Response to the comments from Anonymous Referee #2 for the manuscript:
"Uncertainties in climate change projections covered by the ISIMIP and CORDEX model subsets from CMIP5" by Ito et al.

We would like to appreciate your careful review and constructive comments and suggestions for improving our manuscript. We almost agree with them. We have made modifications through our manuscript according to the responses. Please check our detailed responses below. The numbers of page and line are corresponding to the number in the original file (https://www.geosci-model-dev-discuss.net/gmd-2019-143/gmd-2019-143.pdf).

In this modification, we added a CMIP5 model of CSIRO-Mk3L-1-2, to the original 49 models for the historical run we analyzed. It is because there is no member of r1i1p1 by CSIRO-Mk3L-1-2, but there is r1i2p1 as well as CESM1-WACCM which was already used. The results did not change from the original manuscript by this modification. I apologize for the change.

In this revision, McSweeney and Jones (2016) have referred to as MJ2016 except for the first reference.

Thank you once again for your review.
We would be glad to respond to any further comments you may have.

--- General comments

This manuscript aims to quantify the spread of CMIP5 projections and biases covered by the subsets of models used in the ISIMIP and CORDEX experiments. The first section of the results examines the spread of model performance in reproducing the temperature and precipitation over the historical period (1986-2005), relative to a range of observational and reanalysis products. The rest of the results examines the spread in projected end-of-21st-century changes in annual mean temperature and precipitation, and how it compares to the spread covered by randomly selected subsets. The main findings are that (i) the small ensembles used in ISIMIP and CORDEX generally perform well over the historical period but are not optimal in minimizing historical biases, and (ii) the ISIMIP ensemble outperforms the CORDEX and randomly selected ensembles in covering the full CMIP5 range of projected temperature changes, but both ISIMIP and CORDEX cover a smaller spread of precipitation changes than randomly selected subsets.

This manuscript presents a valuable study to put the CORDEX and ISIMIP subsets in the context of the full CMIP5 ensemble. At this stage, it is mostly descriptive and would greatly benefit from a more comprehensive discussion, including the benefits/limitations of the metrics used, and the implications of its findings. Please clarify how this specific study sets itself apart from existing studies such as McSweeney & Jones (2016), and how your results fit into the context of the existing literature. Minor adjustments to language and sentence structure are needed to improve the readability of the manuscript.

We appreciate your useful comments to improve our manuscript. We have made our response to the general comments as the response to the following specific comments:
- the benefits/limitations of the metrics used and the implications of its findings (#11, #15)
- how this specific study sets itself apart from existing studies such as McSweeney & Jones (2016) and how your results fit into the context of the existing literature (#2, #14).

Also, we have added a section to discuss the results and provide our considerations.
Please find below.

--- Specific comments

☐ Section 1 Introduction

1. P2 L. 19: Please specify what these previous studies have found, and why we have yet to reach a consensus on the method to select small ensembles. Also, revisiting these papers in a discussion section would provide the necessary context to interpret the results and whether the methodology used in this study is distinct from, or improves upon, the ones used in these studies.

In the previous study, the condition for selecting subsets depends on their purpose. For example, whether the model performance is considered, which climatological or extreme variables are used and which region is interested. Thus, we have yet to reach a consensus. Our purpose in this study, however, is to

indicate the property of ISIMIP and CORDEX subsets for the ability to reproduce the present-day temperature and precipitation and for their future change, and is not suggestions of model selection methodology (P3 L23-25). In P2 L19-21, we have described that, although there are various methods, it is most desirable for the methods to select subsets of GCMs that have smaller biases in the historical climate simulations and cover the widest possible uncertainty range of future projections. We have discussed whether the current subsets in ISIMIP and CORDEX are such a subset.
We have just modified the related sentence as the response to "why we have yet to reach a consensus on the method to select small ensembles";

(P2, L19) "... Gobiet 2016). The optimum method, however, remains to be determined because the interests depend on the studies, for instance, how the model performance is considered, which climatological or extreme variables are used and which region is interested."

