https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

Assessing the Hydrological Impact Sensitivity to Climate Model Weighting Strategies

Mehrad Rahimpour Asenjan[1], Francois Brissette[1], Richard Arsenault[1], Jean-Luc Martel[1]

[1]Hydrology, Climate, and Climate Change Laboratory (HC^3), École de technologie supérieure, Université du Québec, 1100 Notre-Dame Street West, Montreal, Quebec, Canada H3C 1K3

Corresponding author: Mehrad Rahimpour Asenjan (mehrad.rahimpour-asenjan.1@ens.etsmtl.ca)

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

Abstract

Climate change impact studies rely on ensembles of General Circulation Model (GCM) simulations. Combining ensemble members is challenging due to uncertainties in how well each model performs. The concept of model democracy where equal weight is given to each model, is common but criticized for ignoring regional variations and dependencies between models. Various weighting schemes address these concerns, but their effectiveness in impact studies, which integrate GCM outputs with separate impact models, remains unclear.

This study evaluated the impact of six weighting strategies on future streamflow projections using a pseudo-reality approach, where each GCM is treated as “the true” climate. The analysis involved an ensemble of 22 CMIP6 climate simulations and used a hydrological model across 3,107 North American catchments. Since climate model outputs often undergo bias correction before being used in hydrological models, this study implemented two approaches: one with bias correction applied to precipitation and temperature inputs, and one without. Weighting schemes were evaluated based on biases relative to the pseudo-reality GCM for annual mean temperature, precipitation and streamflow.

Results show that unequal weighting schemes produce significantly better precipitation and temperature projections than equal weighting. For streamflow projections, unequal weighting offered minor improvement only when bias correction was not applied. However, with bias correction, both equal and unequal weighting delivered similar results. While bias correction has limitations, it remains essential for realistic streamflow projections in impact studies. A pragmatic strategy may be to combine model democracy with selective model exclusion based on robust performance metrics.

This study provides insights on how weighting affects hydrological assessments. It emphasizes the need for careful approaches and further research to manage uncertainties in climate change

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

impact studies. These findings will help improve the accuracy of climate projections and improve the reliability of hydrological impact assessments in a changing climate.

Keywords: Climate model weighting, climate change, hydrology, uncertainty, impact assessment, streamflow projections, pseudo-reality

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

1. Introduction

To assess the impacts of climate change on hydrology, researchers often rely on projections from global and regional climate models (GCMs and RCMs) (Chen et al., 2012; Hagemann et al., 2013; Reshmidevi et al., 2018). Typically, outputs from these models are post-processed (i.e., downscaled and/or bias-corrected) before being used by hydrologic models to simulate future hydrologic conditions (e.g., Raulino et al., 2021). The varying spatial and temporal resolutions, along with differences in the representation of physical processes and feedback mechanisms among GCMs lead to diverse climate sensitivities and a broad range of future climate projections. This variability is widely recognized as a primary source of uncertainty (Hausfather et al., 2022; Li et al., 2023; Murphy et al., 2004; Prein et al., 2020; Stainforth et al., 2007), further complicated by other sources of uncertainty (Merrifield et al., 2020; H. Wang et al., 2020).

Using ensembles of climate models is widely accepted as the best strategy to tackle this uncertainty (Giuntoli et al., 2018; Tebaldi & Knutti, 2007). However, there is no consensus on the most effective method to integrate the outcomes from multiple GCMs. Traditionally, these simulations have been combined by treating each climate model as equally plausible (e.g. Lawrence et al., 2021), a practice known as "model democracy." This approach assumes all models are equally capable of simulating past and future climates (Chen et al., 2017; Knutti, 2010). While model democracy has been successful in replicating the mean state of the observed historical climate (Reichler & Kim, 2008), its applicability and reliability in future impact assessments remain uncertain.

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

Model democracy is critiqued primarily for two reasons. First, GCMs' performance in reproducing climatic patterns varies by location and variable (Abramowitz et al., 2019), suggesting model democracy might not be the best choice in regions where some models are more reliable (Knutti et al., 2013; Lorenz et al., 2018). Second, averaging equally weighted models assumes independence within an ensemble. However, this assumption is often proven incorrect, especially in ensembles like CMIP5 and CMIP6 (Sanderson et al., 2017), since simulations from the same research group may differ only in resolution, and there has been extensive sharing among climate modeling centers, including shared coding and parameterization schemes (Eyring et al., 2019; Knutti et al., 2010). Consequently, the number of truly independent models in these ensembles is likely lower than it appears (Merrifield et al., 2020), which can skew results by duplicating similar information and adding little knowledge to the ensemble (Knutti et al., 2017; Wang et al., 2019).

To mitigate these issues, several studies have explored assigning different weights to climate model simulations based on historical performance, demonstrating more accurate projections compared to simple averaging (e.g. Lorenz et al., 2018; Palmer et al., 2023; Yuan et al., 2020). Others have accounted for model interdependence in their weighting schemes (Brunner et al., 2019; Di Virgilio et al., 2022; Easterling et al., 2017; Liang et al., 2020; Massoud et al., 2019; Sanderson et al., 2015, 2017). However, selecting an ideal subset of climate simulations that considers interdependence is challenging and somewhat subjective (Herger et al., 2018), with a risk of reduced accuracy due to inappropriate weighting (Weigel et al., 2010).

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

In hydrological impact studies, a common method to weight or select GCMs assesses their capability to accurately depict historical climate conditions like temperature and precipitation (Chen et al., 2017; Kolusu et al., 2021; Massoud et al., 2019; Padrón et al., 2019; Ruane & McDermid, 2017). While some studies highlight the benefits of weighting (e.g., Massoud et al., 2019), others note that weighting climate models only slightly affect streamflow projections derived from GCMs (e.g., Chen et al., 2017; Kolusu et al., 2021). A key concern is that weights based solely on climate model outputs like temperature and precipitation for impact assessments may not fully account for the complex relationship between climate variables and hydrological responses, potentially limiting the true utility of climate models in representing hydrological changes (Wang et al., 2019; Wootten et al., 2023).

Recent impact assessment studies have utilized streamflow values to weigh simulations (Castaneda-Gonzalez et al., 2023; Dong et al., 2021; Wang et al., 2019; Yang et al., 2017). For instance, Castaneda-Gonzalez et al., 2023, found that unequal weights improve the accuracy of representing mean annual and seasonal hydrographs during the reference period. Wang et al., (2019), noted that assigning different weights to climate models improves simulation accuracy and reduces biases in average predictions using raw GCM outputs during the reference period. However, the impact of weighting diminished once bias correction was applied to the GCM outputs. While improving accuracy during the reference period is vital, comprehensive out-of-sample testing is crucial to validate these methods for future projection periods (Abramowitz et al., 2019; Herger et al., 2018).

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

The effectiveness of climate model weighting is often benchmarked against equal weights (model democracy) by evaluating their performance in reproducing observed climate variables over a reference period. Without future observations, assessing the robustness of weighting schemes for future projections, especially for streamflow projections, is challenging. Thus, most studies on the efficiency of climate model weighting for future streamflow projections focus on whether unequal weights produce different future projections.

One approach to overcome this challenge is pseudo-reality (or model-as-truth) testing, which involves selecting a climate model simulation as a "pseudo-reality" and treating it as true observed data for both reference and future periods. By comparing different weighting schemes against this pseudo-reality, researchers can infer their effectiveness for future projections. This approach has been used in several studies (Chen et al., 2020; Hernanz et al., 2022; Mendoza Paz & Willems, 2023).

This study aims to assess the impact of various GCM weighting strategies on hydrological impact assessments through a model-as-truth (pseudo-reality) experiment. It investigates six different weighting schemes, including equal and random weighting as benchmarks, to determine the significance of individual climate simulations in hydrological impact studies. By conducting multiple iterations of the pseudo-reality method and considering factors like climate variables (e.g., temperature and precipitation) and geographic regions. This comprehensive approach seeks to thoroughly examine the sensitivity of climate model weighting and ensemble averages. The objective is to understand the complex interactions between weighting schemes and their

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

effects on hydrological assessments, thus providing valuable insights into the broader implications of climate change impact studies.

The research is driven by two primary objectives: first, to explore the influence of different climate model weighting techniques on the accuracy and reliability of hydrological assessments; and second, to ascertain how the selection of assessment criteria influences the outcomes of climate model weighting in hydrological assessments. Through these objectives, the study seeks to provide meaningful insights into the effectiveness and dependability of various weighting techniques in enhancing the accuracy of hydrological assessments related to climate change.

2. Materials and Methods

2.1 Study Area and Data

In this study, catchments were selected from the comprehensive HYSETS database, which includes data from 14,425 catchments across North America (Arsenault et al., 2020b). For our analysis, 3,107 catchments were chosen to ensure coverage across the entire North American continent. The selection criteria included a minimum drainage area of 500 km² and at least 10 years of data availability, as dictated by the requirements of the hydrological models used. The spatial distribution of these catchments is illustrated in Figure 1. Additionally, the meteorological data required for our study were obtained from the ERA5 reanalysis dataset. This dataset has

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

been demonstrated to perform comparably to observation station data in hydrological modelling, without the problems related to missing data, thus ensuring complete temporal coverage (Tarek et al., 2020).

