https://doi.org/10.5194/hess-2020-318
Preprint. Discussion started: 7 July 2020

Hydrology and Earth System Sciences
Discussions

Identifying robust bias adjustment methods for extreme precipitation in a pseudo-reality setting

Torben Schmith[1], Peter Thejll[1], Peter Berg[2], Fredrik Boberg[1], Ole Bøssing Christensen[1], Bo Christiansen[1], Jens Hesselbjerg Christensen[1,3,4], Christian Steger[5], Marianne Sloth Madsen[1]

[1] Danish Meteorological Institute, Lyngbyvej 100, 2100 Copenhagen Ø, Denmark
[2] Swedish Meteorological and Hydrological Institute, Hydrology Research Unit, Norrköping, Sweden
[3] Physics of Ice, Climate and Earth, Niels Bohr Institute, University of Copenhagen, 2100 Copenhagen Ø, Denmark
[4] NORCE Norwegian Research Centre, Bjerknes Centre for Climate Research, 5007 Bergen, Norway
[5] Deutscher Wetterdienst, Frankfurter Straße 135, 63067 Offenbach, Germany

Correspondence: Torben Schmith (ts@dmi.dk)

Abstract

Severe precipitation events occur rarely and are often localized in space and of short duration; but they are important for societal managing of infrastructure. Therefore, there is a demand for estimating future changes in the statistics of these rare events. These are usually projected using Regional Climate Model (RCM) scenario simulations combined with extreme value analysis to obtain selected return levels of precipitation intensity. However, due to imperfections in the formulation of the physical parameterizations in the RCMs, the simulated present-day climate usually has biases relative to observations. Therefore, the RCM results are often bias-adjusted to match observations. This does, however, not guarantee that bias-adjusted projected results will match future reality better, since the bias may change in a changed climate. In the present work we evaluate different bias adjustment techniques in a changing climate. This is done in an inter-model cross-validation setup, in which each model simulation in turn plays the role of pseudo-reality, against which the remaining model simulations are bias adjusted and validated. The study uses hourly data from present-day and RCP8.5 late 21st century from 19 model simulations from the EURO-CORDEX ensemble at 0.11° resolution, from which fields of selected return levels are calculated for hourly and daily time scale. The bias adjustment techniques applied to the return levels are based on extreme value analysis and include analytical quantile-matching together with the simpler climate factor approach. Generally, return levels can be improved by bias adjustment, compared to obtaining them from raw scenarios. The performance of the different methods depends of the time scale considered. On hourly time scale, the climate factor approach performs better than the quantile-matching approaches. On daily time scale, the superior approach is to simply deduce future return levels from observations and the second best choice is using the quantile-mapping approaches. These results are found in all European sub-regions considered.

https://doi.org/10.5194/hess-2020-318
Preprint. Discussion started: 7 July 2020

Hydrology and Earth System Sciences
Open Access
EGU
Discussions

1 Introduction

Severe precipitation events occur either as stratiform day-long precipitation of moderate intensity or as localized cloudbursts lasting a few hours only. Such extreme events may cause flooding with the risk of loss of life and damage to infrastructure. It is expected that future changes in the radiative forcing from greenhouse gases and other forcing agents will influence the large scale atmospheric conditions, such as air mass humidity, vertical stability, and typical low pressure tracks. Therefore also the statistics of the occurrence of severe precipitation events will most likely change.

Global climate models (GCMs) is the main tool for estimating future climate conditions. A GCM is a global representation of the atmosphere, the ocean and the land surface, and the interaction between these components. The GCM is then forced with observed greenhouse gas concentrations, atmospheric compositions, land use, etc. to represent the past and present climate, and with stipulated scenarios of future concentrations of radiative agents to represent the future climate.

Present state-of-the art GCMs from the Coupled Model Intercomparison Project Phase 5 (CMIP5, Taylor et al. 2012) and the recent Coupled Model Intercomparison Project Phase 6 (CMIP6, Eyring et al. 2016) typically have a grid spacing of around 100 km or even more. This resolution is too coarse to describe the effect of regional and local features, such as mountains, coast lines and lakes and to adequately describe convective precipitation systems (Eggert et al. 2015). To model the processes on smaller spatial scales, dynamical downscaling is applied. Here, the atmospheric and surface fields from a GCM simulation are used as boundary conditions for a regional climate model (RCM) over a smaller region with a much finer grid spacing, at present typically around 10 km or even less.

The ability of present-day RCMs to reproduce observed extreme precipitation statistics on daily and sub-daily time scales is essential and has been of concern. Earlier studies analysing this topic have mostly focused on a particular country, probably due to the lack of sub-daily observational data covering larger regions, such as e.g. Europe. Thus, Hanel and Buishand (2010), Kendon et al. (2014), Olsson et al. (2015) and Sunyer et al. (2017) studied daily and hourly extreme precipitation in different European countries and reached similar conclusions: first that the bias of extreme statistics decreases with smaller grid spacing of the model, and second that extreme statistics for 24 h duration are satisfactorily simulated with a grid spacing of 10 km, while 1 h extreme statistics exhibits biases even at this resolution. Recently, Berg et al. (2019) have evaluated high resolution RCMs from the EURO-CORDEX ensemble (Jacob et al. 2014) and came up with similar conclusions for several countries across Europe: RCMs underestimate hourly extremes and give an erroneous spatial distribution.

Extreme convective precipitation of short duration is thus one of the more challenging phenomena to describe physically in RCMs. The reason is that convective events take place on a spatial scale comparable to the RCM grid spacing of presently around 10 km. Therefore, the convective plumes cannot be directly modelled. Instead, the effects of convection are parametrised, i.e. modelled as processes on larger spatial scales. Thus, the inability to reproduce these short duration extremes can be explained by the imperfect parametrization of sub-grid scale convection, which generally leads to too early onset of convective rainfall in the diurnal cycle and subsequent dampening of the build-up of convective available potential energy (CAPE, Trenberth et al. 2003).

https://doi.org/10.5194/hess-2020-318
Preprint. Discussion started: 7 July 2020

Hydrology and Earth System Sciences
Discussions

Thus, even RCMs with their small grid spacing may exhibit systematic biases for variables related to convective precipitation. If there is a substantial bias, we should consider adjusting for this bias. Bias adjustment techniques are thoroughly discussed, including requirements and limitations, in Maraun (2016) and Maraun et al. (2017). There are two main bias adjustment approaches. In the *delta-change* approach, a transformation is established from the present to the future climate in the model run. This transformation is then applied to the observations to get the projected future climate. In the *bias correction* approach, a transformation is established from present model climate data to the observed climate and this transformation is then applied to the future model climate to obtain the projected future climate.

Both adjustment approaches come in several flavours. In the simplest one, the transformation consists of an adjustment of the mean, in the case of precipitation by multiplying the mean by a factor. In the more elaborate flavour, the transformation is defined by quantile-matching, preserving also the higher moments. Quantile-matching adjustment can use either empirical quantiles or analytical distribution functions. The ability of quantile-matching to reduce bias has been demonstrated for daily precipitation in present-day climate using observations, which are split into training and verification parts (Piani et al. 2010; Themeßl et al. 2011).

Bias adjustment techniques originate in the field of weather and ocean forecast modelling, where output is adjusted for model deficiencies and local features. Applying similar bias adjustment techniques to climate model simulations, however, has a complication not present in weather and ocean forecast applications: Climate models are set up and tuned to present-day conditions and verified against observations, but then applied to future changed conditions without any possibility to directly verify the model's performance under these conditions. Consequently, showing that bias adjustment works for present-day climate is a necessary but not sufficient condition for the adjustment to work in the changed climate.

In practical applications of bias adjustment methods to climate simulations, it is generally assumed that the bias of the model is unchanged from the present-day climate to the future climate (stationarity). Only a few examples has pointed out directly how to validate this cornerstone assumption (see however Buser et al. (2010) and Boberg and Christensen (2012)) and therefore it is not obvious that applying bias adjustment improve projections of future climate characteristics. We also note that the bias adjustment methods themselves may influence the climate change signal of the model, depending on the bias and the correction method used (Haerter et al. 2011; Berg et al. 2012; Themeßl et al. 2012).

