Dear editor,

We appreciate very much the comment from the referees and from our collegues posting a SC. These have been extremely useful for improving the manuscript. Below, you find all comments comments with our responses in *italic*. Section numbers and line numbers refer to the marked-up version of the revised manuscript.

Anonymous Referee #1:

This is an interesting contribution involving a lot of work. I have a few general issues that the authors should address in their revisions, followed by some specific comments. Firstly - there needs to be a better discussion about the possible problems in using the pseudo-reality setting for assessment of precipitation extremes. Most models have a tendency to increase the probability of occurrence of rainfall, thereby increasing the size of the sample that could potentially constitute extremes. The authors have avoided this issue to some extent by performing a pseudo-reality assessment. I believe some discussion should be included as this could create difficulties in taking the findings from here to real applications.

> *We have added/modified the intro (lines 172-180) about different validation approaches and their pros and con's. We recognize that models do have a tendency to increased probability of rainfall. As for the last part of the comment, we determine our POT threshold by having three events/year instead of having a fixed threshold. Therefore, we always have the same pool of extremes, regardless of model and present-day/end-21st-century.*

Secondly, the paper is coming across as a bit of a report (and I sympathise with the authors as they do have a lot of information to present). Perhaps a more creative discussion for differences in mountaineous areas versus not, coastal areas versus not, and daily durations versus hourly would be useful. I note the spatial resolution is 11km. Daily extremes should be simulated better at this resolution.

> *Thanks for the advice. We have worked through the text and realize that maybe you think of Section 4.1. Therefore, we have extended the description of Figures 2 and 3. Furthermore, these figures have been modified, caused by a suggestion from another referee.*

Also, no mention is made of the causative GCMs that are interpolated using the RCMs. There are different extent of biases in these. Some discussion should be included on this as well.

> *We have introduced some text on this in section 2, mentioning good performance of GCMs and the argument for using 'ensemble of opportunity' in favour of selection procedures.*

Thirdly, the authors have missed with publications on this topic by Jingwan Li. Relevant papers are: Li, J., et al. (2017). "A comparison of methods for estimating climate change impact on design rainfall using a high-resolution RCM." Journal of Hydrology 547: 413-427. Li, J., et al. (2017). "A comparison of methods to estimate future subdaily design rainfall." Advances in Water Resources 110: 215-227. Li, J., et al. (2018). "Addressing the mischaracterization of extreme rainfall in regional climate model simulations – A synoptic pattern based bias correction approach." Journal of Hydrology 556: 901-912. Li, J., et al. (2018). "Can Regional Climate Modeling Capture the Observed Changes in Spatial Organization of Extreme Storms at Higher Temperatures?" Geophysical Research Letters 45(9): 4475-4484.

I am a co-author on these papers hence have a conflict here. But I think these are very relevant to what the authors are attempting to do here, as she used an even finer resolution RCM with a high density of observed gauges at the same time resolution (hourly). The bias correction approach she adopted acknowledged the bias in simulating convection within the RCMs as well as the quantile bias convective and non-convective rainfall were exhibiting.

> *We were not aware of these papers. We are now referring to the two papers "A comparison …" in the introduction (line 158). Our manuscript evaluates basic adjustment methods only. We know that there is a myriad of special-designed adjust methods, including the one described in the paper "Addressing the mischaracterization … ". We have added a section (5.5, lines 688-714) discussing which methods were/were not included in our study. The paper "Can Regional Climate Modeling Capture …" about the spatial extent of extreme precipitation events is in our opinion not within the scope of our manuscript.*

Now to the specific comments:

line 142 - missing section marker

> *Thanks, has been fixed.*

line 225 - there is another way to create the partial series sample. It is to acknowledge that there may be a bias in the proportion of events that are say convective. If this proportion is biased, one is forming a biased sample effectively by selecting the series the way adopted here. This issue is the focus of Li, J., et al. (2018). "Addressing the mischaracterization of extreme rainfall in regional climate model simulations – A synoptic pattern based bias correction approach." Journal of Hydrology 556: 901-912.

> *In the manuscript we evaluate the basic methods (see line 690). The work described in the suggested paper is not within our scope (see also above)*.

line497 - If the proportion of convective extreme events increases in the future (as it is expected to) then ignoring any bias in the representation of convection as discussed above, will create a non-stationary bias. This can be addressed though using the above mentioned approach.

The aim of our work is to evaluate the simple bias adjustment methods for extremes, as also explained above. More sophisticated methods are not included in this study, but the suggested paper can go into the discussion on future work.

Referee #2:

General comments

In their contribution, Schmith et al. (2020) discuss the robustness of different bias-adjusting methods for (sub)daily rainfall extremes. This yields interesting results and strong links with the context of convection-permitting models and emergent constraints. Yet, there are some aspects about whom I'd like a deeper discussion.

We appreciate this positive overall judgement of our manuscript and are positive towards adding more discussion to it.

The first aspect is the practical use of this study. This is foremost linked with the choice of bias-adjusting methods. Although the use of return periods is perfectly justified from a hydrological point of view, I've seen few studies that actually use bias adjustment directly on the return periods. As such, I'd like to see a larger discussion on the choice of bias-adjusting methods.

Our aim has been to evaluate basic adjustment methods. We have added a new subsection in the discussion (lines 588-714) summarizing the more elaborate quantile mapping methods.

Given a well-justified choice, I understand the use of these simple methods, yet I'd like to see more discussion on how this relates with more complicated, but related bias-adjustment methods, such as e.g. CDF-t (Michelangeli et al., 2009), standard QM, QDM (Cannon et al., 2015), : : : Would it be possible to discuss possible consequences for the use of these methods for the adjustment of subdaily precipitation extremes? This could fit in the second paragraph of Section 5.1, which seems rather limited and abrupt at this point.

In a new sub-section (lines 588-714) we discuss the use of more elaborate methods. We emphasize that these methods build on alternative, but not necessarily more correct, assumptions. It would be interesting to test these methods in our framework, but we reserve this to future publications. We also note that our investigation do not generally find that the more elaborate methods (quantile mapping) outperform the simpler climate factor approach.

A last point related to the practical use is that I missed a more thorough explanation of why the observations perform well, why this version of quantile mapping performs poorly. Although this is discussed slightly in Section 4.3, I wonder if more details or, if possible, practical guidelines could be given in the discussion.

A thorough reveal of causes for some models performing well would require quite some extra analysis which cannot be accommodated within this manuscript. We may speculate that the cause *of observations performing so well as projection is related to the poor signal-to-noise ratio, as seen in Fig. 4. The relatively poor performance of the quantile-matching methods could be caused by the many extreme value distributions to be estimated, each of which are very uncertain. We have added a block of text on this in the Conclusions section.*

A second aspect is that some concepts in the Introduction seem to be accepted as-is, whereas they could deserve a deeper discussion. A first example of this is the discussion of stationarity in the introduction. The references are limited in time, whereas more recent papers expanded this subject, such as Kerkhoff et al. (2014) and Van Schaeybroeck and Vannitsem (2016) on the type of bias relationship and Chen et al. (2015), Velázquez et al. (2015), Wang et al. (2018) and Hui et al. (2019), who discussed the uncertainty introduced by bias nonstationarity. As the stationarity of the bias is an important part of the discussion, I think the paper could benefit from these perspectives.

In the original submitted manuscript, stationarity was mentioned and briefly discussed in the introduction. We have written a new discussion and updated the references (lines 136-147).

A second, smaller example is the use of a delta change based method. While the method isn't completely discredited, there has been some discussion whether it's use for climate change is not too dependent on the assumption that the temporal structure of the time series will not change from present to future (e.g. Johnson and Sharma (2011), Kerkhoff et al. (2014)). It would thus be interesting to read a deeper discussion on the limitations of the methods

We are aware of the assumption about unchanged temporal structure of time series in the delta change approach, though this is only 100% true in the simplest version of a shift of the mean, in the quantile mapping version temporal structure may change. Furthermore, our MOS of extreme levels do not yield any time series as output. Therefore, we think that a discussion as suggested is not relevant for our manuscript.

Specific comments

L. 37: 'quantile-mapping' is used here, whereas in the remainder of the abstract (and the paper) 'quantile-matching' is used. I'd suggest to edit this for coherence, but to also use 'quantile mapping' throughout the paper, as it has been the most used term for this type of bias adjustment during the last few years.

Certainly, the nomenclature should be consistent throughout. We have followed your advice and replaced 'quantile matching' to 'quantile mapping' throughout.

L. 75-82: this paragraph is very scarce on references. Although some of the necessary references are given in the discussion, I think it would be good to also have the reference to the papers about CPMs in this paragraph.

