Identifying robust bias adjustment methods for European extreme precipitation in a multi-model 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], Marianne Sloth Madsen[1], Christian Steger[5]

[1] Danish Meteorological Institute, Copenhagen , 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, Copenhagen, Denmark
[4] NORCE Norwegian Research Centre, Bjerknes Centre for Climate Research, Bergen, Norway
[5] Deutscher Wetterdienst, 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 occurrence of these rare events. These are often projected using data from Regional Climate Model (RCM) 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; these biases can be in the mean and/or in the higher moments. Therefore, the RCM results are adjusted to account for these deficiencies. However, this does not guarantee that adjusted projected results will match future reality better, since the bias may not be stationary in a changing climate. In the present work we evaluate different 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-observations, against which the remaining model simulations are adjusted and validated. The study uses hourly data from historical and RCP8.5 scenario runs from 19 model simulations from the EURO-CORDEX ensemble at 0.11° resolution. Fields of return levels for selected return periods are calculated for hourly and daily time scales based on 25 years long time slices representing present-day (1981-2005) and end-21st-century (2075-2099). The adjustment techniques applied to the return levels are based on extreme value analysis and include climate factor and quantile mapping approaches. Generally, we find that future return levels can be improved by adjustment, compared to obtaining them from raw scenario model data. The performance of the different methods depends on the time scale considered. On hourly time scale, the climate factor approach performs better than the quantile mapping approaches. On daily time scale, the superior approach is to simply deduce future return levels from pseudo-observations and the second best choice is using the quantile mapping approaches. These results are found in all European sub-regions considered. Applying the inter-model cross-validation against model ensemble medians instead of individual models does not change overall conclusions much.

1 Introduction

Severe precipitation events occur typically either as stratiform precipitation of moderate intensity or as intense localized cloudbursts lasting up to 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, the formation of convective systems, and typical low pressure tracks. Therefore also the statistics of the occurrence of severe precipitation events will most likely change.

Global climate models (GCMs) are 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 forcing 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.

An alternative to dynamical downscaling is statistical downscaling. Here large-scale circulation patterns (e.g. the North Atlantic Oscillation) are related to small-scale variables, such as precipitation mean at a station. One assumes that the large-scale circulation pattern is modelled well by the GCM and therefore the approach is called perfect prognosis. Using the relationship with the small-scale variables, calibrated on observations, one can obtain modelled local-scale variables (present-day and future) from the modelled large-scale patterns. A recent overview of these methods and validation of them can be found in Gutiérrez et al. (2019).

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 substantial biases even at this resolution. Recently, Berg et al. (2019) evaluated high resolution RCMs from the EURO-CORDEX ensemble (Jacob et al., 2014) also used here and reached similar conclusions for several countries across Europe: RCMs underestimate hourly extremes and give an erroneous spatial distribution.

85

86 Extreme convective precipitation of short duration is thus one of the more challenging phenomena to
87 represent physically accurate in RCMs. The reason is that convective events take place on a spatial scale
88 comparable to the RCM grid spacing of presently around 10 km. Therefore, the convective plumes cannot
89 be directly modelled. Instead, the effects of convection are parametrised, i.e. modelled as processes on
90 larger spatial scales (Arakawa, 2004). Thus, the inability to reproduce these short duration extremes can be
91 explained by the imperfect parametrization of sub-grid scale convection (Prein et al., 2015), which generally
92 leads to too early onset of convective rainfall in the diurnal cycle and subsequent dampening of the build-
93 up of convective available potential energy (Trenberth et al., 2003).

94

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

104

105 Both adjustment approaches come in several flavours. In the simplest one, the transformation consists of
106 an adjustment of the mean, in the case of precipitation by multiplying the mean by a factor. In the more
107 elaborate flavour, the transformation is defined by quantile mapping, preserving also the higher moments.
108 Quantile mapping can use either empirical quantiles or analytical distribution functions. The ability of
109 quantile mapping to reduce bias has been demonstrated for daily precipitation in present-day climate using
110 observations, which are split into calibration and validation samples (Piani et al., 2010; Themeßl et al.,
111 2011).

112

113 Bias adjustment techniques originate in the field of weather and ocean forecast modelling, where they are
114 known as model output statistics (MOS). Here output from a forecast model is adjusted for model
115 deficiencies and local features not explicitly resolved by the model. Applying similar adjustment techniques
116 to climate model simulations, however, has a complication not present in forecast applications: Climate
117 models are set up and tuned to present-day conditions and verified against observations, but then applied
118 to future changed conditions without any possibility to directly verify the model's performance under these
119 conditions. Therefore, showing that bias adjustment works for present-day climate is a necessary but not
120 sufficient condition for the adjustment to work in the changed climate.

121

122 A central concept of adjustment methods is the assumption of *stationarity* of the bias. For bias correction
123 this means that the transformation from model to observations is unchanged from the present-day climate
124 to the future climate, while for delta-change the transformation from present-day climate to future climate
125 is unchanged from model to observations. In the ideal case of stationarity being fulfilled, the adjustment
126 methods will work perfectly and produce perfect future projections. If stationarity is not fulfilled,
127 adjustment may improve projections, or in the worst cases they may degrade projections, compared to

128 using raw model output. We also note that the adjustment methods themselves may influence the climate
129 change signal of the model, depending on the bias and the method used (Berg et al., 2012; Haerter et al.,
130 2011; Themeßl et al., 2012).

131

132 Stationarity has been debated in recent years in the literature (e.g. Boberg and Christensen, 2012; Buser et
133 al., 2010). Kerkhoff et al. (2014) review and discuss two hypotheses: 1) constant bias: unchanged between
134 present-day and future (i.e. stationarity) and 2) constant relation: bias varies linearly with the signal. Van
135 Schaeybroeck and Vannitsem (2016) used a pseudo-reality setting with a simplified model and found large
136 changes in the bias between present-day and future for many variables and violation of both constant bias
137 and constant relation hypothesis. Chen et al. (2015) concluded that precipitation bias is clearly non-
138 stationary over North America in that variations in bias is comparable to the climate change signal.
139 Velázquez et al. (2015) used a pseudo-reality setting involving two models and concluded that constancy of
140 bias was violated for both precipitation and temperature on monthly time scale. Hui et al. (2019) used a
141 pseudo-reality setting with GCMs and found significant non-stationarity of bias for annual and seasonal
142 temperatures. Besides, they point to a large effect on non-stationarity from internal variability.

143

144 To thoroughly validate adjustment methods, both a calibration dataset and an independent dataset for
145 validation are needed. There are two different approaches to obtain this. In split-sample testing, the
146 observations are divided into calibration and validation parts, often in the form of a cross-validation (e.g.
147 Gudmundsson et al., 2012; Li et al., 2017a, 2017b; Refsgaard et al., 2014; Themeßl et al., 2011). A variant is
148 differential split-sample testing (Klemeš, 1986), where the split in calibration/and validation parts is based
149 on climatological factors, such as wet and dry years, encompassing climate changes and variations into the
150 validation.

151

152 An alternative approach, which we use here, is *inter-model cross-validation*, as pursued by Maraun (2012),
153 Räisänen and Räty (2013) and Räty et al. (2014) and others. The rationale is here that the members in a
154 multi-model ensemble of simulations represent different descriptions of physics of the climate system, with
155 each of them being not too far from the real climate system. Thus, one member of the ensemble
156 alternatively plays the role of *pseudo-observations*, against which the remaining adjusted models are
157 validated. Thus, the trick is that we know both present and future pseudo-observations.

158

159 The advantage of inter-model cross-validation, is that the adjustment methods are calibrated under
160 present-day conditions and validated under future climatic conditions. Therefore, it embraces modelled
161 physical changes between present and future climate, as for instance a shift in the ratio between stratiform
162 and convective precipitation. In this respect it is a more realistic setting than validation based on split-
163 sample test. Also, model and pseudo-observations have the same spatial scale, thus avoiding comparing
164 pointwise observations with area-averaged model data, as is done in the split-sample testing. On the other
165 hand, the method assumes that the modelled present-day is not too different from observations. If this is
166 violated, the method will give too optimistic error estimates compared to what can be expected in the real
167 World. Please cf. also further discussion in Section 5.2.

168

169 Inter-model cross-validation has been applied on daily precipitation to evaluate different adjustment
170 methods (Räty et al., 2014). Here we apply a similar methodology European-wide to extreme precipitation

on hourly and daily time scales. This has been made possible with the advent of the EURO-CORDEX, a large ensemble of high-resolution RCM simulations with precipitation at hourly time-resolution. Being more specific, we 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 adjusted return levels perform better, according to the inter-model cross-validation, than using raw 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 adjustment methods?

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 conclusions (Section 6).

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 years 1850-2005 followed by the RCP8.5 scenario for the years 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 model code 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.

