Authors' responses to comments

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 appreciate the editor's and the three anonymous referees' valuable suggestions and comments on the manuscript. These suggestions are helpful to improve this manuscript. We have carefully studied and responded to all comments point-by-point as follows. For clarity, all comments are given in *italics* and responses are given in plain text. The manuscript has been modified accordingly.

Responses to Editor's comments

We would like to thank the editor for reviewing this manuscript. Our responses are as follows.

> *Thank you for posting your responses to the three referees' reports. The reviews raised some important comments and suggestions that I urge you to consider as I believe they will improve the quality of the manuscript. Based on my own reading, I find this to be an interesting paper that could fit the scope of HESS and would be of interest to the community. I invite you to upload a revised manuscript, incorporating the proposed changes and additions, and making any other modifications where you see fit ('major revision').*

We appreciate that the editor is in favor of the content of this research. We have thoroughly studied the comments from all referees, and the manuscript has been revised accordingly. Specific responses to referees' comments have been listed below. Revised manuscript and specific changes to the manuscript have also been attached at the end of the present response.

> *In addition to the comments from the reviewers, I kindly ask you to add "Author contribution", "Data availability" and "Competing interests" sections to the manuscript, as indicated in the guidance for authors.*

Thanks for the suggestion. We have added relevant information at the end of revised manuscript followed the guidance for authors [P15, L21-P16, L2].

Responses to Referee #1's comments

We sincerely appreciate the referee's comments and suggestions on the manuscript. Our responses

are as follows.

> *This is a well written paper that studies the added value of weighting GCMs within an ensemble as a function of hydrological performance rather than as a function of climatological performance as usually done. The paper discusses some interesting aspects (e.g. the difference in outcome if weighting according to precipitation or temperature under different hydrological regimes) and comes to the conclusion that if raw GCM data is to be used, ensembles should be weighted based on streamflow rather than temperature or precipitation. In exchange, there is not much added value with streamflow-based weighting if the underlying GCMs are duly bias corrected. This outcome is not entirely surprising (see detailed comments) but I think it is nevertheless interesting for the readers of HESS and thus worth publishing.*

We appreciate that the referee is in favor of the content of this research. Detailed comments have been replied as follows and addressed in the revision.

> *Detailed comments*
>
> *In this paper, the GCM weighing is tested for large catchments (» 10'000 km^2) that are simulated with a lumped model (GR4J) at a daily time step. With such a lumped model, it can a priori be assumed that the most important aspect of climate inputs for hydrological model performance and for future simulations is the actual precipitation and temperature bias. In fact, there is a whole body of hydrological literature on the importance of correct area-average precipitation estimates, which should perhaps be linked to this study. A starting point is the work of Lebel et al. 1987. Since the model is lumped, spatial differences between meteorological inputs derived from GCMs cannot show up in the simulation results otherwise than affecting the catchment scale average values (i.e. the bias). Differences between GCM outputs in temporal variability do most likely not show up because they are dampened by the model. The authors argue that the response of a catchment to climate input is nonlinear. This holds in general but if such a simple model is used, no surprising outcomes can be expected (not much difference between climate-based weighting and hydrological weighting in absence of major meteorological biases). This is a limitation of the study: major differences between climate-based weighting and hydrological weighting can a priori not be expected in the bias corrected set-up with such a simple model. This has to be discussed in sufficient detail in the paper and highlighted also in the perspectives. Finally: I am not an expert on bias correction methods. Accordingly I can only assume that this part of the work is state-of-the-art.*

Thanks for the comment. We agree with the reviewer that when using a lumped model, the nonlinear relationship between the climate variables and the impact variable (streamflow) may not be sufficiently revealed. Spatial differences between different climate simulations only affect the basin-averaged inputs to the hydrological model but not directly affect the process of runoff generation

and streamflow routing (Lebel et al., 1987). Temporal variations of climate simulations may be partially reduced by the lumped hydrological model as well. With the help of other more sophisticated hydrological models (such as distributed models), the differences between climate-based weights and streamflow-based weights may become more obvious. For the experiment of raw GCM-simulated streamflows, the weights based on streamflow show better performances than those based on climate variables. This may be related to large differences among climate simulations. But in the experiment of streamflows simulated using bias-corrected GCM outputs, no much discrepancy in the performances between unequal and equal weighting may be partly because only a simple hydrological model is used. In other words, the remaining differences among corrected climate simulations may not be well presented in streamflow simulations when a lumped hydrological model is used in such large watersheds. All the analyses above have been presented in the Discussion section of revised manuscript [P14, L23-P15, L2].

In addition, the daily bias correction (DBC) method used in this study has been applied in many recent studies (e.g., Chen et al., 2017; Li et al., 2019). It can be considered as a superior bias correction method in terms of correcting the bias of precipitation frequency and the bias in the distributions of precipitation amounts and temperature.

Responses to Referee #2's comments

We would like to thank the referee for the time taken in reviewing our paper. Please find the point-by-point responses below. We have made 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 for the professional summary of the work. All the comments and suggestions have been replied to below and 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 are 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). This point has been underscored in the revised manuscript [P2, L10-13 & P13, L12-14].

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. To be sure, we agree that these risks and problems of bias correction should still be addressed in the manuscript. Relevant corrections have been made in the revised manuscript [P9, L18-20 & P14, L7-8].

As suggested by the referee, we have also 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 have been presented in the detailed

responses to specific comments below.

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 has been fixed to "Thus, the equal weighting method may still be a viable and conservative choice when bias correction to GCM simulations is conducted in hydrological climate change impact studies" [P1, L23-25].

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 has been shortened to "The outlet of the Manicouagan-5 River is the Daniel-Johnson Dam" [P4, L9-10], and the data collection of the observed streamflow of Manicouagan-5 has been stated more clearly [P4, L20-21].

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. 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 directly based on streamflow series. In this way, weighting based on streamflow simulations can synthesize GCMs' performances in both temperature and precipitation and circumvent 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, L2-3. Herein, the used variable has been stated more clearly [P6, L20-21].

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

As stated in P9, L18-20, 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 most weighting methods still follow the same assumption as the bias correction and do not overcome the potential problem (especially for the performance-based weighting methods). 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 has been rephrased in the revised manuscript [P9, L18-23].

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). In model-as-truth experiments, the output of one climate model was regarded as the "truth" and the outputs of the remaining 28 climate models were used as simulations to this "truth" model. Then the weights were re-calculated for these remaining models. Since there is a "truth" result for the future period in this case, the performances of weighting methods can be evaluated in terms of reproducing the future "truth". Note that each climate model was regarded as truth in turn, and the results are gathered for each climate simulation playing as the truth.

Figure R1 shows the results of out-of-sample testing over the Xiangjiang watershed for biases of weighted multi-model mean hydrological indices, which are the same as those in Fig. 5. The left and right sides of each stick respectively represent the biases at the reference and future periods when one climate model is regarded as the truth. Similar to Fig. 5, the bias of weighted mean being closer to 0 means that the corresponding weighting method performs better. In general, the results of out-of-sample testing are similar to those of using historical observations. For the experiment of streamflows simulated by raw GCM outputs, Fig. R1a-c shows that unequally weighted means more or less become closer to the truth simulation than those of equal weighting in both reference and future periods. The unequal streamflow-based weights can help to reduce the biases. In particular, the three methods with the most differentiated weights (REA, UREA and CPI) reduce more biases of annual streamflow when compared with other methods (i.e., the ranges of the biases calculated by these three methods are narrower and closer to 0 when different simulations are used as the truth). In addition, although the biases in the future period tend to be larger than those in the reference period, the weighted means still have a slight improvement in most cases. However, for the experiment of using bias-corrected GCM outputs to simulate streamflows, as shown by the similar patterns among equal and unequal weighting methods (Fig. R1d-f), the unequally weighted multi-model means have similar biases to those of equal weighting in both reference and future periods. In addition, the results of out-of-sample testing over the Manicouagan-5 watershed are shown in Fig. R2, and generally, they are also similar to the results of using observations (Fig. 6).

All results and analyses above have been added as a separate sub-section in the section of Results of the revised manuscript [P11, L25-P12, L17 & Fig. 9-10].

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 to some other weighting methods, 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 various performances of different metrics may be due to the usage of weighting methods by end-users, which we failed to state clearly in the original version of the manuscript. Thus, this sentence has been corrected [P10, L9-11] and this point has been discussed in the Discussion section [P14, L12-18].

In addition, using different metrics may result in different performances of a weighting method. However, the main focus of this study is on 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, the discussion on the metric adopted for the weighting methods has been added in the revised version of manuscript [P14, L18-22].

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 simultaneously inputted to the impact model. 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 P1, L23-25, the equal weighting is only a conservative option for handling multi-model ensembles in hydrological impact studies.

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 has been fixed to "In this experiment, compared to the equal weighing method, unequal weighting methods do not bring about much disparateness to the results of hydrological impacts. It is still viable to attend to the bias-corrected ensembles with the equal weighting method" [P13, L24-27].

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, L34-P14, L3). 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 has been 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 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, L33-P14, L5].

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 are superior to model weighting methods 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 has been modified as follows [P14, L6-12].

Weights are generally assigned to climate simulations in a static way with an assumption that weights in the future period are the same as those in the reference 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 suggested using calibrated multiple metrics because it can improve the rationality of weighted multi-model mean (Knutti et al., 2017; Lorenz et al., 2018), while some argued 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 has been modified correspondingly to mention the use of multiple metrics and its potentials to strengthen weighted results [P14, L18-22].

Responses to Referee #3's comments

We would like to express gratitude to the referee for reviewing this paper and offering valuable suggestions. Please find the point-by-point responses below.

> *This study applies different combinations of bias-correction (BC) and model weighting (MW) to post-process climate and hydrological projections in two catchments. Both BC and MW are receiving sustained attention in the community, and so far only few studies combine both. What is important to stress, is that although the underpinnings of these two approaches are quite different, their aim is arguably quite similar: close the gap between simulations and observations. This leads me to comment on the two main findings of the study:*

We would like to thank the referee for the time taken in reviewing our manuscript and for the professional summary of the work. All comments have been replied to below and has been addressed in the revision.

> *Finding 1: "when using raw GCM outputs with no bias correction, streamflow-based weights better represent the mean hydrograph and reduce the bias of annual streamflow" P1L19-20: in my view, this is a natural consequence of applying MW, and in a way, it means that MW is used to correct for/mitigate climate model biases.*

Thanks for the comment. We agree with the referee that MW is used to mitigate biases, but this is not the specific focus of this study and we failed to state the conclusion clear enough. Actually, in this sentence, we intended to emphasize the advantages of streamflow-based weights over the weights calculated using climate variables (i.e. temperature and precipitation in this study). As stated in P13, L4-9 when dealing with the raw GCM-simulated streamflows, biases in multi-model mean of annual streamflow are reduced more by the weights based on the impact variable (streamflow), comparing with the weights based on climate variables. Herein, we have modified the expression

of Finding 1 to make this point clearer [P1, L19-21].

> *Finding 2: "when applying bias correction to GCM simulations before driving the hydrological model, the climate simulations become rather close to the observed climate, so that compared to equal weighting, the streamflow-based weights do not bring significant differences in the multi-model ensemble mean" P1L21-23: my interpretation is that employing successively two techniques with the same purpose makes the second technique redundant. Reducing the biases in the climate simulations, and then applying MW, makes it extremely difficult for the MW to discriminate between good and poor models. I recognise that BC is applied to the climate simulations and MW to the hydrological simulations, but since all the climate simulations are run through the same hydrological model, calibrated presumably with the forcing dataset also used to perform the BC, the differences in the streamflow simulations are minimal (as shown in Figure 3c and especially 4c). This lack of differences explains why the different weighting methods lead to similar results under current climate (the simulations are almost the same, so how they are combined makes little difference).*

We agree with the referee that in this study, MW loses the ability to discriminate the performances of climate simulations after the bias correction. This is also a finding of this study, which was mentioned in P13, L14-18. We have modified this sentence to make this point clearer.

In fact, MW is not designed for dealing with hydrological simulations but a necessary process to handle the ensemble of multiple climate simulations. Even after bias correction, there still exist some differences between climate simulations. In order to obtain evaluation of climate change impacts, it is unavoidable to choose a MW method to synthesize the simulation results from the ensemble (whether or not bias correction is done). Thus, MW is an indispensable process. Actually, both BC and MW are common procedures in regional impact studies. Although it is common to use equal weighting for bias-corrected ensembles, whether unequal weighting is a better choice remains to be investigated (Alder and Hostetler, 2019). The results of this study show that when the bias correction is done in impact studies, unequal weighting does not bring much difference to the impact evaluation. This supports the usage of equal weighting for bias-corrected ensembles. Nonetheless, we still think that with further development of weighting methods (e.g., more aggressive or multi-objective weighting methods), unequal weighting maybe can help to bring different or more reasonable consequences. The discussion on the weighting methods for the bias-corrected ensembles has been modified in the revised manuscript [P13, L10-27].

> *Overall, I suggest shifting the focus from current climatic conditions (for which no climate model and hence MW or BC is necessary) to future conditions (which rely on climate model simulations, which may need BC/MW). In my view, the focus is currently too much on the current conditions. For instance, in the abstract, the authors write "when applying bias correction to GCM simulations before driving the hydrological model, the climate simulations*

become rather close to the observed climate". This is true because of the nature of bias-correction, and was shown in previous studies (e.g., Hakala et al., 2018). What the grey area in Figures 3d and 4d tells us, however, is that under future conditions, there is substantial spread among the hydrological simulations, although the driving GCM simulations have been bias-corrected (likely because of the different sensitivities of the climate models).

Thanks for the comment. We agree with the referee that more attention should be paid to the future projections. In the revised manuscript, future simulations are only evaluated in the form of uncertainty (Section 4.4), since there is no observation in the future period to be compared with. In order to partly overcome this problem, we have added the out-of-sample testing in the revised manuscript following the suggestion of referee #2. In out-of-sample testing, the output of each climate model was regarded as the "truth" in turn and the outputs of the remaining 28 climate models were used as simulations to this "truth" model. Then the weights were re-calculated for these remaining models. Since there is a "truth" result for the future period in this case, the performances of weighting methods can be evaluated in terms of reproducing the future "truth".

Figure R1 shows the biases of weighted multi-model mean indices over the Xiangjiang watershed in the out-of-sample testing. The left and right sides of each stick respectively represent the bias at the reference and future periods when one climate model is regarded as truth. **In general, the results of out-of-sample testing are similar to the results using historical observations** (Fig.5). For the streamflow simulated by raw GCM outputs, Fig.R1a-c shows that unequally weighted means more or less become closer to the truth simulation than those of equal weighting in the reference period. The unequal weighting methods can help to reduce the biases in the reference period. Among seven unequal weighting methods, the three methods with the most differentiated weights (REA, UREA and CPI) reduce more biases. In addition, although the biases in the future period tend to be larger than those in the reference period, the weighted means still have a slight improvement in most cases. For the streamflow simulated by bias-corrected GCM outputs (Fig.R1d-f), the multi-model means generalized by unequal weights have similar biases to those of equal weighting. The results of out-of-sample testing over the Manicouagan-5 watershed are shown in Fig.R2, and generally, they are also similar to the results of using observations (Fig.6). The detailed results and analyses of out-of-sample testing has been added in the revised manuscript [P11, L25-P12, L18].

In addition, it is true that when using bias-corrected GCM outputs to simulate streamflows, the differences between ensemble members have been greatly reduced during the reference period while there are still considerable differences in the future period (which had been mentioned in P9, L15-16). This may be because the bias of climate models is nonstationary (Hui et al., 2019; Chen et al., 2015). However, the sentence in the abstract is only an explanation to the results of Finding 2 instead of a focus of this study, but we failed to state this logic clear enough. Therefore, this sentence has been modified to make the focus of this study clearer [P1, L21-23].

Is there any way to apply MW based on these projected changes, and not based on the streamflow simulations under current climate? In other words, are some of these projections more reliable than others and/or are some projections interdependent, and should be downweighted?

We thank the referee for this suggestion. Actually, the REA method in this study includes projected future changes when assigning weights. The REA considers both the similarity of a climate simulation to the observation in the reference period and its convergence to the weighted multi-model mean in the future period. Although the weights calculated by the REA method are most differentiated for the bias-corrected ensemble (as Fig. 2 shows), they still bring little impacts on the final results of the multi-model mean. In addition, the PI method considers independency among climate simulations when determining weights, but it only relies on reference values which have been tuned by the bias-correction method. The ability of independent criterion may fail because of the bias correction. This point has been discussed in the revised manuscript [P13, L20-27].

In summary, my impression is that Finding 1 is relevant but quite foreseeable. I think that Finding 2 is to a great extent due to the experimental design, in particular to the decision to apply BC and MW successively. I encourage the authors to rethink how to best combine MW and BC, for instance by using different periods and/or criteria for the MW.

We appreciate the comments from the referee. As presented in the last response, the out-of-sample testing has been added in the discussion as a complement. In addition, we have improved our expression that the main focus of this study is to investigate the influences of MW methods on the evaluation of climate change impacts (when the bias correction is or is not done), and to study whether the weighting determined based on the impact variable (streamflow) can induce more reasonable results [P3, L21-28]. This investigation is necessary because MW is a procedure to generalize the results of ensembles and the best way to do it remains questionable. This explanation to the usage of MW and BC has been discussed more clearly in the Discussion section [P13, L10-14].

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., and Lucas-Picher, P.: Assessing the limits of bias-correcting climate model outputs for climate change impact studies, Journal of Geophysical Research: Atmospheres, 120, 1123-1136, https://doi.org/10.1002/2014jd022635, 2015.