Ok, we have introduced the appropriate references

L. 84-91: The terminology in this paragraph could be reconsidered. Although it is debatable whether or not to consider delta change as a bias adjustment approach (the latest textbook, Maraun and Widmann (2018), is on the edge), it feels very strange to read 'bias correction' as a subset of 'bias adjustment' approaches. The use of 'bias adjustment' as a replacement of 'bias correction' has been rising during the last few years, as it is clearer that the methods are statistical and cannot correct all climate model biases. Thus, I would withhold from the use of 'bias correction'. Better terminology seems MOS, with delta change and bias adjustment as possible subcategories, or bias adjustment with delta change and bias adjustment s.s., although the exact choice is personal.

It is indeed difficult to find a coherent terminology - with Maraun&Widman, there is a 'Babylonian confusion'. We have decided to use the generic term 'adjustment' (sometimes bias adjustment' to prevent confusion) with sub-categories 'bias correction' and 'delta change' throughout the revised manuscript. In the main headline, though, we keep 'bias correction' as the generic term for better readability.

L. 253- 286: Although the method described here is indeed based on the same principles as XCDF-t as used by Kallache et al. (2011) and Laflamme et al. (2016), it's not entirely clear how the new method is created by adapting the former. I think the link between both methods should be more detailed, so users can retrace it more easily and infer the strengths and limitations. Especially as it is specifically mentioned that the method 'will be adapted to our needs below', the adaptation seems rather limited.

Our method was originally inspired by XCDF-t, but we make the more direct approach and define transformations, which are the used to correct the return levels. To avoid any confusion, we have chosen to remove the first lines of section 3.3.2

L. 448-453: the explanation of the use of the index by Maurer et al. (2013) should be expanded. Firstly, it's unclear to me where the terminology 'measure of relative spread' is derived from, as it is not named as such in the original paper. Secondly, the interpretation of the R-values is not discussed, although this is quite important: values < 1 indicate that the difference in biases is smaller than the mean bias of both periods, whereas values >1 indicate that the difference in biases is larger, which could have a potentially large impact. As both values are quite far < 1, the bias seems quite stationary, but in your discussion you state that the 24h duration is 'less stationary'. Without giving this numerical explanation, this statement is hard to interpret correctly.

We have expanded the explanation of R, and its interpretation, as suggested. Certainly, both R-values are below 1. However, it is the limit of R=0 which is a sign of a stationary bias factor and this is the basis of our interpretation and discussion.

L. 504-505: This last sentence does not seem to fit with the rest of the paragraph. I think that, with some rewriting, this could become clearer.

This reference doesn't really belong here, so we have deleted this sentence.

Technical comments

we will adhere to the technical comments given below

L. 48: ‘Global climate models (GCMs) is : : :’ -> are *done*

L. 110-111: ‘Only a few examples has : : :’ -> have *done*

L. 112-113: ‘: : : applying bias adjustment improve projections’ -> improves *done*

L. 142: the section marker should be corrected *ok*

L. 194: I can’t find the source of this problem, should not be referenced with co-authors. The official webpage by Springer (https://link.springer.com/book/10.1007%2F978-1-4471-3675-0#about) only mentions one author (Stuart Coles) and there is no mention of other authors elsewhere in the book. So unless I’m missing something, I think the more correct reference is Coles (2001). *Yes, correct, has been changed.*

L. 232-243: ‘Hosking and Wallis (1987) : : : warns : : : . Instead, he recommends : : :’. Shouldn’t these sentences be plural, or are you referring to ‘the paper’ in these sentences instead of ‘the authors’? *Probably one should refer to the authors, we have corrected*

L. 254: ‘Kallache et al. (2011) and Laflamme et al. (2016) applies’ -> apply, as this verb is referring to multiple papers and authors. *done*

L. 265: ‘ths’ -> ’the’ *done*

Figure 6 and Figure 8: Would it be possible to remove the underscores from the plot titles? *Done*

References

Cannon, A. J., Sobie, S. R., and Murdock, T. Q.: Bias correction of GCM precipitation by quantile mapping: How well do methods preserve changes in quantiles and extremes?, Journal of Climate, 28, 6938–6959, https://doi.org/10.1175/JCLI-D-14-00754.1, 2015

Chen, J., Brissette, F. P., and Lucas-Picher, P.: Assessing the limits of bias-correcting climate model outputs for climate change impact studies, Journal of Geophysical Research: Atmospheres, 120, 1123–1136, https://doi.org/10.1002/2014JD022635, 2015

Hui, Y., Chen, J., Xu, C.-Y., Xiong, L., and Chen, H.: Bias nonstationarity of global climate model outputs: The role of internal climate variability and climate model sensitivity, International Journal of Climatology, 39, 2278–2294, https://doi.org/10.1002/joc.5950, 2019

Johnson, F. and Sharma, A.: Accounting for interannual variability: A comparison of options for water resources climate change impact assessments, Water Resources Research, 47, W04 508, https://doi.org/10.1029/2010WR009272, 2011

Kerkhoff, C., Künsch, H. R., and Schär, C.: Assessment of bias assumptions for climate models, Journal of Climate, 27, 6799–6818, https://doi.org/10.1175/JCLI-D-13-00716.1, 2014

Maraun, D. and Widmann, M.: Statistical Downscaling and Bias Correction for Climate Research, Cambridge University Press, https://doi.org/10.1017/9781107588783, 2018

Michelangeli, P.-A., Vrac, M., and Loukos, H.: Probabilistic downscaling approaches: Application to wind cumulative distribution functions, Geophysical Research Letters, 36, L11 708, https://doi.org/10.1029/2009GL038401, 2009

Van Schaeybroeck, B. and Vannitsem, S.: Assessment of calibration assumptions under strong climate changes, Geophysical Research Letters, 43, 1314–1322, https://doi.org/10.1002/2016GL067721, 2016

Velázquez, J. A., Troin, M., Caya, D., and Brissette, F.: Evaluating the time-invariance hypothesis of climate model bias correction: implications for hydrological impact studies, Journal of Hydrometeorology, 16, 2013–2026, https://doi.org/10.1175/JHM-D-14-C50159.1, 2015

Wang, Y., Sivandran, G., and Bielicki, J. M.: The stationarity of two statistical downscaling methods for precipitation under different choices of cross-validation periods, International Journal of Climatology, 38, e330–e348, https://doi.org/10.1002/joc.5375, 2018

Referee #3

Overall comment

Overall, I recommend a better embedding of the manuscript in the current literature, both in introduction (e.g. much work has been done on comparing different bias correction methods, which could be included) and the section 5.1 could easily be expanded. I also would like to see expansion on why different methods give different results. There seems to be no analysis or discussion of what features of different methods contribute to greater or lesser skill. In my view the manuscript would be improved if this were addressed.

We will meet this advice of a more thourough embedding in the relevant. This will be followed by adhering to suggestions given by in particular referee #2. To disentangle why different methods give different results requires more analysis requires extensive analysis and has to be left to future work. We have given an appetizer of this kind of work in section 4.3.

Minor comments

105-106: It is true that future model performance cannot be tested directly. However, split-sample testing is probably the best tool we have for this, particularly when a suspected climate change signal is present in recent historical data.

as we see it, split-sample testing is an alternative to our approach; not neccesarily the best one. We have included a paragraph in the introduction discussing different validation approaches and their pros and con’s (in lines 172-180), in accordance with suggestion from referee #1.

Figure 2,3: I find the colour scale used in these figure inappropriate. Yes, extreme precipitation events are projected to increase, but the scale make the increases look quite alarming. A percentage scale, and/or scale starting at zero would be more appropriate.

We have reacted to this piece of advice by showing instead maps of present-day and maps of the relative change

372-373, this sentence describing relative errors is a little unclear, I would suggest writing "Relative errors from the OBS method are in the range of 20%-40%" or similar.

Done

395 and elsewhere: I'd use "percentiles" rather than "fractiles", e.g. 95th percentile rather than 0.95 fractile

We agree that percentile is more widely used according to Google; therefore we have followed this advice. We have also changed all relative measures to percent throughout.

The writing is generally of a high quality, but with a few corrections needed, such as:
48: "GCMs are" *yes, thank you*
182: "statistics are" *yes, thank you*
I recommend a thorough proofread to catch any other corrections

Short comment #1

Comment on 'Identifying robust bias adjustment methods for extreme precipitation in a pseudo-reality setting' T. Kelder, R. L. Wilby, T. Marjoribanks, L. Slater

Torben Schmith and co-authors address a complex, but important topic. Climate model corrections typically assume stationary biases between simulated and observed extreme precipitation but, in practice, such biases may well be nonstationary (i.e. distributions may shift significantly in the future). Robust evaluation of bias correction methods is hampered by the inability to analyse future model biases, since there are obviously no observations of the future. To address this issue, the authors use model simulations as a pseudo-reality of the present and future climate to evaluate the robustness of various bias correction methods within these 'virtual' worlds.
The authors processed a large amount of data from the EURO-CORDEX ensemble and we commend them for this interesting research and their purposeful discussion of findings. The paper concludes by recommending a preferred bias correction method for climate projection. We offer a few suggestions and raise some issues for further elaboration by the authors.

