Answers to Referee#2

General response:

The authors would like to thank the referee for the time and effort put into this review, which has been useful in improving our manuscript. We have carefully read the comments, and provide a detailed response to all comments, below.

The paper investigates the impacts of 1.5°C and 2.0°C global warming on temperature and precipitation extremes over Europe. With this aim, the authors use an ensemble of dynamically downscaled simulations from the HAPPI project. The analysis focusses on four climate indices from ETCCDI.

The paper covers an interesting and relevant topic that could be a useful addition to the literature. Unfortunately, several aspects in terms of methods, analyses and results are unclear and need further explanation. In fact, the manuscript needs a lot of improvement in order to increase the clarity and readability. Please refer to the main points and specific comments below. In particular, the authors do not provide enough convincing evidence to prove the benefits of their method. Nevertheless, I suppose that there are sufficient arguments for it. Therefore, I suggest a major revision for this manuscript.

Answer: We will revise the introduction and methodology sections to emphasize the paper's objective of introducing a new dataset. In addition, we will increase the clarity and readability, by improving several sentences in response to the comments we received from this referee and the others. The method that the referee mentions, was developed by the HAPPI consortium. We will insert a more extensive description and motivation for this approach, thereby providing a better background. Finally, we also will provide more details on the assumptions and implications of using the HAPPI GCM simulations for regional downscaling using the regional climate model (RCM) REMO. We hope this will satisfy these main concerns of the referee.

MAIN POINTS:

1) Text: The whole manuscript needs a thorough proofreading and language check. Some sentences are incomprehensible; others are just too long and should be split in two to enhance readability. In addition, several phrases/descriptions are not consistent throughout the paper and therefore may confuse the reader. Please also refer to specific comments below.

A: The manuscript will be checked for language and revised to improve the flow, to include more detailed and consistent descriptions, and reduce sentence lengths. We will revise the introduction and methodology sections to emphasize the paper's objective of introducing a new dataset. Also, we use more consistent definitions for simulation periods, and other descriptions, to improve clarity.

2) Figures: The figures and their captions need some general improvement.

A: All Figures and captions will be improved as suggested below.

2a) Why are some land areas in Figures 2-4 grey?

A: The grey boxes are masked out areas. On land they refer to grid boxes that do not match our criteria of 20 or more occurrences of ATG28 during the current period. This will be stated explicitly in the text and caption in the revised paper.

2b) You should use the same format for all spatial plots, e.g. * One colour bar including units * Label individual plots * Same domain * Add more information in the figure itself (warming level, ensemble, . . .).

A: We used different plotting tools and will redo the plots with one tool following the suggestions.

2c) Figure 6: The bars are difficult to distinguish. Non-overlapping bars might be preferable. The distributions for both GCMs are very different. For NorESM, they are quite narrow, while they are much broader for ECHAM6. What are the implications of such large discrepancies? Please discuss. Why do you measure consecutive dry days in weeks?

A: This figure will be redone with non-overlapping bars. With regard to the different distribution widths, there seems to be a dependency on the forcing model. In our investigations, we see NorESM forced ensemble simulate wetter conditions, whereas ECHAM6 forced ensembles simulate warmer and drier conditions. This evaluation will be added in as supplementary material. This likely explains the differences in CDD distributions. The implication is that the absolute values coming from the models should be treated with caution, nevertheless we see a partially significant shift in CDD distributions of both ensembles. This demonstrates a qualitative (not quantitative) change towards longer dry periods. Lastly, CDD is measured in terms of weeks in the plots as it is more intuitive to interpret 8 weeks without precipitation greater that 1mm over a region than it is to read 56 days. In addition, it reduces the ink-to-data ratio in terms of visualization.

3) Is the main objective of this paper to present a new data set or to present mainly new results? Either way, both are not properly presented and need more details.

A: The aim of this paper is to present a new dataset. In addition, we provide 4 examples of how adaptation-relevant information can be derived from this dataset. We will make this clearer by improving the introduction and methodology sections. The latter will have a dedicated sub-section describing the HAPPI protocol with more details (see below).

4) Missing details/explanations: Some points (especially in the Methods section) need a better/more detailed explanation to be understandable. Your descriptions are too short and raise more questions than they answer.

A: Overall, the section on Methods will be substantially improved, as discussed above. We address below the point by point comments made on missing details and explanations also for the other sections.

4a) L58/59: Why did you use 20 years for the pre-industrial period, but only 10 years for all other simulations? From a climatological perspective, 10 years are rather short to enable a climatological view.