2. P3 L. 23-36 Please clarify what is specific to this study: is some aspect of the methodology new? Is this performing an existing analysis to a new set of data? Is the added value of the manuscript to specifically address whether region-specific subsets (CORDEX) outperform a globally consistent sub-set (ISIMIP)? Stating this explicitly would improve the value and readability of the manuscript.

The methodology is not new. The analysed subset has been changed from the subset analysed in McSweeney and Jones (2016) by following the updated selection in ISIMIP. The added value of our manuscript is following, which have been added to P4 L2:

(P4 L2) "The ability for the ISIMIP subset was not mentioned by MJ2016 and thus we investigated that in region by region. We analysed four GCMs selected in ISIMIP2b (unless specified otherwise, hereafter refers to as ISIMIP) here. Thus, assessment of the projections was also updated from MJ2016. The GCMs used in CORDEX have been assessed by region in previous studies, but are limited (e.g., Haensler et al. 2013 for Africa; Bartók et al. 2017 for Europe; Karmalkar 2018 for North America). Even simple assessment conducted is needed for the present CORDEX. Furthermore, uniform assessment across regions permits to discuss the difference of characteristics among the regions and the possibility of heterogeneous scenario as mentioned above. By using the subsets from the two programs, we can explore the difference between the original subset in CORDEX and the subset selected with assuming CORDEX CORE, which is helpful information for the model selection in CORDEX CORE."

From an additional analysis in this revision, we suggest that nine models are needed to solve the heterogeneity of the uncertainty. This result can provide suggestions to the next generations of model selections.

(P9 L18) "The current CORDEX subsets can capture both uncertainties for temperature and precipitation in the regions with a relatively large ensemble. However, it is found that changing the number of models from the current CORDEX members to nine members can capture more than half of the full uncertainty in both projections of temperature and precipitation in more than 85% of all regions, with a possibility of 50%. Furthermore, the same is also shown as for the ISIMIP subset, but for 70% of all regions. Focusing the uncertainty in the future projections, this result proposes that the current number of models need to be changed to discuss a similar uncertainty range among the regions."

☐ Section 2.1

3. P4 L. 18 Why focus on land area only? Regional precipitation, including over the seas/ocean, is relevant for impact studies. Please clearly state the scope (and the application) of this study.

The impacts of climate change appear over the land and ocean as you mentioned. The reason why focusing on land is that the assessment sectors in ISIMIP are mainly over land, and it is important for both programs because of the relevance to human activities. We have added the sentence below:
(P4, L18) "… we focused on the global land area, considering the importance for both programs because of the relevance to human activities."

4. P4 L. 22 Can you justify why you excluded low-precipitation models from the precipitation analysis? I understand these models significantly bias your ensemble but this undermines the stated

aim of the study (i.e. to quantify the spread of the full CMIP5 ensemble covered by the ISIMIP and CORDEX subsets). You arbitrarily reduced the model spread covered in this study. Please at least provide more information as to how many models were excluded (and thus the size of your remaining ensemble), why 0.1mm/day was chosen as a threshold, and the reasoning to exclude these models from the precipitation study but to keep them for temperature (if the argument is that the climate they produce is too unrealistic to be a plausible representation of today's climate).

We have expressed the future change of precipitation as a change ratio of the future precipitation to the present-day precipitation. The expression, which has been often used, is highly sensitive in dry grids. Even if the change amount is quantitatively small, the ratio is extremely large. Such a large ratio leads to a large regional average. The large ratio by a small change in dry grids is difficult to explain the validity, and thus we took the dry grids out of consideration. The threshold can be permitted the exclusion of grids with the ratio of 100% over around the Sahara, and be suppressed the exclusion under 5% of all analyzed grids. We added the following sentence to explain the exclusion:

(P4, L22) "The future change of precipitation expressed in a ratio here. That is the change ratio tend to be large at too dry grid even when the change is quantitatively extreme small. Such a large ratio is difficult to explain its meanings physically. By applying the threshold, the grid indicating an extremely large ratio, for instance, 100% were excluded. The total number of the excluded grids is approximately 5% of all target grids as an average over the used members."

