The Dilemma of Including 'Hot' Models in Climate Impact Studies: A Hydrological Study

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

5   [1]Hydrology, Climate and Climate Change Laboratory, École de technologie supérieure, Université du Québec, 1100 Notre-Dame Street West, Montreal, Quebec, Canada H3C 1K3

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

Abstract. Efficient adaptation strategies to climate change require estimating future impacts and the uncertainty surrounding this estimation. Over- or under-estimating future uncertainty may lead to maladaptation. Hydrological impact studies

10   typically use a top-down approach in which multiple climate models are used to assess the uncertainty related to climate model structure and climate sensitivity. Despite ongoing debate, impact modelers have typically embraced the concept of "model democracy" in which each climate model is considered equally fit. The newer CMIP6 simulations, with several models showing a climate sensitivity larger than that of CMIP5 and larger than the likely range based on past climate information and understanding of planetary physics, have reignited the model democracy debate. Some have suggested that

15   hot models be removed from impact studies to avoid skewing impact results toward unlikely futures.

This large-sample study looks at the impact of removing hot models on the projections of future streamflow over 3,107 North American catchments. More precisely, the variability of future projections of mean, high, and low flows is evaluated using an ensemble of 19 CMIP6 GCMs, 5 of which are deemed "hot" based on their global equilibrium climate sensitivity (ECS). The results show that the reduced ensemble of 14 climate models provides streamflow projections with reduced

20   future variability for Canada, Alaska, the Southwest US, and along the Pacific coast. Elsewhere, the reduced ensemble has either no impact or results in increased variability of future streamflow, indicating that outlier climate models do not necessarily provide outlier projections of future impacts. These results emphasize the delicate nature of climate model selection, especially based on global fitness metrics that may not be appropriate for local and regional assessments.

1. Introduction

25 Understanding the impact of climate change on water resources and hydrology is crucial for developing effective strategies for mitigation and adaptation (Eyring et al., 2019; Miara et al., 2017). The output of hydrological (e.g. Karlsson et al. 2016), water quality (Prajapati et al., 2023) and sediment transport (Sabokruhie et al., 2021) impact assessment studies is dependent on the choice of the future climate change projections. Hydrologists primarily use climate projection outputs from GCMs
30 (e.g. Tabari, 2020) to study these impacts. The Coupled Model Intercomparison Project (CMIP) provides standardized metadata from coordinated simulations by different climate modeling groups (Meehl et al., 2007). The more recent CMIP6 (Eyring et al., 2016) is gradually replacing the widely used CMIP5 from the last decade (Hirabayashi et al., 2021; Martel et al., 2022; Zhang et al., 2023).

35 The concept of "model democracy" has been widely used in impact studies (e.g. Collins et al., 2013; IPCC, 2014), despite criticism (Knutti, 2012). This approach considers climate simulations independent and equally plausible, and uses the ensemble mean and spread to define climate model uncertainty. Research has shown that the average of equally-weighted projections outperforms single models in simulating mean climatic patterns (Chen et al., 2017; Reichler and Kim, 2008). However, this approach may be less effective for CMIP6 ensemble as the validity of some simulations is under question
40 (Hausfather et al., 2022).

The CMIP6 ensemble includes a subset of "hot models" that predict greater warming than previous predictions made by CMIP5 (e.g. Kreienkamp et al., 2020). These hot models have a climate sensitivity that exceeds the expected plausible range, which is based on observations and our understanding of planetary physics. They also exhibit a higher equilibrium climate
45 sensitivity (ECS), a measure of the steady-state temperature increase in the event of doubled carbon dioxide (CO_2) concentrations in the atmosphere (Flynn and Mauritsen, 2020; Zelinka et al., 2020). The ECS values' range in CMIP6 models has increased to 1.8–5.6 °C compared to 2.1–4.7 °C in CMIP5, with an increase in multimodel mean of 3.9 °C in CMIP6 from 3.3 °C in CMIP5 (Zelinka et al., 2020).

50 However, a plethora of evidence based on observations and our understanding of planetary physics indicate that we can confidently restrict the range of future warming and, more importantly, give less weight to extreme estimates (Liang et al., 2020; Tokarska et al., 2020). Recently, more research has been focused on constraining the ECS based on historical and paleoclimatic data (Knutti et al., 2017; Sherwood et al., 2020) or emergent constraints (Cox et al., 2018; Nijsse et al., 2020). For example, Sherwood et al. (2020) used multiple lines of evidence and concluded that the likely (with a 66% chance) ECS
55 value is between 2.6°C and 4.1°C. Consequently, the most recent reports published by the Intergovernmental Panel on Climate Change (IPCC) have narrowed the likely ECS range to 2.5 and 4°C (IPCC, 2021).

Climate change impact studies that include models with high ECS may be biased and may overestimate the magnitude of impacts (Hausfather et al., 2022). Using the full ensemble of CMIP6 projections without restricting the "hot models" may no longer be the most appropriate option for impact studies (Ribes et al., 2021). For instance, Shiogama et al. (2021) proposed a subset selection method that involves screening out hot models as the first step. On the other hand, Palmer et al. (2022) found that models with higher sensitivity better represent some key climatic processes over Europe. While they were unable to provide robust physical explanations for their findings, it is worth noting that at the regional scale, hot models may provide valuable information that may be more important than the global warming trend for impact modelers, adding to the complexity of selecting models for regional impact studies.

The decision to weight climate models for impact studies remains controversial, but it is difficult to ignore the potential pitfalls of using hot models in these studies (Hausfather et al., 2022). This study aims to evaluate how including or excluding hot models in a multi-model ensemble affects the results of a large-scale hydrological climate change impact study. This influence is measured in terms of the magnitude and uncertainty of various streamflow metrics for 3107 North-American catchments.

2. Materials and Methods

The data for this study was obtained from the HYSETS database, which contains hydrometeorological data from various sources for over 14,000 catchments in North America (Arsenault et al., 2020). The database includes all necessary data for the reference period of this study, including catchment boundaries (in the form of shapefiles), streamflow observations, weather observations (from stations as well as multiple gridded and reanalysis datasets), and static catchment descriptors such as area, slope, elevation, land-use fractions, and soil properties. This study used the ERA5 reanalysis dataset for meteorological data, which was found to be a reliable alternative to gauge observations in a previous large-scale comparison study over the same study area (Tarek et al., 2020). To ensure representativeness, a subset of HYSETS catchments was selected using filters. First, catchments with drainage areas below 500 km^2 were excluded because daily hydrological models would be inappropriate for modeling hydrological processes at smaller scales. Next, catchments required at least ten years of data to ensure sufficient data for calibrating hydrological models and bias-correcting climate models. After calibration, only catchments with a Kling-Gupta Efficiency (KGE) value above 0.5 were kept, resulting in a selection of 3,107 catchments, as shown in Figure 1.

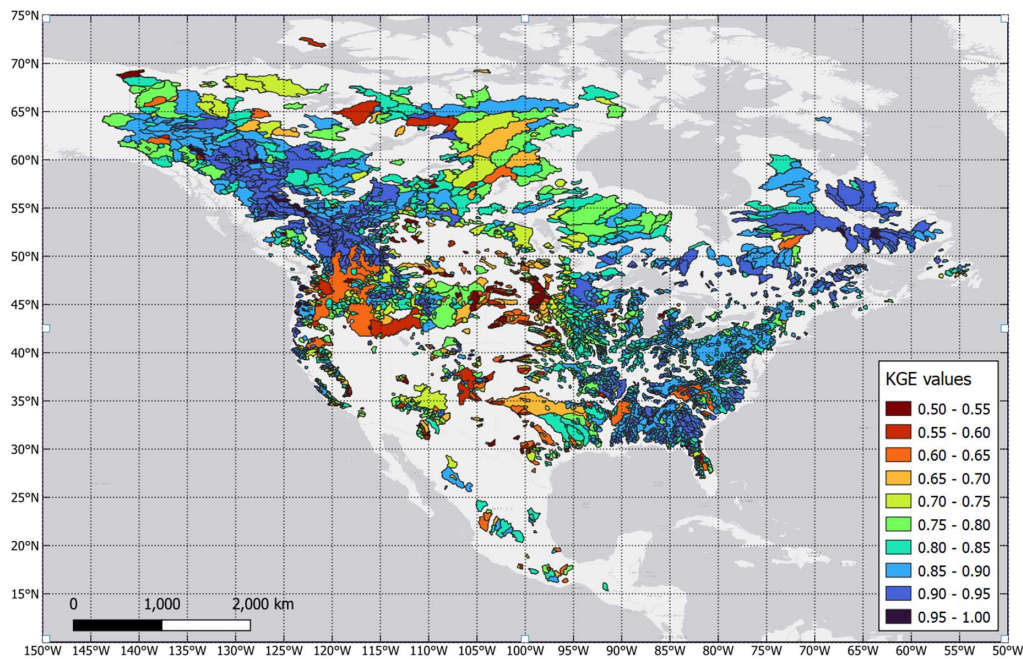



Figure 1 Catchment location. The color scale corresponds to the hydrological model KGE calibration score over the reference period.

