INTERNATIONAL JOURNAL OF CLIMATOLOGY

The VALUE perfect predictor experiment: evaluation of temporal variability

Journal:	*International Journal of Climatology*
Manuscript ID	JOC-16-0593.R1
Wiley - Manuscript type:	VALUE special issue
Date Submitted by the Author:	17-May-2017
Complete List of Authors:	Maraun, Douglas; University of Graz, Wegener Center for Climate and Global Change Huth, Radan; Charles University, Faculty of Science, Dept. of Physical Geography and Geoecology; Institute of Atmospheric Physics, Dept.of Climatology Gutiérrez, José; National Research Council (CSIC), Instituto de Física de Cantabria; San Martin, Daniel; Predictia Intelligent Data Solutions SL, N.A. Dubrovsky, Martin; Institute of Atmospheric Physics, Dept.of Climatology Fischer, Andreas; Federal Office of Meteorology and Climatology (MeteoSwiss), Climate Services Hertig, Elke; University of Augsburg, Institute for Geography Soares, Pedro; Instituto Dom Luiz,Universidade de Lisboa, DEGGE Bartholy, Judit; Eotvos Lorand Tudomanyegyetem, Department of Meteorology Pongracz, Rita; Eotvos Lorand Tudomanyegyetem, Department of Meteorology Widmann, Martin; University of Birmingham, School of Geography, Earth and Environmental Sciences Casado, María; AEMET, Desarrollo y Aplicaciones Ramos, Petra; Delegacion Territorial de AEMET en Andalucía, Ceuta y Melilla, N.A. Bedia, Joaquin; Predictia Intelligent Data Solutions SL, N.A.
Keywords:	Regional climate, Downscaling, Evaluation, Validation, Temporal variability, Spells, Interannual variability, long-term trends

The VALUE perfect predictor experiment: evaluation of temporal variability

Douglas Maraun[1], Radan Huth[2,3], Jose M. Gutierrez[4], Daniel San Martin[5], Martin Dubrovsky[3], Andreas Fischer[6], Elke Hertig[7], Pedro M. Soares[8], Judit Bartholy[9], Rita Pongracz[9], Martin Widmann[10], Maria J. Casado[11], Petra Ramos[12] and Joaquin Bedia[5]

[1] Wegener Center for Climate and Global Change, University of Graz, Brandhofgasse 5, 8010 Graz, Austria
[2]Dept. of Physical Geography and Geoecology, Faculty of Science, Charles University; Albertov 6, 128 43 Praha 2, Czech Republic
[3] Institute of Atmospheric Physics Czech Academy of Sciences, Bocni II 1401, 141 31 Prague, Czech Republic
[4] Institute of Physics of Cantabria (IFCA), University of Cantabria, Avenida de los Castros, Santander 39005, Spain
[5] Predictia Intelligent Data Solutions SL, Avda. los Castros s/n, Building I+D S345, 39005, Santander, Spain
[6] Federal Office of Meteorology and Climatology MeteoSwiss, Operation Center 1, 8085 Zurich-Airport, Switzerland
[7] Institute of Geography, Augsburg University, Alter Postweg 118, 86159 Augsburg
[8] Instituto Dom Luiz, Faculdade de Ciencias, Universidade de Lisboa, 1749-016 Lisbon, Portugal
[9] Dept. of Meteorology, Eotvos Lorand University, Pazmany st. 1/a, H-1117 Budapest, Hungary
[10] School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham, B15 2TT, UK
[11] Agencia Estatal de Meteorologia (AEMET), C/ Leonardo Prieto Castro, 8 Ciudad Universitaria, 28040 Madrid, Spain
[12] Delegacion Territorial de AEMET en Andaluca, Ceuta y Melilla, Avda. Americo Vespucio, n 3 bajo., 41092 Sevilla, Spain

May 17, 2017

Temporal variability is an important feature of climate, comprising systematic variations such as the annual cycle, as well as residual temporal variations such as short-term variations, spells and variability from interannual to long-term trends. The EU-COST Action VALUE developed a comprehensive framework to

evaluate downscaling methods. Here we present the evaluation of the perfect predictor experiment for temporal variability. Overall, the behaviour of the different approaches turned out to be as expected from their structure and implementation. The chosen regional climate model adds value to reanalysis data for most considered aspects, for all seasons and for both temperature and precipitation. Bias correction methods do not directly modify temporal variability apart from the annual cycle. However, wet day corrections substantially improve transition probabilities and spell length distributions, whereas interannual variability is in some cases deteriorated by quantile mapping. The performance of perfect prognosis statistical downscaling methods varies strongly from aspect to aspect and method to method, and depends strongly on the predictor choice. Unconditional weather generators tend to perform well for the aspects they have been calibrated for, but underrepresent long spells and interannual variability. Long-term temperature trends of the driving model are essentially unchanged by bias correction methods. If precipitation trends not well simulated by the driving model, bias correction further deteriorates these trends. The performance of PP methods to simulate trends depends strongly on the chosen predictors.

1 Introduction

Downscaling is a common - often necessary - step in assessing regional climate change and its impacts: the resolution of global coupled atmosphere-ocean general circulation models (GCMs) is typically too coarse to represent many regional- or local-scale climate phenomena. Therefore the output of GCMs is downscaled to provide high resolution simulations over a limited target area. The EU Cooperation in Science and Technology (COST) Action ES1102 VALUE was established to comprehensively evaluate different downscaling methods (Maraun et al., 2015). Three experiments have been defined: a so-called perfect predictor experiment to isolate downscaling skill in present climate; a GCM predictor experiment to evaluate the overall skill to simulate present-day regional climate; and a pseudo reality experiment to evaluate the skill of downscaling methods to represent future climates.

In a community effort, researchers from 16 European institutions participated in the perfect predictor experiment, and more than 50 different statistical downscaling methods have been evaluated at 86 stations across Europe. The evaluation comprises the representation of marginal aspects (such as the mean or variance; (Gutiérrez and coauthors, 2017)), temporal aspects (such as spell length distributions; this contribution), spatial aspects (such as spatial decorrelation lengths; (Widmann and coauthors, 2017)), and multivariable aspects (such as the relationship between temperature and precipitation; Page et al., in preparation). Extreme events as well as an evaluation conditional on relevant synoptic and regional phenomena have been, owing to their importance, considered separately by Hertig and coauthors (2016) and Soares and coauthors (2017). Here we present the evaluation of temporal aspects.

To illustrate different aspects of temporal variability, Figure 1 shows a selected year of precipitation at the participating rain gauge in Graz, Austria. On 18th of July (orange spike), several districts were flooded. The city's streams burst their banks following the heavy rainfalls prior to the event, but a major contributor was the long wet spell in the end of June (red

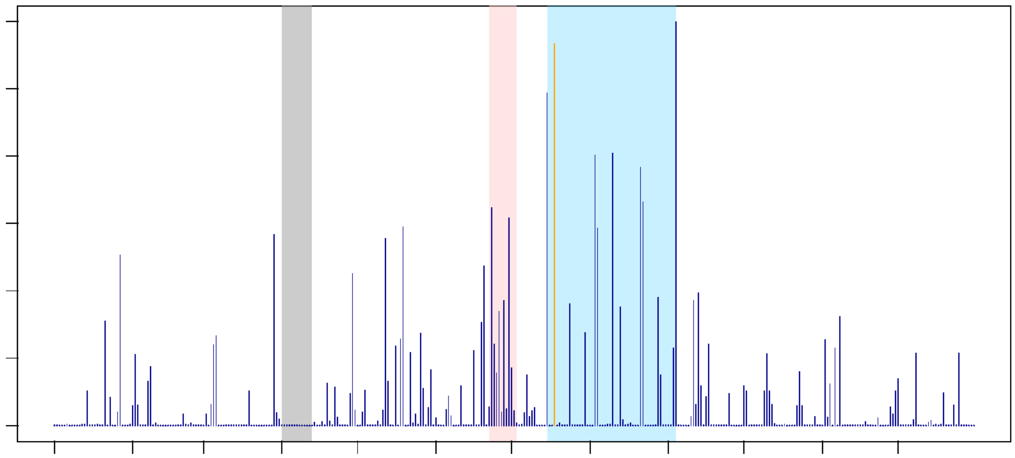

Figure 1: Daily precipitation totals in Graz, 2009. Shading: see text.

shading). Southeast of Graz, the overall event caused several thousand landslides. Total rainfall in June exceeded the climatological mean by more than 60%. Also annual rainfall was about 47% higher than normal (Klein Tank et al., 2002), indicating substantial interannual variability. A pronounced seasonality of all aspects of precipitation is directly apparent. In late winter and early spring, precipitation amounts are low compared to summer. Also the probability of consecutive wet days is low resulting in long dry spells (grey shading). Most dry-wet and wet-wet transitions occur in late spring and early summer, the highest rainfall amounts are observed in late summer (blue shading).

In general, temporal variability involves a wide range of time scales, from the diurnal cycle through day-to-day variations, spells (dry, wet, warm, cold, etc.), and interannual variations to long-term trends. The variability can be broadly separated into systematic variations - the diurnal and annual cycle as well as forced long-term trends - and residual temporal variations, whose characteristics are determined by the large-scale driving processes and by local memory. For instance, temporal dependence in precipitation may stem directly from memory caused by soil-moisture feedbacks, or indirectly from the duration of passing cyclones and anti-cyclones. Temporal aspects of local climate are often essential for impact studies in various sectors such as water (e.g., preconditions of flooding, Froidevaux et al. (2015); dry spells, Stoll et al. (2011)), agriculture (e.g., dry spells Calanca (2007); seasonality, Rosenzweig et al. (2001)), health (Semenza et al., 1996, e.g., heatwaves) and energy (Rosenzweig et al., 2011, e.g., seasonality).

In VALUE we evaluate the performance of different downscaling methods to represent temporal variability. Apart from dynamical downscaling with regional climate models (RCMs, Rummukainen, 2010), different statistical approaches exist (Fowler et al., 2007; Maraun et al., 2010; Wilks, 2010; Maraun, 2016): perfect prognosis (PP) statistical downscaling methods, which are calibrated purely on observations and typically take their predictors from large-scale fields of the free atmosphere; model output statistics (MOS) methods, which are calibrated between model data and observations (in climate science, these are typically bias correction methods); and unconditional weather generators, which are calibrated on local data and do not include any meteorological predictors.

The basic driver of the residual, regional-scale temporal variability is the propagation of planetary and synoptic waves, which is essentially prescribed by GCMs. This continental-scale variability is modulated by regional-scale dynamical processes, influences of the orography, and feedback mechanisms such as soil-moisture-temperature, soil-moisture-precipitation feedbacks and snow-albedo feedbacks (Schär et al., 1999; Seneviratne et al., 2006; Fischer et al., 2007; Hall et al., 2008). As a result, regional-scale temporal variability simulated by RCMs may diverge from the prescribed large-scale variability (Alexandru et al., 2007). Local temporal variability is often - in particular for precipitation and wind - not fully determined by larger-scale variability, but exhibits additional - essentially random - fluctuations. PP statistical downscaling inherits the variability of the large-scale predictors and typically does not add any local short-term variations. Some methods, however, explicitly model local variability by randomisation (von Storch, 1999; Chandler and Wheater, 2002; Volosciuk et al., 2017). Such stochastic models might simply generate white noise, but may also include weather generators (see below) to model short-term temporal dependence by Markov-chain-type components (Maraun et al., 2010). Also bias correction typically does not explicitly add local temporal variability to the driving model, but only subtly modulates temporal variability via its effect on the marginal distribution. For instance wet day frequencies are adjusted, which indirectly affects the representation of spells (Rajczak et al., 2016). Some bias correction methods also attempt to explicitly adjust the temporal structure (Vrac and Friederichs, 2015; Cannon, 2016, e.g.) but at the cost of destroying the temporal consistency with the driving dynamical model. Unconditional weather generators (i.e., weather generators that do not use meteorological predictors) do not provide sequences which are synchronised with the driving models. Instead, the only temporal structure they represent is explicitly modelled, typically by Markov chains (Maraun et al., 2010). Most statistical models - PP and MOS - have an explicit description of the annual cycle, e.g., by being calibrated to each calender day, month or season individually, or (in case of PP) by including the day-of-the year as predictor.

Of the temporal aspects studied in this paper, perhaps the annual cycle has been the most frequent target of validation: many RCM studies as well as studies of both kinds of statistical downscaling (PP and MOS) and of WGs include a validation of the annual cycle, although it usually is not their main topic (e.g. Frei et al., 2003; Moberg and Jones, 2004; Kilsby et al., 2007; Turco et al., 2011; Schindler et al., 2007; Soares et al., 2012; Warrach-Sagi et al., 2013; Kalognomou et al., 2013; Martynov et al., 2013; Keller et al., 2015; Favre et al., 2016). Also studies evaluating precipitation (dry/wet) spells and precipitation transition probabilities (wet/wet, dry/wet) as well as interannual variability have been relatively numerous (e.g. Semenov et al., 1998; Charles et al., 1999; Giorgi et al., 2004; Kilsby et al., 2007; Jacob et al., 2007; Schmidli et al., 2007; Frost et al., 2011; Bürger et al., 2012; Turco et al., 2011; Hu et al., 2013; Gutmann et al., 2014; Keller et al., 2015; Rajczak et al., 2016). Much less attention has, on the other hand, been paid to validation of temperature spells and day-to-day temperature changes; only a few studies have been published that focus on these characteristics (Huth et al., 2001; Bürger et al., 2012; Vautard et al., 2013; Huth et al., 2015; Lhotka and Kyselý, 2015).

The vast majority of validation studies addressing also temporal issues focused on a single downscaling approach or, at best, provide a comparison for models from one family such as Kotlarski et al. (2014); Gutmann et al. (2014). Exceptions are Wilby et al. (1998), who where the first to systematically evaluate temporal aspects in PP methods and unconditional weather generators; the STARDEX project, which assessed temporal aspects of extreme events in PP and a simple MOS method (Haylock et al., 2006; Goodess et al., 2010); the study by Frost et al. (2011), who compared the representation of spell lengths and interannual variability in an RCM, a bias correction method, a PP method and two weather generators; the study by Hu et al. (2013), who carried out a similar intercomparison for a PP method and two weather generators; the study by Bürger et al. (2012), who compared extreme spells in several PP and MOS methods; and the recent study by Huth et al. (2015), which investigated temporal aspects in both statistical and dynamical downscaling methods. But all these studies still include only a rather limited range of methods.

Even though extremely important for climate change studies (Pielke and Wilby, 2012), evaluation studies of trends in downscaled data are scarce (Benestad and Haugen, 2007; Lorenz and Jacob, 2010; Bukovsky, 2012; Ceppi et al., 2012; Huth et al., 2015). These studies broadly indicate a rather limited ability of downscaling methods to reproduce trends.

In brief, a substantial research gap exists. The performance of many downscaling and bias correction methods to represent temporal aspects - both individually and relative to each other - is largely unknown. This study takes a first step to close this gap. In a perfect predictor experiment we analysed the performance of one raw RCM and 48 statistical methods to represent day-to-day variability, spells, seasonality, interannual and long-term variability including trends. Aspects of temporal variability specifically addressing extreme events, such as long heatwaves or meteorological drought, are addressed in the companion paper on extreme events (Hertig and coauthors, 2016, in this issue). The considered experiment was conducted for daily values, hence we cannot evaluate sub-daily variations.

VALUE is a community effort, the participation in this experiment (and its evaluation) was unpaid. The participating methods thus form an ensemble of opportunity. In particular no systematic set of predictor variables or domains has been prescribed. Thus statements about optimal predictor choice are limited to a few comparisons of similar (or identical) methods with different predictors. A detailed set of metadata has, however, been collected for all participating methods. These meta data describe structural aspects of all methods and often allow for quite detailed interpretations of the individual performance. In the paper we will discuss selected examples in more detail, and additionally give a broad overview of the different model families. The metadata and complete results for individual methods are available from the VALUE portal `www.value-cost.eu/validationportal` for further investigation.

The aim of the perfect predictor experiment is to evaluate the isolated skill of the raw RCM and the statistical models. Consequently, this study cannot give a conclusive assessment of the skill to simulate regional future climates. The skill of a full regional modelling system, comprising the full modelling chain from GCM to RCM and/or statistical model, as well as the downscaling performance in future climates will be considered in additional experiments (Maraun et al., 2015).

In the following section we will briefly review the experimental setup, the considered diagnostics and the participating methods. In Section 3 we will present the results for different diagnostics and methods. An overall discussion of the results will follow in the final section.

2 Experiment, Diagnostics and Methods

The experimental design follows the VALUE perfect predictor experiment with station data as target. As (approximately) perfect predictors and perfect boundary conditions, we use ERA-Interim data from 1 Jan 1979 to 31 Dec 2008 (Dee et al., 2011). The MOS methods use ERA-Interim data at their native resolution of 0.75° as input, the PP methods ERA-Interim predictors at 2°, which resembles a typical GCM resolution. Furthermore, most MOS methods also use ERA-Interim, downscaled with the RCM RACMO (van Meijgaard et al., 2008), as input to represent a typical RCM bias correction situation. Apart from the resolution, some important differences between these two MOS settings exist: in the first case, internal variability at the grid-box scale is closely tied to real world internal variability, whereas the RCM develops its own internal variability within the RCM domain. Furthermore, observed temperatures have been assimilated into the ERA-Interim reanalysis; the resulting predictors are thus essentially bias free at the grid-box scale and differences with station observations mainly result from the scale gap. RCM temperatures inside the domain, however, are only mildly constrained by the boundaries and are thus typically affected by biases. Precipitation is in both cases calculated by model parameterisations, without any reference to observed precipitation. It is thus affected by scale-gap and biases.

As predictand data, time series from 86 stations from the publicly available ECA data base were used (Klein Tank et al., 2002). These stations were selected to cover the different European climates, covering mediterranean, maritime, continental, alpine and sub-polar climates. For details refer to Gutiérrez and coauthors (2017) and the supplementary information.

In this manuscript, we consider daily maximum and minimum temperature and daily precipitation only. A dedicated analysis of other variables will be carried out separately for a set of stations in Germany (Page et al., in preparation). For the statistical methods a five-fold cross validation with non-overlapping 6-year blocks is carried out. Further details about the protocol can be found in Maraun et al. (2015), Gutiérrez and coauthors (2017) and on `www.value-cost.eu/validation#Experiment_1a`.

Index	Variables	Performance measure	Resolution	Description
short-term variability				
ACF1	T_{max}, T_{min}	bias	seasonal	lag-1 autocorrelation
ACF2	T_{max}, T_{min}	bias	seasonal	lag-2 autocorrelation
WWprob	precipitation	bias	seasonal	probability of wet-wet transition
WDprob	precipitation	bias	seasonal	probability of wet-dry transition
Spells				
WarmSpellMean	T_{max}	bias	seasonal	mean of the warm (> 90th percentile) spell length distribution
ColdSpellMean	T_{min}	bias	seasonal	mean of the cold (< 10th percentile) spell length distribution
WetSpellMean	precipitation	bias	seasonal	mean of the wet (≥ 1mm) spell length distribution
DrySpellMean	precipitation	bias	seasonal	mean of the dry (< 1mm) spell length distribution
Interannual to long-term variability				
VarY	T_{max}, T_{min}, precipitation	rel. error	seasonal	variance of seasonally/annually averaged data
Cor.1Y	T_{max}, T_{min}, precipitation	bias	seasonal	correlation with observations of seasonally/annually averaged data
Cor.7Y	T_{max}, T_{min}, precipitation	correlation	seasonal	correlation with observations of seasonally/annually averaged and filtered dat
Trend	T_{max}, T_{min}, precipitation	trends themselves	seasonal	long-term (relative) trend of seasonally/annually averaged data
Annual cycle				
AnnualCycleAmp	T_{max}, T_{min}	bias	annual	Amplitude of the annual cycle
AnnualCycleRelAmp	precipitation	rel. error	annual	Relative amplitude of the annual cycle
AnnualCyclePhase	T_{max}, T_{min}	circular bias	annual	Phase of highest peak [2]

Table 1: Diagnostics considered. Diagnostics only shown in the supplementary information are plotted in grey. For details see `http://www.value-cost.eu/validationportal/app#!indices` and click on "details" for the underlying R-Code (note that registration is required).