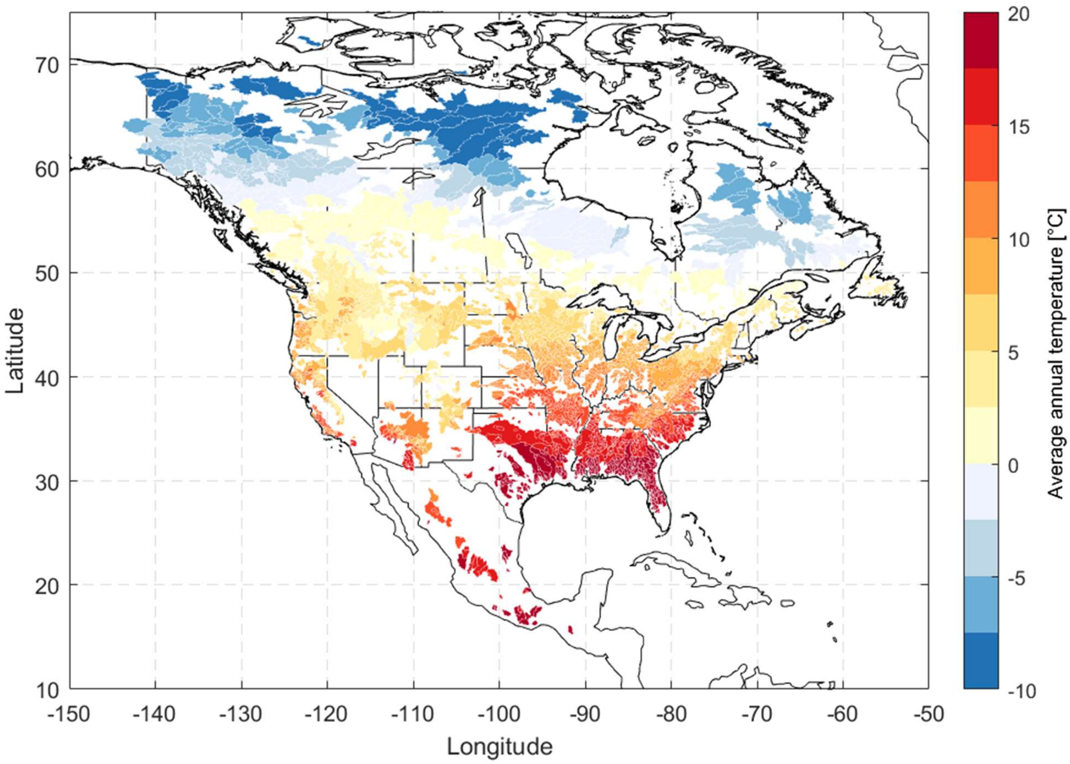


Figure 1. Map of the 3,107 catchments used in this study. The color code represents the mean annual temperature over each catchment. In the case of nested catchments, the smaller ones were plotted on top of larger catchments.

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

2.2 Modelling Chain

Following the standard procedures for hydrological climate change impact analysis, a top-down hydroclimatic modeling chain was used (as outlined in Arsenault et al., 2020a; Rahimpour Asenjan et al., 2023). Precipitation and temperature data were extracted from 22 CMIP6 climate models under the SSP5-8.5 scenario for both the reference and future periods. Table 1 lists the 22 CMIP6 GCMs used in this study, along with their corresponding Equilibrium Climate Sensitivity (ECS) values. ECS is a metric indicating the expected rise in Earth's average surface temperature in response to a doubling of carbon dioxide concentrations in the atmosphere, relative to pre-industrial levels, upon reaching equilibrium. The reference period for this analysis is 1971–2000, with future climate projections covering the period 2071–2100. Figure 2 displays the projected changes in temperature and precipitation between the reference (1971-2000) and future (2071-2100) periods for all 22 GCMs. The ECS values among the GCMs varied between 1.83 to 5.62 °C, highlighting the diverse responses of different models to climate change scenarios and emphasizing the potential significance of weighting model selection.

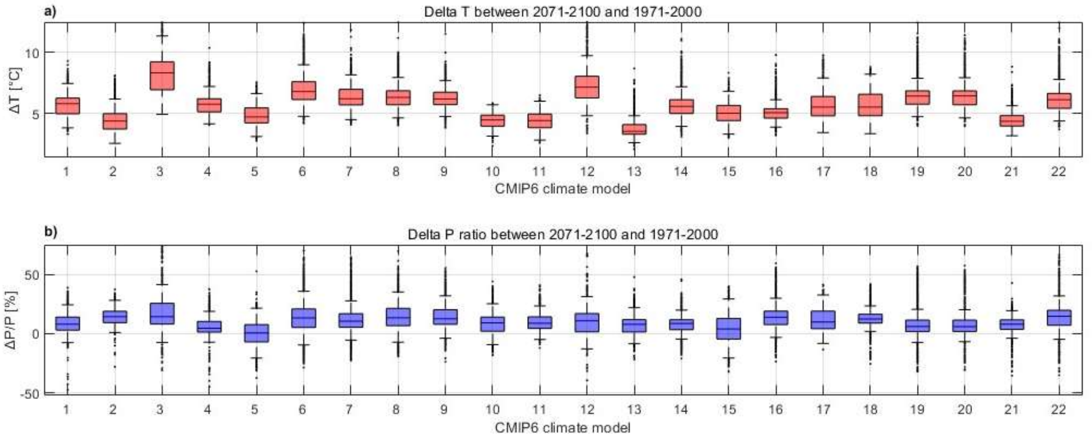
Table 1. The 22 GCMs selected in this study and their corresponding ECS. ECS values were taken from either 1- Tokarska et al., (2020) or 2-Hausfather et al., (2022)

GCM	ECS	Modeling center	ID Number
CanESM5	5.62[1]	CCCma	3
NESM3	4.68[1]	NUIST	22
IPSL-CM6A-LR	4.52[1]	IPSL	12

EC-Earth3-Veg	4.3[1]	EC-Earth-Consortium	8
EC-Earth3-CC	4.23[2]	EC-Earth-Consortium	7
EC-Earth3	4.2[1]	EC-Earth-Consortium	6
EC-Earth3-Veg-LR	4.2[2]	EC-Earth-Consortium	9
GFDL-CM4_gr1	3.89[2]	NOAA-GFDL	20
GFDL-CM4_gr2	3.89[2]	NOAA-GFDL	19
ACCESS-ESM1-5	3.88[1]	CSIRO	4
KIOST-ESM	3.36[2]	KIOST	13
MRI-ESM2-0	3.14[1]	MRI	16
MPI-ESM1-2-HR	3.02[2]	DKRZ	5
BCC-CSM2-MR	3.01[1]	BCC	1
MPI-ESM1-2-LR	2.98[2]	MPI-M	15
FGOALS-g3	2.87[2]	CAS	2
GFDL-ESM4	2.62[1]	NOAA-GFDL	21
NorESM2-LM	2.60[1]	NCC	17
MIROC6	2.57[1]	MIROC	14
NorESM2-MM	2.49[2]	NCC	18
INM-CM5-0	1.92[1]	INM	11

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

INM-CM4-8	1.83[1]	INM	10



165 **FIGURE 2: Projected temperature (a) and precipitation (b) changes between the reference (1971-2000) and future (2071-2100) periods over all 3,107 catchments for all 22 GCMs.**

The study involves two experiments. In the first experiment, uncorrected (raw) GCM data is used. For the second experiment, the multivariate bias correction (MBCn) method (Cannon, 2018) is applied to the climate data, noting that bias correction was performed exclusively using pseudo-reality GCM data. Subsequently, both the raw and bias-corrected climate data were used as inputs for a pre-calibrated hydrological model, which then generated streamflow simulations.

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

The HMETS lumped rainfall–runoff model was used for simulating streamflow (Martel et al., 2017). The HMETS model has demonstrated effective performance in previous hydrological studies (e.g., Tarek et al., 2021). It was calibrated using the Kling-Gupta Efficiency (KGE) objective function (Kling et al., 2012; Kling & Gupta, 2009) with streamflow observation data and ERA5 data spanning 1981-2018. The calibration process's duration varied depending on the availability of streamflow data for each catchment, requiring at least 10 years of observation data, including a 2-year warm-up period, entailed and 10,000 model evaluations using the SCE-UA (Shuffled Complex Evolution - University of Arizona; Duan et al., 1994) algorithm.

2.3 Overview of the weighting strategies

In this study, six weighting methods were employed to aggregate the outcomes of the hydrological model, as detailed in Table 2. These methods are described below.

Table 2. Weighing methods used in this study

Method	Description	References
RAC	Evaluating how closely models match observational series in terms of annual cycles	Wang et al., 2019
REA	Weights are assigned based on independence and convergence, considering the models' consistency and convergence towards collective projections.	(Giorgi & Mearns, 2002)
Skill	Weights are assigned based on the skill of reproducing the annual means, prioritizing models with higher skill.	(Sanderson et al., 2017)
BMA	Weights are assigned based on Bayesian model averaging of equilibrium climate sensitivity (ECS) value	Massoud et al., 2023

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

Equal	Weights are assigned equally
Random	Weights are assigned randomly to models for benchmarking and comparison

2.3.1 Representation of the Annual Cycle (RAC)

The Representation of the Annual Cycle (RAC) skill score measures the similarity between a climate simulation series and the pseudo-reality series in terms of their annual cycles, as defined in equation 1. It calculates the correlation coefficient (*r*) between the monthly observed and simulated series, with the maximum correlation (r_0) set to 1 for this study. Additionally, the parameter $\sigma = \sigma_s / \sigma_o$ represents the ratio between the standard deviations of the monthly simulated series and the monthly observed series. The RAC method aims to quantify the degree of resemblance between the simulated and observed annual cycles (Wang et al., 2019).

$$RAC_i = \frac{4\,(1+r)^4}{(\sigma + 1/\sigma)^2 (1+r_0)^4} \quad (1)$$

2.3.2 Reliability Ensemble Averaging (REA)

The Reliability Ensemble Averaging (REA) technique assigns weights to GCMs based on the model's performance criterion, which evaluates how accurately it reproduces historical observations, and the model convergence criterion, assessing the extent to which a GCM aligns with the multi-model mean in future projections. This indicates its consistency and convergence toward collective model projections (Giorgi & Mearns, 2002).

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

The REA framework evaluates the reliability of a GCM based on several factors, including natural climate variability (ε), determined from the range between the maximum and minimum 20-year moving averages of yearly observations, as shown in equation 2. It also considers the bias (β_i)of a simulation compared to the observational climatological means and the distance (D_i) between the projected change by a given model and the REA-weighted mean change. If the absolute value of the bias or distance is smaller than the climate variability (ε), indicating that the model's deviation falls within natural variability, the climate simulation is considered reliable. This reliability condition is expressed as ε/$|\beta_i|$ or ε/$|D_i|$ being set to 1. The parameters *m* and *n* represent the weights assigned to the performance and convergence criteria, respectively, with both set to 1 in this study.

$$REA_i = \{[\frac{\epsilon}{abs\ (\beta_i)}]^m \times [\frac{\epsilon}{abs\ (D_i)}]^n\}^{1/mn} \quad (2)$$

2.3.3 Skill

The "Skill" weighting method assesses model performance relative to historical climate data to allocate weights to each model within ensemble Projections (Massoud et al., 2019; Wootten et al., 2020). Models that more accurately reflect observed data receive higher weights, thus having a greater influence within the ensemble. The weights, $W_{m,skill}(i)$, are calculated according to equation 3 (Sanderson et al., 2017), based on the RMSE distances ($\delta_{i(obs)}$) between each climate simulation and the pseudo-reality scenario. The index i corresponds to each individual model within the ensemble. The radius of model quality, D_q, determines the degree

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

to which models with lower skill are down-weighted, fixed at 0.9 similar to Massoud et al. (2019). By adjusting model weights according to their skill levels, this method favors models with superior performance while reducing the impact of less skillful ones.

$$W_{m,skill}(i) = e^{-(\frac{\delta_i}{Dq})^2} \quad (3)$$

2.3.4 Bayesian Model Averaging (BMA)

Bayesian Model Averaging (BMA) optimizes the likelihood function to ensure that the combination of models best matches the target distribution (Massoud et al., 2020). In this study, the ECS values are estimated by the IPCC AR6 as the target distribution, represented by a gamma distribution with a range of 2.5–4 °C and a peak near 3 °C (similar to Massoud et al., 2023). For each test, a variety of combinations (n=15,000) of model weights is systematically sampled to find those that result in model combinations with the highest likelihood of matching the desired target field.

