Reviewer 1

I would like to thank the authors for considering the comments from me and the other reviewers. The inclusion of an description of the HAPPI protocol explains a lot and will avoid future misunderstandings. Unfortunately, the new text introduces some new unclear points that have to be resolved; and the paper doesn't make the impression of being carefully read through. Mainly, there are two unresolved things: the shift in distribution of ATG28 and the frequency of dry periods. I, and other reviewers, commented on this in the last review, but I don't think the authors could motivate well enough in their response why they didn't agree with our comments. If the authors don't want to change the text they should be able to motivate why, and explain how the analysis supports their conclusions.

I don't want to come across a unnecessary disagreeable, but as a reviewer I am, in a way, responsible for this paper and that it is of good quality when its published. Comments follow below:

A: We thank the reviewer for the additional comments, and we appreciate the opportunity to clarify several issues and improve the paper.

L81-82: "a weighted sum of RCP2.6 and RCP4.5 is calculated with a global mean temperature response of 2.05°C" I don't understand this. What is weighted? Two time slices of RCP2.6 and RCP4.5? but RCP2.6 never reaches 2.05. Is it the sum of RCP2.6 and RCP4.5 at the end of the century? What is the level of warming in RCP4.5 at the end of the century? How is the sum weighted?

A: It is a weighted sum of the multi-model global mean temperature responses between RCP2.6 and RCP4.5. Mitchell et al. (2017) used the following formula W1 × RCP2.6 + W2 × RCP4.5 with W1 = 0.41 and W2 = 0.59, which results in a 2.05°C global mean temperature response. We have changed the sentence to:

"Therefore, Mitchell et al. (2017) calculated a weighted sum of the RCP2.6 and RCP4.5 multi-model is calculated with a global mean temperature response using the following formula: W1 × RCP2.6 + W2 × RCP4.5 with W1 = 0.41 and W2 = 0.59. The results adds up to of 2.05°C, which is exactly 0.5°C more compared to the chosen 1.5°C period."

L83: "sea ice extend" → "sea ice extent"

A: Changed accordingly.

L127-128: "All four climate indices are calculated from the daily mean precipitation, temperature, and/or dew-point temperature output of the model; for each year and ensemble member" All four indices are not calculated from the daily mean precipitation, temperature, and/or dew-point temperature. I would suggest changing to: "All four climate indices are calculated for each year and ensemble member"

A: Changed as suggested

L164: "For each of the climate indices /.../ computed as follows". Since the change of the indices are calculated in different ways, I don't think this is the best way to start this section. I would go for something like: "The future change of the climate indices are computed as follows:"

A: Changed as suggested.

L165-166: This seems to be a very complicated way to explain what you are doing. Why not just? “For ATG28 we calculate the differences of the 5th, 50th and 95th percentiles between the current period and the projected periods”.

A: Changed as suggested.

L166-167: “20 exceedances” Per month, year, or in the whole 10 years?

A: This refers to 20 exceedances during the current period, which corresponds to the entire 10 years.

L172: RX5day also show relative change (Fig. 4), it is not computed as a simple subtraction, and not similar to ATG28. Don’t use “similar to ATG28” it is only confusing.

A: We deleted this part and split the sentence into two:

“Differences for RX5day are computed by subtracting the ensemble mean of the current decade from the 1.5°C and 2.0°C periods. Statistical significance for RX5day was calculated using a Mann-Whitney-U-test and only results are shown with a significance at the 95% level.”

L177-178: Mann-Whitney is not mentioned in the description of ATG28, but it is in RX5day. Either mention Mann-Whitney in the description of ATG28, if you use it, or change to RX5day. Don’t use “similar to ATG28” it is only confusing.

A: The sentence is changed to:

“Differences in the distribution of CDD are calculated with a Mann-Whitney U-Test with a significance at the 95% level, determining whether samples from the two periods are drawn from a population with the same distribution.”

L184: Also add an explanation of grey boxes over ocean.

A: Added one sentence:

“Ocean boxes are masked out, because any change is very closely related to the prescribed SSTs.”

