Evaluating climate model ensembles design for hydrological impact assessment: uncertainty attribution, transferability, and weighting

by

Mehrad Rahimpour ASENJAN

MANUSCRIPT-BASED THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE IN PARTIAL FULFILLEMENT FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
Ph.D.

MONTREAL, NOVEMBER 18TH, 2025

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

.

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mr. Francois Brissette, Thesis Supervisor
Department of Construction Engineering, École de Technologie Supérieure

Mr. Ricardo Izquierdo, President of the Board of Examiners
Department of Electrical Engineering at École de Technologie Supérieure

Mr. Richard Arsenault, Internal Evaluator
Department of Construction Engineering, École de Technologie Supérieure

Mr. Alex J. Cannon, External Evaluator
Climate Research Division at Environment and Climate Change Canada

THIS THESIS WAS PRENSENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND PUBLIC

MONTREAL, OCTOBER 30TH, 2025

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to Professors François Brissette and Richard Arsenault for their exceptional support and guidance throughout my PhD. I am especially honored to be Professor Brissette's last doctoral student; his mentorship, thoughtful advice, and steady encouragement have been a constant source of inspiration. His critical insights and high standards pushed me to grow as a researcher, while his generosity with his time and wisdom shaped the very foundation of this thesis. I am equally grateful to Professor Arsenault for his invaluable guidance, constructive feedback, and unwavering support, which have been essential to the success of this work.

I would also like to thank Jean-Luc Martel for his helpful feedback and support during key moments of this journey. His input has been greatly appreciated.

My heartfelt thanks go to my family, whose constant love and encouragement have carried me through the ups and downs of this experience. Their belief in me has been an endless source of strength. I'm equally grateful to my friends, whose presence, kindness, and conversations added joy and balance to this intense chapter of my life.

I am sincerely thankful to my colleagues and friends Pouya Sabokruhie, Amirreza Azimi, and Sajjad Azami for their support and encouragement. Their camaraderie, insightful exchanges, and motivation made this journey both more productive and more enjoyable.

To all of you, thank you for being part of this journey.

Évaluation de la conception des ensembles de modèles climatiques pour l'évaluation des impacts hydrologiques: attribution, transférabilité et pondération de l'incertitude

Mehrad Rahimpour ASENJAN

RÉSUMÉ

Comprendre les impacts des changements climatiques sur la disponibilité en eau et les extrêmes hydrologiques est essentiel pour une planification efficace des ressources hydriques. Les évaluations d'impact hydrologique reposent largement sur des ensembles de modèles climatiques globaux (GCM) pour quantifier les changements futurs et leurs incertitudes. L'utilisation d'ensembles multi-modèles (MME) pose toutefois plusieurs défis méthodologiques : choix des modèles, attribution des incertitudes et pondération des modèles constituant l'ensemble. L'utilisation d'un sous-ensemble réduit de GCMs implique des compromis entre faisabilité computationnelle et représentativité. De même, la pondération des modèles selon leur performance ou leur indépendance influence les résultats et les limites d'incertitude. Pourtant, malgré l'importance de ces choix, il existe peu de consensus sur les meilleures pratiques, et leur influence sur les projections hydrologiques demeure peu explorée.

Pour y répondre, trois objectifs guident cette thèse : (1) quantifier l'impact de la sélection de GCM basée sur des indices climatiques sur le transfert d'incertitude vers les projections hydrologiques ; (2) examiner les effets hydrologiques de l'inclusion ou exclusion de modèles à forte sensibilité climatique ; (3) comparer l'effet de différentes pondérations sur l'incertitude des projections de débits futurs. L'objectif n'est pas de promouvoir une stratégie unique, mais de comprendre comment les choix de conception d'ensembles influencent la propagation de l'incertitude climatique dans l'espace hydrologique.

La première analyse étudie le transfert d'incertitude climatique vers les sorties hydrologiques via des méthodes d'échantillonnage (ex. : algorithme KKZ) pour sélectionner des modèles selon des indices de température et de précipitation. L'expérience, menée sur 3 540 bassins nord-américains avec 20 GCM de l'ensemble CMIP5, deux approches de corrections de biais et trois modèles hydrologiques, montre que des ensembles réduits bien conçus peuvent conserver l'essentiel de la dispersion observée. Toutefois, ce transfert est non uniforme et non linéaire : de petites variations de précipitations peuvent produire de fortes différences de débits, surtout aux extrêmes, influencées par la structure du modèle hydrologique et les caractéristiques du bassin.

