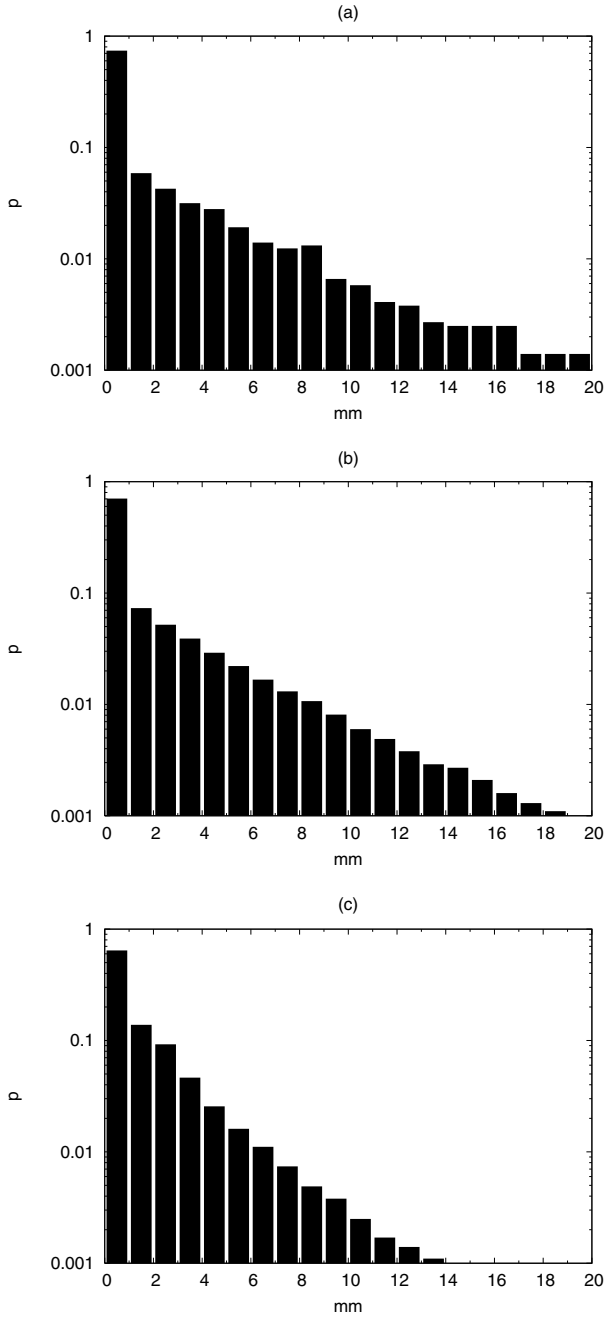

Fig 10. Histograms for 1961–2000 climatological distribution of daily precipitation amount for station Nordby, Denmark in AMJ season: (a) observed; (b) predicted using statistical downscaling of seasonal precipitation and stochastic generation of daily precipitation; (c) raw multi-model output.

Crop yield models also require input of daily values of a number of other variables, such as minimum and maximum temperature, evaporation and solar radiation. Daily time series of these variables can be modelled by a first-order autoregressive process, where the variables are assumed to follow normal distributions with mean and variance conditioned on the occurrence of precipitation and downscaled from the dynamical seasonal forecast model.

Our main focus has been on the derivation of the methodology. The examples have demonstrated that with the outlined statistical methods, we are able to downscale precipitation and 2-m temperature and obtain skill scores that, although modest, are significantly better than skill scores based on climatology. The downscaled ensemble mean predictions are generally more skilful for the observing stations than the raw model output, and the downscaled ensemble predictions are statistically more consistent with observations (flatter rank histograms) than the raw model output. For the single station that we tested, we found that the probability for a wet day in a season as well as the daily persistence could be predicted with skill comparable to that found for prediction of seasonal mean precipitation. Based on these predictions we can stochastically generate daily precipitation time series, the climatological distribution of which agrees well with that observed.

