**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.

## Author Response to Referee 2

Dear Referee 2,

Thank you very much for taking the time to review this manuscript and for your efforts to further improve it. Please find below a point-by-point response to your valuable comments.

Your comment: *The paper "A circulation-based atlas of the CMIP5 and 6 models for regional climate studies in the northern hemisphere" by Swen Brands analysis regional atmospheric circulation patterns based on Lamb (1972) for a set of CMIP5 and CMIP6 models. In general, the paper is a rich information source and an important contribution to the climate science community. In my view there are only some minor points that could be addressed differently or better.*

Response: Many thanks for taking the time to review this study and for your positive feedback. Please find below a point-to-point list of responses to your valuable comments.

Your comment: *In section 6 it might be worth mentioning that this study can serve as a basis for studies on future changes in LWT. It could be interesting to investigate, e.g. if certain LWT changes can be traced back to model families or model configurations (AOGCMs vs. ESMs).*

Response: You are right, this interesting idea is mentioned in more general terms in lines 743-745 of the revised manuscript.

Your comment: *Lines 575-594: The explanation of Table 2 and the metric behind is a bit too brief in my view. It seems an interesting analysis, but I cannot fully follow the different steps resulting in the numbers of Table 2. A bit more detailed explanation here, would be good.*

Response: Thanks for your valuable comment. In the revised manuscript, the procedure used to obtain the different mesh sizes is now explained with more detail (see lines 643-653) and an additional boxplot is introduced to visualize the relationship between the horizontal mesh size of the atmospheric model components and the coupled models' performance (see Figure 13, panel a).

Your comment: *Lines 621-622: I would just drop the name Commonwealth models here. It is used only once thereafter in the following sentence and does not improve the text (it might even confuse more than it helps).*

Response: The "Commonwealth" term has been completely removed from the revised manuscript.

**Author response to Dr. Roland Séférian**

Dear Dr. Séférian,

Thank you very much for taking the time to review the present study. Please find below a point-to-point list of responses to your very valuable comments and suggestions.

Your comment: *This work provides a relevant and timely analysis of CMIP5 and CMIP6 models with respect to the representation of circulation in the northern hemisphere. On top this work goes well beyond a routine assessment of global model performance because some of those global models will be used to drive lateral boundary conditions of regional models or to derive climatic impact-drivers at regional/local scale. While it also shows the general improvement from CMIP5 to CMIP6 in this aspect, this work shine light on some deficiency within the current generation of models. With the objective the author tries to map model performance on two axes: the complexity and the resolution, both of which are difficult to separate.*

Response: Thank you very much for your interpretation of the study. The principal aim of the study is to provide a performance estimate for the GCM configurations participating in CMIP5 and 6, which is based on recurrent regional atmospheric circulation patterns as suggested in Maraun et al. (2017). The secondary aim is to provide a simple approach to measure the complexity of these models, which might then be used as additional GCM selection criterion apart from model performance (see Section 3.3, Table 1 and Figure 13b). To my knowledge, this point has been rarely taken into account in regional climate studies so far. The approach proposed here should be interpreted as a reasonable starting point to measure model complexity and is, of course, open to further modifications and specifications in the future. A corresponding Python function is available at github and can be edited by everyone interested to do so. Importantly, the present study is only *one* among many other studies currently taken into account for GCM selection in regional climate initiatives.

Your comment: *In consequence, the high-level picture of the analysis emerging for this work strongly tights to the Table 1 — where we spotted some errors. For instance, it is indicated that CNRM-CM6-1 and CNRM-CM6-1-HR included online chemistry onboard whereas the description of these model configurations in Voldoire et al. (2019) doesn't support this feature. Same goes, for IPSL models and for GFDL-CM4 which are characterized as 'ESMs' in Table 1 while they do not fit the current understanding of what is an Earth system models (see Jones (2019)). As shown in Séférian et al. (2020), GFDL-CM4 indeed included marine biogeochemistry but only in a stylized manner (reduced complexity marine biogeochemical models). In consequence, there are no biophysical feedbacks represented in GFDL-CM4 whereas it does in GFDL-ESM4.*

Response: Many thanks for revising Table 1, I very much appreciate your comments on this table, since it is key for the understanding of the study. Regarding CNRM-CM6-1 and CNRM-CM6-1-HR, an interactive atmospheric chemistry model was erroneously added in this table because the source attribute of the netCDF output reads as follows (the following example is for CNRM-CM6-1):

u'CNRM-CM6-1 (2017): aerosol: prescribed monthly fields computed by TACTIC_v2 scheme atmos: Arpege 6.3 (T127; Gaussian Reduced with 24572 grid points in total distributed over 128 latitude circles (with 256 grid points per latitude circle between 30degN and 30degS reducing to 20 grid points per latitude circle at 88.9degN and 88.9degS); 91 levels; top level 78.4 km) atmosChem: **OZL_v2** land: Surfex 8.0c ocean: Nemo 3.6 (eORCA1, tripolar primarily 1deg; 362 x 294 longitude/latitude; 75 levels; top grid cell 0-1 m) seaIce: Gelato 6.1'

Since “atmosChem: OZL_v2”, I interpreted this as an interactive component model, as is normally the case if a model is specified for a given realm and the term “prescribed” is missing (compare with the *atmosChem* entry with the the *aerosol* entry above). I had noticed that this was in disagreement with Voldoire et al. (2019), but unfortunately gave preference to the aforementioned source attribute which I interpreted wrongly.

