An intercomparison of a large ensemble of statistical downscaling methods over Europe: Results from the VALUE perfect predictor cross-validation experiment

J.M. Gutiérrez[1]*, D. Maraun[2], M. Widman[3], R. Huth[4], E. Hertig[5], R. Benestad[6], O. Roessler[7], J. Wibig[8], R. Wilcke[9], S. Kotlarski[10], D. San Martín[1,11], S. Herrera[12], J. Bedia[1], A. Casanueva[12], R. Manzanas[1], M. Iturbide[1], M. Vrac[13], M. Dubrovsky[14], J. Ribalaygua[15], J. Pórtoles[15], O. Räty[16], J. Räisänen[16], B. Hingray[17], D. Raynaud[17], M. J. Casado[19], P. Ramos[19], T. Zerenner[20], M. Turco[21], T. Bosshard[22], P. Štěpánek[23], J. Bartholy[24], R. Pongracz[24], D.E. Keller[10,25], A.M. Fischer[10], R.M. Cardoso[26], P.M.M. Soares[26], B. Czernecki[27], C. Pagé[28]

[1]Meteorology Group. Instituto de Física de Cantabria, CSIC-Univ. of Cantabria, Spain.
[2]Wegener Center for Climate and Global Change, University of Graz, Austria.
[3]School of Geography, Earth and Environmental Sciences, University of Birmingham, UK.
[4]Dept. of Physical Geography and Geoecology, Faculty of Science, Charles University; and Institute of Atmospheric Physics, Czech Academy of Sciences, Czech Republic.
[5]Institute of Geography, University of Augsburg, Germany.
[6]The Norwegian Meteorological Institute, Norway.
[7]Department of Geography / Oeschger Centre for Climate Change Research, University of Bern, Switzerland.
[8]Department of Meteorology and Climatology, University of Lodz, Poland.
[9]Rossby Centre, Swedish Meteorological and Hydrological Institute, Sweden.
[10]Federal Office of Meteorology and Climatology MeteoSwiss, Switzerland.
[11]Predictia Intelligent Data Solutions, SME. Spain.
[12]Meteorology Group. Dpto. de Matemática Aplicada y Computación. Univ. of Cantabria, Spain.
[13]Laboratoire des Sciences du Climat et de l'Environnement (LSCE-IPSL/CNRS), France.
[14]Institute of Atmospheric Physics, Czech Academy of Sciences, Czech Republic.
[15]Fundación para la Investigación del Clima (FIC), Spain.
[16]University of Helsinki (UHEL), Finland.
[17]Univ. Grenoble Alpes, CNRS, IRD, Grenoble INP, IGE, France
[19]Agencia Estatal de Meteorologia (AEMET), Spain.
[20]Meteorological Institute. University of Bonn, Germany.
[21]Department of Applied Physics, University of Barcelona, Spain.
[22]Swedish Meteorological and Hydrological Institute (SMHI), Sweden.
[23]Global Change Research Institute CAS, Czech Republic.
[24]Eötvös Loránd University (ELU), Hungary.
[25]Center for Climate Systems Modeling (C2SM), ETH Zurich, Switzerland.
[26]Instituto Dom Luiz, Universidade de Lisboa (IDL), Portugal.
[27]Adam Mickiewicz University, Poland.
[28]CECI, CERFACS - CNRS, France.

*_Corresponding author. Email_: gutierjm@unican.es

Abstract

VALUE is an open European collaboration to intercompare downscaling approaches for climate change research, focusing on different validation aspects (marginal, temporal, extremes, spatial, process-based, etc.). Here we describe the participating methods and first results from the first experiment, using "perfect" reanalysis (and RCM reanalysis-driven) predictors to assess the intrinsic performance of the methods for downscaling precipitation and temperatures over a set of 86 stations representative of the main climatic regions in Europe. This study constitutes the largest and most comprehensive to date intercomparison of statistical downscaling downscaling methods, covering the three common downscaling approaches (perfect prognosis, model output statistics —including bias correction— and weather generators) with a total of over fifty downscaling methods representative of the most common techniques.

Overall, most of the downscaling methods greatly improve raw model biases and no approach or technique seems to be superior in general, since there is a large method-to-method variability. The main factors most influencing the results are the seasonal calibration of the methods (e.g. using a moving window) and their stochastic nature. The particular predictors used also played an important role in cases where the comparison was possible, both for the validation results and for the strength of the predictor-predictand link, indicating the local variability explained. However, the present study cannot give a conclusive assessment of the skill of the methods to simulate regional future climates, and further experiments will be soon performed in the framework of the EURO-CORDEX initiative (where VALUE activities have merged and follow on).

Finally, research transparency and reproducibility has been a major concern and substantive steps have been taken. In particular, the necessary data to run the experiments is provided at `http://www.value-cost.eu/data` and data and validation results are available from the VALUE Validation Portal for further investigation: `http://www.value-cost.eu/validationportal`.

KEY WORDS: *downscaling*, *bias adjustment*, *perfect prognosis*, *model output statistics*, *weather generators*, *validation*, *reproducibility*, *CORDEX*.

1. Introduction

Global Climate Models (GCMs) are the primary tools to simulate multi-decadal climate dynamics and to generate global climate change projections under different future emission scenarios (Taylor et al. 2011). However, these models have a coarse resolution (typically a few hundred kilometers) and suffer from substantial systematic biases when compared with observations (Flato et al. 2013, Sec. 9.6). Therefore, they are unable to provide actionable information at the regional and local spatial scales required in impact and adaptation studies. In order to bridge this gap, two main downscaling approaches have been developed since the early 1990s (Leung et al. 2003; Maraun et al. 2010): Dynamical downscaling (based on Regional Climate Models, RCMs) and Empirical/Statistical Downscaling (ESD, based on statistical models). The relative merits and limitations of both dynamical and statistical downscaling —and combinations of them,— have been widely discussed in the literature (see, e.g,. Fowler et al. 2007; Maraun et al. 2010; Winkler et al. 2011; Takayabu et al. 2016) and it is nowadays recognized that they are complementary in many practical applications.

Dynamical downscaling is carried out running one or several RCMs on a relatively fine grid (e.g. 10-20 km) over a limited domain (e.g. Europe) initialized and driven at the boundaries by the coarse GCM outputs to be downscaled (Giorgi and Mearns 1991; Rummukainen 2010, for a review). These models are able to generate regional physically-consistent predictions for a suite of climate variables (particularly those less affected by model parameterizations), but still may suffer from significant biases (see, e.g. Kotlarski et al. 2014; Casanueva et al. 2016b) which require statistical post-processing before they can be used in impact applications. Ensembles of RCMs have been extensively intercompared in the framework of a series of subsequent community intercomparison initiatives considering increasing resolutions, e.g. PRUDENCE ($0.44°$, Christensen and Christensen 2007), ENSEMBLES ($0.22°$, Christensen et al. 2010) and, more recently, EURO-CORDEX ($0.11°$, Jacob et al. 2014), all focusing on Europe. These experiments include control simulations driven by "perfect" reanalysis boundary conditions —to evaluate the intrinsic performance of the different RCMs,— and GCM driven simulations under different (historical and future) scenarios.

ESD methods rely on statistical models linking informative GCM outputs (predictors) to the local observed predictand of interest over a particular domain (Benestad et al. 2008; Maraun and Widmann 2017). These models are first trained (and tested, e.g., cross-validated) using model and observed data during a representative historical period, and later applied to new (e.g. future) GCM data to obtain the downscaled local predictions. According to the nature of predictors in the training phase, three main approaches for ESD exist (see, e.g. Maraun et al. 2010): 1) Perfect Prognosis (PP), 2) Model Output Statistics (MOS) —including the increasingly popular Bias Correction (BC) techniques,— and 3) Weather Generators (WG). On the one hand, under the PP approach, quasi-observed predictors from reanalysis are used to train the statistical models based on their temporal (e.g. daily or monthly) correspondence with observations in the historical training period. Therefore, predictor variables well represented by both reanalyses and GCMs, and accounting for a major part of the variability in the predictands, are typically chosen in this approach (usually large-scale variables at different vertical levels), whereas variables directly influenced by model parameterizations and/or orography (e.g. precipitation) are usually discarded (Wilby et al. 2004). As a result, one of the most time-consuming tasks of this approach is the selection of a suitable combination of predictors, defined over a particular geographical domain which encompasses the main synoptic phenomena influencing the climate of the region of interest. On the other hand, under the MOS approach, model outputs from the GCM are directly used for training, thus correcting systematic biases against observations. In particular, simple MOS alternatives based on BC techniques are becoming increasingly popular in climate change applications to adjust both GCM and RCM outputs (see, e.g., Themeßl et al. 2012). These techniques adjust the model output distribution towards the observed one to ensure resemblance to the local climatology. The main advantage of MOS techniques is their simplicity, since no predictor/domain screening is typically required (e.g. GCM output for the target variable from the closest model gridbox is commonly considered as the unique predictor). Finally, WG is a third approach which does not explicitly include GCM predictors in the training phase (Wilks and Wilby 1999). The simplest form of WGs are Markov-like processess fitted to the local observed data, which are able to reproduce the local temporal and marginal statistical properties from a set of parameters derived from basic climatological statistics (e.g. autocorrelation, wet-day frequency, mean and standard deviation). The global climate change signals from the GCMs are later temporally disaggregated by producing daily time series from the WG with the parameters transformed according to the projected statistics.

As a result of the intensive research activity carried out in this field during the last two decades, a large number of studies exist mostly describing specific ESD methods and/or applications in different regions of the world, using different validation methodologies and/or experimental frameworks. There are also some intercomparison studies focusing on particular approaches, either PP (Haylock et al. 2006; Frost et al. 2011; Teutschbein et al. 2011; Hu et al. 2013; Gutiérrez et al. 2013; San-Martín et al. 2017), MOS (Teutschbein and Seibert 2013; Gutmann et al. 2014), or WG (Semenov et al. 1998; Hartkamp et al. 2003). Moreover, a few multi-approach intercomparison studies are also available, starting with the pioneering work by Wilby et al. (1998) who analyzed PP and WG methods, and following with the more recent PP and MOS comparisons by Bürger et al. (2012) and Vaittinada Ayar et al. (2016). However, limited comprehensive information is yet available at a continental level (e.g. over Europe) for the informed application of the different ESD approaches for climate change impact and adaptation studies.

The EU Cooperation in Science and Technology (COST) Action ES1102 VALUE (2012-2015, Maraun et al. 2015, http://www.value-cost.eu) has been the first international initiative to create a community for statistical downscaling intercomparison, providing a common experimental framework and developing community validation tools for different validation aspects (marginal statistics, temporal structure, extremes, spatial coherence, process based). Several experiments have been designed to isolate specific points in the downscaling procedure where problems may occur (Maraun et al. 2015). In this paper we describe the cross-validation experiment with "perfect" reanalysis —and RCM reanalysis driven— predictors to downscale precipitation and temperatures over Europe, with over 50 different participating methods, covering the three downscaling approaches (PP, MOS and WG) and the common techniques (quantile mapping, analogs, linear and generalized linear regression, weather typing, etc.). The present paper describes in detail the contributing methods, analyzing the selection and transformation of predictors and the geographical extent, and their influence on the resulting predictor-predictand relationships. This paper also focuses on key method characteristics (e.g. deterministic/stochastic) and implementation details (e.g. seasonal/annual train) which may be relevant for the analysis of the validation results (and is used as metadata in this work). In this contribution we only focus on validation results for the marginal distributions (biases in the mean and the standard deviation of the distributions), but other validation aspects are analyzed in the different papers of this special issue.

Overall, this work constitutes the most comprehensive to date intercomparison of downscaling methods on a continental scale over Europe. We want to remark that the final goal is not ranking the different methods according to their performance, but providing an indication of the relative merits and limitations of the different approaches and families of techniques. Thus, some clearly poor performing methods have been also included to illustrate problems. We want to remark that this experiment alone is not sufficient to evaluate the limitations of (MOS) bias correction techniques (see Maraun et al. 2017, for more details). Moreover, it also does not fully validate PP techniques since further results using GCM predictors are needed to evaluate whether well-represented predictors have been used and the PP assumption is valid. Moreover, this work provides no information on the the extrapolation capabilities (to future climates) of the different MOS and PP techniques (although the reproduction of reanalysis trends is analyzed in Maraun et al. 2018; in this special issue). These problems will be analyzed in subsequent community-open experiments using GCM predictors from historical and future scenarios, which will be open for participation in the framework of the EURO-CORDEX initiative (where VALUE activities have merged and follow on).

Research transparency and reproducibility has been a major concern in this work and substantive steps have been taken to improve the reproducibility of the methods and results, and to promote awareness within the downscaling scientific community. In particular, the necessary data to run the experiments is provided at http://www.value-cost.eu/data, and both the downscaled data and the individual validation results are available at the VALUE validation portal http://www.value-cost.eu/validationportal.

The paper is organized as follows. The experimental framework followed and the predictor and predictand data used are described in Sec. 2. Section 3 presents the methods contributing to this study (a brief description and specific implementation details for each method are given in Annex 1). It also describes the selection of the predictors and data preparation and analyzes the predictor-predictand link established by the different methods. Sections 4 and 5 presents the validation results for precipitation and temperatures, respectively, focusing on the biases in the mean and the standard deviation resulting from the methods. Information regarding transparency and reproducibility of results is given in Sec. 6. Finally, the main conclusions obtained are reported in Sec. 7.

2. Experimental Framework and Data

In this section we briefly describe the experimental framework. In order to promote research transparency and reproducibility the data described in this section is available at http://www.value-cost.eu/data. Further information on the VALUE experimental design is given in Maraun et al. (2015).

a. *Predictands: Local observations*

A subset of stations covering the different European climates and regions with a homogeneous density was selected to enable a comprehensive validation revealing relative strengths and weaknesses of different methods. To keep the exercise as open as possible, the downloadable (blended) ECA&D stations (Klein Tank et al. 2002) was selected and downloaded (on September 2014). A subset of 86 stations was selected with the help of local experts in the different countries, building on high-quality stations with no more than 5% of missing values in the analysis period (1979-2008); see `http://www.value-cost.eu/WG2_dailystations` for more details. The resulting set of stations is listed in Table 1 and graphically displayed in Figure 1. The Köppen–Geiger climate type (see, e.g. Kottek et al. 2006) shown in Table 1 has been calculated directly from the data using MeteoLab `http://meteo.unican.es/trac/MLToolbox/wiki`. Figure 1 shows the eight PRUDENCE (Christensen and Christensen 2007) sub-regions used to combine and summarize the individual validation results at a sub-regional level along the paper. These regions are British Isles, Iberia, France, Central Europe, Scandinavia, the Alps, the Mediterranean, and Eastern Europe.

The resulting dataset (including daily data for precipitation and temperatures for the analysis period) is publicly available in text (csv) format at `http://www.value-cost.eu/data`.

b. *Reanalysis Predictors*

ERA-Interim (Dee et al. 2011) was selected by the CORDEX initiative as the reference reanalysis for the coordinated downscaling experiments. Therefore, in order to be aligned with this initiative, VALUE also used ERA-Interim to drive the experiment with "perfect" predictors. Although reanalysis uncertainty has been recently reported as an additional source of uncertainty for statistical downscaling (Brands et al. 2012), the effect on the downscaled results is relevant only in the tropics (Manzanas et al. 2015). Therefore, this factor was not tested in VALUE.

In order to keep the experimental framework as controlled as possible and to facilitate the work of the contributing groups, we generated a reference predictor dataset downloading ERA-Interim data from ECMWF's MARS. The dataset includes a reduced number of commonly used predictors, degraded to a common 2° grid and post-processed by computing daily means from the original 6 hourly fields when required (see Table 2). This reference dataset includes most of the circulation and thermodynamic predictors at different pressure levels (including some surface predictors), typically used in downscaling applications in different European regions (Huth 1999; Benestad 2002; Huth 2002; Timbal et al. 2003; Huth 2005; HanssenBauer et al. 2005; Gutiérrez et al. 2013; Hertig et al. 2014; San-Martín et al. 2017), excluding redundancy as much as possible. For instance, vorticity and divergence have been considered as potential predictors in the literature (see, e.g. Hessami et al. 2008), but they were excluded from the standard set since they reported similar results to geopotential or wind directions in some studies (Gutiérrez et al. 2013). However, some contributors used in-house ERA-Interim datasets instead, for convenience or because they needed extra predictors (see Sec. b for more details).

Since MOS methods typically work with the direct model output at the nearest gridbox to the target station, we also compiled surface precipitation (PRC) and minimum (TMIN) and maximum (TMAX) temperature from the original ERA-Interim dataset at 0.75°. In order to illustrate the effect of the model resolution on the results, in the analysis we will consider raw ERA-Interim outputs at two different resolutions: 2 and 0.75° (hereafter referred to as ERAINT-200 and ERAINT-075, respectively).

c. *RCM Predictors (for MOS methods)*

Since MOS methods are typically applied to both GCM and RCM outputs, a second (optional) predictor dataset for MOS methods was produced considering daily surface precipitation (PRC) and minimum (TMIN) and maximum (TMAX) temperature from a state-of-the-art RCM (the RACMO2 model) driven in climatic mode by ERA-Interim (see Meijgaard et al. 2012, for a detailed description of the model). This simulation was produced in the framework of the EURO-CORDEX project (Jacob et al. 2014) using 40 hybrid coordinate full vertical levels on a regional $0.11°$ domain over Europe. RACMO2 ranked among the best performing RCMs over Europe in this reanalysis-driven experiment (Kotlarski et al. 2014).

The MOS techniques contributing to this additional experiment are indicated with a check mark in the second column (labelled as 'R') of Tables 3 and 4. This experiment will allow analyzing the advantages and shortcomings of these methods when downscaling to finer spatial scales. Note that a weak day-to-day correspondence with observations is expected for the RACMO2 outputs, since temporal synchrony with observations is only induced by the boundary conditions with prescribed reanalysis values. This must be taken into account when analyzing the results of non-distributional MOS methods, which are better suited for nudged climate simulations with a strong synchrony with observations (see, e.g. Eden et al. 2014).

d. *Cross-validation Approach*

In order to appropriately assess and compare the performance of different downscaling methods with "perfect" predictor data, we applied a cross-validation approach to avoid model overfitting and artificial skill. Cross-validation allows us to test whether the relationship established between predictor and predictand remains valid outside the training period (e.g. in a test period). The most popular and simplest of these approaches is data splitting, which considers independent data for training (e.g. 80% of the available data) and validation (e.g. the remaining 20%). However, this can yield spurious effects due to the particular partition performed. K-folding methods attempt to produce a more rigorous validation through the use of multiple calibration/validation period combinations. This is done by partitioning the available data ($n = 30$ years in our study) into k non-overlapping "folds" or subsets, each containing n/k elements. The downscaling methods are then calibrated and validated k times, considering in turn each of the folds as a test set and training the method with the remaining $k - 1$ ones. The resulting k test series are typically joined and validated together in a single series spanning the whole analysis period. This approach also permits analyzing the variability of the k validation results and estimating confidence intervals for model performance (see, e.g. Gutiérrez et al. 2013).

In general, the selection of k is subject to a number of factors depending on the particular application. In the present case, low k values result in longer validation periods which may be desirable to better characterize model performance, but limit at the same time the data available for training which may only capture part of the climatological distribution. As a compromise, five folds (5-fold cross-validation) were considered, each containing 6 consecutive years (1979-1984, 1985-1990, 1991-1996, 1997-2002, 2003-2008) for validation. All contributed methods followed this approach and joined together the results downscaled for the five test periods into a unique series —covering the whole thirty-years period— which was uploaded to the VALUE validation portal and automatically validated to assess model performance. More details are given in Sec. 6.

e. *Validation Measures*

We analyze minimum and maximum temperatures (TMIN and TMAX) and precipitation (PRC). For the latter we consider separately occurrence and amount, analyzing the variables R01 (Relative wet-day frequency, PRC $>=$ 1mm) and SDII (mean wet-day precipitation, a.k.a. Simple Day Intensity Index), although we also consider the total precipitation amount (PRCTOT) in some illustrative cases. In this paper we focus on general validation aspects involving the observed and predicted marginal distributions. In particular we validate the biases in the mean and the standard deviation, although additional results on distributional similarity (Kolmogorov-Smirnov and Cramer-von Mises tests) have been computed (not shown) and are available through the VALUE portal `http://www.value-cost.eu/validationportal` for further research. The bias in the mean is computed as the difference (for TMIN, TMAX) or ratio (for R01, SDII and PRCTOT) between the downscaled and the observed mean values, whereas the bias in the standard devidation is always obtained as the ratio.

Moreover, in order to analyze the strength of the daily predictor-predictand link (informative for non-distributional MOS and PP methods), we computed the correlation of the daily downscaled and observed series (using the ranked Spearman and Pearson correlations for precipitation and temperatures, respectively). Further validation analyses of aspects such as the representation of the temporal structure, extremes, key processes and multivariate relationships, are analyzed in detail in separate papers of this special issue.

3. Downscaling Approaches and Methods

a. *Description of Contributing Methods*

Tables 3 and 4 show the statistical downscaling methods contributing to this work for precipitation and (minimum and maximum) temperatures, respectively, under the same experimental framework (see Sec. 2). This constitutes the largest and most diverse ensemble of ESD methods analyzed to date, with a total of 45/49 methods for precipitation and temperatures, respectively (28 methods have been applied to both precipitation and temperatures, shaded areas). The detailed description of each of these methods and the implementation details for reproducibility (when available) are given in Annex 1.

These methods are first organized according to the three main approaches: MOS, PP and WGs — conditional WGs (including some model predictor) are listed under the corresponding PP or MOS category, depending on how they are calibrated.— Note that the first three rows indicate the raw model data (RAW) for ERA-Interim (both at $2°$ and $0.75°$ resolution), and for the (ERA-Interim driven) RACMO2 RCM at $0.11°$ resolution. As a second categorization, the methods are organized within each approach according to the families of techniques used, typically transfer functions (TF), analogs (A), and weather types (WT) for both PP and MOS, and additive/multiplicative scaling (S), parametric quantile mapping (PM), and empirical quantile mapping (QM) for MOS methods. This organization groups together similar methods (same approach and technique) and allows for a better intercomparison of model results (this order will be used in all figures in this paper). These families are described in further detail below.

