Authors' responses to comments

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

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

We appreciate the editor's and the two referees' reviews on the manuscript. These comments 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 correspondingly.

Responses to Editor's comments

We sincerely appreciate the editor for reviewing this manuscript again. We have carefully studied and responded to the comments from the editor and referee #3, since referee #2 has been satisfied with our revision. Please find our specific responses below.

> *[Title] I suggest replacing the phrase "a more reasonable" with "a better" or a similar phrase.*

Thanks for the suggestion on the title. We have substituted the phrase "more reasonable" with "better".

> *[5, 16-17] In the DT method, are the precipitation and temperature corrected using a cross-correlation, or does each variable is treated independently?*

We are sorry for the unclarity here. In the DT method, each variable is treated independently. In other words, the dependence between precipitation and temperature was not taken into account in this study. Admittedly, it may be more reasonable to use a bias correction method considering the inter-variable dependence. However, our previous study (Chen et al., 2018) showed that the use of a more complicated method does not manifest much advantage over the use of the independent bias correction method for these two watersheds. This has been clarified in the revised manuscript [P5, L22-24].

> *[Section 3.2] A figure illustrating the GR4J and CemaNeige models will be useful for the readers. This figure can be presented as Supplementary Material or in the main text.*

Thanks for this helpful suggestion. Figure R1 has been added for illustrating structure of the GR4J-6 hydrological model [Figure S1 in the Supplement].

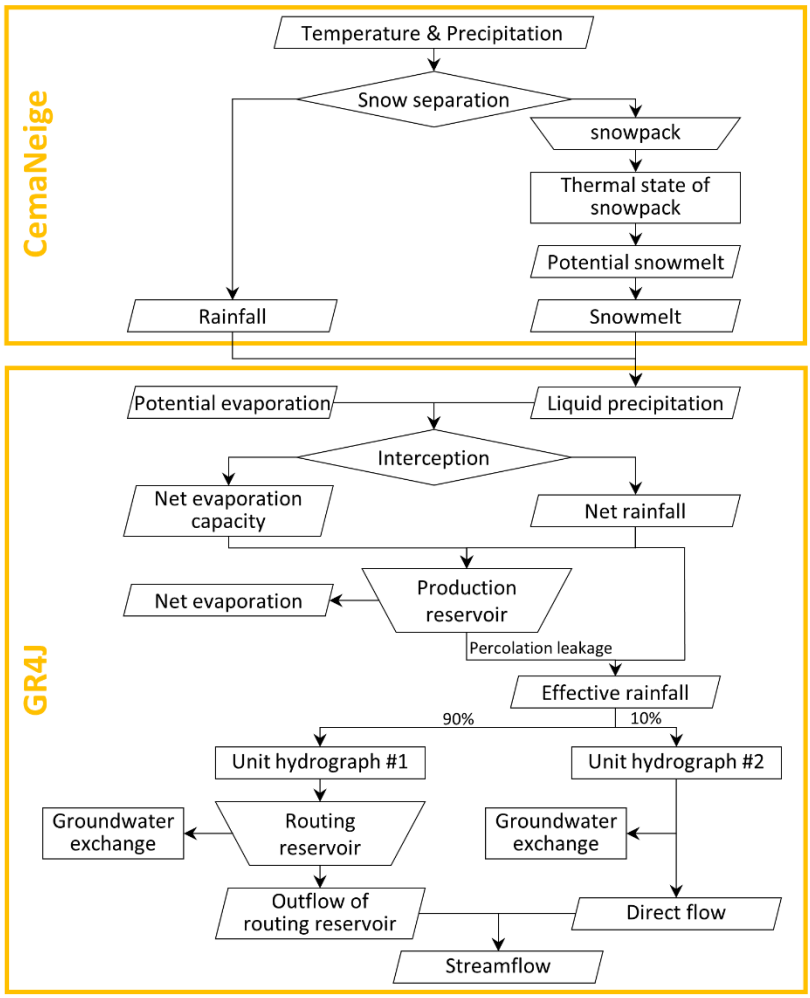

Figure R1. The flowchart of the GR4J-6 hydrological model.

[Figure 8] Can be presented as Supplementary Material.

We appreciate the editor's suggestion. This figure has been moved to the Supplement [Figure S5].

[Section 5] I am missing in the discussion a paragraph about the generalization of the results - what about implications to other climate environments (arid or humid)? urban vs. rural catchments? etc. This should be discussed.

We agree with the editor that it is necessary to further discuss the implications of results for watersheds in other climate regions.

If weights are determined on climate variables for watersheds in different climate regions, they may manifest different performances on the hydrological impacts. For example, for arid watersheds, whose hydrological regime is more characterized by the intense flow and evaporation, a proper combination for the weights based on temperature and precipitation may be needed in order to obtain a better quantification. For urban watersheds, stormwater contributes to their runoff and weights based on precipitation intensity may be more advantageous. Nevertheless, based on the results of

this study, using impact variables to determine weights may help to circumvent the problem of trade-off and choice of climate variables. However, specific advantages of weights based on impact variables and influences of bias correction on the performances of weighting methods in other types of watersheds still deserve further research.

These points have been added in the Discussion section of the revised manuscript [P15, L4-12].

[12 20-28] These sentences repeat what is written already in the introduction.

We agree with the editor that most parts of these sentences repeat the introduction. Here, we intended to state that the results of this study reflect the two problems when unequal weighting methods are used in studies of climate change impacts. Thus, these sentences have been rephrased to curtail the repeating part and to make this point clearer [P12, L21-26].

[13 10-14] Can be deleted, already mentioned in the introduction.

Thanks for the suggestion. We have removed these sentences as suggested [P13, L9-10].

[13 26-27] "Therefore, it is still viable to attend to the bias-corrected ensembles with the equal weighting method" (and similar phrasing in the abstract). This is an awkward way to conclude the results. In fact, the results imply that likely in most cases using bias-correction and equal-weighing is sufficient for hydrological impact studies.

Thanks for the comment. According to the suggestion, this sentence has been modified as "In hydrological impact studies, it is likely that using equal weighting is viable and sufficient in most cases when bias correction has been applied" [P13, L22-23]. Similarly, the conclusion in the Abstract has also been modified accordingly [P1, L24-25].

[Figure 9] The values along the y-axis of a-d, b-e, and c-f should be identical to allow comparison between the plots.

We appreciate the editor's suggestion on this figure. The values along the y-axis have been set to the same for each pair of sub-figures as suggested [Figures 8 and S7].

[Figure 10] Can be moved to the Supplementary Material.

Thanks for the suggestion. This figure has been moved to the Supplement [Figure S7].

Responses to Referee #3's comments

We sincerely appreciate the referee's comments on the manuscript. We have studied the comments and made relevant corrections to the manuscript. Please find the point-by-point responses below.

I thank the authors for addressing my comments. I must admit that I find the manuscript and

their reply a bit difficult to follow, in particular because of grammatical errors - I apologise in advance if some of the points I raise below are the result of a misunderstanding. I am concerned that future readers will also find this study difficult to read and understand, which is unfortunate at this stage of the review process.

We would like to thank the referee for reviewing our manuscript. We feel sorry that some grammatical errors in the last reply led to some confusions. We have carefully reviewed the manuscript and fixed the grammatical errors to make our statements clearer. Please find our specific responses below.

In the abstract, the authors now state:

1. *"when using raw GCM outputs, streamflow-based weights better represent the mean hydrograph and reduce more biases of annual streamflow than the weights calculated using climate variables". This is an interesting and potentially impactful result. But where is it shown? I cannot find a Figure or Table supporting the above statement. Table 3 shows the entropy of weights computed using temperature and precipitation, and Figure S1 the weights based on temperature and precipitation, but as I understand it, this does not demonstrate the benefits of computing weights using streamflow instead of temperature/precipitation. In their reply to my comment, the authors refer to P13, L14-18, which is part of the discussion and does not refer to any Figures or Tables in the manuscript. Please clarify. Please also note that the above sentence is not grammatically correct.*

Sorry for the unclear sentence. In our last response, the citation to "*P13, L14-18*" was used to demonstrate that the results of this study confirm the point of the referee (i.e. "*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*"), and it did not serve to show results for the experiment of using raw GCM outputs to simulate streamflow.

We feel sorry that we did not clearly point out relevant results in our last response. Actually, better performances of streamflow-based weights on mean hydrograph are presented in Figures 3, 4 & S3 for both watersheds and described in Section 4.2 *Impacts on the hydrological regime* (especially in P9, L4-10). Fewer biases of annual streamflow are presented in Figures 5, 6 & S4 and described in Section 4.3 *Bias in multi-model mean* (especially in P10, L9-11 & 24-27). Also, the advantages of streamflow-based weights were also discussed in the first paragraph of the Discussion section. Herein, we have added citation to relevant figures in this paragraph to make this analysis clearer [P12, L26-P13, L8].