However, there is still room for improvement. For specific target regions it is very likely that the skill of the downscaling procedure can be improved by fine-tuning the predictor region and possibly by including additional predictors. We chose to base the downscaling on predictor patterns that were derived using an SVD analysis of the cross-covariance between model output and observations. Alternatively, one could use a canonical correlation analysis instead. Although we found in the examples that the downscaling generally led to increased predictive skill compared to the raw model output, there were still cases in which the raw model output was more skilful than the downscaled predictions. Direct inclusion of the raw model in the predictions could possibly further increase the predictive skill, e.g. by including a residual term in the predictor as suggested by Kharin and Zwiers (2001).

Evaluation of crop yield predictions using precipitation downscaled by the methods outlined in the present paper can be found in Cantelaube and Terres (2005) and Marletto et al. (2005).

5. Acknowledgment

This work was supported by the Commission of the European Union under contract No. EVK-CT-1999-00024 DEMETER.

References

Anderson, J. L. 1996. A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate* **9**, 1518–1530.

Bretherton, C. S., Smith, C. and Wallace, J. M. 1992. An intercomparison of methods for finding coupled patterns in climate data. *J. Climate* **5**, 541–560.

Cantelaube, P. and Terres, J. M. 2005. Use of seasonal weather forecasts in crop yield modelling. *Tellus* **57A**, this issue.

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

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

Feddersen, H. 2003. Predictability of seasonal precipitation in the Nordic region. *Tellus* **55A**, 385–400.

Feddersen, H., Navarra, A. and Ward, M. N. 1999. Reduction of model systematic error by statistical correction for dynamical and seasonal predictions. *J. Climate* **14**, 1974–1989.

Gershunov, A., Barnett, T. P., Cayan, D. R., Tubbs T. and Goddard, L. 2000. Predicting and downscaling ENSO impacts on intraseasonal precipitation statistics in California: The 1997/98 event. *J. Hydrometeorol.* **1**, 201–210.

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

Hamill, T. M. 2001. Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.* **129**, 550–560.

Hamill, T. M. and Colucci, S. J. 1997. Verification of eta-RSM short-range ensemble forecasts. *J. Atmos. Sci.* **125**, 1312–1327.

Huth, R., Kyselý, J. and Dubrovský, M. 2001. Time structure of observed, GCM-simulated, downscaled, and stochastically generated daily temperature series. *J. Climate* **6**, 4047–4061.

Kang, I.-S., Lee, J.-Y. and Park, C.-K. 2004. Potential predictability of summer mean precipitation in a dynamical seasonal prediction system with systematic error correction. *J. Climate* **17**, 834–844.

Katz, R. W. 1977. Precipitation as a chain-dependent process. *J. Appl. Meteorol.* **16**, 671–676.

Katz, R. W. 1996. Use of conditional stochastic models to generate climate change scenarios. *Climatic Change* **32**, 237–255.

Kharin, V. V. and Zwiers, F. W. 2001. Skill as a function of time scale in ensembles of seasonal hindcasts. *Climate Dyn.* **17**, 127–141.

Klein, W. H., Lewis, B. M. and Enger, I. 1959. Objective prediction of five-day mean temperatures during winter. *J. Atmos. Sci.* **16**, 672–682.

Laursen, E. V. 2004. DMI Daily Climate Data Collection 1873-2003, Denmark and Greenland. DMI Technical Report No. 04-03. The Danish Meteorological Institute, Copenhagen, Denmark.

Marletto, V., Zinoni, F., Criscuolo, L., Fontana, G., Marchesi, S. and co-authors. 2005. Evaluation of downscaled DEMETER multi-model ensemble seasonal hindcasts in a northern Italy location by means of a model of wheat growth and soil water balance. *Tellus* **57A**, 488–497.

Mearns, L. O., Rosenzweig, C. and Goldberg, R. 1997. Mean and variance change in climate scenarios: methods, agricultural applications, and measures of uncertainty. *Climatic Change* **35**, 367–396.

Michaelsen, J. 1987. Cross-validation in statistical climate forecast models. *J. Climate Appl. Meteor.* **26**, 1589–1600.

