INTERNATIONAL JOURNAL OF CLIMATOLOGY
*Int. J. Climatol.* (in press)
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/joc.1318

# DOWNSCALING HEAVY PRECIPITATION OVER THE UNITED KINGDOM: A COMPARISON OF DYNAMICAL AND STATISTICAL METHODS AND THEIR FUTURE SCENARIOS

MALCOLM R. HAYLOCK,[a,*] GAVIN C. CAWLEY,[a] COLIN HARPHAM,[a] ROB L. WILBY[b] and CLARE M. GOODESS[a]
[a] *Climatic Research Unit, School of Environmental Sciences, University of East Anglia, UK*
[b] *Environment Agency, Trentside Offices, Nottingham, UK*

*Received 12 September 2005*
*Accepted 15 January 2006*

## ABSTRACT

Six statistical and two dynamical downscaling models were compared with regard to their ability to downscale seven seasonal indices of heavy precipitation for two station networks in northwest and southeast England. The skill among the eight downscaling models was high for those indices and seasons that had greater spatial coherence. Generally, winter showed the highest downscaling skill and summer the lowest. The rainfall indices that were indicative of rainfall occurrence were better modelled than those indicative of intensity. Models based on non-linear artificial neural networks were found to be the best at modelling the inter-annual variability of the indices; however, their strong negative biases implied a tendency to underestimate extremes. A novel approach used in one of the neural network models to output the rainfall probability and the gamma distribution scale and shape parameters for each day meant that resampling methods could be used to circumvent the underestimation of extremes. Six of the models were applied to the Hadley Centre global circulation model HadAM3P forced by emissions according to two SRES scenarios. This revealed that the inter-model differences between the future changes in the downscaled precipitation indices were at least as large as the differences between the emission scenarios for a single model. This implies caution when interpreting the output from a single model or a single type of model (e.g. regional climate models) and the advantage of including as many different types of downscaling models, global models and emission scenarios as possible when developing climate-change projections at the local scale. Copyright © 2006 Royal Meteorological Society.

KEY WORDS: downscaling; precipitation; United Kingdom; extremes; climate-change scenarios

## 1. INTRODUCTION

General circulation models (GCMs) are our most important tools for estimating the climate under scenarios of greenhouse gas emissions. However, the discrepancy between the scale at which the models deliver output and the scale that is required for most impact studies has led to the development of downscaling methodologies. The field of downscaling is divided into two approaches: the nesting of high-resolution regional climate models (RCMs) in the GCMs (dynamical) and the statistical representation of desired fields from the coarse resolution GCM data (statistical).

While there have been several reviews of statistical downscaling methodologies (Giorgi and Mearns, 1991; Hewitson and Crane, 1996; Wilby and Wigley, 1997; Wilby *et al*., 1998), there have been few comprehensive comparisons of model performance. A notable exception is Wilby *et al*. (1998), who downscaled daily precipitation at six US regions using six downscaling models, including weather generators, resampling methods and artificial neural networks (ANN). They found the neural network models to be the least skilful at reproducing observed rainfall due to poor simulation of wet-day occurrence. The six models were applied

* Correspondence to: Malcolm R. Haylock, Climatic Research Unit, University of East Anglia, Norwich, NR4 7TJ, UK;
e-mail: M.Haylock@uea.ac.uk

to output from a GCM climate-change experiment to show that changes in the downscaled precipitation were generally smaller than the changes in the GCM precipitation. Since Wilby *et al*. (1998), there has been a large increase in the number of studies involving RCMs, but few studies comparing these dynamical methods with statistical models. Wilby *et al*. (2000) applied NCEP-reanalysis downscaled precipitation from two models to a hydrological model and compared the daily precipitation, runoff and temperature with observations in the Animas River basin, Colorado. They found comparable performance from a linear regression statistical downscaling model and elevation-corrected output from a RCM, which were both better than the raw NCEP precipitation. Similarly, Murphy (1999) found a linear regression statistical model to have comparable skill to a RCM in downscaling monthly precipitation and temperature at 976 European stations. Kidson and Thompson (1998) also found similar skill between statistical and dynamical methods when they downscaled daily precipitation and minimum and maximum temperature using a statistical regression technique and a single 50 km RCM at 78 stations in New Zealand.

Since the early 1990s, it has been known that the largest changes in the climate under enhanced greenhouse conditions were likely to be seen in changes to the extremes (Gordon *et al*., 1992); however, none of the above studies have addressed downscaling heavy rainfall. The European Union STARDEX project (**STA**tistical and **R**egional dynamical **D**ownscaling of **EX**tremes for European regions–Goodess *et al*., 2005) was devised to overcome these gaps in the literature and provide recommendations on downscaling methodologies to stakeholders.

STARDEX had two main objectives:

- To rigorously and systematically inter-compare and evaluate statistical, dynamical and statistical-dynamical downscaling methods for the reconstruction of observed extremes and the construction of scenarios of extremes for selected European regions
- To identify the more robust downscaling techniques and to apply them to provide reliable and plausible future scenarios of temperature and precipitation-based extremes for selected European regions.

This paper addresses both these objectives by presenting validation of the ability of six statistical and two dynamical downscaling models to reproduce observed heavy precipitation at selected UK stations. Six of the models were then applied to output from climate-change experiments of the Hadley Centre GCM HadAM3P (Pope *et al*., 2000) to construct possible scenarios of UK heavy precipitation for the end of the twenty-first century.

The paper is divided into the following sections: Section 2 presents the data and methodologies used in the study; Section 3 introduces the downscaling models; Section 4 presents validations of the models' ability to downscale heavy precipitation; Section 5 presents scenarios of precipitation using the downscaling models applied to a GCM; and Section 6 concludes the study.

## 2. DATA AND METHODOLOGY

A network of stations in two UK regions with contrasting climates was chosen. The use of two contrasting regions, the near-continental southeast England (SEE) and maritime northwest England (NWE), was used to encourage the development of downscaling methodologies that could be applied to a variety of climates. We selected stations in the two regions that had at least 80% non-missing daily precipitation observations for at least 80% of the years 1958–2000. This resulted in 28 stations in SEE and 15 in NWE, the locations of which are shown in Figure 1.

The annual cycle of precipitation and number of days with precipitation $>1$ mm for the two regions are shown in Figure 2. While both regions receive higher precipitation and more rainy days in the winter months, NWE is noticeably wetter being more exposed to the prevailing westerly airflow from the North Atlantic.

The period of analysis was divided into separate calibration (for the statistical models only) and validation periods. These were 1958–1978 and 1994–2000 for calibration and 1979–1993 for validation. This validation period was chosen as it was for this period that data were available for the two RCMs run under observed

 

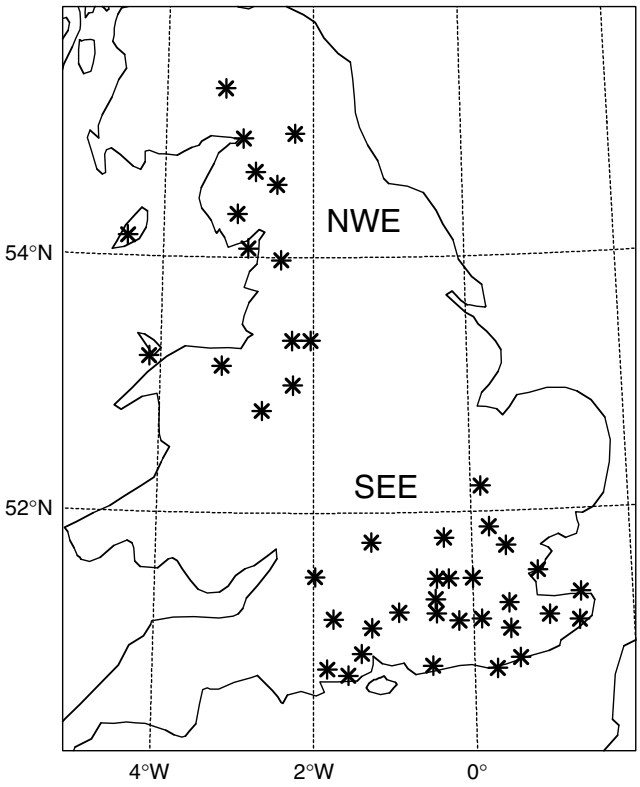


Figure 1. Location of stations in the two study regions southeast England (SEE) and northwest England (NWE)

climate conditions nested in ERA-15 reanalysis data (Gibson *et al*., 1997). These model runs were made available by the MERCURE project (Machenhauer *et al*., 1996). Statistical downscaling models were designed and calibrated using predictors from the NCEP-reanalysis project (Kalnay *et al*., 1996). Although it would be desirable to have the statistical and dynamical models forced by the same boundary conditions, experiments using the RCMs nested in the NCEP reanalyses were not available, nor was ERA-40 data at the time of this study, for training the statistical models. Still, a comparison of the storm tracks in the ERA-15 and NCEP reanalyses was carried out by Hodges *et al*. (2003), who concluded that lower-tropospheric NH storm tracks were very similar in the two data sets, with ERA-15 having slightly more intense systems.