2.3.5 Equal weights and random weights

Equal weights and random weights are used as benchmarks for comparison in this study. Equal weights allocate the same importance to each model in the ensemble, ensuring all models contribute equally to the final outcome. Random weights are assigned from a uniform distribution between 0 and 1. The weights for each catchment and experiment are randomized,

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

using one of the 22 GCMs as the pseudo-reality. Both equal and random weights are normalized to sum to 1.

2.4 Experiment Design

The main methodological steps are depicted in Figure 3. Specifically, Figure 3a illustrates the steps for evaluating the performance of each weighting method for both future precipitation and temperature, while Figure 3b shows similar steps for future streamflows. Given the potential risk of selecting one of the 22 GCMs as the pseudo-reality, where picking an outlier could skew results, each GCM is alternately used as the pseudo-reality, with the remaining 21 GCMs evaluated against it. The steps in Figure 3 are carried out for each catchment and for each of the six weighting methods, necessitating a total of 18,642 repetitions (3,107 catchments x 6 weighting methods). Weights are determined based on the similarity between each of the remaining GCMs and the one chosen as pseudo-reality over the reference period (1971-2000), with all weights normalized to sum to one. For the future period (2071-2100), weighted precipitation and temperature estimates are derived using the weights from the reference period. The bias between these weighted estimates and those from the pseudo-reality GCM is calculated for each catchment. This process is repeated 22 times, once for each GCM as pseudo-reality, resulting in 22 bias (b_i) values for each weighting scheme. To assess the performance of each weighting scheme, the median of these 22 bias values is used.

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

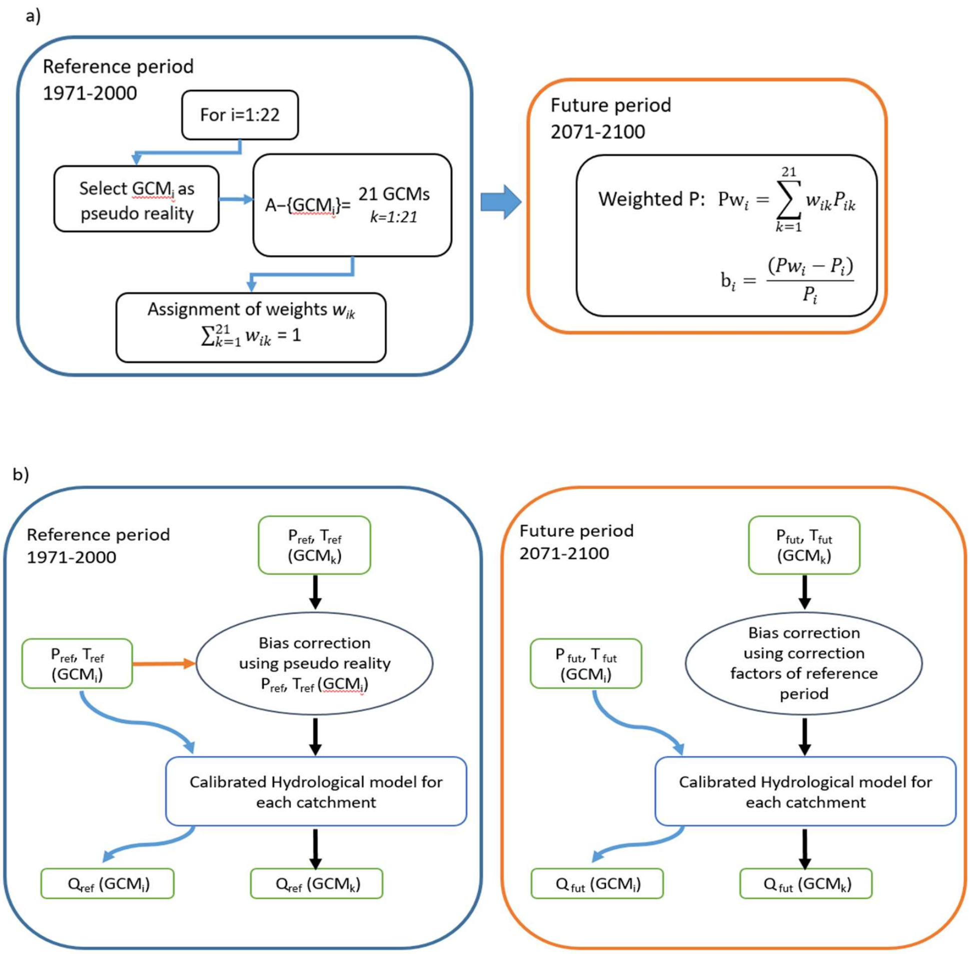


Figure 3. Main methodological steps (a) for the evaluation of the performance of each weighing method for precipitation (shown as P) and temperature (not shown). A = {GCM$_1$, GCM$_2$, GCM$_3$, …, GCM$_{22}$}, and bias = median {b$_1$, b$_2$, b$_3$, …, b$_{22}$}. Additional methodological (b) steps for the evaluation of the performance of each weighing method for streamflow metrics.

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

To assess the impact of weighting on streamflow values, it is necessary to include an additional step of bias correction, as detailed in Figure 3b. Numerous studies have indicated the necessity of bias-correcting precipitation and temperature values to obtain realistic outputs from impact model such as streamflows (Cannon et al., 2020; Dinh & Aires, 2023; Maraun, 2016). A significant issue is that GCMs do not directly produce streamflow values. While they do generate runoff values at each computational grid point, these values are not routed through a catchment outlet, which is essential for accurately simulating streamflows. Furthermore, the resolution of GCMs is often too coarse to effectively represent water fluxes in the stream network. To address this, a calibrated hydrological model (as previously described) was employed to generate streamflow for each catchment using precipitation (P) and temperature (T) data from the chosen pseudo-reality GCM (GCM_i). For the other 21 GCMs (GCM_k), precipitation and temperature values were bias-corrected to align with those of the pseudo-reality GCM_i, following standard practices in impact studies. This adjustment allows for the computation of streamflow values using the bias-corrected P and T with the calibrated hydrological model. It is important to note that the hydrological model is calibrated using observation data. However, the input data for the hydrological model comprises GCM data, which has been bias-corrected against the pseudo-reality, aligning with our aim to represent the pseudo-reality accurately. While the absolute performance of the hydrological model is important, our primary focus remains on accurately capturing the key underlying hydrological processes. As long as these processes are reasonably represented, the model's absolute

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

performance may not be of critical concern. This approach may deviate from conventional practices but is deemed necessary for the purpose of this study.

After applying bias correction, the streamflow characteristics of the 21 GCMs (GCM_k) should closely resemble those of the pseudo-reality GCM_i. Streamflow weights for each weighting method are determined based on two approaches: 1) assigning a 50%-50% weight to each precipitation and temperature, assuming that GCMs with precipitation and temperature characteristics closest to the pseudo-reality GCM should be weighted more heavily, and 2) basing them on streamflows computed using uncorrected precipitation and temperature. Although these two methods are expected to yield similar results, the non-linear response of hydrological models to precipitation and temperature may lead to differing weights. For the future period, pseudo-reality streamflow is generated using the pseudo-reality GCM P and T in the hydrological model, just as in the reference period. For the 21 remaining GCMs, P and T outputs are bias-corrected with the same factors used for the reference period, and streamflow projections are computed using the hydrological model. Streamflow biases are calculated as outlined in Figure 3b.

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

4. Results

4.1. Climate Variable Sensitivity to Weighting Methods

Figure 4 presents the results for all six weighting schemes for mean annual precipitation (prcptot). Specifically, it plots the difference between the median absolute bias of each method and that of equal weighting, represented as a colored circle centered on the centroid of each catchment. For Equal Weighting the median bias value is directly plotted. The median value is taken from the distribution of 22 values, corresponding to the 22 GCMs. Each model is taken in turn as the pseudo-reality, with weighting applied to the remaining 21 GCMs, as discussed in the methodology and presented in Figure 3. A bias of 0 for Equal Weighting (Figure 4-a) indicates a perfect prediction of the pseudo-reality. For the other five methods, a value of 0 signifies performance on par with Equal Weighting (equal biases). A red color indicates that the weighting method performs better than equal weighting, and a blue color indicates the opposite. This approach helps us discern the deviation of each method from Equal Weighting, aiding in understanding their relative effectiveness. Supplementary material Figures S1 and S2 show the median bias for each method.