One approach to partly overcome the above challenge and evaluate the total performance of bias adjustment methods is *inter-model cross-validation*, as pursued by Maraun (2012), Räisänen and Räty (2013) and Räty et al. (2014). The rationale is that the members in a multi-model ensemble of simulations represent different descriptions of physics of the climate system, with each of them being not too far from the real climate system. In the cross-validation exercise, one member of the ensemble in turn plays the role of *pseudo-reality*, against which the remaining bias-adjusted models are evaluated. Thus, the trick is that we know both present and future pseudo-reality.

https://doi.org/10.5194/hess-2020-318
Preprint. Discussion started: 7 July 2020

Inter-model cross-validation has been applied on daily precipitation to evaluate different adjustment methods (Räty et al. 2014). Here we apply a similar methodology European-wide to extreme precipitation on hourly and daily time scale. This has been possible with the advent of the EURO-CORDEX, a large ensemble of high-resolution RCM simulations with precipitation in hourly time-resolution. Being more specific, we will apply the standard extreme value analysis to the ensemble of model data for present-day and end-21st-century conditions to estimate return levels for daily and hourly duration. Then we will apply inter-model cross validation on these return levels in order to address the following questions:

1. Do bias-adjusted return levels perform better, according to the inter-model cross-validating, than using un-corrected model data from scenario simulations?
2. Is there any difference in performance between different adjustment methods?
3. Are there systematic differences in point 1 and 2, depending on the daily and hourly duration?
4. Are there regional differences across Europe in the performance of the different techniques?

Giving qualified answers to these questions can serve as important guidelines for analysis procedures for obtaining future extreme precipitation characteristics.

The rest of the paper contains a description of the EURO-CORDEX data (Section 2) and a description of methods used (Section 3). Then follow the results (Section 4), a discussion of these (Section 5) and finally a summary (Section **Fejl! Henvisningskilde ikke fundet.**).

2 The EURO-CORDEX data

The model simulations used here have been performed within the framework of EURO-CORDEX (Jacob et al. (2014) ; http://euro-cordex.net), which is an international effort aimed at providing RCM climate simulations for a specific European region (see Figure 1) in two standard resolutions with a grid spacing of 0.44° (EUR-44, ~50 km) and 0.11° (EUR-11, ~12.5 km), respectively. All GCM simulations driving the RCMs follow the CMIP5 protocol (Taylor et al. 2012) and are forced with historical forcing for the period 1951-2005 followed by the RCP8.5 scenario for the period 2006-2100 (until 2099 only for HadGEM-ES).

We analyse precipitation data in hourly time-resolution from 19 different GCM-RCM combinations from the EUR-11 simulations shown in Table 1 and we analyse two 25 year long time slices from each of these simulations: a present-day time slice (years 1981-2005) and an end-21st-century time slice (years 2075-2099).

All GCM-RCM combinations we use are represented by one realization only, and therefore the data material used represents 19 different possible realisations of climate model physics, though acknowledging that some GCMs/RCMs might originate from the same or similar ancestor and therefore may not be fully independent. The EURO-CORDEX ensemble includes a few simulations, which do not use the standard EUR-11 grid. These were not included in the analysis, since they should have been re-gridded to the EUR-11 grid which would dampen extreme events, thus introducing an unnecessary error source.

https://doi.org/10.5194/hess-2020-318
Preprint. Discussion started: 7 July 2020

Table 1 Overview of the 19 EURO-CORDEX GCM-RCM combinations used. The rows show the GCMs while the columns show the RCMs. The full names of the RCMs are SMHI-RCA4, CLMcom-CCLM4-8-17, KNMI-RACMO22E, DMI-HIRHAM5, MPI-CSC-REMO2009 and CLMcom-ETH-COSMO-crCLIM-v1-1. Each GCM-RCM combination used is represented by a number (1, 3 or 12) indicating which realization of the GCM is used for the particular simulation.

GCM / **RCM**	RCA	CCLM	RACMO	HIRHAM	REMO	COSMO
ICHEC-EC-EARTH	r12		r1	r3		
MOHC-HadGEM2-ES	r1		r1	r1		
CNRM-CERFACS-CNRM-CM5	r1			r1		
MPI-M-MPI-ESM-LR	r1	r2		r1	r1	r1
IPSL-IPSL-CM5A-MR	r1					
NCC-NorESM1-M	r1			r1		r1
CCCma-CanESM2		r1				
MIROC-MIROC5		r1				

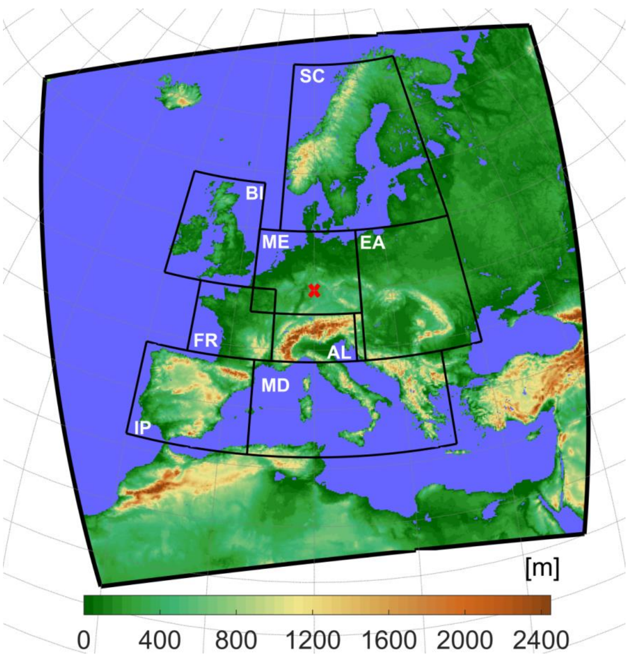

Figure 1 Map showing the EURO-CORDEX region (outer frame) with elevation in colours. PRUDENCE sub-regions (Christensen and Christensen 2007) used in the analysis are also shown: BI = British Isles, IP = Iberian Peninsula, FR = France, ME = Mid-Europe, SC = Scandinavia, AL = Alps, MD = Mediterranean, EA = Eastern Europe. Red cross marks point used in Figure 4.

https://doi.org/10.5194/hess-2020-318
Preprint. Discussion started: 7 July 2020

Hydrology and Earth System Sciences
Open Access
EGU
Discussions

3 Methods

3.1 Duration

Extreme precipitation statistics is often described as function of the time scale involved as intensity-duration-frequency or depth-duration-frequency curves (e.g. Overeem et al. 2008). We consider two time scales or *durations*. One is a duration of 1 h, which is simply the time series of hourly precipitation sums available in each RCM grid point. The other is a duration of 24 h, where a 24 h sum is applied in a sliding window with a one hour time stepping. We will sometimes refer to these as hourly and daily duration, respectively. Our daily duration corresponds to the traditional climatological practice of reporting daily sums but allows heavy precipitation events to occur over two consecutive days. We also emphasize that the duration, as defined here, is not the actual length of precipitation events in the model data, but is merely a concept to define time scales.

3.2 Extreme value analysis

Extreme value analysis (EVA) is about estimating high quantiles of a statistical distribution from observations. The theory relies on fundamental convergence properties of time series of extreme events; for details we refer to Coles et al. (2001).