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				

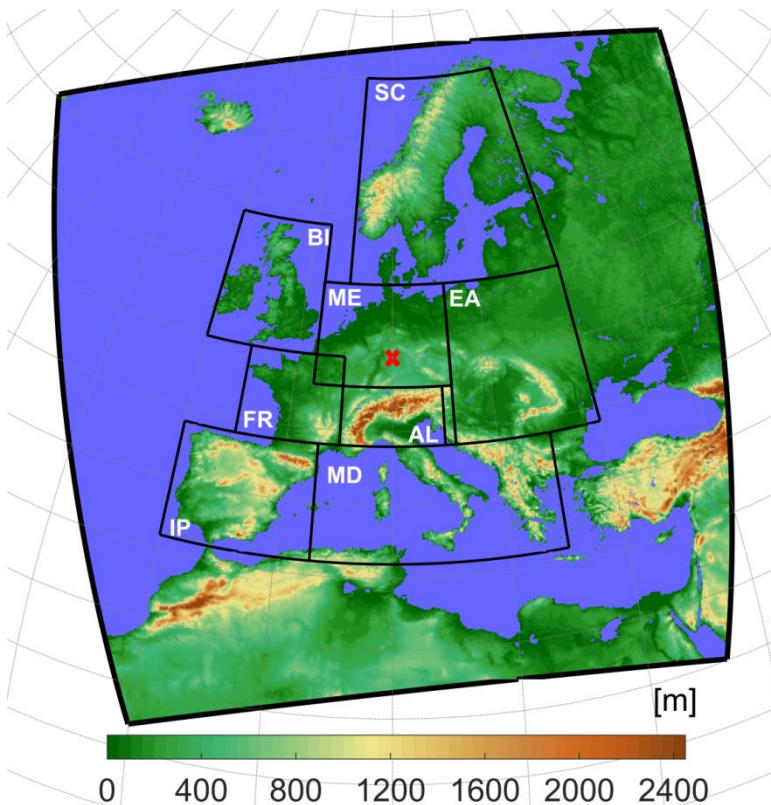



Figure 1. Map showing the EURO-CORDEX region (outer frame) with elevation in colours. PRUDENCE sub-regions (Christensen and Christensen, 2007) used in the analysis are also shown: BI = British Isles, IP = Iberian Peninsula, FR = France, ME = Mid-Europe, SC = Scandinavia, AL = Alps, MD = Mediterranean,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

Generally, GCM results are quite comparable to reality, and many validation studies of GCMs exist, also with an eye on Europe (e.g. McSweeney et al., 2015). We are aware of the use in some papers of selection procedures for selecting how to choose sub-sets of available GCMs (e.g. McSweeney et al., 2015; Rowell, 2019). There is, however, no simple quality index that can be generally applied. Any discrimination of GCMs depends on area, season, and the meteorological field and property being investigated (Gleckler et al., 2008; e.g. their Fig. 9). Furthermore, these tests and selection procedures are based on subjective

228 criteria and come with major caveats that impact the uncertainty range largely (Madsen et al., 2017). We
229 therefore choose, in accordance with most other similar studies, to use an 'ensemble of opportunity' for
230 the present study.
231

3 Methods

3.1 Duration

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

3.2 Extreme value analysis

244 Extreme value analysis (EVA) provides methodologies to estimate high quantiles of a statistical distribution
245 from observations. The theory relies on fundamental convergence properties of time series of extreme
246 events; for details we refer to Coles (2001).
247

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

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

257 The parameter σ is the scale and is a measure of the width of the distribution. The parameter ξ is the shape
258 and describes the character of the upper tail of the GPD-distribution; $\xi > 0$ implies a heavy tail which
259 usually is the case for extreme precipitation events, while $\xi < 0$ implies a thin tail. Note that, quite
260 confusingly, an alternative sign convention of ξ occurs in the literature (e.g. Hosking and Wallis, 1987).
261

262 If we now consider an arbitrary level x with $x > x_0$, the average number of exceedances per year of x will
263 be
264

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

266

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

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

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

270

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

272   from which

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

274

275

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

280

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

288

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

293

294   In practical applications of EVA the parameters are estimated with large uncertainties due to limited length
295   of the time series. The threshold has the smallest relative uncertainty, the scale has a larger relative
296   uncertainty, and the shape has the largest relative uncertainty. Therefore, also the relative uncertainty of
297   the return levels increase with increasing T, as can be seen from Eq. 2.

298

299   ## 3.3   Bias adjustments and extreme value analysis
300   The delta-change and bias correction approaches were introduced in general terms in Section 1. Now we
301   will formulate EVA-based analytical quantile mapping based versions of the two approaches. In what
302   follows O_T is the T-year return levels estimated from present-day pseudo-observations, while C_T (control)
303   and S_T (scenario) denote the corresponding return levels, estimated from present-day and end-21st-century
304   model data, respectively. Finally, P_T (projection) denotes the end-21st-century return level after bias-
305   adjustment has been applied.

306

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.

3.3.2 Analytical quantile mapping based on EVA

In the EVA-based quantile mapping, 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 the 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

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

For the delta-change approach (DC), the modelled GPD distribution functions for present-day and end-21st-century conditions are quantile mapped 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 mapped to obtain the model bias, which is then applied, using 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

345 verifying predictions, where the performance of the prediction should be superior to the performance of
346 reference predictions, such as persistence or climatology.
347
348 We choose two reference methods. One reference is to simply use, for a given model, the return level
349 calculated from (pseudo-)observations as the projected return level (OBS),

$$P_T = O_T$$

350
351 Another reference is to use the raw scenario model output data without any adjustment (SCE):

$$P_T = S_T.$$

353
354 For an overview of methods, see Table 2
355
356 Table 2. Overview of methods used in the inter-comparison

OBS	(Pseudo-)observations (Reference method)
SCE	Raw RCM scenario (Reference method)
FAC	Climate factor on return levels
DC	Quantile mapped delta-change based on EVA
BC	Quantile mapped bias correction based on EVA

357
358
359

3.4 The inter-model cross-validation procedure in detail

361 The inter-model cross-validation goes in detail as follows: Each of the N models are successively regarded
362 as being pseudo-observations. The individual adjustment methods are calibrated on the present-day parts
363 of the pseudo-observations and model return levels (present-day and end-21st-century), as appropriate
364 depending on whether it is a bias correction or delta-change method. The calibration is done as described
365 above. The adjustment methods are then applied to present-day observation and model data, again as
366 appropriate, to obtain end-21st-century adjusted return levels. These are then validated against the end-
367 21st-century return level from pseudo-observations.
368
369 The basic validation metric will be the relative error of end-21st-century return levels for a given duration
370 and return period T:
371
372

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

373
374 i.e. the absolute difference between the projected return level P_T obtained from using adjustment and the
375 validation return level V_T estimated from end-21st-century pseudo-observations, divided by the validation
376 return level. This metric is calculated for every grid point and for every combination of model/pseudo-
377 observations. Since we have $N = 19$ model simulations in the ensemble, we have $N \times (N - 1) = 342$
378 different combinations for validating each adjustment method and make statistics of the relative error. This
379 quantifies the average performance of the different methods.
380

381 User-end scenarios are often constructed as the median or mean from ensembles. We also tested this in
382 the inter-model cross-validation setup. The calibration is performed as before on each of the remaining
383 models and adjusted return levels for the end-21st-century calculated. But then the median of these
384 adjusted future return levels is calculated and this is validated against the future pseudo-observations.
385 Note that this gives only $N = 19$ different combinations and therefore a less robust statistics compared to
386 above.
387

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 relative change from present-day to end-21st-century (right). In each grid point, values are the median return level over all 19 model simulations.

397 Figure 2 displays the geographical distribution of the 10-year return level for precipitation intensity of 1 h
398 duration, calculated as the median return level over all 19 model simulations. The smallest return levels are
399 mainly found in the arid North African region and to some extent in the Norwegian Sea, while the largest
400 return levels are found in southern Europe and in the Atlantic northwest of the Iberian Peninsula.

401 Mountainous regions, such as the Alps and western Norway stand out as have higher return levels than
402 their surroundings. This supports that the models are not totally unrealistic in modelling extreme
403 precipitation.
404
405 There is a general increase in the range of 20-40 % from present-day to end-21st-century climatic
406 conditions. The relative changes are geographically quite uniform across the area. For instance, no evident
407 difference between land and sea appears. Likewise do the mountainous regions not stand out from the
408 surroundings.
409
410

Return level, Duration: 24 h, Return period: 10 y

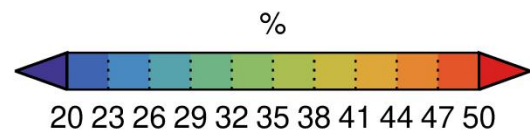
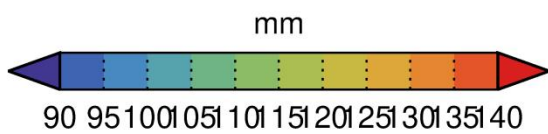
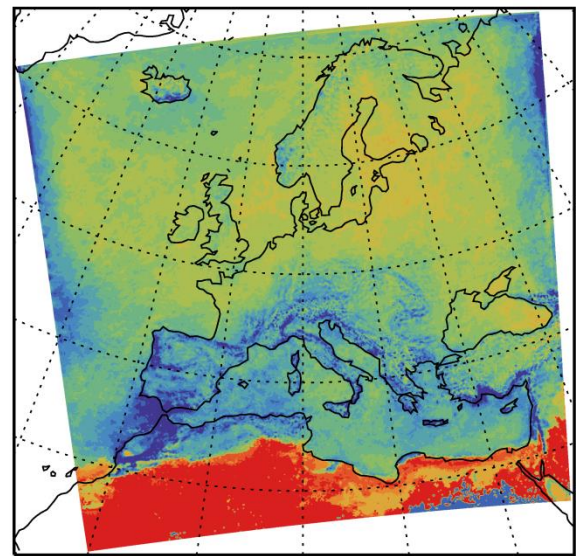
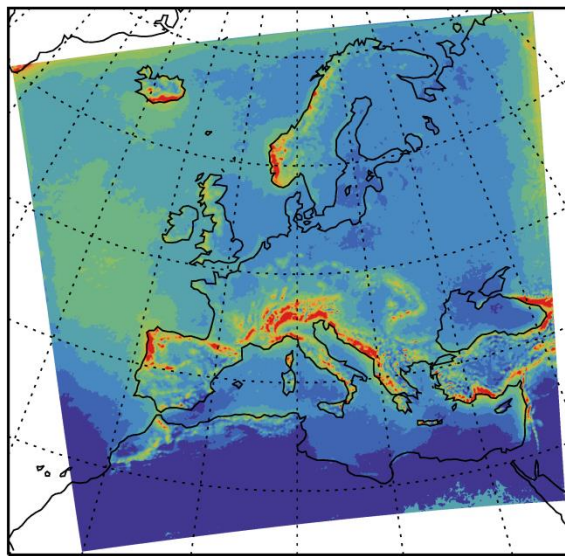
Present-day