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

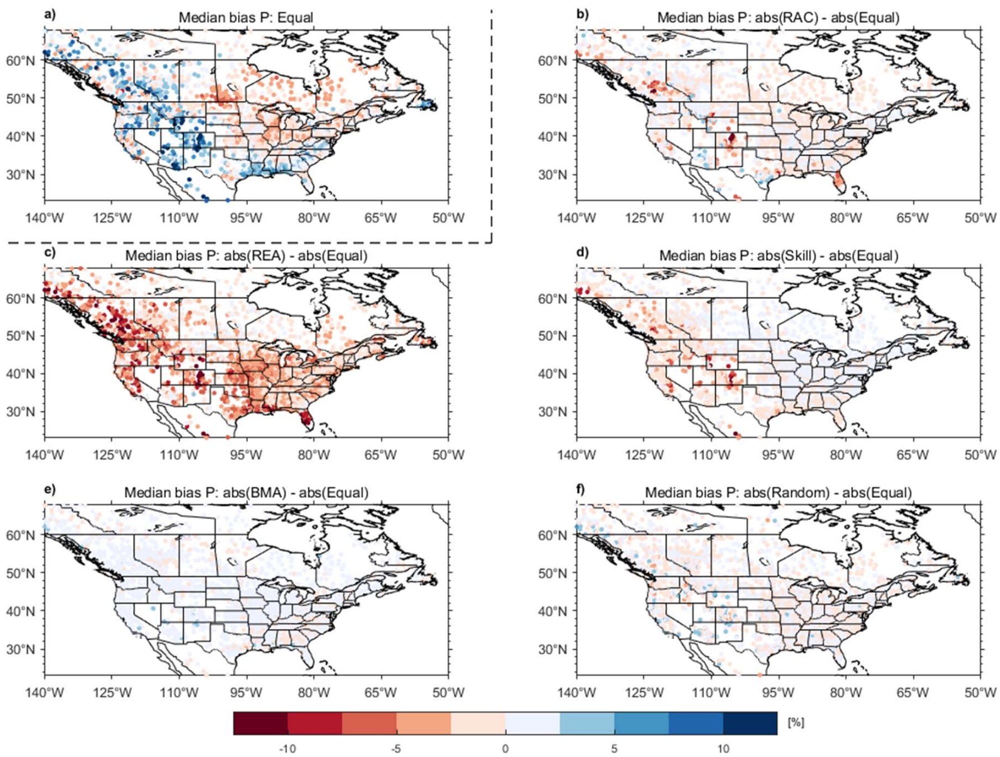


Figure 4. Difference in median absolute precipitation (prcptot) bias across all catchments for the future period (2071-2100). Equal weighting (a) is presented as the actual bias value, while the biases from all other methods (b-f) are expressed as differences relative to the equal weighting bias.

Results highlight the superior performance of the REA weighting scheme compared to other methods tested. The skill method performs better than equal weighting in the western half of the domain but slightly poorer in the eastern half. The other three methods produce results similar to equal weighting, even though BMA tends to be slightly worse and RAC slightly better. Furthermore, the analysis reveals a distinct regional performance pattern, with western and

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

southeastern catchments having a negative bias, and the opposite in the rest of the domain. The results underscore the importance of regional considerations in evaluating the effectiveness of different weighting schemes, confirming that geographic factors play a crucial role in the accuracy of climate projections.

Figure 5 presents results for mean annual temperature (tas) using the same format as Figure 4. In this case, the SKILL method outperforms the others, closely followed by the REA method. The other four methods (RAC, BMA of climate sensitivity, equal weights, Random) yield very similar results. The SKILL and RAC methods demonstrate particularly better performance over the Rockies, British Columbia and Alaska. The largest biases are observed in Northern Canada and Alaska. It should be noted that, despite the sharp color gradient observed in Figure 5, the overall median absolute biases remain small, always less than 0.25 (less than 25% of the original value) for all catchments.

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

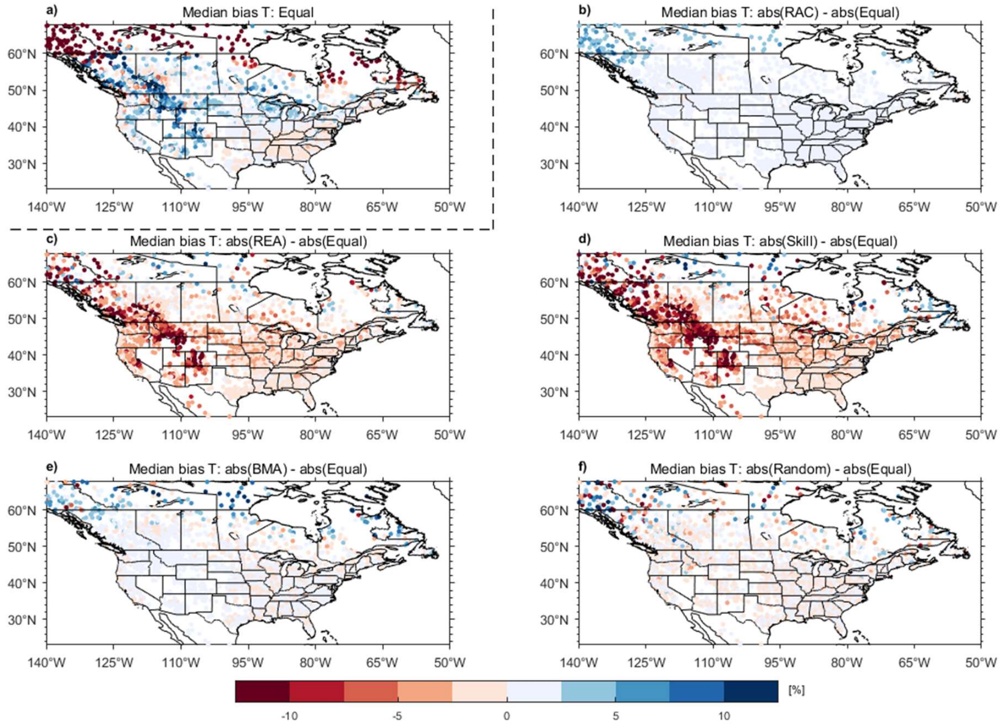


Figure 5. Same as Figure 4, but for mean annual temperature (tas).

Figure 6 presents the median bias for mean annual streamflow (Qm) in the same format as Figures 4 and 5. In this figure, the GCM weighting is equally based on uncorrected precipitation and temperature values over the reference period. To derive daily streamflow, precipitation and temperature data were bias-corrected to match those of the chosen GCM, considered the pseudo-reality. These corrected values were then utilized as inputs to the hydrological model, as detailed in the methodological section. The results indicate that all weighting methods yield nearly identical outcomes. This suggests that unequal weighting of climate models does not offer any significant advantage over the use of equal weights. Similar results are observed for the mean

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

of the maximum and minimum annual discharge values, as shown in supplementary material Figures S3 and S4.

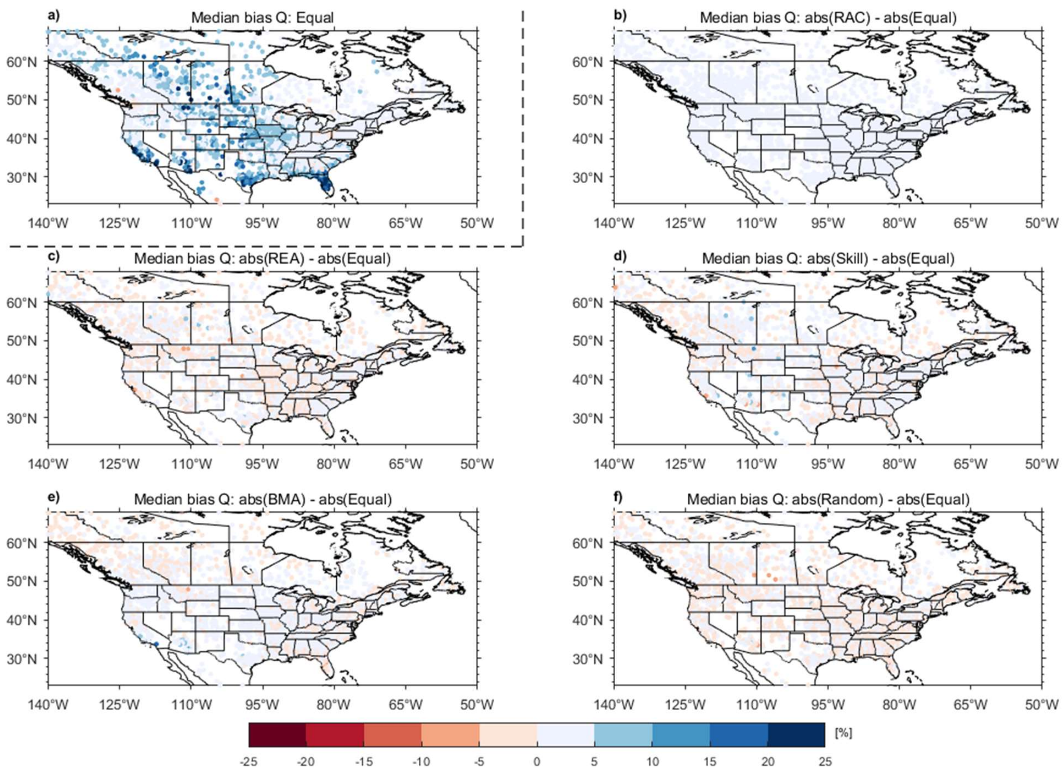


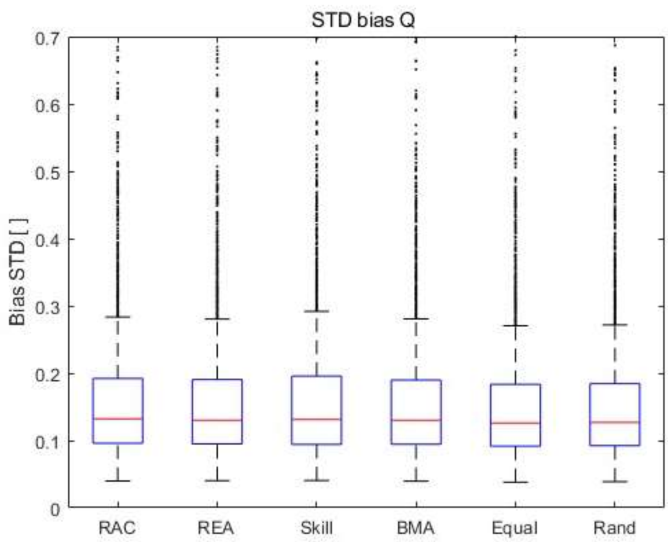
Figure 6. Same as Figure 4, but for mean annual streamflow (Qm).

Figures 4-6 have used the median as the representative metric to evaluate the distribution of 22 values, each derived from treating one of the 22 GCMs as the pseudo-reality target. While a good median performance is considered an important asset, it does not provide a complete assessment of performance. To gain a more comprehensive understanding of the performance

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

of each weighting method, Figure 7 displays the standard deviation of the distribution of the 22 bias values for Qm. The findings indicate that the standard deviation for all weighting methods is nearly identical. This strongly indicates that the performance of the weighting methods is comparable, regardless of which GCM is selected as the pseudo-reality. These results corroborate the findings from Figure 6, showing that equal weighting provides similar results to more complex weighting methods.