Tables 3 and 4 also provide some metadata information about the structural properties of the methods (full metadata is available in the VALUE validation portal, `http://www.value-cost.eu/validationportal`). In particular the column 'ST' indicates the stochastic or deterministic nature of the method ('yes' for stochastic ones, which contributed 100 realizations for the validation process). 'MS' and 'MV' indicate whether the methods are suitable for multi-site and multi-variable problems, respectively; those methods using Principal Components (PCs) as predictors are marked as *'yes'* in the 'MS' column to indicate that some spatial coherence could be imprinted by the predictors. Finally, 'SE' and 'AC' indicate the explicit inclusion of seasonal and autocorrelation model components, respectively. The former is typically achieved by training the models separately for each of the calendar months (or with a 30-day moving window in some MOS methods, see Annex 1 for details). The latter is typically achieved using first-order Markov chains (conditioning the prediction to the previous predicted value) and has only been used in the contributed WG methods. As a result, the temporal structure of all PP and MOS methods in this experiment is driven by the particular model predictors used, i.e. directly from the raw model precipitation and temperature series for MOS methods. This (metadata) information must be taken into account when comparing the evaluation results of different methods, since a method can exhibit good performance for a particular aspect as a result of model construction or fitting (an interesting discussion on fair comparison is given in Casanueva et al. 2016a).

The participating MOS methods (#4-25, #4-23, for precipitation and temperatures, respectively) comprehensively span the range of widely used methods, from simple local scaling methods (labelled as 'S'), to standard parametric ('PM') and empirical ('QM') quantile mapping techniques. More specific BC methods, such as the trend preserving ISI-MIP bias correction methods, or a circulation-conditioned quantile mapping method (EQM-WT) are also included in this study. These methods are usually referred to as distributional MOS methods, in order to remark that they work by transforming the distribution of daily model outputs (the whole distribution or some statistics) towards the observed one. Moreover, the analysis also covers some more experimental recent MOS developments (#21-23, #22) such as stochastic regression (Wong et al. 2014) and analog- and regression-based MOS methods (Turco et al. 2011, 2017), which exploit the (weak) temporal correspondence existing in climatic RCM simulations to establish a link with observations. All the MOS methods in this study are single-site and single-variable, with the exception of MOS-AN and DBS, respectively, the latter providing inter-variable consistency by downscaling temperature conditional on the wet/dry state of the corresponding precipitation series. Finally, two particular methods for precipitation (FIC02P and FIC04P, #24-25) are based on a sequential application of PP and MOS techniques —the input to these methods is the output of the corresponding FIC01P and FIC03P PP results.— Note that these methods are not directly comparable with the rest of MOS techniques, but provide valuable information on the potential added value of mixed downscaling approaches (e.g. applying BC methods to correct systematic biases of PP outputs). The same situation occurs for temperatures with FIC02T, which takes as input FIC01T PP results.

Empirical quantile mapping techniques ('QM') constitute the largest family in this approach, with over ten contributing methods. It is important to remark that some methods are slightly different implementations of the same basic technique (with different number of adjusted percentiles, or extrapolation options; see Annex 1 for details). In particular the empirical methods EQM, QM-DAP, EQM-WIC658, and QMBC-BJ-PR are slightly different versions of the standard empirical quantile mapping approach (see, e.g. Déqué 2007). Parametric quantile mapping techniques ('PM') mainly differ in the distribution function(s) used to calibrate the data. For instance, in the case of precipitation, a gamma distribution is used in EQM, a double-gamma is used in DBS and in Ratyetal-M9 —which is a simplified version of the former,— the optimum among five distributions is used in BC and, finally, gamma and generalized Pareto are used in GPQM to adjust separately the extremes values. It is also important to notice that most MOS methods have been trained separately for each month (or considering a moving window), with the exception of GQM, GPQM, EQM, EQM-WT, and the MOS-GLM/REG/AN family. Therefore, interesting conclusions could be obtained by comparing the results of QM methods taking into account the different configurations and implementations.

The participating PP methods (#26-42,#24-46, for precipitation and temperatures, respectively) broadly represent the most popular and widely used families of techniques —analogs (A), transfer function / regression (TF) and weather-type (WT) methods— in fairly standard implementations in most of the cases. The analog techniques are the only multivariate methods (multi-site and/or multi-variable). However, some of the TF methods use PCs as predictors (Sec. b), which may provide some imprinted spatial inter-consistency due to their spatial character. Those cases are indicated with a '*yes*' multi-site code in Tables 3 and 4. Nonparametric regression methods (e.g. neural networks) are among the most notorious missing families in this study. In some studies these methods have shown to outperform linear models (see, e.g. Gaitan et al. 2014), but there are also studies showing the opposite. Therefore, the VALUE intercomparison framework could provide a better understanding on the added value and limitations of these techniques. The only contributing machine learning technique is the MO-GP method, which applies genetic programming to obtain general symbolic regression equations from data. Therefore, an interesting follow-on of the project would be including new nonlinear machine learning methods in the intercomparison.

The family of Analog (A) methods includes two different variants of the standard technique, considering raw fields with no seasonal restriction (ANALOG), and anomalies with seasonal restriction (ANALOG-ANOM). FIC and ANALOG-MP/SP methods are more elaborated two-step analog methods considering nested global/local domains and predictors. ANALOG-MP/SP are probabilistic methods which include here a stochastic component to produce the 100 realizations of the predictand from the probabilistic prediction available each day. Similarity is quantified by Euclidean distance in all cases with the exception of ANALOG-MP/SP, which use the Teweless-Wobus score. WT-WG is a simple stochastic weather typing approach simulating temperature/precipitation from gaussian/binomial-gamma distributions within each weather type (obtained using only SLP in this study).

The contributing Transfer Function (TF) methods are different variants of Multiple Linear Regression (MLR) techniques and Generalized Linear Models (GLM) and Vectorized GLMs for precipitation. GLMs are an extension of linear regression allowing for non-normal predictand distributions (see Chandler 2005, for an introduction), which have been used for downscaling precipitation in a number of studies (see, e.g., Chandler and Wheater 2002; Abaurrea and Asín 2005). Although MLR has been applied to downscale daily precipitation in previous studies (see, e.g., Hessami et al. 2008; Chen et al. 2014; San-Martín et al. 2017), these techniques are not suitable to model daily precipitation, even after transforming precipitation —using e.g. squared or cubic root values— to make the data more normal. However, we have included them in the present work for illustrative purposes, in order to highlight the associated problems. The different MLR and GLM methods (#26-42, #24-46) have been trained on a daily basis to establish the link with local data —with the exception of the ESD family (#39-42, for temperatures, in italics), trained using monthly aggregated data and providing monthly values.— ESD methods are used here for illustrative purposes and results are only shown for suitable validation scores (mean bias).

Most of the TF methods are deterministic, but there are also some stochastic implementations. A particular method is provided in both deterministic (GLM-DET) and stochastic (GLM) variants, using the expected value in the former case and simulating from the resulting binomial/gamma in the latter, including also an implementation conditioned on weather types (GLM-WT). GLM-P combines logistic (binomial GLM) and exponential regression to simulate occurrence and amount, but only considers a stochastic version of the former one (i.e. only occurrence is simulated). A more sophisticated Vectorised GLM method (Vaittinada Ayar et al. 2016) is used in SWG, which also simulates the predicted values from the resulting conditioned binomial/gamma distributions. On the other hand, stochastic MLR versions (MLR-ASW/AAW) are based on variance inflation using white noise. Note these methods can be compared with the simple deterministic scaling variance inflation versions (MLR-ASI/AAI) in order to analyze the benefit of the stochastic component.

The participating WG methods (#43-48, #47-52, for precipitation and temperatures, respectively) include variants of the Richardson model (Richardson 1981), simulating daily time-series of precipitation, minimum and maximum temperature using Markov chains (order one for SS-WG and one to three for MARFI) and autoregressive models. Moreover, the analysis also covers a recent non-parametric weather generator (GOMEZ) based on nearest neighbor resampling.

b. *Selection of Predictors and Data Preparation*

Selection of predictors and data preparation is a key task for statistical downscaling, in particular for PP methods. Whereas this task is quite simple for distributional MOS methods —which operate directly with model precipitation/temperature, typically on the nearest gridbox, as the single predictor,— the selection of informative predictors for PP methods is a key factor both for model performance and for ability to extrapolate under climate change conditions (Huth 2004; Gutiérrez et al. 2013; San-Martín et al. 2017). Therefore, a region-dependent screening of suitable predictors over different (large or small) domains covering the area of study is usually performed as a first step of the downscaling process. In some cases, this task is automatically performed applying some variable selection method, such as stepwise screening, which is applied in most TF methods as described in Annex 1 (e.g. MLR-RSN/RAN/AAN/AAI/AAW/ASI/ASW). Therefore, the final set of predictors used in these methods may change from variable to variable and from station to station.

Several studies have shown the convenience of combining circulation and thermodynamic predictors in order to include signal-carrying predictors linked to changes in the radiation budget, avoiding to model future climate from changes in circulation alone (Wilby et al. 1998; Huth 2004). Therefore, the final decision about the predictors to be used in a particular region needs to be based on the physical understanding of the problem. The predictors must also be skillfully predicted by GCMs in terms of the statistical characteristics of the large scales (e.g., spatial and temporal structures). Ideally, they should also exhibit a strong link with the local variable in order to represent the large-scale dependency.

Table 5 shows the particular combinations of predictors used by the different PP methods in this study (Tables 3 and 4). Besides the standard variables shown in Table 2, some contributors have considered additional predictors, such as ten meter zonal and meridional wind direction (U10, V10), two meter dewpoint temperature (TD), vertical velocity (VV), relative humidity (R), or thickness between two pressure levels (TH). Only a few methods (WT-WG, FIC01P and the ESD family) use either circulation or thermodynamic predictors alone, whereas the rest of methods build on combinations of circulation predictors and middle-troposphere temperature and/or humidity, which have been found among the best predictors for temperatures and precipitation, respectively. Note that the predictors are inhomogeneous (i.e. not directly related to the target variable to be downscaled) in all cases for precipitation, and in most of the cases for temperature (with the exception of those including $T2$ as predictor which may be considered an homogeneous predictor for minimum and maximum temperatures).

Table 5 also shows the size of the domain used to define the predictors, ranging from continental scale, to smaller national-wide domains, and to local information at the nearest gridboxes (or combinations of them). Note that most of the PP methods consider national or continental-wide information since the use single gridbox information for large-scale variables is not recommended due to the minimum skillful scale of climate models (see, e.g. Takayabu et al. 2016, for more details). The resulting data (values of the predictors for multiple gridboxes) is preprocessed in different ways before using it to train the downscaling methods. Ten regression methods applied individual or combined EOF analysis to obtain PC predictors in order to reduce the spatial redundancy. The rest of regression methods consider raw, standardized, or anomaly point-wise values and in most of the cases they apply a step-wise procedure for predictor selection (see Annex 1 for details). In this case, the models resulting for different stations may be based on different (local) predictors, normally at gridboxes close to the particular station. This constitutes a key factor when validating spatial aspects of the predictions (see Widmann et al. 2017, in this special issue).

Most analog methods consider national-wide information (with the exception of one, which is a applied at a continental level) and use raw data, anomalies, standardized values or PCs to compute the similarity of different fields. Four of the methods are two-step implementations which consider different large-scale and local predictors in nested national- and gridbox-wide domains, respectively.

Finally, we want to remark that although the ESD family of methods is based on common EOFs (both reanalysis and GCM fields are used to compute the EOFs, Benestad et al. 2008) the approach used in this paper applies standard EOFs obtained from ERA-Interim.

c. *Strength of the Predictor-Predictand Link*

PP and non-distributional MOS methods build on a synchronous daily link established between predictor(s) and predictand in the training phase. The strength of this link indicates the local variability explained by the method as a function of the large-scale predictors. In order to provide a quick diagnostic of this strength for the different methods, Figures 2 and 3 show the daily Spearman and Pearson correlations for the downscaled and observed daily precipitation and maximum temperature values, respectively. The results for the raw model outputs (indicated as ERAINT-200, -075 and RACMO22E) are included in the first three columns of the figures and show the comparison with the local observations considering the raw model values at the closest gridbox. Note that these figures are only informative for PP and non-distributional MOS methods since, on the one hand, WG methods have no daily correspondence with the observed data —they are purely stochastic and use no model predictors— and, on the other hand, distributional MOS methods broadly preserve the temporal structure of the raw model predictor. Therefore, distributional MOS and WG results are included in the figures for illustrative purposes, in order to contrast the expected results and to identify potential problems.

As expected, distributional MOS methods closely reproduce the correlation of the corresponding model predictors in most of the cases. The most notorious deviation is the EQM-WIC658 model, which in principle is similar to other implementations of the empirical quantile mapping (e.g. EQM) and therefore is suspected of having an error. There are also noticeable differences for the CDFt model (the case using ERA-Interim, particularly for temperature), which may be due to the particular approach followed to correct the data (see Annex 1 for more details) or to a problem with temporal arrangement of the downscaled data. Furthermore, the ISI-MIP model exhibits smaller correlation than the raw model output for precipitation, which could be explained by the two step process followed, adjusting first the monthly values and then the daily residuals. Finally, on the other hand, WG methods exhibit close to zero correlations in all cases, as expected.

Note that the RACMO2 model (and the MOS results obtained using this predictor, with gray shading in the figures) show smaller correlations than ERA-Interim —which exhibits similar results for the two resolutions considered.— This is due to the climatic nature of the simulation, since day-to-day correspondence with observations is only prescribed at the boundaries of the regional simulation domain. Therefore, the non-distributional MOS analog (MOS-AN) and transfer function methods (MOS-GLM/REG, VGLMGAMMA) —which exploit the (weak) temporal correspondence existing between RACMO2 outputs and observations— exhibit smaller correlations than the PM and QM techniques. When applied to ERA-Interim, these techniques result in similar (MOS-GLM/REG), or even higher (MOS-AN), correlations when compared with their PP counterparts (GLM/MLR and ANALOG, respectively). Note that in this case, the methods are in fact homogeneous (using precipitation or temperature as predictor) versions of the PP methods (considering a single gridbox instead of PCs in the case of MOS-GLM/REG). The higher correlation of the MOS-AN for precipitation in this case is explained by the use of model precipitation as single predictor (Widmann et al. 2003), which is superior to the predictors used by the PP ANALOG version (see Table 5). However, differently to the analog MOS version, the MOS regression methods (MOS-GLM/MLR) result in very small correlations when applied to the weakly synchronized RCM outputs. This may be due to the use of a single gridbox, more sensitive to the weak temporal correspondence with observations.

The range of correlations corresponding to the PP methods are mainly due to the different predictor settings used and to the deterministic/stochastic character of the methods. For instance, the stochastic versions ANALOG-MP/SP, VGLMGAMMA, GLM-P, GLM, GLM-WT, WT-WG and SWG exhibit smaller correlations due to the stochastic component. Moreover, linear regression methods using white noise variance correction (MLR-ASW/AAW) exhibit smaller correlation than the standard (MLR-AAN) or the inflation variance correction (MLR-ASI/AAI) implementations (see Annex 1 for details). It is noticeable that the stochastic GLM method still preserves a strong correlation when compared to the deterministic implementation (GLM-det), indicating that most of the information given by the predictors is still retained in the stochastic implementation. Regarding the analog methods, the smallest correlations are obtained with the method using anomalies (ANALOG-ANOM). Finally, the last two PP methods (WT-WG and SWG) exhibit low correlation values, since they have been designed to have a strong stochastic component weakly forced by the predictors. In particular, the correlation of the WT-WG method is similar to that of the un-conditioned WGs, indicating that this method is purely stochastic (the weather types obtained solely from SLP do not play a relevant conditioning role in this case). Therefore, they can be thought of as weather generators weakly conditioned on circulation.

In the case of maximum temperature, high correlations are obtained in general in all cases. The different correlations observed in the linear regression methods are mainly explained by the different predictor settings used. In particular, those methods including the "homogeneous" predictor two-meter temperature (MO-GP, MLR-T, MLR, MLR-WT) exhibit larger correlations (particularly during winter), due to the stronger connection of this predictor with local surface temperature. Note that this is not an indication of better performance of the model for climate change applications, since upper-air predictors may be more robust.

Finally, regional and seasonal differences are observed in the link strength when looking at the results aggregated over the eight Prudence regions considered (shown by the colored bars in Figs. 2 and 3). For precipitation, both MOS and PP methods mostly preserve the rank of ERA-Interim (and RACMO, for MOS) regional results for the different seasons. In summer the highest correlations are obtained for the Alps and the weakest in Mediterranean and Iberian Peninsula regions, whereas in winter correlations are larger in Central Europe and smaller in Eastern Europe and British Isles. For the case of maximum temperature, correlations are higher for Eastern and Central Europe and smaller for Iberia and the British Isles (for summer) and Iberia and the Alps (for winter). There exist also some cases where the PP methods show some differences with respect to the ERA-Interim and MOS results. For instance, the Alps are among the regions with highest winter correlations for PP methods, contrary to the ERA-Interim results. In other cases, PP methods enlarge the regional variability of results. For instance, the regional correlations of summer maximum temperatures have larger spread for PP methods, mainly due to the small correlation obtained for the British islands, particularly for regression methods.

4. Validation Results for Precipitation

In this section we present the first validation results obtained for precipitation, focusing on general marginal distributional aspects. Figures 4 and 5 show the relative biases for R01 (relative wet-day frequency) and SDII (mean wet-day precipitation), respectively (predicted over observed mean values).

The results for the raw model outputs (indicated as ERAINT-200, -075 and RACMO22E) are included in the first three columns and show the comparison with the local observations considering the raw model values at the closest gridbox. Model outputs tend to overestimate wet-day frequency and underestimate precipitation amount. Moreover, the resulting biases decrease when increasing the resolution, being largest for the 2.0° ERA-Interim and smaller for the 0.11° RACMO2. Overall, most of the downscaling methods greatly improve model biases (both for ERAINT-200 and RACMO2 predictors, the latter shaded in the figures) and no downscaling approach or technique seems to be superior in general. An exception is found for the four linear regression methods (from MLR-RAN to MLR-ASI), which exhibit very large biases, even larger than those corresponding to the raw model outputs. A similar behavior is also found for the GLM-P method, with larger biases than the rest of GLM implementations. These results clearly illustrate the inadequacy of linear regression methods for downscaling precipitation values. Note that the nonlinear regression method (MO-GP) presents smaller biases, though still larger than for the rest of methods. On the other hand, GLMs exhibit small biases, particularly in the vectorized versions (VGLMGAMMA, SWG) and the version conditioned to weather types (GLM-WT), all including some sort of seasonality, either imposed by training the model separately for each month, or indirectly conditioning the model to twelve different catalogues of weather types. On the other hand, the different analog implementations present similar biases, with exception of FIC01P (using only geopotential fields) which exhibits larger biases for winter SDII. Moreover, in this case, training the methods separately for each month/season does not clearly improve the results, since the analog method trained with year around data (ANALOG) exhibits similar biases than to the rest of (seasonally trained) analog implementations. This different behavior may be a consequence of the fact that, as opposed to regression methods, analog methods do not explicitly calibrate the mean value towards the observations.

Regarding the MOS methods, similar results are obtained for the different families of techniques, although there is an outstanding group of methods with very small biases, formed by the empirical quantile methods (QM) including a seasonal component (see Table 3) —with the exception of CDFt, which systematically overestimate precipitation intensity.— The key role of the seasonal component can be seen comparing EQM and EQMs methods, only differing in the 31-day moving window used to train the latter. Therefore, similarly to the regression techniques (TF), seasonal calibration is beneficial for QM methods. On the other hand, the results are similar for the two predictor settings —ERA-Interim and RACMO2— with slightly smaller (and more centered) biases for the latter, particularly for scaling and parametric quantile mapping methods. Those methods with no seasonal component (e.g. GQM, GPQM, EQM) show compensating DJF and JJA biases. It is also interesting to note that the particular FIC02P/04P methods (which apply the parametric BC quantile method to outputs from FIC01/03, respectively) improve the performance of the corresponding PP counterparts (see, e.g. wet frequency for FIC01P) and also show better results than the direct application of the BC method to ERA-Interim. This indicates that the PP method (applied to ERA-Interim) produces more realistic local precipitation results, well suited for a parametric correction.

The WG techniques exhibit small biases, with the exception of the MARFI family which systematically over- and under-estimate wet-day frequency and amount, respectively.

When looking at the regional variability of results (horizontal color bars in Figures 4 and 5), there is a high method-to-method regional variability. The largest/smallest biases are found in the Mediterranean/British Isles for ERA-Interim. However, these regional differences are greatly reduced in all downscaled results, although some methods still exhibits large biases in the Mediterranean region during Summer. Figure 6 shows the individual station results for winter (DJF) and summer (JJA) for PRCTOT (total precipitation) relative biases, with southernmost stations at the bottom and northernmost stations at the top. This figure shows systematic bias patterns across stations for each particular methods (greeny or browny vertical bars), although there are some stations where most of the downscaling methods exhibit a similar systematic bias. For instance, most of the methods overestimate total precipitation for station number 12 (Roma-Ciampino).

Finally, Figure 7 shows the results of the relative biases for the standard deviation of daily precipitation (downscaled over observed standard deviations), manifesting the deficiencies already reported for some of the methods (e.g. CDFt and the linear regression family). Among the MOS methods, some techniques tend to systematically underestimate (e.g. ISI-MIP) or overestimate (DBS, Ratyetal-M9) variability and, again, the best performing methods are QMs including a seasonal component. As opposite to the bias in the mean, larger biases are found here for RACMO2 than for ERA-Interim downscaled values for some of the MOS methods. Regarding PP methods, analog techniques tend to systematically underestimate variability. It is also shown that regression-based deterministic methods (including GLM-det) can only reproduce a small part of the observed variance. Moreover, as opposite to the two GLM stochastic versions (GLM and GLM-WT), the MLR stochastic versions (GLM-P and MLR-ASW, the former simulating precipitation occurrence and the latter inflating with white noise) fail to recover the observed variability. Again, FIC02P/04P methods seem to correct the deficiencies of their PP counterparts.