There are two main methodologies in EVA to obtain estimates of the high percentiles and the corresponding return levels. In the *classical*, or *block maxima*, method, a generalised extreme value distribution is fitted to the series of maxima over a time block, usually a year. Alternatively, in the *peak-over-threshold* (POT) or *partial-duration-series* method, which is used here, all peaks with maximum above a (high) threshold, x_0, are considered. The peaks are assumed to occur independently at an average rate per year of λ_0. To ensure independence between peaks, a minimum time separation between peaks is specified. Theory tells us, that when the threshold goes to infinity, the distribution of the exceedances above the threshold, $x - x_0$, converges to a generalised Pareto distribution, whose cumulative distribution function is

$$\mathcal{G}(x - x_0) = 1 - \left(1 + \xi \frac{x - x_0}{\sigma}\right)^{-\frac{1}{\xi}}, x > x_0$$

The parameter σ is the scale and is a measure of the width of the distribution. The parameter ξ is the shape and describes the character of the upper tail of the GPD-distribution; $\xi > 0$ implies a heavy tail which usually is the case for extreme precipitation events, while $\xi < 0$ implies a thin tail. Note that, quite confusingly, an alternative sign convention of ξ occurs in the literature (e.g. Hosking and Wallis 1987).

If we now consider an arbitrary level x with $x > x_0$, the average number of exceedances per year of x will be

$$\lambda_x = \lambda_0 \left[1 - \mathcal{G}(x - x_0)\right]. \qquad (1)$$

The T-year return level, x_T, is defined as the precipitation intensity which is exceeded on average once every T years

$$\lambda_{x_T} T = 1$$

and by combining with (1) we get an expression for the return level x_T

https://doi.org/10.5194/hess-2020-318
Preprint. Discussion started: 7 July 2020

Hydrology and Earth System Sciences
Discussions

$$\lambda_0[1 - \mathcal{G}(x_T - x_0)]T = 1,$$

from which

$$x_T = \mathcal{G}^{-1}\left(1 - \frac{1}{\lambda_0 T}\right) + x_0. \qquad (2)$$

Data points to be included in the POT analysis can be selected in two different ways. Either the threshold x_0 is specified and λ_0 is then a parameter to be determined or, alternatively, λ_0 is specified and x_0 determined as a parameter. We choose the latter approach, since it is most convenient when working with data from many different model simulations.

Choosing λ_0 is a point to consider: a too high value would include too few data points in the estimation and a too low value implies the risk that the exceedances $x_T - x_0$ cannot be considered as GPD-distributed. We choose $\lambda_0 = 3$ in accordance with Berg et al. (2019), which gives 75 data points for estimation for the 25 years period. Hosking and Wallis (1987) investigated the estimation of parameters of the GPD-distribution and based on this warns against using the often applied maximum likelihood estimation for a sample size below 500. Instead, he recommends probability-weighted moments and we have followed this advice here.

We required a minimum of 3 and 24 h separation between peaks for 1 and 24 h duration, respectively. This is in accordance with Berg et al. (2019) and furthermore, synoptic experience tells us that this will ensure that neighbouring peaks are from independent weather systems. We found only a weak influence of these choices on the results of our analysis.

3.3 Bias adjustments and extreme value analysis

The delta-change and bias correction approaches were introduced in general terms in Section 1. Now we will formulate EVA-based analytical quantile-mapping based versions of the two approaches. In what follows O_T is the T-year return levels estimated from (pseudo-)observations during the present-day period, while C_T (control) and S_T (scenario) denote the corresponding return levels, estimated from present-day and end-21st-century model data, respectively. Finally, P_T (projection) denotes the end-21st-century return level after bias-adjustment has been applied.

3.3.1 Climate factor on the return levels (FAC)

The simplest adjustment approach is to assume a climate factor on the return level (FAC)

$$P_T = \underbrace{S_T/C_T}_{\substack{Delta-change\\ climate\ factor}} \cdot O_T = \underbrace{O_T/C_T}_{\substack{Bias\ correction\\ climate\ factor}} \cdot S_T$$

We note that the delta-change and bias correction approach are identical for the FAC method.

https://doi.org/10.5194/hess-2020-318
Preprint. Discussion started: 7 July 2020

3.3.2 Analytical quantile matching based on EVA

Kallache et al. (2011) and Laflamme et al. (2016) applies a transformation methodology for extreme values, based on analytical quantile-matching and applicable for both the block- and the POT-methods, which will be adapted to our needs below.

In the EVA-based quantile-matching, two POT-based extreme value distributions with different parameters are matched. Being more specific, we want to construct a transformation $x \to y$ defined by requiring that exceedance rates above x and y, respectively, are equal for any x:

$$\lambda_x = \lambda_y.$$

This implies, according to (1), that

$$\lambda_{0x}[1 - \mathcal{G}_x(x - x_0)] = \lambda_{0y}[1 - \mathcal{G}_y(y - y_0)],$$

where $\mathcal{G}_x$ is ths GPD distribution of the exceedances $x - x_0$ and λ_{0x} the associated exceedance rate, and $\mathcal{G}_y$ and λ_{0y} are the similar entities for y.

To simplify, we let $\lambda_{0x} = \lambda_{0y}$(see Section 3.2) and therefore get

$$\mathcal{G}_x(x - x_0) = \mathcal{G}_y(y - y_0),$$

from which we obtain the transformation

$$y = y_0 + \mathcal{G}_y^{-1}\big(\mathcal{G}_x(x - x_0)\big). \quad (3)$$

For the delta-change approach (DC), the modelled GPD distribution functions for present-day and end-21st-century conditions are quantile-matched and the transformation obtained this way is then applied to return levels determined from present-day (pseudo-)observations O_T. Thus the corresponding projected T-year return level is according to Eq. (3)

$$P_T = S_0 + \mathcal{G}_S^{-1}\big(\mathcal{G}_C(O_T - C_0)\big),$$

where $\mathcal{G}_C$ and $\mathcal{G}_S$ are the GPD cumulative distribution functions for the modelled present-day (control) and end-21st-century (scenario) data, respectively, and C_0 and S_0 are the corresponding threshold values.

For the bias correction approach (BC), the present-day (control) and (pseudo-)observed GPD cumulative distribution functions are quantile-matched to obtain the model bias, which then is applied, according to eq. (3), to modelled end-21st-century (scenario) return levels.

$$P_T = O_0 + \mathcal{G}_O^{-1}\big(\mathcal{G}_C(S_T - C_0)\big),$$

where $\mathcal{G}_O$ is the GPD cumulative distribution function for the observations and O_0 the corresponding threshold.

3.3.3 Reference adjustment methods

The performance of the bias adjustment methods described above will be compared with the performance of two reference adjustment methods, which are defined below. This is a similar to what is practice when verifying predictions, where the performance of the prediction should be superior to the performance of reference predictions, such as persistence or climatology.

https://doi.org/10.5194/hess-2020-318
Preprint. Discussion started: 7 July 2020

We choose two reference methods. One reference is to simply use, for a given model, the return level calculated from (pseudo-)observations as the projected return level (OBS),

$$P_T = O_T$$

Another reference is to use the scenario model output without any bias adjustment (SCE):

$$P_T = S_T.$$

For an overview of methods, see Table 2

Table 2 Overview of methods used in the inter-comparison

OBS	(Pseudo-)observations (Reference)
SCE	Unadjusted RCM scenario (Reference)
FAC	Climate factors on return levels
DC	Quantile-matched delta-change based on EVA
BC	Quantile-matched bias correction based on EVA

4 Results

4.1 Modelled return levels for present-day and end-21st-century conditions

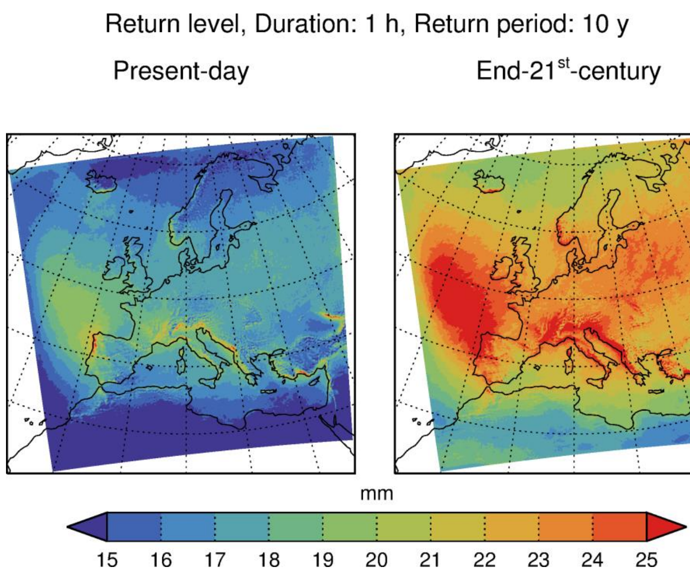

Figure 2 Geographical distribution of the 10 year-return level of precipitation intensity for 1 hour duration for present-day (left) and end-21st-century (right). In each grid point, values are the median return level over all 19 model simulations.