L189: “no shift in the distribution”. Maybe I just don’t understand what you mean, but I still don’t see this. When I look at figs 2 d-f and 3 d-f I see that in some areas the 95th percentile increases more than the 50th and 5th percentiles. To me this means, not only a shift in the distribution, but also a change in the shape of the distribution. I made a comment about this in my previous review. That you “have little reason to think that the shape of the distribution is changing” is not a satisfactory reply to that. I see a reason that the shape is changing: for temperature extremes warming is not linear and it is known that different parts of the distribution changes in different ways. A recent example of this, using HAPPI data, is Lewis et al. (2019) (https://www.sciencedirect.com/science/article/pii/S2212094719300556). They identify temperature hotspots where the tail of the temperature distribution increases with mean land surface warming at a faster rate than the rest of the temperature distribution. Two of these hotspots are central Europe and the Mediteranean. It would be easy for you to calculate the distance between the 5th and 95th percentiles for the current, 1.5 and 2 simulations respectively. That would give you some indication.

A: We think you are referring to the line: “there is no change in the shape of the distribution” and thank you for pointing this out. A (significant) shift can be seen everywhere, where there are yellow to red colours. We computed the distance between the percentiles as you requested. For many areas, we only see changes of +-1 or 2 days. But indeed we can see a change in shape for the

Northern part of Spain, the French Atlantic Coast and parts of Eastern Europe. Will change the sentence to:

"For the Northern parts of Spain, at the French Atlantic coast and parts of Eastern Europe, we can note stronger changes for the 95th compared to the 5th percentile. This means that the shape of the distribution changes towards more high extreme values. For most of the other regions there is only little or no change in the shape of the distribution of ATG28."

Fig 2 Explain grey areas over land. Change to a figure of higher resolution in the next version.

A: Explanation added. Better resolved versions are available. We apologize for the poor quality in the draft version.

Fig 3: See comments on Fig 2.

A: See above

Fig 4: Change to a figure of higher resolution in the next version.

A: See above

L225: What do mean by "To account for"?

A: We rephrased the sentence to:

"Figure 5 shows spatial differences the in 50-year return period of daily rainfall intensity (RI50yr) across Europe"

L238: "statically" → "statistically"

A: Changed accordingly

L239-240: This is a highly confusing sentence. It seems like you are comparing your p-value to 0.05 or a number smaller than that (alpha), and if the p-value is greater than that the difference is significant. However, if alpha is much smaller than 0.05, for example 0, all p-values will be greater. Also, you should have a fixed alpha, either its equal to 0.05 or something else. Furthermore, shouldn't your p-value be smaller than 0.05 to be significant? Consider rewriting to something like: "When the resulting p-values of the test are smaller than or equal to a significance level of 0.05, the null hypothesis is rejected indicating that the distributions differ."

A: We have simplified the sentence as suggested.

L242-243: This is repeating what is written above, remove.

A: We removed the sentence, which was accidentally repeated.

L250: "1.5°C vs. 2.0°C" Do you mean "1.5°C and 2.0°C"?

A: Yes. Changed as suggested.

L252-254: I still don't agree that you can conclude that the dry periods will occur more often, but after going through the reviewers comments and the responses again I start to see where the misunderstanding comes from. Remember that you only have the longest dry period for each grid point, you don't know how many dry periods you have. Therefore you can't say anything about the frequency of the dry periods. It could be that a prolongation of the longest dry period means that all dry periods will be longer (same or increased frequency). It could also be that several short dry period are replace by one long (decreased frequency). We don't know that. You could, however,

argue that within a region longer dry periods will be more probable. I suspect this is what you mean, correct me if I'm wrong. Let's take the example of the Iberian Peninsula in ECHAM. It's clear that there is a shift in the distribution between "hist" and "2.0". The longest dry period will be longer, and the chance that a gridpoint will experience a dry period of, let's say, 10 weeks is much bigger in "2.0" than in "hist". Does this mean that longer dry periods will occur more often (more frequent)? Not necessarily. Youcould say that its more likely in the future that the longest dry period somewhere on the Iberian Peninsula will be longer than 10 weeks, than today. The common definition of frequency is how often something happens in time, not in space. Remember also that the different gridpoints are not independent. It's likely that several of the longest dry periods in different gridpoints occur at the same time. It could be that all of the longest dry periods occur at the same time.

A: We would like to thank the reviewer for taking the time to elaborate on the CDD index study. Using the example of the Iberian Peninsula helped identify where our misunderstandings have arisen. We agree to eliminate any reference to 'frequency' as it has obviously lead to this misunderstanding. Both the reviewer and authors agree that the longest dry period of the three regions in Fig.6 is longer under 2.0°C than current conditions. The authors use of the term frequency was in regard to the changes in CDD distribution, which the reviewer has articulated ought to be written in terms of probability or likelihood, and as such the text has been changed (see L251, L254).

