INTERNATIONAL JOURNAL OF CLIMATOLOGY RMetS Royal Meteorological Society

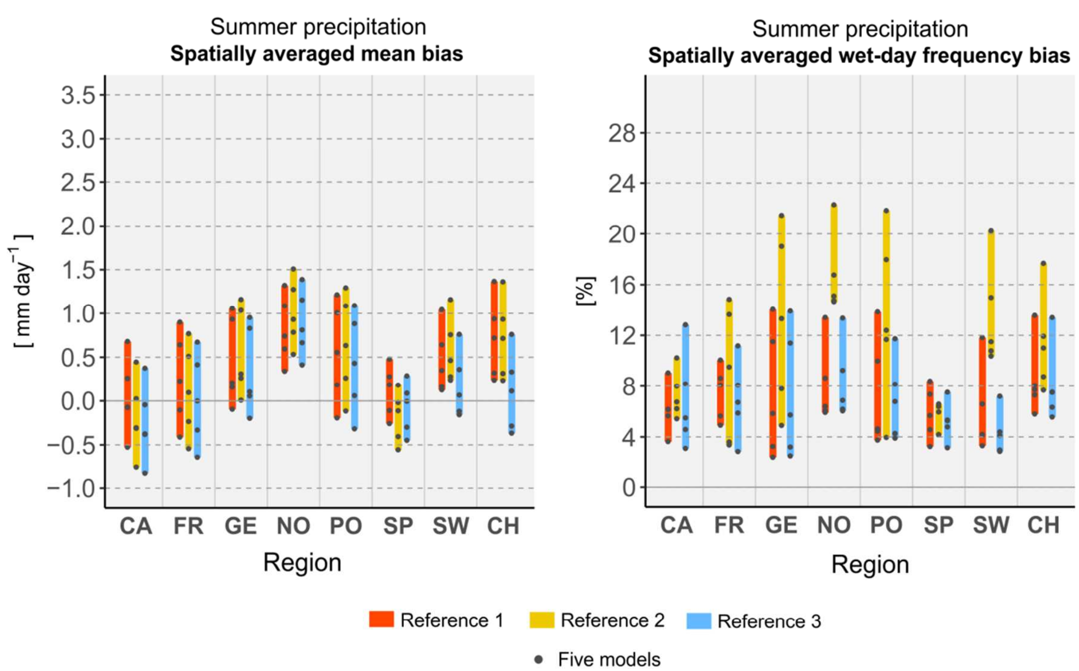
Observational uncertainty and regional climate model evaluation: A pan-European perspective

Journal:	*International Journal of Climatology*
Manuscript ID	JOC-17-0256.R1
Wiley - Manuscript type:	VALUE Special Issue
Date Submitted by the Author:	n/a
Complete List of Authors:	Kotlarski, Sven; Federal Institute of Meteorology and Climatology MeteoSwiss, Climate Szabó, Péter; Hungarian Meteorological Service, Climate and Ambient Air Herrera, Sixto; Universidad de Cantabria, Meteorology Group, Dpto. de Matemática Aplicada y Ciencias de la Computación Räty, Olle; Univ Helsinki, Finland, Department Keuler, Klaus; Brandenburg University of Technology, Chair Environmental Meteorology Soares, Pedro; Instituto Dom Luiz,Universidade de Lisboa, DEGGE Cardoso, Rita; Instituto Dom Luiz, Universidade de Lisboa, DEGGE Bosshard, Thomas; Swedish Meteorological and Hydrological Institute, Hydrology Department Pagé, Christian; UMR CNRS 5318 CECI – CERFACS, Climate modeling and Global change Boberg, Fredrik; Danish Meteorological Institute, Danish Climate Centre Gutiérrez, José; National Research Council (CSIC), Meteorology Group, Instituto de Física de Cantabria Isotta, Francesco; Federal Institute of Meteorology and Climatology MeteoSwiss, Climate Jaczewski, Adam; National Research Institute, Institute of Meteorology and Water Management Kreienkamp, Frank; Deutscher Wetterdienst, Klima- und Umweltberatung Liniger, Mark Andrea; Federal Office of Meteorology and Climatology MeteoSwiss, Climate Lussana, Cristian; Norwegian Meteorological Institute, Observation and Climate Pianko-Kluczyńska, Krystyna; National Research Institute, Institute of Meteorology and Water Management
Keywords:	RCM evaluation, observations, uncertainty, Europe, CORDEX
Country Keywords:	Germany, France, Switzerland, Spain, Sweden

Observational uncertainty and regional climate model evaluation: A pan-European perspective

Sven Kotlarski*, Péter Szabó, Sixto Herrera, Olle Räty, Klaus Keuler, Pedro M. Soares, Rita M. Cardoso, Thomas Bosshard, Christian Pagé, Fredrik Boberg, José M. Gutiérrez, Francesco A. Isotta, Adam Jaczewski, Frank Kreienkamp, Mark A. Liniger, Cristian Lussana, Krystyna Pianko-Kluczyńska

Five state-of-the-art reanalysis-driven regional climate model experiments are evaluated against three different observational reference datasets for two variables (temperature and precipitation) and for eight sub-regions of the European continent. Overall, we find the influence of observational uncertainty to be smaller than model uncertainty. For individual regions and seasons, however, model evaluation can considerably depend on the chosen reference and final model ranks can be strongly influenced.



Observational uncertainty and regional climate model evaluation: A pan-European perspective

Sven Kotlarski[1], Péter Szabó[2], Sixto Herrera[3], Olle Räty[4], Klaus Keuler[5], Pedro M. Soares[6], Rita M. Cardoso[6], Thomas Bosshard[7], Christian Pagé[8], Fredrik Boberg[9], José M. Gutiérrez[10], Francesco A. Isotta[1], Adam Jaczewski[11], Frank Kreienkamp[12], Mark A. Liniger[1], Cristian Lussana[13], Krystyna Pianko-Kluczyńska[11]

[1]Federal Office of Meteorology and Climatology MeteoSwiss, Switzerland
[2]Hungarian Meteorological Service, Hungary
[3]Meteorology Group, Dpto. de Matemática Aplicada y Ciencias de la Computación. Universidad de Cantabria, Spain
[4]University of Helsinki, Finland
[5]Brandenburg University of Technology, Germany
[6]Instituto Dom Luiz, Faculdade de Ciências, Universidade de Lisboa, Portugal
[7]Swedish Meteorological and Hydrological Institute, Sweden
[8]UMR CNRS 5318 CECI – CERFACS, France
[9]Danish Meteorological Institute, Denmark
[10]Meteorology Group, Instituto de Física de Cantabria (CSIC-Univ. de Cantabria), Spain
[11]Institute of Meteorology and Water Management – National Research Institute, Poland
[12]Deutscher Wetterdienst, Germany
[13]Norwegian Meteorological Institute, Norway

Abstract

The influence of uncertainties in gridded observational reference data on regional climate model (RCM) evaluation is quantified on a pan-European scale. Three different reference datasets are considered: the coarse-resolved E-OBS dataset, a compilation of regional high-resolution gridded products (HR) and the European-scale MESAN reanalysis. Five high-resolution ERA-Interim driven RCM experiments of the EURO-CORDEX initiative are evaluated against each of these references over eight European sub-regions and considering a range of performance metrics for mean daily temperature and daily precipitation. The spatial scale of the evaluation is 0.22°, i.e. the grid spacing of the coarsest dataset in the exercise (E-OBS).

While the three reference grids agree on the overall mean climatology, differences can be pronounced over individual regions. These differences partly translate into RCM evaluation uncertainty. Still, for most cases observational uncertainty is smaller than RCM uncertainty. For individual sub-regions and performance metrics, however, observational uncertainty can dominate. This is especially true for precipitation and for metrics targeting the wet-day frequency, the pattern correlation and the distributional similarity. In some cases also the spatially averaged mean bias can be considerably affected.

An illustrative ranking exercise highlights the overall effect of observational uncertainty on RCM ranking. Over individual sub-domains, the choice of a specific reference can modify RCM ranks by up to four levels (out of five RCMs). For most cases, however, RCM ranks are stable irrespective of the reference. These results provide a two-fold picture: model uncertainty dominates for most regions and for most performance metrics considered, and observational uncertainty plays a minor role. For individual cases, however, observational uncertainty can be pronounced and needs to be definitely taken into account. Results can to some extent also depend on the treatment of potential precipitation undercatch in the observational reference.

Keywords

RCM evaluation, observations, uncertainty, Europe, CORDEX

1. Introduction

The existence and availability of reliable high-quality observational data is essential for climate monitoring. It is furthermore the basis for the development, evaluation and application of both physically-based and statistical weather and climate models. This includes downscaling approaches that translate large-scale atmospheric features into higher-resolved and even point-scale information (e.g., Fowler et al., 2007). Observations are already used during model development, but also model calibration and initialization often heavily rely on an existing observational reference (e.g., Bellprat et al., 2012). As such, the quality of any model-derived weather or climate product can be expected to depend on the quality of the underlying observations. The same is true for model evaluation exercises that assess and inter-compare the performance of one or several modelling systems by comparison against observation-based records (e.g., Christensen et al., 2010, Kotlarski et al., 2014). Consequently, uncertainties in the observational reference directly translate into uncertainties of model evaluation results.

Observational uncertainties themselves can be large and originate from multiple sources. Already raw observations are likely to suffer from inaccuracies due to residual non-climatic influences (Hartmann et al., 2013, Hegerl et al., 2001, McMillan et al., 2012). Such influences include malfunctions and error margins of measurement devices and, in case of long-term records, replacements of the device, relocations of the measurement site or physical changes of the surrounding landscape. For the case of precipitation, site measurements are furthermore subject to systematic biases due to the local deformation of the wind field by the gauge and wetting and evaporation losses. This systematic undercatch is pronounced for windy conditions and for snowfall and can result in an important underestimation of true precipitation sums (e.g., Adam and Lettenmaier, 2003, Cheval et al., 2010, Frei et al., 2003, Groisman and Legates, 1994, Sevruk, 1985, Wolff et al., 2015). Some of the mentioned inaccuracies can be reduced by postprocessing the raw measurement records, e.g. by applying data homogenization procedures (Begert et al., 2005) or a dedicated precipitation undercatch correction (Richter, 1995). Additionally, representativity issues arise for point measurements, i.e. the question to what extent a point record reflects conditions for a larger area, for instance the mean conditions over a climate model grid box obtained through averaging all subgrid variabilities in space (e.g., Osborn and Hulme, 1997).

To avoid the latter complication, climate model evaluation wherever possible relies on gridded reference datasets that are obtained by a spatial analysis and interpolation of point measurements onto a regular grid yielding area-representative grid cell mean values. Additionally, gridded remote sensing products and model-derived reanalyses are used. In any case, the gridding procedure itself involves assumptions and uncertainties with corresponding effects on the final product. For gridded datasets obtained by spatial interpolation of point measurements problems arise especially in regions with sparse data coverage, complex topography and for variables with a high spatio-temporal climatic variability (e.g., Wagner et al., 2007). Spatial variance, for instance, is mostly underestimated by gridded products (Beguería et al., 2016) and trends can be affected by a temporally changing network density (e.g., Frei, 2014, Hofstra et al., 2009). Sampling issues due to random natural climate variability, i.e. the fact that the observed record is only one possible realization of the analysis period's climate, can introduce further uncertainties (e.g., Addor and Fischer, 2015, Mahlstein et al., 2015).

In summary, any available observation-based record is unlikely to reflect the true state of atmospheric quantities but only some approximation of it. A number of studies exist that quantify the related observational uncertainty by comparing several observation-based reference datasets for specific variables and regions (e.g., the recent works by Awange et al., 2016, Berg et al., 2016, Dunn et al., 2014, Gbambie et al., 2017, Gervais et al., 2014, Herold et al., 2016, Hofstra et al., 2009, Isotta et al., 2015, Kyselý and Plavcová, 2010, Palazzi et al., 2013, Rauthe et al., 2013, Schneider et al., 2014; Tanarhte et al., 2012). In evaluation exercises these shortcomings of the reference inevitably influence the performance assessment of climate models and introduce uncertainties in the evaluation results. Previous works have addressed this issue by employing multiple reference data sources for global and regional climate model (GCM, RCM) evaluation (Addor and Fischer, 2015, Bellprat et al., 2012, Brienen et al., 2016, Bucchignani et al., 2016, Casanueva et al., 2013, Cheneka et al., 2016, Davin et al., 2016, Di Luca et al., 2012, Gómez-Navarro et al., 2012, Haslinger et al., 2013, Kotlarski et al., 2005, Kotlarski et al., 2012, Maraun et al., 2012, Prein and Gobiet, 2017, Ring et al., 2016, Sunyer et al., 2013). Besides quantifying the influence of observational uncertainty on individual model performance scores, two of these studies (Gómez-Navarro et al., 2012 and Sunyer et al., 2013) also explicitly address the modification of model ranks when changing the observational reference.

Most of the mentioned works consider geographic domains of limited extent only, such as individual river catchments or countries, and focus on precipitation. At this point, we refrain from listing the individual results but note that (1) even in regions covered by dense observational networks observational uncertainty can be large and can be comparable to RCM uncertainty (measured by the spread between individual RCM experiments) and that (2) observational uncertainty can have the potential to influence the outcome of climate model weighting and ranking exercises. Among the mentioned works a particularly relevant study is the one by Prein and Gobiet (2017) who, focusing on precipitation, inter-compared a large number of gridded observational datasets over parts of the European continent and used this observational ensemble to evaluate state-of-the-art RCM experiments. They found that observational uncertainty can be of similar magnitude as RCM biases, particularly in regions of low station density and for high temporal and spatial resolution statistics.

In the present work we build upon and complement these previous studies by quantifying observational uncertainty on a pan-European scale not only for precipitation but also for temperature and by assessing its influence on RCM evaluation in a well-defined performance assessment framework. We explicitly include an illustrative model ranking exercise and relate observational spread to RCM spread. Our main objective is to illustrate the influence of observational uncertainty on RCM evaluation and RCM ranking for different European sub-regions, for two variables and for a range of performance scores reflecting different model bias characteristics.

2. Data and Methods

2.1 Observational Reference Data

To sample observational uncertainty we employ three observational reference grids that are available (1) for both mean temperature and precipitation, (2) at a daily resolution, (3) for the common 18-year long evaluation period 1989-2006, and (4) at a grid spacing comparable to or higher than the current RCM resolution for multi-decadal climate projections. Note that the latter criterion does not necessarily imply a higher effective resolution of the observational datasets compared to the RCMs. Depending on the underlying network density the effective resolution of the data could be considerably lower than the nominal grid spacing (e.g., Beguería et al., 2016, Isotta et al., 2015, Prein and Gobiet, 2017) . The three observational reference grids represent an "ensemble of opportunity", i.e. we consider datasets that are readily available, that fulfil the above-mentioned criteria and that include the evaluation of climate models in their intended range of application. We hence accept inter-dependencies of the three datasets that could arise, for instance, from the use of the same station series for gridding or calibration purposes or from similar gridding concepts. In particular, we combine reference datasets that result from an explicit gridding procedure of observations with a reanalysis-based product. We also do not intend to provide final explanations for differences among the three reference datasets. This would imply a much more detailed analysis of the influence of the gridding process and of different network densities on the final gridded product and would go beyond the scope of the present work. These aspects are covered by the accompanying study of Herrera et al. (2017). Furthermore, note that we use the term *observations* for results from both gridding processes and reanalysis procedures. This contrasts with other, more direct definitions of *observations* based on actual station data or remote sensing results. We hence do not explicitly differentiate between *observational uncertainty* and *gridding uncertainty* and use the former term to capture both.

2.1.1 E-OBS

The gridded E-OBS dataset (Haylock et al., 2008; version 15) covers the entire European land surface and is based on the ECA&D (European Climate Assessment and Dataset) station data plus more than 2000 further stations from additional archives. We here use the daily temperature and precipitation grids of the rotated 0.22° version (approx. 25 km grid spacing). For several years E-OBS has now been a standard reference for RCM evaluation over the European continent. Known deficiencies of E-OBS relate to remaining inhomogeneities in the station series and to the dataset's quality in regions of sparse station density. The latter particularly affects the representation of daily extremes (e.g., Bellprat et al., 2012; Herrera et al., 2012; Hofstra et al., 2009, Hofstra et al., 2010, Lenderink, 2010, Maraun et al., 2012) and the effective spatial resolution which is presumably lower than the nominal 0.22° grid spacing (e.g., Hanel and Buishand, 2011; Kyselý and Plavcová, 2010). The systematic undercatch of rain gauges (e.g., Sevruk, 1986) has not been corrected for, i.e., E-OBS likely underestimates true precipitation sums.

2.1.2 National High-Resolution Grids (HR)

Our second observational reference is a compilation of national/regional high-resolution temperature and precipitation grids that are available for parts of the European continent only (Fig. 1). This dataset has been assembled within the COST Action VALUE (Maraun et al., 2015). It covers modified sets of regions and datasets compared to the recent work of Prein and Gobiet (2017), including one additional country (Poland), an updated version of the Norwegian and the German dataset and the consideration of Switzerland only instead of the entire Alps, employing a different high-resolution observational grid. In overlapping boarder regions covered by two national datasets only one of them has been considered[1]. In the following a brief description of each dataset is provided. Except for the Swedish product, none of the precipitation grids explicitly accounts for the systematic undercatch of rain gauges.

Spain (SP): For peninsular Spain and the Balearic Islands an improved 3-dimensional areal representative version (AA-3D) of the Spain02 gridded dataset at 0.22° grid spacing on a rotated grid is used (Herrera et al. 2012; 2016). Spain02 is based on a very dense and quality-controlled station network consisting of 2756 and 237 stations for precipitation and temperature, respectively. The interpolation and gridding procedure is the same as applied for E-OBS.

Poland (PO): The AA-3D methodology used for the Spanish grid was extended to build an observational grid for Poland based on a quality-controlled observational station dataset provided by the Institute of Meteorology and Water Management - National Research Institute, Center for Poland's Climate Monitoring; see Herrera et al. (2017) for further details. This dataset comprises 197 stations for precipitation and 123 for temperature. Station data were homogenized prior to the

[1] In the following pairs of overlapping countries/regions the bold country/region has been considered: **NO**/SW, **SP**/FR, **CH**/GE, **FR**/GE, **CH**/FR, **PO**/CA, **GE**/PO. In case of the Carpathian dataset, which extends far into Poland, this means a substantial cut-off at its northern boundary.

gridding by applying the MASH v3.03 procedure (e.g., Szentimrey, 2013) to the daily data (Lakatos et al. 2013).

France (FR): The France national high-resolution analysis SAFRAN is available at an hourly time step and on a grid of 8 km spacing (Durand et al., 1993; Quintana-Seguí et al. 2008; Vidal et al., 2010). It is based on observations at more than 4000 sites collected by Météo-France as well as on operational Numerical Weather Prediction analyses along with some climatological data. It covers all water basins affecting Metropolitan France including Corsica. Prior to its use within the present work SAFRAN was conservatively interpolated to the rotated 0.11° EURO-CORDEX grid.

Sweden (SW): The daily gridded PTHBV dataset provides daily precipitation and temperature data at 4 km grid spacing and covers Sweden plus some adjacent regions. The product is based on more than 350 (800) stations for temperature (precipitation) and has been constructed by optimal interpolation with a climatological background field that accounts for wind-orography effects (Johansson and Chen, 2003). In the present work it is the only dataset that has been corrected to account for the systematic undercatch of rain gauges. The correction is based on gauge type, precipitation type (rain or snow), wind classification and exposure of the gauges (Berg et al., 2016).