A: The pre-industrial period is only used as a baseline to define 0°C global mean warming. The definition is coming from the HAPPI protocol that every group doing global simulations followed. We are aware that there are several slightly different definitions of this particular period. The IPCC Special Report on 1.5°C warming lists several of these definitions. We will add a new sub-section on HAPPI to the manuscript (see below) to make these definitions more clear. Also the ten-year period is coming from the HAPPI protocol and is discussed in Mitchell et al. (2017). We will add the motivation for the ten-year period in the HAPPI sub-section.

4b) L58: Which period did you use for the future climate simulations? There are a few references given, but it should be specifically described which methods have been used here and how they are applied.

A: The method will be explained in more detail in the sub-section about HAPPI (we included a rewritten section in the response to referee#1).

4c) Did you compare the simulation for a current decade to observations and/or reanalyses to check how realistic they are? This would be very important.

A: We agree that comparing simulation results of climate models to observations is always very important to gain trust in models. The model version on this domain has been extensively evaluated using Re-analysis as boundary conditions in many other studies, and the relevant papers are already cited. In this paper, we want to demonstrate the benefits of using a large ensemble when looking at small changes in terms of global mean temperature - therefore an analysis against observations or reanalysis is less important. Especially since we only consider projected changes. We did comparisons in terms of quick views though, and concluded that the results were of comparable quality as, e.g., historical simulations from CORDEX with CMIP5 boundary conditions. However, as this point has also been raised by other referees, we will include a quick analysis on the general performance as supplement material.

4d) L59-61: How are the warming levels (1.5 and 2.0) calculated? Did you use RCP2.6 for 1.5 warming and RCP4.5 for 2.0 C warming?

A: The warming level has been calculated from a CMIP5 ensemble mean global mean temperature response. In case of the 1.5°C period RCP2.6 was used. The 2.0°C period is calculated using a weighted mean between RCP2.6 and RCP4.5. A more extensive explanation is given in the new subsection on the HAPPI experiment (please refer to response to referee#1).

4e) L73: The regional ensemble consists of 125 members, correct?

A: We have 125 members (100 from ECHAM6 and 25 from NorESM) for each of the three periods (current, 1.5° and 2.0° periods).

4f) L122-124: How exactly did you define CDD? Why did you calculate the CDD for the PRUDENCE regions and not on grid-point basis as the other indices? Is CDD the maximum number per year or over the 10-year period?

A: As we explain in the paper, the threshold is less than 1 mm per day (lines 123-124). We calculated the maximum duration for the entire 10 years of each ensemble member. The CDD analysis is computed for the Prudence regions and not 'per-grid-box' as for the other indices used in this study because applications drought indices are relevant over larger areas, whereas in the cases of the other indices considering high temperatures and heavy precipitation, analysis of individual grid-boxes are more relevant as they have more local applications. We will add this to the text.

L122 will be rephrased: "Lastly, the Consecutive Dry Days (CDD), defined as the maximum number of consecutive days with a daily precipitation amount of less than 1mm over a region (Karl et al., 1999; Peterson et al., 2001) is calculated for the entire 10 year period of each ensemble member. The CDD is calculated for each of PRUDENCE regions (Christensen et al., 2007), illustrated in Figure 1, because drought indicators are relevant over large areas.

4g) L130-145: Why did you choose different statistical methods to investigate the individual climate indices?

A: It is a common procedure to employ different statistical methods for different climate indices depending on the physical variables from which it computed from, for example, temperature or precipitation. This is because some statistical tests assume a given underlying distribution shape, for example temperature generally follows a normal distribution, whereas precipitation does not. Thus,

different statistical tests ought to be employed for climate indices derived from temperature versus precipitation. In addition, the use and application of the different indices also warrant different approaches and methods.

4h) L130-145: The application of the significance measures is unclear, please rewrite. There are two methods used for two different parameters (ATG28 and RX5day). There seems to be some confusion on what is used for CDD (L143 says CDD similar to ATG28, but the method for CDD seems to be similar to RX5day, namely Mann-Whitney). No information is given for RI50yr. Anyway, the paper provides only information on the significant changes for CDD, but not for the other parameters. This should be remedied.

A: We will rewrite section 2.3 and dedicate one sub-section to each indicator to better distinguish the used methods between them.

4i) Why do you think that all differences between the two ensembles are due to the different ensemble sizes (e.g. L172-175)? They could also result from the driving GCMs. It might be useful to include the results for the smaller sub-sample of ECHAM6 that you mentioned in the text.

A: As we stated already in the paper, analysed this with a random sub-set of 25 ECHAM6 members and found similar, noisier patterns like in the NorESM ensemble. This supports our conclusion regarding ensemble size. However, to provide this actual information to the readers, we will add the ECHAM6 sub-set plots as supplement, in the revised paper.

4j) Table 1: Partly, the smaller ensemble (NorESM) generates more significant results. How does this relate to the hypothesis that a larger ensemble size is beneficial?