Table 1 lists the diagnostics we considered: the indices to measure a specific aspect of temporal variability, the corresponding performance measure to quantify the mismatch with observations and the temporal resolution (seasonal, annual) at which the evaluation has been carried out. In two cases, we assessed correlations between observed and downscaled local time series, namely at the interannual and seven year time scales. In this case, the diagnostic consists of a performance measure - the correlation - only.

Detailed descriptions of these diagnostics can be found in the supplementary information. The code used to calculate these diagnostics is available from
`http://www.value-cost.eu/validationportal/app#!indices` (registration required).

In this analysis, we compare methods from the PP, MOS and unconditional weather generator approaches with raw ERA-Interim output, and dynamically downscaled ERA-Interim. Tables 2 and 3 list the methods participating in the experiment (many methods are identical for the different variables, but in several cases differences exist in the implementation for different variables. Therefore, we decided not to list the methods in a single table). The MOS methods are listed prior to the PP methods to ease comparison with the raw RCM and ERA-Interim data.

PP methods are calibrated purely on observed predictors and predictands. The statistical model is then applied to climate model predictors. In a climate change context, the approach is based on three major assumptions Maraun and Widmann (2018): first, that the GCM predictors are perfectly simulated (hence the name) in present and future climate. As a consequence, predictors are typically taken from large-scale fields of the free atmosphere. Second, the predictors should be informative of local variability and climate change. And third, the model structure should well describe local variability, and allow for at least moderate extrapolations under climate change. Our evaluation experiment employs perfect predictors to isolate downscaling skill in present climate. It can therefore be used to assess whether the chosen predictors are informative of local variability and observed changes, and whether the model structure well describes observed local variability and changes. The perfect prognosis assumption and performance under future climate change, however, cannot be assessed.

The participating PP methods broadly represent widely used approaches - analogue, regression and weather-type methods. Some of regression methods apply variance inflation (MLR-ASI, MLR-AAI, GLM-P), some are stochastic (see Tables). The ESD methods downscale at the monthly scale, thus no diagnostics are considered that involve daily values. The ESD-EOF implementation differs from the standard ESD version in that the predictand values are filtered by PCA Benestad et al. (2015b).

All stochastic methods use, conditionally on the predictors, independent noise, i.e., they do not have an explicit Markov component implemented to simulate short-term persistence. For precipitation, some of the participating PP methods have been included for illustrative purposes only (MLR-RAN, MLR-RSN, MLR-ASW, MLR-ASI). In fact, it is well known that simple multiple linear regression methods are not suitable to model daily precipitation. Yet they do participate in the intercomparison to highlight the problems associated with them (marked in grey in Table 3). Two of the stochastic methods (GLM and SWG) are based on generalised linear models, with a logistic regression for the occurrence process, and a generalised linear regression on the gamma distribution parameters for the amounts process. GLM-WT and WT-WG condition the distribution parameters for occurrence and amounts on weather types.

MOS methods are calibrated between model simulations and observations. The approach can thus in principle adjust biases (in fact, in climate science, these are almost exclusively bias correction methods, i.e., predictor and predictand have the same physical dimension), but has to be calibrated individually to the chosen model. MOS is based on three major assumptions (which make up the so-called stationarity assumption), similar to those of the PP approach Maraun and Widmann (2018) : first, the predictors have to be credibly (but not necessarily bias free) simulated. Second, the predictors need to be representative of the local variable. And third, as in PP, the structure of the transfer function needs to be suitable. Again, the first assumption cannot be tested with perfect predictors, only the second and third, and only for present day climate.

The participating MOS methods comprehensively span the range of widely used methods, and also cover some more experimental recent developments such as stochastic bias correction (VGLMGAMMA Wong et al., 2014). None of the participating MOS methods modifies residual temporal dependence directly, but only indirectly via changes in the marginal distribution. The CDFt method calibrates a statistical distribution also in the validation period. As this is only 6 years in our experiment (in a climate change experiment, one would typically use a 30 year time slice), we expect a broad spread for the resulting performance measures due to sampling variability.

Unconditional weather generators are not conditioned on meteorological predictors, but stochastically simulate marginal and temporal aspects, sometimes also spatial. They are calibrated to observed weather statistics. Under climate change, the model parameters (or the observed weather statistics) are adjusted by so-called change factors derived from climate models. The underlying assumptions are thus similar to those for MOS Maraun and Widmann (2018): first, the change factors have to be credibly simulated, and all relevant change factors have to be included; second, the simulated change factors have to representative of local changes; and third, the model structure has to be suitable. In the chosen experiment, no change factors are applied between calibration and validation period; thus only the suitability of the model structure can be evaluated. Some climatic statistics may have changed between calibration and validation period, but resulting systematic biases cancel out under cross-validation.

The SS-WG and MARFI unconditional weather generators are of the Richardson type Richardson (1981), i.e., they use a Markov chain to simulate precipitation occurrence, and an autoregressive model to simulate temperature. A major difference between the tow is the wet-day threshold: the SS-WG uses 1 mm, the MARFI models use 0.5 mm (note that the evaluation indices are in any case based on a 1 mm threshold). The GOMEZ weather generators are based on resampling.

Diagnostics have been calculated for each method and each station. They can be downloaded from the VALUE portal (`www.value-cost.eu/validationportal/app#!validation`). For stochastic methods, an ensemble of 100 realisations have been uploaded. The performance measures have been derived for each realisation and then averaged across the ensemble.

When interpreting the evaluation results, it has to be acknowledged whether a specific index is calibrated or emerges from the model. For instance, a good representation of the annual cycle could result from including meteorological predictors that describe the annual cycle, or trivially from fitting a statistical model separately to each month. In particular, weather generators by construction resemple many marginal and temporal aspects. In this study, only spell lengths and interannual variability are not calibrated. In Tables 2 and 3 we therefore also list whether short-term dependence (AC) and seasonality (SE) are calibrated or not. For further details on the contributing methods see Gutiérrez and coauthors (2017) or the VALUE portal (`www.value-cost.eu/validationportal/app#!downscalingmethod`).

3 Results

Figure 2 illustrates selected temporal aspects for precipitation in Graz, Austria, and how corresponding model performance has been quantified in this study. The top panel shows the dry spell length distribution. Observations are shown in bold solid black, the results for five different statistical methods are shown in color. Methods in red and orange are MOS, in blue PP, and the method shown in magenta is an unconditional weather generator. One index that can be derived from the distribution is the mean spell length (which is quantified in this study for all the participating methods and all selected weather stations). Dashed vertical lines show this index for observations and statistical models. The performance of a model is given by the difference between the modelled and observed mean, i.e., the mean spell length bias. Similarly, the bottom panel shows the annual cycle of daily mean precipitation. Here, two indices are considered: first, the relative amplitude (for temperature the absolute amplitude) defined as the difference between maximum and minimum value (horizontal dashed lines), relative to the mean of these two values. Second, the phase of the annual cycle, defined as the day of the annual cycle maximum[4] (vertical dashed lines). The performance for the first is measured as the relative error between modelled and observed relative amplitude, for the second as the circular bias between modelled and observed phase (circular in the sense that the difference between, say, 31st of December and 1st of January is -1 day, not 364 days).

In the following, we present the results, separately for temperature and precipitation. To keep the number of figures at a reasonable level, we selected a suite of relevant diagnostics for short-term variability, spells, monthly to interannual variability, and the annual cycle. Often, only one season is shown, in case of temperature, only either daily minimum or maximum temperature. A more comprehensive catalog of plots can be found in the supplementary information. The figures for all diagnostics are organised similarly, see Fig. 3 as an example. In this example, one diagnostic is shown for daily maximum and minimum temperature. In the top row, the observed indices are shown - here auto-correlation of daily maximum (left) and minimum (right) temperatures. Note that correlations on interannual and 7-year time scales have no corresponding observed indices, consequently no maps are drawn. The two panels below show the performance measures for these indices (top: maximum temperature, bottom: minimum temperature). Each box-whisker-plot represents one method: the raw driving data (ERA-Interim at the 2° resolution used as predictor for PP methods, at the native 0.75° resolution and the RACMO2 RCM), the MOS methods, the PP methods and the unconditional weather generators. The individual box-whisker-plots summarise the results for all 86 stations: the boxes give the 25%-75% range, the whiskers the maximum value within 1.5 times the interquartile range; values outside that range are plotted individually. The thick colored horizontal bars show the medians for the individual PRUDENCE regions (Christensen and Christensen, 2007). Note that the number of stations entering these calculations differs from region to region (ranging from 3 in France to 21 in Scandinavia, typically around 10). A red asterisk indicates that values lie outside the plotted range. Results for individual

[4]In some cases, the annual cycle of precipitation has two maxima. We will discuss below how the phase is defined in this case.

stations are - depending on the index - substantially affected by noise, but the median over all considered stations in general provides a robust estimate of the overall performance of a given method. Furthermore, the diagnostic is solely defined between observations and simulations, thus no observed indices exist.

For a given index, all methods are shown for which the index may sensibly be calculated. That is, methods producing only monthly output are not shown for any indices based on daily values. Otherwise, all indices are presented, even though a method might not be designed to reproduce them. Such results are not intended to denounce specific methods, but rather to highlight the consequences of using a method in such a context. These situations will be made explicit to avoid misinterpretation of the results.

As mentioned in the introduction, the methods participating in the experiment form an ensemble of opportunity. Also we have a list of candidate predictors for each method, but the actually selected set of predictors might be much lower for individual stations. To fully attribute differences in model performance to the approach, the particular implementation and the choice of predictors, dedicated sensitivity studies would be required. In many cases, conclusions may be drawn for groups of methods. For instance, all analog methods often behave similarly independent of the different predictors and implementations. Thus, conclusions about analog-type methods as a whole can often be drawn. A discussion of differences within this type, however, would be very speculative, because the individual methods often differ both in the implementation and choice of predictors. The level of detail in our interpretation will thus differ from case to case. In some cases, any discussion would be too speculative - we then restrict ourselves to a description of the findings.

3.1 Temperature

short-term variability Figure 3 shows the results for lag-1 autocorrelation of summer daily maximum and minimum temperature as a measure of short-term persistence. The top row shows observations for daily maximum (left) and minimum (right) temperature. The corresponding plots for winter can be found in the supplementary information. For T_{max}, summer persistence is relatively evenly distributed across Europe; for T_{min}, persistence is notably lower over many regions. The bottom panels show the performance of the individual models.

The spatial averaging of ERA-Interim results in a moderate overestimation of summer persistence of T_{max} (upper panel), these biases are reduced by the RCM. Almost all MOS methods inherit the skill of the predictor data set, in particular the added value of the RCM. The regression based MOS method (MOS-REG) includes averaging across several grid boxes and thus overestimates persistence. All analog methods underestimate persistence of temperature. The reason might be twofold: first, the spatial predictor variability might be strongest for circulation-based predictors. Thus, analogs may be selected that best constrain circulation (and in turn precipitation, see Section 3.2). And second, large-scale analogs might be sufficiently dissimilar at local scales to deteriorate day-to-day variations. Understanding this problem requires further detailed analysis. The ANALOG-ANOM method uses predictors defined at a continental scale, which likely explains the low performance.

As expected, all deterministic regression models overestimate persistence, as not all local variability is explained by large-scale predictors. This problem cannot be mitigated by inflated regression (MLR-ASI, MLR-AAI). All stochastic regression models randomise with white noise

(MLR-ASW, MLR-AAW; though conditional on the predictors) and thus underestimate persistence. The low performance of the SWG method may partly be explained by the use of continental-scale predictors in combination with a stochastic white-noise randomisation. The WT-WG method performs worst, as it is stochastic and additionally uses only sea level pressure as predictor. For the Iberian Pensinula and the UK, ERA-Interim overestimates summer persistence of T_{max}, the RCM reduces the bias. Conversely, for Eastern Europe ERA-Interim is almost bias free, but the RCM reduces persistence. This performance is again inherited by many statistical methods.

For T_{min} (lower panel), the performance is consistently worse for all approaches, whith a strong tendency to overestimate summer persistence. The RCM, however, performs slightly worse than ERA-Interim. The relative performance across most other methods is similar to that for T_{max}. The ISIMIP method, driven with ERA-Interim, is a notable exception - it has the lowest bias of all MOS methods. Most MOS methods leave the persistence bias essentially unchanged, the methods driven with reanalysis data have a lower bias, the methods driven with the RCM a higher. Interestingly, however, some QM-based bias correction methods moderately improve the representation of persistence indirectly by adjusting marginal distributions. The persistence of summer T_{min} is overestimated in the British Isles. But in contrast to the overall behaviour, this bias is reduced by the RCM (and again, this reduction is inherited by the MOS methods). The performance for most methods is best in the Alps.

Spells Overall, the performance to simulate spells is similar to the performance to simulate short-term variability. The results for summer temperature spells are shown in Figure 4, measured in terms of the mean spell length. Recall that temperature-related spells are not defined by exceedances of absolute thresholds (e.g., 30°C), but by the 90th percentile of daily maximum temperature, which varies from station to station and will be much lower in Scandinavia than in the Mediterranean (Table 1). The longest summer warm spells occur in Scandinavia, the shortest in the western Mediterranean. Summer cold spells are generally much shorter shortest in Northern Europe, and longest in the Mediterranean.

ERA-Interim simulates slightly too long warm spells of T_{max} (upper panel), in particular for the area averaged version. The RCM, again, adds value. MOS inherits the predictor performance (by construction, as the percentile-based spells are invariant to bias correction). Owing to the predictor averaging, the regression based MOS (MOS-REG) again performs considerably worse. Also the behavior of the PP methods is broadly consistent with that for short-term persistence: analog methods and stochastic white noise methods (MLR-ASW, MLR-AAW, WT-WG, SWG) simulate too short spells. This holds in particular WT-WG, driven only with sea level pressure. Weather generators slightly underestimate mean spell lengths, in particular those who underestimate short-term persistence. Persistence of summer warm spells of T_{max} is consistently overestimated over the Mediterranean, a bias which is much improved by the RCM.

The persistence for summer cold spells of T_{min} (lower panel), consistent with the results for short-term persistence, is generally too high. The RCM deteriorates the performance of ERA-Interim. This performance is, again trivially, unchanged by the MOS methods. The PP methods perform similar as for warm spells, though with a tendency towards higher persistence. All weather generators perform well, consistent with the results for short-term persistence. Cold spells of summer T_{min} are too long for the British Isles and (but to a lesser extent) the Mediterranean. Performance is best for the Alps.

Seasonality The amplitude of the annual cycle of T_{max} (Figure 5) is small towards the Atlantic and the Mediterranean, and large in the continental climates of eastern Scandinavia and Eastern Europe. It peaks in July in continental central and eastern Europe, and slightly later in August towards the Atlantic. ERA-Interim slightly underestimates the amplitude of the seasonal cycle (upper panel) - likely linked to its resolution, as the further averaging increases the bias. The RCM in general adds value, but also increase spread across stations. Being seasonally trained, most MOS methods trivially capture the annual cycle well. Note, however, that also the quantile mapping methods without an explicity annual cyle perform well (GPQM, EQM, EQM-WT) for most stations. The authors do not understand the strong drop in performance of the MOS-REG method when driven with the RCM instead of ERA-Interim. Most PP methods perform reasonably well, even those without seasonal training, because the physical link between the predictors (including temperature) and the predictand is close. Only the WT-WG method sticks out: it is not seasonally trained and uses only sea level pressure as predictor. Thus, seasonality in circulation patterns is captured, but not the changes in temperature within these patterns. The weather generators perform well by construction.

The phase of the seasonal cycle (lower panel) is captured by most methods. ERA-Interim peaks a day too late, the RCM increases the spread across stations. MOS methods perform well, even those with an explicit model of the seasonal cycle (GPQM, EQM, EQM-WT) are within ±2 days (apart from the MOS-REG method, when driven with the RCM). The analog methods perform reasonably well, although the version without seasonal training (ANALOG) has a comparably broad spread across seasons. For regression models, no seasonal training is required if the predictors are standardised (e.g., MLR-AAN, MLR-AAI compared to MLR-RAN). Biases in the ESD methods are caused by the monthly resolution of the data. Again, weather generators perform well by construction.

Interannual Variability and Long-Term Trends Interannual variability of summer daily maximum temperature, measured by the variance of summer mean values, is lowest in the Mediterranean and Scotland, and consistently higher in Central and Eastern Europe and Scandinavia (Figure 6). ERA-Interim slightly underestimates interannual variability, again likely linked to the area averaging. The performance varies widely across stations. The RCM adds moderate value (high in the Mediterranean), but also spread. Simple additive MOS (RaiRat-M6) leaves interannual variability unchanged. Variances of the daily distribution are underestimated by ERA-Interim (see Gutiérrez and coauthors (2017)). The resulting correction by quantile mapping inflates interannual variability, in particular for the Mediterranean, where it is overestimated by around 50%. MOS-REG underestimates interannual variability, in particular when driven with ERA-Interim, because it uses predictors averaged over several grid-boxes.

All analog methods underestimate interannual variability, consistent with the results for short-term persistence. The ANALOG-ANOM method searches for continental-scale analogs within a one-month window around the calendar day of interest - this likely restricts the number of analogs and in turn also the represented variability. Interestingly, most regression methods dramatically underestimate interannual variability. The worst performing methods are those without a seasonal cycle and non-standardised predictors (MLR-RAN), those without temperature predictors (ESD-EOFSLP, ESD-SLP, WT-WG) and those with white noise randomisation (MLR-ASW, MLR-AAW, WT-WG, SWG). Note also that both the ESD methods and the SWG method are defined on continental-scale predictors, which may not be suitable to capture local variations. Inflated regression by construction slightly increases the variance at interannual scales. WGs do not model long-term variations and thus underestimate interannual variability.

In addition to considering the variance at the interannual scale, we also investigate the correlation between the downscaled time series and observations at the interannual scale. Prior to calculating correlations, the time series are linearly detrended. This analysis provides additional insight into the predictors required to explain longer-term variations. These correlations can only be calculated when simulated and observed time series are in synchrony. The RCM develops its own internal variability and thus reduces synchronicity. Therefore we have not shown results for the RCM and RCM-driven MOS. Equivalently, the unconditional weather generators are not in synchrony with observations and hence not shown. Correlations for ERA-Interim and essentially all deterministic MOS methods are high. It is not clear to the authors why CDFt and EQMWIC658 are so little synchronised - they deterministically transform the ERA-Interim predictors and should thus only marginally affect the temporal sequence.

Also PP methods perform well in general. Exceptions are the ANALOG-ANOM method, the ESD methods, the WT-WG and the SWG method. Recall that ANALOG-ANOM takes analogs from a 30 day window around the calender day of interest - the identified analogs might therefore have a rather strong mismatch at the local scale and thus destroy synchronicity. Also, analogs of this method are defined over the whole European domain, which might result in additional discrepancies at the local scales. The ESD methods, which use either 2m-temperature or sea level pressure as predictor, perform worse compared to other regression models; again, also the ESD method uses predictors defined over the whole of Europe. The WT-WG and SWG methods perform rather bad, likely because they are based on white noise randomisation. The WT-WG additionally only uses sea level pressure as predictand, the SWG predictors are defined at the continental scale.

