Accepted manuscript. International Journal of Climatology (Jan 2019). DOI: 10.1002/joc.6024

# Validation of spatial variability in downscaling results from the VALUE perfect predictor experiment

M. Widmann[1], J. Bedia[2,3], J.M. Gutiérrez[4], T. Bosshard[5], E. Hertig[6], D. Maraun[7], M.J. Casado[8], P. Ramos[8], R.M. Cardoso[9], P.M.M. Soares[9], J. Ribalaygua[10], C. Pagé[11], A. Fischer[12], S. Herrera[2], and R. Huth[13]

[1]School of Geography, Earth and Environmental Sciences, University of Birmingham, UK
[2]Dept. Applied Mathematics and Computing Science, University of Cantabria, Santander, Spain
[3]Predictia Intelligent Data Solutions S.L., Santander, Spain
[4]National Research Council (CSIC), Instituto de Física de Cantabria, Santander, Spain
[5]Swedish Meteorological and Hydrological Institute (SMHI), Norrköping, Sweden
[6]Dept. of Geography, University of Augsburg, Germany
[7]Wegener Center for Climate and Global Change, University Graz, Austria
[8]Agencia Estatal de Meteorología (AEMET), Madrid, Spain
[9]Instituto Dom Luiz (IDL), Faculdade de Ciências, Universidade de Lisboa, Portugal
[10]Fundación para la Investigación del Clima (FIC), Spain
[11]Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique (CERFACS), Toulouse, France
[12]Federal Office of Meteorology and Climatology (MeteoSwiss), Zurich, Switzerland
[13]Institute of Atmospheric Physics, Charles University, Prague, Czech Republic

February 5, 2019

Accepted manuscript. International Journal of Climatology (Jan 2019). DOI: 10.1002/joc.6024

# Contents

## Abstract

The spatial dependence of meteorological variables is crucial for many impacts, e.g. droughts, floods, river flows, energy demand, and crop yield. There is thus a need to understand how well it is represented in downscaling products. Within the COST Action VALUE we have conducted a comprehensive analysis of spatial variability in the output of over 40 different downscaling methods in a perfect predictor setup. The downscaling output is evaluated against daily precipitation and temperature observations for the period 1979-2008 at 86 sites across Europe and 53 sites across Germany. We have analysed the dependency of correlations of daily temperature and precipitation series at station pairs on the distance between the stations. For the European dataset we have also investigated the complexity of the downscaled data by calculating the number of independent spatial degrees of freedom. For daily precipitation at the German network we have additionally evaluated the dependency of the joint exceedance of the wet day threshold and of the local 90th percentile on the distance between the stations. Finally we have investigated regional patterns of European monthly precipitation obtained from rotated principal component analysis.

We analysed Perfect Prog methods, which are based on statistical relationships derived from observations, as well as Model Output Statistics approaches, which attempt to correct simulated variables. In summary we found that most Perfect Prog downscaling methods, with the exception of multi-site analog methods and a method that explicitly models spatial dependence yield unrealistic spatial characteristics. RCM-based Model Output Statistics methods showed good performance with respect to correlation lengths and the joint occurrence of wet days, but a substantial overestimation of the joint occurrence of heavy precipitation events. These findings apply to the spatial scales that are resolved by our observation network, and similar studies with higher resolutions, which are relevant for small hydrological catchment, are desirable.

# 1 Introduction

Projections for future climate change are primarily based on simulations with coupled atmosphere-ocean general circulation models (GCMs). Their relatively coarse horizontal resolution of around 100 km means that not all relevant atmospheric processes can be realistically modelled, which leads to errors on the resolved scales. Moreover, the output does not have the spatial resolution often needed for impact and adaptation studies. In order to overcome these problems downscaling (DS) methods are routinely used, either based on high-resolution regional climate models (RCMs), on statistical methods, or on a combination of both (Maraun and Widmann, 2018; Ekstroem et al., 2015; Hewitson et al., 2014; Maraun et al., 2010).

The spatial structure of the output from DS methods is highly relevant when the results are used to assess impacts that are determined by spatial aggregation of meteorological variables. Typical examples for which a realistic representations of spatial variability matters are river flow and floods (Arnaud et al., 2002; Segond et al., 2007; Viviroli et al., 2009), droughts (Trambauer et al., 2015), glacier mass balance (Machguth et al., 2009), ecosystem composition (Monestiez et al., 2001), crop yields (Holzkämper et al., 2012), energy consumption and production, as well as weather-related health problems. For instance an over- or underestimation of correlations between precipitation timeseries at different locations within a river catchment would typically lead to an over- or underestimation of high and low river flow conditions.

Within the COST action VALUE a comprehensive validation framework for DS methods has been designed and implemented (Maraun et al., 2015). The user-relevant aspects of DS output identified in the framework are marginal distributions including extremes, temporal variability, and intervariable relationships, all considered at individual locations, as well as spatial variability. The performance of DS methods with respect to the aspects defined at individual stations within Europe has been investigated in the companion papers in this special issue (Gutiérrez et al., 2018; Hertig et al., 2018; Maraun et al., 2018). Here we analyse specifically how well the different DS methods represent the spatial structure of precipitation and temperature fields over Europe. As pointed out in Maraun et al. (2015) it is usually not the spatial pattern of the long-term mean but the structure of the individual events that is relevant for impacts, because it includes for instance the information on whether all locations within a river catchment tend to receive precipitation at the same time, or whether it is likely that some areas stay dry when there is precipitation in others. It can be useful to remove the effect of the climatological mean on individual events and to analyse the residual spatial variability, i.e. to express the data as deviations from the long-term mean.

More formally speaking, when considering a meteorological variable simultaneously at different locations we are dealing with a multivariate dataset given by the values at the different locations, and the goal when validating spatial variability is to investigate the similarity of the observed and downscaled data clouds. To a first order approximation the datasets are characterised by their multivariate long-term temporal means, i.e. by the patterns of the climatological mean. For the observations it is mainly influenced by the meridional gradient and local differences in the radiation budget, the proximity to the oceans, the mean large-scale atmospheric circulation, and topography. These factors influence meteorological processes such as atmospheric stability, convection, flow convergence, frontal passages, or Foehn, which affect the spatial structure of individual weather events as well as of the long-term mean. It can be expected that almost all statistical DS method reproduce the mean temperature and precipitation fields quite well by construction, for instance by estimating anomalies around the observed mean in the case of regression-based methods or by adjusting distributions. The skill of DS methods with respect to representing the mean has been analysed to some extent in Gutiérrez et al. (2018), albeit without explicitly investigating the spatial pattern of the bias of the long-term means. The mean bias in the raw output of regional models has been investigated in many publications (e.g. Kotlarski et al., 2014; Isotta et al., 2015). Moreover, as already mentioned, it is mostly the structure of the residual spatial variability that is impact-relevant. We therefore focus in our analysis on the spatial structure of the residual variability, mainly on the daily timescale.

For multivariate Gaussian data the structure of the variability around the mean is fully captured by the covariance matrix, and for normalised data by the correlation matrix. It is thus a natural starting point to investigate the similarity of the observed and the downscaled covariances or correlations between different locations. As correlations are a direct measure for the strength of linear relationships between timeseries we will consider those. We will also investigate the probabilities for joint exceedances of thresholds, which are of practical relevance for impact modelling and which for non-Gaussian data do not directly follow from the covariance matrix. We note that multivariate data can alternatively be described by a combination of their marginal distributions, which are investigated in Gutiérrez et al. (2018), and copulas that analytically express the dependence structure. However, for brevity this approach is not taken here. In addition we will analyse the overall complexity, and the representation of regional patterns. Details on our validation approach are given in the method section.

In spite of the importance of the spatial structure of daily values for climate impacts, only a few studies have validated the spatial aspects of standard deterministic Perfect Prog (PP) downscaling products. Correlations between timeseries at different locations, including their dependency on distance, have been analysed (Easterling, 1999; Kettle and Thompson, 2004; Huth et al., 2008, 2015), and homogeneous regions have been investigated by cluster analysis (Huth, 2002). These studies, most of which focus on temperature, indicate that PP methods that use large-scale predictors overestimate spatial correlations, whereas local analog methods underestimate them. Huth et al. (2015) additionally included two RCMs in the method comparison and found no systematic over- or underestimation for them. A comparison of some PP and MOS methods, as well as RCMs, undertaken by Ayar et al. (2016) included some analysis of spatial variability of daily precipitation based on the leading Principal Component (PC) loading patterns and on correlations of daily patterns. The study found a mixed performance of the RCMs and MOS with better skill in winter than in summer, and in general low performance for PP methods. The analog method showed as expected realistic PC loadings but failed to capture the individual daily patterns.

In addition, stochastic PP methods that explicitly model spatial structure have been developed and analysed. Frost et al. (2011) evaluated correlations of occurrence and amount of daily precipitation at different locations obtained from a Nonhomogeneous Hidden Markov Model (NHMM) for occurrence combined with conditional multiple regression for amounts, and from GLIMCLIM, a conditional multisite weather generator based on a generalised linear model, and found that both substantially underestimated intersite correlations. Hu et al. (2013) obtained similar results for GLIMCLIM, but found in contrast that a NHMM performed well. The difference can be a result of both the predictor choice, or the specific regional climate. A further method type are conditional multisite weather generators for precipitation constrained by the observed dependences between sites, which were found to represent the observed properties well (Cannon, 2008; Wilks, 2012).

Disaggregation methods for precipitation investigated in Ferraris et al. (2003) show substantial over- and underestimations of intersite correlations with no method performing systematically better than others. However, advanced stochastic models for precipitation that include a disaggregation step based on two-dimensional, latent Gaussian fields showed realistic spatial characteristics (Paschalis et al., 2013).

Recently several analog methods in which the analogs are based on a coarse resolution representation of the predictand variable rather than on the large-scale atmospheric circulation have been developed. There are different implementations depending on how model biases are treated and on how the downscaled field is constructed from a pool of analog situations; for a description of the frequently used 'localised constructed analog method' (LOCA) and a discussion of other variants see Pierce et al. (2014). They are implemented such that a common analog is chosen for adjacent locations and thus yield realistic spatial fields by construction if individual analogs are used and fairly realistic fields if weighted means of multiple analogs are used. An intercomparison of bias corrected constructed analogs (BCCA), of methods combining bias correction for monthly or daily fields and spatial disagregation (BCSDm, BCSMd), and of an asynchronous regression method is presented in Gutmann et al. (2014), who found that all methods but BCSDm substantially overestimate spatial correlations. The reason for the good performance of BCSDm is that in contrast to the other methods it inherits the spatial variability from the observations, rather than from the driving model.

Recent developments also include multisite MOS methods. Bárdossy and Pegram (2012) found that RCM precipitation had too low intersite correlations and formulated a matrix and a sequential recorrelation method to adjust the spatial structure, with the former applicable to match Pearson correlations and the latter to reproduce more general copula-based representations of the multivariate structure. The correction methods led to a realistic spatial structure, with the exception of an underrepresentation of clustering of extreme precipitation, allow for changes in the spatial dependences in a future climate, and mainly preserve the temporal structure of the RCM output. Cannon (2018) developed a multivariate quantile mapping method that yields the observed multivariate distribution, applied it to correct spatial RCM precipitation fields, and demonstrated realistic spatial characteristics of the corrected fields. There are also parametric quantile mapping methods that interpolate the observed distribution parameters to high spatial resolution (Mamalakis et al., 2017), but as they do not model the spatial structure of variability they are essentially singlesite MOS methods.

In the context of ensemble weather forecasting postprocessing methods have been used that rearrange the simulated data in time so they have the same rank structure as the observations in a training period (known as Schaake Shuffle), which leads to a reproduction of the spatial and intervariable dependence structure of the training data (Clark et al., 2004). The method has been employed to provide input for hydrological forecasts (Voisin et al., 2011) and to postprocess atmospheric reanalyses (Vrac and Friederichs, 2015). A drawback that makes its application in a climate change context problematic is that it is constrained to reproduce the temporal rank structure of the training dataset. Vrac (2018) has suggested a rank-based resampling method that relaxes this condition and also introduces stochasticity by generating as many multivariate corrected outputs as the number of statistical dimensions (i.e. number of grid-cells × number of climate variables). This study has also demonstrated how to apply the method in a climate change context. However, further research on the usefulness of the method for climate change studies is needed, for instance because the reshuffling breaks the physical consistency between large-scale atmospheric states and the postprocessed variables, and will usually modify the climate change signal.

Our analysis extends these studies by considering a large number of downscaling methods (47 for precipitation and 45 for temperature) and by systematically comparing them with respect to several measures of spatial variability, using validation datasets over Europe and Germany. The structure of DS methods can be expected to have a strong influence on the spatial variability of their output. Singlesite methods, which are fitted to individual target locations, might for instance yield a realistic spatial structure if the predictors explain a large fraction of the local variability, but might overestimate spatial correlations if small-scale variability is substantial and not adequately represented. A detailed analysis of the variance explained by each downscaling method is provided in Gutiérrez et al. (2018). Multisite DS methods, which simultaneously use several locations for model fitting, might either achieve realistic spatial variability through the common influence of predictors or through explicit constraints on the multivariate structure of noise components or of the final output. In our study we compare downscaling methods of different types which will allow us to investigate whether some types exhibit a common behaviour with respect to spatial variability. We note that the VALUE perfect predictor experiment uses an ensemble of opportunity in which most of the methods are fitted on single sites, reflecting the dominance of such methods in DS applications. In particular, no method explicitly models spatial dependence in the European-wide experiment, although for some methods, spatial dependence results as a consequence of the use of common predictors (e.g. regression methods using PCs) or of the method characteristics (e.g. some analog methods using the same analog day for all sites). However, for the additional experiment over Germany, two regression methods that explicitly consider spatial dependence have contributed to the study.

Section 2 starts with a discussion of the observations used for validation as well as of the downscaled data, including a brief overview of the different types of downscaling methods and of the experimental setup. It then continues with an explanation of the different measures for spatial variability employed to validate and compare the downscaling methods. Section 3 will present the validation results in separate subsections for each validation measure. Summary and conclusions will be given in section 4.

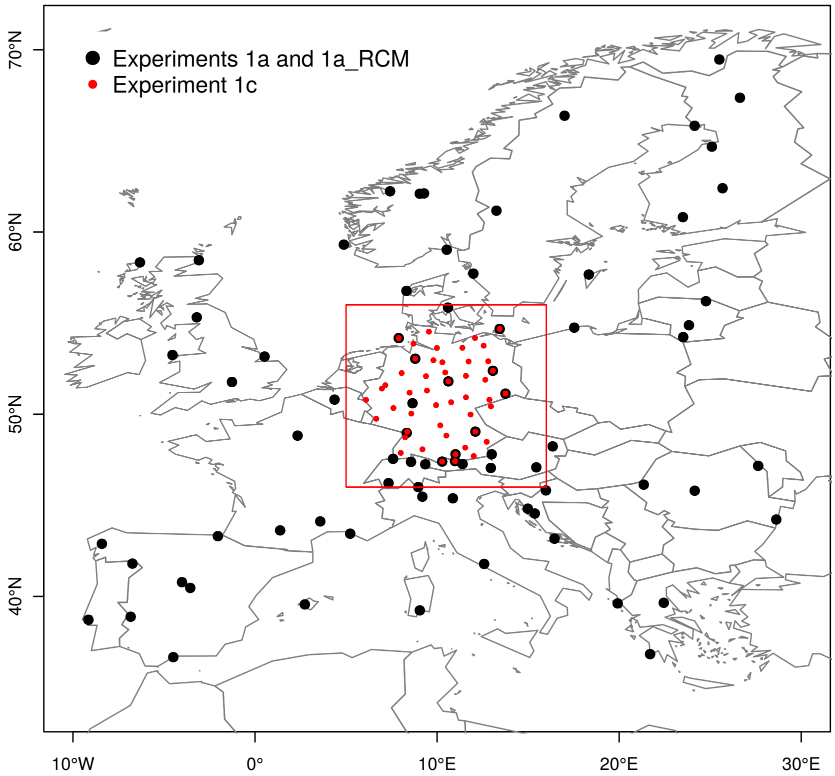


Figure 1: Locations of the reference stations for the European experiments (1a and 1a-RCM, black circles, VALUE-ECA-86-v2 dataset) and the German experiment (1c, red, VALUE-ECA-53-Germany-spatial-v1 dataset).