90 Table 1 presents the list of 19 CMIP6 GCMs selected for this study. This list includes 5 hot models, defined by their ECS greater than 4.1. These models are: CanESM5 (ECS: 5.6), NESM3 (ECS: 4.7), IPSLCM6A-LR (ECS: 4.5), EC-Earth3-veg (ECS: 4.3), EC-Earcth3 (ECS: 4.2). This study will be able to compare the uncertainty generated by the entire ensemble (19 models) to that of a reduced ensemble (14 models) obtained by removing the 5 hot models.

95

100

Table 1. The 19 GCMs selected in this study and their corresponding ECS.

GCM	ECS
CANESM5	5.62
NESM3	4.68
IPSL-CM6A-LR	4.52
EC-Earth3-Veg	4.3
EC-Earth3	4.2
ACCESS-ESM1-5	3.9
GFDL-CM4_gr1	3.87
GFDL-CM4_gr2	3.87
MRI-ESM2-0	3.14
BCC-CSM2-MR	3.01
MPI-ESM1-2-HR	2.97
MPI-ESM1-2-LR	2.97
FGOALS-g3	2.9
GFDL-ESM4	3.87
NorESM2-LM	2.60
MIROC6	2.57
NorESM2-MM	2.5
INM-CM5-0	1.92
INM-CM4-8	1.83

105 The impact study in this paper uses a traditional top-down hydroclimatic modeling chain consisting of one shared socioeconomic pathway (SSP8.5), 19 CMIP6 GCMs, one bias correction method, and one hydrological model. The study focuses solely on GCM uncertainty and doesn't consider other components, such as alternative SSPs, bias correction methods, or hydrological models, which would add uncertainty to future projections. These have been explored in previous

studies (e.g. Wilby and Harris, 2006; Chen et al., 2011; Giuntoli et al., 2018; Troin et al., 2022), and are outside the scope of
110   this work. The reference period is based on the 1971-2000 time frame, while the future climate is based on 2070-2099.

Climate models are mathematical representations of the Earth's climate system, based on current understanding of its physics
and chemistry. They are formulated using simplifying assumptions and parameterizations, but may not fully capture the
complexity of the real climate system due to limited observations and understanding. As a result, climate models can be
115   biased when compared to observations, due to factors such as model resolution, errors in reference datasets, and sensitivity
to initial conditions. To ensure realistic impact simulations in impact studies, it is important to bias-correct climate model
outputs. In this work, Cannon's (2018) N-dimensional multivariate bias correction (MBCn) method was used to correct
biases in daily precipitation and temperature. MBCn is considered the most advanced and efficient quantile-based
multivariate bias correction method, as reported by studies such as Chen et al. (2018), Su et al. (2020), and Cannon et al.
120   (2020). MBCn transfers the distribution of observational data to the corresponding distribution from the climate model while
preserving its projection trends, crucial for climate change impact studies (Maraun, 2016). No downscaling was performed
since this study was conducted at the catchment scale.

In this study, the GR4J lumped rainfall-runoff model (Perrin et al., 2003) was chosen to simulate streamflows. The model
125   was selected due to the large number of catchments, which made it infeasible to use more complex, distributed models.
Additionally, lumped models use averaged temperature and precipitation at the catchment scale, which is more consistent
with the scale of GCMs, eliminating the need for downscaling. Lumped models have been shown to perform well in
simulating streamflows at catchment outlets (e.g. dos Santos et al., 2018; Reed et al., 2004). The GR4J model is simple,
efficient, and high-performing compared to other lumped conceptual models. It uses precipitation, potential
130   evapotranspiration (PET), and catchment surface area as inputs. To account for snow accumulation in some catchments, the
GR4J model is linked with the CemaNeige snow module (Valéry et al., 2014), resulting in a 6-parameter model (GR4J_CN).
The GR4J_CN model combination has been used in many studies, including climate change impact studies, and has been
shown to perform well under a wide range of conditions (e.g. Riboust et al., 2019; Tarek et al., 2020; Wang et al., 2019).

135   The hydroclimatic modeling chain described above generated 19 different 30-year time series of daily streamflow for the
2070-2099 future period, each corresponding to one of the 19 GCMs listed in Table 1. Three streamflow metrics were
extracted from each 30-year time series, representing mass balance (Qmean) and high (Qmax) and low (Qmin) flows:

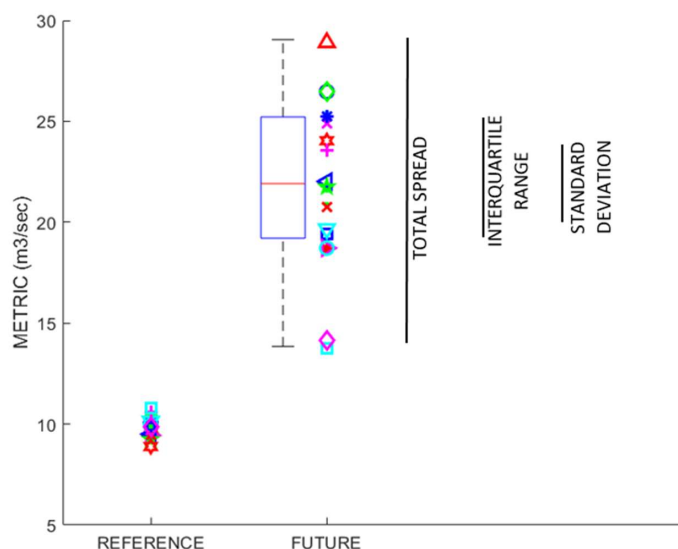
- Qmean: obtained by averaging daily streamflow over the 30-year period.

- Qmax: obtained by averaging the 30 annual maximum simulated streamflows.

140   • Qmin: obtained by averaging the 30 annual minimum simulated streamflows.

These metrics will be used to assess the impact of removing hot climate models across a range of flow conditions.

Figure 2 presents the three dispersion metrics used in this study to compare the spread (or uncertainty) of future projections of streamflow metrics. For the three streamflow metrics, 19 values from the original ensemble and 14 from the reduced

145 ensemble for both the reference and future periods are extracted. The spread of the streamflow projections over the reference period is small, but it is not zero due to imperfect bias correction and the hydrology model's strong non-linear response to precipitation and temperature inputs. The spread is comparatively much larger in the future period, mainly due to differences in sensitivity and structure of the climate models.

150 **Figure 2. Representation of the dispersion metrics used in this paper. Each marker represents one of the 19 climate models.**

Total spread (TS) is defined as the full range of future streamflow responses:

$$TS = metric_{max} - metric_{min}$$

155 The interquartile range (IQR) is defined as the distance between the 75th and 25th quantiles of the distribution as shown by the blue rectangle in the boxplot in Figure 2.

$$IQR = Q_{75} - Q_{25}$$

Finally, the standard deviation (σ) is the standard mathematical measure of dispersion. In the case of a normal distribution,
160 the standard deviation and interquartile range are perfectly correlated, but this may not be the case for a skewed distribution.

All three metrics have units of m³/s and are therefore dependent on catchment size and, to a lesser extent, mean annual precipitation. To account for this, the metrics will be presented in a non-dimensional form:

$$TS_{nd} \quad = \frac{TS_{14}}{TS_{19}}$$

165 Where TS_{19} and TS_{14} respectively represent the total spread for the full and reduced ensemble. TS_{nd} varies between 0 and 1, with TS_{nd}=1 meaning that no reduction in total spread was obtained by removing the five warm models from the ensemble, and TS_{nd}= 0 signifies that the total spread of the reduced ensemble has been totally eliminated.

Similarly, for the interquartile range ratio, we find:

170

$$IQR_{nd} \quad = \frac{IQR_{14}}{IQR_{19}}$$

However, in this case, the potential values vary in the 0 to ∞ range. More practically, a value below 1 indicates that the IQRhas been reduced by removing the five hot models from the ensemble, whereas a value larger than 1 shows the opposite. The latter is possible if the removed models are somewhat close to the median of the ensemble.