Germany (GE): The high resolution daily gridded HYRAS dataset has been produced as part of the KLIWAS research programme (*Impacts of climate change on waterways and navigation – searching for options of adaptation*; www.kliwas.de). It covers the period 1951 to 2006 and is available at 5 km grid spacing for all river catchments in Germany as well as adjacent river basins with drainage towards Germany (i.e. the entire Rhine, Danube and Elbe catchments). More detailed information about the dataset and its underlying station network, which consists of up to 1000 and 6200 stations for temperature and precipitation, respectively, is provided by Rauthe et al. (2013) and Frick et al. (2014).

Carpathians (CA): The CARPATCLIM gridded observational dataset (Lakatos et al., 2013) covers parts of 9 countries along the Carpathian Mountains and is based on raw station time series that were exchanged along the borders to ensure data homogeneity (temperature: 258 stations, precipitation: 727 stations). Quality control and homogenization were carried out at daily resolution using the MASH software (Szentimrey, 2004). The MISH package (Szentimrey and Bihari, 2007) was employed for spatial interpolation. The publicly available CARPATCLIM dataset for 11 variables is provided at daily temporal resolution and 0.1° grid spacing for the period 1961-2010 (www.carpatclim-eu.org). Note that, in contrast to the other national/regional grids, CARPATCLIM does not represent areal grid cell averages but point estimates for the grid cell centers.

Norway (NO): The gridded seNorge version 2 (seNorge2) dataset is based on two modified optimal interpolation schemes (Gandin, 1965), one for temperature and one for precipitation, in which the prior distribution is estimated from in-situ observations (Lussana et al., 2016, Uboldi et al., 2008). The input data used are original non-homogenized station series from the Norwegian Climate Database (480 and 920 stations on average for temperature and precipitation, respectively). Data for both variables are provided at daily resolution on a 1 km grid.

Switzerland (CH): For the region of Switzerland, the TabsD (temperature; MeteoSwiss, 2013a) and RhiresD (precipitation; MeteoSwiss, 2013b) datasets at 2 km grid spacing are used. Both datasets rely on a large but temporally varying number of station series (temperature: 93, precipitation: about 520) and were produced accounting for the special requirements of interpolating station data in topographically complex terrain (e.g., Frei, 2014).

2.1.3 EURO4M MESAN

The European Reanalysis and Observations for Monitoring project (EURO4M) has produced several gridded datasets for Europe, among others a High Resolution Limited Area Model (HIRLAM) reanalysis at a grid spacing of 0.2° (approx. 22 km) using 3D-VAR data assimilation (Dahlgren and

Gustafsson, 2012). Several simulated surface fields – including near-surface air temperature and precipitation - have afterwards been downscaled with the MESAN system to a 0.05° grid (approx. 5 km) using optimal interpolation techniques (Häggmark et al., 2000) and assimilating further surface observations. Depending on the region, the number of stations used for assimilation and interpolation is partly larger and partly smaller or comparable to E-OBS (see Fig. 1 in Prein and Gobiet, 2017). For precipitation, the surface observations assimilated in the MESAN downscaling step were not corrected for the measurement bias of rain gauges. Hence the final EURO4M MESAN precipitation product – although originating from simulated precipitation of the HIRLAM model – has to be assumed to be undercatch-affected.

2.2 RCM Data

The RCM simulations that are evaluated in the present work originate from the EURO-CORDEX initiative (Jacob et al., 2014) and have been carried out at a grid spacing of 0.11° on a rotated grid under the CORDEX simulation protocol. All experiments cover a full European domain (see Kotlarski et al., 2014) and were driven by the ERA-Interim reanalysis (Dee at al., 2011) at the lateral domain boundaries. We hence evaluate the so-called *perfect boundary experiments* instead of the GCM-driven historical control runs. Such an evaluation places a stronger focus on the downscaling performance itself as potentially strong biases in the GCM-derived boundary forcing are avoided. In total, five simulations are used (Table 1) that form a subset of those experiments considered in the EURO-CORDEX standard evaluation (Kotlarski et al., 2014). Note that two of the five RCMs employed (HIRHAM 5 and RACMO 2.2E) as well as the reanalysis-model (MESAN; see above) originate from the numerical weather prediction model HIRLAM and partly share the same code. Hence, their respective outputs cannot be considered to be fully independent of each other.

2.3 Analysis Domain and Analysis Grid

The analysis domain of the present work consists of the eight regions covered by HR (Fig. 1; Section 2.1.2) and samples an important part of continental-scale climate variability in Europe. To enable a consistent comparison on a grid cell level the higher-resolved HR, MESAN and RCM data (including elevation) were conservatively aggregated to the rotated 0.22° E-OBS grid, i.e. to the coarsest grid considered in this work, prior to the analysis. This enables a grid-cell-by-grid-cell comparison and avoids the additional interpolation of E-OBS to the higher-resolved RCM grid. This procedure is also beneficial in case that the effective resolution of a certain dataset is smaller than its nominal grid spacing; spatial aggregation would then more accurately represent the effective resolution of the data. For temperature an additional elevation correction from the aggregated HR, MESAN and RCM elevation to the elevation of the corresponding E-OBS grid cell was carried out assuming a spatially and temporally uniform lapse rate of 0.0065 °C m^{-1}.

2.4 Performance Metrics

The performance of the RCMs was evaluated on the common 0.22° analysis grid and separately for each of the eight sub-regions. Seven different metrics were chosen which describe different aspects of model performance. Five of these metrics were computed for both temperature and precipitation and one further metric was calculated for temperature or precipitation only, resulting in six metrics for each variable.

For each observational reference dataset the metrics were calculated for every climate model j, season k, and analysis region r. For the sake of simplicity, those indices are omitted in the following. We define O_n and X_n to be daily observational and climate model data, respectively, at a particular grid point n within the analysis region r that contains a total of N grid points. Further, overbars denote the temporal mean over all time steps in the analysis period that fall into the season k, two overbars denote temporal and spatial mean, and yearly seasonal means are denoted by the index y.

The performance of the climatological seasonal mean averaged over a sub-region was evaluated by the bias given as

$$\boldsymbol{BIAS} = \frac{1}{N}\sum_{n=1}^{N}(\bar{X}_n - \bar{O}_n) \qquad \text{(Eq. 1)}$$

Moderate extremes at the upper end of the distribution were evaluated by the mean absolute error of the 99th percentile:

$$\boldsymbol{MAE99} = \frac{1}{N}\sum_{n=1}^{N}|P^{99}(X_n) - P^{99}(O_n)| \qquad \text{(Eq. 2)}$$

with P^{99} denoting the percentile function for the 99th percentile. For precipitation, all-day percentiles (including the dry days) were used. Note the absolute nature of MAE99 and the fact that, in contrast to the BIAS metric, under- and overestimations of P^{99} at individual grid cells within a given sub-region do not compensate each other.

The similarity of the spatial pattern of climatological seasonal means was assessed using pattern correlation as defined by the Pearson product-moment coefficient of linear correlation

$$\boldsymbol{PACO} = \frac{cov(\bar{X}_n, \bar{O}_n)}{sd(\bar{X}_n)sd(\bar{O}_n)} \quad , n = 1..N \qquad \text{(Eq. 3)}$$

with cov and sd representing the spatial covariance and standard deviation, respectively.

The interannual variability of seasonal means was evaluated using the ratio of interannual variability (RIAV). The spatial and temporal means of a season were first calculated for every year separately, and the standard deviations were then related according to

$$\boldsymbol{RIAV} = \frac{sd(\bar{\bar{X}}_y)}{sd(\bar{\bar{O}}_y)} \qquad \text{(Eq. 4)}$$

The Cramér-von Mises Test (CMT; Anderson, 1962, Lunneborg, 2005) was used to evaluate the similarity of the cumulative distribution functions of daily values. In the case of precipitation, only the wet days were considered (wet-day threshold of 1mm day^{-1}). In order to remove the influence of the bias in the mean (which is evaluated already by the BIAS metric) the climate model data were first corrected for the mean bias. For temperature and precipitation this was done by additive and multiplicative correction, respectively. After the bias correction, the CMT was applied to every grid point separately resulting in a probability value for rejection p_n. Using a significance level of 0.05, the fraction of grid-points with non-rejection (i.e., the null-hypothesis of the two distributions being similar cannot be rejected at a probability of 0.05) was calculated. The latter represents the final Cramér-von Mises performance metric CM. In mathematical terms, this can be described as follows:

$$p_n = CMT(X_n, O_n) \qquad \text{(Eq. 5)}$$

$$\begin{matrix} c_n = 1 & if\ p_n > 0.05 \\ c_n = 0 & if\ p_n \le 0.05 \end{matrix} \qquad \text{(Eq. 6)}$$

$$\boldsymbol{CM} = \frac{1}{N}\sum_{n=1}^{N} c_n \qquad \text{(Eq. 7)}$$

Note that this simple version of the metric neglects a potential spatial autocorrelation of the test statistic and does not consider field significance (e.g., Ivanov et al. 2017a and 2017b). Two further metrics were only calculated for either temperature or precipitation. For temperature only, the mean absolute error of the 1st percentile was used to evaluate moderately cold extremes:

$$\boldsymbol{MAE01} = \frac{1}{N}\sum_{n=1}^{N}|P^1(X_n) - P^1(O_n)| \qquad \text{(Eq. 8)}$$

For precipitation only, the mean absolute bias in the wet-day frequency was evaluated by

$$\boldsymbol{WDFREQ} = \frac{1}{N}\sum_{n=1}^{N}|wdfr(X_n) - wdfr(O_n)| \qquad \text{(Eq. 9)}$$

with $wdfr()$ being the wet-day frequency [%] for a given grid point and a given season for a wet-day threshold of 1 mm day^{-1}.

2.5 Uncertainty Intercomparison

A dedicated comparison framework was employed to quantify the relation between *observational uncertainty* (the influence of the choice of the reference dataset on the evaluation) and *model uncertainty* (the effect of the choice of a specific RCM on the evaluation). In case observational uncertainty is large, model evaluation against one specific reference dataset has to be considered as non-robust and evaluation exercises need to definitely take into account observational uncertainty.

Let $P_{i,j}$ be the value of given performance metric for a given variable, sub-region and season when employing reference dataset i ($i \in \{1,2,3\}$) for evaluating RCM j ($j \in \{1,2,3,4,5\}$). Observational uncertainty is defined as the mean standard deviation of the metric's values when comparing an RCM against each of the three reference datasets:

$$U_{OBS} = \frac{\sum_{j=1}^{5}\sqrt{\frac{1}{2}\sum_{i=1}^{3}\left(P_{i,j} - \frac{1}{3}\sum_{i=1}^{3}P_{i,j}\right)^2}}{5} \qquad \text{(Eq. 10)}$$

Correspondingly, model uncertainty is defined as the mean standard deviation of the respective metric's values when comparing all $\binom{5}{3} = 10$ three-member RCM sub-ensembles against a given reference dataset:

$$U_{MOD} = \frac{\sum_{i=1}^{3}\left(\frac{1}{10}\sum_{n=1}^{10}\sqrt{\frac{1}{2}\sum_{j\in S_n}\left(P_{i,j} - \frac{1}{3}\sum_{j\in S_n}P_{i,j}\right)^2}\right)}{3} \qquad \text{(Eq. 11)}$$

where $S_n = \{(1,2,3); (1,2,4); (1,2,5); (1,3,4); (1,3,5); (1,4,5); (2,3,4); (2,3,5); (2,4,5); (3,4,5)\}$. Three-member sub-ensembles are chosen to be consistent with U_{obs}. The ratio

$$R = \frac{U_{OBS}}{U_{MOD}} \qquad \text{(Eq. 12)}$$

for a given metric, variable, sub-region and season then defines the ratio of observational and model uncertainty. If this ratio is larger than 1 observational uncertainty is larger than model uncertainty and, hence, presents an important contribution to overall evaluation uncertainty and should be considered in evaluation exercises. Note that in our case model uncertainty is defined via the spread among different re-analysis driven RCMs. When evaluating RCM experiments that are driven by different GCMs at their lateral boundaries (i.e. the kind of experiments employed for regional climate projections) this spread and, hence, model uncertainty can be expected to be larger.

As mentioned earlier, the observational references except for HR over Sweden have not been corrected for precipitation undercatch and might underestimate true precipitation sums which can have an effect on the uncertainty intercomparison. In a dedicated sensitivity analysis we therefore carried out a modified uncertainty analysis for precipitation. For this purpose, a bulk correction of 20 % was applied to all observational references (E-OBS, MESAN and HR, except for HR over sub-region SW), i.e. daily precipitation amounts were multiplied by a factor of 1.2. This bulk correction might underestimate the undercatch in winter in some regions and overestimate it in summer. It should only be considered as a rough estimate employed to address the principle sensitivity of our uncertainty analysis with respect to the undercatch issue. Simulated precipitation amounts were not modified. Uncertainty ratios R were re-computed employing the undercatch-corrected observations.

2.6 Ranking Framework

As model selection and weighting schemes (e.g., Christensen et al., 2010) are commonly based on the assessment of a climate model's ability to simulate the present-day climate (Räisänen et al., 2007), part of the uncertainty in these schemes arises from differences between the available reference datasets. To test how the relative performance of the RCMs depends on the selected reference, a simple scheme combining the performance metrics introduced in Section 2.4 was used. First, to ensure that smaller values indicate better RCM performance absolute values were considered for BIAS while RIAV, PACO and CM were transformed according to

$$P' = |1 - P| \quad \text{(Eq. 13)}$$

with P being the value of the respective performance metric. MAE99, MAE01 and WDFREQ were used as computed according to Eqs. 2, 8 and 9, respectively. For a consistent combination of the metrics the values were furthermore normalized (Santer et al., 2009; Rupp et al., 2013) to obtain the respective score $S \in [0,1]$ for a given model j and performance metric m (indices for season k, region r and reference dataset i omitted):

$$S_{j,m} = 1 - \frac{P_{j,m} - \min(P_m)}{\max(P_m) - \min(P_m)} \quad \text{(Eq. 14)}$$

with $\min(P_m)$ and $\max(P_m)$ denoting the minimum/maximum value of the five $P_{j,m}$ (five RCMs j) for the case considered. Note that, in contrast to the performance metric $P_{j,m}$, the larger the value of the score $S_{j,m}$ the better the performance of a particular RCM for a given performance metric. For each reference dataset and each variable, the final overall normalized scores were then calculated separately for each RCM j and region r by taking an average over K seasons and M performance metrics:

$$\bar{S}_{j,r} = \frac{1}{MK} \sum_{m=1}^{M} \sum_{k=1}^{K} S_{j,k,m,r} \quad \text{(Eq. 15)}$$

Thus, equal weight is given to each performance score. The RCM simulations were then ranked according to the obtained $\bar{S}_{j,r}$ values separately for temperature and precipitation (M=6 in Eq. 15). If there are no systematic differences in the relative RCM performance (i.e., $S_{j,k,m,r}$ tends to vary randomly for a given model j and a given region r) $\bar{S}_{j,r}$ is expected to approach 0.5. Combined temperature and precipitation ranks were computed by considering both temperature and precipitation metrics in Eq. 15 (M=12).

A similar scheme with a slightly different set of performance metrics was compared to a more sophisticated scheme by Rupp et al. (2013) and was found to yield qualitatively similar results. One should note that model ranking is inherently subjective (Overland et al., 2011) and depends on the selected climatic aspects, error measures as well as the temporal and spatial scales considered. However, for illustrational purposes the selected scheme is considered sufficient.

3. Reference Data Uncertainty

In order to provide a first impression on the differences among the three reference datasets which will ultimately determine differences in the RCM evaluation exercise we here present a comparison of E-OBS, MESAN and HR in terms of the spatial distribution of climatological seasonal mean values. This comparison is directly relevant for the BIAS metric but might also concern metrics such as

MAE99, MAE01 or PACO. For the comparison we assume HR as reference (due to its highest underlying network density) and display the differences of E-OBS and MESAN with respect to HR.

Figure 2 shows the spatial distribution of seasonal mean temperature in HR and the corresponding deviations of E-OBS and MESAN. All three datasets agree on the general continental-scale temperature gradients and on large-scale mean values (not explicitly shown but deducible from Figure 2). Differences, however, appear over individual sub-regions and are obviously connected to the merging of different regional grids in the HR dataset and to complex orography. Over the Spanish Highlands, the Scandinavian Alps, Switzerland, south-western France and the Carpathians both E-OBS and MESAN can considerably deviate from HR in both seasons. Over the Carpathians these differences are systematic in the sense that HR provides the highest temperatures. When moving to Poland, i.e. into a region covered by a different sub-regional dataset in HR, a close agreement between E-OBS, MESAN and HR is obtained in both seasons. Over south-western France, in contrast, HR systematically shows lower temperatures. Over most parts of Spain MESAN yields lower winter temperatures than HR with differences partly larger than 2°C. Again, this bias pattern disappears when moving into France where mean winter temperatures in HR and MESAN closely agree. E-OBS shows the highest temperatures over FR in both seasons with differences to HR often larger than 0.5°C. This might be connected to the fact that most French station data underlying E-OBS represents larger urban settings possibly affected by the urban heat island effect (see, e.g., http://www.ecad.eu/download/stations.txt). Further consistent features are higher temperatures in E-OBS and MESAN over parts of Scandinavia and the European Alps.

Regarding mean seasonal precipitation all reference datasets again agree on the basic continental scale patterns and on large-scale mean values (not explicitly shown but deducible from Figure 3). A noticeable difference in comparison to HR is the general underestimation of precipitation by both E-OBS and MESAN in both seasons and for most parts of the analysis domain. Deviations can be as large as 50% (e.g. Poland and Sweden in MESAN with respect to HR). Exceptions are the complex coastline of Western Norway, where E-OBS provides higher precipitation sums than HR in both winter and summer, and Spain, where MESAN precipitation is comparable to HR and over parts of the country even higher in summer. The same is true over parts of the Carpathians. Over France, MESAN and HR are in very close agreement, which is likely connected to the good station coverage in MESAN over this region and which supports findings by Isotta et al. (2015) and Prein and Gobiet (2017) for (south-eastern) France. The general picture of highest precipitation sums in HR and drier conditions in E-OBS and MESAN might be a direct consequence of the higher underlying network density in HR and the fact that more high-elevation stations are sampled. Over Sweden, a further reason for lower precipitation sums in E-OBS and MESAN especially in wintertime is presumably the applied undercatch correction in the PTHBV dataset underlying HR in this region (see Section 2.1.2).

4. Model Evaluation

In the following the results of the model evaluation exercise are presented separately for both variables (temperature and precipitation) and for both seasons (winter and summer). The analysis allows for a separate assessment of each performance metric, each observational reference and each sub-region. For the sake of clarity and according to the objectives of this work we do not explicitly identify the five individual RCM experiments (see Table 1 for their identification though).

4.1 Temperature

Figures 4 and 5 present the temperature evaluation results for the winter and the summer season, respectively. In most cases the spatially averaged model biases (BIAS) approximately agree for the three reference datasets. In winter (Fig. 4) a cold model bias prevails and, depending on sub-region and RCM, can amount to more than -2 °C. The range of model biases (given by the vertical extent of the bars) is largest over sub-region CH, where two of the five RCMs are subject to pronounced cold biases. This is likely related to the strong topographic variability of this domain, the pronounced differences in RCM orographies and to the fact that the applied lapse rate correction is based on the simplifying assumption of a global lapse rate being stationary in both time and space. Over France and Norway cold biases are most pronounced when evaluating against E-OBS, which is in line with the higher winter temperatures in E-OBS compared to HR and MESAN over these sub-domains (cf. Fig. 2).