# 2 Data and methods

## 2.1 Observations and downscaled data

The predictands for the DS methods are observations for daily precipitation as well as for daily minimum and maximum temperature at 86 stations across Europe. This VALUE-ECA-86-v2 dataset is a subset of the publicly available ECA dataset (Tank et al., 2002) and covers the period 1979 - 2008. Besides the European-wide experiment (referred to as experiment_1a, or simply exp_1a), which is the common experiment for the different validation studies, we also present here the results of an experiment based on a denser ECA subset of 53 stations within Germany for the same variables (referred to as experiment_1c, or simply exp_1c), which was designed to focus on spatial validation aspects. Details on data availability are given in Gutiérrez et al. (2018). Both networks are shown in Fig. 1.

The downscaling methods that have been considered in our study for precipitation are listed in Table 1, those used for temperature in Table 2. The columns '1a' and '1c' indicate the methods contributing to each of the experiments. All downscaling methods have been calibrated following a five-fold cross validation with non-overlapping consecutive 6-year blocks. Further details about the methods and the experimental setup can be found in Maraun et al. (2015), Gutiérrez et al. (2018), and on `www.value-cost.eu/validation#Experiment_1a`.

We distinguish between PP and MOS methods (see e.g. Maraun et al., 2010)). For the former the statistical relationships are derived from observations whereas MOS methods are fitted using predictors from RCMs (or global climate models). PP methods represent real-world links between large-scale predictors and the local predictand, and thus in applications to output from climate models they require realistically simulated predictors – hence the name 'Perfect Prog(nosis)'. MOS methods represent relationships between simulated and observed variables, are therefore model-specific, and do not only represent downscaling relationships but can also correct model biases. Unconditional weather generators (WGs), which are statistical models that produce timeseries with temporal characteristics similar to observations without any predictors are a third group of methods listed under 'WG'. Conditional WGs, which include meteorological predictors that influence the properties of the timeseries, should not be categorised as a separate group to MOS and PP, because depending on the setup for model fitting they either follow the PP or MOS approach, and are thus listed under either PP or MOS.

The PP methods are validated in a perfect predictor setup using predictors from the ERA-Interim Reanalysis (Dee et al., 2011) for the period 1979 - 2008 on a coarse-grained 2° resolution, which is similar to typical output from global climate models. The PP assumption for the predictors is thus met by construction. The MOS methods for the European experiment exp_1a are directly applied to the ERA-Interim data on both the original 0.75° and on the coarse-grained resolution. We have conducted an additional European experiment exp_1a_RCM for which the MOS predictors are taken from the RACMO RCM (van Meijgaard et al., 2008) driven by perfect boundary conditions from ERA-Interim on the original 0.75° resolution. For the German experiment exp_1c we have used MOS predictors from ERA-Interim on the original 0.75° resolution.

The PP methods used here cover the widely used approaches, i.e. analog, regression and weather type methods; the MOS methods cover frequently used quantile mapping methods as well as recently developed stochastic MOS.

Information on the structural elements of the DS methods that may influence the spatial characteristics of the ouput are also given in tables 1 and 2. The 'MS' column indicates whether the DS model has been fitted simultaneously for multiple (or all) locations ('yes') or individually for each location ('no'). The 'EX' column lists whether the statistical model has explicit constraints on the structure of spatial variability ('yes'), for instance on correlations for adjacent locations. The 'ST' column indicates whether the DS output contains stochastic noise ('yes'). The final column 'PC' states whether or not principal components have been used as predictors. As already mentioned almost all of the methods are fitted and applied at single sites, with only some analog methods being applied to multiple sites. Note that methods that are fitted at individual sites might still be used for multiple sites if for instance realistic spatial patterns can be expected through the influence of the predictors.

All methods participating in the European experiment are fully described in Annex 1 of Gutiérrez et al. (2018). We now describe the two additional methods, GLM-BN-DET and DSCLIM-D, contributing only to the German experiment. GLM-BN-DET is a multivariate extension of the GLM-DET method, which explicitly models the spatial structure of precipitation occurrence by considering a dependence graph linking marginally and/or conditionally dependent stations.

This graph allows to obtain a probabilistic model (a Bayesian network) which encodes all the dependences displayed in the graph by means of an appropriated factorisation of the joint probability distribution. This model allows simulating spatially consistent precipitation occurrences. Moreover, for each particular station, the model determines the set of stations (Markov blanket) exerting a spatial influence. For each station, this set is included as spatial predictors (in addition to the large-scale information) in the binomial/gamma GLM model thus the model yields spatially consistent precipitation amounts. Details on this particular methodology are given in Cano et al. (2004). DSCLIM-D is based on weather typing, combined with linear regression and weather analogs. The method has been introduced by Boé et al. (2006), but the version used here differs in some details. The implementations for temperature and precipitation are slightly different, and for brevity we explain only the latter case. DSCLIM-D uses a clustering method to determine weather types (10 in this implementation) in the SLP field. For each day the Euclidean distances of the SLP field to all the weather types are calculated and used as predictors for the square root of the precipitation anomaly at a given location in a multiple linear regression. The mean of the estimated precipitation over all stations in the target area is then used to define a set of analog days from which the downscaled local precipitation is chosen. The set is defined by the days in the fitting period that belong to the same weather type as well as have averaged precipitation in the same decile as the estimated averaged precipitation. We note that comparing deciles is similar to quantile mapping or inflated regression. In the deterministic version of the method, which is used here, one analog precipitation field is randomly selected, the stochastic version used several analogs.

| **Type** | **Code** | **Tech** | **1a** | **1c** | **MS** | **EX** | **ST** | **PC** |
|---|---|---|---|---|---|---|---|---|
| **MOS** | Ratyetal-M6 | S | × | - | no | no | no | no |
| | Ratyetal-M7 | S | × | - | no | no | no | no |
| | ISI-MIP | S/PM | × | × | no | no | no | no |
| | DBS | PM | × | × | no | no | no | no |
| | Ratyetal-M9 | PM | × | - | no | no | no | no |
| | BC | PM | × | × | no | no | no | no |
| | GQM | PM | × | × | no | no | no | no |
| | GPQM | PM | × | × | no | no | no | no |
| | EQM | QM | × | × | no | no | no | no |
| | EQMs | QM | × | - | no | no | no | no |
| | EQM-WT | QM/WT | × | × | no | no | no | no |
| | QMm | QM | × | × | no | no | no | no |
| | QMBC-BJ-PR | QM | × | - | no | no | no | no |
| | CDFt | QM | × | - | no | no | no | no |
| | QM-DAP | QM | × | - | no | no | no | no |
| | EQM-WIC658 | QM | × | - | no | no | no | no |
| | Ratyetal-M8 | QM | × | - | no | no | no | no |
| | MOS-AN | A | × | - | yes | no | no | no |
| | MOS-GLM | TF | × | - | no | no | yes | no |
| | VGLMGAMMA | TF/WG | × | - | no | no | yes | no |
| | FIC02P | PM/A/TF | × | × | no | no | no | no |
| | FIC04P | PM/A/TF | × | × | no | no | no | no |
| **PP** | FIC01P | A/TF | × | × | yes | no | no | no |
| | FIC03P | A/TF | × | × | yes | no | no | no |
| | ANALOG-ANOM | A | × | - | yes | no | no | no |
| | ANALOG | A | × | × | yes | no | no | yes |
| | ANALOG-MP | A | × | × | yes | no | yes | no |
| | ANALOG-SP | A | × | - | yes | no | yes | no |
| | MO-GP | TF | × | - | no | no | no | no |
| | GLM-P | TF | × | × | no | no | yes[a] | no |
| | MLR-RAN | TF | × | × | no | no | no | no |
| | MLR-RSN | TF | × | × | no | no | no | no |
| | MLR-ASW | TF | × | - | no | no | yes | no |
| | MLR-ASI | TF | × | × | no | no | no | no |
| | GLM-DET | TF | × | × | no | no | no | yes |
| | GLM | TF | × | - | no | no | yes | yes |
| | GLM-WT | TF/WT | × | × | no | no | yes | yes |
| | GLM-BN-DET | TF | - | × | yes | yes | no | yes |
| | DSCLIM-D | A/WT | - | × | yes | no | no | no |
| | WT-WG | WT | × | - | no | no | yes | yes |
| | SWG | TF | × | - | no | no | yes | yes |
| **WG** | SS-WG | WG | × | - | no | no | yes | no |
| | MARFI-BASIC | WG | × | - | no | no | yes | no |
| | MARFI-TAD | WG | × | - | no | no | yes | no |
| | MARFI-M3 | WG | × | - | no | no | yes | no |
| | GOMEZ-BASIC | WG | × | - | no | no | yes | no |
| | GOMEZ-TAD | WG | × | - | no | no | yes | no |

Table 1: Participating methods for precipitation for the European (exp1a) and German experiment (exp1c). Techniques: A: analog; S: scaling; PM: parametric quantile mapping; QM: empirical quantile mapping; TF: regression-like transfer function; WT: weather typing; WG: weather generator. Columns 1a and 1c indicate whether the methods have participated in the European and German experiment. MS: Multisite fitting: MS; EX: Explicitly modelled spatial structure; ST: Stochastic noise; PC: PCs used as predictors. [a] Only occurrence is randomised, amounts are based on inflated regression (in this case, the results are based on a single realisation).

| **Type** | **Tech** | **Code** | **MS** | **EX** | **ST** | **PC** |
|---|---|---|---|---|---|---|
| **MOS** | RaiRat-M6 | S | no | no | no | no |
| | RaiRat-M7 | S | no | no | no | no |
| | RaiRat-M8 | S | no | no | no | no |
| | SB | S | no | no | no | no |
| | ISI-MIP | S/PM | no | no | no | no |
| | DBS | PM | no | no | no | no |
| | GPQM | PM | no | no | no | no |
| | EQM | QM | no | no | no | no |
| | EQMs | QM | no | no | no | no |
| | EQM-WT | QM/WT | no | no | no | no |
| | QMm | QM | no | no | no | no |
| | QMBC-BJ-PR | QM | no | no | no | no |
| | CDFt | QM | no | no | no | no |
| | QM-DAP | QM | no | no | no | no |
| | EQM-WIC658 | QM | no | no | no | no |
| | RaiRat-M9 | QM | no | no | no | no |
| | DBBC | QM | no | no | no | no |
| | DBD | QM | no | no | no | no |
| | MOS-REG | TF | no | no | no | no |
| | FIC02T | PM/A/TF | no | no | no | no |
| **PP** | FIC01T | A/TF | yes | no | no | no |
| | ANALOG-ANOM | A | yes | no | no | no |
| | ANALOG | A | yes | no | no | yes |
| | ANALOG-MP | A | yes | no | yes | no |
| | ANALOG-SP | A | yes | no | yes | no |
| | MO-GP | TF | no | no | no | no |
| | MLR-T | TF | no | no | no | no |
| | MLR-RAN | TF | no | no | no | no |
| | MLR-RSN | TF | no | no | no | no |
| | MLR-ASW | TF | no | no | yes | no |
| | MLR-ASI | TF | no | no | no | no |
| | MLR-AAN | TF | no | no | no | no |
| | MLR-AAI | TF | no | no | no | no |
| | MLR-AAW | TF | no | no | yes | no |
| | MLR-PCA-ZTR | TF | no | no | no | yes |
| | MLR | TF | no | no | no | yes |
| | MLR-WT | TF/WT | no | no | no | yes |
| | WT-WG | WT | no | no | yes | yes |
| | SWG | TF | no | no | yes | yes |
| **WG** | SS-WG | WG | no | no | yes | no |
| | MARFI-BASIC | WG | no | no | yes | no |
| | MARFI-TAD | WG | no | no | yes | no |
| | MARFI-M3 | WG | no | no | yes | no |
| | GOMEZ-BASIC | WG | no | no | yes | no |
| | GOMEZ-TAD | WG | no | no | yes | no |

Table 2: Participating methods for temperature for the European experiment (exp1a). Techniques: A: analog; S: scaling; PM: parametric quantile mapping; QM: empirical quantile mapping; TF: regression-like transfer function; WT: weather typing; WG: weather generator. Multisite fitting: MS; EX: Explicitly modelled spatial structure; ST: Stochastic noise; PC: PCs used as predictors.

## 2.2 Validation measures

We now discuss the different validation measures on which the method comparison is based. All computations have been done in R and the codes are publicly available at Santander Meteorology Group (2016).

### 2.2.1 Correlations

Pairwise cross-correlations among all pairs of stations ($n \times \frac{n-1}{2}$ pairs, $n$ being the number of stations) are computed for the different target variables and seasons (Spearman for precipitation and Pearson for temperatures), and for experiments 1a and 1a-RCM ($n = 86$) and 1c ($n = 53$). For the temperature data the seasonal cycle of each data series is removed prior to correlation analysis by subtracting the climatological mean for each particular day of the year based on the whole analysis period 1979-2008. The mean is based on a circular moving average with a window width of 31 days centred around the target day. The precipitation data are used in their original form. In both cases, no detrending has been used. In addition to the visual comparison of correlation matrices we calculate the correlation matrix distance (CMD, Herdin et al., 2005). It measures the similarity between two correlation matrices and is defined as one minus the inner product of the normalised vectorized matrices. For matrices that are identical up to a scaling factor, the CMD is zero and for very different matrices, for which the associated vectors are orthogonal, the CMD is one.

Station correlograms are then derived by plotting the cross-correlation value for each station pair against their respective (great circle) geographical distances. As the resulting cloud of points may hinder a quick assessment of the dependency of the correlations on distance, we fitted reference curves to each correlogram using a local polynomial fit (“loess”, degree 2), allowing for a better comparability between downscaling methods and against the reference data. The local fit was preferred to other correlogram global fitting models commonly used in geostatistics (e.g. exponential or spherical; see e.g. Hengl, 2007)), as it does not require *a priori* assumptions about the structure of the correlations. It is therefore suitable for different kinds of correlation structures and flexible enough to allow for a direct comparison across different downscaling methods and experiments. As a measure for the overall behaviour of the fitted curves we then calculated correlation lengths (CL) for certain representative thresholds, as the abscissa of the point of intersection of the correlation threshold with the fitted line. We tested different thresholds, and the final values used are given in Table 3. The CL biases for the predictions were calculated as the difference of the CL for a given method and the CL of the observations (Table 4). This bias is a simple measure for the difference in the correlation structure between the predictions and the observations.