175 Finally, for the standard deviation the following ratio is used:

$$\sigma_{nd} \quad = \frac{\sigma_{14}}{\sigma_{19}}$$

where a value below 1 indicates a smaller standard deviation for the reduced ensemble, and the opposite for a value above 1. σ_{nd} has the same possible range of values as IQR_{nd} (0 to ∞).

3. Results

180 Figure 3 presents the box plots of projected temperature increases for each of the 3107 catchments and for each climate model. The box plots provide a visual representation of key elements of the temperature increase distribution. The median of the distribution is shown as the red line near the centre of the blue rectangle, which delimits the interquartile range (Q75 and Q25 for the upper and lower end of the rectangle). The whiskers represent the minimum and maximum of the distribution, and the red crosses are considered statistical outliers. Results indicate that the distribution of projected temperature increases

185 generally follows the same order as the ECS values presented in Table 1. However, there are some differences, which are not unexpected as global-scale (ECS) values are compared to regional-scale (ΔT) values. The five hot models are ranked as the first, second, third, fifth, and sixth hottest regional models based on median values (considering that GFDL-CMA gr1 and gr2, respectively fourth and fifth, are actually the same model with different spatial resolutions).

190

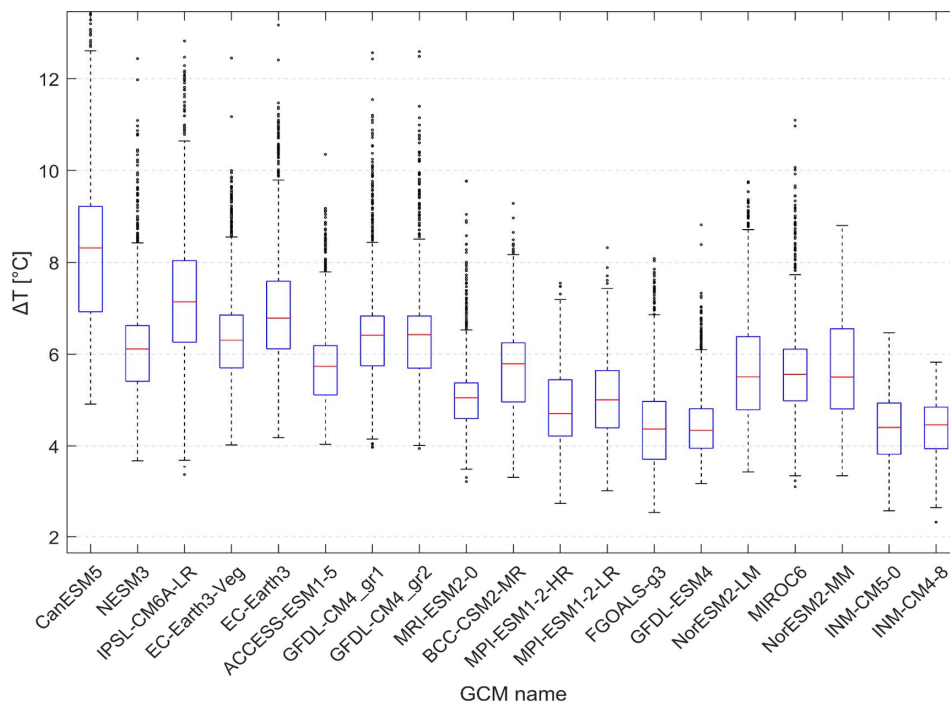



Figure 3. Distribution of projected temperature increase (ΔT) for the 19 CMIP6 selected model for the 2070-2099 future period, compared to the 1971-2000 reference period. Each boxplot represents the distribution of projected increases for the 3107 study catchments. The climate models are ordered in terms of their global-scale ECS values, starting with the largest to the left. A few
195 **catchment outliers are beyond the Y-axis limits.**

Figure 4 presents the boxplots of the projected changes in relative precipitation between the future and reference periods ($\frac{P_{fut}-P_{ref}}{P_{ref}}$). The boxplots depict the distribution of the projected precipitation changes for each of the 3107 catchments. Results indicate that the hot models, identified by their ECS values, are also among the models with the largest projected

200 changes in relative precipitation. Specifically, the five hot models are all within the group of the eight wettest models. The models with more modest increases in precipitation (e.g., MPI-ESM, ACCESS) are also among the cooler models. This trend is expected, as a warmer atmosphere can hold more moisture (up to 7% per °C, according to the Clausius-Clapeyron relationship), leading to more precipitation. Increased precipitation may mitigate the anticipated impacts of warmer models, such as increased evapotranspiration.

205

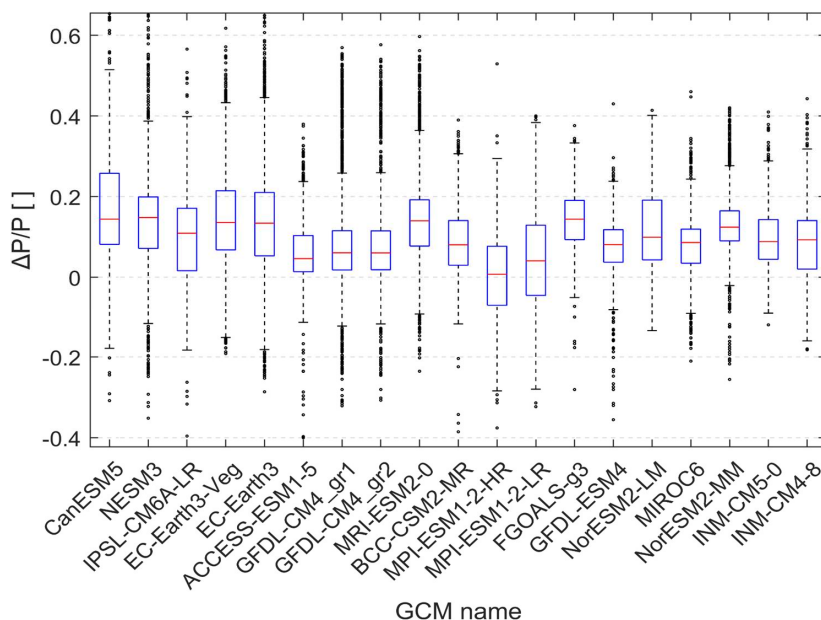



Figure 4. Distribution of projected relative annual precipitation increase (ΔP/P) for the 19 CMIP6 selected models for the 2070-2099 future period. Each boxplot represents the distribution of projected increases for the 3107 study catchments. The climate models are ordered in terms of their global-scale ECS values, starting with the largest to the left. A few catchment outliers are beyond the Y-axis limits.

Figure 5 presents the total spread ratio for mean (Q_{mean}), annual max (Q_{max}), and min (Q_{min}) streamflow obtained by removing the 5 hot models from the 19-member ensemble. A dark red color indicates no reduction in total spread with the reduced ensemble, while lighter colors indicate a reduction. It can be seen that there is a clear spatial pattern that is relatively similar for all three streamflow metrics. The largest reductions in total spread are seen in the northern regions as well as in the US southeast, and along the US Pacific coast for Q_{mean} and Q_{min}. For all other regions of the US, no reduction in total spread is observed. The reduced spread observed in the northern regions is smaller for Q_{max}. Despite these trends, a lot of variability remains present, with neighbouring catchments sometimes showing contrasting behaviour. More specifically, 57.0% of the catchments see a decrease in total spread for Q_{mean}, 53.3% for Q_{max} and 61.7% for Q_{min}.