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.
Hui, Y., Chen, J., Xu, C. Y., Xiong, L., and Chen, H.: Bias nonstationarity of global climate model outputs: The role of internal climate variability and climate model sensitivity, International Journal of Climatology, 39, 2278-2294, https://doi.org/10.1002/joc.5950, 2019.
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.
Li, L., Shen, M., Hou, Y., Xu, C.-Y., Lutz, A. F., Chen, J., Jain, S. K., Li, J., and Chen, H.: Twenty-first-century glacio-hydrological changes in the Himalayan headwater Beas River basin, Hydrology and Earth System Sciences, 23, 1483-1503, https://doi.org/10.5194/hess-23-1483-2019, 2019.
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.

Figures

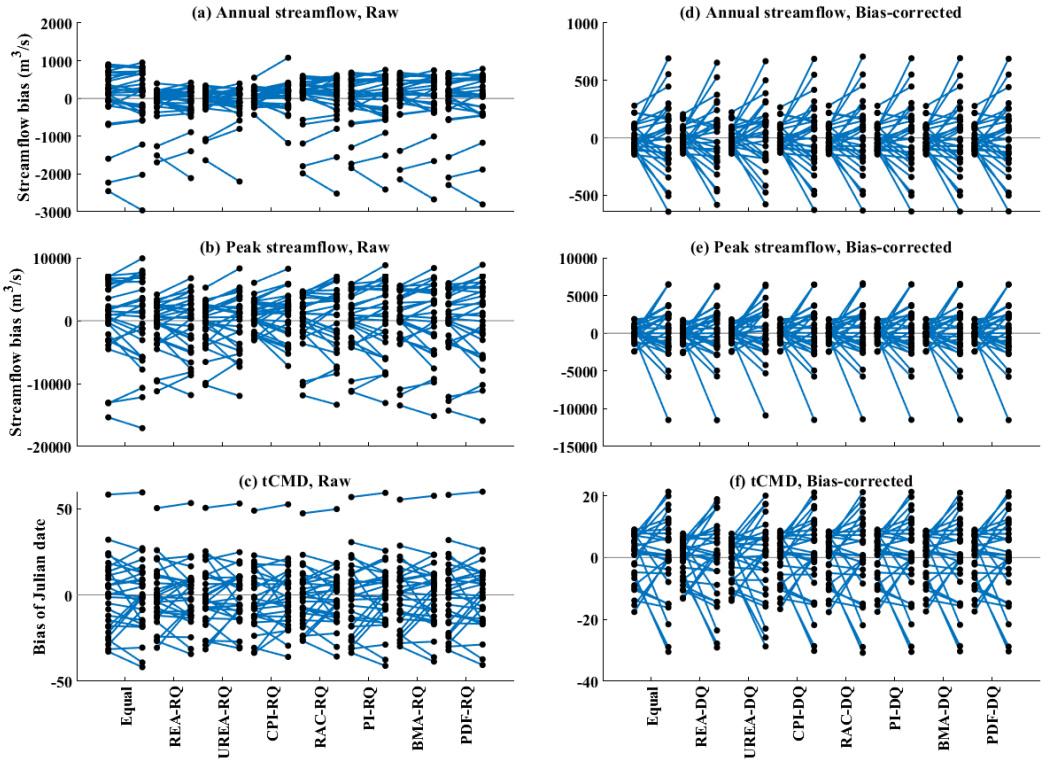

Figure R1. Bias in mean annual streamflow, mean peak streamflow and mean center of timing of annual flow (tCMD) of weighted multi-model ensemble mean in the out-of-sample testing over the Xiangjiang watershed. Twenty-nine sticks of each weighting method represent the results when each of 29 climate models was regarded as the "truth" in turn, and the left and right points in each stick represent the bias for the reference and future periods, respectively.

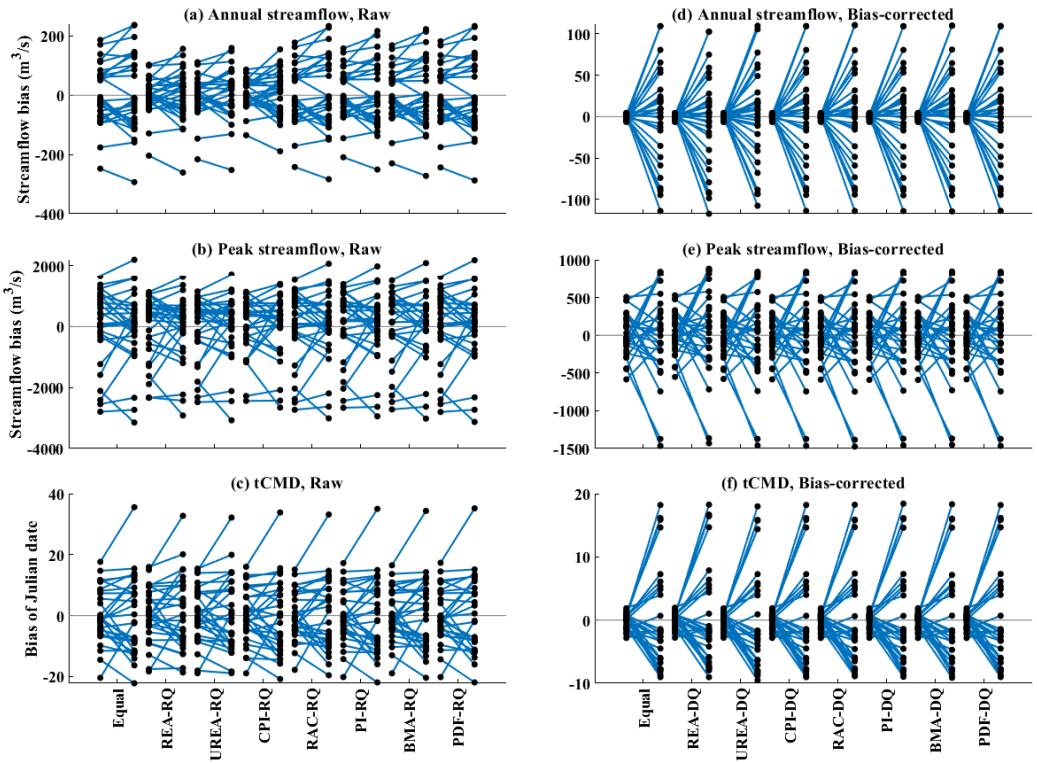

Figure R2. The same as the Fig. R1 but for the Manicouagan-5 watershed.

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

Hui-Min Wang[1], Jie Chen[1*], Chong-Yu Xu[1,2], Hua Chen[1], Shenglian Guo[1], Ping Xie[1], Xiangquan Li[1]

[1]State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, Wuhan, 430072, China
[2]Department of Geosciences, University of Oslo, Oslo, Norway

Correspondence to: Jie Chen (jiechen@whu.edu.cn)

Abstract: With the increase in the number of available global climate models (GCMs), pragmatic questions come up when using them to quantify the climate change impacts on hydrology: Is it necessary to weight GCM outputs in the impact studies, and if so, how to weight them? Some weighting methods have been proposed based on the performances of GCM simulations with respect to reproducing the observed climate. However, the process from climate variables to hydrological responses is nonlinear, and thus the assigned weights based on their performances in climate simulations may not be correctly translated to hydrological responses. Assigning weights to GCM outputs based on their ability to represent hydrological simulations is more straightforward. Accordingly, the present study assigns weights to GCM simulations based on their ability to reproduce hydrological characteristics and investigates their influence on the quantification of hydrological impacts. Specifically, eight weighting schemes are used to determine the weights of GCM simulations based on streamflow series simulated by a lumped hydrological model using raw or bias-corrected GCM outputs. The impacts of weighting GCM simulations are investigated in terms of reproducing the observed hydrological regimes for the reference period (1970-1999) and quantifying the uncertainty of hydrological changes for the future period (2070-2099). The results show that when using raw GCM outputs ~~with no bias correction~~, streamflow-based weights better represent the mean hydrograph and reduce ~~the bias~~more biases of annual streamflow than the weights calculated using climate variables. However, when applying bias correction to GCM simulations before driving the hydrological model, the streamflow-based unequal weights do not bring significant differences in the multi-model ensemble mean and uncertainty of hydrological impacts, since bias-corrected climate simulations become rather close to observations. ~~the climate simulations become rather close to the observed climate, so that compared to equal weighting, the streamflow-based weights do not bring significant differences in the multi-model ensemble mean and uncertainty of hydrological impacts. Since bias correction has been an indispensable procedure in hydrological impact studies~~ Thus, the equal weighting method may still be a viable and conservative choice when bias correction to GCM simulations is conducted in hydrological climate change impact studies. ~~for the studies of hydrological climate change impacts.~~

1 Introduction

Multi-model ensembles (MMEs) consisting of climate simulations from multiple global climate models (GCMs) have been widely used to quantify future climate change impacts and the corresponding uncertainty (Wilby and Harris, 2006; IPCC, 2013; Chiew et al., 2009; Chen et al., 2011; Tebaldi and Knutti, 2007). The number of climate models has increased rapidly, resulting in the obviously growing size of MMEs. For example, the Coupled Model Inter-comparison Project Phase 5 (CMIP5) archive contains 61 GCMs from 28 modeling institutes, with some GCMs providing multiple simulations (Taylor et al., 2012). Due to the lack of consensus on the proper way to combine ~~the~~ simulations of a MME, the prevailing approach is ~~that of~~the model democracy ("one model one vote") for the sake of simplicity, where each member in an ensemble is considered to have equal ability to simulate historical and future climates. The model democracy method has been applied to many global and regional climate change impact studies (e.g., IPCC, 2014; Minville et al., 2008; Maurer, 2007). Although it has been reported that the equal average of a multi-model ensemble often outperforms any individual model in regards to the reproduction of the mean state of observed historical climate ~~(Gleckler et al., 2008; Reichler and Kim, 2008; Pincus et al., 2008), the use of unequal weights has been recommended in some studies (Xu et al., 2010; Giorgi and Mearns, 2002; Murphy et al., 2004).~~(Gleckler et al., 2008; Reichler and Kim, 2008), whether the equal weighting is a better strategy for hydrological impact studies remains to be investigated (Alder and Hostetler, 2019).

Several studies have raised concerns about the strategy of model democracy, due to the following two reasons (Lorenz et al., 2018; Knutti et al., 2017; Cheng and AghaKouchak, 2015). First, GCM simulations in an ensemble do not have identical skills at representing historical climate observations. They may perform differently in simulating future climate. GCM performances may also vary by their variables and locations (Hidalgo and Alfaro, 2015; Abramowitz et al., 2019), which further challenges the rationality of model democracy in regional impact studies. Second, equal weights imply that the individual members in an ensemble are independent of each other. However, some climate models share common modules, parts of codes, parameterizations and so on (Knutti et al., 2010; Sanderson et al., 2017). Some pairs of GCMs submitted to the CMIP5 database only differ in the spatial resolution (e.g. MPI-ESM-MR and MPI-ESM-LR; see Giorgetta et al., 2013). The replication or overlapping in these GCMs may lead to the inter-dependence of MMEs, resulting in common biases towards the replicating section and inflating confidence in the projection uncertainty (Sanderson et al., 2015; Jun et al., 2012).

With the intention of improving climate projections and reducing the uncertainty, some weighting approaches have been proposed to assign unequal weights to climate model simulations according to their performances with respect to reproducing some diagnostic metrics of historical climate observations (Murphy et al., 2004; Sanderson et al., 2017; Cheng and AghaKouchak, 2015). For example, Xu et al. (2010) apportioned weights for GCMs based on their biases to the observed data in terms of two diagnostic metrics (climatological mean and inter-annual variability) for producing probabilistic climate projections. Lorenz and Jacob (2010) used errors in the trends of temperature to evaluate climate projections and determine weights. Other criteria have also been introduced into model weighting as a complement to the performance criterion. Some

examples are the convergence of climate projections for a future period (Giorgi and Mearns, 2002) and the interdependence among climate models (Sanderson et al., 2017).

Despite the different diagnostic metrics or definitions of model performances employed in these weighting methods, weights are commonly determined with respect to the ability of climate simulations at reproducing observed climate variables, such as temperature and precipitation (e.g., Chen et al., 2017; Wilby and Harris, 2006; Xu et al., 2010). However, for the impact studies, the relationship between climate variables and the impact variable is often not straightforward or explicit. In other words, the process from climate variables to their impacts may not be linear (Wang et al., 2018; Risbey and Entekhabi, 1996; Whitfield and Cannon, 2000). For example, Mpelasoka and Chiew (2009) reported that in Australia, a small change in annual precipitation can result in a several-times change in annual runoff. Thus, the weights calculated in the climate world may not be effective in the impact field.

In addition, a number of climate variables may determine the climate change impacts on a single environmental sector. For example, the runoff generation in a watershed is usually determined by precipitation, temperature, and other climate variables. Thus, it is not an easy task to determine the relative importance of each climate variable in impact studies, which is the other challenge to combining sets of weights based on different climate variables into a single set of weights for impact simulations. Previous studies have usually assumed that all variables are equally important and had an equal weight assigned to each climate variable (Xu et al., 2010; Chen et al., 2017; Zhao, 2015). However, these climate variables are usually not equally important in the impact field. For example, precipitation may be more important than temperature for a rainfall-dominated watershed, but could be different for a snowfall-dominated watershed. Thus, it may be more straightforward to calculate the weights for GCMs based on their ability to reproduce the single impact variable instead of multiple climate variables. Such a method would integrate the synthetic ability of GCMs in terms of simulating multiple climate variables to that of one impact variable. In addition, this method could also circumvent the previous problem of potential nonlinearity between climate variables and the impact variable.

Accordingly, the objectives of this study are to assign weights to GCM simulations ~~indirectly~~ according to their ability to represent hydrological observations, and to assess the impacts of these weighting methods on the assessment of hydrological responses to climate change. The case study was conducted over two watersheds with different climatic and hydrological characteristics. Since both bias correction and model weighting are common procedures in regional and local impact studies, this study considers two experiments (raw and bias-corrected GCM outputs) to simulate streamflows and investigate the performances of weighting methods. Seven weighting methods were used to assign unequal weights ~~based on streamflow series~~for streamflows simulated ~~using the~~by raw ~~GCM outputs~~ or bias-corrected ~~outputs.~~GCMs, respectively. The impacts of unequal weights are then assessed and compared to the equal weighting method in terms of multi-model ensemble mean and uncertainty related to the choice of a climate model.

2 Study area and data

2.1 Study Area

This study was conducted over two watersheds with different climate and hydrological characteristics: the rainfall-dominated Xiangjiang watershed and the snowfall-dominated Manicouagan-5 watershed (Figure 1). The Xiangjiang River is one of the largest tributaries of the Yangtze River in central-southern China, and its drainage area is about 94 660 km^2 (Figure 1a). A catchment with a surface area of about 52 150 km^2 above the Hengyang gauged station was used in this study. The catchment is heavily influenced by the East Asian Monsoon, which causes a humid subtropical climate with hot and wet summers and mild winters. The average temperature over the catchment is about 17 °C with the coldest month averaging about 7 °C. The average annual precipitation is about 1570 mm, of which 61% falls in the wet season from April to August. The daily averaged streamflow at the Hengyang gauged station is around 1400 m^3s^{-1}. The annual average of summer peak streamflow is about 4420 m^3s^{-1}, mainly due to summer extreme rainfalls.

The Manicouagan-5 watershed, located in the center of the Province of Quebec, Canada, is the largest sub-basin of the Manicouagan watershed (Figure 1b). Its drainage area is about 24 610 km^2, most of which is covered by forest. ~~At the~~The outlet of the Manicouagan-5 River is the Daniel-Johnson Dam~~, the longest multiple arch buttress dam in the world~~. The Manicouagan-5 watershed has a continental subarctic climate characterized by long and cold winters. The average temperature over the watershed is about -3 °C, with nearly half of the year having a daily temperature below 0 °C. The average annual precipitation is about 912 mm, evenly distributed over each year. The average discharge at the outlet of the Manicouagan-5 River is about 530 m^3s^{-1}. Snowmelt contributes to the peak discharge during May, whose annual average is about 2200 m^3s^{-1}.

2.2 Data

This study used daily maximum and minimum temperatures and precipitation from observation and GCM simulations for both watersheds. The observed meteorological data for the Xiangjiang watershed were collected from 97 precipitation gauges and 8 temperature gauges. Streamflow series were collected from the Hengyang gauged station. For the Manicougan-5 watershed, the observed meteorological data were extracted from the gridded dataset of Hutchinson et al. (2009), which is interpolated from daily station data using a thin-plate smoothing spline interpolation algorithm. ~~The hydrological data were collected from the Daniel Johnson Dam using mass balance calculations.~~ Streamflow series were the inflows of the Daniel Johnson Dam, which were calculated using mass balance calculations. All the observation data for both watersheds cover the historical reference period (1970-1999).

For the climate simulations, maximum and minimum temperatures and precipitation of 29 GCMs were extracted from the CMIP5 archive over both watersheds (Table 1). All simulations cover both the historical reference period (1970-1999) and the future projection period (2070-2099). One Representative Concentration Pathway (RCP8.5) was used in terms of climate projections in the future period. RCP8.5 was selected because it projects the most severe increase in greenhouse gas emissions among the four RCPs, and it is often used to design conservative mitigation and adaptation strategies (IPCC, 2014).

3 Methodology

To begin the process of calculating the weight for each GCM simulation, a multi-model ensemble constructed by 29 CMIP5 GCMs was utilized to drive a calibrated hydrological model over the two watersheds. Two experiments were designed to generate the ensembles of streamflow simulations. The first experiment drives the hydrological model using raw GCM outputs with no bias correction, while the second drives the hydrological model using bias-corrected climate simulations. Although it is not common to use raw GCM simulations for hydrological impact studies, the rationale for using them in this study is to examine the impacts of bias correction on weighting GCMs. The bias correction may adjust the relative performances between climate simulations and thus affect the determination of the relative weight for each ensemble member. Based on the ensemble of hydrological simulations from GCM outputs, eight weighting methods were employed to determine the weights of each GCM and to combine ensemble members for the assessment of hydrological climate change impacts. More detailed information is given below.