To characterise decadal scale variations, we considered correlations between simulated and observed time series at the 7-year scale. The seasonal aggregated time series are filtered with a 7-year Hamming filter. Correlations are calculated on the filtered time series without any further detrending. The choice of 7 years is a compromise between the desired information about long time scales, and the limited length of the time series. The effective number of data points is thus low for each series (of the order of 5 per series), but still a coherent picture emerges when investigating larger regions.

Figure **??** presents the results for summer (top panel) and winter (bottom panel) daily maximum temperature. The results are overall similar to those for interannual variability. Correlations are in general slightly lower during summer, in particular for ESD-SLP and WT-WG (driven by sea level pressure only) for which correlations are consistently negative. Correlations are lower on the Iberian Peninsula, for winter for the whole Mediterranean.

Finally, we investigate the representation of long-term temperature trends by the different methods. Figure 8 displays the results for winter daily maximum temperatures in selected regions. Of course, no results for weather generators are shown, as these do not include any predictors or change factors to represent long-term changes. Note that in this experiment it is not relevant whether the trends are statistically significant, because long-term variations are imprinted by the ERA-Interim predictors - the right predictor choice should therefore capture large-scale forced trends. It is, however, relevant whether the simulated trends are statistically distinguishable from the observed trends. Thus, we calculated 95% confidence intervals of the trend estimates, marked as grey shading in the panels. As trends differ very much across Europe, we calculated average trends across the PRUDENCE regions. The variations of trends within a region is indicated by whiskers; these denote 1.96 times the variance of all trend estimates across the region.

Observed winter trends are highest in Scandinavia and lowest in the Mediterranean, which is consistent with polar amplification. ERA-Interim performs mostly fine, but overestimates trends in Central Europe, the Alps and the Meditrerranean (but note that the underlying ECA-D data are not homogenised, so a definite answer as to which trends are more realistic is impossible). The RCM underestimates trends in particular in Scandinavia, but also in the Alps and the Mediterranean. These trends are inherited by additive bias correction (RaiRat-M6), but notably modified by many quantile mapping methods due to inflation of daily variances. Note that also the ISI-MIP method, which is designed to perserve mean trends, modifies trends in some regions. These trend variations are substantial, but within the range of uncertainty of the observed trend estimates. The performance of PP methods again depends mainly on the predictor choice. Methods using only sea level pressure or temperature (but not both; ESD-EOFSLP, ESD-SLP, ESD-T2, WT-WG) tend to perform badly, although filtering of stations by PCA appears to strongly increase the link with the temperature predictor on decadal scales (ESD-EOFT2). The ANALOG-ANOM, again, uses rather narrowly defined analogs (continental scale, within one month), the SWG method combines a white-noise stochastic approach with continental-scale predictors. The best performing methods (ANALOG-MP, ANALOG-SP, MO-GP, MLR, MLR-WT) all include circulation predictors and 2m temperature. Note, however, that 2m temperature is likely not well simulated by GCMs (see the discussion in Section 4).

Summer trends of daily maximum temperatures (see supplementary information) are highest in Eastern Europe and the Alps. ERA-Interim in general captures these trends, but underestimates them in the Alps and overestimates them in the Mediterranean. The RCM underestimates summer trends everywhere, in particular in the Alps where the simulated trend is not consistent with the observations. The performance of the statistical post-processing methods is similar to that for winter.

3.2 Precipitation

short-term variability As a measure of persistence in precipitation, we consider wet-wet and dry-wet transition probablities (Figure 9. Short-term persistence in precipitation amounts has not been investigated. Winter Wet-wet transition probabilities (top left panel) are low in southern Europe and high along the Atlantic coasts as well as in high mountains. Winter dry-wet transition probabilities (top right panel) are generally lower than wet-wet probabilities, with low values in southern Europe.

Because it represents area average precipitation, ERA-Interim overestimates wet-wet probabilities, in particular when further averaged. Here the RCM adds substantial value. MOS methods perform consistently well. Interestingly, the simple rescaling by the method RaiRat-M6 appears to perform en par with explicit wet day corrections by quantile mapping (note that the BC method only treats zero precipitation as dry). MOS-AN defines analogs based on simulated large-scale precipitation fields - these may not discriminate well between local dry and wet days. MOS-GLM and VGLMGAMMA are both stochastic methods with white noise randomisation and consequently simulate too weak wet persistence. The 4-grid-box-averaging of the MOS-GLM input appears to considerably improve the performance though. Yet difficulties in regression-based MOS techniques are evident from the low performance of MOS-GLM when driven with RCM data: the RCM strongly perturbs the local day-to-day correspondence between observations and simulation, which is required for a successful calibration.

The analog methods perform well for wet-wet transitions, most deterministic regression models fail. In fact, simple linear regression models (MLR-RAN/RSN/ASW/ASI) are by construction not capable of simulating daily precipitation variability - still the corresponding results are included for illustration and comparison. Only the deterministic generalised linear model (GLM) performs reasonably well. Most stochastic methods with white noise randomisation (GLM-WT, WT-WG, SWG) slightly underestimate wet-day-persistence, in particular WT-WG, which uses only sea level pressure, but no humidity predictors. The stochastic GLM with predictors of the circulation as well as temperature and specific humidity at cloud base is the best performing PP method. Interestingly, the structurally similar GLM-P (at least for the occurrence process) method with similar predictors performs substantially worse. One reason might be that the former defines predictors at the synoptic scale, the latter at the grid-box scale. For wet-day occurrence, vertical velocities are important which can be determined from horizontal convergence or divergence. Grid box pressure or velocities, however, do not carry such information. Still, further analyses comparing different predictor choices are required to fully understand the performance of specific predictors.

Dry-wet transition probabilities are well represented by ERA-Interim. The RCM has a slightly positive bias. Surprisingly, however, MOS appears to reduce dry-wet transitions (by wet day adjustments). Thereby it induces a negative bias for ERA-Interim, but removes the positive RCM bias. Only for the UK, the positive RCM bias is even increased by many methods. Stochastic MOS (MOS-GLM, VGLMGAMMA) simulate too many dry-wet transitions, but the averaging of simulated precipitation across grid-boxes seems to substantially improve the problem (MOS-GLM-E vs. VGLMGAMMA-E). The performance of the different PP methods depends strongly on both their structure and the chosen predictors. The authors do not fully understand the differences in performance of different implementations. The two best performing methods are ANALOG-ANOM and GLM. Both methods include circulation based predictors (which should indirectly give information about lifting) and, at least indirectly, measures of relative humidity (dew point temperature depression; specific humidity in combination with temperature). Other methods, however, include similar predictors, but perform worse. Recall, however, that we only know the candidate predictors used for calibration, not the finally selected predictors at the given stations. The SS-WG and GOMEZ weather generators slightly overestimate dry-wet transitions, even though this aspect is explicitly calibrated. Recall that the MARFI weather generator uses a wet-day threshold of 0.5 mm, resulting in a strong overestimation of dry-wet transitions when evaluated against a 1 mm threshold.

Spells The behaviour of mean spell lengths - as well as the corresponding method performance - is closely tied to that of transition probabilities (Figure 10). Mean winter wet-spell lengths (top left) are high along the along the Atlantic west coasts and mountain ranges, and short in Eastern Europe and the Mediterranean. Summer dry spells (top right) are short in Central and Northern Europe, and long in the Mediterranean.

ERA-Interim underestimates winter wet spells because of spatial averaging (upper panel).

At first sight, the RCM adds no value. Yet the RCM reduces the ERA-Interim bias of too many wet-days Gutiérrez and coauthors (2017) as well as the bias in too high a wet-wet transition probability (see above). As a result, the RCM implicitly adds value in the subsequent bias correction, in particular over the Iberian Peninsula. Quantile mapping without seasonal training (GQM, GPQM, EQM) overestimates winter wet spell lengths. Interestingly, conditioning on weather types (EQM-WT) essentially has the same effect as an explicit seasonal training (EQMs), indicating that biases are circulation dependent and translate into seasonally-dependent biases, because the frequency of weather types changes throughout the year. The MOS-AN, MOS-GLM and VGLMGAMMA perform very similar as with regard to short-term persistence. In particular the averaging of predictors across 4 grid boxes in the stochastic methods (MOS-GLM-E vs. VGLMGAMMA-E) seems to be crucial to increase skill. The performance of the PP methods scatters widely, as already for short term persistence. Only the ANALOG-ANOM and GLM perform well. The SS-WG and GOMEZ Weather generators slightly underestimate wet spell lengths. Again, the MARFI weather generator sticks out because of the different wet day threshold.

The performance for summer dry spells is overall similar to that for winter wet spells. ERA-Interim spells are again too short, but here the RCM adds substantial value, likely due to a reduction of the area-average-related drizzle effect. MOS appears to increase the length of dry spells as a consequence of the wet day correction. For ERA-Interim this leads to unbiased results, whereas the RCM performance is deteriorated towards too long dry spells. This problem occurs in particular for quantile mapping methods, which are not seasonally trained (GQM, GPQM, EQMs, EQM-WT). Analog methods perform slightly better for dry- than for wet spells, the GLM performs worse than for wet spells, but still reasonably well. Weather generators perform slightly better for dry- than for wet spells. Owing to the different wet-day threshold, the MARFI weather generator is slightly more biased and has a much higher spread across stations. In general, the length of dry spells is overestimated in the Mediterranean and France.

Seasonality Seasonality of precipitation is measured by the relative amplitude (defined as the difference between precipitation in the maximum and minimum of the seasonal cycle, relative to the annual mean) and phase (defined as the position of the maximum of the seasonal cycle). Although the calculation is identical to that of the seasonal cycle of temperature, some details will be relevant in particular for precipitation. In fact, the seasonal cycle of precipitation has two peaks in many regions, sometimes even shoulders or peaks that may be artefacts of sampling variability. Following Favre et al. (2016), we therefore filter the seasonal cycle by four harmonics - this model is flexible enough to capture smooth - likely physical - variations, but at the same time filters out residual noise (see Figure 2). The amplitude of the seasonal cycle is simply defined as the difference between maximum and minimum. For the phase definition, further steps have been carried out. They are a compromise between being simple and transparent, but at the same time capturing the complex seasonal behaviour. First, secondary peaks with an amplitude (defined as the difference between the closest local minimum and the peak itself) of less than 10% of the total amplitude have been removed, as well as neighboring peaks with a minimum in between that is less than 10% of the total amplitude lower than the mean height of the two peaks. The two peaks are then replaced by a single peak by averaging their height as well as phase. The first step removes all minor peaks, the second step removes dips in an overall broad maximum, which are both likely an artefact of sampling variability. Visual inspection of observed seasonality for all 86 stations corroborates that this definition conforms with expert judgment. We then record the phase of the remaining highest and second highest peak for observations and all simulations. The observed phase is then defined as that of the highest peak. The simulated phase is defined as the phase of that of the two highest peaks, which is closest to the observed. The latter definition avoids that, if highest and second highest peak have similar height and are swapped in the simulation, an artifically large phase bias is calculated. Apart from this phase definition we considered other measures for characterising the timing of the seasonal cycle, but rejected all other possibilites. We considered, e.g,. correlations between simulated and observed seasonal cycle, but this measure is difficult to interpret in terms of an actual mismatch in timing. Additionally, we also considered to calculate phases of secondary peaks, but concluded that a plain and transparent presentation of performance across Europe would be difficult.

Seasonality of precipitation (Figure 11) has a strong north-south gradient, ranging from less than 50% of annual mean precipitation in central-west Europe to more than 200% in southern Spain and southern Greece. The annual cycle peaks in winter along the Atlantic and the Mediterranean, and in summer in Central and eastern Europe and eastern Scandinavia. Reanalysis and RCM underestimate the amplitude of the annual cycle, although the RCM adds considerable value. MOS generally performs well, although methods without seasonal training (GQM, GPQM, EQM, EQM-WT) overestimate the relative amplitude by about 20%. Note, however, that conditioning the correction on weather types (EQM-WT) substantially reduces this bias. PP performance again depends on the method-type, the treatment of seasonality, and the choice of predictors. The analog methods perform reasonably well, linear regression models all underrepresent the relative amplitude (MLR-RAN/RSN/ASW/ASI). The good performance of the GLM method indicates that a sensible model structure and predictor choice (circulation and humidity) may allow to capture the seasonal cycle without an explicit model. The phase of the seasonal cycle is well captured by most methods. The bad performance of WT-WG indicates that sea level pressure alone does not determine the seasonal cycle.

Interannual Variability and long-term trends Interannual variability of precipitation varies unsystematically in space (Figure 12). Values, however, tend to be higher at higher elevations. As for temperature, reanalysis data underrepresent interannual variability, especially at low resolution. But in contrast to temperature, the RCM succeeds in reducing the overall bias, in particular over the Mediterranean. Deterministic MOS methods suffer strongly from variance inflation, which in cases doubles the internnual variance. Regression based MOS by contrast tends to underestimate interannual variability, consistent with the driving model. The performance of PP methods, again, varies considerably. Note, however, that all well performing methods include not only circulation-based predictors, but also measures of humidity (ANALOG-ANOM, ANALOG, ANALOG-SP, GLM-det, GLM, GLM-WT). Weather generators, as expected, underestimate interannual variability - even more so for the MARFI weather generator because of the different wet-day threshold.

Interannual correlations are, as expected, lower for precipitation than for temperature: only about 50% of the local variability ($\sim 0.7^2$) seems to be explained by the area average, the rest is due to local variability. Deterministic MOS methods do not modify this correlation (again, we cannot explain the performance of EQM-WIC658). For the stochastic MOS methods, the value of averaging simulated precipitation across neighboring grid boxes is evident (compare MOS-GLM-E and VGLMGAMMA-E). All PP methods explain substantially less of the interannual variability than the grid-box ERA-Interim. The worst performing methods are ANALOG-ANOM (analogs searched within 30 day window only, continental scale predictors and analogs), MLR-ASW (Gaussian white noise radomisation), WT-WG (stochastic, only sea level pressure as predictors) and SWG (stochastic, continental scale predictors). Note the substantial difference between the - structurally similar - GLM and SWG models. GLM defines predictors on a national scale, SWG on a continental scale.

Seven year correlations betweeen simulations and observations are similar to interannual correlations; they are much higher though in winter than in summer (see supplementary information).

Finally, we investigate the performance in representing relative trends in seasonal mean precipitation. Figure 13 presents the results for summer and selected regions. All observed trends are essentially zero and insignificant, with moderately positive values in Central Europe. We nevertheless show the results to demonstrate the behaviour of the different methods. ERA-Interim captures the observed trends in some regions, but simulates a zero trend for Central Europe, and a negative trend for the Alps. The RCM simulates positive trends for the British Isles, Central Europe, Scandinavia and the Alps, although all these are within the range of sampling uncertainty. The MOS methods tend to inflate the wrong RCM trends, as well as the wrong negative ERA-Interim trends in the Alps. Many PP methods capture observed trends quite well, although the performance changes substantially - and not for obvious reasons - from region to region. Idenifying necessary predictors appears to be much less straight forward than in case of temperature trends.

4 Discussion and Conclusions

We have systematically evaluated how different types of downscaling and bias correction approaches represent temporal aspects. These aspects comprise systematic seasonal variations and residual temporal dependence such as short-term persistence, spell length distributions and interannual to long-term variability variability. Additionally, we considered long-term trends, which are a superposition of long-term internal climate variability and forced trends. Our results complement, corroborate and extend earlier findings, in particular by Frost et al. (2011), Hu et al. (2013), Benestad and Haugen (2007) and Huth et al. (2015).

Overall, the behaviour of the different approaches turned out to be as expected from their structure and implementation. For the interpretation of the results, it has to be acknowledged whether a particular aspect of a model is explicitly calibrated - a good performance is then more or less trivial - or emerges from the model, e.g., by well chosen meteorological predictors.

A summary of the results (apart from correlations and long-term trends) can be found in Figure 14. The raw ERA-Interim data are typically biased compared to observed station data, stronger so for the spatially aggregated 2° version. Note, however, that these discrepancies are not neccesarily bias in the sense of model errors, but simply reflect the scale-gap between area averages and point values (Volosciuk et al., 2015). The chosen RCM adds value to reanalysis data for most considered aspects, for all seasons and for both temperature and precipitation. Note, however, that we included just one RCM in our validation study. One should be careful in generalising these results because RCMs may differ considerably in their ability to reproduce temporal characteristics (Kotlarski et al., 2014; Huth et al., 2015).

The MOS methods considered in this intercomparison do not explicitly change the residual temporal dependence (and it is questionable whether they should explicitly do so, as such changes would destroy the temporal consistency with the driving model). However, quantile mapping approaches modifying the marginal distribution (including wet day probabilities) do indirectly improve temporal variability. For temperature, some implementations slightly improve short-term persistence, but in particular for precipitation, the representation of transition probabilities as well as wet and dry spells is substantially improved. Interestingly, dry-wet transitions and dry-spell lengths are much better for the bias-corrected RCM than for bias-corrected reanalyses, even though the added value of the RCM for these indices was marginal only. Interannual and long-term variability is typically inflated by MOS. Moderately for temperature, but substantially for precipitation. These findings corroborate earlier results of adverse inflation effects by quantile mapping (Maraun, 2013). long-term trends are inherited from the driving model, but may be substantially deteriorated by further variance inflation. The annual cycle is improved by almost all MOS methods - but recall that most methods are seasonally trained. Conditioning on weather types (EQM-WT) seems to a successful - and physically more defensible - variant to better represent the annual cycle. In any case, our results clearly show that - for many but not all temporal aspects - dynamical downscaling prior to the bias correction substantially improves the results compared to a direct bias correction from the global model[5]. The reason of course is that the bias correction does not improve the representation of meso-scale processes. Thus, depending on the context, dynamical downscaling may be advisable or even essential.

The performance of the participating PP methods varies strongly from aspect to aspect and method to method. Analogue methods show difficulties representing temperature variability, but perform quite well for precipitation variability. Two reasons may contribute to the low performance for temperature: first, predictors describing circulation and humidity have much stronger spatial-temporal variability than temperature fields and therefore dominate the definition of the analogs. Second, predictors and analogs are often defined on large scales. Locally, differences between actual weather and analogs may be substantial. Thus, even if analogs may describe a smooth temperature evolution at large scales, the resulting local sequence might be too noisy.

Deterministic linear regression models perform fairly well for temperature, but overestimate short-term persistence and spell lengths. White noise randomisation deteriorates the representation of these aspects. Linear regression models, in any variant, are far too simplistic for precipitation downscaling. They strongly overestimate wet-wet transitions and the length of wet spells, while stochastic methods underestimate these aspects. Biases for dry-wet transitions and dry-spell lengths tends to be opposite to those for wet-wet transitions and wet-spell lenghts, but they are substantial for almost all PP methods. Only a stochastic generalised linear model with suitable predictors has shown to perform well (GLM). A structually similar model (SWG) - with similar predictor variables, but defined on the continental scale - performs notably bad. The representation of the annual cycle depends strongly on the individual method; whether or not a method is seasonally trained plays a minor role - the choice of reasonable predictors seems to be a key factor. For temperature, temperature related predictors are required; for precipitation, circulation and humidity based predictors. There is evidence that biases in interannual variability of temperature mainly depend on the

[5]Note in this context, that the ERA-Interim is an "ideal" GCM in the sense that it is forced to closely follow the observed large-scale weather.

method type (again, analog methods and white noise randomisation underestimate internal variability), on the predictor variables (all well performing methods combine circulation and temperature predictors) and the domain size (all methods using continental-size predictor domains perform badly). For precipitation, the inclusion of predictors that represent both circulation and humidity appears to be crucial. long-term trends in temperature are captured by models with surface temperature predictors (see the critical discussion below), for precipitation no conclusions can be drawn based on the available ensemble, and the rather low signal-to-noise ratio. Overall, white-noise randomisation with continental-scale predictors turned out to perform weakly. Apparently, the variance explained by predictors at such large scales is rather low, such that the residual white noise is too strong to retain the overall temporal dependence.