A: We disagree with the reviewer, as it cannot be concluded on the basis of the CDD in Table 1 only that the NorESM smaller ensemble provides more robust results. This could be by chance, as the NorESM model could lead to relatively dry projections, for instance, leading to changes that are more significant. We have looked at four different indices (precipitation and temperature), and across the board the larger ensemble has less noise, regardless of the change, at the same level of warming.

4k) Compare your results to previous studies (see interactive comment by Laura Suarez-Gutierrez for more details).

A: We have provided responses to the comments from Laura Suarez-Gutierrez. We will include comparisons to other studies, but many of Laura Suarez-Gutierrez suggestions are beyond the scope of our paper. Please see our responses to her comments.

4l) The authors should not oversell their results (or should argue more convincingly). E.g. Impact of the ATG28 increase (L241-246): What does such a change really mean w.r.t health issues? I assume that the number of days above the threshold is already high around the Mediterranean? Does a change of O(10days) drastically change the base level and/or the potential health impacts in this region? Furthermore, is a resolution of 50km sufficient to derive change estimates for local adaptation measures? The authors do not provide enough convincing evidence for the benefits of their method. Nevertheless, I suppose that there are sufficient arguments for it, but this should be better phrased.

A: The chosen threshold is relevant in particular for sudden cardiac death exposure. The number of days in the baseline is of similar order. Please bare in mind that this indicator is not dependant on temperature alone, but that humidity also plays an important role. We agree that 50km might not be sufficient to inform local adaptation measures, especially if one thinks about cities. But in this regard

our results would serve as a lower bound of expected changes, because we have no urban heat island effects in our model. We will rework the discussion section to make it more convincing.

SPECIFIC COMMENTS

Data set, dataset or data-set?

A: We will use the term dataset and change the manuscript accordingly.

NOResm or NorESM?

A: We will use the term NorESM and change the manuscript accordingly.

L17: "measures at a" – Delete "a".

A: We will correct as recommended.

L20-22: This sentence is incomprehensible, especially the first part.

A: We will rephrase that sentence as follows:

"Identifying regional climate change impacts for different global mean temperature targets is increasingly relevant to both the private and public sectors. In the private sector, investors demand financial disclosure associated with climate change risks and opportunities (Goldstein, et al., 2018). In the public sector, policy makers rely on climate information build on internationally agreed limits to develop national climate action policies."

L25: "current generation global climate simulations" –> "current generation of global climate simulations"

A: We will change as recommended.

L45-47: Please rephrase sentence.

A: We will merge the sentence with the following sentence:

"Here, we develop two regional climate datasets of 25 and 100 members to create a large ensemble of RCM simulation which are particularly suitable to study extremes. Earlier studies such as Leduc et al. (2019) have successfully demonstrated the usefulness of such an approach."

L55-56: You cannot downscale an RCM using GCM simulations.

A: We will rephrase that sentence:

"To create a data set for regional climate impact studies for Europe under 1.5°C and 2.0°C global warming the regional climate model REMO has been used to dynamically downscale two GCM ensembles following the HAPPI experiment protocol by Mitchell et al. (2017)."

L59: What does "CMIP5 mean SST anomaly" actually mean? All CMIP5 models and members, or just some, only ECHAM6/NorESM? Is there a reference?

A: This point will be addressed in a dedicated section explaining the HAPPI protocol in more detail (see answer to referee#1).

In the HAPPI protocol all CMIP5 models are included in the averaged SST for the current and the projected periods of 1.5°C and 2.0°C respectively. The SST anomalies of the 1.5°C projected period are calculated by subtracting the averaged current SST from the averaged SSTs of the 1.5°C

projection. The SST anomalies are then added to the observed SSTs of 2006-2015. The 2.0°C SST anomaly is computed in a similar manner.

L66/67: “from the core domain defined by CORDEX the entire domain has 121x129 grid boxes” – Please reword.

A: We will rephrase:

“The European CORDEX domain for REMO covers 121x129 grid boxes. To exclude the sponge zone, where the REMO simulations are relaxed towards the GCM solutions, a core domain of 106x103 grid boxes, following the CORDEX definition, is used for the analyses.”

L94: “recommend” –> “recommended”

A: We will changed as recommended.

L97-100: You defined an abbreviation for each climate index. Use them more consequently throughout the text.

A: We will make changes made as recommended.

L99: “precipitation intensity at the 50-yr return period” – Do you mean precipitation intensity with a 50-yr return period? Please use a consistent explanation throughout the text.

A: Yes, we mean precipitation intensity that occurs every 50 years. We will provide an explanation, and use the definition more consistently throughout the text.

L111: What do you mean with “annual sum maximum”?