Additionally, similar outcomes are observed for the mean of the maximum and minimum annual discharge values, as detailed in the supplementary material (Figures S3 and S4). This consistency across different metrics and figures reinforces the conclusion that the choice of weighting method does not significantly affect the assessment of GCM performance in predicting future streamflow.



https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

Figure 7. Boxplot for the standard deviation of the distribution of the 22 bias values of mean annual streamflow (Qm).

Results from Figures 6 and 7 show that the bias correction step which is almost always used for precipitation and temperature prior to computing streamflow removes the advantage of some weighting methods as was seen for precipitation and temperature (Figures 4 and 5).

4.2 Streamflow Weighting without Bias Correction

To delve deeper into the matter of bias correction, a repeat of the streamflow weighting experiment was conducted without applying any bias correction. The weighting was carried out based on two different approaches:

- The first approach was based on uncorrected precipitation and temperature values, similar to the procedure in Figure 6, but without subsequent bias correction before calculating streamflows.
- The second approach relied on streamflows computed over the reference period using the non-bias-corrected precipitation and temperature values.

Omitting the bias correction of precipitation and temperature values before computing streamflows was expected to result in a broader range of streamflow outcomes. As explained in the methodology, GCMs exhibiting the smallest deviations in precipitation and temperature

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

when compared to the target pseudo-reality GCM are likely to produce streamflows closer to the pseudo-reality, thus receiving heavier weighting.

The outcomes of this experiment are showcased in Figure 8 (for the first approach) and Figure 9 (for the second approach), both of which illustrate the median bias for the mean annual streamflow discharge in the same format as Figure 6. The results from both figures are very similar, as hypothesized in the methodology, and are therefore discussed together. It is observed that the REA weighting method, with the Skill method trailing closely, results in biases that are mostly lower than those resulting from equal weighting, although the improvements are relatively modest. The other three weighting methods give results that are very similar to that of equal weighting.

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

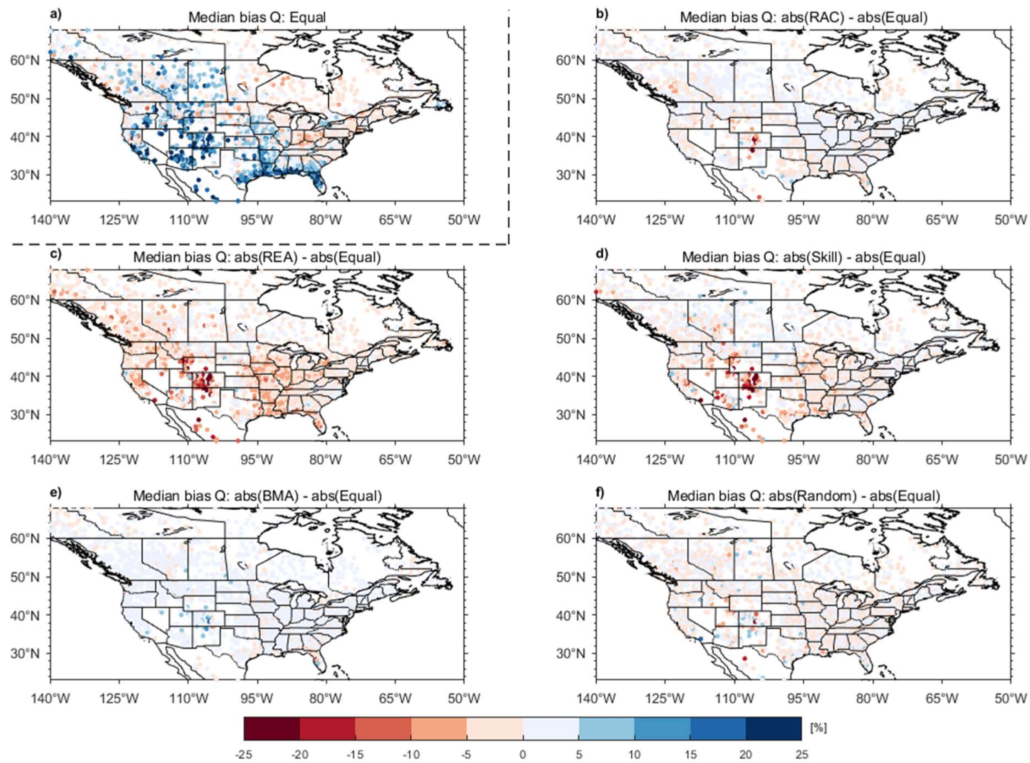
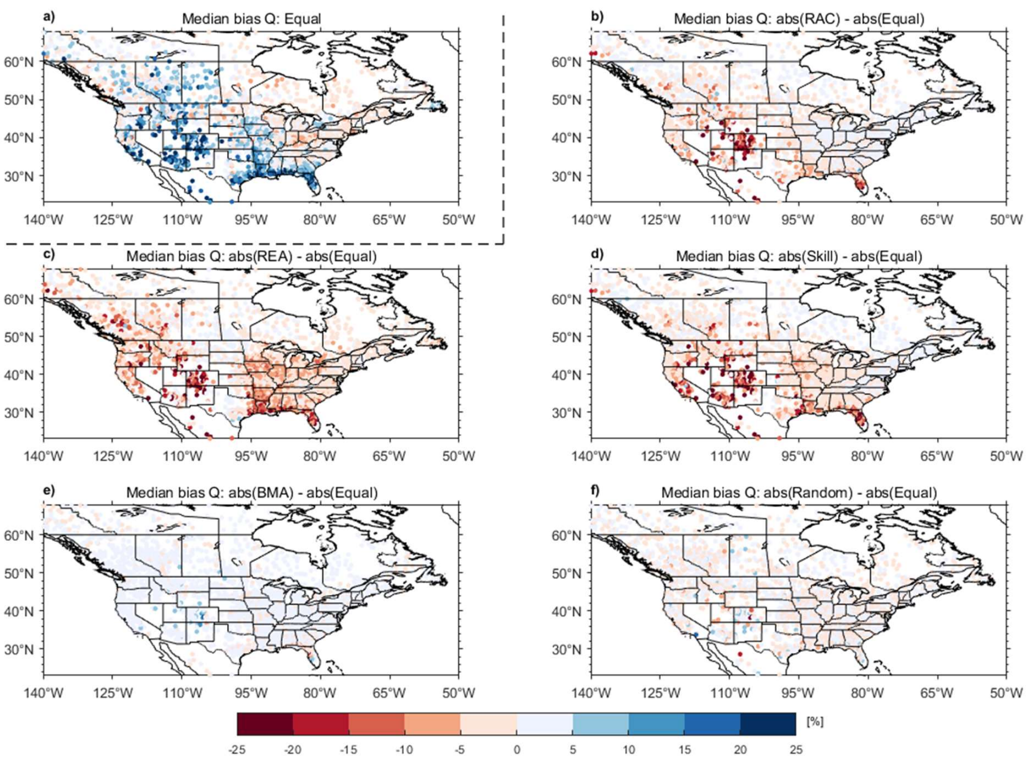


Figure 8. Same as Figure 4, but for mean annual streamflow (Qm) and using the first approach.

400 In both experiments, the biases are considerably larger than those observed in Figure 6. This pattern underscores the importance of bias correction in achieving more accurate projections of streamflow. It also suggests that the bias correction process effectively standardizes all temperature and precipitation projections against each other, thereby nullifying any potential benefits of employing more complex weighting methods over simple equal weighting.

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

A slight improvement is observed when weighting is based on streamflow performance rather than prcptot and tas. This improvement is likely due to the inherently nonlinear nature of the relationship between precipitation, temperature and streamflow. Streamflow-based weights are unaffected by the nonlinear relationship between climate and impact variables, and thus reflect the degree of agreement between GCM simulations and observed streamflow more accurately.



https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

Figure 9. Same as Figure 4, but for mean annual streamflow (Qm) and using the second approach

4. Discussion

4.1 Discussion overview

To assess the potential impacts of climate change and its associated uncertainties, climate change impact studies commonly utilize ensembles of climate simulations. Many of these studies treat all GCM projections as equally probable, in part because assigning weights to projections can be subjective and introduces an additional layer of uncertainty into the impact analysis. Selecting an appropriate weighting method is indeed challenging, and considerable debate about the best way to weigh climate models in impact assessment studies. A primary concern is that weights are often determined based on past performance, which may not be a reliable indicator of future accuracy.

One approach to assessing future accuracy is the pseudo-reality method. In this study, pseudo-reality was employed to examine the effects of six weighting strategies on future streamflow projections, aiming to provide more comprehensive insights into the impacts of various weighting methods. The pseudo-reality approach provides future projections against which other GCM projections can be assessed. Using each of the 22 GCMs as the pseudo-reality target in turn is an important methodological step to average out the impact of selecting a GCM with

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

either low or high-sensitivity. The median results presented above, therefore, provide a valid estimate of the expected performance of each weighting method.

4.2 Evaluation of Weighting Methods in Hydrological Impact Studies

In hydrological impact studies, the use of weights is an implicit practice. While the most common approach is equal weighting, binary weights (0 or 1) are also employed to either include or exclude specific climate projections, such as excluding SSP1-2.6 scenarios, for example. The goal of applying unequal weighting is to enhance the accuracy of the ensemble mean and to improve reliability through a more accurate assessment of the uncertainty associated with GCMs. In this context, our findings suggest that in the absence of a bias correction step, applying unequal weighting—particularly the Reliability Ensemble Averaging (REA) method—results in better projections for future precipitation, temperature, and streamflows. This improvement is consistent regardless of whether the weights are based on precipitation and temperature data or on streamflow data, with a notable enhancement for weights based on streamflow. These results align with previous studies, such as those by Castaneda-Gonzalez et al., (2023) and Wang et al., (2019). The results also show that the best weighting method for temperature (skill) differs from that for precipitation (REA), even though the latter also performs well for temperature. This introduces an additional layer of complexity when choosing a weighting approach.