La deuxième analyse évalue l'exclusion des modèles à forte sensibilité climatique (« modèles chauds ») sur 3107 bassins nord-américains, en utilisant 19 GCM de l'ensemble CMIP6 dont cinq « modèles chauds ». Leur retrait réduit l'incertitude dans certaines régions (Alaska, sud-ouest des É.-U., parties du Canada) mais l'augmente ailleurs, soulignant l'importance d'évaluer les GCM sur des critères régionaux, et pas seulement globaux.

Enfin, la thèse teste plusieurs schémas de pondération via une expérience dans une pseudo-réalité, où chaque GCM est traité comme la « vraie » réalité future. Six méthodes sont

appliquées à 22 GCM couplés à un modèle hydrologique sur 3 107 bassins. Les pondérations inégales basées sur les observations historiques améliorent les projections climatiques, mais l'effet sur les débits est atténué, surtout après correction de biais, laquelle réduit la sensibilité aux choix de pondération.

Cette thèse apporte des perspectives nouvelles sur la conception des ensembles de modèles climatiques pour les évaluations d'impact hydrologique. Elle souligne que la construction d'ensembles doit intégrer le comportement pertinent pour l'impact (variabilité, saisonnalité des débits) et trouver un équilibre entre diversité, efficacité computationnelle et pertinence pour l'application visée.

Mots-clés: changement climatique, modélisation hydrologique, propagation des incertitudes, conception d'ensembles, pondération, correction de biais, modèles chauds, pseudo-réalité, débits

Evaluating Climate Model Ensembles design for Hydrological Impact Assessment: Uncertainty Attribution, Transferability, and Weighting

Mehrad Rahimpour ASENJAN

ABSTRACT

Understanding the impacts of climate change on water availability and hydrological extremes is critical for effective water resources planning. Hydrological impact assessments rely heavily on global climate model (GCM) ensembles to quantify future changes and their associated uncertainties. The use of multi-model ensembles (MMEs), however, presents several methodological challenges, including model selection, uncertainty attribution, and ensemble weighting. Selecting a reduced subset from an ever-growing pool of GCMs introduces methodological trade-offs between computational feasibility and ensemble representativeness. Similarly, weighting the individual GCMs by performance or by independence affects the outcome as well as its uncertainty limits. Yet, despite the critical role of these decisions, there is little consensus on best practices, and the influence of these design strategies on hydrological projections remains underexplored.

To tackle these issues, three specific research objectives are pursued in this thesis: (1) to quantify the impact of GCM selection based on climate indices on uncertainty transferability to hydrological projections; (2) to examine the hydrological implications of including or excluding high-sensitivity climate models in multi-model ensembles; and (3) to compare the effects of different GCM weighting schemes on the uncertainty of future streamflow projections. Rather than promoting a single optimal strategy, the objective is to understand how different ensemble design choices affect the propagation of climate uncertainty into hydrological space.

The first analysis investigates the transferability of climate uncertainty to hydrological outputs by applying sampling methods such as the KKZ algorithm to sub-select climate models based on temperature and precipitation indices. This experiment was conducted across 3,540 North American catchments using 20 CMIP5 GCMs, two bias correction methods and three conceptual hydrological models. Results show that when carefully designed, reduced ensembles can retain most of the spread observed in streamflow projections derived from the full ensemble. However, the translation of uncertainty is non-uniform and nonlinear, meaning small differences in climate inputs, particularly precipitation, may result in large variations in streamflow, especially for high and low flow regimes.

Secondly, the thesis examines the effect of excluding high Equilibrium Climate Sensitivity (ECS) models, referred to as “hot” models, on projected streamflow. Exclusion of these models reduces the spread of projected streamflow changes in some regions such as Alaska, southwestern U.S., and parts of Canada, but increased it in others, highlighting the need to evaluate GCMs using region-specific, rather than global, criteria.

Finally, the thesis assesses the performance of weighting schemes in GCMs through a pseudo-reality experiment, where each of the GCMs is, in turn, simulated as the “true” future. This allows an objective comparison of weighting performance against a known target in future where the true reality is unknown. The analysis applies six weighting approaches to an ensemble of 22 CMIP6 GCMs, coupled with a hydrological model across 3,107 North American catchments. Results indicate that unequal weighting by historical temperature and precipitation improves climate variable projections' quality. But for streamflow, these improvements are blunted, particularly if bias correction has been applied to inputs.