5. P4 L. 32 Please include a definition of the skill score used here, or a reference to a published paper using the exact same skill score. In Taylor (2001), two examples of skill scores are used, to illustrate that the skill score can be adjusted depending on whether you value high correlation or matching the variability most. In addition, it is explicitly stated that the value of Ro should be reported every time a skill score is used.

We appreciate your pointing out. The definition below has been added to P4, L32,

(P4, L32) "… we used the skill score proposed by Taylor (2001) (hereafter referred to as skill score) as follows:

$S=4(1+R)/\{(\sigma+\sigma^{-1})^2(1+R_0)\}$, (1)

where R is the spatial correlation coefficient between referred observation and simulation, σ is the standard deviation of simulation normalized by the reference spatial pattern and R_0 is the maximum correlation attainable. The value of R_0 was assumed to 1 here."

6. P5 L5 Is R the min-max range of a given subset? Please clarify. Using 'uncertainty range' is misleading; it sounds like you are sampling your ensemble. If I understand correctly, you generate 10,000 values of RSub, then look at ensemble spread (in Fig 4).

The value, Rsub which we used here, is the max.-min. ranges of the uncertainty estimated from the ISIMIP subset, the CORDEX subset, or the 10,000 random subset samples from $CMIP_{Full\_Future}$. The corresponding parts have been modified as follows:

(P5 L4) "The FRC from the regional averages (FRA) was defined as the fraction of the maximum-minimum range of the uncertainty in the regional averaged projections from a subset of $CMIP_{Full\_Future}$ (R_{Sub}) to the range from $CMIP_{Full\_Future}$ (R_{Full}), as follows:
(Equation 2)
The range of R_{Sub} was computed from the ISIMIP and CORDEX subsets and also arbitrary subset samples we generated. From the comparison with the arbitrary samples, we can investigate how well the ISIMIP and CORDEX subsets captured the uncertainty range of projections. McSweeney and Jones (2016) presented the comparison using their 500 samples as 'representation'. Our arbitrary samples were generated by randomly selected n models without repetition from $CMIP_{Full\_Future}$ 10,000 times, where n is the sample size of subsets in ISIMIP (n = 4) or CORDEX (n depends on the regions; see Table 1). Then, the variance of the FRA was estimated from the 10,000 random samples of the subset of $CMIP_{Full\_Future}$ and compared with the FRA from the ISIMIP and CORDEX subsets."

☐ Section 3

7. This section is entitled ‘Results and Discussion’ but mostly contains the description of the results. Regardless of whether it is included in this section or a separate section, the manuscript needs a more comprehensive discussion (see other comments below).

We have made an additional section in this revision for the discussions. Please check the responses below.

☐ Section 3.1

8. P.5, L. 28-29 Please include the top 50% ensembles in the supplementary material, so that future model selection can rely on your analysis to select less biased models.

We appreciate your constructive suggestion. We have added the high performance subsets as an supplementary material. The material has been refereed in P5, L19 (“The models included in the high performance subset is shown in Supplement 3.”).

10. P5 L29-30 Can you suggest why the spread is different between the two ensembles over the Northern Hemisphere? In Fig.1, the spread for ISIMIP is sometimes significantly larger, sometimes significantly smaller than CORDEX. Could this be due to the number of models in CORDEX? Do some of the regions have models that overlap across ISIMIP and CORDEX? Simply stating that they are different in some regions is not very informative.

We appreciate your accurate indication. As you pointed out, we have found that part of the characteristics of the difference in the spread has a relationship to the overlapping of used model members. The sentence has been modified and added more explanation:

(P5 L29-30) “The difference in the spread between the ISIMIP and CORDEX subsets has a characteristic in region-by-region and part of them relates to the overlapping of model members used across ISIMIP and CORDEX. For example, in five regions of Central and South America, Europe, Africa and South Asia, the CORDEX subsets include more than three of four ISIMIP models and the ensemble is large in CORDEX than in ISIMIP (Supplement 1). As the result, the variance of biases estimated from the CORDEX subset covers that from ISIMIP. Especially in Europe, the difference of the variance between the CORDEX and ISIMIP subsets is large and the models not included in the ISIMIP subset are found to make the variance increase. Focusing on the regions where the CORDEX subsets include only two models in the ISIMIP subset, the variance from the CORDEX subset tends to be larger than that from the ISIMIP subset, especially in the regions with large ensemble in the CORDEX subsets, like North America, SEA and Australasia. By contrast, the variance from the CORDEX subsets is relatively small in the regions with small ensemble in the CORDEX subsets, like MENA and Central Asia. In East Asia, the variance is small in CORDEX despite using seven models in contrast to four models in ISIMIP. Thus the biases are found to be similar to each other in CORDEX-East Asia.”