In addition, in order to fix grammatical errors, we have corrected the sentence in the abstract as "when using raw GCM outputs to simulate streamflows, streamflow-based weights have a better performance in reproducing observed mean hydrograph than climate-variable-based weights" [P1,

L19-21].

2. "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." As also stressed by reviewer 2, this is expected, as by construction, bias-correction forces climate simulations to look like observations, i.e. artificially reduces differences between them (e.g., Hakala et al., 2018), thereby making it difficult to differentiate between good and poor models. Hence, on its own, this result does not justify publication in my view.

We appreciate the referee's comment on this conclusion. Sorry that we failed to make this point clear enough in the previous version. As stated in P8, L15-16 & P9, L11-15, we agree that in the experiment of using bias-corrected GCM outputs to simulate streamflows, the ability of weighting methods to differentiate performances among different GCMs is affected by the bias correction, which is also observed in Hakala et al. (2018). This is the reason why the results of unequal weighting are similar to those of equal weighting in this experiment. This is also discussed in the Discussion section (P13, L7-13). However, this does not mean that the performances of model weighting methods in this case do not deserve research. Although the equal weighting is often used for bias-corrected ensembles by default in many regional hydrological impact studies, whether this method is viable or sufficient remains unclear. This study focuses on this problem and can offer a reference for the use of weighting methods in hydrological impact studies. In addition, bias correction methods are usually applied to climate variables for hydrological impacts studies. Even though most of bias correction methods can reduce the bias of climate model simulations in terms of a few statistical metrics, no bias correction methods are perfect to remove all the biases. However, hydrological simulations reflect the overall performance of climate simulations, and small biases in climate simulations (in terms of a few metrics) may result in large biases in hydrological simulations, especially taking into account the fact that the climate to hydrological process is nonlinear. Thus, if unequal weighting methods consider criteria that are different to the bias correction, they may have potentials to induce better quantification. This problem deserves further studies.

These points have been clarified in the revised manuscript [P12, L19-21 & P13, L22-29].

The authors conducted additional analysis based on a pseudo-reality experiment to explore the consequence of model weighting under future climate. I thank them for their effort. The results are consistent with those based on current climatic conditions: bias-correction makes it very difficult to distinguish between good and poor models, and there are essentially no benefits to implement a weighting scheme, as the weighting schemes considered do not reduce biases more than an equal weight strategy (see Figures 9d-f and Figures 10d-f). Again, I believe that sequentially applying bias-correction and model weighting based on bias-corrected

simulations is a flawed approach.

Thanks for the comments on the out-of-sample testing. We feel that we need to explain better our motivation for conducting this study. On the one hand, we agree that unequal weighting methods do not bring significant improvements on the multi-model mean compared to the equal weighting when using bias-corrected ensembles. On the other hand, we think that model weighting should not be regarded as a supplementary process but a necessary process in impact studies. No matter whether bias correction is done before driving the hydrological model or not, a decision on the weighting methods is always necessary in order to obtain multi-model mean or uncertainty evaluation. Albeit the normal and default choice of equal weighting, whether this choice is viable and sufficient remains unclear, and this study focuses on this question. We have clarified this point in the Discussion section of the revised manuscript [P12, L19-21].

Furthermore, I am concerned that the results of this study might be misinterpreted. Readers might believe that there are no benefits to use weighting when the climate simulations have been bias-corrected. It is possible, however, that there are benefits of combining bias-correction and model weighting, but I would argue that the model weighting should be done using other criteria than those considered for bias-correction. For instance, the model weighting could reduce the influence of (or exclude) models with erroneous behaviour (for instance climate models creating snow towers, as it is the case for EURO-CORDEX members) or are unable to capture key processes such as atmospheric rivers or atmospheric patterns (e.g., NAO). This could potentially constrain the ensemble (e.g., Padrón et al., 2018) and could be a successful application of model weighting and bias correction. But this would require substantial additional analysis.

Thanks for the comment. We are sorry that we might fail to explain our motivation and main findings clear enough in the previous version. This study does not mean to demonstrate that unequal weighting is totally not beneficial for bias-corrected ensembles. As stated in the P15, L27-29, this study concludes that equal weighting method is not perfect but a viable and conservative choice so far to deal with bias-corrected GCMs in hydrological impact studies.

In addition, this study used two weighting methods (REA and PI) which consider criteria that are different from the bias correction. However, both methods still do not bring differences to the final results. This has been discussed in P13, L15-20. Actually, as stated in the responses above, we agree with the referee that future development of model weighting may have potentials to induce better quantification of hydrological impacts, and our study can offer a reference for this. As stated in P14, L13-14, considering dynamic reliability of GCM simulations in model weighting may help to induce a better quantification of hydrological impacts. In addition, considering criteria different to bias correction methods in the model weighting may also help to induce a better quantification of hydrological impacts.

This has been clarified in the revised manuscript [P13, L27-29 & P14, L13-14].

> *Overall, I think that the paper lacks a clear vision and a clear message. In their title, the authors ask "Does the weighting of climate simulations result in a more reasonable quantification of hydrological impacts?" But then they do not clarify what they mean by "reasonable"...*

We appreciate the referee's the comment on the title. First of all, we have changed the phrase "a more reasonable" to "a better" as suggested by the editor. In addition, following many studies in model weighting, "better" in the title means a multi-model ensemble mean which is more similar to the observation and less uncertainty in the ensemble. Both objectives have been studied in this work. Accordingly, this point has been stated in the Discussion section [P12, L19-21].

References

Chen, J., Li, C., Brissette, F. P., Chen, H., Wang, M., and Essou, G. R. C.: Impacts of correcting the inter-variable correlation of climate model outputs on hydrological modeling, Journal of Hydrology, 560, 326-341, https://doi.org/10.1016/j.jhydrol.2018.03.040, 2018.

Hakala, K., Addor, N., and Seibert, J.: Hydrological Modeling to Evaluate Climate Model Simulations and Their Bias Correction, Journal of Hydrometeorology, 19, 1321-1337, https://doi.org/10.1175/jhm-d-17-0189.1, 2018.

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

Hui-Min Wang[1], Jie Chen[1*], Chong-Yu Xu[2,1], 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~~in using them to quantify ~~the~~ climate change impacts on hydrology: Is it necessary to unequally 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 of GCMs 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 to simulate streamflows, streamflow-based weights have a better performance in reproducing observed mean hydrograph than climate-variable-based weights. ~~when using raw GCM outputs, streamflow based weights better represent the mean hydrograph and reduce more biases of annual streamflow than the weights calculated using climate variables.~~ However, when ~~applying~~ bias correction is applied 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. ~~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.~~ Thus, it is likely that using bias correction and equal weighting is viable and sufficient for hydrological impact studies.

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 simulations of a MME, the prevailing approach is 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), 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 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 for streamflows simulated by raw or bias-corrected 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~~climatic 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^{3}s^{-1}. The annual average of summer peak streamflow is about 4420 m^{3}s^{-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. The outlet of the Manicouagan-5 River is the Daniel-Johnson Dam. 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^{3}s^{-1}. Snowmelt contributes to the peak discharge during May, whose annual average is about 2200 m^{3}s^{-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. 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~~weights 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 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, separately. 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). Each variable was corrected independently and inter-variable dependence was not considered in this study. Previous study has showed that the use of a more complicated method does not manifest much advantage over the use of the independent bias correction method for these two watersheds (Chen et al., 2018).

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 groundwater exchange 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. A brief flowchart of GR4J-6 model is shown in the Supplement (Fig. S1).

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) criterion (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 calculated for each of streamflow data series simulated by raw GCM outputs and bias-corrected outputs. 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~~In specific, for the entropy of weights (Eq. (1)), ~~the entropy~~it 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:

$$E = -\sum_{i=1}^{N} w_i \ln w_i \tag{1}$$

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~~Then, 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 model weighting on the uncertainty of hydrological impacts related to the choice of GCMs are investigated ~~in terms of~~through 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~~for the 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 are calculated 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 the Supplement (Fig. S2). 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~~they are based on bias-corrected temperature and precipitation.

