University of East Anglia

An intercomparison of statistical downscaling methods for Europe and European regions – assessing their performance with respect to extreme temperature and precipitation events

C.M. Goodess, C. Anagnostopoulou, A. Bárdossy, C. Frei, C. Harpham, M.R. Haylock, Y. Hundecha, P. Maheras, J. Ribalaygua, J. Schmidli, T. Schmith, K. Tolika, R. Tomozeiu and R.L. Wilby

2005 (published as CRU RP11 in 2012)

Climatic Research Unit
School of Environmental Sciences
University of East Anglia

Climatic Research Unit Research Publication 11

(CRU RP11)

About the Climatic Research Unit

www.cru.uea.ac.uk

The Climatic Research Unit (CRU) is widely recognised as one of the world's leading institutions concerned with the study of natural and anthropogenic climate change. CRU is part of the School of Environmental Sciences at the University of East Anglia in Norwich. The aim of the Climatic Research Unit is to improve scientific understanding in three areas:

- past climate history and its impact on humanity;
- the course and causes of climate change;
- prospects for the future.

About Climatic Research Unit Research Publications

www.cru.uea.ac.uk/publications/

The majority of CRU's research output is published in the peer-reviewed science literature (journals and books), listed in full on our publications webpage. CRU Research Publications (CRU RPs) are occasional research reports that, due to various factors such as length, style or intended audience, are not suitable for publishing in the science literature. The first nine CRU RPs were originally published in hard copy form but are now available in electronic form, while from CRU RP10 onwards they are available only from our website in electronic form.

CRU Research Publications series

www.cru.uea.ac.uk/publications/crurp/

CRU RP1 (1973) Lamb HH
The seasonal progression of the general atmospheric circulation affecting the North Atlantic and Europe.

CRU RP2 (1974) Collected abstracts of the International CLIMAP Conference held in Norwich in May 1973
Mapping the atmospheric and oceanic circulation and other climatic parameters at the time of the Last Glacial Maximum, about 17000 years ago.

CRU RP3 (1974) Lamb HH
The current trend of world climate – a report on the early 1970s and a perspective.

CRU RP4 (1975) Wright PB
An index of the Southern Oscillation.

CRU RP5 (1977) Lamb HH
Understanding climatic change and its relevance to the world food problem.

CRU RP6 (1978) Douglas KS, Lamb HH & Loader C
A meteorological study of July to October 1588: the Spanish Armada storms.

CRU RP6a (1979) Douglas KS & Lamb HH
Weather Observations and a Tentative Meteorological Analysis of the Period May to July 1588.

CRU RP7 (1980) Perry AH & Fearnside T
Northern Hemisphere pentad (5-day) mean sea level pressure values for the period 1951–70 and comparisons with earlier epochs.

CRU RP8 (1985) Jones PD, Ogilvie AEJ & Wigley TML
Riverflow data for the United Kingdom: reconstructed data back to 1844 and historical data back to 1556.

CRU RP9 (1988) Santer BD
Regional validation of General Circulation Models.

CRU RP10 (2004) Osborn TJ, Briffa KR, Schweingruber FH & Jones PD
Annually resolved patterns of summer temperature over the Northern Hemisphere since AD 1400 from a tree-ring-density network.

CRU RP11 (2005) Goodess CM, Anagnostopoulou C, Bárdossy A, Frei C, Harpham C, Haylock MR, Hundecha Y, Maheras P, Ribalaygua J, Schmidli J, Schmith T, Tolika K, Tomozeiu R & Wilby RL
An intercomparison of statistical downscaling methods for Europe and European regions – assessing their performance with respect to extreme temperature and precipitation events.

An intercomparison of statistical downscaling methods for Europe and European regions – assessing their performance with respect to extreme temperature and precipitation events

Goodess[1], C.M., Anagnostopoulou[2], C., Bárdossy[3], A., Frei[4], C., Harpham[5], C., Haylock[1], M.R., Hundecha[3], Y., Maheras[2], P., Ribalaygua[6], J., Schmidli[4], J., Schmith[7], T., Tolika[2], K., Tomozeiu[8], R. and Wilby[9], R.L.

[1] Climatic Research Unit, School of Environmental Sciences, University of East Anglia, Norwich, NR4 7TJ, UK: c.goodess@uea.ac.uk

[2] University of Thessaloniki, Greece; [3] Institut für Wasserbau, University of Stuttgart, Germany; [4]Atmospheric and Climate Science ETH, Switzerland; [5]Kings College London, UK; [6]Fundación para la Investigación del Clima, Spain; [7]Danish Meteorological Institute, Denmark; [8]Servizio Meteorologico Regional, ARPA-Emilia Romagna, Italy; [9]Environment Agency, UK.

Abstract

As part of the European Union-funded STARDEX project, a systematic and rigorous intercomparison was undertaken of 22 statistical downscaling methods, focusing on 10 indices of extreme temperature and precipitation. A case-study approach was taken, encompassing six European regions and Europe as a whole (the latter utilising a new data set of almost 500 daily station time series for the period 1958-2000). Before the STARDEX statistical downscaling methods were applied to output from global climate models, their performance was assessed using reanalysis data, as described here. The use of common data sets, calibration/validation periods

and test statistics provides a rigorous experimental framework for answering such well-defined questions as: is there any systematic difference in performance of the methods between different seasons, indices and regions; or between direct methods in which the seasonal indices of extremes are downscaled and indirect methods in which daily time series are generated and the seasonal indices then calculated from these. The extent to which these questions can be addressed is limited by the variation in skill from method-to-method, index-to-index, season-to-season and station-to-station, with the latter dominating. This variability means that it is not possible to identify a consistently superior method in the majority of cases. Hence a major recommendation is to use a range of the better statistical downscaling methods for the construction of scenarios of extremes, just as it is recommended good practice to use a range of global and regional climate models in order to reflect a wider range of the uncertainties. Thus downscaling uncertainties should always be considered, alongside other uncertainties including choice of GCM and impact model.

1. Introduction

Climate scenarios drive all top-down climate impact assessment studies. The value of such studies is, therefore, limited by the availability of appropriate and reliable climate scenarios. Thus there is a growing demand for scenarios with higher and higher spatial and temporal resolutions for increasingly specialised applications within many different socio-economic sectors, including agriculture, forestry, water resources, energy, transport, tourism and public health. Recent events, such as the August 2002 floods in Central and Eastern Europe and the severe heatwaves experienced across many parts of Europe in August 2003, graphically illustrate the losses of life and high economic damages which can be caused by extreme weather events. According to estimates by Munich Re, for example, the August 2002 floods

were responsible for economic losses of 21.1 billion Euro and insured losses of 3.4 billion Euro, together with over 100 fatalities (Munich Re, 2002). Events such as this also demonstrate the need for scenarios of weather extremes as well as mean climate. At the same time there is a need to quantify and, where possible, reduce the uncertainties associated with climate scenarios (Karl *et al*., 1999; Beersma *et al*., 2000; Cramer *et al*., 2000; Meehl *et al*., 2000).

The mismatch in scales between model resolution and the increasingly small scales required by impact analysts can be overcome by downscaling, i.e., 'sensibly projecting the large-scale information on the regional scale' (von Storch *et al*., 1993). Two major approaches to downscaling, statistical (based on the application of relationships identified in the observed climate, between the large-scale and smaller-scale, to climate model output) and dynamical (using physically-based Regional Climate Models (RCMs)) were developed and tested 5-10 years ago by a number of different research groups, and shown to offer good potential for the construction of high-resolution scenarios (Hewitson and Crane, 1996; Wilby *et al*., 1998; Giorgi and Mearns, 1999; Mearns *et al*., 1999; Murphy, 1999; Zorita and von Storch, 1999; Murphy, 2000). In both cases, however, the focus during this period was on changes in mean climate rather than on daily extremes, leaving considerable scope for further development and refinement of the methodologies. This was the major focus of three European Union funded projects running from 2001/2002 to 2004/2005: MICE, PRUDENCE and STARDEX (Christensen *et al*., 2005a). While PRUDENCE focused on the development and use of RCMs (Christensen *et al*., 2005b) and MICE on the use of GCM and RCM output in impacts studies (Hanson *et al*., 2005), STARDEX focused on the development and assessment of improved statistical

downscaling methods for Europe with emphasis on the ability to construct scenarios of extremes.

The generic advantages and disadvantages of dynamical and statistical downscaling for the construction of scenarios of extremes are summarised in Tables 1 and 2 respectively. For some climate impact study applications, statistical downscaling may provide a more appropriate approach than dynamical downscaling, in particular, where station or point values of extremes are required or when computational resources are limited. Statistical downscaling methodologies can also provide information about the performance of GCMs and RCMs with respect to their ability to reproduce large-scale circulation and other predictor variables together with their relationships with surface climate (the predictands in statistical downscaling). Understanding and quantifying predictor/predictand relationships in observed data sets, an important step in statistical downscaling, can help to identify potential sources of climate model bias and increase confidence in simulated changes of surface climate. Thus many aspects of the work undertaken in the STARDEX project inform and complement work undertaken in PRUDENCE.

More than 20 different statistical downscaling methods (Section 3.1) were developed and evaluated in the STARDEX project using daily temperature and precipitation time series from six European case-study regions (Iberian Peninsula (Western Iberia and Southeast Spain), Greece, the Alps, the German Rhine catchment, UK (Northwest UK and Southeast England) and Northern Italy (Emilia Romagna)) and for Europe as a whole (Section 2.1). In the latter case, a new dataset of almost 500 daily station records was used. For model calibration, predictor variables were constructed from reanalysis data (Section 2.2). The focus was on 10 indices of extremes describing frequency, magnitude and persistence characteristics of extreme

temperature and precipitation events (Section 2.3). The need to work with reasonably large sample sizes means that 'moderate' extremes are considered, e.g., 90th percentile values rather than 95th or 99th.

In order to ensure a fair and consistent intercomparison and evaluation of performance, a standard verification procedure was used, including standard calibration/validation periods and performance statistics (Section 3.2).

The validation analyses presented in Section 4 address well-defined questions including: is there any systematic difference in performance of the methods between different seasons, indices and regions; or between direct methods in which the seasonal indices of extremes are downscaled and indirect methods in which daily time series are generated and the seasonal indices then calculated from these. Finally, in Section 5, the relative confidence in the different statistical downscaling methods is summarised and recommendations made concerning their use for climate scenario construction and further development.

2. The STARDEX Datasets

2.1 Observed station data

A European-wide data set of observed daily maximum and minimum temperature and precipitation for 495 stations for the period 1958-2000 was developed by Fundación para la Investigación del Clima (FIC) for use in the STARDEX project (FIC, 2005). The dataset has good spatial coverage over most of Europe (Figure 1). Station data were provided by the European Climate Assessment (ECA) project (http://eca.knmi.nl) and by national meteorological services from 14 European countries. Quality control analyses of the daily temperature and precipitation time series values were undertaken by FIC in order to identify erroneous,

e.g., negative rainfall or Tmin>Tmax, or spatially incoherent values (FIC, 2005). An iterative homogenisation procedure based on the approach of Moberg and Alexandersson (1997) was also applied to the mean annual maximum and minimum temperature series. Due to the relatively sparse distribution of the stations, detected inhomogeneities were not corrected, but flagged as suspect. In addition to the quality control work undertaken by FIC, considerable work on quality control and homogenisation was previously undertaken by the ECA on all temperature and precipitation series in their data set (Wijngaard *et al*., 2003), including those subsequently incorporated in the STARDEX dataset.

The European-wide data set was complemented by higher-density datasets produced by the relevant STARDEX partner for each of the six case-study regions (Table 3). For each of these regions, a subset of stations drawn from the FIC European-wide data set was also identified and used, for example, to intercompare locally- and European-wide developed downscaling methods (see Section 3.2).

2.2 Reanalysis data

The predictor variables, including sea level pressure (SLP), 500 hPa geopotential height, 1000-500 hPa thickness field and relative/specific humidity and temperature at different pressure levels, used to calibrate and validate the STARDEX statistical downscaling models (Table 4) were derived from the National Centers for Environmental Prediction (NCEP) reanalyses (Kalnay *et al*., 1996). A number of studies have evaluated the reliability of these variables over Europe and the North Atlantic. Reid *et al*. (2001), for example, showed that mean SLP is generally well simulated over the domain of interest to STARDEX, although large differences compared with gridded and station data sets do occur outside this domain, over

Greenland and the Barents Sea. Other studies have compared NCEP upper-air temperatures with satellite observations (Basist and Chelliah, 1997; Shah and Rind, 1998), and shown good agreement. However, a global analysis indicates that, prior to 1979, NCEP upper-air temperatures do not agree with radiosonde and satellite data (Santer *et al*., 1999). Despite these issues, the NCEP reanalysis data set was considered the most appropriate for use in STARDEX, having the major advantages of comparable spatial resolution to the current generation of GCMs and spanning the 40-year period for which suitable daily observed data were available (see Section 2.1). Although a comparably long reanalysis data set (ERA-40) is now available from the European Centre for Medium-range Weather Forecasts, at the start of the STARDEX project, only 15 years of reanalysis data (ERA-15) were available from this source.

Since some NCEP variables are provided on a Gaussian grid, rather than a regular latitude/longitude grid, natural neighbour interpolation, a sophisticated weighted average method (as implemented in the Natgrid software routine, part of the NCAR software library, see http://ngwww.ucar.edu/ngdoc/ng4.3/ngmath/natgrid/intro.html), was used to interpolate all potential predictor variables to a standard 2.5˚ latitude by 2.5˚ longitude grid.

2.3 Indices of extremes

A set of 10 core indices of extremes (six for precipitation and four for temperature) was identified for use in the STARDEX project (Table 5). Many of the indices are based on thresholds defined using percentile values rather than fixed values. This makes them transferable across the range of climatic regimes experienced across Europe. However, such 'fixed-bin' approaches do have some

limitations, e.g., when exploring the contribution of extreme events to overall trends (Michaels *et al*., 2004). In order to ensure reasonable sample sizes and to avoid major difficulties in trend analysis (Frei and Schär, 2001), the focus is on 'moderate' extremes, i.e., 90th and 10th percentile values, rather than the far tails of the distributions. The core set was carefully chosen to encompass magnitude (e.g., Tmax 90th percentile), frequency (e.g., number of days with precipitation exceeding the 90th percentile) and persistence (e.g., longest dry spell length and heat wave duration) of extremes. A software package for calculating these indices (and over 40 more), together with documentation providing a detailed definition of each index, is available from the STARDEX web site: http://www.cru.uea.ac.uk/cru/projects/stardex/. Core indices calculated for the FIC European-wide dataset (Section 2.1) can also be downloaded from this site.

It should be noted that many other definitions of extremes are available. Those used here are, however, highly appropriate for the STARDEX purposes of developing and evaluating statistical downscaling methods for the construction of scenarios of extremes. As well as being rather moderate, they are defined primarily from a climatic perspective rather than an impacts perspective. This is not to say that they are irrelevant for impacts purposes. The greatest 5-day total rainfall, for example, is likely to be relevant to flooding episodes on smaller catchments, although a longer aggregation period may be more appropriate for larger catchments.

3. The STARDEX Experimental Approach

3.1 The STARDEX statistical downscaling methods

A range of the various possible approaches to statistical downscaling (Wilby *et al*., 1998; Zorita and von Storch, 1999) was developed and evaluated by STARDEX partners:

- multiple linear regression (MLR);
- canonical correlation analysis (CCA);
- artificial neural networks (ANN);
- multivariate autoregressive (MAR) modelling;
- conditional re-sampling (CR) and other analogue-based methods;
- methods based on a 'potential precipitation circulation index' (PPCI) and 'critical circulation patterns';
- a conditional weather generator (CWG); and,
- local scaling (LOC, LOCI) and dynamical scaling (DYN, DYNI).

In all, 22 different methods were developed and tested by the 12 STARDEX partners. These are summarised in Table 4, which lists the predictors and predictands used and provides a very brief outline of each method. A standard naming nomenclature is used for these methods: the first component is the STARDEX institution which developed the method, while the second is the technique, followed by, where appropriate, the sub-technique (e.g., KCL_ANN_RBF is the Radial Basis Function Artificial Neural Network technique developed by Kings College London). The methods can be divided into 'indirect' methods, where daily time series are downscaled and indices of extremes (Table 5) calculated, and 'direct' methods where the indices of extremes are employed as predictands. The downscaling methods

themselves are described in detail in STARDEX Deliverable D15, while the methodologies used for selecting potential predictor variables are described in STARDEX Deliverable D10 (both available from http://www.cru.uea.ac.uk/cru/projects/stardex/).

The development of the STARDEX statistical downscaling methods is underpinned by detailed analyses of observed data undertaken in the earlier stages of the project, including analyses of trends in the indices of extremes (see STARDEX Deliverable D9 available from http://www.cru.uea.ac.uk/cru/projects/stardex/ and Schmidli and Frei, 2005) and exploration of relationships between these indices and their trends, and potential predictor variables (Haylock and Goodess, 2004; Maheras *et al*., 2004).

Good predictor variables are defined in STARDEX Deliverable D10 as:

- having strong, robust and physically-meaningful relationships with the predictand;
- having stable and stationary relationships with the predictand;
- explaining low-frequency variability and trends;
- being at an appropriate spatial scale (in terms of both physical processes and GCM performance); and
- well reproduced by GCMs.

A number of different methods were used to select the most appropriate predictor variables for use in each STARDEX study region including: stepwise multiple regression, compositing, correlation analysis, principal components analysis and CCA. These more traditional methods proved more useful than novel methods such as a genetic algorithm (Holland, 1975) approach (see STARDEX Deliverable D10 for further discussion on predictor selection).