Figure 2 displays the geographical distribution of the 10-year return level for precipitation intensity of 1 h duration, calculated as the median return level over all 19 model simulations. There is a general increase

https://doi.org/10.5194/hess-2020-318
Preprint. Discussion started: 7 July 2020

Hydrology and Earth System Sciences
Open Access
Discussions

from present-day to end-21st-century climatic conditions. The smallest return levels are mainly found in the arid North African region and to some extent in the Norwegian Sea, while the largest return levels are found in southern Europe and in the Atlantic northwest of the Iberian Peninsula. Mountainous regions, such as the Alps and western Norway have higher return levels than their surroundings.

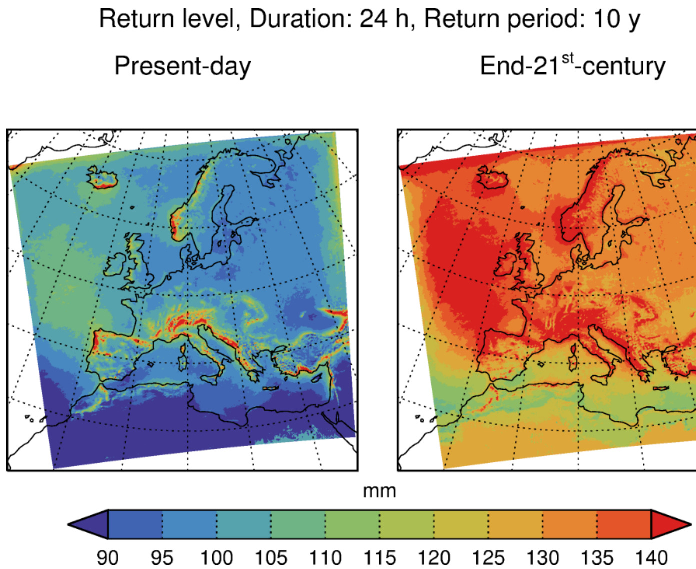

Figure 3 As Figure 2 but for 24 h duration

We also show in Figure 3 the median 10-year return level for 24 h duration, and this shows similar qualitative characteristics: For both durations the return levels generally increase from present-day to end-21st-century conditions, although the effect is more pronounced for 1 h duration.

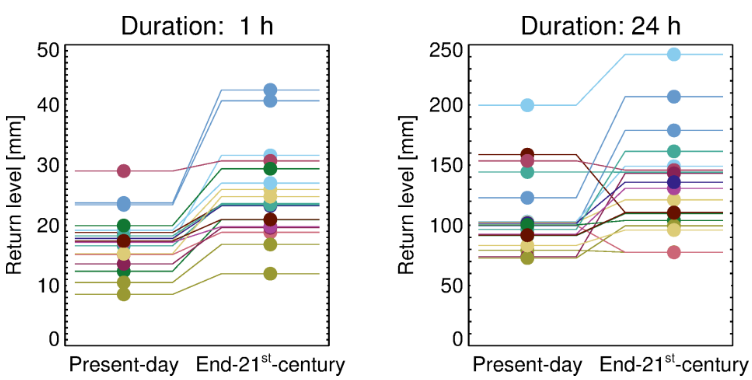

Figure 4 Modelled return levels at 50N/10E (northern Germany, marked with 'X' in Figure 1) for present and future for 10 y return period and 1 h and 24 h durations. Different colours represent the 19 different GCM-RCM simulations listed in Table 1.

https://doi.org/10.5194/hess-2020-318
Preprint. Discussion started: 7 July 2020

Hydrology and Earth System Sciences
Discussions

To get a more detailed impression of the data, Figure 4 shows return levels and their changes from present-day to end-21st-century for a grid point in Northern Germany for all 19 model simulations. For 1 h duration (left panel) return values increase from present-day to end-21st-century in all cases. For 24 h duration (right panel) typically the return levels increase from present-day to end-21st-century but with some exceptions. For both durations, we also note the large spread in return levels within the ensemble. The spread is much higher than the change between present and future for most models; in other words: a poor signal to noise ratio.

4.2 Inter-model cross-validation

4.2.1 Validation metrics

Results of the inter-model cross-validation are presented in this section. The basic verification metric will be the relative error of future return levels for a given duration and return period T, defined as

$$RE = |P_T - V_T|/V_T$$

i.e. the absolute difference between the projected return level P_T obtained from applying bias adjustment and the verification return level V_T estimated from end-21st-century pseudo-reality, divided by the verification return level. This metric is calculated for every grid point and for every model/pseudo-reality combination. Since we have $N = 19$ model simulations in the ensemble, we can make $N \times (N-1) = 342$ evaluations of each bias adjustment method and make statistics of the relative error. This quantifies the average performance of the different bias adjustment methods.

In the following, we will present results using two different types of display. First, we will use spatial maps of the median relative error, calculated from all model/pseudo-reality combinations. Second, we will, for each adjustment method and for each model/pseudo-reality combination, calculate the median relative error over each of the eight PRUDENCE sub-regions defined in Christensen and Christensen (2007) and shown on Figure 1. For each region we will illustrate the distribution of the relative error across all model/pseudo-reality combinations by showing the median and the 0.05/0.95-percentiles of this distribution.

4.2.2 Results for 1 h duration

Figure 5 shows the median, across all model/pseudo-reality combinations, of RE for all five methods for 1 h duration and 10 y return period.

https://doi.org/10.5194/hess-2020-318
Preprint. Discussion started: 7 July 2020

Hydrology and Earth System Sciences
Open Access
Discussions

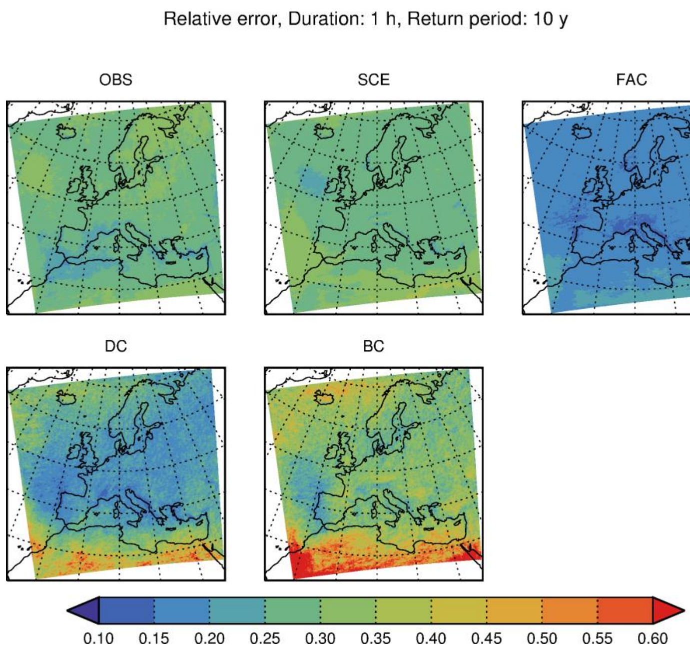

Figure 5 Geographical distribution of the relative error of end-21st-century 10 year return level for 1 h duration precipitation intensity from the inter-model cross-validation. Colours show the median of the relative error calculated over all model/pseudo-reality combinations. Panels are for the different bias correction methods.