It should be noted that the CDD is calculated after a spatial mean of the Prudence region is taken and thus representative of the whole domain, not calculated per grid-box as the reviewer has written. This may have also lead to some of the misunderstanding. We have elaborated on the methodology as a result.

L253: I think correct English is "indistinguishable from" not "indistinguishable compared to"

A: Yes, changed as suggested.

L253-254: This is a ambiguous sentence. It's true that you can deduce that region 2, will suffer from more frequent and longer drought periods than experienced before. But compared to regions 6 and 8? Even if the increase in 2 is larger than in 6 and 8, the dry periods in 6 and 8 could still be longer than in 2 (I know it isn't, but still). I guess what you're after is that the longest dry period changes more in some regions (e.g region 2) and less in others (e.g. regions 6 and 8). L256-258: With the same kind of argument you could say that the British Isles would benefit from an temperature increase of 2 instead of 1.5. From Fig 6 its obvious that several regions would benefit from a lower temperature increase. Wouldn't you agree that the Iberian Peninsula would benefit more than the Mediterranean from a reduce warming? Although the change at 1.5 is already significant.

A: We have removed the term 'suffer from' and reformulated the sentence to state the facts.

L287: Please insert "RI50" somewhere here.

A: Added as suggested.

Reviewer 2

The current version clearly improves the paper. However, several parts are still not properly described/discussed, although several points were already raised during the previous review round. further revisions are necessary before your study can be considered for publication. Please refer to the main points and specific comments below.

A: We thank the reviewer for the comments, and we respond to the individual comments, below.

MAIN POINTS

---

1) HAPPI protocol: The description of the HAPPI approach is much improved. However, it still leaves out many aspects on the validity and limits of the approach, for instance with respect to the climate variability. This was already requested during the previous review process, but is not sufficiently addressed in the current version of the manuscript. For example, you could extend the Supplementary and include a comparison of the temporal variability (and not only the annual cycle) between the HAPPI simulations and observations.

A: We added a section in the discussion addressing this point, where we discuss the findings of Fischer et al. (2018) on AMIP vs. Coupled ensembles. This includes the aspects of atmospheric vs. oceanic variability and where we argue that our data should be mostly used for statistic based on daily or multi-daily data rather than seasonal or yearly.

2) Section 2.4:

2a) It would be worth a try to combine this section with the previous one (2.3). The reader would then have a direct overview of the individual indices (including calculation, changes and significance). Additionally, you could avoid the repetitive use of some phrases.

A: We agree with Reviewer#2 and merged the corresponding parts of section 2.4 with 2.3.

2b) Are the changes calculated from the ensemble mean or median?

A: This depends on the index. In case of ATG28 among others the 50th percentile is computed, which corresponds to the median. In case of RI50yr the relative difference is based on the mean. We added this information to the text.

2c) The calculation of significance is still unclear. There is no significance test for RI50yr, correct? The description for CDD is confusing: L177 states that differences are calculated similar to ATG28 with a Mann-Whitney U-Test, but you did not use this test for ATG28 but for RX5day… See also main comment 4h) of my first review.

A: For RI50yr no significance test has been done. We have changed this section in response to Reviewer#1 by deleting the confusing “similar to” statements. In addition, merging section 2.3 with 2.4 as suggested in comment 2a) should make clearer how each index and the corresponding change has been computed.

3) Discussion and conclusions: This part is still insufficient. In the previous round, the reviewers raised the important point on how your results relate/compare to other simulations (e.g. the MPI grand ensemble) and to other studies. Laura Suarez-Gutierrez even provided some examples. Nonetheless,

you do not consider any of these suggestions. There is not a single reference throughout the whole discussion and conclusion section. Instead, you argue that such a discussion is out of the scope of your paper. However, in order to provide convincing evidence to prove the benefits of your method this must not be neglected. It is even mandatory.

A: We agree that the discussion should include these examples. During the first round we misinterpreted Laura Suarez-Gutierrez comments and thought that we should include a full discussion on different GCM modelling approaches. We thank Reviewer#2 for this clarification and expanded the discussion and conclusion part accordingly.

4) Language: The language has improved significantly. Nevertheless, further revisions are needed. There are still some incomprehensible sentences, rather colloquial wording and superficial descriptions. Please also refer to the specific comments below.

SPECIFIC COMMENTS

---

1) References in the text should be sorted either chronologically or alphabetically (see e.g. L30f, L65).

A: We have double-checked the references. ESD also offers the option of sorting by relevance, which we followed in the specific case of L30. L65 has been split and in all other cases we followed chronological sorting (which often coincides with relevance anyway).