The STARDEX statistical downscaling methods (Table 4) range from standard linear regression methods (MLR (Wilks, 1995; Draper and Smith, 1981), used by four STARDEX groups), through methods focusing on spatial patterns (CCA (Barnett and Preisendorfer, 1987; von Storch *et al*., 1993; Gyalistras *et al*., 1994), also used by four STARDEX groups), to non-linear neural network methods (ANN) and other less-widely used approaches including some analogue-based methods. Three STARDEX groups evaluated a number of different ANN approaches (Table 4). Kings College London (KCL), for example, worked with Radial Basis Function (RBF) (Broomhead and Lowe, 1988; Moody and Darken, 1989) ANNs and Multi Layer Perceptron (MLP) models (Rumelhart and McClelland, 1986) in order to simulate multi-site precipitation (Harpham and Wilby, 2004; 2005). Prior to STARDEX, there had been relatively few examples of ANNs used for downscaling and even fewer applications to downscaling multi-site precipitation extremes (McGinnis, 1997; Cavazos, 1999; Crane and Hewitson, 1998; Zorita and von Storch, 1999; Cavazos, 2000; Schoof and Pryor, 2001; Olsson *et al*., 2001; Trigo and Palutikof, 2001). Thus the STARDEX study provides the most systematic evaluation of ANN methods for the latter purpose to date.

The first ANN approach evaluated in STARDEX is MLP models. These models have a single hidden layer, and the non-linear transformation of the linear sums is catered for by the activation functions in the middle and output layer (Harpham and Wilby, 2004; 2005). The goal in training an MLP model is to find the values of the weight vectors that minimise the error between that network output and the desired target value. The most common method for training an MLP network is the error back-propagation algorithm (Rumelhart and McClelland, 1986) and this is

the approach used in STARDEX by KCL (Harpham and Wilby, 2004; 2005) and the Aristotle University of Thessaloniki (AUTH).

The RBF network developed by KCL consists of two layers (Harpham and Wigley, 2004; 2005). In the first layer, the basis functions, which provide the non-linear behaviour (Bishop, 1995), are determined by unsupervised training using *K*-means clustering of the predictor input vector alone. Singular Value Decomposition is then used to estimate the second layer weight vectors. This distinction between the first and second layer weights is considered to be a particular advantage of the RBF approach since suitable parameters can be chosen for the hidden nodes without having to perform a full non-linear optimisation of the network (Harpham and Wilby, 2005).

For the first study region (see Section 3.2) the MLP and RBF results from KCL were very similar, so for the second study region KCL substituted the MLP with a Genetic Algorithm (GA) RBF to provide further comparison. The RBF networks training algorithm determines the network structure either by *a priori* or *a posteriori* knowledge. Networks configured in this manner are not guaranteed to have an optimal structure and may result in under-fitting or over-fitting of the training data resulting in poor generalisation ability. Only by adopting a trial-and-error approach can this be corrected. The GA-RBF addresses the *ad hoc* approach applied to configuring the standard RBF network structure by using a genetic algorithm to optimise the network structure/parameters. The GA-RBF configuration used in STARDEX (Harpham, 2004), together with optimising the basis centres (including the number of basis centres) and their associated width, introduces the optimisation of the basis function type at a particular node.

A Bayesian approach to MLP modelling was taken by the University of East Anglia (UEA) (Cawley *et al*., 2003) in order to avoid over-fitting the training data

(Buntine and Weigend, 1991; Mackay, 1992a,b). In addition to using the usual sum-of-squares (SSE) error metric in the training process, UEA also tested a Bernoulli/Gamma misfit term (Williams, 1998), which does not make the implicit assumption of a Gaussian noise process. Reflecting the Bayesian approach, outputs of a committee of 20 networks were combined to create the daily series, using Monte-Carlo (MC) simulation in one version of the model (Table 4).

The ability to model multi-site time series while retaining the spatial correlation of the observed series is one potential advantage of the ANN approach to statistical downscaling (Harpham and Wilby, 2004; 2005). This is also an advantage of the multivariate autoregressive (MAR) model developed by the University of Stuttgart (USTUTT) (Table 4). This approach is based on a modified version of the space-time model described by Bárdossy and Plate (1992). The important development within the STARDEX project is that the spatial covariance structure of the variable concerned (temperature or precipitation) is taken into account and maintained in downscaling. The model can also be used to generate areal values of precipitation and temperature on grids using the spatial structure from observation locations, which can then be used in hydrological impacts studies. The model parameters are conditioned on circulation patterns defined using a fuzzy rule-based classification scheme (Bárdossy *et al.*, 1995, 2002) and moisture flux is also taken into account in downscaling precipitation.

Another group of STARDEX approaches (Table 4) is based on analogue or re-sampling approaches (Palutikof *et al.*, 2002). The ADGB_HYPER4 method, for example, is a novel approach based on pre-selection of 'potentially extreme days' and then random re-sampling in the four-dimensional hyper-space of the predictor variable principal components. Only days with extreme precipitation are downscaled,

not full daily time series. A two-step analogue approach is also taken by FIC in the FIC_ANAL2 method (Table 4). In this case, a set of analogues ('n' most similar days) is selected in the first step from a reference dataset based on similarity of the geostrophic fluxes at 1000 and 500 hPa. In the second step, precipitation and temperature are obtained by searching in the 'n' days population for relationships between the predictands and additional predictors. For precipitation, the six most similar days are averaged. For temperature, MLR is performed using only the 'n' days population with forward and backward stepwise selection of predictors. The additional potential predictors used here are low tropospheric thickness, average temperature of the preceding days and a sinusoid function of the day of the year.

A combination of regression and analogue selection is also used in the Centre National de la Recherche Scientifique (CNRS) downscaling method. Here, however, regression is used in the first step to construct a 'Potential Precipitation Circulation Index' (PPCI), i.e., a linear regression of daily precipitation against the anomaly pattern correlation of a day's large-scale circulation (the 700 hPa geopotential height field) with the precipitation regime clusters identified by Plaut *et al*. (2001). The PPCI values are divided into 20 bins, and then a day randomly selected from the appropriate bin. A broadly similar approach is used in the KCL conditional re-sampling (CR) method (Wilby *et al*., 2003). In this case, the SDSM (Statistical DownScaling Model) single site conditional weather generator (Wilby *et al*., 2002; 2003) is used to downscale an area-average precipitation series (referred to as the marker site). Wet-day amounts are then re-sampled from the empirical distribution of area averages conditional on the large-scale atmospheric forcing and a stochastic error term. The actual wet-day amount is determined by mapping the modelled normal cumulative distribution value onto the observed cumulative distribution at the marker

site. The marker site is then cross-referenced to the actual amount at each station within the region.

A conditional weather generator (CWG) (Goodess and Palutikof, 1998) was developed by the Danish Meteorological Institute (DMI). First, a surface pressure pattern is obtained as the average pressure difference between wet and dry days observed at a given station. A circulation index is then calculated by regressing the daily surface pressure field on this pattern. The circulation index is divided into a number of quantiles, usually between 5 and 10, and for each quantile the following precipitation quantities are calculated: the probability for wet/dry days, the probabilities for a wet/dry day following a dry/wet day, and the two gamma distribution parameters for precipitation amount. A two-state Markov Chain process combined with random sampling from the gamma distribution (Wilks and Wilby, 1999) is then used to generate the daily precipitation series.

Two variants of local scaling (LOC and LOCI) and of dynamical scaling (DYN and DYNI) were developed by the Eidgenössische Technische Hochschule (ETH). The common idea of all four methods is to use GCM-simulated precipitation as a predictor for regional/local precipitation. These STARDEX methods are modifications of the procedures of Widmann and Bretherton (2003). Thus the GCM-simulated precipitation is rescaled with a spatially varying factor which compensates for long-term biases of the GCM at the station being downscaled. While the original schemes (LOC and DYN) aim only at correcting the bias in mean precipitation, the STARDEX modifications (LOCI and DYNI) include a bias correction for precipitation frequency and intensity. The two local scaling methods use GCM precipitation as their only predictor, while the two dynamical scaling methods include

the 1000 hPa geopotential field as an additional predictor which results in flow-dependant scaling factors.

3.2 Principles for verification and intercomparison of the STARDEX statistical downscaling methods

In order to ensure a rigorous and systematic approach to validation and intercomparison of the STARDEX statistical downscaling methods, a set of principles for this work was agreed. All partners used data from common predictand (Section 2.1) and predictor (Section 2.2) datasets. The core set of 10 indices of extremes was used, together with mean daily precipitation and mean maximum and minimum temperature – giving 13 indices in total (Table 5). A common verification or independent validation period was also chosen: 1979-1993, for compatibility with the 'perfect-boundary condition', i.e., ERA-15 forced, RCM simulations undertaken in the MERCURE project (Machenhauer *et al*., 1998). The remaining period of data, 1958-1978 and 1994-2000 was used for model calibration or training. A common set of verification statistics for the comparison of observed and downscaled annual series of seasonal indices was identified: Root Mean Square Error (RMSE) between the observed and simulated index with the bias removed; Spearman rank-correlation coefficient (CORR); and BIAS (the mean difference between the simulated and observed indices). In all cases reported here, these statistics were calculated for the 1979-1993 verification period.

The STARDEX statistical downscaling methods (Table 4) use a range of approaches suitable for different applications. Some provide multi-site information, for example. The latter methods are more applicable to denser regional station

networks than European wide. Thus it was decided that it would be inappropriate to use a single case-study for verifying and intercomparing all the methods. Instead, the matrix shown in Table 6 was devised. Three groups (UEA, FIC and DMI) applied their methods European-wide, i.e., to the FIC dataset of 495 stations (Figure 1). The other nine groups undertook initial development of their method(s) in the region with which they were most familiar, e.g., ETH from Zurich, Switzerland initially developed their scaling methods for the Alps. In addition to this 'primary' region, eight of the nine groups also applied their method(s) in a 'secondary' region with a contrasting climatic regime, e.g., the UK in the case of ETH. Each group developed their method(s) using the full set of stations available for their primary case-study region and, in most cases, the subset of stations from the FIC dataset for their secondary region (see Tables 3 and 6). However, the intercomparisons described in Section 4 are based on the regional subsets of stations (see Table 3) in all but one case.

This case-study approach allows a number of different intercomparisons to be undertaken by sampling the matrix shown in Table 6 either horizontally or vertically. Downscaling methods from six different groups can be directly compared in the case of Alpine precipitation, for example. Servizio Meteorologico Regional, ARPA_Emilia Romagna (ARPA) and AUTH both applied regression and CCA-based downscaling methods in the Greek and Northern Italy case-study regions – allowing another interesting set of comparisons to be made. Many other such intercomparisons were undertaken.

Having set up this rigorous experimental framework, suitable approaches for handling the many combinations of different methods (22 – see Table 4), regions

(seven – see Table 3), indices (13 – see Table 5) and seasons (four) had to be devised. These approaches and the major results are described in the next section.

4. Results of the STARDEX Intercomparison Exercise

The experimental matrix shown in Table 6 allows a number of specific questions to be addressed:

- Is there any systematic difference in performance of the methods between different seasons?
- Is there any systematic difference in performance of the methods between different indices?
- Is there any systematic difference in performance of the methods between different regions?
- Do direct methods in which the seasonal indices of extremes are downscaled perform better than indirect methods in which daily time series are generated and the seasonal indices then calculated from these?
- Do the regionally-developed methods perform better than the European-wide methods?
- Can a single 'best' method be identified?

These questions were addressed by undertaking a series of regional analyses which are reported in detail in STARDEX Deliverable D12 (available from http://www.cru.uea.ac.uk/projects/stardex) with the major conclusions summarised in the following sections. A preliminary inspection of the downscaled results for all

regions indicated that the variation in skill from station-to-station dominates the variations in skill from index-to-index, from method-to-method and from season-to-season. This is clearly demonstrated in Figure 2, which shows box-and-whisker plots of the Spearman correlation skill for four different downscaling methods applied to Alpine precipitation. Thus in order to address the questions above, results were averaged across neighbouring stations, as well as across the different indices, seasons and methods, as appropriate for addressing each question. While this was a pragmatic approach, designed to draw general conclusions from a large amount of downscaled data, it does not preclude more detailed analyses using different averaging methods or individual results (see Section 5).

4.1 Is there any systematic difference in performance between the different seasons?

In the majority of cases, performance is best in winter and worst in summer, particularly with respect to precipitation. This is illustrated by Figures 3 and 4 which show Spearman correlations averaged across stations and precipitation indices for the UK (14 downscaling methods) and the Iberian Peninsula (six methods) respectively. Exceptions to this pattern can, however, be identified – the DMI_CWG method, for example, performs worst for autumn in the Iberian Peninsula (Figure 4). All the other methods applied to the Iberian Peninsula perform better in autumn than spring (but see Section 4.3). For the UK, the difference in performance between spring and autumn is less clear, although in a number of cases, it is better for autumn than spring. For Greece, however, correlations are lower in autumn than all other seasons. This is demonstrated in Figure 5, which indicates consistently lower correlations (including a

few negative correlations) in autumn for precipitation indices averaged across four stations from western Greece. These lower correlations could be attributed to the higher variability of the geopotential heights during autumn compared with the other seasons (Maheras *et al*., 2002).

4.2 Is there any systematic difference in performance of the methods between different indices?

As might be expected, performance is better for temperature than precipitation, and, particularly for precipitation, also tends to be better for the mean indices (pav, txav and tnav – see Table 5) than the indices of extremes. In the German Rhine study region, for example, correlations averaged across 10 stations and five methods are greater than +0.7 for temperature in most cases (with the exception of heat wave duration) (Figure 6), but only reach a maximum of about +0.65 for precipitation (Figure 7). Figure 6 indicates that performance for all of the temperature indices of extremes is somewhat poorer than for average maximum and minimum temperature. The models, by virtue of their calibration processes, still tend to gravitate towards the central tendency of the training data sets. In the case, of precipitation, however, performance is considerably poorer for the indices of extremes than average precipitation (Figure 7).

With respect to precipitation, the maximum number of dry days (pxcdd), a measure of persistence, seems to be better simulated than the indices of extremes which focus more on the magnitude of events. This is clearly the case for the German Rhine in all seasons (Figure 7) and for western Greece in winter (Figure 5) and the Alps in winter and autumn (Figure 2). These three figures also indicate that the

number of precipitation events greater than the 90th percentile (pnl90) tends to be better simulated than the percentage of precipitation from such events (pfl90) and the 90th percentile value itself (pq90). Thus in general, the occurrence process seems to be better captured than the amounts process. This is likely to be because the downscaling schemes always include some predictors based on large-scale circulation which are likely the capture patterns prohibiting the occurrence of precipitation (hence the better results for pxcdd) whereas the magnitude of intense precipitation is more likely to depend on much smaller-scale mechanisms (hence the poorer results for pq90).

4.3 Is there any systematic difference in performance of the methods between different regions?

Comparison of Figures 3 and 4 indicates that a similar range of downscaling methods tends to give somewhat higher correlations for the UK case-study region than for the Iberian Peninsula. Within these two regions (sub-regional results not shown), performance with respect to precipitation tends to be worse for the SE Spain sub-region, one of the driest regions in Europe and subject to more localised and Mediterranean influences (Goodess and Palutikof, 1998) than the Western Iberian region, which is more affected by larger-scale Atlantic influences (Goodess and Jones, 2002). Given these differences, further work is recommended on verifying these two sub-regions separately. The finding that autumn precipitation is better simulated than spring precipitation (Section 4.1) is, for example, not necessarily applicable to the SE Spain sub-region where autumn precipitation events are mainly due to convection, although also related to easterly air masses tracking over a warm

Mediterranean Sea (Goodess and Jones, 2002). Sub-regional contrasts in performance are also evident for Greece, particularly with respect to precipitation – except in spring, performance is generally better for the four western stations (Figure 5) than for the four eastern stations (not shown).

Regional differences in performance are most clearly identified by looking at results for the three methods that were applied to the European-wide dataset (Table 6). The best performing of these methods is the FIC_ANAL2 method (see Sections 4.5 and 4.6). For temperature, performance of this method is consistent across the European domain, with correlations generally exceeding +0.8 in the case of average minimum temperature (Figure 8). In winter, however, a few Scandinavian stations have large negative correlations, and the correlations fall to +0.6 to +0.8 at a few locations in summer. For precipitation, performance is poorer and less consistent across Europe, as illustrated by the results for average precipitation shown in Figure 9. In winter, correlations tend to be higher on west-coast locations and drop off towards the eastern edges of the domain. In summer, the lowest correlations (-0.2 to +0.2 for average precipitation) occur over southern Spain and in an area to the east of the Alps.

4.4 Do direct methods perform better than indirect methods?

It is difficult to make any general statement on the performance difference between methods that directly downscale seasonal indices and those that downscale daily series of precipitation. While indirect methods appear to do better than direct methods for UK and Iberian Peninsula precipitation (Figures 3 and 4), for example,

this is just as likely to be due to the superior performance of ANN-based methods (Section 4.6) and/or the use of different predictor variables.

A more direct comparison is possible between the indirect USTUTT_MAR and direct USTUTT_MLR methods in the German Rhine. The seasonal skill scores (RMSE and CORR) of these two methods are shown in Table 7 for the precipitation indices, averaged across the full regional set of 100 stations. These results indicate that the direct MLR method performs slightly better than the indirect MAR method in summer, and *vice versa*, in spring and autumn. In winter, it is not possible to identify which of the two methods performs better.

Comparisons between direct (ARPA_CCA and ARPA_MLR) and indirect (AUTH_CCA and AUTH_MREG) methods can also be made for Greece (Figure 5) and Northern Italy (Figure 10). The indirect CCA and regression methods seem to perform slightly better than the direct methods in the case of Greek precipitation (Figure 5). Similarly, the indirect CCA approach appears consistently slightly better for Northern Italian temperature, while the differences between the direct and indirect regression approaches are less consistent (Figure 10).