5. Validation Results for Temperatures

Figures 8 and 9 show the results for the mean biases (downscaled minus observed mean values) of daily maximum and minimum temperatures, respectively. The raw model outputs from ERA-Interim largely under/over estimate maximum/minimum temperatures in almost all regions, whereas RACMO2 exhibits smaller biases but still with a large regional variability (in this case, the model tend to underestimate both minimum and maximum temperatures). Overall, most of the downscaling methods greatly improve model biases and again no downscaling approach or technique seems to be superior in general. There are a few methods exhibiting very large biases (WT-WG for both TMAX and TMIN, and CDFt, MLR-ASW for TMIN) which indicate some problem with the configuration of the method or with the particular execution. Moreover, the MOS-REG regression method exhibits large biases when applied to RACMO2 model (much larger than when applied to ERA-Interim). Therefore, in this case the synchrony of the climatic run with observations is too weak to allow for a suitable implementation of this type of MOS regression technique (at least considering information only on the nearest gridbox). Moreover, the methods SB and EQM-WIC658 exhibit large biases (particularly during winter) which can not be explained from the definition of the method (other similar techniques show small biases) and could be an indication of some problem in the application of the method.

In general, the family of methods exhibiting larger biases are the analog techniques, but this could be explained because they do not explicitly calibrate the mean during training. Among these methods, the best results are obtained with ANALOG and ANALOG-SP, both using 2m temperature as predictor, which seems to have an important role in this case. However, the particular choice of the predictor cannot explain the differences among the regression techniques. Similarly to the case of precipitation, a key factor explaining the variability of MOS results is the seasonal training of the methods (e.g. EQM vs EQMs). Note that the cross-validated mean bias for simple linear scaling methods (additive and/or multiplicative; e.g. RaiRat-M6 and RaiRat-M7) should be zero by construction (in cases with no missing data). In this work, the seasonal cross-validated results obtained for these methods are different from zero (although very small) due to the two-month moving window used to compute the scaling factors (see details in Annex 1, describing the methods).

The above figures do not show relevant regional differences for the biases of the downscaled methods with the exception of the analog methods where the regional biases observed seem to be related to the predictors used. Figure 10 gives further information showing the individual station results (sorted as in Table 1) for daily maximum temperature for winter (DJF, top) and summer (JJA, bottom). Besides of revealing the bad performing methods, these figures show the systematic biases exhibited for the methods to under/over estimate across the different stations (e.g. the analog methods).

Finally, Fig. 11 shows the results for the standard deviation (relative biases, downscaled divided by the observed daily standard deviations) of daily minimum temperatures during winter and daily maximum temperatures during summer. This figure allows to clearly differentiate those MOS scaling methods not correcting the variance of the downscaled results (RaiRat-M6/M7, SB, ISI-MIP), which show the same biases as the input reanalysis or RACMO2 models. Moreover, as expected, those quantile mapping techniques trained annually exhibit larger biases on the seasonal variances than those seasonally trained (see, e.g. EQM and EQMs). These two factors explain most of the variability of the MOS results (together with the model deficiencies already reported). Regarding the PP methods, the analog techniques tend to systematically underestimate the variance, with FIC01T being the worst of this group. Note that the results of this method are later used as input for the FIC02T, which correct for this deficiency applying a quantile mapping approach. In the case of the linear regression methods, all deterministic implementations underestimate the observed variance, in correspondence with the daily correlation values shown in Figure 3. However, as expected, those methods correcting the seasonal variance (using inflation and white noise, as in MLR-ASW and ASI, respectively —note that a problem was already reported for MLR-ASW results for minimum temperature—) show more centered results (although there are some outlier stations where the variance is over-estimated). Note that variance correction at an annual basis (e.g., MLR-AAI/AAW) yields seasonal biases comparable to other deterministic linear regression implementations. Finally, the deterministic symbolic regression method MO-GP is a multi-objective method optimizing several statistics, including standard deviation. As a result it exhibits smaller seasonal variance biases (even though it is trained on an annual basis), and larger mean biases (see Figures 9 and 8), than the regression methods. This could be beneficial for climate change applications since it requires no postprocessing but, again, a more comprehensive assessment of other aspects is needed.

6. Promoting Transparency and Reproducibility of Results

Research transparency and reproducibility is a major concern in the different experimental disciplines (see a string of freely available nature articles on reliability and reproducibility of published research at `http://go.nature.com/huhbyr`). For instance, a recent survey over 1500 scientists recently reported by Baker (2016) revealed that the work published in different research fields (including Earth and Environment) were mostly not reproducible (over two-thirds). As a result, there is growing alarm about results that cannot be reproduced. In VALUE substantive steps were taken in order to improve transparency and reproducibility of results, and to promote awareness within the downscaling scientific community.

The main difficulties for research reproducibility identified include 1) access restrictions to raw input data and/or results, 2) poor experimental design information, 3) lack of code availability, and 4) incomplete documentation of the particular configuration and implementation used (data preprocessing, method configuration and specific parameter values, training options, etc.). In some cases, the steps involved in the downscaling process are very technical and they are not always appropriately documented in practical applications, thus making difficult the reproducibility of the results.

The following actions have been undertaken in VALUE in order to avoid the above mentioned problems:

- 1) All the data needed for the experiments described in this paper has been collected and made available at `http://www.value-cost.eu/data`. Moreover, the daily downscaled data and the resulting validations for each of the methods and experiments are also publicly available under the liberal Creative Commons Attribution (CC BY) License (http://creativecommons.org/licenses/by/4.0). Contributors also get access to VALUE_PRIVATE published data, which is made available internally to the consortium for quality check and verification purposes before publication (more info in `http://www.value-cost.eu/terms`).

- 2) The experimental framework was designed and published (Maraun et al. 2015) in advance of the open call for contribution to this validation experiment.

- 3) The code used for the validation framework (from data loading to computing all the validation measures) has been coded in R and is publicly available from `http://github.com/SantanderMetGroup/`

R_VALUE. In addition, the packages and/or code needed to reproduce the results for some of the downscaling methods are publicly available (see Annex 1). Other methods use proprietary software and cannot be replicated; however, we decided to also include this information to favor method inter-comparability, but requiring the open publication of the results (both predictions and validation results), which was mandatory for all methods contributing to this paper.

- 4) Furthermore, a metadata description vocabulary for statistical downscaling methods has been defined and implemented in the VALUE Validation Portal, providing information on the method characteristics and implementation details needed to properly analyze the results.

Finally, transparency is further promoted by the VALUE Validation Portal http://www.value-cost.eu/validationportal, providing public access to the metadata and data for all contributing methods and also for all validation results, as well as tools to filter and visualize (both in tabular and graphical formats) the results.

7. Conclusions

In this paper we present the ensemble statistical downscaling methods produced in the VALUE collaboration, which covers the three common downscaling approaches (perfect prognosis, model output statistics ?including bias correction? and weather generators) with a total of over fifty downscaling methods. We also present the first results from the inter-comparisson experiment under the same cross-validation experimental framework using "perfect" predictors. Additional experiments with GCM data will follow to contribute to the EURO-CORDEX initiative. Appropriate metadata on the main model characteristics (e.g. deterministic or stochastic nature) and implementation details (predictors, geographical domain, monthly/seasonal training, etc.) have been collected in order to properly analyze the results.

Overall, most of the downscaling methods greatly improve model biases and no downscaling approach or technique seems to be superior in general, due to the large method-to-method variability of results. Some bad performing methods have been identified as potentially failed methods due to different problems giving some clues about future quality checks to be implemented in the VALUE validation portal. Our results also show the inadequacy of linear regression methods for downscaling daily precipitation values, which is still used in some applications (see, e.g., Jeong et al. 2012; Chen et al. 2014). Regarding the MOS methods, empirical quantile methods including a seasonal component form an outstanding group of methods with very small biases. However, there are particular PP and WG methods with a similar performance. In this work we found that, in agreement with previous studies (Reiter et al. 2017), introducing a seasonal component (e.g. training the methods separately each calendar season, month or moving window) improves the results. However, we found that all implementations (even a daily moving window) resulted in a relevant performance improvement, differently to Reiter et al. (2017), where seasons were recommended for calibration. The deterministic or stochastic nature of the method was the most relevant factor (together with seasonal training) for explaining the variability of results for biases in the standard deviation.

In this work we have also tested some new experimental developments, such as stochastic and analog- or regression-based MOS methods, applied to RCM climatic runs driven by reanalysis. The results seem to be promising for precipitation but not for temperatures, apparently due to the weak synchrony between RCM outputs and observations. Some promising results have been also obtained when combining PP and MOS methods. In particular, parametric quantile methods are shown to produce better results when applied to the outputs of an analog method (using ERA-Interim predictor data), than when applied directly to ERA-Interim. This indicates that the PP method produces more realistic local precipitation results than ERA-Interim, well suited for a parametric bias correction. A similar result is obtained with MOS methods when applied to RCM climatic runs driven by reanalysis than to the reanalysis outputs directly. However, these first validation results should be interpreted with caution, since a good performance in terms of bias may not be an indication of a better performance of the model for climate change applications. Therefore, a comprehensive validation analysis of different aspects is needed in order to properly assess the performance of this technique (e.g. temporal and extreme aspects, as described in the companion papers of this special issue). Moreover, the present study cannot give a conclusive assessment of the skill of downscaling methods to simulate regional future climates, and further experiments (Maraun et al. 2015) will be soon performed in the framework of the EURO-CORDEX initiative, thus completing the analysis initiated in the present manuscript.

Finally, in order to favor research reproducibility, the experimental framework is precisely defined, all datasets needed for this experiment are publicly distributed `http://www.value-cost.eu/datasets` and, in some cases, the packages and/or code to reproduce the results are publicly available. A metadata description vocabulary has been defined and implemented in the VALUE Validation Portal `http://www.value-cost.eu/validationportal`, which provides metadata information for all contributing methods (approach, technique, predictors, method configuration, etc.). Transparency is also promoted by the VALUE Validation Portal `http://www.value-cost.eu/validationportal`, which provides public access to data and metadata information for all contributing methods and also for all validation results, as well as tools to filter and visualize (both in tabular and graphical formats) the results. In particular, most of the figures of the paper can be reproduced with these tools.

8. Acknowledgments

This work have been performed in the framework of the VALUE is funded via the EU COST Action ES1102, under FP7 programme. We thank the VALUE community for their input to this framework, helping in the coding the validation routines and helping to select the representative stations at a national level. In particular we would like to thank Andreas Gobiet, Constantin Mares, Robertas Alzbutas, Mandy Vlachogianni, Adam Jaczewski, Peter Thejll, Patrick Willems, Ivan Pilaš, Meron Teferi Taye, and Fredrik Boberg. We acknowledge the data providers in the ECAD project (data and metadata are available at http://www.ecad.eu) and ECMWF for allowing us to re-distribute ERA-Interim daily data (regridded at a 2° resolution) within VALUE (registration is public) for the standard set of predictors. We also acknowledge KNMI for making publicly available the RACMO2 0.11° resolution ERA-Interim driven simulations within the EURO-CORDEX initiative.

JMG and SH acknowledge partial funding from MULTI-SDM project (MINECO/FEDER, CGL2015-66583-R). BH and DR acknowledge COMPLEX project (FP7-ENV-2012. Number: 308601). MT was supported by HOPE project (MINECO, CGL2014-52571-R). Participation of MD and RH was funded by the Ministry of Education, Youth, and Sports of the Czech Republic, contracts LD12029 and LD12059, respectively.

The authors gratefully acknowledge helpful comments by the anonymous reviewers.

9. ANNEX 1. Description of methods

This annex includes the detailed description of the methods used in this work (Tables 3 and 4). They are organized alphabetically within each downscaling approach (MOS, PP and WG) in the following sections.

a. *MOS Methods*

- **BC (only precipitation):** Parametric bias-correction method using the optimum among five theoretical distributions (Gamma, Weibull, Classical Gumbel, Reversed Gumbel and Log-logistic, all of them with four parameters) for each station on a monthly basis (Monjo et al. 2014).
  Implementation: In-house R code.

- **CDFt:** The CDFt approach links the local-scale CDF of the variable of interest to the associated large-scale CDF through a "quantile-quantile" approach performed between the future large- and local-scale CDFs (and not between present CDFs as in the classical quantile-quantile method). To do so, the future local-scale CDF is first estimated based on the assumption of a mathematical transformation to link the evolution of the large-scale CDF to the evolution of the local-scale one. Hence, CDFt is a variant of quantile-quantile but CDFt accounts for the CDF changes from the calibration to the projection (or future) time periods (Vrac et al. 2012).

- **DBS:** Distribution based parametric quantile mapping (Yang et al. 2010, 2015). The cumulative distribution of precipitation and temperature are fitted by double-gamma and normal distributions, respectively. A wet-day correction is applied to precipitation series. In case of too many wet-days in the predictor data, all wet-days below a derived threshold are removed so that the wet-day frequency is the same as in the predictand data. In case of too few wet-days in the predictor data, the wet-day correction is done by adding wet-days to already existing wet-spell, starting with the longest ones. Temperature correction is done conditional on the wet/dry state of the corresponding precipitation series. The parameters were seasonally calibrated for every month in the annual cycle.
  Implementation: In-house FORTRAN code.

- **DBD/DBBC (only temperatures):** Bias is calculated separately for all percentiles (from 1 to 99) and a polynomial function of second degree is fitted as a function of the temperature values (DBD) or the probabilities (DBBC). In the validation period the model temperature (or the corresponding percentile) is used to calculate the bias to be subtracted in the adjustment process. The difference between DBD and DBBC is that the bias is connected with temperature and percentile values, respectively. In both cases, calculations are performed for each season separately.
  Implementation: In-house Matlab code.

- **EQM/EQMs/EQM-WT:** Implementation of Empirical Quantile Mapping (EQM) adjusting 99 percentiles and linearly interpolating inside this range every two consecutive percentiles; outside this range a constant extrapolation (using the correction obtained for the 1st or 99th percentile) is applied (Déqué 2007). In the case of the precipitation, when the predicted dry frequency is larger than the observed one the frequency adaptation proposed by Themeßl et al. (2012) is applied. In order to explicitly model the seasonal cycle, the variant EQMs considers a 31 day moving window centered on every calendar day to calibrate the method. EQM-WT is a state-dependent version of EQM, conditioning the training to 12 Weather Types defined using a k-means algorithm (k=12) applied to the daily SLP over Europe. For the experiment with the RACMO2 RCM predictors, SLP is taken from the RCM model and smoothed to a 1° resolution.
  Implementation: EQM is implemented in the *downscaleR* (Bedia et al. 2016) R package (*biasCorrection* function) with the options *method = "eqm"* and *extrapolation="constant"*, including *precipitation = TRUE* and *pr.threshold = 1* for precipitation. For EQMs the extra argument *window = c(30, 1)* was included. This package is freely available without restriction.

- **EQM-WIC658:** Implementation of the empirical quantile mapping method (Déqué 2007) sorting the values into bins with adjustable width (e.g. 0.1°) and applying a linear interpolation between two percentiles (bins); out of range values are adjusted using constant extrapolation (using the correction obtained for the minimum or maximum). In order to cope with the seasonal cycle, a 31 day moving window centered on every calendar day is used to calibrate the method. More details in Wilcke et al. (2013).

- **FIC02P/04P (only precipitation):** Parametric bias correction technique (method BC above) applied to FIC01P/03P results. The method is applied separately for each month (Monjo et al. 2014).

- **FIC02T (only temperatures):** Parametric bias correction technique considering Gaussian distributions applied to FIC01T results. The method is applied separately for each station and for each month (Monjo et al. 2014).

- **GQM/GPQM:** Gamma/Gaussian parametric Quantile Mapping (GQM) to approximate the empirical distribution of precipitation intensity / temperature. In the case of the precipitation, the frequency adaptation proposed by Themeßl et al. (2012) is previously applied to calibrate precipitation occurrence (a 1mm threshold is used). The Generalized Pareto Quantile Mapping (GPQM) version considers a Generalized Pareto to adjust separately the extremes values (over the 95th percentile) as in Gutjahr and Heinemann (2013).
  Implementation: GQM/GPQM are implemented in the *downscaleR* (Bedia et al. 2016) R package (*biasCorrection* function) with the options *method = "gqm"/"gpqm"*, including *precipitation = TRUE* and *pr.threshold = 1* for precipitation. This package is freely available without restriction.

- **ISI-MIP:** The trend preserving ISI-MIP method proposed by Hempel et al. (2013) in the framework of the Inter-Sectoral Impact Model Intercomparison Project (ISI-MIP). This method works in a two-step approach. First, the monthly mean is adjusted with a linear/multiplicative scaling and, then, the resulting daily anomalies are corrected by means of a parametric (gaussian/exponential) quantile mapping, for temperatures/precipitation, respectively. In the case of precipitation also a frequency adjustment is included for both the monthly and daily components. This method has been designed to simultaneously adjust groups of variables (precipitation-snow, temperatures, wind speed and components).
  Implementation: ISI-MIP is implemented in the *downscaleR* (Bedia et al. 2016) R package (*isimip* function). This package is freely available without restriction.

- **MOS-AN (only precipitation):** MOS implementation of the analog method considering precipitation as the single predictor, and trained across different zones (similar to the Prudence regions) computing similarity using Euclidean distances of the precipitation fields. As a benchmark this method has been applied directly to ERA-Interim precipitation, with "perfect" (day-to-day) synchrony with observations. When applied to the ERA-Interim driven RCM simulation, this method exploits the marginal temporal synchrony within the RCM domain given by the synchronous forcing at the boundary (Turco et al. 2011, 2017). Note that this method is best suited for for nudged RCM simulations.
  Implementation: In-house Matlab code.

- **MOS-REG/GLM (only temperatures/precipitation):** MOS implementation of linear (and generalized linear) methods considering as predictor the mean of the predicted temperature (precipitation) at the four nearest gridboxes (Herrera et al. 2017).
  Implementation: In-house Matlab code.

- **QM-DAP:** Implementation of the empirical quantile mapping method (Déqué 2007) smoothing the final corrections (obtained for individual percentiles) with a low-pass Gaussian filter (over 20 percentiles) to reduce noise in the individual percentile values. Each month was treated separately and a time window including the previous and following month was applied. To preserve reasonable extrapolated values (in the tails of the distribution), changes between the last percentiles (likely to be very noisy) were limited to certain values (such as a coefficient of 1.5 for maximal extrapolated value, compared to the last percentile, and a ratio of 3.0 as a change between the last two percentile values). More details in Štěpánek, P. et al. (2016).
  Implementation: In-house R code, incorporated in ProClimDB software (www.climahom.eu).

- **QMm:** Equidistant empirical quantile mapping. Empirical CDFs are calculated for the observation and the calibration and validation periods. The probabilities are calculated for bins with widths set for the resolution of the observational data (e.g. $0.1°$). For each day in the validation period, the probability obtained from the validation CDF is used in the observational and calibration CDFs to obtain the corresponding data values (Li et al. 2010). The difference between the observed and calibration data is used as the correction term for the validation. In the case of precipitation and in order to reduce the models drizzle effect, the percentile of the dry days of the validation period is matched to the observations, i.e. the precipitation in the validation period which corresponds to a percentile lower than the observational percentile is set to zero. In order to account for the seasonal cycle, the CDFs are constructed for a 31 day window centered on each day of the year. **Implementation:** In-house FORTRAN code. Available upon request to R.M. Cardoso.

- **QMBC-BJ-PR:** Implementation of the empirical quantile mapping method (Déqué 2007) adjusting 101 percentiles (including the minimum and maximum values) and using constant interpolation (with the mean of the two correction factors) between every two consecutive percentiles. Out of range values are adjusted using constant extrapolation (using the correction obtained for the minimum or maximum). The calibration is performed separately for each month. More details in Pongrácz et al. (2014); Bartholy et al. (2015).
  Implementation: In-house FORTRAN code.

- **Ratyetal-M6-M9 (only precipitation):** Monthly bias correction of daily precipitation implemented as in Räty et al. (2014). Methods M6 and M7 adjust the mean and standard deviation using linear and power scaling functions, respectively. M8 is a non-parametric quantile mapping with smoothing tailored for precipitation. The smoothing parameter value a=0.02 in Eq. (5) of Räty et al. (2014). M9 is a simplified version of the DBS method monthly transfer functions are estimated by fitting separate gamma distributions below and above the 95th percentile of daily precipitation (Yang et al. 2010); this constitutes a simplified version of the DBS method where the wet-day correction is only applied when there are too many wet days in the predictor data. No correction is done if the modeled wet day frequency is smaller than the observed one. In this sense M9 is less sophisticated than the actual DBS version. A 0.1 mm threshold was used to define wet-days. All methods use three-month time window when deriving the monthly corrections (e.g. data from December-January-February used for the correction applied in January).
  Implementation: In-house Fortran code.

- **RaiRat-M6-M9 (only temperatures):** Monthly varying bias correction of temperature following (Räisänen and Räty 2013). M6 adjusts only the mean value, M7 mean and standard deviation, and M8 mean, standard deviation and skewness. M9 uses a non-parametric quantile mapping approach with smoothing parameter D = 0.05 in Eq. (5) of (Räisänen and Räty 2013). A two-month data window is used in deriving the corrections (e.g. from mid-April to Mid-June for the correction applied in May) in all these methods.
  Implementation: In-house Fortran code.

- **SB (only temperatures)** A local scaling method where mean bias calculated separately for each season is subtracted from simulations in validating period.
  Implementation: In-house Matlab code.

- **VGLMGAMMA (only precipitation):** A stochastic single-site MOS approach to predict precipitation occurrence and amounts conditionally on simulated daily precipitation as predictor. Precipitation occurrence is modeled via a logistic regression; precipitation amounts on wet days based on a vector generalised linear model that expresses the rate and shape parameters of the 2-parameter gamma distribution as a function of simulated daily precipitation. Temporal dependence is not explicitly modelled but only imprinted by the predictor, i.e., individual occurrences and amounts are conditionally independent (Wong et al. 2014; Volosciuk et al. 2017).
  Implementation: In-house R code.

b. *PP Methods*

- **ANALOG:** Standard analog technique using Euclidean distance considering the complete fields to compute distances (Gutiérrez et al. 2013; San-Martín et al. 2017). The method has been trained across different zones covering Europe (similar to the Prudence regions) and has no seasonal component.

The method used raw predictor values applying a compression preprocess keeping the PCs explaining 95% of the total variance.
Implementation: *MeteoLab* public Matlab toolbox (`http://meteo.unican.es/trac/MLToolbox/wiki`) using *downTrain* function with parameters em method.type = 'analogs', *AnalogsNumber = 1*, *resampling='no'*.
These results (for the case of the ERA-Interim predictors) can be also reproduced (and modified) online using the statistical downscaling portal **http://meteo.unican.es/downscaling**, which builds on MeteoLab, and includes as illustrative examples the same standard predictor data and the VALUE regions used in this study. This package is freely available without restriction.