First we look at the reference methods. The OBS method has relative errors in the approximate interval 0.2-0.4. Lowest values are found in the Mediterranean, western France and the Atlantic west of the Mediterranean; highest values in the Atlantic west of Ireland and in Scandinavia. The SCE method has errors in the interval 0.25-0.45, lowest values in the Atlantic west of Ireland; largest values over parts of the Atlantic and northern Africa. Of the two reference methods, the OBS method outperforms SCE in the south, while the opposite is true in the north.

The relative error of FAC is below 0.2 in most places. It is everywhere smaller than the relative error of the reference methods OBS and SCE. The DC method has a relative error comparable to (e.g. Western France, Western Iberia and Eastern Atlantic) or larger than (in particular in Northern Africa) that of FAC. That said, the concept of relative error should be used with care in an arid region, such as Northern Africa. But from this result, it is not justified to use the more complicated DC, in favour of the simpler FAC. Finally, the relative error of BC is everywhere above both DC and FAC, indicating the poorest performance of all methods considered.

https://doi.org/10.5194/hess-2020-318
Preprint. Discussion started: 7 July 2020

Hydrology and Earth System Sciences
Open Access
Discussions

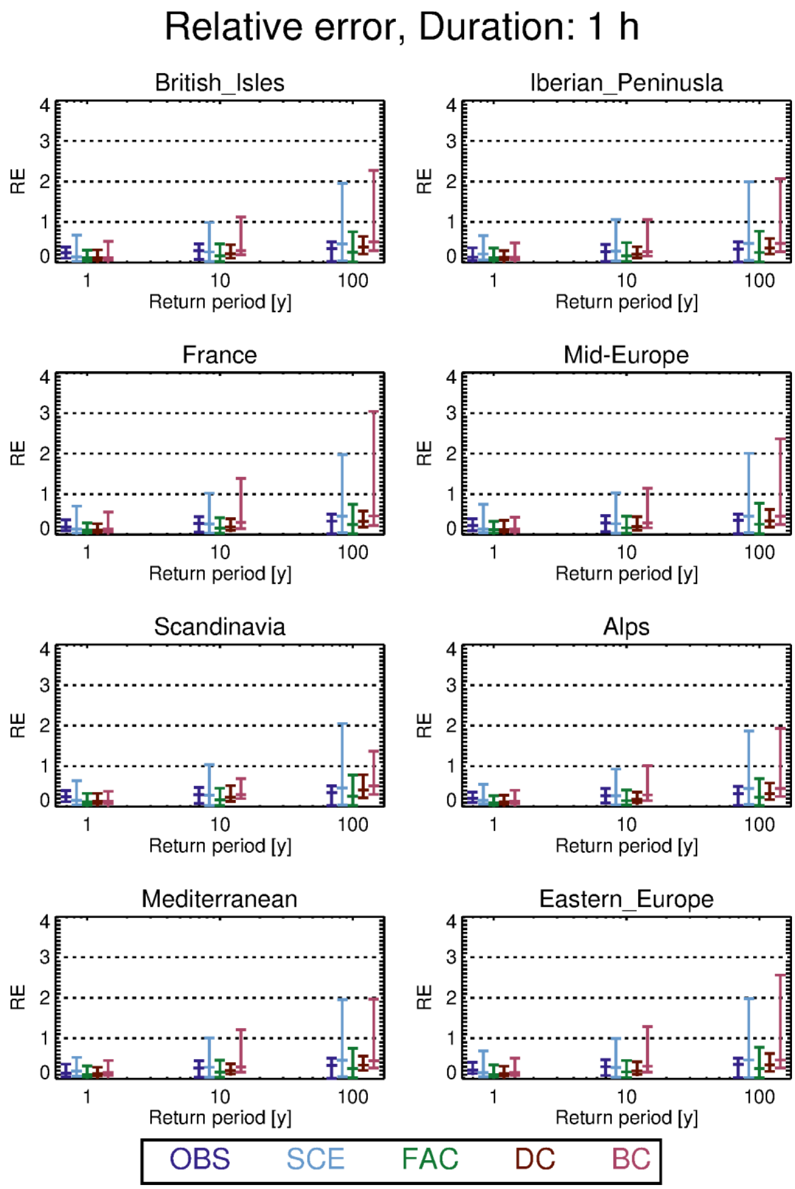

Figure 6 Statistical distribution (median and .05/.95-fractiles) of the relative error of the inter-model cross-validation for 1 hour duration for 1 y, 10 y and 100 y return periods. Panels represent PRUDENCE sub-regions shown in Figure 1. Each colour represents an adjustment method (see Table 2).

The statistical distribution of the relative error is shown in Figure 6 for the eight PRUDENCE sub-regions (see Figure 1). We first note that the distribution of relative error is shifted towards higher values for larger return periods, as expected. Next, we note that the two reference methods, OBS and SCE, behave differently. SCE generally has a little larger median relative error, but the .95-fractile is much larger for SCE than for OBS, in particular for large return periods. Thus, OBS overall performs better than SCE, meaning

https://doi.org/10.5194/hess-2020-318
Preprint. Discussion started: 7 July 2020

Hydrology and Earth System Sciences
Discussions

that using present-day pseudo-observations to estimate projected end-21st-century return levels yields better relative error than using raw modelled scenario data.

The FAC method generally has the best overall performance, both in terms of median and .95-fractile of the relative error. Of the two quantile-matching methods, the DC method has a slightly poorer performance than FAC, both in terms of the median and the .95-fractile of the relative error. Finally, BC has poorer performance than DC, when comparing the median of the relative error and in particular for the .95-fractile.

In summary, for 1 h duration, the method with the best performance is using a climate factor on the return levels (FAC). This method outperforms both reference methods and the more sophisticated methods based on quantile-matching, DC and BC, the latter having the poorest overall performance of them all.

4.2.3 Results for 24 h duration

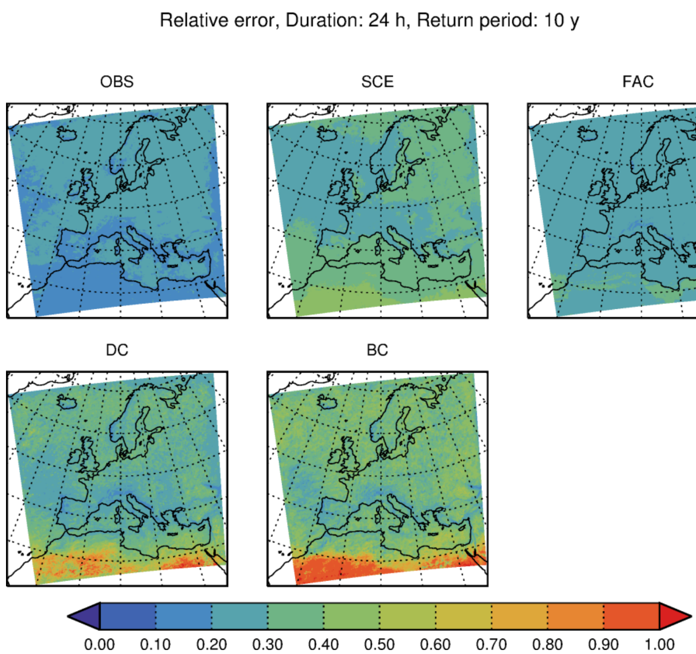

Figure 7 As Figure 5 but for 24 h duration.