Unconditional weather generators tend to perform well for the aspects they have been calibrated for: they only slightly underestimate short-term temperature persistence and wet-wet transitions, but slightly overestimate dry-wet transitions. Nevertheless also many non-calibrated aspects are faily well represented. Temperature spell lengths are slightly underestimated, in particular for winter cold spells and summer warm spells. Wet spell lengths are well represented, dry spell lengths underestimated. Only interannual variability is substantially underrepresented. These effects are well known issues (Wilks and Wilby, 1999) and are relevant also for decadal variability. Seasonality is, by construction, well simulated.

Overall, the performance is similar in different seasons - but recall that in particular most MOS methods and all weather generators are calibrated to do so. These explicit seasonal models, however, may be questioned for being used in a future climate: seasonally varying biases indicate that seasonal biases may also change differently on long time scales.

Our findings highlight a series of open research questions, and the need for a range of improvements. MOS methods perform overall very well. Some key issues, however, remain to be addressed: the inflation (or potentially deflation) of interannual and long-term variability and trends is of course directly tied to the simplicity of quantile mapping compared to MOS methods in weather forecasting and the PP methods presented here: whereas the latter express physical relationships between large and local scales at least rudimentarily as regression models and thereby can distinguish between forced and local internal variability, quantile mapping adjusts only long-term distributions of daily values without any physical basis. This calibration is especially problematic when a scale gap between predictand and predictor is to be bridged (Maraun, 2013). The reason for the calibration, of course, is that regression models cannot easily be calibrated in a free running climate model, which is not in synchrony with observations Maraun et al. (2010). More research is needed to understand the link between biases in short-term variability and long-term variability. Some methods have been developed to separate variability on different scales, and to adjust them independently, other methods have been developed to preserve climate model trends to various degrees (Li et al., 2010; Haerter et al., 2011; Hempel et al., 2013; Pierce et al., 2015). The physical assumptions underlying these different methods need to be better understood. In any case, our results show that any bias correction relies on climate models that simulate realistic trends. In case of downscaling to a finer resolution, it might be useful to separate the bias correction from the downscaling, i.e., apply a correction against gridded observational data, and then implement a stochastic downscaling model against point data (Volosciuk et al., 2017). Regression based MOS methods have been presented as further alternatives (MOS-REG/GLM, VGLMGAMMA), but these cannot be calibrated to standard climate model simulations. The results show that even typical RCM hindcast simulations (where the RCM is driven with a reanalysis, MOS-REG-R and MOS-GLM-R) are not sufficiently synchronous to ensure a successful calibration. A way out might be to condition bias correction on weather types, such as demonstrated by EQM-WT.

Various research strands are possible and necessary to better understand and to improve PP methods. For analog methods, in particular in case of temperature, a way forward could be based on defining the analogs not on a single day, but rather on a sequence of days (e.g. Beersma and Buishand, 2003). Such approaches, however, require long time series. Note, however, that analog methods cannot represent substantial climatic changes, where no analogs might be available to sample from Gutiérrez et al. (2013). An obvious improvement of regression models is a better representation of residual variability - for temperature the in linear models for temperature, and generalised linear models for precipitation. Here, conditional weather generators are promising that extend the white noise randomisation (both for temperature and precipitation) by a Markov component. For instance, one may include not only meteorological predictors, but also simulated predictand values from previous days as predictors (Chandler and Wheater, 2002; Yang et al., 2005).

The crucial questions regarding the PP approach are, however, not an improvement in model structure, but a better understanding of predictor choice. Unfortunately, the available model ensemble did not allow for a stringent identification of suitable predictors. Nevertheless, the results highligh a couple of issues. Note that these are questions of physics more than of statistics. First, what is a suitable domain size? The GLM-P and GLM methods include a structurally similar rainfall occurrence process and a - at first sight - similar set of predictors. But the GLM method performs far better than GLM-P in simulating all occurrence-related aspects. A major difference between the two implementations is that GLM uses synoptic scale predictors, whereas GLM-P relies on grid-box predictors. Precipitation occurrence is controlled by relative humidity and vertical velocity. The latter is typically represented by predictors of the horizontal circulation. The underlying reasoning is that horizontal divergence and convergence determines vertical descent and ascent. Convergence and divergence, in turn, may be implicit in large-scale pressure fields, but they are not represented by grid-box pressure values. Thus, the choice of predictor variables depends on the domain size. Many methods with limited performance, in particular for temperature, where based on continental-scale predictors. Thus, there is evidence that such predictor domains are simply too large to successfully represent local variability. Here one has to trade-off between downscaling across large areas and precision at local scales. In fact, we see the main strength of PP methods not in competing with RCMs across whole continents, but rather in providing tailored region-specific projections.

Second, which predictors are required for representing long-term trends? We demonstrated that model performance for the same set of predictors differed substantially for short-term persistence and long-term changes. The reason of course is that downscaling methods are calibrated to day-to-day-variability, but are intended to work on long-term variability (Huth et al., 2015). For temperature, a combination of temperature and circulation predictors appeared to faily well explain long-term trends. Precipitation, however, is a more complex nonlinear process, and no method convincingly captured trends in all considered regions. A further complicating issue is the low signal to noise ratio: all trends, and all misrepresentations, are still within the sampling uncertainty.

Weather generators do have an explicit model of the short-term temporal dependence, but those variants participating in this intercomparison did not include any meteorological predictors. As a result, these methods underestimated long-term variability - it was not explicity modelled. Also here improvements are possible, e.g., by conditioning the weather generator on monthly aggregates (being generated by the separate monthly WG or taken from the driving data - e.g. GCM, RCM or reanalysis) to improve the representation of interannual variability (Dubrovský et al., 2004).

This study was based on a perfect predictor setting to isolate downscaling skill. Therefore, we did not investigate the performance with imperfect predictors or boundary conditions from free running GCMs. Downscaling methods - apart from unconditional weather generators - to a large extent inherit the errors in representing temporal variability of the driving models (Hall, 2014). The downscaling performance may, therefore, drop considerably, when driven by imperfect forcing from a GCM. For MOS, the issue is rather subtle: marginal biases in present climate are by construction removed, hence it is difficult to identify fundamental GCM errors such as the misrepresentation of the large-scale circulation and its temporal structure. Thus, also non-calibrated aspects, in particular the temporal aspects, should thus be evaluated.

For PP one typically assumes that large-scale predictors from the free atmosphere fulfill the PP assumption. This assumption should be tested for GCMs. Again, evaluating temporal aspects might be more informative than evaluating marginal aspects - often, predictors are based on anomalies, such that mean biases are implicitly removed. But even more, many PP predictors are not defined at large scales, and not chosen from the free atmosphere. For instance, those methods that best represented temperature trends all relied on 2m-temperature. In the reanalysis, which has been used as predictors, temperature observations have been assimilated into the model, such that grid-box variability and long-term are likely correctly represented in data rich regions. Local surface feedbacks that modulate temperature variability are thus implictly accounted for. But a free running GCM will likely not correctly represent these feedbacks, such that GCM simulated 2m temperature will likely not fulfill the PP assumption. Similar arguments apply for grid box values of, e.g., 10m winds.

Even though we investigated the performance to represent observed trends, we can only draw limited conclusions about representing future trends. MOS relies on credibly simulated grid box trends - the ERA-Interim trends are approximately correct by construction, the RCM show substantial deficiencies. But also for PP methods, our findings are far from being conclusive. For temperature, as discussed before, the PP assumption for relevant predictors may not be fullfilled. For precipitation, simply no conclusions are possible because of the low signal-to-noise ration. In any case, a method performing badly with perfect predictors will not perform better with imperfect predictors. Passing this evaluation is therefore a necessary, but not a sufficient requirement for a method to be applicable under climate change conditions.

This discussion shows that further studies are required to establish the skill of downscaling under simulated future conditions. The VALUE community is planning additional experiments Maraun et al. (2015): GCM predictor experiments to asses the performance under imperfect predictors, and pseudo reality experiments to establish statistical downscaling skill in simulated future climates. Additionally, we have identified a range of open questions that can be addressed within our perfect predictor experiment, in particular related to the predictor choice of PP methods. The metadata and complete results for individual methods are available from the VALUE portal `www.value-cost.eu/validationportal`. They can be downloaded and further analysed. Additionally, we encourage dedicated sensitivity studies based on the ensemble at hand.

Appendix

Similarly to the portrait diagram in Sillmann et al. (2013), Figure 14 summarises the performance of the different methods for different indices in one (color-coded) value. To make these comparable across methods and indices, a reference scale has to be defined. This scale cannot simply be measured in terms of the best and worst performing methods for an index, as such a scale would only measure relative performance, not absolute performance. For instance, one would not be able to distinguish an index that is well represented from one that is poorly represented by all methods. Sillmann et al. (2013) define the variability of an index in space as reference scale. But this scale cannot be applied to a single series, and it cannot distinguish between indices that are well modelled by al methods across space (e.g., the seasonal cycle) and indices that are badly modelled (e.g., interannual variability). Thus, we attempt to define natural scales for different types of indices:

- For biases in mean temperature, we define twice the standard deviation of daily variability as scale. For Gaussian distributed variables, this range spans roughly 95% of the probability mass.
- For biases of temperature indices, which may be expressed as anomalies (such as the 20 year return value or the amplitude of the seasonal cycle), we chose the actual modulus of the anomaly (i.e., the difference of the return value and mean temperature, or the amplitude itself) as reference scale.
- For relative biases, which assume only positive values (such as for temperature variance, precipitation intensity or mean spell length), a natural scale is the observed value itself.
- For the phase of the seasonal scale we (somewhat arbitrarily) define one month as a reference scale.

References

A. Alexandru, R. de Elia, and R. Laprise. Internal variability in regional climate downscaling at the seasonal scale. *Mon. Wea. Rev.*, 135(9):3221–3238, 2007. doi: 10.1175/MWR3456.1.

J. Bartholy, R. Pongrácz, and A. Kis. Projected changes of extreme precipitation using multi-model approach. *Q.J. Hung. Meteorol. Serv.*, 119:129–142, 2015.

J. Bedia, M. Iturbide, S. Herrera, R. Manzanas, and J. Gutiérrez. downscaler: Climate data manipulation, bias correction and statistical downscaling. http://github.com/SantanderMetGroup/downscaleR/wiki, 2016.

J.J. Beersma and T.A. Buishand. Multi-site simulation of daily precipitation and temperature conditional on the atmospheric circulation. *Clim. Res.*, 25:121–133, 2003.

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

R.E. Benestad and J.E. Haugen. On complex extremes: flood hazards and combined high spring-time precipitation and temperature in Norway. *Climatic Change*, 85(3-4):381–406, 2007.

R.E. Benestad, D. Chen, A. Mezghani, L. Fan, and K. Parding. On using principal components to represent stations in empirical-statistical downscaling. *Tellus A*, 67, 2015b.

M.S. Bukovsky. Temperature trends in the NARCCAP regional climate models. *J. Climate*, 24:3985–3991, 2012.

G. Bürger, T. Q. Murdock, A. T. Werner, S. R. Sobie, and A. J. Cannon. Downscaling extremes - an intercomparison of multiple statistical methods for present climate. *J. Climate*, 25: 4366–4388, 2012.

P. Calanca. Climate change and drought occurrence in the Alpine region: How severe are becoming the extremes? *Glob. Planet. Change*, 57(1):151–160, 2007.

A.J. Cannon. Multivariate Bias Correction of Climate Model Output: Matching Marginal Distributions and Intervariable Dependence Structure. *J. Climate*, 29(19):7045–7064, 2016.

P. Ceppi, S.C. Scherrer, A.M. Fischer, and C. Appenzeller. Revisiting Swiss temperature trends 19592008. *Int. J. Climatol.*, 32:203–213, 2012.

R. E. Chandler and H. S. Wheater. Analysis of rainfall variability using generalized linear models: A case study from the west of Ireland. *Wat. Resour. Res.*, 38(10):1192, 2002.

S.P. Charles, B.C. Bates, and J.P. Hughes. A spatiotemporal model for downscaling precipitation occurrence and amounts. *J. Geophys. Res.*, 104(D24):31,657–31,669, 1999.

J. H Christensen and O. B. Christensen. A summary of the PRUDENCE model projections of changes in European climate by the end of this century. *Clim. Change*, 81:7–30, 2007.

D.P. Dee, S.M. Uppala, A.J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M.A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A.C.M. Beeljars, L. van den Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A.J. Geer, L. Haimberger, S.B. Healy, H. Hersbach, E.V. Hólm, L. Isaksen, P. Kållberg, M. Köhler, M. Matricardi, A.P. McNally, B.M. Monge-Sanz, J.-J. Morcrette, B.-K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J.-N. Thépaut, and F. Vitart. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quart. J. Royal Meteorol. Soc.*, 137:553–597, 2011.

M. Dubrovský, J. Buchtele, and Z. Žalud. High-frequency and low-frequency variability in stochastic daily weather generator and its effect on agricultural and hydrologic modelling. *Clim. Change*, 63(1):145–179, 2004.

A. Favre, N. Philippon, B. Pohl, E.-A. Kalognomou, C. Lennard, B. Hewitson, G. Nikulin, A. Dosio, H.-J. Panitz, and R. Cerezo-Mota. Spatial distribution of precipitation annual cycles over South Africa in 10 CORDEX regional climate model present-day simulations. *Clim. Dynam.*, 46:1799–1818, 2016.

E.M. Fischer, S.I. Seneviratne, P.L. Vidale, D. Lüthi, and C. Schär. Soil moisture-atmosphere interactions during the 2003 European summer heat wave. *J. Climate*, 20:5081–5099, 2007.

H. J. Fowler, S. Blenkinsop, and C. Tebaldi. Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. *Int. J. Climatol.*, 27:1547–1578, 2007.

C. Frei, J. H. Christensen, M. Deque, D. Jacob, R. G. Jones, and P. L. Vidale. Daily precipitation statistics in regional climate models: Evaluation and intercomparison for the european alps. *Journal of Geophysical Research-Atmospheres*, 108(D3), 2003.

P. Froidevaux, J. Schwanbeck, R. Weingartner, C. Chevalier, and O. Martius. Flood triggering in Switzerland: the role of daily to monthly preceding precipitation. *Hydrol. Earth Syst. Sci.*, 19(9):3903–3924, 2015.

A.J. Frost, S.P. Charles, B. Timbal, F.H.S. Chiew, R. Mehrotra, K.C. Nguyen, R.E. Chandler, J.L. McGregor, G. Fu, D.G.C. Kirono, et al. A comparison of multi-site daily rainfall downscaling techniques under Australian conditions. *J. Hydrol.*, 408(1):1–18, 2011.

F. Giorgi, X. Bi, and J. Pal. Mean, interannual variability and trends in a regional climate change experiment over Europe. I. Present-day climate (19611990). *Clim. Dynam.*, 22: 733–756, 2004.

C.M. Goodess, C. Anagnostopoulou, A. Bárdossy, C. Frei, C. Harpham, M.R. Haylock, Y. Hundecha, P. Maheras, J. Ribalaygua, J. Schmidli, T. Schmith, K. Tolika, R. Tomozeiu, and R.L. Wilby. An intercomparison of statistical downscaling methods for Europe and European regions assessing their performance with respect to extreme weather events and the implications for climate change applications. Technical report, Climatic Research Unit, 2010.

J.M. Gutiérrez and coauthors. An intercomparison of a large ensemble of statistical downscaling methods for europe: Overall results from the value perfect predictor cross-validation experiment. *Int. J. Climatol., subm.*, 2017.

J.M. Gutiérrez, D. San-Martín, S. Brands, R. Manzanas, and S. Herrera. Reassessing statistical downscaling techniques for their robust application under climate change conditions. *J. Climate*, 26(1):171–188, 2013.

E. Gutmann, T. Pruitt, M.P. Clark, L. Brekke, J.R. Arnold, D.A. Raff, and R.M. Rasmussen. An intercomparison of statistical downscaling methods used for water resource assessments in the united states. *Wat. Resour. Res.*, 50(9):7167–7186, 2014.

J.O. Haerter, S. Hagemann, C. Moseley, and C. Piani. Climate model bias correction and the role of timescales. *Hydrol. Earth Syst. Sci.*, 15(3):1065–1079, 2011.

A. Hall, X. Qu, and J.D. Neelin. Improving predictions of summer climate change in the united states. *Geophys. Res. Lett.*, 35:L01702, 2008.

Alex Hall. Projecting regional change. *Science*, 346(6216):1461–1462, 2014.

M. R. Haylock, G. C. Gawley, C. Harpham, R. L. Wilby, and C. M. Goodess. Downscaling heavy precipitation over the UnitedKingdom: A comparison of dynamical and statistical methods and their future scenarios. *Int. J. Climatol.*, 26(10):1397–1415, 2006.

S. Hempel, K. Frieler, L. Warszawski, J. Schewe, and F. Piontek. A trend-preserving bias correction - the ISI-MIP approach. *Earth Syst. Dynam.*, 4:219–236, 2013.

S. Herrera, M. Turcu, and J.M. Gutiérrez. A mos-regression technique for temporally-coherent bias correction of regional climate model simulations. *Clim. Dynam.*, submitted, 2017.

E. Hertig and coauthors. Validation of extremes from the perfect-predictor experiment of the cost action value. *Int. J. Climatol.*, 2016.

E. Hertig and J. Jacobeit. Assessments of Mediterranean precipitation changes for the 21st century using statistical downscaling techniques. *Int. J. Climatol.*, 28:1025–1045, 2008.

Y. Hu, S. Maskey, and S. Uhlenbrook. Downscaling daily precipitation over the yellow river source region in china: a comparison of three statistical downscaling methods. *Theor. Appl. Climatol.*, 112(3-4):447–460, 2013.

R. Huth. Statistical downscaling of daily temperature in central europe. *J. Climate*, 15: 1731–1742, 2002.

R. Huth, J. Kyselý, and M. Dubrovský. Time structure of observed, GCM-simulated, downscaled, and stochastically generated daily temperature series. *J. Climate*, 14:4047–4061, 2001.

R. Huth, J. Miksovsky, P. Stepanek, M. Belda, A. Farda, Z. Chladova, and P. Pisoft. Comparative validation of statistical and dynamical downscaling models on a dense grid in central Europe: temperature. *Theor. Appl. Climatol.*, 120(3-4):533–553, MAY 2015. ISSN 0177-798X. doi: 10.1007/s00704-014-1190-3.

D. Jacob, L. Bärring, O. B. Christensen, J. H. Christensen, M. de Castro, M. Déqué, F. Giorgi, S. Hagemann, M. Hirschi, R. Jones, E. Kjellström, G. Lenderink, B. Rockel, E. Sánchez, C. Schär, S. I. Seneviratne, S. Somot, A. van Ulden, and B. van den Hurk. An intercomparison of regional climate models for Europe: model performance in present-day climate. *Clim. Change*, 81:31–52, 2007.

E.-A. Kalognomou, C. Lennard, M. Shongwe, I. Pinto, A. Favre, M. Kent, B. Hewitson, A. Dosio, G. Nikulin, H.-J. Panitz, and M. Büchner. A Diagnostic Evaluation of Precipitation in CORDEX Models over Southern Africa. *J. Climate*, 26(23):9477–9506, 2013. doi: 10.1175/JCLI-D-12-00703.1.