The focus of the exercise was on heavy precipitation. Therefore, we used a suite of seven precipitation indices calculated for each season (Table I). The use of climate indices for monitoring extremes is following

Table I. Abbreviations, names and descriptions of the seven indices of daily precipitation used in the study

| Index | Name | Description |
|---|---|---|
| pav | Mean precipitation | Average precipitation on all days |
| pint | Precipitation intensity | Average precipitation on days with >1 mm |
| pq90 | Precipitation 90th percentile | 90th percentile of precipitation on days with >1 mm |
| px5d | Maximum 5-day precipitation | Maximum precipitation from any five consecutive days |
| pxcdd | Maximum consecutive dry days | Maximum number of consecutive days with <1 mm |
| pfl90 | Fraction of total from heavy events | Fraction of total precipitation from events > long-term 90th percentile |
| pnl90 | Number of heavy events | Number of events > long-term 90th percentile |

 

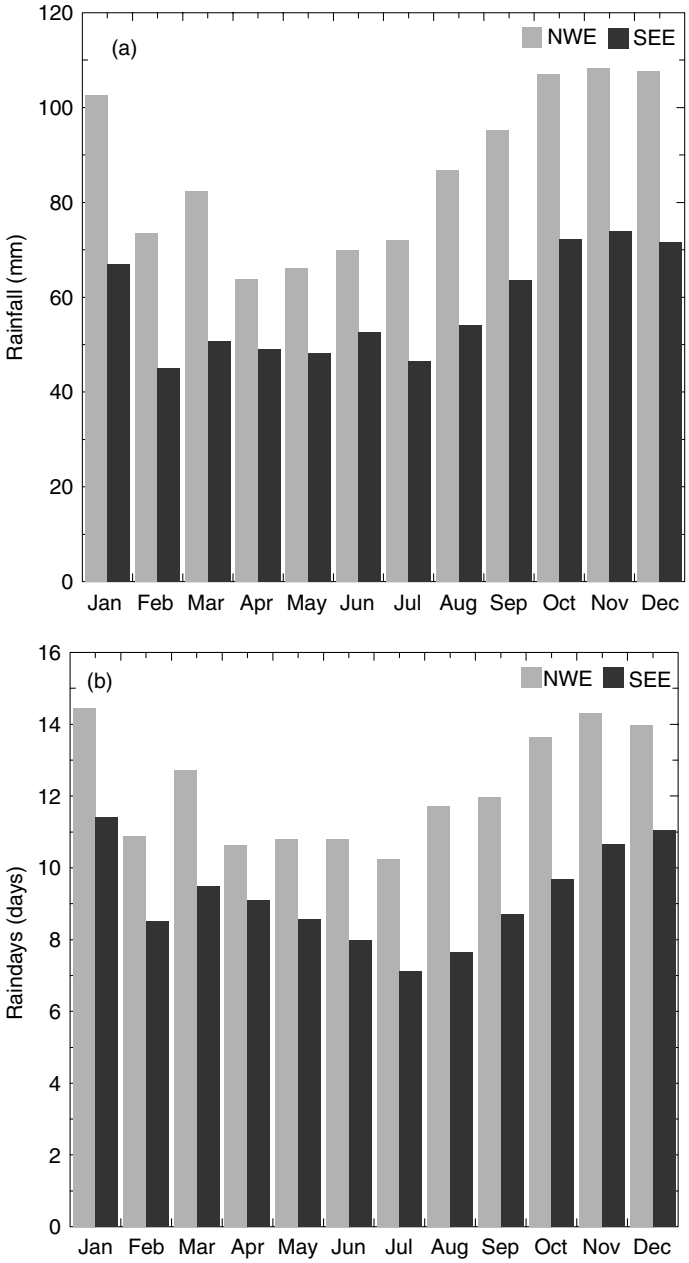


Figure 2. (a) monthly precipitation and (b) number of days with rain >1 mm averaged across all stations and years for the two regions

the recommendations of Nicholls (1995) and the indices were designed in agreement with the methodologies of Nicholls and Murray (1999): that they be applicable to a variety of climates; that they are designed to maximise their independence (low correlation); that only raindays are used to calculate percentiles; and that they consider the fraction of total rainfall from extreme events. Four of the seven indices relate to very wet events: 90th percentile; maximum 5-day total; number of heavy events; and proportion of total from heavy events. One index describes very dry events (maximum number of consecutive dry days) and two indices describe changes to the entire distribution (mean daily precipitation and precipitation intensity). The threshold of 1 mm for a wet day was used because previous studies have found that lower thresholds can

 

be sensitive to problems such as under reporting of small precipitation amounts and changes in the units of measurement (e.g. Hennessy *et al*., 1999). A dry day is defined as having less than 1 mm precipitation. These indices were used for validation of the downscaling models (Section 4) as well as presentation of changes in extremes under future emission scenarios (Section 5). FORTRAN software for calculating these indices, along with 52 other indices of precipitation and temperature, can be downloaded from the STARDEX web site (http://www.cru.uea.ac.uk/cru/projects/stardex/).

## 3. DOWNSCALING METHODOLOGIES

The eight downscaling models can be classified as either statistical or dynamical. The statistical models are all regression based and are constructed by deriving empirical relationships between the large-scale GCM predictors and the station-scale predictands. Although two of the models contain a stochastic element, there are no purely stochastic weather generators (e.g. WGEN; Richardson, 1981) that generally use Markov processes to model rainfall occurrence and possibly amount. Nor are there any circulation pattern downscaling models that relate observed rainfall to a weather pattern classification scheme (Wilby, 1995), although large-scale circulation predictors are used in all models. Dynamical models are high-resolution RCMs, nested in the coarser resolution GCMs.

### *3.1. Statistical models*

The statistical models can be subdivided into *direct* and *indirect* methods. Direct methods explicitly model the seasonal indices of heavy precipitation using seasonal predictors. Their advantage is that they are generally less demanding numerically due to the smaller data requirements. Indirect methods model daily precipitation using daily predictors from which the precipitation indices are calculated. Their advantage is that we have access to the downscaled daily precipitation data that can be used in impact models and for further analyses.

Statistical models can also be divided into single and multi-site methods. Single-site methods model each station independently. Multi-site methods model all sites simultaneously, thereby possibly maintaining inter-station relationships, e.g. correlation. Of our six statistical models, four are multi-site methods. However, in validating the models we did not consider inter-station behaviour, as we were primarily interested in the ability of the models to downscale rainfall at a single station.

#### *3.1.1. Direct methods.*

*CCA*. Canonical correlation analysis (CCA) is a multi-site method used to model the seasonal precipitation indices directly using seasonal means of circulation variables. Four predictors were chosen by examining correlations between the observed precipitation indices and the potential predictors in the NCEP reanalyses. These are sea-level pressure, relative humidity at 700 hPa, specific humidity at 700 hPa and temperature at 700 hPa. Predictors were examined over the region 35° to 70 °N and 20 °W to 15 °E, however, the size of the region had little impact on the skill of the model. More details about the selection of potential predictors can be found in Haylock and Goodess (2004). For each season and precipitation index, a CCA was carried out using all 15 possible combinations ($2^4 - 1$) of the four predictors. The best set of predictors was selected using cross validation in the training period, whereby each year was removed and the model trained on the remaining years. The missing year was then hindcast and the skill measured by averaging across all stations the Spearman rank correlation between the observed and hindcast indices. Therefore, the predictor set varies between indices and seasons but is the same for all the stations in the region.

The canonical patterns and series were calculated using a singular value decomposition of the cross-covariance matrix of the principal components (PCs) of the predictor and predictand fields. This is numerically more stable than the more common method of working with the joint variance-covariance matrix (Press *et al*., 1986) and also incorporates the pre-filtering of the data by using just the significant PCs (Barnett and Preisendorfer, 1987). Bretherton *et al*. (1992) discuss the benefits of this methodology further and compare it with other methods of finding coupled modes. The number of PCs retained for the analysis was selected

 

by a Monte Carlo process, whereby 1000 PC analyses were carried out using data randomly resampled in time from the original series (Preisendorfer *et al.*, 1981). In each of the 1000 analyses, the eigenvalues were calculated. Each of the eigenvalues of the real observations was then compared against the distribution of the 1000 randomly generated values to determine if they were greater than the rank 50 eigenvalue (equivalent to $p < 0.05$). Therefore, the number of eigenvectors retained was different for each predictor, predictand and season.

Once the best set of predictors was chosen using the above-mentioned cross validation procedure in the calibration period, the model was calibrated using these predictors for the entire calibration period 1958–1978 and 1994–2000 for application to the validation period. This was the method adopted for all the statistical models.

*3.1.2. Indirect methods.*

Four of the downscaling methods employ ANN. These regression-based models employ non-linear transfer functions to map the predictors to the predictands. The predictors were daily values of 26 variables comprising surface pressure, temperature and humidity as well as upper air measures of wind speed and direction, vorticity, divergence, humidity, temperature and geopotential height. The predictors were available over several UK grid boxes at concurrent and forward lagged time steps, giving over 200 possible predictors for each region. The final indirect method, a conditional resampling method, also uses these same predictors.

*MLPK*. MLPK is a multi-site multi-layer perceptron (MLP) ANN (Bishop, 1995). The MLP model maps a set of inputs (the predictors) to a set of outputs (the predictands) via a single set of nodes (the *hidden layer*) using non-linear transformations of a weighted sum of the inputs. The weights were chosen to minimise the error of the output compared with observations using a sum-of-squares error metric. Minimising the sum-of-squares error leads to a model that estimates the conditional mean of the target distribution. Predictors were chosen from the full set of over 200 by a stepwise multiple linear regression (SMLR) procedure. Predictors were chosen using the area-average rainfall and then the same predictors were used to model each of the stations independently. Two other predictor selection methods were tested: a genetic algorithm method based on the application of Darwinian evolution theory (Holland, 1975); and a compositing procedure. The SMLR method performed best with the genetic algorithm yielding far too many predictors for regression analysis. A two-stage modelling process was adopted whereby rainfall occurrence and amounts were modelled separately. If the occurrence model predicted a probability of rainfall greater than 0.4, then the amounts model was used to give the expected rainfall amount, otherwise modelled rainfall for that day was set to zero. Further details of this model are given in Harpham and Wilby (2005).