For 24 h duration (see Figure 7), OBS has the lowest median relative error (lower than 0.3) in most regions of all the adjustment methods, while SCE has higher relative error in the interval 0.3-0.6 approximately, with the highest values in North Africa. FAC has relative errors between OBS and SCE. Of the quantile-matching methods, DC has relative errors in the interval 0.2-0.8 approximately, larger than FAC in most places, and finally BC has, as for 1 h duration, the largest median relative errors of all the methods.

https://doi.org/10.5194/hess-2020-318
Preprint. Discussion started: 7 July 2020

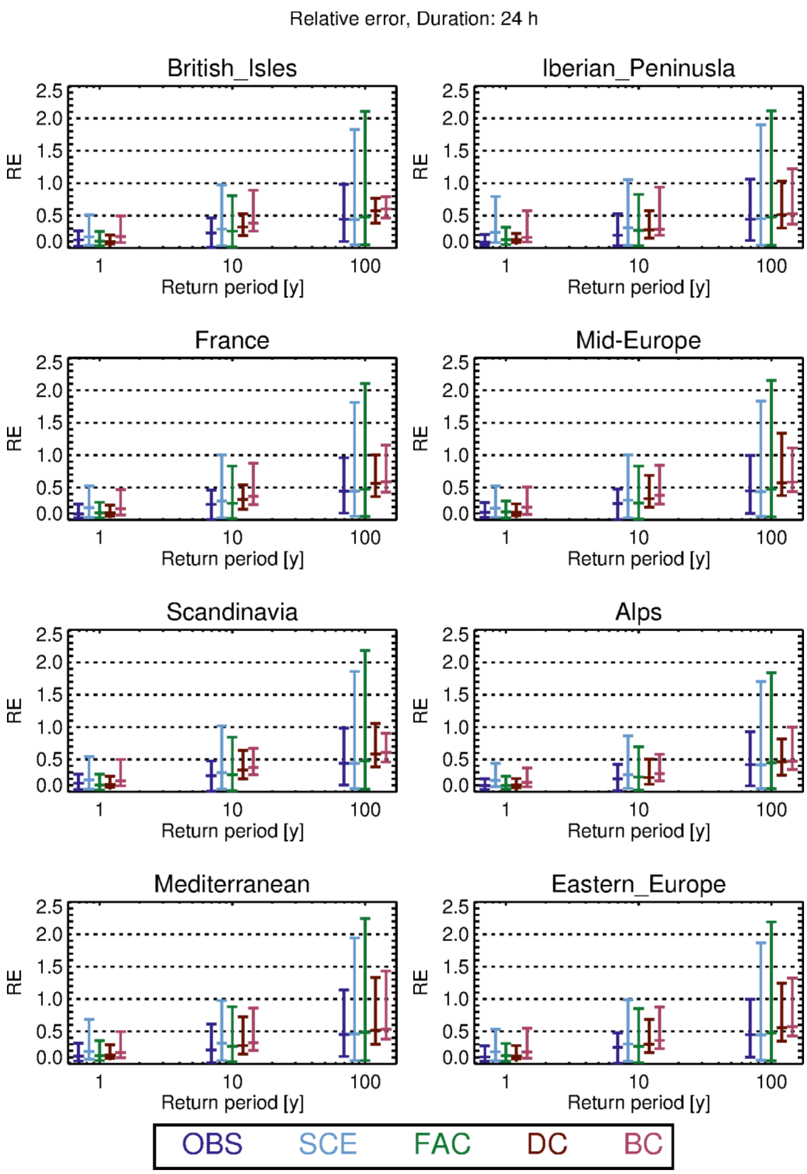

Figure 8 As Figure 6 but for 24 h duration

As for the 1 h duration, we also compare the entire statistical distribution of the relative error of the different adjustment methods for all three return periods (Figure 8), and again, both median and .95-fraction of the relative error increases for larger return periods, as expected. Further, OBS seems, surprisingly, to have a small median relative error and the smallest .95-fractile of all methods considered for all sub-regions. SCE has a median not too different from that of OBS, but the .95-fractile is much larger.

https://doi.org/10.5194/hess-2020-318
Preprint. Discussion started: 7 July 2020

Similar characteristics hold for FAC. The quantile-matching methods DC and BC have slightly larger median values, but the .95-fractile is smaller than for FAC. All these characteristics hold for all sub-regions.

4.3 Further analysis on conditions for skill

To get further insight into the difference in performance between hourly and daily precipitation, we consider the relationship between the bias factor for present-day $B_P = \frac{C}{O}$ and end-21st-century $B_F = \frac{S}{V}$ for all model/pseudo-reality combinations (see Figure 9).

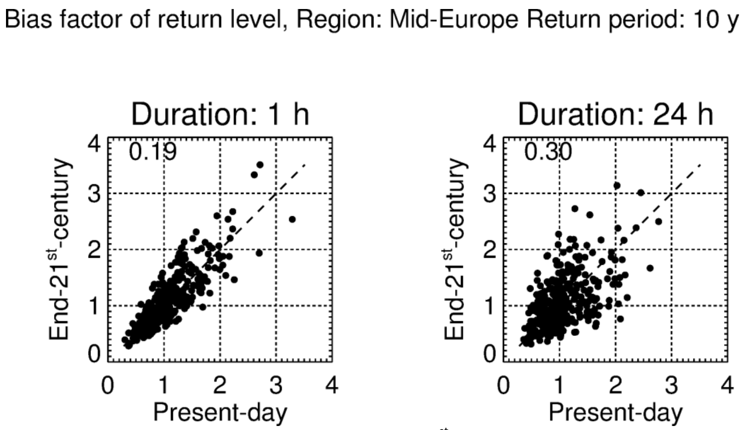

Figure 9 Relationship between present-day and end-21st-century bias factors of 10-year return levels for Mid-Europe sub-region for all pseudo-observation/model combinations. Left panel: 1 h duration and right panel: 24 h duration. Numbers in upper left corners are the ***R*** measure of relative spread. See text for details.

In this figure, the relationship between present-day and end-21st-century bias factors appears more pronounced for 1 h duration than for 24 h duration. That said, it must be borne in mind that if the point (x, y) is in the plot, so is the point $(1/y, 1/x)$, and this implies an inherent tendency to a fan-like spread of points from $(0,0)$, as seen on both plots.

Therefore, to quantify the relationship we use the measure of the relative spread introduced by Maurer et al. (2013):

$$R = \langle \frac{|B_F - B_P|}{(B_F + B_P)/2} \rangle,$$

where $\langle \cdot \rangle$ means averaging over model/pseudo-reality combinations. These R-values, given in the upper lefter corner of each panel, also support the partial relationships described above, and a stronger one for hourly duration.

These relations are important since they could explain the generally good performance of the FAC adjustment methods seen in the previous section. Suppose that $B_P = B_F$, then

$$P_T = \frac{S_T}{C_T} O_T = S_T \frac{O_T}{C_T} = S_T B_P = S_T B_F = S_T \frac{V_T}{S_T} = V_T$$

and the FAC method will therefore adjust perfectly.

https://doi.org/10.5194/hess-2020-318
Preprint. Discussion started: 7 July 2020

Hydrology and Earth System Sciences
Open Access
Discussions

We also note that daily data, due to the summation, would have less erratic behaviour than hourly and therefore we would expect any relationship to be less masked by noise for daily data than for hourly data from purely statistical grounds. Therefore, any explanation to why it is opposite should probably be found in physics or details of modelling. We will discuss this further in Section 5.3.

5 Discussion

5.1 Relation with other studies

The study by Räty et al. (2014) touches upon related issues to ours. However, our study includes smaller temporal scales (hourly and daily) than does their study and higher return periods (up to 100 years vs. the .999-fractile of daily precipitation corresponding to a return period of around 3 years). Nevertheless, the two studies agree in their main conclusion; namely that applying a bias adjustment seems to offer an additional level of realism to the processed data series, including in the climate projections, as compared to using unadjusted model results. The two studies also both support the somewhat surprising conclusion that, using present-day observations as the scenario gives a skill comparable to that of the bias adjustment methods.

Another relevant study to discuss here is Laflamme et al. (2016) who apply the BC method similar to ours to daily data from different model runs and concludes that “downscaled results are highly dependent on RCM and GCM model choice”. Finally, Kallache et al. (2011) obtained good result with the BC in a training/verification split of historical data.

5.2 Convection in RCMs

The grid spacing of present state-of-the-art RCMs available in large ensembles, such as CORDEX, is around 10 km, and at this resolution it is necessary to describe convection through parameterizations. This is obviously an important deficit for our purpose, since this could represent a systematic bias in all our simulations and therefore violate our underlying assumptions that the individual model simulations and the real-world observations behave approximately similar in a physical sense.