### 2.2.2 Spatial degrees of freedom

We determine the number of independent spatial degrees of freedom (DOF) that are associated with the observations and with the downscaling products. DOFs quantify the complexity of time- and space-dependent datasets and are based on the correlation or covariance matrix. In addition to describing the dependency

| **Var.** | **Exps. 1a and 1a-RCM** | **Exp. 1c** |
|---|---|---|
| Precip | 0.35 | 0.50 |
| Tmin | 0.50 | 0.65 |
| Tmax | 0.50 | 0.65 |

Table 3: Correlation thresholds used for calculating correlation lengths in the European experiments (1a and 1a-RCM) and in the German experiment (1c).

| | **Exps. 1a and 1a-RCM** | | | | | **Exp. 1c** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Var.** | *annual* | *DJF* | *JJA* | *MAM* | *SON* | *annual* | *DJF* | *JJA* | *MAM* | *SON* |
| Precip | 495 | 527 | 429 | 475 | 540 | 404 | 546 | 310 | 393 | 417 |
| Tmin | 741 | 822 | 647 | 771 | 653 | 569 | 695 | 462 | 541 | 475 |
| Tmax | 873 | 1005 | 785 | 893 | 870 | 698 | 788 | 697 | 653 | 668 |

Table 4: Correlation length (CL) values (in km) calculated from the correlograms of the reference station datasets (VALUE-ECA-86-v2 for experiments 1a and 1a-RCM, and VALUE-ECA-53-Germany-spatial-v1 for experiment 1c).

of the correlations on distance by a single number (CL) we thus also use a single number to capture a key property of the correlation matrices themselves and then calculate its biases.

One possible way to define complexity is to consider the eigenvalue spectrum of the covariance or correlation matrix. Consider a situation where the timeseries at all locations are perfectly correlated, which means there would be only one independent variable. In this case one PC (e.g. Hannachi et al., 2007) would explain all the variance, i.e. the first eigenvalue of the covariance matrix would be equal to the total variance and all other eigenvalues would be zero. If, in the other extreme case, the timeseries at all locations were independent, the eigenvalue spectrum would be completely flat, as no correlations between the station records could be exploited to construct any PCs that explain more variance than an individual station record. Roughly speaking, the steepness of the eigenvalue spectrum can thus be taken as an indication for the complexity of the data, with a steep (flat) spectrum being associated with low (high) complexity.

An alternative way to define the complexity of a space- and time-dependent field $\psi_i(t)$ is to consider the timeseries of the spatial sum of the squares of the values at the individual locations $i$, i.e.

$$E(t) = \sum_{i=1}^{n} \psi_i^2(t) \tag{1}$$

with $n$ being the number of locations. For independent variables $E(t)$ has a $\chi^2$-distribution with $N$ degrees of freedom, for dependent variables the distribution is well approximated by a $\chi^2$ distribution with fewer degrees of freedom. A useful measure of complexity is obtained by asking how many independent variables are needed to obtain approximately the same $\chi^2$ distribution, which is defined by its mean and variance, as for the timeseries of the sum of squares of the dependent variables.

This approach has been reviewed by Bretherton et al. (1999) who have shown that for normally-distributed PCs the $\chi^2$ and the eigenvalue approaches are equivalent if, as suggested in earlier studies, the degrees of freedom (DOF) are calculated from the eigenvalue spectrum by:

$$DOF = \frac{(\sum_i \lambda_i)^2}{\sum_i \lambda_i^2} \tag{2}$$

where $\lambda_i$ is the $i$-th eigenvalue and the summation is over all the eigenvalues.

In this paper we follow the computationally easier eigenvalue approach and calculate the independent spatial degrees of freedom according to equation 2. The normality assumption has been checked in the reference observation dataset VALUE-ECA-86-v2 (see Sec. 2.1), by comparing the empirical distribution function of each PC against the cumulative distribution function of the normal distribution using the Kolmogorov-Smirnov test, implemented in the function `ks.test` of the R package `stats` (R Core Team, 2018). All PCs were found to be indistinguishable from a normal distribution at the 5% significance level. The singular value decomposition implementation used in the (function `svd` in the R package `stats` R Core Team, 2018)) cannot handle missing values in the covariance matrix, and a few methods yielding missing values for all data in some stations did thus not yield results (this will be later indicated in the corresponding figure captions).

For consistency with the analysis of correlation lengths (Sec. 2.2.1), we base the DOFs on the eigenvalues of the correlation rather than the covariance matrix. In other words, we calculate the DOFs for standardised data, where the timeseries at each location have the same variance. The seasonal cycle is subtracted in the same way as for the correlation analysis. The DOFs for the observations, which are the reference for calculating DOF biases, are given in Table 5.

| | DJF | MAM | JJA | SON |
|---|---|---|---|---|
| precip | 30.02 | 41.51 | 48.64 | 36.05 |
| tmin | 6.56 | 7.66 | 9.43 | 8.86 |
| tmax | 5.65 | 6.56 | 7.55 | 6.91 |

Table 5: Degrees of freedom (DOF) for daily precipitation, minimum and maximum temperature from the VALUE-ECA-86-v2 observation dataset.

### 2.2.3 Joint threshold exceedances

The correlation-based analyses discussed above investigate the strength of linear relationships between the timeseries at different locations. However, for users of the downscaled data it may often be also relevant to know whether the probabilities for joint exceedance of a certain threshold at different locations are realistic in the downscaled data. Typical examples are the joint occurrence of precipitation or of heavy precipitation. For brevity, we restrict the analysis of such joint threshold exceedances to precipitation. This is the most challenging case since temperature fields are typically much smoother and spatially homogeneous. Therefore, we consider two typical cases: the wet day threshold of 1 mm/day and exceedance of thresholds for high precipitation, namely the local 90th percentile.

The most direct way to analyse the dependence between the data $X_i, X_j$ at a pair of stations $\{i, j\}$ for exceeding a threshold $x_{0i}$ at location $i$ and $x_{0j}$ at location $j$, is subtracting the product of marginals $P(x_i \geq x_{0i}) \cdot P(x_j \geq x_0 j)$ from the joint probability $P(x_i \geq x_{0i}, x_j \geq x_{0j})$. Their difference is zero only in case that $P(x_i \geq x_{0i})$ and $P(x_j \geq x_{0j})$ are totally independent and the larger the value, the more dependent they are. However, this difference would not only be influenced by the dependence for threshold exceedance, but also by the marginal probabilities at each of the stations, and is thus not a useful measure for the dependence itself.

A more suitable framework is based on the Mutual Information (MI) which measures the dependence between two random variables $X, Y$ and is unaffected by their marginal distributions. It is a standard approach in probability and information theory (see e.g. Hlinka et al., 2014), and for discrete random variables is defined as:

$$MI(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \cdot \log \left( \frac{p(x,y)}{p(x) \cdot p(y)} \right) \tag{3}$$

$MI$ is zero if the two events are independent, i.e. if $p(X,Y) = p(X) \cdot p(Y)$, non-negative ($MI(X,Y) \geq 0$) and symmetric ($MI(X,Y) = MI(Y,X)$).

In our analysis we consider the binary variables $\Psi_i$ at the locations $i$ which state whether the precipitation $x_i$ is above or below the threshold $x_{0i}$, i.e. $\psi_i = 1$ if $x_i \geq x_{0i}$ and $\psi_i = 0$ if $x_i < x_{0i}$. Following the definition above we then calculate for each pair of locations $i, j$ the MI for these binary variables

$$MI_{i,j} = MI(\Psi_i, \Psi_j) = \sum_{\psi_i \in [0,1]} \sum_{\psi_j \in [0,1]} (p(\psi_i, \psi_j) \cdot \log \left( \frac{p(\psi_i, \psi_j)}{p(\psi_i) \cdot p(\psi_j)} \right) \tag{4}$$

We calculate $MI$ for the dry-wet threshold $x_{0i} = 1$mm/d as well as for a high precipitation threshold defined as the 90th percentile ($P90_i$) of the observed daily precipitation (including dry days) at each station, i.e. $x_{0i} = P90_i$.

Following the methodology for correlograms (see section 2.2.1), we plot each $MI_{ij}$ against the distance of the locations $i, j$ and fit a degree-2 loess curve to the resulting plots. We then define MI thresholds for calculating the MI lengths (MILs) for observations and for the different downscaling methods. For the dry-wet binary variable based on $x_{0i} = 1$mm/d we use MI thresholds that depend on the experiment and season in order to obtain observed MILs that are similar (within a few kilometers) to the observed CLs, which makes it easier to assess whether MI yields information about the methods that is not already included in the CLs. The respective values are given in Table 6. For the high precipitation threshold $x_{0i} = P90_i$ we use a constant MI threshold of 0.1. Analogous to the correlation analysis MIL biases are calculated for the different downscaling methods, seasons and experiments by subtracting the respective observed MIL.

### 2.2.4 Regionalisation

Note that in this study, we apply the term regionalisation in the sense of spatial clustering, i.e. in the sense of finding regions with common variability. In order to achieve a regionalisation of the station data, orthogonally rotated (Varimax criterion, S-mode) principal component analysis (RCPA, e.g. Richman, 1986;

| **Experiments** | *annual* | *DJF* | *JJA* | *MAM* | *SON* |
|---|---|---|---|---|---|
| 1a, 1aRCM | 0.18 | 0.14 | 0.20 | 0.18 | 0.20 |
| 1c | 0.24 | 0.22 | 0.24 | 0.22 | 0.24 |

Table 6: MI thresholds used to calculate the MI lengths for the precipitation occurrence (1 mm threshold in the European experiments (1a and 1a-RCM) and the German experiment (1c).

Hannachi et al., 2007) is applied separately for each season to the correlation matrices calculated from detrended monthly timeseries.

The decision on the number of PCs to be rotated is based on the criterion that each retained PC has to be representative for at least one input variable, following Philipp et al. (2007). A rotated PC is considered representative for a given station if the loading of this PC at this station is larger than the loadings of the other PCs at this station by at least one standard deviation of all loadings at this station; additionally, this loading has to be statistically significant at the 5% level. Each station is assigned to the region (as defined by RPCA) for which it has the highest PC loading.

The number of PCs is determined from observations. Then the same number of PCs is used for the PCAs of the output from the downscaling methods. Following a standard approach the observed and the downscaled groupings are compared using the Adjusted Rand Index (ARI, Hubert and Arabie, 1985; Santos and Embrechts, 2009). The ARI is based on how pairs of objects, which in our case are pairs of locations, are classified as being either in the same or in different groups, which in our case are homogeneous regions. When comparing two classifications $U$ and $V$ there are four options for each pair and we denote the number of pairs for each option as:

- $a$ number of pairs that are in the same group in both classifications
- $b$ number of pairs that are in the same group in $U$ and in different groups in $V$
- $c$ number of pairs that are in the same group in $V$ and in different groups in $U$
- $d$ number of pairs that are in different groups in $U$ and in different groups in $V$

With these definitions, and $n$ being the number of objects, the ARI can be expressed as

$$ARI = \frac{\binom{n}{2}(a+d) - [(a+b)(a+c) \, + \, (c+d)(b+d)]}{\binom{n}{2}^2 \, - \, [(a+b)(a+c) \, + \, (c+d)(b+d)]} \, . \tag{5}$$

Its value increases with the agreement of the two classifications; 0 indicates no agreement and the maximum is 1 for identical classifications.

As already mentioned in Sec. 2.2.2 the singular value decomposition routine used for PCA cannot handle missing values, and therefore the regionalisation could not be calculated for a few methods.

# 3 Results

## 3.1 Example situation

Before we present the results of the statistical analyses we give an example for observed and downscaled precipitation on a specific day and for a few selected methods to illustrate the different characteristics of downscaling methods (Fig. 2). We chose 15. August 1998, because on this day there was frontal precipitation (over parts of the Scandinavia and the Baltic) as well as convective precipitation (over the Iberian Peninsula and parts of northern Italy). The distinction is based on the analysis of pressure charts and vertical temperature profiles (not shown).

The precipitation observations show low to medium values at most stations in Northern Spain and at one station in northern Italy, while the values in Scandinavia and the Baltic are medium to high. The ERA Interim reanalysis partly underestimates the amplitudes, shows a continuous rain band south of the Alps whereas only one station has recorded rainfall in this region, and does not simulate the convective precipitation in central Iberia. In comparison the RACMO regional model simulates the intensities in some regions better, for instance over Iberia and Scandinavia, but shows the well-known drizzle effect with light precipitation over large areas, as well as an unrealistic rain band north of the Alps and over parts of Germany and France. We note that satellite pictures showed convection over Germany, which however was not associated with precipitation. As expected the two quantile mapping methods EQMs (empirical) and RATY (parametric) inherit the partly unrealistic spatial structure from RACMO but change the specific values, with the EQMs intensities being in general closer to the observations than those from RATY.

The ANALOG-ANOM method captures well the fact that the convective precipitation only occurs at some locations and that the frontal precipitation is more homogeneous in space. The individual locations at which the convective precipitation occurs are partly different to the observations, which is an expected consequence of the stochastic nature of occurrence of convection. The values for the convective precipitation are close to the observed ones, whereas the intensity of the frontal precipitation is underestimated.

The MLR-RAN method (PP, multiple linear regression using large-scale predictors) unrealistically yields precipitation at all locations with the exception of some stations close to the eastern boundary of the analysis domain. For the stations where precipitation was observed the intensities are roughly in the right range. For the WT-WG method (weather generator conditioned on weather types) one can either plot individual realisations or the average over a simulated ensemble (100 realisations in this case). The individual realisations (not shown) have a much too low spatial coherency. This indicates that the random variability component, which is sampled individually at each location, is large compared to the fraction of variability that is conditional on the weather types. Here we show the conditioned component, i.e. the averaged values, which is as expected, too smooth, with precipitation occurring almost everywhere and values at the locations with observed precipitation being often too low.

In summary, the examples suggest that the methods that either inherit the spatial structure from an RCM (EQMs and RATY) or use observed spatial structures (ANLOG-ANOM) yield relatively realistic spatial patterns. In contrast conditioning precipitation at single sites on large-scale predictors (MLR-RAN, WT-WG) leads to fields that are too smooth when only the conditioned component is considered (MLR-RAN, averaged WT-WG), or not smooth enough when the stochastic component is added (individual realisations of WT-WG).

## 3.2 Correlations

Selected examples of pairwise cross-correlation matrices for winter (DJF) are displayed in Figs. 3a and 3b for precipitation and maximum temperature respectively. The 86 European stations (Fig. 1) are arranged so that station pairs with a small distance are near to the diagonal while distant pairs are near the upper-left corner. The geographic distances (measured along a great circle) are shown in the upper triangle in the first matrix of each panel, while the observed correlations are shown in the lower triangle.

In general, all methods are able to reproduce to some extent the correlation structure of both temperature and precipitation, with the exception of WT-WG. The WT-WG correlations shown are the average of the correlations for individual realisations (in contrast to Fig. 2 where correlations for ensemble-averaged values are shown), and despite the conditioning of the weather generator on weather types it yields almost uncorrelated values for all stations, regardless of their distance. This is explained by the weak conditioning imposed by the only predictor (SLP) used in this method, which explains only a very small fraction of the variance and results in an almost purely stochastic method (see also Gutiérrez et al. (2018)). The correlations of raw ERA-Interim and RACMO output are both in good overall agreement with the observations. However, the results for the different methods differ in detail. For instance, MLR-RAN systematically yields too high positive and negative correlations for distant station pairs, while EQMs and in particular ANALOG-ANOM reproduce most aspects of the structure well. The latter has the highest CMD value for precipitation (0.988) and maximum temperature (0.992).