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

To assess the effectiveness of the REA method, a test was conducted where model weights were inverted relative to their REA-calculated values. This meant that models assigned the least weight became the most heavily weighted, and vice versa. In theory, this should provide the worst possible weights and result in the largest possible biases. After inverting the weights (1/W) and renormalizing them to sum to 1, the resulting median bias values were evaluated. The inversion of REA weights results in notably increased bias values, as indicated by the darker colors in Figure 10. This observation serves to underscore the effectiveness of the REA method.

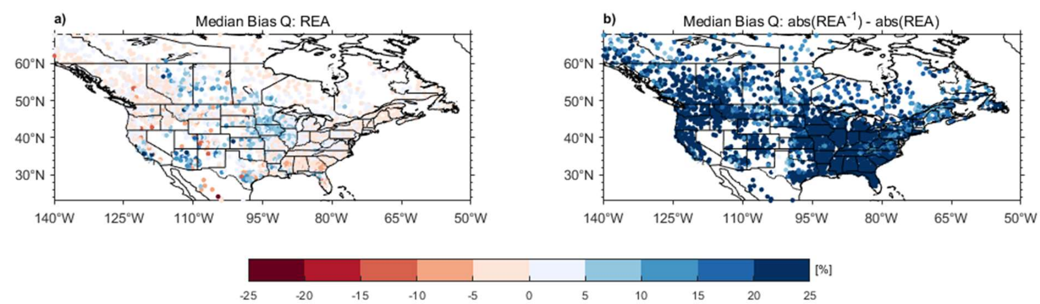


Figure 10. Same as figure 9 with a) REA and b) inverted REA weights

Conversely, the findings of this study suggest that when bias correction is applied, equal and unequal weighting methods lead to similar outcomes regarding streamflow projections. Weights were determined before applying bias correction because, after bias correction, all precipitation and temperature time series would closely align with the pseudo-reality time series, essentially leading to equal weights (Shin et al., 2020). Performing bias correction prior to running the hydrological model normalizes all climate projections over the reference period, effectively

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

diminishing the initial performance advantage of certain climate models. Looking towards future periods, the effectiveness of bias correction is influenced by the climate sensitivity of each GCM and the internal variability of the climate system (Chen et al., 2020), which can negate all benefits derived from computed weights.

Bias correction is often considered a necessary but flawed tool. Without it, impact studies would yield unrealistic streamflow projections. This process introduces several challenges (Maraun, 2016), including added uncertainty, the potential misrepresentation of extremes, the assumption that biases remain constant over time, and concerns regarding the manipulation of physically consistent data. Even advanced bias correction methods, such as the MBCn, which preserves the delta change signal and maintains multivariate properties and was used in this study, cannot fully overcome these issues. In hydrology, streamflow results from complex, non-linear interactions between precipitation and temperature, indicating that even minor modifications to time series can lead to significant changes in streamflow. Despite these challenges, bias correction remains indispensable for addressing issues related to climate model resolution, parameterization, and the imperfect representation of physical processes (Chen et al., 2021).

4.3 Embracing Model Democracy as a Middle-Ground Strategy

If unequal weighting does not significantly enhance hydrological impact studies, as shown in this study, then advocating for the principle of model democracy is justifiable, at least from a practical

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

perspective. This approach simplifies the modeling process by eliminating the need to assign weights within the impact study modeling chain.

A middle-ground strategy involves adopting a model democracy approach after excluding some poorly performing GCMs. This method can be equated to a binary [0, 1] weighting approach. Di Virgilio et al. (2022) have supported this as an advantageous strategy. However, the effectiveness of this approach necessitates careful selection of model performance metrics and dependence on reliable observations. These considerations are crucial for improving the accuracy and robustness of future change estimates and the uncertainties associated with them, as noted by Singh & AchutaRao (2020). The exclusion of GCMs might also be guided by factors other than performance, such as excluding models with a climate sensitivity considered too high (Hausfather et al., 2022; Rahimpour Asenjan et al., 2023), or based on more specific criteria, like omitting GCMs that do not physically represent the North American Great Lakes for a study focused on that region.

4.4 Implication of ensemble size on random weighting

An intriguing finding from this study was that random weighting yielded results comparable to those of equal weighting. For random weighting, a uniform distribution between 0 and 1 was used, and the weights were then normalized to ensure their sum was 1. This finding can be attributed to the large number of GCMs in the ensemble, as it is recognized that the ensemble mean from a large sample of GCMs typically is better than any individual GCM (e.g. Crawford et

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

al., 2019; Ganguly & Arya, 2023). In other words, the number of GCMs is large enough to compensate for the inclusion of poorly performing GCMS.

To investigate the impact of GCM ensemble size, an experiment test was conducted with a reduced ensemble of 7 randomly selected GCMs (one-third of the remaining 21 models, after choosing one as the pseudo-reality). The results of this experiment, depicted in Figure 11, demonstrate that using random weights in this smaller ensemble performed worse than equal weighting, as shown by the darker blue colors compared to Figure 9-b. This supports the previously mentioned hypothesis.

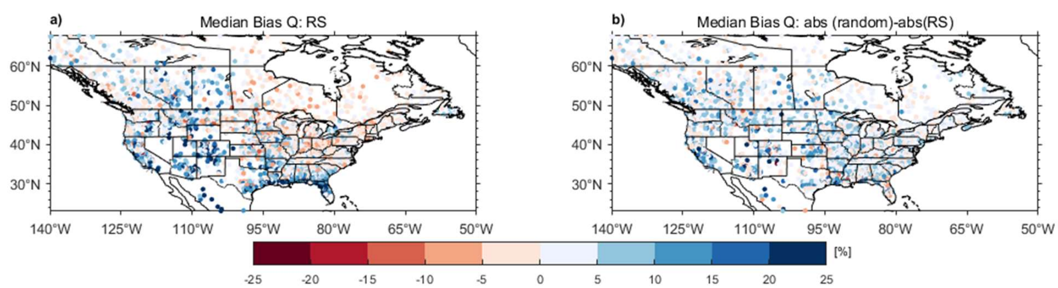


Figure 11. Similar to figure 9, comparing two scenarios: a) Using 7 randomly selected and equally weighted GCMs, and b) the difference in median streamflow bias when using 7 randomly selected GCMs with random and equal weights.

In addition, a single trial of random weight was used. Ideally, multiple trials with different sets of random weights would have been performed to ensure that no bias was introduced. However, given the large number of GCMs in the ensemble and the extensive number of catchments in this

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

study, any significant impact is highly unlikely. The fact that the spatial coherences of the random weights' results were the same as that of other methods supports this assertion.

4.5 Limitation and future work

Weighting methods in climate impact studies involve subjective decisions in selecting diagnostic metrics, translating them into performance measures, and normalizing these into weights. It is essential to recognize these subjective uncertainties since inappropriate weighting methods can either compromise the accuracy of projections or mask underlying uncertainties. In this study, precipitation (prcptot) and temperature (tas) were used for weighting purposes because they are critical inputs to all hydrological models and directly influence streamflow outputs. Another subjective choice was how to combine these variables. It was chosen to treat them equally, with each contributing 50% of the final weights, though this decision was also subjective as well. Impact studies relying on climate variables for weighting, face uncertain trade-offs, often due to nonlinear relationships with streamflow. Relying solely on a single diagnostic metric, such as the climatological mean, for weight determination raises concerns about whether reducing bias in one metric would be beneficial for others. In addition, some models may receive disproportionately high (or low) weights due to their high similarity (or discrepancy) to observations over the reference period. As Shin et al. (2020) noted, this can be particularly noticeable with precipitation, and some form of smoothing scheme might be necessary. Employing a suite of metrics or calibrating multiple metrics could improve the rationale behind

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

the weighted multi-model mean, yet uncertainties in these methods continue to be a subject for further research.

In this study, we utilized the lumped hydrological model HMETS due to the large-sample nature of our research, which made the use of a process-based model impractical. For the hydrological model calibration, observed precipitation, temperature, and streamflow data were used. This approach was necessitated by the challenges associated with using GCM data for hydrological model calibration, primarily because the daily sequences in observations and GCM outputs are not correlated. Using this hydrological model with the pseudo-reality GCM without any prior bias correction is somewhat unconventional and will likely result in mean annual streamflows that are biased, possibly to a significant degree, compared to streamflow observations. However, the pseudo-reality approach requires only the generation of somewhat realistic streamflows, since all other GCMs will be assessed against this reality, and even bias-corrected against this pseudo-reality, thus providing a correct assessment of the weighting strategy. An alternative strategy allowing for direct hydrological model calibration against GCM data has been proposed by Ricard et al. (2023). However, this approach has not yielded streamflow results as reliable as those obtained through direct observation-based calibration.

To further assess these impacts, all methodological steps outlined in Figure 3b were conducted using another hydrological model, the GR4J model (Perrin et al., 2003) linked with the CemaNeige snow module (Valéry et al., 2014). Using this model produced very similar results and led us to the same conclusions with respect to climate model weighting (results not shown).

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

5. Conclusion

This study offers a comprehensive analysis of how weighting members within an ensemble of 22 CMIP6 climate models affects streamflow projections across a large sample of 3,107 North American catchments. Six weighting schemes, including random and equal approaches, were established. Assessing the efficiency of weighting for future conditions presents a challenge due to the absence of future precipitation, temperature, and streamflow data. Therefore, to validate the weighting methods, the study employed the pseudo-reality approach. Each of the 22 GCMs was treated as the pseudo-reality in turn, thus providing future temperature and precipitation data against which the efficiency of the weighting could be evaluated. Future streamflows were generated using the pseudo-reality GCM in conjunction with a hydrological model.

The results indicated that weighting the ensemble led to improved projections of future precipitation and temperature. The optimal weighting method varied between precipitation and temperature. In terms of streamflow projections, the REA weighting method resulted in modest improvements in streamflow predictions compared to equal weighting when no bias correction was performed. No weighting method outperformed equal weighting once bias correction was applied to the precipitation and temperature time series. This is likely due to the complex nonlinear interactions that lead to streamflow. Consequently, using equal weighting of GCMs (model democracy) seems to be a valid strategy for hydrological impact assessment, and especially so when bias correction of climate model outputs is considered necessary.

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

Code and data availability.

The hydrometeorological data used in this study were obtained from the HYSETS database: https://doi.org/10.17605/OSF.IO/RPC3W (Arsenault et al., 2022). The CMIP6 GCM model outputs are available from the Earth System Grid Federation (ESGF) portal at Lawrence Livermore National Laboratory: https://esgf-node.llnl.gov/search/cmip6/ (ESGF, 2022). The processed data and the codes used in this work are available from the authors upon reasonable request.