4.5 Do the regionally-developed methods perform better than the European-wide methods?

The most direct comparison possible here, is between regionally-developed CCA methods (ARPA_CCA, AUTH_CCA and, for the UK only, UEA_CCA) and their European-wide applications (UEA_CCA for Europe as a whole and the Iberian

Peninsula). Figure 5 indicates that the regionally-based CCA methods are superior with respect to western Greek precipitation.

However, for other regions and methods, there is no clear indication that the locally-developed methods perform consistently better than the methods developed based on the European-wide data set. For German Rhine precipitation, the performance of the European-wide FIC_ANAL2 method is comparable with that of the locally developed USTUTT_MAR and USTUTT_MLR methods in winter and autumn, and even slightly better for some indices (Figure 11). In spring, all three methods show more or less similar performance, while FIC_ANAL2 tends to perform better for most indices in summer. For the majority of German Rhine precipitation indices and seasons (Figure 11), the FIC_ANAL2 method is clearly superior to the other two European-wide methods (UEA_CCA and DMI_CWG). Of the three methods applied to German Rhine temperature (UEA_CCA, FIC_ANAL2 and USTUTT_MLR), the FIC_ANAL2 method is consistently better (results not shown). For Greek temperature (results not shown), performance of the FIC_ANAL2 method is comparable to that of the locally-developed MLR and CCA methods. The FIC_ANAL2 method also performs reasonably well in comparison with locally-developed methods for Alpine precipitation (Figure 2) and temperature (results not shown).

4.6 Can a single 'best' method be identified?

The discussion in Section 4.5 indicates that the performance of the FIC_ANAL2 method is generally good with respect to temperature (Figure 8). Only one other method was applied to European-wide temperature (UEA_CCA4) and this

latter method (Figure 12) is clearly less skilful than FIC_ANAL2. FIC_ANAL2 also performs well for temperature extremes (such as the 10th percentile of minimum temperature, Figure 13). Although it performs better than UEA_CCA and DMI_CWG with respect to precipitation across Europe (Figure 9), the skill, particularly for precipitation extremes (such as the greatest 5-day rainfall total, Figure 14) is substantially less than for temperature.

The variation in performance between seasons (Figures 3, 4 and 7), stations (Figure 2) and indices (Figures 2, 5 and 11) makes it particularly difficult to identify a 'best' or 'better' method for downscaling precipitation. In fact, it is impossible to identify a consistently 'best' method. If the correlations for Iberian precipitation shown in Figure 4 are averaged across seasons (Figures 15a), it can be concluded that the ANN (KCL_ANN_RBF and KCL_ANN_GA_RBF) methods are superior to the other methods applied in this region. However, a different conclusion emerges if the biases are considered (Figure 15b). Since each of the indices has different units, biases cannot be averaged across indices. Therefore, each bias was converted to a rank compared to other models for each index, season and station. The ranks were then averaged across all indices, seasons and stations. Methods with lower biases therefore have a higher rank. It is evident from Figure 15 that methods with the highest correlations tend to have higher biases and methods with the lowest correlations (DMI_CWG, UEA_CCA4 and KCL_CR) have lower biases. A similar pattern is also evident in precipitation results for the UK (not shown) and in a number of other cases. For example, pxcdd in Greece tends to have higher correlations than other indices of extremes (Figure 5), but also higher biases (not shown). Similarly, for Alpine temperature (results not shown), the FIC_ANAL2 method tends to give higher correlations and lower RMSE than the University of Berne (UNIBE) CCA

method, but the latter gives the lowest biases. The question as to whether correlations or biases are a more important test statistic when trying to identify the better performing downscaling methods is discussed in Section 5.

5. Conclusions and Recommendations on the Use of Statistical Downscaling Methods

The STARDEX intercomparison of statistical downscaling methods for Europe and European regions focused on performance with respect to extreme temperature and precipitation events and benefited from being undertaken in a rigorous experimental framework. Thus the NCEP Reanalysis-based verification analyses described here, were conducted using a common set of principles (Section 3.2). The STARDEX experimental matrix (Table 6) allows a number of well-defined questions to be addressed in Section 4. However, the extent to which these questions can be answered unequivocally is limited by the variation in skill from method-to-method, index-to-index, season-to-season and station-to-station. The latter variation in skill is found to dominate (Figures 2 and 15). Neighbouring stations can display very different behaviour. However, for the same station, contrasts in performance between seasons or indices can be as high or higher than inter-station contrasts.

The variability in skill tends not to be systematic, hence it is difficult or impossible to identify a single best method in most cases. Attempts to do this are further complicated by the finding that methods/indices with the highest correlations are often not those with the lowest biases or RMSE (Section 4.6). More emphasis is given in this paper to correlation results than to RMSE or bias. This is because systematic differences in performance tend to be easier to identify with respect to

correlations than the other statistics, and also because it may be easier to correct for biases than other types of error, and the ability to capture inter-annual variability is considered to be an important characteristic of statistical models which are to be applied in the climate change context (Wilby *et al.*, 2004). Under prediction of low-frequency climate variability is, however, a characteristic of many statistical downscaling methods (Wilby and Wigley, 2000; Wilby *et al*., 2004), including the STARDEX methods (as demonstrated for Alpine precipitation in Figure 16), and is often referred to as the 'overdispersion' problem (Katz and Parlange, 1998).

Despite the difficulties discussed above, a number of conclusions concerning the relative performance of the STARDEX statistical downscaling methods can be drawn from the analyses presented in Section 4, including:

- performance is generally better for temperature than precipitation, better for means than extremes, and best in winter and worst in summer;
- however, there are always exceptions to the rules, for example, in Greece, the poorest precipitation results are for autumn;
- the FIC_ANAL2 European-wide method performs well for temperature, as well as or better than locally-developed methods;
- CCA methods seem to perform better when applied locally rather than European wide;
- the performance of ANN methods is generally quite good, particularly with respect to precipitation correlations (e.g., for the Iberian Peninsula) which reflect inter-annual skill;

- it is particularly difficult to make statements about whether ‘direct’ or ‘indirect’ methods are consistently better for downscaling indices of extremes; and,
- for precipitation extremes, persistence, notably the length of the longest dry spell, is better represented than magnitude or frequency characteristics.

Since it is not possible to identify a consistently superior method in the majority of cases, a major recommendation from the STARDEX verification studies is to use a range of the better statistical downscaling methods for the construction of scenarios of extremes, just as it is recommended good practice to use a range of global and regional climate models in order to reflect a wider range of the uncertainties (Mearns *et al*., 2003). This implies a need for new easy-to-use statistical downscaling tools based on the STARDEX improved methodologies.

The variability in performance also underlines the importance of undertaking rigorous verification studies for each statistical downscaling application. While verification must always be appropriate to the proposed application, the finding that statistical downscaling techniques struggle to reproduce even moderate extremes in many cases, indicates that testing against more severe events would not make much sense.

One issue that has not been addressed here, is how to judge whether performance is sufficiently good to proceed to scenario construction. For many regions and indices of extremes considered here, it is arguable that the skill is unacceptably low for summer precipitation and that scenarios should not be constructed in these cases. This may reflect a simple lack of predictability of extreme event behaviour at local scales. Although no significance testing of the skill scores

presented here has been undertaken it is likely that many of the downscaling models show statistically insignificant skill for many of the indices, especially in summer and for some problematic regions. Thus the analyses presented here could usefully be extended by excluding summer results when averaging across seasons and only comparing methods where the threshold for significance has been achieved. Similarly, given the poorer performance of the downscaling results for SE Spain compared with Western Iberia, it would be preferable to analyse these sub-regions separately.

It is worth noting that the skill measures used here will all penalise methods with a stochastic component unless the ensemble range of such methods is employed in the comparison. This is highlighted in a more detailed intercomparison of the KCL precipitation results for the UK (Harpham and Wilby, 2005). It is shown that the KCL_ANN methods out-perform the KCL_CR method for the STARDEX indices (because the former methods are deterministic and the latter is stochastic), but this is reversed when it comes to comparing quantiles which reflect the distribution of daily event magnitudes.

The Reanalysis-based verification studies described here reflect the complexity of undertaking rigorous systematic methodological intercomparisons. Many more verification analyses could be undertaken using the large volumes of data generated by STARDEX (much of which will be publicly available from October 2005). However, these studies represent only one step in the process of constructing reliable and robust climate change scenarios using statistical downscaling (Wilby *et al*., 2004). As part of the STARDEX project, considerable work was also done on predictor selection (including assessment of the sensitivity of methods to the choice of predictors and their spatial domains) before finalising the improved statistical

downscaling methods listed in Table 4. Having verified the downscaling methods for the present day using Reanalysis data, the next step is to determine whether the predictor variables are sufficiently well reproduced by GCMs (see STARDEX Deliverable D13 available from http://www.cru.uea.ac.uk/cru/projects/stardex/). Only then, is it reasonable to apply the statistical downscaling models using output from control and perturbed GCM simulations. One of the final steps of the STARDEX work is to assess the statistically downscaled changes in European temperature and precipitation extremes focusing on issues such as their consistency and robustness, including their consistency with GCM and RCM simulated changes, focusing on those climate models used within the PRUDENCE project (Christensen *et al*., 2005b). The significance of the downscaling uncertainty relative to the other uncertainties, including the choice of GCM, must also be addressed.

Acknowledgments

This work was supported by the European Commission as part of the STARDEX (STAtistical and Regional dynamical Downscaling of EXtremes for European regions) contract EVK2-CT-2001-00115. The following organisations are thanked for the provision of daily climate data used within the STARDEX project: the European Climate Assessment project (http://eca.knmi.nl/); Zentralanstalt für Meteorologie und Geodynamik, Austria; Ceský Hydrometeorologický ústav, Czech Republic; Danmarks Meteorologiske Institut, Denmark; Ilmatieteen Laitos, Finland; Meteo France, France; Deutscher Wetterdienst, Germany; Országos Meteorológiai Szolgálat, Hungary; Koninklijk Nederlands Meteorologisch Instituut, Netherlands; Meteorologisk Institutt, Norway; Instituto de Meteorologia, Portugal; Russian

Meteorological and Hydrological Institute, Russia; Instituto Nacional de Meteorología, Spain; Sveriges Meteorologiska Och Hydrologiska Institute, Sweden; and, MeteoSchweiz, Switzerland. The following STARDEX participants are thanked for their contributions to the work described here: Stefano Alberghi, ADGB; Lucia Benito, FIC; Carlo Cacciamani, ARPA-SMR; Hans Caspary, FTS; Gavin Cawley, UEA; Bo Christiansen, DMI; Dimitrios Gyalistras, UNIBE; Phil Jones, UEA; Mike Lincoln, UEA; Antonella Morgillo, ARPA-SMR; Valentina Pavan, ARPA-SMR; Guy Plaut, CNRS; Eric Simonnet, CNRS; Luis Torres, FIC; and, Ennio Tosi, ADGB.

References

Bárdossy, A., Duckstein, L. and Bogardi, I., 1995: 'Fuzzy rule-based classification of atmospheric circulation patterns', *International Journal of Climatology*, **15**, 1087-1097.

Bárdossy, A. and Plate, E.J., 1992: 'Space-time model for daily rainfall using atmospheric circulation patterns', *Water Resources Research,* **28**, 1247-1259.

Bárdossy, A., Stehlík, J. and Caspary, H.-J., 2002: 'Automated objective classification of daily circulation patterns for precipitation and temperature downscaling based on optimized fuzzy rules', *Climate Research,* **23**, 11-22.

Barnett, T. P. and Preisendorfer, R., 1987: 'Origin 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.

Basist, A.N. and Chelliah, M., 1997: 'Comparison of tropospheric temperatures derived from the NCEP/NCAR reanalysis, NCEP operational analysis, and the microwave sounding unit', *Bulletin of the American Meteorological Society*, **78**, 1431-1447.

Beersma, J., Agnew, M.D., Viner, D. and Hulme, M., 2000: *Climate Scenarios for Water-Related and Coastal Impacts,* Proceedings of the EU Concerted Action Initiative ECLAT-2 Workshop 3, KNMI, Netherlands, May 10-12th 2000, Climatic Research Unit, Norwich, UK, 140pp.

Bishop, C.M., 1995: *Neural Networks for Pattern Recognition*, Oxford: Clarendon Press.

Broomhead, D.S. and Lowe, D., 1988: 'Multivariable function interpolation and adaptive networks', *Complex Systems*, **2**, 321-355.

Buntine, W.L. and Weigend, A.S., 1991: 'Bayesian back-propagation', *Complex Systems*, **5**, 603-643.

Cavazos, T., 1999: 'Large scale circulation anomalies conducive to extreme precipitation events and derivation of daily rainfall in northeastern Mexico and southeastern Texas', *Journal of Climate,* **12**, 1506-1523.

Cavazos, T., 2000: 'Using self-organizing maps to investigate extreme climate events: An application to wintertime precipitation in the Balkans', *Journal of Climate*, **13**, 1718-1732.

Cawley, G.C., Haylock, M., Dorling, S.R., Goodess, C. and Jones, P.D., 2003: 'Statistical downscaling with artificial neural networks', *Proceedings of the European Symposium on Artificial Neural Networks* (ESANN-2003), pp.167-172.

Christensen, J.H. et al., 2005a: 'Editorial', *Climatic* Change, this volume.

Christensen, J.H. et al., 2005b: 'Evaluating the performance and utility of regional climate models in climate change research: Reducing uncertainties in climate change projections – the PRUDENCE approach', *Climatic Change*, this volume.

Cramer, W., Doherty, R., Hulme, M. and Viner, D., 2000: *Climate Scenarios for Agricultural and Ecosystem Impacts,* Proceedings of the EU Concerted Action

Initiative ECLAT-2 Workshop 2, Potsdam, Germany October 13th - 15th, 1999, Climatic Research Unit, Norwich, UK, 120pp.

Crane, R.G. and Hewitson, B.C., 1998: 'Doubled CO_2 precipitation changes for the Susquehanna Basin: down-scaling from the Genesis general circulation model', *International Journal of Climatology*, **18**, 65-76.

Draper N.R. and H. Smith, 1981: *Applied Regression Analysis*, Wiley, New York.

FIC (Fundación para la Investigación del Clima), 2005: *The STARDEX European-wide Daily Dataset*, STARDEX Technical Note, available from http://www.cru.uea.ac.uk/cru/projects/stardex/.

Frei, C. and Schär, C., 1998: 'A precipitation climatology of the Alps from high-resolution rain-gauge observations', *International Journal of Climatology*, **18**, 873-890.

Frei, C. and Schär, C., 2001: 'Detection probability of trends in rare events: Theory and application to heavy precipitation in the Alpine region', *Journal of Climate*, **14,** 1564-1584.

Giorgi, F. and Mearns, L.O., 1999: 'Introduction to special section: Regional climate modeling revisited', *Journal of Geophysical Research*, **104**, 6335-6352.

Goodess, C.M. and Palutikof, J.P., 1998: 'Development of daily rainfall scenarios for southeast Spain using a circulation-type approach to downscaling', *International Journal of Climatology*, **18**, 1051-1083.

Goodess, C.M. and Jones, P.D., 2002: 'Links between circulation and changes in the characteristics of Iberian rainfall', *International Journal of Climatology*, **22**, 1593-1615.

Goodess, C.M., Hanson, C., Hulme, M. and Osborn, T.J., 2003: 'Representing climate and extreme weather events in integrated assessment models: A review of existing methods and options for development', *Integrated Assessment*, **4**, 145-171.

Gyalistras, D., von Storch, H., Fischlin, A. and Beniston, M., 1994: 'Linking GCM-simulated climatic changes to ecosystem models: case studies of statistical downscaling in the Alps', *Climate Research*, **4**, 167-189.

Hanson, C.E., Palutikof, J.P., Livermore, M.T.J., Barring, L., Bindi, M., Corte-Real, J., Duaro, R., Giannakopoulos, C., Holt, T., Kundzewicz, Z., Leckebusch, G., Radziejewski, M., Santos, J., Schlyter, P., Schwarb, M., Stjernquist, I. and Ulbrich. U., 2005: 'Modelling the impact of climate extremes: An overview of the MICE project', *Climatic Change*, this volume.

Harpham, C., 2004: *Development of a Novel Radial Basis Function Network Using Genetic Algorithms*, PhD Thesis, University of Derby.

Harpham, C. and Wilby, R.L., 2004: 'Multi-site simulation of daily precipitation amounts using Artificial Neural Networks', *Proceedings of the British Hydrological Society International Hydrology Conference*, London.

Harpham, C. and Wilby, R.L., 2005: 'Multi-site downscaling of heavy daily precipitation occurrence and amounts', *Journal of Hydrology*, in press.

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

Hewitson, B.C. and Crane, R.G., 1996: 'Climate downscaling: techniques and application', *Climate Research*, **7**, 85-95.

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

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

Karl, T.R., Nicholls, N. and Ghazi, A. (eds), 1999: 'Weather and climate extremes: Changes, variations and a perspective from the insurance industry', *Climatic Change*, **42**, 1-349.

Katz, R.W. and Parlange, M.B., 1998: 'Overdispersion phenomenon in stochastic modelling of precipitation', *Journal of Climate*, **11**, 591-601.

Machenhauer, B., Windelband, M., Botzet, M., Hesselbjerg, J., Déqué, M., Jones, G.R., Ruti, P.M. and Visconti, G., 1998: *Validation and Analysis of Regional Present-day Climate and Climate Change Simulations over Europe*, Max-Planck Institute of Meteorology Report No. 275, 87pp.

Mackay, D.J.C., 1992a: 'Bayesian interpolation', *Neural Computation*, **4**, 415-447.