4.2 Impacts on the hydrological regime

The weights determined by 8 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. S3a,b). Particularly, for the weights based on raw temperature, 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~~in this case, their ~~generated~~ ensemble mean streamflows are significantly inferior to that generated by equal weights (Fig. S3a). 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. S3b). This reflects the advantage of ~~weighting~~ streamflow ~~series~~-based weights 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~~in biases brings about similar weights for all ~~GCM simulated time series when~~ensemble members in the experiment of using bias-corrected GCM-simulated streamflows. Thus, the multi-model ensemble means of monthly mean streamflow constructed by all unequal weighting ~~method~~methods 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~~choice 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 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~~weighting methods still follow the same assumption. In other words, the bias non-stationarity implies that climate models differ in their ability to simulate the climate for the reference and future ~~period~~periods. The weights calculated in the reference period may not be applicable in the future period. The results of this study also ~~proved~~prove this, as all of the weighting methods project similar ensemble means of 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 are used to calculate weights, the multi-model mean streamflows perform better than or similar to those ~~simulated~~calculated using ~~GCM~~weights based on raw temperature and precipitation ~~data~~ (Fig. S3c). 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~~performances 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 be because only the mean value (climatological mean or monthly mean series) was used as the evaluation metric when ~~determining~~ weights are determined, 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. S4a-c). When ~~using~~ raw temperature series ~~of GCMs~~are used 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~~in the experiment of 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 (Figure 6), 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. S4d-f), a certain improvement on mean indices was also observed (the only exception is raw precipitation-based PDF weights)~~. Compared~~, but compared to the weights calculated using streamflow series, nearly all streamflow-based weights ~~based on GCM-simulated streamflows~~ reduce more biases than those based on temperature and precipitation. However, ~~when using~~if bias-corrected GCM-simulated streamflows are used (Figure 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 were 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 S5 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~~through 29 values simulated by ~~each~~ climate ~~simulation~~simulations, 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~~In the experiment of 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~~for 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. S6a-d). When ~~calculating~~ weights are calculated 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 the 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~~The uncertainty of hydrological impacts in terms of four hydrological indices over the Manicouagan-5 watershed~~.~~ is shown in the Supplement (Fig. S5). For weights calculated using raw GCM-simulated streamflows (Fig. S5a-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 are used (Fig. S5e-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~~the 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~~assessment 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 8 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. 8a-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. 8d-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 the Supplement (Fig. S7), and generally, they are also similar to the results of using observations (Fig. 6).

5 Discussion

Model weighting is a necessary process in dealing with multi-model ensembles in impact studies. No matter whether bias correction is applied before driving the hydrological model or not, a decision on the weighting methods is always necessary in order to obtain multi-model mean or uncertainty evaluation. Besides the common equal weighting, many studies have proposed unequal weighting methods in order to obtain a better quantification on climate change, such as more reliable multi-model ensemble mean or constrained uncertainty (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 studies only use climate variables to determine weights, which may cause two problems for impact studies: uncertain trade-off between multiple variables and nonlinear relationship between climate and impact variables. Actually, the results of this study reflect these problems. ~~In addition to the equal weighting method, which is a normal strategy for handling multi-model ensembles, many studies have 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 temperature in the experiment of raw GCM-simulated streamflows in the Xiangjiang watershed (Fig. S3), which lead to obviously biased multi-model mean hydrographs at the reference period. 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~~In this case, 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 (Fig. S3), 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~~based on raw GCM-simulated streamflows reduce more biases of the mean annual streamflow than those based on raw temperature and precipitation (Figures 5 and 6). 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, in the experiment of ~~simulating streamflows~~ using raw GCMs to simulate streamflows, weights calculated based on streamflows not only circumvent the above two problems, they also bring about ~~fewer~~smaller biases in mean annual streamflow for the multi-model means.

~~Since bias correction methods are routinely applied to GCM outputs for hydrological impact assessments, this study considered two experiments where raw and bias corrected GCM simulated streamflows were separately used to determine weights. 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).~~In addition, this study considered the differences in performances of weighting methods when the bias correction method is applied or not. 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 is also observed in other studies (e.g., Chen et al., 2017; Hakala et al., 2018). 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~~independence 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.~~ The out-of-sample testing also manifests the same phenomena. ~~Therefore, it is still viable to attend to the bias-corrected ensembles with the equal weighting method.~~ Therefore, in hydrological impact studies, it is likely that using equal weighting is viable and sufficient in most cases when bias correction has been applied. Admittedly, even though most of bias correction methods can reduce the bias of climate model simulations in terms of a few statistical metrics, no bias correction methods are perfect to remove all the biases. Besides, hydrological simulations reflect the overall performance of climate simulations, and small biases in climate simulations (in terms of a few metrics) may results in large biases in hydrological simulations, especially taking nonlinear processes from climate to hydrology worlds into account. Thus, if unequal weighting methods consider criteria that are different to the bias correction, they may have potentials to induce a better quantification of hydrological impacts. This problem deserves further studies.

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. 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. 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 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 (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). The way to deal with the dynamic reliability of climate models in weighting methods 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. 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 is used, 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.

In addition, the other limitation is that this study only considered two watersheds in humid regions. If weights are determined on climate variables for watersheds in different climate regions, they may manifest different performances on the hydrological impacts. For example, for arid watersheds, whose hydrological regime is more characterized by the intense flow and evaporation, a proper combination for the weights based on temperature and precipitation may be necessary in order to obtain a better quantification. For urban watersheds, stormwater contributes to their runoff and weights based on precipitation intensity may be more advantageous. Nevertheless, based on the results of this study, using impact variables to determine

weights may help to circumvent the problem of trade-off and choice of climate variables. However, specific advantages of weights based on impact variables and influences of bias correction on the performances of weighting methods in other types of watersheds still deserve site-specific research.

6 Conclusion

In order to weight climate models based on the impact variable and to quantify its influences on the impact assessment, this study assigns weights to an ensemble of 29 CMIP5 GCMs over two watersheds through a group of weighting methods based on GCM-simulated streamflow time series. Streamflow series are simulated by separately inputting the raw and bias-corrected GCM simulations to hydrological models. 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 mean hydrographs and reducing the biases of multi-model ~~means~~mean 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~~are brought close to the observations by the bias correction method. Consequently, the weights assigned to climate simulations become similar to each other, resulting in similar multi-model means and uncertainty of hydrological impacts for all 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 to end-users with a detailed explanation of the weighting procedure, as well as the results of using equal weighting method ~~to end-users~~.

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.

Chen, J., Li, C., Brissette, F. P., Chen, H., Wang, M., and Essou, G. R. C.: Impacts of correcting the inter-variable correlation of climate model outputs on hydrological modeling, Journal of Hydrology, 560, 326-341, https://doi.org/10.1016/j.jhydrol.2018.03.040, 2018.

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.

Hakala, K., Addor, N., and Seibert, J.: Hydrological Modeling to Evaluate Climate Model Simulations and Their Bias Correction, Journal of Hydrometeorology, 19, 1321-1337, https://doi.org/10.1175/jhm-d-17-0189.1, 2018.

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.

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.

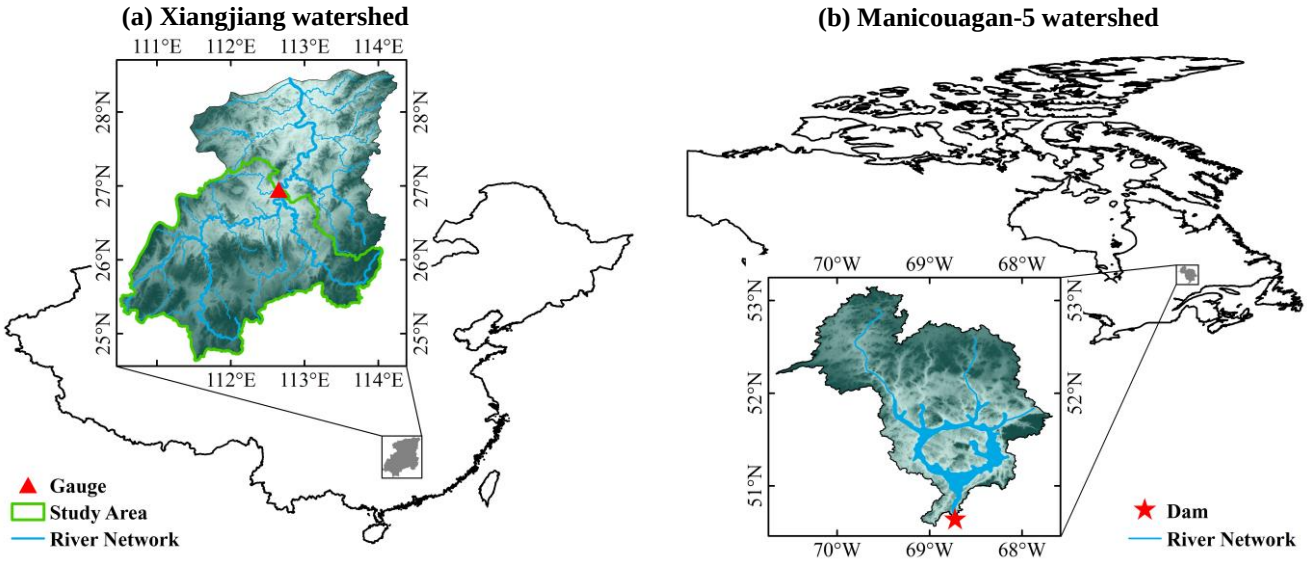

Figure 1. Locations of the (a) Xiangjiang and (b) Manicouagan-5 watersheds. (The study area in the Xiangjiang watershed is one of its sub-basins as the green boundary.)