1. Given that the analysis is based on an ensemble of climate model experiments, the logic should be explained for treating model-to-model biases in extreme precipitation as equivalent to model-to-observation biases. The paper acknowledges the limited ability of _10km resolution model simulations at representing convective processes. Hence, more explanation is needed for an unfamiliar reader on why model experiments

can be used to draw conclusions about the best bias correction methods on hourly timescales, if one cannot trust the model simulations to realistically represent convective processes.

Acknowledging that the models represent convection imperfect, we are actually better off evaluating the bias correction methods between models than between model and observation. We are here addressing the statistical nature of the corrections, not the physical processes which bias correction methods are not suitable for anyway. We do not promote, naively applying these methods to hourly data from these models. However, the presented methods can in the future be applied to convection permitting model simulations that better represent the convective process, and results from our current manuscript would apply equally to that case. We have added a sentence about this in lines 636-640 of the revised manuscript.

2. Related to #1, a few cautionary remarks could be made about some of the GCMs used to drive the CORDEX experiments (see: Liepert and Lo, 2013). The realism of the downscaled extreme precipitation depends on the realism of the boundary forcing. Use of an ‘ensemble of opportunity’ is not unusual, but some studies narrow the choice of candidate models (and hence uncertainty) based on physical realism tests (e.g. McSweeney et al., 2015; Rowell, 2019).

We only partly agree with this. The large-scale atmospheric state is certainly determined by the boundary forcing; though, the RCM is able to modulate it. Distribution of precipitation intensities are to a large extent determined by the RCM (see e.g. (Christensen and Kjellström 2020)). This is particularly true for the high-extreme end of the spectrum.

We are aware of the use of selection procedures put forward in the cited papers. There is, however, no simple quality index that can be generally applied. Any discrimination of GCMs depends depend 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 criteria and come with major caveats that impact the uncertainty range largely (Madsen et al. 2017). We therefore choose, in accordance with most other similar studies, to use ‘ensemble of opportunity’ for the present study. We now discuss that in lines 235-243.

3. In the inter-model cross-validation setup, every model/pseudo-reality combination is used. This setup can be useful for assessing relationships between present and future bias correction factors (e.g. Fig. 9), but does not mimic climate projections, where the ensemble mean, and range are typically used. In the present setup, a future projection is treated as a deterministic prediction, rather than a probabilistic projection. Perhaps use of the climate ‘pseudo-observed’ run might be favoured over future predictions simply because there is less variability in the present climate? How sensitive are the results to taking the mean of all ensemble members minus the ‘pseudo-reality’ member (e.g. Fig. 3 in Räty et al. 2014)? This has the added benefit of involving much fewer permutations (and hence calculations).

This is a good idea, which we have now implemented in our analysis suite. Results of this are included in the revised manuscript.

4. The range of the projection matters. For example, Fig. 4 shows that there are

future scenarios that exceed the present climate range. Hence, the worst-case 10-year precipitation event from the 'pseudo-obs' range would not include plausible future 10-year events. Therefore, more qualification is needed in the Abstract and Conclusions to guard against this possibility and the potentially misleading assertion that "the superior approach is to simply deduce future return levels from observations". Overall, the headline findings of the research could be presented in more nuanced ways, especially within the Abstract.

We are afraid that we do not understand the central statement of this point ("Hence, the worst-case …). Therefore, we are not able to comment on it.

5. The Abstract and Introduction assert that "Severe precipitation events are usually projected using Regional Climate Model (RCM) scenario simulations." We gently remind the authors that statistical downscaling is also widely used for projecting severe precipitation events and suggest that more inclusive wording be used.

We agree that this suggestion is appropriate and have added a paragraph in the introduction (lines 68-74).

References

Liepert, B.G. and F. Lo, 2013: CMIP5 update of 'Inter-model variability and biases of the global water cycle in CMIP3 coupled climate models'. Environ. Res. Lett., 8(2), p.029401, https://iopscience.iop.org/article/10.1088/1748-9326/8/2/029401/meta.

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

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.

Rowell, D.P., 2019: An observational constraint on CMIP5 projections of the East African Long Rains and Southern Indian Ocean warming. Geophys. Res. Lett., 46(11), 6050-6058, https://doi.org/10.1029/2019GL082847.

Other changes

For improved readability, we now use 'calibration' throughout, instead of changing between 'training' and 'calibration'. Similarly for 'validation'/'verification', and for 'pseudo-reality'/'pseudo-observations' (except in a few cases).

We have moved most parts of former subsection 4.2.1 to create a new subsection 3.4 where the whole inter-model cross-validation procedure incl. validation metrics is described in detail.

In the discussion section, we have swapped the works of LaFlamme and Kallache, to obtain chronology in the text.

Our added references:

Christensen, O. B., and E. Kjellström, 2020: Partitioning uncertainty components of mean climate and climate change in a large ensemble of European regional climate model projections. *Clim. Dyn.*, **54**, 4293–4308, https://doi.org/10.1007/s00382-020-05229-y.

Gleckler, P. J., K. E. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. *J. Geophys. Res.*, **113**, D06104, https://doi.org/10.1029/2007JD008972.

Madsen, M. S., P. L. Langen, F. Boberg, and J. H. Christensen, 2017: Inflated Uncertainty in Multimodel-Based Regional Climate Projections. *Geophys. Res. Lett.*, **44**, 11,606-11,613, https://doi.org/10.1002/2017GL075627.

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]~~, Christian Steger[5]~~, Marianne Sloth Madsen[1], Christian Steger[5]

[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 occurrence of these rare events. These are ~~usually~~ often projected using ~~information based on~~data from 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; these biases can be in the mean and/or in the higher moments. Therefore, the RCM results ~~are often bias-adjusted to match observations~~ are adjusted to account for these deficiencies. However, ~~T~~this~~This~~ does~~, however,~~ not guarantee that ~~bias-~~adjusted projected results will match future reality better, since the bias may ~~change~~ not be stationary in a changing~~edchanged~~ 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~~observations, against which the remaining model simulations are ~~bias~~ adjusted and validated. The study uses hourly data from ~~present-day~~historical and RCP8.5 ~~late 21st century~~ scenario runs 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~~scales. The ~~bias~~ adjustment techniques applied to the return levels are based on extreme value analysis and include climate factor and ~~analytical~~ quantile-~~matching~~ mapping~~together with the simpler climate factor approach~~ approaches. Generally, we find that future return levels can be improved by ~~bias~~ adjustment, compared to obtaining them from raw scenario~~s~~ model data. The performance of the different methods depends ~~of~~ on the time scale considered. On hourly time scale, the climate factor approach performs better than the quantile-~~matching~~ 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~~h~~ change overall conclusions much.

1 Introduction

Severe precipitation events occur typically either as stratiform ~~day-long~~ 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) ~~is~~ 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 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~~reached 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~~ represent physically accurate 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,~~(Arakawa 2004). Thus, the inability to reproduce these short duration extremes can be explained by the imperfect parametrization of sub-grid scale convection~~,~~ (Prein et al. 2015), 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 (Trenberth et al. 2003).

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 in a statistical sense ~~bias.~~ before any further data analysis.~~.~~ ~~Bias~~ Such adjustment techniques are thoroughly discussed, including requirements and limitations, in Maraun (2016) and Maraun et al. (2017). There are basically 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~~mapping, preserving also the higher moments. Quantile-~~matching~~ mapping ~~adjustment~~ can use either empirical quantiles or analytical distribution functions. The ability of quantile-~~matching~~ mapping to reduce bias has been demonstrated for daily precipitation in present-day climate using observations, which are split into ~~training~~ calibration and ~~verification~~ validation ~~parts~~ samples (Piani et al. 2010; Themeßl et al. 2011).

Bias adjustment techniques originate in the field of weather and ocean forecast modelling, where they are known as model output statistics (MOS). ~~where~~ Here output from a forecast model is adjusted for model deficiencies and local features not explicitly resolved by the model. 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. Therefore~~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~~ A central concept of adjustment methods is the assumption of *stationarity* of the bias. For bias correction this ~~is~~means that the ~~bias of the~~transformation from model to observations is unchanged from the present-day climate to the future climate, ~~(stationarity)~~while for delta-change the transformation from present-day climate to future climate is unchanged from model to observations. In the ideal case of stationarity being fulfilled, the adjustment methods will work perfectly and produce perfect future projections. If stationarity is not fulfilled, adjustment may improve projections, or in the worst cases they may degrade projections, compared to using raw model output.

