Response to Anonymous Referee #2

We would like to thank the referee for the thorough review and constructive comments. We address the comments below. The reviewer's comments are in black, our responses in red.

Overarching Assessment

This paper compares high-resolution GCM simulations that were produced for High- ResMIP with Euro-CORDEX simulations at a similar resolution. The study is limited to a comparison of the daily precipitation distribution in the two ensembles. Overall, it is novel and interesting, but has a tendency towards making oversimplified statements that needs to be corrected.

We have revised the entire manuscript. We have performed more analyses by showing spatial distribution of precipitation seasonal means for the two ensembles, as well as Taylor diagrams, as suggested by Reviewer 1, which strengthen our conclusions. We also now systematically include both EUR-11 and EUR-44 to the comparison with PRIMAVERA, with now an emphasis on EUR-11. While EUR-44 is closer to PRIMAVERA in terms of horizontal grid spacing, EUR-11 is the latest state-of-the-art EURO-CORDEX ensemble, which is more appropriate to compare with the state-of-the-art high-resolution GCM simulations (EUR-44 is a bit older). We have also increased the size of the EUR-11 ensemble, as more simulations have now become available, giving an overview of the performance of the latest EURO-CORDEX ensemble. We have removed what we called the "synthetic datasets", but have included a discussion section on observational uncertainty. Finally, we have moved the figures related to the sensitivity tests to additional material. To represent the ensemble spread, we now use the inter-quartile range (not a bootstrap resampling anymore), as it is more representative of the ensemble size. And we now use a student t-test to build the pies, instead of p-values. More details are given below in response to each comment.

General Comments

1. Without using additional observations, and without the presence of additional high- quality datasets for use, I understand that it is hard to include observational uncertainty. Applying a broad correction to illustrate this uncertainty is a little crude though. I do think it should be included, but lacking a better option, I think you need to better empha- size that this is by no means the ideal way to include an estimate of the observational uncertainty and it is also uncertain. Also, I think calling it a "synthetic dataset" is an un- fortunate choice in terminology, as it is not fake or insincere, and I suggest you rethink that choice. At line 359, for example, you could simply say "They are further away from observations, but closer if our measure of uncertainty due to undercatch is considered."

We would like to thank the reviewer for this suggestion. As also suggested by Reviewer 1, we have now removed these "synthetic datasets" from our analyses. We still keep them in a pie-chart figure in the supplementary material to discuss observational uncertainty (section 4.2).

2. You cite Roberts et al. 2018 (BAMS), but there's a point they make that I think needs to be highlighted in your discussion as well (see their paragraph that spans page 2342- 2343). That is, that high-resolution GCMs are likely to provide improved information at the synoptic scale, and because an RCM's representation of the large-scale can only be as good as that from the GCM, this implies that high-resolution GCMs may provide better boundary

conditions for even higher resolution RCMs. See also Gutowski et al. (https://doi.org/10.1175/BAMS-D-19-0113.1) for a similar, relevant follow on discussion to Roberts et al. This point would be relevant, for instance, around line 479, where you make a statement of expectations that I do not agree with and which needs references or a better backup discussion if you intend include it.

We have decided to keep the sentence 'The performance of PRIMAVERA was not logically expected', but we have reformulated the paragraph and now provide a discussion regarding this statement. We also now include a reference to Gutowski et al, thank you for pointing to it. It is now written in the first paragraph of section 5.2.

3. The authors pose 2 questions around line 150, but only ever address question 1. I suggest either addressing the 2nd question or stating here that this paper will only address question 1.

We have changed this paragraph and now only specifically address question 1.

4. There are a number of statements made in the text that I think are too specific or overly general based on what I see in the presented data. See the "specific comments" section for instances that I particularly think need to be corrected.

This issue was also reported by Reviewer 1. We have revised our entire results and discussion sections.

Specific Comments

Title: I suggest you change the title to state that you are assessing European CORDEX RCMs, as the conclusions could be very different over different regions.

Done.

Abstract: The abstract should state that the PRIMAVERA simulations are a part of HighResMip.