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

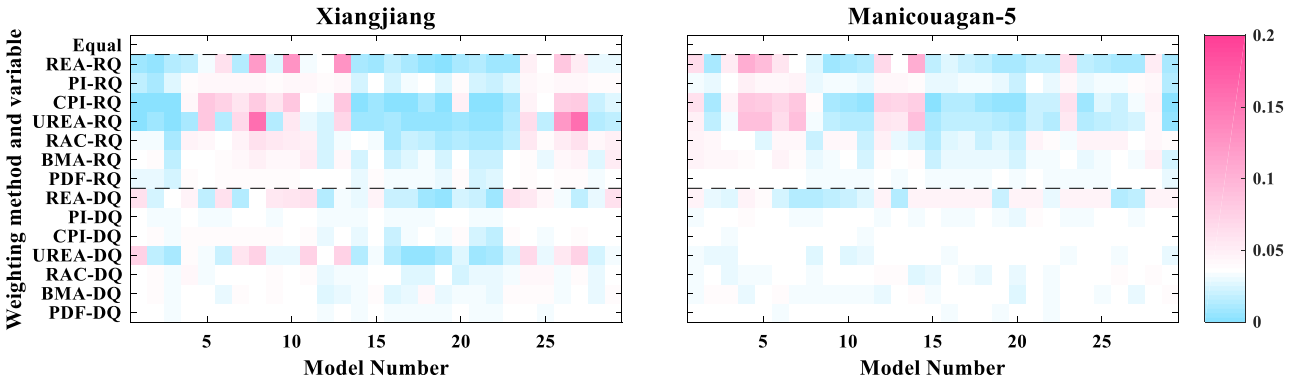

Figure 2. Weights assigned by equal weighting and 7 unequal weighting methods based on raw climate model data-simulated streamflow (RQ) and bias corrected data-simulated streamflow (DQ) for two watersheds. (Equal weight is presented in white, weights greater than equal are presented in red, and weights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

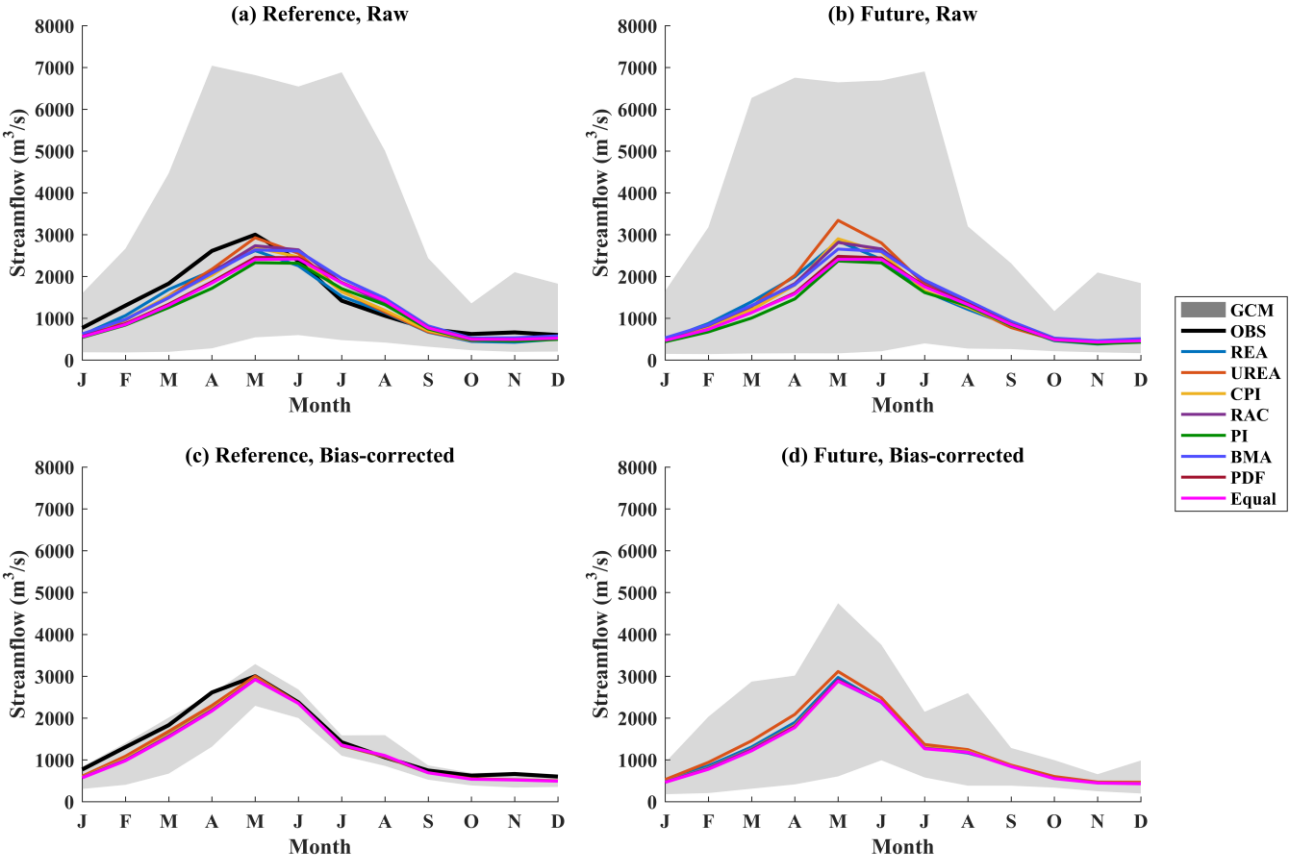

Figure 3. The envelope of monthly mean streamflows simulated by 29 raw and bias-corrected GCM outputs and the multi-model ensemble means of monthly mean streamflows weighted by 8 weighting methods based on GCM-simulated streamflows over the Xiangjiang watershed for the reference and future periods (OBS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

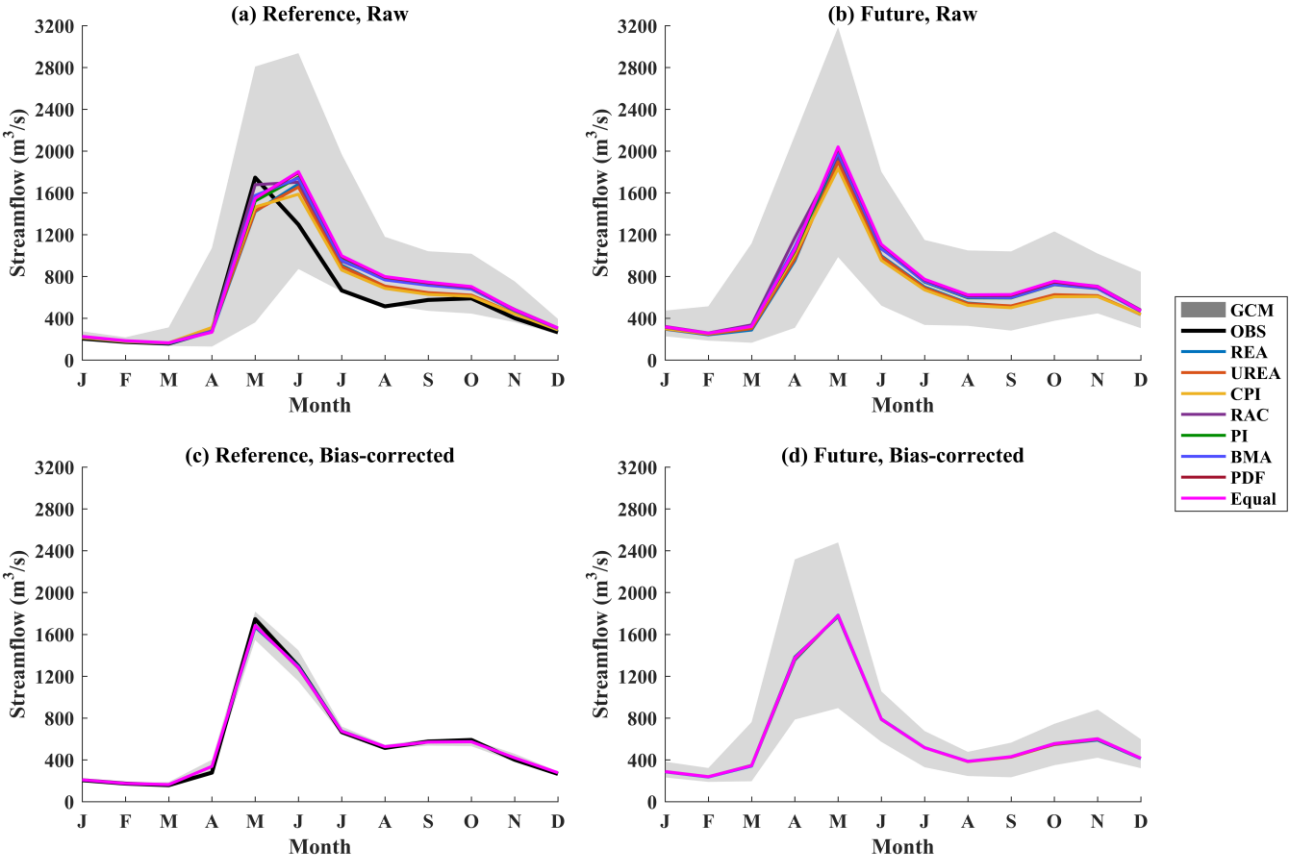

Figure 4. The same as Fig. 3 but for the Manicouagan-5 watershed.