In the summer season (Fig. 5) notable differences of the BIAS metric when comparing against different reference datasets are apparent for sub-regions CA, FR, NO, SP and CH. For the other regions, spatially averaged model biases mostly agree. A similar finding is obtained for MAE99 and MAE01. In the latter case, however, the evaluation against MESAN yields considerably larger summer model biases than for E-OBS and HR in the topographically complex sub-regions NO and CH. Note the extremely large ranges of MAE01 in the winter season with differences between the RCMs of more than 9 °C in sub-region CH, independently of the reference dataset. These large biases are found in two of the five RCMs only. In general, the large model spread over CH indicates difficulties of the RCMs to reliably reproduce minimum temperatures over regions of complex topography.

Concerning the RIAV metric the evaluation results are robust with respect to the choice of reference in both winter and summer with minor exceptions only. Model uncertainty as expressed by the vertical extent of the bars is generally much larger than the influence of the reference dataset. The situation is different though for the PACO metric. Here, the choice of the reference can have an important influence on the evaluation results. Correlation coefficients are high in general (> 0.8 in all cases) owing mainly to the pronounced influence of topography on spatial temperature patterns which is, in principle, represented by both the RCMs and by the observations. Depending on sub-region and season, reference data uncertainty can however strongly dominate. Use of the HR dataset as reference leads to lower correlation coefficients in winter in sub-regions FR and SP. The same is true for FR, PO and SP in summer. These results suggest differences in the spatial pattern of seasonal mean temperatures in the three reference datasets even for regions of pronounced topography and even for the aggregated evaluation scale of 0.22°.

For the distribution-based CM metric, the choice of the reference has an important effect in a few cases only and model uncertainty mostly dominates. The choice of the reference dataset markedly influences CM over CA, FR and SP in winter and CA and CH in summer.

4.2 Precipitation

For precipitation, a pronounced dependency of the BIAS metric on the choice of the reference can be found in both seasons but depending on the sub-region (Figs. 6 and 7). In winter and for sub-region SP, positive model biases with respect to E-OBS can partly translate into negative biases with respect to HR, reflecting the higher precipitation sums in HR compared to E-OBS over most parts of sub-region SP (cf. Figure 3). The same is true for SW and CH in summer. In a few cases the BIAS ranges for the three reference datasets only slightly overlap and reference data uncertainty is of a similar magnitude as model uncertainty (for instance, sub-regions CA, PO and SW in winter). In the last case (SW in winter) a possible reason is the undercatch correction of the Swedish HR dataset that potentially reduces positive model biases compared to the non-corrected E-OBS and MESAN data.

For MAE99, i.e. for the upper tail of the daily precipitation distribution, reference data uncertainty has a larger magnitude than for the BIAS metric (note the different y-axis scales in the upper left and upper middle panels) but is clearly dominated by model uncertainty, especially in summertime. A completely different result is obtained for the spatially averaged absolute bias of the wet day frequency WDFREQ. While model biases with respect to E-OBS and HR approximately agree, the use of the MESAN reanalysis as reference is in most cases associated with larger biases that are partly outside the bias range obtained for E-OBS and HR. The reason is a considerably lower wet day frequency in MESAN compared to E-OBS and HR and a generally positive wet day frequency bias of the RCMs. This bias, and hence WDFREQ, is therefore largest when using MESAN as reference.

In wintertime and over sub-regions PO, SW and CH the MESAN reanalysis is furthermore associated with larger RIAV values (Fig. 6, lower left panel), i.e. a more pronounced overestimation of interannual precipitation variability. All other cases show similar RIAV ranges regardless of the reference employed and model uncertainty clearly dominates. For PACO the results considerably depend on the sub-region. As a general picture, PACO values are systematically lower compared to temperature which reflects the less pronounced control of topography on the spatial pattern of mean seasonal precipitation. The PACO ranges for the three reference datasets are similar in many cases but there are exceptions. The use of E-OBS, for instance, leads to considerably lower values over sub-regions CA, FR and SP in winter while HR is associated with a lower pattern correlation for PO but higher values for sub-region SW. In summer, MESAN is associated with lower correlations over CA, and HR with higher correlations over CA and SW. Overall, however, model uncertainty dominates for the PACO metric.

A different picture is obtained for the distribution-based CM metric (lower right panels). The range of CM values for a given reference dataset is generally high, but especially the use of MESAN as reference can be associated with much lower values compared to E-OBS and HR, i.e. with a lower fraction of grid cells passing the CM test. This feature affects all sub-regions in winter and sub-regions GE, NO, PO, SW and CH in summer. It is obviously associated with the much higher WDFREQ value when using MESAN as a reference, i.e. with the lower wet day frequency in MESAN. Note that CM only considers the wet-day distribution (see Section 2.4) and is not directly affected by wet-day frequency biases. The close relation between both metrics hence indicates that model biases in the wet-day frequency when comparing against MESAN come along with biases in the precipitation distribution for wet days only, i.e. that at least the complete lower tails of the two all-day distributions (model and reference) considerably differ from each other.

5. Observational Versus Model Uncertainty

We here present the results of the uncertainty intercomparison introduced in Section 2.5. This analysis can be seen as a summary of the comparison between observational uncertainty (offset of the three vertical bars for a given performance metric, season and sub-region) and model uncertainty (vertical extent of the bars) provided in Chapter 4 and apparent from Figures 4 to 7.

For temperature (Fig. 8) uncertainty ratios smaller than one are obtained in most cases, i.e. observational uncertainty is typically smaller than model uncertainty. But exceptions to this general pattern are possible, and also the magnitude of the uncertainty ratio primarily depends on the performance metric considered. For the seasonal mean model bias (BIAS) ratios are consistently smaller than 0.5, indicating a model uncertainty being twice as large as observational uncertainty. With reference to the scores describing the tails of the daily values (MAE99 and MAE01) and the frequency distribution (CM), observational uncertainty is also smaller than model uncertainty with the exception of Spain for MAE01 in summer. Ratios for RIAV are below one throughout all sub-regions and both seasons with typically somewhat larger values in winter. In contrast to all other performance metrics, the ratios for the pattern correlation PACO are close to or larger than one in at least half of the cases, i.e. observational uncertainty dominates. This is in particular true for sub-regions FR, GE and SP during summer.

As a general pattern observational uncertainty, i.e. the choice of the reference data, tends to be more important for precipitation (Fig . 9) than for temperature. Uncertainty ratios for WDFREQ and CM are close to or even larger than one in most cases. Maximum values larger than three are obtained for winter in sub-regions CA, PO and SW (WDFREQ) and for sub-region CA (CM). For the cases of WDFREQ and CM these high values are clearly related to low WDFREQ values in the MESAN reference, which constitute an outlier within the observational ensemble. They are probably related to specifics of the MESAN spatial interpolation and not to shortcomings in the underlying station observations. Except for a few cases summer ratios are smaller than their winter counterparts, indicating a smaller contribution of observational uncertainty in summer. This is mainly due to the fact that MESAN deviates stronger from E-OBS and HR in winter than in summer. A clearly dominating observational uncertainty is also found for PACO in sub-region CA (both seasons) as well as in PO and SW (winter only). The same is true for the winter BIAS in sub-regions CA, PO and SW. The latter are, however, outliers since for the BIAS metric ratios close to 0.5 are obtained for most other cases, i.e. model uncertainty clearly dominates. Also for MAE99 and RIAV ratios smaller than one are obtained with the exception of CA (MAE99) and PO and SW (RIAV) in winter

The influence of a potential precipitation undercatch in the observational references on the uncertainty analysis can derived from Figure 10. For most performance metrics the uncertainty ratios are not or only slightly modified compared to the original results (Fig. 9). The most important change is obtained for the MAE99 metric which is especially sensitive as it considers absolute biases at the upper tail of the distribution. Here, uncertainty ratios are increasing in many cases. Roughly, the same stands for further measures based on daily data (WDFREQ and CM). For sub-region SW, specifically, observational uncertainty for MAE99 grows in both winter and summer as only two of the references (E-OBS and MESAN) were corrected for the undercatch compared to the original analysis. This results in larger inter-observational differences, in a larger observational uncertainty and, hence, in a larger uncertainty ratio. For the BIAS metric over SW in winter, undercatch correction brings the three references closer together (not shown), resulting in a decreasing observational uncertainty and a decreasing uncertainty ratio.

6. Model Ranking

To assess the influence of observational uncertainty on model ranking we first show the results for $S_{j,m}$ (Eq. 14; simply denoted as S hereafter) separately for temperature and precipitation when averaged over all seasons and regions (Fig. 11). For illustrational purposes, the actual RCM ranks based on S are also shown. The individual performance metrics show a varying degree of variation in S between the reference datasets (horizontal variation within a given panel). BIAS and MAE99 have a similar normalized error pattern and almost identical ranks for all reference datasets. Model C, for instance has the best performance for both metrics in terms of temperature, independently of the reference dataset. In contrast, model D shows the worst performance for temperature but the best for precipitation. In contrast to these cases of agreement between reference datasets, scores for CM (precipitation) and PACO (temperature) show noticeable differences when employing E-OBS, MESAN or HR as reference. Unsurprisingly, variations are even larger when individual regions are considered (not shown). Concerning the performance of a given model for different performance metrics (vertical variation within a given panel) model C, for instance, has the highest S values (and the best ranking) for most temperature performance metrics, while model D shows the best performance in the case of precipitation. While not the worst performing model in all cases, model E often shows the lowest S and ranks poorly accordingly, regardless of the reference dataset considered. The fact that the dependence of the evaluation results on the reference dataset in turn depends on the metric considered confirms findings from previous works (e.g., Santer et al., 2009, Rupp et al., 2013) and should be kept in mind when interpreting the ranking results.

To illustrate the results for the full ranking scheme, Fig. 12 presents the overall normalized score $\bar{S}_{j,r}$ of Eq. 15 (denoted as $\bar{S}$ hereafter) for each sub-region together with the actual RCM ranks. As an overall picture RCM ranks are similar, independently of the reference dataset employed. However, differences in $\bar{S}$ between the reference datasets can be non-negligible depending on the region and RCM considered. On average, differences in $\bar{S}$ between the reference datasets are largest over sub-regions GE and PO, although individual RCMs also stand out in other regions such as SP (model C) or

FR (model D). On the other hand, Switzerland (CH) shows only small differences in the overall scores between the reference datasets. Furthermore, variations in the actual ranks depending on the reference dataset employed are apparent. These differences tend to be smallest in CH, NO and SW, where the intermodel differences in $\bar{S}$ are relatively large compared to the differences between the reference datasets. In other sub-regions a change of the reference dataset can lead to larger changes in the model ranks (e.g., the rank of model C in SP can change by four levels). This shows that model ranking becomes more dependent on the reference dataset when spatial details are considered. Finally, although the best performing RCM depends on the region and the reference, a noticeable feature is the systematically poor performance of model E in comparison to other models. Model E has the lowest rank in almost all cases regardless of the reference, and values of $\bar{S}$ rarely approach 0.5 for this model.

7. Summary and Conclusions

The objective of the present work was to illustrate the effect of uncertainties in gridded observational reference datasets on RCM evaluation for two variables (temperature and precipitation) on a pan-European scale. For this purpose we made use of three different gridded observational reference datasets (E-OBS v15, national/regional high-resolution grids (HR), EURO4M MESAN) and five reanalysis-driven RCM experiments carried out within the EURO-CORDEX initiative. Our well-defined performance assessment framework considers a range of performance metrics for eight different sub-regions of the European continent and includes an illustrative model ranking scheme. Note that the ensemble of reference grids is an ensemble of opportunity and is likely subject to inter-dependencies arising, for instance, from the use of common station time series in the interpolation or assimilation procedure. In general, an extension of the observational ensemble by, for instance, satellite-based products or by new upcoming datasets could alter the derived observational uncertainties and, hence, the overall evaluation results. The same would be true for an extension of the set of RCMs considered or for a different sampling of available RCM experiments.

A comparison of climatological seasonal mean values as represented by the three reference grids alone yields a general agreement concerning the continental-scale patterns, but also differences on regional scales. These depend on the variable, region and season considered and translate into differences in RCM performance scores. Largest differences in seasonal mean temperature occur over regions of pronounced topography, such as Spain, the European Alps, Scandinavia and the Carpathians. Except for the latter case, the high-resolution HR dataset typically shows lowest temperatures which might be related to a better sampling of high-elevation stations by HR. For the case of precipitation both MESAN and E-OBS typically underestimate mean seasonal precipitation as provided by HR.

For most performance metrics and especially for temperature, the influence of the choice of the observational reference on model evaluation is rather weak and is smaller than model uncertainty. This is especially true for winter temperature, where only the pattern correlation (PACO) and to some extent the distribution-based Cramér-von Mises score (CM) show notable dependencies on the reference dataset employed. However, winter PACO values are still larger than 0.8 for each individual sub-region and for any combination of RCM and reference dataset. Hence, spatial temperature patterns are, in a general sense, well represented by the RCMs independently of the specific reference employed. The same is true for the summer season which, however, is subject to slightly larger reference data uncertainty for PACO and for the mean absolute error of the 1st daily percentile (MAE01). For precipitation the influence of observational uncertainty is larger than for temperature. It often dominates model uncertainty especially for the absolute bias in the wet-day frequency (WDFREQ) and for the Cramér-von Mises score (CM) in winter. But even the spatially averaged measures of seasonal mean bias (BIAS) and ratio of interannual variability (RIAV) can be considerably affected by the choice of the reference observational product. The fact that most observational references are not corrected for rain gauge undercatch has some influence on the final uncertainty analysis but does not change the general picture. Note that observational uncertainty being smaller than model uncertainty does not necessarily imply that uncertainties in observations are negligible and without influence. They can still be relevant, for instance, in model development or model bias correction.

When employing a simple and illustrative model ranking scheme on these results it is found that RCM ranking can depend on the reference dataset employed, and more often for precipitation than for temperature. In individual cases, final model ranks can differ by up to four (out of five models) depending on the choice of the reference dataset. These findings are in line with previous works (e.g., Gómez-Navarro et al., 2012; Prein and Gobiet, 2016) which suggests that uncertainties related to the reference data should ideally be taken into account when assessing climate model performance in the present-day climate. However, if a focus is laid on temperature only the three reference datasets agree to a large extent, indicating the suitability of each individual product for climate model evaluation purposes. Furthermore, spatio-temporally averaged temperature and precipitation climates are very similar among the three references (see the BIAS metric), and model uncertainty clearly dominates in these case. Also note that all datasets employed in the present work were aggregated to the comparatively low E-OBS grid spacing of 0.22° prior to the analysis, including the high-resolution HR data. This spatial aggregation might to some extent mask the added value of HR but is required in the context of the present work. The full benefits of the higher-resolved HR data and their underlying dense station network will however only become apparent when evaluating, for instance, very high resolution RCM experiments at the convection-resolving scale (e.g. Ban et al., 2014).

Considering the ranking exercise itself, one should keep in mind that the ranking scheme applied here is likely to suffer from commonly known limitations (Overland et al., 2011; Santer et al., 2009, Rupp et al., 2013) and that the results are specific for the selected RCMs and performance metrics. On the other hand, it has been previously shown that only small uncertainties in the ranking and weighting of models can result in strong differences and potentially misleading signals (Weigel et al., 2010).

Acknowledgments

The present work has been carried out as part of the EU-COST Action VALUE (Validating and Integrating Downscaling Methods for Climate Change Research; ES1102). We gratefully acknowledge the providers of RCM and observational data. For the high-resolution national/regional grids these are the University of Cantabria (SP), the Institute of Meteorology and Water Management - National Research Institute (PO), Météo-France/CERFACS (FR), The Swedish Meteorological and Hydrological Institute (SE), Deutscher Wetterdienst (GE), the Hungarian Meteorological Service (CA), the Norwegian Meteorological Institute (NO) and Federal Office of Meteorology and Climatology MeteoSwiss (CH). Furthermore, we acknowledge the E-OBS dataset from the EU-FP6 project ENSEMBLES (http://ensembles-eu.metoffice.com) and the data providers in the ECA&D project (http://eca.knmi.nl). The MESAN dataset was provided by the Swedish Meteorological and Hydrological Institute. All analysis were performed on the computing infrastructure of the Swiss National Supercomputing Centre CSCS. We furthermore thank the climate modelling groups of the EURO-CORDEX initiative for producing and making available their model output. The contribution of Olle Räty was partly funded by the Vilho, Yrjö and Kalle Väisälä Foundation of the Finnish Academy of Science and Letters.

References

Adam JC, Lettenmaier DP. 2003. Adjustment of global gridded precipitation for systematic bias. Journal of Geophysical Research 108: 4257 (D9). doi: 10.1029/2002JD002499.

Addor N, Fischer EM. 2015. The influence of natural variability and interpolation errors on bias characterization in RCM simulations. Journal of Geophysical research – Atmospheres, 120. doi: 10.1002/2014JD022824.

Anderson TW. 1962. On the Distribution of the Two-Sample Cramer-von Mises Criterion. The Annals of Mathematical Statistics 33(3): 1148-1159. doi:10.1214/aoms/1177704477.

Awange JL, Ferreira VG, Forootan E, Khandu, Andam-Akorful SA, Agutu NO, He XF. 2016. Uncertainties in remotely sensed precipitation data over Africa. International Journal of Climatology 36: 303-323. doi: 10.1002/joc.4346.

Ban N, Schmidli J, Schär C. 2014. Evaluation of the convection-resolving regional climate modeling approach in decade-long simulations. Journal of Geophysical Research 199 (13): 7889-7907. doi: 10.1002/2014JD021478.

Begert M, Schlegel T, Kirchhofer W. 2005. Homogeneous temperature and precipitation series of Switzerland from 1864 to 2000. International Journal of Climatology 25: 65-80. doi: 10.1002/joc.1118.

Beguería S, Vicente-Serrano SM, Tomás-Burguera M, Maneta M. 2016. Bias in the variance of gridded data sets leads to misleading conclusions about changes in climate variability. International Journal of Climatology 36(9): 3413-3422. doi: 10.1002/joc.4561.

Bellprat O, Kotlarski S, Lüthi D, Schär C. 2012. Exploring Perturbed Physics Ensembles in a Regional Climate Model. Journal of Climate 25: 4582-4599. doi: 10.1175/JCLI-D-11-00275.1.

Berg P, Norin L, Olsson J. 2016. Creation of a high resolution precipitation data set by merging gridded gauge data and radar observations for Sweden. Journal of Hydrology 541: 6-13. doi: 10.1016/j.jhydrol.2015.11.031.

Brienen S, Früh B, Walter A, Trusilova K, Becker P. 2016. A Central European precipitation climatology – Part II: Application of the high-resolution HYRAS data for COSMO-CLM evaluation. Meteorologische Zeitschrift 25(2): 195-214. doi: 10.1127/metz/2016/0617.

Bucchignani E, Montesarchio M, Zollo AL, Mercogliano P. 2016. High-resolution climate simulations with COSMO-CLM over Italy: performance evaluation and climate projections for the 21st century. International Journal of Climatology 36: 735-756. doi: 10.1002/joc.4379.

Casanueva A, Herrera S, Fernández J, Frías MD, Gutiérrez JM. 2013. Evaluation and projection of daily temperature percentiles from statistical and dynamical downscaling methods. Natural Hazards and Earth System Sciences 13: 2089-2099. doi: 10.5194/nhess-13-2089-2013.