*MLPS and MLPR*. These two single-site methods use a MLP ANN with a single hidden layer. Daily rainfall in the United Kingdom is generally dominated by frontal activity, which, due to its skewness, is better modelled using the Gamma distribution (Stern and Coe, 1984). Therefore, we used a Gamma function error metric, as distinct from the MLPK model that uses a sum-of-squares. In addition, the discrete nature of the occurrence of precipitation/dry day was incorporated into the error metric with a Bernoulli term (Williams, 1998). The models were then trained to estimate the probability of rainfall as well as the shape and scale parameters of a Gamma distribution for the amount of rainfall.

With over 200 possible predictors at each station, we employed a Bayesian regularisation scheme to avoid over fitting (Williams, 1995), which incorporates into the error metric a term that penalises overly complex models. We also used automatic relevance determination (ARD; Mackay, 1992) to determine which of the predictors were the most important. Further details of this model can be found in Cawley *et al*. (2003).

The difference between the MLPS and MLPR models was the way we interpreted the three output parameters: the probability of rainfall and the shape and scale parameters. For MLPS, we calculated an amount of rainfall for each day by multiplying together the three parameters, which gives the expected amount of rainfall and tends towards the conditional mean. We are primarily interested in heavy precipitation, which is unlikely to be adequately captured by an estimate of the conditional mean of the data. Therefore, for MLPR we randomly decided if it was a wet day (based on the probability of rainfall). Then, if it was a wet

 

day we chose an amount by selecting a random percentile from the Gamma distribution using the modelled shape and scale parameters. When assessing the skill of the MLPR model (Section 4), we generated 1000 possible realisations and averaged the skill across these. Similarly, when generating scenarios for this model (Section 5) we averaged across 1000 realisations.

In addition, the MLPS and MLPR models were run 20 times, each time selecting possibly different predictors and converging on slightly differing outputs. This was because the models were initialised with random predictor weights. This gave 20 different estimates of the three rainfall parameters. We created separate realisation for each of the 20 models and averaged the skill or scenario projections across these.

To aid distinguishing the three MLP models in the following discussion, it is useful to remember MLPS is a *simple* interpretation of the output parameters to give the expected rainfall whereas MLPR uses *resampling* to generate possible daily rainfall series. MLPK is the multi-site method.

*RBF*. Radial basis function (RBF) is a multi-site ANN. Similarly to the MLP ANN, it maps a set of inputs to a set of outputs using a non-linear *hidden layer*. However, the functions that comprise the hidden layer are quite different. MLP models are based on functions that divide the input data space into *hyperplanes* (which can be thought of as two-dimensional planes for a three-dimensional input space), optimised depending on the outputs. In an RBF network, the hidden layer consists of radial functions: functions whose response decreases monotonically from a central point. These functions divide the input data space into localised groups (which can be thought of as three-dimensional spheres for a three-dimensional input space). The radial functions are chosen from the inputs themselves independently of the outputs, which makes these networks much faster to train. In this case, a convergent K-means clustering algorithm was used to determine the hidden layer parameters (Anderberg, 1973). A two-stage amounts and occurrence modelling process was adopted similar to that used in the MLPK model. Further details of this model are given in Harpham and Wilby (2005).

*SDSM*. Statistical downscaling method (SDSM) is a two-step conditional resampling methodology. This multi-site method first downscales area-averaged precipitation using a combination of regression-based methods and a stochastic weather generator. Secondly, precipitation at individual sites is resampled from their distributions depending on the downscaled area-average precipitation. The resampling method is unique among all the models in that it preserves area-average precipitation and the spatial covariance of the daily rainfall. However, a consequence of this is that downscaled daily rainfall will never be greater than the maximum observed value, although multi-day totals can be greater. More detail regarding this model, including evaluations of its performance, can be found in Wilby *et al*. (2002) and Wilby *et al*. (2003).

### *3.2. Dynamical models*

*HadRM3*. HadRM3 is the third generation RCM developed by the Hadley Centre. Based on their atmospheric GCM, HadAM3 (Pope *et al*., 2000), this 19-level model has a horizontal resolution of 50 km $\times$ 50 km and four soil moisture levels. It has a comprehensive representation of atmospheric and land surface physics as well as an explicit sulphur cycle to estimate the concentration of sulphate aerosol particles from sulphur dioxide emissions.

During the STARDEX project, the Hadley Centre changed the version of their regional model from HadRM3H to HadRM3P. The main changes were in the representation of clouds and precipitation. Since we did not have access to either the results of the newer model nested in reanalyses data or the older model nested in the climate-change GCM, there is a difference in the Hadley Centre RCM we have validated (HadRM3H) and the one we have used in the climate-change experiments (HadRM3P). However, changes to the moisture parameterisations, although critical to this study, were undertaken to correct biases that were only observed in selected parts of the globe outside of Europe (Jenkins, Jones and Mitchell, personal communication).

*CHRM*. CHRM is an adaptation of HRM, the operational weather forecasting model of the German and Swiss meteorological services (Luthi *et al*., 1996; Vidale *et al*., 2003). This 20-level model has a horizontal resolution of 55 km $\times$ 55 km and three soil moisture levels. It has a full package of physical parameterisations including a mass-flux scheme for moist convection (Tiedtke, 1989) and Kessler-type

 

microphysics (Kessler, 1969; Lin *et al*., 1983). Vidale *et al*. (2003) analysed in detail the output of this model forced by observed boundary conditions and discussed important sensitivities of the model to hydrological parameterisations. In particular, they found an unnatural persistence of dry and hot conditions during summer due to insufficient evaporation.

## 4. VALIDATION OF DOWNSCALING MODELS

To validate the downscaling models, we calculated the Spearman (rank) correlation, bias and debiased root-mean-square error (RMSE) for each of the precipitation indices and seasons over the validation period 1979–1993. The statistical models were calibrated using the period 1958–1978 and 1994–2000. The correlation gives a validation of the inter-annual variability of the models independent of any bias or incorrect variance. Modelling the inter-annual variability is particularly important as it indicates that the models are correctly reproducing the predictor-predictand relationships. This provides confidence that changes in the predictors under climate-change conditions will produce correct changes in the predictands. As with all statistical downscaling methods, an important assumption is that the predictor-predictand relationships remain stationary in the future. The non-parametric Spearman correlation was used instead of the Pearson correlation to minimise the effect of outliers from the possibly non-Gaussian distributed indices. We also calculated the Pearson correlation (not shown) and found the results to be almost identical. From here on, when we mention correlation we are referring to the Spearman correlation. The bias indicates how the mean of the modelled indices compares with the observations independent of problems with inter-annual variability. It could be argued that the bias is the most important skill score if we are only interested in the average change in the rainfall indices under climate-change conditions. However, some of the direct methods that explicitly model the indices (e.g. CCA), will, by their design, have a low bias in the verification period. This is therefore misleading with regard to their true skill. The debiased RMSE measures the average error in the modelled indices once biases are removed. This indicates problems with incorrectly modelled inter-annual variability including variance.

Figures 3 and 4 show the correlation for each season for SEE and NWE respectively. Each of the four frames shows the correlation for each model and precipitation index for a single season. The correlations are presented as box-and-whisker plots showing the median, inter-quartile range (box), 5th and 95th percentiles (whiskers) for the correlations across all stations in the region. We have also shown the correlation for a simple model 'AREA' that uses the all-station average daily rainfall as the downscaled rainfall at each station. The purpose of including this model is to estimate a limit to the skill that can be obtained by a 'perfect GCM': a GCM that reproduces area-average rainfall accurately. Any model that performs better than the AREA model is doing so because of its ability to downscale the rainfall and not just model the area average. The AREA model also gives us an estimate of the spatial coherence of the rainfall indices; when the correlation skill of the AREA model is high, there is less difference between the inter-annual variability of the stations and the area average and so the inter-station correlation is higher. Note that we are not proposing the AREA model as a downscaling model but including it for diagnosing the spatial properties of the rainfall indices. Although we show below that area-average rainfall is a powerful predictor of station rainfall, it is generally poorly simulated in GCMs. In a comparison of GCM precipitation in five models to observe, Giorgi and Francisco (2000) found model biases mostly in the ∼40 to 80% range, with some areas exceeding 100%. However, recent work in STARDEX using rescaled GCM precipitation to model precipitation at the RCM scale over the Alps (Schmidli *et al*., 2005) shows that post-processing of the GCM precipitation can yield downscaling models with comparable skill to statistical models using other GCM predictors.