Mackay, D.J.C., 1992b: 'A practical Bayesian framework for backprop networks', *Neural Computation*, **4**, 448-472.

Maheras, P., Flocas, H.A., Anagnostopoulou, C., and Patrikas, I., 2002: 'On the vertical structure of composite surface cyclones in the Mediterranean region', *Theoretical and Applied Climatology*, **71**, 199-217.

Maheras, P., Tolika, K., Anagnostopoulou, C., Vafiadis, M., Patrikas, I. and Flocas, H., 2004: 'On the relationships between circulation types and changes in rainfall variability in Greece', *International Journal of Climatology*, **24**, 1695-1712.

McGinnis, D.L., 1997: 'Estimating climate-change impacts on Colorado Plateau

snowpack using downscaling methods', *Professional Geographer*, **49**, 117-125.

Mearns, L.O., Bogardi, I., Giorgi, F., Matyasovszky, I. and Palecki, M., 1999: 'Comparison of climate change scenarios generated from regional climate model experiments and statistical downscaling', *Journal of Geophysical Research*, **104**, 6603-6621.

Mearns, L.O., Giorgi, F., Whetton, P., Pabon, D., Hulme, M. and Lal, M., 2003: *Guidelines for Use of Climate Scenarios Developed from Regional Climate Model Experiments*, Intergovernmental Panel on Climate Change (IPCC) Task Group on Data and Scenario Support for Impacts and Climate Analysis (TGICA), available from: http://ipcc-ddc.cru.uea.ac.uk/guidelines/dgm_no1_v1_10-2003.pdf.

Meehl, G.A., Zwiers, F., Evans, J., Knutson, T., Mearns, L. and Whetton, P., 2000: 'Trends in extreme weather and climate events: Issues related to modeling extremes in projections of future climate change', *Bulletin of the American Meteorological Society*, **81**, 427-436.

Michaels, P.J., Knappenberger, P.C., Frauenfeld, O.W. and Davie, R.E., 2004: 'Trends in precipitation on the wettest days of the year across the contiguous USA', *International Journal of Climatology*, **24**, 1873-1882.

Moberg, A., and Alexanderson, H., 1997: 'Homogenisation of Swedish temperature data. Part II: Homogenised gridded air temperature compared with a subset of global gridded air temperature since 1861', *International Journal of Climatology*, **17**, 35-54.

Moody, J.E. and Darken, C.J., 1989: 'Fast learning in networks of locally-tuned processing units', *Neural Computation*, **1**, 281-294.

Munich Re, 2002: *Flooding in Central and Eastern Europe – August 2002*. MRNatCatposter 31, www.munichre.com/default_e.asp.

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

Murphy, J.M., 2000: 'Predictions of climate change over Europe using statistical and dynamical downscaling techniques', *International Journal of Climatology*, **20**, 489-501.

Olsson, J., Uvo, C.B. and Jinno, K., 2001: 'Statistical atmospheric downscaling of short-term extreme rainfall by neural networks', *Physics and Chemistry of the Earth Part B – Hydrology, Oceans and Atmosphere*, **26**, 695-700.

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', *Journal of Climate*, **15**, 3529-3548.

Plaut, G., Schuepbach, E. and Doctor, M., 2001: 'Heavy precipitation events over a few Alpine sub-regions and the links to large-scale circulation: 1971-1995', *Climate Research*, **17**, 285-302.

Reid, P.A., Jones, P.D., Brown, O., Goodess, C.M. and Davies, T.D., 2001: 'Assessments of the reliability of NCEP circulation data and relationships with surface climate by direct comparisons with station based data', *Climate Research*, **17**, 247-261.

Rumelhart, D.E. and McClelland, J.L. (eds.), 1986: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, **1**, Cambridge, MA:MIT Press.

Santer, B.D., Hnilo, J.J., Wigley, T.M.L., Boyle, J.S., Doutriaux, C., Fiorino, M., Parker, D.E. and Taylor, K.E., 1999: 'Uncertainties in observationally based estimates of temperature change in the free atmosphere', *Journal of Geophysical Research – Atmospheres*, **104**, 6305-6333.

Schmidli, J. and Frei, C., 2005: 'Trends of heavy precipitation and wet and dry spells

in Switzerland during the 20th century', *International Journal of Climatology*, in press.

Schoof, J.T. and Pryor, S.C., 2001: 'Downscaling temperature and precipitation: A comparison of regression-based methods and artificial neural networks', *International Journal of Climatology*, **21**, 773-790.

Shah, K.P. and Rind, D., 1998: 'Comparing upper tropospheric and lower stratospheric temperatures: microwave sounding unit, radiosonde, COSPAR International Reference Atmosphere, and National Centers for Environmental Prediction – National Center for Atmospheric Research reanalysis monthly mean climatologies', *Journal of Geophysical Research – Atmospheres*, **103**, 31569-31591.

Trigo, R.M. and Palutikof, J.P., 2001: 'Precipitation scenarios over Iberia: A comparison between direct GCM output and different downscaling techniques', *Journal of Climate*, **14**, 4422-4446.

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', *Journal of Climate*, **6**, 1161-1171.

Widmann, M., and Bretherton, C.S., 2003: 'Statistical precipitation downscaling over the Northwestern Unites States using numerically simulated precipitation as a predictor', *Journal of Climate*, **16**, 799-816.

Wijngaard, J.B., Klein Tank, A,M.G. and Konnen, G.P., 2003: 'Homogeneity of 20th century European daily temperature and precipitation series'. *International Journal of Climatology*, **23**, 679-692.

Wilby, R.L. and Wigley, T.M.L., 2000: 'Precipitation predictors for downscaling observed and general circulation model relationships', *International Journal of*

Climatology, **20**, 641-661.

Wilby, R.L., Wigley, T.M.L., Conway, D., Jones, P.D., Hewitson, B.C., Main, J. and Wilks, D.S., 1998: 'Statistical downscaling of general circulation model output: A comparison of methods', *Water Resources Research*, **34**, 2995-3008.

Wilby, R.L., Dawson, C.W. and Barrow, E.M., 2002: 'SDSM – a decision support tool for the assessment of regional climate change impacts', *Environmental Modelling and Software*, **17**, 147-159.

Wilby, R.L., Tomlinson, O.L. and Dawson, C.W., 2003: 'Multi-site simulation of precipitation by conditional resampling', *Climate Research*, **23**, 183-194.

Wilby, R.L., Charles, S.P., Zorita, E., Timbal, B., Whetton, P. and Mearns, L.O., 2004: *Guidelines for Use of Climate Scenarios Developed from Statistical Downscaling Methods*, Intergovernmental Panel on Climate Change (IPCC) Task Group on Data and Scenario Support for Impacts and Climate Analysis (TGICA), available from: http://ipcc-ddc.cru.uea.ac.uk/guidelines/StatDown_Guide.pdf.

Wilks D., 1995: *Statistical Method in the Atmospheric Sciences*, Academic Press.

Wilks, D.S. and Wilby, R.L., 1999: 'The weather generation game: a review of stochastic weather models', *Progress in Physical Geography*, **23**, 329-357.

Williams, P.M., 1998: 'Modelling seasonality and trends in daily rainfall data', in M.I. Jordan, M.J. Kearns and S.A. Solla (eds.), *Advances in Neural Information Processing Systems – Proceedings of the 1997 Conference*, **10**, pp.985-991, MIT Press.

Zorita, E. and von Storch, H., 1999: 'The analog method as a simple statistical downscaling technique: Comparison with more complicated methods', *Journal of Climate*, **12**, 2474-2489.

Table 1: Summary of the advantages and disadvantages of the direct use of regional climate model output to construct scenarios of extremes (Goodess *et al*., 2003). 4 = advantage, 8 = disadvantage, ? = advantage/disadvantage of the method is uncertain.

4	Provides physically-consistent multi-variate information
4	Higher spatial resolution than GCMs should reduce some biases (e.g., more intense extremes)
8	Relatively short (e.g., 30 year) runs make it difficult to assess multi-decadal natural variability
8	Runs may not be available for time periods of interest (e.g., 2020s)
8	Relatively few simulations/ensembles available
8	Affected by biases in the underlying GCM
?	Added value of higher spatial resolution needs to be demonstrated

Table 2: Summary of the advantages and disadvantages of statistical downscaling for the construction of scenarios of extremes (Goodess *et al.*, 2003). 4 = advantage, 8 = disadvantage, ? = advantage/disadvantage of the method is uncertain.

4	Provides station/point values of extremes
4	Less computer intensive than dynamical downscaling
4	Can be applied to GCM and/or RCM output
8	Assumes that predictor/predictand relationships will be unchanged in the future (the stationarity issue)
8	Requires long/reliable observed data series
8	Affected by biases in the underlying GCM
?	May be possible to 'correct' predictors for systematic model biases
?	Scenarios may indicate changes which differ substantially in magnitude, and even in direction, from those based directly on model output
?	Ideally, downscaling methods should reflect the underlying physical mechanisms and processes, but statistical downscaling is unlikely, for example, to treat convective rainfall events in a physically realistic way
?	Sensitive to specific methodology, choice of predictor variables, etc.

Table 3: Summary of data available for the STARDEX regional case-study regions.

Region	*Number of stations in the full regional dataset (and STARDEX group providing the data)*	*Number of stations in the regional subset*	*Stations in the regional subset*
Iberian Peninsula:			
Western Iberia	11 (UEA)	11	Portugal: Beja, Coimbra, Lisboa Geofisica, Santarem, Pegoes, Alvega, Mora, Penhas Douradas, Portalegre Spain: Badajoz/Talavera, Alcuescar
SE Spain	5 (UEA)	5	Albacete/Los Llano., Valencia, Alicante Ciudad Ja., Murcia/Alcantarilla, Murcia/San Javier
Greece:	22 (AUTH)	4 Western Greece 4 Eastern Greece	Agrinio, Ioannina, Kalamata, Kerkyra Alexandroupoli, Mytilini, Samos, Rodos
Alps:	N Alps: 27 grid points, Ticino: 15 grid points, 0.5º precipitation grid, Frei and Schär, 1998 (ETH) 21 temperature (UNIBE)	10	Austria: Innsbruck-Univ. France: Nice, Montelimar Germany: Muenchen Italy: Bologna, Lazzaro Alberoni, Bobbio Switzerland: Arosa, Locarno-Monti, Zuerich
German Rhine:	100 (USTUTT-IWS)	10	Feldberg/Schw., Karlsruhe, Mannheim, Deuselbach, Koeln-Wahn, Giessen, Wuertzburg, Saarbruecken-E., Kahler Asten, Nuernberg-Kra.
UK:			
NW UK	15 precipitation (UEA)	3	Eskdalemuir, Ringway, Shawbury
SE England	28 precipitation (UEA)	3	Cambridge, Goudhurst, Oxford
Northern Italy: (Emilia Romagna)	39 temperature (ARPA) 59 precipitation (ARPA)	8	Bobbio, Lazzaro Alberoni, Bedonia, Bologna, Alfonsine, Parma, Firenzuola, Verghereto

Table 4: Summary of the STARDEX improved statistical downscaling methodologies

Method	***Predictand(s) (Unless otherwise indicated, predictands are station series)***	***Predictor(s)*** ***(See STARDEX Deliverable D10 for selection procedure)***	***Description*** ***(See STARDEX Deliverable D15 for details)***
ADGB_HYPER4	Regional DP index	GPH anomalies at 500 hPa, RH at 700 hPa, geostrophic wind at 500 hPa & precipitable water	Random sampling within the 4-dimensional hyperspace of the 4 predictors which defines conditions for high precipitation
ARPA_CCA	PIE, TIE	SLP, SH at 1000, 950, 850 and 700 hPa, and T at 850 hPa	Canonical Correlation Analysis
ARPA_MLR	PIE, TIE	Z500: First 4 PCs of 500 hPa GPH anomalies T850: First 4 PCs of 850 hPa T	Multiple Linear Regression
AUTH_ANN	DP, DT	500 hPa GPH & 1000-500 hPa thickness	Artificial Neural Network
AUTH_CCA	DP, DT	500 hPa GPH & 1000-500 hPa thickness	Canonical Correlation Analysis
AUTH_MREG	DP, DT	Circulation types for 500 hPa, 1000-500 hPa thickness	Multiple Linear Regression
CNRS_PPCI	DP	Large Scale Circulation patterns defined using 700 hPa GPH	Random selection of an analogue within a set of training days having the same 'Potential Precipitation Circulation Index' category
DMI_CWG	DP	SLP	Conditional weather generator, conditional on quantile values of a circulation index, in which precipitation occurrence and amount are modelled separately
ETH_DYN	DP – station data or mesoscale grids	Grid-box precipitation	As ETH_LOC, but with flow-dependent scaling factors
ETH_DYNI	DP – station data or mesoscale grids	Grid-box precipitation	As ETH_LOCI, but with flow-dependent scaling factors
ETH_LOC	DP – station data or mesoscale grids	Grid-box precipitation	Local scaling of GCM simulated precipitation

ETH_LOCI	DP – station data or mesoscale grids	Grid-box precipitation	Local scaling of GCM simulated precipitation with correction of precipitation frequency and intensity bias
FIC_ANAL2 *(2SA in some figures)*	DP, DT	Geostrophic fluxes at 1000 & 500 hPa, low tropospheric humidity and thickness	Two-step analogue method, in which (1) the 'n' most similar days to the day being simulated are selected from a reference data set and (2) regression is performed using predictand/predictor relationships from the 'n' days data set
KCL_ANN_GA_RBF	DP	The SDSM set of predictors	Genetic algorithm used to optimise the Radial Basis Function network structure and parameters
KCL_ANN_IRBF	DP	The SDSM set of predictors	Individual Radial Basis Function artificial neural network model (i.e., applied to individual sites in each region)
KCL_ANN_MLP	DP	The SDSM set of predictors	Multi Layer Perceptron artificial neural network model
KCL_ANN_RBF	DP	The SDSM set of predictors	Radial Basis Function artificial neural network model (applied across all sites for each region)
KCL_CR	DP	The SDSM set of predictors	Conditional resampling of area average precipitation, conditional on the large-scale atmospheric forcing and a stochastic error term, and daily precipitation amounts at a 'marker site' (generated using SDSM).
UEA_ANN_GAMMA *(GAM in some figures)*	DP	The SDSM set of predictors	Bayesian multilayer perceptron artificial neural networks, using the hybrid Bernoulli/Gamma data misfit term
UEA_ANN_GAMMAMC	DP	The SDSM set of predictors	Bayesian multilayer perceptron artificial neural networks, using the hybrid Bernoulli/Gamma data misfit term and Monte-Carlo simulation
UEA_ANN_SSE	DP	The SDSM set of predictors	Bayesian multilayer perceptron artificial neural networks, using the sum-of-squares data misfit term
UEA_CCA	PIE	CCA1: MSLP CCA4: MSLP + GPH, RH, T at 500, 700 & 850 hPa	Canonical Correlation Analysis
UNIBE_CCA	DT	SLP and GPH, T, SH & RH at 100, 850, 700, 500 and 300 hPa	Canonical Correlation Analysis
USTUTT_MAR	DP	Objective circulation patterns and: - eastward moisture flux at 700	Multivariate Auto-Regressive model

		hPa (for precipitation) - GPH at pressure level corresponding the circulation pattern	
USTUTT_MLR	PIE, TIE	GPH, RH, T, divergence and vorticity at several levels, eastward moisture flux at 700 hpa level and objective circulation patterns	Multiple Linear Regression

DP = daily precipitation
DT = daily temperature
PIE = STARDEX core indices of precipitation extremes
TIE = STARDEX core indices of temperature extremes

GPH = Geopotential height
MSLP = Mean sea level pressure
PC = Principal Component
RH = Relative humidity
SDSM = Statistical DownScaling Model (Wilby *et al*., 2002)
SH = Specific humidity
T = Temperature

Table 5: The STARDEX 10 core indices of extremes and three mean indices.

Precipitation related indices of extremes	
pq90	90th percentile of rainday amounts (mm/day)
px5d	Greatest 5-day total rainfall
pint	Simple daily intensity (rain per rainday)
pxcdd	Maximum number of consecutive dry days
pfl90	% of total rainfall from events > long-term 90th percentile
pnl90	Number of events > long-term 90th percentile of raindays
Temperature related indices of extremes	
txq90	Tmax 90th percentile (ºC)
tnq10	Tmin 10th percentile (ºC)
tnfd	Number of frost days Tmin < 0 °C
txhw90	Heat wave duration (days)
Mean indices	
pav	Precipitation average (mm/day)
txav	Average Tmax (ºC)
tnav	Average Tmin (ºC)

Table 6: The case-study regions in which STARDEX statistical downscaling methods were applied. x = method applied to the European-wide data set (Figure 1). 1 = partner's primary region, 2 = partner's secondary region.

Group_Method(s)	*Europe*	*Iberian Peninsula*	*Greece*	*Alps*	*German Rhine*	*NW and SE UK*	*Northern Italy*
UEA_CCA	x						
FIC_ANAL2	x						
DMI_CWG	x						
UEA_CCA and ANN		2				1	
KCL_ANN and CR		2				1	
UNIBE_CCA				1			
CNRS_PPCI		2		1			
ARPA-SMR_CCA and MLR			2				1
ETH_DYN and LOC				1		2	
USTUTT/FTS_MAR and MLR				2	1		
AUTH_ANN, CCA and MREG			1				2

Table 7: Summary of the seasonal skill scores of the indirect USTUTT_MAR and direct USTUTT_MLR downscaling methods for precipitation indices averaged over 100 German Rhine stations.