Cheneka BR, Brienen S, Fröhlich K, Asharaf S, Früh B. 2016. Searching for an added value in downscaled seasonal hindcasts over East Africa: COSMO-CLM forced by MPI-ESM. Advances in Meteorology 2016: 1-17. doi: 10.1155/2016/4348285.

Cheval S, Baciu M, Dumitrescu A, Breza T, Legatesb DR, and Chende V. 2010. Climatologic adjustments to monthly precipitation in Romania. International Journal of Climatology 31: 704-714. doi:10.1002/joc.2099.

Christensen JH, Kjellström E, Giorgi F, Lenderink G, Rummukainen M. 2010. Weight assignment in regional climate models. Climate Research 44: 179-194. doi: 10.3354/cr00916.

Dahlgren P, Gustafsson N. 2012. Assimilating host model information into a limited area model. Tellus 64A: 15836. doi: 10.3402/tellusa.v64i0.15836.

Davin EL, Maisonnave E, Seneviratne SI. 2016. Is land surface processes representation a possible weak link in current Regional Climate Models? Environmental Research Letters 11: 074027. doi: 10.1088/1748-9326/11/7/074027.

Dee DP, Uppala SM, Simmons AJ, Berrisford P, Poli P, Kobayashi S, Andrae U, Balmaseda MA, Balsamo G, Bauer P, Bechtold P, Beljaars ACM, van de Berg L, Bidlot J, Bormann N, Delsol C, Dragani R, Fuentes M, Geer AJ, Haimberger L, Healy SB, Hersbach H, Hólm EV, Isaksen L, Kallberg P, Köhler M, Matricardi M, McNally AP, Monge-Sanz BM, Morcrette J-J, Park B-K, Peubey C, de Rosnay P, Tavolato C, Thépaut J-N, Vitart F. 2011. The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Quarterly Journal of the Royal Meteorological Society 137: 553–597. doi: 10.1002/qj.828.

Di Luca A, de Elía R, Laprise R. 2012. Potential for added value in precipitation simulated by high-resolution nested Regional Climate Models and observations. Climate Dynamics 38: 1229-1247. doi: 10.1007/s00382-011-1068-3.

Dunn RJH, Donat MG, Alexander LV. 2014. Investigating uncertainties in global gridded data sets of climate extremes. Climate of the Past 10: 2171-2199. doi: 10.5194/cp-10-2171-2014.

Durand Y, Brun E, Mérindol L, Guyomarc'h G, Lesaffre B, Martin E. 1993. A meteorological estimation of relevant parameters for snow models. Annals of Glaciology 18: 65–71.

Frei C, Christensen JH, Déqué M, Jacob D, Jones RG, Vidale PL. 2003. Daily precipitation statistics in regional climate models: Evaluation and intercomparison for the European Alps. Journal of Geophysical Research 108(D3): 4124. doi: 10.1029/2002JD002287.

Frei C. 2014. Interpolation of temperature in a mountainous region using nonlinear profiles and non-Euclidean distances. International Journal of Climatology 34: 1585-1605. doi: 10.1002/joc.3786.

Frick C, Steiner H, Mazurkiewicz A, Riediger U, Rauthe M, Reich T, Gratzki A. 2014. Central European high-resolution gridded daily data sets (HYRAS): Mean temperature and relative humidity. Meteorologische Zeitschrift 23(1): 15-32. doi: 10.1127/0941-2948/2014/0560.

Fowler HJ, Blekinsop S, Tebaldi C. 2007. Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. International Journal of Climatology 27: 1547-1578. doi: 10.1002/joc.1556.

Gandin LS. 1965. Objective analysis of meteorological fields (Ob″ektivnyi analiz meteorologicheskikh polei), Translated from Russian by the Israel Program for Scientific Translations, Jerusalem. Quarterly Journal of the Royal Meteorological Society. doi: 10.1002/qj.49709239320.

Gbambie ASB, Poulin A, Boucher MA, Arsenault R. 2017. Added value of alternative information in interpolated precipitation datasets for hydrology. Journal of Hydrometeorology 18: 247-264. doi: 10.1175/JHM-D-16-0032.1.

Gervais M, Tremblay LB, Gyakum JR, Atallah E. 2014. Representing extremes in a daily gridded precipitation analysis over the United States: Impacts of station density, resolution, and gridding methods. Journal of Climate 27: 5201-5218. doi: 10.1175/JCLI-D-13-00319.1.

Gómez-Navarro JJ, Montávez JP, Jerez S, Jiménez-Guerrero P, Zorita E. 2012. What is the role of the observational dataset in the evaluation and scoring of climate models? Geophysical Research Letters 39: L24701. doi: 10.1029/2012GL054206.

Groisman PY, Legates DR. 1994. The accuracy of United States precipitation data, Bulletin of the American Meteorological Society 75: 215– 227.

Häggmark L, Ivarsson K-I, Gollvik S, Olofsson P-O. 2000. MESAN, an operational mesoscale analysis system. Tellus 52A: 2-20.

Hanel M, Buishand A. 2011. Analysis of precipitation extremes in an ensemble of transient regional climate model simulations for the Rhine basin. Climate Dynamics 36: 1135–1153.

Hartmann DL, Klein Tank AMG, Rusticucci M, Alexander LV, Brönnimann S, Charabi Y, Dentener FJ, Dlugokencky EJ, Easterling DR, Kaplan A, Soden BJ, Thorne PW, Wild M, Zhai PM. 2013. Observations: Atmosphere and Surface. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker TF, Qin D, Plattner G-K, Tignor M, Allen SK, Boschung J, Nauels A, Xia Y, Bex V, Midgley PM (eds.)]. Cambridge University Press: Cambridge, UK and New York, NY, USA.

Haslinger K, Anders I, Hofstätter M. 2013. Regional climate modelling over complex terrain: an evaluation study of COSMO-CLM hindcast model runs for the Greater Alpine Region. Climate Dynamics 40: 511-529. doi: 10.1007/s00382-012-1452-7.

Haylock MR, Hofstra N, Klein Tank AMG, Klok EJ, Jones PD, New M. 2008. A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. Journal of Geophysical Research 113: D20119. doi:10.1029/2008JD010201.

Hegerl GC, Jones PD, Barnett PD. 2001. Effect of observational sampling error on the detection and attribution of anthropogenic climate change. Journal of Climate, 14: 198–207.

Herold N, Alexander LV, Donat MG, Contractor S, Becker A. 2016. How much does it rain over land? Geophysical Research Letters 43: 341-348. doi: 10.1002/2015GL066615.

Herrera S, Gutiérrez JM, Ancell R, Pons MR, Frías MD, Fernández J. 2012. Development and analysis of a 50-year high-resolution daily gridded precipitation dataset over Spain (Spain02). International Journal of Climatology 32: 74–85. doi:10.1002/joc.2256.

Herrera S, Fernández J, Gutiérrez JM. 2016. Update of the Spain02 gridded observational dataset for EURO-CORDEX evaluation: assessing the effect of the interpolation methodology. International Journal of Climatology 36: 900-908. doi:10.1002/joc.4391.

Herrera S, Jaczewski A, Kotlarski S, Gutiérrez JM, Soares PMM. 2017. Sensitivity analysis of observational gridded datasets to the density of stations and the interpolation methodology. In preparation.

Hofstra N, Haylock M, New M, Jones PD. 2009. Testing E-OBS European high-resolution gridded data set of daily precipitation and surface temperature. Journal of Geophysical Research 114: D21101. doi: 10.1029/2009JD011799.

Hofstra N, New M, McSweeney C. 2010. The influence of interpolation and station network density on the distribution and extreme trends of climate variables in gridded data. Climate Dynamics 35: 841–858.

Isotta FA, Vogel R, Frei C. 2015. Evaluation of European regional reanalyses and downscalings for precipitation in the Alpine region. Meteorologische Zeitschrift 24: 15-37. doi: 10.1127/metz/2014/0584.

Ivanov M, Warrach-Sagi K, Wulfmeyer V. 2017a. Field significance of performance measures in the context of regional climate model evaluation. Part 1: temperature. Theoretical and Applied Climatology, in press. doi: 10.1007/s00704-017-2100-2.

Ivanov M, Warrach-Sagi K, Wulfmeyer V. 2017b. Field significance of performance measures in the context of regional climate model evaluation. Part 2: precipitation. Theoretical and Applied Climatology, in press. doi: 0.1007/s00704-017-2077-x.

Jacob D, Petersen J, Eggert B, Alias A, Christensen OB, Bouwer LM, Braun A, Colette A, Déqué M, Georgievski G, Georgopoulou E, Gobiet A, Menut L, Nikulin G, Haensler A, Hempelmann N, Jones C, Keuler K, Kovats S, Kröner N, Kotlarski S, Kriegsmann A, Martin E, van Meijgaard E, Moseley C, Pfeifer S, Preuschmann S, Radermacher C, Radtke K, Rechid D, Rounsevell M, Samuelsson P, Somot S, Soussana J-F, Teichmann C, Valentini R, Vautard R, Weber B, Yiou P. 2014. EURO-CORDEX: new high-resolution climate change projections for European impact research. Regional Environmental Change 14: 563-578. doi: 10.1007/s10113-013-0499-2.

Johansson B, Chen D. 2003. The influence of wind and topography on precipitation distribution in Sweden: statistical analysis and modelling. International Journal of Climatology 23(12): 1523–1535.

Kotlarski S, Block A, Böhm U, Jacob D, Keuler K, Knoche R, Rechid D, Walter A. 2005. Regional climate model simulations as input for hydrological applications: evaluation of uncertainties. Advances in Geosciences 5: 119-125.

Kotlarski S, Hagemann S, Krahe P, Podzun R, Jacob D. 2012. The Elbe river flooding 2002 as seen by an extended regional climate model. Journal of Hydrology 472-473: 169-183. doi: 10.1016/j.jhydrol.2012.09.020.

Kotlarski S, Keuler K, Christensen OB, Colette A, Déqué M, Gobiet A, Goergen K, Jacob D, Lüthi D, van Meijgaard E, Nikulin G, Schär C, Teichmann C, Vautard R, Warrach-Sagi K, Wulfmeyer V. 2014. Regional climate modeling on European scales: a joint standard evaluation of the EURO-CORDEX RCM ensemble. Geoscientific Model Development 7: 1297-1333. doi: 10.5194/gmd-7-1297-2014.

Kyselý J, Plavcová E. 2010. A critical remark on the applicability of E-OBS European gridded temperature data set for validating control climate simulations. Journal of Geophysical Research 115: D23118. doi: 10.1029/2010JD014123.

Lakatos M, Szentimrey T, Bihari Z, Szalai S. 2013. Creation of a homogenized climate database for the Carpathian region by applying the MASH procedure and the preliminary analysis of the data. Idojaras 117: 143–158.

Lenderink G. 2010. Exploring metrics of extreme daily precipitation in a large ensemble of regional climate model simulations. Climate Research 44: 151-166. doi: 10.3354/cr00946.

Lunneborg CE. 2005. Cramer-Von Mises Test. Wiley StatsRef: Statistics Reference Online. doi: 10.1002/9781118445112.stat06554.

Lussana C, Tveito OE, Uboldi F. 2016. seNorge v2.0, Temperature: An observational gridded dataset of temperature for Norway. MET Report No. 14/2016. Available from https://www.met.no/sokeresultat/_/attachment/inline/243074f4-09bf-4f63-b98a-f329b3661ce4:c586d2b116d185dc2ac000a1eca6cd98f2f5bdbd/MET-report-14-2016.pdf (last access: 18 July 2017).

Mahlstein I, Spirig C, Liniger MA, Appenzeller C. 2015. Estimating daily climatologies for climate indices derived from climate model data and observations. Journal of Geophysical Research – Atmospheres 120: 2808-2818. doi: 10.1002/2014jD022327.

Maraun D, Osborn TJ, Rust HW. 2012. The influence of synoptic airflow on UK daily precipitation extremes. Part II: regional climate model and E-OBS data validation. Climate Dynamics 39: 287-301. doi: 10.1007/s00382-011-1176-0.

Maraun D, Wigmann M, Gutiérrez JM, Kotlarski S, Chandler RE, Hertig E, Wibig J, Huth R, Wilcke RAI. 2015. VALUE: A framework to validate downscaling approaches for climate change studies. Earth's Future 3. doi: 10.1002/2014EF000259.

McMillan H, Krueger T, Freer J. 2012. Benchmarking observational uncertainties for hydrology: rainfall, river discharge and water quality. Hydrological Processes 26: 4078-4111. doi: 10.1002/hyp.9384.

MeteoSwiss. 2013a: Daily Mean, Minimum and Maximum Temperature: TabsD, TminD ,TmaxD. Available from www.meteoswiss.admin.ch/content/dam/meteoswiss/de/service-und-publikationen/produkt/raeumliche-daten-temperatur/doc/ProdDoc_TabsD.pdf (last access: 18 July 2017).

MeteoSwiss. 2013b: Daily Precipitation (final analysis): RhiresD. Available from www.meteoswiss.admin.ch/content/dam/meteoswiss/de/service-und-publikationen/produkt/raeumliche-daten-niederschlag/doc/ProdDoc_RhiresD.pdf (last access: 18 July 2017).

Osborn T, Hulme M. 1997. Development of a relationship between station and grid-box rainday frequencies for climate model evaluation. Journal of Climate 10: 1885-1908. doi: 10.1175/1520-0442.

Overland JE, Wang M, Bond NA, Walsh JE, Kattsov VM, Chapman WL. 2011. Considerations in the selection of global climate models for regional climate projections: the Arctic as a case study. Journal of Climate 24: 1583–1597. doi: 10.1175/2010JCLI3462.1.

Palazzi E, von Hardenberg J, Provenzale A. 2013. Precipitation in the Hindu-Kush Karakoram Himalaya: observations and future scenarios. Journal of Geophysical Research: Atmospheres 188: 85-100. doi: 10.1029/2012JD018697.

Prein AF, Gobiet A. 2017. Impacts of uncertainties in European gridded precipitation observations on regional climate analysis. International Journal of Climatology 37: 305-327. doi: 10.1002/joc.4706.

Quintana-Seguí P, Le Moigne P, Durand Y, Martin E, Habets F, Baillon M. 2008. Analysis of near-surface atmospheric variables: validation of the SAFRAN analysis over France. Journal of Applied Meteorology and Climatology 47: 92-107. doi: 10.1175/2007JAMC1636.1.

Rauthe M, Steiner H, Riediger U, Mazurkiewicz A, Gratzki, A. 2013. A Central European precipitation climatology? Part I: Generation and validation of a high-resolution gridded daily data set (HYRAS). Meteorologische Zeitschrift 22 (3): 235-256. doi: 10.1127/0941-2948/2013/0436.

Richter D. 1995. Ergebnisse methodischer Untersuchungen zur Korrektur des systematischen Messfehlers des Hellmann-Niederschlagmessers. Berichte des Deutschen Wetterdienstes 194. Selbstverlag des Deutschen Wetterdienstes, Offenbach am Main. Available from http://www.dwd.de/DE/leistungen/pbfb_verlag_berichte/pdf_einzelbaende/194_pdf.pdf (last access: 18 July 2017).

Ring C, Mannig B, Pollinger F, Paeth H. 2016. Uncertainties in the simulation of precipitation in selected regions of humid and dry climate. International Journal of Climatology 36: 3521-3538. doi: 10.1002/joc.4573.

Rupp DE*,* Abatzoglou JT*,* Hegewisch KC*,* Mote *PW*. 2013*.* Evaluation of CMIP5 20th century climate simulations for the Pacific Northwest USA*.* Journal of Geophysical Research: Atmospheres 118*:* 10.884-10.906, doi:10.1002/jgrd.50843.

Räisänen J. 2007. How reliable are climate models? Tellus A 59: 2-29. doi: 10.1111/j.1600-0870.2006.00211.x.

Santer BD, Taylor KE, Gleckler PJ, Bonfils C, Barnett TP, Pierce DW, Wigley TML, Mears C, Wentz FJ, Brüggemann W, Gillett NP, Klein SA, Solomon S, Stott PA, Wehner MF. 2009. Incorporating model quality information in climate change detection and attribution studies. Proceedings of the National Academy of Sciences 106 (35): 14778-14783. doi: 10.1073/pnas.0901736106.

Schneider U, Becker A, Finger P, Meyer-Christoffer A, Ziese M, Rudolf B. 2014. GPCC's new land surface precipitation climatology based on quality-controlled in situ data and its role in quantifying the global water cycle. Theoretical and Applied Climatology 115: 15-40. doi: 10.1007/s00704-013-0860-x.

Sevruk B. 1985. Systematischer Niederschlagsmessfehler in der Schweiz. In: Der Niederschlag in der Schweiz, Beiträge zur Geologie der Schweiz - Hydrologie, Vol. 31, pp. 65– 75, Bundesamt für Wasser und Geologie, Bern, Switzerland.

Sevruk B. 1986. Correction of precipitation measurements summary report. In: Correction of precipitation measurements. Züricher Geographische Schriften 23: 13–23.

Sunyer MA, Sorup HJD, Christensen OB, Madsen H, Rosbjerg D, Mikkelsen PS, Arnbjerg-Nielsen. 2013. On the importance of observational data properties when assessing regional climate model performance of extreme precipitation. Hydrology and Earth System Sciences 17: 4323-4337. doi: 10.5194/hess-17-4323-2013.

Szentimrey T. 2004. Multiple Analysis of Series for Homogenization (MASH); Verification procedure for homogenized time series". Proceedings of the Fourth Seminar for Homogenization and Quality Control in Climatological Databases, Budapest, Hungary. WMO, WCDMP-No. 56, pp. 193-201.

Szentimrey T, Bihari Z. 2007. Mathematical background of the spatial interpolation methods and the software MISH (Meteorological Interpolation based on Surface Homogenized Data Basis). Proceedings from the Conference on Spatial Interpolation in Climatology and Meteorology, Budapest, Hungary, 2004, COST Action 719, COST Office, 17–27. Lubin D, Massom R. 2006. Polar Remote Sensing. Volume I: Atmosphere and Oceans. Praxis Publishing Ltd: Chichester, UK.

Szentimrey T. 2013. Multiple Analysis of Series for Homogenization (MASH v3.03). CARPATCLIM Deliverable D2.10. Annex 3 – Description of MASH and MISH algorithms. Hungarian Meteorological Service. Available from http://www.carpatclim-eu.org/docs/mashmish/mashmish.pdf (last access: 03 April 2017).

Tanarhte M, Hadjinicolaou P, Lelieveld J. 2012. Intercomparison of temperature and precipitation data sets based on observations in the Mediterranean and the Middle East. Journal of Geophysical Research 117: D12102. doi: 10.1029/2011JD017293.

Uboldi F, Lussana C, Salvati M. 2008. Three-dimensional spatial interpolation of surface meteorological observations from high-resolution local networks. Meteorological Applications 15: 331-345. doi:10.1002/met.76.

Vidal JP, Martin E, Franchistéguy L, Baillon M, Soubeyroux JM. 2010. A 50-year high-resolution atmospheric reanalysis over France with the Safran system. International Journal of Climatology 30 (11): 1627-1644.

Wagner S, Kunstmann H, Bardossy A. 2007. Uncertainties in water balance estimations due to scarce meteorological information: a case study for the White Volta catchment in West Africa. In: Quantification and Reduction of Predictive Uncertainty for Sustainable Water Resources Management (Proceedings of Symposium HS2004 at IUGG2007, Perugia, July 2007). IAHS Publ. 313.

Weigel AP, Knutti R, Liniger MA, Appenzeller C. 2010. Risks of model weighting in multimodel climate projections. Journal of Climate 23(15): 4175-4191. doi: 10.1175/2010JCLI3594.1.