Stationarity has been debated in recent years in the literature (e.g. Buser et al. 2010; Boberg and Christensen 2012). Kerkhoff et al. (2014) review and discuss two hypotheses: 1) constant bias: unchanged between present-day and future (i.e. stationarity) and 2) constant relation: bias varies linearly with the signal. Van Schaeybroeck and Vannitsem (2016) used a pseudo-reality setting with a simplified model and found large changes in the bias between present-day and future for many variables and violation of both constant bias and constant relation hypothesis. Chen et al. (2015) concluded that precipitation bias is clearly non-stationary over North America in that variations in bias is comparable to the climate change signal. Velázquez et al. (2015) used a pseudo-reality setting involving two models and concluded that constancy of bias was violated for both precipitation and temperature on monthly time scale. Hui et al. (2019) used a pseudo-reality setting with GCMs and found significant non-stationarity of bias for annual and seasonal temperatures. Besides, they point to a large effect on non-stationarity from internal variability.
~~Only a few examples has pointed out directly how to validate this cornerstone assumption (see however Buser et al. and Boberg and Christensen) and Boberg and Christensen) 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 ~~-~~method used (Haerter et al. 2011; Berg et al. 2012; Themeßl et al. 2012).

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

An alternative ~~One~~ approach, which we use here, ~~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) and others~~and also used here~~. The rationale is here 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~~Thus, one member of the ensemble ~~in turn~~alternatively plays the role of *pseudo-*

~~reality~~observations, against which the remaining ~~bias-~~adjusted models are ~~evaluated~~validated. Thus, the trick is that we know both present and future pseudo-~~reality~~observations.

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

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 scales. This has been made possible with the advent of the EURO-CORDEX, a large ensemble of high-resolution RCM simulations with precipitation at~~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~~validation, than using ~~un-corrected~~ 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 ~~techniques~~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 ~~a summary~~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 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~~ 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				

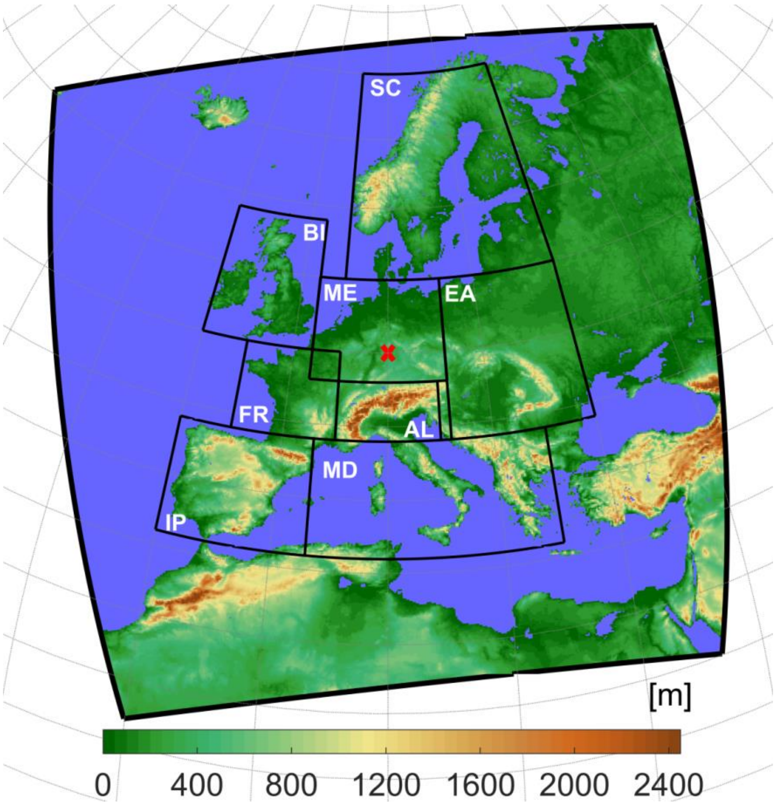

Figure 1. Map showing the EURO-CORDEX region (outer frame) with elevation in colours. PRUDENCE sub-regions (Christensen and Christensen 2007) used in the analysis are also shown: BI = British Isles, IP = Iberian Peninsula, FR = France, ME = Mid-Europe, SC = Scandinavia, AL = Alps, MD = Mediterranean, 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 criteria and come with major caveats that impact the uncertainty range largely (Madsen et al. 2017). We therefore choose, in accordance with most other similar studies, to use an 'ensemble of opportunity' for the present study.

3 Methods

3.1 Duration

Extreme precipitation statistics ~~is~~ are often described as a 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~~ calculated in a sliding window with a one hour time step~~ping~~. 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~~provides methodologies to estimate~~ 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

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

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

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

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 periods. Hosking and Wallis (1987) investigated the estimation of parameters of the GPD-distribution and based on this warn~~s~~ against using the often applied maximum likelihood estimation for a sample size below 500. Instead, ~~he~~ they recommend~~s~~ 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.

3.3.2 Analytical quantile ~~matching~~ mapping 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~~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~~s~~ 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~~ 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-~~matched~~ mapped~~d~~ to obtain the model bias, which ~~then~~ is then applied, ~~according to~~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 verifying predictions, where the performance of the prediction should be superior to the performance of reference predictions, such as persistence or climatology.

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 raw scenario model output data 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~~ Raw RCM scenario (Reference)

FAC	Climate factors on return levels
DC	Quantile-~~matched~~ mapped delta-change based on EVA
BC	Quantile-~~matched~~ mapped bias correction based on EVA

3.4 The inter-model cross-validation procedure in detail

The inter-model cross-validation goes in detail as follows: Each of the N models are successively regarded as being pseudo-observations. The individual adjustment methods are calibrated on the present-day parts of the pseudo-observations and model return levels (present-day and end-21st-century), as appropriate depending on whether it is a bias correction or delta-change method. The calibration is done as described above. The adjustment methods are then applied to present-day observation and model data, again as appropriate, to obtain end-21st-century adjusted return levels. These are then validated against the end-21st-century return level from pseudo-observations.

The basic validation metric will be the relative error of end-21st-century return levels for a given duration and return period T:

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

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

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

4 Results

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

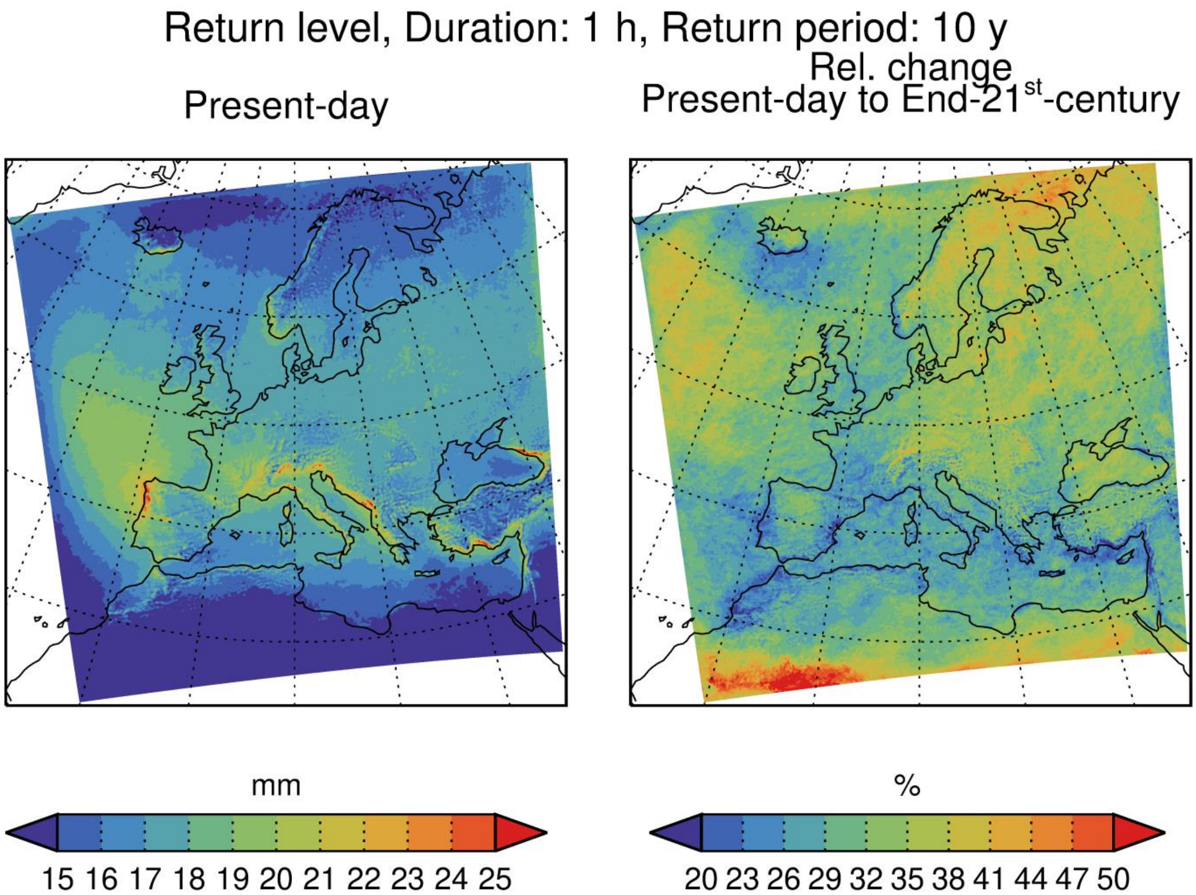

Figure 2. Geographical distribution of the 10 year-return level of precipitation intensity for 1 hour duration for present-day (left) and relative change from present-day to end-21st-century (right). In each grid point, values are the median return level over all 19 model simulations.