Done.

Line 25: For the abstract, I suggest removing the reference to bins and making the sentence "We perform this exercise for the distribution of daily precipitation over Europe. . ."

We have replaced that sentence by 'The focus of the evaluation is on the distribution of daily precipitation at 50km scale under current climate conditions'.

Line 34-35: Instead of "PRIMAVERA appear to be closer to observations. However, when we apply an averaged precipitation undercatch error of 20%, CORDEX become closer to these synthetic datasets.", you might just say that PRIMAVERA may be closer to the observations in this regard, but that uncertainty exists in the observations due to a potential undercatch error, especially in heavy precipitation. I suggest this, because the reader at this point does not understand why you have chosen the 20% value, and because I disagree with the terminology choice for "synthetic datasets".

Rephrased as suggested.

Line 54: The word tuned is used differently by different people and parts of the com- munity; therefore, I suggest you be more specific. I, for one, associate "tuning" with the modification of specific parameter numbers (e.g., changing grassland albedo for your region to something that is known to be more appropriate for your region). It might be more all-encompassing to say that parameterization schemes can be chosen based on their appropriateness for the region and tuning can be completed to better match regional observations. Or, more generally, that an RCM's configuration can be cus- tomized to focus on and best simulate the most relevant climate processes of a given region.

The sentence was rephrased as suggested.

Line 59: Remove "so-called". It has two meanings, the latter of which is not appropriate here and could be misconstrued. 1: a common name for something. 2: a word that is used to describe something that is not suitable or not correct

Agreed. We just referred to the common name for these models, so we removed it.

Line 110: While I partly agree with this statement, it isn't universally correct. I would say that it has "not always facilitated the communication".

We have removed that sentence.

Line 111: 2-step nesting isn't relevant to this sentence (and not always needed), and convection-resolving simulations are not new, although they have only more recently become more mainstream in climate. I suggest revising this sentence.

We agree and have now removed this paragraph. We now mention the convection-resolving simulations in the discussion section (5.2) as the new target for future RCM simulations.

Line 118: this implies that RCMs have not also become more complex; however, they too have been moving towards increasing complexity (e.g. Turuncogle and Sannino 2017, https://doi.org/10.1007/s00382-016-3241-1)

Agreed. The sentence does not mention RCMs anymore and we now also mention regional ESM earlier on together with the suggested reference, thank you.

Line 167: I do not understand this statement, please rephrase: "and concludes with an opening"

This sentence has been removed.

Fig. 1: In the caption you reference parts a and b, but there are no a and b labels in the figure. Please modify one or the other for consistency.

Thank you for spotting this. We have changed the caption accordingly.

Line 250: it is stated that 1000 bootstrap samples are used. Did the authors test this number and its effect on the distributions generated? It is not a very large number of samples,

especially considering the input data, and the statement at line 283 suggests to me that it is in fact much too low a number of samples to reasonably sample all combinations. Please justify the use of this particular number of samples or use many more.

We tried two numbers of samples: 10^3 and 10^6. Using 10^6 did not particularly change the results, and it did not change our main conclusions. However, we understand the concern and now that we include more EUR-11 simulations, which include ~40 members, the number of combinations becomes very large, and we agree that the bootstrapping method may reach its limits. We have therefore replaced the bootstrapping test by a Welch's t-test performed for each bin. This method shows to be stricter than our initial bootstrapping method, and we are confident that it is more robust. We have altered the method section accordingly.

Line 255: How were these bins chosen? Are they representative of percentiles or thresholds that impacts users are concerned about? They seem random to me, and 60mm/day seems quite high for the high category for some regions.

We agree that the chosen values were not appropriate for each season and region. We have now changed the intervals to relative intervals defined for each region and season using the observational datasets (see Fig. 1). The low intensity interval now represents the lowest 40% of observational mean precipitation. The moderate intensity interval represents the middle bins cumulating the next 50% of mean precipitation. The high intensity interval represents the highest 10% of mean precipitation. We added vertical bars in the graphs to show the chosen intervals for each season and region.
This new representation changes our results (see new Fig. 8), mainly in the highest intervals, which now include more bins than with the fixed threshold (especially in winter). Our main conclusions remain similar. We re-wrote our results section with these new graphs which we believe are more robust.

