Author Response to Referee 1

Dear Referee 1,

Thank you very much for taking the time to review this manuscript and for your valuable corrections and suggestions.

Your comment: *The paper provides an interesting and highly relevant analysis of CMIP5 and CMIP6 models with respect to the representation of circulation in the northern hemisphere. It also shows the general improvement from CMIP5 to CMIP6 in this aspect. The analysis criteria are especially interesting for e.g. the regional climate modelling community by having an additional evaluation criteria to the commonly used temperature and precipitation analysis.*

I recommend to accept the manuscript after taking some minor points into account.

Response: Many thanks for your interest in the study and for your positive feedback. For the revised manuscript, 10 additional GCMs and 2 additional members of CNRM-CM6-1 have been added to the evaluation, although this was not requested by any of the referees, making it even more exhaustive. Also, with the help of a small survey sent out to all modelling teams, the documentation about the components of the participating GCMs has been confirmed and further extended. Please find below a point-to-point list to your valuable comments and suggestions.

Your comment: *Abstract, line 2: In many applications relevant for decision making, and particularly when deriving future projections with the delta-change method, they are assumed to be perfect. --> Isn't the delta-change method rather assuming that the model biases are constant than assuming that models are perfect?*

Response: I have been thinking quite a bit about this sentence as well. What I mean here is that stakeholders not familiar with climate science, and most importantly politicians, run the risk of using delta change estimates (or multi-model mean values thereof) as if they were deterministic predictions actually to occur in the future, and would then base their decision making and ultimately legislation on this premise. GCM errors and the stationarity assumption you mention are technical issues stakeholders are normally not aware of. A solution on how such technical questions should influence practical decision making is difficult. However, there is no need for lengthy discussions in the abstract and, following your advice, I have downweighted and simplified this sentence to: “In most applications relevant for decision making, they are assumed to provide a plausible range of possible future climate states.” (see lines 2-3 in the revised manuscript)

Your comment: *Line 8: Both approaches, however, are in principle unable to correct errors resulting from a wrong representation of the large-scale circulation in the global model. --> Dynamical downscaling, at least to some extent within their regional domain, can correct errors in the large-scale circulation.*

Response: Following your advice, this passage reads as follows in the revised manuscript (see lines 8-9): “For both approaches, however, it is difficult to correct errors resulting from a wrong representation of the large-scale circulation in the global model.”

Your comment: *Line 14: The latest model generation --> add (CMIP6).*

Response: “(CMIP6)” has been added here

Your comment: *Introduction, line 50: they do not correct errors inherited from a wrong representation of the large-scale atmospheric circulation --> As already stated above, I think this is a bit too strongly formulated. I'd rather say "correction of errors inherited from a wrong representation of the large-scale atmospheric circulation is challenging".*

Response: You are right, this sentence now reads as follows: “Now while downscaling methods are able to imprint the effects of the local climate factors on the coarse resolution GCM, the correction of errors inherited from a wrong representation of the large-scale atmospheric circulation is challenging (Prein et al., 2019)” (see lines 49-51 of the revised manuscript)

Your comment: *Line 70: the three aforementioned regions --> Which regions are you referring to?*

Response: Here I refer to Greenland and the surrounding seas, the southwestern U.S. and the Gobi desert. For the revised manuscript, this sentence was removed from the Introduction section.

Your comments:

Applied Data and Usage: Line 88: integrations for given model --> integrations for ***a*** *given model*

Line 101: and the considerations of other model developers --> and the considerations of other model ***developments****.*

Line 104: metadata provided the model output files --> metadata provided by the model output files.

Line 111: but also the by the -> but also by the

Line 118: Roberts et al. (2019)) --> Roberts et al., 2019)

Methods: Line 196: being the the standard --> being the standard

Response: Thanks for careful reading, all these errors have been corrected in the revised manuscript.

Your comment: *Line 198: Is CRMSE used for the ranking as well?*

Response: The CRMSE is used here instead of the MAE since the original version of the Taylor-Diagram works with anomaly fields, i.e. removes the pattern mean value from observations and model data prior to calculating the error statistics (Taylor 2001).

Your comment: *Model contributions from ...: (This is a very useful overview!). Considering the EC-EARTH model: Do you think the good performance can be explained by its relationship to the ERA5 reanalysis in terms of model parts? Maybe it's worth adding a note on that. When you compare to JRA-55 you see that the performance of EC-EARTH drops (but it still outperforms many other models). Maybe this can also explain the additional outliers mentioned in line 575.*

Response: EC-Earth's atmospheric component was derived from ECMWF's Integrated Forecasting System, which was also used to produce the ERA-Interim reanalysis (ERA5 is not used in the present study). This might explain why the performance for EC-Earth is slightly better when compared with ERA-Interim instead of JRA-55. However, this effect is small and notably shifts in the model ranks only in those regions where the two reanalyses substantially differ from each other. In fact, the outliers you mention are mainly located in these 3 regions. As mentioned by you, the overall results depicted in Figure 11 do not change if JRA-55 is used as reference reanalysis instead of ERA-Interim. This is pointed out in lines 637-42 of the revised manuscript and visualized in the supplementary material (see figs-refjra55/as-figure-11-but-wrt-jra55.pdf therein).

Your comments: *line 512: not argument --> not* ***an*** *argument; line 520: it had to excluded --> it had to* ***be*** *excluded; line 582: to obtain the size of combined --> to obtain the size of* ***the*** *combined; line 604: been run been to --> been run to*

Response: Thanks for careful reading. The aforementioned text passages were removed or corrected in the revised manuscript.

Your comment: *Summary, discussion and conclusions, line 671: Select the most favourable model --> Although the proposed method is objective, I don't think it will allow the user to select "the most favourable model". First of all, it only covers a certain aspect (representation of circulation frequencies), and taking other performance scores into account (e.g. temperature biases) will give a different model ranking. Further, the ranking provided is based on annual frequencies. Looking at seasonal frequencies will probably also provide different rankings. So in the end, the selection of "the most favourable model" will be a subjective user decision depending on the weight he gives on different aspects. In summary, the performance atlas provided in the paper provides a very useful* ***additional*** *source for model selection, but will not provide a singular basis for that decision.*

Response: The applied methods surely only cover certain aspects of model behaviour. However, as you state below, since the Lamb Weather Types are well known to be associated with typical regional precipitation, temperature and wind patterns, they constitute a good overarching concept. The text passage you mention above no longer appears in the revised manuscript.

Your comment: *General: As far as I understood, the LWT classification only takes pressure gradients into account. Did you also look at biases in pressure, e.g. the monthly SLP pressure bias in the models? Is there a relationship between the ranking you calculate and the pressure bias, e.g. models with a large pressure bias perform not well. Or is it possible that models with a large pressure bias nevertheless show a good representation of LWT patterns?*

Response: From the results of many previous studies, I would say yes, there is a relationship, but I did not specifically assess this issue in the context of the present study. I would expect only a weak relationship between the bias of to the point-wise mean SLP and the MAE of the LWT frequencies because LWTs are defined by pressure gradients rather than absolute values. However, I might of course be wrong and it would be worthwhile to look into these relationships in the future, also in regard with temperature and precipitation biases.

Once again, many thanks for your valuable comments and suggestions and for your efforts to improve the manuscript.