We now investigate the dependency of correlations on distance more systematically by comparing correlograms and CL values. The former are shown for some example methods in Fig. 4 for the European station network (experiments 1a and 1aRCM) and in Fig. 5 for the high-density German network (experiment 1c). In addition to the actual correlations these figures include the fitted curves and the CLs (vertical lines). As expected the observed correlations (upper-left panels) decline with distance and for the European dataset level off around zero. The fact that the correlations show approximately an exponential decrease in Fig. 4 but a more linear decrease in Fig. 5 is due to the different size of the analysis domains. In experiment 1c there are some missing CL values for temperatures, because due to the small analysis domain and the smooth topography the temperature records are highly correlated for all station pairs and in some cases the fitted line is therefore above the corresponding correlation threshold (0.65, Table 3) for all distances. In contrast precipitation has a higher degree of spatial heterogeneity and CLs are obtained in all cases.

For the European data (Fig. 4) ERA-Interim tends to slightly overestimate the correlations in both seasons and reproduces the observed slight difference between the seasons. RACMO has values closer to reality, but does not capture the observed seasonal difference. Both MOS methods (EQMs-R and Ratyetal-M8-R) further reduce the correlations compared to the raw RCM but to a different extent, and the lack of a seasonal difference remains. As expected, the analog method (ANALOG-ANOM), which selects an entire analog field reproduces the observed correlations. The PP example method (MLR-RAN), which uses large-scale predictors, overestimates correlations. As already shown in Fig. 3a the weather generator conditioned on weather types (WT-WG) strongly underestimates correlations when individual realisations are considered. For the ensemble average (dashed lines) the correlations are too high in winter and still substantially too low in summer. The deficiencies of this method have also been reported in Gutiérrez et al. (2018).

It can be seen in Fig. 5 that in Germany and on the shorter distances, which are resolved well by the high-density network, the observed seasonal differences are larger than in the European case, with higher correlations in winter. All example methods do now also show a seasonal difference. As in the European case ERAINT overestimates correlations. The MOS-corrected ERA-Interim precipitation (EQM-R) leads to fairly realistic correlations, as does one of the PP methods (DSCLIM-D), while the other ones either overestimate (GLM-DET) or underestimate (GLM-BN-DET) correlations. As explained in section 2.1, the latter is an extension of the former, explicitly including a model for spatial dependence (based on probabilistic networks).

We now look at the full set of methods with respect to the precipitation CL bias for the European (Fig. 6) and German datasets (Fig. 7). In Fig. 6 ERAINT has a positive CL bias, which gets reduced when the reanalysis is dynamically downscaled with RACMO, as already seen in the previous figures. Most deterministic MOS methods do reduce the bias both in the reanalysis-driven (*-E) and RACMO-driven (*-R) case, with the former still having higher CLs than the latter, as for the raw numerical models. Many MOS methods that are based on quantile mapping have very low CL biases, while some of the scaling approaches (e.g. Ratyetal-M7) have slightly higher biases. Consistent with the previous plots, the stochastic methods (MOS-GLM, VGLMGAMMA) have substantial negative CL biases for the individual realisations. The bias for the ensemble mean is positive for MOS-GLM, while it is negative for VGLMGAMMA, suggesting that for the latter the distributions are not constrained closely enough by the predictands.

The PP methods in Fig. 6 show a wide range of positive and negative biases. Positive biases occur for regression methods with large-scale predictors (MLR-RAN, MLR-RSN, MLR-ASI, GLM-DET) because the predictors for different stations are similar (e.g. PCs from ERA-Interim fields). The FIC01P method, which is a combination of an analog method and postprocessing using a transfer function, has also a positive bias. In contrast, negative biases are visible for methods that use local predictors, e.g. information taken from the gridcell covering the target station, for instance some of the linear models (GLM, GLM-WT, GLM-P) and the 'multi-objective genetic programming method' (MO-WT). The ANALOG method, which is based on regional-scale predictors shows a negative CL bias. Individual realisations of some stochastic methods (ANALOG-M, ANALOG-SP, GLM-P) have also negative biases. Biases close to zero are achieved with one analog method (ANALOG-ANOM) and a regression method with noise added (MLR-ASW).

When the CL biases on shorter distances are considered (Fig. 7) the raw ERA-Interim precipitation shows again a positive bias, while biases close to zero are obtained for MOS methods based on quantile mapping. For the PP methods the positive biases of regression methods using large-scale predictors and the negative bias for those using local predictors remain. The ANALOG method is now almost bias-free, in contrast to the European case. The reason is that the predictors are neither global, nor completely local, but based on the division of the whole domain in a number of sub-domains with each containing several stations. The selection of analog dates is common for all stations within a sub-domain, thus guaranteeing the spatial consistency within sub-domains, whereas different dates can be chosen for different sub-domains. As Germany lies within one sub-domain and Europe covers several subdomains the CL bias is close to zero for experiment 1c (sampling effects remain) and negative for experiment 1a. The second method that is bias-free is a hybrid method (DSCLIM-D) which combines a weather type based transfer function and an analog approach.

For the European dataset we also consider the CL bias for minimum and maximum temperature (Figs. 8 and 9). As temperature fields are smoother than precipitation fields, we use a correlation threshold of 0.5 rather than 0.35, which was used for the European precipitation data. The results for minimum and maximum temperatures are very similar. The MOS results are fundamentally different from the precipitation case. While for precipitation many MOS methods did reduce the CL bias relative to the raw models (both for ERAINT and RACMO), for temperature there is for almost all MOS methods no reduction of the positive model bias. The reason might be that precipitation is an intermittent process for which debiasing the marginal distribution affects correlations more strongly than for the continuous temperature timeseries. The high biases for CDFt-E and MOS-REG-R need further investigation. The CDFt method was also found to behave differently to other MOS methods with respect to the temporal correlation between predictions and observations (Gutiérrez et al., 2018), trends (Maraun et al., 2018) and extreme events (Hertig et al., 2018). We note that this method is different from the other MOS techniques in the sense that it also uses the predictand distribution in the validation period (see Gutiérrez et al. (2018), Appendix A.1 for the full method description), which may lead to a high sampling variability in our experimental setup. The CDFt data passed our standard quality test, but the correlation vs. distance plots for CDFt for maximum and minimum temperatures and experiment 1a showed an unusual behaviour with no clear link between correlations and distance, and thus a technical error for downscaled temperatures using CDFt-E cannot be ruled out.

As for precipitation the PP methods show again in general higher biases than the MOS methods, and some analog methods perform well, whereas others do not. A noticeable difference is the smaller number of methods with negative biases for temperature. Although the set of methods is not identical, there are some methods used for both predictor variables that have large negative biases for precipitation but small biases for temperature, namely ANALOG-SP and MO-GP. A potential reason is that for those methods the predictors constrain temperature better than precipitation.

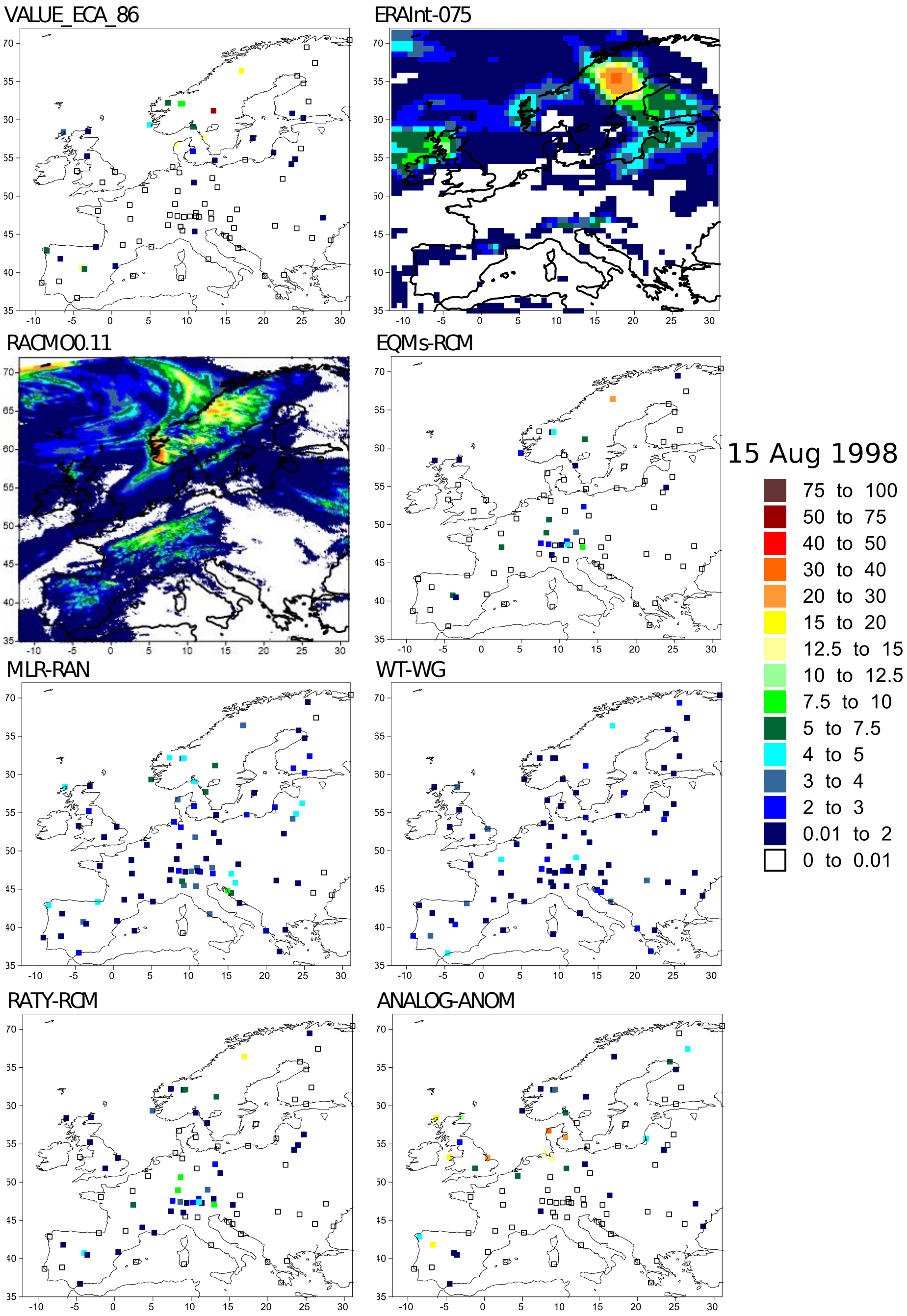
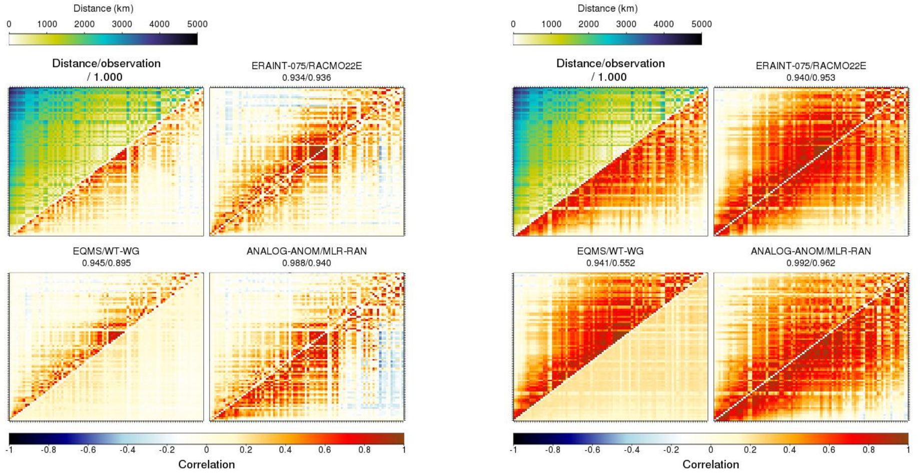


Figure 2: Observed (VALUE-ECA-86-v2, top-left panel) and downscaled precipitation on 15. August 1998 (mm/d). The second and third panels (from top to bottom, and left to right) show the 24h accumulated precipitation from the ERA-Interim reanalysis (ERAint-075 panel) and from the RACMO RCM (0.11 degree horizontal resolution, RACMO 0.11 panel) driven by ERA-Interim The downscaling methods are labelled by their codes (Table 1), with the "-RCM" suffix indicating MOS methods used in experiment 1a-RCM.



(a) Daily DJF precipitation (Spearman's $\rho$ correlation coefficient)

(b) Daily DJF maximum temperature (Pearson's $r$ correlation coefficient)

Figure 3: Pairwise cross-correlation matrices for winter for the 86 locations of the VALUE-ECA-86-v2 dataset. In each panel, the first matrix represents the geographic distances between pairs of stations (above the diagonal) and the correlations of the observations (below the diagonal). The remaining matrices display the correlations for two different methods indicated by the panel titles with the values for the first (second) method given above (below) the diagonal. The number under the method names is one minus the correlation matrix distance between the method and the observation correlation matrices.

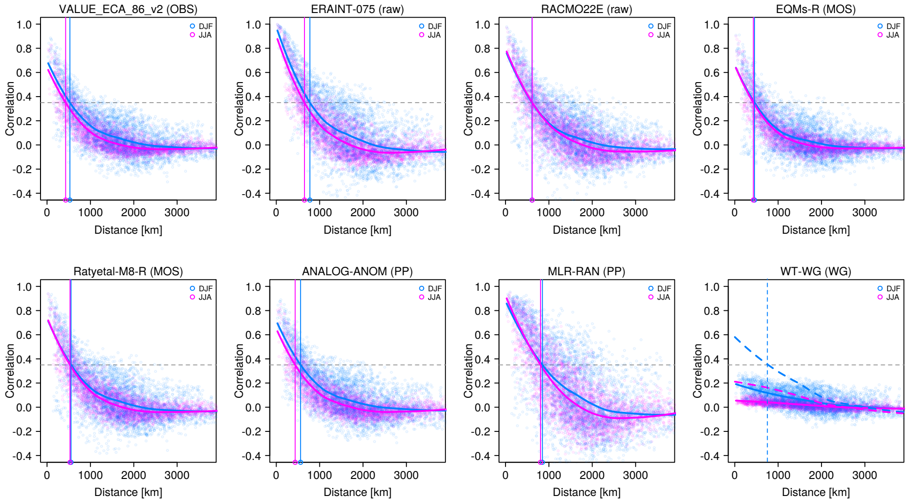


Figure 4: Correlograms for daily precipitation for JJA and DJF showing correlations of the timeseries for each pair of stations against their geographical distances (European experiment, exp1a). For the stochastic WT-WG method the fitted curves of the *averaged* option and the corresponding CL value are indicated by dashed lines (individual values are omitted for clarity).