Rel. change
Present-day to End-21st-century

mm

90 95 100 105 110 115 120 125 130 135 140

%

20 23 26 29 32 35 38 41 44 47 50

411
412 Figure 3. As Figure 2 but for 24 h duration
413
414 We also show in Figure 3 the median 10-year return level for 24 h duration. Again, the largest return levels
415 are found in southern Europe and northwest of the Iberian Peninsula. Also, the mountainous regions stand
416 out with higher return levels even more pronounced than for 1 h duration. The return levels generally
417 increase from present-day to end-21st-century conditions with around the same percentage as for 1 h
418 duration and also geographically homogeneous.
419 .
420
421

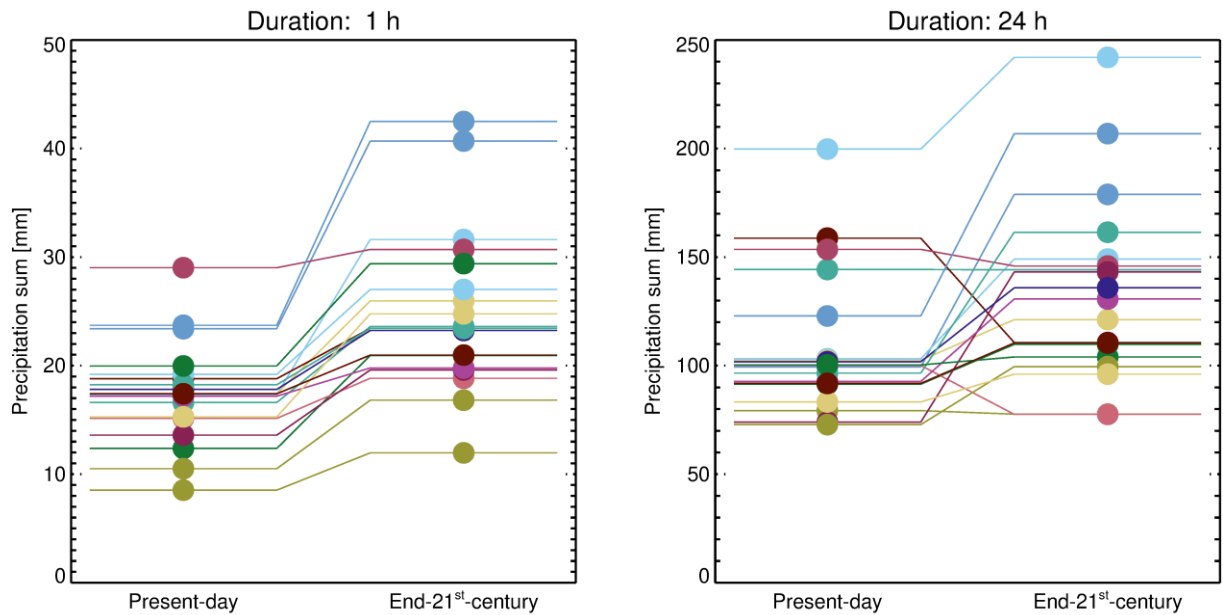



Figure 4. Modelled return levels at 50N/10E (northern Germany, marked with 'X' in Figure 1) for present and future for 10 y return period and 1 h and 24 h durations. Different colours represent the 19 different GCM-RCM simulations listed in Table 1.

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. This behaviour is common to all regions. 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. This is probably a combined effect of different climate signals in different models and natural variability (Aalbers et al., 2018).

4.2   Inter-model cross-validation

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 combinations of model/pseudo-observations. Second, we will, for each adjustment method and for each combination of model/pseudo-observations, 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 combinations of model/pseudo-observations by showing the median and the 5/95-percentiles of this distribution.

4.2.1   Results for 1 h duration

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

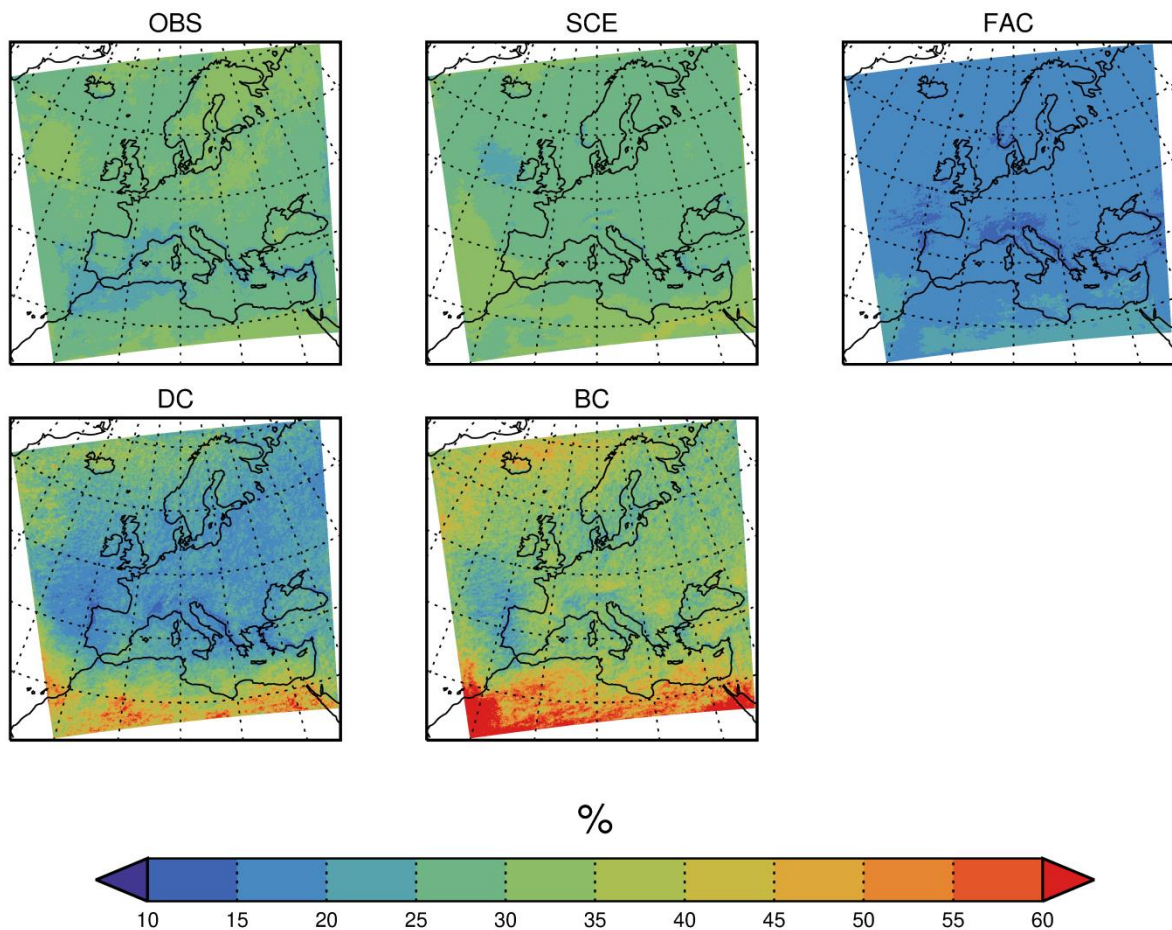
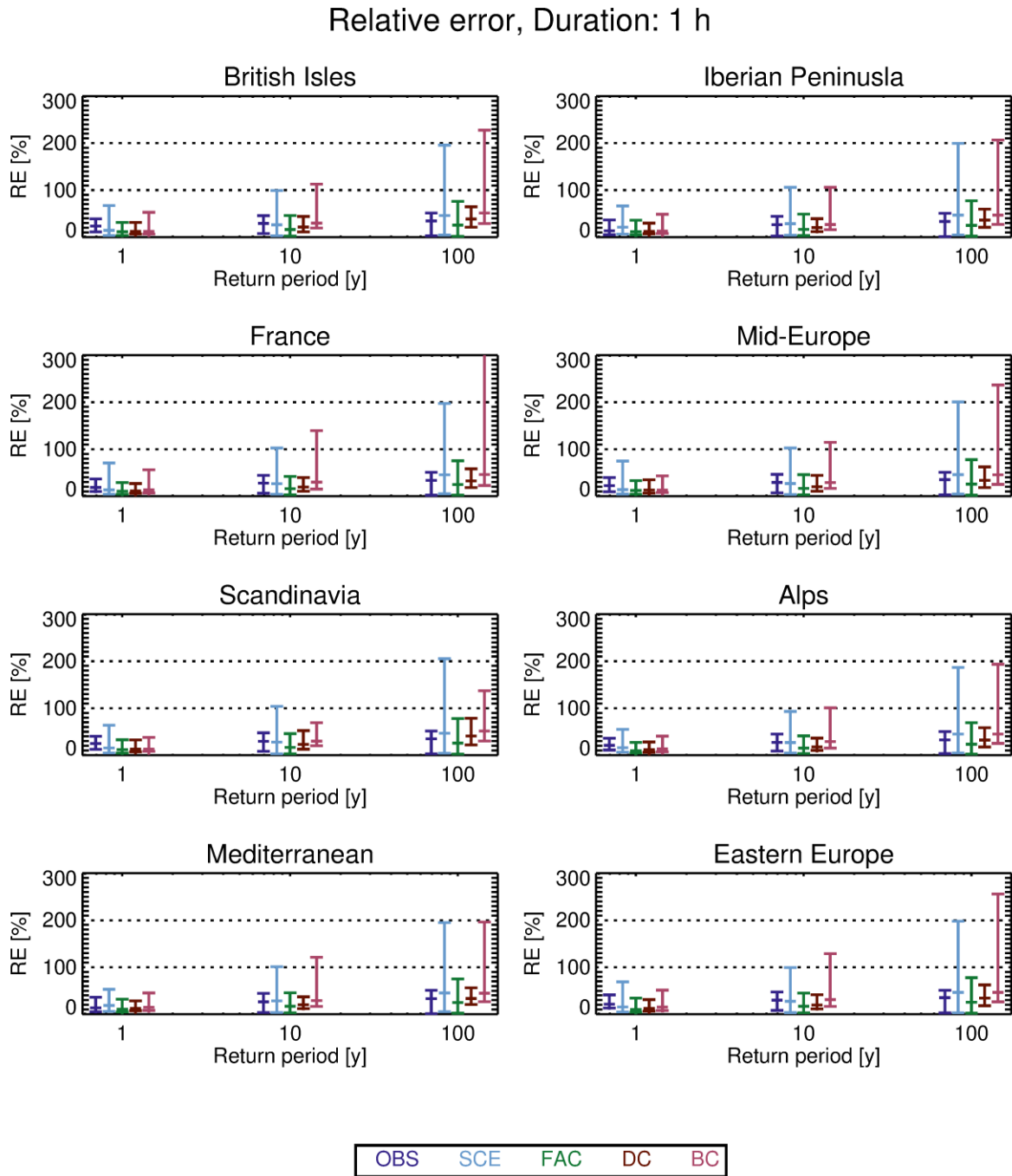