Wolff MA, Isaksen K, Petersen-Overleir A, Odemark K, Reitan T, Braekkan R. 2015. Derivation of a new continuous adjustment function for correcting wind-induced loss of solid precipitation: results of a Norwegian field study. Hydrology and Earth System Sciences 19: 951-967. doi: 10.5194/hess-19-951-2015.

Figures

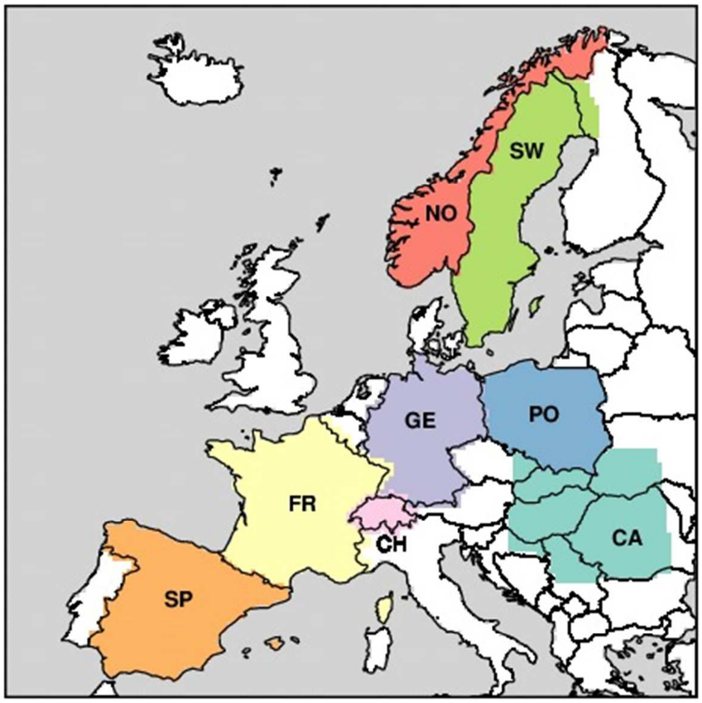


Fig. 1: The eight sub-regions considered for RCM evaluation. SP: Spain, FR: France, CH: Switzerland, GE: Germany, NO: Norway, SW: Sweden, PO: Poland, CA: Carpathians.

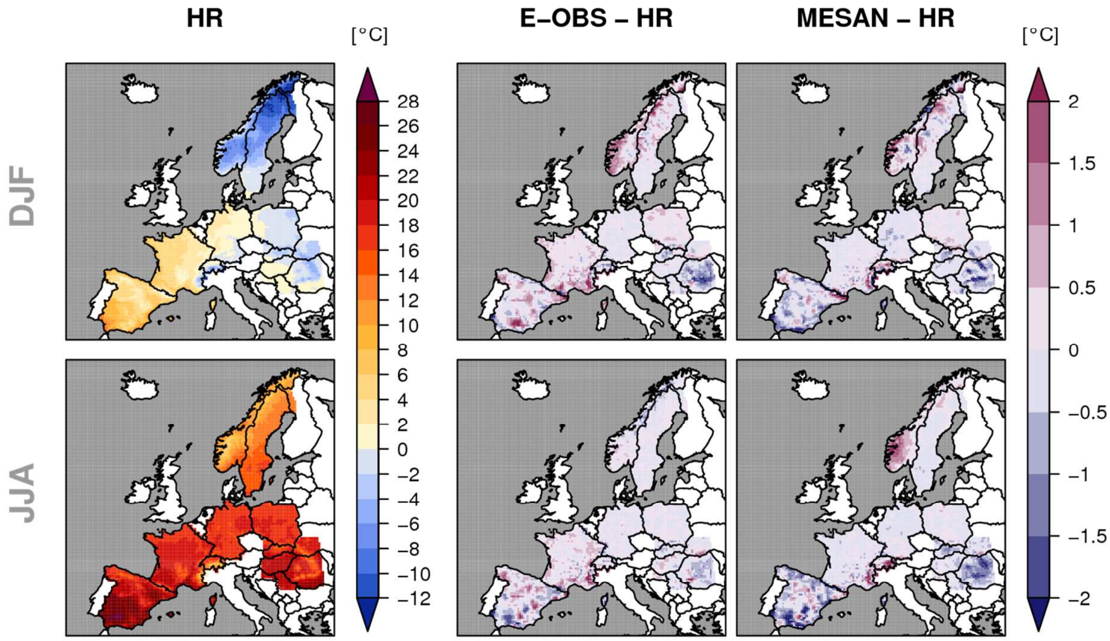
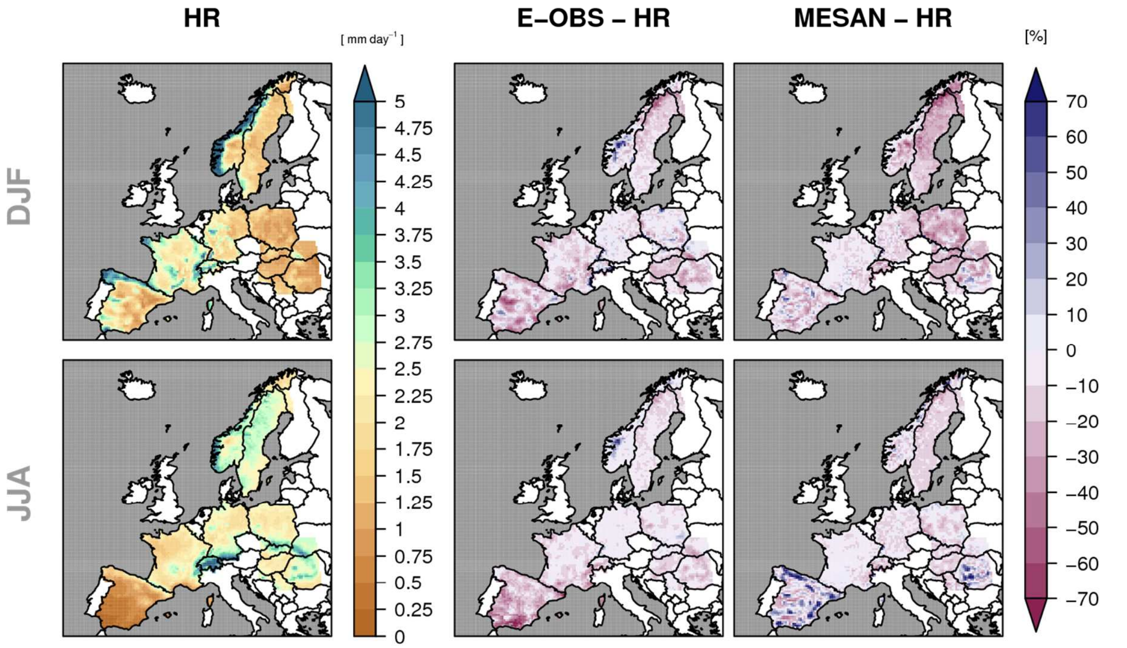


Fig. 2: Spatial pattern of seasonal mean temperature [^{o}C] in HR in the period 1989-2006 (left column) and difference between E-OBS and HR (middle column) and MESAN and HR (right column). Upper row: Winter (DJF), lower row: Summer (JJA).



942
943
944 **Fig. 3**: As Figure 2 but for mean seasonal precipitation [mm day^{-1}]. Differences between E-OBS and HR and
945 between MESAN and HR are given in [%].

946

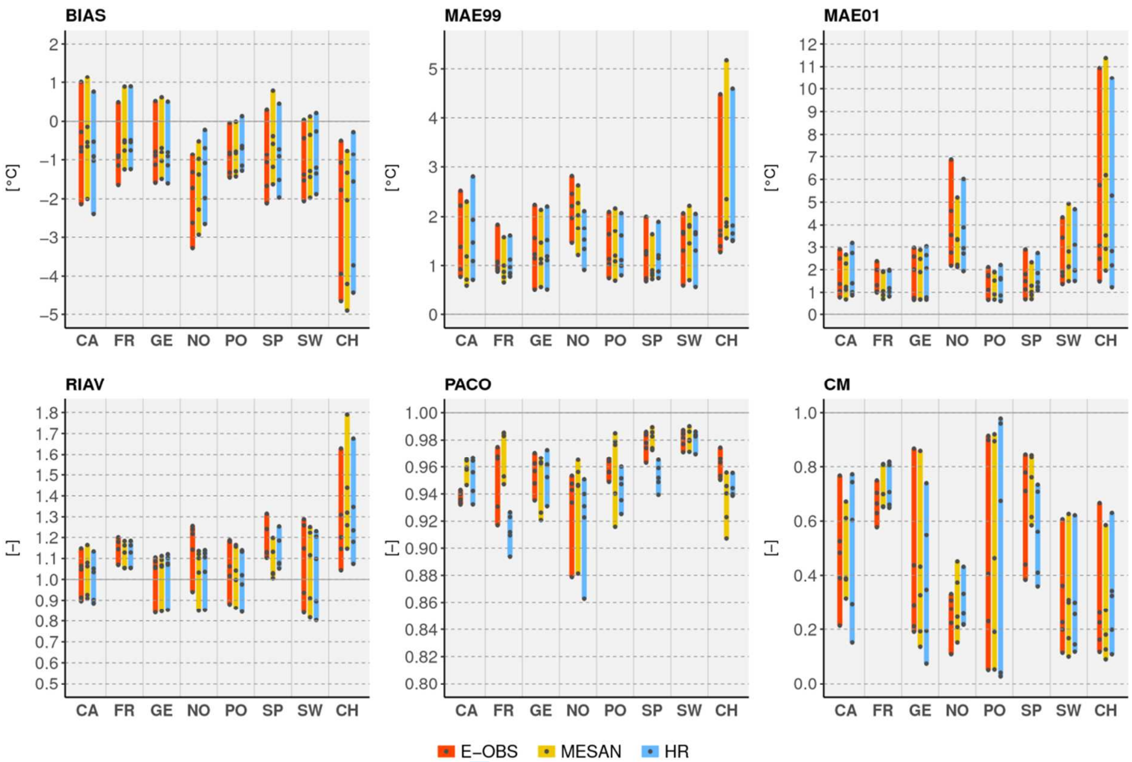


Fig. 4: Evaluation results for winter (DJF) temperature. The six panels correspond to the six performance metrics considered, the colours refer to the three observational references. Each set of three bars corresponds to one sub-region (x-axis). The five dots within each bar refer to the evaluation results for the five individual RCMs, whereas the bars themselves depict the model spread in terms of the minimum-maximum range.

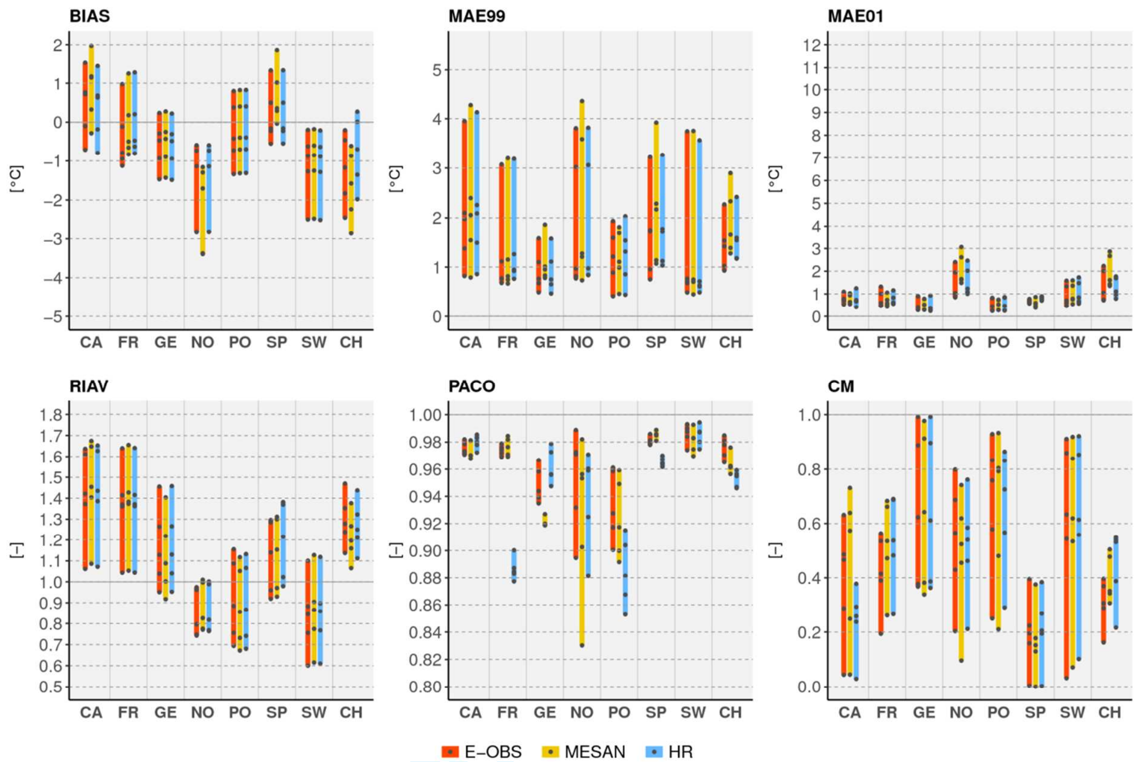


Fig. 5: As Figure 4 but for summer (JJA) temperature.

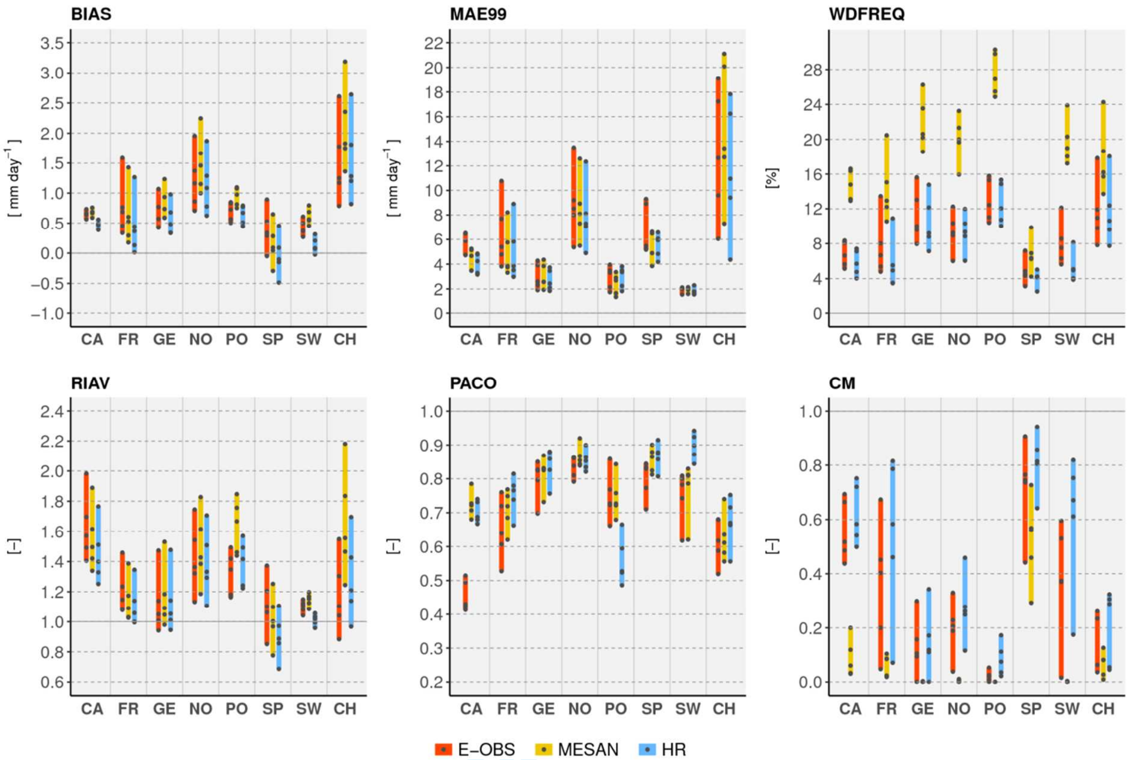
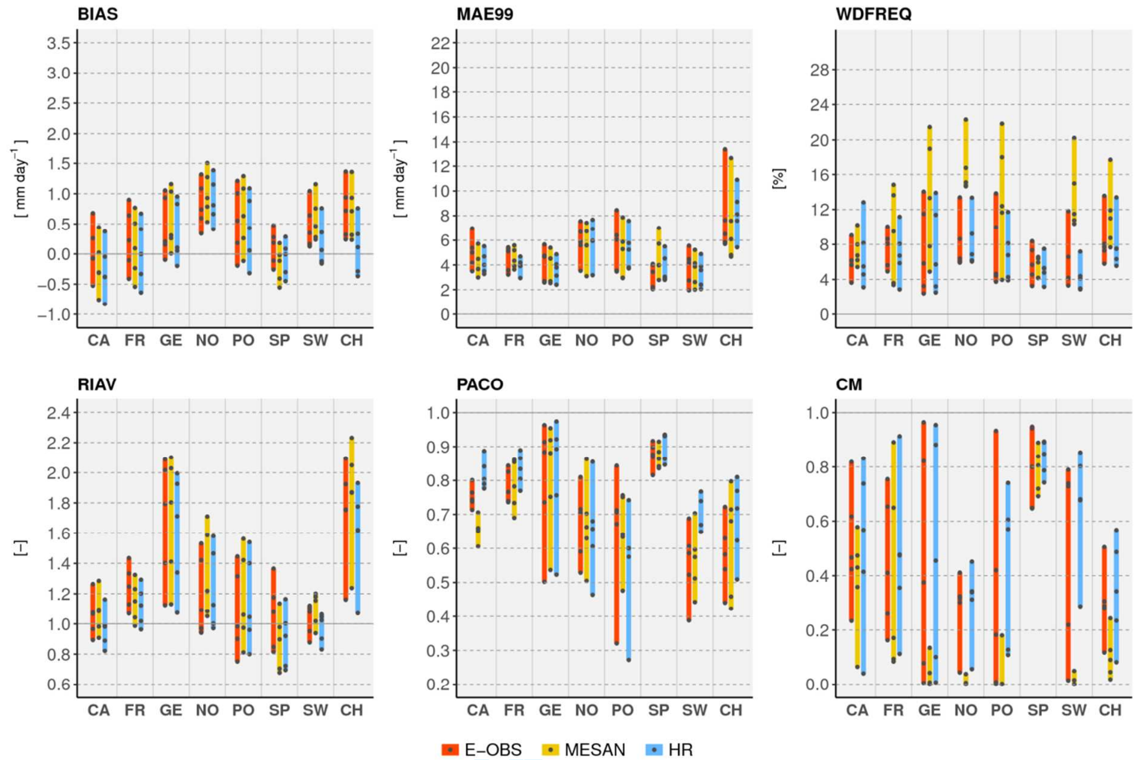


Fig. 6: As Figure 4 but for winter (DJF) precipitation.

962



963

964 **Fig. 7**: As Figure 4 but for summer (JJA) precipitation.