Index	*Method*	*Winter*				*Spring*				*Summer*				*Autumn*			
		Mean RMSE	*Min CORR*	*Max CORR*	*Mean CORR*	*Mean RMSE*	*Min CORR*	*Max CORR*	*Mean CORR*	*Mean RMSE*	*Min CORR*	*Max CORR*	*Mean CORR*	*Mean RMSE*	*Min CORR*	*Max CORR*	*Mean CORR*
pav	USTUTT_MAR	0.49	0.11	0.91	0.66	0.50	0.39	0.90	0.71	0.70	-0.10	0.75	0.41	0.55	0.05	0.86	0.46
	USTUTT_MLR	0.45	0.32	0.90	0.71	0.51	0.30	0.90	0.64	0.63	0.16	0.88	0.52	0.49	-0.13	0.86	0.50
pint	USTUTT_MAR	1.09	-0.15	0.76	0.41	1.05	-0.38	0.73	0.35	1.42	-0.55	0.85	0.15	1.28	-0.28	0.64	0.16
	USTUTT_MLR	1.03	-0.14	0.82	0.41	1.13	-0.34	0.82	0.28	1.44	-0.46	0.77	0.12	1.26	-0.44	0.73	0.16
pq90	USTUTT_MAR	3.02	-0.30	0.79	0.33	3.03	-0.31	0.80	0.29	3.98	-0.51	0.78	0.12	3.63	-0.55	0.78	0.22
	USTUTT_MLR	3.15	-0.60	0.81	0.29	3.22	-0.28	0.74	0.23	3.84	-0.68	0.76	0.20	3.89	-0.43	0.71	0.14
px5d	USTUTT_MAR	16.75	-0.09	0.78	0.35	14.60	-0.23	0.84	0.47	19.50	-0.49	0.70	0.13	17.77	-0.21	0.74	0.33
	USTUTT_MLR	15.84	-0.06	0.81	0.41	15.54	-0.33	0.80	0.39	18.32	-0.29	0.79	0.29	18.64	-0.58	0.81	0.19
pxcdd	USTUTT_MAR	4.15	-0.12	0.73	0.38	3.74	-0.06	0.82	0.47	5.76	-0.56	0.68	0.18	5.56	0.14	0.84	0.49
	USTUTT_MLR	4.21	-0.24	0.74	0.31	3.94	-0.39	0.87	0.40	4.94	-0.15	0.90	0.45	5.94	-0.12	0.77	0.41
pfl90	USTUTT_MAR	0.14	-0.33	0.78	0.29	0.14	-0.51	0.81	0.20	0.15	-0.65	0.72	0.10	0.14	-0.43	0.85	0.22
	USTUTT_MLR	0.13	-0.08	0.80	0.33	0.15	-0.30	0.69	0.14	0.15	-0.61	0.66	0.11	0.16	-0.51	0.51	0.04
pnl90	USTUTT_MAR	1.98	-0.28	0.86	0.42	1.89	-0.05	0.90	0.45	1.90	-0.56	0.88	0.21	1.81	-0.54	0.80	0.30
	USTUTT_MLR	1.92	-0.07	0.90	0.48	1.84	-0.12	0.86	0.42	1.80	-0.22	0.77	0.32	1.75	-0.50	0.81	0.21

Figure Captions

Figure 1: Location of the 495 stations in the STARDEX European-wide dataset.

Figure 2: Seasonal box-and-whisker plots of the Spearman correlation skill for four statistical downscaling methods for six precipitation indices and 10 Alpine stations.

Figure 3: Spearman correlations for 14 statistical downscaling methods and four seasons, averaged across the seven precipitation indices and 6 UK stations. Methods are from the following groups: DMI (CWG), ETH (DYN, DYNI, LOC, LOCI), FIC (2SA), KCL (IRBF, MLP, RBF, CR) and UEA (GAM, SSE, CCA4, CCA1).

Figure 4: Spearman correlations for six statistical downscaling methods and four seasons, averaged across the seven precipitation indices and 16 Iberian Peninsula stations. Methods are from the following groups: DMI (CWG), KCL (GA-RBF, RBF, CR), FIC (2SA), and UEA (CCA4).

Figure 5: Spearman correlations for eight statistical downscaling methods, four seasons and the seven precipitation indices, averaged across four stations from Western Greece. It = ARPA methods (see Table 4).

Figure 6: Spearman correlations for four seasons and the six temperature indices, averaged across three statistical downscaling methods (FIC_ANAL2, UEA_CCA4 and USTUTT_MLR) and 10 German Rhine stations.

Figure 7: Spearman correlations for four seasons and the seven precipitation indices, averaged across five statistical downscaling methods (DMI_CWG, FIC_ANAL2, UEA_CCA4, USTUTT_MAR and USTUTT_MLR) and 10 German Rhine stations.

Figure 8: Spearman correlations for average minimum temperature (tnav) downscaled using the FIC_ANAL2 method for the European-wide dataset, for winter (upper panel) and summer (lower panel).

Figure 9: Spearman correlations for average precipitation (pav) downscaled using the FIC_ANAL2 method for the European-wide dataset, for winter (upper panel) and summer (lower panel).

Figure 10: Spearman correlations for the six temperature indices, five statistical downscaling methods and four seasons, averaged over eight Northern Italian (Emilia Romagna) stations.

Figure 11: Spearman correlations for the seven precipitation indices, five statistical downscaling methods and four seasons, averaged over 10 German Rhine stations.

Figure 12: Spearman correlations for average minimum temperature (tnav) downscaled using the UEA_CCA4 method for the European-wide dataset, for winter (upper panel) and summer (lower panel).

Figure 13: Spearman correlations for the 10th percentile of minimum temperature (tnq10) downscaled using the FIC_ANAL2 method for the European-wide dataset, for winter (upper panel) and summer (lower panel).

Figure 14: Spearman correlations for the greatest 5-day rainfall total (px5d) downscaled using the FIC_ANAL2 method for the European-wide dataset, for winter (upper panel) and summer (lower panel).

Figure 15a: Spearman correlations for six statistical downscaling methods, averaged across the seven precipitation indices, four seasons and 16 Iberian Peninsula stations.

Figure 15b: Rank of absolute bias for six statistical downscaling methods, averaged across the seven precipitation indices, four seasons and 16 Iberian Peninsula stations. Lower biases are given a higher rank.

Methods are from the following groups: DMI (CWG), KCL (GA-RBF, RBF, CR), FIC (2SA), and UEA (CCA4). The vertical bars and lines shown the 5th and 95th percentiles of the distributions of the correlations and bias ranks.

Figure 16: Seasonal box-and-whisker plots of the ratio of downscaled standard deviation over observed standard deviation for four statistical downscaling methods for six precipitation indices and 10 Alpine stations.

Figure 1

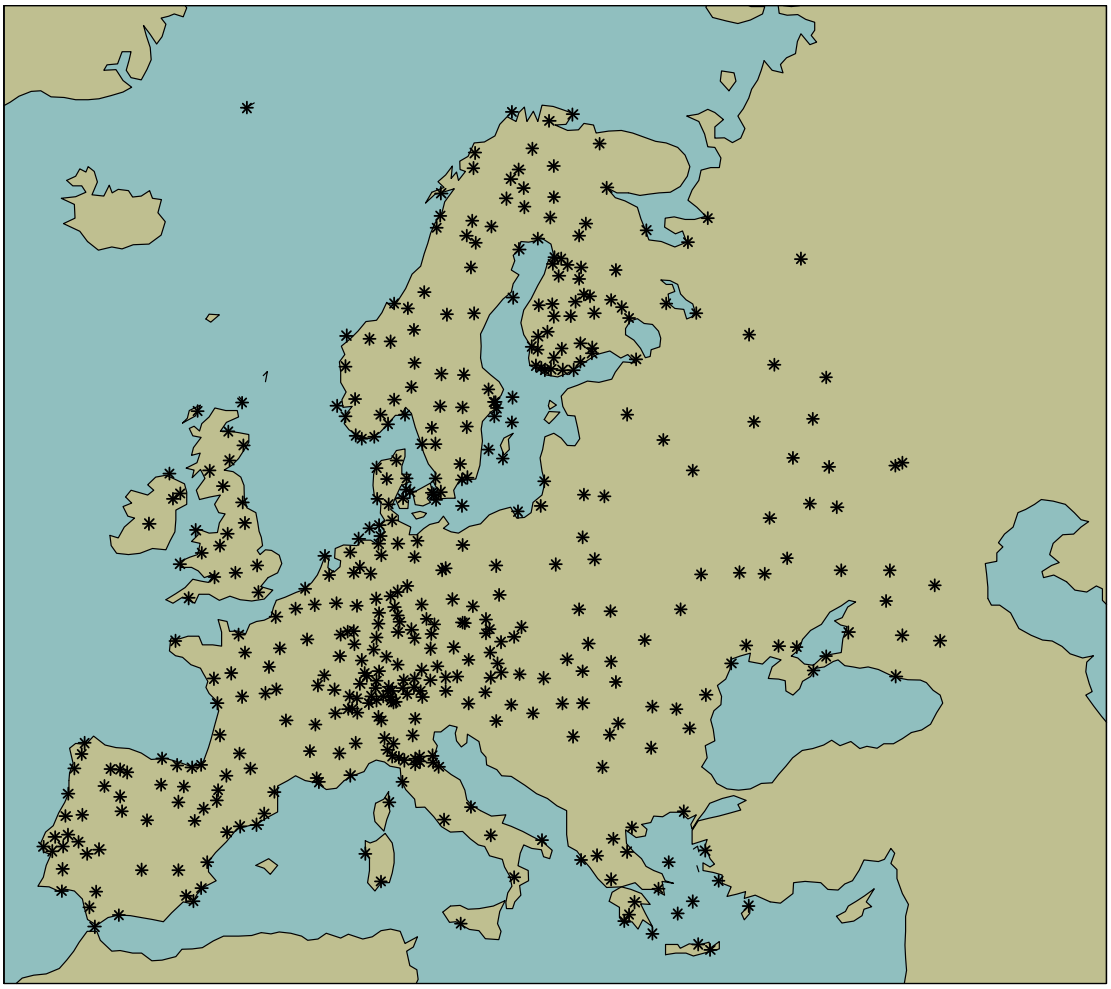
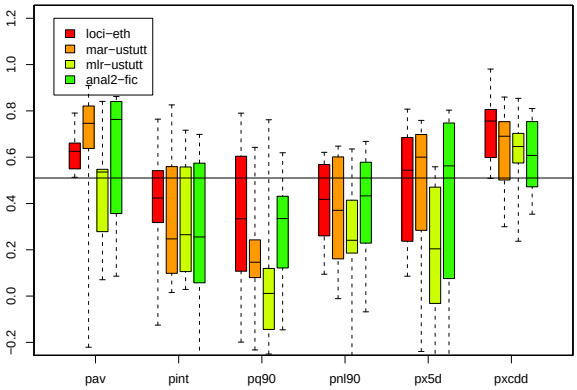
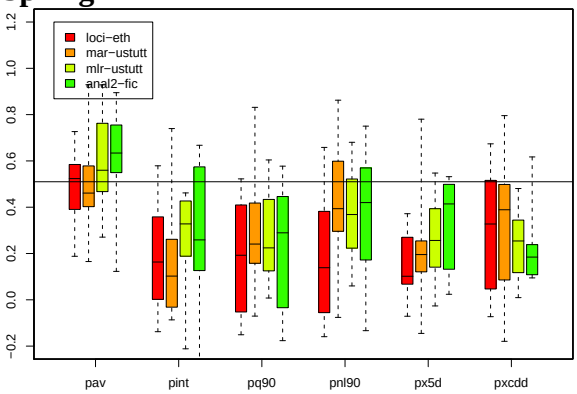


Figure 2

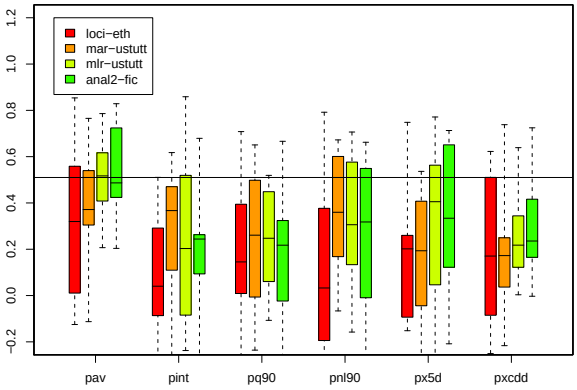
Winter



Spring



Summer



Autumn

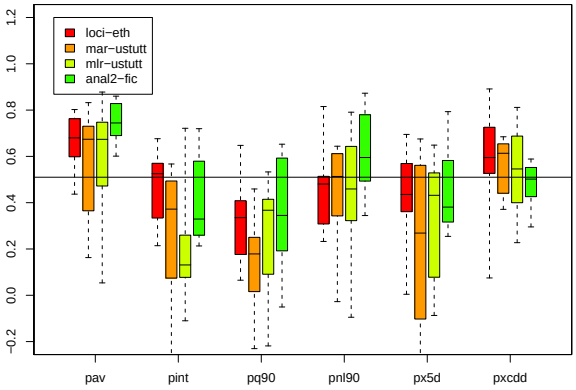


Figure 3:

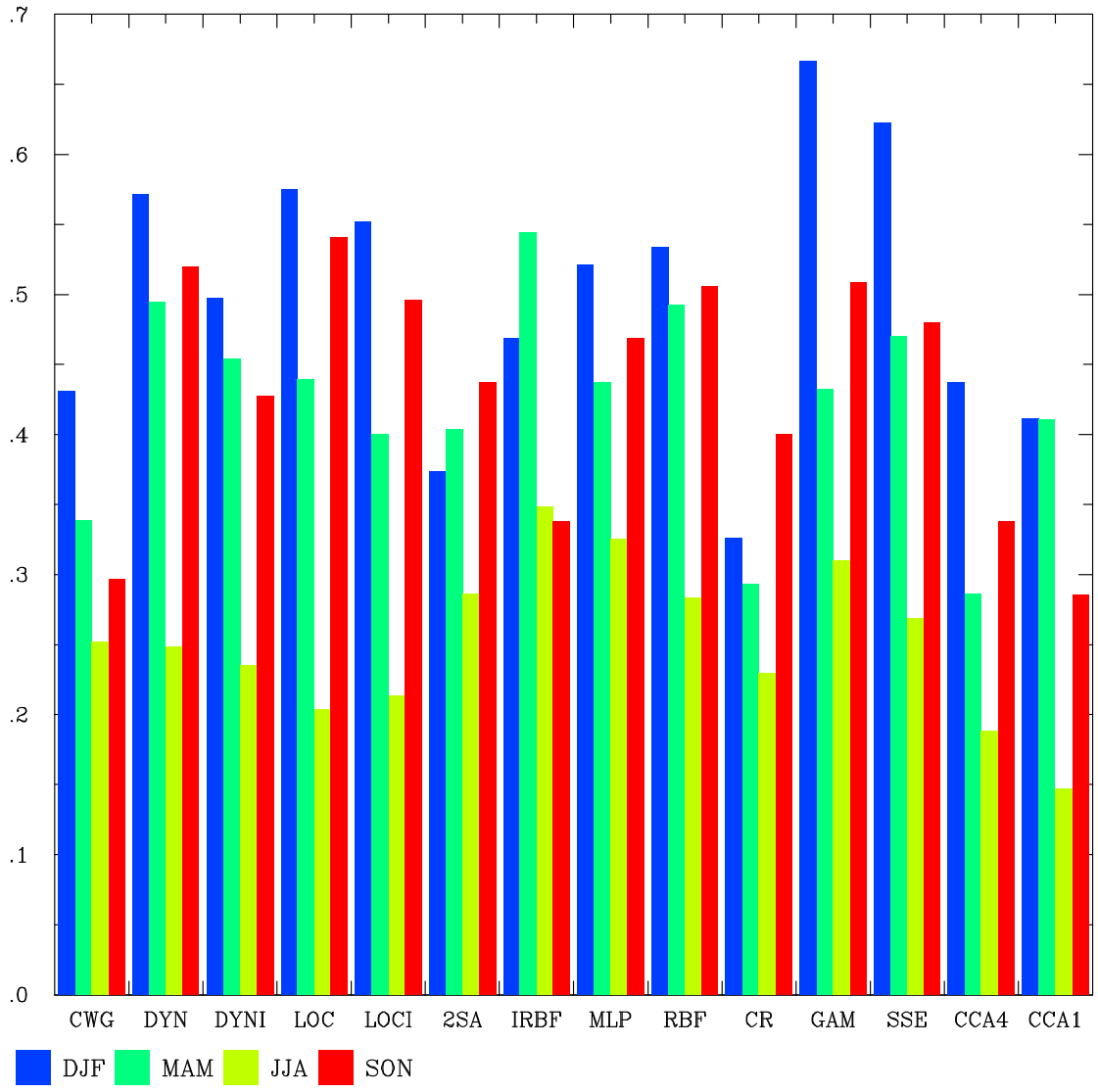

.7
.6
.5
.4
.3
.2
.1
.0
CWG
DYN
DYNI
LOC
LOCI
2SA
IRBF
MLP
RBF
CR
GAM
SSE
CCA4
CCA1
DJF
MAM
JJA
SON

Figure 4

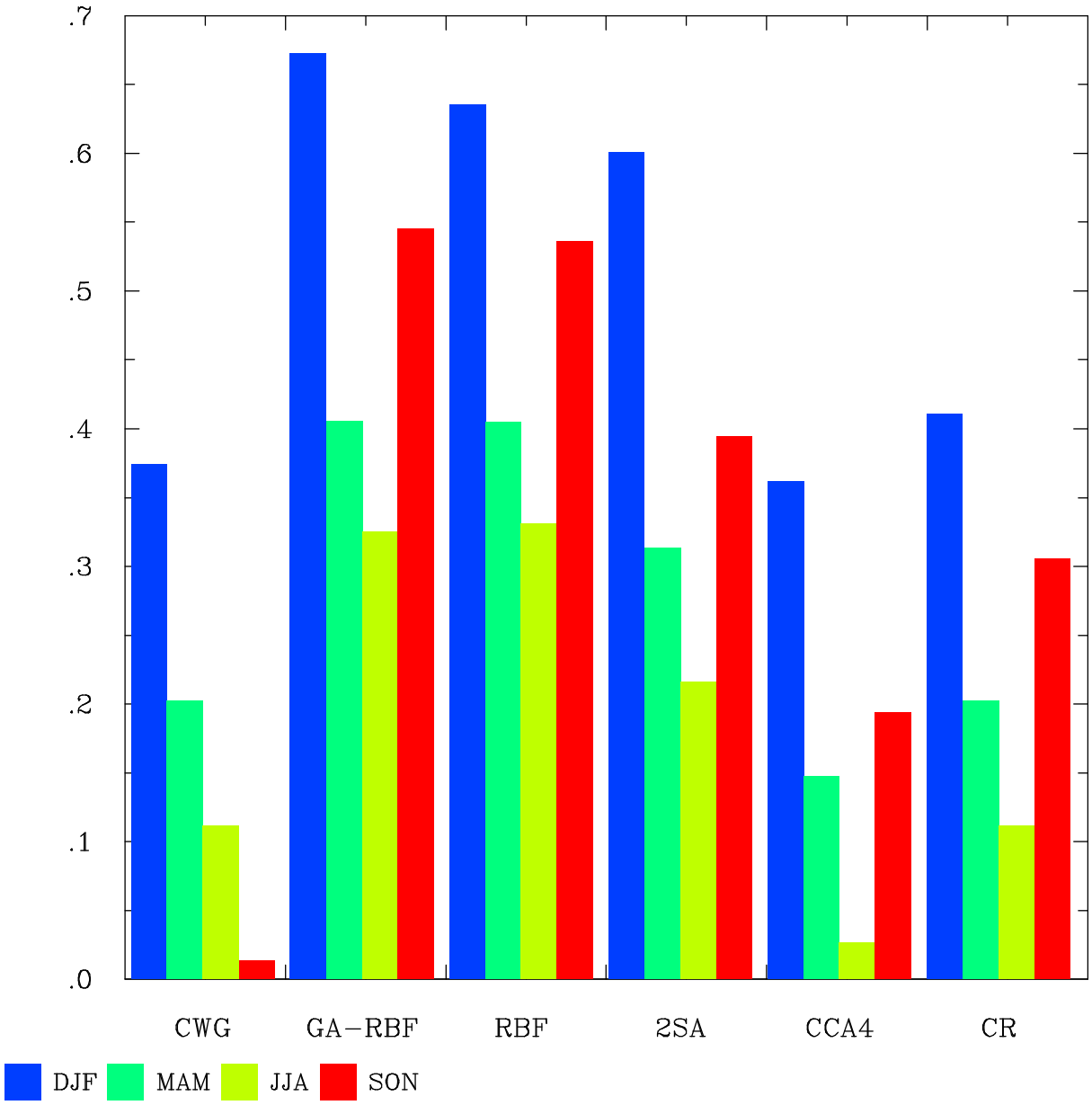
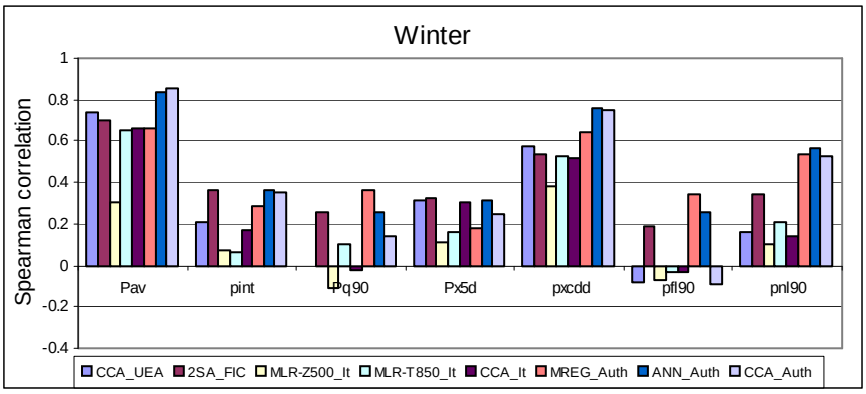
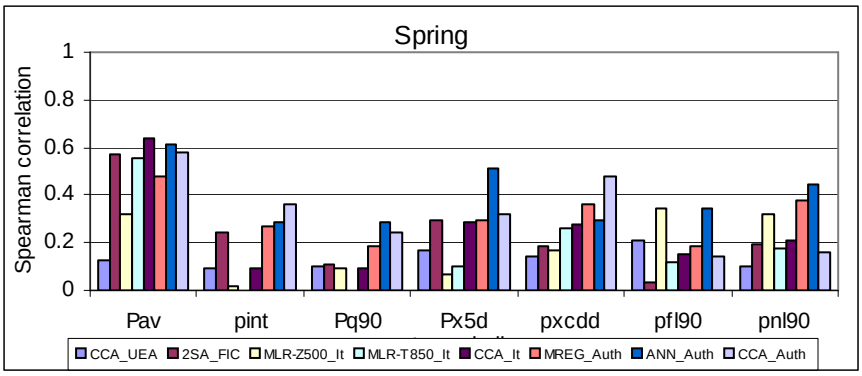
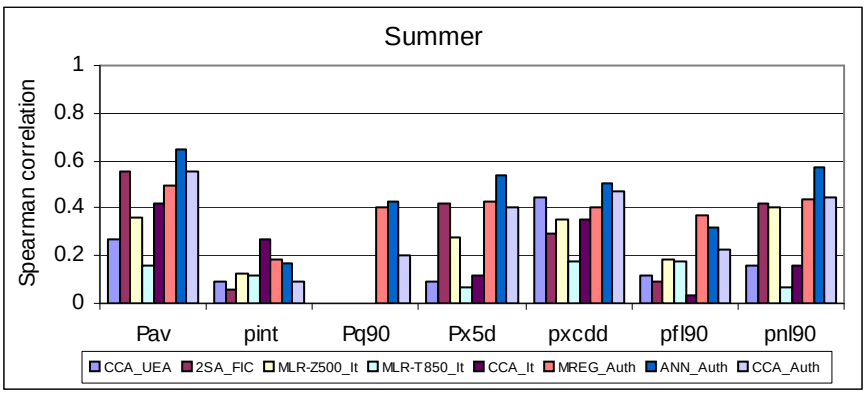


Figure 5


Winter
Spearman correlation
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
Pav
pint
Pq90
Px5d
pxcdd
pfl90
pnl90
CCA_UEA
2SA_FIC
MLR-Z500_It
MLR-T850_It
CCA_It
MREG_Auth
ANN_Auth
CCA_Auth


Spring
Spearman correlation
1
0.8
0.6
0.4
0.2
0
Pav
pint
Pq90
Px5d
pxcdd
pfl90
pnl90
CCA_UEA
2SA_FIC
MLR-Z500_It
MLR-T850_It
CCA_It
MREG_Auth
ANN_Auth
CCA_Auth


Summer
Spearman correlation
1
0.8
0.6
0.4
0.2
0
Pav
pint
Pq90
Px5d
pxcdd
pfl90
pnl90
CCA_UEA
2SA_FIC
MLR-Z500_It
MLR-T850_It
CCA_It
MREG_Auth
ANN_Auth
CCA_Auth

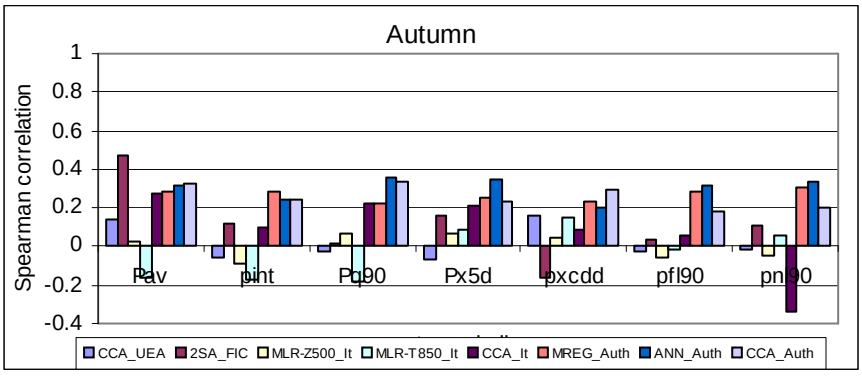

Autumn
Spearman correlation
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
Pav
pint
Pq90
Px5d
pxcdd
pfl90
pnl90
CCA_UEA
2SA_FIC
MLR-Z500_It
MLR-T850_It
CCA_It
MREG_Auth
ANN_Auth
CCA_Auth

Figure 6

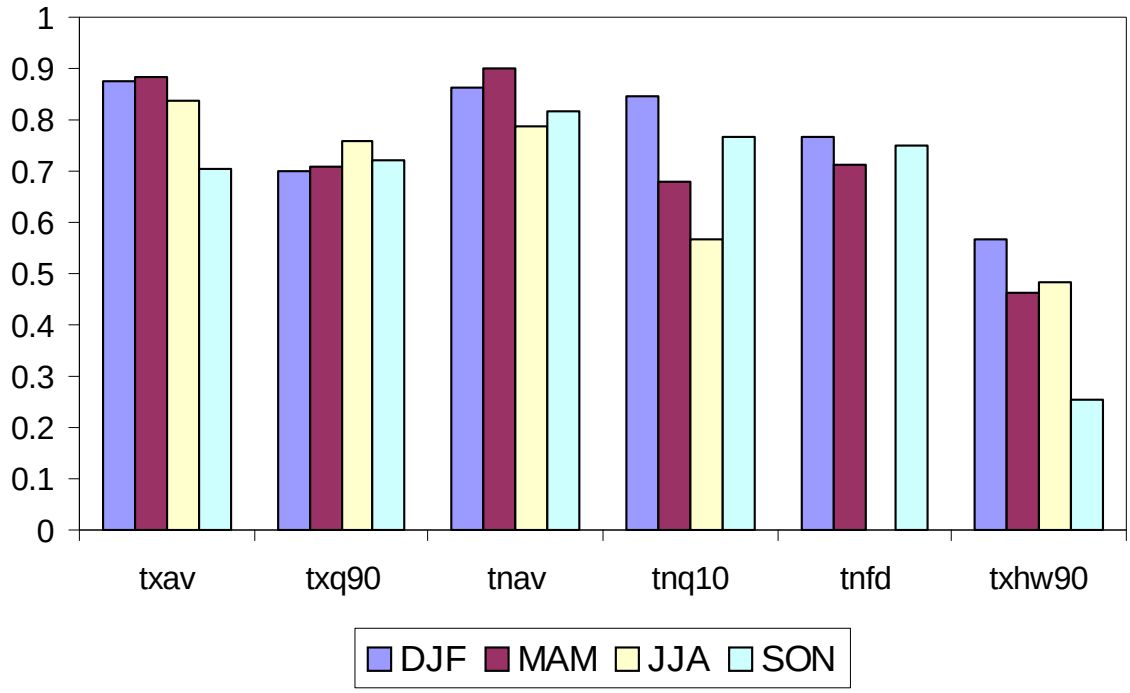

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
txav
txq90
tnav
tnq10
tnfd
txhw90
DJF
MAM
JJA
SON

Figure 7

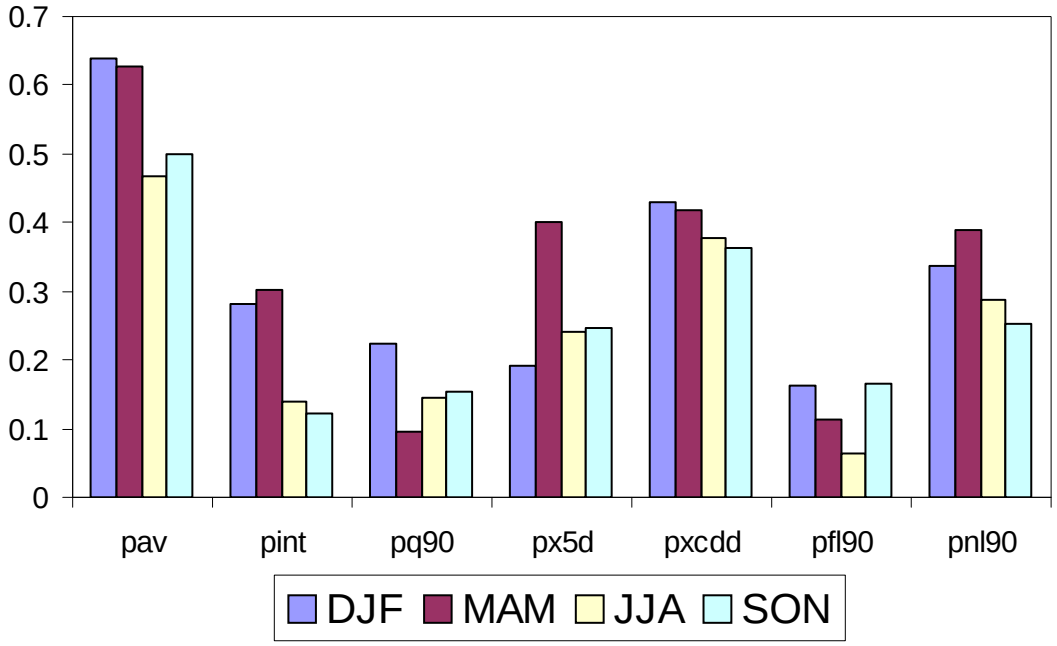

0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
pav
pint
pq90
px5d
pxcdd
pfl90
pnl90
DJF
MAM
JJA
SON