A: This is a spelling mistake. It should be “annual sum”. We will revise:

“The index for the annual maximum of the five-day precipitation sum (RX5day) is used to characterise heavy precipitation events, which can be relevant for flood generation in river basins.

L121: “between 100 and 100 years” – Please correct.

A: This will be corrected to ‘10 and 100 years’.

L130: What do you mean with “historical”? Pre-industrial or current? Same in L141, L205, L215.

A: This is an inconsistency. It should be “current” everywhere. The text will be changed accordingly.

L132: “Only areas with more than 20 non-zero data points” – Isn’t the calculation done at every single grid point?

A: Yes, the calculation is done on every grid point. The data points refer to number of exceedance in the current period. The formulation will be changed:

“Only grid boxes with more than 20 exceedances over threshold in the current period were included in the analysis of ATG28 in order to allow for confidence interval calculations for the shown percentiles using order statistics.”

L152: “in the median around the Mediterranean” – Please reword.

A: This will be rephrased as follows:

“Around the Mediterranean the increase in ATG28 during the 1.5° C period is mostly moderate with up to 9 days in the median whereas changes in the 2.0°C period are reaching 18 days and more.”

L154: Is a spatial resolution of 0.44 high enough to resolve complex topography? What about EUR-11?

A: We agree that EUR-11 would be better to resolve complex topography, however, with the current generation of HPC computers, such a large ensemble of RCM simulations would not have been possible to conduct on much higher resolution than 0.44°.

L176: "interior of the simulation domain" – Please reword.

A: We will rephrase:

"Apart from artificial effects due to the boundary conditions, the strongest signal within the core domain appears over the Baltic Sea, with an increase of up to 15% in RX5day under a 2.0°C increase in GMT."

L190-191: Please reword sentence.

A: We will rephrase as follows:

"The relative change (in percent) in RI50yr across Europe are presented in Figure 5."

L204: You never defined/explained p-values.

A: Please see our answer to L204 below

L204: "Both the distributions 1.5C and 2.0C" – Please reword.

A: We will rephrase the entire section to read as follows:

"In this section, the changes in the Consecutive Dry Days (CDD) distributions for the 1.5°C and 2.0°C periods compared to the current period are presented. To distinguish whether these changes in the distributions are statically significant we employ the Mann-Whitney U-Test. Where the resulting p-values of the test are greater or equal to a significance level, alpha, of 0.05 or smaller, the null hypothesis is rejected indicating the distributions differ. The p-values for each of the PRUDENCE regions (Christensen et al., 2007), shown in Fig. 1, are presented in Table 1."

L215: "shift towards longer periods of dry days" – This depends on your definition of CDD. If CDD is the maximum number of consecutive dry days, your results only show that the longest dry period is getting longer in a warmer climate. That does not mean that all dry periods will be longer.

A: Our formulation of CDD is the maximum consecutive days in 10 years. In this analysis, we are determining whether the shift in CDD distributions are significant. In Figure 6, the percentage of the entire CDD distribution of a given duration is presented.

L215 reformulation:

"Over this region, one can see an increase in duration of the longest dry period and that they occur more often (Figure 6)."

L216: "the Alps and Eastern European region" –> "the Alps and Eastern Europe"

A: We will rephrase accordingly.

L218: "suffer from more frequent and longer drought periods" – Again, this depends on your definition of CDD. If you used the common one (CDD being the length of the longest dry period), you cannot say anything about the frequency of dry periods.

A: In Figure 6, we have plotted the distribution of CDD and are thereby studying the frequency of dry periods. Over region 2 one can see that the length of the longest dry period increases as the reviewer states and one can also see that the number of longer CDD increases. The authors argue 'more frequent and longer drought periods' in this case is correct.

L222: "adaption" should be "adaptation" Figure 6: Use ECHAM6 instead of ECHAM.

A: We will change accordingly.

L235: "10 x 100 years" –> "100 x 10 years"

A: We will hange accordingly.

L247: "RX5day" –> "(RX5day)"

A: We will change the sentence to:

"The RX5day shows a general increase over Europe which is more pronounced under higher global mean warming."

L251: "such precipitation extremes" –> "such as precipitation extremes"

A: "as" will be included.

L263: "historically similar" – Strange wording.

A: See our answer to L263-264, below.

L263-264: "pre-industrial period" – I thought you were calculating differences between future and current climate and not between future and pre-industrial?

A: There was a mix-up in the sentence. It should always be current vs. future. The text will be changed accordingly, as:

"The changes to CDD distributions show that Spain will experience significantly more drought conditions in the future compared to the current period, even at a 1.5°C increase in GMT. For Italy, drought conditions associated with the 1.5°C simulations show non-significant changes, yet those associated with the 2.0°C simulations are significantly different to the current period, thus showing possible consequences of exceeding the 1.5°C GMT target of the Paris agreement."