D. Keller, A.M. Fischer, C. Frei, M.A. Liniger, C. Appenzeller, and R. Knutti. Implementation and validation of a Wilks-type multi-site daily precipitation generator over a typical Alpine river catchment. *Hydrol. Earth Syst. Sci.*, 19:2163–2177, 2015.

D.E. Keller, A.M. Fischer, M.A. Liniger, C. Appenzeller, and R. Knutti. Testing a weather generator for downscaling climate change projections over Switzerland. *Int. J. Climatol.*, 2016.

C. G. Kilsby, P. D. Jones, A. Burton, A. C. Ford, H. J. Fowler, C. Harpham, P. James, A. Smith, and R. L. Wilby. A daily weather generator for use in climate change studies. *Env. Mod. Soft.*, 22:1705–1719, 2007.

A.M.G. Klein Tank, J.B. Wijngaard, G.P. Können, R. Böhm, G. Demarée, A. Gocheva, M. Mileta, S. Pashiardis, L. Hejkrlik, C. Kern-Hansen, R. Heino, P. Bessemoulin, G. Müller-Westermeier, M. Tzanakou, S. Szalai, T. Pálsdóttir, D. Fitzgerald, S. Rubin, M. Capaldo, M. Maugeri, A. Leitass, A. Bukantis, R. Aberfeld, A.F.V. van Engelen, E. Forland, M. Mietus, F. Coelho, C. Mares, V. Razuvaev, E. Nieplova, T. Cegnar, J.A. López, B. Dahlström, A. Moberg, W. Kirchhofer, A. Ceylan, O. Pachaliuk, L.V. Alexander, and P. Petrovic. Daily dataset of 20th-century surface air temperature and precipitation series for the european climate assessment. *Int. J. Climatol.*, 22(12):1441–1453, 2002.

S. Kotlarski, K. Keuler, O.B. Christensen, A. Colette, M. Déqué, A. Gobiet, K. Goergen, D. Jacob, D. Lüthi, E. van Meijgaard, G. Nikulin, C. Schär, C. Teichmann, R. Vautard, K. Warrach-Sagi, and V. Wulfmeyer. Regional climate modelling on European scales: A joint standard evaluation of the EURO-CORDEX RCM ensemble. *Geosci. Model. Dev. Discuss.*, 7:217–293, 2014.

O. Lhotka and J. Kyselý. Spatial and temporal characteristics of heat waves over Central Europe in an ensemble of regional climate model simulations. *Clim. Dynam.*, 45:2351–2366, 2015.

H. Li, J. Sheffield, and E.F. Wood. Bias correction of monthly precipitation and temperature fields from Intergovernmental Panel on Climate Change AR4 models using equidistant quantile matching. *J. Geophys. Res.*, 115:D10101, 2010.

P. Lorenz and D. Jacob. Validation of temperature trends in the ENSEMBLES regional climate model runs driven by ERA40. *Clim. Res.*, 44:167–177, 2010.

D. Maraun. Bias correction, quantile mapping and downscaling: Revisiting the inflation issue. *J. Climate*, 26:2137–2143, 2013.

D. Maraun. Bias correcting climate change simulations - a critical review. *Curr. Clim. Change Rep.*, 2(4):211–220, 2016. doi: 10.1007/s40641-016-0050-x.

D. Maraun and M. Widmann. *Statistical Downscaling and Bias Correction for Climate Research.* Cambridge University Press, 2018.

D. Maraun, F. Wetterhall, A. M. Ireson, R. E. Chandler, E. J. Kendon, M. Widmann, S. Brienen, H. W. Rust, T. Sauter, M. Themeßl, V. K. C. Venema, K. P. Chun, C. M. Goodess, R. G. Jones, C. Onof, M. Vrac, and I. Thiele-Eich. Precipitation downscaling under climate change. Recent developments to bridge the gap between dynamical models and the end user. *Rev. Geophys.*, 48:RG3003, 2010.

D. Maraun, M. Widmann, J. M. Gutierrez, S. Kotlarski, R. E. Chandler, E. Hertig, J. Wibig, R. Huth, and R. A. I. Wilcke. VALUE: A framework to validate downscaling approaches for climate change studies. *Earth's Future*, 3:1–14, 2015.

A. Martynov, R. Laprise, L. Sushama, K. Winger, L. Separovic, and B. Dugas. Reanalysis-driven climate simulation over CORDEX North America domain using the Canadian Regional Climate Model, version 5: model performance evaluation. *Clim. Dynam.*, 41: 29733005, 2013.

A. Moberg and P.D. Jones. Regional climate model simulations of daily maximum and minimum near-surface temperatures across Europe compared with observed station data 1961-1990. *Clim. Dynam.*, 23:695–715, 2004.

R. Monjo, G. Chust, and V. Caselles. Probabilistic correction of RCM precipitation in the Basque Country (Northern Spain). *Theor. Appl. Climatol.*, 117(1-2):317–329, 2014.

C: Obled, G. Bontron, and R. Garçon. Quantitative precipitation forecasts: a statistical adaptation of model outputs through an analogues sorting approach. *Atmos. Res.*, 63(3): 303–324, 2002.

R.A. Pielke and R.L. Wilby. Regional climate downscaling: What's the point? *EOS*, 93(5): 52–53, 2012.

D.W. Pierce, D.R. Cayan, E.P. Maurer, J.T. Abatzoglou, and K.C. Hegewisch. Improved Bias Correction Techniques for Hydrological Simulations of Climate Change. *J. Hydrometeorol.*, 16(6):2421–2442, 2015.

R. Pongrácz, J. Bartholy, and A. Kis. Estimation of future precipitation conditions for Hungary with special focus on dry periods. *Időjárás*, 118(4):305–321, 2014.

J. Räisänen and O. Räty. Projections of daily mean temperature variability in the future: cross-validation tests with ENSEMBLES regional climate simulations. *Clim. Dynam.*, 41: 1553–1568, 2013.

J. Rajczak, S. Kotlarski, and C. Schär. Does quantile mapping of simulated precipitation correct for biases in transition probabilities and spell lengths? *J. Climate*, 29:1605–1615, 2016.

O. Räty, J. Räisänen, and J.S. Ylhäisi. Evaluation of delta change and bias correction methods for future daily precipitation: intermodel cross-validation using ENSEMBLES simulations. *Clim. Dynam.*, 42(9-10):2287–2303, 2014.

D. Raynaud, B. Hingray, I. Zin, S. Anquetin, S. Debionne, and R. Vautard. Atmospheric analogues for physically consistent scenarios of surface weather in Europe and Maghreb. *Int. J. Climatol.*, 37(4):2160–2176, 2017.

C.W. Richardson. Stochastic simulation of daily precipitation, temperature, and solar radiation. *Wat. Resour. Res.*, 17(1), 1981.

C. Rosenzweig, A. Iglesias, X. B. Yang, P. R. Epstein, and E. Chivian. Climate change and extreme weather events: Implications for food production, plant diseases, and pests. *Global Change and Human Health*, 2(2):90–104, 2001.

C. Rosenzweig, W.D. Solecki, S.A. Hammer, and S. Mehrotra, editors. *Climate Change and Cities: First Assessment Report of the Urban Climate Change Research Network*. Cambridge University Press, 2011.

M. Rummukainen. State-fo-the-art with regional climate models. *Wiley Int. Rev. Clim. Change*, 1:82–96, 2010. DOI: 10.1002/wcc.8.

D. San-Martín, R: Manzanas, S. Brands, S. Herrera, and J.M. Gutiérrez. Reassessing model uncertainty for regional projections of precipitation with an ensemble of statistical downscaling methods. *J. Climate*, 30(1):203–223, 2017.

C. Schär, D. Lüthi, U. Beyerle, and E. Heise. The soilprecipitation feedback: A process study with a regional climate model. *J. Climate*, 12(3):722–741, 1999. doi: 10.1175/1520-0442(1999)012¡0722:TSPFAP¿2.0.CO;2.

A. Schindler, D. Maraun, and J. Luterbacher. Validation of the present day annual cycle in heavy precipitation over the British Islands simulated by 14 RCMs. *J. Geophys. Res.*, 117: D18107, 2007.

J. Schmidli, C. M. Goodess, C. Frei, M. R. Haylock, Y. Hundecha, J. Ribalaygua, and T. Schmith. Statistical and dynamical downscaling of precipitation: An evaluation and comparison of scenarios for the european alps. *Journal of Geophysical Research-Atmospheres*, 112(D4), 2007.

M. A. Semenov, R. J. Brooks, E. M. Barrow, and C. W. Richardson. Comparison of the wgen and lars-wg stochastic weather generators for diverse climates. *Clim. Res.*, 10(2):95–107, 1998.

J.C. Semenza, C.H. Rubin, K.H. Falter, J.D. Selanikio, W.D. Flanders, H.L. Howe, and J.L. Wilhelm. Heat-related deaths during the july 1995 heat wave in chicago. *N. Engl. J. Med.*, 335(2):84–90, 1996.

S. Seneviratne, D. Lüthi, M. Litschi, and C. Schär. Land-atmosphere coupling and climate change in Europe. *Nature*, 443:205–209, 2006.

J. Sillmann, V.V. Kharin, X. Zhang, F.W. Zwiers, and D. Bronaugh. Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate. *J. Geophys. Res.*, 118(4):1716–1733, 2013.

P. Soares and coauthors. Process based evaluation of the VALUE perfect predictor experiment of statistical downscaling methods. *Int. J. Climatol., subm.*, 2017.

P.M.M. Soares, R.M. Cardoso, P.M.A. Miranda, P. Viterbo, and M. Belo-Pereira. Assessment of the ENSEMBLES regional climate models in the representation of precipitation variability and extremes over Portugal. *J. Geophys. Res.*, 117(D7):D071114, 2012.

P. Štěpánek, P. Zahradníček, A. Farda, P. Skalák, M. Trnka, J. Meitner, and K. Rajdl. Projection of drought-inducing climate conditions in the czech republic according to euro-cordex models. *Clim. Res.*, 70(2-3):179–193, 2016.

S. Stoll, H.J. Hendricks Franssen, M. Butts, and W. Kinzelbach. Analysis of the impact of climate change on groundwater related hydrological fluxes: a multi-model approach including different downscaling methods. *Hydrology and Earth System Sciences*, 15(1):21–38, 2011.

M. Turco, P. Quintana-Seguí, M.C. Llasat, S. Herrera, and J.M. Gutiérrez. Testing MOS precipitation downscaling for ENSEMBLES regional climate models over Spain. *J. Geophys. Res.*, 116(D18), 2011.

M. Turco, M.C. Llasat, S. Herrera, and J.M. Gutiérrez. Bias correction and downscaling of future RCM precipitation projections using a MOS-Analog technique. *J. Geophys. Res.*, 122(5):2631–2648, 2017.

P. Vaittinada Ayar, M. Vrac, S. Bastin, J. Carreau, M. Déqué, and C. Gallardo. Intercomparison of statistical and dynamical downscaling models under the EURO-and MED-CORDEX initiative framework: present climate evaluations. *Clim. Dynam.*, 46(3-4):1301–1329, 2016.

E. van Meijgaard, L. H. van Ulft, W. J. van de Berg, F. C. Bosveld, B. J. J. M. van den Hurk, G. Lenderink, and A. P. Siebesma. The KNMI regional atmospheric climate model RACMO version 2.1. Technical Report 302, Royal Dutch Meteorological Institute, KNMI, Postbus 201, 3730 AE, De Bilt, The Netherlands, 2008.

R. Vautard, A. Gobiet, D. Jacob, M. Belda, A. Colette, M. Déqué, J. Fernández, M. García-Díez, K. Goergen, I. Güttler, T. Halenka, T. Karacostas, E. Katragkou, K. Keuler, S. Kotlarski, S. Mayer, E. van Meijgaard, G. Nikulin, M. Patarcic, J. Scinocca, S. Sobolowski, M. Suklitsch, C. Teichmann, K. Warrach-Sagi, V. Wulfmeyer, and P. Yiou. The simulation of European heat waves from an ensemble of regional climate models within the EURO-CORDEX project. *Clim. Dynam.*, 41:2555–2575, 2013.

C. Volosciuk, D. Maraun, V.A. Semenov, and W. Park. Extreme precipitation in an atmosphere general circulation model: impact of horizontal and vertical model resolutions. *J. Climate*, 28(3):1184–1205, 2015.

C. Volosciuk, D. Maraun, M. Vrac, and M. Widmann. A combined statistical bias correction and stochastic downscaling method for precipitation. *Hydrol. Earth Syst. Sci.*, 21(3):1693–1719, 2017.

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

M. Vrac and P. Friederichs. Multivariate-intervariable, spatial, and temporal-bias correction. *J. Climate*, 28(1):218–237, 2015.

M. Vrac, P. Drobinski, A. Merlo, M. Herrmann, C. Lavaysee, L. Li, and S. Somot. Dynamical and statistical downscaling of the French Mediterranean climate: uncertainty assessment. *Nat. Haz. Earth Syst. Sci.*, 12:2769–2784, 2012.

K. Warrach-Sagi, T. Schwitalla, V. Wulfmeyer, and H.-S. Bauer. Evaluation of a climate simulation in Europe based on the WRF–NOAH model system: precipitation in Germany. *Clim. Dynam.*, 41(3):755–774, 2013. doi: 10.1007/s00382-013-1727-7.

M. Widmann and coauthors. Validation of spatial variability in downscaling results from the value perfect predictor experiment. *Int. J. Climatol., subm.*, 2017.

R.L. Wilby, T.M.L. Wigley, D. Conway, P.D. Jones, B.C. Hewitson, J. Main, and D.S. Wilks. Statistical downscaling of general circulation model output: A comparison of methods. *Wat. Resour. Res.*, 34(11):2995–3008, 1998.

R.A.I. Wilcke, T. Mendlik, and A. Gobiet. Multi-variable error correction of regional climate models. *Clim. Change*, 120(4):871–887, 2013.

D. S. Wilks and R. L. Wilby. The weather generation game: a review of stochastic weather models. *Prog. Phys. Geogr.*, 23(3):329–357, 1999.

D.S. Wilks. Use of stochastic weathergenerators for precipitation downscaling. *Wiley Interdisciplinary Reviews: Climate Change*, 1(6):898–907, 2010.

G. Wong, D. Maraun, M. Vrac, M. Widmann, J. Eden, and T. Kent. Stochastic model output statistics for bias correcting and downscaling precipitation including extremes. *J. Climate*, 27:6940–6959, 2014.

C. Yang, R. E. Chandler, and V. S. Isham. Spatial-temporal rainfall simulation using generalized linear models. *Wat. Resour. Res.*, 41:W11415, 2005.

W. Yang, J. Andréasson, L.P. Graham, J. Olsson, J. Rosberg, and F. Wetterhall. Distribution-based scaling to improve usability of regional climate model projections for hydrological climate change impacts studies. *Hydrol. Res.*, 41(3-4):211–229, 2010.

W. Yang, M. Gardelin, J. Olsson, and T. Bosshard. Multi-variable bias correction: application of forest fire risk in present and future climate in Sweden. *Natural hazards and earth system sciences*, 15(9):2037–2057, 2015.

T. Zerenner, V. Venema, P. Friederichs, and C. Simmer. Downscaling near-surface atmospheric fields with multi-objective Genetic Programming. *Env. Mod. Soft.*, 84:85–98, 2016.

Acknowledgements

VALUE has been funded as EU COST Action ES1102. Participation of M. Dubrovsk and R. Huth in VALUE was supported by the Ministry of Education, Youth, and Sports of the Czech Republic under contracts LD12029 and LD12059, respectively.

Code	Tech	ST	AC	SE	Predictors	Domain	Reference
MOS							
RaiRat-M6	S	no	no	yes	temperature	gridbox	Räisänen and Räty (2013)
RaiRat-M7	S	no	no	yes	temperature	gridbox	Räisänen and Räty (2013)
RaiRat-M8	S	no	no	yes	temperature	gridbox	Räisänen and Räty (2013)
SB	S	no	no	yes	temperature	gridbox	
ISI-MIP	S/PM	no	no	yes	temperature	gridbox	Hempel et al. (2013)
DBS	PM	no	no	yes	temperature	gridbox	Yang et al. (2010, 2015)
GPQM	PM	no	no	no	temperature	gridbox	Bedia et al. (2016)
EQM	QM	no	no	no	temperature	gridbox	Bedia et al. (2016)
EQMs	QM	no	no	yes	temperature	gridbox	Bedia et al. (2016)
EQM-WT	QM/WT	no	no	no	temperature	gridbox	Bedia et al. (2016)
QMm	QM	no	no	yes	temperature	gridbox	Li et al. (2010)
QMBC-BJ-PR	QM	no	no	yes	temperature	gridbox	Pongrácz et al. (2014) Bartholy et al. (2015)
CDFt	QM	no	no	yes	temperature	gridbox	Vrac et al. (2012)
QM-DAP	QM	no	no	yes	temperature	gridbox	Štěpánek et al. (2016)
EQM-WIC658	QM	no	no	yes	temperature	gridbox	Wilcke et al. (2013)
RaiRat-M9	QM	no	no	yes	temperature	gridbox	Räisänen and Räty (2013)
DBBC	QM	no	no	yes	temperature	gridbox	
DBD	QM	no	no	yes	temperature	gridbox	
MOS-REG	TF	yes	no	no	temperature	4 gridboxes	Herrera et al. (2017)
FIC02T	PM/A/TF	no	no	yes	temperature	gridbox	
PP							
FIC01T	A/TF	no	no	yes	Z1000+500	nat. > gridb.	
ANALOG-ANOM	A	no	no	yes	SLP/TD/T2/U+V+Z850	continental	Vaittinada Ayar et al. (2016)
ANALOG	A	no	no	no	SLP/T2/T850+700+500/Q850+500/Z500	national	Gutiérrez et al. (2013) San-Martín et al. (2017)
ANALOG-MP	A	no	no	yes	Z1000+500 > U+V600/T850	nat. > gridb.	Obled et al. (2002) Raynaud et al. (2017)
ANALOG-SP	A	no	no	yes	Z1000+500 > T2/T2-TD	nat. > gridb.	Obled et al. (2002) Raynaud et al. (2017)
MO-GP	TF	no	no	no	full standard set	gridbox	Zerenner et al. (2016)
MLR-T	TF	no	no	no	T2/SLP/U+V10m/T+Q+U+V850+700+500	gridbox	
MLR-RAN	TF	no	no	no	Z500/T850	gridbox	Huth (2002); Huth et al. (2015)
MLR-RSN	TF	no	no	yes	Z500/T850	gridbox	Huth (2002); Huth et al. (2015)
MLR-ASW	TF	yes	no	yes	Z500/T850	gridbox	Huth (2002); Huth et al. (2015)
MLR-ASI	TF	no	no	yes	Z500/T850	gridbox	Huth (2002); Huth et al. (2015)
MLR-AAN	TF	no	no	yes	Z500/T850	gridbox	Huth (2002); Huth et al. (2015)
MLR-AAI	TF	no	no	yes	Z500/T850	gridbox	Huth (2002); Huth et al. (2015)
MLR-AAW	TF	yes	no	yes	Z500/T850	gridbox	Huth (2002); Huth et al. (2015)
MLR-PCA-ZTR	TF	no	no	yes	Z850/T850/R850	continental	Hertig and Jacobeit (2008)
ESD-EOFSLP	TF/WT	no	no	yes	SLP	continental	Benestad et al. (2015a)
ESD-EOFT2	TF/WT	no	no	yes	T2	continental	Benestad et al. (2015a)
ESD-SLP	TF/WT	no	no	yes	SLP	continental	Benestad et al. (2015a)
ESD-T2	TF/WT	no	no	yes	T2	continental	Benestad et al. (2015a)
MLR	TF	no	no	no	SLP/T2/T850+700+500/Q850+500/Z500	national	Gutiérrez et al. (2013)
MLR-WT	TF/WT	yes	no	yes	SLP/T2/T850+700+500/Q850+500/Z500	national	Gutiérrez et al. (2013)
WT-WG	WT/WG	yes	no	no	SLP	national	Gutiérrez et al. (2013)
SWG	TF/WG	yes	no	yes	SLP/T2/TD/U+V+Z850	continental	Vaittinada Ayar et al. (2016)
WG							
SS-WG	WG	yes	yes	yes	NA	NA	Keller et al. (2015, 2016)
MARFI-BASIC	WG	yes	yes	yes	NA	NA	
MARFI-TAD	WG	yes	yes	yes	NA	NA	
MARFI-M3	WG	yes	yes	yes	NA	NA	
GOMEZ-BASIC	WG	yes	yes	yes	NA	NA	
GOMEZ-TAD	WG	yes	yes	yes	NA	NA	

Table 2: Participating methods for temperature. Techniques: S: additive correction; PM: parametric quantile mapping; QM: empirical quantile mapping; A: analog method; TF: regression-like transfer function; WT: weather typing; WG: weather generator. Explicitly modelled: ST: stochastic noise, AC: autocorrelation, SE: seasonality. SLP: sea level pressure, T2: 2m-temperature, T: temperature, TD: dew point temperature, Z: geopotential height, Q: specific humidity, R: relative humidity, U,V,Z: velocities. A > indicates a two-step method. For the full VALUE standard set of predictors and further details on the methods see Gutiérrez and coauthors (2017) or `http://www.value-cost.eu/validationportal/app#!downscalingmethod`.