Also, we have modified the sentence about the spread of the temperature bias:

(P6 L9-11) “The spread of B(T(ISIMIP)) is covered by that of B(T(CORDEX)) in the same four regions as the bias in the precipitation except for Europe, because of the overlapping of used model members. The spreads of B(T(ISIMIP)) and B(T(CORDEX)), however, resemble each other compared with the precipitation bias, indicating that CORDEX used models with a quantitatively similar performance to ISIMIP, despite using more models than ISIMIP except for Central Asia.”

11. P6 L1-3 This paragraph would benefit from an earlier explanation about how the skill and bias metrics are different and the insight gained by using both. Include some interpretation of why the model ensembles that perform relatively well in a bias metric perform less well in the skill metric. (same comment for P6 L11)

Thank you for your suggestion. The skill score quantifies the similarity of the spatial pattern by a correlation coefficient and a standard deviation. The bias evaluates the quantity itself by the regional average of the difference from the observation. Thus there is a case with large positive and negative biases in each grid even when the spatial average is small, that is to say, the spatial pattern is different from the observation. The ensemble showing a small bias and a low score represents the quantity closed to the observation as the spatial average but a low similarity of the pattern. Therefore both metrics are

needed to assess how well the ensemble represents the reality. We have added the following sentences in each part:

(P4, L32) "In addition to the skill score, we use the model bias to evaluate the quantity itself. The usage of the two metrics enables the assessment of both the spatial pattern and the quantity."
(P6, L2) "That is to say, ISIMIP and CORDEX subsets include the member showing a low similarity of the spatial pattern to the observation."
(P6, L11) "Therefore, relative to $CMIP_{highS}$, the subsets can quantitatively represent the observed temperature as a regional average well but the spatial pattern represented by some members in the subsets has not much resembled the observation."

12. P6, L. 15-16 Please include the top 50% ensembles in the supplementary material. In addition, please include in the discussion whether the 'best performing models' perform well both in temperature and precipitation, and whether selecting according to high skill or low bias makes a difference. As you state that a better ensemble can be selected, please give the evidence from your results that this can be done robustly.

We appreciate your constructive suggestion. We have added the top 50% models as Supplement 3 (Response #8). The comparison between the top 50% ensembles for the bias and skill score is interesting. From Supplement 3, when we focus on one variable of either temperature or precipitation, 13 models in 25 all high-performance models are included in both subsets of high-performance models for the bias and skill score. Thus, the model with a small bias indicates a high score with 50% of the possibility. We have described this explanation to Section of discussion which we have added in this revision:

(P8 L23) "Focusing on one variable of either temperature or precipitation, 13 models in 25 all high-performance models are included in both subsets of high-performance models for the bias and skill score (Supplement 3). In addition to the two indices of bias and skill score for one variable, the models indicating the high performance for both two variables of temperature and precipitation is 0 at the minimum number in Southeast Asia and the Arctic and 9 at the maximum number in Africa. The averaged number over the regions is approximately 4. Therefore, although the model with a small bias indicates a high score with 50% of the possibility, it is difficult to select models with a high performance for both variables of temperature and precipitation."

In addition, explanation and discussion were not enough for the description of selecting a better ensemble. We have added the limitation.

(P8 L23) " … a much better model subset, regarding to biases and skill scores, can be selected with making use of the advantage of the small number of models. However, such a selection can be conducted when there are no constraints of data availability which was the main constraint to select the current subsets in ISIMIP and CORDEX and when we use one variable of either temperature or precipitation."