2) L11/12: "… for periods with levels of 1.5°C and 2.0°C global warming …" – Please reword.

A: Deleted "levels of".

3) L20/21: "… relevant to both the private and the public sector."

A: Changed "sectors" to "sector".

4) L32: "However, Mitchell et al. (2016) argue that …"

A: Changed accordingly.

5) L42-46: Please reword sentences and remove doubling of "bridge the gap between GCM model output and regional climate impact assessment".

A: We have rephrased this paragraph to:

"Regional climate impact assessments often require a much higher resolution than GCMs (e.g., Giorgi and Jones, 2009). To bridge this gap, dynamical downscaling with Regional Climate Models (RCMs) is one option, which provides physically consistent high-resolution climate information (Jacob et al., 2014; Giorgi and Gutowski, 2015; Gutowski et al., 2016). Here, the RCM REMO (Jacob et al., 2012) is used to dynamically downscale simulations from two GCMs of the HAPPI consortium. Two regional climate datasets of 25 and 100 members are developed, to create a large ensemble of RCM simulation, which are particularly suitable to study extremes."

6) L59: Could you add a reference for the IPCC Special Report?

A: Added as requested.

7) L59-61: Please reword sentence.

A: Section has been rephrased to:

"The HAPPI protocol by Mitchell et al. (2017) has been set up to inform the IPCC Special Report on 1.5°C Warming (IPCC, 2018). The idea is that large ensembles (>50 members) of GCM simulations will allow studying extreme events, even for the small differential warming between a current decade (2006-2015) and two future decades under 1.5°C and 2.0°C global warming."

8) L65/66: Reasoning unclear, please elaborate.

A: We rephrased the sentence to:

"All simulations were conducted in atmosphere-only mode in order to increase ensemble size, because of lower computational costs (Mitchell et al., 2017). In addition, He and Soden (2016) have shown that atmosphere-only mode simulations can provide more accurate regional projections, because they do not suffer from systematic biases such as SST drifts."

9) L71/72: "… temperature response for 2091-2100 compared to 1861-1880 is 1.55°C under RCP2.6."

A: Changed accordingly.

10) L75: "as" instead of "because"?

A: Changed accordingly.

11) L89-92: Could you include the difference between sponge zone and core domain in Figure 1?

A: Added to the plot.

12) L92-96: Please split sentence.

A: Changed accordingly

13) L128:/129: "… 1000 data points … 250 data points …" – Maybe include 100/25 members x 10 years?

A: Added information on members

14) Figure 1: Not every reader might be familiar with the PRUDENCE regions. Please explain abbreviations for individual regions in the figure caption. Some abbreviations are difficult read. Could you increase the readability?

A: We changed the caption of Figure 1 by adding the names of the PRUDENCE regions and improved readability inside the figure.

15) L172-174: Please split sentence.

A: Sentence has been rephrased in response to Referee #1

16) Caption Figure 2: "as" instead of "because"?

A: Changed accordingly.

17) L206: "are" instead of "is"

A: Changed accordingly.

18) Figure 4: The caption does not match the figure, you mixed up rows and columns.

A: We forgot to update the text after redoing the figure. The caption has been fixed.

19) Figure 5: The caption does not match the figure, you mixed up rows and columns. Additionally, there is a doubling of "in percent" and "in %".

A: We forgot to update the text after redoing the figure. The caption has been fixed.

20) L239-243: There is a doubling of the sentence “Where the p-value is greater or equal to …”

A: Sentence has been removed.

21) L239: “Where the resulting p-values are greater or equal to a significance level …” – The null hypothesis is that the distributions are the same, correct? Then it should be “Where the p-values are less or equal to a significance level…”. Otherwise you marked the wrong numbers in Table 1.

A: Thank you for pointing this out. Indeed, it is “less than”. This section has been rewritten on request of Reviwer#1.

22) Table 1: Why did you put the PRUDENCE regions in parentheses? “Hist” should be “current”.

A: Changed as suggested.

Reviewer 3

The paper has already gone through one round of review and in particular the description of methodology is in much better shape now than the initial submission.

A: We thank the reviewer for the second review, and this positive comment, and we provide replies to the specific comments, below.

The duality of the paper trying to both presenting the data set and examples on how it can be used still detracts somewhat from both. Although I realize that this is likely beyond the scope of the paper I think the analysis would have been more interesting if the indices were also compared to observations and/or reanalysis.