3.1 Bias correction

Since the raw outputs of GCMs are often too coarse and biased to be directly input into hydrological models for impact studies, bias correction is commonly applied to GCM outputs prior to the runoff simulation (Wilby and Harris, 2006; Chen et al., 2011; Minville et al., 2008). A distribution-based bias correction method, the daily bias correction (DBC) method of Chen et al. (2013), was used in this study. DBC is the combination of the local intensity scaling (LOCI) method (Schmidli et al., 2006) and the daily translation (DT) method (Mpelasoka and Chiew, 2009). The LOCI method was used to adjust the wet-day frequency of climate model simulated precipitation. A threshold was determined for the reference period to ~~insure~~ensure that the simulated precipitation occurrence is identical to the observed precipitation occurrence. The same threshold was then used to correct the wet-day frequency for the future period. The DT method was used to correct biases in the frequency distribution of simulated precipitation amounts and temperature. The frequency distribution was represented by 100 percentiles ranging from the 1^{st} to the 100^{th}, and the correction factors were calculated for each percentile. The same correction factors were then employed to correct the distributions for the future period. The use of distribution-based biases facilitates the use of different correction factors for different levels of precipitation. Some studies have shown the advantages of distribution-based bias correction over other correction methods in the assessment of hydrological impacts (Chen et al., 2013; Teutschbein and Seibert, 2012).

3.2 Runoff simulation

The runoff was simulated using a lumped conceptual hydrological model, GR4J-6, which couples a snow accumulation and melt module, CemaNeige, with a rainfall-runoff model, GR4J (Arsenault et al., 2015). The CemaNeige model divides the precipitation into liquid and solid according to the daily temperature range, and generates snowmelt depending on the thermal state and water equivalent of the snowpack (Valéry et al., 2014). CemaNeige has two free parameters: the melting rate and the

thermal state coefficient. The GR4J model consists of a production reservoir and a routing reservoir (Perrin et al., 2003). A portion of net rainfall (liquid precipitation with evaporation subtracted) goes into the production reservoir, whose leakage forms the effective rainfall when combined with the other proportion of net rainfall. The effective rainfall is then divided into two flow components. Ninety percent of the effective rainfall routes via a unit hydrograph and enters into the routing reservoir. The other 10% generates the direct flow through the other unit hydrograph. There is ~~a~~ groundwater exchange ~~between~~with neighbouring catchments in the direct flow and the outflow nonlinearly generated by the routing reservoir. Four free parameters in GR4J must be calibrated: the maximum capacity of the production reservoir, the groundwater exchange coefficient, the one-day-head maximum capacity of the routing reservoir and the time base of unit hydrograph.

The time periods of the observed data used for hydrological model calibration and validation are presented in Table 2. The shuffled complex evolution optimization algorithm (Duan et al., 1992) was employed to optimize the parameters of GR4J-6 for both watersheds. The optimized parameters were chosen to maximize the Nash-Sutcliffe Efficiency (NSE) criteria (Nash and Sutcliffe, 1970). The selected sets of parameters yield NSEs greater than 0.87 for both calibration and validation periods, indicating the reasonable performance of GR4J-6 and the high quality of the observed datasets for both watersheds.

3.3 Weighting Methods

Raw and bias-corrected climate simulations were input to the calibrated GR4J-6 model to generate raw and bias-corrected streamflow data series, respectively. Eight weighting methods were then employed to determine the weight of each hydrological simulation, including the equal weighting method (model democracy) and 7 unequal weighting methods. All of the unequal weighting methods are described in detail in the supplementary material so they are only briefly presented herein. Seven unequal weighting methods consist of two multiple criteria-based weighting methods and five performance-based weighting methods. The two multiple criteria-based weighting methods are the reliability ensemble averaging method (REA) and the performance and interdependence skill (PI). The REA method considers both the bias of a GCM to observation in the reference period (performance criterion) and its similarity to other GCMs in the future projection (convergence criterion) (Giorgi and Mearns, 2002). The PI method weights an ensemble member according to its bias to historical observation (performance criterion) and its distance to other ensemble members in the reference period (interdependence criterion) (Knutti et al., 2017; Sanderson et al., 2017). The biases and distances in the REA and PI methods were calculated based on the diagnostic metric of the climatological mean of streamflow.

The five performance-based weighting methods are the climate prediction index (CPI), upgraded reliability ensemble averaging (UREA), the skill score of the representation of the annual cycle (RAC), Bayesian model averaging (BMA), and the evaluation of the probability density function (PDF). All of these methods only consider the differences of climate simulations to historical observation, but they differ in the metrics or algorithms used to determine weights. The CPI assigns weights based on the biases in the climatological mean and assumes that the simulated climatological mean follows a Gaussian distribution (Murphy et al., 2004). UREA considers biases in both the climatological mean and the inter-annual variance to determine weights (Xu et al., 2010). Both the RAC and BMA calculate weights based on monthly series. The RAC defines a skill score

in simulating the annual cycle according to the relationship among the correlation coefficient, standard deviations and centered root-mean-square error (Taylor, 2001). BMA combines the results of multiple models through the Bayesian theory (Duan et al., 2007; Raftery et al., 2005; Min et al., 2007). The PDF determines weights according to the overlapping area of probability density function between daily simulations and observations (Perkins et al., 2007).

Using all eight methods, the weights were ~~respectively~~ calculated for each of streamflow data series simulated by raw GCM outputs and bias-corrected outputs. ~~For a comparison, the above methods were also used to calculate weights based on raw or bias-corrected temperature and precipitation series in terms of performances in simulating observed temperature and precipitation.~~ For a comparison, raw and bias-corrected temperature and precipitation series were also individually used to calculate climate-based weights using the above weighting methods.

3.4 Data Analysis

The extent of inequality of each set of weights was first investigated by the entropy of weights (Déqué and Somot, 2010). The entropy of weights reflects the extent of how a weighting method discriminates the relative reliability between GCM simulations. Next, in order to investigate the impacts of weighting GCM simulations for hydrological impact studies, unequal weights were used to combine the ensemble of hydrological simulations. The impacts of unequal weights were compared to the results obtained using the equal weighting method. The comparison focuses on three aspects: (1) the simulation of reference and future hydrological regimes; (2) the bias of the multi-model ensemble mean during the reference period; and (3) the uncertainty of changes in hydrological indices between future and reference periods.

Specifically, when using the entropy of weights (Eq. (1)), the entropy reaches a maximum value when the weights are equally distributed among ensemble members. A smaller entropy indicates a larger difference among the weights of ensemble members. Thus, the entropy reflects the extent of inequality for a set of weights:

$$E = -\sum_{i=1}^{N} w_i \ln w_i \quad (1)$$

where w_i is the weight assigned to the ith ensemble member, and N is the total number of ensemble members.

Since weighting methods are usually proposed to reduce biases in the ensemble of climate simulations, the multi-model ensemble means determined by these weights are then evaluated in terms of the representation of observation during the reference period. The multi-year averages of three hydrological indices were calculated for each streamflow simulation: (1) annual streamflow; (2) peak streamflow; and (3) the center of timing of annual flow (tCMD: the occurrence day of the midpoint of annual flow). The multi-model mean indices were then obtained based on the weights assigned to each simulation and compared to the indices of observation.

The influences of weighting on the uncertainty of hydrological impacts related to the choice of GCMs are investigated in terms of the changes in four hydrological indices between the reference and future periods: (1) mean annual streamflow; (2) mean streamflow during the high flow period; (3) mean streamflow during the low flow period; and (4) mean peak streamflow (the periods of high and low flow are shown in Table 2). The Monte-Carlo approach was introduced to sample the uncertainty

for unequally weighted ensembles (Wilby and Harris, 2006; Chen et al., 2017). The hydrological indices were randomly sampled one thousand times based on the calculated weights. For example, if a climate model simulation is assigned a weight of 0.2, the hydrological index simulated by that climate simulation has a probability of 20% to be chosen as the sample in each Monte-Carlo experiment.

4 Results

4.1 Weights of GCMs

Figure 2 presents the weights calculated based on the streamflow data series simulated by raw GCM outputs and bias-corrected outputs for 8 (one equal and 7 unequal) weighting methods over two watersheds. These results show the ability of different weighting methods to distinguish the performance or reliability of individual ensemble members. The entropy of weights was also calculated to quantify the extent of this disproportion for each set of weights (Table 3). Some weighting methods tend to aggressively discriminate the reliability of GCMs and assign differentiated weights to ensemble members, while other methods assign similar weights to each of them. Specifically, when calculating weights based on raw GCM-simulated streamflows, REA, UREA and CPI produce the weights that most radically discriminate ensemble members among all eight weighting methods for both watersheds. The RAC method generates less differentiated unequal weights, followed by the BMA and PI methods, but weights assigned by the PDF method closely resemble the equal weighting method. However, when calculating weights based on bias-corrected GCM-simulated streamflows, the inequality of weights is reduced, and all the unequal weighting methods receive a lower entropy of weights for both watersheds (Table 3). Most sets of these weights become similar to the equal weighting method, with the exception of REA and UREA for the Xiangjiang watershed, and REA for the Manicouagan-5 watershed (Fig. 2). This result was expected, as the bias correction method brings all GCM simulations to be close to the observations. The differences among GCM simulations become greatly reduced.

In addition, the weights based on the raw and bias-corrected temperature and precipitation time series of GCM simulations were also calculated and are shown in Fig. S1. For the weights based on the raw temperature and precipitation, REA, UREA and CPI still generate the most unequal weights among these weighting methods over both watersheds, as Table 3 indicates. Again, the weights become equalized when calculating weights based on bias-corrected temperature and precipitation.

4.2 Impacts on the hydrological regime

The weights determined by eight weighting methods were first utilized to combine GCM-simulated streamflow series. Figure 3 shows the weighted multi-model mean of monthly mean streamflow for the Xiangjiang watershed. The gray envelope represents the range of monthly mean streamflow simulated using 29 GCM simulations. At the reference period, streamflows simulated by raw GCMs cover a wide range (Figure 3a). However, the equal-weighted multi-model mean streamflow performs better than most of the streamflow series simulated by individual GCMs with respect to reproducing the observed streamflow;

even so, the equal-weighted ensemble mean still underestimates the streamflow before the peak (January – May) and overestimates it after the peak (June – September).

For the ensemble mean combined by unequal weights, the three weighting methods that generate highly differentiated weights (REA, UREA and CPI) outperform the equal weighting method with respect to reproducing the observed monthly mean streamflow. The BMA and RAC methods improve the performance of streamflow simulations before the peak at the cost of performance after the peak, while an opposite pattern is observed when using the PI method. The PDF method generates an ensemble mean of monthly mean streamflows almost identical to that of the equal weighting method. This is an expected result, as the PDF method assigns almost identical weights to all GCM simulations.

Weights calculated based on the raw temperature and precipitation of GCM outputs were also used to construct the ensemble mean of monthly mean streamflows (Fig. S2a,b). Particularly, the ensemble mean hydrographs combined using the REA, UREA and CPI methods largely deviate from the observation. Although REA, UREA and CPI generate highly differentiated weights when based on GCM raw temperatures, their generated ensemble mean streamflows are significantly inferior to that generated by equal weights (Fig. S2a). In addition, when using raw precipitation to calculate weights, the weighting methods perform worse than or similar to those calculated based on streamflow series (Fig. S2b). This reflects the advantage of weighting streamflow series in terms of reproducing the observed mean hydrograph.

The bias correction method can reduce the biases of precipitation and temperature in representing the mean monthly streamflow for the reference period, as indicated by the narrowed envelope (Figure 3c), although a small amount of uncertainty is still observed. The reduction of biases brings about similar weights for all GCM-simulated time series when using bias-corrected GCM-simulated streamflows. Thus, the multi-model ensemble means of monthly mean streamflow constructed by all unequal weighting method are very similar to those constructed by the equal weighting method, as shown in Figure 3c.

For the bias-corrected GCM-simulated streamflow at the future period (Figure 3d), a larger uncertainty related to the use of climate models is observed, as indicated by the wider envelope of the mean monthly streamflow. This may be because the bias of GCM outputs is non-stationary. All bias correction methods are based on a common assumption that the bias of climate model outputs is constant over time. ~~However, this assumption may not be true, due to natural climate variability and climate sensitivity (Hui et al., 2019). In addition, if the bias of climate model outputs is not stationary, it is unnecessary to use unequal weighting methods.~~ However, this assumption may not always be true because of natural climate variability and climate sensitivity to various forcings (Hui et al., 2019; Chen et al., 2015), and most weight methods still follow the same assumption. In other words, the bias non-stationarity implies that climate models differ in their ability to simulate ~~precipitation and temperatures~~the climate for the future period. The weights calculated ~~at~~in the reference period may not be applicable ~~at~~in the future period. The results of this study also proved this ~~conclusion~~, as all of the weighting methods project similar ensemble means of ~~mean~~ monthly mean streamflows for the future period.

Figure 4 presents the same information as Figure 3 but for the Manicouagan-5 watershed. Nearly half of the monthly mean streamflow time series simulated by raw GCM outputs have delayed peak (June) compared to the observed one (May) at the reference period, which leads to the delayed peak streamflow of the weighted multi-model mean streamflows for all weighting

methods (Figure 4a). Nonetheless, when using raw GCM-simulated streamflow series to calculate weights, the multi-model mean streamflows perform better than or similar to those simulated using GCM raw temperature and precipitation data (Fig. S2c). However, for the bias-corrected streamflow series, the uncertainty of monthly streamflows simulated by individual bias-corrected GCMs is largely reduced and the problem of delayed peak streamflow is corrected (Figure 4c). Similar to the case in the Xiangjiang watershed, all unequally weighted multi-model mean streamflows are identical to that of the equal weighting method. For the future period, although the uncertainty of single bias-corrected GCM-simulated streamflows increases (Figure 4d), there are still very little differences among the future multi-model mean streamflows combined by different weighting methods.

4.3 Bias in multi-model mean

In order to quantify the performance of weighting methods with respect to reproducing the multi-model ensemble mean, biases of the multi-model ensemble mean relative to corresponding observation were calculated for the reference period in terms of three hydrological indices (mean annual streamflow, mean peak streamflow and mean center of timing of annual flow; tCMD). A smaller bias represents a better performance. Figure 5 presents the biases of weighted multi-model mean indices over the Xiangjiang watershed. For the streamflows simulated using raw GCM outputs, the weighting methods show varied performance in terms of reproducing observed indices (Figure 5a-c). Except for the PI method, the unequal-weighted multi-model means more or less outperform the equal weighting method in terms of reducing biases in mean annual streamflow and mean center timing, while an opposite result is observed in mean peak streamflow. ~~This may because most weighting methods only consider the mean value (climatological mean or monthly mean series) when evaluating GCMs, and none of them include peak or extreme values.~~ This may be because only the mean value (climatological mean or monthly mean series) was used as the evaluation metric when determining weights, while peak or extreme values were not considered. Additionally, weights calculated based on the raw temperature and precipitation of GCM outputs were used to calculate multi-model mean indices for comparison (Fig. S3a-c). When using raw temperature series of GCMs to determine weights, they often bring about more biases in mean annual streamflow and tCMD. The weights based on raw precipitation show some superiority in reducing bias in mean peak streamflow. However, when using bias-corrected GCM-simulated streamflows to calculate weights (Figure 5d-f), the biases in multi-model mean indices are much less varied among different weighting methods. This is similar to the previous results of hydrological regimes.

For the case in the Manicouagan-5 watershed, twenty-five of the 29 streamflow series simulated by raw GCMs have larger mean annual streamflows and mean peak streamflows than those of the observations, and 26 series generate delayed tCMD. This leads to the overestimation of multi-model mean indices for all weighting methods (Figure 6a-c). Compared to the equal weighting method, all unequal weighting methods overcome this overestimation more or less. The three weighting methods that generate highly differentiated weights (REA, UREA and CPI) notably reduce biases for all three hydrological indices. For most weights calculated based on raw temperature and precipitation of GCM outputs (Fig. S3d-f), a certain improvement on mean indices was also observed (the only exception is raw precipitation-based PDF weights). Compared to weights calculated

using streamflow series, nearly all weights based on GCM-simulated streamflows reduce more biases than those based on temperature and precipitation. However, when using bias-corrected GCM-simulated streamflows (Fig. 6d-f), again, all weighting methods generate very similar mean indices to the equal weighting method, since the biases among different GCM-simulated streamflows have been largely reduced by the bias correction method.

4.4 Impacts on uncertainty

In addition to the multi-model ensemble mean, the impacts of weighting GCM simulations on uncertainty of hydrological responses also need to be assessed. Thus, this study also evaluated how unequal weighting methods affect the uncertainty of hydrological impacts related to the choice of GCMs. Figures 7 and 8 present the box plots of changes in 4 hydrological indices (mean annual streamflow, mean streamflow during the high/low flow periods and mean peak streamflow) between the reference and future periods. The box plots of the equal weighting method are depicted using 29 values simulated by each climate simulation, while the box plots of 7 unequal weighting methods are constructed using 1,000 values sampled by the Monte-Carlo approach based on assigned weights. For example, a simulation with 2-times the weight as another one will occur 2-times as often as that one in the 1,000 samples of Monte-Carlo experiments. While the 1,000 samples still only consist of the 29 values, the occurrence of each value reflects its possibility to be chosen and presents the uncertainty related to the choice of GCMs determined by assigned weights.

Figure 7 presents the uncertainty of hydrological changes for the Xiangjiang watershed. When using raw GCM-simulated streamflows (Figure 7a-d), depending on the weighting methods, unequal weights show the varying effects on the uncertainty. Both the PDF and PI methods suggest similar uncertainties to those of the equal weighting method for all four hydrological indices. The BMA and RAC methods generate slightly larger uncertainty for the change in mean annual streamflow and slightly smaller uncertainty of the change in low streamflow. The two weighting methods that generate the most differentiated weights (REA and UREA) largely reduce the uncertainty and increase the changes of the upper and lower probabilities for all four hydrological variables. The impacts of weights calculated based on raw GCM temperature and precipitation series were also analyzed (Fig. S4a-d). When calculating weights based on raw temperature, REA, UREA and CPI tend to aggressively reduce the uncertainty in mean high streamflow and peak streamflow. Precipitation-based weights show similar influences on uncertainty as weights based on streamflows. However, for the bias-corrected GCM-simulated streamflows (Figure 7e-h), the uncertainty of changes in the four hydrological indices is similar among all weighting methods.