Code	Tech	ST	AC	SE	Predictors	Domain	Reference
MOS							
Ratyetal-M6	S	no	no	yes	precipitation	gridbox	Räty et al. (2014)
Ratyetal-M7	S	no	no	yes	precipitation	gridbox	Räty et al. (2014)
ISI-MIP	S/PM	no	no	yes	precipitation	gridbox	Hempel et al. (2013)
DBS	PM	no	no	yes	precipitation	gridbox	Yang et al. (2005, 2015)
Ratyetal-M9	PM	no	no	yes	precipitation	gridbox	Räty et al. (2014)
BC	PM	no	no	yes	precipitation	gridbox	Monjo et al. (2014)
GQM	PM	no	no	no	precipitation	gridbox	Bedia et al. (2016)
GPQM	PM	no	no	no	precipitation	gridbox	Bedia et al. (2016)
EQM	QM	no	no	no	precipitation	gridbox	Bedia et al. (2016)
EQMs	QM	no	no	yes	precipitation	gridbox	Bedia et al. (2016)
EQM-WT	QM/WT	no	no	no	precipitation	gridbox	Bedia et al. (2016)
QMm	QM	no	no	yes	precipitation	gridbox	Li et al. (2010)
QMBC-BJ-PR	QM	no	no	yes	precipitation	gridbox	Pongrácz et al. (2014) Bartholy et al. (2015)
CDFt	QM	no	no	yes	precipitation	gridbox	Vrac et al. (2012)
QM-DAP	QM	no	no	yes	precipitation	gridbox	Štěpánek et al. (2016)
EQM-WIC658	QM	no	no	yes	precipitation	gridbox	Wilcke et al. (2013)
Ratyetal-M8	QM	no	no	yes	precipitation	gridbox	Räty et al. (2014)
MOS-AN	A	no	no	no	precipitation	gridbox	Turco et al. (2011, 2017)
MOS-GLM	TF	yes	no	no	precipitation	4 gridboxes	Herrera et al. (2017)
VGLMGAMMA	TF/WG	yes	no	yes	precipitation	gridbox	Wong et al. (2014)
FIC02P	PM/A/TF	no	no	yes	precipitation	gridbox	
FIC04P	PM/A/TF	no	no	yes	precipitation	gridbox	
PP							
FIC01P	A/TF	no	no	yes	Z1000+500	nat. > gridb.	
FIC03P	A/TF	no	no	yes	U+V10m/U+V500/R850+700 > R850/Q700	nat. > gridb.	
ANALOG-ANOM	A	no	no	yes	SLP/TD/T2/U+V+Z850	continental	Vaittinada Ayar et al. (2016)
ANALOG	A	no	no	no	SLP/T2/T850+700+500/Q850+500/Z500	national	Gutiérrez et al. (2013) San-Martín et al. (2017)
ANALOG-MP	A	no	no	yes	Z1000+500 > U+V600/T850	nat. > gridb.	Obled et al. (2002) Raynaud et al. (2017)
ANALOG-SP	A	no	no	yes	Z1000+500 > T2/T2-TD	nat. > gridb.	Obled et al. (2002) Raynaud et al. (2017)
MO-GP	TF	no	no	no	full standard set	gridbox	Zerenner et al. (2016)
GLM-P	TF	yes[3]	no	no	Z500/T850	gridbox	
MLR-RAN	TF	no	no	no	Z500/T850	gridbox	
MLR-RSN	TF	no	no	yes	Z500/T850	gridbox	
MLR-ASW	TF	yes	no	yes	Z500/T850	gridbox	
MLR-ASI	TF	no	no	yes	Z500/T850	gridbox	
GLM-det	TF	no	no	no	SLP/T2/T850+700+500/Q850+500/Z500	national	San-Martín et al. (2017)
GLM	TF	yes	no	no	SLP/T2/T850+700+500/Q850+500/Z500	national	San-Martín et al. (2017)
GLM-WT	TF/WT	yes	no	yes	SLP/T2/T850+700+500/Q850+500/Z500 (WT: only SLP)	national	San-Martín et al. (2017)
WT-WG	WT/WG	yes	no	no	SLP	national	San-Martín et al. (2017)
SWG	TF/WG	yes	no	yes	SLP/T2/TD/U+V+Z850	continental	Vaittinada Ayar et al. (2016)
WG							
SS-WG	WG	yes	yes	yes	NA	NA	Keller et al. (2015, 2016)
MARFI-BASIC	WG	yes	yes	yes	NA	NA	
MARFI-TAD	WG	yes	yes	yes	NA	NA	
MARFI-M3	WG	yes	yes	yes	NA	NA	
GOMEZ-BASIC	WG	yes	yes	yes	NA	NA	
GOMEZ-TAD	WG	yes	yes	yes	NA	NA	

Table 3: Participating methods for precipitation. Techniques: S: scaling; PM: parametric quantile mapping; QM: empirical quantile mapping; A: analog method; TF: regression-like transfer function; WT: weather typing; WG: weather generator. Explicitly modelled: ST: stochastic noise, AC: autocorrelation, SE: seasonality. SLP: sea level pressure, T2: 2m-temperature, T: temperature, TD: dew point temperature, Z: geopotential height, Q: specific humidity, R: relative humidity, U,V,Z: velocities. A > indicates a two-step method. Methods included for ilustrative purposes are marked in grey. For the full VALUE standard set of predictors and further details on the methods see Gutiérrez and coauthors (2017) or `http://www.value-cost.eu/validationportal/app#!downscalingmethod`.

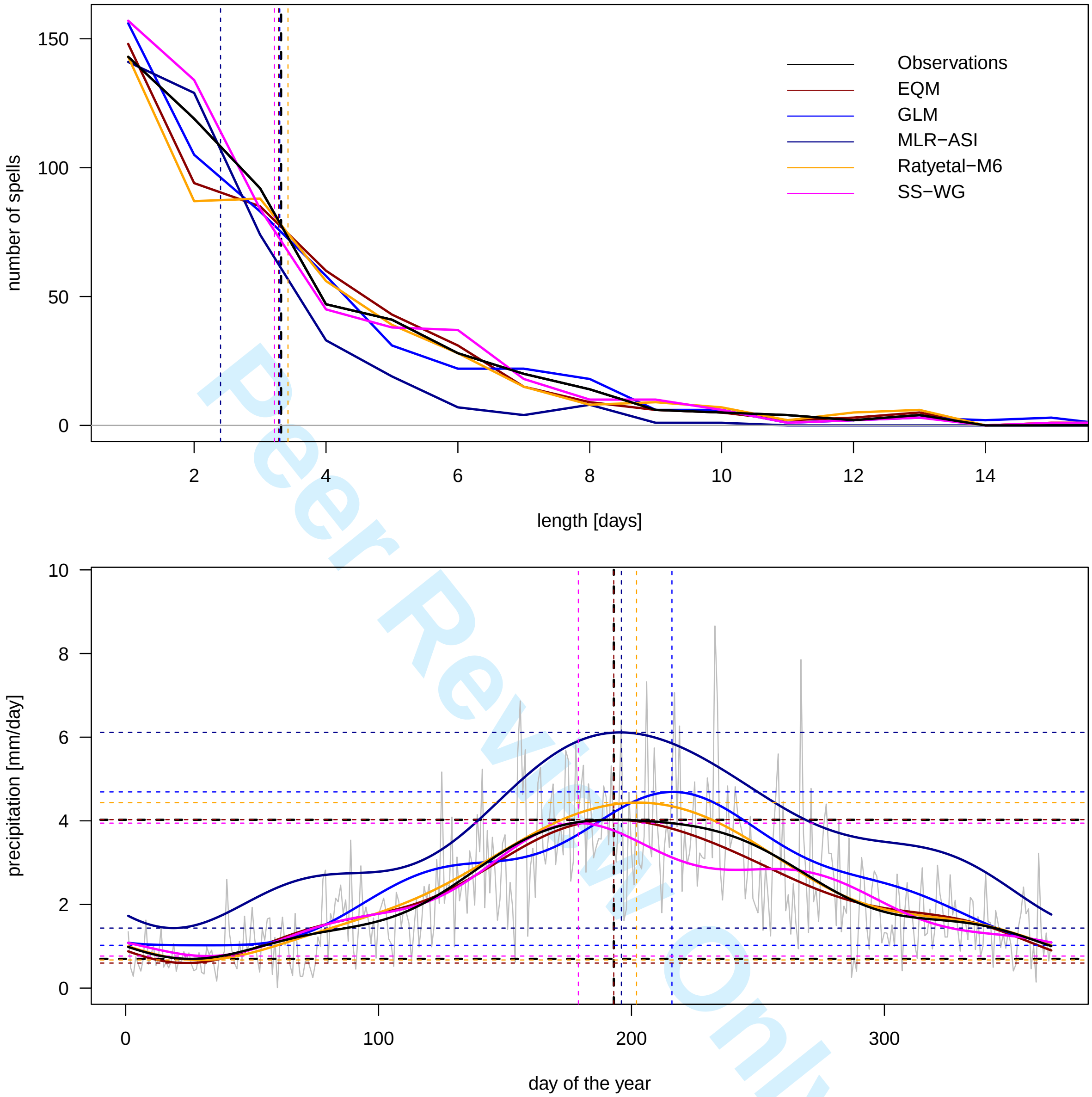

Figure 2: Illustration of selected aspects for daily precipitation, Graz, Austria. Top: dry spell length distribution. Bottom: annual cycle. Black: observations, red: EQM, orange: Ratyetal-M6, blue: MLR-SDSM, dark blue: MLR-ASI, magenta: SS-WG. Top, vertical dashed lines: mean spell length; bottom, vertical dashed lines: phase of annual cycle maximum; bottom, horizontal lines: minimum and maximum of annual cycle.

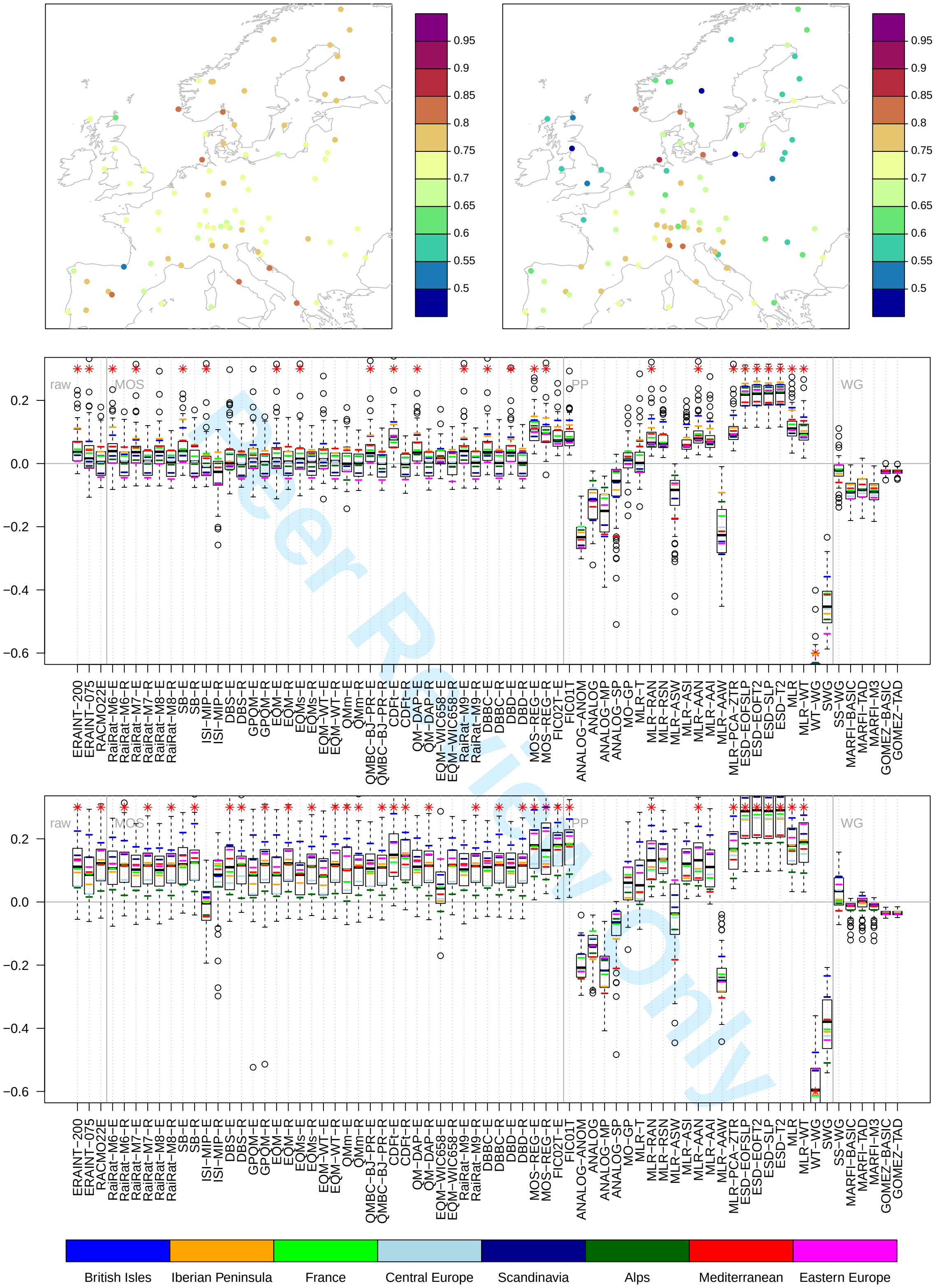

Figure 3: AC1 for summer T_{max} (left/top) and T_{min} (right/bottom). Top row: observed relationships for summer. Bottom rows: bias of the individual methods. For each method, box-whisker-plots summarise the information for all considered stations. Boxes span the 25-75% range, the whiskers the maximum value within 1.5 times the interquartile range, values outside that range are plotted individually. A red asterisk indicates that values lie outside the plotted range. The suffixes in the names of the MOS methods indicate whether a method has been driven with ERA-Interim (-E) or the RCM (-R).

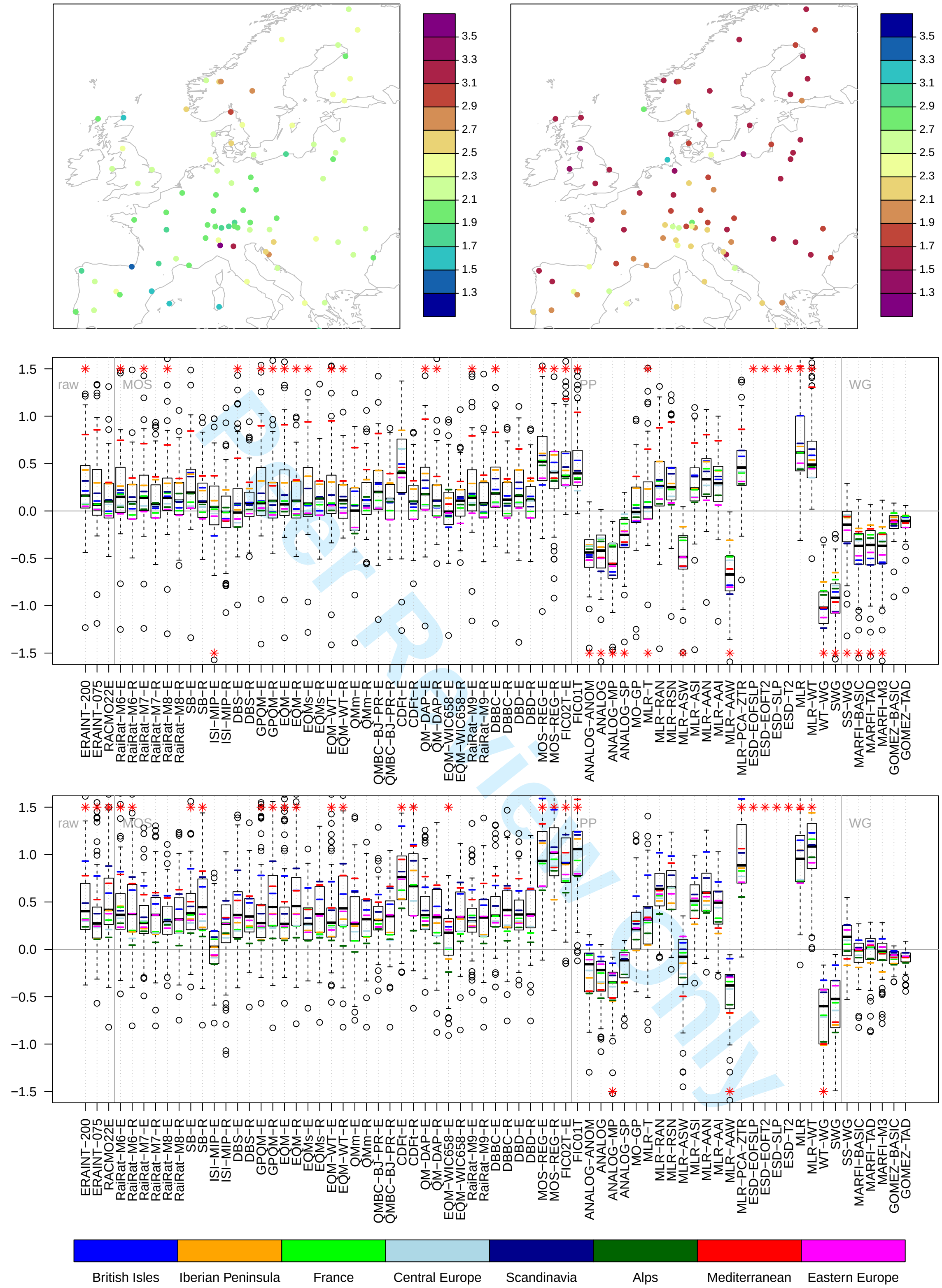

Figure 4: As Fig.3, but for summer WarmSpellMean [days] of T_{max} (top/left) and summer ColdSpellMean [days] of T_{min} (bottom/right)

