# Reviewer 2 response

Recommendation: Major revision

The authors assessed CMIP6 models in terms of their performance and diversity in simulating several variables, e.g., temperature, precipitation, and circulation, over Europe. Based on the assessment, they created sub-sets of CMIP6 models, which can be used for downscaling or impacts assessments. The approach can also be applied to other regions of the world. The topic is important and falls within the scope of the journal. The manuscript is generally well written. My major concern includes: the assessment of CMIP6 models did not well consider the link between the model's ability to simulate historical climate and future climate change. The assessments are overly dependent on subjective assessment criteria. Detailed comments are laid out below.

We thank the reviewer for this overall positive and constructive response.

Major comments:

1. No link was established in terms of the model's ability to simulate the historical climate and the projected changes. Thus, the models that can better reproduce historical climate may not necessarily generate a more reliable projection of future climate. After excluding the least realistic models, the filtered CMIP6 models show higher sensitivity. Is the result reasonable?

It is correct that we do not attempt to explicitly link baseline performance to the credibility of future projections. What we do suggest, is that there are a number of issues around using climate model projections from models which do not behave realistically in terms of key large scale regional climate characteristics in the baseline climate. Here the question is not whether well performing (better) models can offer a (more) reliable projection, but whether those models that we know to be particularly unrealistic in terms of the key large scale climate characteristics that determine the regional weather, and its variability can offer useful information about projected future climate to the climate impacts community. An increasing body of literature does link short comings in the ability of a model to realistically represent an observed baseline to being an indicator that the models' future projections are less reliable (e.g., Whetton et al., 2007; Overland et al., 2011; Lutz et al., 2016; Ruane and McDermid, 2017; Jin, Wang and Liu, 2020; Chen et al., 2022).

Having identified models that we consider particularly unrealistic to arrive at a filtered subset, we then explore what that means for the range of future projections. We find that the better-performing filtered subset happens to contain a higher proportion of higher sensitivity models. This study is not intended to present an emergent constraint, but an exploration of how the performance-based filtering impacts projection range compared with other sub-selection approaches. We do not conclude that the upper-end of the projection range is more credible for Europe – indeed this would not be a reasonable result as the reviewer asks, but we do think that the identified relationship between filtered ensembles and climate sensitivity highlights a tension with other potential selection approaches, such as selecting models based on global historical trends, or matching IPCC distributions of climate sensitivity. Our intention is to expose this tension for potential users of these simulations, over Europe.

These findings are complemented by a recent study that takes account of regional temperature trends which, finds that for some European areas (e.g., France), constraining the CMIP6 ensemble based on regional temperature trends, or a combination of regional and global temperature trends finds that projected summer temperature changes are shifted towards high sensitivities rather than the lower sensitivities suggested by global analyses (Qasmi and Ribes, 2022; Ribes et al., 2022). We find that the higher sensitivity models that are part of our filtered ensemble may still provide a useful projection for the European region.

We propose to clarify in our manuscript that our result is to highlight this tension between selecting subsets based on regional performance and selecting subsets based on other criteria e.g., to represent the IPCCs plausible range of climate sensitivity.

Chen, Z. et al. (2022) 'Observationally constrained projection of Afro-Asian monsoon precipitation', Nature Communications, 13(1), p. 2552. doi: 10.1038/s41467-022-30106-z.
Jin, C., Wang, B. and Liu, J. (2020) 'Future Changes and Controlling Factors of the Eight Regional Monsoons Projected by CMIP6 Models', Journal of Climate. Boston MA, USA: American Meteorological Society, 33(21), pp. 9307–9326. doi: 10.1175/JCLI-D-20-0236.1.
Lutz, A. F. et al. (2016) 'Selecting representative climate models for climate change impact studies: an advanced envelope-based selection approach', International Journal of Climatology. John Wiley & Sons, Ltd, 36(12), pp. 3988–4005. doi: https://doi.org/10.1002/joc.4608.
Overland, J. E. et al. (2011) 'Considerations in the Selection of Global Climate Models for Regional Climate Projections: The Arctic as a Case Study', Journal of Climate. Boston MA, USA: American Meteorological Society, 24(6), pp. 1583–1597. doi: 10.1175/2010JCLI3462.1.
Qasmi, S. and Ribes, A. (2022) 'Reducing uncertainty in local temperature projections', Science Advances. American Association for the Advancement of Science, 8(41), p.