- **ANALOG-ANOM:** For a given day to be downscaled, the ANALOG-ANOM (Vaittinada Ayar et al. 2016) determines the day in the calibration period which has the closest atmospheric situation. This is determined by a similarity metric (here a Euclidean distance) between the predictor set of the day to be downscaled and the predictor set of the day in the calibration period, considering the whole European domain. For this method, the predictors are fields of daily anomalies with respect to the annual cycle computed from cubic regression smoothing splines fitted on the empirical daily annual cycle. Moreover, a seasonal restriction is applied: the selected analogs have to be in a +/-15 day-window around the climatological day of interest.

- **ANALOG-MP/SP:** Two versions of the analog model developed by Obled et al. (2002), optimized for the multivariate prediction of weather variables over the European region (Raynaud et al. 2016). For each prediction day, the probabilistic prediction is obtained from the 30 best atmospheric analogs selected in the atmospheric archive (selection in a calendar window of +/- 30 days). A two-level stepwise analogy is used for the analog selection: The first analogy level leads to 100 analogs from which are identified the 30 best final ones thanks to the 2nd analogy level. In both MP and SP versions, the first level of analogy is based on the shapes of 1000 and 500 hPa geopotential fields over a spatial domain centred on the target station (or centered on the region in the case of the multisite experiment). The analogy criterion is the Teweless-Wobus Score (Teweless, 1954). The 2nd level of analogy relies on a thermodynamic mesoscale predictor (analogy criterion is the RMSE). In ANALOG-SP, the predictor (T-Td at 2m) is the same for the three predictants (precip., Tmin, Tmax). The values of local temperature obtained with each analog day are post corrected using the difference between the mesoscale 2m temperature of a given target day with the one of the analog. In ANALOG-MP, the 2nd analogy predictor is predictand specific (VV600 for precip., T850 for temp. variables). In the present work, ANALOG-MP/SP include a stochastic process to produce the 100 required realizations of the predictand from the probabilistic prediction computed for each day.

- **ESD-EOFSLP/EOFT2/SLP/T2 (only temperatures):** Multiple linear regression method using monthly aggregated predictor and predictand data. It is important to remark that the ESD package is not designed to downscale daily values, but parameters describing the seasonal distribution of daily (or hourly) data, and combine this with a weather generator to produce time series. In this contribution this method has been trained on a monthly basis (using monthly aggregated data) in the traditional way, but this package is more flexible and it is typically calibrated differently when applied to GCM data. In that case, common EOFs (representative of both reanalysis and GCMs, Benestad et al. 2015b) are used as predictors and normally PCA are used as predictands for groups of stations which are subject to similar weather phenomena (multi-site application), although the method can be also applied to downscale more general information, such as the occurrence of intense local 24-hour precipitation events over seasonal intervals (Benestad and Mezghani 2015).
Implementation: ESD is implemented in the *esd* R package (Benestad et al. 2015a). This package is freely available without restriction.

- **FIC01P/03P (only precipitation):** FIC01P is a two-step analog methods. In a first step, the 30 closest analogs are computed for each test day based on Z1000 and Z500. Every analogue is defined in a three-windows nested grid (for short, medium, and large scale) with different weights. For this experiment we have used 42 main windows, each one with 3 nested windows, covering Europe. Instead of considering the weighted (according to similarity) mean observations of the analog days (p_i), the second step performs a pooling and ranking of the analog days month by month (900 values for each month) and computes the mean of consecutive blocks of 30 days q_i according to their mean values. Afterwards the values p_i are substituted by the new values q_i following a rank order, i.e. maximum by maximum, and so on (see, Ribalaygua et al. 2013, for more details). FIC03P is a version of FIC01P using near surface Wind, Wind at 500 hPA, relative humidity at 850 hPa and relative humidity at 700 hPa for computing the analogues. Moreover, we sort the n selected analogues for each problem day using their relative humidities at 850 hPa values and we weight the precipitation of the analogue day using the relation between the specific humidities at 700 hPa of the problem day and the analogue day. Then, as in FIC01P, we reassignate the previously daily simulated precipitation of a month, by using the distribution of the used analogue days for the whole month.

- **FIC01T (only temperatures):** A two-step analog method with the same first step as FIC01P with the same predictors, but considering 150 analogs for each test day. The second step consisting of a multiple linear regression using 1000-850 thickness, 1000-500 thickness and daily solar radiation (calculated as a function of the day of the year and the latitude of the station) as regressors; the regression is fitted considering the analog days. More details in Ribalaygua et al. (2013).

- **GLM-DET/GLM/GLM-WT (only precipitation):** Standard two-stage implementation of Generalized Linear Models (GLMs) for precipitation, in which a GLM with Bernoulli error distribution and logit canonical link-function (also known as logistic regression) is used to downscale daily precipitation occurrence (as characterized by a threshold of 0.1mm) and a GLM with gamma error distribution and log canonical link-function is applied to downscale daily precipitation amount (San-Martín et al. 2017). The method is trained across different zones covering Europe (similar to the PRUDENCE regions) with no seasonal component. The predictors are the 20 leading PCs (15 for GLM-WT) of the joined predictor fields (which account for 75-90% of the explained variance across the different zones). Particular methods are provided in both deterministic (GLM-DET) and stochastic (GLM) variants, using the expected value in the former case and simulating from the resulting binomial/gamma in the latter. An implementation conditioned to weather types (GLM-WT) is also used, considering 12 weather types defined using a k-means algorithm (k=12) applied to the daily SLP (this variable is excluded from the predictor set in this case).
  Implementation: *MeteoLab* public Matlab toolbox (`http://meteo.unican.es/trac/MLToolbox/wiki`) using *downTrain* function with parameters *type = 'glm'*, *ThresholdPrecip = 0.1*, *NumberOfNearestNeighbours = 0*, *NumberOfPCs = 15*, *SimOccurrence = 'true'*, *SimAmount = 'true'*, *minrainydays = 5*.
  These results (for the case of the ERA-Interim predictors) can be also reproduced (and modified) online using the statistical downscaling portal **http://meteo.unican.es/downscaling**, which builds on MeteoLab, and includes as illustrative examples the same standard predictor data and the VALUE regions used in this study. This package is freely available without restriction.

- **MLR/MLR-WT (only temperatures):** (Gutiérrez et al. 2013) Multiple linear regression trained across different zones covering Europe (similar to the Prudence regions) with no seasonal component. The predictors are the 15 leading PCs of the joined predictor fields (which account for 75-90% of the explained variance across the different zones considered). MLR-WT is a state-dependent version of MLR, conditioning the training to 12 Weather Types defined using a k-means algorithm (k=12) applied to the daily SLP over Europe (this variable is excluded from the predictor set).
  Implementation: *MeteoLab* public Matlab toolbox (`http://meteo.unican.es/trac/MLToolbox/wiki`) using *downTrain* function with parameters *type = 'linear_regression'*, *NumberOfNearestNeighbours = 0*, and *NumberOfPCs = 15*.
  These results (for the case of the ERA-Interim predictors) can be also reproduced (and modified)

online using the statistical downscaling portal **http://meteo.unican.es/downscaling**, which builds on MeteoLab, and includes as illustrative examples the same standard predictor data and the VALUE regions used in this study. This package is freely available without restriction.

- **MLR-PCA-ZRT (only temperatures):** Linear regression model using s-mode PCs as predictors. The selection of predictors has been automated by iterating through all possible predictor combinations, minimizing the mean squared error and maximizing the time series correlation in calibration and validation (Hertig and Jacobeit 2008; Hertig et al. 2013; Jacobeit et al. 2014). The selection was done for each station separately and models were developed for each month separately.
Implementation: In-house Fortran (for PC calculation) and R code (using "lm" for regression).

- **MLR-RSN/RAN/AAN/AAI/AAW/ASI/ASW:** Multiple linear pointwise regression (with stepwise screening) using gridpoint raw data (or anomalies), trained at an annual (or seasonal) basis and including optional variance corrections in the form of inflation or addition of white noise. The first letter of the code refers to the raw (R) or anomaly (A) data used as predictors, the second letter refers to the annual (A) or seasonal (S) training, and the third letter refers to inflation (I) or white noise (W) variance correction (N for no correction). More details in Huth (2002); Huth et al. (2015).
Implementation: In-house Fortran code.

- **MLR-T/GLM-P:** These methods have been implemented following the Statistical DownScaling Method SDSM, which builds on linear regression (Wilby et al. 2002). The parameters of the regression model are obtained by the least squares method from standardized variables. The link between the predictands and predictors is either an unconditional model, used for temperature (MLR-T), or a conditional model, used for precipitation (GLM-P), being the conditioning variable the probability of wet-day occurrence. The GLM-P method uses a logistic regression to estimate the probability of wet-day occurrence and an exponential regression to calculate the total daily precipitation amounts (Kilsby et al. 1998). Rainfall occurs when the probability of wet-day occurrence is greater than or equal to a uniform random number like in Wilby et al. (2002), thus incorporating an additional stochastic process. The selection of predictors changes from one site to another and from one variable to another and is based on a step-wise approach building on the adjusted determination coefficient.
Implementation: In-house C code.

- **MO-GP:** Multi-objective Genetic Programming (MOGP) performs a symbolic regression building a tree (six levels at most) with arithmetic functions and if-statements, i.e., not only the parameters but also the structure of the regression models are generated by GP. The multi-objective approach aims at a simultaneous optimization of RMSE, bias, standard deviation, selected quantiles and, for precipitation, the number of precipitation days. MOGP is applied individually for each station and variable. Except for precipitation, the predictors are interpolated from the four closest GCM grid cells to the location of the respective station. Precipitation is taken at the GCM grid box closest to a station. The MOGP code is based on the Strength Pareto Evolutionary Algorithm (SPEA) by Zitzler and Thiele (1999) and the GPLAB by Silva and Almeida (2003). SPEA returns not one single regression model for each station and variable but a set of Pareto optimal models. From each set of potential downscaling models one has been selected that optimizes a trade-off between all objectives. The automatic selection results in 8 predictors on average for precipitation and 6 for temperature. More details can be found in Zerenner et al. (2016).
Implementation: In-house MATLAB code (based on the GPLAB).

- **SWG:** A two-step approach is implemented to model precipitation in a Vectorised Generalized Linear Models (VGLM). First, the rainfall occurrence is modeled through a logistic regression, allowing to characterize the probability of rainfall occurrence for a given day conditionally on atmospheric predictors. Then, the probability density function (pdf) of the rain intensity (given that it rains) is assumed to be a Gamma distribution whose logarithms of the shape and rate parameters are linear functions of the large-scale predictors. For temperature, a single step is used, where temperature is supposed to follow a Gaussian distribution with the mean and the logarithm of the standard deviation linearly dependent on the predictors (Vaittinada Ayar et al. 2016).

- **WT-WG:** Gaussian/binomial-gamma distributions are fitted to the observed temperature/precipitation values within each of the 100 weather types obtained applying k-means to the SLP fields. These distributions are obtained to simulate downscaled values. More details in Gutiérrez et al. (2013); San-Martín et al. (2017).
  Implementation: *MeteoLab* public Matlab toolbox (`http://meteo.unican.es/trac/MLToolbox/wiki`) using *downTrain* function with parameters *method.type* = *'WT'*. This package is freely available without restriction.

c. *WG methods*

- **GOMEZ-BASIC/TAD:** Non-parametric weather generator based on a nearest neighbors resampling technique making no assumption on the distribution of the variables being generated. To represent the interdiurnal variability, each term (except for the first one) of the synthetic time series is derived from followers of K terms (selected from the learning observed series) closest to the previously generated term; the first term in the series is selected randomly from all available terms, which are within 10 days from January 1. The distance between individual terms is based on the Mahalanobis distance, in which precipitation is a binary variable (0 stands for the dry day, 1 for the wet day). Two versions of the generator were used in the VALUE experiment. In BASIC temperature is represented by TMAX and TMIN. In temperature is represented by TAVG and DTR (defined above in description of MARFI).

- **MARFI-BASIC/TAD/M3:** Parametric multivariate stochastic Richardson-type Richardson (1981) weather generator, which is a flexible follower of the Met&Roll generator (Dubrovský 1997; Dubrovský et al. 2004). Precipitation occurrence is modeled by Markov chain (order may vary between 1 and 3) and precipitation amount on wet day is sampled from the Gamma distribution. Standardized values of the temperature variables are modeled by the first-order bi-variate autoregressive model, in which the means and standard deviations of the two variables are conditioned on the state (wet or dry) of the day. Three versions of the settings were used in the experiment. In BASIC the two temperature variables are TMAX and TMIN, order of the Markov chain is one. TAD is similar to BASIC, but temperature is represented by TAVG (defined as an average of TMAX and TMIN) and DTR =(TMAX-TMIN) transformed (using quantile-mapping) into normally distributed variable. M3 is the same as BASIC, but a third-order Markov chain is used to model wet day occurrence.

- **SS-WG:** Multi-variate Richardson-type (Richardson 1981) weather generator simulating daily time-series of precipitation, minimum and maximum temperature (Keller et al. 2015, 2016). First, daily precipitation occurrence is modelled based on a first-order two-state Markov chain using 1mm/day as a wet threshold. Precipitation intensities are simulated from a mixture model of two exponential distributions. To ensure inter-variable consistency, the parameters of the temperature statistics are conditioned on the precipitation state. Synthetic temperature time-series are simulated using a first-order autoregressive model (AR1). All WG parameters are determined for each station and each month separately.

References

Abaurrea, J. and J. Asín, 2005: Forecasting local daily precipitation patterns in a climate change scenario. *Climate Research*, **28 (3)**, 183–197, doi:10.3354/cr028183, URL http://www.int-res.com/abstracts/cr/v28/n3/p183-197/.

Baker, M., 2016: 1500 scientists lift the lid on reproducibility. *Nature News*, **533 (7604)**, 452, doi:10.1038/533452a, URL http://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970.

Bartholy, J., R. Pongrácz, and A. Kis, 2015: Projected changes of extreme precipitation using multi-model approach. *Idojaras*, **119 (2)**, 129–142, URL https://hungary.pure.elsevier.com/en/publications/projected-changes-of-extreme-precipitation-using-multi-model-appr.

Bedia, J., M. Iturbide, S. Herrera, R. Manzanas, and J. Gutiérrez, 2016: *downscaleR: Climate data manipulation, bias correction and statistical downscaling*. URL http://github.com/SantanderMetGroup/downscaleR/wiki, r package version 2.0-0.

Benestad, R., I. Hanssen-Bauer, and D. Chen, 2008: *Empirical-Statistical Downscaling*. World Scientific, URL http://www.worldscientific.com/worldscibooks/10.1142/6908.

Benestad, R., A. Mezghani, and K. Parding, 2015a: *esd: Climate analysis and empirical-statistical downscaling (ESD) package for monthly and daily data.* URL http://rcg.gvc.gu.se/edu/esd.pdf, r package version 1.0.

Benestad, R. E., 2002: Empirically downscaled temperature scenarios for northern Europe based on a multi-model ensemble. *Climate Research*, **21 (2)**, 105–125, doi:10.3354/cr021105, URL http://www.int-res.com/abstracts/cr/v21/n2/p105-125/.

Benestad, R. E., D. Chen, A. Mezghani, L. Fan, and K. Parding, 2015b: On using principal components to represent stations in empirical–statistical downscaling. *Tellus A: Dynamic Meteorology and Oceanography*, **67 (1)**, 28 326, doi:10.3402/tellusa.v67.28326, URL http://www.tandfonline.com/doi/abs/10.3402/tellusa.v67.28326.

Benestad, R. E. and A. Mezghani, 2015: On downscaling probabilities for heavy 24-hour precipitation events at seasonal-to-decadal scales. *Tellus A: Dynamic Meteorology and Oceanography*, **67 (1)**, 25 954, doi:10.3402/tellusa.v67.25954, URL http://www.tandfonline.com/doi/abs/10.3402/tellusa.v67.25954.

Brands, S., J. M. Gutiérrez, S. Herrera, and A. S. Cofiño, 2012: On the Use of Reanalysis Data for Downscaling. *Journal of Climate*, **25 (7)**, 2517–2526, doi:10.1175/JCLI-D-11-00251.1, URL http://www.meteo.unican.es/en/node/73004.

Bürger, G., T. Q. Murdock, A. T. Werner, S. R. Sobie, and A. J. Cannon, 2012: Downscaling Extremes—An Intercomparison of Multiple Statistical Methods for Present Climate. *Journal of Climate*, **25 (12)**, 4366–4388, doi:10.1175/JCLI-D-11-00408.1, URL http://journals.ametsoc.org/doi/abs/10.1175/JCLI-D-11-00408.1.

Casanueva, A., S. Herrera, J. Fernández, and J. M. Gutiérrez, 2016a: Towards a fair comparison of statistical and dynamical downscaling in the framework of the EURO-CORDEX initiative. *Climatic Change*, 1–16, doi:10.1007/s10584-016-1683-4, URL http://link.springer.com/article/10.1007/s10584-016-1683-4.

Casanueva, A., et al., 2016b: Daily precipitation statistics in a EURO-CORDEX RCM ensemble: added value of raw and bias-corrected high-resolution simulations. *Climate Dynamics*, **47 (3-4)**, 719–737, doi:10.1007/s00382-015-2865-x, URL `https://link.springer.com/article/10.1007/s00382-015-2865-x`.

Chandler, R. E., 2005: On the use of generalized linear models for interpreting climate variability. *Environmetrics*, **16 (7)**, 699–715, doi:10.1002/env.731, URL `http://onlinelibrary.wiley.com/doi/10.1002/env.731/abstract`.

Chandler, R. E. and H. S. Wheater, 2002: Analysis of rainfall variability using generalized linear models: A case study from the west of Ireland. *Water Resources Research*, **38 (10)**, 10–1–10–11, doi:10.1029/2001WR000906, URL `http://onlinelibrary.wiley.com/doi/10.1029/2001WR000906/abstract`.

Chen, J., F. P. Brissette, and R. Leconte, 2014: Assessing regression-based statistical approaches for downscaling precipitation over North America. *Hydrological Processes*, **28 (9)**, 3482–3504, doi:10.1002/hyp.9889, URL `http://onlinelibrary.wiley.com/doi/10.1002/hyp.9889/abstract`.

Christensen, J., E. Kjellstrom, F. Giorgi, G. Lenderink, and Rummukainen M, 2010: Weight assignment in regional climate models. *Climate Research*, **44 (2-3)**, 179–194, URL `http://www.int-res.com/abstracts/cr/v44/n2-3/p179-194/`.

Christensen, J. H. and O. B. Christensen, 2007: A summary of the PRUDENCE model projections of changes in European climate by the end of this century. *Climatic Change*, **81 (1)**, 7–30, doi:10.1007/s10584-006-9210-7, URL `http://link.springer.com/article/10.1007/s10584-006-9210-7`.

Dee, D. P., et al., 2011: The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, **137 (656)**, 553–597, doi:10.1002/qj.828, URL `http://onlinelibrary.wiley.com/doi/10.1002/qj.828/abstract`.

Déqué, M., 2007: Frequency of precipitation and temperature extremes over France in an anthropogenic scenario: Model results and statistical correction according to observed values. *Global and Planetary Change*, **57 (1–2)**, 16–26, doi:10.1016/j.gloplacha.2006.11.030, URL `http://www.sciencedirect.com/science/article/pii/S0921818106002748`.

Dubrovský, M., 1997: Creating Daily Weather Series with Use of the Weather Generator. *Environmetrics*, **8 (5)**, 409–424, doi:10.1002/(SICI)1099-095X(199709/10)8:5⟨409::AID-ENV261⟩3.0.CO;2-0, URL `http://onlinelibrary.wiley.com/doi/10.1002/(SICI)1099-095X(199709/10)8:5<409::AID-ENV261>3.0.CO;2-0/abstract`.

Dubrovský, M., J. Buchtele, and Z. Žalud, 2004: High-Frequency and Low-Frequency Variability in Stochastic Daily Weather Generator and Its Effect on Agricultural and Hydrologic Modelling. *Climatic Change*, **63 (1-2)**, 145–179, doi:10.1023/B:CLIM.0000018504.99914.60, URL `http://link.springer.com/article/10.1023/B:CLIM.0000018504.99914.60`.

Eden, J. M., M. Widmann, D. Maraun, and M. Vrac, 2014: Comparison of GCM- and RCM-simulated precipitation following stochastic postprocessing. *Journal of Geophysical Research: Atmospheres*, 2014JD021732, doi:10.1002/2014JD021732, URL `http://onlinelibrary.wiley.com/doi/10.1002/2014JD021732/abstract`.

Flato, G., et al., 2013: Evaluation of Climate Models. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, T. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. Allen, J. Boschung,

A. Nauels, Y. Xia, V. Bex, and P. Midgley, Eds., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 741–866, URL www.climatechange2013.org, dOI: 10.1017/CBO9781107415324.020.

Fowler, H. J., S. Blenkinsop, and C. Tebaldi, 2007: Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. *International Journal of Climatology*, **27 (12)**, 1547–1578, doi:10.1002/joc.1556, URL http://onlinelibrary.wiley.com/doi/10.1002/joc.1556/abstract.

Frost, A. J., et al., 2011: A comparison of multi-site daily rainfall downscaling techniques under Australian conditions. *Journal of Hydrology*, **408 (1–2)**, 1–18, doi:10.1016/j.jhydrol.2011.06.021, URL http://www.sciencedirect.com/science/article/pii/S0022169411004525.

Gaitan, C. F., W. W. Hsieh, A. J. Cannon, and P. Gachon, 2014: Evaluation of Linear and Non-Linear Downscaling Methods in Terms of Daily Variability and Climate Indices: Surface Temperature in Southern Ontario and Quebec, Canada. *Atmosphere-Ocean*, **52 (3)**, 211–221, doi:10.1080/07055900.2013.857639, URL http://dx.doi.org/10.1080/07055900.2013.857639.

Giorgi, F. and L. O. Mearns, 1991: Approaches to the simulation of regional climate change: A review. *Reviews of Geophysics*, **29 (2)**, 191–216, doi:10.1029/90RG02636, URL http://onlinelibrary.wiley.com/doi/10.1029/90RG02636/abstract.