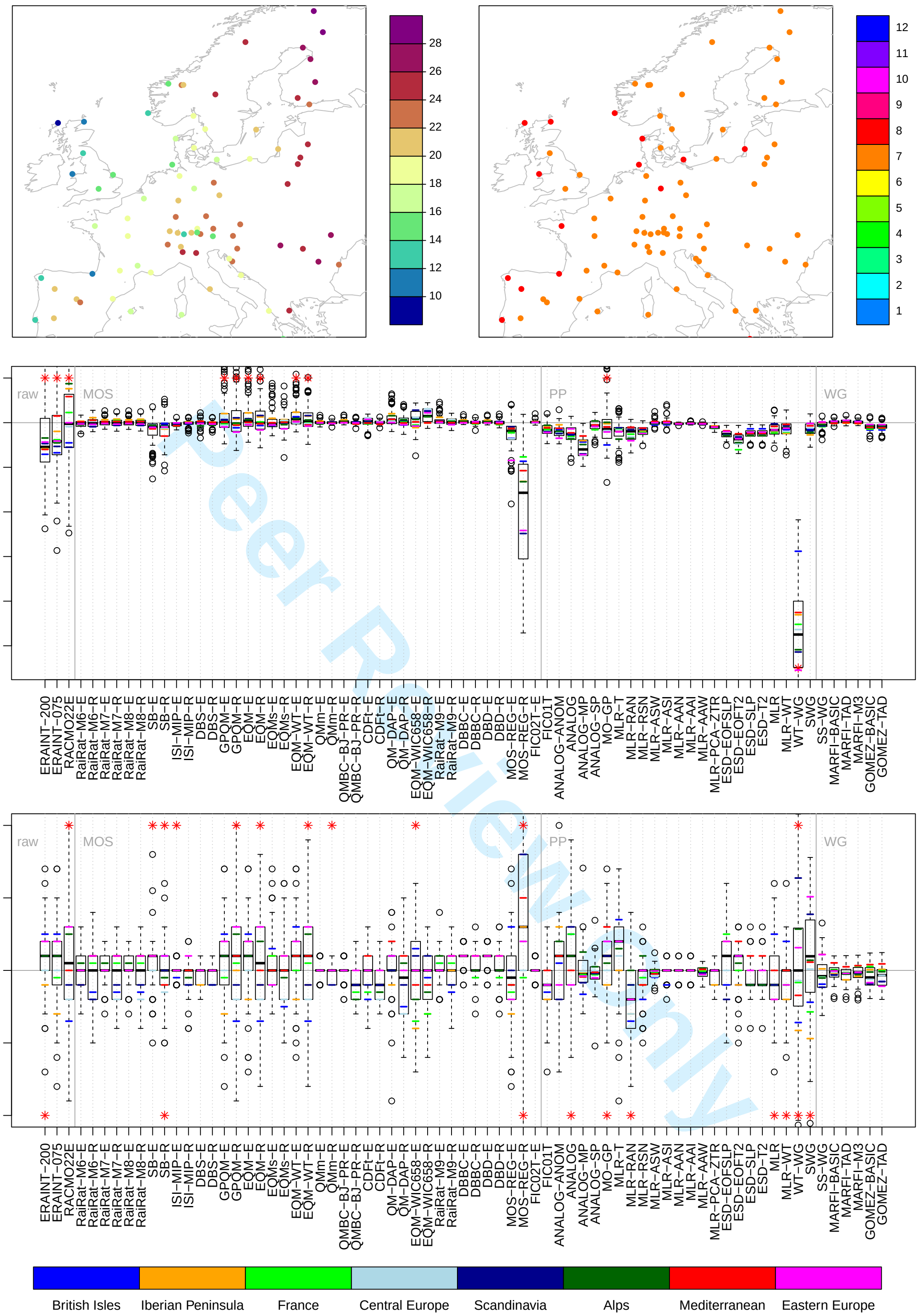

Figure 5: As Fig.3, but for the amplitude [K] (left/top) and phase [days] (right/bottom) of the annual cycle for T_{max}.

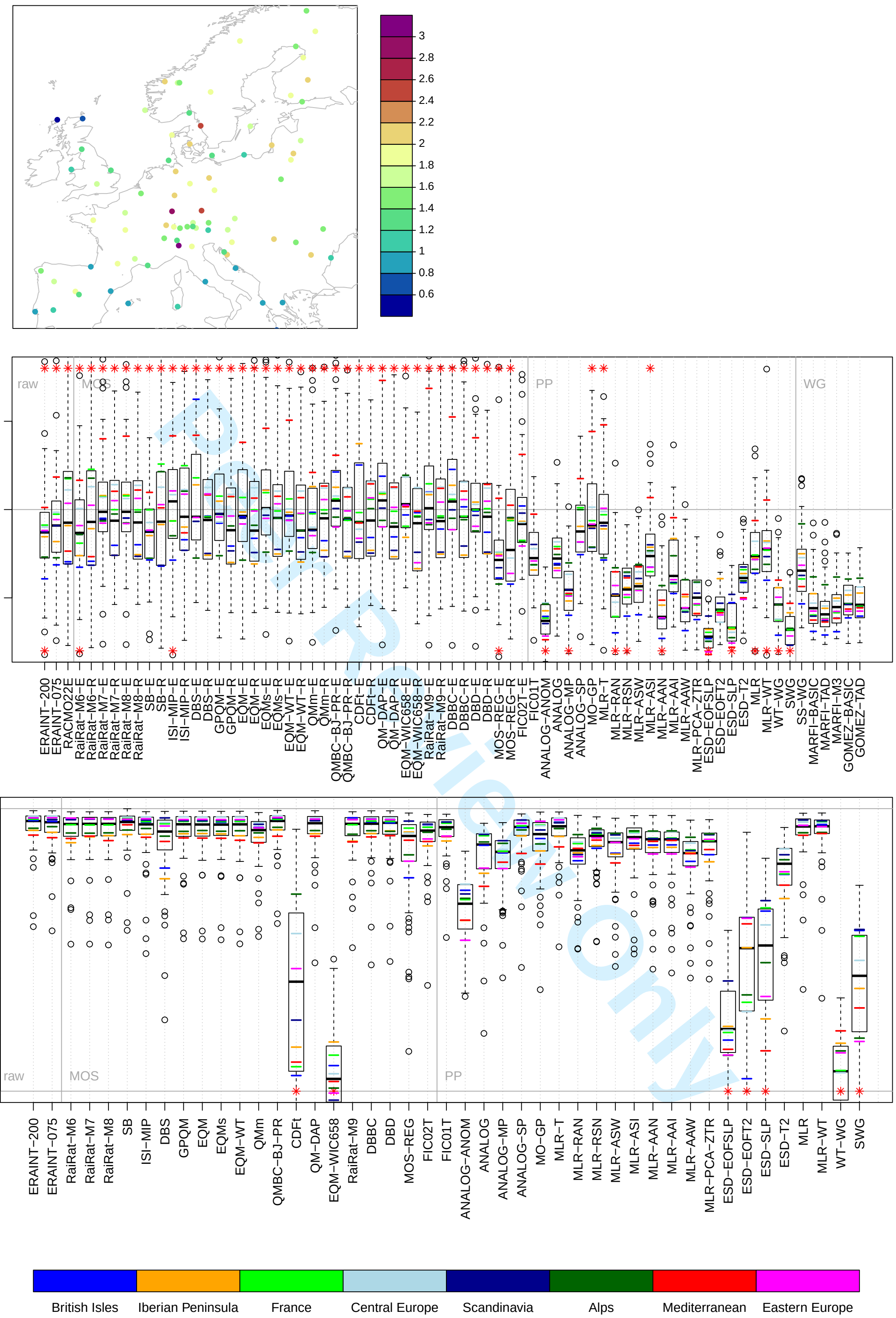

Figure 6: As Fig.3, but for summer VarY [K^2] (map/top) and Cor.1Y (no map/bottom) of T_{max}.

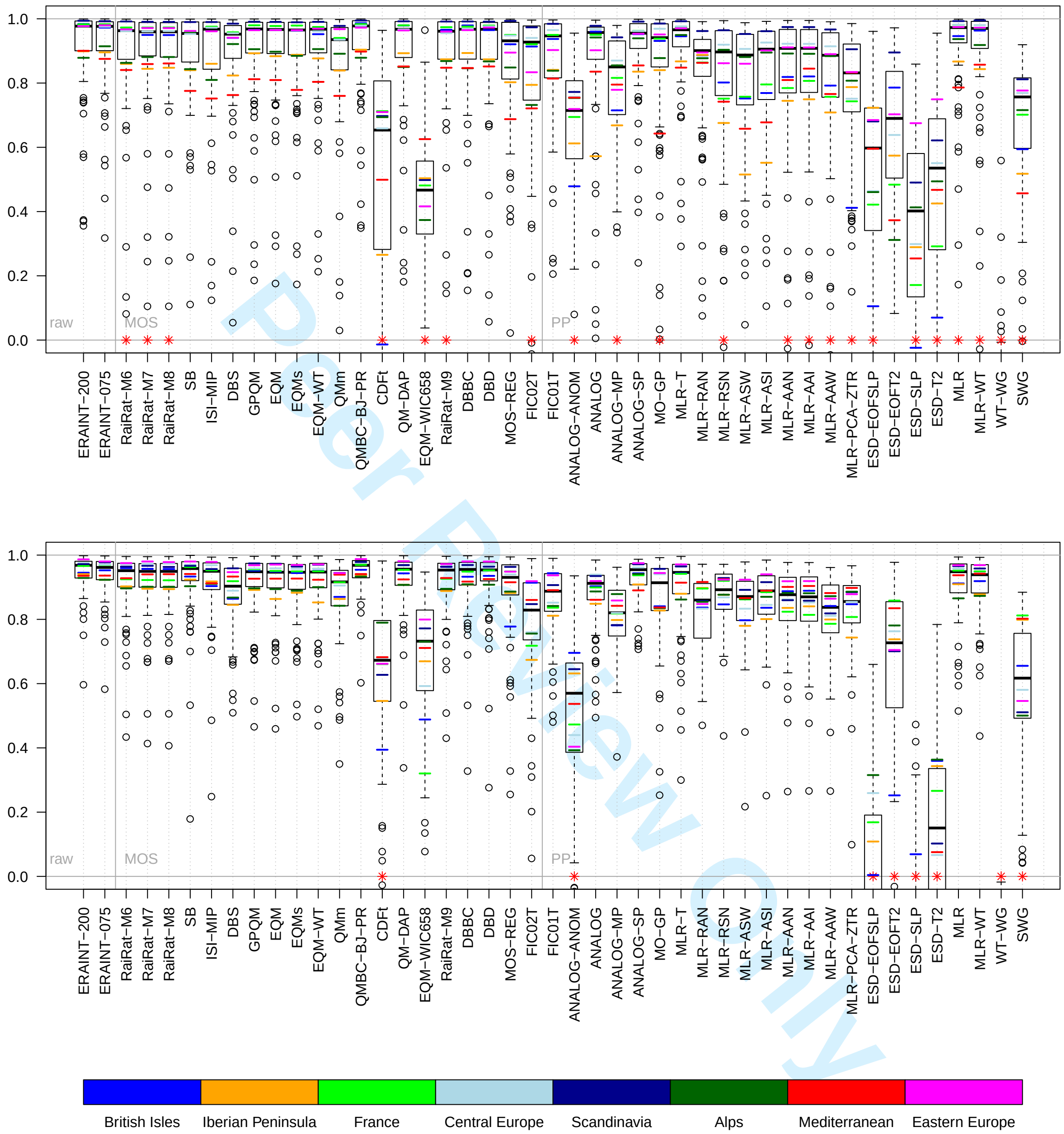

Figure 7: As Fig.3, but for Cor.7Y and T_{max}. Top: DJF; bottom: JJA.

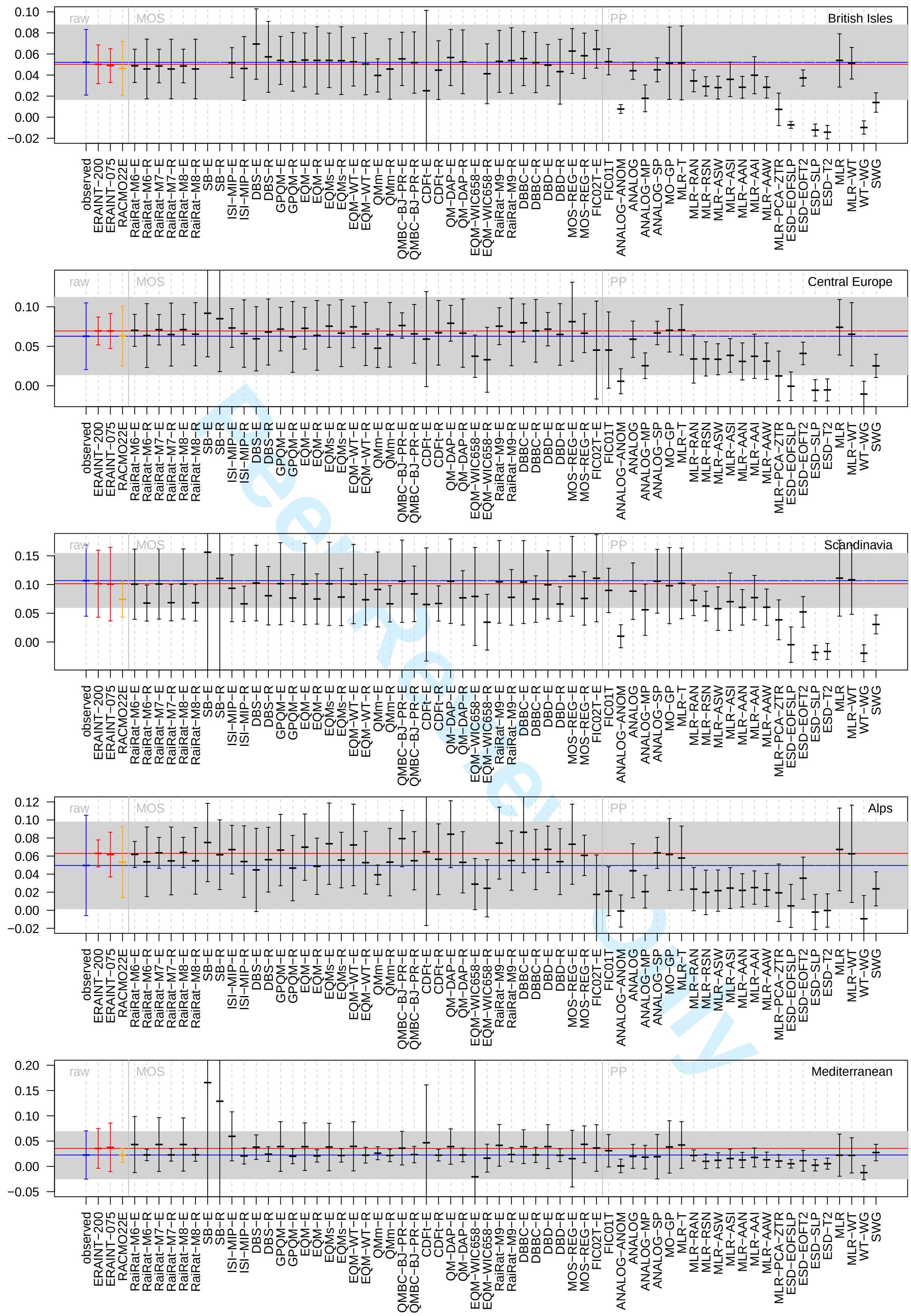

Figure 8: As Fig.3, but for the trend [K] in DJF mean T_{max}.

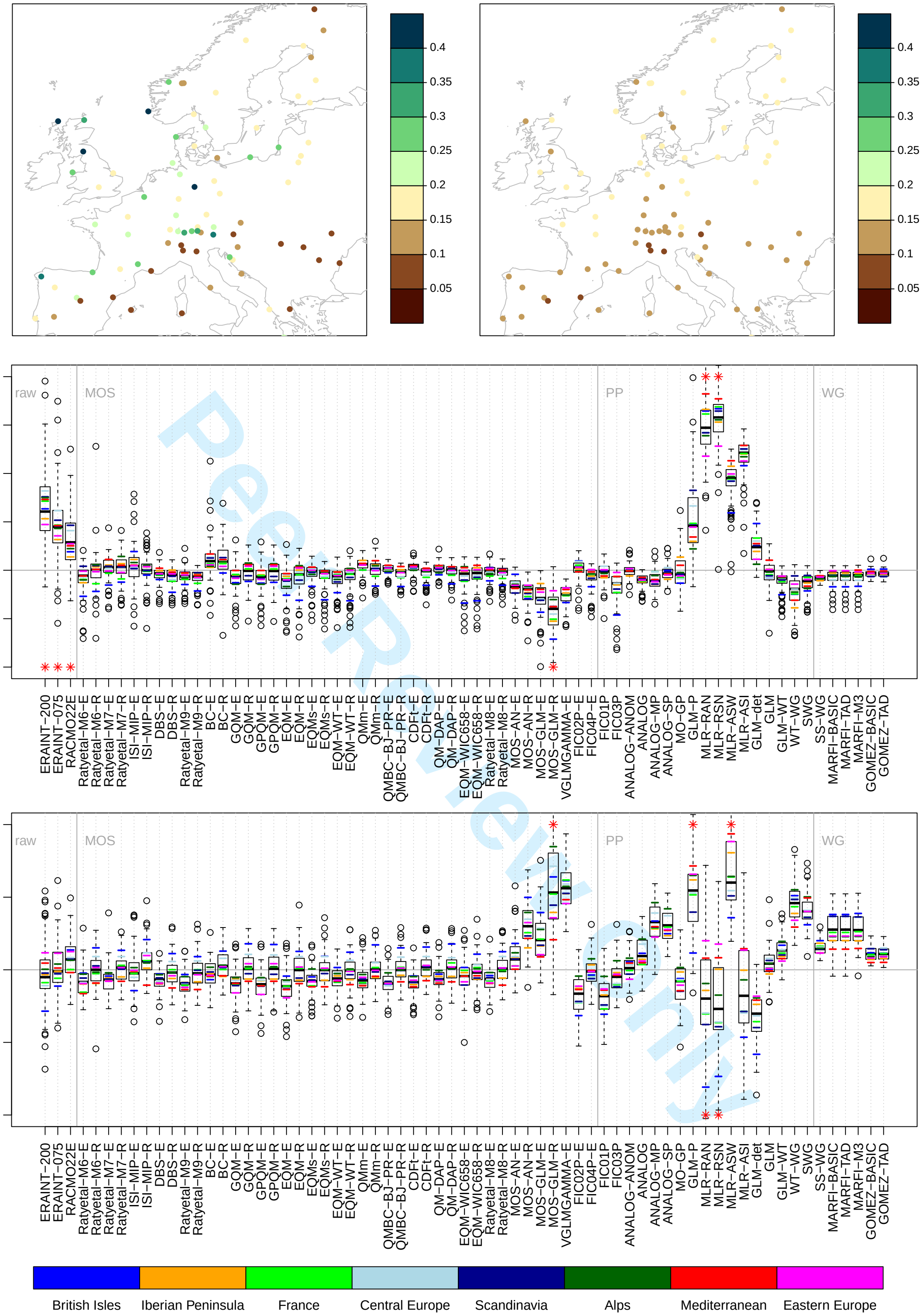

Figure 9: As Fig.3, but for winter WWProb (left/top) and DWProb (right/bottom).