Figure 5. Geographical distribution of the relative error of end-21st-century 10 year return level for 1 h duration precipitation intensity from the inter-model cross-validation. Colours show the median of the relative error calculated over all model/pseudo-observations combinations. Panels are for the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

First we look at the reference methods. Relative errors from the OBS method are in the range of 20-40 %. 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 25-45 %, lowest values in the Atlantic west of Ireland; largest values over parts of the Atlantic and northern Africa. The two reference methods give rather similar results, but the OBS method slightly outperforms SCE in the south, while the opposite is true in the north.

The relative error of FAC is below 20 % 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,

467 the relative error of BC is everywhere above both DC and FAC, indicating the poorest performance of all
468 methods considered.
469

470
471 Figure 6. Statistical distribution (median and 5th/95th percentile) of the relative error of the inter-model cross-validation for 1 hour
472 duration for 1 y, 10 y and 100 y return periods. Panels represent PRUDENCE sub-regions shown in Figure 1. Each colour represents
473 a adjustment method (see Table 2).
474

475   The statistical distribution of the relative error is shown in Figure 6 for the eight PRUDENCE sub-regions

476   (see Figure 1). We first note that the distribution of relative error is shifted towards higher values for larger

477   return periods, as expected. Next, we note that the two reference methods, OBS and SCE, behave

478   differently. SCE generally has a little larger median relative error, but the 95th percentile is much larger for

479   SCE than for OBS, in particular for large return periods. Thus, OBS overall performs better than SCE,

480   meaning that using present-day pseudo-observations to estimate projected end-21st-century return levels

481   yields better relative error than using raw modelled scenario data.

482

483   The FAC method generally has the best overall performance, both in terms of median and 95th percentile of

484   the relative error. The DC method has a slightly poorer performance than FAC, both in terms of the median

485   and the 95th percentile of the relative error. Finally, BC has poorer performance than DC, when comparing

486   the median of the relative error and in particular for the 95th percentile.

487

488   In summary, for 1 h duration, the method with the best performance is using a climate factor on the return

489   levels (FAC). This method outperforms both reference methods and the more sophisticated methods based

490   on quantile mapping, DC and BC, the latter having the poorest overall performance of them all. Note that

491   DC is comparing GPDs from the same model, whereas BC is comparing GPDs from different models. If the

492   difference, in terms of GPD parameters, between two models in the present-day climate is typically larger

493   than the difference between the same model in present-day and end-21st-century climate, it can explain

494   the different results.

495

496

497   **4.2.2   Results for 24 h duration**

498

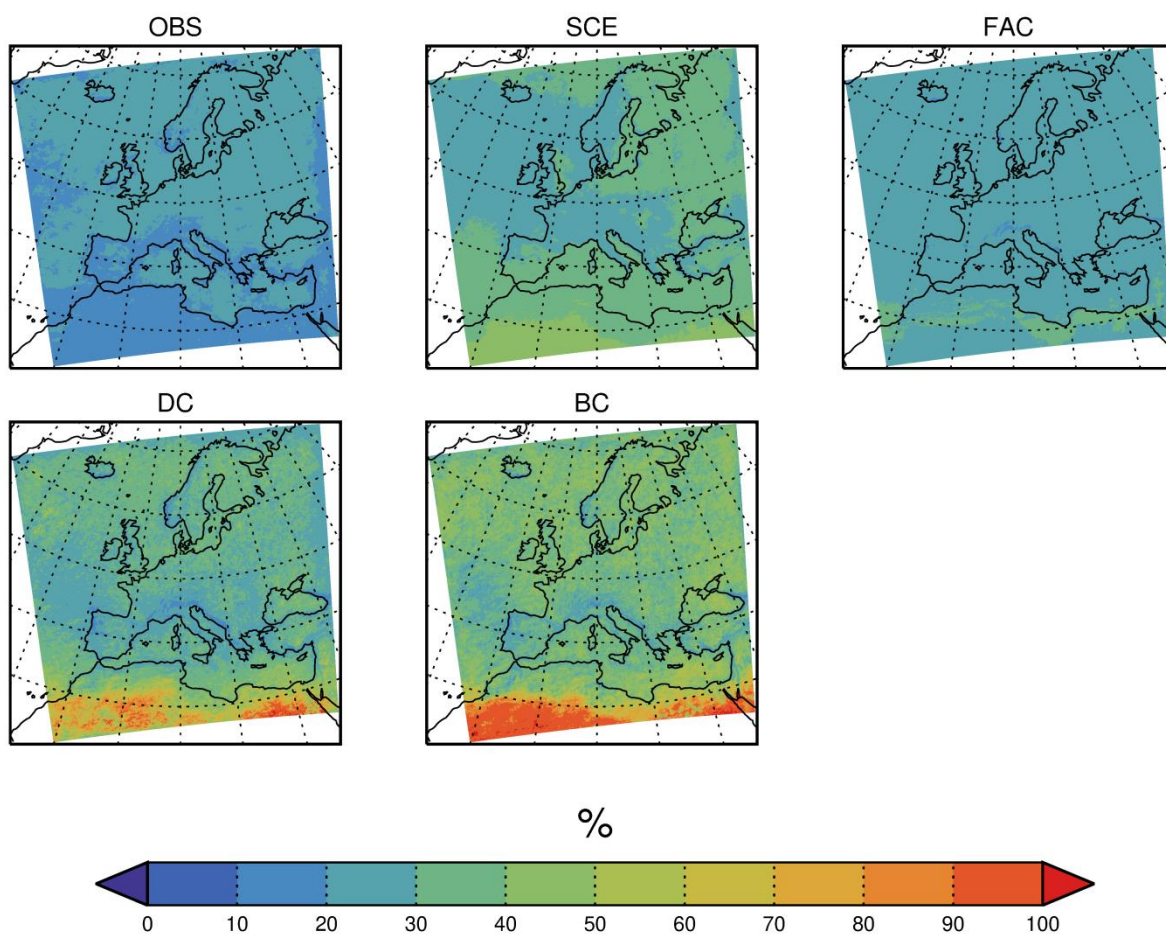
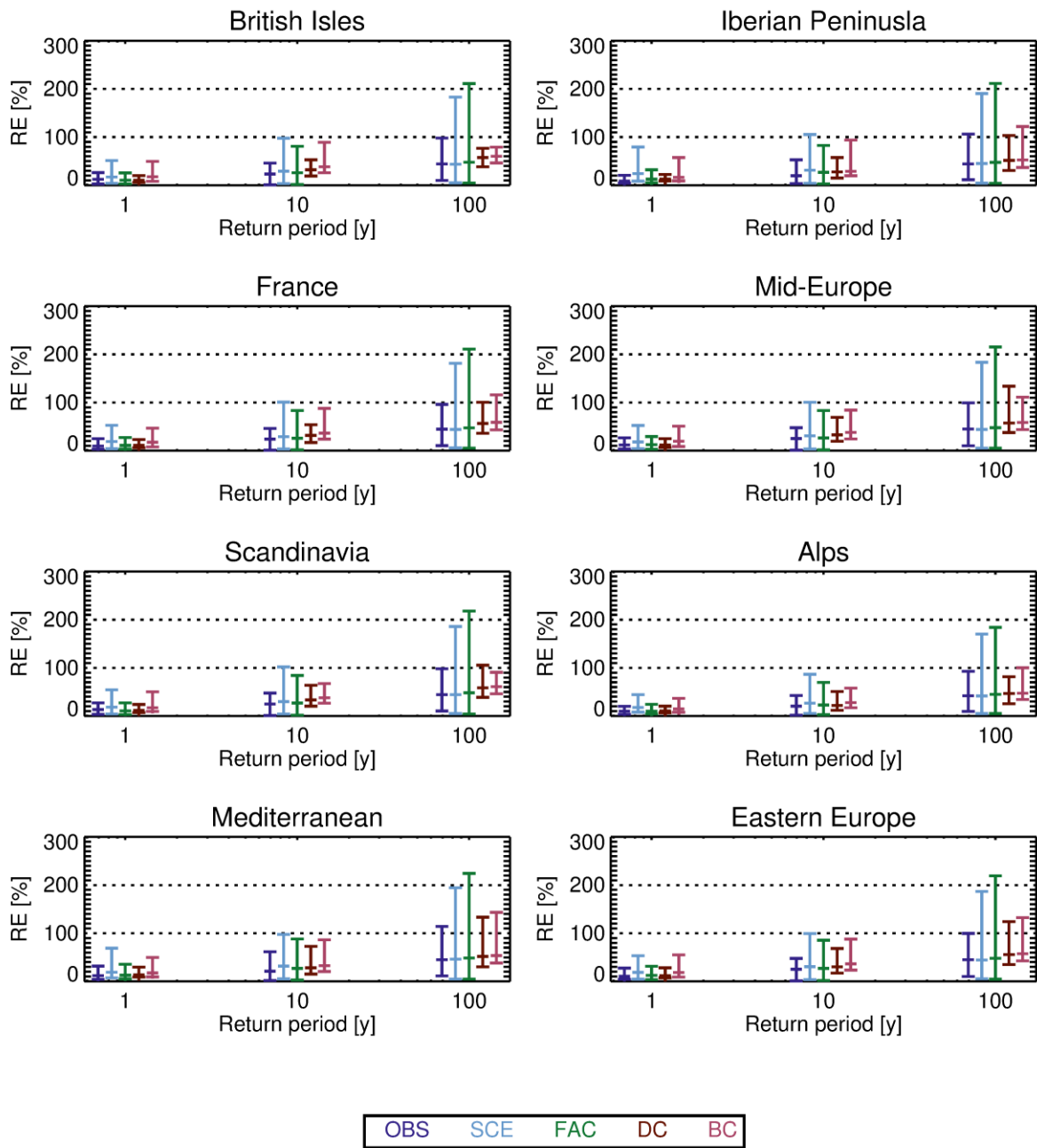