Figure 8 presents the uncertainty of hydrological impacts in terms of four hydrological indices over the Manicouagan-5 watershed. For weights calculated using raw GCM-simulated streamflows (Figure 8a-d), only UREA clearly reduces the uncertainty for mean annual streamflow. The REA, UREA and CPI methods reduce the uncertainty for mean low streamflow and decrease its value of upper probability. There are few differences in the uncertainty of mean high streamflow and peak streamflow among all weighting methods. However, when using bias-corrected GCM-simulated streamflows (Figure 8e-h), again, the uncertainty of changes in all four hydrological indices is very similar among most of the weighting methods. Only CPI suggests slight increases in changes of the lower probability.

4.5 Out-of-sample Testing

In the above assessments for weighting methods except their impacts on uncertainty, the weighting methods are mostly evaluated in terms of their performances to simulate observations in the reference period. This kind of assessments has been referred to as "in-sample" testing (Herger et al., 2018). But the performances of weighting methods in the future period ("out-of-sample") may also need to be investigated. However, there is no observations to be compared with in the future period. Thus, an out-of-sample testing was then performed by conducting model-as-truth experiments (Herger et al., 2018; Abramowitz et al., 2019). In model-as-truth experiments, the output of each climate model was regarded as the "truth" in turn and the outputs of the remaining 28 climate models were used as simulations to this "truth" model. Then, the weights were re-calculated for these remaining models. Since there is a "truth" at the future period in this case, the performances of weighting methods can be evaluated in terms of reproducing the future "truth".

Figure 9 shows the results of out-of-sample testing over the Xiangjiang watershed for biases of weighted multi-model mean hydrological indices, which are the same as those in Fig. 5. The left and right sides of each stick respectively represent the biases at the reference and future periods when one climate model is regarded as the truth. Similar to Fig. 5, the bias of weighted mean being closer to 0 means that the corresponding weighting method performs better. In general, the results of out-of-sample testing are similar to those where historical observations are used. For the experiment of streamflows simulated by raw GCM outputs, Fig. 9a-c shows that unequally weighted means more or less become closer to the truth simulation than those of equal weighting for both reference and future periods. The unequal streamflow-based weights can help to reduce the biases. In particular, the three methods with the most differentiated weights (REA, UREA and CPI) reduce more biases of annual streamflow when compared with other methods, in that the ranges of the biases calculated by these three methods are narrower and closer to 0 when different simulations are used as the truth. In addition, although the biases in the future period tend to be larger than those in the reference period, the weighted means still have a slight improvement in most cases. However, for the experiment of using bias-corrected GCM outputs to simulate streamflows, as shown by the similar patterns among equal and unequal weighting methods (Fig. 9d-f), the unequally weighted multi-model means have similar biases to those of using equal weighting method at both reference and future periods. In addition, the results of out-of-sample testing over the Manicouagan-5 watershed are shown in Fig. 10, and generally, they are also similar to the results of using observations (Fig. 6).

5 Discussion

In addition to the equal weighting method, which is a normal strategy for handling multi-model ensembles, many studies have ~~also~~ proposed various unequal weighting methods for impact studies (e.g., Giorgi and Mearns, 2002; Sanderson et al., 2017; Xu et al., 2010; Min et al., 2007; Murphy et al., 2004). Most of these methods calculate weights based on the reliability of GCM simulations relative to observed climates, or at least adopt their reliability as one of their weighting criteria. In other words, the performances of GCM simulations are usually evaluated by comparing them to observed climate using certain

metrics. However, this method may have two problems. First, the trade-off between multiple climate variables related to the impact variable remains uncertain, which leads to difficulty in obtaining a single set of weights for impact studies. Second, the relationship between climate variables and the impact variable is often non-linear and not explicit, which may jeopardize the validity and reasonableness of climate-based weights in the impact studies. Some examples are the weights based on ~~raw GCM~~ temperature in the experiment of raw GCM-simulated streamflows in the Xiangjiang watershed, which lead to obviously biased multi-model mean hydrographs at the reference period. ~~However,~~But using the weights calculated based on raw GCM precipitation does not lead to such biases. This may be because the runoff generation in the Xiangjiang watershed is dominated more by rainfall than temperature. Therefore, weights calculated using temperature may not reflect a GCMs' reliability that is relevant to hydrological responses. On the contrary, for the snow-dominated Manicouagan-5 watershed, the snowmelt-driven spring flood is an important characteristic of its hydrological regime, and both temperature and precipitation conditions have large influences on this process. Thus, weights based on temperature and precipitation do not lead to obviously biased multi-model mean hydrographs~~.~~ in this case. Furthermore, over both watersheds, most weights calculated using raw GCM-simulated streamflows reduce more biases of the mean annual streamflow than those based on raw temperature and precipitation. This is as expected, because weights based on streamflows directly reflect how GCM simulations conform to the observed streamflow and are not affected by the non-linear relationship between climate variables and impact variables. ~~Generally~~Generally, in the experiment of simulating streamflows using raw GCMs, weights calculated based on streamflows not only circumvent the above two problems, they also bring about fewer biases in mean annual streamflow for the multi-model means.

Since bias correction methods are routinely applied to GCM outputs for hydrological impact assessments ~~of climate change~~, this study considered two experiments where raw and bias-corrected GCM-simulated streamflows were ~~used~~ separately used to determine weights. ~~The weights were correspondingly assigned to two types of streamflows and their impacts on hydrological responses to climate change were compared.~~ The performances of weighting methods are separately examined for the two experiments. Although the equal weighting is often used by default to combine bias-corrected ensembles in hydrological impact studies, whether unequal weighting is necessary still remains to be investigated (Alder and Hostetler, 2019). As shown in Figures 3 and 4, biases in the simulated mean monthly streamflows are greatly reduced for the reference period after bias correction. ~~This reduction in biases directly affects the determination of weights. When using bias-corrected GCM simulations, all of the weighting methods assign similar weights to ensemble members. Furthermore, two experiments revealed different performances of unequal weights in quantifying hydrological impacts. For the experiment with raw GCM-simulated streamflows, the impacts of unequal weighting vary with the choice of weighting methods. With bias-corrected GCM simulated streamflows, the results are totally different than in the first experiment. Not only are all the weighted multi-model means of monthly mean streamflows similar to those of the equal weighting method, the uncertainty of the hydrological impacts is also similar among all of the weighting methods. This is because performance related weighting methods assign similar weights to all simulations. Since bias correction has been an indispensable procedure for hydrological impact studies, and unequal weighting methods do not have a large influence on impact results, the model democracy approach is still~~

~~recommended in dealing with multi-model ensembles.~~ This change in biases affects the ability of most unequal weighting methods to discriminate the performances of climate simulations. In this experiment, all of the weighting methods assign similar weights to all simulations (as indicated by the decline of entropy of weights calculated by each weighting method). This is because climate simulations become rather close to each other in the reference period, and all weighting methods except REA in this study only rely on reference performances (which means that they lose the ability to discriminate the performances of climate simulations). As to the REA method, even though it considers future projections in its convergence criterion when calculating weights and its weights are still the most differentiated for the bias-corrected ensemble (as shown in Fig. 2), they bring little impacts on the final results of the multi-model mean. In addition, the PI method considers independency among simulations, but it only relies on reference values which have been tuned by the bias-correction method. The ability of independent criterion may be affected because of the bias correction. In general, in this experiment, compared to the equal weighing method, unequal weighting methods do not bring about much disparateness to the results of hydrological impacts. The out-of-sample testing also manifested the same phenomena. Therefore, it is still viable to attend to the bias-corrected ensembles with the equal weighting method.

Despite the choices of variables used to calculate weights, the establishment of any weighting method involves subjective choices of diagnostic metrics, its translation to performance measurement, and normalization to weights (Knutti et al., 2017; Santer et al., 2009). For example, in the RAC method, the correlation coefficient and standard deviation are used as diagnostic metrics, and GCM skills are measured through the translation of a fourth-order formulation. The skill scores are then divided by their sum to be normalized. Any of these steps can ultimately affect the property of a weighting method. For example, the REA, UREA and CPI methods are inclined to generate more differentiated weights, while other methods assign more similar weights to ensemble members. ~~However, all~~All of these aspects in weighting methods are often ~~arbitrary~~predefined without detailed examination or based on expert experience and, thus, can actually introduce several layers of subjective uncertainty. An improper weighting method may even cause a risk of reducing projection accuracy (Weigel et al., 2010), and extremely aggressive weighting may conceal the uncertainty rather than reduce it (Chen et al., 2017). Thus, notwithstanding the equal weighting is not a perfect solution, model weighting methods should be used with cautions and the results of equal weighting should be presented along with those of unequal weighting methods.

Moreover, some risks may exist in the usage of weighting methods in impact studies. Firstly, weights are generally assigned to climate simulations in a static way (i.e. weights in the ~~reference~~future period are the same as those in the ~~future~~reference period). ~~This approach is based on the assumption that the performances of GCM simulations are stable and stationary.~~ 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 (Weigel et al., 2010; Miao et al., 2016), and models with better performance in the reference period do not necessarily provide more realistic signals of climate change (Reifen and Toumi, 2009; Knutti et al., 2010). ~~Therefore, the assumption of stationary GCM performances may be questionable. Secondly, performance measurement in most weighting methods only depends on one diagnostic metric, such as the long-term mean state (e.g., 30-year climatological mean in REA, PI, UREA, and CPI methods).~~The way to deal

with the dynamic reliability of climate models deserves further studies. Secondly, many researchers and end-users in hydrological impacts only consider one diagnostic metric to determine weights, such as the climatological mean (e.g., Wilby and Harris, 2006; Chen et al., 2017). It is not clear whether reducing the bias of one specific metric can transfer to other metrics. The weights calculated using the raw GCM-simulated streamflows in the Xiangjiang watershed are one negative example, where the bias in mean annual streamflow is reduced while the bias in the mean peak streamflow is enlarged. Some studies have also shown similar problems (Jun et al., 2012; Santer et al., 2009). For example, Jun et al. (2012) demonstrated that there is little relationship between a GCMs' ability to reproduce mean temperature state and trend of temperature. ~~Overall, notwithstanding the potential gains obtained by weights calculated using streamflows, the equal weighting method remains an easily available and conservative way to handle the ensemble of multiple climate models for hydrological impact studies~~Actually, a set of metrics can be introduced to determine weights (e.g., Sanderson et al., 2017). Some studies suggested using calibrated multiple metrics because it can improve the rationality of weighted multi-model mean (Knutti et al., 2017; Lorenz et al., 2018), while some argued that multiple metrics form another level of uncertainty within weighting methods (Christensen et al., 2010). Thus, the best way to choose proper metrics and synthesize performances in multiple metrics still remains in doubt and deserves further research.

There is a limitation in the hydrological modeling in this study. Only large watersheds were considered, as well as a lumped hydrological model. When using a lumped model, the nonlinear relationship between the climate variables and the impact variable (streamflow) may not be sufficiently revealed. Spatial differences between different climate simulations only affect the basin-averaged inputs to the hydrological model but not directly affect the process of runoff generation and streamflow routing (Lebel et al., 1987). Temporal variations of climate simulations may be partially reduced by the lumped hydrological model as well. With the help of other more sophisticated hydrological models (such as distributed models), the differences between climate-based weights and streamflow-based weights may become more obvious. For the experiment of raw GCM-simulated streamflows, the weights based on streamflow perform better than those based on climate variables. This may be related to large differences among climate simulations. But in the experiment of streamflows simulated using bias-corrected GCM outputs, that no much discrepancy is seen in the performances between unequal and equal weighting may be partly because only a simple hydrological model is used. In other words, the remaining differences among corrected climate simulations may not be well presented in streamflow simulations when a lumped hydrological model is used in such large watersheds.

6 Conclusion

In order to weight climate models based on ~~runoff simulation~~the impact variable and to quantify its ~~influence~~influences on ~~hydrological~~the impact assessment, this study assigns weights to an ensemble of 29 CMIP5 GCMs ~~were weighted by~~ over two watersheds through a group of weighting methods based on ~~the their~~ GCM-simulated ~~streamflows~~streamflow time series. ~~Raw~~Streamflow series are simulated by separately inputting the raw and bias-corrected GCM simulations ~~were used~~ to ~~drive~~

hydrological models ~~to obtain hydrological simulations~~. Using streamflows to determine weights is straightforward and can avoid the difficulty of combining weights based on multiple climate variables for impact studies. The influences of these unequal weights on the assessment of hydrological impacts were then investigated and compared to the common strategy of model democracy.

This study concludes that for the streamflows simulated using raw GCM outputs without bias correction, using unequal weights has some advantages over the equal weighting method in simulating observed hydrographs and ~~in~~ reducing the biases of multi-model means in mean annual streamflow. In particular, the weights calculated based on streamflows can reduce more biases of multi-model mean annual streamflow and better reproduce observed hydrographs, compared with the weights calculated based on climate variables. However, when using bias-corrected GCM outputs to simulate streamflow, GCM simulations were brought close to the observations by the bias correction method. ~~The~~Consequently, the weights assigned to climate simulations ~~consequently~~ become similar to each other, resulting in similar multi-model means ~~or~~and uncertainty of hydrological impacts for all ~~of the~~ unequal weighting methods. Therefore, the equal weighting method is still a conservative and viable option for combining the bias-corrected multi-model ensembles. Or, if an unequal weighting method is applied, it is better to present it with a detailed explanation of the weighting procedure, as well as the results of using equal weighting method to end-users.

~~Since bias-correction or downscaling has been an indispensable procedure when assessing climate change impacts on hydrology, the equal weighting method is still recommended, or at least, the equal weighting results should be provided to end users along with the unequal weighting results, as well as a detailed explanation of the weighting procedure.~~

Data availability

The climate simulation data can be accessed from the CMIP5 archive (https://esgf-node.llnl.gov/projects/esgf-llnl/, last access: 3 June 2019). The observation data in the Xiangjiang and Manicouagan-5 are not publicly available due to the restrictions of data providers, but can be requested by contacting the corresponding author.

Author contributions

JC conceived the original idea, and HMW and JC designed the methodology. JC and HC collected the data. HMW developed the model code and performed the simulations, with some contributions from XL. HMW, JC, CYX, SG and PX contributed to the interpretation of results. HMW wrote the paper, and JC, CYX, SG and PX revised the paper.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (Grant No. 51779176, 51539009, 91547205), the Overseas Expertise Introduction Project for Discipline Innovation (111 Project) funded by Ministry of Education and State Administration of Foreign Experts Affairs P.R. China (Grant No. B18037), the Thousand Youth Talents Plan from the Organization Department of CCP Central Committee (Wuhan University, China) and the Research Council of Norway (FRINATEK Project 274310). The authors would like to acknowledge the World Climate Research Program Working Group on Coupled Modelling, and all climate modeling institutions listed in Table 1 for making GCM outputs available. We also thank Hydro-Québec and the Changjiang Water Resources Commission for providing observation data in the Manicouagan-5 and Xiangjiang watersheds, respectively.

References

Abramowitz, G., Herger, N., Gutmann, E., Hammerling, D., Knutti, R., Leduc, M., Lorenz, R., Pincus, R., and Schmidt, G. A.: ESD Reviews: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing, Earth System Dynamics, 10, 91-105, https://doi.org/10.5194/esd-10-91-2019, 2019.

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.

Arsenault, R., Gatien, P., Renaud, B., Brissette, F., and Martel, J.-L.: A comparative analysis of 9 multi-model averaging approaches in hydrological continuous streamflow simulation, Journal of Hydrology, 529, 754-767, https://doi.org/10.1016/j.jhydrol.2015.09.001, 2015.

Chen, J., Brissette, F. P., Poulin, A., and Leconte, R.: Overall uncertainty study of the hydrological impacts of climate change for a Canadian watershed, Water Resources Research, 47, W12509, https://doi.org/10.1029/2011wr010602, 2011.

Chen, J., Brissette, F. P., Chaumont, D., and Braun, M.: Performance and uncertainty evaluation of empirical downscaling methods in quantifying the climate change impacts on hydrology over two North American river basins, Journal of Hydrology, 479, 200-214, https://doi.org/10.1016/j.jhydrol.2012.11.062, 2013.

Chen, J., Brissette, F. P., and Lucas-Picher, P.: Assessing the limits of bias-correcting climate model outputs for climate change impact studies, Journal of Geophysical Research: Atmospheres, 120, 1123-1136, https://doi.org/10.1002/2014jd022635, 2015.

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.

Cheng, L., and AghaKouchak, A.: A methodology for deriving ensemble response from multimodel simulations, Journal of Hydrology, 522, 49-57, https://doi.org/10.1016/j.jhydrol.2014.12.025, 2015.

Chiew, F. H. S., Teng, J., Vaze, J., Post, D. A., Perraud, J. M., Kirono, D. G. C., and Viney, N. R.: Estimating climate change impact on runoff across southeast Australia: Method, results, and implications of the modeling method, Water Resources Research, 45, https://doi.org/10.1029/2008wr007338, 2009.

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.

Déqué, M., and Somot, S.: Weighted frequency distributions express modelling uncertainties in the ENSEMBLES regional climate experiments, Climate Research, 44, 195-209, https://doi.org/10.3354/cr00866, 2010.

Duan, Q., Sorooshian, S., and Gupta, V.: Effective and efficient global optimization for conceptual rainfall-runoff models, Water Resources Research, 28, 1015-1031, https://doi.org/10.1029/91WR02985, 1992.