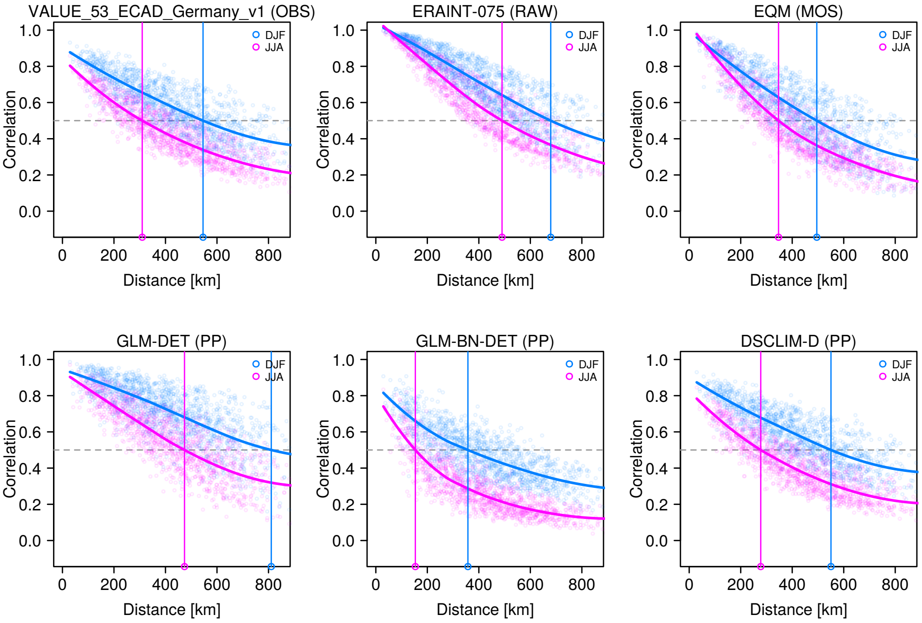


Figure 5: Same as Fig. 4 but for selected methods used in the German experiment (exp1c). The correlations for the reference observations (VALUE-ECA-53-Germany-spatial-v1) are shown in the upper left panel.

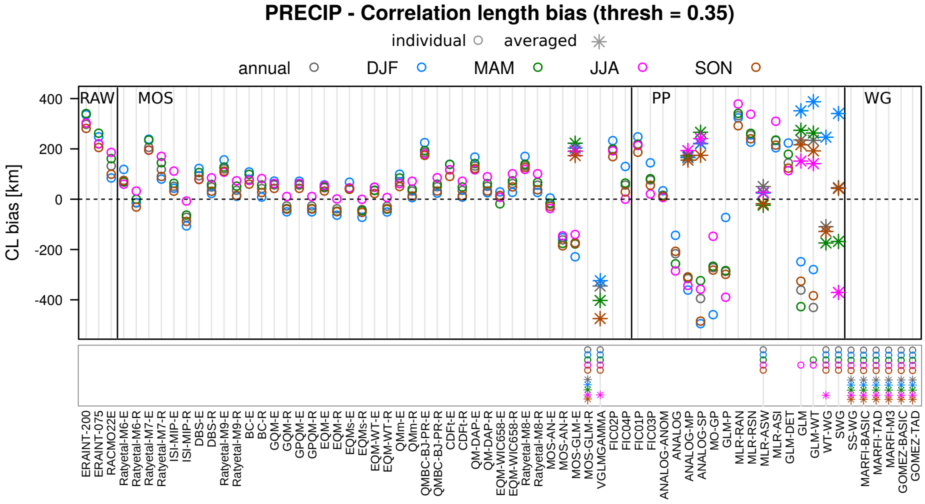


Figure 6: Correlation length (CL) biases for daily precipitation from the downscaling methods tested in experiments 1a (suffix $-E$ for MOS methods) and 1a-RCM (suffix $-R$) with respect to the reference values based on the VALUE-ECA-86-v2 dataset (Table 4). For the stochastic methods, the results of both the member-averaged (asterisks) and individual (circles) approaches are shown. The box in the lower part of the figure shows the seasons/approaches for which the CL cannot be calculated due to very low correlations.

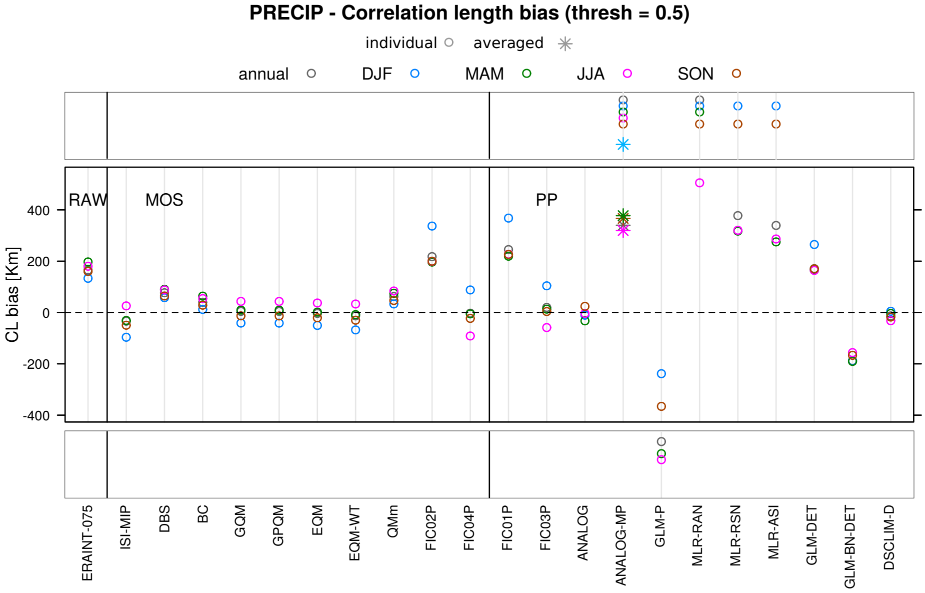


Figure 7: Same as Fig. 6 but for the German experiment (exp1c). The boxes in the lower/upper part of the figure show the seasons and approaches for which CL distance cannot be calculated. Upper box: the fitted correlogram line is entirely above the threshold. Lower box: the fitted correlogram line is entirely below the threshold.

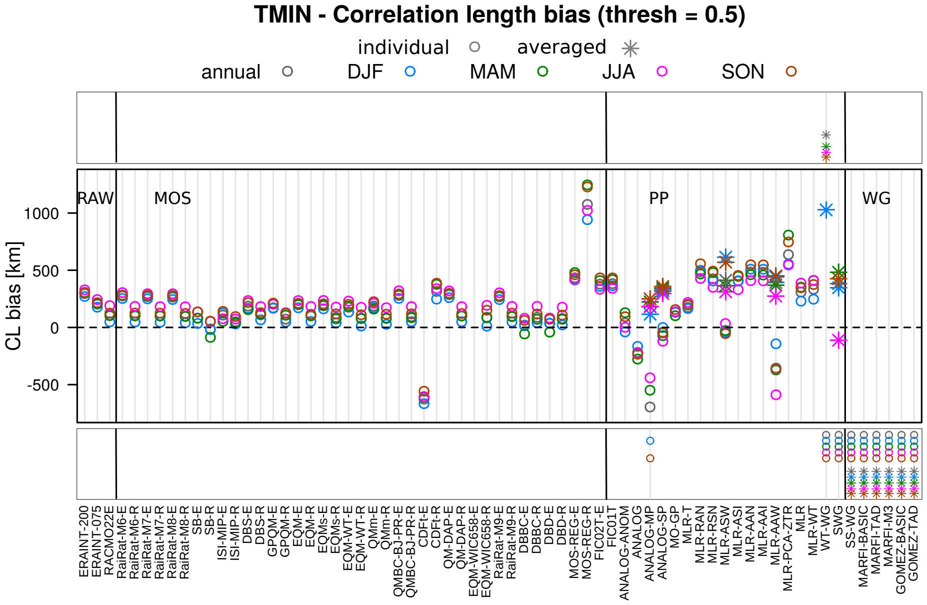


Figure 8: Same as Fig. 6 but for minimum temperature.

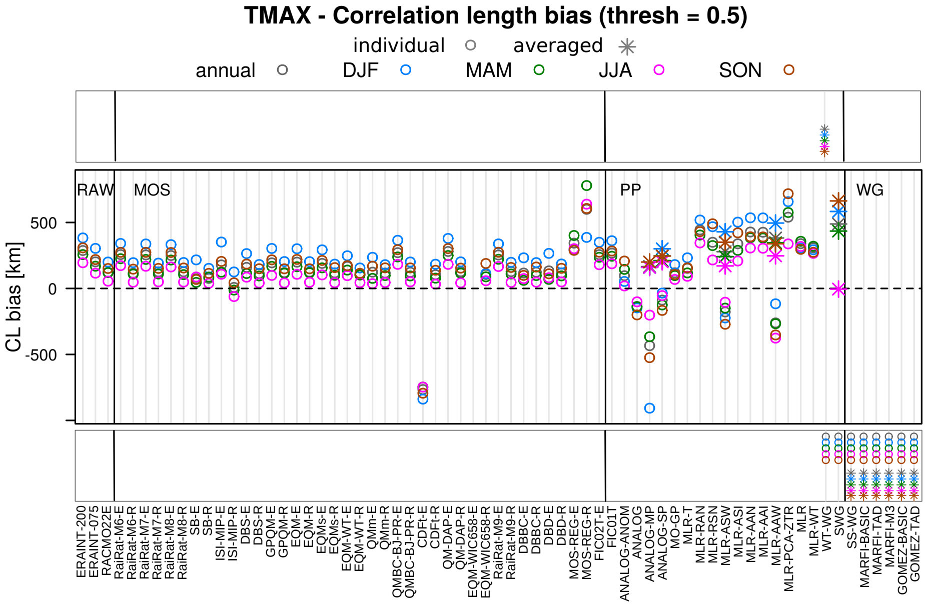


Figure 9: Same as Fig. 8, but for maximum temperature.

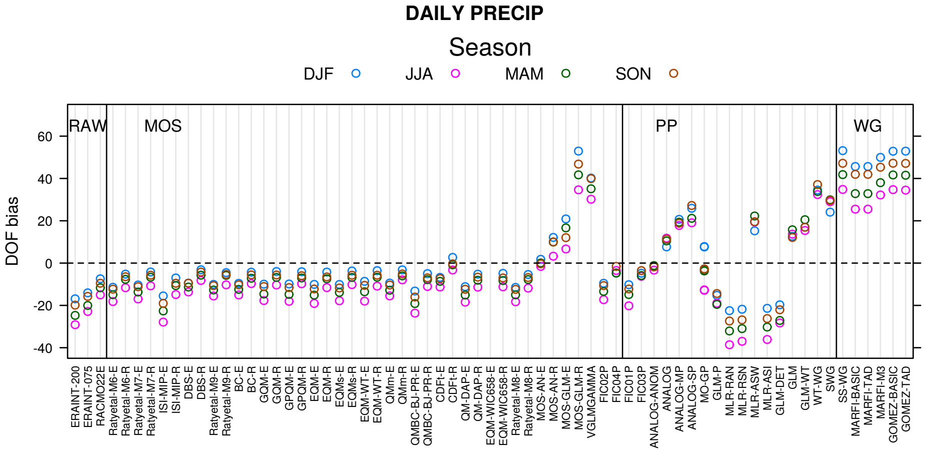


Figure 10: Bias of the spatial degrees of freedom (DOF) for daily precipitation from the methods included in the European experiments (exp1a and exp1a-RCM).

## 3.3 Spatial degrees of freedom

The DOF biases for precipitation, which express differences in the dimensionality of the fields, are shown in Fig. 10. Almost all MOS methods have a negative bias and thus underestimate complexity. The underestimation is strongest in summer, where convective, and thus small-scale, precipitation is more important than in the other seasons. Compared to the raw model results, most MOS methods reduce the absolute bias. The exception are some of the stochastic methods (MOS-GLM, VGLMGAMMA), which strongly overestimate complexity. The MLR-based PP methods also underestimate complexity, whereas some of the analog methods have a small bias and others overestimate it. The weather generators show a strong overestimation.

The DOF biases for temperature are shown in Fig. 11. For almost all downscaling methods they are substantially smaller than for precipitation, with many MOS and some PP methods leading to biases smaller than 2. The exception are some WG methods (SS-WG, GOMEZ-BASIC, GOMEZ-TAD), which show biases of up to 40. During summer and autumn the DOF biases for minimum temperature are larger than those for maximum temperature. In contrast to the precipitation case the biases for the MOS-corrected models are very similar to those of the raw models.

Most methods with a positive (negative) CL bias, i.e. those for which correlations drop too slowly (too quickly), have a negative (positive) DOF bias. One clear exception is CDFt-E for temperature, which is in line with other MOS methods with respect to the underestimation of the DOFs, but as mentioned in Sec. 3.2 has a large negative CL bias, which may be due to technical errors. We note that reordering the stations would not affect the DOFs, but would lead to erroneous correlograms if not taken into account when calculating the distances between station pairs. There are also some MOS methods that have a slightly positive CL bias despite their negative DOF bias.

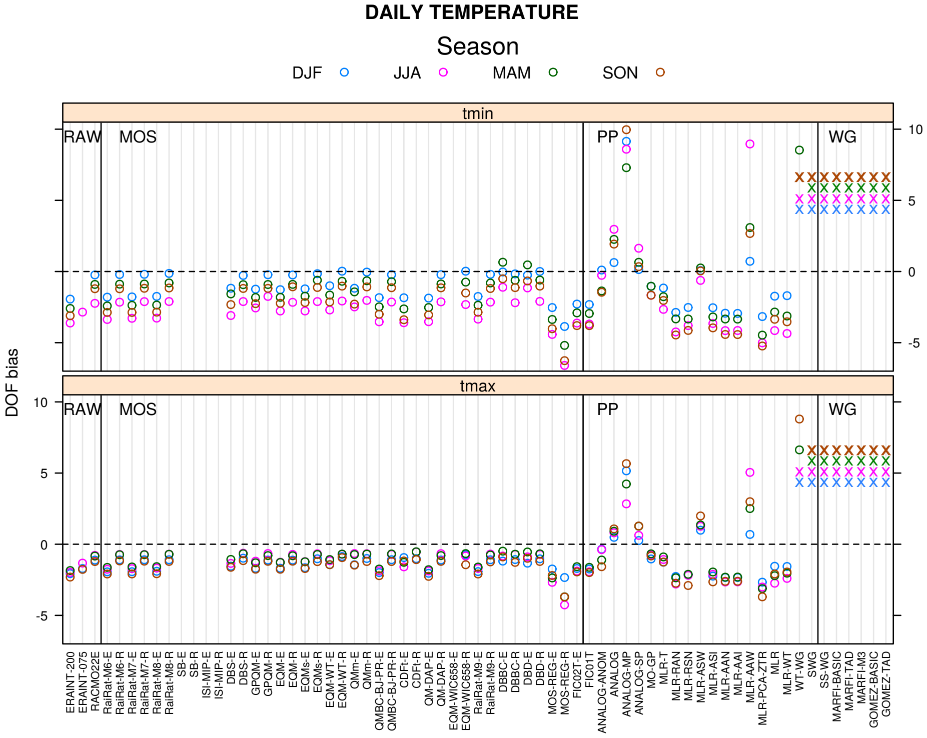


Figure 11: Same as Fig. 10, but for daily temperature. The methods marked with a cross (×, coloured according to the season) are out of range with positive bias of more than 10 degrees of freedom. The methods without results are those having missing values in the covariance matrix (see Sect. 2.2.2).