☐ Section 3.2

13. P6, L.25 Please place this into context by mentioning other studies that have looked at emergent constraints, even if it's only in specific regions (e.g. Bracegirdle et al, 2018 for Southern Ocean winds; Bracegirdle and Stephenson 2013 for Arctic warming).

As you noted, we have added related previous studies and modified the sentence,

(P6 L25) "… suggesting that the bias and skill score are not good emergent constraints to reduce the uncertainty of ΔT in this study though the previous studies have showed the reduction of the uncertainty (e.g. Smith and Chandler 2010; Bracegirdle and Stephenson 2013; Bracegirdle et al., 2013; Simpson et al. 2016)"

☐ Section 3.3

14. In general, this section is confusing. It would benefit from clearly stating what is being compared, and referring to specific aspects of Fig 4 to support your statements. Specifically: P8 L1-2: Please clarify which metric you use to make that statement (i.e. the total coverage on the y-axis). I got

confused because the performance of FRACORDEX remains low compared to FRARandom_C, even as the number of models increases. Please state explicitly where you are comparing it to the full range, or to Random_C (3 sentences later). P8 L16: Please specify which FRA you are talking about: the median of FRARandom_C? Or FRACORDEX? Or both? P8 L19-22: This is an interesting point, but if you make the point that increasing the number of models produces a higher FRA, please show the evidence for it. The latter part of the description is unclear so adding the technical details and a figure would make a stronger point.

We have modified each sentence you pointed out as what follows:

(P8 L1-2) "A relatively high coverage, above ~50%, is shown on FRA_{CORDEX} for both changes of temperature and precipitation in eight regions when using nine models or more, except for temperature in Antarctica (Fig. 4a, b): that is to say, the CORDEX subset captures more than half of the range from $CMIP_{Full\_Future}$."
(P8 L16) "…and thus the large model ensemble results in an increase in FRA_{CORDEX} and FRA_{Random_C}."
(P8 L19-22) We have added the figure for the change of FRA with the number of models not only in Central Asia but also in the other regions. Please check Figures 5 and 6.

☐ Summary and Conclusions

15. This section provides a general summary of the findings, but would benefit from providing context as to how these results compared to other studies (e.g. those cited in the introduction), and how the findings advance the general understanding of the field. In addition, statements in P9 L4-5 and P9 L 15-16 seems to indicate CORDEX performance to be bad relative to randomly selected ensembles, while P9 L8-9 states 'relatively wide coverage of both uncertainties'. Please clarify so that the message is not ambiguous. For example, it is ok to state that CORDEX is not performing well compared to the randomly selected ensembles, but is marginally better than ISIMIP at sampling uncertainties in projected change in precipitation.

Regarding the comparison with other studies, as mentioned in P7 L9, "... global consistent four models used in ISIMIP2b, which are taken into consideration of the ability of reproduction, still remains difficult to capture the uncertainties in regional precipitation change, as in McSweeney and Jones (2016) which analysed for five models in the fast track." The result of assessing the CORDEX subset was not able to compare with other studies because of the difference in the variables, part of the regions and seasons. For results from an additional analysis conducted in this revision, we referred to the approach in McSweeney and Jones (2016) but the results couldn't compare each other. It is because "They focused on a subset covering the uncertainty in each grid most widely over the globe or regions and investigated how the coverage changes with the number of models. On the other hand, in this study, to consider making better use of the current subsets, we investigated how the coverage changes with changing the number of models from the current model members." (Added to Section 3.3)

Thanks for your suggestion on the ambiguous statement. We have added the following statement to P9 L14 and have modified the statement on P9 L15-16.
(P9 L14) "The CORDEX subset is not performing well compared to the randomly selected samples but is marginally better than ISIMIP at covering uncertainties in the projected change in precipitation when a large model ensemble used."
(P9 L15-16) "The region-specific model subset, like CORDEX, captures coverage of both uncertainties well compared to the global common subset, but large model members are needed."