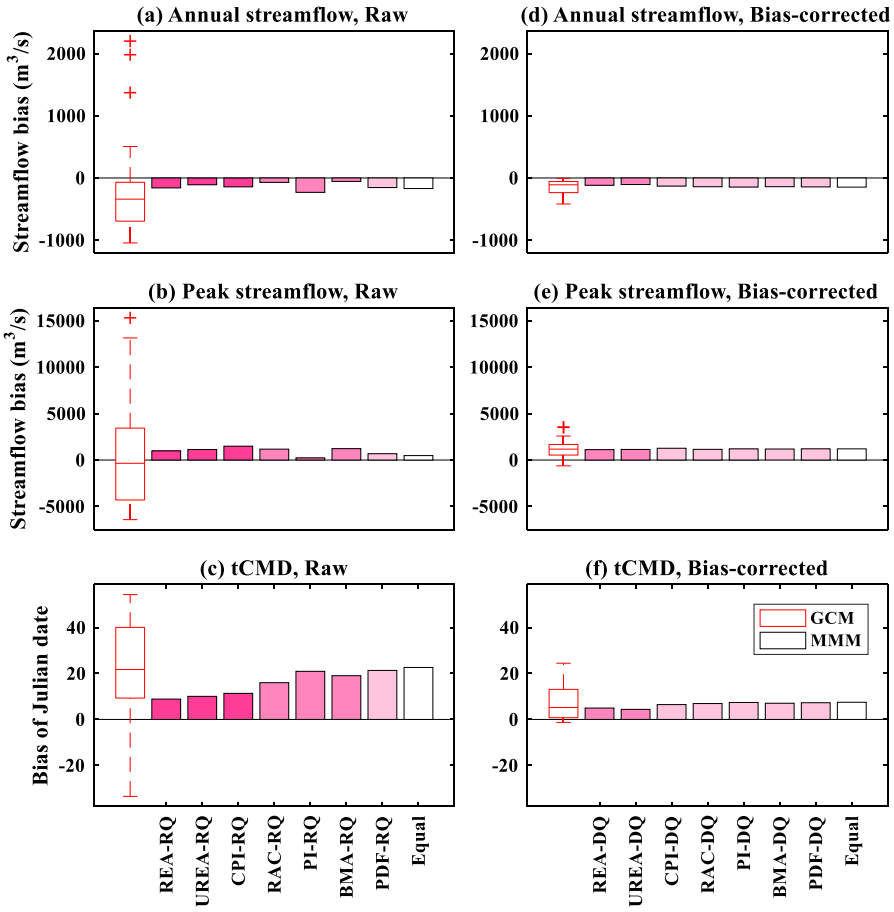

Figure 5. Bias in mean annual streamflow, mean peak streamflow and mean center of timing of annual flow (tCMD) simulated using 29 raw or bias-corrected GCM outputs and the multi-model means (MMM) combined by weights based on raw (RQ) and bias-corrected (DQ) GCM-simulated streamflows in the Xiangjiang 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.)

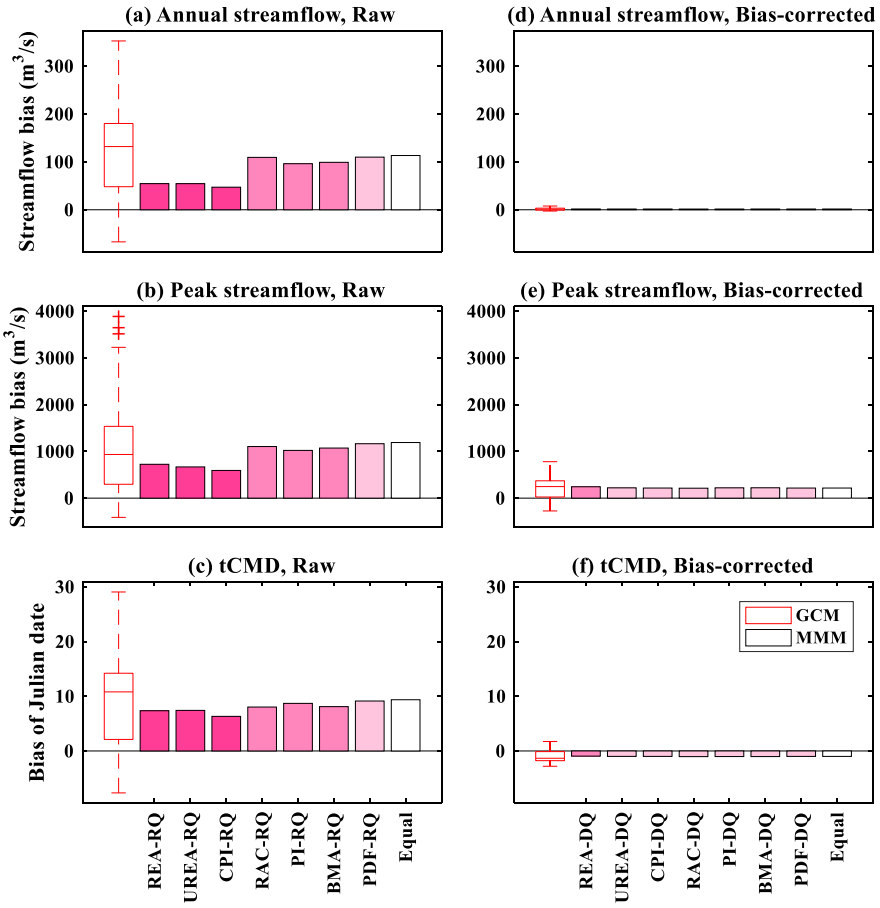

Figure 6. The same as Fig. 5 but for the Manicouagan-5 watershed.

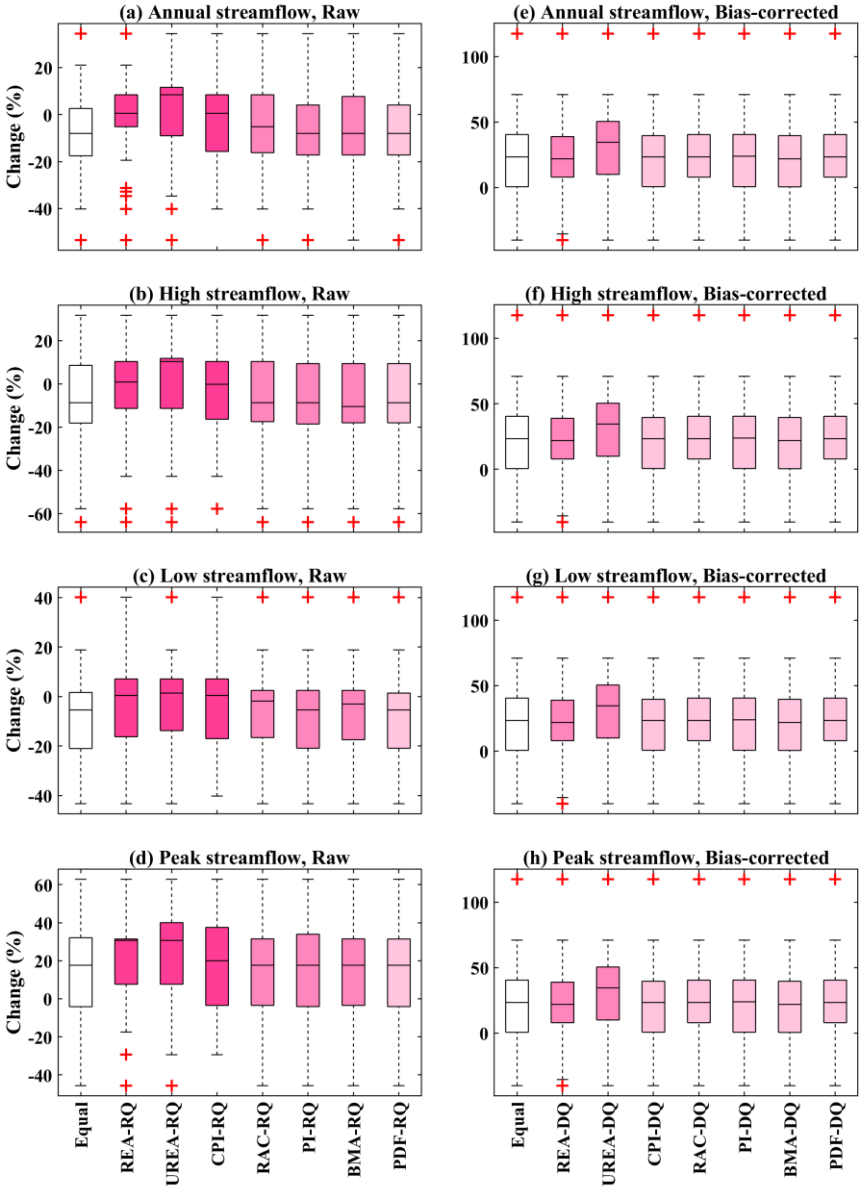

Figure 7. Box plot of changes in four hydrological indices calculated by raw or bias-corrected GCM-simulated streamflows in the Xiangjiang watershed. The changes of hydrological variables were sampled through the Monte-Carlo approach based on the weights calculated using raw (RQ) or bias-corrected (DQ)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