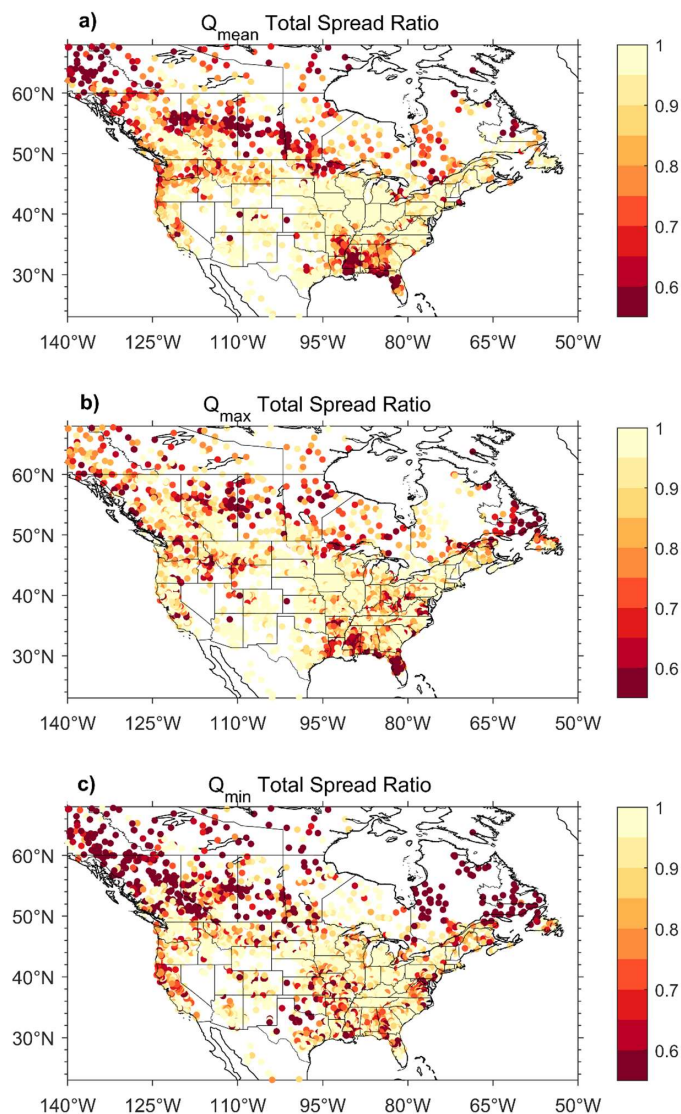



Figure 5. Total spread ratio for Q_mean (a), Q_max (b), and Q_min (c) resulting from the removal of the five hot models.

Figure 6 presents data from Figure 5 in the form of boxplots to better illustrate the range of total spread reduction. It shows
225   that the median total spread ratio is relatively high for all three streamflow metrics: Q_{mean} (0.96), Q_{max} (0.95) and Q_{min} (0.93).

This is primarily because a significant number of catchments see no reduction in total spread (43%, 46.7%, and 38.3% respectively). However, there is a significant reduction in total spread observed in many catchments, and this decrease is strongly dependent on the geographical location of the catchments, as previously seen in Figure 5. Additionally, it can be seen that removing the hot models has a greater impact on Qmin than on the other two metrics.

230

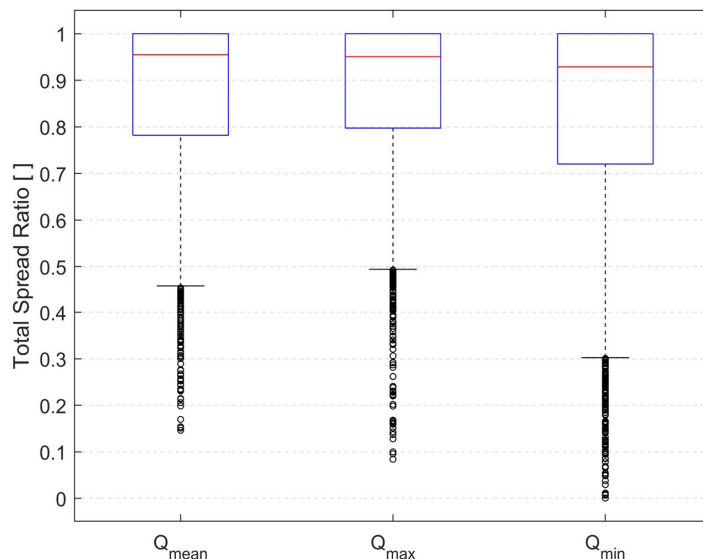
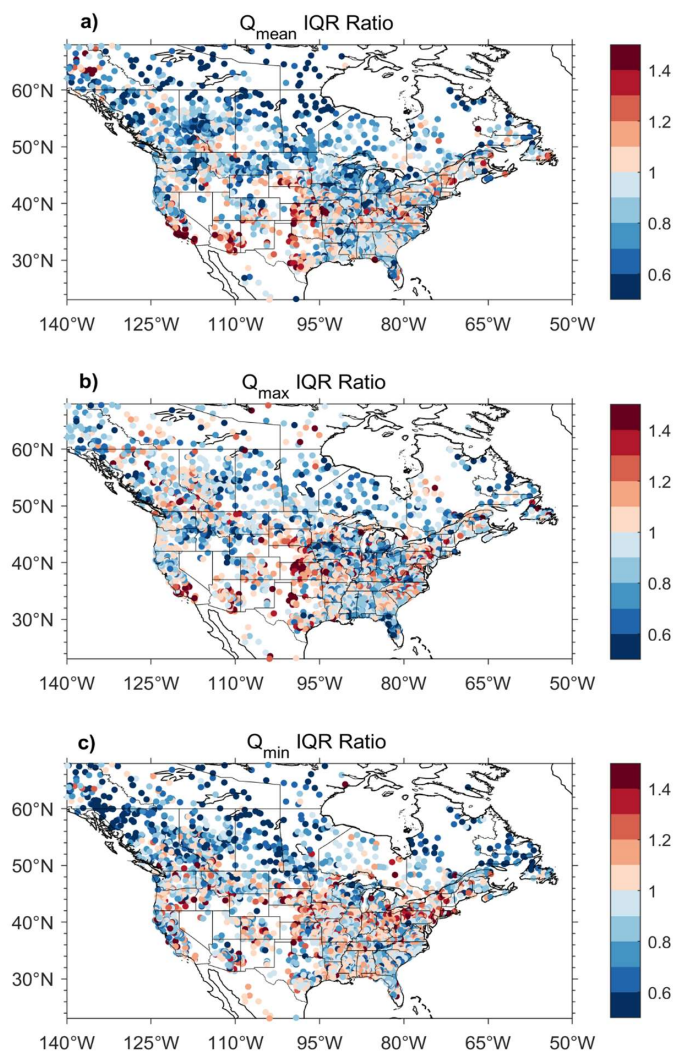



Figure 6 Boxplots of total spread ratio ($TS_{nd} = \frac{TS_{14}}{TS_{19}}$) for Qmean, Qmax and Qmin resulting from the removal of the 5 hot models from the ensemble

235 The total spread ratio is heavily impacted by outliers and may not accurately represent the overall spread of models. Figures 7 and 8 present the interquartile range (IQR) ratio for the three streamflow metrics. A red color (IQR > 1) indicates that the model spread has increased following the removal of the hot models whereas a blue color (IQR < 1) corresponds to a decrease. Results indicate that removing the hot models consistently reduces IQR ratios in Canada for Q_{mean} and Q_{min} , and to a lesser extent for Q_{max}. However, in CONUS, the results are more complex with a lot of regional variability. Removing
240 outlier models in the US southwest results in an increase in IQR for both Q_{mean} and Q_{max}.

Overall, as shown in the boxplots of Figure 8, removing the hot models likely reduces the spread, with 68%, 62% and 60% of catchments seeing a reduction in IQR range for Q_{mean}, Q_{max} and Q_{min} respectively. These values are larger than those obtained for total spread.

245   **Figure7. Change in the IQR ratio for Q_{mean} (a), Q_{max} (b), and Q_{min} (c) resulting from the removal of the five hot models.**

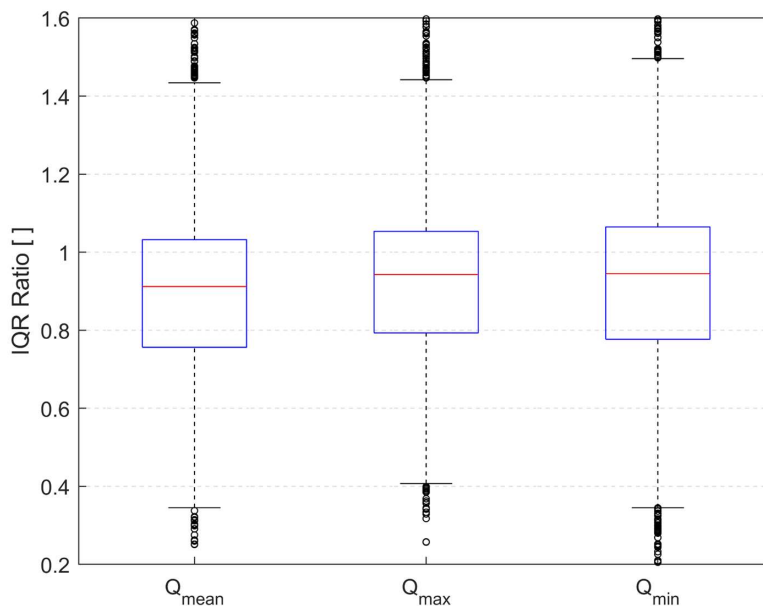



Figure 8. Boxplots of change in the interquartile range ratio (IQR_{nd}) for Qmean, Qmax and Qmin resulting from the removal of the 5 hot models. A few outliers are beyond the Y-axis limits.