With the advent of convective-permitting models, a more realistic modelling of convective precipitation events is within reach and a change in the characteristics of such events is seen (Kendon et al. 2017; Lenderink et al. 2019; Prein et al. 2015). This next generation of convection-permitting RCMs with a grid spacing of a few km allows a much better representation of the diurnal cycle and convective systems as a whole (Prein et al. 2015). With that in mind, we foresee redoing the analysis when a suitable ensemble of convective-permitting RCM simulations becomes available.

https://doi.org/10.5194/hess-2020-318
Preprint. Discussion started: 7 July 2020

5.3 Stationarity of bias

The success of applying bias adjustment to climate model simulations is linked to the biases being stationary, i.e. present and future biases being more or less identical. In Section 4.3 we showed (in Figure 9) that this was the case for 1 h duration and less so for 24 h duration in our pseudo-reality setting. Such a relationship is an example of an emergent constraint (Collins et al. 2012). This is a model-based concept, originally introduced to explain that models which have a too warm (cold) present-day climate tend to have a relatively warmer (colder) future climate. The reason for this is that it is the same underlying physics which generates the present-day and future temperatures (Christensen and Boberg 2012). It has also been shown that on monthly time scales, the precipitation bias in Scandinavia depends on the precipitation (Christensen et al. 2008).

We suggest that our observed emergent constraints could be explained in a similar manner; namely as a result of the Clausius-Clapeyron relation linking atmospheric temperature changes to changes in its humidity content and thereby precipitation changes. The change prescribed by the Clausius-Clapeyron equation is usually termed the thermodynamic contribution. In addition to this, there is a dynamic contribution and this may explain the differences between the hourly and daily relation seen in Figure 9. The rationale is that hourly extremes are entirely due to convective precipitation events with almost no dynamic contribution (Lenderink et al. 2019), while daily extremes are a mixture of convective events and large-scale strong precipitation, of which the latter has a more significant dynamic contribution (Pfahl et al. 2017), causing the less marked emergent constraint for the daily time scale. This interpretation is also supported in Figure 4, in which daily precipitation sees some 'crossovers' (future return level smaller than present), whereas hourly precipitation does not have any crossovers.

5.4 The spatial scale

In the definition of model bias it is tacitly assumed that the observational dataset has the same spatial resolution as the model data. In practice, however, it is rarely possible to separate the bias from a spatial scale mismatch. For instance, if we compare modelled precipitation, which represents averages over a grid box, with rain gauge data, which represent a point, there can be a quite substantial mismatch for extreme events (Eggert et al. 2015; Haylock et al. 2008). Therefore, if the bias is adjusted towards such point values, it may lead to further complications (Maraun 2013).

Sometimes though, it is desirable to include the scale mismatch in the bias adjustment. Many impact models, e.g. hydrological models, are tuned to perform well with local observational data as input. This presents an additional challenge if this impact model is to be driven by climate model data for climate change studies, since the climate model will have biases in its climate characteristics (mean, variability, etc.) compared to those of the observed data. Applying the bias adjustment step, the hydrological model can rely on its calibration to observed conditions (Refsgaard et al. 2014; Haerter et al. 2015).

6 Conclusions

https://doi.org/10.5194/hess-2020-318
Preprint. Discussion started: 7 July 2020

Based on hourly precipitation data from a 19-member ensemble of climate simulations we have investigated the benefit of bias adjusting extreme precipitation return levels on hourly and daily time scales and evaluated the different methods. This is done in a pseudo-reality setting, where one model simulation in turn from the ensemble plays the role of observations extending into the future. The return levels obtained from each of the remaining model simulations are then bias adjusted in the present-day period, using different adjustment methods. Then the same adjustment methods are applied to end-21st-century model data to obtain projected return levels, which are then compared with the corresponding pseudo-realistic future return levels.

The main result of this inter-comparison is that applying bias adjustment methods improves projected extreme precipitation return levels, compared to using the un-adjusted model runs. Can an overall superior adjustment methodology be appointed? For hourly duration, the method to recommend (having the smallest relative error) is the simple climate factor approach FAC, which is better in terms of the relative error than the more complicated analytical quantile mapping methods based on EVA, DC and, in particular, BC. For daily duration, the OBS method performs surprisingly well, having the smallest .95-fractile relative error. Furthermore, the quantile methods perform better than FAC, with DC having the smallest relative error. These conclusions hold regardless of the sub-region considered.

Finally, we registered emergent constraints between present-day and end-21st-century biases. This was more pronounced for hourly than for daily time scale. This could be caused by hourly precipitation being more directly linked to the Clausius-Clapeyron response, but this requires more clarification in future work.

Data availability. The hourly EURO-CORDEX precipitation data are not part of the standard suite of CORDEX and are therefore not produced nor shared by all modelling groups. The data used in this study may be obtained upon request from each modelling group. The IDL code used in the analysis can be obtained from TS.

Author contribution. TS and PT designed the analysis with contribution from other co-authors and programmed the analysis software. PB, FB, OBC and PT prepared the data. TS prepared the manuscript with contributions from PT, PB, FB, OBC, BC, JHC, CS, and MSM.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The work was supported by the European Commission through the Horizon 2020 Programme for Research and Innovation under the EUCP project (Grant Agreement 776613). Part of the funding was provided by the Danish State through the Danish Climate Atlas. Some of the simulations were performed in the COPERNICUS C3S project C3S_34b (PRINCIPLES). We acknowledge the World Climate Research Programme's Working Group on Regional Climate, and the Working Group on Coupled Modelling, former coordinating body of CORDEX and responsible panel for CMIP5. We also thank the climate modelling groups (listed in Table 1 of this paper) for producing and making their model output available. We also acknowledge the Earth System Grid Federation infrastructure an international effort led by the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison, the European Network for Earth System Modelling and other partners in the Global Organisation for Earth System Science Portals

(GO-ESSP). It is appreciated that Geert Lenderink, KNMI, Claas Teichmann, GERICS and Heimo Truhetz, University of Graz made model data of hourly precipitation available for analysis.

References

Berg, P., H. Feldmann, and H.-J. Panitz, 2012: Bias correction of high resolution regional climate model data. *J. Hydrol.*, **448–449**, 80–92, https://doi.org/10.1016/j.jhydrol.2012.04.026.

Berg, P., O. B. Christensen, K. Klehmet, G. Lenderink, J. Olsson, C. Teichmann, and W. Yang, 2019: Summertime precipitation extremes in a EURO-CORDEX 0.11° ensemble at an hourly resolution. *Nat. Hazards Earth Syst. Sci.*, **19**, 957–971, https://doi.org/10.5194/nhess-19-957-2019.

Boberg, F., and J. H. Christensen, 2012: Overestimation of Mediterranean summer temperature projections due to model deficiencies. *Nat. Clim Change*, **2**, 433–436.

Buser, C., H. Künsch, and C. Schär, 2010: Bayesian multi-model projections of climate: generalization and application to ENSEMBLES results. *Clim. Res.*, **44**, 227–241, https://doi.org/10.3354/cr00895.

Christensen, J. H., and O. B. Christensen, 2007: A summary of the PRUDENCE model projections of changes in European climate by the end of this century. *Clim. Change*, **81**, 7–30, https://doi.org/10.1007/s10584-006-9210-7.

Christensen, J. H., and F. Boberg, 2012: Temperature dependent climate projection deficiencies in CMIP5 models. *Geophys. Res. Lett.*, **39**, 24705, https://doi.org/10.1029/2012GL053650.

——, ——, O. B. Christensen, and P. Lucas-Picher, 2008: On the need for bias correction of regional climate change projections of temperature and precipitation. *Geophys. Res. Lett.*, **35**, https://doi.org/10.1029/2008GL035694.

Coles, S., J. Bawa, L. Trenner, and P. Dorazio, 2001: *An introduction to statistical modeling of extreme values*. Springer,.

Collins, M., R. E. Chandler, P. M. Cox, J. M. Huthnance, J. Rougier, and D. B. Stephenson, 2012: Quantifying future climate change. *Nat. Clim. Change*, **2**, 403–409, https://doi.org/10.1038/nclimate1414.