Gutiérrez, J. M., D. San-Martín, S. Brands, R. Manzanas, and S. Herrera, 2013: Reassessing Statistical Downscaling Techniques for Their Robust Application under Climate Change Conditions. *Journal of Climate*, **26 (1)**, 171–188, doi:10.1175/JCLI-D-11-00687.1, URL http://journals.ametsoc.org/doi/abs/10.1175/JCLI-D-11-00687.1.

Gutjahr, O. and G. Heinemann, 2013: Comparing precipitation bias correction methods for high-resolution regional climate simulations using COSMO-CLM. *Theoretical and Applied Climatology*, **114 (3-4)**, 511–529, doi:10.1007/s00704-013-0834-z, URL http://link.springer.com/article/10.1007/s00704-013-0834-z.

Gutmann, E., T. Pruitt, M. P. Clark, L. Brekke, J. R. Arnold, D. A. Raff, and R. M. Rasmussen, 2014: An intercomparison of statistical downscaling methods used for water resource assessments in the United States. *Water Resources Research*, **50 (9)**, 7167–7186, doi:10.1002/2014WR015559, URL http://onlinelibrary.wiley.com/doi/10.1002/2014WR015559/abstract.

HanssenBauer, I., C. Achberger, R. E. Benestad, D. Chen, and E. J. Frland, 2005: Statistical downscaling of climate scenarios over Scandinavia. *Climate Research*, **29 (3)**, 255–268, doi:10.3354/cr029255, URL http://www.int-res.com/abstracts/cr/v29/n3/p255-268/.

Hartkamp, A. D., J. W. White, and G. Hoogenboom, 2003: Comparison of three weather generators for crop modeling: a case study for subtropical environments. *Agricultural Systems*, **76 (2)**, 539–560, doi:10.1016/S0308-521X(01)00108-1, URL http://www.sciencedirect.com/science/article/pii/S0308521X01001081.

Haylock, M. R., G. C. Cawley, C. Harpham, R. L. Wilby, and C. M. Goodess, 2006: Downscaling heavy precipitation over the United Kingdom: a comparison of dynamical and statistical methods and their future scenarios. *International Journal of Climatology*, **26 (10)**, 1397–1415, doi:10.1002/joc.1318, URL http://onlinelibrary.wiley.com/doi/10.1002/joc.1318/abstract.

Hempel, S., K. Frieler, L. Warszawski, J. Schewe, and F. Piontek, 2013: A trend-preserving bias correction – the ISI-MIP approach. *Earth Syst. Dynam.*, **4 (2)**, 219–236, doi:10.5194/esd-4-219-2013, URL http://www.earth-syst-dynam.net/4/219/2013/.

Herrera, S., M. Turco, and J. M. Gutiérrez, 2017: A MOS-Regression Technique for Temporally-Coherent Bias Correction of Regional Climate Model Simulations. *Climate Dynamics*, **submitted**.

Hertig, E. and J. Jacobeit, 2008: Downscaling future climate change: Temperature scenarios for the Mediterranean area. *Global and Planetary Change*, **63 (2–3)**, 127–131, doi:10.1016/j.gloplacha.2007.09.003, URL http://www.sciencedirect.com/science/article/pii/S0921818107001749.

Hertig, E., S. Seubert, A. Paxian, G. Vogt, H. Paeth, and J. Jacobeit, 2013: Changes of total versus extreme precipitation and dry periods until the end of the twenty-first century: statistical assessments for the Mediterranean area. *Theoretical and Applied Climatology*, **111 (1-2)**, 1–20, doi:10.1007/s00704-012-0639-5, URL http://link.springer.com/article/10.1007/s00704-012-0639-5.

Hertig, E., S. Seubert, A. Paxian, G. Vogt, H. Paeth, and J. Jacobeit, 2014: Statistical modelling of extreme precipitation indices for the Mediterranean area under future climate change. *International Journal of Climatology*, **34 (4)**, 1132–1156, doi:10.1002/joc.3751, URL http://onlinelibrary.wiley.com/doi/10.1002/joc.3751/abstract.

Hessami, M., P. Gachon, T. B. M. J. Ouarda, and A. St-Hilaire, 2008: Automated regression-based statistical downscaling tool. *Environmental Modelling & Software*, **23 (6)**, 813–834, doi:10.1016/j.envsoft.2007.10.004, URL http://www.sciencedirect.com/science/article/pii/S1364815207001867.

Hu, Y., S. Maskey, and S. Uhlenbrook, 2013: Downscaling daily precipitation over the Yellow River source region in China: a comparison of three statistical downscaling methods. *Theoretical and Applied Climatology*, **112 (3-4)**, 447–460, doi:10.1007/s00704-012-0745-4, URL http://link.springer.com/article/10.1007/s00704-012-0745-4.

Huth, R., 1999: Statistical downscaling in central Europe: evaluation of methods and potential predictors. *Climate Research*, **13 (2)**, 91–101, doi:10.3354/cr013091, URL http://www.int-res.com/abstracts/cr/v13/n2/p91-101/.

Huth, R., 2002: Statistical Downscaling of Daily Temperature in Central Europe. *Journal of Climate*, **15 (13)**, 1731–1742, doi:10.1175/1520-0442(2002)015⟨1731:SDODTI⟩2.0.CO;2, URL http://journals.ametsoc.org/doi/abs/10.1175/1520-0442(2002)015%3C1731%3ASDODTI%3E2.0.CO%3B2.

Huth, R., 2004: Sensitivity of Local Daily Temperature Change Estimates to the Selection of Downscaling Models and Predictors. *Journal of Climate*, **17 (3)**, 640–652, doi:10.1175/1520-0442(2004)017⟨0640:SOLDTC⟩2.0.CO;2, URL http://journals.ametsoc.org/doi/abs/10.1175/1520-0442%282004%29017%3C0640%3ASOLDTC%3E2.0.CO%3B2.

Huth, R., 2005: Downscaling of humidity variables: a search for suitable predictors and predictands. *International Journal of Climatology*, **25 (2)**, 243–250, doi:10.1002/joc.1122, URL http://onlinelibrary.wiley.com/doi/10.1002/joc.1122/abstract.

Huth, R., J. Mikšovský, P. Štěpánek, M. Belda, A. Farda, Z. Chládová, and P. Pišoft, 2015: Comparative validation of statistical and dynamical downscaling models on a dense grid in central Europe: temperature. *Theoretical and Applied Climatology*, **120 (3-4)**, 533–553, doi:10.1007/s00704-014-1190-3, URL http://link.springer.com/article/10.1007/s00704-014-1190-3.

Jacob, D., et al., 2014: EURO-CORDEX: new high-resolution climate change projections for European impact research. *Regional Environmental Change*, **14 (2)**, 563–578, doi:10.1007/s10113-013-0499-2.

Jacobeit, J., E. Hertig, S. Seubert, and K. Lutz, 2014: Statistical downscaling for climate change projections in the Mediterranean region: methods and results. *Regional Environmental Change*, **14 (5)**, 1891–1906, doi:10.1007/s10113-014-0605-0, URL `http://link.springer.com/article/10.1007/s10113-014-0605-0`.

Jeong, D. I., A. St-Hilaire, T. B. M. J. Ouarda, and P. Gachon, 2012: Comparison of transfer functions in statistical downscaling models for daily temperature and precipitation over Canada. *Stochastic Environmental Research and Risk Assessment*, **26 (5)**, 633–653, doi:10.1007/s00477-011-0523-3, URL `https://link.springer.com/article/10.1007/s00477-011-0523-3`.

Keller, D. E., A. M. Fischer, C. Frei, M. A. Liniger, C. Appenzeller, and R. Knutti, 2015: Implementation and validation of a Wilks-type multi-site daily precipitation generator over a typical Alpine river catchment. *Hydrol. Earth Syst. Sci.*, **19 (5)**, 2163–2177, doi:10.5194/hess-19-2163-2015, URL `http://www.hydrol-earth-syst-sci.net/19/2163/2015/`.

Keller, D. E., A. M. Fischer, M. A. Liniger, C. Appenzeller, and R. Knutti, 2016: Testing a weather generator for downscaling climate change projections over Switzerland. *International Journal of Climatology*, n/a–n/a, doi:10.1002/joc.4750, URL `http://dx.doi.org/10.1002/joc.4750`.

Kilsby, C. G., P. S. P. Cowpertwait, P. E. O'Connell, and P. D. Jones, 1998: Predicting rainfall statistics in England and Wales using atmospheric circulation variables. *International Journal of Climatology*, **18 (5)**, 523–539, doi:10.1002/(SICI)1097-0088(199804)18:5⟨523::AID-JOC268⟩3.0.CO;2-X, URL `http://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-0088(199804)18:5<523::AID-JOC268>3.0.CO;2-X/abstract`.

Klein Tank, A. M. G., et al., 2002: Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. *International Journal of Climatology*, **22 (12)**, 1441–1453, doi:10.1002/joc.773, URL `http://onlinelibrary.wiley.com/doi/10.1002/joc.773/abstract`.

Kotlarski, S., et al., 2014: Regional climate modeling on European scales: a joint standard evaluation of the EURO-CORDEX RCM ensemble. *Geosci. Model Dev.*, **7 (4)**, 1297–1333, doi:10.5194/gmd-7-1297-2014, URL `http://www.geosci-model-dev.net/7/1297/2014/`.

Kottek, M., J. Grieser, C. Beck, B. Rudolf, and F. Rubel, 2006: World Map of the Köppen-Geiger climate classification updated. *Meteorologische Zeitschrift*, **15 (3)**, 259–263, doi:10.1127/0941-2948/2006/0130.

Leung, L. R., L. O. Mearns, F. Giorgi, and R. L. Wilby, 2003: Regional Climate Research. *Bulletin of the American Meteorological Society*, **84 (1)**, 89–95, doi:10.1175/BAMS-84-1-89, URL `http://journals.ametsoc.org/doi/abs/10.1175/bams-84-1-89`.

Li, H., J. Sheffield, and E. F. Wood, 2010: Bias correction of monthly precipitation and temperature fields from Intergovernmental Panel on Climate Change AR4 models using equidistant quantile matching. *Journal of Geophysical Research: Atmospheres (1984–2012)*, **115 (D10)**, doi:10.1029/2009JD012882, URL `http://onlinelibrary.wiley.com/doi/10.1029/2009JD012882/abstract`.

Manzanas, R., S. Brands, D. San-Martín, A. Lucero, C. Limbo, and J. M. Gutiérrez, 2015: Statistical Downscaling in the Tropics Can Be Sensitive to Reanalysis Choice: A Case Study for Precipitation in the Philippines. *Journal of Climate*, **28 (10)**, 4171–4184, doi:10.1175/JCLI-D-14-00331.1, URL `http://journals.ametsoc.org/doi/abs/10.1175/JCLI-D-14-00331.1`.

Maraun, D. and M. Widmann, 2017: *Statistical Downscaling and Bias Correction for Climate Research by Douglas Maraun*. Cambridge University Press, URL `/core/books/statistical-downscaling-and-bias-correction-for-climate-research/4ED479BAA8309C7ECBE6136236E3960F`.

Maraun, D., et al., 2010: Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Reviews of Geophysics*, **48 (3)**, n/a–n/a, doi:10.1029/2009RG000314, URL `http://onlinelibrary.wiley.com/doi/10.1029/2009RG000314/abstract`.

Maraun, D., et al., 2015: VALUE: A framework to validate downscaling approaches for climate change studies. *Earth's Future*, **3 (1)**, 2014EF000 259, doi:10.1002/2014EF000259, URL `http://onlinelibrary.wiley.com/doi/10.1002/2014EF000259/abstract`.

Maraun, D., et al., 2017: Towards process-informed bias correction of climate change simulations. *Nature Climate Change*, **in press**.

Meijgaard, E. v., L. H. v. Ulft, G. Lenderink, S. R. d. Roode, E. L. Wipfler, R. Boers, and R. M. A. Timmermans, 2012: *Refinement and Application of a Regional Atmospheric Model for Climate Scenario Calculations of Western Europe*. Programme Office Climate changes Spatial Planning, URL `http://www.wur.nl/de/Publicatie-details.htm?publicationId=publication-way-343237303937`.

Monjo, R., G. Chust, and V. Caselles, 2014: Probabilistic correction of RCM precipitation in the Basque Country (Northern Spain). *Theoretical and Applied Climatology*, **117 (1-2)**, 317–329, doi:10.1007/s00704-013-1008-8, URL `http://link.springer.com/article/10.1007/s00704-013-1008-8`.

Obled, C., G. Bontron, and R. Garçon, 2002: Quantitative precipitation forecasts: a statistical adaptation of model outputs through an analogues sorting approach. *Atmospheric Research*, **63 (3–4)**, 303–324, doi:10.1016/S0169-8095(02)00038-8, URL `http://www.sciencedirect.com/science/article/pii/S0169809502000388`.

Pongrácz, R., J. Bartholy, and A. Kis, 2014: Estimation of future precipitation conditions for Hungary with special focus on dry periods. *Idojaras*, **118 (4)**, 305–321, URL `https://hungary.pure.elsevier.com/hu/publications/estimation-of-future-precipitation-conditions-for-hungary-with-sp`.

Räisänen, J. and O. Räty, 2013: Projections of daily mean temperature variability in the future: cross-validation tests with ENSEMBLES regional climate simulations. *Climate Dynamics*, **41 (5-6)**, 1553–1568, doi:10.1007/s00382-012-1515-9, URL `http://link.springer.com/article/10.1007/s00382-012-1515-9`.

Räty, O., J. Räisänen, and J. S. Ylhäisi, 2014: Evaluation of delta change and bias correction methods for future daily precipitation: intermodel cross-validation using ENSEMBLES simulations. *Climate Dynamics*, **42 (9)**, 2287–2303, doi:10.1007/s00382-014-2130-8, URL `http://dx.doi.org/10.1007/s00382-014-2130-8`.

Raynaud, D., B. Hingray, I. Zin, S. Anquetin, S. Debionne, and R. Vautard, 2016: Atmospheric analogues for physically consistent scenarios of surface weather in Europe and Maghreb. *International Journal of Climatology*, doi:10.1002/joc.4844, URL `http://onlinelibrary.wiley.com/doi/10.1002/joc.4844/abstract`.

Reiter, P., O. Gutjahr, L. Schefczyk, G. Heinemann, and M. Casper, 2017: Does applying quantile mapping to subsamples improve the bias correction of daily precipitation? *International Journal of Climatology*, n/a–n/a, doi:10.1002/joc.5283, URL `http://onlinelibrary.wiley.com/doi/10.1002/joc.5283/abstract`.

Ribalaygua, J., L. Torres, J. Pórtoles, R. Monjo, E. Gaitán, and M. R. Pino, 2013: Description and validation of a two-step analogue/regression downscaling method. *Theoretical and Applied Climatology*, **114 (1-2)**, 253–269, doi:10.1007/s00704-013-0836-x, URL `http://link.springer.com/article/10.1007/s00704-013-0836-x`.

Richardson, C. W., 1981: Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resources Research*, **17 (1)**, 182–190, doi:10.1029/WR017i001p00182, URL `http://onlinelibrary.wiley.com/doi/10.1029/WR017i001p00182/abstract`.

Rummukainen, M., 2010: State-of-the-art with regional climate models. *Wiley Interdisciplinary Reviews: Climate Change*, **1 (1)**, 82–96, doi:10.1002/wcc.8, URL `http://onlinelibrary.wiley.com/doi/10.1002/wcc.8/abstract`.

San-Martín, D., R. Manzanas, S. Brands, S. Herrera, and J. M. Gutiérrez, 2017: Reassessing Model Uncertainty for Regional Projections of Precipitation with an Ensemble of Statistical Downscaling Methods. *Journal of Climate*, **30 (1)**, 203–223, doi:10.1175/JCLI-D-16-0366.1, URL `http://journals.ametsoc.org/doi/10.1175/JCLI-D-16-0366.1`.

Semenov, M. A., R. J. Brooks, E. M. Barrow, and C. W. Richardson, 1998: Comparison of the WGEN and LARS-WG stochastic weather generators for diverse climates. *Climate Research*, **10 (2)**, 95–107, doi:10.3354/cr010095, URL `http://www.int-res.com/abstracts/cr/v10/n2/p95-107/`.

Silva, S. and J. Almeida, 2003: GPLAB - A Genetic Programming Toolbox for MATLAB. *Nor3dic MATLAB Conference (NMC-2003)*, URL `https://www.cisuc.uc.pt/publication/show/1290`.

Takayabu, I., H. Kanamaru, K. Dairaku, R. Benestad, H. v. Storch, and J. H. Christensen, 2016: Reconsidering the Quality and Utility of Downscaling. *Journal of the Meteorological Society of Japan. Ser. II*, **94A**, 31–45, doi:10.2151/jmsj.2015-042.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2011: An Overview of CMIP5 and the Experiment Design. *Bulletin of the American Meteorological Society*, **93 (4)**, 485–498, doi:10.1175/BAMS-D-11-00094.1, URL `http://journals.ametsoc.org/doi/abs/10.1175/BAMS-D-11-00094.1`.

Teutschbein, C. and J. Seibert, 2013: Is bias correction of regional climate model (RCM) simulations possible for non-stationary conditions? *Hydrol. Earth Syst. Sci.*, **17 (12)**, 5061–5077, doi:10.5194/hess-17-5061-2013, URL `http://www.hydrol-earth-syst-sci.net/17/5061/2013/`.

Teutschbein, C., F. Wetterhall, and J. Seibert, 2011: Evaluation of different downscaling techniques for hydrological climate-change impact studies at the catchment scale. *Climate Dynamics*, **37 (9-10)**, 2087–2105, doi:10.1007/s00382-010-0979-8, URL `http://link.springer.com/article/10.1007/s00382-010-0979-8`.

Themeßl, M. J., A. Gobiet, and G. Heinrich, 2012: Empirical-statistical downscaling and error correction of regional climate models and its impact on the climate change signal. *Climatic Change*, **112 (2)**, 449–468, doi:10.1007/s10584-011-0224-4, URL `http://link.springer.com/article/10.1007/s10584-011-0224-4`.

Timbal, B., A. Dufour, and B. McAvaney, 2003: An estimate of future climate change for western France using a statistical downscaling technique. *Climate Dynamics*, **20 (7-8)**, 807–823, doi:10.1007/s00382-002-0298-9, URL `https://link.springer.com/article/10.1007/s00382-002-0298-9`.

Turco, M., M. C. Llasat, S. Herrera, and J. M. Gutiérrez, 2017: Bias correction and downscaling of future RCM precipitation projections using a MOS-Analog technique. *Journal of Geophysical Research: Atmospheres*, **122 (5)**, 2631–2648, doi:10.1002/2016JD025724, URL http://onlinelibrary.wiley.com/doi/10.1002/2016JD025724/abstract.

Turco, M., P. Quintana-Seguí, M. C. Llasat, S. Herrera, and J. M. Gutiérrez, 2011: Testing MOS precipitation downscaling for ENSEMBLES regional climate models over Spain. *Journal of Geophysical Research*, **116 (D18)**, doi:10.1029/2011JD016166, URL http://www.meteo.unican.es/en/node/73023.

Štěpánek, P., Zahradníček, P., Farda, A., Skalák, P., Trnka, M., Meitner, J., and Rajdl, K., 2016: Projection of drought-inducing climate conditions in the Czech Republic according to Euro-CORDEX models. *Climate Research*, **70 (2-3)**, 179–193, URL http://www.int-res.com/abstracts/cr/v70/n2-3/p179-193/.

Vaittinada Ayar, P., M. Vrac, S. Bastin, J. Carreau, M. Déqué, and C. Gallardo, 2016: Intercomparison of statistical and dynamical downscaling models under the EURO- and MED-CORDEX initiative framework: present climate evaluations. *Climate Dynamics*, **46 (3-4)**, 1301–1329, doi:10.1007/s00382-015-2647-5, URL http://link.springer.com/article/10.1007/s00382-015-2647-5.

Volosciuk, C., D. Maraun, M. Vrac, and M. Widmann, 2017: A combined statistical bias correction and stochastic downscaling method for precipitation. *Hydrol. Earth Syst. Sci.*, **21 (3)**, 1693–1719, doi:10.5194/hess-21-1693-2017, URL http://www.hydrol-earth-syst-sci.net/21/1693/2017/.

Vrac, M., P. Drobinski, A. Merlo, M. Herrmann, C. Lavaysse, L. Li, and S. Somot, 2012: Dynamical and statistical downscaling of the French Mediterranean climate: uncertainty assessment. *Nat. Hazards Earth Syst. Sci.*, **12 (9)**, 2769–2784, doi:10.5194/nhess-12-2769-2012, URL http://www.nat-hazards-earth-syst-sci.net/12/2769/2012/.

Widmann, M., C. S. Bretherton, and E. P. Salathé, 2003: Statistical Precipitation Downscaling over the Northwestern United States Using Numerically Simulated Precipitation as a Predictor. *Journal of Climate*, **16 (5)**, 799–816, doi:10.1175/1520-0442(2003)016⟨0799:SPDOTN⟩2.0.CO;2, URL http://journals.ametsoc.org/doi/abs/10.1175/1520-0442(2003)016%3C0799%3ASPDOTN%3E2.0.CO%3B2.

Wilby, R. L., S. P. Charles, E. Zorita, B. Timbal, P. Whetton, and L. O. Mearns, 2004: Guidelines for Use of Climate Scenarios Developed from Statistical Downscaling Methods. Supporting Material, Intergovernmental Panel on Climate Change. URL http://www.ipcc-data.org/guidelines/dgm_no2_v1_09_2004.pdf, accessed on 28 Aug 2013.

Wilby, R. L., C. W. Dawson, and E. M. Barrow, 2002: sdsm — a decision support tool for the assessment of regional climate change impacts. *Environmental Modelling & Software*, **17 (2)**, 145–157, doi:10.1016/S1364-8152(01)00060-3, URL http://www.sciencedirect.com/science/article/pii/S1364815201000603.

Wilby, R. L., T. M. L. Wigley, D. Conway, P. D. Jones, B. C. Hewitson, J. Main, and D. S. Wilks, 1998: Statistical downscaling of general circulation model output: A comparison of methods. *Water Resources Research*, **34 (11)**, 2995–3008, doi:10.1029/98WR02577, URL http://onlinelibrary.wiley.com/doi/10.1029/98WR02577/abstract.