Many thanks also for pointing out that ocean biogeochemistry in GFDL-CM4 is represented by a reduced complexity marine biogeochemical model without biophysical feedbacks. In the revised manuscript, the respective complexity integer was consequently set to 1.

In order to avoid such errors, the complexity codes provided in Table 1, column 7 of the revised manuscript have been confirmed and corrected by the corresponding modelling teams by means of e-mail correspondence.

Considering Collins et al. (2011), Jones (2020) and personal e-mail correspondence I have had with two modelling groups, I have come to the conclusion that there exist at least 3 different definitions for the term “Earth System Model”:

1. “ESM could also be defined as adding other than pure physical processes of ocean and atmosphere to the classical GCM.” (personal communication with the EC-Earth group)

which is qualitatively identical to:

“These models are now known as Earth System Models (ESMs) to denote that they simulate more than just the “physical” elements of the world's weather” (Jones 2020, page 1)

2. “There is no strict definition of which processes at what level of complexity are required before a climate model becomes an Earth system model [...] however typically the term “Earth system” is used for those models that at least include terrestrial and ocean carbon cycles.” (Collins et al. 2011, page 1051, this definition was used in the first version of the manuscript)

3. The ESM term is “mostly about fashion, less about content” (personal communication with a developer from one important GCM group)

Since the ESM term is not clearly defined, it is avoided in the revised manuscript. Instead, the manuscript now simply refers to “more” or “less” complex models.

Your comment: *Regarding the axis of the resolution, providing the nominal resolution would help to compare model between each other. The nominal resolution has been reported by the modelling groups to CMIP6 for each component/realm.*

Response: The nominal resolution reported to CMIP6 is only approximate and this is why the present study works with the mesh size of the atmosphere and ocean grids, as indicated in the source attributes or directly by the data arrays in the netCDF files. This approach is also approximate, but likely more exact than reported nominal resolutions. Also, the nominal resolution of the model versions participating in CMIP5 has not been reported, which is another reason for the use of the alternative approach.

Your comment: *Apart from these remarks/on Table 1, we would like to provide a couple of suggestions that could be useful for this work.*

*As this work focus on the performance over the historical period, it might be relevant to provide some information on how the model has been tuned/calibrated. At least to know if this set of metrics has been used as a target to prepare the model for CMIP5 and for CMIP6. Such questions tend to emerge now in the literature (see Spafford and MacDougall, in review ni GMDD) because of their implication on routine performance benchmark.*

Response: This is a very interesting issue, which however is very difficult to trace back to all model configurations used in the present study, inlcuding those from CMIP5. For the model family performing best here (EC-Earth3), Klaus Wyser and Ralf Döscher confirmed that: "In the EC-Earth3 tuning process, regional SLP patterns were not a target." via e-mail correspondence. A further look at Döscher et al. (2021) reveals that "The atmospheric component of EC-Earth has been tuned with the goal of achieving a reasonably small radiative imbalance at the top of the atmosphere". A more systematic assessment of this issue is interesting, but out of the scope of the present study.

Your comment: *On the other hand, the paper is not clear on the treatment of the model realization. As shown in Olonscheck et al (2020), large ensemble of realization may improve the comparison with the observation. Considering the magnitude of the internal variability of the atmospheric circulation feature, considering additional information on available model member might help. With that said, comparing model with different ensemble size might complicate the picture but discussing the impact of the member on the overall model performance and ranking would be a very valuable outcome of the paper*.

Response: I am afraid that this might be a misunderstanding since already in the first version of the manuscript a total of 70 alternative runs from 12 distinct GCMs were assessed to estimate the role of internal variability (see Figure 11 in manuscript version 1). This long list has now been extended to 72 alternative runs from 13 distinct GCMs. Namely, 2 additional CNRM-CM6-1 members were included in Figure 12 of the revised manuscript. As you can see from these figures, the effect of internal variability on the overall results is negligible.

Your comment: We hope that the author will find these comments and suggestions useful/relevant.

Response: I am very grateful to your valuable corrections and suggestions. Thank you very much for taking the time to review this study.

The references cited in this response letter are listed in the revised manuscript.