Author Contribution

The experiments were designed by FB and MRA, and they were carried out by MRA. The findings were analyzed and interpreted by MRA, and FB. The paper was written by MRA and FB, with significant contributions from JLM and RA. JLM and RA also provided editorial feedback on the paper's early draughts.

Competing interests.

The authors declare that they have no conflict of interest.

Acknowledgments

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

This research has been supported by the Natural Sciences and Engineering Research Council of Canada (grant no. RGPIN-2020-07242).

References

Abramowitz, G., Herger, N., Gutmann, E., Hammerling, D., Knutti, R., Leduc, M., Lorenz, R., Pincus, R., & Schmidt, G. A. (2019). ESD Reviews: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing. *Earth System Dynamics*, *10*(1), 91–105. https://doi.org/10.5194/esd-10-91-2019

Arsenault, R., Brissette, F., Chen, J., Guo, Q., & Dallaire, G. (2020a). NAC ^{2} H: The North American Climate Change and Hydroclimatology Data Set. *Water Resources Research*, *56*(8). https://doi.org/10.1029/2020WR027097

Arsenault, R., Brissette, F., Martel, J.-L., Troin, M., Lévesque, G., Davidson-Chaput, J., Gonzalez, M. C., Ameli, A., & Poulin, A. (2020b). A comprehensive, multisource database for hydrometeorological modeling of 14,425 North American watersheds. *Scientific Data*, *7*(1), Article 1. https://doi.org/10.1038/s41597-020-00583-2

Brunner, L., Lorenz, R., Zumwald, M., & Knutti, R. (2019). Quantifying uncertainty in European climate projections using combined performance-independence weighting. *Environmental Research Letters*, *14*(12), 124010. https://doi.org/10.1088/1748-9326/ab492f

Cannon, A. J. (2018). Multivariate quantile mapping bias correction: An N-dimensional probability density function transform for climate model simulations of multiple variables. *Climate Dynamics*, *50*(1), 31–49. https://doi.org/10.1007/s00382-017-3580-6

Cannon, A. J., Piani, C., & Sippel, S. (2020). Chapter 5—Bias correction of climate model output for impact models. In J. Sillmann, S. Sippel, & S. Russo (Eds.), *Climate Extremes and Their Implications for Impact and Risk Assessment* (pp. 77–104). Elsevier. https://doi.org/10.1016/B978-0-12-814895-2.00005-7

Castaneda-Gonzalez, M., Poulin, A., Romero-Lopez, R., & Turcotte, R. (2023). Weighting climate models for hydrological projections: Effects on contrasting hydroclimatic regions. *Climatic Change*, *176*(12), 170. https://doi.org/10.1007/s10584-023-03643-9

Chen, H., Xu, C.-Y., & Guo, S. (2012). Comparison and evaluation of multiple GCMs, statistical downscaling and hydrological models in the study of climate change impacts on runoff. *Journal of Hydrology*, *434–435*, 36–45. https://doi.org/10.1016/j.jhydrol.2012.02.040

Chen, J., Arsenault, R., Brissette, F. P., & Zhang, S. (2021). Climate Change Impact Studies: Should We Bias Correct Climate Model Outputs or Post-Process Impact Model Outputs? *Water Resources Research*, *57*(5), e2020WR028638. https://doi.org/10.1029/2020WR028638

Chen, J., Brissette, F. P., & Caya, D. (2020). Remaining error sources in bias-corrected climate model outputs. *Climatic Change*, *162*(2), 563–582.

Chen, J., Brissette, F. P., Lucas-Picher, P., & Caya, D. (2017). 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

Crawford, J., Venkataraman, K., & Booth, J. (2019). Developing climate model ensembles: A comparative case study. *Journal of Hydrology*, *568*, 160–173. https://doi.org/10.1016/j.jhydrol.2018.10.054

Di Virgilio, G., Ji, F., Tam, E., Nishant, N., Evans, J. P., Thomas, C., Riley, M. L., Beyer, K., Grose, M. R., Narsey, S., & Delage, F. (2022). Selecting CMIP6 GCMs for CORDEX Dynamical Downscaling: Model Performance, Independence, and Climate Change Signals. *Earth's Future*, *10*(4), e2021EF002625. https://doi.org/10.1029/2021EF002625

Dinh, T. L. A., & Aires, F. (2023). Revisiting the bias correction of climate models for impact studies. *Climatic Change*, *176*(10), 140. https://doi.org/10.1007/s10584-023-03597-y

Dong, F., Javed, A., Saber, A., Neumann, A., Alberto Arnillas, C., Kaltenecker, G., & Arhonditsis, G. (2021). A flow-weighted ensemble strategy to assess the impacts of climate change on watershed hydrology. *Journal of Hydrology*, *594*, 125898. https://doi.org/10.1016/j.jhydrol.2020.125898

Duan, Q., Sorooshian, S., & Gupta, V. K. (1994). Optimal use of the SCE-UA global optimization method for calibrating watershed models. *Journal of Hydrology*, *158*(3), 265–284. https://doi.org/10.1016/0022-1694(94)90057-4

Easterling, D. R., Kunkel, K. E., Wehner, M. F., & Sun, L. (2017). Precipitation change in the United States. Climate Science Special Report: Fourth National Climate Assessment, Vol. I, U.S. Global Change Research Program,. *Weather and Climate Extremes*, *11*, 207–230. https://doi.org/10.7930/J0H993CC

Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Caldwell, P., Collins, W. D., Gier, B. K., Hall, A. D., Hoffman, F. M., Hurtt, G. C., Jahn, A., Jones, C. D., Klein, S. A., Krasting, J. P., Kwiatkowski, L., Lorenz, R., Maloney, E., Meehl, G. A., … Williamson, M. S. (2019). Taking climate model evaluation to the next level. *Nature Climate Change*, *9*(2), 102–110. https://doi.org/10.1038/s41558-018-0355-y

Ganguly, T., & Arya, D. S. (2023). A novel framework for a multimodel ensemble of GCMs and its application in the analysis of projected extremes. *International Journal of Climatology*, *43*(15), 7308–7325. https://doi.org/10.1002/joc.8266

Giorgi, F., & Mearns, L. O. (2002). 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*(10), 1141–1158. https://doi.org/10.1175/1520-0442(2002)015<1141:COAURA>2.0.CO;2

Giuntoli, I., Villarini, G., Prudhomme, C., & Hannah, D. M. (2018). Uncertainties in projected runoff over the conterminous United States. *Climatic Change*, *150*(3), 149–162. https://doi.org/10.1007/s10584-018-2280-5

Hagemann, S., Chen, C., Clark, D. B., Folwell, S., Gosling, S. N., Haddeland, I., Hanasaki, N., Heinke, J., Ludwig, F., Voss, F., & Wiltshire, A. J. (2013). Climate change impact on available water resources obtained using multiple global climate and hydrology models. *Earth System Dynamics*, *4*(1), 129–144. https://doi.org/10.5194/esd-4-129-2013

Hausfather, Z., Marvel, K., Schmidt, G. A., Nielsen-Gammon, J. W., & Zelinka, M. (2022). Climate simulations: Recognize the 'hot model' problem. *Nature*, *605*(7908), 26–29.

https://doi.org/10.1038/d41586-022-01192-2

Herger, N., Abramowitz, G., Knutti, R., Angélil, O., Lehmann, K., & Sanderson, B. M. (2018). Selecting a climate model subset to optimise key ensemble properties. *Earth System Dynamics*, *9*(1), 135–151. https://doi.org/10.5194/esd-9-135-2018

Hernanz, A., García-Valero, J. A., Domínguez, M., & Rodríguez-Camino, E. (2022). Evaluation of statistical downscaling methods for climate change projections over Spain: Future conditions with pseudo reality (transferability experiment). *International Journal of Climatology*, *42*(7), 3987–4000. https://doi.org/10.1002/joc.7464

Kling, H., Fuchs, M., & Paulin, M. (2012). Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *Journal of Hydrology*, *424–425*, 264–277. https://doi.org/10.1016/j.jhydrol.2012.01.011

Kling, H., & Gupta, H. (2009). On the development of regionalization relationships for lumped watershed models: The impact of ignoring sub-basin scale variability. *Journal of Hydrology*, *373*(3), 337–351. https://doi.org/10.1016/j.jhydrol.2009.04.031

Knutti, R. (2010). The end of model democracy?: An editorial comment. *Climatic Change*, *102*(3–4), 395–404. https://doi.org/10.1007/s10584-010-9800-2

Knutti, R., Abramowitz, G., Collins, M., Eyring, V., Gleckler, P. J., Hewitson, B., & Mearns, L. (2010). *Good Practice Guidance Paper on Assessing and Combining Multi Model Climate Projections* (T. Stocker, Q. Dahe, G.-K. Plattner, M. Tignor, & P. Midgley, Eds.; pp. 1–15). https://elib.dlr.de/66594/

Knutti, R., Masson, D., & Gettelman, A. (2013). Climate model genealogy: Generation CMIP5

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

and how we got there: CLIMATE MODEL GENEALOGY. *Geophysical Research Letters*, *40*(6), 1194–1199. https://doi.org/10.1002/grl.50256

Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E. M., & Eyring, V. (2017). A climate model projection weighting scheme accounting for performance and interdependence: Model Projection Weighting Scheme. *Geophysical Research Letters*. https://doi.org/10.1002/2016GL072012

Kolusu, S. R., Siderius, C., Todd, M. C., Bhave, A., Conway, D., James, R., Washington, R., Geressu, R., Harou, J. J., & Kashaigili, J. J. (2021). Sensitivity of projected climate impacts to climate model weighting: Multi-sector analysis in eastern Africa. *Climatic Change*, *164*(3–4), 36. https://doi.org/10.1007/s10584-021-02991-8

Lawrence, D. J., Runyon, A. N., Gross, J. E., Schuurman, G. W., & Miller, B. W. (2021). Divergent, plausible, and relevant climate futures for near- and long-term resource planning. *Climatic Change*, *167*(3), 38. https://doi.org/10.1007/s10584-021-03169-y

Li, D., DeConto, R. M., & Pollard, D. (2023). Climate model differences contribute deep uncertainty in future Antarctic ice loss. *Science Advances*, *9*(7), eadd7082. https://doi.org/10.1126/sciadv.add7082

Liang, Y., Gillett, N. P., & Monahan, A. H. (2020). Climate Model Projections of 21st Century Global Warming Constrained Using the Observed Warming Trend. *Geophysical Research Letters*, *47*(12), e2019GL086757. https://doi.org/10.1029/2019GL086757

Lorenz, R., Herger, N., Sedláček, J., Eyring, V., Fischer, E. M., & Knutti, R. (2018). Prospects and Caveats of Weighting Climate Models for Summer Maximum Temperature Projections

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

Over North America. *Journal of Geophysical Research: Atmospheres*, *123*(9), 4509–4526. https://doi.org/10.1029/2017JD027992

Maraun, D. (2016). Bias Correcting Climate Change Simulations—A Critical Review. *Current Climate Change Reports*, *2*(4), 211–220. https://doi.org/10.1007/s40641-016-0050-x

Martel, J.-L., Demeester, K., Brissette, F., Poulin, A., & Arsenault, R. (2017). HMETS-A simple and efficient hydrology model for teaching hydrological modelling, flow forecasting and climate change impacts. *International Journal of Engineering Education*, *33*, 1307–1316.