### 3.4 Joint threshold exceedances

The methodology for the joint threshold exceedances analysis is very similar to that for correlation (see Sec. 2.2.3), and we therefore do not show the MI matrices and MI vs. distance diagrams. The characteristic MI lengths for the reference observations exceeding the wet day threshold are presented in Tab. 7 and for exceeding the local 90th percentile in Table 8. As in the case of the correlograms, lower MIL values indicate a faster loss of mutual dependence as a function of distance, while higher MIL values indicate a stronger dependence between stations. For both thresholds there is a marked seasonal dependence, with the minimum in summer and the maximum in winter. For the 90th percentile autumn values are also high. The MILs obtained from the European and the German observational datasets were similar (Table 7).

The high-density German dataset is better suited than the European dataset for calculating MILs for both thresholds, as it has a larger number of station pairs within the distance ranges relevant for calculating the MILs for both thresholds, and thus provides more robust results. We therefore restrict the MIL analysis to experiment 1c. This has the additional advantage that we avoid a potential loss of robustness in the summer results arising from locations with no precipitation for the whole season, which may occur in some parts of Southern Europe. The biases for the wet day threshold with respect to the observed reference values are shown in Fig. 12 and for the 90th percentile threshold in Fig. 13.

For the wet day threshold all MOS methods slightly overestimate the dependence. The exceptions are FIC02P, which strongly overestimates it, and FIC04P, which in most seasons slightly underestimates it. All MOS methods but FIC02P reduce the bias compared to the raw reanalysis data. Among the PP methods ANALOG and DSCLIM-D (which contains an analog step) are bias-free apart from sampling effects, and the individual realisations of ANALOG-MP has also a very low bias. The MLR methods overestimate the dependence, whereas GLM-P strongly underestimate it.

The different downscaling methods perform similarly with respect to the MIL biases for the wet day threshold and to the CL biases (Fig. 7). Both show a bias reduction by most MOS methods, and the same sign and relative size of the bias for both quantities. Too strong (weak) correlations of the timeseries are thus associated with too high (low) dependences of the occurrence of wet or dry days.

The overall picture is different for the 90th percentile threshold. Almost all MOS methods show the same overestimation of dependence as the raw reanalysis data. In the PP group the analog methods and GLM-BN-DET and DSCLIM-D have very low biases, whereas the MLR methods very strongly overestimate dependences for heavy precipitation.

Accepted manuscript. International Journal of Climatology (Jan 2019). DOI: 10.1002/joc.6024

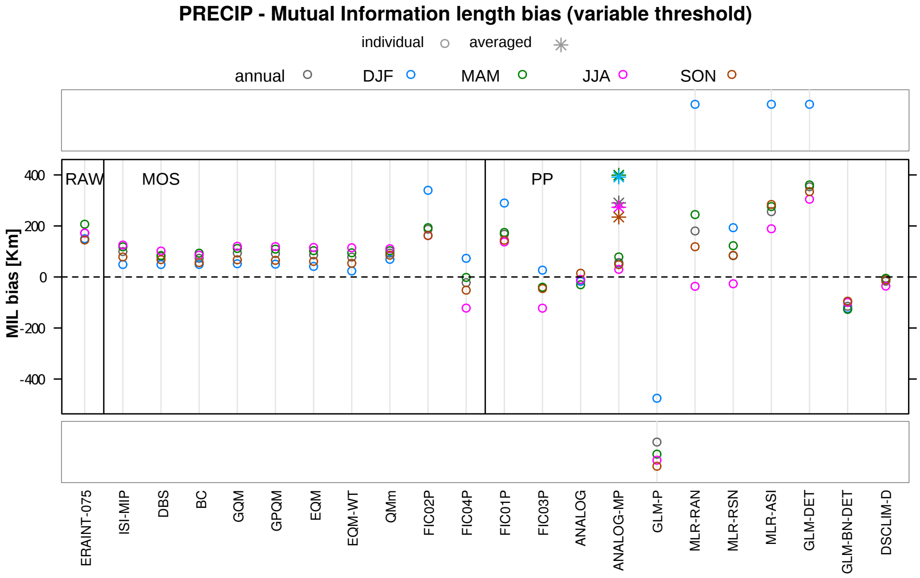


Figure 12: Mutual Information Length (MIL) biases for the exceedance of the wet day precipitation threshold, obtained from experiment 1c with respect to the values from the reference observations (VALUE-53-ECAD-Germany-v1 dataset, Table 7). Values in the boxes in the upper and lower part of the figure indicate methods for which the MI Length value cannot be calculated due to the MI values being to high or too low (as for correlations in Fig. 7).

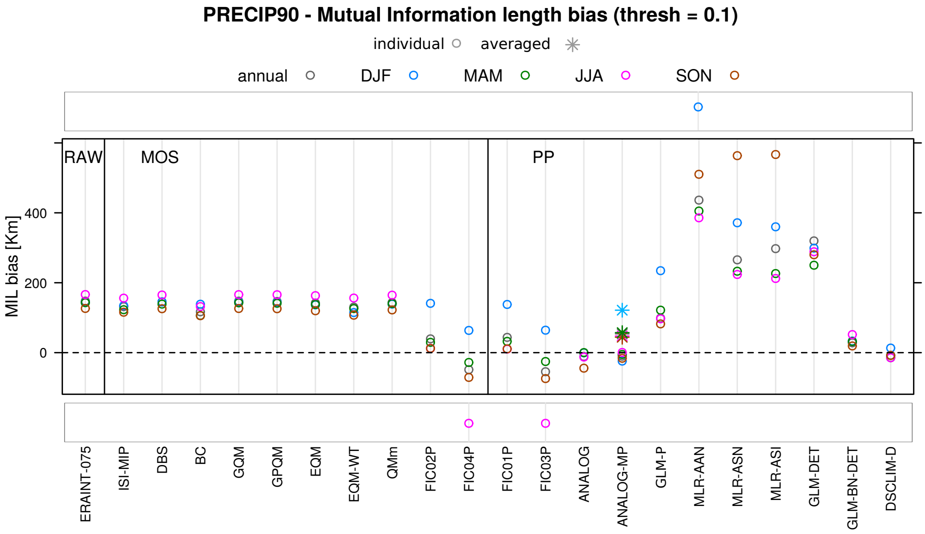


Figure 13: Same as Fig. 12 but for the exceedance of the $90^{th}$ percentile of daily precipitation obtained from experiment 1c.

| **Experiments** | *annual* | *DJF* | *JJA* | *MAM* | *SON* |
|---:|---:|---:|---:|---:|---:|
| 1a, 1aRCM | 359 | 554 | 216 | 324 | 340 |
| 1c | 338 | 528 | 256 | 360 | 345 |

Table 7: Mutual Information Length (MIL) values (in km) calculated for exceedance of the wet day threshold of daily precipitation in the reference station datasets (VALUE-ECA-86-v2 for experiments 1a and 1a-RCM and VALUE-ECA-53-Germany-spatial-v1 for experiment 1c), using the MI thresholds displayed in Tab. 6.

| *annual* | *DJF* | *JJA* | *MAM* | *SON* |
|---:|---:|---:|---:|---:|
| 191 | 284 | 109 | 183 | 234 |

Table 8: Mutual Information Length (MIL) values (in km) calculated for exceedance of the $90^{th}$ percentile of daily precipitation in the reference station dataset of experiment 1c (VALUE-ECA-53-Germany-spatial-v1, using a fixed threshold of 0.1 for all seasons. Note that only the experiment 1c (German dataset) has been used in this case as reference (see Sec. 3.4).

### 3.5 Regionalisation

The number of PCs retained for rotation is shown in Table 9 along with the cumulative fraction of variance explained for the observed daily precipitation, minimum and maximum temperature at the 86 European stations. As expected a higher number of PCs is needed to explain a certain fraction of the variability of precipitation compared to temperature, as the spatial patterns of the former contain more small-scale structures. Also, more PCs are needed to represent precipitation well in summer and spring than in autumn and winter, due to the higher contribution of small-scale, convective precipitation in the former seasons, and the dominance of large-scale, stratiform precipitation in the latter. The fact that the retained PCs do not explain all the variance in the datasets is one of the potential reasons for differences between the rotated EOFs in the observations and the downscaling results.

| **Var.** | *DJF* | *MAM* | *JJA* | *SON* |
|---|---|---|---|---|
| Precip | 13 (78.3) | 15 (71.4) | 19 (71.4) | 13 (71.8) |
| Tmin | 6 (85.8) | 6 (85.1) | 6 (82.3) | 6 (81.4) |
| Tmax | 6 (87.2) | 6 (87.3) | 6 (86.5) | 5 (82.0) |

Table 9: Number of principal components retained for rotation and cumulative variance (in parentheses, %) for precipitation, minimum and maximum temperature at the 86 stations of the ECA-VALUE-86-v2 observation dataset.

For temperature 5-6 PCs are retained and thus 5-6 regions are identified. The regions for maximum temperature in the different seasons are shown in Fig. 14. Europe is divided roughly intp northern Europe, north-western Europe, south-western Europe, central and southern Europe, eastern Europe, and south-eastern Europe. The boundaries between the regions are to some extent seasonally dependent. They are also not always simply connected geographical regions, as for instance in autumn and spring one station in northern Italy is grouped together with the south-western stations, or in winter the UK, Germany and the Alpine regions contain stations associated with different rotated PCs. Similar regions are found for minimum temperature, but there are also some differences, for instance a distinct central alpine region for minimum temperature in winter (not shown).

Fig. 15 shows the ARI for minimum and maximum temperatures, which is used as performance measure to judge the ability of the downscaling methods to capture the observed regions of similar temperature variations. It can be seen that the single-site WG based methods (GOMEZ-BASIC, GOMEZ-TAD, MARFI-BASIC, MARFI-TAD, MARFI-M3, SS-WG) are not able to reproduce the regions at all due to the generation of synthetic time series at one specific location without considering spatial relationships. WG methods that include atmospheric covariates (WT-WG, SWG) perform somewhat better by indirectly incorporating spatial information carried by the covariates. There is no systematic difference between MOS and PP methods. The ARI mostly lies between about 0.3 and 0.9 and varies more between seasons than between methods. The best performance is achieved for spring to autumn, whereas in winter the lowest ARI values are systematically attained. The lower performance in winter might partly be explained by region-specific phenomena (for instance inversion), which

Accepted manuscript. International Journal of Climatology (Jan 2019). DOI: 10.1002/joc.6024

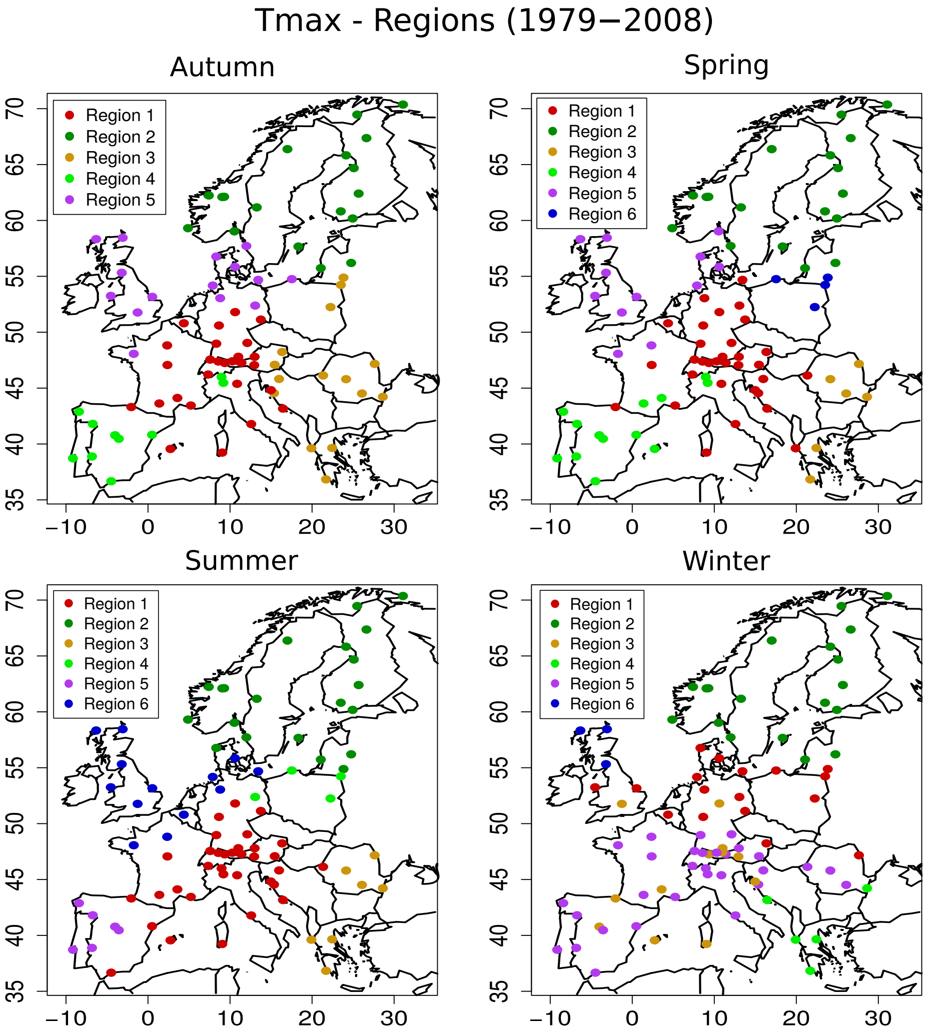


Figure 14: Regions derived from rotated PCA of seasonally detrended monthly maximum temperatures in the period 1978-2008, considering the 86 stations of the VALUE-ECA-86-v2 observational dataset.

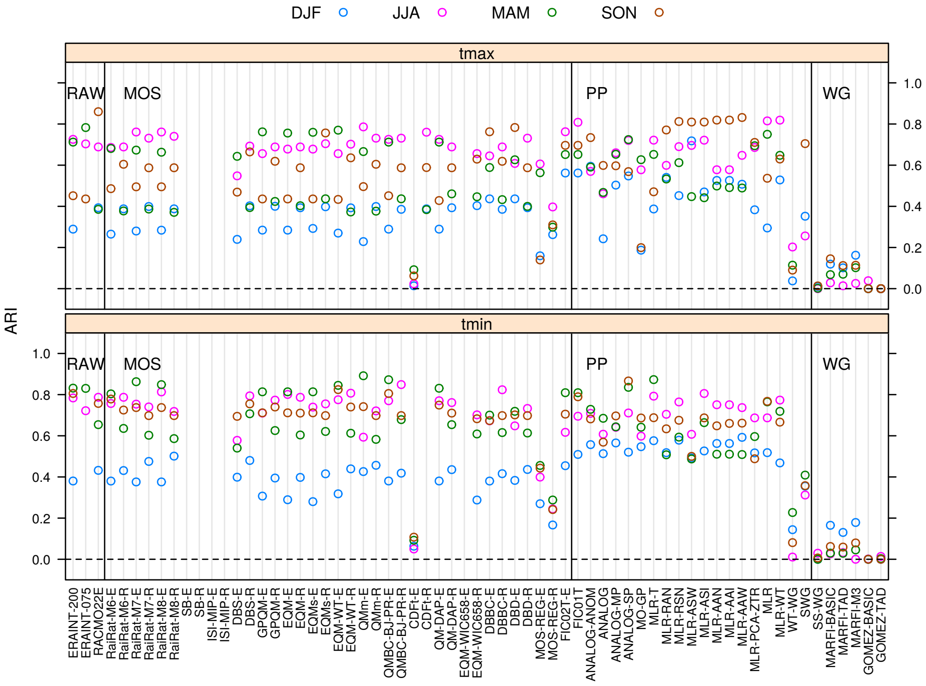


Figure 15: Adjusted Rand Index (ARI) for minimum (top) and maximum (bottom) temperatures obtained from the European experiments (exp1a and exp1a-RCM). ARI measures the agreement between the regionalisations for the observations (VALUE-ECA-86-v2 stations) and the downscaling output, ranging from 0 (no agreement) to 1 (perfect agreement). The methods without results are those having some missing values in the covariance matrix, as indicated in Section 2.2.2.

are not adequately captured by the downscaling methods. The ARI for analog methods, which by construction lead to a realistic spatial structure of the daily fields, is not higher than for many other methods. The monthly means to which the rotated PCA is applied, might be somewhat different from the true monthly means, and the questions arises to what extent the results of the rotated PCA describe robust statistical properties, and to what extent they might be influenced by the individual realisations. The ARI for precipitation is shown in Fig. 16 and lies between about 0.2 and 0.6, but with no seasonal structure to it. Like for temperature, WGs are not able to map the regions and no superior performance of multi-site methods arises (not shown).