Figure 7. As Figure 5 but for 24 h duration.

For 24 h duration (see Figure 7), OBS has the lowest median relative error (less than 30 %) in most regions of all the adjustment methods, while SCE has higher relative error in the interval 30-60 % approximately, with the highest values in North Africa. FAC has relative errors in-between those of OBS and SCE. Of the quantile mapping methods, DC has relative errors in the interval 20-80 % 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.

508
509 Figure 8. As Figure 6 but for 24 h duration

510

511 As for the 1 h duration, we also compare the entire statistical distribution of the relative error of the

512 different adjustment methods for all three return periods (Figure 8), and again, both median and 95th

513 percentile of the relative error increases for larger return periods, as expected. Further, OBS seems,

514 surprisingly, to have a small median relative error and the smallest 95th percentile of all methods

515 considered for all sub-regions. SCE has a median not too different from that of OBS, but the 95th percentile

516   is much larger. Similar characteristics hold for FAC. The quantile mapping methods DC and BC have slightly

517   larger median values, but the 95th percentile is smaller than for FAC. All these characteristics hold for all

518   sub-regions.

519

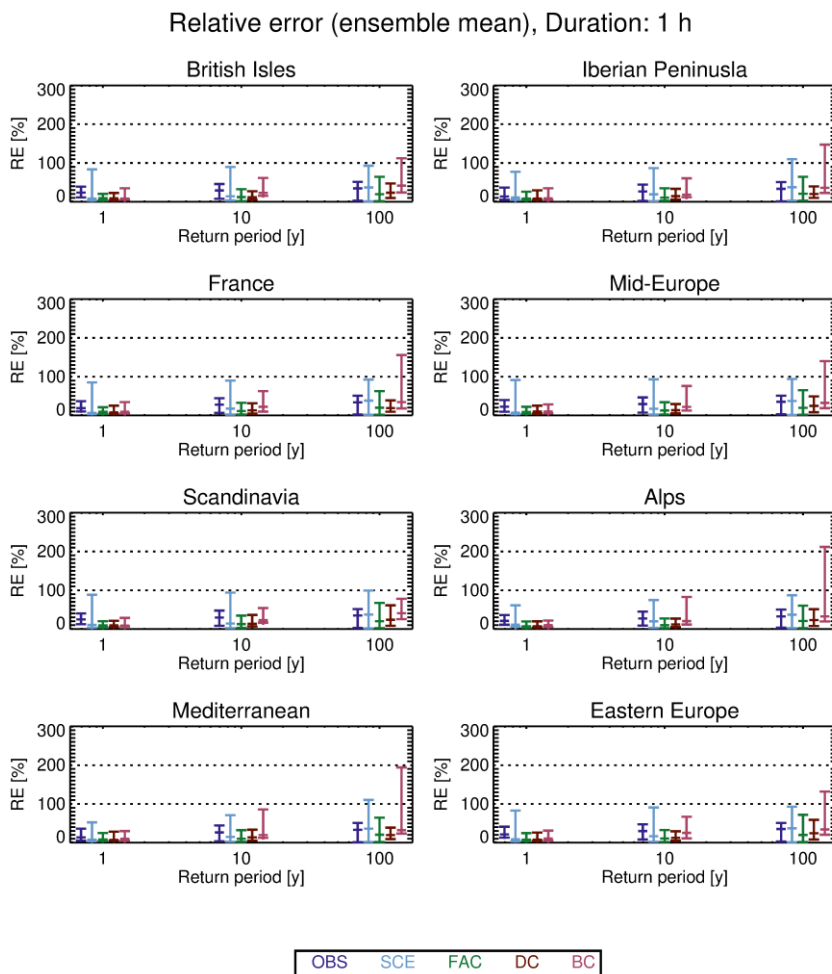
4.2.3   Ensemble median

521   Also inter-model cross-validation of pseudo-observations against model ensemble median, as described in

522   Section 3.4, was carried out. For duration 1 h, distribution of the relative error is shown in Figure 9. By

523   comparing with Figure 6, the distribution of the relative error does not change much overall. However, for

524   many of the sub-regions considered and for the longer return periods, the FAC and BC have a smaller 95th

525   percentile for cross-validation against model ensemble means, than against individual models.

526
527   Figure 9. As Figure 6 but for inter-model cross-validation against ensemble medians.
528

529   Also for 24 h duration the distribution of the relative errors does not change much when shifting to

530   validation against ensemble median (not shown).

