Replies to Referee #2

Does the weighting of climate simulations result in a more reasonable quantification of hydrological impacts?

Hui-Min Wang, Jie Chen, Chong-Yu Xu, Hua Chen, Shenglian Guo, Ping Xie, Xiangquan Li

We would like to thank the reviewer for the time taken in reviewing this paper. All comments are all valuable to improve this manuscript. Please find the point-by-point responses below. For clarity, comments are given in *italics*, and our responses are given in plain text. We will make the revisions to the manuscript as suggested.

> *Summary and General Comments*
>
> *The manuscript by Wang et al investigates the impact of multiple ensemble weighting techniques on the simulations of hydrological impacts for two different river basins. The authors compare the results from a hydrological impact model driven by weighted and unweighted GCM projections. In addition, the authors compare the results from bias-correcting the GCM output before weighting or not. They conclude that weighting the bias corrected GCM output has not a large effect while differences are larger when using raw output, improving the representation of the mean hydrograph and reducing the annual streamflow bias. The authors conclude that the equal weighting method is a conservative approach and still viable given the small effect weighting has on a bias corrected ensemble.*

We would like to express gratitude to the referee for reviewing this manuscript and offering precious comments and suggestions. All the comments and suggestions have been replied to below and will be addressed in the revision of manuscript.

> *Overall the paper is well and comprehensively written and the analysis extensive. The fact that weighting the bias corrected ensemble has a very small effect is not surprising. Given that bias correcting and weighting for performance have the same goal, bringing the ensemble closer to observed values, I do not understand why you would do both? Some of the risks and disadvantages for weighting, which are all true, also apply for bias correction (e.g. Sippel et al. 2016, Maraun et al. 2017). Both tools need to be applied carefully and have their pitfalls. For instance, it has been shown that out-of-sample testing is crucial for any kind of weighting or sub-sampling (e.g. Herger et al. 2018, Abramowitz et al. 2019), which is still missing so far in this study. In that sense I am not convinced that the authors come to the correct conclusion, even though their arguments are generally not wrong (see below). Weighting a GCM ensemble will conserve dependencies between different variables in a physically consistent way, and in cases where this is important, it might be preferred over bias correction. However, all the risks*

> *the authors mention apply, and the study shows nicely that there are still many open questions on how to use these methods properly. I would recommend to rephrase some of the discussion and conclusions more carefully and also account for the assumptions and risks in bias correction.*

We agree with the referee that bias correction methods have similar goals as most of the weighting methods, which is to bring ensembles closer to observed values. Nevertheless, they still have different traits and functions. The bias correction directly deals with the biases of climate simulations and bridges the gap between the coarse outputs of climate models and data requirements of hydrological models. The model weighting assigns relative reliability to each climate simulation and aggregates multi-model ensembles. There are some differences between climate simulations whether the bias correction is done or not. In this case, a model weighting method always needs to be determined in order to obtain the overall impact evaluation and relevant uncertainty. Although the equal weighting is usually used in hydrological impact studies, it still deserves detailed investigation whether an unequal weighting method is necessary for bias-corrected ensembles. This problem is also mentioned in other studies (Alder and Hostetler, 2019; Chen et al., 2017). In addition, we agree with the referee that the bias correction has some similar risks to the model weighting. But the main goal of this study is to investigate the effects of model weighting when the bias correction is or is not conducted. We agree that these risks and problems of bias correction should be still addressed in the manuscript.

As suggested by the referee, we have added out-of-sample testing when evaluating performances of different weighting methods. The performances of weighting methods in this case are similar to the previous results that are based on historical observation. This confirms the conclusion of this study. Detailed results and relevant analysis will be presented in the revised manuscript.

> *Specific comments*

> *P1, L23-25: This conclusion is a bit far-fetched and ignores the independence issue nicely described on page 2, around line 20.*

Thanks for the referee's comment on this conclusion. We agree that this conclusion neglects the strengths of model weighting in impact studies and excessively trusts the function of bias correction. Actually, whether the bias correction overmatches the model weighting is not the research problem of this study, and we should focus on the effects of model weighting in two conditions (whether the bias correction is done or not). Thus, this conclusion will be fixed to "The equal weighting method may still be a viable and conservative choice when bias correction is conducted in the studies of hydrological climate change impacts".