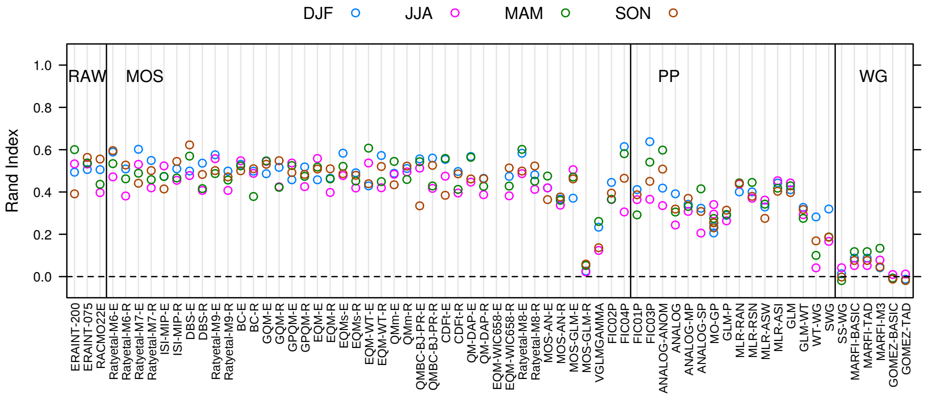


Figure 16: Same as Fig. 15, but for precipitation.

# 4 Summary and conclusions

We have evaluated the spatial variability of the output from over 40 downscaling methods for the period 1979-2008 at a European-wide network of 86 stations, and at a high-resolution network of 53 stations in Germany. Predictors for the PP methods and boundary conditions for the RACMO regional model have been taken from the ERA-Interim reanalysis. MOS methods have been applied to the reanalysis as well as to the RACMO output. We have analysed the dependency of correlations of daily temperature and precipitation series at station pairs on the distance between the stations. For the European dataset we have also investigated the complexity of the downscaled data by calculating the number of independent spatial degrees of freedom. For daily precipitation at the German network we have additionally evaluated the dependency of the joint exceedance of the wet day threshold and of the local 90th percentile on the distance between the stations. Finally we have investigated regional patterns of European monthly precipitation and temperature obtained from rotated principal component analysis.

The results for correlation lengths and degrees of freedom based on the European network are summarised in Fig. 17. Findings related to joint threshold exceedances are not included in the figure because they are based on the German predictand data and a different set of methods. Results from the regionalisation are not included because they are derived from monthly rather than daily data. The figure shows the relative bias calculated as the ratio of the bias and the observed value for the correlation lengths or the degrees of freedom. This normalisation makes it easier to compare the values for differen seasons, and for correlation lengths and degrees of freedom. For the bias in the degrees of freedom we have swapped the sign because a bias in correlation lengths is usually associated with a bias of the opposite sign in the degrees of freedom. The summary figure and the detailed results presented earlier show that there is a very large spread in how well the different downscaling methods represent the characteristics of the observations, ranging from close to reality to very unrealistic.

For all three predictand variables the raw models have positive biases in correlation length and negative biases in the number of degrees of freedom. The biases for the RACMO model are smaller than those for the reanalysis, which demonstrates the benefit of the explicit representation of smaller spatial scales. It is likely that these biases are not fully due to model deficiencies because the spatial scales of the data are different. Observations averaged over the gridcells can have higher correlations between two locations than local values, and the number of degrees of freedom of spatial averages can be lower. Likewise the dependence of the exceedance of thresholds at different locations, for which the models showed a positive bias, might be higher for area means than for local values. Nevertheless the biases represent actual errors if the gridcell values are used as direct estimates for local values.

As can be seen in Fig. 17 most MOS methods substantially reduce the positive biases in correlation length for precipitation, whereas there is no clear improvement for temperature. This difference might be due to the fact that precipitation is an intermittent process with many zero values, for which correcting the simulated marginal distribution affects correlations and threshold exeedances more strongly than for the continuous temperature timeseries. The

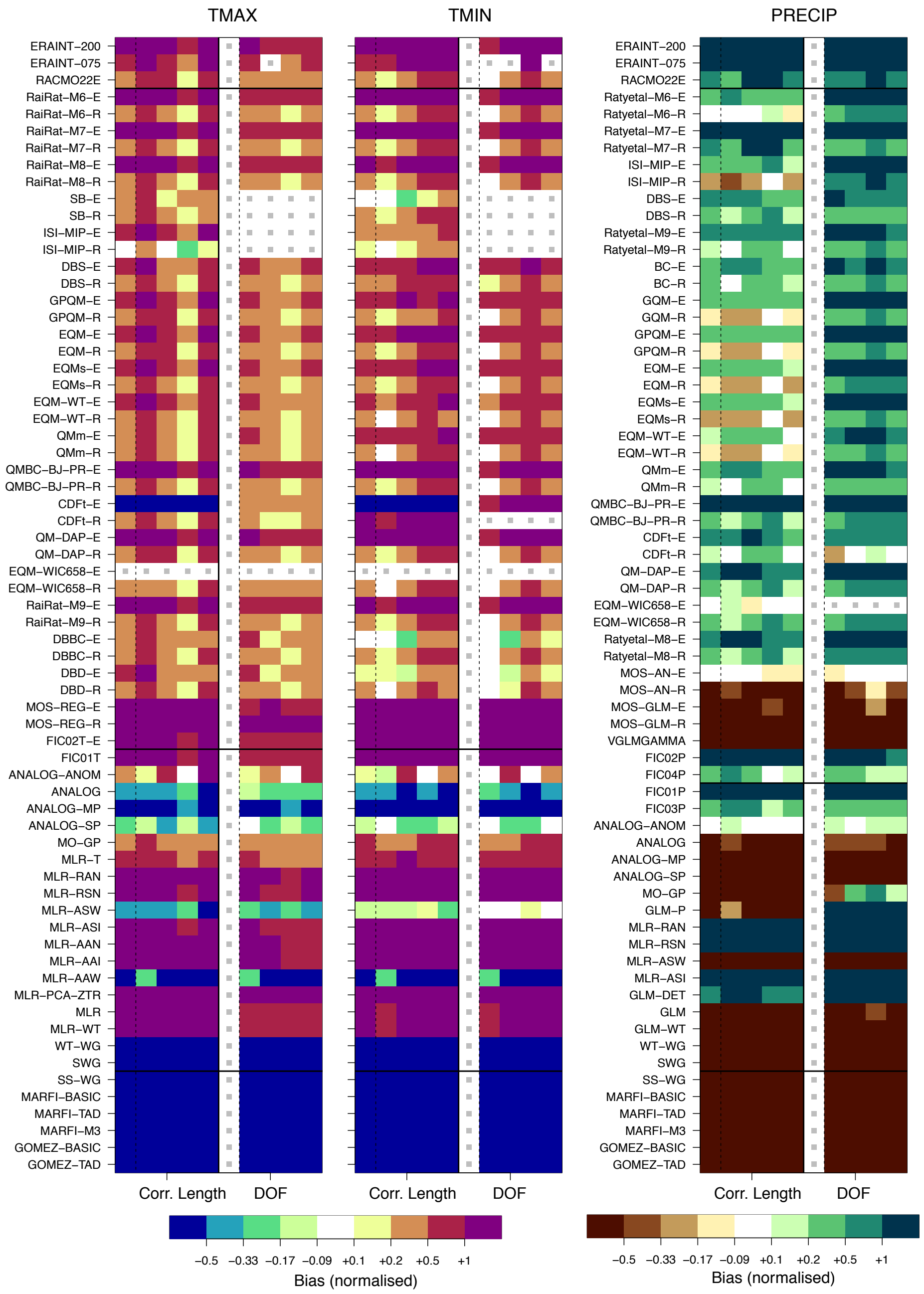


Figure 17: Relative biases in correlation length and independent spatial degrees of freedom (with sign swapped) based on the European network for daily maximum and minimum temperature, and precipitation. The columns indicate the seasons (annual, DJF, MAM, JJA, SON). For the degrees of freedom no annual values have been calculated.

bias in the degrees of freedom is not reduced as much. It was also shown that MOS methods reduce the positive bias in the dependence for wet threshold exceedance, but not for the exceedance of the 90th percentile of local daily precipitation. High-resolution, convection-permitting RCMs combined with MOS might represent the spatial characteristics of heavy precipitation events considerably better, but are still not widely used in climate change studies because they are computationally expensive (Prein et al., 2015). The value added by the regional model is still present after the MOS postprocessing (methods with suffix '-R' perform better than those with suffix '-E'). For temperature the seasonality of the biases is similar for the raw model and for the MOS-corrected values. The biases in correlation length and in the degrees of freedom are for minimum temperature in general slightly higher than those for maximum temperature.

Fig. 17 and the specific findings in the main section also show that for all predictand variables MOS methods perform in general better than PP methods, however with some noteworthy exceptions. Deterministic PP methods that are based on multiple linear regression and large-scale predictors tend to strongly overestimate spatial correlations and also dependences of threshold exceedances, while some other PP methods, for instance MO-GP and GLM-P, which use local predictors, underestimate the joint variability between the stations, in particular for precipitation. Given the different predictors used for different PP methods it is possible that the results are strongly influenced by the predictor choice rather than by the structure of the statistical model. Analog methods yield, as expected, realistic spatial characteristics apart from sampling effects if a common analog date is selected for all locations, whereas they underestimate links between the stations if analogs are defined locally. In addition to the analog methods the GLM-BN-DET method, which explicitly models spatial dependence, performes very well with respect to the joint exceedance of the local 90th percentile of daily precipitation, but somewhat underestimates the joint exceedance of the wet-day threshold and of correlation lengths. Within the set of PP methods analysed in our study multisite analog methods are thus the only ones that are clearly suitable in applications where a realistic representation of spatial variability is important. In climate change applications it needs to be carefully checked however whether their use is justified, as potential changes of the character of the analogs with respect to the predictor variable, and potential new weather situation that are not well represented by the analogs may make it difficult to capture the climate change signal. Furthermore, the temporal sequence of the downscaled series might be unrealistic (Maraun et al., 2018).

The stochastic PP and MOS methods considered in the study yield timeseries that are too independent between the stations. There are two potential contributions to this. First, the local variability that is explained by large-scale predictors, and thus leads to links between locations, could be underestimated due to the choice of statistical model and predictors. Second, the local noise is independently added at different locations, and thus cannot include potential links in the unexplained variability. The unconditional, local weather generators, which generate timeseries that are completely uncorrelated between the locations, trivially fail to generate realistic spatial fields. Recently multisite weather generators have been developed, and it has been demonstrated that they can capture the spatial characteristics of precipitation at the catchment scale well (e.g. Keller et al., 2015). If parameter changes in a future climate can be credibly estimated, for instance by conditioning them on predictor variables, such multisite weather generators can in principle be applied for climate change studies.

As can be seen in Fig. 17 in most cases positive (negative) biases in the correlation length are associated with negative (positive) biases in the degrees of freedom, and the ranking of the magnitudes is similar. This might be expected as both measures are based on correlations and capture aspects of the spatial complexity of the fields, with low (high) complexity likely to be associated with large (small) correlation lengths and a low (high) number of degrees of freedom. However, there are some exceptions. For temperature the only method for which the association is not found is CDFt-E, which as discussed earlier might be due to technical problems with the method. The other exception are some of the MOS methods for precipitation, which have small negative biases for the correlation lengths (see also Fig. 6) but also negative biases for the degrees of freedom. This shows that although both measures usually yield essentially the same information, subtleties in the correlation structure can exist that lead to both biases having the same sign. This situation can occur because the correlation lengths are dominated by station pairs with distances that lead to correlations near the correlation threshold, whereas the degrees of freedom are based on the entire correlation matrix. Although both approaches require the calculation of the correlation matrix, calculating the degrees of freedom is more straightforward because only the eigenvalue spectrum is required, whereas determining the correlation lengths requires the calculation of correlations as a function of distance, fitting of a smooth function, and involves a subjective correlation threshold.

In summary we found that most PP downscaling methods yield unrealistic spatial characteristics, regardless of whether large-scale or local predictors were used, and therefore should not be applied for multisite downscaling if the spatial characteristics of the results are relevant. The exception are multisite analog methods and a method that explicitly models spatial dependence, which performed well. The raw RCM clearly improves the skill compared to the driving reanalysis. Adjusting the marginal distributions through MOS further reduces biases in correlation lengths for precipitation and joint occurrence of wet days, but does neither reduce the underestimation of complexity as measured by degrees of freedom, nor the substantial overestimation of the joint occurrence of heavy precipitation events, while the improvements through the RCM are in most cases retained. Whether the spatial characteristics of the output of these methods is realistic enough for a given application needs to be carefully considered in each individual case. Moreover, a good performance in a perfect predictor setup is no guarantee that the methods will perform well when driven with GCM simulations for the present climate or that the climate change signal is realistically represented (e.g. Maraun et al., 2017).

Despite the satisfying skill of some statistical downscaling methods, our results show that providing downscaled meteorological fields with realistic spatial characteristics remains a challenge. In principle the common influence of predictors in singlesite PP methods could lead to realistic spatial patterns, but in the methods considered here it does not. The better skill of the RCM and of MOS methods compared to most PP methods shows that explicit physical modelling with local statistical post-processing is in general a better approach for obtaining realistic spatial fields than deriving full spatial fields from large-scale predictors (with the exceptions mentioned above). However none of the methods considered is able to produce output with a highly realistic spatial structure, including the dependences for the exceedance of high precipitation thresholds. There is thus still a clear need for increasing the resolution of RCMs used in climate change studies, because the explicit physical modelling of small-scale processes can be expected to improve the spatial characteristics of the raw model output and of MOS-corrected fields, as well as lead to more realistic climate change signals if regional processes affect climate change. Multisite weather generators and multisite MOS have also the potential to yield realistic spatial fields, but depend either on the assumption that the spatial dependence does not change over time, or on ways to estimate and include changes in the dependence.