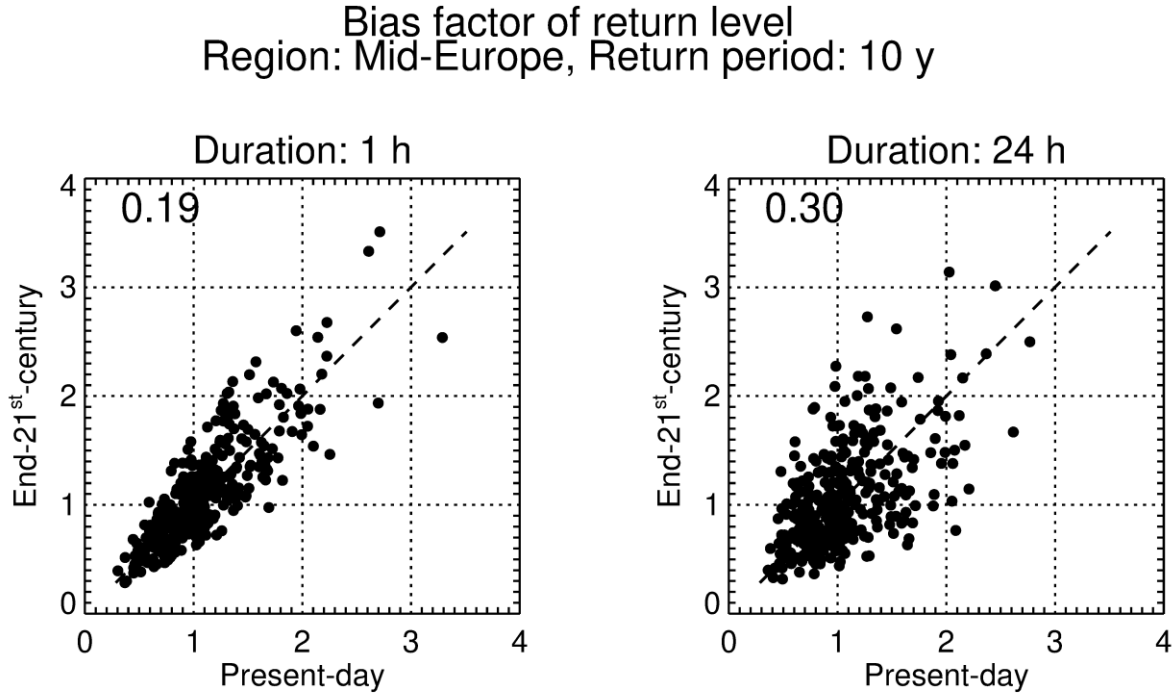
4.3   Further analysis on conditions for skill

532

533 To get further insight into the difference in performance between hourly and daily precipitation, we

534 consider for a given return period the relationship between the bias factor for present-day $B_{P,T} = \frac{C_T}{O_T}$ and

535 end-21st-century $B_{F,T} = \frac{S_T}{V_T}$ for all model/pseudo-observations combinations (see Figure 10).

536

537
538 Figure 10. Relationship between present-day and end-21st-century bias factors of 10-year return levels for Mid-Europe sub-region
539 for all pseudo-observation/model combinations. Left panel: 1 h duration and right panel: 24 h duration. Numbers in upper left
540 corners are the **R** indices. See text for details.

541

542 In this figure, the relationship between present-day and end-21st-century bias factors appears more

543 pronounced for 1 h duration than for 24 h duration. That said, it must be borne in mind that if the point

544 (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

545 points from $(0,0)$, as seen on both plots.

546

547 To quantify the strength of the above relationship, we define an index:

548 $$R = \langle \frac{|B_F - B_P|}{(B_F + B_P)/2} \rangle,$$

549 where $\langle \cdot \rangle$ means averaging over combinations of model/pseudo-observations. This index is an extension of

550 the index introduced by Maurer et al. (2013). It is the ensemble average of the relative absolute difference

551 between the present-day and future bias. A value of $R = 0$ means these biases are equal, i.e. perfect

552 stationarity; and the smaller the value of R, the closer to stationarity (in an ensemble sense).

553

554 Values of R are given in the upper left corner of each panel of Figure 10 and they also support the partial

555 relationships described above, and a stronger one for hourly duration. These relations are important since

556 they could explain the generally good performance of the FAC method seen in the previous section.

557 Suppose that $B_{P,T} = B_{F,T}$, then

558 $P_T = \frac{S_T}{C_T} O_T = S_T \frac{O_T}{C_T} = S_T B_P = S_T B_F = S_T \frac{V_T}{S_T} = V_T$

559

and the FAC method will therefore adjust perfectly.

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) and higher return periods (up to 100 years vs. the 99.9[th] percentile 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 both support, in agreement with our study, the somewhat surprising conclusion that using present-day (pseudo-)observations as the scenario gives a skill comparable to that of the bias adjustment methods.

Kallache et al. (2011) proposed a correction method for extremes, CDF-t, and obtained good validation result with calibration/validation split of historical data from Southern France. The CDF-t method was applied by Laflamme et al. (2016) on daily New England data and concludes that "downscaled results are highly dependent on RCM and GCM model choice".

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 similarly in a physical sense. Thus, we do not promote naively applying the presented adjustment methods to hourly data from these models. Instead, the present work should be seen as a statistical exercise and the methods can in the future be applied to convection permitting model simulations that better represent the convective process. The results from the present work would apply equally to that case.

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

599 whole (Prein et al., 2015). With that in mind, we foresee redoing the analysis when a suitable ensemble of
600 convective-permitting RCM simulations becomes available.
601

5.3 Stationarity of bias

603 The success of applying bias adjustment to climate model simulations is linked to the biases being
604 stationary, i.e. present and future biases being more or less identical. In Section 4.3 we showed (in Figure
605 10) that this was the case for 1 h duration and less so for 24 h duration in our pseudo-reality setting. Such a
606 relationship is an example of an emergent constraint (Collins et al., 2012). This is a model-based concept,
607 originally introduced to explain that models which have a too warm (cold) present-day climate tend to have
608 a relatively warmer (colder) future climate. The reason for this is that it is the same underlying physics
609 which generates the present-day and future temperatures (Christensen and Boberg, 2012).
610

611 We suggest that our observed emergent constraints could be explained in a similar manner; namely as a
612 result of the Clausius-Clapeyron relation linking atmospheric temperature changes to changes in its
613 humidity content and thereby precipitation changes. The change prescribed by the Clausius-Clapeyron
614 equation is usually termed the thermodynamic contribution. In addition to this, there is a dynamic
615 contribution and this may explain the differences between the hourly and daily relation seen in Figure 10.
616 The rationale is that hourly extremes are entirely due to convective precipitation events with almost no
617 dynamic contribution (Lenderink et al., 2019), while daily extremes are a mixture of convective events and
618 large-scale strong precipitation, of which the latter has a more significant dynamic contribution (Pfahl et al.,
619 2017), causing the less marked emergent constraint for the daily time scale. This interpretation is also
620 supported in Figure 4, in which daily precipitation sees some 'crossovers' (future return level smaller than
621 present), whereas hourly precipitation does not have any crossovers.
622

5.4 The spatial scale

624 In the definition of model bias it is tacitly assumed that the observational dataset has the same spatial
625 resolution as the model data. In practice, however, it is rarely possible to separate the bias from a spatial
626 scale mismatch. For instance, if we compare modelled precipitation, which represents averages over a grid
627 box, with rain gauge data, which represent a point, there can be a quite substantial mismatch for extreme
628 events (Eggert et al., 2015; Haylock et al., 2008). Therefore, if the bias is adjusted towards such point
629 values, it may lead to further complications (Maraun, 2013).
630

631 Sometimes though, it is desirable to include the scale mismatch in the bias adjustment. Many impact
632 models, e.g. hydrological models, are tuned to perform well with local observational data as input. This
633 presents an additional challenge if this impact model is to be driven by climate model data for climate
634 change studies, since the climate model will have biases in its climate characteristics (mean, variability, etc.)
635 compared to those of the observed data. Applying the adjustment step, the hydrological model can rely on
636 its calibration to observed conditions (Haerter et al., 2015; Refsgaard et al., 2014).
637

5.5 Adjustment methods not included in the study

639 Only the basic adjustment methods have been included in our study. The simple climate factor approach

640 has been applied in numerous hydrological applications (DeGaetano and Castellano, 2017; Sunyer et al.,
641 2015) and others. We also wanted to test quantile mapping approaches, which in extreme value theory
642 takes the form of a parametric transfer function. This we have applied in two flavours in the spirit of (Räty
643 et al. (2014). Finally, we wanted to benchmark against the 'canonical' benchmark methods: observations
644 and raw model output.
645
646 There is a myriad of more specialised methods, each tailored to account for a particular deficit of the
647 simpler methods. First, there is the issue whether it for precipitation is more reasonable to map relative
648 quantile changes rather than absolute ones (Cannon et al., 2015). It has also been argued that a bias
649 correction method should preserve long-term trends, i.e. the 'climate signal' and only adjust the shorter
650 time scales, as extensively discussed in (Cannon et al., 2015). Then multivariate methods have been argued
651 for and applied in order to preserve relationships between variables (Cannon, 2018), and nested methods
652 to account for different biases for different time scales (Mehrotra et al., 2018). Also methods to correct for
653 systematic displacement of variable features in complex terrain have been suggested and applied (Maraun
654 and Widmann, 2015). Finally, Li et al. (2018) adjusts stratiform and convective precipitation separately
655 instead of adjusting the total precipitation. In this way, any future change in the ratio between the two
656 types of precipitation is accounted for.
657
658 It could be interesting to examine the above methods in future studies, though we acknowledge it would
659 be a quite extensive work. We can at present only guess about the outcome of such work but the more
660 refined methods may not perform too well in the inter-model cross-validation setting. The reason for this
661 suspicion is that these methods, while being more elaborate, in most cases also have more parameters to
662 be estimated, implying a higher risk of overfitting. An argument in favour of this is that the present study
663 shows that the more elaborate quantile mapping methods DC og BC do not outperform the simpler FAC
664 method.