250   Figures 9 and 10 present the results for the standard deviation ratio. The results are similar to those of the IQR ratio, but are less noisy. Figure 9 shows clearer regional patterns of spread reduction and increases. For Q_{mean}, removing the hot models results in increases in standard deviation ratios for the north central, north-east, and southwest of the US. Patterns are similar for Q_{max} and Q_{min}, but with more noise. Figure 10 shows that roughly two-thirds of catchments see a decrease in spread, while one-third see an increase.

255

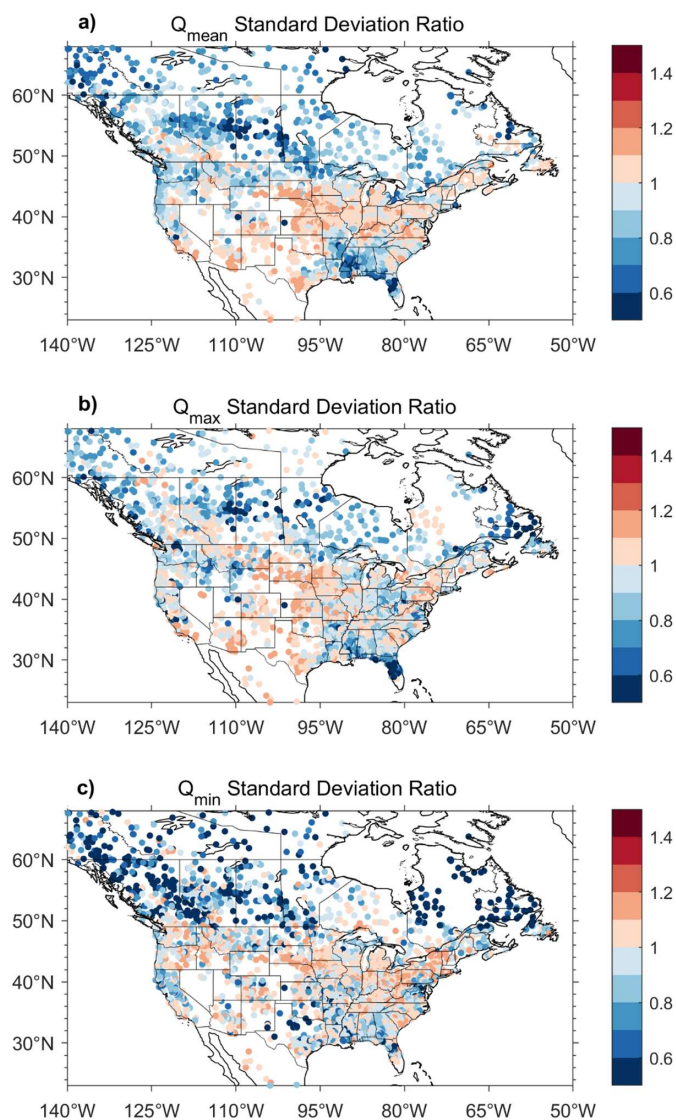



Figure 9. Standard deviation ratio for Q_{mean} (a), Q_{max} (b), and Q_{min} (c) resulting from the removal of the five hot models.

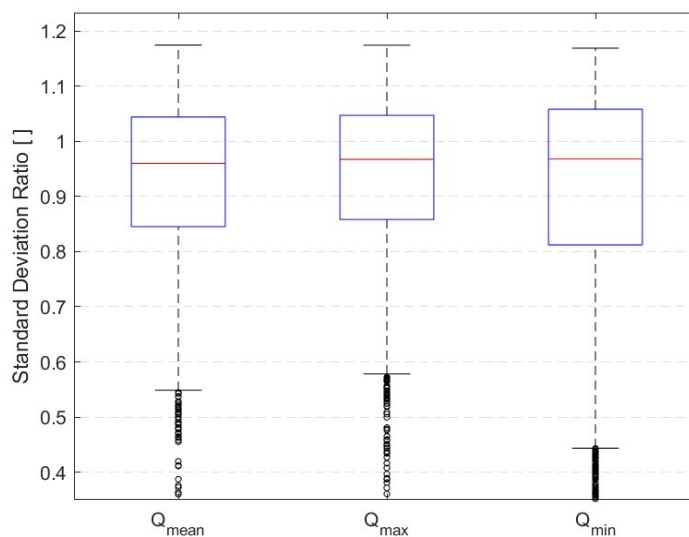



Figure 10. Boxplots of the standard deviation ratio for Q_{mean}, Q_{max}, and Q_{min} resulting from the removal of the five hot models.

260

4. Discussion

Uncertainty is a key factor in assessing the impact of climate change. Different models and techniques, including various climate models, can lead to diverse climate projections and scenarios. Climate change interacts with other stressors, such as land use change and population growth, in complex and unpredictable ways, making it important to accurately address
265 uncertainty in climate impact studies to develop effective adaptation measures. Incorrectly representing uncertainty can lead to poor adaptation.

Results show that removing the "hot models" is likely to reduce the spread of three streamflow metrics. Between 60% and 75% of catchments show a decrease in the spread of future streamflow projections, indicating that the hot models are outliers
270 or further from the mean than the average model. In such cases, keeping the hot models would result in an overestimation of future streamflow uncertainty. However, removing the hot models also led to an increase in the spread in certain regions, indicating overconfidence in the results. This means that while the hot models are outliers with respect to ECS, they may not be outliers with respect to impact studies. Generally, a reduction in spread was evident in northern regions such as Canada and Alaska, as well as the coast of California and the southeastern region of the US.

275 A reduction in the spread of future streamflow is expected when removing the hot models or reducing the number of climate models. A bootstrap methodology was used to determine if the changes in spread were due to a reduction in the number of

models. This was conducted by selecting a random sample of 14 (out of 19) models 100 times and computing the average STD ratio. This was repeated for all catchments and the aggregated results are shown in Figure 11.

The results indicate that removing five random models results in a decrease in the standard deviation ratio more than 75% of

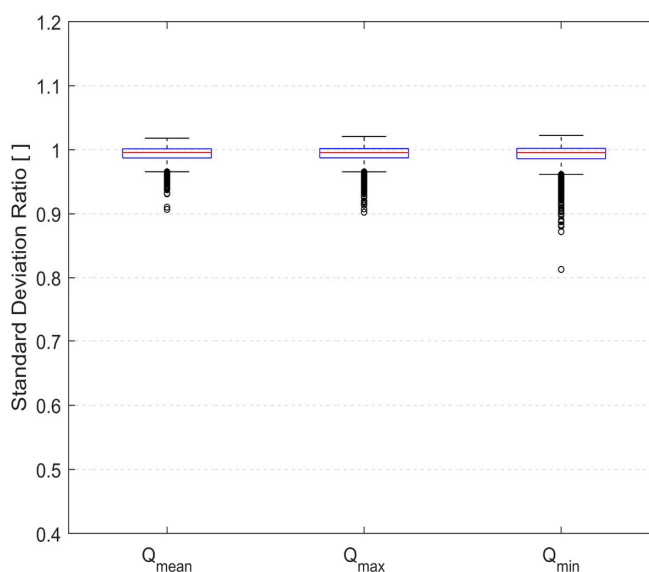
280   the time for all three streamflow metrics, but the median spread reduction ratio for this spread metric is extremely small (about 0.99 for all three streamflow metrics). This shows that removing the 5 hot models has a much larger impact than removing 5 random models. The spread reduction observed in many catchments is therefore not solely related to a reduction in the number of models.

285

Figure 11. Boxplots of the average standard deviation ratio for Q_{mean}, Q_{max}, and Q_{min} resulting from the removal of 5 random models, after sampling 100 random combinations of 5 models.

At first glance, there is a strong physical reasoning for removing climate models with equilibrium climate sensitivity (ECS)

290   exceeding values expected from current data and understanding of planetary physics (Ribes et al., 2021; Shiogama et al., 2021). However, it should be noted that most impact studies are conducted at the regional or local scale and these models may not be considered outliers at these scales. This study found that while globally hot models may still be among the hottest in the study domain, they are not consistently the hottest, raising questions about whether their global behavior should automatically eliminate them from regional studies.