eabo6872. doi: 10.1126/sciadv.abo6872.
Ribes, A. et al. (2022) 'An updated assessment of past and future warming over France based on a regional observational constraint', Earth Syst. Dynam. Discuss., 2022(March), pp. 1–29. doi: 10.5194/esd-13-1397-2022.
Ruane, A. C. and McDermid, S. P. (2017) 'Selection of a representative subset of global climate models that captures the profile of regional changes for integrated climate impacts assessment', Earth Perspectives, 4(1), p. 1. doi: 10.1186/s40322-017-0036-4.
Whetton, P. et al. (2007) 'Assessment of the use of current climate patterns to evaluate regional enhanced greenhouse response patterns of climate models', Geophysical Research Letters. John Wiley & Sons, Ltd, 34(14). doi: https://doi.org/10.1029/2007GL030025.

2. Quantitative measures are preferred for model evaluation. Visual inspection hinders the inter-comparison of various studies to a certain degree as different people may have different judgments on "satisfactory", "unsatisfactory", and "Inadequate". I'm wondering to what extent the results will be different if the authors use objective assessment criteria only.

We understand the reviewers point that in the case of more subjective criteria, to a certain degree people may have different judgements. We have used a combination of quantitative and qualitative measures where we have found them appropriate. One point to note is that 'quantitative' is not always synonymous with 'objective' – e.g., the choice of metric and threshold for classification involves subjective judgements.

There are two main reasons for our use of qualitative measures – firstly is to account for the variety of characteristics in errors that different models display and allow us to judge their implications and significance. If we look at the climatological circulation assessment, we find that the RMSE calculated in parallel with the quantitative assessment doesn't always lead to the same classification as the visual inspection – in this case because some patterns of error are more concerning to us than others – errors in magnitude of the mean circulation (feature in broadly correct locations but with errors in magnitude) are less concerning than cases where features are incorrectly located. Visual inspection allows us to understand the characteristic of the error and consider its impact on other aspects of the model.

Figure 1 shows some examples of where bias alone and/or a RMSE threshold for windspeed would not be suitable to determine the classification of the models. BCC-CSM2-MR is classified as satisfactory for DJF circulation (Fig.1b and e). This is because although there are some errors in windspeed magnitude over western and central Europe, the pattern of large- scale circulation is reasonably well captured (as compared to ERA5 in Fig.1a). The BCC-CSM-MR model has a similar regional RMSE as BCC-ESM1 (Fig.1f), however this model is classified as unsatisfactory due to a lack of south westerly winds over the northern UK and Scandinavia (Fig,1c and f). This is

also highlighted by the negative bias in windspeed over these areas, indicating that the winds are too weak (Fig.1f). The ACCESS-ESM1-5 model (Fig 1.d and g), is also classified as unsatisfactory despite a lower regional RMSE than BCC-CSM2-MR, this is due to the wind direction being too westerly in the North Atlantic and over the UK and northern Europe. The windspeeds over Scandinavia are too weak, while the windspeed over the UK and central Europe is too strong (Fig.1g).

A quantitative metric might be designed to capture these characteristics on which our judgement is made, but this may 'miss' another error characteristic that subsequently appears in another model.

The second reason for using visual inspection is that the process of examining the fields offers us a much better understanding of model characteristics, which does not arise from summary statistics. In the study presented we have often shown quantitative metrics which were used in parallel with visual inspection.