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 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 stand out as have higher return levels than their surroundings. This supports that the models are not totally unrealistic in modelling extreme precipitation.

There is a general increase in the range of 20-40% from present-day to end-21st-century climatic conditions. The relative changes are geographically quite uniform across the area. For instance, no evident difference between land and sea appears. Likewise do the mountain~~e~~ous regions not stand out from the surroundings.

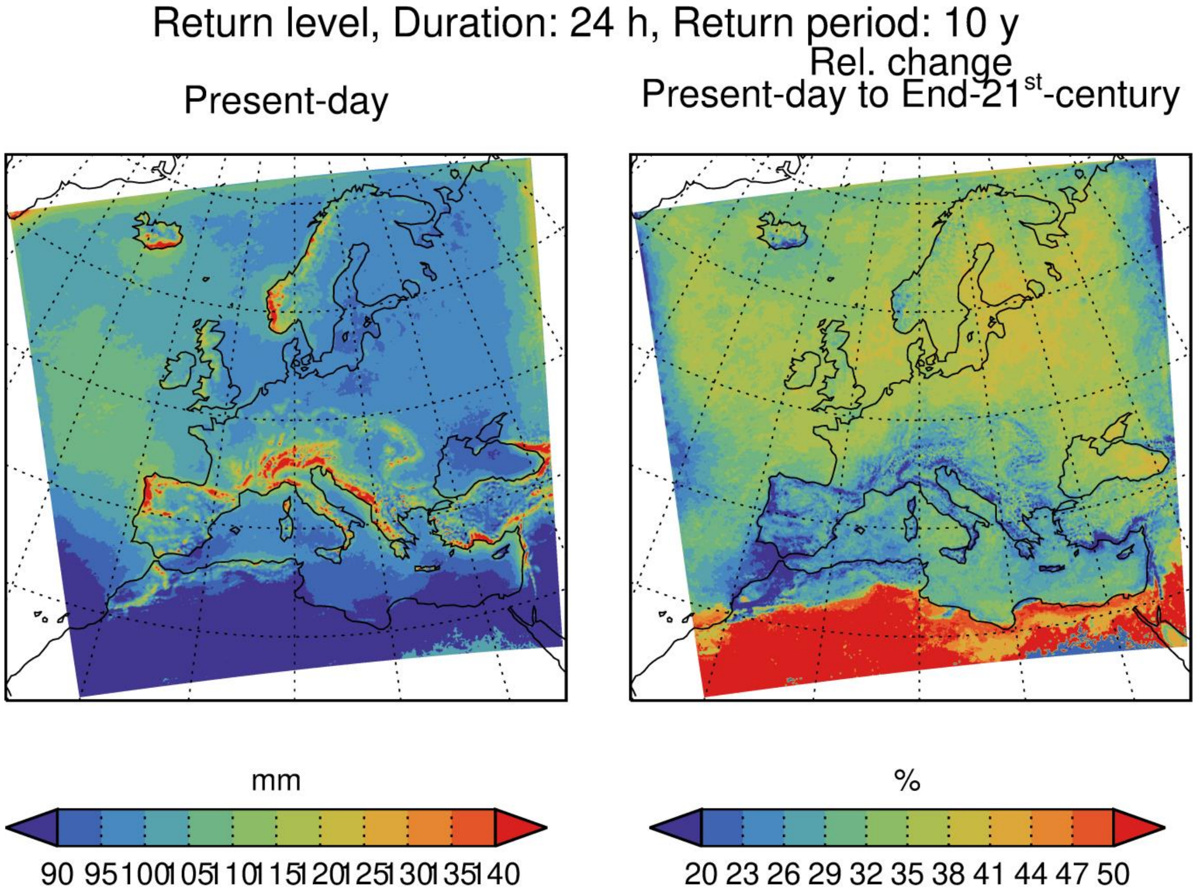

Figure 3. As Figure 2 but for 24 h duration

We also show in Figure 3 the median 10-year return level for 24 h duration. Again, the largest return levels are found in southern Europe and northwest of the Iberian Peninsula. Also, the ~~montaneous~~mountainous regions stand out with higher return levels even more pronounced than for 1 h duration. ~~, and this shows similar qualitative characteristics: For both durations t~~The return levels generally increase from present-day to end-21st-century conditions~~, although the effect is more pronounced~~ with around the same percentage as for 1 h duration and also geographically homogeneous.

.

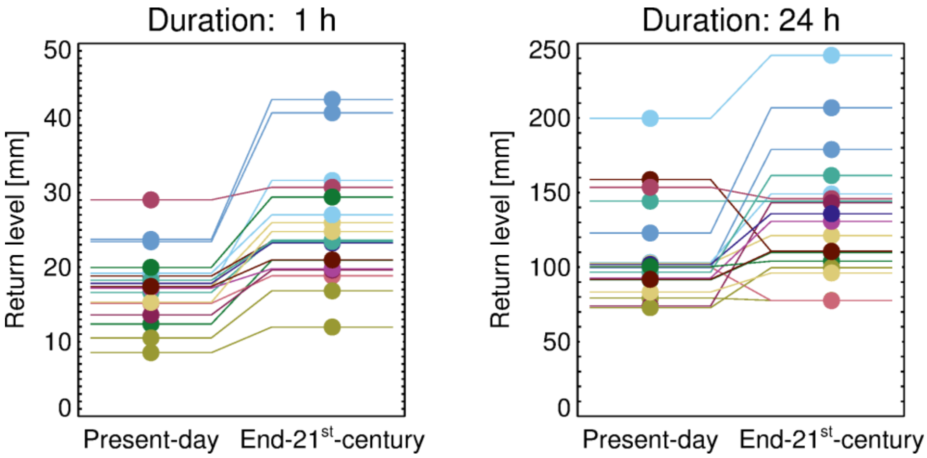

Figure 4. Modelled return levels at 50N/10E (northern Germany, marked with 'X' in Figure 1) for present and future for 10 y return period and 1 h and 24 h durations. Different colours represent the 19 different GCM-RCM simulations listed in Table 1.

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. 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~~

$$\cancel{RE = |P_T - V_T|/V_T}$$

~~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 combinations of model/pseudo-~~reality~~ observations~~combinations~~. Second, we will, for each adjustment method and for each combination of model/pseudo-~~reality~~ observations~~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 combinations of ~~-~~model/pseudo-~~reality~~ observations~~combinations~~ by showing the median and the ~~0.0~~5/~~0.~~95-percentiles of this distribution.

