Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC1
https://doi.org/10.5194/gmd-2020-418-RC1, 2021

Comment on "A circulation-based performance atlas of the CMIP5 and 6 models for regional climate studies in the northern hemisphere"

Anonymous Referee #1

Referee comment on "A circulation-based performance atlas of the CMIP5 and 6 models for regional climate studies in the Northern Hemisphere mid-to-high latitudes" by Swen Brands, Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2020-418-RC1, 2021

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:

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?
- 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.
- Line 14: The latest model generation --> add (CMIP6).

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".
- Line 70: the three aforementioned regions --> Which regions are you referring to?

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
- Line 198: Is CRMSE used for the ranking as well?

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 reanalyses 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-AERTH drops (but it still outperforms many other models). Maybe this can also explain the additional outliers mentioned in line 575.
- 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

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.

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?