295

In this study, the climate performance of these models (such as their ability to represent climatic, hydroclimatic, or hydrological metrics) was not evaluated. The goal was to examine the impact of removing 5 hot models from a 19-member ensemble. However, it is important to note that judging climate models based solely on their ECS values may result in the removal of models that have desirable characteristics at the regional scale (e.g. Palmer et al., 2022). Additionally, keeping

300   hot models may also be useful from an impact perspective as they may provide a clearer picture of future changes, as internal variability is less likely to obscure changes. This is similar to the rationale behind using high-emission scenarios in impact studies, such as SSP8.5, even though they may not be considered realistic scenarios anymore (e.g. Hausfather and Peters, 2020). It is important to consider worst-case scenarios when analysing potential outcomes, as high levels of greenhouse gas emissions, or high model sensitivity, such as those projected in SSP8.5 or high ECS models, are not unrealistic, even though

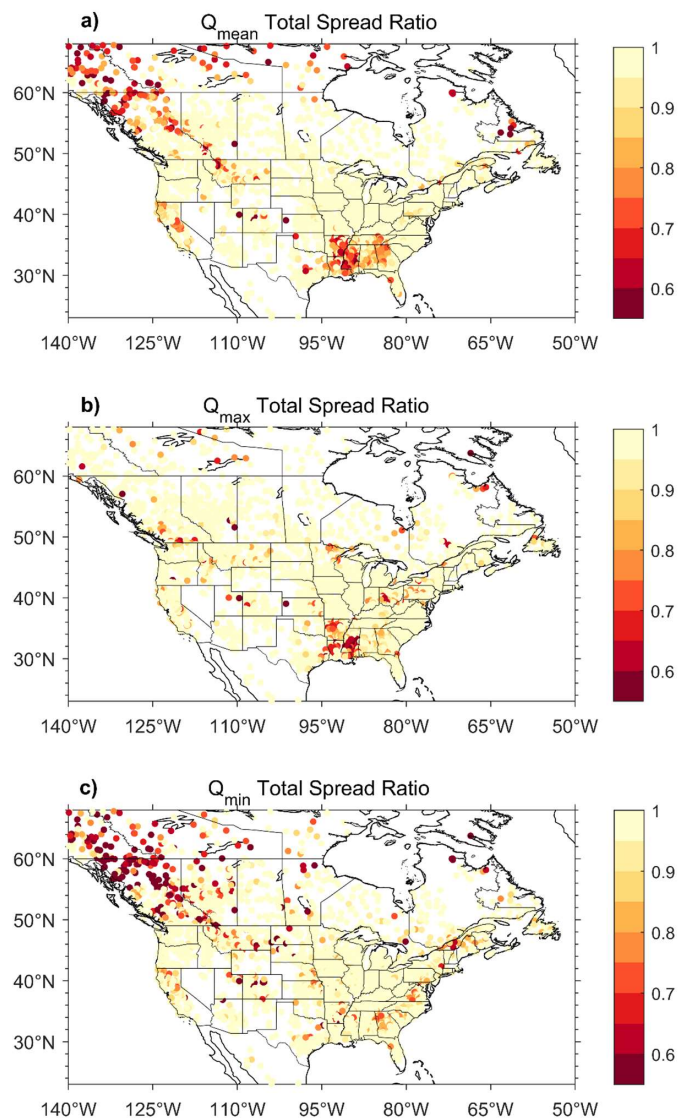
305   they may be less likely. While it is valuable to consider these high-end scenarios, it should be made clear that they are indeed worst-case scenarios.

In this study, the question of whether to remove the "hot models" for impact studies is complex. Results showed that for about one-third of all catchments, removing these models increased the future uncertainty of streamflow. This suggests that

310   these "hot outliers" may not always be "hydrological outliers" when put through a hydrological modeling process. Hydrological models are well-known for being highly non-linear integrators of weather variables such as temperature and precipitation, and these results align with findings from other studies that have demonstrated the complex relationship between climate model projections and hydrological projections (e.g. Chen et al., 2016; Ross and Najjar, 2019). The fact that the CMIP6 hot climate models tend to be wet models may also be a factor in these results, as increased evapotranspiration

315   could be offset by increased precipitation, leading to somewhat average results for the wrong reasons.

The regional impact of model importance is shown by comparing Figures 12 and 13, which demonstrate the total spread ratio resulting from removing a single climate model and creating an 18-member ensemble. CanESM5 (Figure 12) and NESM3 (Figure 13) have the highest global sensitivity in this study. Removing CanESM5 leads to a clear reduction of total spread in

320   Alaska and Yukon (for Qmean and Qmin) and in the Southeast USA for Qmax, indicating that CanESM5 is an outlier in these regions. Conversely, removing NESM3 does not result in significant decreases in spread over most of the study domain, demonstrating that it is not an outlier in most regions. This underscores the strong regional differences among globally identified hot models.

325 **Figure 12. Total spread ratio for Q_{mean} (a), Q_{max} (b), and Q_{min} (c) resulting from the removal of a single climate model (CanESM5).**

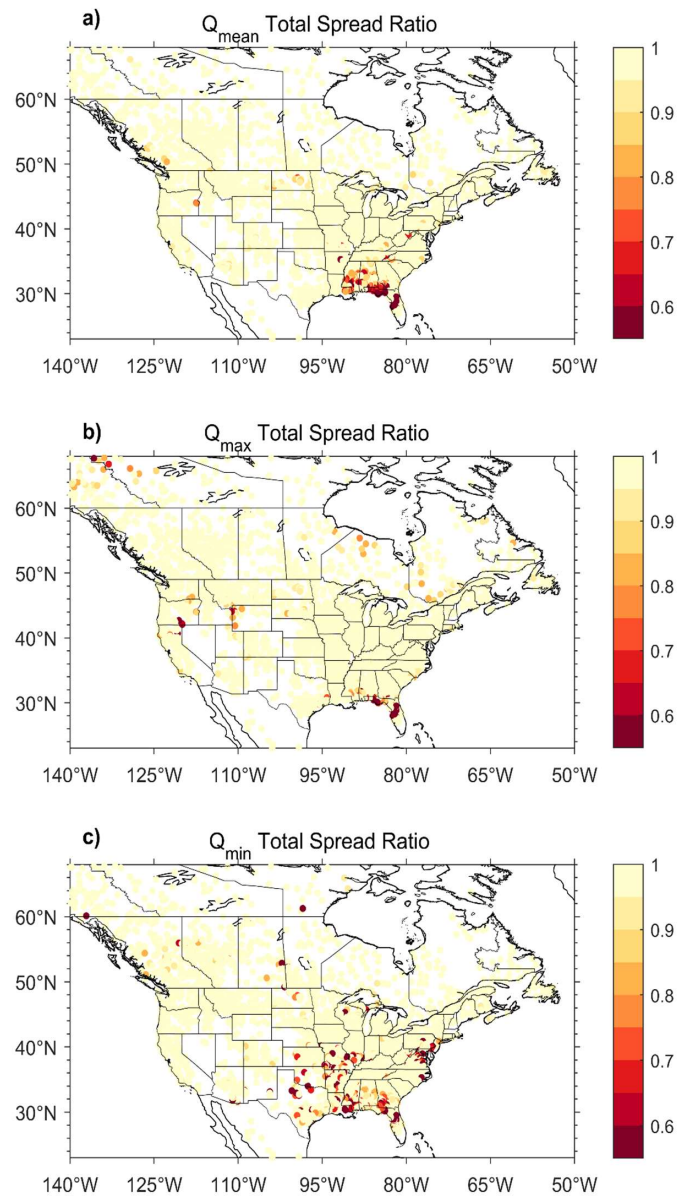



Figure 13. Total spread ratio for Q_{mean} (a), Q_{max} (b), and Q_{min} (c) resulting from the removal of a single climate model (NESM3).

The only uncertainty in this study is that originating from GCM/EMSs. As stated earlier, in most impact studies, additional
330   sources of uncertainty would also be incorporated. Additional greenhouse gases emissions scenarios would be selected as
well as other impact models (e.g. hydrology models). Downscaling and additional bias correction may be performed. These
additional components are likely to generate additional uncertainty which may, in some cases, dwarf that of climate models.
As such, many of the differences observed in this paper between the original and reduced climate model ensembles may
have little impact on the final uncertainty estimation. For example, for low flows, many studies have shown that most of the
335   uncertainty lies within the hydrology models (e.g. Giuntoli et al., 2018; Krysanova et al., 2018; Trudel et al., 2017) and
removing climate models would have no impact on uncertainty.