~~4.2.2~~4.2.1 Results for 1 h duration

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

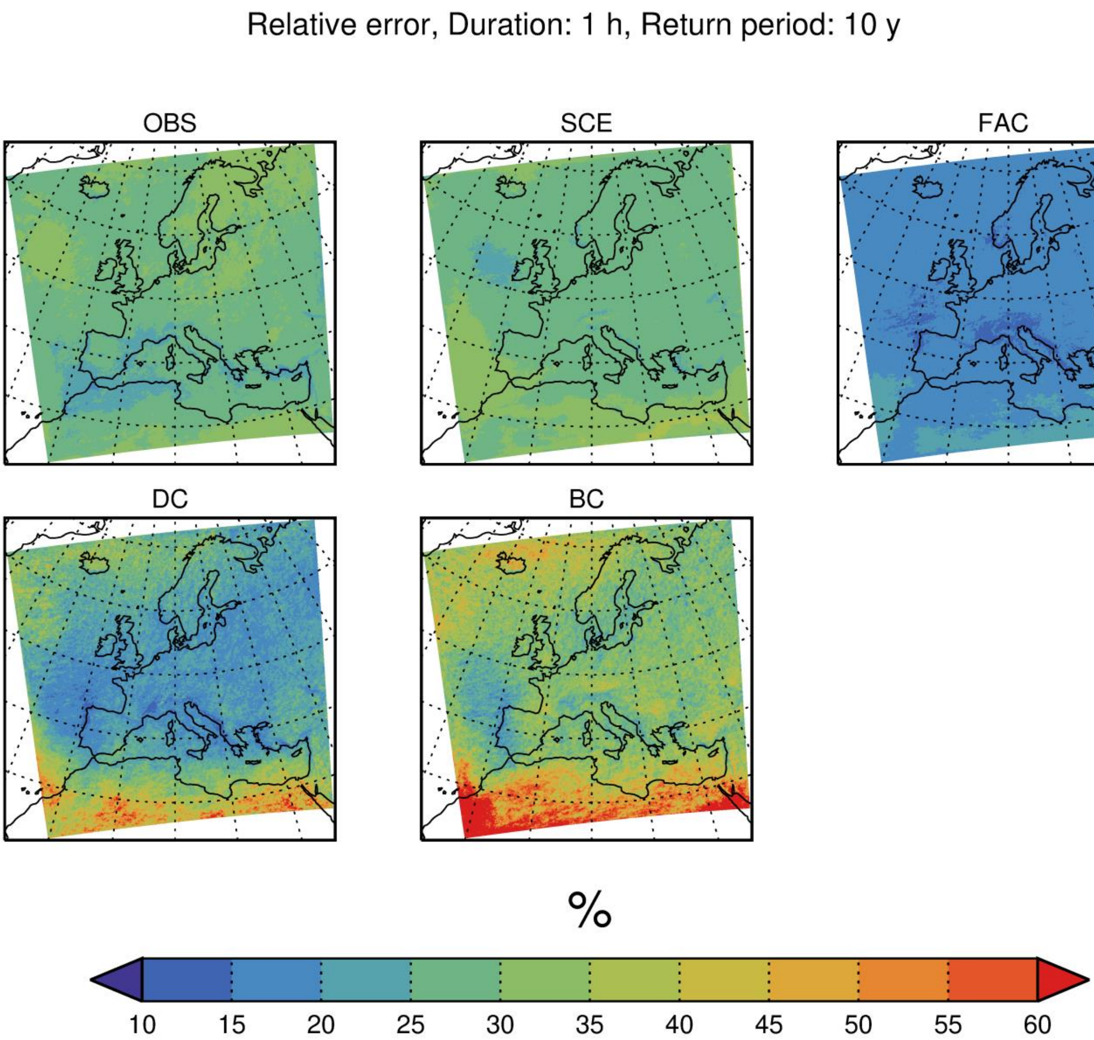

Figure 5. Geographical distribution of the relative error of end-21st-century 10 year return level for 1 h duration precipitation intensity from the inter-model cross-validation. Colours show the median of the relative error calculated over all model/pseudo-~~reality~~ observations combinations. Panels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

First we look at the reference methods. ~~The~~ Relative errors from the OBS method ~~has relative errors in the approximate interval~~are in the range of ~~0.~~20-~~0.~~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 ~~0.~~25-~~0.~~45%, lowest values in the Atlantic west of Ireland; largest values over parts of the Atlantic and northern Africa. The two reference methods give on the whole rather similar results, but ~~Of the two reference methods,~~ the OBS method slightly outperforms SCE in the south, while the opposite is true in the north.

The relative error of FAC is below ~~0.~~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, the relative error of BC is everywhere above both DC and FAC, indicating the poorest performance of all methods considered.

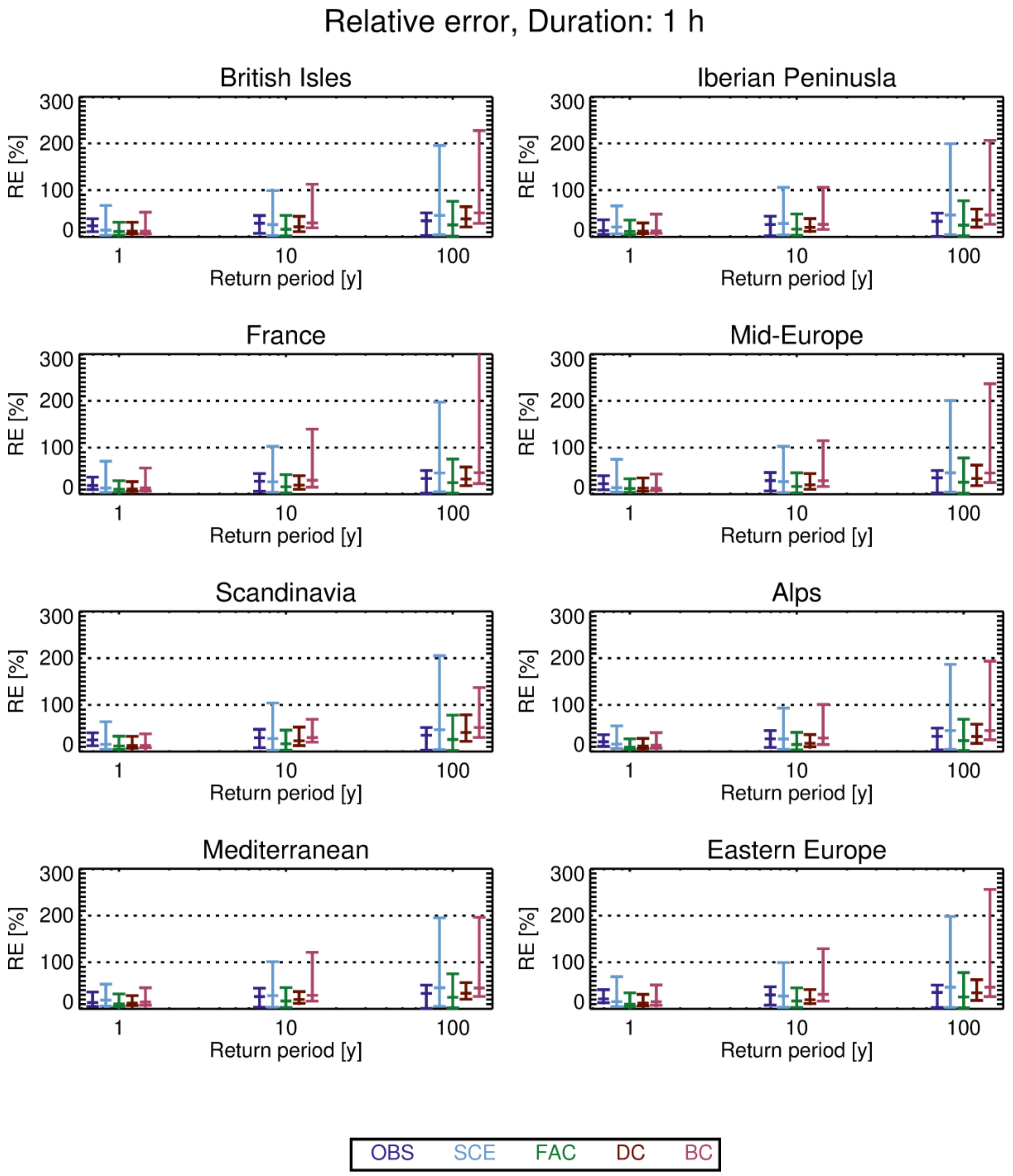

Figure 6~~.~~ Statistical distribution (median and ~~.0~~5th/~~.~~95th ~~-~~ ~~fractiles~~percentile) of the relative error of the inter-model cross-validation for 1 hour duration for 1 y, 10 y and 100 y return periods. Panels represent PRUDENCE sub-regions shown in Figure 1. Each colour represents a~~n~~ 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 ~~.~~95th ~~fractile~~ percentile is much larger for SCE than for OBS, in particular for large return periods. Thus, OBS overall performs better than SCE, -meaning 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 ~~.~~95th ~~-fractile~~percentile of the relative error. ~~Of the two quantile-matching mapping methods, t~~Thethe DC method has a slightly poorer performance than FAC, both in terms of the median and the ~~.~~95th ~~-fractile~~percentile of the relative error. Finally, BC has poorer performance than DC, when comparing the median of the relative error and in particular for the ~~.~~95th ~~-fractile~~percentile.

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~~mapping, DC and BC, the latter having the poorest overall performance of them all. Note that DC is comparing GPDs from the same model, whereas BC is comparing GPDs from different models. If the difference, in terms of GPD parameters, between two models in the present-day climate is typically larger than the difference between the same model in present-day and end-21st-century climate, it can explain the different results.