965

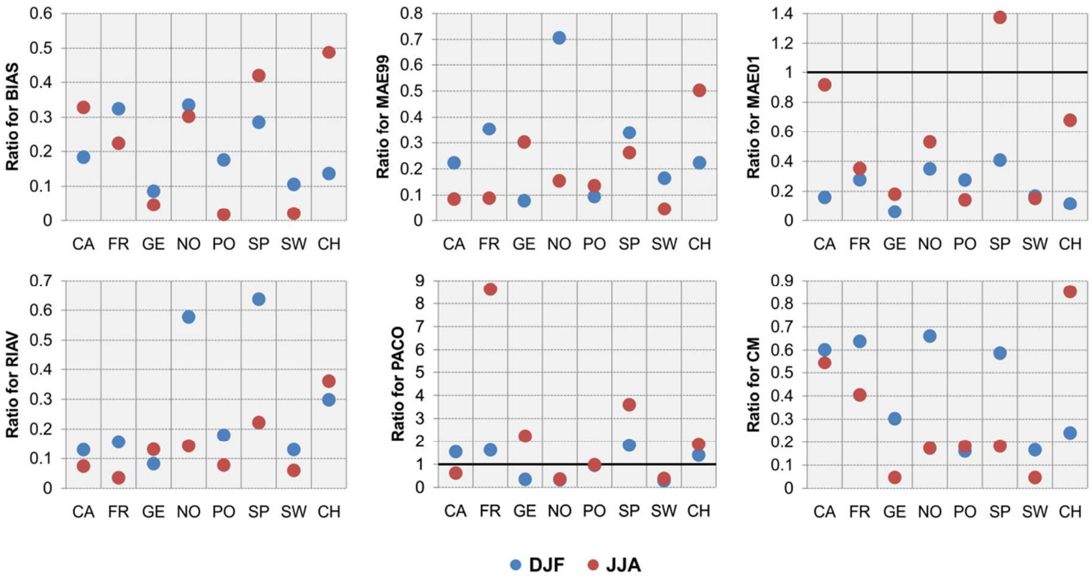


Fig. 8: Uncertainty intercomparison for temperature. The six panels refer to the six performance metrics considered, the two colours to the seasons. An uncertainty ratio R larger (smaller) than one (thick horizontal line) corresponds to a dominating observational (model) uncertainty for the respective case.

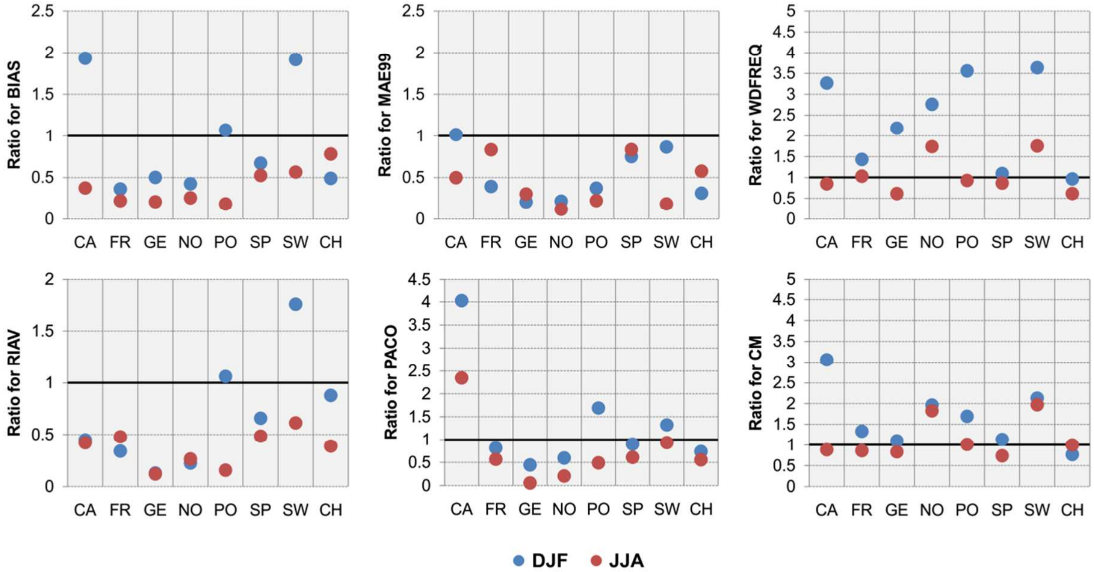


Fig. 9: As Figure 8 but for precipitation.

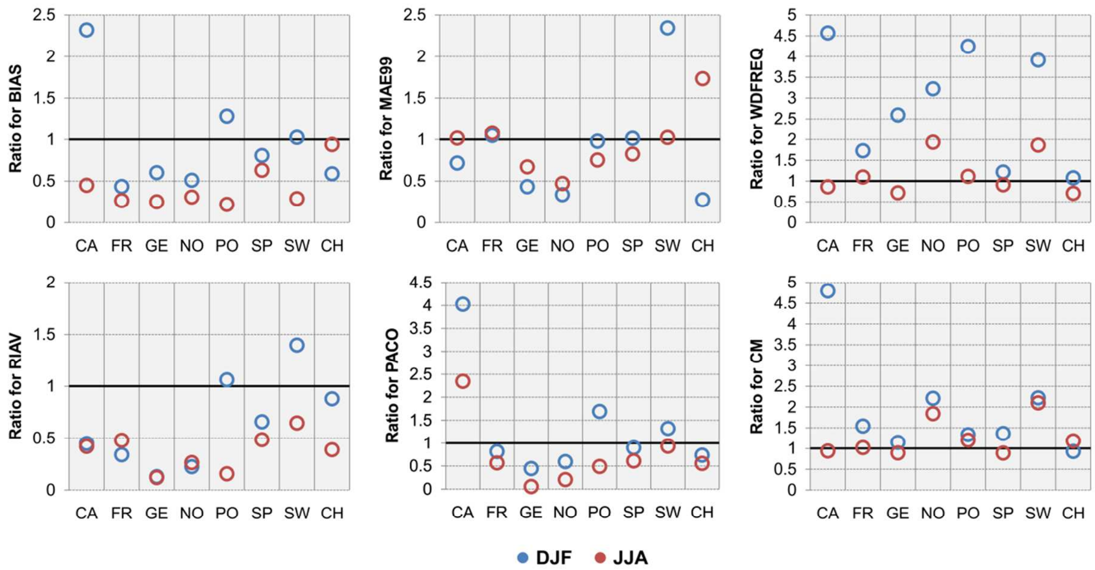


Fig. 10: As Figure 9 but for corrected precipitation: 20% were added to all daily precipitation amounts in all three observational references except for HR over sub-region SW. Open circles instead of filled ones are used for better separation from Fig. 9.

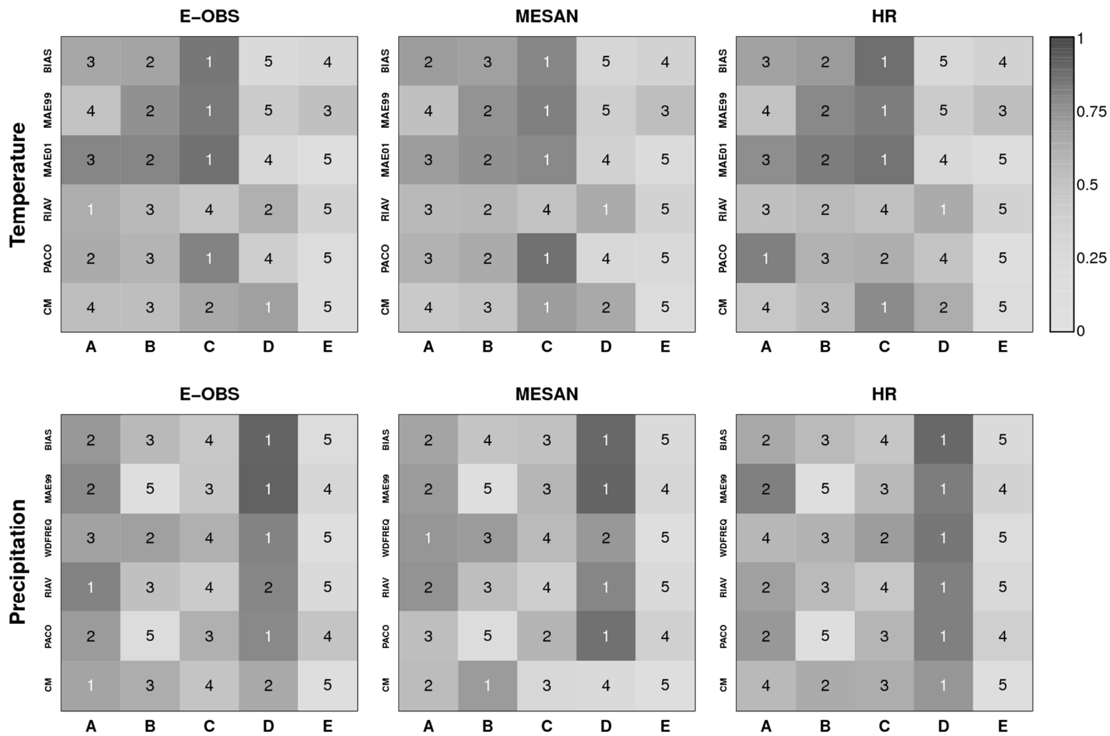


Fig. 11: Normalized performance scores (shading) for individual performance metrics, when averaged over all seasons and regions. The upper row shows the results for temperature and the lower row for precipitation. Numbering inside the shaded boxes indicates the actual RCM rank for each case. In each panel, the individual rows indicate the performance metric, the individual columns the five RCMs considered.

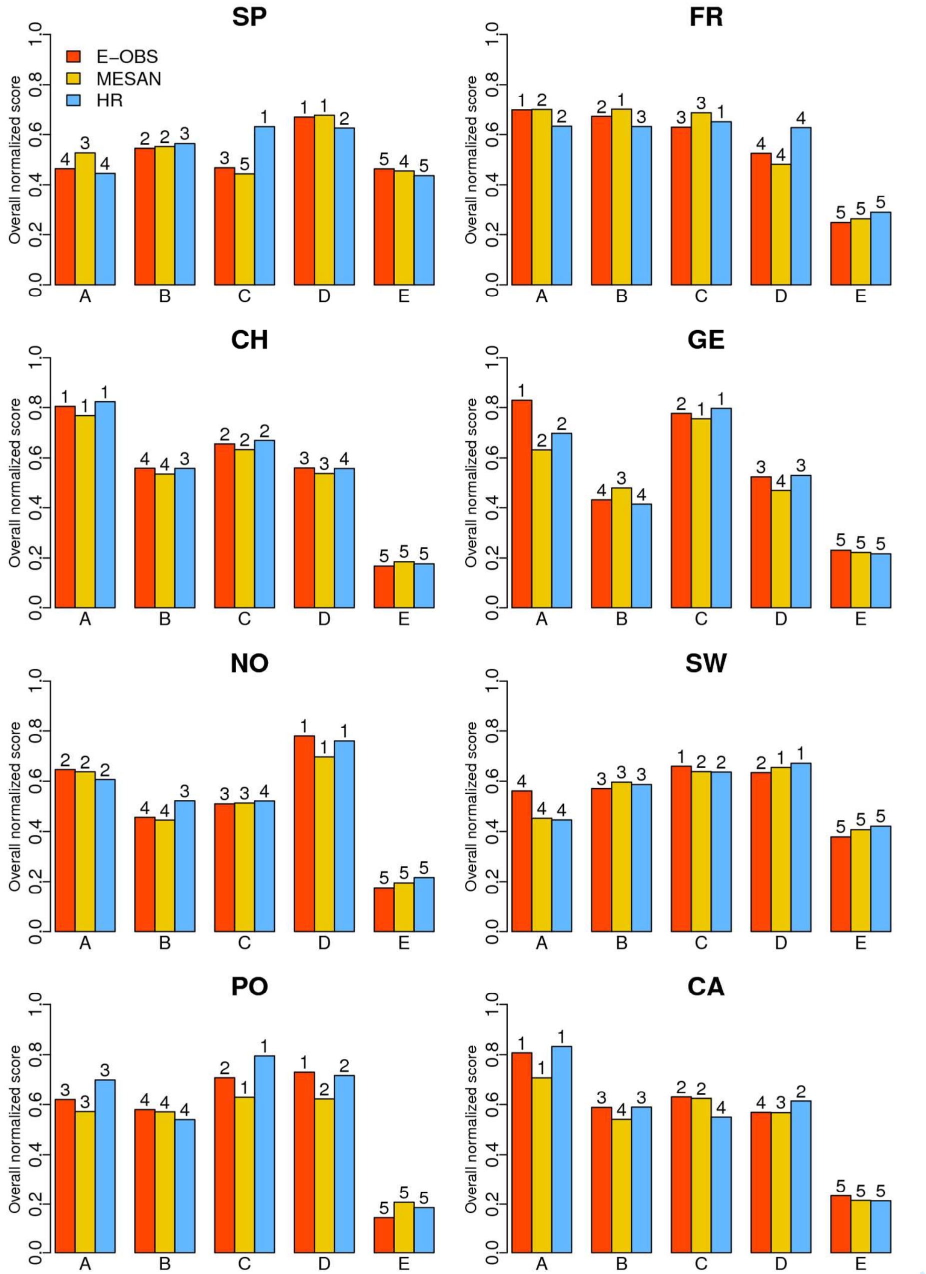
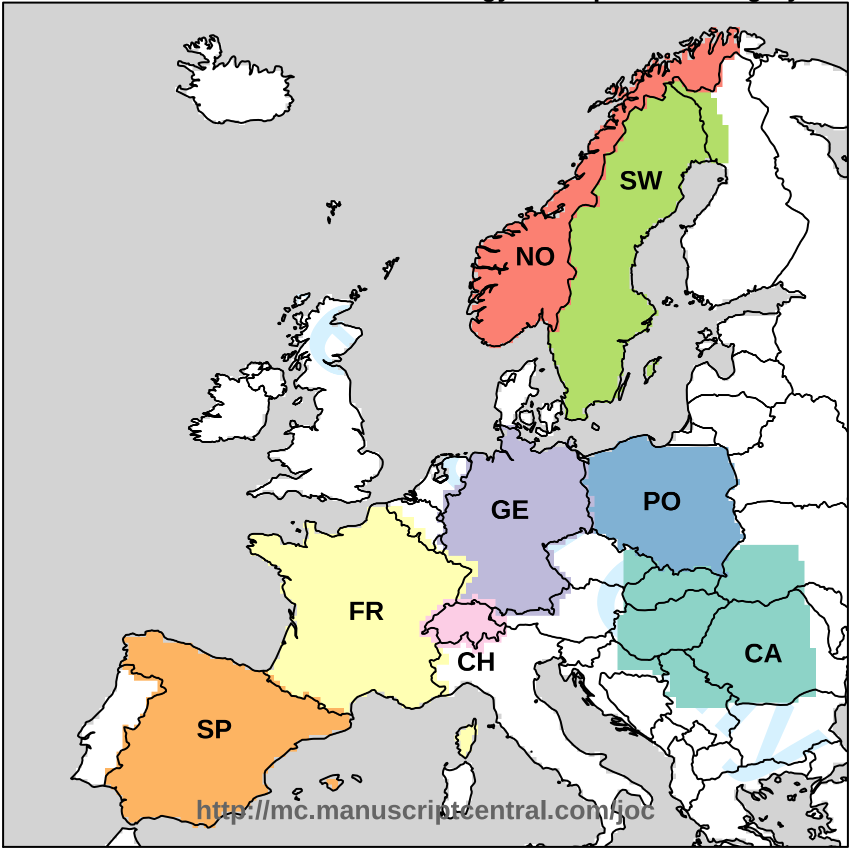


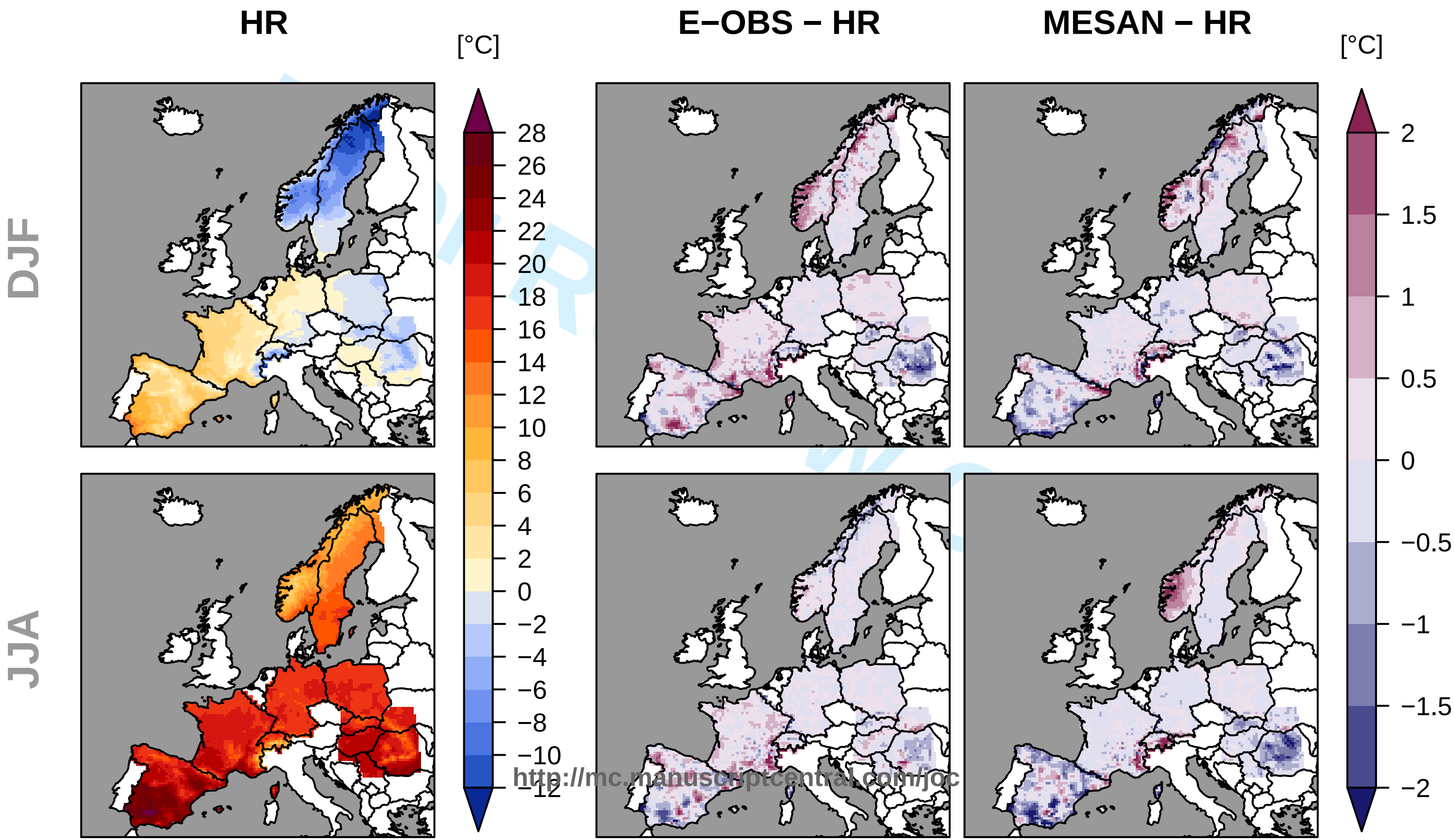
Fig. 12: Overall (combined temperature and precipitation) normalized performance scores for each sub-region. The numbering above the bars indicates the actual RCM ranks separately for each reference dataset.