Figures 3 and 4 show some notable similarities. Firstly, the indices are not equally well modelled: pav and pxcdd are the two indices with the highest correlations and pfl90, pq90 and pint are the least well reproduced. This is quite consistent across regions and seasons. One would expect that indices reflecting the entire distribution (e.g. pav) to be better modelled and those representing the extremes (e.g. pq90) to be less so. However, this is contradicted by the fact that average precipitation intensity (pint) has among the lowest correlations and pxcdd has some of the highest. The pint index is a measure of rainfall intensity with

 

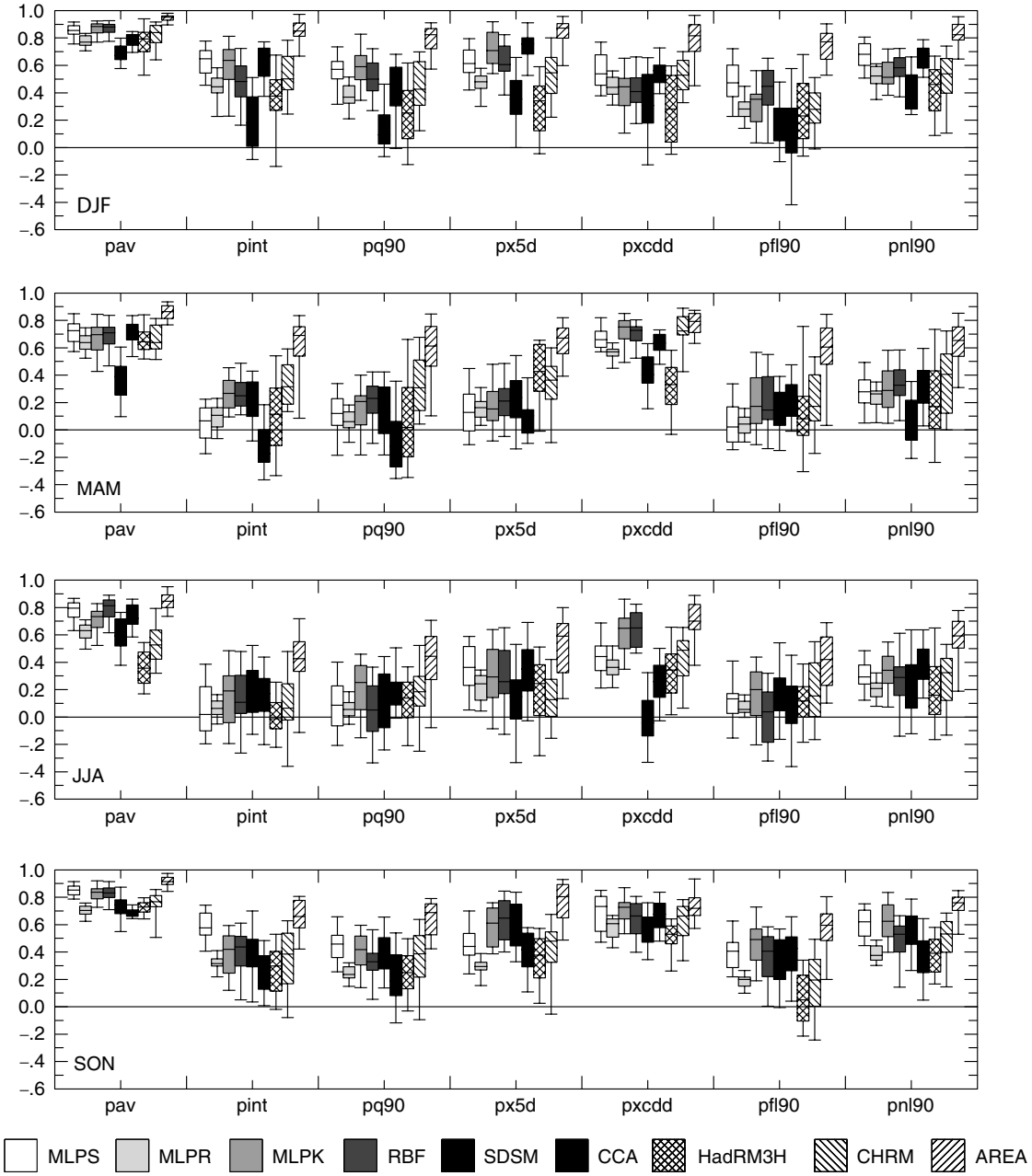


Figure 3. Correlation of modelled and observed precipitation indices for each season for SEE. Bars show inter-quartile range and median with 5th and 95th percentiles indicated by outer range

the effect of the number of wet days removed, whereas pxcdd is a count of dry days with no information about intensity. This implies that the models are doing better at determining the occurrence of rainfall than the intensity. Additionally, the correlations for the AREA model show that those indices and seasons with higher spatial coherence (higher AREA correlation) are better reproduced in the models. The AREA model generally has the highest correlations, suggesting that the area-average rainfall from a 'perfect GCM' is a better predictor than other variables with regard to inter-annual variability. This is more so for SEE than NWE and is also more apparent in DJF and MAM. Therefore, in SEE during DJF and MAM the

 

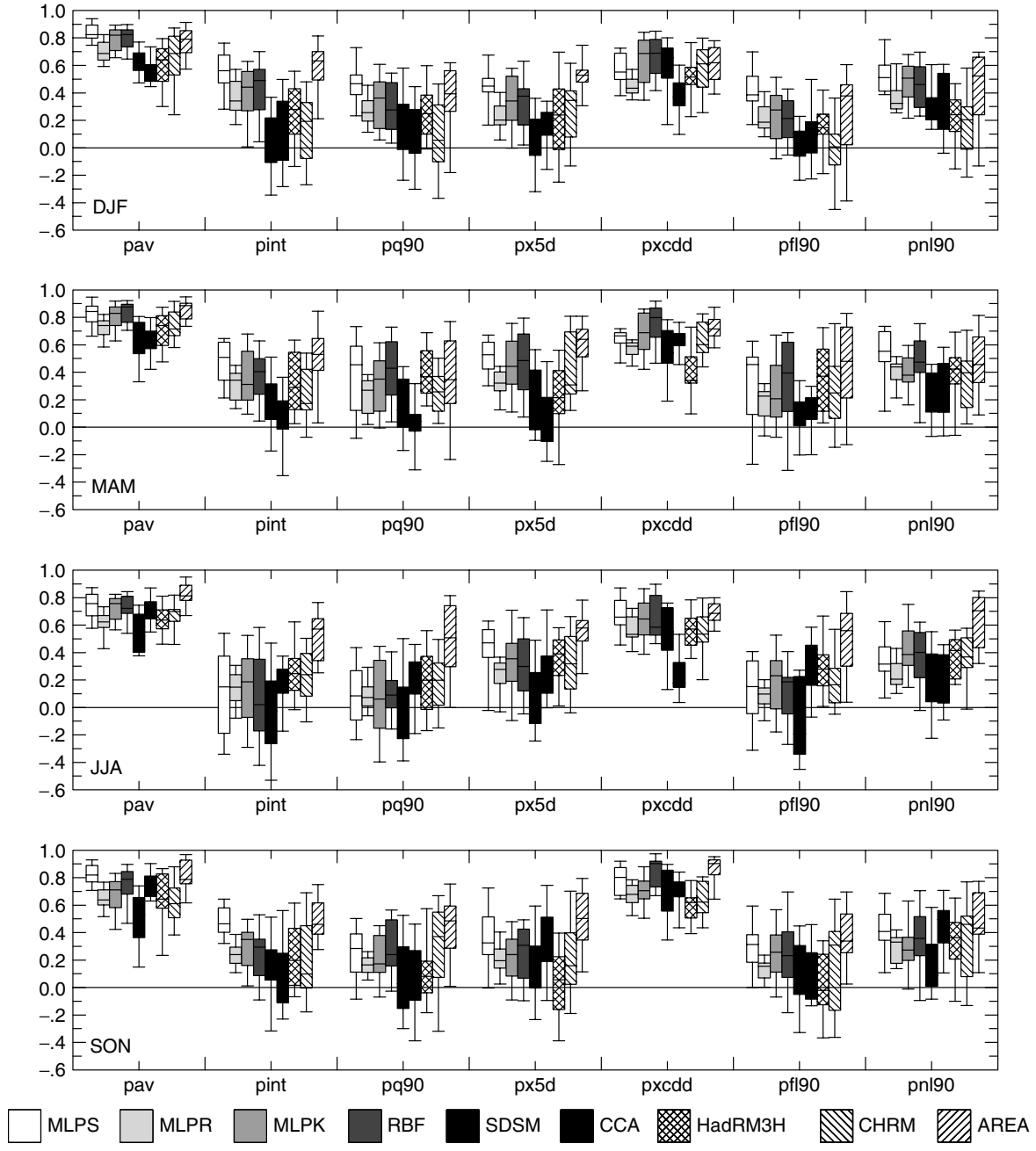


Figure 4. As for Figure 3 but for NWE

spatial coherence is high but the models are not performing as well as one might expect. This suggests that there is possibly some process that is driving the inter-annual variability of the precipitation in SEE in DJF and MAM that is being missed by the models or not captured by the predictors. A second notable result evident in Figures 3 and 4 is that the spread of correlations among the stations is very large for each model. In many cases, the range extends into negative correlations, although this is never the case for pav. This indicates that there is some site-specific behaviour that is not being captured by the regional scale predictors.

 

There are some notable differences between the correlations in the two regions. Firstly, the correlation varies between the seasons more in SEE than in NWE. In SEE, the correlations are generally higher in DJF and SON than in MAM and JJA. Secondly, the correlations in SEE are generally higher than NWE in DJF and SON and lower in MAM and JJA. Both these observations can be explained by the fact that the model skill is probably dependent on the degree to which inter-annual variability is being driven by the regional scale circulation. In those regions and times of the year when precipitation is more dominated by localised convection, we would expect the skill to be lower.

While Figures 3 and 4 show the relative correlation performance of the models for each season, and index, it is difficult to draw general conclusions about which models perform the best. We therefore assigned the models a rank based on their correlation relative to the other models for each station, index and season. We then averaged the rank across all stations, indices and seasons to arrive at an overall measure of correlation skill. This is shown in Figure 5. Models with a higher rank performed, on average, better than those with a lower rank. For NWE, the best models were the three neural network models MLPS, MLPK and RBF, followed by the two regional models, then MLPR, CCA and SDSM. For SEE, the three leading models were the same, although reordered, followed by CHRM, CCA, SDSM, HadRM3H and MLPR. In both regions, as previously noted, the AREA correlations outperform all the models, indicating that the area-average precipitation from a perfect GCM is the best predictor.