16. Please include a more comprehensive discussion of your methods and results, including:
Two metrics for "good performance" are used in parallel throughout the study (low bias and high skill score). Please comment as to how similar/distinct these two metrics are, and on the insights gained by using both (qualitatively or quantitatively). Similarly, how different are the 'top 50%' ensembles? i.e. does using skill or bias for selection of the best performing models significantly affect the ensemble?

We appreciate your comments.
First, how similar/distinct these two metrics are, and on the insights gained by using both?

As described in Response #11, we can evaluate the abilities to represent the spatial pattern and the quantity itself by using the two metrics. How similar these two metrics are can be estimated by how many the models selected by each metric overlap. Because 13 models in 25 all high-performance models are included in both subsets of high-performance models for the bias and skill score, the similarity is not so high, around 50%. The number of overlapped models is described in Section of discussion added in this revision. (P8 L23)

Second, how different are the 'top 50%' ensembles?
How different are the top 50% ensembles have been shown in Response #11. Please check. The model with a small bias indicates a high score with 50% of the possibility. Thus a significant influence appears on the selected ensembles.

17. The results section 3.1 focuses mostly on whether the ISIMIP and CORDEX fall within the observational spread (e.g. L. 2-3 on page 6). It would be helpful to distinguish whether this is mainly due to a large spread in the model ensembles, or whether a systematic bias is seen in certain regions (e.g. Fig 1 shows model ensembles overestimate precipitation in most regions). Also, please include a discussion of the expected variance of model ensembles. In coupled models, the timing of climate variability modes is unlikely to match that of observations, so the variance over a 20-year period is likely to be higher in model ensembles than observations.

Here, the observational spread is the spread of the 20-year averaged precipitation calculated from seven observational datasets, not the variance over a specific period. Therefore we cannot discuss the different variance between the model and observations, resulted from the timing of climate variability.

18. In this study, the performance of CORDEX and ISIMIP are considered independently for the temperature and precipitation changes (with precipitation being scaled by the temperature change). Please discuss whether there is any evidence that a selection on one variable (e.g. precipitation) is sufficient to select good performing models, or whether a combined approach is necessary to select models. In climate impacts, people care about the plausibility and diversity of climate sampled, not a single variable.

We appreciate your important suggestion. In this revision, we confirmed a quite small number of models indicating a high performance for both principal variables of temperature and precipitation. In addition, we considered that the evaluation for the simulated principal variables is needed for the studies of ISIMIP and CORDEX, but not possibly sufficient for model selections. Because the large-scale circulation characterized the regional climate, its performance is also important. When we can obtain the reference data, the method used in this study can be applied to the evaluation of the performance. To select subsets in the next generations with the performance considered, it is necessary to construct a combined approach that can take into account multiple variables. We have described this explanation to Section of discussion which we have added:

(P8 L23) "…Therefore, although the model with a small bias indicates a high score with 50% of the possibility, it is difficult to select models with a high performance at the quantity and the spatial pattern for both variables of temperature and precipitation.
In this study, we assessed the current ISIMIP and CORDEX subsets to investigate whether the subset indicates small biases in the historical climatology and covers the uncertainty in the future projections widely using temperature and precipitation. Both variables are most frequently used in future projections and also weather forecasts. The evaluation for such a principal variable is important for the studies of ISIMIP and CORDEX. It should be noted, however, that ISIMIP needs the dataset with reasonable for multiple variables used in their impact assessment and with enable to discuss the uncertainty in the projections. CORDEX requires the dataset with based on a plausible mechanism of the climatology as the input data for RCMs. Thus, there is a possibility that a good subset which we presented based on the model performance for temperature and precipitation would be an option of their future subsets.
Although ISIMIP and CORDEX have tight constraints for model selection at the present, both programs will select the subset showing a reasonable climate based on a plausible mechanism in the future. In the case, two variables of temperature and precipitation are not possibly sufficient for model selections. At least for the regional climatological studies and the assessment of its impact, it is important to reproduce

large-scale circulations which characterize the regional climate. Especially, the spatial pattern of precipitation depends on the accuracy of the circulation. Indeed, model change in ISIMIP from the fast track to ISIMIP2b has already been performed with a consideration of the ability to reproduce ENSO and monsoon (Frieler et al. 2017). The evaluation method used in this study can be applied to the other variables when we can obtain the reference data. For instance, Taylor's skill score which we used to evaluate the pattern of temperature and precipitation can also apply to the pattern of circulation. However, as more variables and evaluation indices are employed, it is more difficult to obtain the CMIP5 models with high accuracy as described above.
It is preferable to select subsets in the next generations based on a combined approach that can consider not only the ability to reproduce the principal variables of temperature and precipitation but also the other ones which are also important to characterize the regional climate. Construction of such an approach would be one of the important tasks for both programs."

