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.