Wilcke, R. A. I., T. Mendlik, and A. Gobiet, 2013: Multi-variable error correction of regional climate models. *Climatic Change*, **120 (4)**, 871–887, doi:10.1007/s10584-013-0845-x, URL http://link.springer.com/article/10.1007/s10584-013-0845-x.

Wilks, D. S. and R. L. Wilby, 1999: The weather generation game: a review of stochastic weather models. *Progress in Physical Geography*, **23 (3)**, 329–357, doi:10.1177/030913339902300302, URL http://ppg.sagepub.com/content/23/3/329.

Winkler, J. A., et al., 2011: Climate Scenario Development and Applications for Local/Regional Climate Change Impact Assessments: An Overview for the Non-Climate Scientist. *Geography Compass*, **5 (6)**, 275–300, doi:10.1111/j.1749-8198.2011.00425.x, URL http://onlinelibrary.wiley.com/doi/10.1111/j.1749-8198.2011.00425.x/abstract.

Wong, G., D. Maraun, M. Vrac, M. Widmann, J. M. Eden, and T. Kent, 2014: Stochastic Model Output Statistics for Bias Correcting and Downscaling Precipitation Including Extremes. *Journal of Climate*, **27 (18)**, 6940–6959, doi:10.1175/JCLI-D-13-00604.1, URL http://journals.ametsoc.org/doi/abs/10.1175/JCLI-D-13-00604.1.

Yang, W., J. Andréasson, L. P. Graham, J. Olsson, J. Rosberg, and F. Wetterhall, 2010: Distribution-based scaling to improve usability of regional climate model projections for hydrological climate change impacts studies. *Hydrology Research*, **41 (3-4)**, 211–229, doi:10.2166/nh.2010.004, URL http://hr.iwaponline.com/content/41/3-4/211.

Yang, W., M. Gardelin, J. Olsson, and T. Bosshard, 2015: Multi-variable bias correction: application of forest fire risk in present and future climate in Sweden. *Nat. Hazards Earth Syst. Sci.*, **15 (9)**, 2037–2057, doi:10.5194/nhess-15-2037-2015, URL http://www.nat-hazards-earth-syst-sci.net/15/2037/2015/.

Zerenner, T., V. Venema, P. Friederichs, and C. Simmer, 2016: Downscaling near-surface atmospheric fields with multi-objective Genetic Programming. *Environmental Modelling & Software*, **84**, 85–98, doi:10.1016/j.envsoft.2016.06.009, URL http://www.sciencedirect.com/science/article/pii/S1364815216302122.

Zitzler, E. and L. Thiele, 1999: Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach. *IEEE Transactions on Evolutionary Computation*, **3 (4)**, 257–271, doi:10.1109/4235.797969.

Table 1: List of stations indicating the order (sorted by latitude), ECA&D IDs, name, longitude, latitude, elevation, country, and Köppen–Geiger climate type.

#	ID	Name	Lon.	Lat.	Elev.	Country	Köppen
1	231	Malaga	-4.49	36.67	7	Spain	Csa
2	63	Methoni	21.70	36.83	51	Greece	Csa
3	214	Lisboa-Geofisica	-9.15	38.72	77	Portugal	Csa
4	229	Badajoz/Talavera-La-Real	-6.83	38.88	185	Spain	Csa
5	175	Cagliari	9.05	39.23	21	Italy	Csa
6	3919	Palma-De-Mallorca	2.74	39.56	8	Spain	BSk
7	59	Corfu	19.92	39.62	11	Greece	Csa
8	62	Larissa	22.45	39.65	72	Greece	BSk
9	3946	Madrid-Barajas	-3.56	40.47	609	Spain	BSk
10	232	Navacerrada	-4.01	40.78	1894	Spain	Csb
11	236	Tortosa-Observatorio-Ebro	0.49	40.82	44	Spain	Csa
12	176	Roma-Ciampino	12.58	41.78	105	Italy	Csa
13	212	Braganca	-6.73	41.80	690	Portugal	Csb
14	1394	Santiago-De-Compostela	-8.41	42.89	370	Spain	Cfb
15	1686	Hvar	16.45	43.17	20	Croatia	Csa
16	234	San-Sebastian-Igueldo	-2.04	43.31	251	Spain	Cfb
17	39	Marseille-Marignane	5.23	43.44	5	France	Csa
18	800	Toulouse-Blagnac	1.38	43.62	151	France	Cfa
19	355	Mont-Aigoual	3.58	44.12	1567	France	Cfb
20	2062	Constanta	28.63	44.22	13	Romania	Cfa
21	219	Bucuresti-Baneasa	26.08	44.52	90	Romania	Cfa
22	1684	Gospic	15.37	44.55	564	Croatia	Cfb
23	1687	Zavizan	14.98	44.82	1594	Croatia	Dfc
24	177	Verona-Villafranca	10.87	45.38	68	Italy	Cfa
25	173	Milan	9.19	45.47	150	Italy	Cfa
26	450	Sibiu	24.15	45.80	444	Romania	Cfb
27	21	Zagreb-Gric	15.98	45.82	156	Croatia	Cfa
28	242	Lugano	8.97	46.00	300	Switzerland	Cfa
29	217	Arad	21.35	46.13	116	Romania	Cfb
30	1662	Sion-2	7.33	46.22	482	Switzerland	Cfb
31	15	Sonnblick	12.95	47.05	3106	Austria	ET
32	32	Bourges	2.37	47.07	161	France	Cfb
33	12	Graz	15.45	47.08	366	Austria	Cfb
34	951	Iasi	27.63	47.17	102	Romania	Cfa
35	243	Saentis	9.35	47.25	2502	Switzerland	ET
36	13	Innsbruck	11.40	47.27	577	Austria	Cfb
37	244	Zueriswitzerland	8.57	47.38	556	Switzerland	Cfb
38	4002	Oberstdorf	10.28	47.40	806	Germany	Cfb
39	58	Zugspitze	10.99	47.42	2964	Germany	ET
40	239	Basel-Binningen	7.58	47.55	316	Switzerland	Cfb
41	14	Salzburg	13.00	47.80	437	Austria	Cfb
42	48	Hohenpeissenberg	11.01	47.80	977	Germany	Cfb
43	322	Rennes	-1.73	48.07	36	France	Cfb
44	16	Wien	16.35	48.23	198	Austria	Cfb
45	38	Paris-14e	2.34	48.82	75	France	Cfb
46	2762	Rheinstetten	8.33	48.97	116	Germany	Cfb
47	4004	Regensburg	12.10	49.04	365	Germany	Cfb
48	3991	Giessen-Wettenberg	8.65	50.60	203	Germany	Cfb
49	17	Uccle	4.37	50.80	100	Belgium	Cfb
50	483	Dresden-Klotzsswitzerlande	13.76	51.13	227	Germany	Cfb
51	274	Oxford	-1.27	51.77	63	UK	Cfb
52	2006	Brocken	10.62	51.80	1142	Germany	Dfc
53	333	Siedlce	22.25	52.25	152	Poland	Dfb
54	54	Potsdam	13.06	52.38	81	Germany	Cfb
55	42	Bremen	8.80	53.05	4	Germany	Cfb
56	351	Waddington	0.52	53.17	68	UK	Cfb
57	350	Valley	-4.53	53.25	11	UK	Cfb
58	468	Helgoland	7.89	54.18	4	Germany	Cfb
59	1020	Lazdijai	23.52	54.23	133	Lithuania	Dfb

Table 1: List of stations indicating the order (sorted by latitude), ECA&D IDs, name, longitude, latitude, elevation, country, and Köppen–Geiger climate type.

#	ID	Name	Lon.	Lat.	Elev.	Country	Köppen
60	3994	Arkona	13.44	54.68	42	Germany	Cfb
61	332	Leba	17.53	54.75	2	Poland	Dfb
62	200	Kaunas	23.83	54.88	77	Lithuania	Dfb
63	272	Eskdalemuir	-3.20	55.32	242	UK	Cfb
64	201	Klaipeda	21.07	55.73	6	Lithuania	Cfb
65	113	Tranebjerg	10.60	55.85	11	Denmark	Cfb
66	1009	Birzai	24.77	56.20	60	Lithuania	Dfb
67	107	Vestervig	8.32	56.77	18	Denmark	Cfb
68	465	Visby	18.33	57.67	42	Sweden	Cfb
69	462	Goteborg	11.99	57.72	5	Sweden	Cfb
70	349	Stornoway	-6.32	58.33	9	UK	Cfb
71	275	Wick	-3.08	58.45	36	UK	Cfb
72	192	Faerder	10.53	59.03	6	Norway	Cfb
73	194	Utsira-Fyr	4.88	59.31	55	Norway	Cfb
74	28	Helsinki-Kaisaniemi	24.95	60.18	4	Finland	Dfb
75	708	Jokioinen-Jokioisten	23.50	60.81	104	Finland	Dfb
76	5585	Salen	13.26	61.17	360	Sweden	Dfc
77	191	Kjoeremsgrende	9.05	62.10	626	Norway	Dfc
78	330	Fokstua	9.28	62.12	952	Norway	Dfc
79	1051	Tafjord	7.42	62.23	15	Norway	Csb
80	29	Jyvaskyla-Lentoasema	25.68	62.40	139	Finland	Dfc
81	7682	Siikajoki-Revonlahti	25.09	64.68	48	Finland	Dfc
82	339	Haparanda	24.14	65.83	5	Sweden	Dfc
83	1427	Jackvik	17.00	66.38	430	Sweden	Dfc
84	30	Sodankyla-Lapin-Ilmatiet	26.63	67.37	179	Finland	Dfc
85	190	Karasjok	25.50	69.47	129	Norway	Dfc
86	195	Vardoe	31.08	70.37	14	Norway	ET

Variable	Code	Levels	Units	Temporal Aggregation
Minimum Temperature	TMIN	-	K	Daily minimum
Maximum Temperature	TMAX	-	K	Daily maximum
Total Precipitation	PRC	-	m	Daily accumulated
Mean Sea Level Pressure	MSL	-	Pa	Daily Mean
2m Temperature	2T	2m	K	Daily mean
Geopotential	Z	250 500 700 850 1000 mb	m2 s-2	Daily Mean
Temperature	T	250 500 700 850 1000 mb	K	Daily Mean
westerly wind component	U	250 500 700 850 1000 mb	m s-1	Daily Mean
southerly wind component	V	250 500 700 850 1000 mb	m s-1	Daily Mean
Specific humidity	Q	250 500 700 850 1000 mb	kg kg-1	Daily Mean

Table 2: Description of the variables, pressure levels, units and temporal aggregation of the common set of predictors used in the reference VALUE dataset.

Table 3: Table of ESD methods contributing to VALUE Experiment 1a for precipitation using ERA-Interim predictors (and RACMO2 RCM predictors additionally, for those methods with a cross in the second column). CODE is the public code of the method as shown in (`http://www.value-cost.eu/validationportal`). APPRO. and TECH. indicate the approach and techniques used, respectively. The codes used for the approaches are: RAW (raw data), MOS (Model Output Statistics), PP (Perfect Prognosis), WG (Weather Generators), and the families of techniques: S (additive/multiplicative scaling), PM (parametric quantile mapping), QM (empirical quantile mapping), WT (weather types), A (analogs), TF (transfer function), WG (Markov-type WGs). ST indicates the stochastic nature of the method (yes for stochastic ones, providing 100 realizations); MS and MV indicate whether the methods are multi-site and multi-variable, respectively (methods using PCs as predictors are indicated with a *yes* in the MS column). Finally, SE and AC indicate the explicit inclusion of seasonal and autocorrelation components, respectively. All methods provide daily data for the 86 stations. The shading indicates the subset of methods applied also for temperatures. (*) Only occurrence is randomized, amounts are based on inflated regression (in this case, a single realization was provided and used for validation).

#	R	INSTITUTION	CODE	APPRO.	TECH.	ST	MS	MV	SE	AC
1	-	ECMWF	ERAINT-200	RAW	-	-	-	-	-	-
2	-	ECMWF	ERAINT-075	RAW	-	-	-	-	-	-
3	X	KNMI	RACMO22E	RAW	-	-	-	-	-	-
4	X	UHEL	Ratyetal-M6	MOS	S	no	no	no	yes	no
5	X	UHEL	Ratyetal-M7	MOS	S	no	no	no	yes	no
6	X	UCAN/CSIC	ISI-MIP	MOS	S|PM	no	no	no	yes	no
7	X	SMHI	DBS	MOS	PM	no	no	yes	yes	no
8	X	UHEL	Ratyetal-M9	MOS	PM	no	no	no	yes	no
9	X	FIC	BC	MOS	PM	no	no	no	yes	no
10	X	UCAN/CSIC	GQM	MOS	PM	no	no	no	no	no
11	X	UCAN/CSIC	GPQM	MOS	PM	no	no	no	no	no
12	X	UCAN/CSIC	EQM	MOS	QM	no	no	no	no	no
13	X	UCAN/CSIC	EQMs	MOS	QM	no	no	no	yes	no
14	X	UCAN/CSIC	EQM-WT	MOS	QM|WT	no	no	no	no	no
15	X	IDL	QMm	MOS	QM	no	no	no	yes	no
16	X	ELU	QMBC-BJ-PR	MOS	QM	no	no	no	yes	no
17	X	LSCE/IPSL	CDFt	MOS	QM	no	no	no	yes	no
18	X	GCRI-CAS	QM-DAP	MOS	QM	no	no	no	yes	no
19	X	SMHI	EQM-WIC658	MOS	QM	no	no	no	yes	no
20	X	UHEL	Ratyetal-M8	MOS	QM	no	no	no	yes	no
21	X	UB	MOS-AN	MOS	A	no	yes	no	no	no
22	X	UCAN/CSIC	MOS-GLM	MOS	TF	yes	no	no	no	no
23	-	UNIGRAZ	VGLMGAMMA	MOS	TF	yes	no	no	yes	no
24	-	FIC	FIC02P	MOS|PP	PM|A|TF	no	no	no	yes	no
25	-	FIC	FIC04P	MOS|PP	PM|A|TF	no	no	no	yes	no
26	-	FIC	FIC01P	PP	A|TF	no	yes	no	yes	no
27	-	FIC	FIC03P	PP	A|TF	no	yes	no	yes	no
28	-	LSCE/IPSL	ANALOG-ANOM	PP	A	no	yes	yes	yes	no
29	-	UCAN/CSIC	ANALOG	PP	A	no	yes	yes	no	no
30	-	CNRS/IGE	ANALOG-MP	PP	A	yes	yes	yes	yes	no
31	-	CNRS/IGE	ANALOG-SP	PP	A	yes	yes	yes	yes	no
32	-	MIUB	MO-GP	PP	TF	no	no	no	no	no
33	-	AEMET	GLM-P	PP	TF	yes(*)	no	no	no	no
34	-	CUNI	MLR-RAN	PP	TF	no	no	no	no	no
35	-	CUNI	MLR-RSN	PP	TF	no	no	no	yes	no
36	-	CUNI	MLR-ASW	PP	TF	yes	no	no	yes	no
37	-	CUNI	MLR-ASI	PP	TF	no	no	no	yes	no
38	-	UCAN/CSIC	GLM-DET	PP	TF	no	*yes*	no	no	no
39	-	UCAN/CSIC	GLM	PP	TF	yes	*yes*	no	no	no
40	-	UCAN/CSIC	GLM-WT	PP	TF|WT	yes	*yes*	no	no	no
41	-	UCAN/CSIC	WT-WG	PP	WT	yes	no	no	no	no
42	-	LSCE/IPSL	SWG	PP	TF	yes	*yes*	no	yes	no
43	-	METEOSWISS	SS-WG	WG	WG	yes	no	yes	yes	yes
44	-	IAP-CAS	MARFI-BASIC	WG	WG	yes	no	yes	yes	yes
45	-	IAP-CAS	MARFI-TAD	WG	WG	yes	no	yes	yes	yes
46	-	IAP-CAS	MARFI-M3	WG	WG	yes	no	yes	yes	yes
47	-	IAP-CAS	GOMEZ-BASIC	WG	WG	yes	no	yes	yes	yes
48	-	IAP-CAS	GOMEZ-TAD	WG	WG	yes	no	yes	yes	yes

Table 4: As Table 3 but for minimum and maximum temperatures. All methods provide daily data for the 86 stations, except the ESD family (#39-42, in *italics*) which provide monthly data. The shading indicates the subset of methods applied also for precipitation.

#	R	INSTITUTION	CODE	APPRO.	TECH.	ST	MS	MV	SE	AC
1	-	ECMWF	ERAINT-200	RAW	-	-	-	-	-	-
2	-	ECMWF	ERAINT-075	RAW	-	-	-	-	-	-
3	X	KNMI	RACMO22E	RAW	-	-	-	-	-	-
4	X	UHEL	RaiRat-M6	MOS	S	no	no	no	yes	no
5	X	UHEL	RaiRat-M7	MOS	S	no	no	no	yes	no
6	X	UHEL	RaiRat-M8	MOS	S	no	no	no	yes	no
7	X	UL	SB	MOS	S	no	no	no	yes	no
8	X	UCAN/CSIC	ISI-MIP	MOS	S|PM	no	no	no	yes	no
9	X	SMHI	DBS	MOS	PM	no	no	yes	yes	no
10	X	UCAN/CSIC	GPQM	MOS	PM	no	no	no	no	no
11	X	UCAN/CSIC	EQM	MOS	QM	no	no	no	no	no
12	X	UCAN/CSIC	EQMs	MOS	QM	no	no	no	yes	no
13	X	UCAN/CSIC	EQM-WT	MOS	QM|WT	no	no	no	no	no
14	X	IDL	QMm	MOS	QM	no	no	no	yes	no
15	X	ELU	QMBC-BJ-PR	MOS	QM	no	no	no	yes	no
16	X	LSCE/IPSL	CDFt	MOS	QM	no	no	no	yes	no
17	X	GCRI-CAS	QM-DAP	MOS	QM	no	no	no	yes	no
18	X	SMHI	EQM-WIC658	MOS	QM	no	no	no	yes	no
19	X	UHEL	RaiRat-M9	MOS	QM	no	no	no	yes	no
20	X	UL	DBBC	MOS	QM	no	no	no	yes	no
21	X	UL	DBD	MOS	QM	no	no	no	yes	no
22	X	UCAN/CSIC	MOS-REG	MOS	TF	no	no	no	no	no
23	-	FIC	FIC02T	MOS|PP	PM|A|TF	no	no	no	yes	no
24	-	FIC	FIC01T	PP	A|TF	no	yes	no	yes	no
25	-	LSCE/IPSL	ANALOG-ANOM	PP	A	no	yes	yes	yes	no
26	-	UCAN/CSIC	ANALOG	PP	A	no	yes	yes	no	no
27	-	CNRS/IGE	ANALOG-MP	PP	A	yes	yes	yes	yes	no
28	-	CNRS/IGE	ANALOG-SP	PP	A	yes	yes	yes	yes	no
29	-	MIUB	MO-GP	PP	TF	no	no	no	no	no
30	-	AEMET	MLR-T	PP	TF	no	no	no	no	no
31	-	CUNI	MLR-RAN	PP	TF	no	no	no	no	no
32	-	CUNI	MLR-RSN	PP	TF	no	no	no	yes	no
33	-	CUNI	MLR-ASW	PP	TF	yes	no	no	yes	no
34	-	CUNI	MLR-ASI	PP	TF	no	no	no	yes	no
35	-	CUNI	MLR-AAN	PP	TF	no	no	no	no	no
36	-	CUNI	MLR-AAI	PP	TF	no	no	no	no	no
37	-	CUNI	MLR-AAW	PP	TF	yes	no	no	no	no
38	-	IGUA	MLR-PCA-ZTR	PP	TF	no	*yes*	no	yes	no
39	-	AMU	*ESD-EOFSLP*	PP	TF|WT	no	*yes*	no	yes	no
40	-	AMU	*ESD-EOFT2*	PP	TF|WT	no	*yes*	no	yes	no
41	-	AMU	*ESD-SLP*	PP	TF|WT	no	no	no	yes	no
42	-	AMU	*ESD-T2*	PP	TF|WT	no	no	no	yes	no
43	-	UCAN/CSIC	MLR	PP	TF	no	*yes*	no	no	no
44	-	UCAN/CSIC	MLR-WT	PP	TF|WT	no	*yes*	no	no	no
45	-	UCAN/CSIC	WT-WG	PP	WT	yes	no	no	no	no
46	-	LSCE/IPSL	SWG	PP	TF	yes	*yes*	no	yes	no
47	-	METEOSWISS	SS-WG	WG	WG	yes	no	yes	yes	yes
48	-	IAP-CAS	MARFI-BASIC	WG	WG	yes	no	yes	yes	yes
49	-	IAP-CAS	MARFI-TAD	WG	WG	yes	no	yes	yes	yes
50	-	IAP-CAS	MARFI-M3	WG	WG	yes	no	yes	yes	yes
51	-	IAP-CAS	GOMEZ-BASIC	WG	WG	yes	no	yes	yes	yes
52	-	IAP-CAS	GOMEZ-TAD	WG	WG	yes	no	yes	yes	yes

Table 5: Details about the predictors, geographical domains and preprocessing transformations used in the different MOS and PP statistical downscaling methods (note that distributional MOS and WG methods are not included since they use precipitation/temperatures at the closest gridbox, or use no predictor, respectively). The first column refers to the codes given in Tables 3 and 4. The last two columns indicate the transformations applied to the predictors (standardization, anomalies over the annual cycle, EOF/PC computation) and the size of the domain used: 'cont' for a single continental domain, 'nat' for multiple nation-wide domains, and 'gb'for information from the closest gridbox (or four gridboxes, '4 gb'). *Two-step* methods are indicated by including a '>' symbol between the two predictor/domain configurations used.