~~4.2.3~~4.2.2 Results for 24 h duration

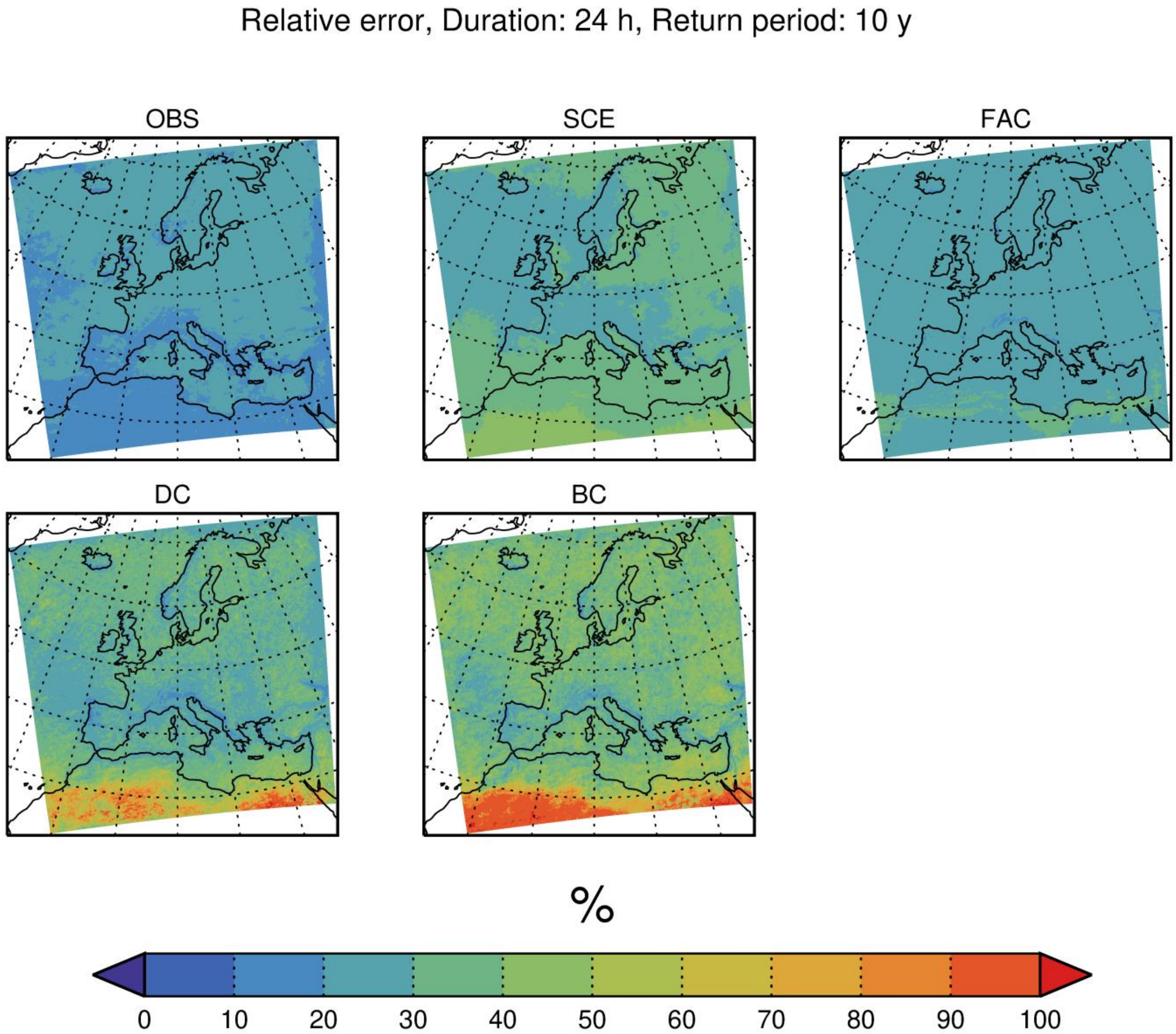

Figure 7. As Figure 5 but for 24 h duration.

For 24 h duration (see Figure 7), OBS has the lowest median relative error (~~lower~~ less than ~~0.~~30%) in most regions of all the adjustment methods, while SCE has higher relative error in the interval ~~0.~~30-~~0.~~60% approximately, with the highest values in North Africa. FAC has relative errors in-between those of OBS and SCE. Of the quantile-~~matching~~ mapping methods, DC has relative errors in the interval ~~0.~~20-~~0.~~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.

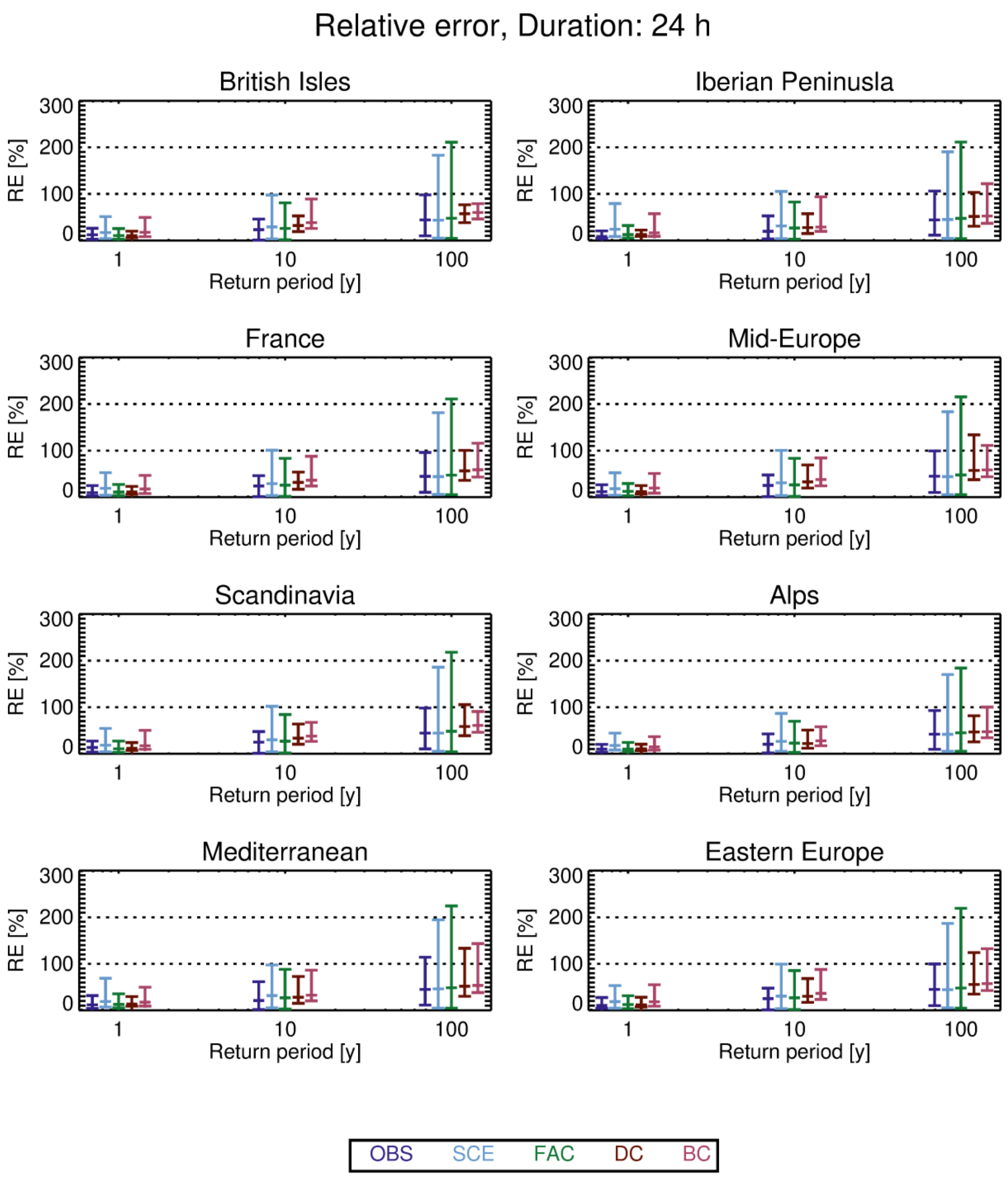

Figure 8. As Figure 6 but for 24 h duration

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

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

4.2.3 Ensemble median

Also inter-model cross-validation of pseudo-observations against model ensemble median, as described in Section 3.4, was carried out. For duration 1 h, distribution of the relative error is shown in Figure 9. By comparing with Figure 6, the distribution of the relative error does not change much overall. However, for many of the sub-regions considered and for the longer return periods, the FAC and BC have a smaller 95th percentile for cross-validation against model ensemble means, than against individual models.