In addition, we described the following sentence to Summary:

(P9 L18) "In this study, we have assessed the subsets using the principal variables of temperature and precipitation. It is not sufficient for selecting subsets in the next generations. We suggest that it is preferable a combined approach that can consider the ability not only for temperature and precipitation but also for the other ones which are also important to characterize the regional climate. Construction of such an approach would be urgently demanded for both programs."

--- Technical corrections
P1 L18 (and after) High performed models -> high performance models or high-fidelity models
We have modified them as the referee mentioned. Thanks.

P1 L20-25 Please rework this section to clarify the meaning. As you have not previously introduced the 10,000 sampling strategy, these two sentences are confusing.
I am sorry for the confusing. The section has been rephrased,
"Compared with the randomly selected 10,000 arbitrary subset samples, the CORDEX subset shows low coverage of the uncertainty for the temperature change projections in some regions, and the ISIMIP subset high coverage in all regions. On the other hand, for the precipitation change projections, the CORDEX subsets show low coverage in half of the regions compared with the arbitrary subsets, but tend to cover the uncertainty widely compared with the ISIMIP subset."

P2 L33 Please rephrase this sentence for better readability. For example: "In addition, paper X and Y showed that combining region-specific subsets covers more uncertainty than a single, globally consistent, subset of models."
The sentence has been rephrased as follows:
"They also illuminated that region-specific subsets generally cover more the uncertainty than globally consistent subsets in 26 global regions."

P5, L11 Please rephrase that last sentence for readability.
The sentence has been rephrased as follows:
"Then, the variance of the FRA was estimated from the 10,000 random subset samples of $CMIP_{Full\_Future}$ and compared with the FRA from the ISIMIP and CORDEX subsets."

P7, L9-12 This sentence needs to be reworked for readability.
The sentence has been rephrased as follows:
"Therefore, the subset of four models used in ISIMIP2b shows the difficulty of capturing the uncertainties in regional precipitation change. This result is the same as stated using the subset of five models used in the fast track of ISIMIP discussed by MJ2016, despite two of the five models changed."

P7, L15 "with those of the 10,000" -> remove "the"
We have modified them as the referee mentioned. Thanks.

P7 L16 “randomly sampled subsets” of what?
We have rephrased to “randomly samples subsets of $CMIP_{Full\_Future}$”.

P9 L17 Be more specific: ‘it depends on the number of models used’ is too vague to be informative -> “FRA increases with the number of models used”, or “regions covered by bigger ensembles generally have higher FRA”. . .
We agree with the comments. The sentence has been modified to “large model menber are needed”.

P13 L8 “areal mean of the reference data” -> normalized by the regional average of GPCC data.
We appreciate your revised. We have modified the sentence as you mentioned.

P14 Figure 2 Why does Antarctica have no top 50% in temperature? Explain that somewhere (main text or figure caption).
We apologize for missing the explanation. The reference data of temperature does not cover the Antarctica, and thus we cannot indicate the results for the top 50%. We have added the sentence below in the caption of Figure 2 and also Supplement 4.
“The top 50% of the CMIP5 models cannot be plotted over Antarctica because of missing the CRU reference data.”

P16 Figure 4 “uncertainty range” -> range Also, why are red dots missing in some regions in Fig 4a?
We have changed to “range”. Red dots look missing because the dots overlap where the coverage is the same between ISIMIP (blue dot) and CORDEX (red dot). “The ISIMIP and CORDEX coverages in (a) overlaps in MENA, N. America and Africa.” is added to the caption.

We have revised our manuscript to address comments from Anonymous Reviewer #2.