We repeated the average ranking exercise for the debiased RMSE and bias statistics, however, we assigned a greater rank to those models with a lower RMSE or bias. In addition, we used the absolute value of the bias, ignoring the sign, as we are primarily interested in the magnitude of the difference between the modelled and observed indices. The average ranks for these two statistics are shown in Figures 6 and 7. The RMSE average ranks are very similar to the correlation averages (Figure 5) with the exception that the CCA model, which performs much better with the third highest score in NWE and second highest score in SEE. The debiased

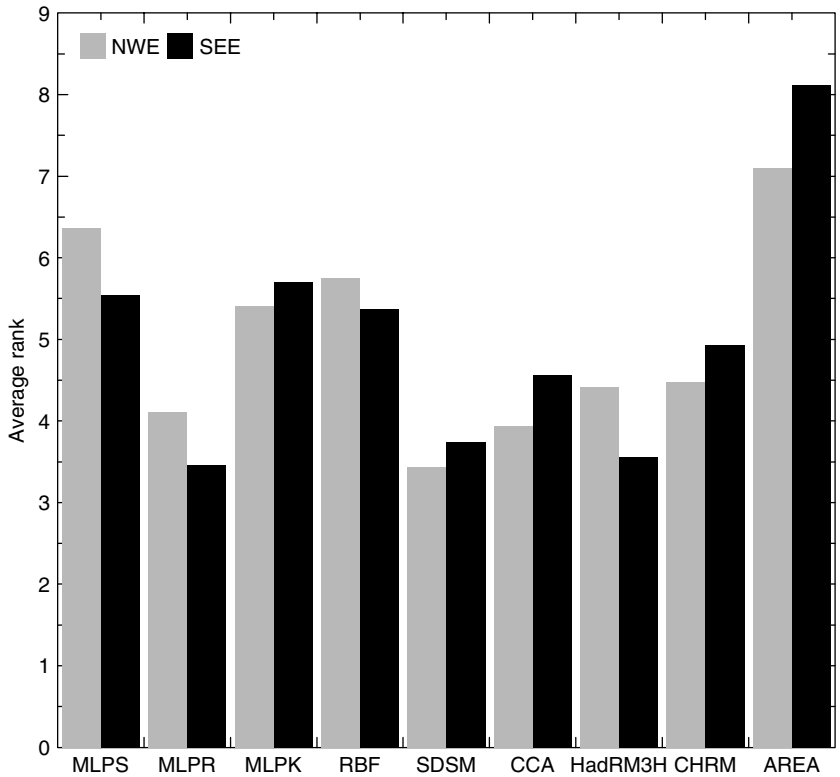


Figure 5. Average rank of correlation between models averaged across all indices, seasons and stations. Higher ranks indicate better performance

 

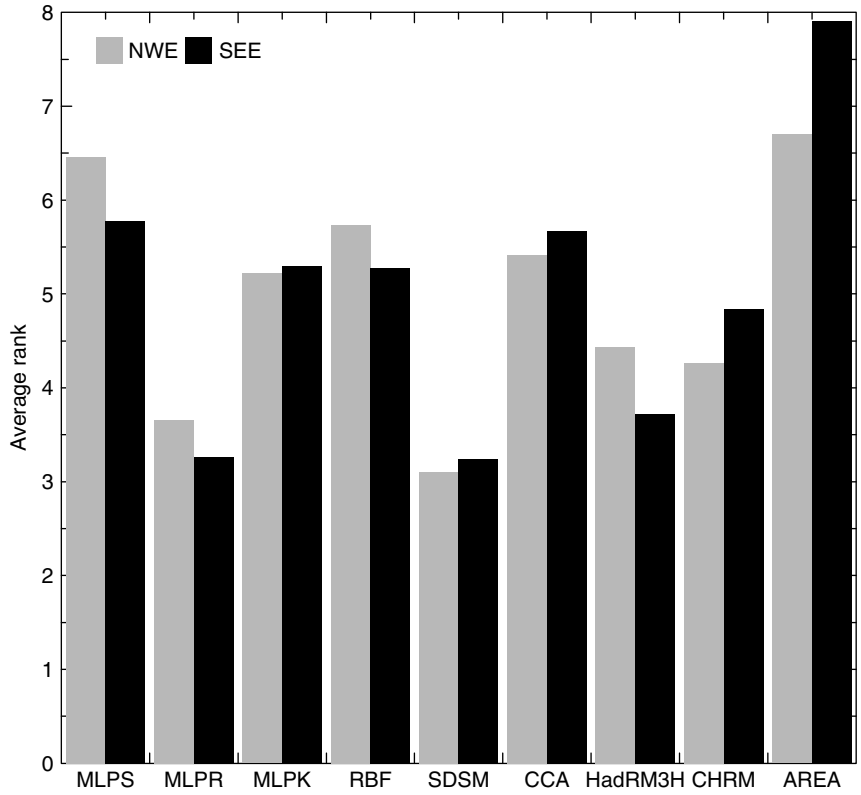


Figure 6. As for Figure 5 but for debiased RMSE

RMS score measures the skill of the reproduction of the inter-annual variability of the indices including the variance. The greater comparative skill of the CCA model with regards RMSE than correlation implies it is doing a better job with the variance than the other models as it's correlation skill is low. Again the AREA skill is greater than all the downscaling models.

The average bias rankings are very different to the correlations. In general, the models with the highest correlation rankings have the lowest bias rankings. The CCA and MLPR models have the lowest biases (highest bias ranks) and the MLPS, MLPK and RBF models have the highest biases (lowest bias ranks). An examination of the box-and-whisker plots of the bias scores (not shown) shows that the three neural network models with high bias consistently underestimate the heavy precipitation indices. A similar result was found by Harpham and Wilby (2005), who showed using quantile–quantile plots of the daily precipitation that the MLPK and RBF models consistently underestimated heavy precipitation and the SDSM model performed much better. This can be explained by the fact that the regression methods without a stochastic component that model daily rainfall (MLPK, MLPS and RBF) model the mean expected rainfall for each day. They will therefore tend to underestimate the extremes. The SDSM model, although modelling mean area-average rainfall for each day, includes a resampling component, which reduces the tendency towards mean values. Similarly, the MLPR model contains a stochastic component that diminishes the tendency to underestimate the extremes. The CCA model, although regression based, is designed to explicitly model the extremes and so would be expected to have little bias. The comparative AREA bias statistics are moderate. This is to be expected since, while the inter-station correlations of the indices may be high, we would expect their means to be different due to the differing exposure, aspect and altitude of the stations.

The three ANN models that tend to underestimate the extremes also have the highest bias for the pxcdd (maximum dry spell length) index. The MLPS model generally underestimates this index because the way the daily rainfall is calculated, by multiplying the probability of rain with the gamma shape and scale parameters,

 

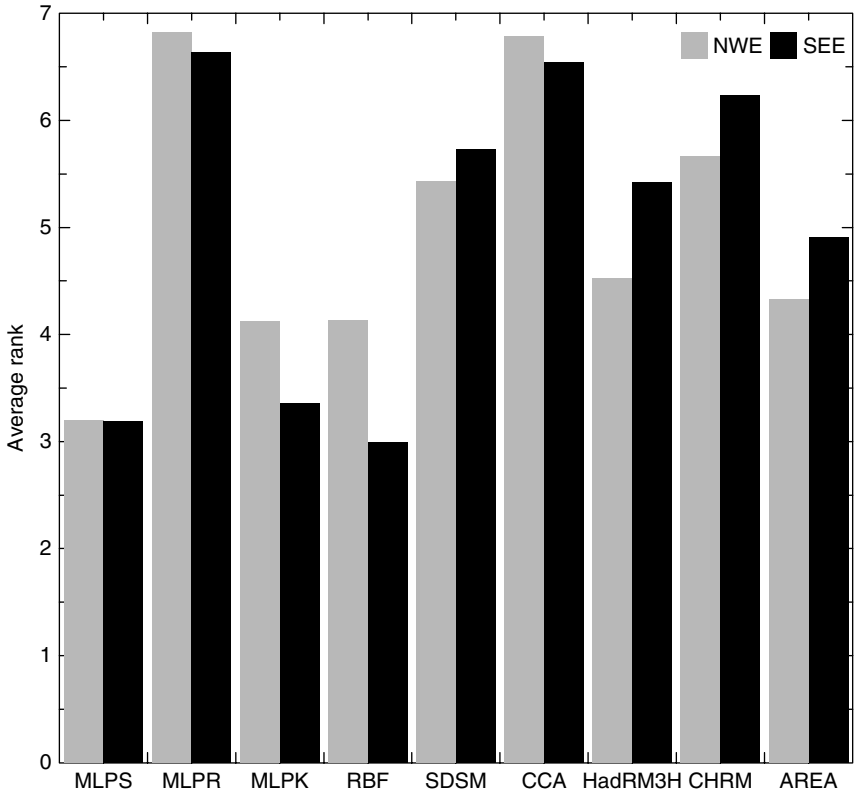


Figure 7. As for Figure 5 but for bias

it always gives a rainfall amount (however small) and so tends to overestimate the number of days with low rainfall. The MLPK and RBF models generally overestimate this index.