Figure 8

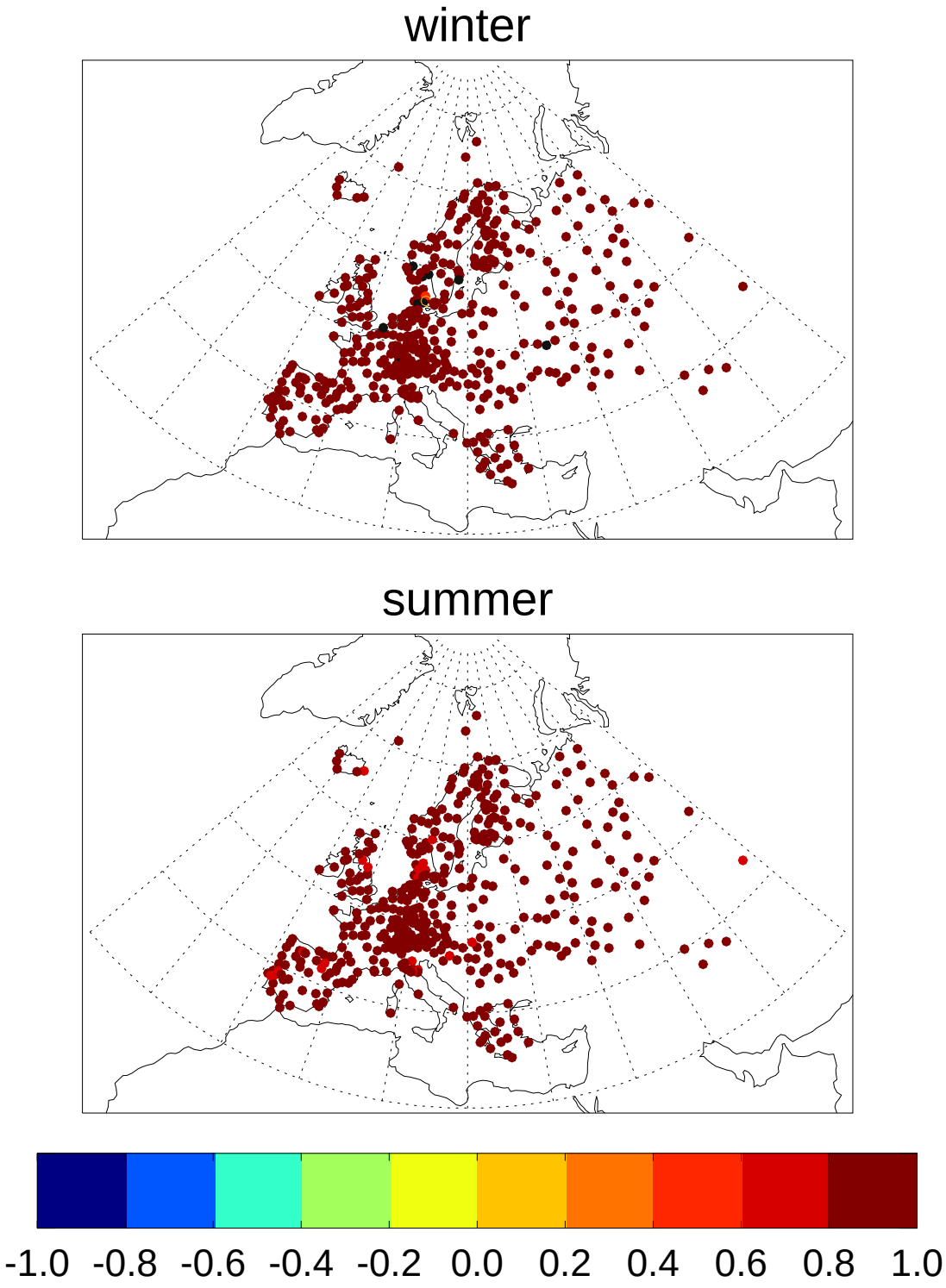


Figure 9

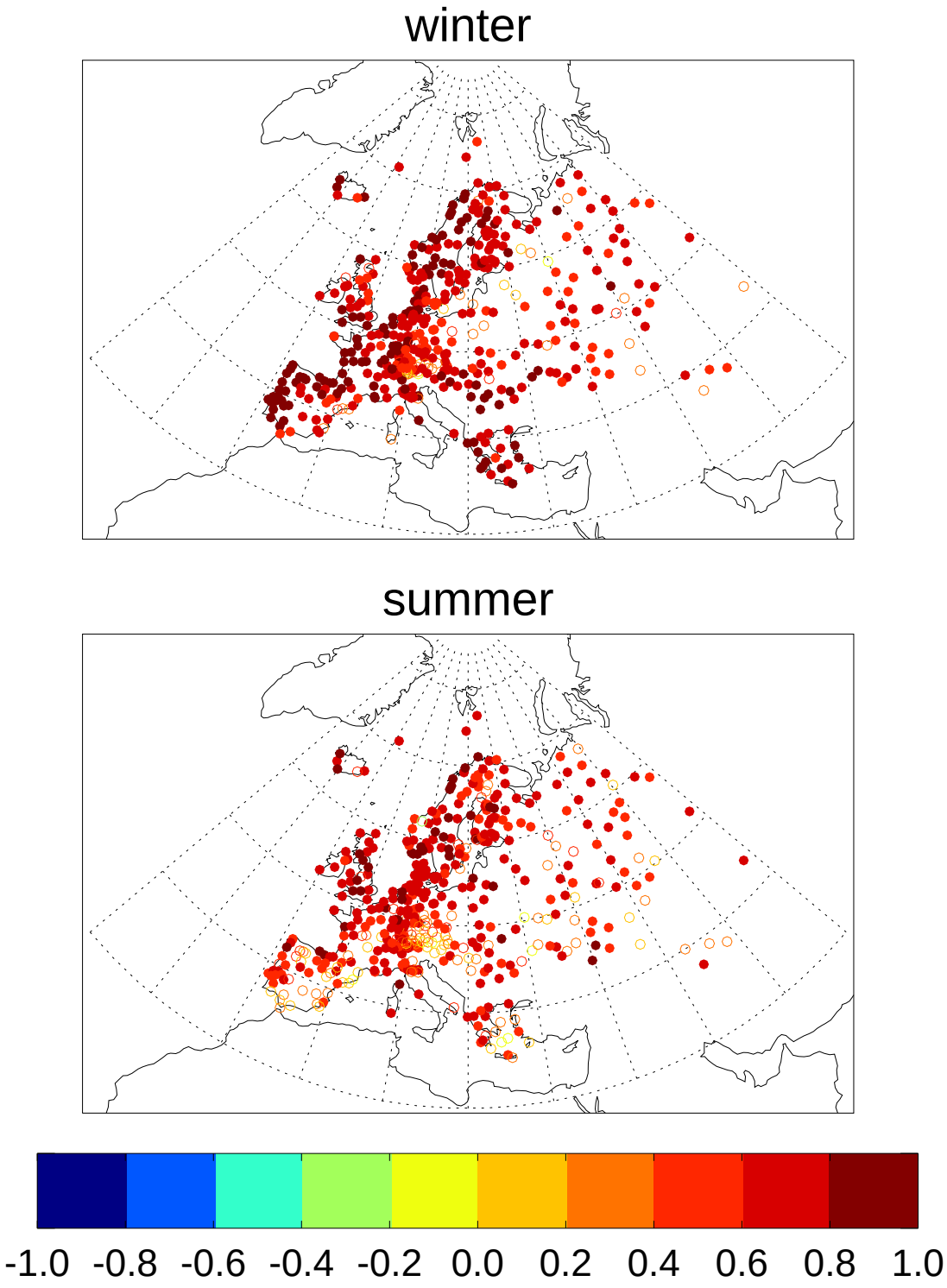
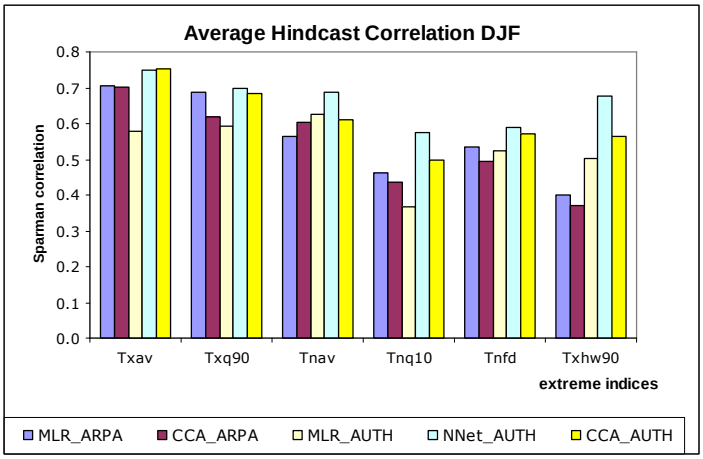
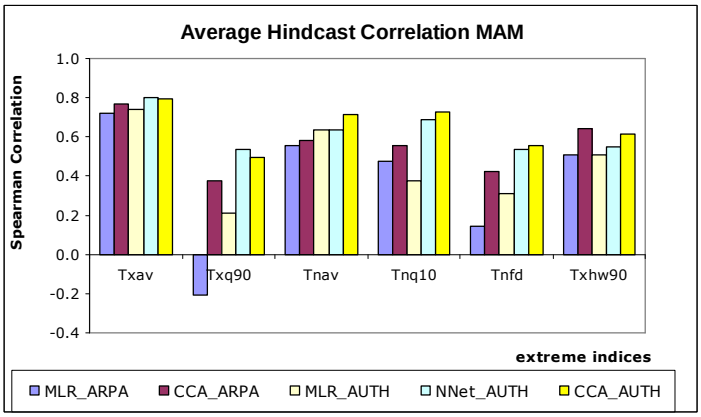
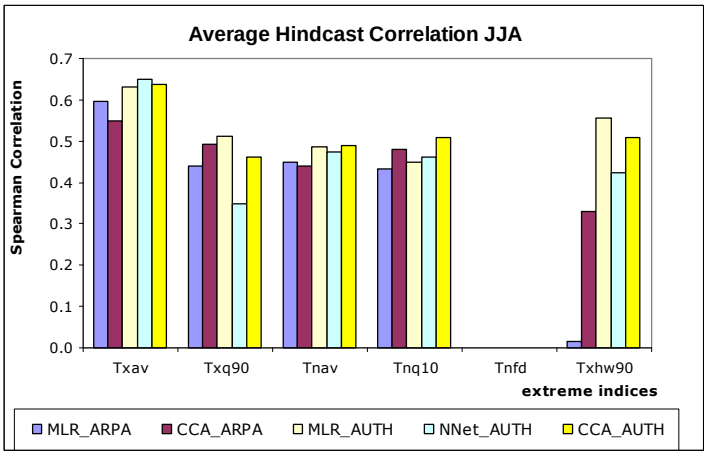


Figure 10


Average Hindcast Correlation DJF
Sparman correlation
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Txav
Txq90
Tnav
Tnq10
Tnfd
Txhw90
extreme indices
MLR_ARPA
CCA_ARPA
MLR_AUTH
NNet_AUTH
CCA_AUTH


Average Hindcast Correlation MAM
Spearman Correlation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
Txav
Txq90
Tnav
Tnq10
Tnfd
Txhw90
extreme indices
MLR_ARPA
CCA_ARPA
MLR_AUTH
NNet_AUTH
CCA_AUTH


Average Hindcast Correlation JJA
Spearman Correlation
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Txav
Txq90
Tnav
Tnq10
Tnfd
Txhw90
extreme indices
MLR_ARPA
CCA_ARPA
MLR_AUTH
NNet_AUTH
CCA_AUTH

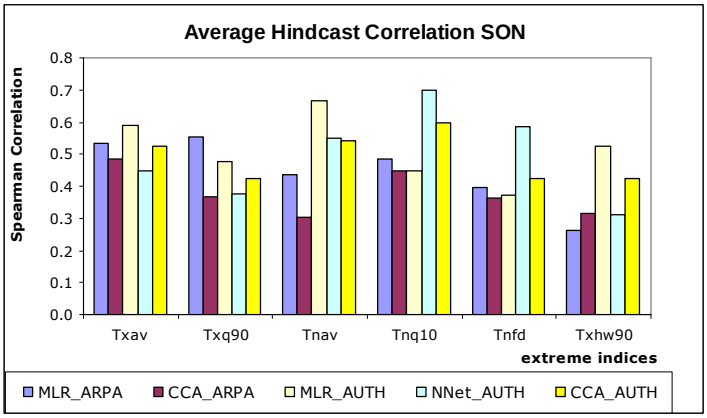
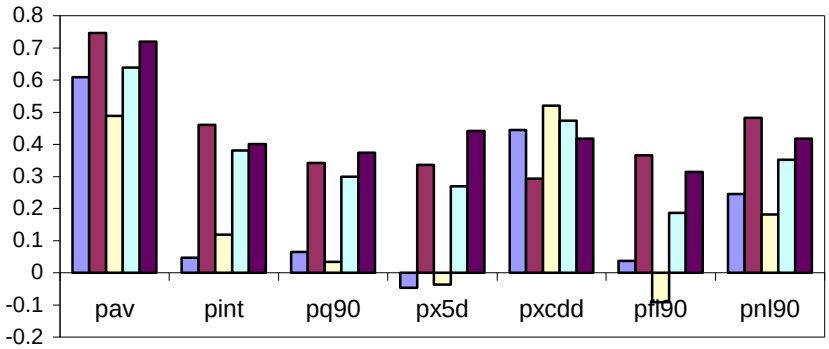
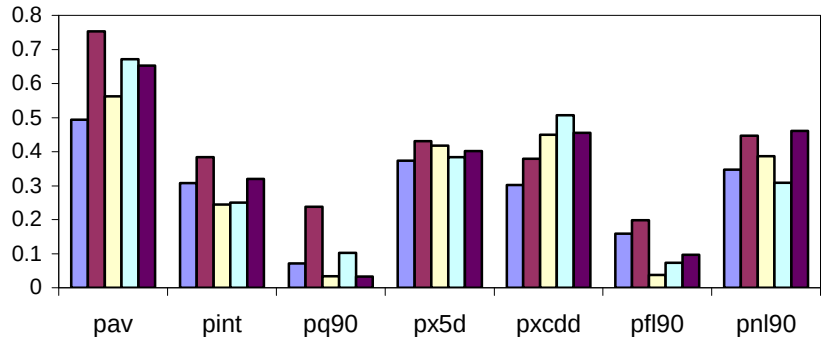

Average Hindcast Correlation SON
Spearman Correlation
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Txav
Txq90
Tnav
Tnq10
Tnfd
Txhw90
extreme indices
MLR_ARPA
CCA_ARPA
MLR_AUTH
NNet_AUTH
CCA_AUTH