Eggert, B., P. Berg, J. O. Haerter, D. Jacob, and C. Moseley, 2015: Temporal and spatial scaling impacts on extreme precipitation. *Atmospheric Chem. Phys.*, **15**, 5957–5971, https://doi.org/10.5194/acp-15-5957-2015.

Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor, 2016: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.*, **9**, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016.

Haerter, J. O., S. Hagemann, C. Moseley, and C. Piani, 2011: Climate model bias correction and the role of timescales. *Hydrol. Earth Syst. Sci.*, **15**, 1065–1079, https://doi.org/10.5194/hess-15-1065-2011.

Haerter, J. O., B. Eggert, C. Moseley, C. Piani, and P. Berg, 2015: Statistical precipitation bias correction of gridded model data using point measurements. *Geophys. Res. Lett.*, **42**, 1919–1929, https://doi.org/10.1002/2015GL063188.

https://doi.org/10.5194/hess-2020-318
Preprint. Discussion started: 7 July 2020

Hydrology and Earth System Sciences
Discussions

Hanel, M., and T. A. Buishand, 2010: On the value of hourly precipitation extremes in regional climate model simulations. *J. Hydrol.*, **393**, 265–273, https://doi.org/10.1016/j.jhydrol.2010.08.024.

Haylock, M. R., N. Hofstra, A. M. G. Klein Tank, E. J. Klok, P. D. Jones, and M. New, 2008: A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. *J. Geophys. Res.*, **113**, https://doi.org/10.1029/2008JD010201.

Hosking, J. R. M., and J. R. Wallis, 1987: Parameter and Quantile Estimation for the Generalized Pareto Distribution. *Technometrics*, **29**, 339, https://doi.org/10.2307/1269343.

Jacob, D., and Coauthors, 2014: EURO-CORDEX: new high-resolution climate change projections for European impact research. *Reg. Environ. Change*, **14**, 563–578, https://doi.org/10.1007/s10113-013-0499-2.

Kallache, M., M. Vrac, P. Naveau, and P.-A. Michelangeli, 2011: Nonstationary probabilistic downscaling of extreme precipitation. *J. Geophys. Res.*, **116**, https://doi.org/10.1029/2010JD014892.

Kendon, E. J., N. M. Roberts, H. J. Fowler, M. J. Roberts, S. C. Chan, and C. A. Senior, 2014: Heavier summer downpours with climate change revealed by weather forecast resolution model. *Nat. Clim. Change*, **4**, 570–576, https://doi.org/10.1038/nclimate2258.

——, and Coauthors, 2017: Do Convection-Permitting Regional Climate Models Improve Projections of Future Precipitation Change? *Bull. Am. Meteorol. Soc.*, **98**, 79–93, https://doi.org/10.1175/BAMS-D-15-0004.1.

Laflamme, E. M., E. Linder, and Y. Pan, 2016: Statistical downscaling of regional climate model output to achieve projections of precipitation extremes. *Weather Clim. Extrem.*, **12**, 15–23, https://doi.org/10.1016/j.wace.2015.12.001.

Lenderink, G., D. Belušić, H. J. Fowler, E. Kjellström, P. Lind, E. van Meijgaard, B. van Ulft, and H. de Vries, 2019: Systematic increases in the thermodynamic response of hourly precipitation extremes in an idealized warming experiment with a convection-permitting climate model. *Environ. Res. Lett.*, **14**, 074012, https://doi.org/10.1088/1748-9326/ab214a.

Maraun, D., 2012: Nonstationarities of regional climate model biases in European seasonal mean temperature and precipitation sums. *Geophys. Res. Lett.*, **39**, n/a-n/a, https://doi.org/10.1029/2012GL051210.

Maraun, D., 2013: Bias Correction, Quantile Mapping, and Downscaling: Revisiting the Inflation Issue. *J. Clim.*, **26**, 2137–2143, https://doi.org/10.1175/JCLI-D-12-00821.1.

——, 2016: Bias Correcting Climate Change Simulations - a Critical Review. *Curr. Clim. Change Rep.*, **2**, 211–220, https://doi.org/10.1007/s40641-016-0050-x.

——, and Coauthors, 2017: Towards process-informed bias correction of climate change simulations. *Nat. Clim. Change*, **7**, 764–773, https://doi.org/10.1038/nclimate3418.

Maurer, E. P., T. Das, and D. R. Cayan, 2013: Errors in climate model daily precipitation and temperature output: time invariance and implications for bias correction. *Hydrol. Earth Syst. Sci.*, **17**, 2147–2159, https://doi.org/10.5194/hess-17-2147-2013.

Hydrology and Earth System Sciences Discussions

Olsson, J., P. Berg, and A. Kawamura, 2015: Impact of RCM Spatial Resolution on the Reproduction of Local, Subdaily Precipitation. *J. Hydrometeorol.*, **16**, 534–547, https://doi.org/10.1175/JHM-D-14-0007.1.

Overeem, A., A. Buishand, and I. Holleman, 2008: Rainfall depth-duration-frequency curves and their uncertainties. *J. Hydrol.*, **348**, 124–134, https://doi.org/10.1016/j.jhydrol.2007.09.044.

Pfahl, S., P. A. O'Gorman, and E. M. Fischer, 2017: Understanding the regional pattern of projected future changes in extreme precipitation. *Nat. Clim. Change*, **7**, 423–427, https://doi.org/10.1038/nclimate3287.

Piani, C., J. O. Haerter, and E. Coppola, 2010: Statistical bias correction for daily precipitation in regional climate models over Europe. *Theor. Appl. Climatol.*, **99**, 187–192, https://doi.org/10.1007/s00704-009-0134-9.

Prein, A. F., and Coauthors, 2015: A review on regional convection-permitting climate modeling: Demonstrations, prospects, and challenges. *Rev. Geophys.*, **53**, 323–361, https://doi.org/10.1002/2014RG000475.

Räisänen, J., and O. Räty, 2013: Projections of daily mean temperature variability in the future: cross-validation tests with ENSEMBLES regional climate simulations. *Clim. Dyn.*, **41**, 1553–1568, https://doi.org/10.1007/s00382-012-1515-9.

Räty, O., J. Räisänen, and J. S. Ylhäisi, 2014: Evaluation of delta change and bias correction methods for future daily precipitation: intermodel cross-validation using ENSEMBLES simulations. *Clim. Dyn.*, **42**, 2287–2303, https://doi.org/10.1007/s00382-014-2130-8.

Refsgaard, J. C., and Coauthors, 2014: A framework for testing the ability of models to project climate change and its impacts. *Clim. Change*, **122**, 271–282, https://doi.org/10.1007/s10584-013-0990-2.

Sunyer, M., J. Luchner, C. Onof, H. Madsen, and K. Arnbjerg-Nielsen, 2017: Assessing the importance of spatio-temporal RCM resolution when estimating sub-daily extreme precipitation under current and future climate conditions. *Int. J. Climatol.*, **37**, 688–705.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An Overview of CMIP5 and the Experiment Design. *Bull. Am. Meteorol. Soc.*, **93**, 485–498, https://doi.org/10.1175/BAMS-D-11-00094.1.

Themeßl, M. J., A. Gobiet, and A. Leuprecht, 2011: Empirical-statistical downscaling and error correction of daily precipitation from regional climate models. *Int. J. Climatol.*, **31**, 1530–1544, https://doi.org/10.1002/joc.2168.

——, ——, and G. Heinrich, 2012: Empirical-statistical downscaling and error correction of regional climate models and its impact on the climate change signal. *Clim. Change*, **112**, 449–468, https://doi.org/10.1007/s10584-011-0224-4.

Trenberth, K. E., A. Dai, R. M. Rasmussen, and D. B. Parsons, 2003: The Changing Character of Precipitation. *Bull. Am. Meteorol. Soc.*, **84**, 1205–1218, https://doi.org/10.1175/BAMS-84-9-1205.