> *P4, L8: not relevant.*

Thanks for the comment. We agree that the introduction to the characteristics of the Daniel-Johnson Dam is redundant, but we also think that it is necessary to mention the Daniel-Johnson dam because the discharge data used for calibrating the hydrological model is collected here and is the inflow of the reservoir. This sentence will be shortened to “The outlet of the Manicouagan-5 River is the Daniel-Johnson Dam”.

P6, L20: The climatological mean of what? Temperature, precipitation, streamflow? All of them together or only individual? That makes a large difference and it has been shown that only using one at a time for PI is risky (Lorenz et al. 2018).

Thanks for the comment on the presentation of methodology. We failed to state it clear enough. In this sentence, the climatological mean is for streamflow as indicated by P7, L1-3. Since GCM’s performances on hydrological simulation are related to multiple variables (such as precipitation and temperatures in this study) and there is no widely accepted way to combine multiple sets of weights into single one, this study proposed to determine weights based on streamflow series. In this way, weighting based on streamflow simulations can synthesize GCMs’ performances in both temperature and precipitation and circumvents the problem of non-linear relationship between climate and impact variables. In addition, calculating weights based on temperature and precipitation is also used in this study for comparison, as stated in P7, L1-3. Herein, the used variable will be stated in this sentence in the revised manuscript for clarity.

P9, L19-23: Yes, but the same assumption applies for bias-correction.

As stated in P9, L15-17, we agree with the referee that the assumption of stationary biases in GCM outputs also applies for the bias correction method. In this sentence, we intend to state that if the weighting methods still follow the same assumption as the bias correction (as most performance-based weighting methods do), there will be no needs to do unequal weighting. However, some other weighting methods contain other criteria that do not follow the same assumption, such as the interdependence criterion in the PI method and the future convergence criterion in the REA method. This point will be rephrased in the revised manuscript.

P10, L2-5: The testing is all done in sample. Out-of-sample testing is needed.

Thanks for the suggestion. We agree with the reviewer and we have done the out-of-sample testing by conducting model-as-truth experiments (Herger et al., 2018). Relevant results and analyses will be added and discussed in the revised version of the manuscript.

P10, L9-11: At least for PI any metric could be considered, the fact that only climatology was used is because the authors chose to do it this way, but is not a property of the method.

We appreciate the referee’s comments on this analysis. We approve of the idea that other metrics can be used in PI method. In fact, different metrics can also be applied into some other weighting methods if users want to do so, even though these methods are designed to use the climatological mean. However, many researches and end-users in hydrological impacts only consider the climatological mean (e.g., Wilby and Harris, 2006; Chen et al., 2017). In this sentence, we were to express that the different performances on different metrics may be due to the usage of weighting methods by end-users, which we failed to state clearly in the submitted version of the manuscript. Thus, this sentence will be corrected to "This may be because only the mean value (climatological mean or monthly mean series) was used as the evaluation metric when determining weights in this study, and peak or extreme values were not considered".

In addition, using different metrics may result in different performances of a weighting method (as stated in P12, L29-31). However, the main focus of this study is the effects of weighting GCM based on their performances in streamflow simulations. Whether other metrics bring about different results needs further research and is beyond the scope of this study. Therefore, in the Discussion section, we will add some discussion on the metric adopted for the weighting methods.

> *P11-P12, L31-4: While these arguments are true, bias-correction has similar problems. Also, it looks to me that the equally weighted ensemble has the same issue?*

We agree with the referee that similar to the model weighting, the bias-correction has the problem of non-linear relationship between climate variables and impact variables. However, the bias-correction does not have the problem of trade-off among different climate variables. This is because bias correction is done for each variable, and corrected variables are then inputted to the impact model at the same time. No trade-off needs to be processed in this procedure. For model weighting methods, how to combine different sets of climate-based weights becomes a question. For example, the weights calculated based on temperature and precipitation need to be combined into a single set when generalizing the hydrological impacts for the two watersheds in this study. In this case, the trade-off between two variables is needed, which may be varied in different watersheds.

Similarly, the equal weighting is the same. The trade-off between different variables is not needed, but it also cannot circumvent the problem of non-linear relationship between climate and impact variables. Therefore, as stated in P13, L18, the equal weighting is only a conservative option for handling multi-model ensembles in impact studies. All these problems will be discussed in the revised manuscript.

> *P12, L27-28: I do not think the results fully support this statement. We might not have found a clearly better way than model democracy, but equal weighting is as at least as arbitrary as weighting.*

We thank the referee for this comment. We agree that this statement is somewhat ambitious for the results. This statement will be fixed to "When GCM outputs are processed by the bias correction, compared to the equal weighing method, unequal weighting methods do not bring about much different impact results”.