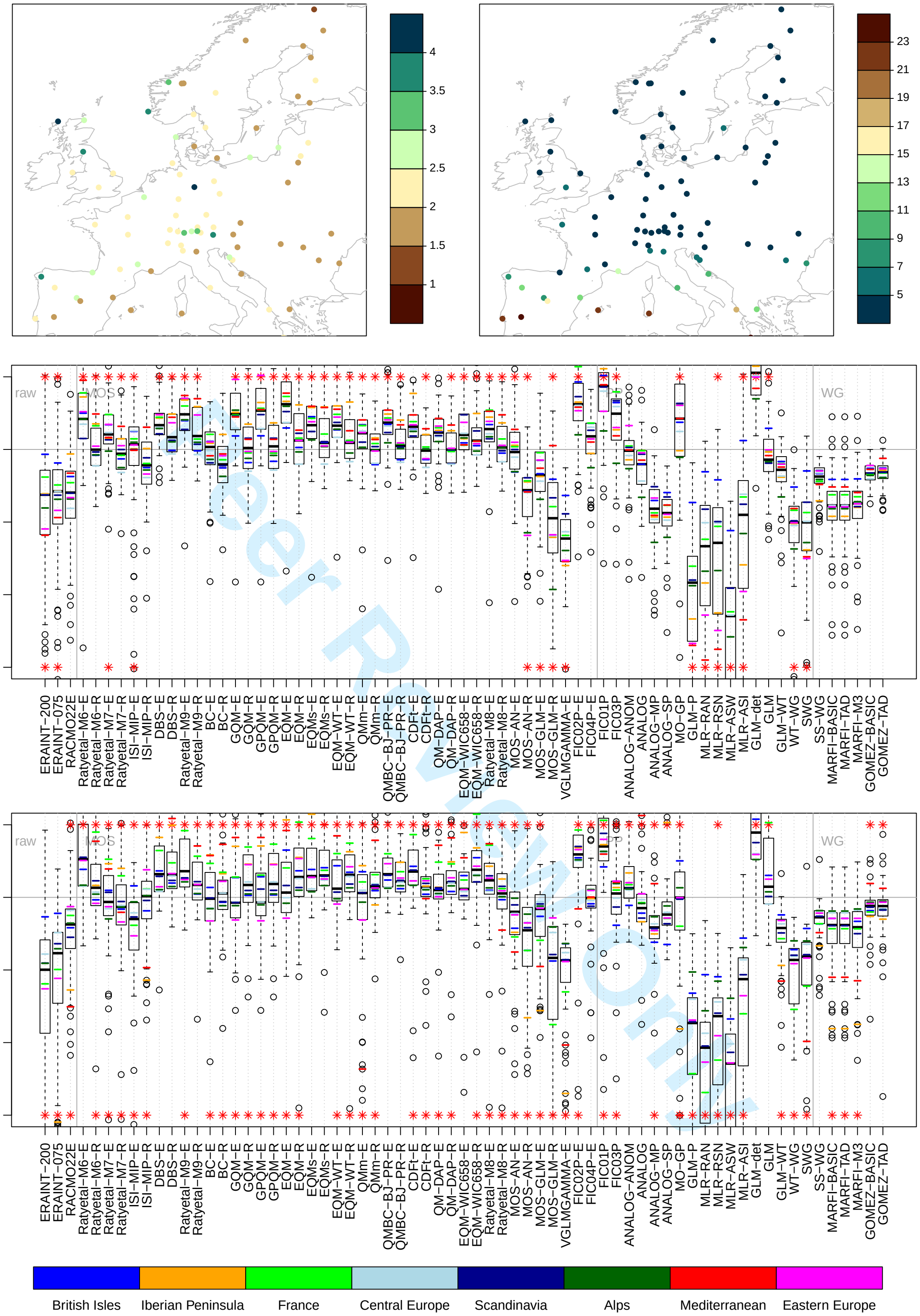

Figure 10: As Fig.3, but for winter WetSpellMean [days] (left/top) and summer DrySpellMean [days] (right/bottom)

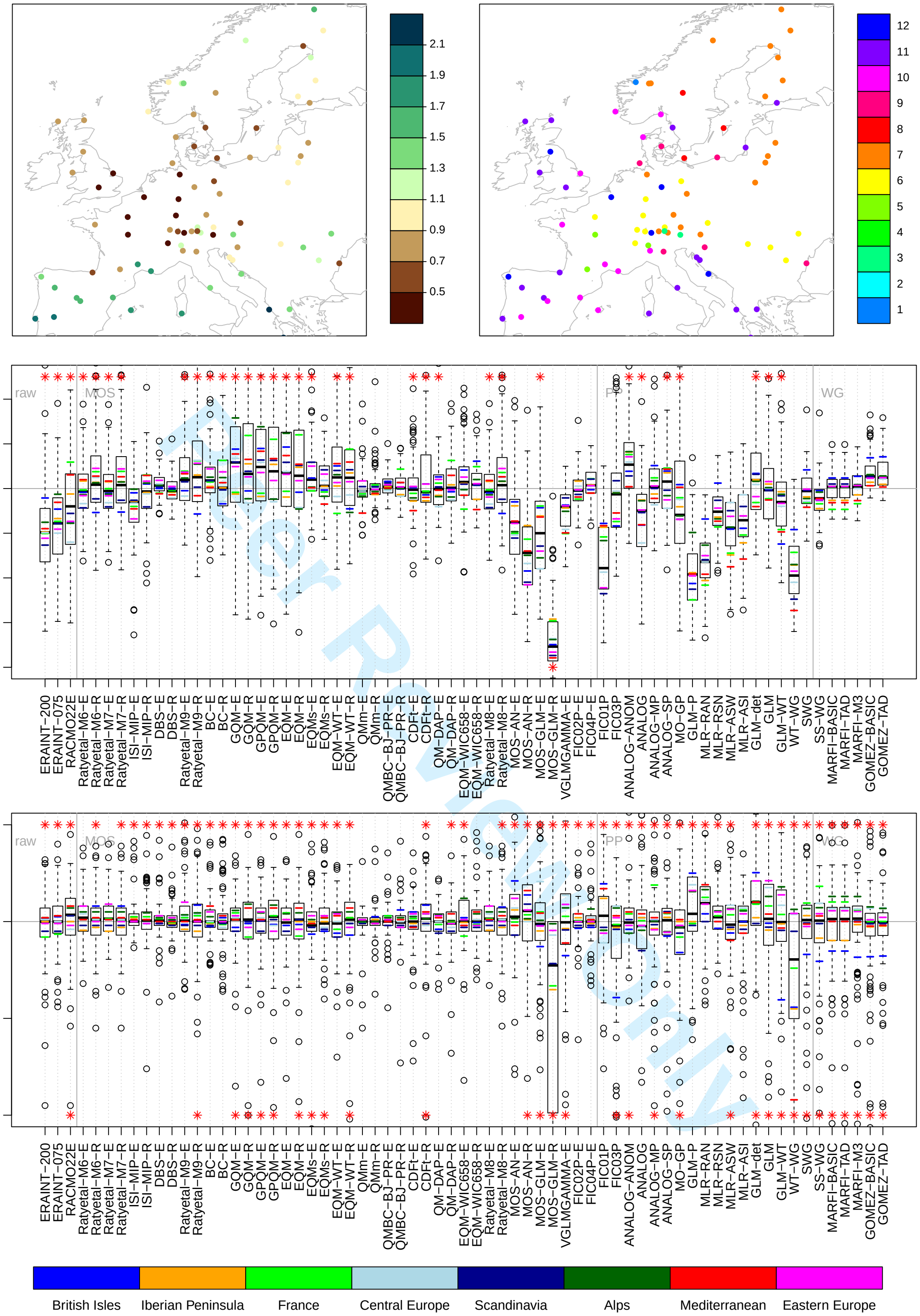

Figure 11: As Fig.3, but for the relative amplitude (left/top) and phase [days] (right/bottom) of the annual cycle of precipitation.

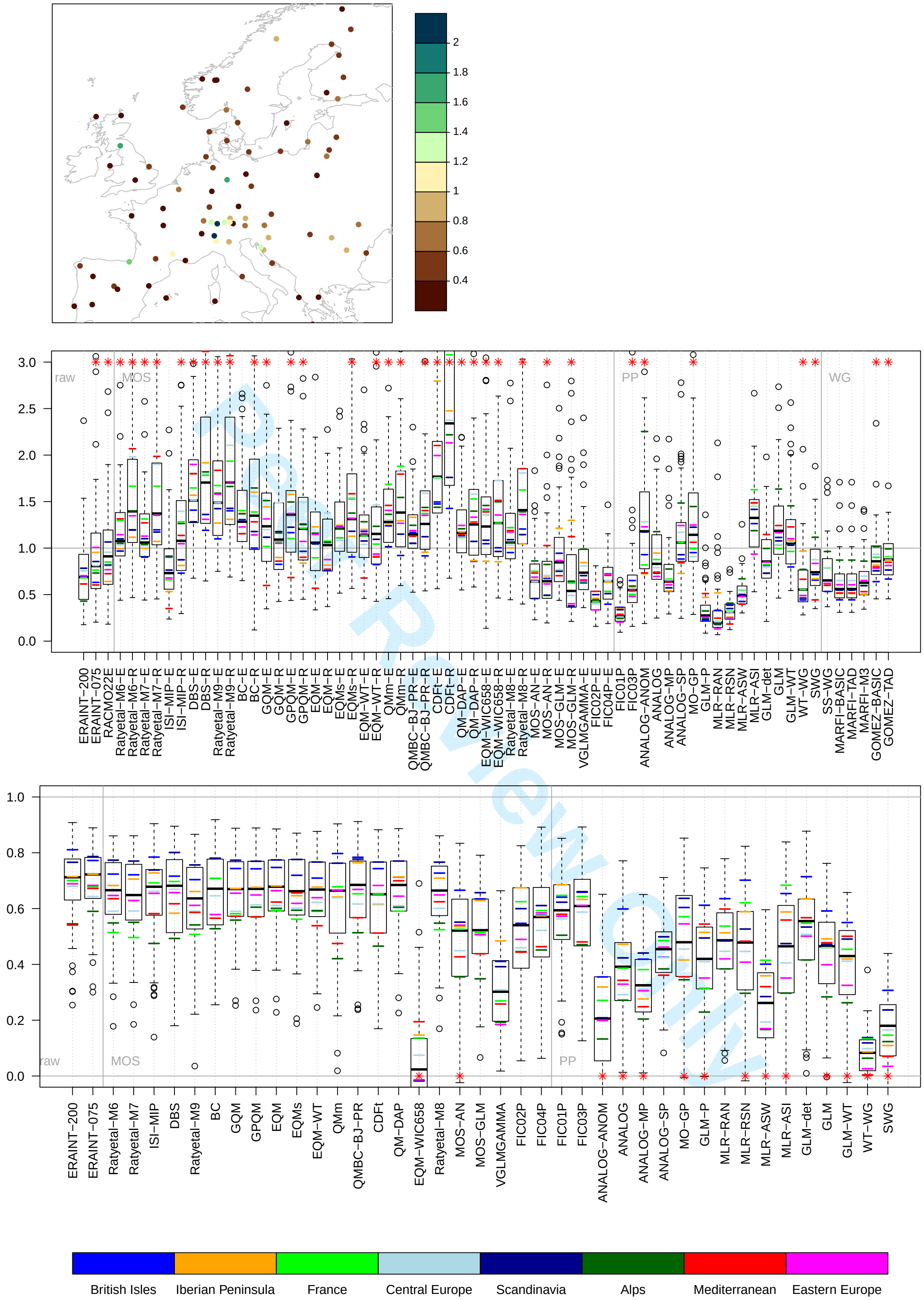

Figure 12: As Fig.3, but for summer VarY [mm^2] (map/top) and Cor.1Y (no map/bottom) of precipitation.

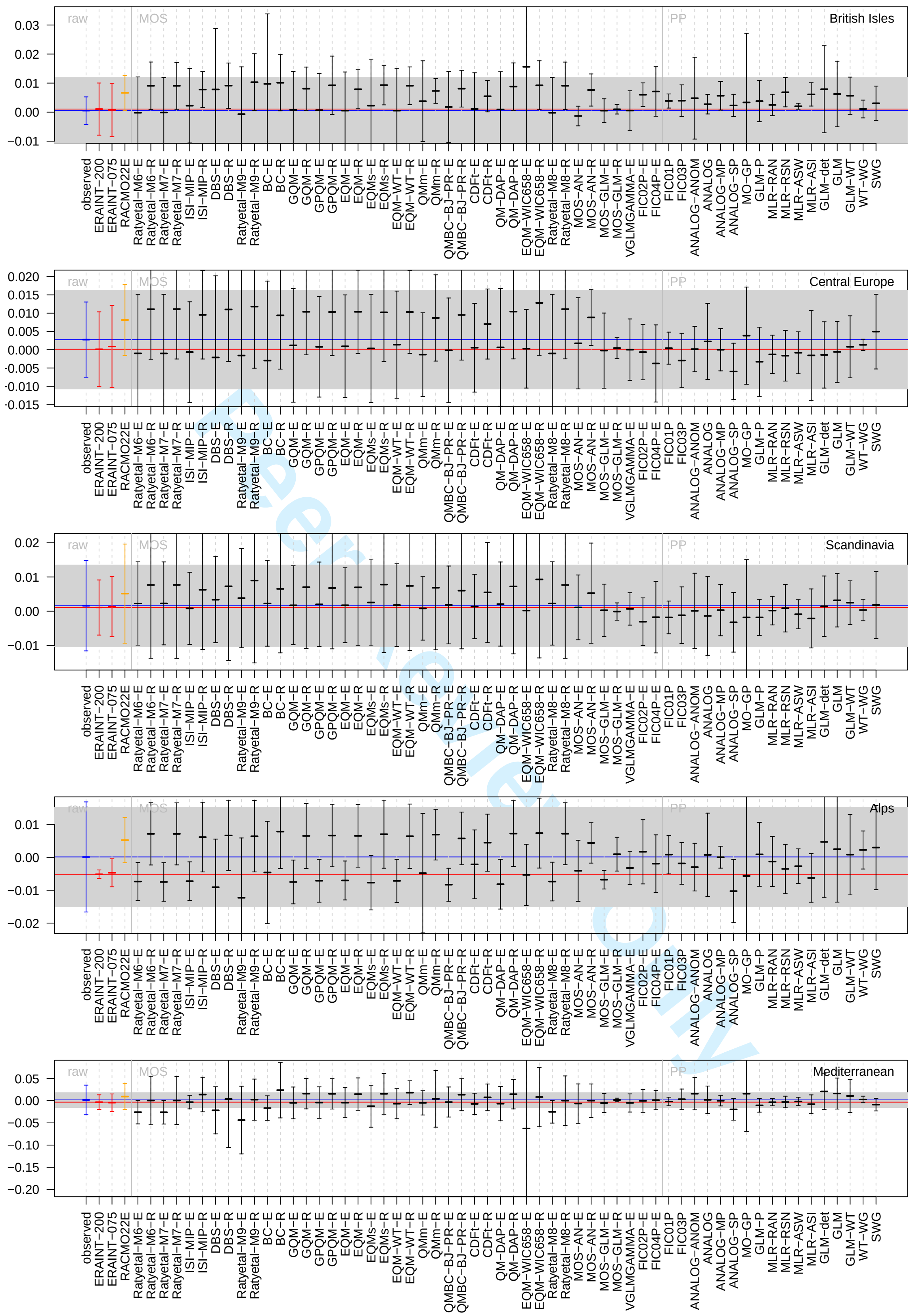

Figure 13: As Fig.3, but for the relative trend in JJA mean precipitation.

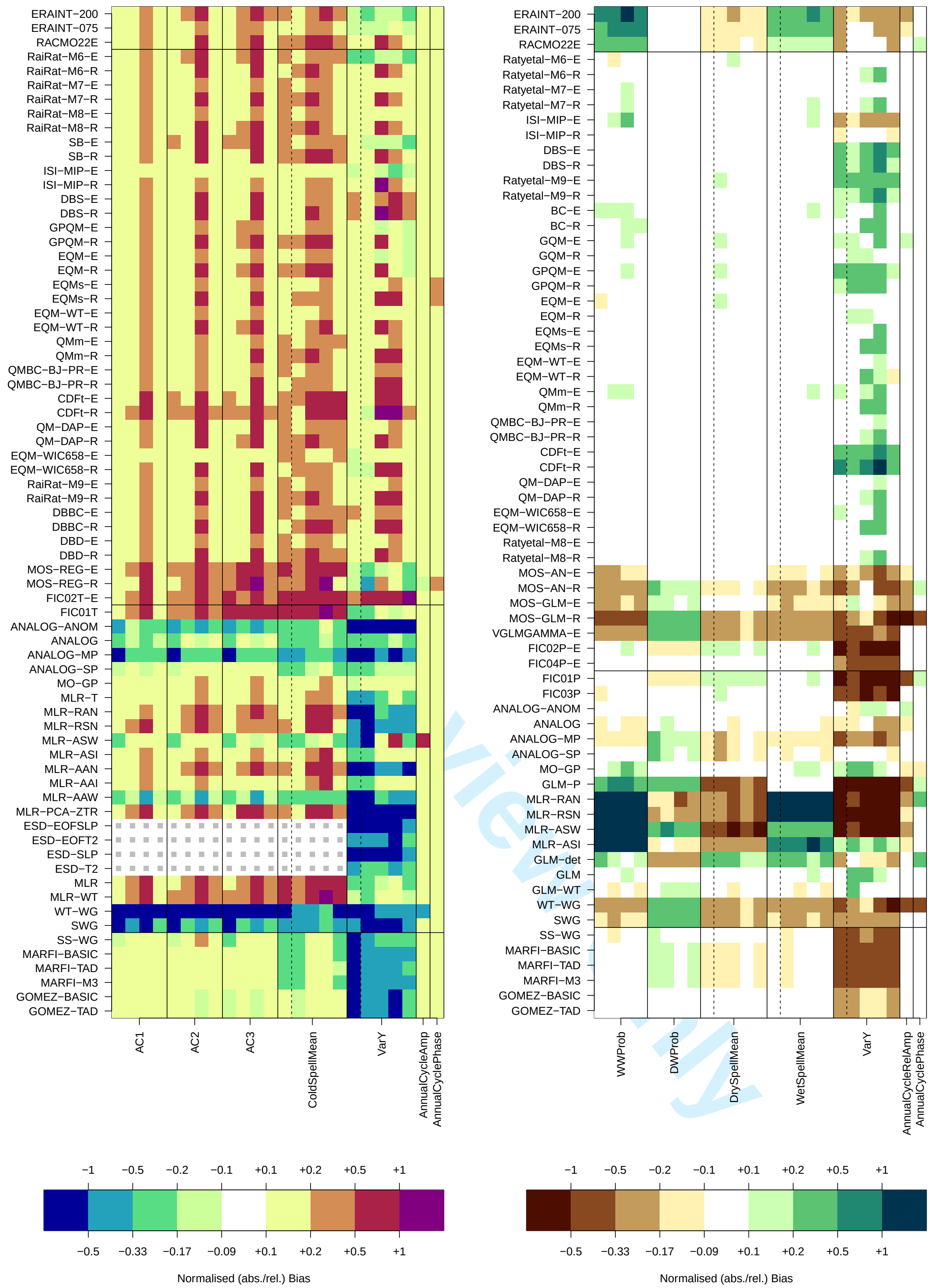

Figure 14: Performance summary. Left: T_{min}, right: precipitation. For each index either the performance for all 4 seasons is shown, or additionally the performance for the whole year (separated by a dashed line), or - in case of the seasonal cycle - ony for the whole year. Grey squares indicate that no values have been calculated. For the scales used for normalisation, see Appendix.