We note that the observation network used in VALUE is designed for validation of a wide range of aspects of downscaling results, and not specifically selected for the analysis of spatial variability. In particular the European network, but also the German one, have station densities that do not well resolve variability within small hydrological catchments. Thus similar studies with a very high station density would be desirable. On very small scales subgrid variability becomes relevant for MOS methods and our results might not be directly transferable because deterministic MOS approaches can be expected to lead to too high dependences in cases where there is substantial subgrid variability (Maraun, 2013).

As our intercomparison is based on an ensemble of opportunity of downscaling methods it would also be very useful to conduct future comparisons of spatial aspects with a set of downscaling methods that does include all methods that are designed to represent spatial variability well. This should include for instance the multisite weather generators and multisite MOS methods mentioned in the introduction. The evaluation of the former in different studies has been inconclusive, while it has been positive for the latter, and a systematic comparison using a common experimental setup would be very helpful for identifying suitable methods and for informing further method development. The methods that explicitly model spatial dependence are more complex, more difficult to calibrate and apply, and more computationally expensive than most of the methods used in our study, which is one of the main reasons they are not frequently used and thus not included. The complexity of these methods also means that they are not necessarly much easier to implement and apply than high-resolution RCMs. Which combination of dynamical and statistical models is best suited for a given application therefore needs careful consideration.

## Acknowledgements

VALUE has been funded as EU COST Action ES1102. Participation of R. Huth in VALUE was supported by the Ministry of Education, Youth, and Sports of the Czech Republic under contract LD12059. JMG and SH acknowledge partial funding from MULTI-SDM project (MINECO/FEDER, CGL2015-66583-R).

# References

Arnaud, P., Bouvier, C., Cisneros, L. and Dominguez, R. (2002), 'Influence of rainfall spatial variability on flood prediction', *J. Hydrol.* **260**(1), 216–230.

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

Bárdossy, A. and Pegram, G. (2012), 'Multiscale spatial recorrelation of RCM precipitation to produce unbiased climate change scenarios over large areas and small', *Water Resour. Res.* **48**(9).

Boé, J., Terray, L., Habets, F. and Martin, E. (2006), 'A simple statistical-dynamical downscaling scheme based on weather types and conditional resampling', *J. Geophys. Res.-Atmos.* **111**(D23).

Bretherton, C. S., Widmann, M., Dymnikov, V. P., Wallace, J. M. and Blade, I. (1999), 'The effective number of spatial degrees of freedom of a time-varying field', *J. Clim.* **12**(7), 1990–2009.

Cannon, A. J. (2008), 'Probabilistic multisite precipitation downscaling by an expanded Bernoulli-Gamma density network', *J. Hydrometeorol.* **9**(6), 1284–1300.

Cannon, A. J. (2018), 'Multivariate quantile mapping bias correction: an N-dimensional probability density function transform for climate model simulations of multiple variables', *Clim. Dynam.* **50**(1-2), 31–49.

Cano, R., Sordo, C. and Gutiérrez, J. M. (2004), Applications of Bayesian Networks in meteorology, *in* J. A. Gámez, S. Moral and A. Salmerón, eds, 'Advances in Bayesian Networks', Springer Berlin Heidelberg, pp. 309–328.

Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B. and Wilby, R. (2004), 'The Schaake shuffle: A method for reconstructing space-time variability in forecasted precipitation and temperature fields', *J. Hydrometeorol.* **5**(1), 243–262.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Holm, E. V., Isaksen, L., Kallberg, P., Koehler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J. J., Park, B. K., Peubey, C., de Rosnay, P., Tavolato, C., Thepaut, J. N. and Vitart, F. (2011), 'The ERA-Interim reanalysis: configuration and performance of the data assimilation system', *Q. J. Roy. Meteor. Soc.* **137**(656, A), 553–597.

Easterling, D. (1999), 'Development of regional climate scenarios using a downscaling approach', *Climatic Change* **41**(3-4), 615–634.

Ekstroem, M., Grose, M. R. and Whetton, P. H. (2015), 'An appraisal of downscaling methods used in climate change research', *Wiley Interdisciplinary Reviews - Climate Change* **6**(3), 301–319.

Ferraris, L., Gabellani, S., Rebora, N. and Provenzale, A. (2003), 'A comparison of stochastic models for spatial rainfall downscaling', *Water Resour. Res.* **39**(12).

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

Gutiérrez, J., Maraun, D., Widmann, M., Huth, R., Hertig, E., Benestad, R., Roessler, O., Wibig, J., Wilcke, R., Kotlarski, S., San-Martín, D., Herrera, S., Bedia, J., Casanueva, A., Manzanas, R., Iturbide, M., Vrac, M., Dubrovsky, M., Ribalaygua, J., Pórtoles, J., Räty, O., Räisänen, J., Hingray, B., Raynaud, D., Casado, M., Ramos, P., Zerenner, T., Turco, M., Bosshard, T., Štěpánek, P., Bartholy, J., Pongracz, R., Keller, D., Fischer, A., Cardoso, R., Soares, P., Czernecki, B. and Pagé, C. (2018), 'An intercomparison of a large ensemble of statistical downscaling methods over Europe: Results from the VALUE perfect predictor cross-validation experiment', *Int. J. Climatol.* (in this issue).

Gutmann, E., Pruitt, T., Clark, M. P., Brekke, L., Arnold, J. R., Raff, D. A. and Rasmussen, R. M. (2014), 'An intercomparison of statistical downscaling methods used for water resource assessments in the United States', *Water Resour. Res.* **50**(9), 7167–7186.

Hannachi, A., Jolliffe, I. T. and Stephenson, D. B. (2007), 'Empirical orthogonal functions and related techniques in atmospheric science: A review', *Int. J. Climatol.* **27**(9), 1119–1152.

Hengl, T. (2007), *A practical guide to geostatistical mapping of environmental variables*, European commission. Joint Research Centre. Publications Office, Luxembourg.

Herdin, M., Czink, N., Ozcelik, H. and Bonek, E. (2005), Correlation matrix distance, a meaningful measure for evaluation of non-stationary MIMO channels, *in* 'Vehicular Technology Conference, 2005. VTC 2005-Spring. 2005 IEEE 61st', Vol. 1, IEEE, pp. 136–140.

Hertig, E., Maraun, D., Bartholy, J., Pongracz, R., Vrac, M., Mares, I., Gutiérrez, J. M., Wibig, J., Casanueva, A. and Soares, P. M. M. (2018), 'Comparison of statistical downscaling methods with respect to extreme events over Europe: Validation results from the perfect predictor experiment of the COST Action VALUE', *Int. J. Climatol.* (in this issue).

Hewitson, B. C., Daron, J., Crane, R. G., Zermoglio, M. F. and Jack, C. (2014), 'Interrogating empirical-statistical downscaling', *Climatic Change* **122**(4), 539–554.

Hlinka, J., Hartman, D., Vejmelka, M., Novotna, D. and Palus, M. (2014), 'Non-linear dependence and teleconnections in climate data: sources, relevance, nonstationarity', *Clim. Dynam.* **42**(7-8), 1873–1886.

Holzkämper, A., Calanca, P. and Fuhrer, J. (2012), 'Statistical crop models: predicting the effects of temperature and precipitation changes', *Clim. Res.* **51**, 11–21.

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

Hubert, L. and Arabie, P. (1985), 'Comparing partitions', *J. Classif.* **2**(2-3), 193–218.

Huth, R. (2002), 'Statistical downscaling of daily temperature in central Europe', *J. Clim.* **15**, 1731–1742.

Huth, R., Kliegrova, S. and Metelka, L. (2008), 'Non-linearity in statistical downscaling: does it bring an improvement for daily temperature in Europe?', *Int. J. Climatol.* **28**(4), 465–477.

Huth, R., Miksovsky, J., Stepanek, P., Belda, M., Farda, A., Chladova, Z. and Pisoft, P. (2015), 'Comparative validation of statistical and dynamical downscaling models on a dense grid in central Europe: temperature', *Theor. Appl. Climatol.* **120**(3-4), 533–553.

Isotta, F. A., Vogel, R. and Frei, C. (2015), 'Evaluation of European regional reanalyses and downscalings for precipitation in the Alpine region', *Meteorol. Z.* **24**, 15–37.

Keller, D., Fischer, A., Frei, C., Liniger, M., Appenzeller, C. and Knutti, R. (2015), 'Implementation and validation of a Wilks-type multi-site daily precipitation generator over a typical Alpine river catchment', *Hydrol. Earth Syst. Sci.* **19**(5), 2163–2177.

Kettle, H. and Thompson, R. (2004), 'Statistical downscaling in European mountains: verification of reconstructed air temperature', *Clim. Res.* **26**(2), 97–112.

Kotlarski, S., Keuler, K., Christensen, O. B., 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. and Wulfmeyer, V. (2014), 'Regional climate modeling on European scales: a joint standard evaluation of the EURO-CORDEX RCM ensemble', *Geosci. Model Dev.* **7**(4), 1297–1333.

Machguth, H., Paul, F., Kotlarski, S. and Hoelzle, M. (2009), 'Calculating distributed glacier mass balance for the Swiss Alps from regional climate model output: a methodical description and interpretation of the results', *J. Geophys. Res.* **114**.

Mamalakis, A., Langousis, A., Deidda, R. and Marrocu, M. (2017), 'A parametric approach for simultaneous bias correction and high-resolution downscaling of climate model rainfall', *Water Resour. Res.* **53**(3), 2149–2170.

Maraun, D. (2013), 'Bias correction, quantile mapping, and downscaling: Revisiting the inflation issue', *Journal of Climate* **26**(6), 2137–2143.

Maraun, D., Huth, R., Gutiérrez, J. M., San-Martín, D., Dubrovsky, M., Fischer, A., Hertig, E., Soares, P. M. M., Bartholy, J., Pongrácz, R., Widmann, M., Casado, M. J., Ramos, P. and Bedia, J. (2018), 'The VALUE perfect predictor experiment: evaluation of temporal variability', *Int. J. Climatol.* (in this issue).

Maraun, D., Shepherd, T. G., Widmann, M., Zappa, G., Walton, D., Gutiérrez, J. M., Hagemann, S., Richter, I., Soares, P. M., Hall, A. et al. (2017), 'Towards process-informed bias correction of climate change simulations', *Nature Climate Change* **7**(11), 764.

Maraun, D., Wetterhall, F., Ireson, A. M., Chandler, R. E., Kendon, E. J., Widmann, M., Brienen, S., Rust, H. W., Sauter, T., Themessl, M., Venema, V. K. C., Chun, K. P., Goodess, C. M., Jones, R. G., Onof, C., Vrac, M. and Thiele-Eich, I. (2010), 'Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user', *Rev. Geophys.* **48**(3). RG3003.

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

Maraun, D., Widmann, M., Gutiérrez, J., Kotlarski, S., Chandler, R., Hertig, E., Wibig, J., Huth, R. and Wilcke, R. (2015), 'VALUE: A framework to validate downscaling approaches for climate change studies', *Earth's Future* **3**(1), 1–14. 2014EF000259.

Monestiez, P., Courault, D., Allard, D. and Ruget, F. (2001), 'Spatial interpolation of air temperature using environmental context: application to a crop model', *Environ. Ecol. Stat.* **8**, 297–309.

Paschalis, A., Molnar, P., Fatichi, S. and Burlando, B. (2013), 'A stochastic model for high-resolution space-time precipitation simulation', *Water Resour. Res.* **49**(12), 8400–8417.

Philipp, A., Della-Marta, P. M., Jacobeit, J., Fereday, D. R., Jones, P. D., Moberg, A. and Wanner, H. (2007), 'Long-term variability of daily North Atlantic-European pressure patterns since 1850 classified by simulated annealing clustering', *J. Clim* **20**(16), 4065–4095.

Pierce, D., Cayan, D., R. and Thrasher, B. (2014), 'Statistical downscaling using localized constructed analogs (LOCA)', *J. Hydrometeorol.* **15**(6), 2558–2585.

Prein, A. F., Langhans, W., Fosser, G., Ferrone, A., Ban, N., Goergen, K., Keller, M., Tölle, M., Gutjahr, O., Feser, F. et al. (2015), 'A review on regional convection-permitting climate modeling: Demonstrations, prospects, and challenges', *Reviews of geophysics* **53**(2), 323–361.

R Core Team (2018), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
**URL:** *https://www.R-project.org/*

Richman, M. (1986), 'Rotation of principal components', *J. Climatol.* **6**(3), 293–335.

Santander Meteorology Group (2016), *R.VALUE: Climate data validation in the framework of the COST action VALUE.* R package version 1.4-14.
**URL:** *https://github.com/SantanderMetGroup/*$R_V ALUE$

Santos, J. M. and Embrechts, M. (2009), On the use of the adjusted rand index as a metric for evaluating supervised classification, *in* 'International Conference on Artificial Neural Networks', Springer, pp. 175–184.

Segond, M. L., Wheater, H. S. and Onof, C. (2007), 'The significance of spatial rainfall representation for flood runoff estimation: A numerical evaluation based on the Lee catchment, UK', *J. Hydrol.* **347**(1), 116–131.

Tank, A., Wijngaard, J., Können, G., Böhm, R., Demarée, G., Gocheva, A., Mileta, M., Pashiardis, S., Hejkrlik, L., Kern-Hansen, C., Heino, R., Bessemoulin, P., Muller-Westermeier, G., Tzanakou, M., Szalai, S., Palsdottir, T., Fitzgerald, D., Rubin, S., Capaldo, M., Maugeri, M., Leitass, A., Bukantis, A., Aberfeld, R., Van Engelen, A., Forland, E., Mietus, M., Coelho, F., Mares, C., Razuvaev, V., Nieplova, E., Cegnar, T., Lopez, J., Dahlstrom, B., Moberg, A., Kirchhofer, W., Ceylan, A., Pachaliuk, O., Alexander, L. and Petrovic, P. (2002), 'Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment', *Int. J. Climatol.* **22**(12), 1441–1453.

Trambauer, P., Werner, M., Winsemius, H. C., Maskey, S., Dutra, E. and Uhlenbrook, S. (2015), 'Hydrological drought forecasting and skill assessment for the Limpopo river basin, Southern Africa', *Hydrol. Earth Syst. Sci.* **19**(4), 1695–1711.

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

Viviroli, D., Zappa, M., Schwanbeck, J., Gurtz, J. and Weingartner, R. (2009), 'Continuous simulation for flood estimation in ungauged mesoscale catchments of Switzerland - Part I: modelling framework and calibration results', *J. Hydrol* **377**, 191–207.

Voisin, N., Pappenberger, F., Lettenmaier, D. P., Buizza, R. and Schaake, J. C. (2011), 'Application of a medium-range global hydrologic probabilistic forecast scheme to the Ohio River Basin', *Weather Forecast.* **26**(4), 425–446.

Vrac, M. (2018), 'Multivariate bias adjustment of high-dimensional climate simulations: The "Rank Resampling for Distributions and Dependences" (R2D2) bias correction', *Hydrol. Earth Syst. Sci. Discuss.* **2018**, 1–33.

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

Wilks, D. S. (2012), 'Stochastic weather generators for climate-change downscaling, part II: multivariable and spatially coherent multisite downscaling: Stochastic weather generators for climate-change downscaling', *Wiley Interdisciplinary Reviews: Climate Change* **3**(3), 267–278.