Figure 11

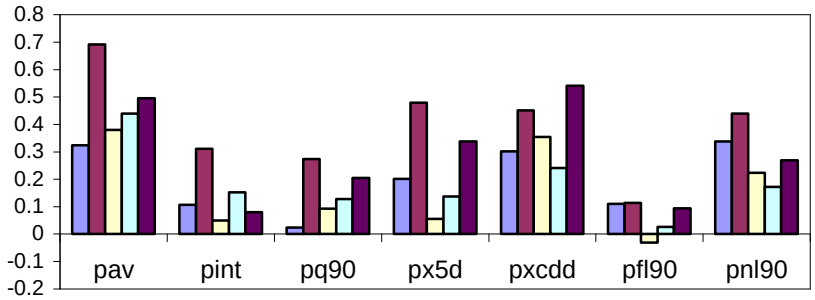
Winter



Spring



Summer



Autumn

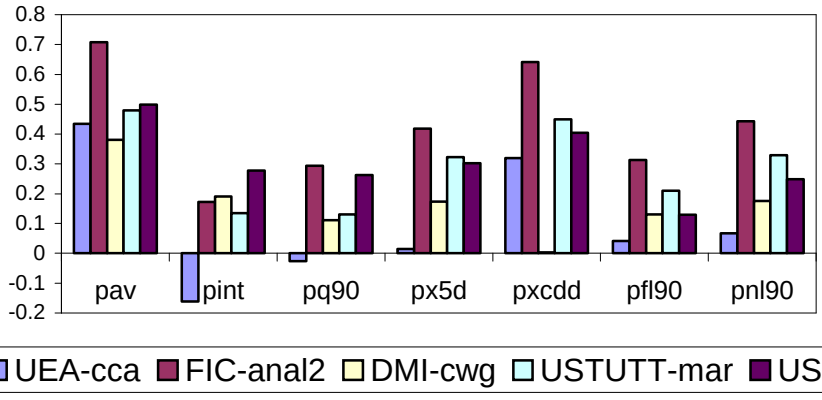


Figure 12

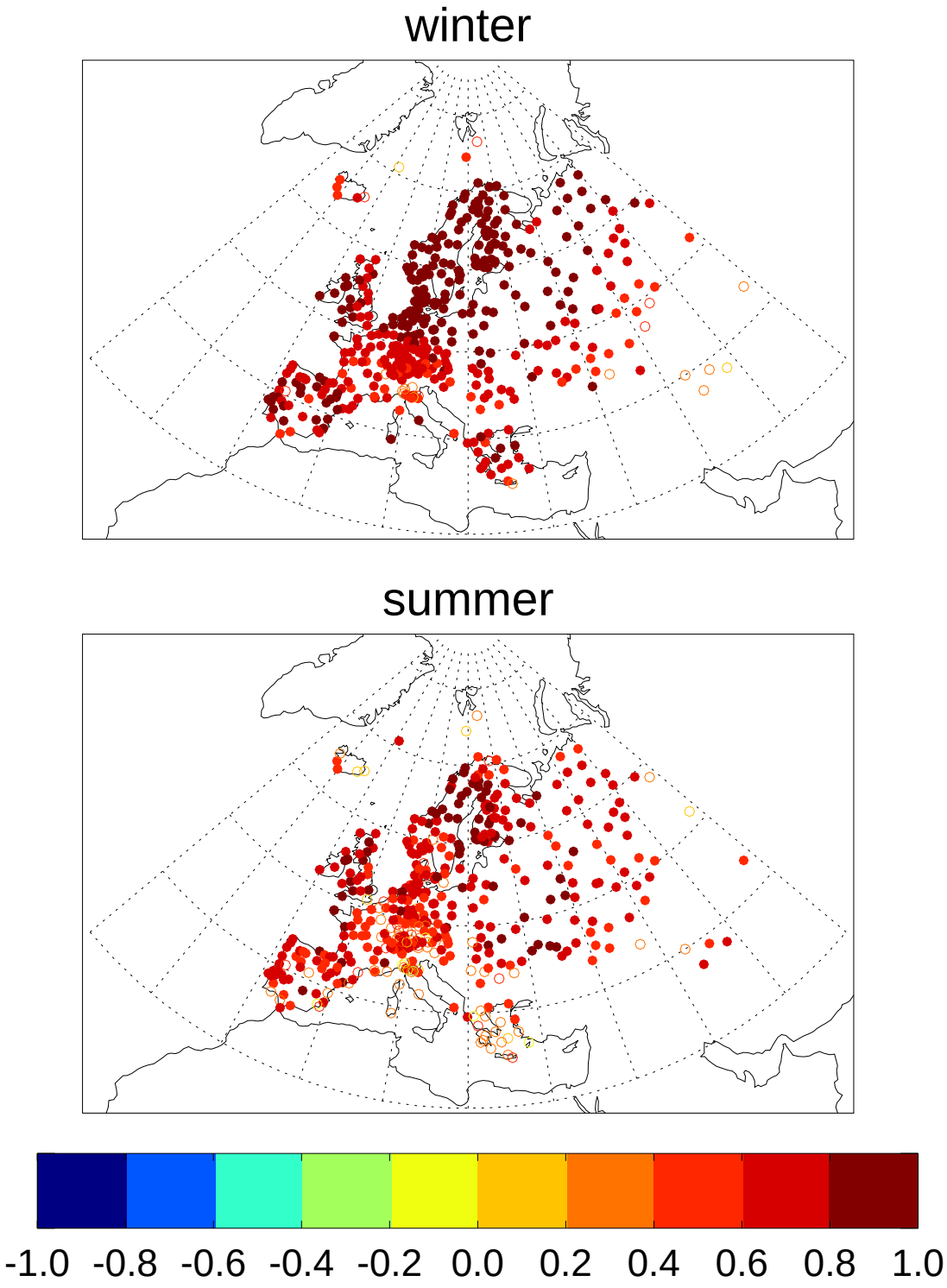


Figure 13

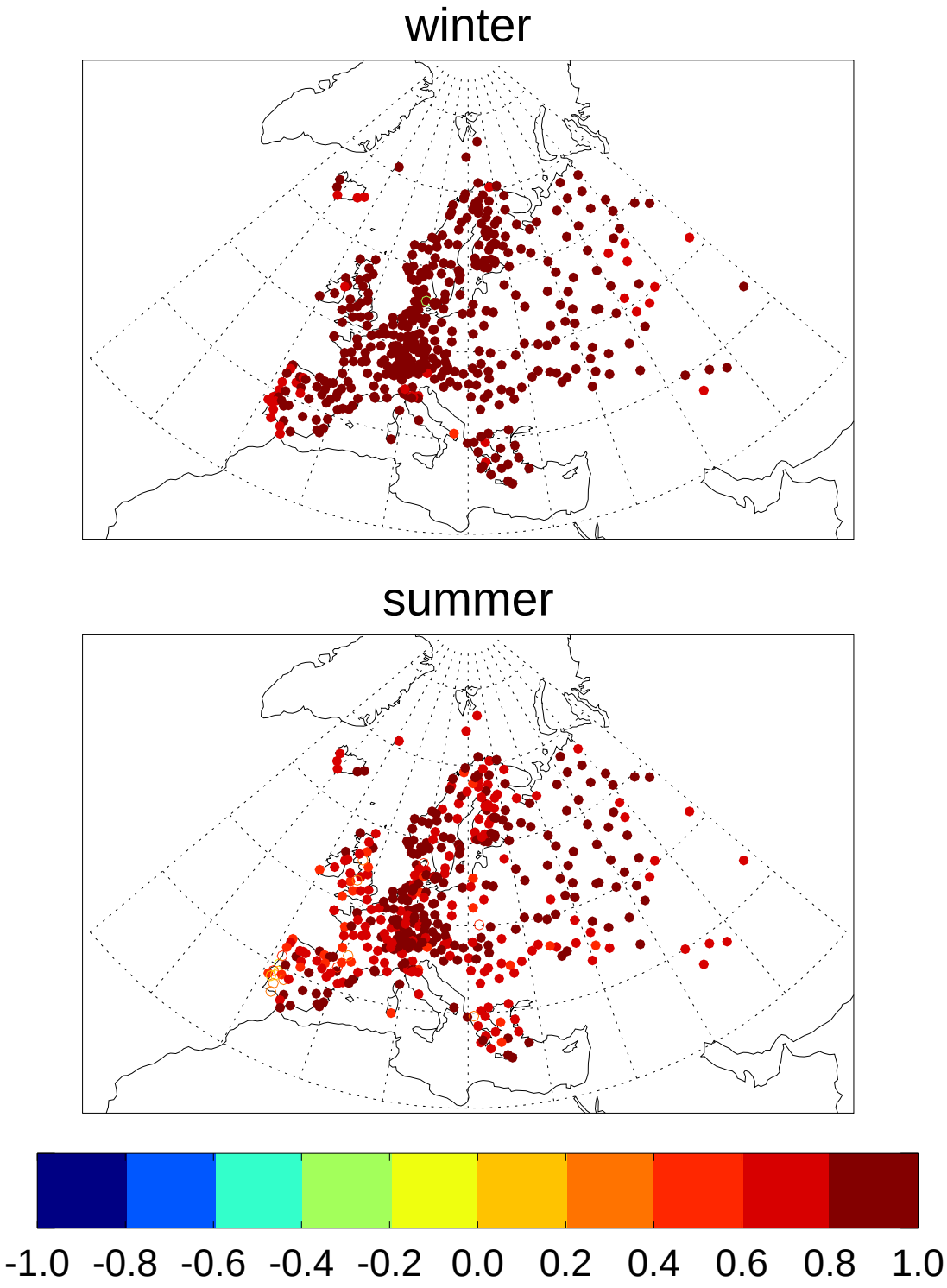


Figure 14

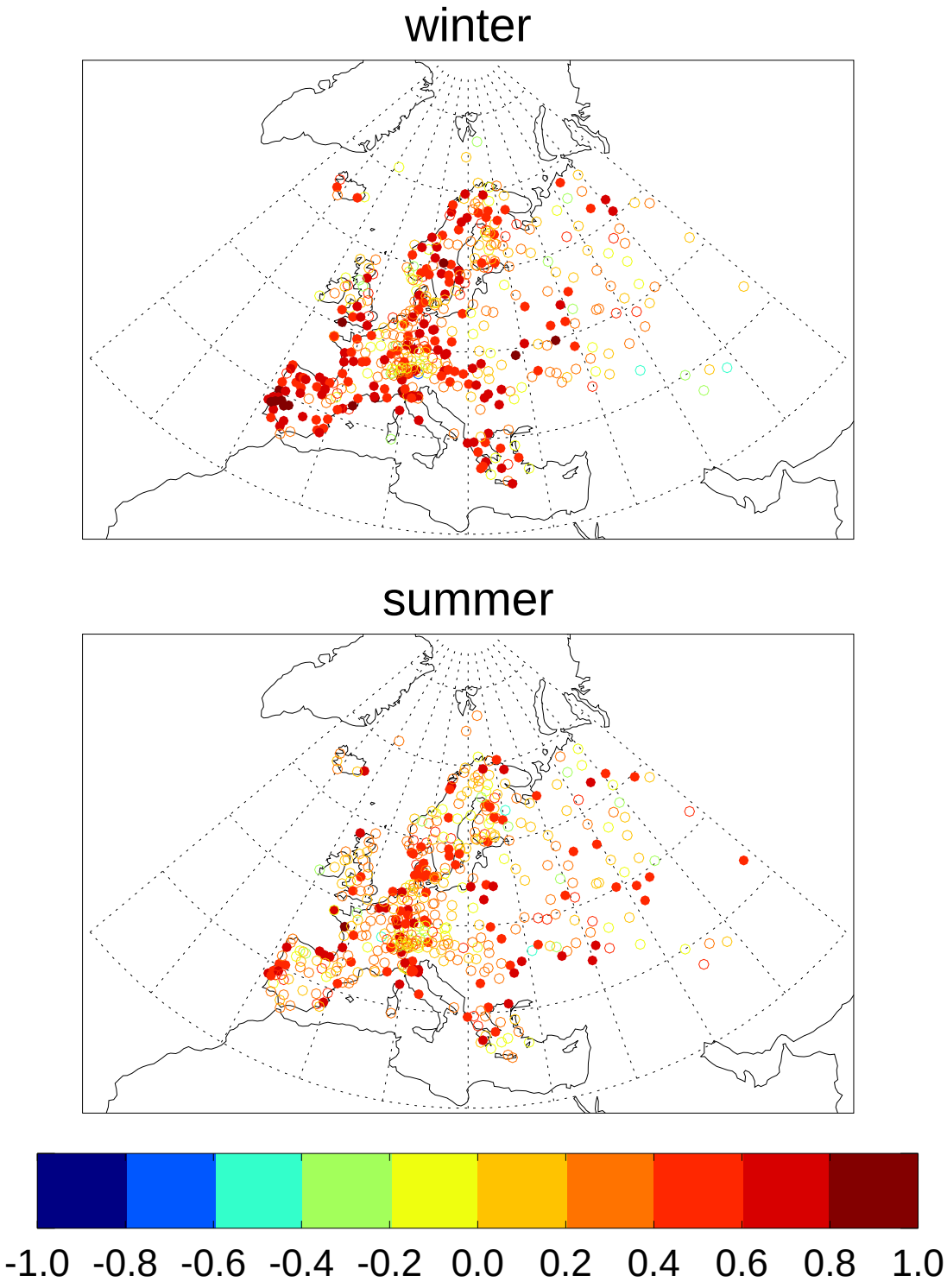


Figure 15a:

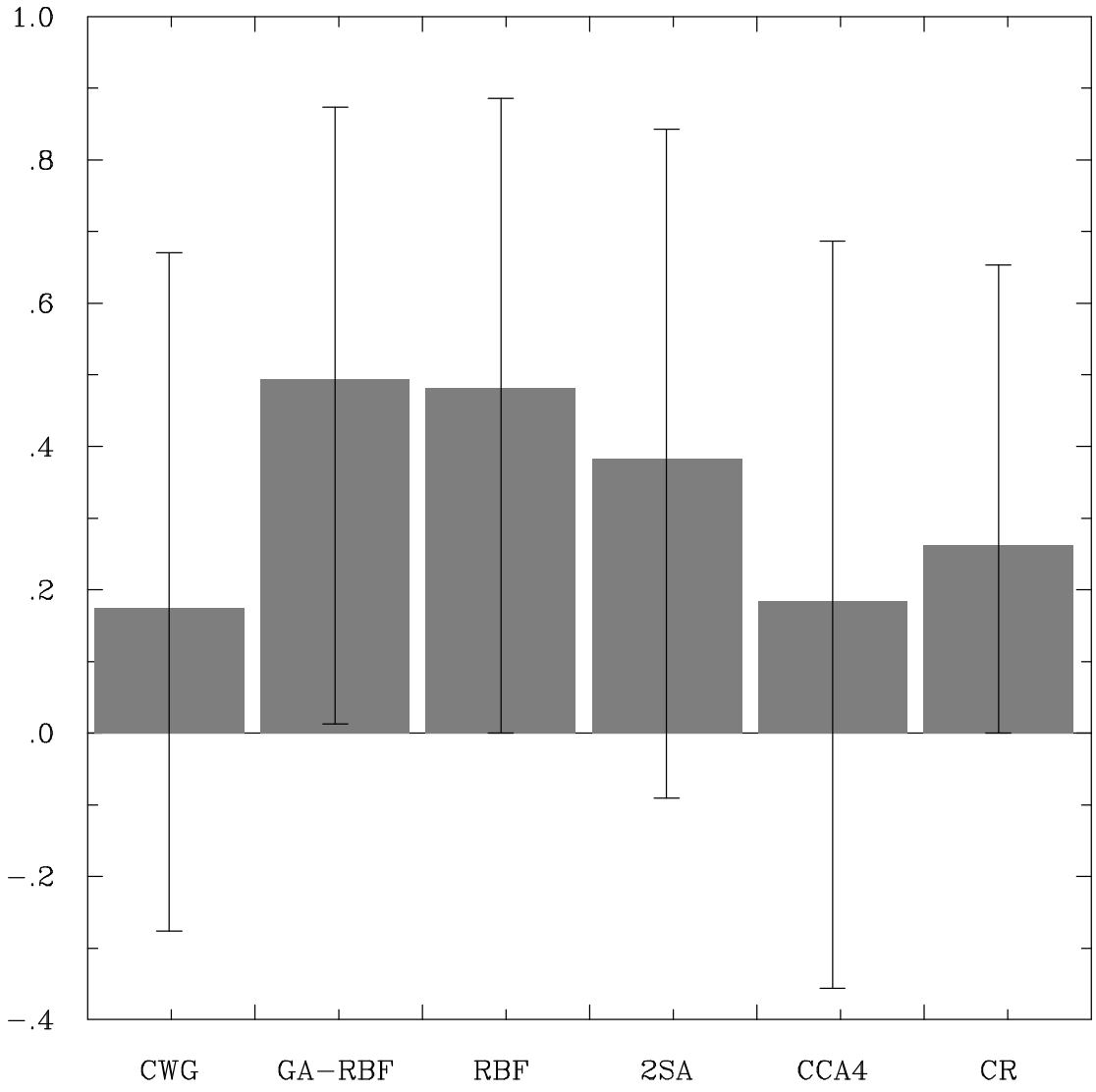


Figure 15b

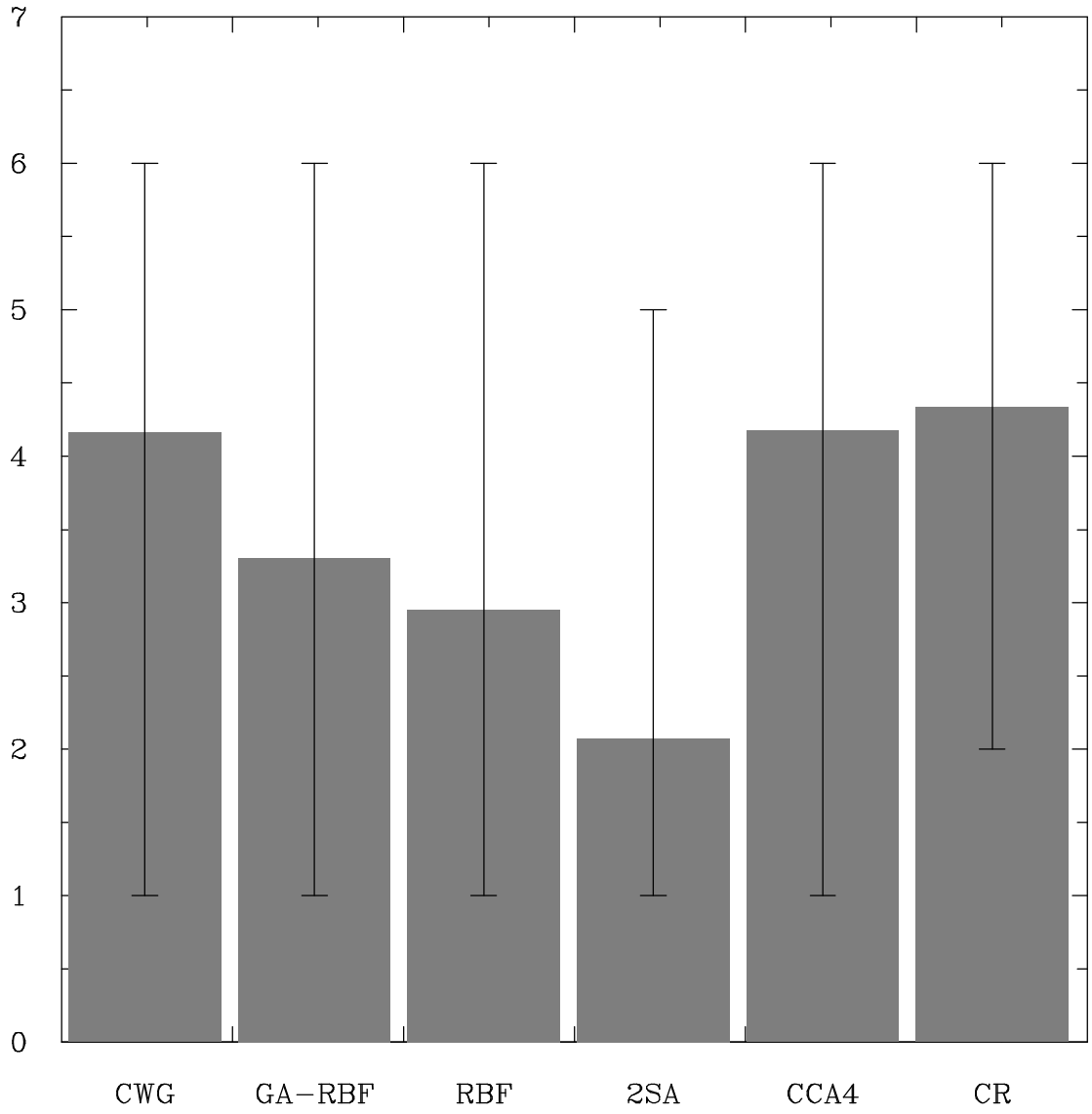
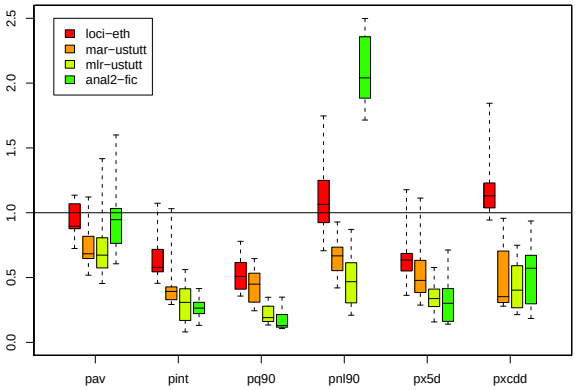
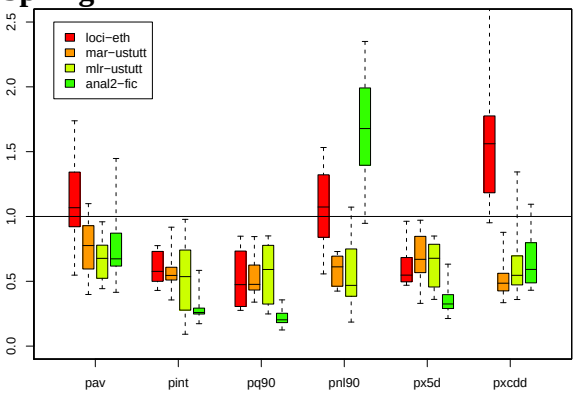


Figure 16

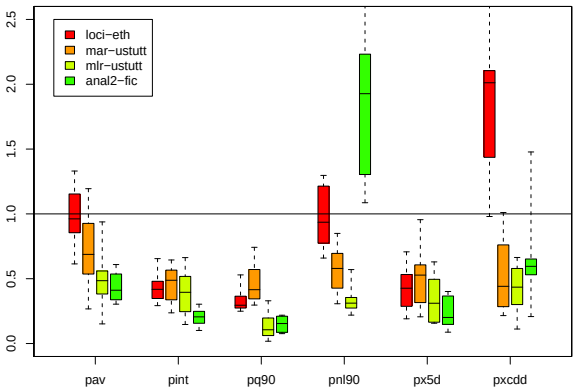
Winter



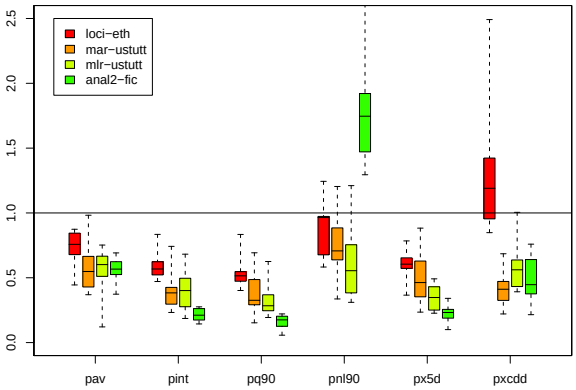
Spring



Summer



Autumn



University of East Anglia

Authors

Clare Goodess is a Senior Research Associate and Research and Administration Manager in the Climatic Research Unit, School of Environmental Sciences, University of East Anglia.

Christina Anagnostopoulou

András Bárdossy

Christoph Frei

Colin Harpham

Malcolm Haylock

Yeshewa Hundecha

Panagiotis Maheras

Jaime Ribalaygua

Juerg Schmidli

Torben Schmith

Konstantia Tolika

Rodica Tomozeiu

Rob Wilby

Climatic Research Unit
School of Environmental Sciences
University of East Anglia