A: As mentioned by the reviewer, a detailed performance measure of each individual index is beyond the scope of our paper as they should be seen as examples of what can be done with the data. In addition, we already added a supplement with some general performance measures to the supplement during the first round of revision.

The presentation of the data set itself can however be easily improved either through writing that you follow a certain standard, e.g. CORDEX and / or write a brief description in the supplement. What are the main variables and output frequency. Are they available for others? The availability notice says where they are stored, but not if they can be accessed.

A: Currently the data is available on request via the DKRZ cloud (swiftbrowser). We follow (as much as possible) the CORDEX and CMOR data conventions. We added a table with all available variables and frequencies to the supplement.

Since the resolution of the RCM is still quite coarse and not that different from the GCM (0.4 vs 1 degrees) I also think it would be useful if the authors discuss if they expect the conclusions to stay the same for higher RCM resolution. Although the pattern will likely be more noisy would e,g, 25 simulations with 0.22 degree resolution look like the 25 member subset? In particular many of the important meso-scale precipitation systems in the Mediterranean region are not resolved in a model with 0.4 degrees resolution.

A: For such large ensembles, there is always a trade-off between resolution, model complexity and ensemble size. In addition, many modelling groups in HAPPI did not have the resources to save the data that is necessary to perform dynamical downscaling. Given that ECHAM6 in the HAPPI experiments is relatively course (T63 or 1.875°), a direct downscaling to much higher resolution than 0.44° should we avoided, because the spatial spin-up of small-scale features inside the domain would become too large (see e.g., Matte, 2017). We agree that some of the important meso-scale systems are missed with 0.44°. Also the prescribed SST may not be appropriate to correctly simulate coastal precipitation especially around the Mediterranean and we briefly discuss this already in our paper and expanded on it (see below). To our best knowledge we do not expect fundamental changes in a 25-member 0.22° ensemble compared to our 0.44° subset. Details might change, but the qualitative patterns should look the same (if we neglect spatial spin-up issues discussed above).

We will add the following lines to the discussion part:

"The relatively course resolution of 0.44° for a dynamical downscaling study over Europe is a compromise between model complexity, resolution and ensemble size. In addition, the course driving data particularly of the ECHAM6 model with T63 or 1.875° horizontal resolution does not allow a much higher resolution for direct downscaling without having a too large spatial spin-up of

small-scale features inside the domain (see, e.g. Matte, 2017). Besides details in some regions such as coastal precipitation around the Mediterranean, we do not expect fundamental qualitative changes to our findings on a higher resolution such as 0.22°."

line 3. "Dataset" is used two times in the same line

A: We changed the sentence to:

"The dataset is a unique and physically consistent, as it is derived from a large ensemble of regional climate model simulations."

line 70. anomaly → increment?

A: Changed accordingly

Line 90-95: Add information on GCM resolution.

A: ECHAM6 was run in T63 (1.875°) horizontal resolution and NorESM in 1.25°x0.94°. We added the resolution information to the text.

Line 97. How are the land use described in REMO for future scenarios?'

A: In REMO we followed the CORDEX approach and kept land-use constant for all periods. We added the information to the text.

"In REMO the same greenhouse gas forcings as for the GCMs were used and no land-use changes were applied."

Line 114. "one initial soil temperature state for every ensemble member in one period." But moisture profile is can vary?

A: We used this procedure for all soil variables. We deleted the word "temperature" from the sentence.

Line 139 "The index for the annual maximum. → minor quibble. Do you call it an index or can you just write The annual maximum …

A: Changed accordingly.

Figure 1: Perhaps obvious but can not see if IP is a subsection of MD (I presume it is)

A: We followed the original definition of the PRUDENCE regions where IP is not a subsection of MD.

line 215 "because the GCMs usually do not resolve these small basins. In these locations the SST is interpolated from the nearest SST value of the GCM, which might not be adequate for the region"

The basin is missing entirely? A larger grid size should just reduce the ocean fraction of the grid?

A: This refers to small sub-basins such as the Adriatic Sea. They are hardly resolved by the CMIP5 models and the question is, if such an SST is matching the regional characteristics. When looking at the HAPPI SSTs there is no West-East gradient detectable for the Adriatic Sea, because of insufficient resolution. We changed the formulation of that sentence and included a reference pointing out the importance of SST for heavy precipitation at the Adriatic Sea.

Matte, D.; Laprise, R.; Thériault, J. M. & Lucas-Picher, P.: Spatial spin-up of fine scales in a regional climate model simulation driven by low-resolution boundary conditions, Climate Dynamics, 2017, 49, 563-574