> *P13, L1-2: Equally weighting is also arbitrary, given that is assumes all models are equally likely and independent, which they are not.*

We appreciate the referee’s comment. As stated in P2, L14-23, equal weighting ignores the differences in the performances and potential dependency of GCMs. But at the same time, unequal weighting methods also have potential problems of reducing projection accuracy and concealing projection uncertainty (as stated in P13, L2-4). Therefore, equal weighting should not be regarded as the final solution but a conservative method, and the weighting methods should be used with cautions for now. Accordingly, this sentence will be corrected as: “All of these aspects in weighting methods are often predefined without detailed examination or based on expert experience and, thus, can actually introduce several layers of subjective uncertainty. Notwithstanding the equal weighting is not a totally perfect solution, model weighting methods should be used with cautions and the results of equal weighting should be presented along with the results of unequal weighting methods”.

> *P13, L5-10: Again, the same applies to bias-correction.*

We agree that the bias correction has the same problem that climate simulations are corrected statically. Herein, we did not intend to say that bias correction methods have some strengths over the model weighting but only to state one problem of present model weighting methods. This could be a focus for future study of model weighting. In order to stress this problem and eliminate vagueness, the statement here will be modified as follows.

Weights are generally assigned to climate simulations in a static way (i.e. weights in the reference period are the same as those in the future period). This usage shares the same assumption with bias-correction methods that the performances of GCM simulations are stable and stationary. However, some studies have shown that model skills are nonstationary in a changing climate, and models with better performance in the reference period do not necessarily provide more realistic signals of climate change. The way to deal with the dynamic reliability of climate models deserves further study.

> *P13, L11: Again, because you chose to only include one metric does not make it a property of the method. At least some of the weighting methods can account for multiple metrics to be included and people argue to do so (e.g. Knutti et al. 2017).*

We agree with the referee that in the PI method, multiple metrics could be used to weight climate simulations. Yet, when introducing multiple metrics, there must be decisions on the relevant diagnostic metrics and the way to synthesize GCM’s overall performances in multiple metrics. Some studies have stated that using calibrated multiple metrics helps to improve the agreement with observation (Knutti et al., 2017; Lorenz et al., 2018), while some argue that multiple metrics form another level of uncertainty within weighting methods (Christensen et al., 2010). These problems deserve further detailed investigation but they are beyond the scope of this study (which is to investigate whether weighting based on streamflow simulations induces better quantification of hydrological impacts). Thus, this sentence will be modified correspondingly to mention the use of multiple metrics and its potentials to strengthen weighted results.

References

Alder, J. R., and Hostetler, S. W.: The Dependence of Hydroclimate Projections in Snow‐Dominated Regions of the Western United States on the Choice of Statistically Downscaled Climate Data, Water Resources Research, 55, 2279-2300, https://doi.org/10.1029/2018wr023458, 2019.

Chen, J., Brissette, F. P., Lucas-Picher, P., and Caya, D.: Impacts of weighting climate models for hydro-meteorological climate change studies, Journal of Hydrology, 549, 534-546, https://doi.org/10.1016/j.jhydrol.2017.04.025, 2017.

Christensen, J. H., Kjellström, E., Giorgi, F., Lenderink, G., and Rummukainen, M.: Weight assignment in regional climate models, Climate Research, 44, 179-194, https://doi.org/10.3354/cr00916, 2010.

Herger, N., Abramowitz, G., Knutti, R., Angélil, O., Lehmann, K., and Sanderson, B. M.: Selecting a climate model subset to optimise key ensemble properties, Earth System Dynamics, 9, 135-151, https://doi.org/10.5194/esd-9-135-2018, 2018.

Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E. M., and Eyring, V.: A climate model projection weighting scheme accounting for performance and interdependence, Geophysical Research Letters, https://doi.org/10.1002/2016gl072012, 2017.

Lorenz, R., Herger, N., Sedláček, J., Eyring, V., Fischer, E. M., and Knutti, R.: Prospects and Caveats of Weighting Climate Models for Summer Maximum Temperature Projections Over North America, Journal of Geophysical Research: Atmospheres, 123, 4509-4526, https://doi.org/10.1029/2017jd027992, 2018.

Wilby, R. L., and Harris, I.: A framework for assessing uncertainties in climate change impacts: Low-flow scenarios for the River Thames, UK, Water Resources Research, 42, W02419, https://doi.org/10.1029/2005wr004065, 2006.