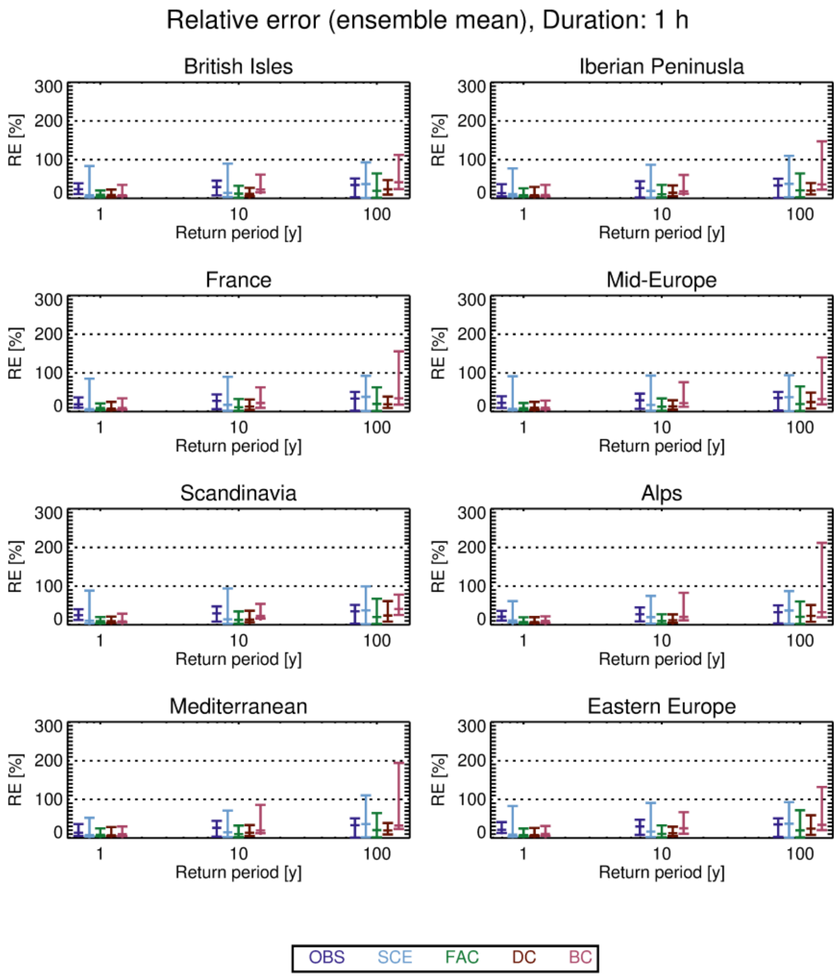

Figure 9. As Figure 6 but for inter-model cross-validation against ensemble medians.

Also for 24 h duration the distribution of the relative errors does not change much when shifting to validation against ensemble median (not shown).

4.3 Further analysis on conditions for skill

To get further insight into the difference in performance between hourly and daily precipitation, we consider for a given return period the relationship between the bias factor for present-day $\cancel{B_F} B_{P,T} = \frac{C_T \cancel{C}}{O_T \cancel{O}}$ and end-21st-century $\cancel{B_F} B_{F,T} = \frac{S_T \cancel{S}}{V_T \cancel{V}}$ for all model/pseudo-~~reality~~ observations combinations (see Figure 10).

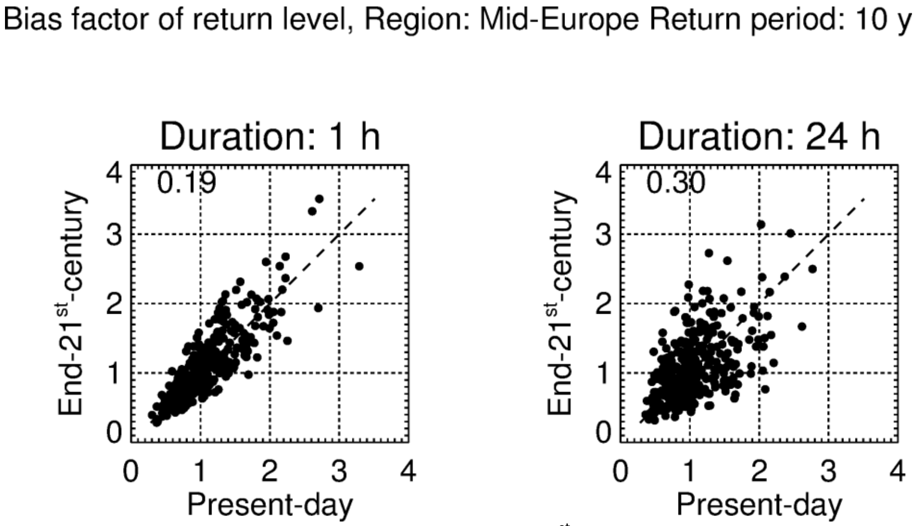

Figure 10. Relationship between present-day and end-21st-century bias factors of 10-year return levels for Mid-Europe sub-region for all pseudo-observation/model combinations. Left panel: 1 h duration and right panel: 24 h duration. Numbers in upper left corners are the $\boldsymbol{R}$ ~~measure of relative spread~~indices. 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, t~~To quantify the strength of the above relationship, we ~~use the measure of the relative spread~~define an index ~~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 combinations of model/pseudo-~~reality~~ observations~~combinations~~. This index is an extension of the index introduced by Maurer et al. (2013). It is the ensemble average of the relative absolute difference between the present-day and future bias. A value of $R = 0$ means these biases are equal, i.e. perfect stationarity; and the smaller the value of R, the closer to stationarity (in an ensemble sense).

~~These~~ Values of R are given in the upper left~~er~~ corner of each panel of Figure 10 and they~~,~~ 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~~ method~~s~~ seen in the previous section. Suppose that $\cancel{B_F} B_{P,T} = \cancel{B_F} B_{F,T}$, 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.

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 ~~.~~99.9th ~~fractile~~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 ~~also~~ 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. ~~Another relevant study to discuss here is~~ The CDF-t method was applied by Laflamme et al. (2016) ~~who apply the BC method similar to ours to~~ on daily New England 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~~ 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)~~(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.

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 10) 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~~ total amount of simulated ~~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 10. 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~~ to observed conditions (Refsgaard et al. 2014; Haerter et al. 2015).

5.5 Adjustment methods not included in the study

Only the basic adjustment methods have been included in our study. The simple climate factor approach has been applied in numerous hydrological applications (Sunyer et al. 2015; DeGaetano and Castellano 2017) and others. We also wanted to test quantile-mapping approaches, which in extreme value theory takes the form of a parametric transfer function. This we have applied in two flavours in the spirit of (Räty et al. (2014). Finally, we wanted to benchmark against the 'canonical' benchmark methods: observations and raw model output.

There is a myriad of more specialised methods, each tailored to account for a particular deficit of the simpler methods. First, there is the issue whether it for precipitation is more reasonable to map relative quantile changes rather than absolute ones (Cannon et al. 2015). It has also been argued that a bias correction method should preserve long-term trends, i.e. the 'climate signal' and only adjust the shorter time scales, as extensively discussed in (Cannon et al. 2015). Then multivariate methods have been argued for and applied in order to preserve relationships between variables (Cannon 2018). Also methods to correct for systematic displacement of variable features in complex terrain have been suggested and applied (Maraun and Widmann 2015). Finally, Li et al. (2018) adjusts stratiform and convective precipitation separately instead of adjusting the total precipitation. In this way, any future change in the ratio between the two types of precipitation is accounted for.

It could be interesting to examine the above methods in future studies, though we acknowledge it would be a quite extensive work. We can at present only guess about the outcome of such work but the more refined methods may not perform too well in the inter-model cross-validation setting. The reason for this suspicion is that these methods, while being more elaborate, in most cases also have more parameters to be estimated, implying a higher risk of overfitting. An argument in favour of this is that the present study shows that the more elaborate quantile mapping methods DC og BC do not outperform the simpler FAC method.

6 Conclusions

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 ~~.~~95th ~~-fractile~~percentile of the 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. 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 ~~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

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

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, https://doi.org/10.1038/NCLIMATE1454.

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.

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

——, S. R. Sobie, and T. Q. Murdock, 2015: Bias Correction of GCM Precipitation by Quantile Mapping: How Well Do Methods Preserve Changes in Quantiles and Extremes? *J. Clim.*, **28**, 6938–6959, https://doi.org/10.1175/JCLI-D-14-00754.1.

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

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.

Coles, S., 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.

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

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.

Gleckler, P. J., K. E. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. *J. Geophys. Res.*, **113**, D06104, https://doi.org/10.1029/2007JD008972.

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

Gutiérrez, J. M., and Coauthors, 2019: An intercomparison of a large ensemble of statistical downscaling methods over Europe: Results from the VALUE perfect predictor cross-validation experiment. *Int. J. Climatol.*, **39**, 3750–3785, https://doi.org/10.1002/joc.5462.

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.

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.

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

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.

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

Klemeš, V., 1986: Operational testing of hydrological simulation models. *Hydrol. Sci. J.*, **31**, 13–24, https://doi.org/10.1080/02626668609491024.

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.

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

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

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

Madsen, M. S., P. L. Langen, F. Boberg, and J. H. Christensen, 2017: Inflated Uncertainty in Multimodel-Based Regional Climate Projections. *Geophys. Res. Lett.*, **44**, 2017GL075627, https://doi.org/10.1002/2017GL075627.

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.

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

Maraun, D., 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.

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

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.

Rowell, D. P., 2019: An Observational Constraint on CMIP5 Projections of the East African Long Rains and Southern Indian Ocean Warming. *Geophys. Res. Lett.*, **46**, 6050–6058, https://doi.org/10.1029/2019GL082847.

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.

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

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.

Van Schaeybroeck, B., and S. Vannitsem, 2016: Assessment of calibration assumptions under strong climate changes. *Geophys. Res. Lett.*, **43**, 1314–1322, https://doi.org/10.1002/2016GL067721.

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