Line 270: you should state here if they are in the supplement or not shown.

We have removed this sentence as we discuss all seasons with the pie-charts.

Fig 4, and other similar figures. The legend placement is sloppy in some of these images. It should be placed where it doesn't cover data.

Thank you, this has been fixed.

Line 330: I disagree with the statement that "there is no systematic difference". I see that there is a distinct systematic difference in most regions regardless of season, and so do you, clearly, because you discuss it in the next paragraph. P clearly has more light rain, C has more heavy rain, and the area under the curve is greater in C than P.

Please be more accurate.

We removed this sentence and rephrased the whole paragraph.

Line 336: the introduction to this sentence is confusing. Do you mean to state that "PRIMAVERA still overestimates low intensity precipitation in all seasons and regions, like CMIP5, although to a lesser extent."? And if so, where's the figure that shows that? I can't tell that it is "to a lesser extent" from the current content.

We have removed this sentence. We have also now replaced the CMIP5 figures by Fig. 2 and 3 that show CMIP5, EUR-11 and PRIMAVERA.

Line 340: At this point in the text, the statement here is not proven. You should add that this will be discussed later.

Agreed. We removed this sentence and have considerably changed the text.

Line 345: What is the sensitivity to the results in this paragraph to undercatch error?

See reply to comment L381 below.

Fig. 7. The C and P indicators on this figure are unfortunately too small for the resolution/quality of the PDF, and they are very hard to read, even when zoomed in, because they are so fuzzy. This needs to be fixed.

We fixed this by having only 4 panels instead of 9 in this figure. It is now Fig. S8.

Line 381: A pie chart including the observational uncertainty would be an interesting addition to the supplementary material, and address my question at line 345.

Thank you for this suggestion. We have included this new figure (Fig. S7). It shows that CORDEX intense precipitation is closer to the observations scaled by a factor 1.2. This is now discussed in a section on observational uncertainty (4.3).

Line 415: Given how similar PRIMAVERA and CORDEX-44 are, this statement is diffi- cult to verify by just eye-balling the difference between this figure and previous figures. CORDEX-44 and CORDEX-11 are not identical and some of the small differences may matter. Could you overlay all 3 datasets?

We have now redone our analyses with both CORDEX-44 and CORDEX-11, with a focus on CORDEX-11, which is a newer ensemble. We have created new figures showing seasonal mean bias (Fig. 4) and Taylor diagrams for all ensembles (Fig. 5). We also now include a pie chart for both CORDEX-44 and CORDEX-11 (Fig. 8), so the differences between CORDEX-11, CORDEX-44 and PRIMAVERA are clearly identified.

Line 445: Not all convective precipitation schemes (CPS) take into account convective inhibition in their triggering function. This makes a difference in the drizzle problem. In my experience, RCMs are more likely to use CPS that include this (in mid-latitudes at least), and GCMs are less likely to. This isn't tuning, it's a configuration choice.

We reworded the "tuning" expression, but we did not speculate further on this issue.

Line 483: "over CMIP5-driven CORDEX simulations for precipitation over Europe". . . this is a bit general. It would be more accurate to say "over CMIP5-driven CORDEX simulations in some regions and seasons by our metrics for precipitation over Europe"

Done.

Line 486: Please re-write this sentence, I do not understand at all what it is trying to say. And, if I guess at what it is trying to say, I do not agree with the statement.

The sentence has been reworded as 'It indicates that the potential improvement of large-scale dynamics in high-resolution GCMs has a positive influence on precipitation distribution.'

Technical Corrections:

I do not think it should be the responsibility of the reviewer to copy-edit the manuscript. This paper is readable and well-organized, but could use some English language copy- editing. I have pointed out important instances of text in the previous section that I think need to be corrected for clarity purposes though, as these effect the understanding of the science presented. Please pay particular attention those.

We have now carefully revised the manuscript.