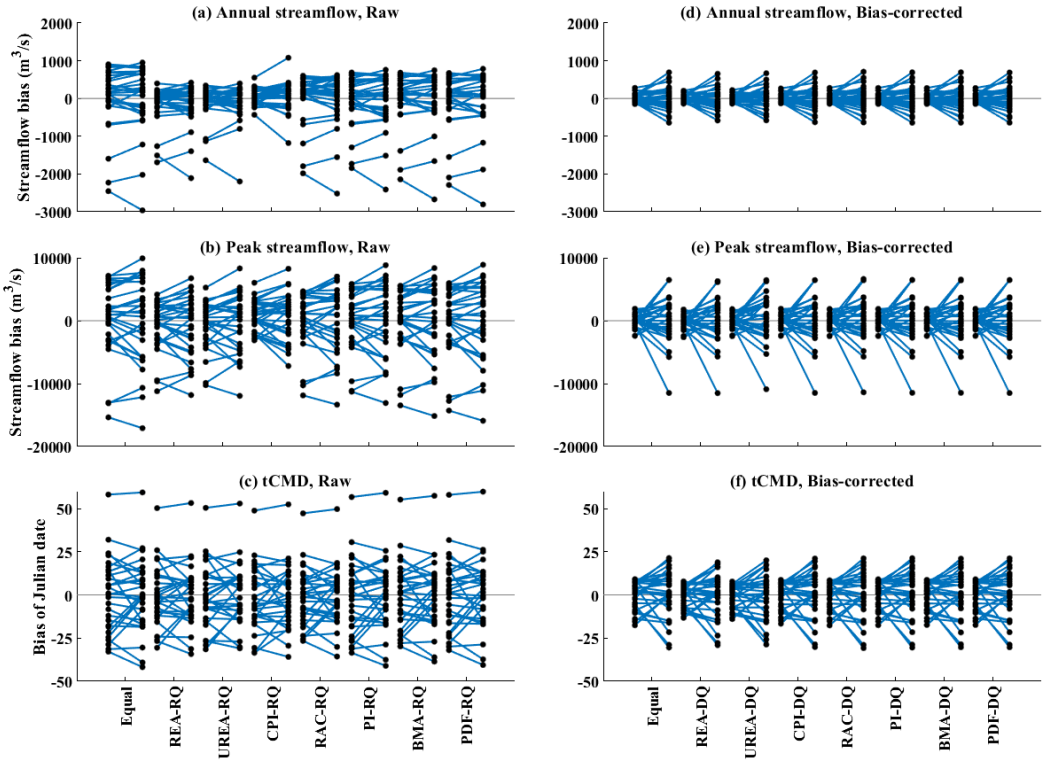

Figure 8. Bias in mean annual streamflow, mean peak streamflow and mean center of timing of annual flow (tCMD) of weighted multi-model ensemble mean in the out-of-sample testing over the Xiangjiang watershed. Twenty-nine sticks of each weighting method represent the results when each of 29 climate models 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.

Supplement of

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

Hui-Min Wang et al.

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

S1 Weighting methods

S1.1 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

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

where $\in$ 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 $\in$, this climate simulation is regarded to be reliable in the corresponding respect (i.e. $\in / \mathrm{abs}(B_i)$ or $\in / \mathrm{abs}(D_i)$ 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.

S1.2 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~~the 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~~the 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 $\sigma_B = 2.8$.

S1.3 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.

S1.4 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.

S1.5 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).

S1.6 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).

S1.7 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.

S2 Hydrological model: GR4J-6

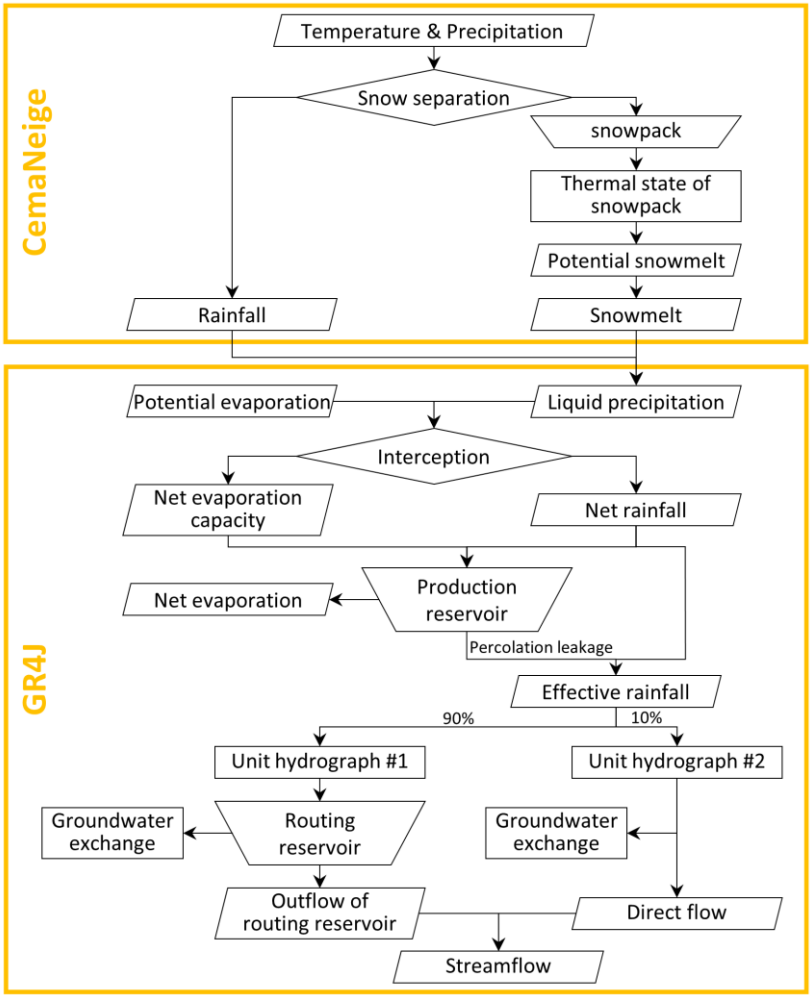

Figure S1.The flowchart of the GR4J-6 hydrological model.

S3 Supplementary results

S3.1 Weights of GCMs

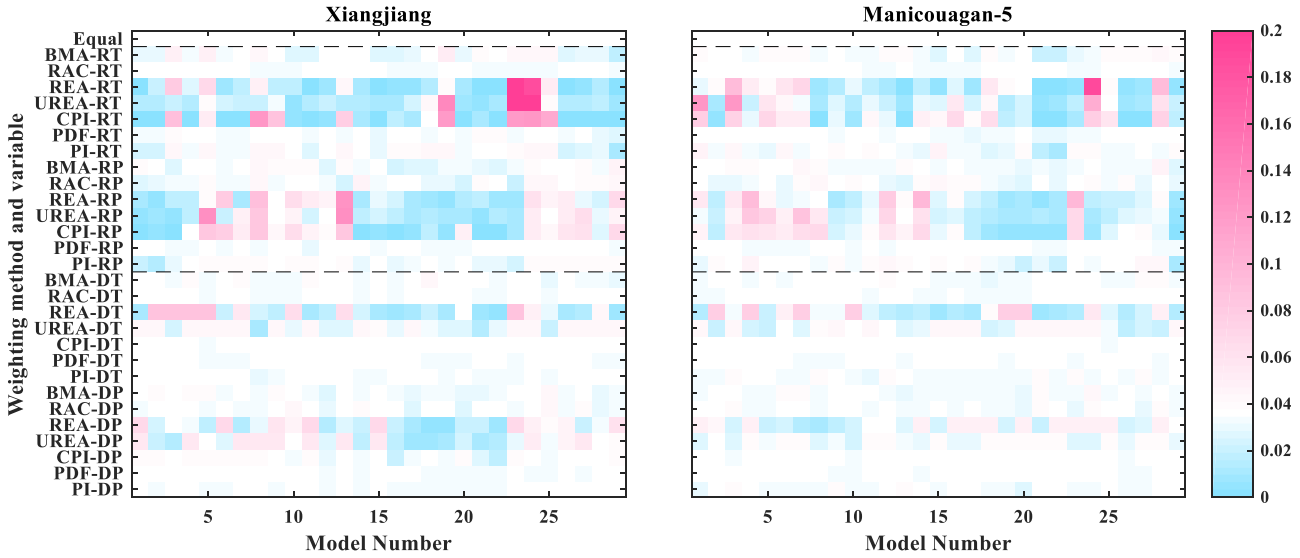

Figure S2. Weights assigned 8 weighting methods based on raw temperature (RT) and precipitation (RP) of GCM outputs and bias-corrected temperature (DT) and precipitation (DP) of GCM outputs for two watersheds. (Equal weight is presented in white, weights larger than equal are presented in red, and weights lower than equal are in blue.)