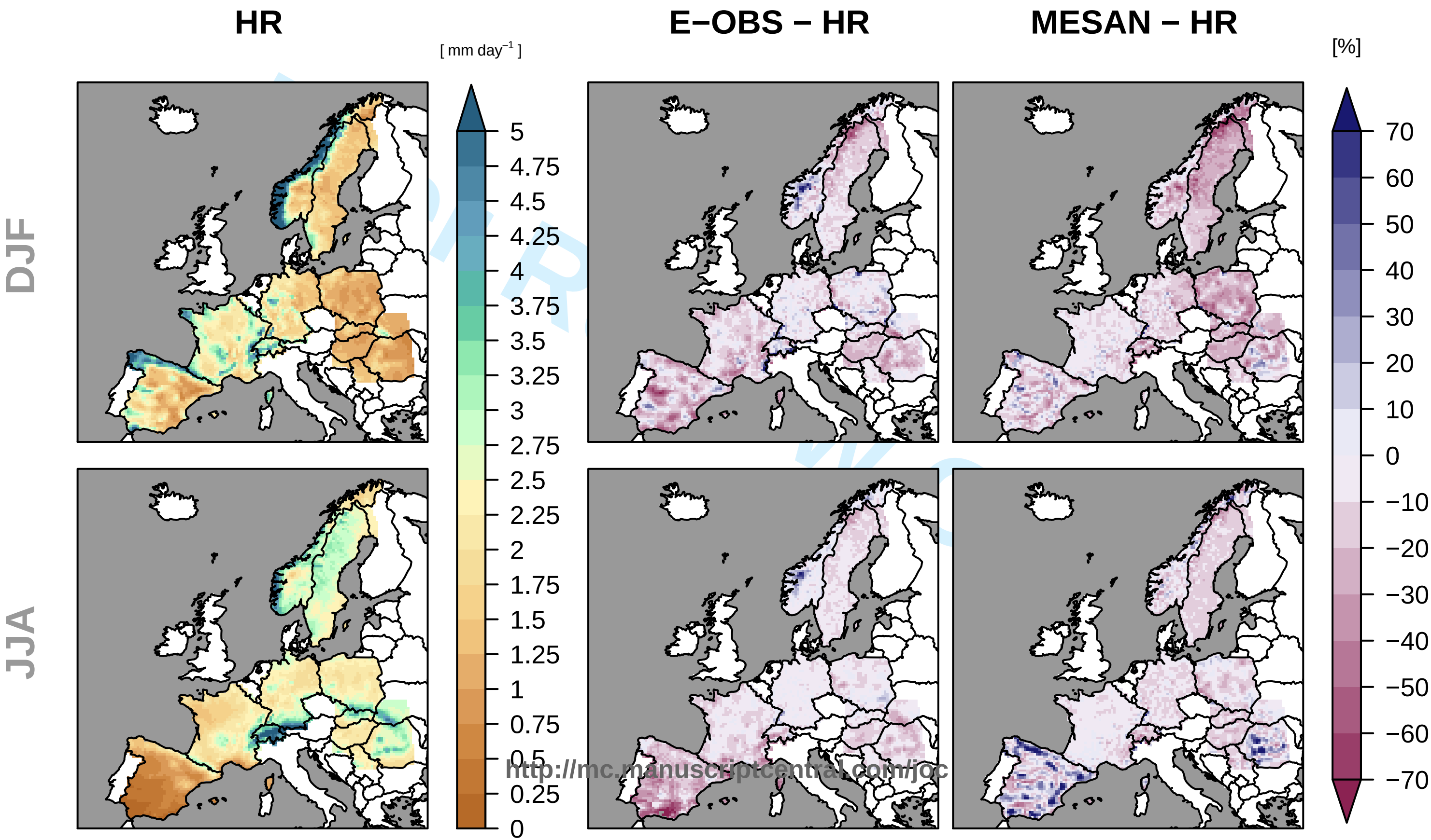
Tables

Table 1: Overview on the employed observational reference and RCM datasets. In this work the individual datasets are simply referred to by their abbreviation (last column).

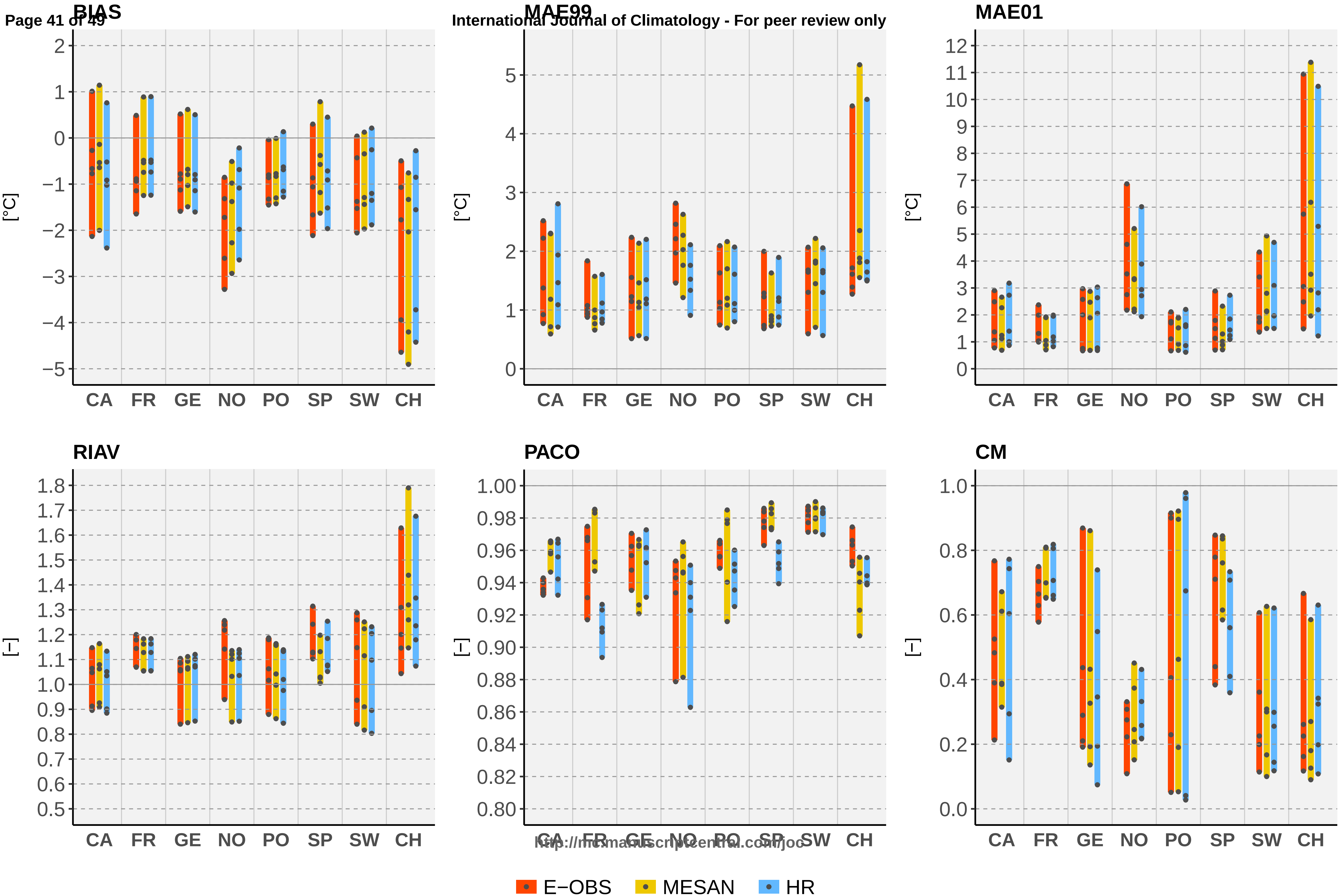
Type of dataset	**Details**		
Observational reference	Name	Description	Abbreviation
	E-OBS v15	Section 2.1.1	**E-OBS**
	National high-resolution grids	Section 2.1.2	**HR**
	EURO4M MESAN	Section 2.1.3	**MESAN**
RCM	Model name and version	Institute/Group	Abbreviation
	CCLM 4.8.17	CLMcom	**A**
	HIRHAM 5	DMI	**B**
	WRF 3.3.1F	IPSL-INERIS	**C**
	RACMO 2.2E	KNMI	**D**
	RCA 4	SMHI	**E**

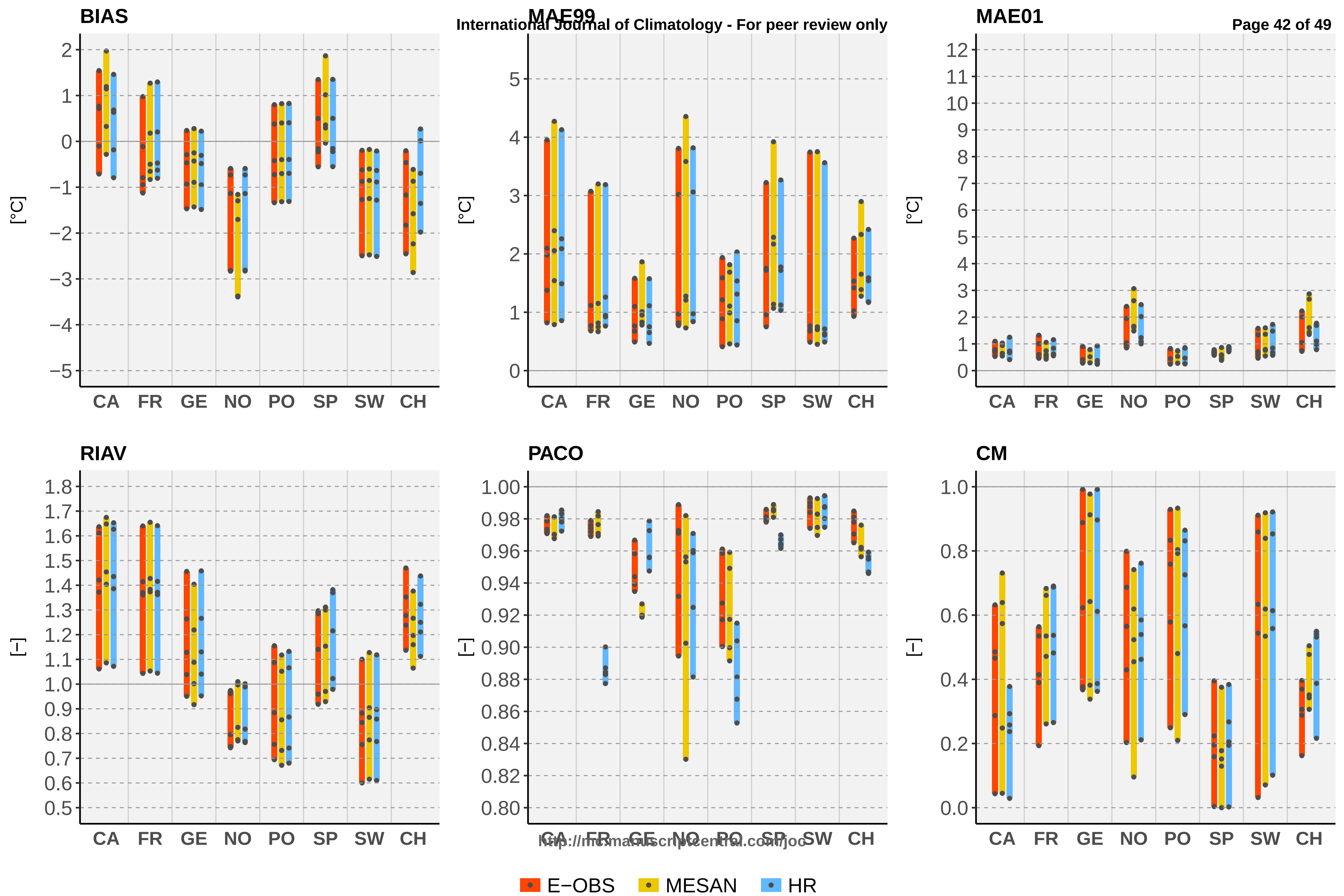

NO
SW
GE
PO
FR
CH
CA
SP

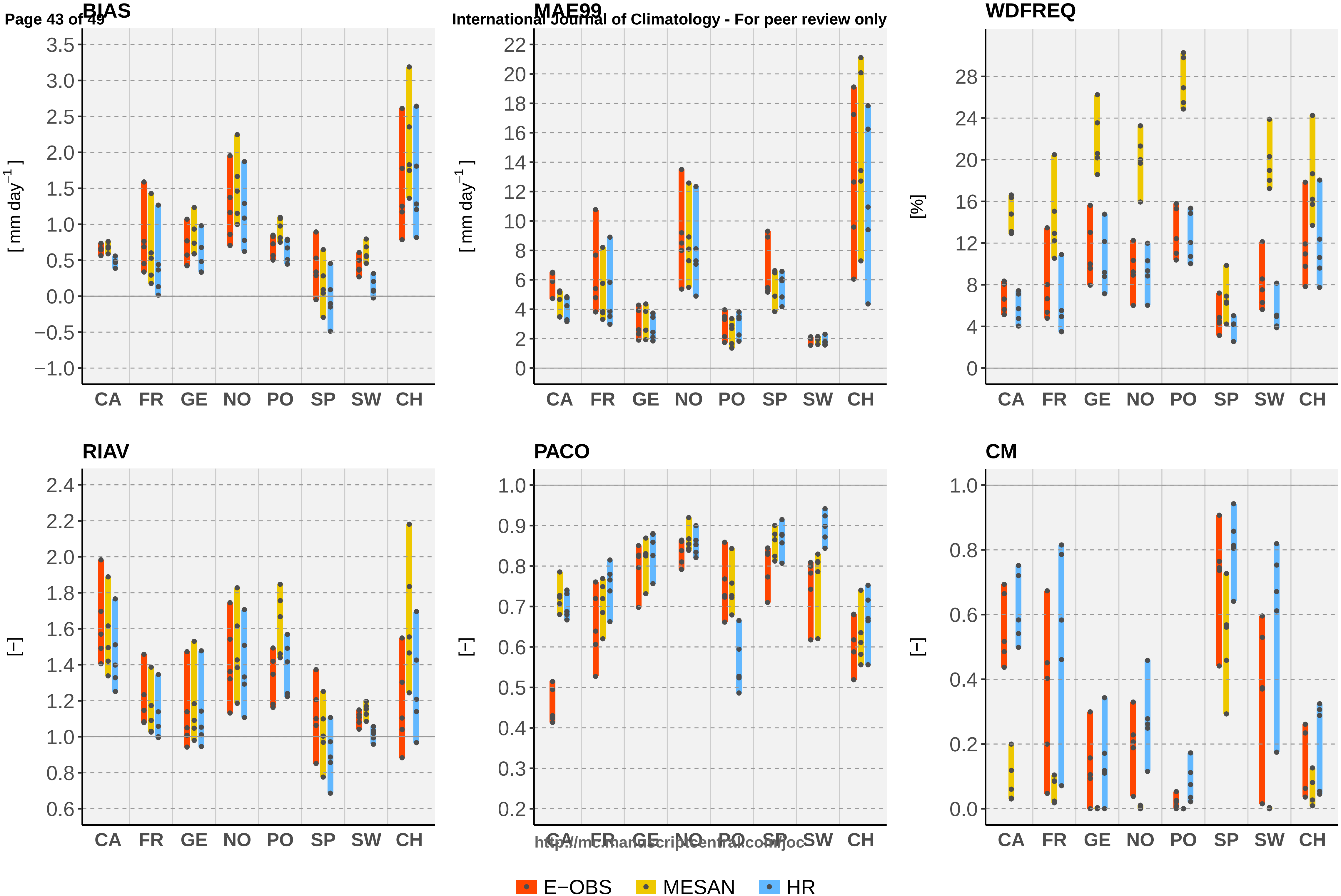

HR
E−OBS − HR
MESAN − HR
[°C]
[°C]
DJF
JJA
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
−2
−4
−6
−8
−10
−12
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2