Morse, A. P., Doblas-Reyes, F. J., Hoshen, M. B., Hagedorn, R., Thomson, M. C. and co-authors. 2005. A forecast quality assessment of an end-to-end probalistic multi-model seasonal forecast system using a malaria model. *Tellus* **57A**, 464–475.

North, G. R., Bell, T. L. and Cahalan, R. F. 1982. Sampling errors in the estimation of empirical orthogonal functions. *Mon. Wea. Rev.* **110**, 699–706.

Palmer, T. N., Alessandri, A., Andersen, U., Cantelaube, P., Davey, M. and co-authors. 2004. Development of a European multi-model ensemble system for seasonal to inter-annual prediction (DEMETER). *Bull. Am. Meteorol. Soc.* **85**, 853–872.

Palutikof, J. P., Goodess, C. M., Watkins, S. J. and Holt, T. 2002. Generating rainfall and temperature scenarios at multiple sites: examples from the Mediterranean. *J. Climate* **15**, 3529–3548.

Parlange, M. B. and Katz, R. W. 2000. An extended version of the Richardson model for simulating daily weather variables. *J. Appl. Meteorol.* **39**, 610–622.

Richardson, C. W. 1981. Stochastic simulation of daily precipitation, temperature, and solar radiation. *J. Appl. Meteorol.* **17**, 182–190.

Rowell, D. P. 1998. Assessing potential seasonal predictability with an ensemble of multidecadal GCM simulations. *J. Climate* **11**, 109–120.

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.

Semenov, M. A. and Barrow, E. M. 1997. Use of a stochastic weather generator in the development of climate change scenarios. *Climatic Change* **35**, 397–414.

Stanski, H. R., Wilson L. J. and Burrows W. R. 1989. Survey of common verification methods in meteorology. World Weather Watch Technical Report No. 8, WMO/TD No. 358.

Stephenson, D. B., Rupa Kumar, K., Doblas-Reyes, F. J., Royer, J.-F. and co-authors. 1999. Extreme daily rainfall events and their impact on ensemble forecasts of the Indian monsoon. *Mon. Wea. Rev.* **127**, 1954–1966.

Stockdale, T. N. 1997. Coupled ocean–atmosphere forecasts in the presence of climate drift. *Mon. Wea. Rev.* **125**, 809–818.

Stockdale, T. N., Anderson, D. L. T., Alves, J. O. S. and Balmaseda, M. A. 1998. Global seasonal rainfall forecasts using a coupled ocean–atmosphere model. *Nature* **392**, 370–373.

Tippett, M. K., Barlow, M. and Lyon, B. 2003. Statistical correction of central southwest Asia winter precipitation simulations. *Int. J. Climatol.* **23**, 1421–1433.

Tuomenvirta, H., Drebs, A., Førland, E., Tveito, O. E., Alexandersson, H. and co-authors. 2001. Nordklim data set 1.0. Technical Report DNMI 08/01. The Norwegian Meteorological Institute, Oslo, Norway.

von Storch, H. 1999. On the use of 'inflation' in statistical downscaling. *J. Climate* **12**, 3505–3506.

von Storch, H., Zorita, E. and Cubasch, U. 1993. Downscaling of global climate change estimates to regional scales: an application to Iberian rainfall in wintertime. *J. Climate* **6**, 1161–1171.

Widmann, M., Bretherton, C. S. and Salathé E. P. Jr. 2003. Statistical precipitation downscaling over the northwestern United States using numerically simulated precipitation as predictor. *J. Climate* **16**, 799–816.

Wilby, R. L. and Wigley, T. M. L. 2000. Precipitation predictors for downscaling: observed and general circulation model relationships. *Int. J. Climatol.* **20**, 641–661.

Wilks, D. S. 1992. Adapting stochastic weather generation algorithms for climate change studies. *Climatic Change* **22**, 67–84.

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

Wilks, D. S. 1999. Multisite downscaling of daily precipitation with a stochastic weather generator. *Climate Res.* **11**, 125–136.