## 5. SCENARIOS OF HEAVY PRECIPITATION

Six of the downscaling models were applied to the output of climate-change experiments using the Hadley Centre high-resolution atmospheric GCM HadAM3P (Pope *et al*., 2000). HadAM3P data were available for an ensemble of three 30-year members using the A2 emission scenario (IPCC, 2000) and one B2 scenario member. For each ensemble member, we had access to HadAM3P data for the control period 1961–1990 and the projected period 2071–2100. After regridding the HadAM3P data to match the NCEP grid with which the downscaling models were trained, the six models were forced by these data and the precipitation indices calculated. In Section 4, the neural network models were calibrated using data with zero mean and unit standard deviation. Therefore, for the control period 1961–1990, each of the predictors was normalised accordingly and then the 1961–1990 means and standard deviations were applied to the 2071–2100 period to convert these predictors to the normalised space. The CCA method did not employ such normalisations. This is because this methodology uses predictors based on large-scale PCs of variables rather than grid point values for the neural networks, which were found to reliably simulate the GCM due to their large-scale nature.

Figures 8 and 9 show the proportional change in each model for SEE and NWE under the A2 scenario. Values above 1.0 correspond to an increase in the precipitation index in 2071–2100 compared with 1961–1990 and values below 1.0 correspond to a decrease. The three 30-year A2 ensemble members were treated as one 90-year experiment and so these changes refer to the proportional change in the means for the two 90-year periods. The spread of values for each model represents the inter-quartile range among the 28 stations (box)

 

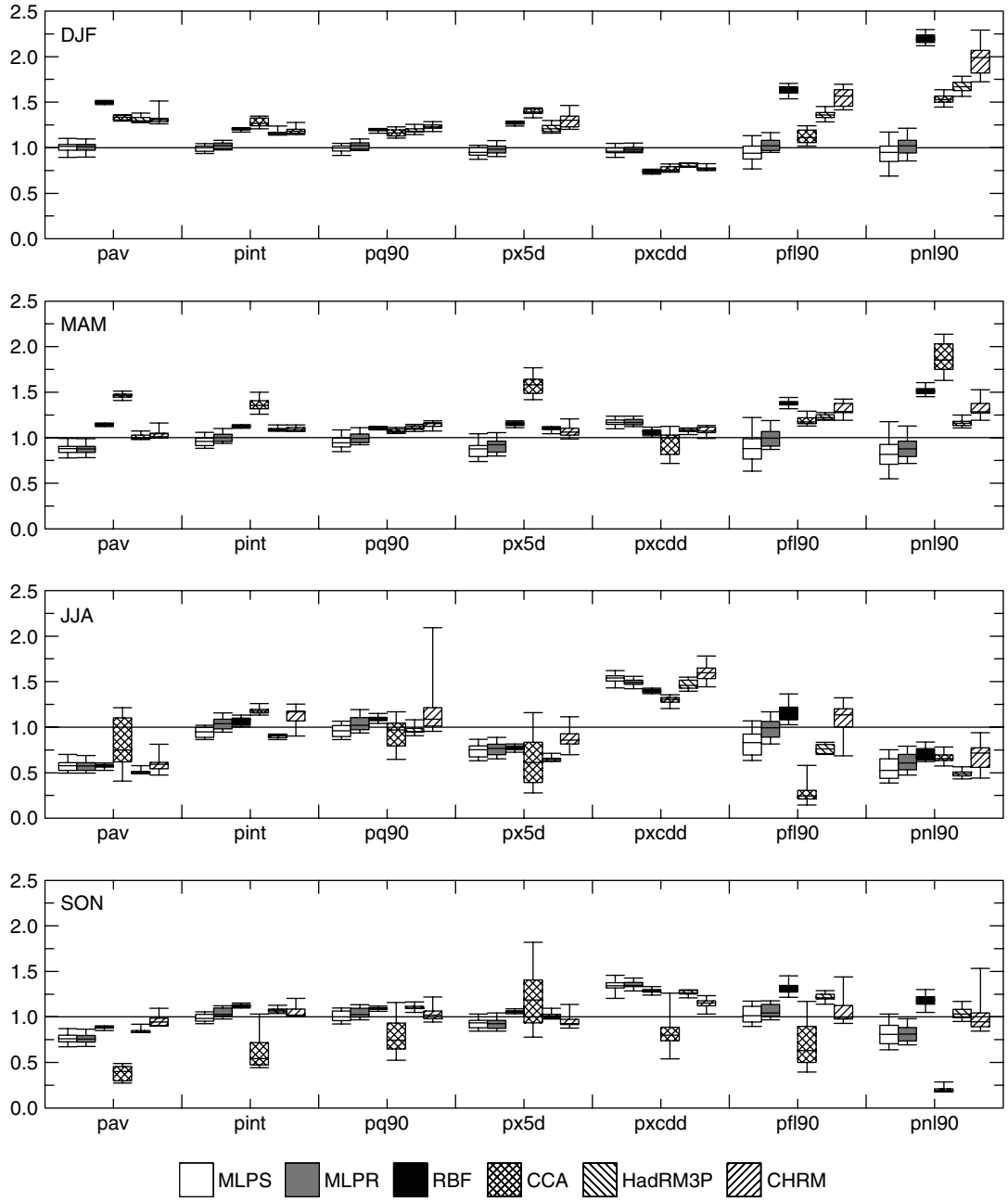


Figure 8. Factor change in the indices for SEE under A2 scenario

and the 5th and 95th percentiles (whisker). These figures show a general agreement among the models to wetter conditions in winter and drier conditions in summer. However, there is a notable spread among the models. In particular, in DJF the MLPS and MLPR models suggest there is little change in the indices, and in SON in SEE the CCA model predicts a larger decrease in the indices than the other models. In the latter case the large spread of values for the CCA model in SON across all the stations suggests that there is more uncertainty in the projections for this season, however, the other models have generally lower ranges. The differences in the values between models are at least as large as the differences between stations for a single

 

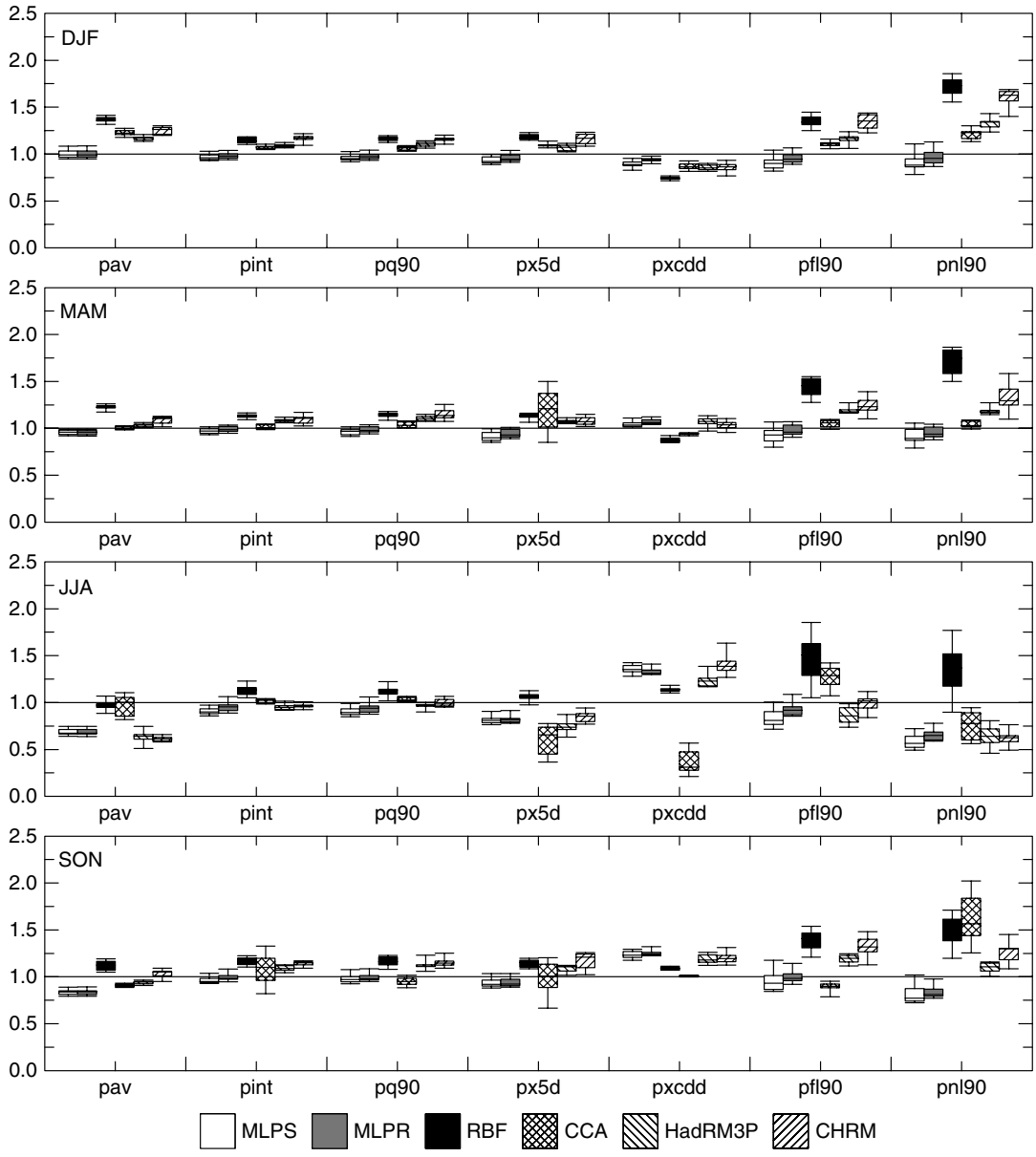


Figure 9. As for Figure 8 but for NWE

model. Similar plots were done for the B2 scenario (not shown) that matched the A2 results except that all the changes were closer to the line of no change. Importantly, the changes from A2 to B2 for a single model were of the same order of magnitude as the spread between models for a single scenario. This highlights the importance of incorporating many types of downscaling models into climate-change projections at local scales.