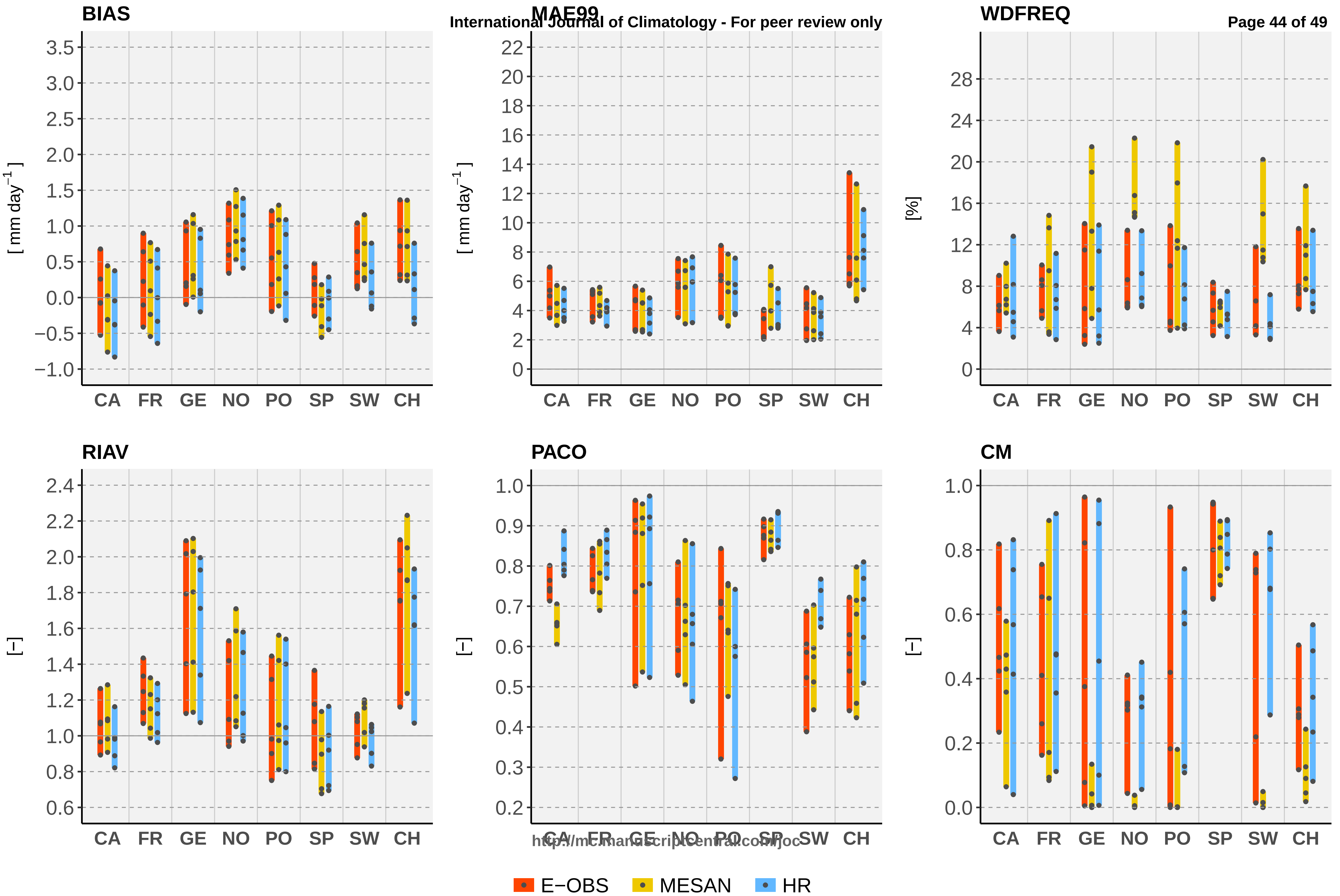


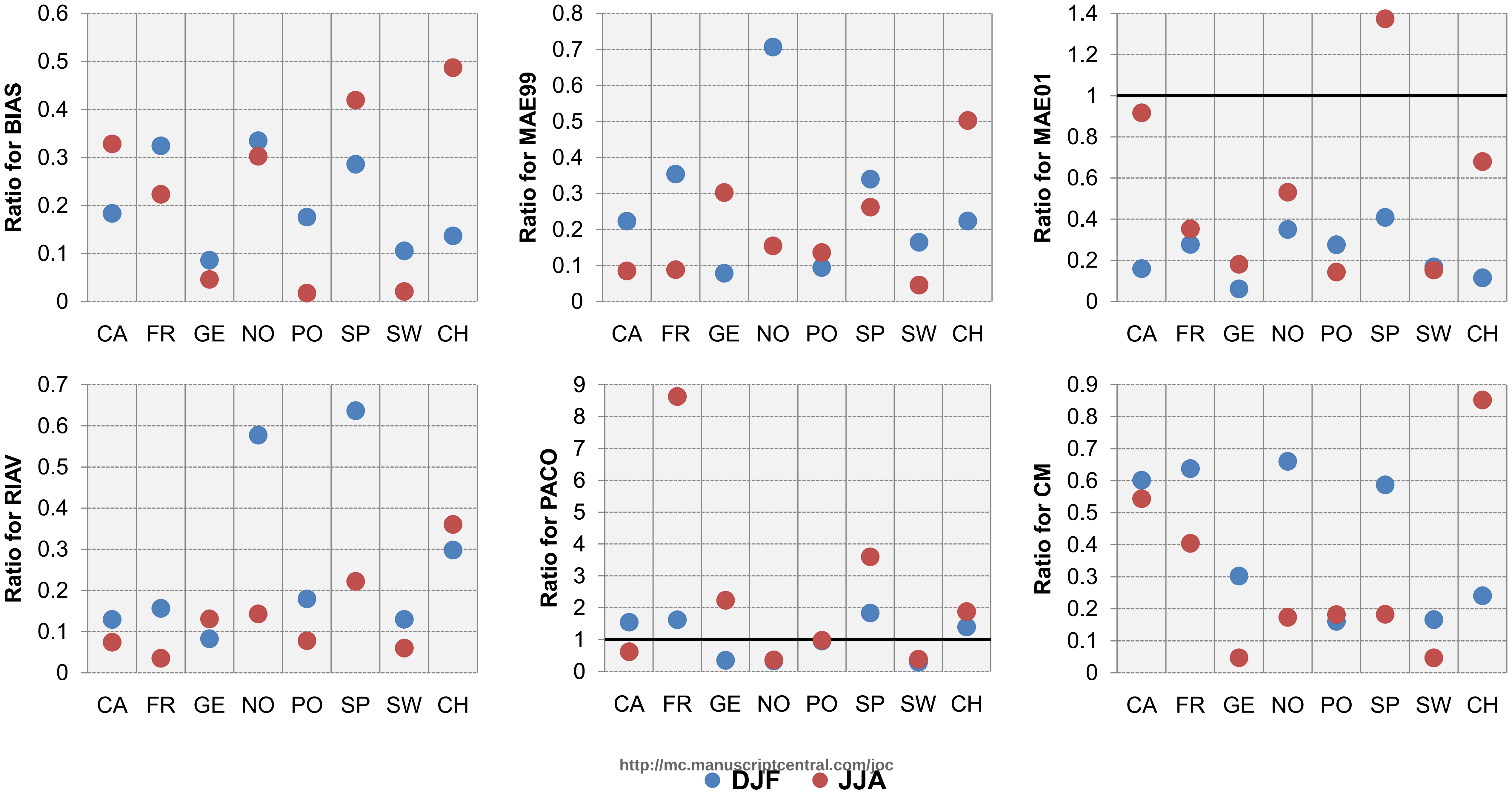
HR
E−OBS − HR
MESAN − HR
[mm day^{-1}]
[%]
DJF
JJA
5
4.75
4.5
4.25
4
3.75
3.5
3.25
3
2.75
2.5
2.25
2
1.75
1.5
1.25
1
0.75
0.5
0.25
0
70
60
50
40
30
20
10
0
−10
−20
−30
−40
−50
−60
−70


BIAS
MAE99
MAE01
RIAV
PACO
CM
[°C]
[–]
CA FR GE NO PO SP SW CH
E−OBS
MESAN
HR

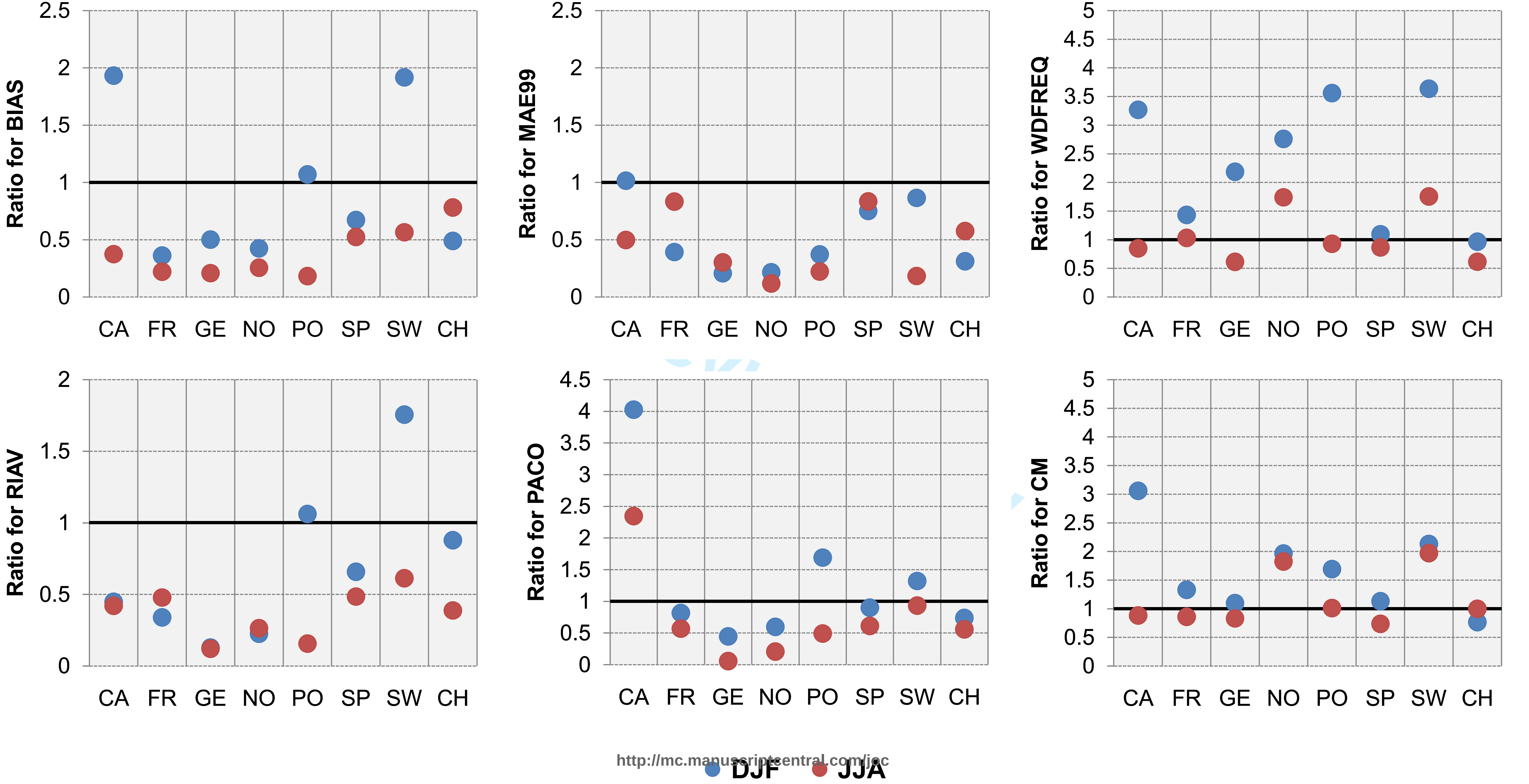

BIAS
MAE99
MAE01
RIAV
PACO
CM
[°C]
[–]
CA FR GE NO PO SP SW CH
E−OBS MESAN HR

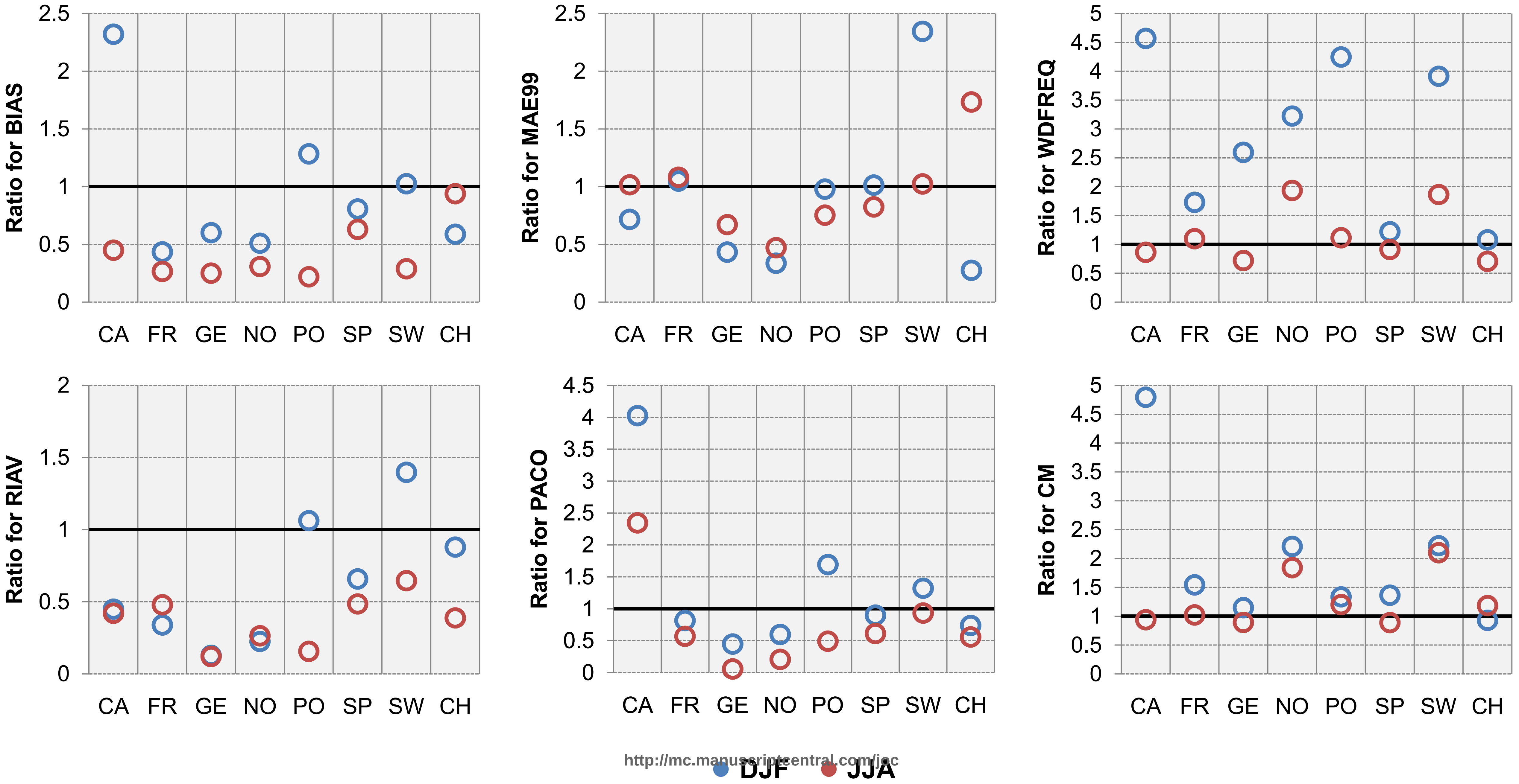

BIAS
MAE99
WDFREQ
RIAV
PACO
CM
[mm day^{-1}]
[%]
[–]
CA FR GE NO PO SP SW CH
E−OBS
MESAN
HR

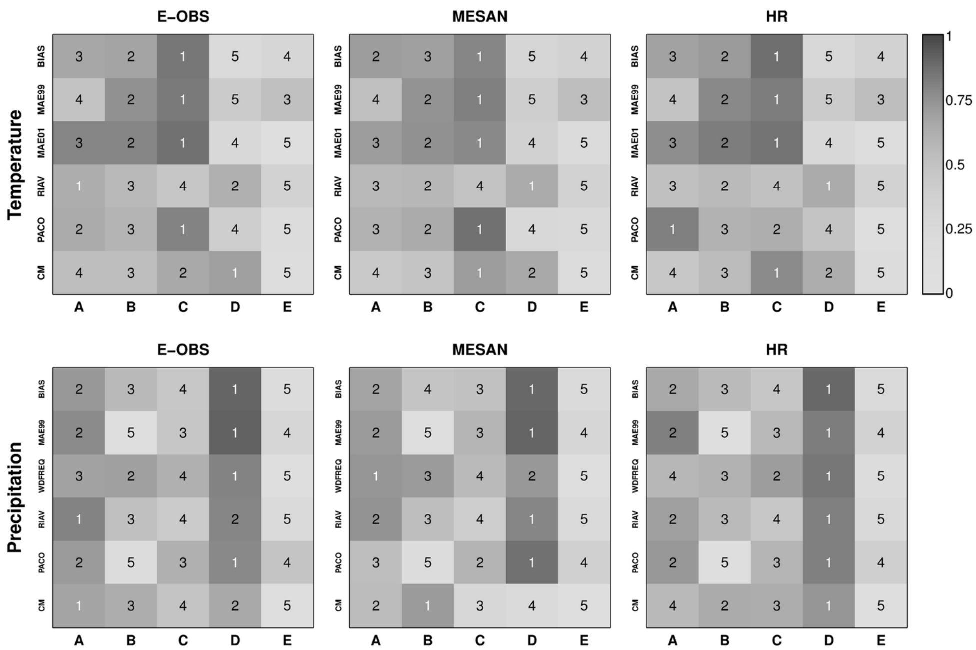

BIAS
[mm day^{-1}]
MAE99
[mm day^{-1}]
WDFREQ
[%]
RIAV
[−]
PACO
[−]
CM
[−]
CA FR GE NO PO SP SW CH
E−OBS
MESAN
HR



Ratio for BIAS
Ratio for MAE99
Ratio for MAE01
Ratio for RIAV
Ratio for PACO
Ratio for CM
CA
FR
GE
NO
PO
SP
SW
CH
DJF
JJA

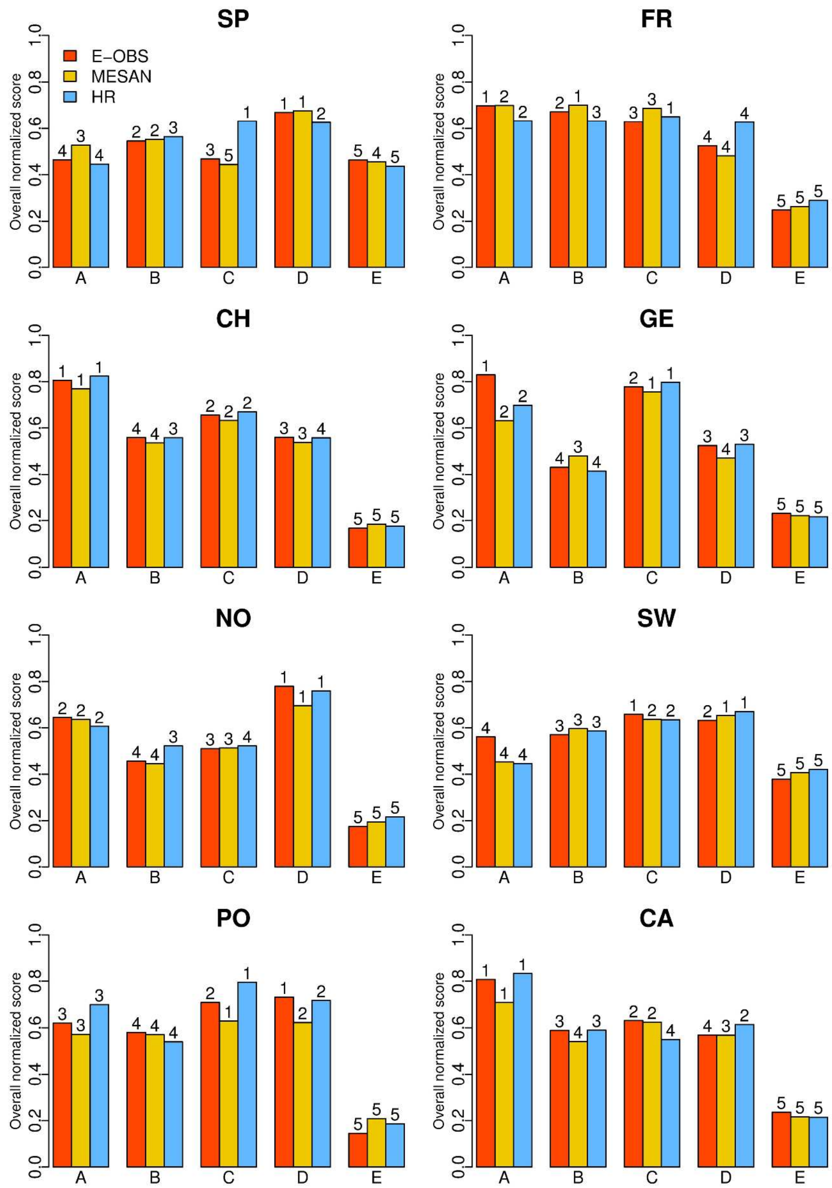

Ratio for BIAS
Ratio for MAE99
Ratio for WDFREQ
Ratio for RIAV
Ratio for PACO
Ratio for CM
CA FR GE NO PO SP SW CH
DJF
JJA


Ratio for BIAS
Ratio for MAE99
Ratio for WDFREQ
Ratio for RIAV
Ratio for PACO
Ratio for CM
CA FR GE NO PO SP SW CH
DJF
JJA



Normalized performance scores (shading) for individual performance metrics, when averaged over all seasons and regions. The upper row shows the results for temperature and the lower row for precipitation. Numbering inside the shaded boxes indicates the actual RCM rank for each case. In each panel, the individual rows indicate the performance metric, the individual columns the five RCMs considered.

152x101mm (300 x 300 DPI)



Overall (combined temperature and precipitation) normalized performance scores for each sub-region. The numbering above the bars indicates the actual RCM ranks separately for each reference dataset.

297x420mm (300 x 300 DPI)

Table 1: Overview on the employed observational reference and RCM datasets. In this work the individual datasets are simply referred to by their abbreviation (last column).

Type of dataset	Details		
Observational reference	Name	Description	Abbreviation
	E-OBS v15	Section 2.1.1	**EOBS**
	National high-resolution grids	Section 2.1.2	**HR**
	EURO4M MESAN	Section 2.1.3	**MESAN**
RCM	Model name and version	Institute/Group	Abbreviation
	CCLM 4.8.17	CLMcom	**A**
	HIRHAM 5	DMI	**B**
	WRF 3.3.1F	IPSL-INERIS	**C**
	RACMO 2.2E	KNMI	**D**
	RCA 4	SMHI	**E**