665 # 6   Conclusions
666
667 Based on hourly precipitation data from a 19-member ensemble of climate simulations we have
668 investigated the benefit of bias adjusting extreme precipitation return levels on hourly and daily time scales
669 and evaluated the different methods. This is done in a pseudo-reality setting, where one model simulation
670 in turn from the ensemble plays the role of observations extending into the future. The return levels
671 obtained from each of the remaining model simulations are then adjusted in the present-day period, using
672 different adjustment methods. Then the same adjustment methods are applied to end-21st-century model
673 data to obtain projected return levels, which are then compared with the corresponding pseudo-realistic
674 future return levels.
675
676 The main result of this inter-comparison is that applying bias adjustment methods improves projected
677 extreme precipitation return levels, compared to using the un-adjusted model runs. Can an overall superior
678 adjustment methodology be appointed? For hourly duration, the method to recommend (having the
679 smallest relative error) is the simple climate factor approach FAC, which is better in terms of the relative
680 error than the more complicated analytical quantile mapping methods based on EVA, DC and, in particular,
681 BC. For daily duration, the OBS method performs surprisingly well, having the smallest 95th percentile of

the relative error. Furthermore, the quantile mapping methods perform better than FAC, with DC having the smallest relative error. These conclusions hold regardless of the sub-region considered. We also cross-validated against model ensemble means; this gave in general similar results without significant changes in the distribution of the relative error.

Finally, we registered emergent constraints between present-day and end-21st-century biases. This was more pronounced for hourly than for daily time scales. 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. PB was funded by the project AQUACLEW, which is part of ERA4CS, an ERA-NET initiated by JPI Climate, and funded by FORMAS (SE), DLR (DE), BMWFW (AT), IFD (DK), MINECO (ES), ANR (FR) with co-funding by the European Commission (Grant Agreement 690462). 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 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. We appreciate constructive comments from referee Jorn van de Velde, from two anonymous referees, and from T. Kelder, R. L. Wilby, T. Marjoribanks, and L. Slater.

References

Aalbers, E. E., Lenderink, G., van Meijgaard, E. and van den Hurk, B. J. J. M.: Local-scale changes in mean and heavy precipitation in Western Europe, climate change or internal variability?, Clim. Dyn., 50(11–12), 4745–4766, doi:10.1007/s00382-017-3901-9, 2018.

727 Arakawa, A.: The Cumulus Parameterization Problem: Past, Present, and Future, J. Clim., 17, 33, 2004.

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

730 Berg, P., Christensen, O. B., Klehmet, K., Lenderink, G., Olsson, J., Teichmann, C. and Yang, W.: Summertime
731 precipitation extremes in a EURO-CORDEX 0.11° ensemble at an hourly resolution, Nat. Hazards Earth Syst.
732 Sci., 19(4), 957–971, doi:10.5194/nhess-19-957-2019, 2019.

733 Boberg, F. and Christensen, J. H.: Overestimation of Mediterranean summer temperature projections due
734 to model deficiencies, Nat. Clim. Change, 2(6), 433–436, doi:10.1038/NCLIMATE1454, 2012.

735 Buser, C., Künsch, H. and Schär, C.: Bayesian multi-model projections of climate: generalization and
736 application to ENSEMBLES results, Clim. Res., 44(2–3), 227–241, doi:10.3354/cr00895, 2010.

737 Cannon, A. J.: Multivariate quantile mapping bias correction: an N-dimensional probability density function
738 transform for climate model simulations of multiple variables, Clim. Dyn., 50(1–2), 31–49,
739 doi:10.1007/s00382-017-3580-6, 2018.

740 Cannon, A. J., Sobie, S. R. and Murdock, T. Q.: Bias Correction of GCM Precipitation by Quantile Mapping:
741 How Well Do Methods Preserve Changes in Quantiles and Extremes?, J. Clim., 28(17), 6938–6959,
742 doi:10.1175/JCLI-D-14-00754.1, 2015.

743 Chen, J., Brissette, F. P. and Lucas-Picher, P.: Assessing the limits of bias-correcting climate model outputs
744 for climate change impact studies, J. Geophys. Res. Atmospheres, 120(3), 1123–1136,
745 doi:10.1002/2014JD022635, 2015.

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

748 Christensen, J. H. and Christensen, O. B.: A summary of the PRUDENCE model projections of changes in
749 European climate by the end of this century, Clim. Change, 81(S1), 7–30, doi:10.1007/s10584-006-9210-7,
750 2007.

751 Coles, S.: An introduction to statistical modeling of extreme values, Springer., 2001.

752 Collins, M., Chandler, R. E., Cox, P. M., Huthnance, J. M., Rougier, J. and Stephenson, D. B.: Quantifying
753 future climate change, Nat. Clim. Change, 2(6), 403–409, doi:10.1038/nclimate1414, 2012.

754 DeGaetano, A. T. and Castellano, C. M.: Future projections of extreme precipitation intensity-duration-
755 frequency curves for climate adaptation planning in New York State, Clim. Serv., 5, 23–35,
756 doi:10.1016/j.cliser.2017.03.003, 2017.

757 Eggert, B., Berg, P., Haerter, J. O., Jacob, D. and Moseley, C.: Temporal and spatial scaling impacts on
758 extreme precipitation, Atmospheric Chem. Phys., 15(10), 5957–5971, doi:10.5194/acp-15-5957-2015, 2015.

759 Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J. and Taylor, K. E.: Overview of the
760 Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geosci.
761 Model Dev., 9(5), 1937–1958, doi:10.5194/gmd-9-1937-2016, 2016.

762   Gleckler, P. J., Taylor, K. E. and Doutriaux, C.: Performance metrics for climate models, J. Geophys. Res.,
763   113(D6), D06104, doi:10.1029/2007JD008972, 2008.

764   Gudmundsson, L., Bremnes, J. B., Haugen, J. E. and Engen-Skaugen, T.: Technical Note: Downscaling RCM
765   precipitation to the station scale using statistical transformations – a comparison of methods, Hydrol. Earth
766   Syst. Sci., 16(9), 3383–3390, doi:10.5194/hess-16-3383-2012, 2012.

767   Gutiérrez, J. M., Maraun, D., Widmann, M., Huth, R., Hertig, E., Benestad, R., Roessler, O., Wibig, J., Wilcke,
768   R., Kotlarski, S., San Martín, D., Herrera, S., Bedia, J., Casanueva, A., Manzanas, R., Iturbide, M., Vrac, M.,
769   Dubrovsky, M., Ribalaygua, J., Pórtoles, J., Räty, O., Räisänen, J., Hingray, B., Raynaud, D., Casado, M. J.,
770   Ramos, P., Zerenner, T., Turco, M., Bosshard, T., Štěpánek, P., Bartholy, J., Pongracz, R., Keller, D. E., Fischer,
771   A. M., Cardoso, R. M., Soares, P. M. M., Czernecki, B. and Pagé, C.: An intercomparison of a large ensemble
772   of statistical downscaling methods over Europe: Results from the VALUE perfect predictor cross-validation
773   experiment, Int. J. Climatol., 39(9), 3750–3785, doi:10.1002/joc.5462, 2019.

774   Haerter, J. O., Hagemann, S., Moseley, C. and Piani, C.: Climate model bias correction and the role of
775   timescales, Hydrol. Earth Syst. Sci., 15(3), 1065–1079, doi:10.5194/hess-15-1065-2011, 2011.

776   Haerter, J. O., Eggert, B., Moseley, C., Piani, C. and Berg, P.: Statistical precipitation bias correction of
777   gridded model data using point measurements, Geophys. Res. Lett., 42(6), 1919–1929,
778   doi:10.1002/2015GL063188, 2015.

779   Hanel, M. and Buishand, T. A.: On the value of hourly precipitation extremes in regional climate model
780   simulations, J. Hydrol., 393(3–4), 265–273, doi:10.1016/j.jhydrol.2010.08.024, 2010.

781   Haylock, M. R., Hofstra, N., Klein Tank, A. M. G., Klok, E. J., Jones, P. D. and New, M.: A European daily high-
782   resolution gridded data set of surface temperature and precipitation for 1950–2006, J. Geophys. Res.,
783   113(D20), doi:10.1029/2008JD010201, 2008.

784   Hosking, J. R. M. and Wallis, J. R.: Parameter and Quantile Estimation for the Generalized Pareto
785   Distribution, Technometrics, 29(3), 339, doi:10.2307/1269343, 1987.

786   Hui, Y., Chen, J., Xu, C., Xiong, L. and Chen, H.: Bias nonstationarity of global climate model outputs: The
787   role of internal climate variability and climate model sensitivity, Int. J. Climatol., 39(4), 2278–2294,
788   doi:10.1002/joc.5950, 2019.

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

796   Kallache, M., Vrac, M., Naveau, P. and Michelangeli, P.-A.: Nonstationary probabilistic downscaling of
797   extreme precipitation, J. Geophys. Res., 116(D5), doi:10.1029/2010JD014892, 2011.

798   Kendon, E. J., Roberts, N. M., Fowler, H. J., Roberts, M. J., Chan, S. C. and Senior, C. A.: Heavier summer
799   downpours with climate change revealed by weather forecast resolution model, Nat. Clim. Change, 4(7),
800   570–576, doi:10.1038/nclimate2258, 2014.

801 Kendon, E. J., Ban, N., Roberts, N. M., Fowler, H. J., Roberts, M. J., Chan, S. C., Evans, J. P., Fosser, G. and
802 Wilkinson, J. M.: Do Convection-Permitting Regional Climate Models Improve Projections of Future
803 Precipitation Change?, Bull. Am. Meteorol. Soc., 98(1), 79–93, doi:10.1175/BAMS-D-15-0004.1, 2017.