Duan, Q., Ajami, N. K., Gao, X., and Sorooshian, S.: Multi-model ensemble hydrologic prediction using Bayesian model averaging, Advances in Water Resources, 30, 1371-1386, https://doi.org/10.1016/j.advwatres.2006.11.014, 2007.

Giorgetta, M. A., Jungclaus, J., Reick, C. H., Legutke, S., Bader, J., Böttinger, M., Brovkin, V., Crueger, T., Esch, M., Fieg, K., Glushak, K., Gayler, V., Haak, H., Hollweg, H.-D., Ilyina, T., Kinne, S., Kornblueh, L., Matei, D., Mauritsen, T., Mikolajewicz, U., Mueller, W., Notz, D., Pithan, F., Raddatz, T., Rast, S., Redler, R., Roeckner, E., Schmidt, H., Schnur, R., Segschneider, J., Six, K. D., Stockhause, M., Timmreck, C., Wegner, J., Widmann, H., Wieners, K.-H., Claussen, M., Marotzke, J., and Stevens, B.: Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the Coupled Model Intercomparison Project phase 5, Journal of Advances in Modeling Earth Systems, 5, 572-597, https://doi.org/10.1002/jame.20038, 2013.

Giorgi, F., and Mearns, L. O.: Calculation of Average, Uncertainty Range, and Reliability of Regional Climate Changes from AOGCM Simulations via the "Reliability Ensemble Averaging" (REA) Method, Journal of Climate, 15, 1141-1158, https://doi.org/10.1175/1520-0442(2002)015<1141:coaura>2.0.co;2, 2002.

Gleckler, P. J., Taylor, K. E., and Doutriaux, C.: Performance metrics for climate models, Journal of Geophysical Research, 113, D06104, https://doi.org/10.1029/2007jd008972, 2008.

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.

Hidalgo, H. G., and Alfaro, E. J.: Skill of CMIP5 climate models in reproducing 20th century basic climate features in Central America, International Journal of Climatology, 35, 3397-3421, https://doi.org/10.1002/joc.4216, 2015.

Hui, Y., Chen, J., Xu, C. Y., Xiong, L., and Chen, H.: Bias nonstationarity of global climate model outputs: The role of internal climate variability and climate model sensitivity, International Journal of Climatology, 39, 2278-2294, https://doi.org/10.1002/joc.5950, 2019.

Hutchinson, M. F., McKenney, D. W., Lawrence, K., Pedlar, J. H., Hopkinson, R. F., Milewska, E., and Papadopol, P.: Development and Testing of Canada-Wide Interpolated Spatial Models of Daily Minimum–Maximum Temperature and Precipitation for 1961–2003, Journal of Applied Meteorology and Climatology, 48, 725-741, https://doi.org/10.1175/2008jamc1979.1, 2009.

IPCC: Evaluation of Climate Models, in: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M. M. B., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. M., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 741-866, 2013.

IPCC: Summary for Policymakers, in: Climate Change 2014 – Impacts, Adaptation and Vulnerability: Part A: Global and Sectoral Aspects: Working Group II Contribution to the IPCC Fifth Assessment Report, edited by: Barros, V. R., Field, C. B., Dokken, D. J., Mastrandrea, M. D., Mach, K. J., Bilir, T. E., Chatterjee, M., Ebi, K. L., Estrada, Y. O., Genova, R. C., Girma, B., Kissel, E. S., Levy, A. N., MacCracken, S., Mastrandrea, P. R., and White, L. L., Cambridge University Press, Cambridge, 1-32, 2014.

Jun, M., Knutti, R., and Nychka, D. W.: Spatial Analysis to Quantify Numerical Model Bias and Dependence, Journal of the American Statistical Association, 103, 934-947, https://doi.org/10.1198/016214507000001265, 2012.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., and Meehl, G. A.: Challenges in Combining Projections from Multiple Climate Models, Journal of Climate, 23, 2739-2758, https://doi.org/10.1175/2009jcli3361.1, 2010.

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.

Lebel, T., Bastin, G., Obled, C., and Creutin, J. D.: On the accuracy of areal rainfall estimation: A case study, Water Resources Research, 23, 2123-2134, https://doi.org/10.1029/WR023i011p02123, 1987.

Lorenz, P., and Jacob, D.: Validation of temperature trends in the ENSEMBLES regional climate model runs driven by ERA40, Climate Research, 44, 167-177, https://doi.org/10.3354/cr00973, 2010.

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.

Maurer, E. P.: Uncertainty in hydrologic impacts of climate change in the Sierra Nevada, California, under two emissions scenarios, Climatic Change, 82, 309-325, https://doi.org/10.1007/s10584-006-9180-9, 2007.

Miao, C., Su, L., Sun, Q., and Duan, Q.: A nonstationary bias-correction technique to remove bias in GCM simulations, Journal of Geophysical Research: Atmospheres, 121, 5718-5735, https://doi.org/10.1002/2015jd024159, 2016.

Min, S. K., Simonis, D., and Hense, A.: Probabilistic climate change predictions applying Bayesian model averaging, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 365, 2103-2116, https://doi.org/10.1098/rsta.2007.2070, 2007.

Minville, M., Brissette, F., and Leconte, R.: Uncertainty of the impact of climate change on the hydrology of a nordic watershed, Journal of Hydrology, 358, 70-83, https://doi.org/10.1016/j.jhydrol.2008.05.033, 2008.

Mpelasoka, F. S., and Chiew, F. H. S.: Influence of Rainfall Scenario Construction Methods on Runoff Projections, Journal of Hydrometeorology, 10, 1168-1183, https://doi.org/10.1175/2009jhm1045.1, 2009.

Murphy, J. M., Sexton, D. M., Barnett, D. N., Jones, G. S., Webb, M. J., Collins, M., and Stainforth, D. A.: Quantification of modelling uncertainties in a large ensemble of climate change simulations, Nature, 430, 768-772, https://doi.org/10.1038/nature02771, 2004.

Nash, J. E., and Sutcliffe, J. V.: River flow forecasting through conceptual models part I — A discussion of principles, Journal of Hydrology, 10, 282-290, https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

Perkins, S. E., Pitman, A. J., Holbrook, N. J., and McAneney, J.: Evaluation of the AR4 Climate Models' Simulated Daily Maximum Temperature, Minimum Temperature, and Precipitation over Australia Using Probability Density Functions, Journal of Climate, 20, 4356-4376, https://doi.org/10.1175/jcli4253.1, 2007.

Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation, Journal of Hydrology, 279, 275-289, https://doi.org/10.1016/s0022-1694(03)00225-7, 2003.

Pincus, R., Batstone, C. P., Hofmann, R. J. P., Taylor, K. E., and Glecker, P. J.: Evaluating the present-day simulation of clouds, precipitation, and radiation in climate models, Journal of Geophysical Research, 113, https://doi.org/10.1029/2007jd009334, 2008.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M.: Using Bayesian Model Averaging to Calibrate Forecast Ensembles, Monthly Weather Review, 133, 1155-1174, https://doi.org/10.1175/mwr2906.1, 2005.

Reichler, T., and Kim, J.: How Well Do Coupled Models Simulate Today's Climate?, Bulletin of the American Meteorological Society, 89, 303-312, https://doi.org/10.1175/bams-89-3-303, 2008.

Reifen, C., and Toumi, R.: Climate projections: Past performance no guarantee of future skill?, Geophysical Research Letters, 36, https://doi.org/10.1029/2009gl038082, 2009.

Risbey, J. S., and Entekhabi, D.: Observed Sacramento Basin streamflow response to precipitation and temperature changes and its relevance to climate impact studies, Journal of Hydrology, 184, 209-223, https://doi.org/10.1016/0022-1694(95)02984-2, 1996.

Sanderson, B. M., Knutti, R., and Caldwell, P.: A Representative Democracy to Reduce Interdependency in a Multimodel Ensemble, Journal of Climate, 28, 5171-5194, https://doi.org/10.1175/jcli-d-14-00362.1, 2015.

Sanderson, B. M., Wehner, M., and Knutti, R.: Skill and independence weighting for multi-model assessments, Geoscientific Model Development, 10, 2379-2395, https://doi.org/10.5194/gmd-10-2379-2017, 2017.

Santer, B. D., Taylor, K. E., Gleckler, P. J., Bonfils, C., Barnett, T. P., Pierce, D. W., Wigley, T. M., Mears, C., Wentz, F. J., Bruggemann, W., Gillett, N. P., Klein, S. A., Solomon, S., Stott, P. A., and Wehner, M. F.: Incorporating model quality information in climate change detection and attribution studies, Proceedings of The National Academy of Sciences of the United States of America, 106, 14778-14783, https://doi.org/10.1073/pnas.0901736106, 2009.

Schmidli, J., Frei, C., and Vidale, P. L.: Downscaling from GCM precipitation: a benchmark for dynamical and statistical downscaling methods, International Journal of Climatology, 26, 679-689, https://doi.org/10.1002/joc.1287, 2006.

Taylor, K. E.: Summarizing multiple aspects of model performance in a single diagram, Journal of Geophysical Research: Atmospheres, 106, 7183-7192, https://doi.org/10.1029/2000jd900719, 2001.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An Overview of CMIP5 and the Experiment Design, Bulletin of the American Meteorological Society, 93, 485-498, https://doi.org/10.1175/bams-d-11-00094.1, 2012.

Tebaldi, C., and Knutti, R.: The use of the multi-model ensemble in probabilistic climate projections, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 365, 2053-2075, https://doi.org/10.1098/rsta.2007.2076, 2007.

Teutschbein, C., and Seibert, J.: Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods, Journal of Hydrology, 456-457, 12-29, https://doi.org/10.1016/j.jhydrol.2012.05.052, 2012.

Valéry, A., Andréassian, V., and Perrin, C.: 'As simple as possible but not simpler': What is useful in a temperature-based snow-accounting routine? Part 2 – Sensitivity analysis of the Cemaneige snow accounting routine on 380 catchments, Journal of Hydrology, 517, 1176-1187, https://doi.org/10.1016/j.jhydrol.2014.04.058, 2014.

Wang, H.-M., Chen, J., Cannon, A. J., Xu, C.-Y., and Chen, H.: Transferability of climate simulation uncertainty to hydrological impacts, Hydrology and Earth System Sciences, 22, 3739-3759, https://doi.org/10.5194/hess-22-3739-2018, 2018.

Weigel, A. P., Knutti, R., Liniger, M. A., and Appenzeller, C.: Risks of Model Weighting in Multimodel Climate Projections, Journal of Climate, 23, 4175-4191, https://doi.org/10.1175/2010jcli3594.1, 2010.

Whitfield, P. H., and Cannon, A. J.: Recent Variations in Climate and Hydrology in Canada, Canadian Water Resources Journal, 25, 19-65, https://doi.org/10.4296/cwrj2501019, 2000.

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.

Xu, Y., Gao, X., and Giorgi, F.: Upgrades to the reliability ensemble averaging method for producing probabilistic climate-change projections, Climate Research, 41, 61-81, https://doi.org/10.3354/cr00835, 2010.

Zhao, Y.: Investigation of uncertainties in assessing climate change impacts on the hydrology of a Canadian river watershed, Thèse de doctorat électronique, École de technologie supérieure, Montréal, 2015.

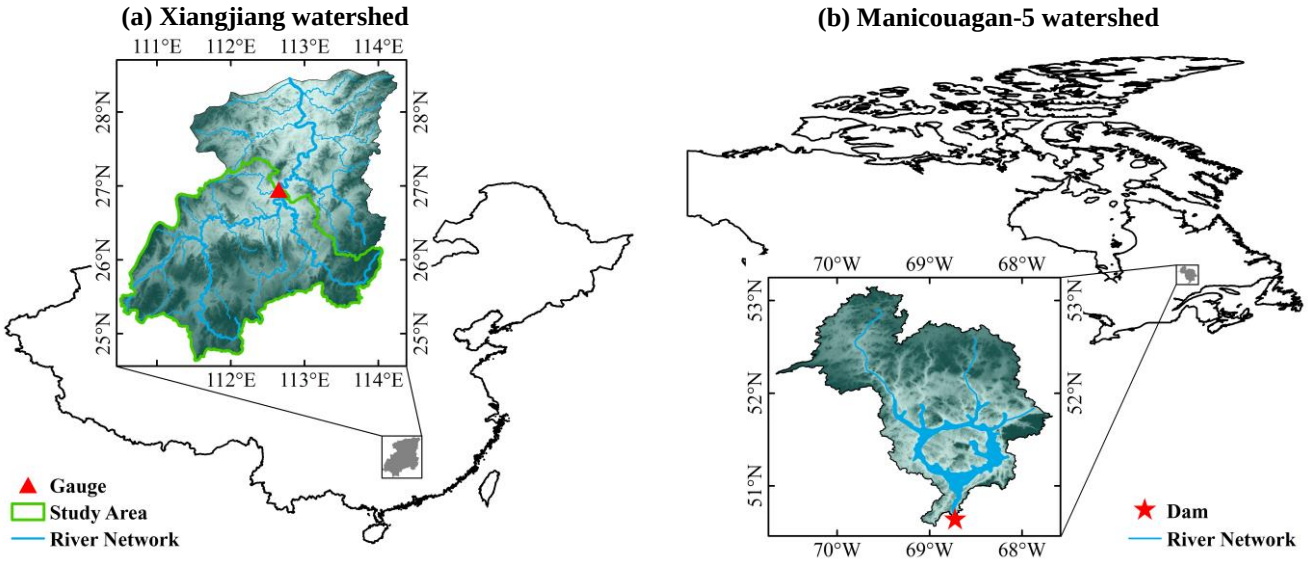

Figure 1. Locations of the (a) Xiangjiang and (b) Manicouagan-5 watersheds. (The study area in the Xiangjiang watershed is one of its sub-basins as the green boundary.)

Table 1. Information about the 29 GCMs used.

No.	Model name	Resolution (Lon. × Lat.)	Institution
1	ACCESS1.0	1.875 × 1.25	Commonwealth Scientific and Industrial Research Organization (CSIRO) and Bureau of Meteorology (BOM), Australia
2	ACCESS1.3	1.875 × 1.25	
3	BCC-CSM1.1	2.8 ×2.8	Beijing Climate Center, China Meteorological Administration
4	BCC-CSM1.1(m)	1.125 ×1.125	
5	BNU-ESM	2.8 × 2.8	College of Global Change and Earth System Science, Beijing Normal University
6	CanESM2	2.8 × 2.8	Canadian Centre for Climate Modelling and Analysis
7	CCSM4	1.25 × 0.94	US National Centre for Atmospheric Research
8	CESM1(CAM5)	1.25 × 0.94	National Science Foundation, Department of Energy, NCAR, USA
9	CMCC-CMS	1.875 × 1.875	Centro Euro-Mediterraneo per I Cambiamenti Climatici
10	CMCC-CM	0.75 × 0.75	
11	CMCC-CESM	3.75 × 3.7	
12	CNRM-CM5	1.4 × 1.4	Centre National de Recherches Météorologiques and Centre Européen de Recherche et Formation Avancée en Calcul Scientifique
13	CSIRO-Mk3.6.0	1.8 ×1.8	Commonwealth Scientific and Industrial Research Organization and Queensland Climate Change Centre of Excellence
14	FGOALS-g2	1.875 × 1.25	LASG, Institute of Atmospheric Physics, Chinese Academy of Sciences, and CESS, Tsinghua University
15	GFDL-CM3	2.5 × 2.0	NOAA Geophysical Fluid Dynamics Laboratory
16	GFDL-ESM2G	2.5 × 2.0	
17	GFDL-ESM2M	2.5 × 2.0	
18	INM-CM4	2.0 × 1.5	Russian Institute for Numerical Mathematics
19	IPSL-CM5A-LR	3.75 × 1.9	Institut Pierre Simon Laplace
20	IPSL-CM5A-MR	2.5 × 1.25	
21	IPSL-CM5B-LR	3.75 × 1.9	
22	MIROC-ESM-CHEM	2.8 × 2.8	Japan Agency for Marine-Earth Science and Technology, Atmosphere and Ocean Research Institute (The University of Tokyo), and National Institute for Environmental Studies
23	MIROC-ESM	2.8 × 2.8	
24	MIROC5	1.4 × 1.4	Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology
25	MPI-ESM-LR	1.875 × 1.875	Max Planck Institute for Meteorology
26	MPI-ESM-MR	1.875 × 1.875	
27	MRI-ESM1	1.125 × 1.125	Meteorological Research Institute
28	MRI-CGCM3	1.1 × 1.1	
29	NorESM1-M	2.5 × 1.875	Norwegian Climate Centre

Table 2. Nash-Sutcliffe Efficiency (NSE) of hydrological models in the calibration and validation periods.

Country	Watershed name	Area (km^2)	High flow	Low flow	Calibration period	NSE calibration	Validation period	NSE validation
China	Xiangjiang	52150	Apr-Jun	Jul-Nov	1975-1987	0.912	1988-2000	0.871
Canada	Manicouagan-5	24610	Mar-Jul	Aug-Feb	1970-1979	0.926	1980-1989	0.881

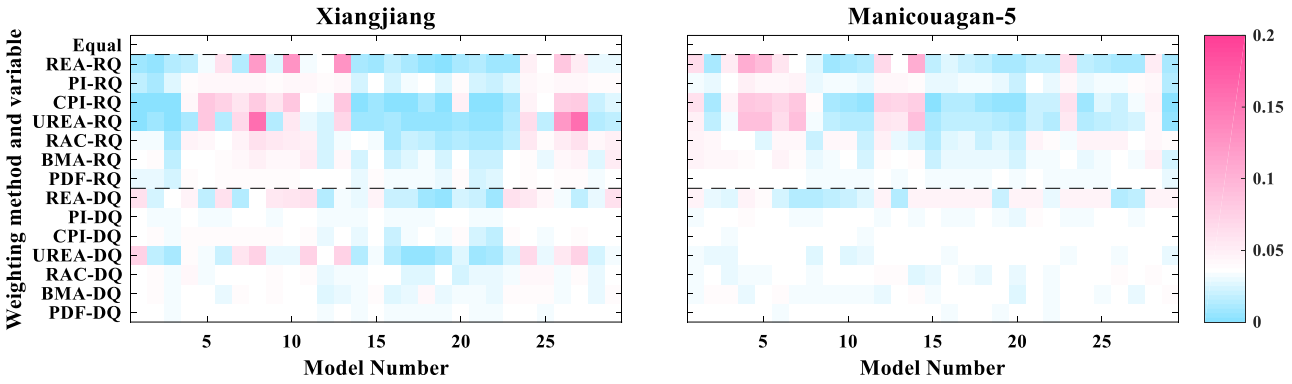

Figure 2. Weights assigned by equal weighting and 7 unequal weighting methods based on raw climate model data-simulated streamflow (RQ) and bias corrected data-simulated streamflow (DQ) for two watersheds. (Equal weight is presented in white, weights greater than equal are presented in red, and weights less than equal in blue.)