The results show that there is no simple answer as to whether or not including hot models in climate change impact studies.
In the absence of any computational limitations, we would recommend using as many climate models as possible and study
340   *at posteriori* the impact of including hot models or not. If a selection of a subset of climate models is necessary (e.g. inability
to use a large ensemble due to limited computational capability or costly to run impact models) removing hot models may be
a reasonable option. Evaluating climate model fitness for impact studies is a difficult endeavour, and in addition to ECS,
additional performance metrics should also carefully be taken into account.

5. Conclusion

345   This study examines the impact of removing a subset of hot climate models on the spread of future projections of streamflow
for 3,107 North American catchments. Three streamflow metrics were considered: mean annual streamflow, as well as the
mean of the annual maximum and minimum streamflow, over the reference period (1971-2000) and future period (2070-
2099).

350   Hot climate models are determined based on their global equilibrium climate sensitivity (ECS), whereas impact studies
typically focus on the local to regional scale. The hot climate models remain among the hottest in our regional evaluation,
but they also tend to be among the wettest, potentially leading to a complex hydrological response.

Our research revealed mixed impacts of removing the hot climate models. A decrease in the variability of projected
355   streamflow metrics was generally observed in Canada and Alaska, the southeast US, and the Pacific coast of the US.
However, in other regions, removing the hot models resulted in no changes, and in some cases, even increases in the
variability of projected flows. This suggests that the hot models are not necessarily hydrological outliers, raising questions
about using global performance metrics rather than regional ones for model selection.

360 The findings of this study emphasize the importance of carefully selecting climate models and the potential risks of including inadequate models in impact studies. In the absence of constraints, it is recommended to use as many climate models as possible in determining impact uncertainty and to assess the impact of subsets of climate models (based on high global equilibrium climate sensitivity or other performance metrics) a posteriori to assess the sensitivity of the impact model to climate model selection. These results highlight the need for further research on climate model fitness and the proper

365 selection of model subsets for impact studies.

Code and data availability.

The hydrometeorological data used in this study was obtained from the HYSETS database, which is available at https://doi.org/10.17605/OSF.IO/RPC3W (Arsenault et al., 2020). The CMIP6 GCM model outputs are accessible through

370 the Earth System Grid Federation Portal at Lawrence Livermore National Laboratory (https://esgf-node.llnl.gov/search/cmip5/). The processed data and the used codes are available via contacting the authors.

Author Contribution

The experiments were designed by FB, and they were carried out by MRA. The findings were analysed 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

375 provided editorial feedback on the paper's early draughts.

Competing interests.

The authors declare that they have no conflict of interest.

Acknowledgments

380 This work was supported by the Natural Science and Engineering Research Council (NSERC) of Canada.

References

Arsenault, R., Brissette, F., Chen, J., Guo, Q., and Dallaire, G.: NAC [2] H: The North American Climate Change and Hydroclimatology Data Set, Water Resour. Res., 56, https://doi.org/10.1029/2020WR027097, 2020a.

Arsenault, R., Brissette, F., Martel, J.-L., Troin, M., Lévesque, G., Davidson-Chaput, J., Gonzalez, M.C., Ameli, A., and

385 Poulin, A.: A comprehensive, multisource database for hydrometeorological modeling of 14,425 North American watersheds. Sci. Data, 7, 243, https://doi.org/10.1038/s41597-020-00583-2, 2020b.

Cannon, A. J., Piani, C., and Sippel, S.: Bias correction of climate model output for impact models, in: Climate Extremes and Their Implications for Impact and Risk Assessment, Elsevier, 77-104, https://doi.org/10.1016/B978-0-12-814895-

2.00005-7, 2020.

390 Chen, J., Li, C., Brissette, F. P., Chen, H., Wang, M., and Essou, G. R.: 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.

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

395 Chen, J., Brissette, F. P., and Lucas-Picher, P.: Transferability of optimally-selected climate models in the quantification of climate change impacts on hydrology, Climate Dynamics, 47(9), 3359-3372, 2016.

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(12), 2011.

Collins, M., Knutti, R., Arblaster, J., Dufresne, J.-L., Fichefet, T., Gao, X., Jr, W.J.G., Johns, T., Krinner, G., Shongwe, M.,
400 Weaver, A.J., Wehner, M., Allen, M.R., Andrews, T., Beyerle, U., Bitz, C.M., Bony, S., Booth, B.B.B., Brooks, H.E., Brovkin, V., Browne, O., Brutel-Vuilmet, C., Cane, M., Chadwick, R., Cook, E., Cook, K.H., Eby, M., Fasullo, J., Forest, C.E., Forster, P., Good, P., Goosse, H., Gregory, J.M., Hegerl, G.C., Hezel, P.J., Hodges, K.I., Holland, M.M., Huber, M., Joshi, M., Kharin, V., Kushnir, Y., Lawrence, D.M., Lee, R.W., Liddicoat, S., Lucas, C., Lucht, W., Marotzke, J., Massonnet, F., Matthews, H.D., Meinshausen, M., Morice, C., Otto, A., Patricola,
405 C.M., Philippon, G., Rahmstorf, S., Riley, W.J., Saenko, O., Seager, R., Sedláček, J., Shaffrey, L.C., Shindell, D., Sillmann, J., Stevens, B., Stott, P.A., Webb, R., Zappa, G., Zickfeld, K., Joussaume, S., Mokssit, A., Taylor, K., and Tett, S.: Long-term Climate Change: Projections, Commitments and Irreversibility 108, 2013.

Cox, P.M., Huntingford, C., and Williamson, M.S.: Emergent constraint on equilibrium climate sensitivity from global
temperature variability, Nature, 553, 319–322, https://doi.org/10.1038/nature25450, 2018.

410 dos Santos, F. M., de Oliveira, R. P., and Mauad, F. F.: Lumped versus distributed hydrological modeling of the Jacaré-Guaçu Basin, Brazil, Journal of Environmental Engineering, 144(8), 04018056. https://doi.org/10.1061/(ASCE)EE.1943-7870.0001397, 2018.

Eyring, V., Bony, S., Meehl, G.A., Senior, C.A., Stevens, B., Stouffer, R.J., and Taylor, K.E.: Overview of the Coupled
Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geosci. Model Dev., 9,
415 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016, 2016.

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., Pendergrass, A.G., Pincus, R., Ruane, A.C., Russell, J.L., Sanderson, B.M., Santer, B.D., Sherwood, S.C., Simpson, I.R., Stouffer, R.J., and Williamson, M.S.: Taking climate model evaluation to the
420 next level, Nat. Clim. Change, 9, 102–110, https://doi.org/10.1038/s41558-018-0355-y, 2019.

Flynn, C.M., and Mauritsen, T.: On the climate sensitivity and historical warming evolution in recent coupled model
ensembles, Atmospheric Chem. Phys., 20, 7829–7842, https://doi.org/10.5194/acp-20-7829-2020, 2020.

Giuntoli, I., Villarini, G., Prudhomme, C., and Hannah, D. M.: Uncertainties in projected runoff over the conterminous
United States, Climatic Change, 150, 149-162, 2018.

425 Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance
criteria: Implications for improving hydrological modelling, Journal of hydrology, 377(1-2), 80-91, 2009.

Hausfather, Z., Marvel, K., Schmidt, G.A., Nielsen-Gammon, J.W., and Zelinka, M.: Climate simulations: recognize the 'hot
model' problem, Nature, 605, 26–29, https://doi.org/10.1038/d41586-022-01192-2, 2022.

Hausfather, Z., and Peters, G. P.: Emissions–the 'business as usual'story is misleading, Nature, 577 (7792), 618-620, 2020.

430 Hirabayashi, Y., Tanoue, M., Sasaki, O., Zhou, X., and Yamazaki, D.: Global exposure to flooding from the new CMIP6
climate model projections, Sci. Rep., 11, 3740, https://doi.org/10.1038/s41598-021-83279-w, 2021.

IPCC: Intergovernmental Panel on Climate Change, Sixth Assessment Report: The Physical Science Basis, 2021.

IPCC: Intergovernmental Panel on Climate Change. Fifth Assessment Report: Climate Change 2014 Synthesis Report, 2014.

Karlsson, I. B., Sonnenborg, T. O., Refsgaard, J. C., Trolle, D., Børgesen, C. D., Olesen, J. E., Jeppesen, E., and Jensen, K.
435 H.: Combined effects of climate models, hydrological model structures and land use scenarios on hydrological
impacts of climate change, *Journal of Hydrology*, *535*, 301–317, https://doi.org/10.1016/j.jhydrol.2016.01.069,
2016

440   Knutti, R.: The end of model democracy?: An editorial comment, Clim. Change, 102, 395–404. https://doi.org/10.1007/s10584-010-9800-2, 2010.