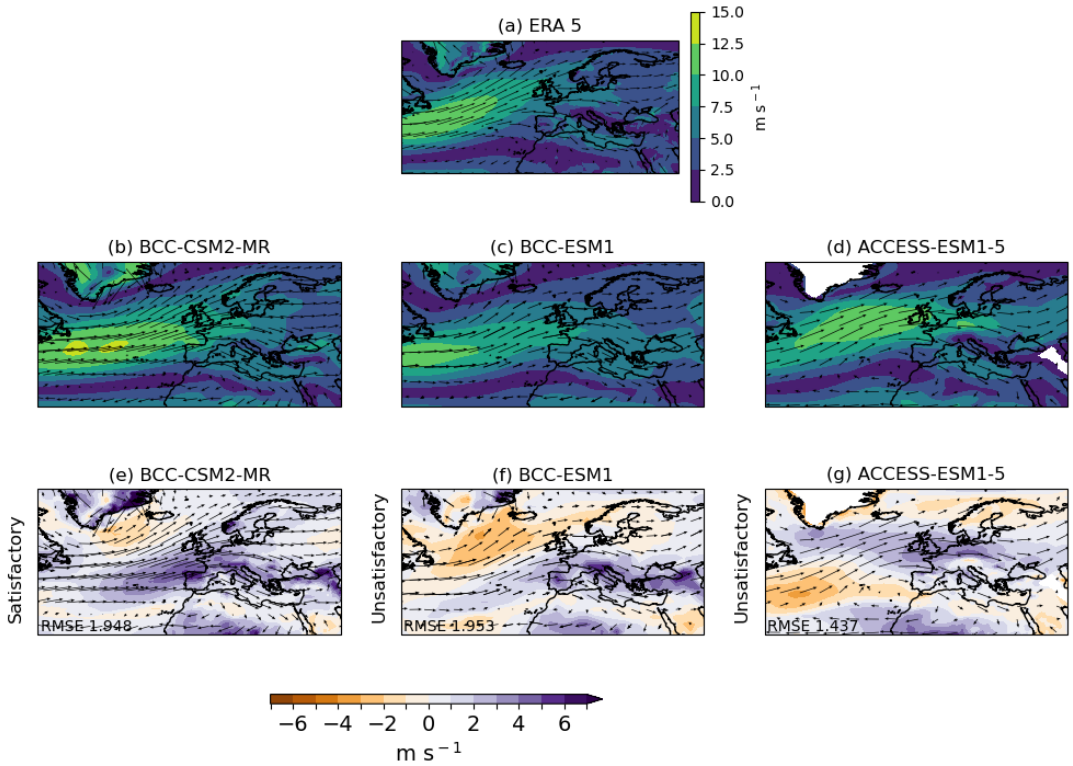


*Figure 1 a) ERA5 DJF wind climatology (windspeed and direction 1995-2014), b)-d) Examples of DJF circulation patterns for model climatology. Vector arrows show absolute wind direction, contours show absolute windspeed, e)-f) As in row above but with contours showing windspeed bias compared to ERA5 (vector arrows still show absolute wind direction).*

How was the RMSE of the zonal mean track calculated? It seems that the authors calculated the zonal mean track and obtained a time series. The RMSE is calculated using the time series derived from models and observation. Please note it makes no sense by comparing the year-to-year variation of the unforced internal variability derived from AOGCMs against the observed one. In this case, the RMSE is largely determined by the phase discrepancy between simulation and observation. Please also check the use of RMSE elsewhere.

Thank you for bringing to our attention that this part of the methodology requires further clarification. The RMSE was not calculated using a time series or via consideration of each model's internal variability. This is the case for all the variables. The zonal mean of the model mean track density from 20W-20E was taken to get a profile of storm number by latitude. Then the RMSE was calculated of the models compared to the profile obtained from ERA5. The RMSE was calculated from 25-80N. There is no timeseries element of this and it is just the RMSE of the zonal mean, model mean track density. At no point is the unforced interval variability of the models compared or used in the RMSE calculations. We will clarify this in the paper's text.

Other comments:

Section 2: It is not clear to me how the CMIP6 models are grouped into classifications. Please clarify how the quantitative and qualitative measures were used and what is the threshold of quantitative measures to group the models. I suggest the authors introduce the "criteria" first and explain the classification definitions based on the criteria.

Models were classified for individual criteria and not grouped into an overall classification (figures 4 and 5 in the manuscript). Models were then sub-selected based on whether they had any red flags (inadequate) and the percentage of orange (unsatisfactory) flags. This is presented as only one example of how the assessment can be used to sub-select models. An alternative approach, for example, would be to only remove models with an inadequate flag. Thank you for this suggestion we will improve the clarity of the main text.

L64: "processed based" -> "process-based"

L70: How the regional processes are linked to future changes?