804 Kerkhoff, C., Künsch, H. R. and Schär, C.: Assessment of Bias Assumptions for Climate Models, J. Clim.,
805 27(17), 6799–6818, doi:10.1175/JCLI-D-13-00716.1, 2014.

806 Klemeš, V.: Operational testing of hydrological simulation models, Hydrol. Sci. J., 31(1), 13–24,
807 doi:10.1080/02626668609491024, 1986.

808 Laflamme, E. M., Linder, E. and Pan, Y.: Statistical downscaling of regional climate model output to achieve
809 projections of precipitation extremes, Weather Clim. Extrem., 12, 15–23, doi:10.1016/j.wace.2015.12.001,
810 2016.

811 Lenderink, G., Belušić, D., Fowler, H. J., Kjellström, E., Lind, P., van Meijgaard, E., van Ulft, B. and de Vries,
812 H.: Systematic increases in the thermodynamic response of hourly precipitation extremes in an idealized
813 warming experiment with a convection-permitting climate model, Environ. Res. Lett., 14(7), 074012,
814 doi:10.1088/1748-9326/ab214a, 2019.

815 Li, J., Evans, J., Johnson, F. and Sharma, A.: A comparison of methods for estimating climate change impact
816 on design rainfall using a high-resolution RCM, J. Hydrol., 547, 413–427, doi:10.1016/j.jhydrol.2017.02.019,
817 2017a.

818 Li, J., Johnson, F., Evans, J. and Sharma, A.: A comparison of methods to estimate future sub-daily design
819 rainfall, Adv. Water Resour., 110, 215–227, doi:10.1016/j.advwatres.2017.10.020, 2017b.

820 Li, J., Sharma, A., Evans, J. and Johnson, F.: Addressing the mischaracterization of extreme rainfall in
821 regional climate model simulations – A synoptic pattern based bias correction approach, J. Hydrol., 556,
822 901–912, doi:10.1016/j.jhydrol.2016.04.070, 2018.

823 Madsen, M. S., Langen, P. L., Boberg, F. and Christensen, J. H.: Inflated Uncertainty in Multimodel-Based
824 Regional Climate Projections, Geophys. Res. Lett., 44(22), 2017GL075627, doi:10.1002/2017GL075627,
825 2017.

826 Maraun, D.: Nonstationarities of regional climate model biases in European seasonal mean temperature
827 and precipitation sums, Geophys. Res. Lett., 39(6), n/a-n/a, doi:10.1029/2012GL051210, 2012.

828 Maraun, D.: Bias Correction, Quantile Mapping, and Downscaling: Revisiting the Inflation Issue, J. Clim.,
829 26(6), 2137–2143, doi:10.1175/JCLI-D-12-00821.1, 2013.

830 Maraun, D.: Bias Correcting Climate Change Simulations - a Critical Review, Curr. Clim. Change Rep., 2(4),
831 211–220, doi:10.1007/s40641-016-0050-x, 2016.

832 Maraun, D. and Widmann, M.: The representation of location by a regional climate model in complex
833 terrain, Hydrol. Earth Syst. Sci., 19(8), 3449–3456, doi:10.5194/hess-19-3449-2015, 2015.

834 Maraun, D., Shepherd, T. G., Widmann, M., Zappa, G., Walton, D., Gutiérrez, J. M., Hagemann, S., Richter, I.,
835 Soares, P. M. M., Hall, A. and Mearns, L. O.: Towards process-informed bias correction of climate change
836 simulations, Nat. Clim. Change, 7(11), 764–773, doi:10.1038/nclimate3418, 2017.

837 Maurer, E. P., Das, T. and Cayan, D. R.: Errors in climate model daily precipitation and temperature output:
838 time invariance and implications for bias correction, Hydrol. Earth Syst. Sci., 17(6), 2147–2159,
839 doi:10.5194/hess-17-2147-2013, 2013.

840 McSweeney, C. F., Jones, R. G., Lee, R. W. and Rowell, D. P.: Selecting CMIP5 GCMs for downscaling over
841 multiple regions, Clim. Dyn., 44(11–12), 3237–3260, doi:10.1007/s00382-014-2418-8, 2015.

842 Mehrotra, R., Johnson, F. and Sharma, A.: A software toolkit for correcting systematic biases in climate
843 model simulations, Environ. Model. Softw., 104, 130–152, doi:10.1016/j.envsoft.2018.02.010, 2018.

844 Olsson, J., Berg, P. and Kawamura, A.: Impact of RCM Spatial Resolution on the Reproduction of Local,
845 Subdaily Precipitation, J. Hydrometeorol., 16(2), 534–547, doi:10.1175/JHM-D-14-0007.1, 2015.

846 Overeem, A., Buishand, A. and Holleman, I.: Rainfall depth-duration-frequency curves and their
847 uncertainties, J. Hydrol., 348(1–2), 124–134, doi:10.1016/j.jhydrol.2007.09.044, 2008.

848 Pfahl, S., O'Gorman, P. A. and Fischer, E. M.: Understanding the regional pattern of projected future
849 changes in extreme precipitation, Nat. Clim. Change, 7(6), 423–427, doi:10.1038/nclimate3287, 2017.

850 Piani, C., Haerter, J. O. and Coppola, E.: Statistical bias correction for daily precipitation in regional climate
851 models over Europe, Theor. Appl. Climatol., 99(1–2), 187–192, doi:10.1007/s00704-009-0134-9, 2010.

852 Prein, A. F., Langhans, W., Fosser, G., Ferrone, A., Ban, N., Goergen, K., Keller, M., Tölle, M., Gutjahr, O.,
853 Feser, F., Brisson, E., Kollet, S., Schmidli, J., Lipzig, N. P. M. and Leung, R.: A review on regional convection-
854 permitting climate modeling: Demonstrations, prospects, and challenges, Rev. Geophys., 53(2), 323–361,
855 doi:10.1002/2014RG000475, 2015.

856 Räisänen, J. and Räty, O.: Projections of daily mean temperature variability in the future: cross-validation
857 tests with ENSEMBLES regional climate simulations, Clim. Dyn., 41(5–6), 1553–1568, doi:10.1007/s00382-
858 012-1515-9, 2013.

859 Räty, O., Räisänen, J. and Ylhäisi, J. S.: Evaluation of delta change and bias correction methods for future
860 daily precipitation: intermodel cross-validation using ENSEMBLES simulations, Clim. Dyn., 42(9–10), 2287–
861 2303, doi:10.1007/s00382-014-2130-8, 2014.

862 Refsgaard, J. C., Madsen, H., Andréassian, V., Arnbjerg-Nielsen, K., Davidson, T. A., Drews, M., Hamilton, D.
863 P., Jeppesen, E., Kjellström, E., Olesen, J. E., Sonnenborg, T. O., Trolle, D., Willems, P. and Christensen, J. H.:
864 A framework for testing the ability of models to project climate change and its impacts, Clim. Change,
865 122(1–2), 271–282, doi:10.1007/s10584-013-0990-2, 2014.

866 Rowell, D. P.: An Observational Constraint on CMIP5 Projections of the East African Long Rains and
867 Southern Indian Ocean Warming, Geophys. Res. Lett., 46(11), 6050–6058, doi:10.1029/2019GL082847,
868 2019.

869 Sunyer, M., Luchner, J., Onof, C., Madsen, H. and Arnbjerg-Nielsen, K.: Assessing the importance of spatio-
870 temporal RCM resolution when estimating sub-daily extreme precipitation under current and future
871 climate conditions, Int. J. Climatol., 37(2), 688–705, 2017.

872 Sunyer, M. A., Gregersen, I. B., Rosbjerg, D., Madsen, H., Luchner, J. and Arnbjerg-Nielsen, K.: Comparison
873 of different statistical downscaling methods to estimate changes in hourly extreme precipitation using RCM
874 projections from ENSEMBLES, Int. J. Climatol., 35(9), 2528–2539, doi:10.1002/joc.4138, 2015.

875   Taylor, K. E., Stouffer, R. J. and Meehl, G. A.: An Overview of CMIP5 and the Experiment Design, Bull. Am.
876   Meteorol. Soc., 93(4), 485–498, doi:10.1175/BAMS-D-11-00094.1, 2012.

877   Themeßl, M. J., Gobiet, A. and Leuprecht, A.: Empirical-statistical downscaling and error correction of daily
878   precipitation from regional climate models, Int. J. Climatol., 31(10), 1530–1544, doi:10.1002/joc.2168,
879   2011.

880   Themeßl, M. J., Gobiet, A. and Heinrich, G.: Empirical-statistical downscaling and error correction of
881   regional climate models and its impact on the climate change signal, Clim. Change, 112(2), 449–468,
882   doi:10.1007/s10584-011-0224-4, 2012.

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

885   Van Schaeybroeck, B. and Vannitsem, S.: Assessment of calibration assumptions under strong climate
886   changes, Geophys. Res. Lett., 43(3), 1314–1322, doi:10.1002/2016GL067721, 2016.

887   Velázquez, J. A., Troin, M., Caya, D. and Brissette, F.: Evaluating the Time-Invariance Hypothesis of Climate
888   Model Bias Correction: Implications for Hydrological Impact Studies, J. Hydrometeorol., 16(5), 2013–2026,
889   doi:10.1175/JHM-D-14-0159.1, 2015.

890
891