Knutti, R., Rugenstein, M.A.A., and Hegerl, G.C.: Beyond equilibrium climate sensitivity, Nat. Geosci., 10, 727–736. https://doi.org/10.1038/ngeo3017, 2017.

Kreienkamp, F., Lorenz, P., and Geiger, T.: Statistically Downscaled CMIP6 Projections Show Stronger Warming for Germany, Atmosphere, 11, 1245, https://doi.org/10.3390/atmos11111245, 2020.

445   Krysanova, V., Donnelly, C., Gelfan, A., Gerten, D., Arheimer, B., Hattermann, F., and Kundzewicz, Z. W.: How the performance of hydrological models relates to credibility of projections under climate change, Hydrological Sciences Journal, 63(5), 696-720, 2018.

Liang, Y., Gillett, N.P., and Monahan, A.H.: Climate Model Projections of 21st Century Global Warming Constrained Using the Observed Warming Trend, Geophys. Res. Lett., 47, e2019GL086757, https://doi.org/10.1029/2019GL086757,

450   2020.

Maraun, D.: Bias correcting climate change simulations-a critical review, Current Climate Change Reports, 2, 211-220, 2016.

Martel, J.-L., Brissette, F., Troin, M., Arsenault, R., Chen, J., Su, T., and Lucas-Picher, P.: CMIP5 and CMIP6 Model Projection Comparison for Hydrological Impacts Over North America, Geophys. Res. Lett., 49, e2022GL098364,

455   https://doi.org/10.1029/2022GL098364, 2022.

Meehl, G.A., Covey, C., Delworth, T., Latif, M., McAvaney, B., Mitchell, J.F.B., Stouffer, R.J., and Taylor, K.E.: THE WCRP CMIP3 Multimodel Dataset: A New Era in Climate Change Research, Bull. Am. Meteorol. Soc., 88, 1383–1394, https://doi.org/10.1175/BAMS-88-9-1383, 2007.

Miara, A., Macknick, J.E., Vörösmarty, C.J., Tidwell, V.C., Newmark, R., and Fekete, B.: Climate and water resource

460   change impacts and adaptation potential for US power supply, Nat. Clim. Change, 7, 793–798. https://doi.org/10.1038/nclimate3417, 2017.

Nijsse, F.J.M.M., Cox, P.M., and Williamson, M.S.: Emergent constraints on transient climate response (TCR) and equilibrium climate sensitivity (ECS) from historical warming in CMIP5 and CMIP6 models, Earth Syst. Dyn., 11, 737–750, https://doi.org/10.5194/esd-11-737-2020, 2020.

465   Palmer, T.E., McSweeney, C.F., Booth, B.B.B., Priestley, M.D.K., Davini, P., Brunner, L., Borchert, L., and Menary, M.B.: Performance based sub-selection of CMIP6 models for impact assessments in Europe (preprint), Earth system change: climate prediction, https://doi.org/10.5194/esd-2022-31, 2022.

Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation. Journal of Hydrology, 279(1–4), 275–289, https://doi.org/10.1016/S0022-1694(03)00225-7, 2003.

470   Prajapati, S., Sabokruhie, P., Brinkmann, M., and Lindenschmidt, K.-E.: Modelling Transport and Fate of Copper and Nickel across the South Saskatchewan River Using WASP—TOXI, *Water*, *15*(2), Article 2, https://doi.org/10.3390/w15020265, 2023.

Reed, S., Koren, V., Smith, M., Zhang, Z., Moreda, F., Seo, D.-J., and Participants, D.: Overall distributed model intercomparison project results, Journal of Hydrology, 298(1–4), 27–60.

475   https://doi.org/10.1016/j.jhydrol.2004.03.031, 2004.

Reichler, T., and Kim, J.: How Well Do Coupled Models Simulate Today's Climate? Bull. Am. Meteorol. Soc., 89, 303–312, https://doi.org/10.1175/BAMS-89-3-303, 2008.

Ribes, A., Qasmi, S., and Gillett, N.P.: Making climate projections conditional on historical observations, Sci. Adv., 7, eabc0671, https://doi.org/10.1126/sciadv.abc0671, 2021.

480   Riboust, P., Thirel, G., Le Moine, N., and Ribstein, P.: Revisiting a simple degree-day model for integrating satellite data: Implementation of SWE-SCA hysteresis, Journal of Hydrology and Hydromechanics, 67(1), 70–81, https://doi.org/10.2478/johh-2018-0004, 2019.

Ross, A. C., and Najjar, R. G.: Evaluation of methods for selecting climate models to simulate future hydrological change, Climatic Change, 157(3), 407-428, 2019.

485   Sabokruhie, P., Akomeah, E., Rosner, T., and Lindenschmidt, K.-E.: Proof-of-concept of a quasi-2d water-quality modelling approach to simulate transverse mixing in rivers, *Water*, *13*(21), 3071, 2021.

Sherwood, S.C., Webb, M.J., Annan, J.D., Armour, K.C., Forster, P.M., Hargreaves, J.C., Hegerl, G., Klein, S.A., Marvel, K.D., Rohling, E.J., Watanabe, M., Andrews, T., Braconnot, P., Bretherton, C.S., Foster, G.L., Hausfather, Z., von

490   der Heydt, A.S., Knutti, R., Mauritsen, T., Norris, J.R., Proistosescu, C., Rugenstein, M., Schmidt, G.A., Tokarska, K.B., and Zelinka, M.D.: An Assessment of Earth's Climate Sensitivity Using Multiple Lines of Evidence, Rev. Geophys., 58, e2019RG000678, https://doi.org/10.1029/2019RG000678, 2020.

Shiogama, H., Ishizaki, N.N., Hanasaki, N., Takahashi, K., Emori, S., Ito, R., Nakaegawa, T., Takayabu, I., Hijioka, Y., Takayabu, Y.N., and Shibuya, R.: Selecting CMIP6-Based Future Climate Scenarios for Impact and Adaptation Studies, SOLA, 17, 57–62, https://doi.org/10.2151/sola.2021-009, 2021.

495   Su, T., Chen, J., Cannon, A. J., Xie, P., and Guo, Q.: Multi-site bias correction of climate model outputs for hydro-meteorological impact studies: An application over a watershed in China, Hydrological Processes, 34, 2575-2598, https://doi.org/10.1002/hyp.13750, 2020.

Tabari, H.: Climate change impact on flood and extreme precipitation increases with water availability, Sci. Rep., 10, 13768, https://doi.org/10.1038/s41598-020-70816-2, 2020.

500   Tarek, M., Brissette, F. P., and Arsenault, R.: 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, 2020.

Tokarska, K.B., Stolpe, M.B., Sippel, S., Fischer, E.M., Smith, C.J., Lehner, F., and Knutti, R.: Past warming trend constrains future warming in CMIP6 models, Sci. Adv., 6, eaaz9549, https://doi.org/10.1126/sciadv.aaz9549, 2020.

505   Troin, M., Martel, J. L., Arsenault, R., and Brissette, F.: Large-sample study of uncertainty of hydrological model components over North America, Journal of Hydrology, 609, 127766, 2022.

Trudel, M., Doucet-Généreux, P. L., and Leconte, R.: Assessing river low-flow uncertainties related to hydrological model calibration and structure under climate change conditions, Climate, 5(1), 19, 2017.

Valéry, A., Andréassian, V., and Perrin, C.: 'As simple as possible but not simpler': What is useful in a temperature-based
510   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., Xu, C.-Y., Chen, H., Guo, S., Xie, P., and Li, X.: 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, 2019.

515   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(2), 2006.

Zelinka, M.D., Myers, T.A., McCoy, D.T., Po-Chedley, S., Caldwell, P.M., Ceppi, P., Klein, S.A., and Taylor, K.E.: Causes of Higher Climate Sensitivity in CMIP6 Models, Geophys. Res. Lett., 47, e2019GL085782, https://doi.org/10.1029/2019GL085782, 2020.

520   Zhang, Y., Liu, H., Qi, J., Feng, P., Zhang, X., Liu, D.L., Marek, G.W., Srinivasan, R., and Chen, Y.: Assessing impacts of global climate change on water and food security in the black soil region of Northeast China using an improved SWAT-CO2 model, Sci. Total Environ., 857, 159482, https://doi.org/10.1016/j.scitotenv.2022.159482, 2023.

525