CODE	PREDICTORS	TRANSFORM	DOMAIN
MOS-GLM	Precip./Temp.	standardized	4 gb
MOS-REG	Precip./Temp.	standardized	4 gb
MOS-AN	Precip.	raw data	nat
VGLMGAMMA	Precip.	standardized	gb
FIC01P	Z1000+500	standardized	nat
FIC03P	U+V10, U+V500, R850+700 > R850, Q700	standardized	nat > gb
FIC01T	Z1000-500 > TH1000-850 + 1000-500	standardized	nat > gb
ANALOG-ANOM	SLP, TD, T2, U850, V850, Z850	anomalies	cont
ANALOG	SLP, T2, T500+700+850, Q500+850, Z500	PCs (95% variance)	nat
ANALOG-MP	Z1000+500 > VV600, T850	raw data	nat > gb
ANALOG-SP	Z1000+500 > T2-TD, T2	raw data	nat > gb
MO-GP	Standard set	raw data	gb
GLM-P	SLP, U+V10, T+Q+U+V850+700+500	standardized	gb
GLM-DET	SLP, T2, T500+700+850, Q500+850, Z500	20 joined PCs	nat
GLM	SLP, T2, T500+700+850, Q500+850, Z500	20 joined PCs	nat
GLM-WT	T2, T500+700+850, Q500+850, Z500 (SLP for WT)	15 joined PCs	nat
MLR-RAN	Z500, T850	raw data	cont
MLR-RSN	Z500, T850	raw data	cont
MLR-ASW	Z500, T850	anomalies	cont
MLR- ASI	Z500, T850	anomalies	cont
MLR-AAN	Z500, T850	anomalies	cont
MLR-AAI	Z500, T850	anomalies	cont
MLR-AAW	Z500, T850	anomalies	cont
MLR-PCA-ZTR	Z850, T850, R850	s-mod PCs	cont
MLR-T	T2, SLP, U+V10, T+Q+U+V850-700-500	standardized	gb
MLR	SLP, T2, T500+700+850, Q500+850, Z500	15 joined PCs	nat
MLR-WT	SLP, T2, T500+700+850, Q500+850, Z500	15 joined PCs	nat
ESD-EOFSLP	SLP	20 PCs	cont
ESD-EOFT2	T2	20 PCs	cont
ESD-SLP	SLP	raw data	cont
ESD-T2	T2	raw data	cont
WT-WG	SLP	15 PCs	nat
SWG	SLP, TD, T2, U850+V850+Z850	2 PCs each	cont

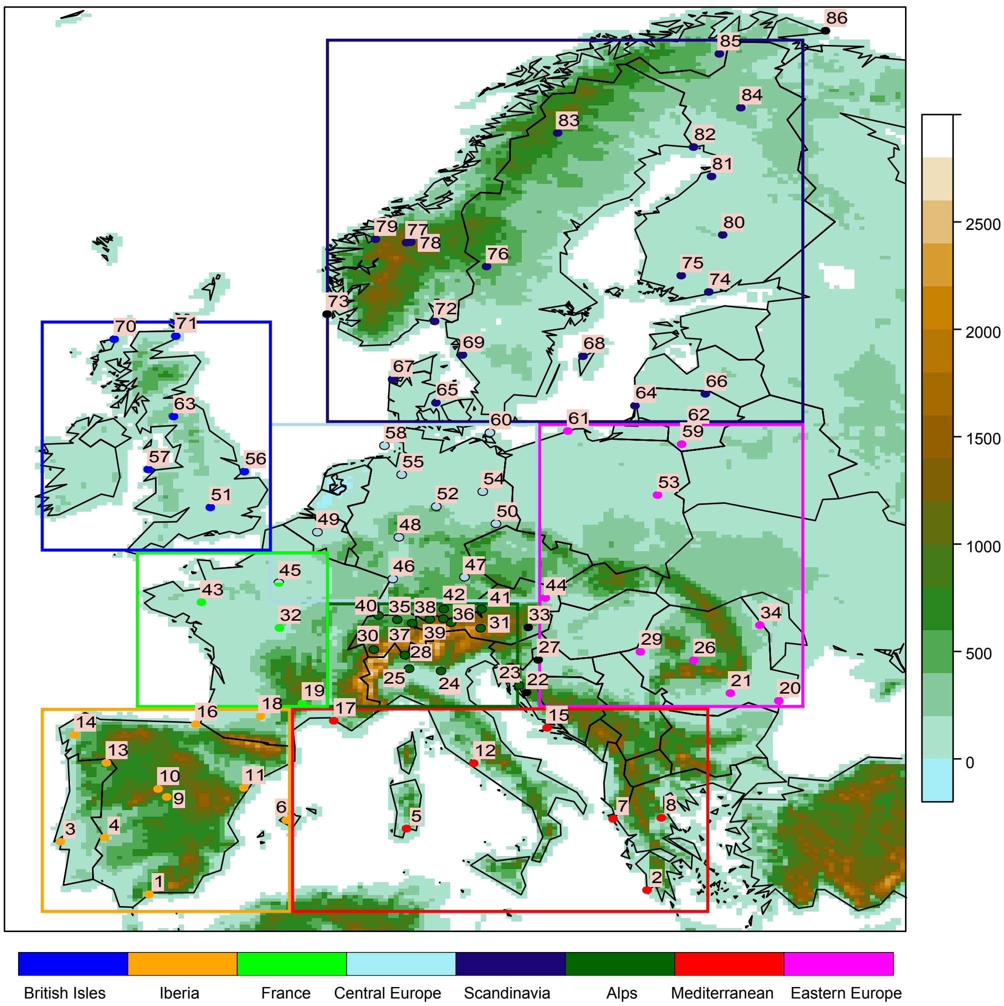

Figure 1: Location of the 86 stations used in the paper, sorted according to latitude (see Table 1). Colors represent the orography (for the EURO-CORDEX 0.11° resolution grid, in meters). The colored boxes (and circles) show the eight PRUDENCE sub-regions (and the corresponding stations); the legend at the bottom of the figure indicates the names of the different regions.

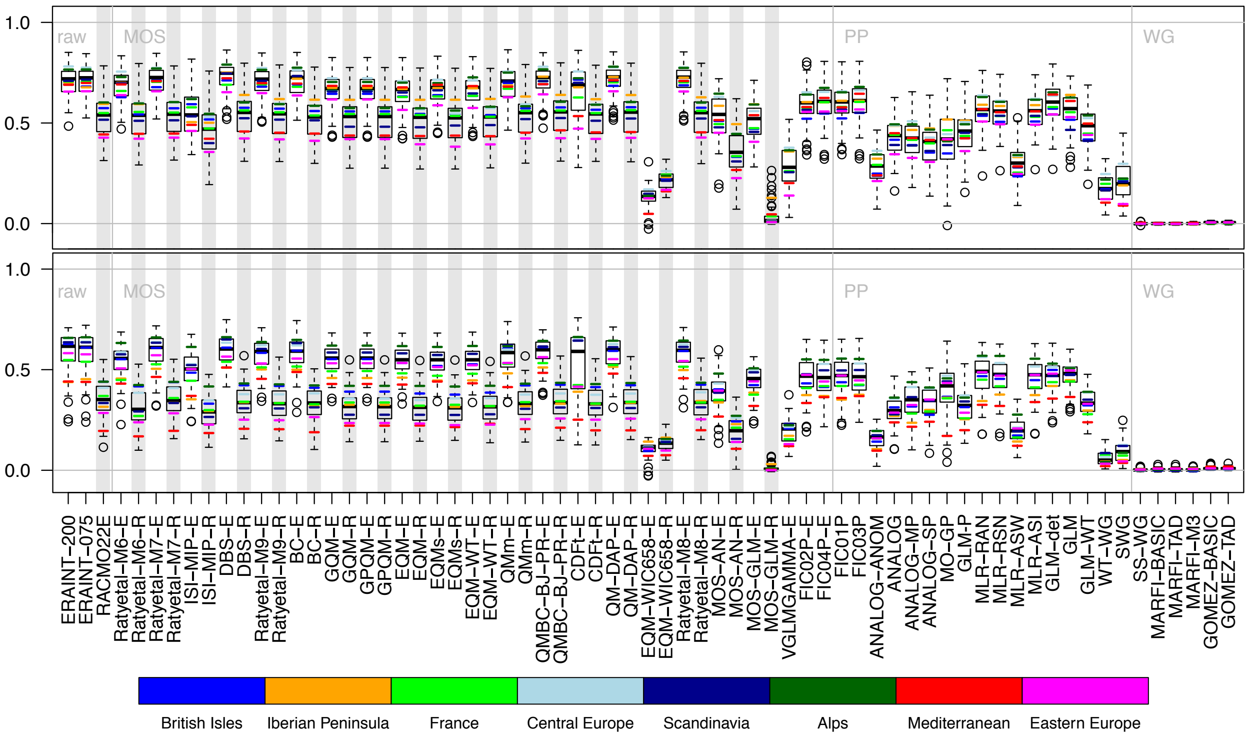

Figure 2: Spearman correlation of downscaled and observed daily precipitation for winter (DJF, top) and summer (JJA, bottom). For each method, the box-whisker-plot summarizes the results of the 86 stations. Boxes span the 25-75% range and the whiskers the minimum/maximum value (within 1.5 times the interquartile range); outliers are plotted individually. Average results over the different Prudence regions are indicated by a colored horizontal bar for each method (see the colors in the bottom legend). Shading indicates the MOS results using RACMO2 predictors (all others use ERA-Interim). The methods are sorted as in Table 3.

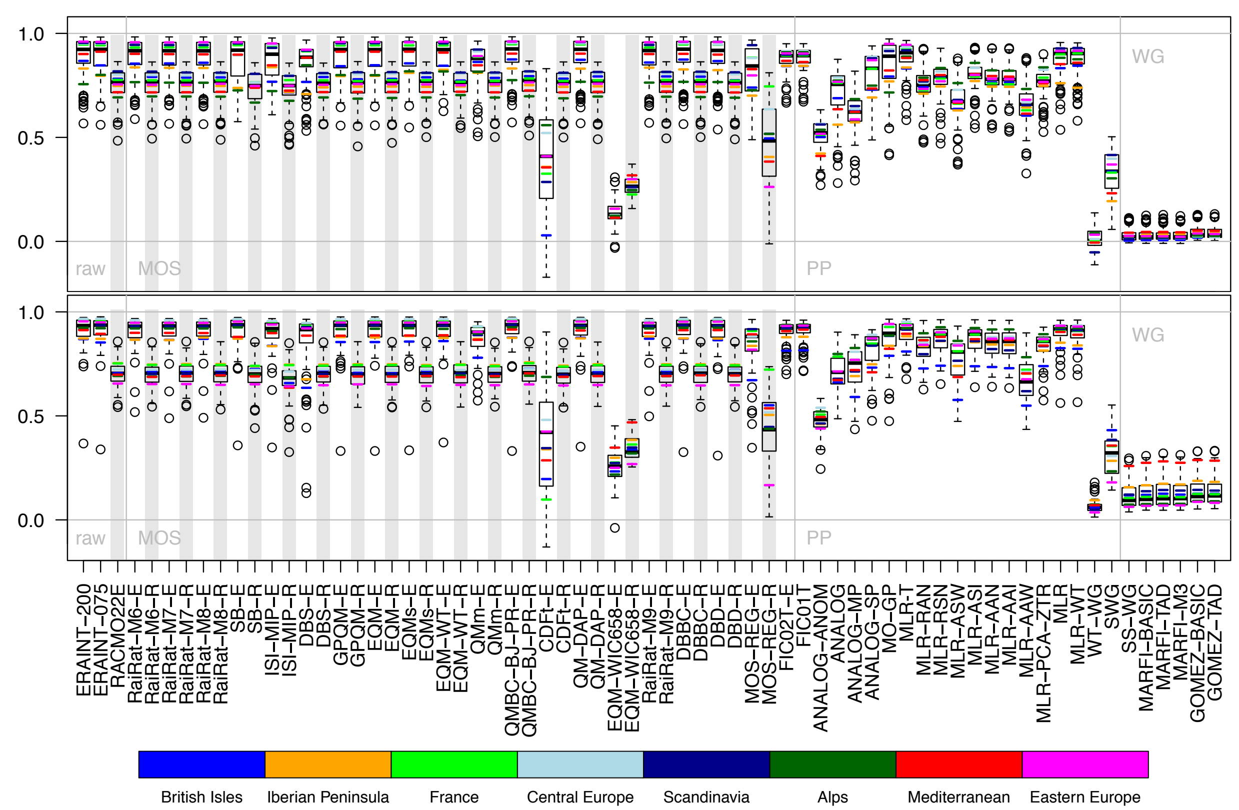

Figure 3: As Figure 2 but for Pearson correlation of downscaled and observed daily maximum temperature.

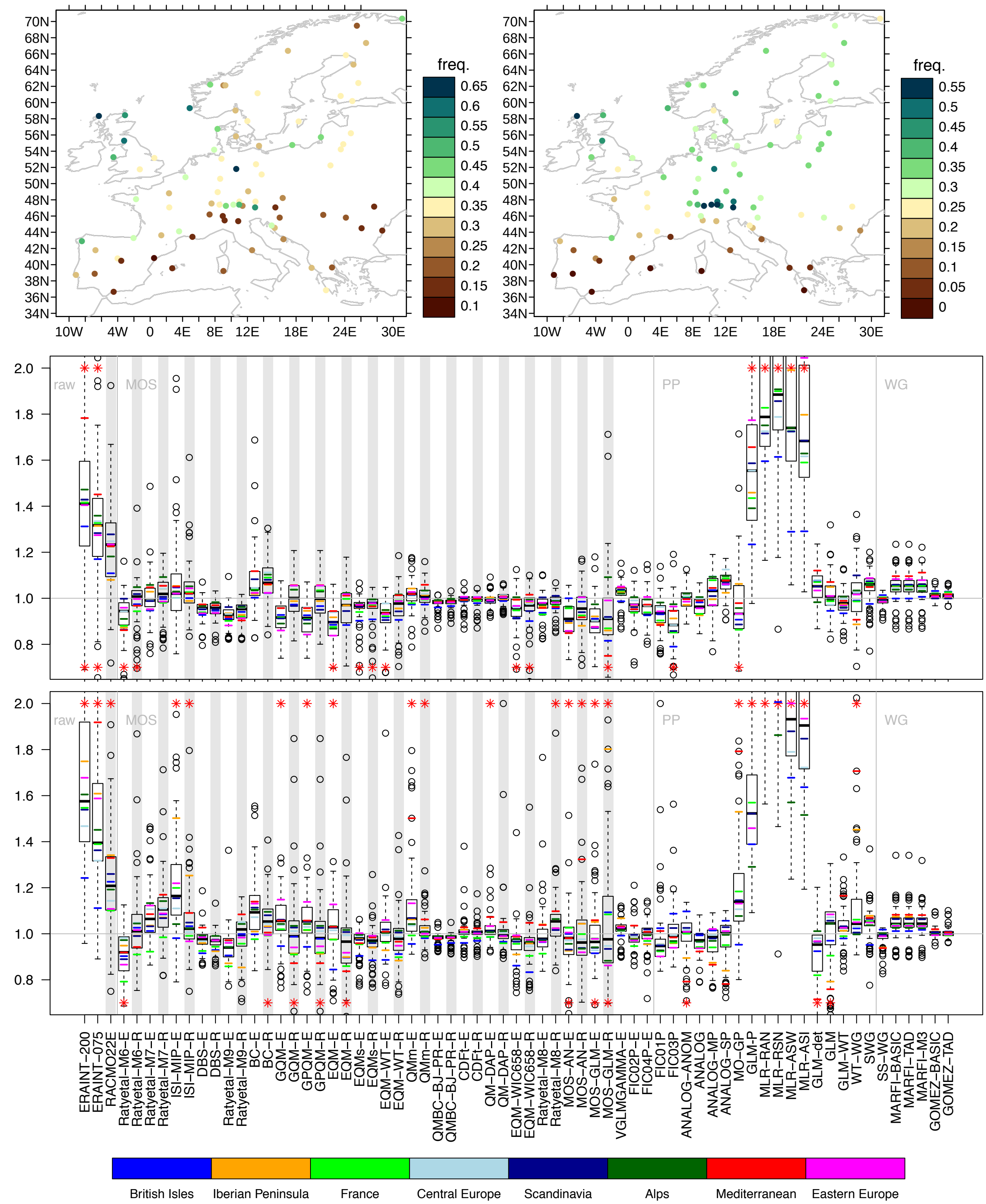

Figure 4: Observed climatological R01 (relative wet-day frequency) values for winter (DJF, top left) and summer (JJA, top right). The biases of the downscaling methods (Table 3) are shown in the middle and bottom panels, for winter and summer, respectively. For each method, the box-whisker-plot summarizes the results of the 86 stations. Boxes span the 25-75% range and the whiskers the maximum value (within 1.5 times the interquartile range); outliers are plotted individually. A red asterisk indicates that values lie outside the plotted range. Average results over the different Prudence regions are indicated for each method (see the colors in the bottom legend). Shades indicate the MOS results using RACMO2 predictors (all others use ERA-Interim). The methods are sorted as in Table 3 (first the raw model outputs, followed by MOS, PP and WG methods).

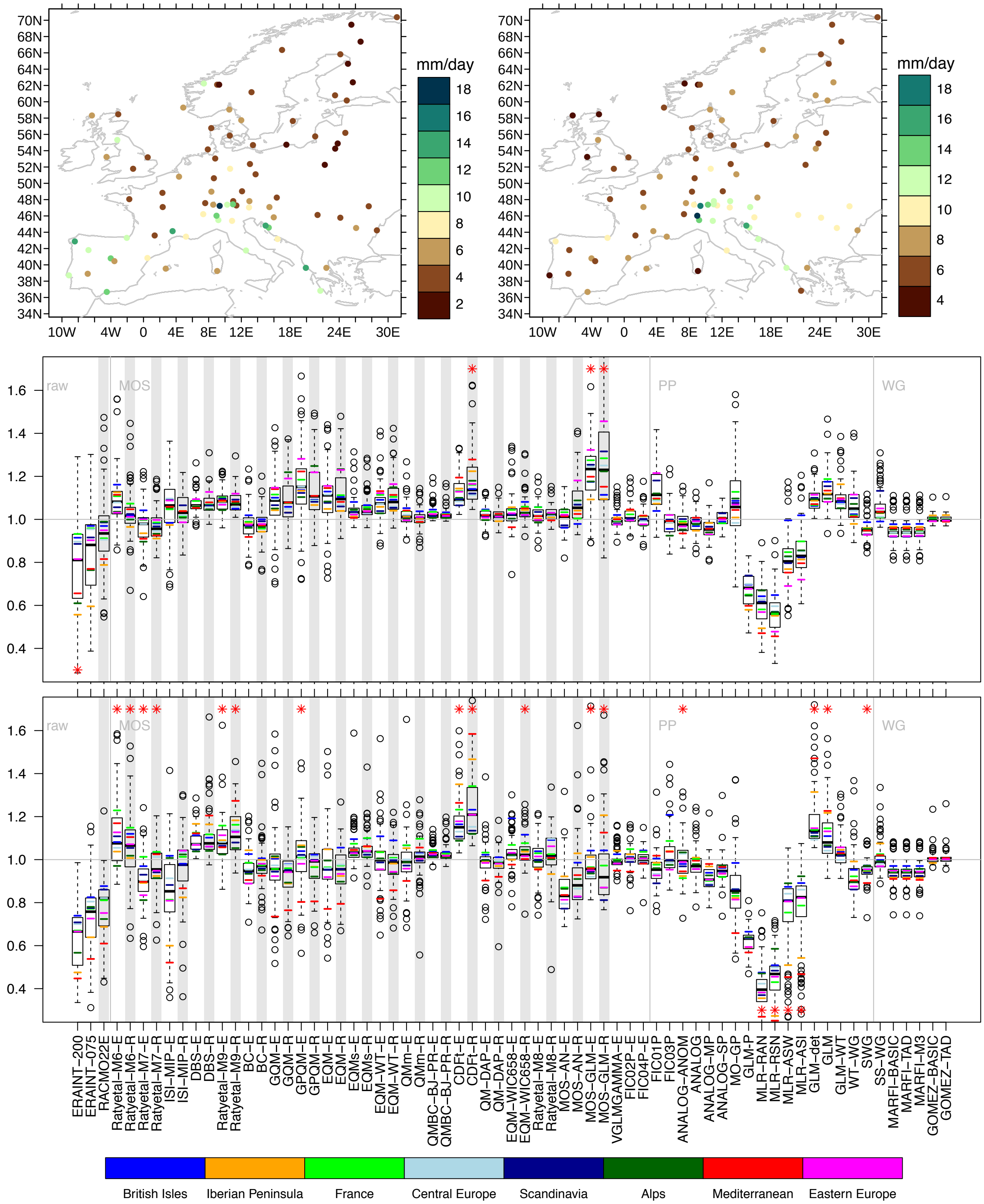

Figure 5: As figure 4, but for SDII (mean wet-day precipitation).