L137: "process base" -> "process-based", "does not use and regional or global warming trends"->"does not use regional or global warming trends". Please carefully

read throughout the manuscript and correct the typos or grammar mistakes. E.g. L202 …

Thank you for noting these errors, these will be corrected, and the final manuscript will be proofread.

L217: What is the temporal resolution of the dataset, monthly mean or daily mean? Which CMIP6 experiment was used for the baseline period? Both the baseline and future periods are only 20 years. The climatological means averaged over 20 years may still contain internal climate variability, e.g., AMO or PDO, which may affect the evaluation and selection of the models to a certain extent.

Thank you for highlighting that this is not clear, we use monthly datasets and the historical experiment for the baseline. We have selected the time periods used in the assessment to align with the European Projections Project (e.g., Brunner et al., 2020).

Brunner, L. et al. (2020) 'Comparing Methods to Constrain Future European Climate Projections Using a Consistent Framework', Journal of Climate, 33(20), pp. 8671–8692. doi: 10.1175/JCLI-D-19-0953.1

L225: Please clarify what reanalysis and observational data were used in this study.

ERA5 was the reanalysis data used for the assessment criteria. The exception is for precipitation where EOBS data was used. This will be clarified in the text.

L254-255: How the circulation pattern is measured? Is the RMSE calculated using two wind speed fields or an RMS vector error between two vector fields? If the RMSE is calculated with wind speed, it does not reflect the errors in wind direction. Instead, the RMSE for vector field can reflect both errors in wind speed and wind direction. Therefore, I suggest the authors use the latter one. Similarly, the difference in wind speed illustrated in Fig. 1 can only describe the errors in wind speed. The same wind speed does not mean the same wind direction. The authors may consider using a vector difference between the model and ERA5. The magnitude of vector difference takes both differences in wind speed and wind direction into account.

The wind speed was used as a measure of the magnitude of error, while the circulation pattern of wind direction and magnitude was assessed visually. Thank you for this suggestion, it may be interesting to use the vector error in addition to the windspeed and see if this is a better indicator of errors in the circulation pattern.

Xu et al, 2016: A diagram for evaluating multiple aspects of model performance in simulating vector fields. Geosci. Model Dev., 9, 4365–4380

L270: Please explain how the "track density" is defined. Please use the degree symbol "°" to represent latitude and longitude here and elsewhere.

The track density is calculated using an objective cyclone tracking and identification method based on 850 hPa relative vorticity (Hodges, 1994, 1995). The method and data are the same used in Priestley et al. (2020). This will be clarified in the manuscript.

Hodges, K. I., 1994: A general method for tracking analysis and its application to meteorological data. Mon. Wea. Rev., 122, 2573–2586, https://doi.org/10.1175/1520-0493(1994)122,2573: AGMFTA.2.0.CO;2.
Hodges, K. I., 1995: Feature tracking on the unit sphere. Mon. Wea. Rev., 123, 3458–3465, https://doi.org/10.1175/1520-0493(1995)123,3458: FTOTUS.2.0.CO;2.
Priestley, M. D. K. et al. (2020) 'An Overview of the Extratropical Storm Tracks in CMIP6 Historical Simulations', Journal of Climate. Boston MA, USA: American Meteorological Society, 33(15), pp. 6315–6343. doi: 10.1175/JCLI-D-19-0928.1

L321: "depending to on" -> "depending on"

L334: "with with" -> "with"

L343-345: How about the range of other quantities, e.g. precipitation and storm track density?

The authors agree that it would be interesting to investigate other variables, it would extent the scope and length of the existing paper considerably to consider projections from the filtered ensemble for all the criteria that have been assessed. This is something that the authors are interested in exploring further and in a more thoroughly in a second follow up paper.

L362: Please clarify what numerical score was given for each group of models.

This information can be added to the manuscript in the supplementary info. Each model was scored individually. Models were classified for individual criteria and not given an overall classification. Models were then sub-selected based on whether they had any red flags (inadequate) and the percentage of orange (unsatisfactory) flags. This is presented as one example of how the assessment can be used to sub-select models. Model sub-selection is always subjective to some extent and the approach will depend on the application.

L644: “35°N-75°” -> “35°N-75°N”

Fig. S4: What does the “??” refer to in the figure caption?

This is a typo, it refers to table 2 in the main manuscript, thank you for noting this, it will be corrected.