S3.2 Impacts on the hydrological regime

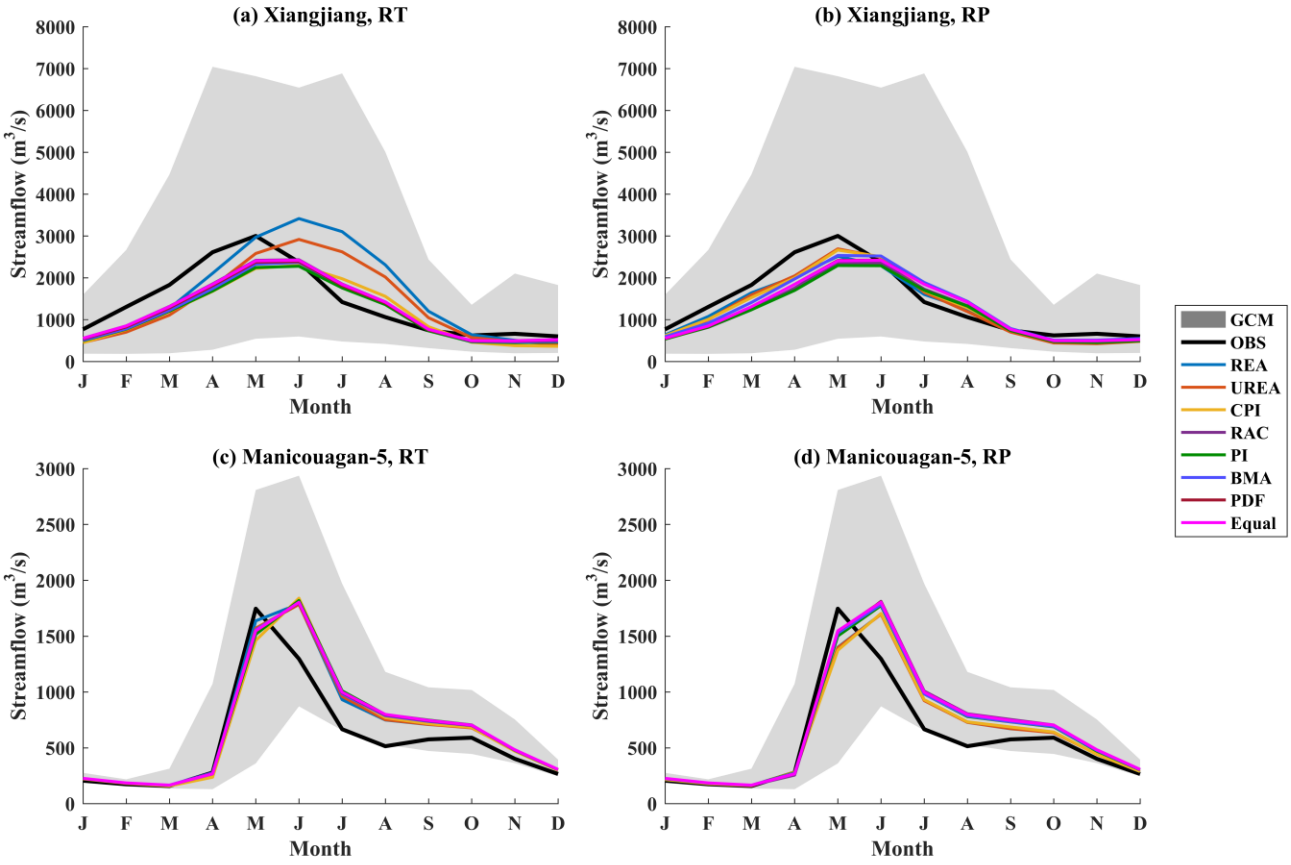

Figure S3. The envelope of monthly mean streamflows simulated by 29 raw and bias-corrected GCM outputs and the multi-model ensemble means of monthly mean streamflows weighted by 8 weighting methods based on raw temperature (RT) and precipitation (RP) of GCM outputs in both watersheds for the reference period (OBS = the hydrograph simulated from meteorological observation).

S3.3 Bias in multi-model mean

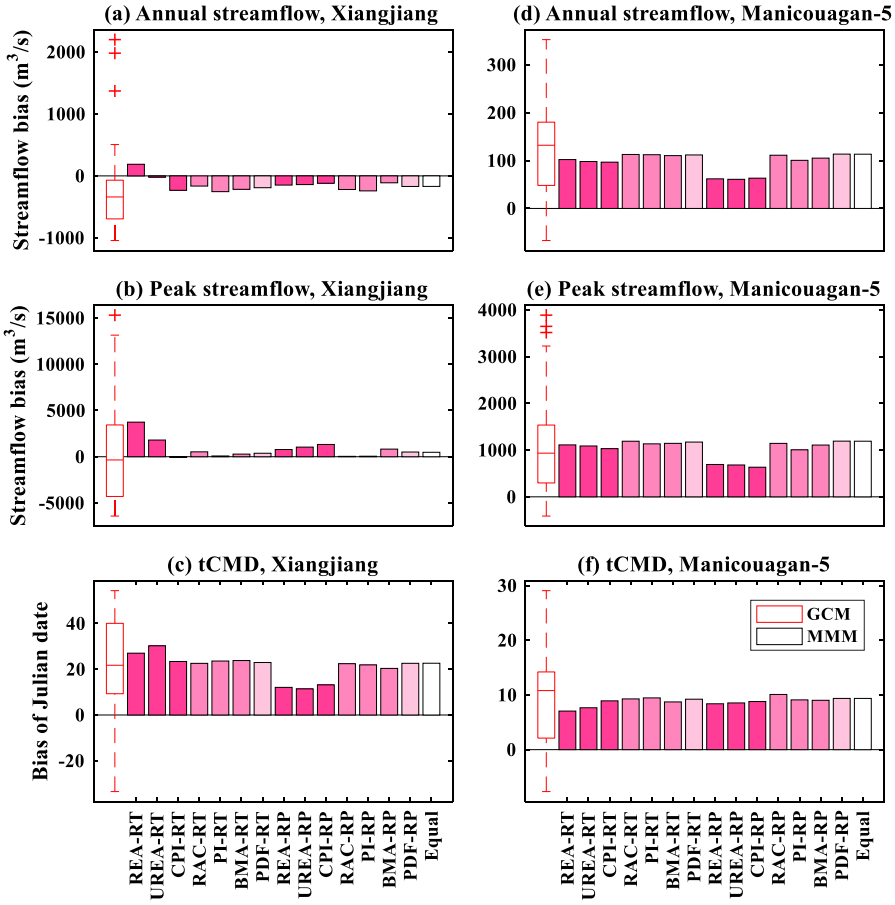

Figure S4. Bias in mean annual streamflow, mean peak streamflow and mean center of timing of annual flow (tCMD) simulated using 29 raw or bias-corrected GCM outputs and the multi-model means (MMM) combined by weights based on raw temperature (RT) and raw precipitation (RP) in both watershed for the reference period. (The depth of pink in bars of MMM represents the level of inequality of weights as indicated in Table 3.)