Massoud, E. c., Espinoza, V., Guan, B., & Waliser, D. e. (2019). Global Climate Model Ensemble Approaches for Future Projections of Atmospheric Rivers. *Earth's Future*, *7*(10), 1136–1151. https://doi.org/10.1029/2019EF001249

Massoud, E. C., Lee, H., Gibson, P. B., Loikith, P., & Waliser, D. E. (2020). Bayesian Model Averaging of Climate Model Projections Constrained by Precipitation Observations over the Contiguous United States. *Journal of Hydrometeorology*, *21*(10), 2401–2418. https://doi.org/10.1175/JHM-D-19-0258.1

Massoud, E. C., Lee, H. K., Terando, A., & Wehner, M. (2023). Bayesian weighting of climate models based on climate sensitivity. *Communications Earth & Environment*, *4*(1), 1–8. https://doi.org/10.1038/s43247-023-01009-8

Mendoza Paz, S., & Willems, P. (2023). The skill of statistical downscaling in future climate with high-resolution climate models as pseudo-reality. *Journal of Hydrology: Regional Studies*, *48*, 101477. https://doi.org/10.1016/j.ejrh.2023.101477

Merrifield, A. L., Brunner, L., Lorenz, R., Medhaug, I., & Knutti, R. (2020). An investigation of weighting schemes suitable for incorporating large ensembles into multi-model ensembles. *Earth System Dynamics*, *11*(3), 807–834. https://doi.org/10.5194/esd-11-807-2020

Murphy, J. M., Sexton, D. M. H., Barnett, D. N., Jones, G. S., Webb, M. J., Collins, M., & Stainforth, D. A. (2004). Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, *430*(7001), Article 7001. https://doi.org/10.1038/nature02771

Padrón, R. S., Gudmundsson, L., & Seneviratne, S. I. (2019). Observational Constraints Reduce Likelihood of Extreme Changes in Multidecadal Land Water Availability. *Geophysical Research Letters*, *46*(2), 736–744. https://doi.org/10.1029/2018GL080521

Palmer, T. E., McSweeney, C. F., Booth, B. B. B., Priestley, M. D. K., Davini, P., Brunner, L., Borchert, L., & Menary, M. B. (2023). Performance-based sub-selection of CMIP6 models for impact assessments in Europe. *Earth System Dynamics*, *14*(2), 457–483. https://doi.org/10.5194/esd-14-457-2023

Perrin, C., Michel, C., & Andréassian, V. (2003). Improvement of a parsimonious model for streamflow simulation. *Journal of Hydrology*, *279*(1), 275–289. https://doi.org/10.1016/S0022-1694(03)00225-7

Prein, A. F., Liu, C., Ikeda, K., Bullock, R., Rasmussen, R. M., Holland, G. J., & Clark, M. (2020). Simulating North American mesoscale convective systems with a convection-permitting climate model. *Climate Dynamics*, *55*(1), 95–110. https://doi.org/10.1007/s00382-017-3993-2

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

Rahimpour Asenjan, M., Brissette, F., Martel, J.-L., & Arsenault, R. (2023). Understanding the influence of “hot” models in climate impact studies: A hydrological perspective. *Hydrology and Earth System Sciences*, *27*(23), 4355–4367. https://doi.org/10.5194/hess-27-4355-2023

Raulino, J. B. S., Silveira, C. S., & Lima Neto, I. E. (2021). Assessment of climate change impacts on hydrology and water quality of large semi-arid reservoirs in Brazil. *Hydrological Sciences Journal*, *66*(8), 1321–1336. https://doi.org/10.1080/02626667.2021.1933491

Reichler, T., & Kim, J. (2008). How Well Do Coupled Models Simulate Today’s Climate? *Bulletin of the American Meteorological Society*, *89*(3), 303–312. https://doi.org/10.1175/BAMS-89-3-303

Reshmidevi, T. V., Nagesh Kumar, D., Mehrotra, R., & Sharma, A. (2018). Estimation of the climate change impact on a catchment water balance using an ensemble of GCMs. *Journal of Hydrology*, *556*, 1192–1204. https://doi.org/10.1016/j.jhydrol.2017.02.016

Ricard, S., Lucas-Picher, P., Thiboult, A., & Anctil, F. (2023). Producing reliable hydrologic scenarios from raw climate model outputs without resorting to meteorological observations. *Hydrology and Earth System Sciences*, *27*(12), 2375–2395. https://doi.org/10.5194/hess-27-2375-2023

Ruane, A. C., & McDermid, S. P. (2017). Selection of a representative subset of global climate models that captures the profile of regional changes for integrated climate impacts assessment. *Earth Perspectives*, *4*(1), 1. https://doi.org/10.1186/s40322-017-0036-4

Sanderson, B. M., Knutti, R., & Caldwell, P. (2015). A Representative Democracy to Reduce

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

Interdependency in a Multimodel Ensemble. *Journal of Climate*, *28*(13), 5171–5194. https://doi.org/10.1175/JCLI-D-14-00362.1

Sanderson, B. M., Wehner, M., & Knutti, R. (2017). Skill and independence weighting for multi-model assessments. *Geoscientific Model Development*, *10*(6), 2379–2395. https://doi.org/10.5194/gmd-10-2379-2017

Shin, Y., Lee, Y., & Park, J.-S. (2020). A Weighting Scheme in A Multi-Model Ensemble for Bias-Corrected Climate Simulation. *Atmosphere*, *11*(8), Article 8. https://doi.org/10.3390/atmos11080775

Singh, R., & AchutaRao, K. (2020). Sensitivity of future climate change and uncertainty over India to performance-based model weighting. *Climatic Change*, *160*(3), 385–406. https://doi.org/10.1007/s10584-019-02643-y

Stainforth, D. a, Allen, M. r, Tredger, E. r, & Smith, L. a. (2007). Confidence, uncertainty and decision-support relevance in climate predictions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *365*(1857), 2145–2161. https://doi.org/10.1098/rsta.2007.2074

Tarek, M., Brissette, F., & Arsenault, R. (2021). Uncertainty of gridded precipitation and temperature reference datasets in climate change impact studies. *Hydrology and Earth System Sciences*, *25*(6), 3331–3350. https://doi.org/10.5194/hess-25-3331-2021

Tarek, M., Brissette, F. P., & Arsenault, R. (2020). Evaluation of the ERA5 reanalysis as a potential reference dataset for hydrological modelling over North America. *Hydrology and Earth System Sciences*, *24*(5), 2527–2544. https://doi.org/10.5194/hess-24-2527-

2020

Tebaldi, C., & Knutti, R. (2007). The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *365*(1857), 2053–2075. https://doi.org/10.1098/rsta.2007.2076

Tokarska, K. B., Stolpe, M. B., Sippel, S., Fischer, E. M., Smith, C. J., Lehner, F., & Knutti, R. (2020). Past warming trend constrains future warming in CMIP6 models. *Science Advances*, *6*(12), eaaz9549. https://doi.org/10.1126/sciadv.aaz9549

Valéry, A., Andréassian, V., & Perrin, C. (2014). '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

Wang, H., Chen, J., Xu, C., Zhang, J., & Chen, H. (2020). A Framework to Quantify the Uncertainty Contribution of GCMs Over Multiple Sources in Hydrological Impacts of Climate Change. *Earth's Future*, *8*(8). https://doi.org/10.1029/2020EF001602

Wang, H.-M., Chen, J., Xu, C.-Y., Chen, H., Guo, S., Xie, P., & Li, X. (2019). Does the weighting of climate simulations result in a better quantification of hydrological impacts? *Hydrology and Earth System Sciences*, *23*(10), 4033–4050. https://doi.org/10.5194/hess-23-4033-2019

Weigel, A. P., Knutti, R., Liniger, M. A., & Appenzeller, C. (2010). Risks of Model Weighting in Multimodel Climate Projections. *Journal of Climate*, *23*(15), 4175–4191.

https://doi.org/10.5194/egusphere-2024-1183
Preprint. Discussion started: 5 June 2024

https://doi.org/10.1175/2010JCLI3594.1

Wootten, A. M., Massoud, E. C., Waliser, D. E., & Lee, H. (2023). Assessing sensitivities of climate model weighting to multiple methods, variables, and domains in the south-central United States. *Earth System Dynamics*, *14*(1), 121–145. https://doi.org/10.5194/esd-14-121-2023

Wootten, A., Massoud, E., Sengupta, A., Waliser, D., & Lee, H. (2020). The Effect of Statistical Downscaling on the Weighting of Multi-Model Ensembles of Precipitation. *Climate*, *8*, 138. https://doi.org/10.3390/cli8120138

Yang, H., Zhou, F., Piao, S., Huang, M., Chen, A., Ciais, P., Li, Y., Lian, X., Peng, S., & Zeng, Z. (2017). Regional patterns of future runoff changes from Earth system models constrained by observation. *Geophysical Research Letters*, *44*(11), 5540–5549. https://doi.org/10.1002/2017GL073454

Yuan, S., Quiring, S. M., Kalcic, M. M., Apostel, A. M., Evenson, G. R., & Kujawa, H. A. (2020). Optimizing climate model selection for hydrological modeling: A case study in the Maumee River basin using the SWAT. *Journal of Hydrology*, *588*, 125064. https://doi.org/10.1016/j.jhydrol.2020.125064