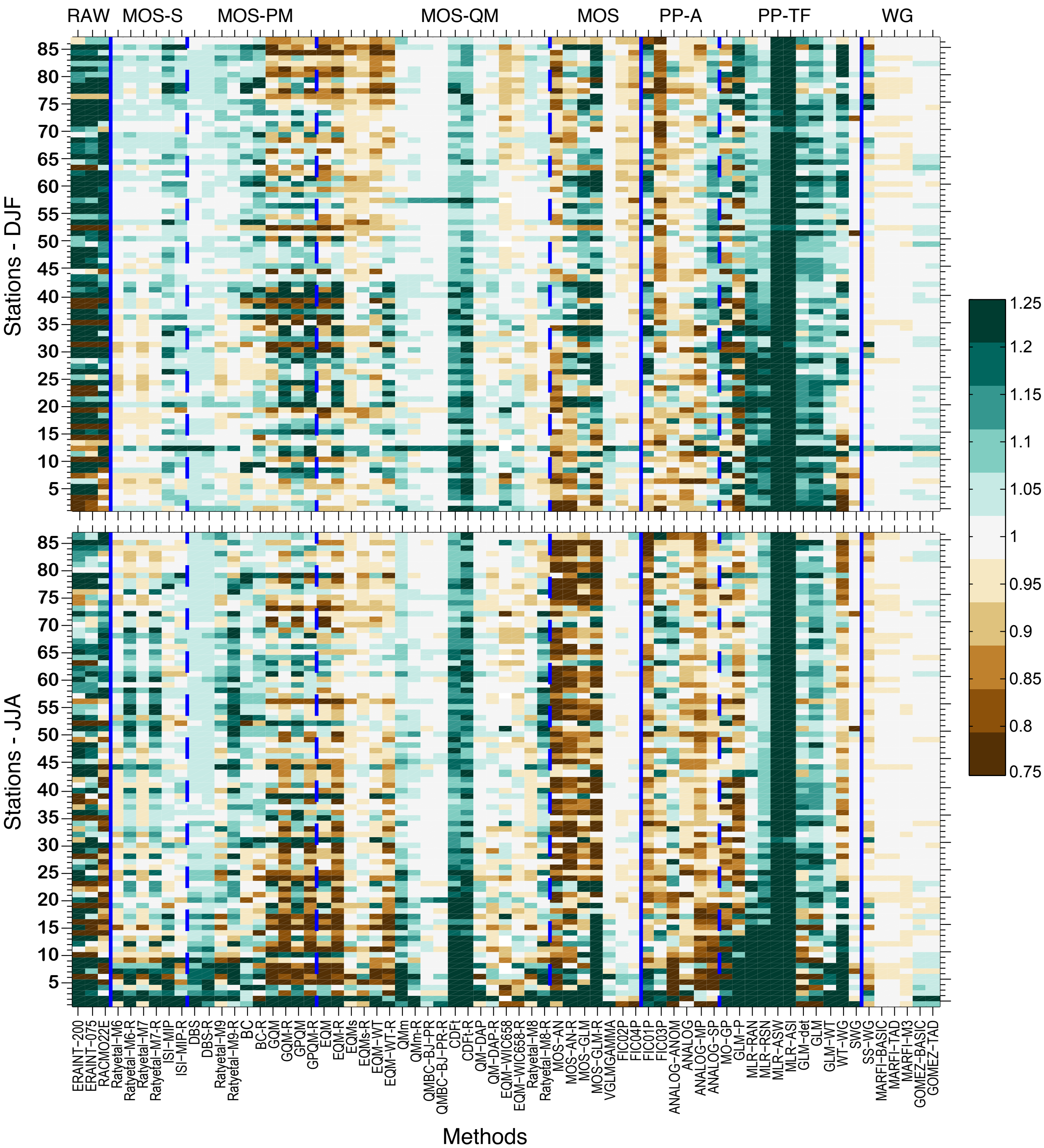

Figure 6: Individual station results (sorted as in Table 1) for total precipitation (PRCTOT) biases for winter (DJF, top) and summer (JJA, bottom). Vertical dashed lines separate the different approaches and techniques (see Table 3).

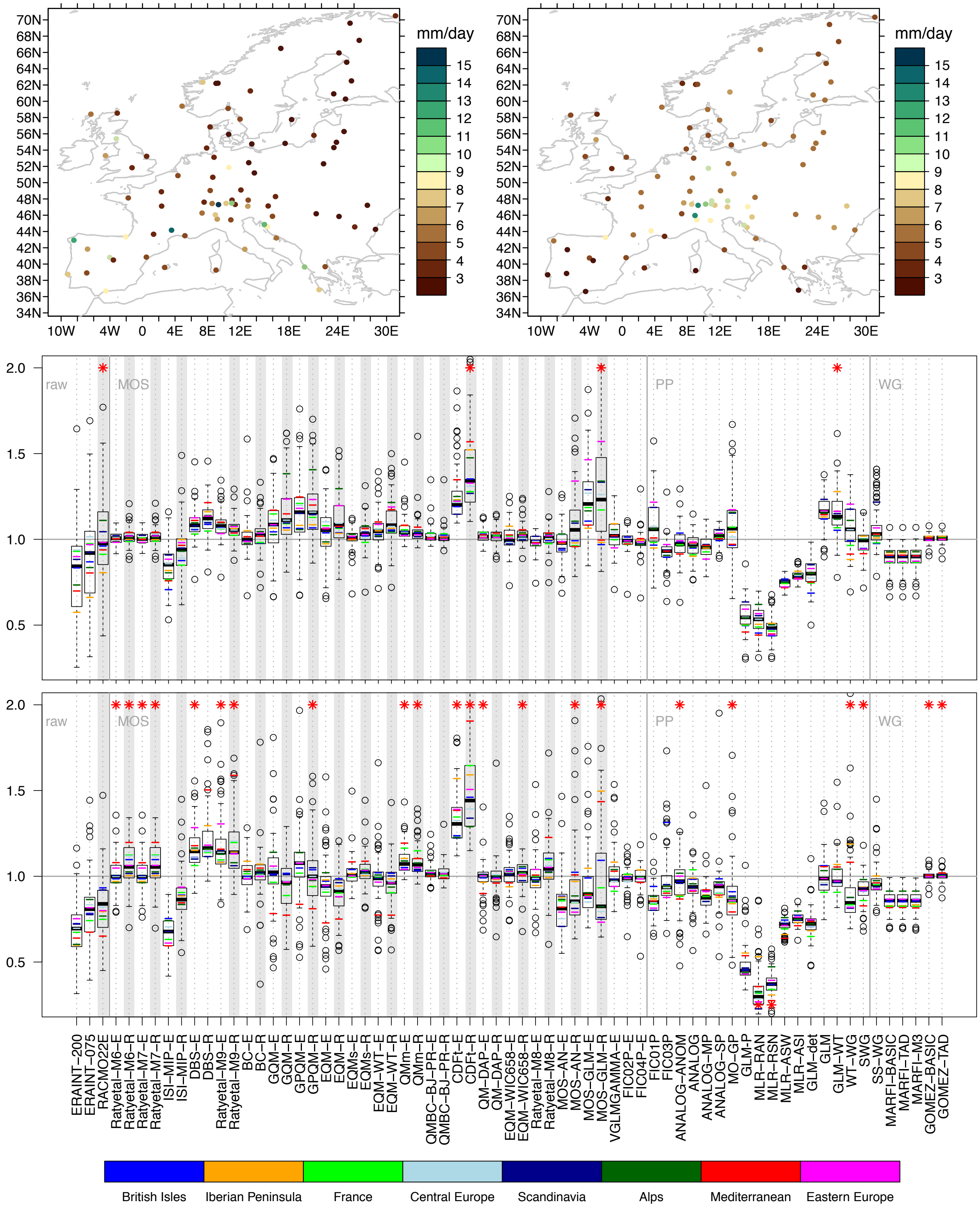

Figure 7: As figure 4, but for the standard deviation of daily precipitation for winter (DJF, top left) and summer (JJA, top right). Relative standard deviation biases (predicted over observed deviations) are shown in the middle and bottom panels, for winter and summer, respectively.

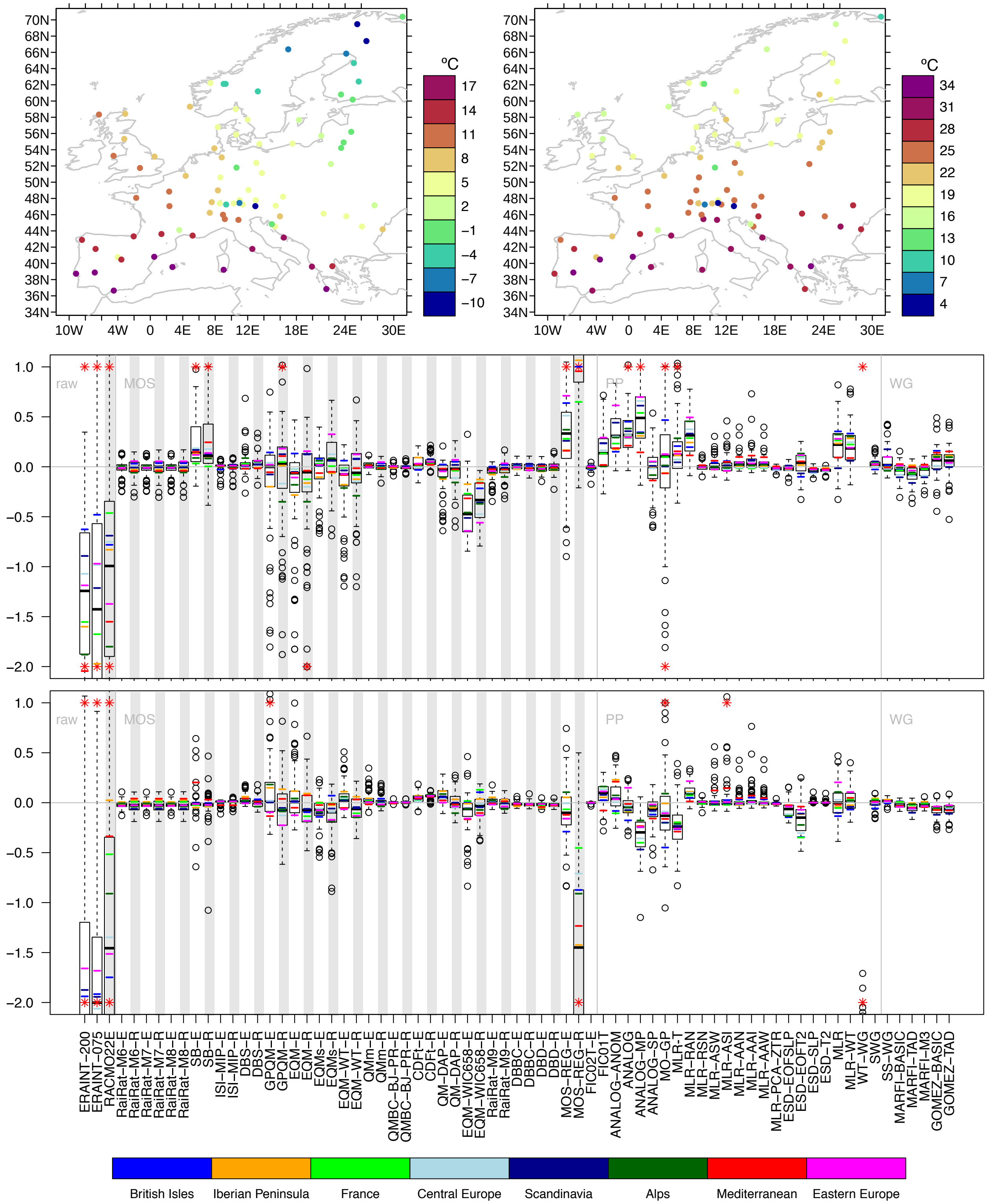

Figure 8: Observed mean climatologies (deg C) of daily maximum temperature for winter (DJF, top left) and summer (JJA, top right). The biases of the downscaling methods (Table 4) are shown in the middle and bottom panels, for winter and summer, respectively. For each method, the box-whisker-plot summarizes the results of the 86 stations. Boxes span the 25-75% range and the whiskers the maximum value (within 1.5 times the interquartile range); outliers are plotted individually. A red asterisk indicates that values lie outside the plotted range. Average results over the different Prudence regions are indicated for each method (see the colors in the bottom legend). Shades indicate the MOS results using RACMO2 predictors (all others use ERA-Interim).

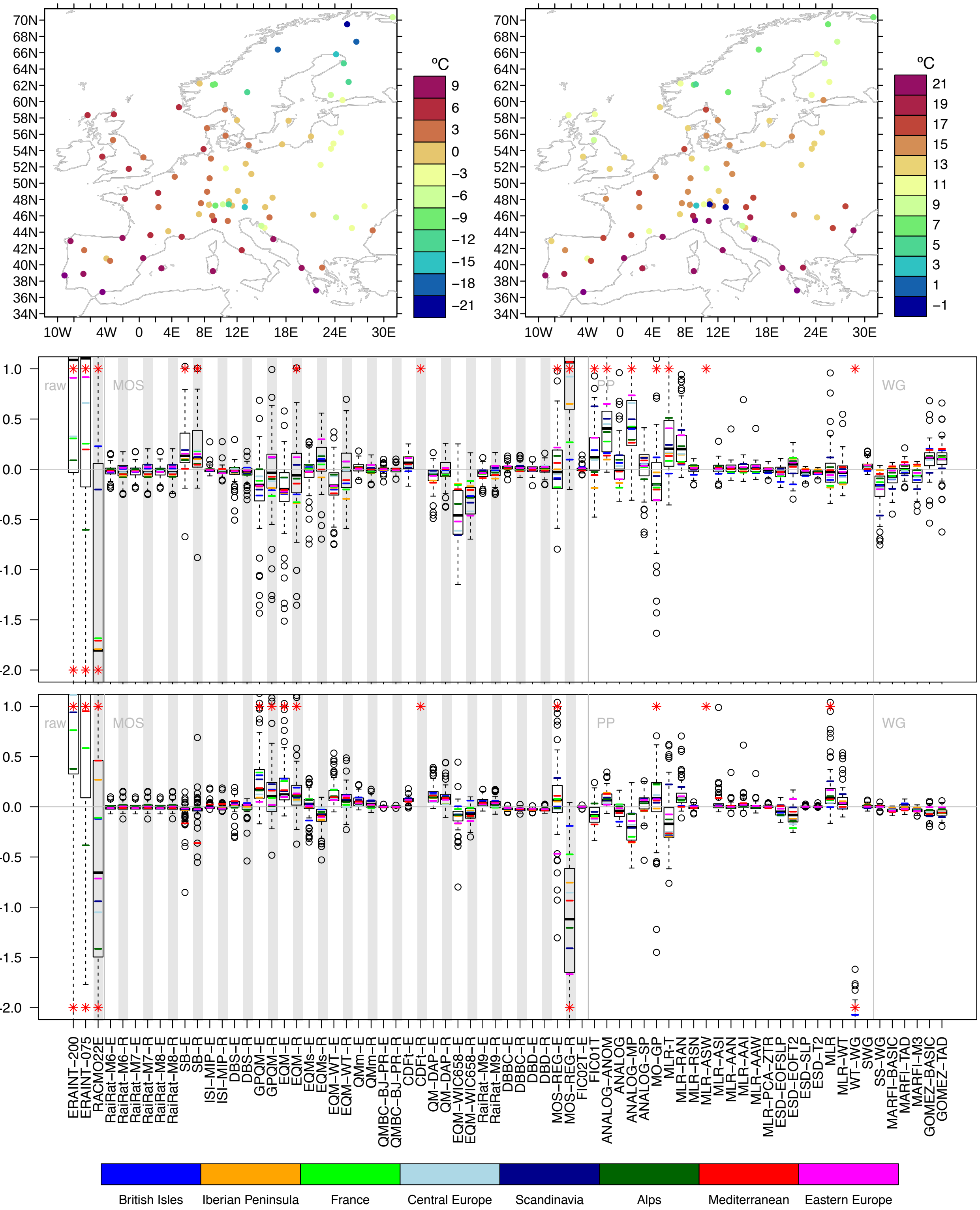

Figure 9: As Figure 8, but for daily minimum temperature.

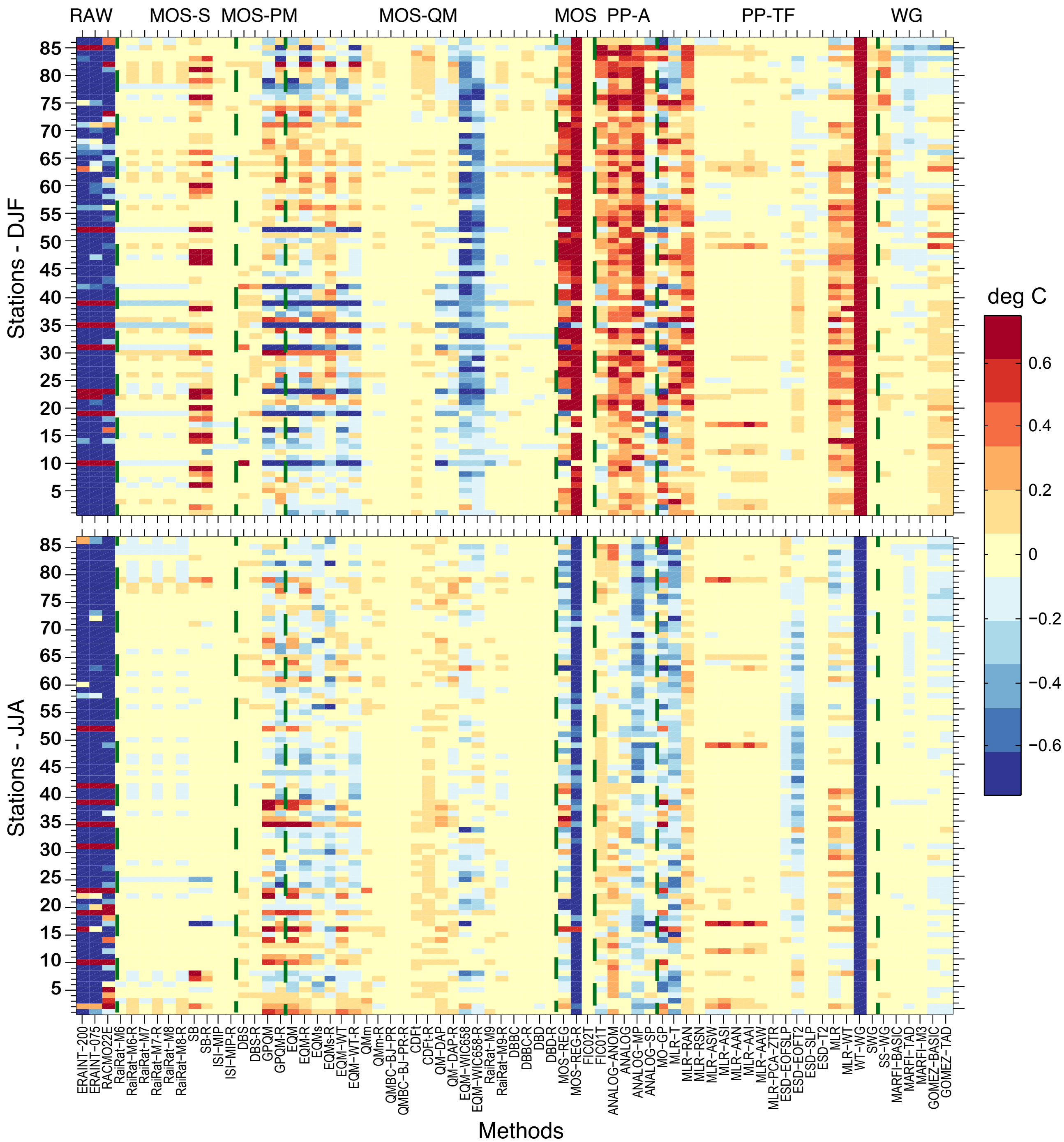

Figure 10: Individual station results (sorted as in Table 1) for daily maximum temperature biases for winter (DJF, top) and summer (JJA, bottom). Vertical dashed lines separate the different approaches and techniques (see Table 4).

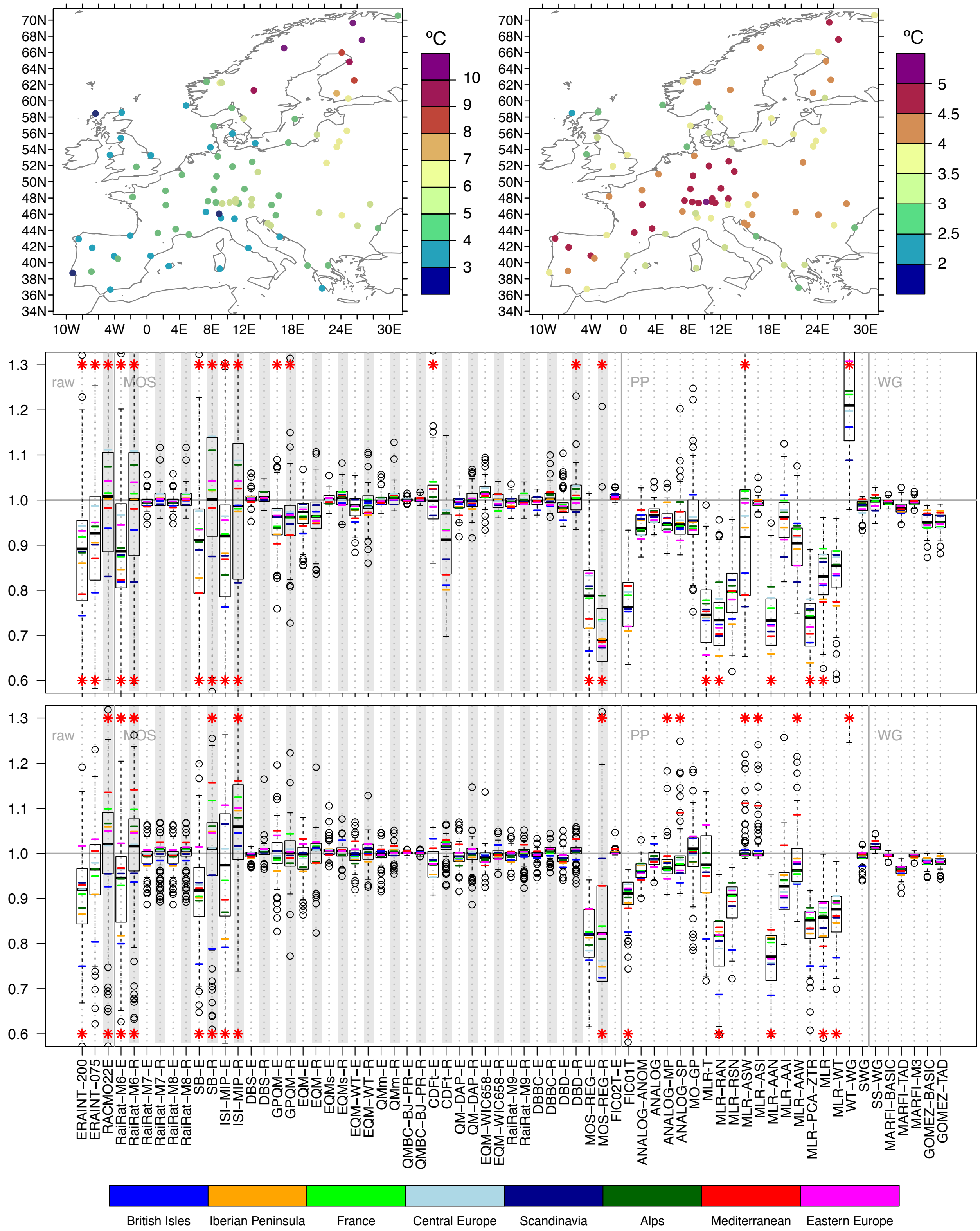

Figure 11: As Figure 8, but for the standard deviation of daily minimum temperature for winter (DJF, top left) and maximum temperature for summer (JJA, top right). Relative standard deviation biases (predicted over observed values) are shown in the middle and bottom panels, for winter and summer, respectively.