Since no one downscaling model, or group of models, stands above all others in terms of performance across all skill scores in the validation exercise (Section 4), we have no way to distinguish among the different projections for the different models. Therefore, we have treated the differing results from the models as a source of uncertainty. In addition, we cannot assign probabilities to the two available emission scenarios and must assume these to be equally likely (IPCC, 2000). With six models and two emission scenarios, we can

 

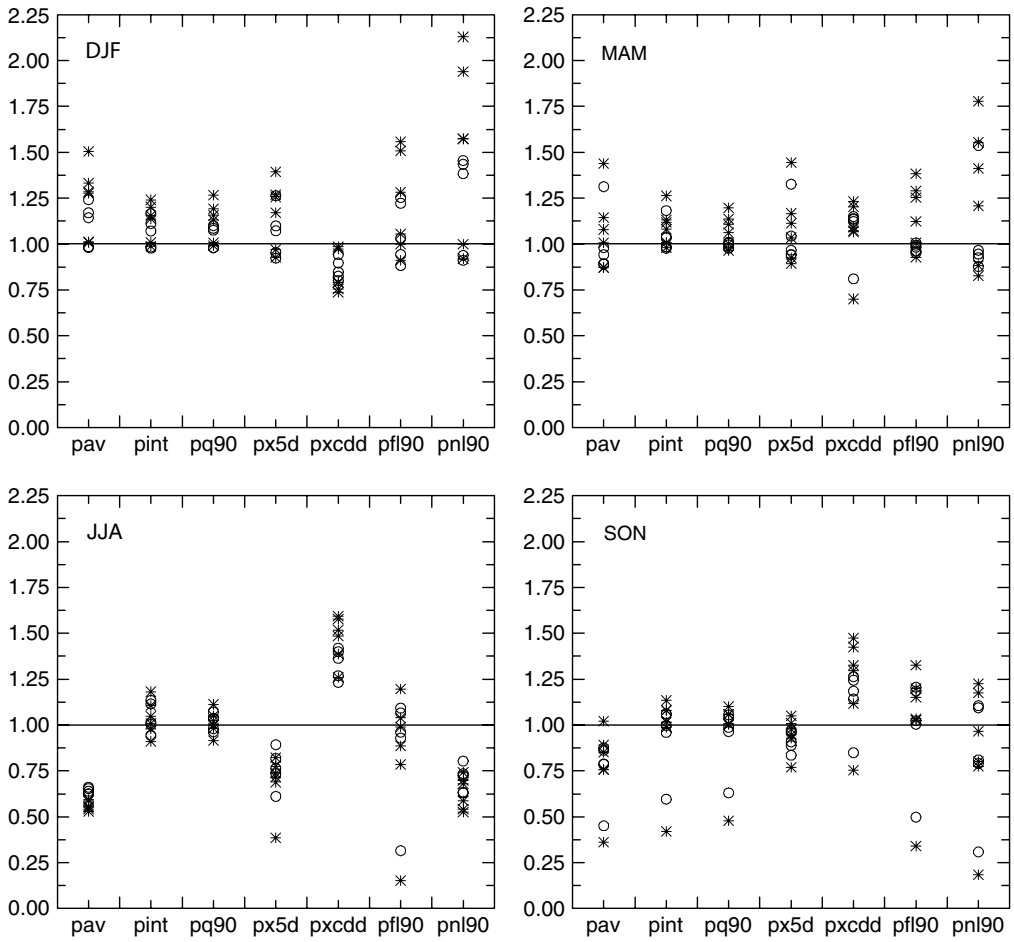


Figure 10. Change in indices using six downscaling models for Oxford under A2 (asterix) and B2 (circles) scenarios

produce a 12-member distribution of modelled changes in each of the indices for each station. Note that this still underestimates the uncertainty as we have used a limited number of downscaling models with only two emission scenarios forced by a limited number of ensembles from a single GCM. The use of only a single GCM is particularly important as it is acknowledged that the GCM is still the largest source of uncertainty in regional climate-change projections (Giorgi *et al*., 2001). Figure 10 shows the changes in the indices for the station Oxford in SEE. For some indices and seasons, the spread is very small (e.g. pav in JJA) but for others it is much larger (e.g. pnl90 in DJF). Importantly, for each index the variability among models is of the same order of magnitude as the variability between the two scenarios. To summarise the changes across all stations, Figure 11 shows the range of changes in each of the indices for all the stations in each region. The previously noted tendency to wetter conditions in DJF is very evident and less so is the tendency to drier conditions in JJA and wetter conditions in MAM. In these three cases, the extremes show similar tendency to the total precipitation.

## 6. CONCLUSIONS

Six statistical and two dynamical downscaling models were compared in their ability to downscale seven seasonal indices of heavy precipitation to the station scale. Models based on non-linear ANNs were found to be the best at modelling the inter-annual variability of the indices; however, their strong negative biases

 

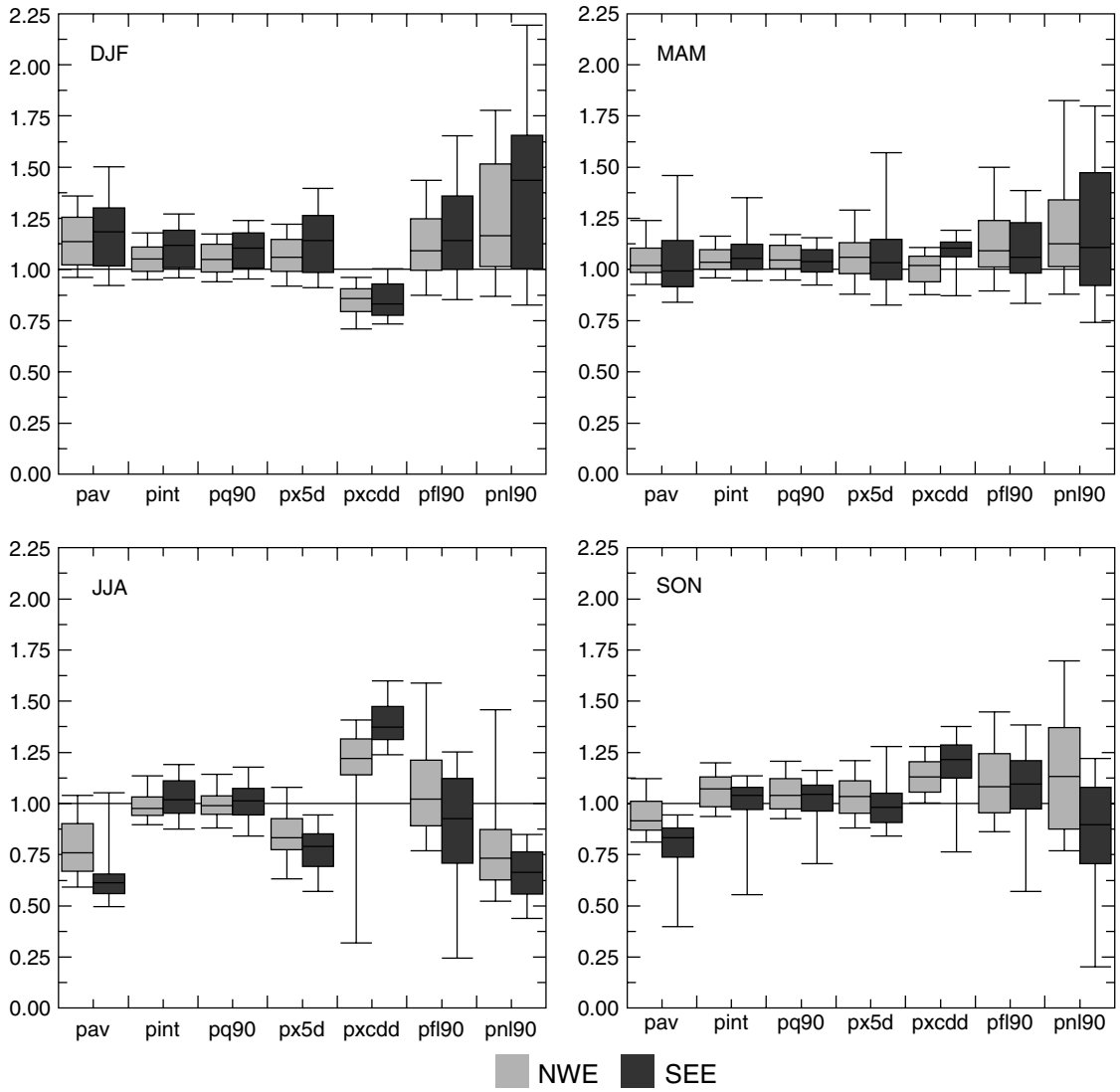


Figure 11. All-model all-scenario change in indices

implied a tendency to underestimate extremes. This was due to the design of these models to reproduce the conditional mean precipitation for each day. A novel approach used in one of the ANN models to output the rainfall probability and the gamma distribution scale and shape parameters for each day meant that resampling methods could be used to circumvent the underestimation of extremes. It also means that Monte Carlo simulations could be used to generate probability distributions for the change in the indices under climate-change conditions, however, this was not explored in this study as only one model generated such probabilistic output. The incorporation of probabilistic output into other downscaling models would be of great benefit to constructing probabilistic scenarios of extremes.

The correlation skill of a simple model that used the daily area-average observed rainfall as the downscaled rainfall was used as a proxy for spatial coherence of the indices. This showed that skill among the eight downscaling models was high for those indices and seasons that had greater spatial coherence. Generally, DJF showed the highest downscaling skill and JJA the lowest. The rainfall indices that were indicative of rainfall occurrence, such as pxcdd and pav, were better modelled than those indicative of intensity.