S3.4 Impacts on uncertainty

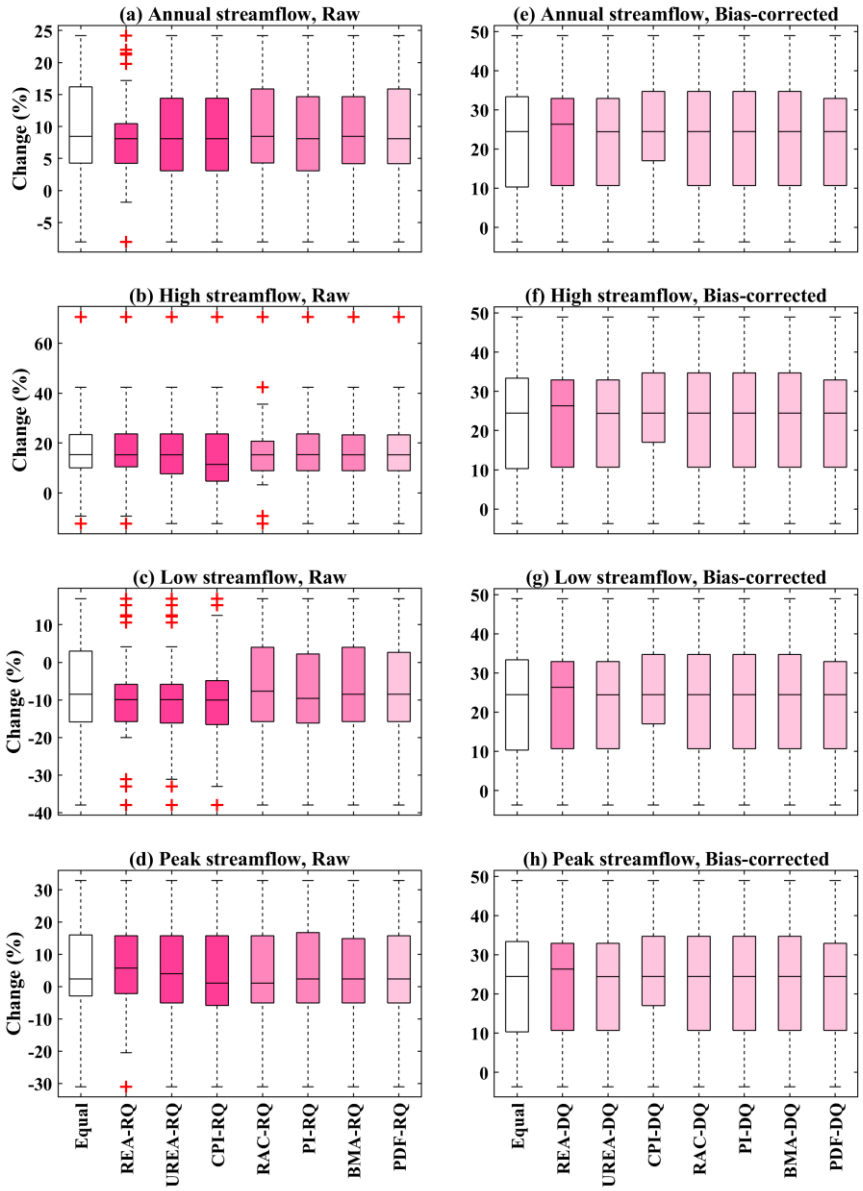

Figure S5. Box plot of changes in four hydrological indices calculated by raw or bias-corrected GCM-simulated streamflows in the Manicouagan-5 watershed. The changes of hydrological variables were sampled through the Monte-Carlo approach based on the weights calculated using raw (RQ) or bias-corrected (DQ) GCM-simulated streamflows. (The depth of pink represents the level of inequality of the weights.).

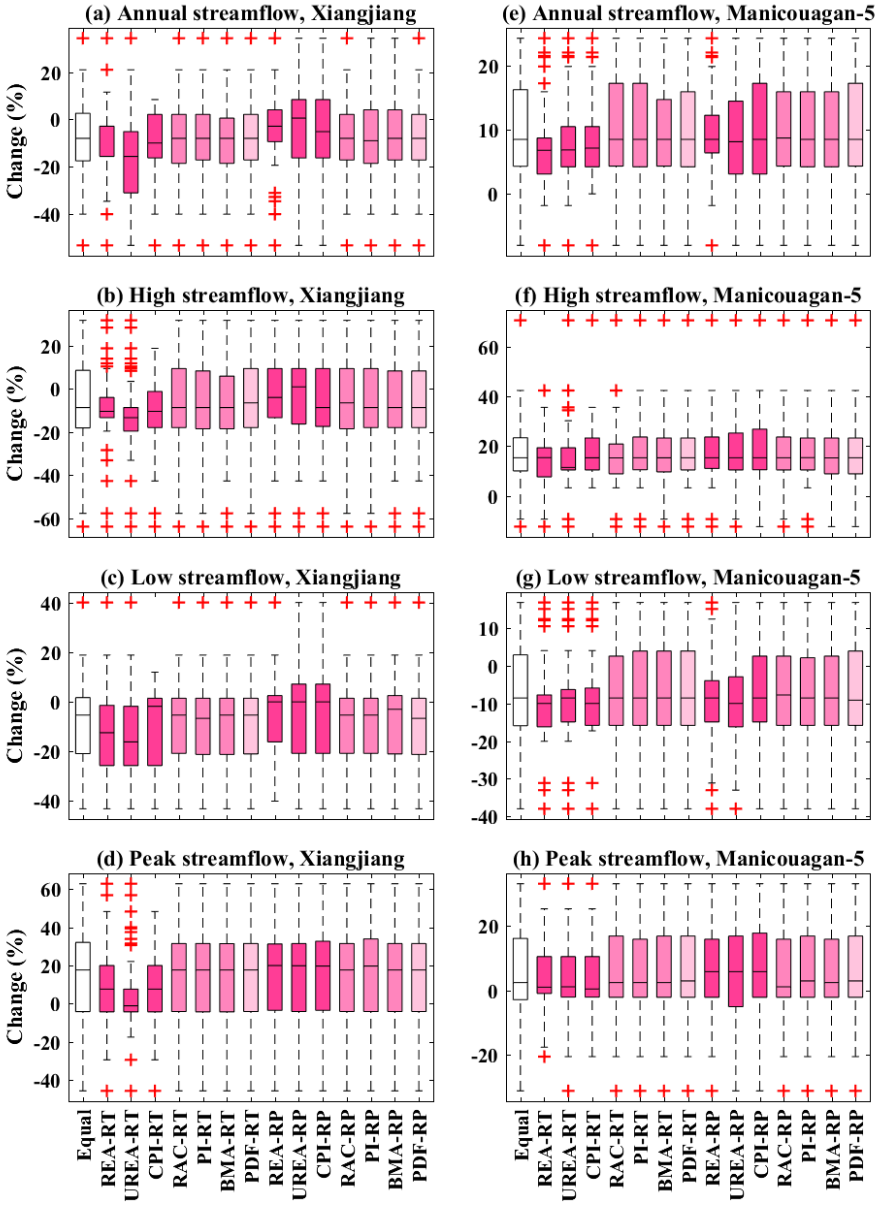

Figure S6. Box plot of changes in four hydrological indices simulated by raw GCM-simulated streamflows over both watersheds. The changes of hydrological variables were sampled through the Monte-Carlo approach based on the weights calculated using raw temperature (RT) and precipitation (RP) of GCM outputs. (The depth of pink represents the level of inequality of weights.)

S3.5 Out-of-sample Testing

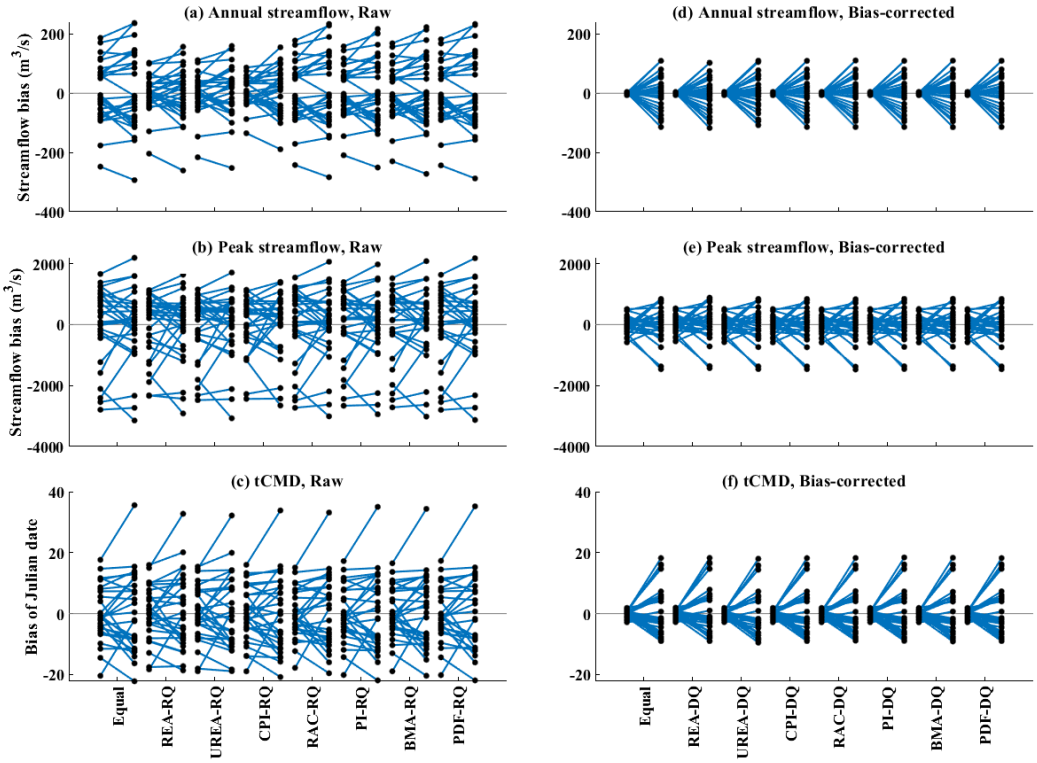

Figure S7. Bias in mean annual streamflow, mean peak streamflow and mean center of timing of annual flow (tCMD) of weighted multi-model ensemble mean in the out-of-sample testing over the Manicouagan-5 watershed. Twenty-nine sticks of each weighting method represent the results when each of 29 climate 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.

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.