Table 3. The entropy of weights calculated by equal weighting and 7 unequal weighting methods based on raw climate model data-simulated streamflow (RQ) and bias corrected data-simulated streamflow (DQ) for two watersheds. The entropy of weights calculated based on raw and bias-corrected temperature (RT and DT) and precipitation (RP and DP) are also presented for comparison.

	Xiangjiang watershed						Manicouagan-5 watershed					
	RT	RP	RQ	DT	DP	DQ	RT	RP	RQ	DT	DP	DQ
REA	2.45	3.04	2.93	3.05	3.18	3.22	2.87	3.11	3.06	3.12	3.30	3.29
PI	3.34	3.35	3.33	3.37	3.37	3.37	3.34	3.34	3.34	3.36	3.36	3.37
CPI	2.46	2.92	2.86	3.37	3.36	3.35	2.99	3.12	3.00	3.37	3.37	3.37
UREA	2.72	3.00	2.73	3.33	3.22	3.15	3.02	3.15	3.10	3.33	3.35	3.36
RAC	3.37	3.35	3.25	3.37	3.36	3.36	3.37	3.36	3.32	3.37	3.36	3.36
BMA	3.34	3.36	3.33	3.36	3.36	3.36	3.35	3.36	3.35	3.37	3.36	3.36
PDF	3.36	3.37	3.36	3.37	3.37	3.37	3.37	3.37	3.37	3.37	3.37	3.37
Equal	3.37						3.37					

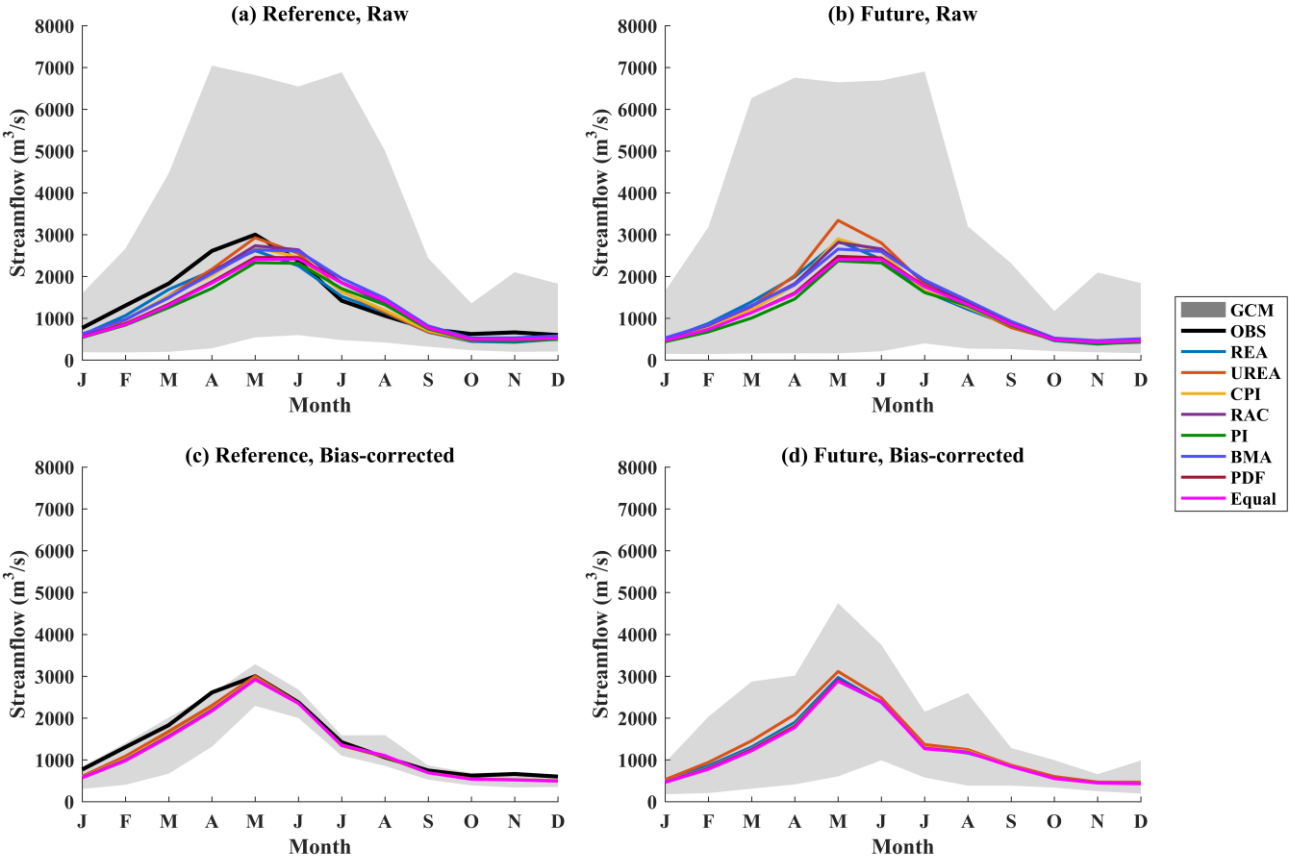

Figure 3. The envelope of monthly mean streamflows simulated by 29 raw and bias-corrected GCM outputs and the multi-model ensemble means of monthly mean streamflows weighted by 8 weighting methods based on GCM-simulated streamflows over the Xiangjiang watershed for the reference and future periods (OBS = the hydrograph simulated from meteorological observation).

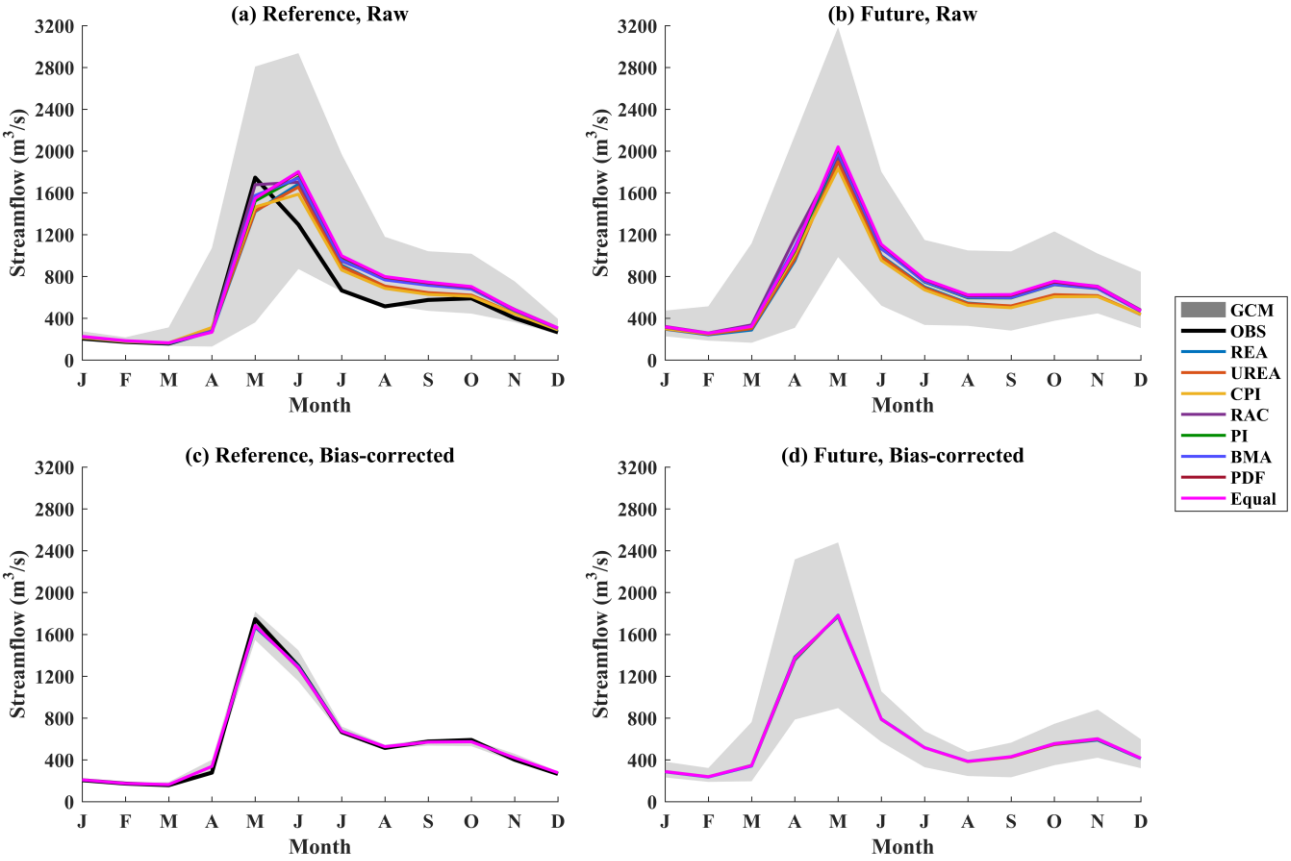

Figure 4. ~~The envelope of monthly mean streamflows simulated by 29 raw and bias corrected GCM outputs and the multi-model ensemble means of monthly mean streamflows weighted by 8 weighting methods based on GCM-simulated streamflows over the Manicouagan-5 watershed for the reference and future periods (OBS = the hydrograph simulated from meteorological observation).~~ The same as Fig. 3 but for the Manicouagan-5 watershed.

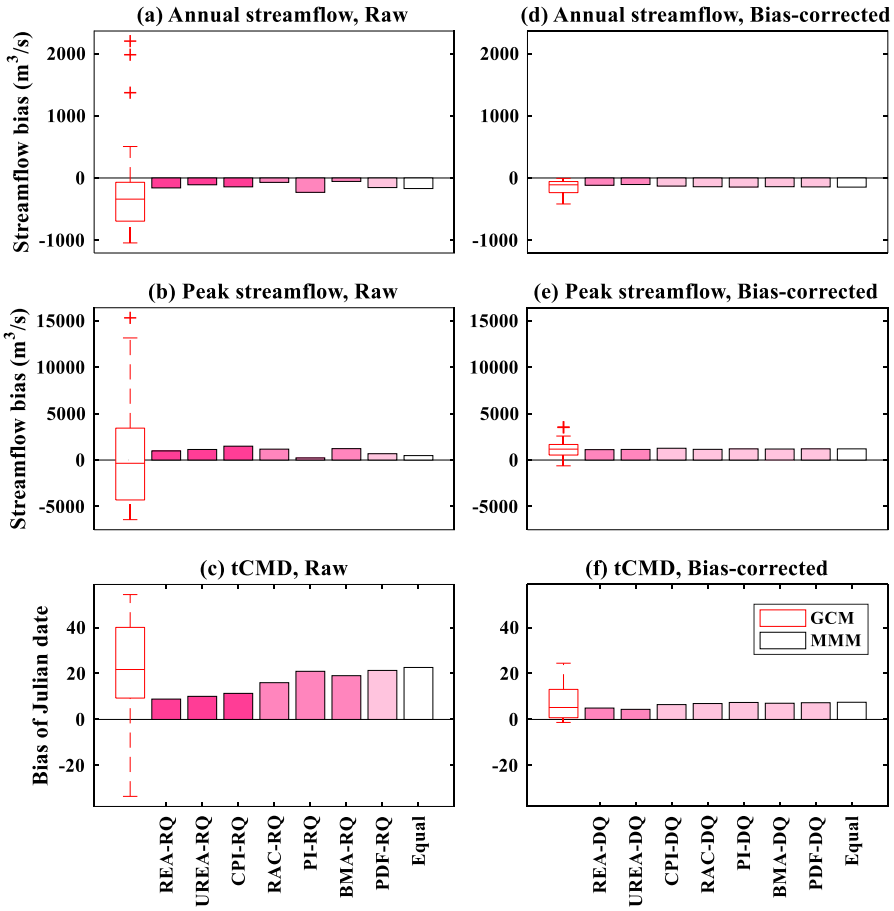

Figure 5. Bias in mean annual streamflow, mean peak streamflow and mean center of timing of annual flow (tCMD) simulated using 29 raw or bias-corrected GCM outputs and the multi-model means (MMM) combined by weights based on raw (RQ) and bias-corrected (DQ) GCM-simulated streamflows in the Xiangjiang watershed in the reference period. (The depth of pink in the MMM bars represents the level of inequality of weights, as indicated in Table 3.)

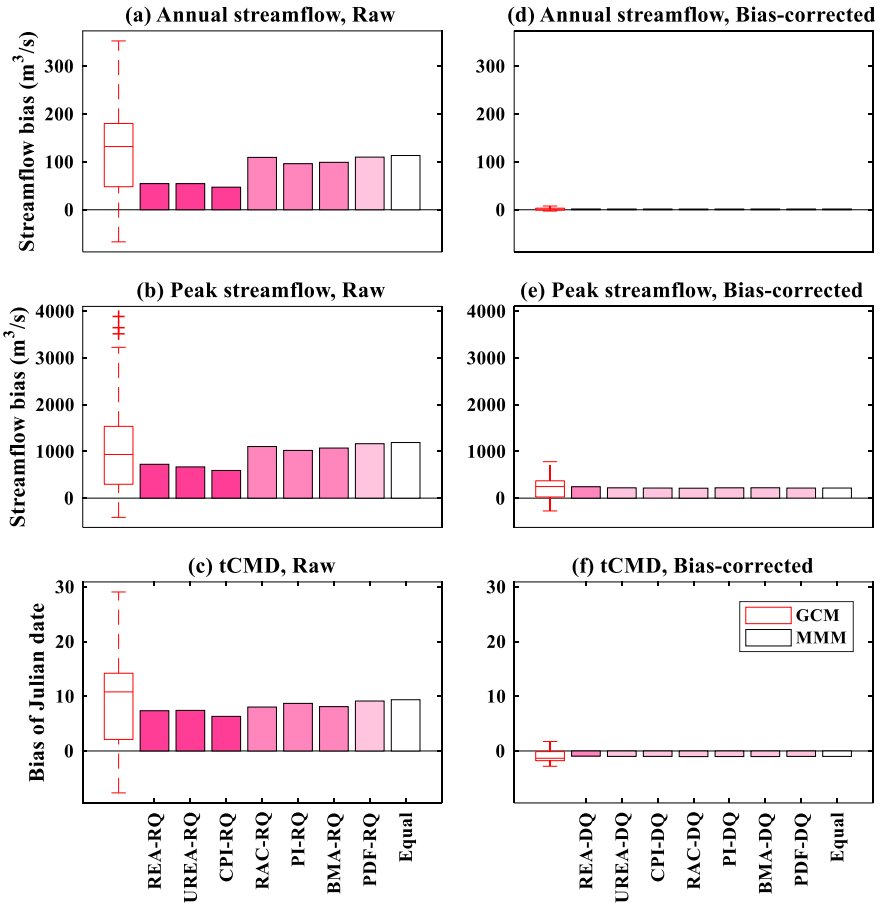

Figure 6. ~~Bias in mean annual streamflow, mean peak streamflow and mean center of timing of annual flow (tCMD) of streamflows simulated using 29 raw or bias-corrected GCMs and the multi-model means (MMM) combined by weights based on raw streamflows (RQ) and bias-corrected streamflows (DQ) in the reference period in the Manicouagan-5 watershed. (The depth of pink in the MMM bars represents the level of inequality for the corresponding set of weights.)~~The same as Fig. 5 but for the Manicouagan-5 watershed.