 

Although four of the statistical models were multi-site models, we have not addressed the skill of the models in maintaining inter-station relationships. We would expect this to be higher in the multi-site models. Incorporating inter-site statistics into the single-site models would be of benefit to users of such downscaled data to ensure consistent scenario time series across stations.

Six of the models were applied to the Hadley Centre GCM HadAM3P forced by emissions according to two SRES scenarios. This revealed that the inter-model differences between the future changes in the downscaled precipitation indices were at least as large as the differences between the emission scenarios for a single model. This implies caution when interpreting the output from a single model or a single type of model (e.g. RCMs) and the advantage of including as many different types of downscaling models, GCMs and emission scenarios as possible when developing climate-change projections at the local scale.

## ACKNOWLEDGEMENTS

This work was funded by the Commission of the European Union under the STARDEX (STAtistical and Regional dynamical Downscaling of EXtremes for European regions) contract (EVK2-CT-2001-00115). Data for HadAM3P GCM was kindly provided by the Hadle Centre. Thanks also to Juerg Schmidli for providing the RCM data from the MERCURE project.

## REFERENCES

Anderberg MR. 1973. *Cluster Analysis for Applications*. Academic Press: New York.

Barnett TP, Preisendorfer R. 1987. Origins and levels of monthly and seasonal forecast skill for United-States surface air temperatures determined by canonical correlation-analysis. *Monthly Weather Review* **115**: 1825–1850.

Bishop CM. 1995. *Neural Networks for Pattern Recognition*. Oxford University Press: Oxford, 508.

Bretherton CS, Smith C, Wallace JM. 1992. An intercomparison of methods for finding coupled patterns in climate data. *Journal of Climate* **5**: 541–560.

Cawley GC, Haylock M, Dorling SR, Goodess C, Jones PD. 2003. Statistical Downscaling With Artificial Neural Networks. In *ESANN-2003, Proceedings of the European Symposium on Artificial Neural Networks*, Bruges, Belgium, 167–172.

Gibson JK, Kållberg P, Uppala S, Hernandez A, Nomura A, Serrano E. 1997. ECMWF re-analysis project report series: 1. ERA-15 description. ECMWF: 74.

Giorgi F, Francisco R. 2000. Evaluating uncertainties in the prediction of regional climate change. *Geophysical Research Letters* **27**: 1295–1298.

Giorgi F, Mearns LO. 1991. Approaches to the simulation of regional climate change–a review. *Reviews of Geophysics* **29**: 191–216.

Giorgi F, Hewitson B, Christensen J, Hulme M, von Storch H, Whetton PH, Jones R, Mearns L, Fu C. 2001. Regional climate information–evaluation and projections. In *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climatic Change*, Houghton JT, Ding Y, Griggs DJ, Noguer M, van der Linden PJ, Dai X, Maskell K, Johnson CA (eds). Cambridge University Press: Cambridge, 881.

Goodess CM, Anagnostopoulou C, B ardossy A, Frei C, Harpham C, Haylock MR, Hundecha Y, Maheras P, Ribalaygua J, Schmidli J, Schmith T, Tolika T, Tomozeiu R, Wilby RL. 2006. An intercomparison of statistical downscaling methods for Europe and European regions – assessing their performance with respect to extreme temperature and precipitation events. *Climatic Change*, Accepted.

Gordon HB, Whetton PH, Pittock AB, Fowler AM, Haylock MR. 1992. Simulated changes in daily rainfall intensity due to the enhanced greenhouse-effect–implications for extreme rainfall events. *Climate Dynamics* **8**: 83–102.

Harpham C, Wilby RL. 2005. Multi-site downscaling of heavy daily precipitation occurrence and amounts. *Journal of Hydrology* **312**(1–4): 235–255.

Haylock M, Goodess C. 2004. Interannual variability of European extreme winter rainfall and links with mean large-scale circulation. *International Journal of Climatology* **24**: 759–776.

Hennessy KJ, Suppiah R, Page CM. 1999. Australian rainfall changes, 1910–1995. *Australian Meteorological Magazine* **48**: 1–13.

Hewitson BC, Crane RG. 1996. Climate downscaling: Techniques and application. *Climate Research* **7**: 85–95.

Hodges KI, Hoskins BJ, Boyle J, Thorncroft C. 2003. A comparison of recent reanalysis datasets using objective feature tracking: Storm tracks and tropical easterly waves. *Monthly Weather Review* **131**: 2012–2037.

Holland J. 1975. *Adaptation in Natural and Artificial Systems*. University of Michigan Press: Ann Arbor, Michigan.

IPCC. 2000. *Emissions Scenarios 2000*. Cambridge University Press: Cambridge, 570.

Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J, Zhu Y, Chelliah M, Ebisuzaki W, Higgins W, Janowiak J, Mo KC, Ropelewski C, Wang J, Leetmaa A, Reynolds R, Jenne R, Joseph D. 1996. The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society* **77**: 437–471.

Kessler E. 1969. On the distribution and continuity of water substance in atmospheric circulation models. *Meteorol. Monogr.* vol 10*, Am. Meteorol. Soc.*, Boston, Mass.

Kidson JW, Thompson CS. 1998. A comparison of statistical and model-based downscaling techniques for estimating local climate variations. *Journal of Climate* **11**: 735–753.

Lin YL, Farley RD, Orville HD. 1983. Bulk parameterization of the snow field in a cloud model. *Journal of Climate and Applied Meteorology* **22**: 1065–1092.

 

Luthi D, Cress A, Davies HC, Frei C, Schar C. 1996. Interannual variability and regional climate simulations. *Theoretical and Applied Climatology* **53**: 185–209.

Machenhauer B, Wildelband M, Botzet M, Jones RG, Déqué M. 1996. Validation of present-day regional climate simulations over Europe: Nested LAM and variable resolution global model simulations with observed or mixed-layer ocean boundary conditions. Max-Planck Institute Report 191. Max-Planck-Institut für Meteorologie: Hamburg, Germany.

Mackay DJC. 1992. Bayesian interpolation. *Neural Computation* **4**: 415–447.

Murphy J. 1999. An evaluation of statistical and dynamical techniques for downscaling local climate. *Journal of Climate* **12**: 2256–2284.

Nicholls N. 1995. Long-term climate monitoring and extreme events. *Climatic Change* **31**: 231–245.

Nicholls N, Murray W. 1999. Workshop on indices and indicators for climate extremes, Asheville, NC, USA, 3–6 June 1997–Breakout group B: Precipitation. *Climatic Change* **42**: 23–29.

Pope VD, Gallani ML, Rowntree PR, Stratton RA. 2000. The impact of new physical parametrizations in the Hadley centre climate model: HadAM3. *Climate Dynamics* **16**: 123–146.

Preisendorfer RW, Zwiers FW, Barnett TP. 1981. Foundations of principal component selection rules. SIO Reference Series 81-4. Scripps Institution of Oceanography: La Jolla, 192.

Press WH, Flannery BP, Teukolsky SA, Vetterling WT. 1986. *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press: Cambridge, 818.

Richardson CW. 1981. Stochastic simulation of daily precipitation, Temperature, and Solar-Radiation. *Water Resources Research* **17**: 182–190.

Schmidli J, Goodess CM, Frei C, Haylock MR, Hundecha Y, Ribalaygua J, Schmith T. 2005. Statistical and Dynamical Downscaling of Precipitation: Evaluation, Intercomparison, and Scenarios for the European Alps, Submitted (2006).

Stern RD, Coe R. 1984. A model-fitting analysis of daily rainfall data. *Journal of the Royal Statistical Society. Series A, (Statistics in Society)* **147**: 1–34.

Tiedtke M. 1989. A comprehensive mass flux scheme for Cumulus parameterization in large-scale models. *Monthly Weather Review* **117**: 1779–1800.

Vidale PL, Luthi D, Frei C, Seneviratne SI, Schar C. 2003. Predictability and uncertainty in a regional climate model. *Journal of Geophysical Research-Atmospheres* **108**(D18): 4586 DOI:10.1029/2002JD002810.

Wilby R. 1995. Simulation of precipitation by weather pattern and frontal analysis. *Journal of Hydrology* **173**: 91–109.

Wilby RL, Dawson CW, Barrow EM. 2002. SDSM–a decision support tool for the assessment of regional climate change impacts. *Environmental Modelling & Software* **17**: 147–159.

Wilby RL, Tomlinson OJ, Dawson CW. 2003. Multi-site simulation of precipitation by conditional resampling. *Climate Research* **23**: 183–194.

Wilby RL, Wigley TML. 1997. Downscaling general circulation model output: a review of methods and limitations. *Progress in Physical Geography* **21**: 530–548.

Wilby RL, Wigley TML, Conway D, Jones PD, Hewitson BC, Main J, Wilks DS. 1998. Statistical downscaling of general circulation model output: A comparison of methods. *Water Resources Research* **34**: 2995–3008.

Wilby RL, Hay LE, Gutowski WJ, Arritt RW, Takle ES, Pan ZT, Leavesley GH, Clark MP. 2000. Hydrological responses to dynamically and statistically downscaled climate model output. *Geophysical Research Letters* **27**: 1199–1202.

Williams PM. 1995. Bayesian regularization and pruning using a Laplace prior. *Neural Computation* **7**: 117–143.

Williams PM. 1998. Modelling seasonality and trends in daily rainfall data. In *Advances in Neural Information Processing Systems–Proceedings of the 1997 Conference*, Jordan MI, Kearns MJ, Solla SA (eds). MIT Press: 985–991.

 