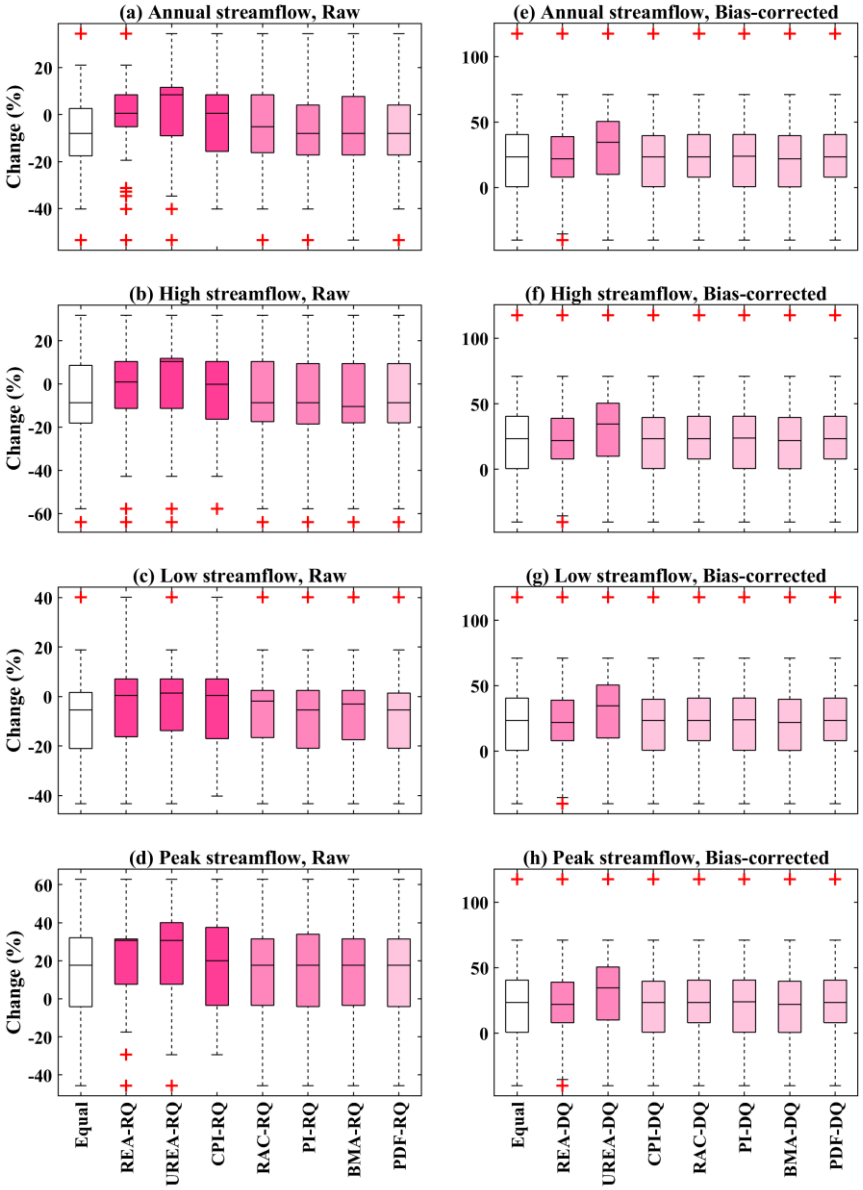

Figure 7. Box plot of changes in four hydrological indices calculated by raw or bias-corrected GCM-simulated streamflows in the Xiangjiang watershed. The changes of hydrological variables were sampled through the Monte-Carlo approach based on the weights calculated using raw (RQ) or bias-corrected (DQ) GCM-simulated streamflows. (The depth of pink represents the level of inequality of the weights.)

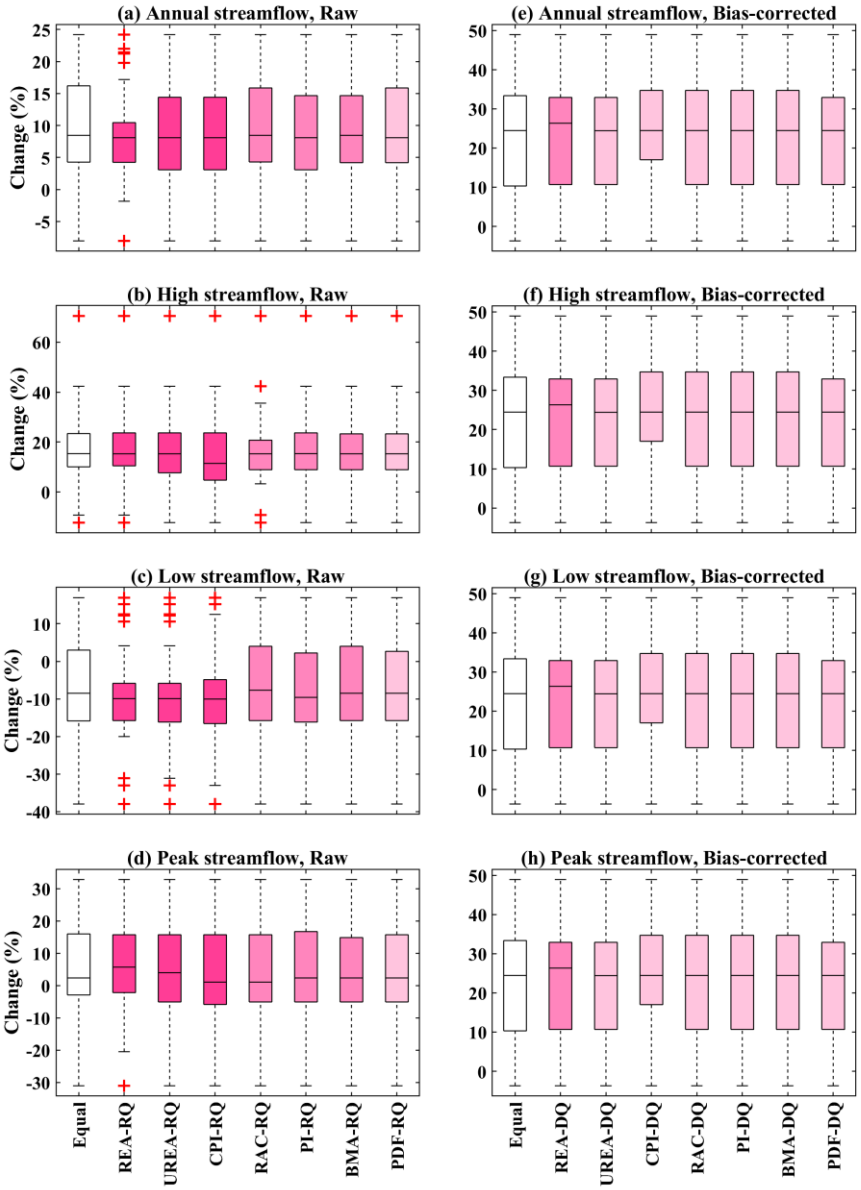

Figure 8. ~~Box plot of changes in four hydrological indices calculated by raw or bias-corrected GCM-simulated streamflows in the Manicouagan-5 watershed. The changes of hydrological variables were sampled through the Monte-Carlo approach based on the weights calculated using raw (RQ) or bias-corrected (DQ) GCM-simulated streamflows. (The depth of pink represents the level of inequality of the weights.)~~ The same as Fig. 7 but for the Manicouagan-5 watershed.

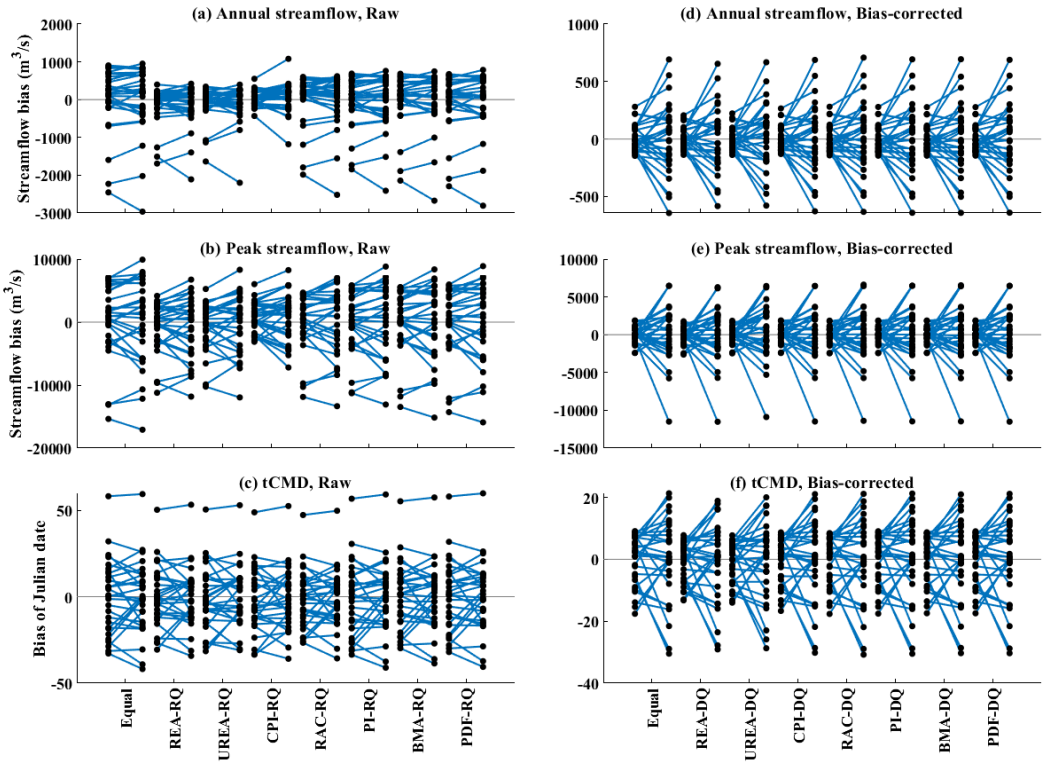

Figure 9. Bias in mean annual streamflow, mean peak streamflow and mean center of timing of annual flow (tCMD) of weighted multi-model ensemble mean in the out-of-sample testing over the Xiangjiang watershed. Twenty-nine sticks of each weighting method represent the results when each of 29 climate models was regarded as the "truth" in turn, and the left and right points in each stick represent the bias for the reference and future periods, respectively.

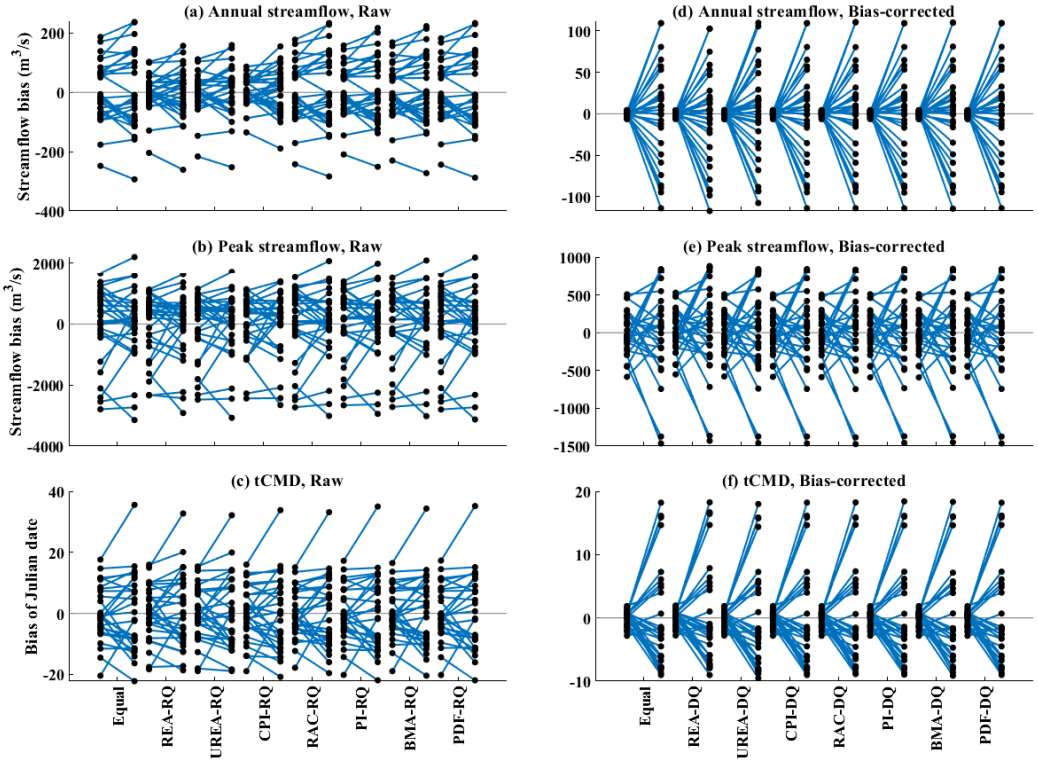

Figure 10. The same as Fig. 9 but for the Manicouagan-5 watershed.

Supplement of

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

Hui-Min Wang et al.

Correspondence to: Jie Chen (jiechen@whu.edu.cn)

1 Weighting methods

1.1 Reliability ensemble averaging (REA)

The reliability ensemble averaging method of Giorgi and Mearns (2002) considers two reliability criteria for a GCM. The first one is the model performance criterion that evaluates the ability of a climate model to simulate historical observation, and the other is the model convergence criterion that examines the difference of a model to the multi-model mean in the future period. The reliability factor of a model is defined as

$$\mathrm{REA}_i = \left\{ \left[\frac{\epsilon}{\mathrm{abs}(B_i)} \right]^m \times \left[\frac{\epsilon}{\mathrm{abs}(D_i)} \right]^n \right\}^{1/mn} \tag{S1}$$

where ϵ represents the natural climate variability estimated by the interval between the maximum and minimum of 20-year moving averages of yearly observation series. B_i is the bias of a simulation to the observation in terms of the climatological mean, and D_i is the distance between the change of a given model and the REA-weighted mean change. In addition, if the absolute value of bias B_i or distance D_i is smaller than climate variability ϵ, this climate simulation is regarded to be reliable in the corresponding respect (i.e. $\epsilon / \mathrm{abs}(B_i)$ or $\epsilon / \mathrm{abs}(D_i)$ is set to 1).The parameters m and n represent the weight assigned to performance and convergence criteria, respectively, and are both set to 1 in this study.

1.2 Weighing scheme accounting for performance and interdependence (PI)

Since many climate models share similar modules or parts of codes, they cannot be regarded as independent of each other as in model democracy. Thus, Knutti et al. (2017) proposed a weighting scheme accounting for both performance and interdependencies (PI). The interdependence score I_i of an ith model is evaluated as

$$\mathrm{I}_i = \frac{1}{1 + \sum_{j \neq i}^{N} e^{-\frac{D_{ij}^2}{\sigma_D^2}}} \tag{S2}$$

where D_{ij} measures the distance between the ith and the jth model in terms of the climatological mean. The uniqueness radius σ_D determines how strongly the model interdependency criterion is stressed. When a model is far from all the other models, its interdependence score becomes larger but no more than 1. The performance score P_i of an ith model is evaluated as

$$\mathrm{P}_i = e^{-\frac{B_i^2}{\sigma_B^2}} \tag{S3}$$

where B_i measures the distance of the ith model to the observation in terms of the climatological mean. The skill radius σ_B determines how strongly the model performance criterion is stressed. The overall score of an ith model is calculated by multiplying its interdependence score and performance score as follows:

$$\mathrm{PI}_i = \mathrm{P}_i \times \mathrm{I}_i \tag{S4}$$

Two parameters, σ_D and σ_B, are measured by the multiples of the median distances across all model pairs, and are chosen by visual inspection based on two standards. First, the choice of σ_D should attempt to guarantee that the group of models that are known to be similar (i.e. MIROC-ESM-CHEM, MIROC-ESM and MIROC5 in this study) should gain an I_i about $1/k$ (k is the number of alike models) (Sanderson et al., 2017). Second, σ_B is sampled via perfect model tests (cross validation), in which each model is alternatively regarded as the truth model and the others are used to calculate the PI weights (Knutti et al., 2017). The determination of σ_D should attempt to guarantee that 80% of the truth models fall into the 10-90% range projected by the corresponding weighted ensemble in the future period. For the Manicougan-5 watershed, $\sigma_D = 0.35$ and $\sigma_B = 2$. For the Xiangjiang watershed, $\sigma_D = 0.25$ and $\sigma_B = 2.8$.

1.3 Representation of the annual cycle (RAC)

The skill score of representation of the annual cycle (RAC) is developed based on the Taylor diagram, which is used to indicate the similarity between a climate simulation series and an observation series (Taylor, 2001). The RAC method can be expressed as the following 4th order formulation.

$$\mathrm{RAC}_i = \frac{4(1+r)^4}{(\sigma + 1/\sigma)^2(1+r_0)^4} \tag{S5}$$

where r is the correlation coefficient between the monthly observed and simulated series, and r_0 is the maximum correlation, which is set to 1 in this study. The parameter $\sigma = \sigma_s/\sigma_o$ is the ratio between the standard deviation of a monthly simulated series and that of a monthly observed series.

1.4 Upgraded reliability ensemble averaging (UREA)

Since the REA method may artificially reduce uncertainty by its convergence criterion and only consider one metric (i.e. climatological mean), Xu et al. (2010) proposed upgraded reliability ensemble averaging (UREA) to eliminate the model convergence criterion and to introduce other statistics. Even though multiple climate variables were simultaneously evaluated by multiplying their skill scores in Xu et al. (2010), this study individually evaluated each variable as follows.

$$\mathrm{UREA}_i = \left[\frac{\epsilon_a}{\mathrm{abs}(B_{a,i})}\right]^{m_1} \times \left[\frac{\epsilon_v}{\mathrm{abs}(B_{v,i})}\right]^{m_2} \tag{S6}$$

where $B_{a,i}$ and $B_{v,i}$ are the biases of a climate simulation in the average and variance, respectively. ϵ_a and ϵ_v represent the natural climate variability in terms of annual average and inter-annual variation, respectively. The variation is measured by the standard deviation for temperature series and by the coefficient of variation for precipitation and runoff series. In addition, if

the absolute value of bias in the average $B_{a,i}$ or variance $B_{v,i}$ is smaller than climate variability $\in$, this climate simulation is regarded to be reliable in the corresponding respect (i.e. $\in_a / \mathrm{abs}(B_{a,i})$ or $\in_v / \mathrm{abs}(B_{v,i})$ is set to 1). The parameters m_1 and m_2 represent the weight assigned to two metrics and are both set to 1 in this study.

1.5 Bayesian model averaging (BMA)

Bayesian model averaging (BMA) is a statistical inference approach to obtain probabilistic forecasts from multi-model ensemble simulations based on Bayes theory. BMA has been used to develop probabilistic predictions for ensembles of weather forecasting models, climate models or hydrological predictions (Duan et al., 2007; Min et al., 2007; Raftery et al., 2005). Denote y as the variable to be predicted, $D = [y_1^o, y_2^o, \dots, y_T^o]$ as the observed series with a length of T, and $f = [f_1, f_2, \dots, f_N]$ as the ensemble of series simulated by climate models. Based on the total probability rule, the probability density function of the prediction $p(y|D)$ can be presented as follows.

$$p(y|D) = \sum_{i=1}^{N} p(f_i|D) \cdot p_i(y|f_i, D) \tag{S7}$$

where each simulation f_i is associated with a conditional probability density function, $p_i(y|f_i, D)$, which represents the conditional distribution of y on f_i, given that f_i is regarded as the best simulation for D. The posterior probability $p(f_i|D)$ represents the likelihood that a simulation is the right simulation. It can also be seen as the weight, $w_i = p_i(y|f_i, D)$, which reflects the capability of a simulation to reproduce the observation. Then, the posterior mean is as follows.

$$E[y|D] = \sum_{i=1}^{N} p(f_i|D) \cdot E[p_i(y|f_i, D)] = \sum_{i=1}^{N} w_i f_i \tag{S8}$$

As the use of BMA in Duan et al. (2007), this study assumed that $p_i(y|f_i, D)$ consists of a Gaussian distribution; monthly data series were then adopted as model simulated series f_i. For the variables that do not follow a Gaussian distribution (i.e. precipitation and streamflow in this study), the Box-Cox transformation was used to transform the variables before the BMA algorithm. This study used the Expectation-Maximization algorithm to solve the BMA weights. More details of this algorithm can be found in Duan et al. (2007).

1.6 Climate prediction index (CPI)

The Climate prediction index (CPI) was introduced by Murphy et al. (2004) to weight climate models based on their relative reliability to correctly simulate climate observation. Assuming that the simulated variable belongs to the Gaussian distribution, the likelihood of a simulated statistic is proportional to the following equation.

$$\mathrm{CPI}_i = \exp\left[-0.5 \frac{(s_i - o_i)^2}{\sigma_{ANN}^2}\right] \tag{S9}$$

where the climatological mean of a simulated series s_i is assumed to have a Gaussian distribution with an expectation of o_i (the observational climatological mean) and a variance simply estimated by σ_{ANN}^2 (the inter-annual variance of the simulated series).

1.7 Evaluation of the probability density function (PDF)

Perkins et al. (2007) proposed a skill score to evaluate climate models' ability to reproduce the probability density functions (PDF) of observation. Expressed formally, the skill score of a climate simulation is given as

$$\mathrm{PDF}_i = \sum_{1}^{K} \mathrm{minimum}(Z_s, Z_o) \tag{S10}$$

where the probability density function of simulated or observed daily series is separated into K bins, and Z_s and Z_o represent the frequency in a given bin, respectively.

References

Duan, Q., Ajami, N. K., Gao, X., and Sorooshian, S.: Multi-model ensemble hydrologic prediction using Bayesian model averaging, Advances in Water Resources, 30, 1371-1386, https://doi.org/10.1016/j.advwatres.2006.11.014, 2007.

Giorgi, F., and Mearns, L. O.: Calculation of Average, Uncertainty Range, and Reliability of Regional Climate Changes from AOGCM Simulations via the "Reliability Ensemble Averaging" (REA) Method, Journal of Climate, 15, 1141-1158, https://doi.org/10.1175/1520-0442(2002)015<1141:coaura>2.0.co;2, 2002.

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.

Min, S. K., Simonis, D., and Hense, A.: Probabilistic climate change predictions applying Bayesian model averaging, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 365, 2103-2116, https://doi.org/10.1098/rsta.2007.2070, 2007.

Murphy, J. M., Sexton, D. M., Barnett, D. N., Jones, G. S., Webb, M. J., Collins, M., and Stainforth, D. A.: Quantification of modelling uncertainties in a large ensemble of climate change simulations, Nature, 430, 768-772, https://doi.org/10.1038/nature02771, 2004.

Perkins, S. E., Pitman, A. J., Holbrook, N. J., and McAneney, J.: Evaluation of the AR4 Climate Models' Simulated Daily Maximum Temperature, Minimum Temperature, and Precipitation over Australia Using Probability Density Functions, Journal of Climate, 20, 4356-4376, https://doi.org/10.1175/jcli4253.1, 2007.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M.: Using Bayesian Model Averaging to Calibrate Forecast Ensembles, Monthly Weather Review, 133, 1155-1174, https://doi.org/10.1175/mwr2906.1, 2005.

Sanderson, B. M., Wehner, M., and Knutti, R.: Skill and independence weighting for multi-model assessments, Geoscientific Model Development, 10, 2379-2395, https://doi.org/10.5194/gmd-10-2379-2017, 2017.

Taylor, K. E.: Summarizing multiple aspects of model performance in a single diagram, Journal of Geophysical Research: Atmospheres, 106, 7183-7192, https://doi.org/10.1029/2000jd900719, 2001.

Xu, Y., Gao, X., and Giorgi, F.: Upgrades to the reliability ensemble averaging method for producing probabilistic climate-change projections, Climate Research, 41, 61-81, https://doi.org/10.3354/cr00835, 2010.

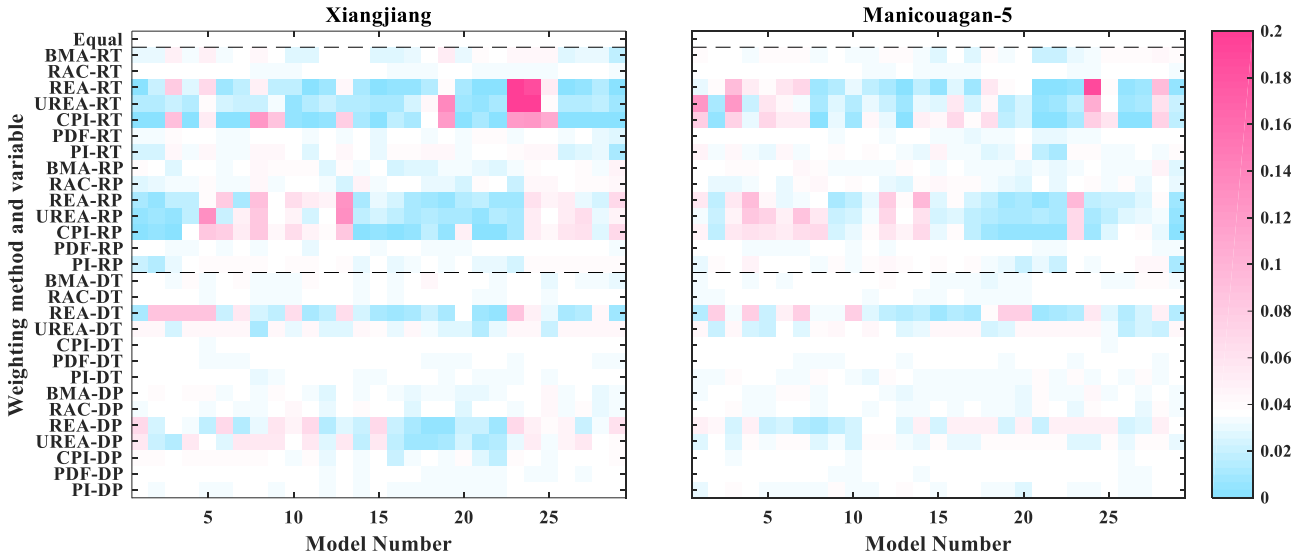

Figure S1. Weights assigned 8 weighting methods based on raw temperature (RT) and precipitation (RP) of GCM outputs and bias-corrected temperature (DT) and precipitation (DP) of GCM outputs for two watersheds. (Equal weight is presented in white, weights larger than equal are presented in red, and weights lower than equal are in blue.)

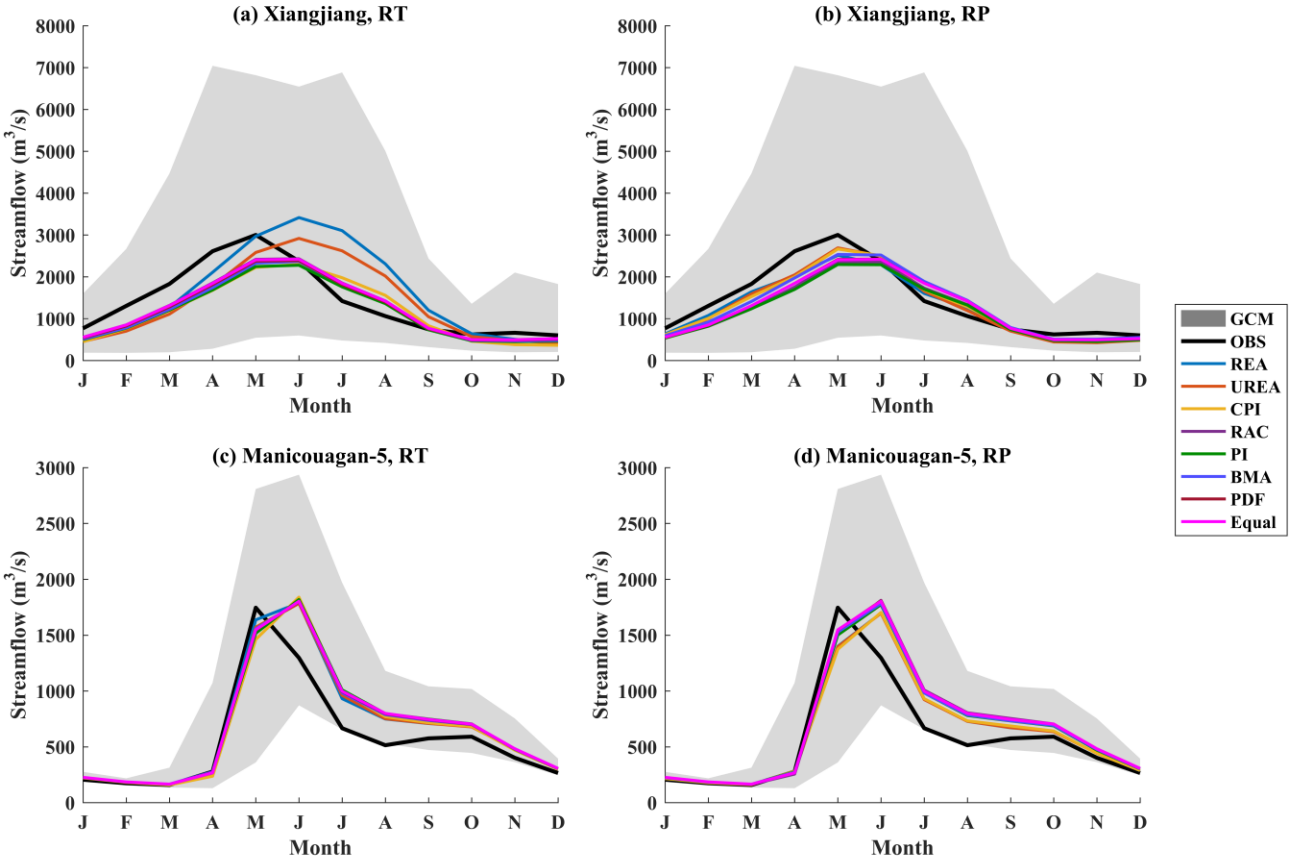

Figure S2. The envelope of monthly mean streamflows simulated by 29 raw and bias-corrected GCM outputs and the multi-model ensemble means of monthly mean streamflows weighted by 8 weighting methods based on raw temperature (RT) and precipitation (RP) of GCM outputs in both watersheds for the reference period (OBS = the hydrograph simulated from meteorological observation).

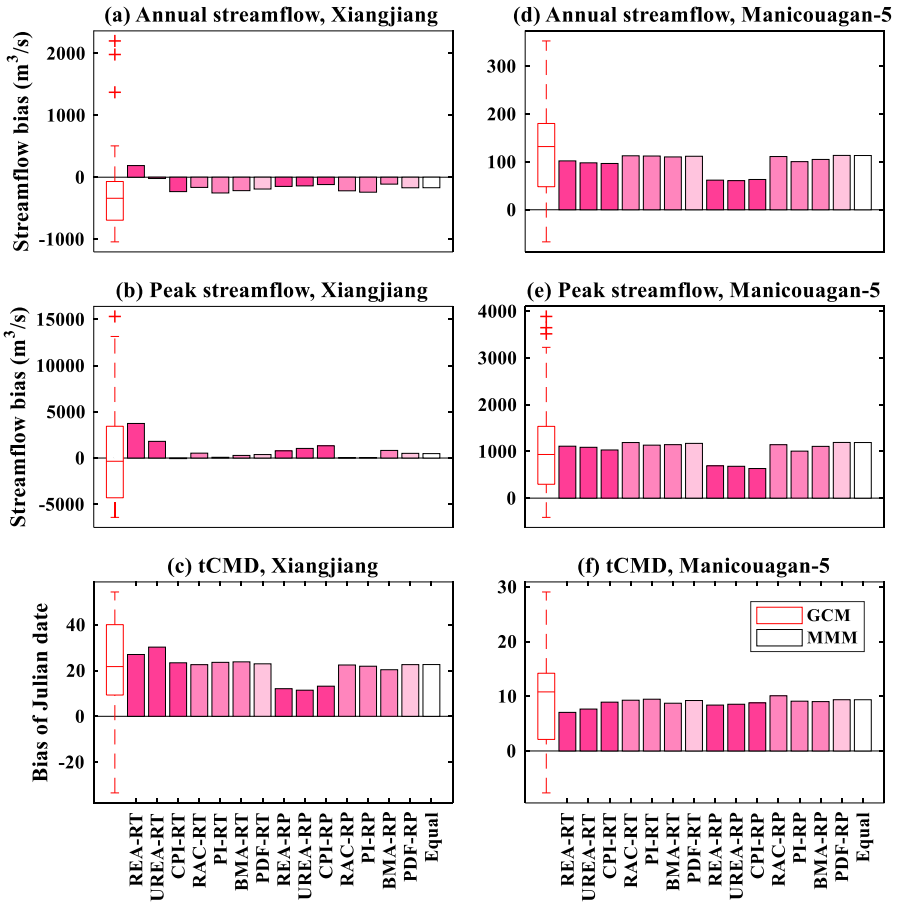

Figure S3. Bias in mean annual streamflow, mean peak streamflow and mean center of timing of annual flow (tCMD) simulated using 29 raw or bias-corrected GCM outputs and the multi-model means (MMM) combined by weights based on raw temperature (RT) and raw precipitation (RP) in both watershed for the reference period. (The depth of pink in bars of MMM represents the level of inequality of weights as indicated in Table 3.)

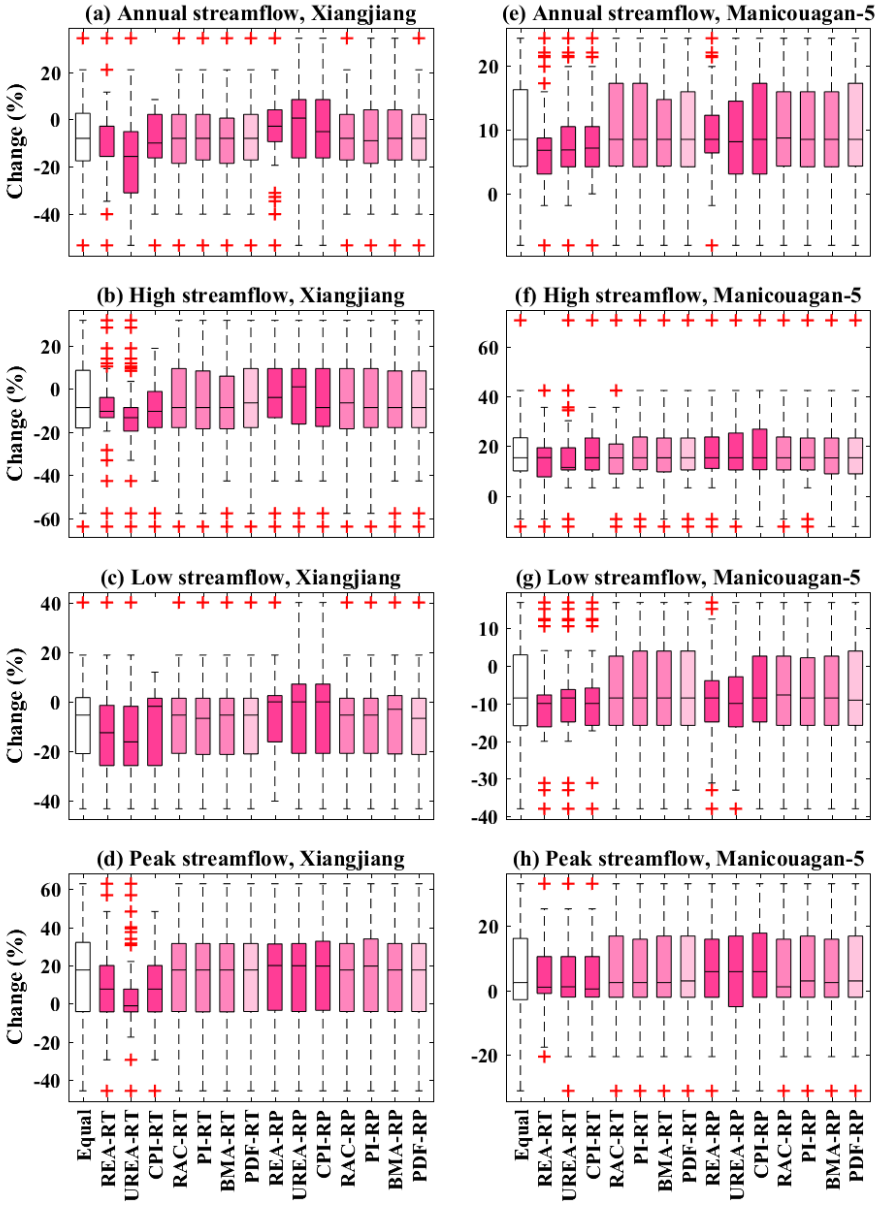

Figure S4. Box plot of changes in four hydrological indices simulated by raw GCM-simulated streamflows over both watersheds. The changes of hydrological variables were sampled through the Monte-Carlo approach based on the weights calculated using raw temperature (RT) and precipitation (RP) of GCM outputs. (The depth of pink represents the level of inequality of weights.)