This thesis provides new insights into the design of climate model ensembles for hydrological impact assessments. It emphasizes that ensemble construction should not be based solely on climate performance metrics, but must incorporate impact-relevant behavior such as streamflow variability and seasonality. The findings advocate for a more pragmatic approach to ensemble design, balancing model diversity, computational efficiency, and relevance to the intended application.

Keywords: climate change, hydrological modeling, uncertainty propagation; ensemble design; model weighting; bias correction; hot models; pseudo-reality; streamflow

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

LIST OF ABREVIATIONS

ACCESS-ESM1-5	Australian Community Climate and Earth System Simulator Earth System Model version 1.5
AR5	Fifth Assessment Report
AR6	Sixth Assessment Report
BCC-CSM2-MR	Beijing Climate Center Climate System Model version 2, Medium Resolution
CANESM5	Canadian Earth System Model version 5
CDF	Cumulative Distribution Function (plural CDFs)
CMIP	Coupled Model Intercomparison Project
CMIP5	Coupled Model Intercomparison Project Phase 5
CMIP6	Coupled Model Intercomparison Project Phase 6
CO_2	Carbon dioxide
DDS	Dynamically Dimensioned Search
$\Delta P/P$	Relative change in precipitation
ΔT	Change in temperature
EC-Earth3	European Earth System Model version 3
EC-Earth3-Veg	EC-Earth3 with dynamic vegetation
ECS	Equilibrium Climate Sensitivity
ERA5	Fifth generation ECMWF atmospheric reanalysis
ESM	Earth System Model (plural ESMs)
FGOALS-g3	Flexible Global Ocean–Atmosphere–Land System Model grid-point version 3
GCM	General Circulation Model (plural GCMs)

GFDL-CM4	Geophysical Fluid Dynamics Laboratory Climate Model version 4
GFDL-ESM4	Geophysical Fluid Dynamics Laboratory Earth System Model version 4
GR4J	Génie Rural à 4 paramètres Journalier
GR4J_CN	GR4J with CemaNeige snow module
HRU	Hydrological Response Unit (plural HRUs)
HYSETS	Hydrometeorological dataset for North America
INM-CM4-8	Institute of Numerical Mathematics Climate Model version 4.8
INM-CM5-0	Institute of Numerical Mathematics Climate Model version 5.0
IPCC	Intergovernmental Panel on Climate Change
IPSL-CM6A-LR	Institut Pierre-Simon Laplace Climate Model version 6A, low resolution
IQR	Interquartile Range
IQR_{nd}	Non-dimensional Interquartile Range ratio
KGE	Kling–Gupta Efficiency
KKZ	Katsavounidis-Kuo-Zhang
LSTM	Long Short-Term Memory
MBC	Multivariate Bias Correction
MBCn	Multivariate Bias Correction (n-dimensional)
MIROC6	Model for Interdisciplinary Research on Climate version 6
ML	Machine Learning
MME	Multi-Model Ensemble
MPI-ESM1-2-HR	Max Planck Institute Earth System Model version 1.2, high resolution
MPI-ESM1-2-LR	Max Planck Institute Earth System Model version 1.2, low resolution

MRI-ESM2-0	Meteorological Research Institute Earth System Model version 2.0
NESM3	Nanjing University of Information Science and Technology Earth System Model version 3
NorESM2-LM	Norwegian Earth System Model version 2, low resolution
NorESM2-MM	Norwegian Earth System Model version 2, medium resolution
NSE	Nash–Sutcliffe Efficiency
NSERC	Natural Sciences and Engineering Research Council of Canada
P	Precipitation
PET	Potential Evapotranspiration
Q_{max}	Mean of annual maximum flows
Q_{mean}	Mean annual streamflow
Q_{min}	Mean of annual minimum flows
QM	Quantile Mapping
RCP	Representative Concentration Pathway (plural RCPs)
RCM	Regional Climate Model (plural RCMs)
RMSE	Root Mean Square Error
SCE-UA	Shuffled Complex Evolution – University of Arizona
σ	Standard deviation
σ_{nd}	Non-dimensional standard deviation ratio
SSP	Shared Socioeconomic Pathway (plural SSPs)
T	Temperature
TCR	Transient Climate Response
TSQM	Two-Stage Quantile Mapping

TS	Total Spread
TS_{nd}	Non-dimensional Total Spread ratio
US	United States

LIST OF SYMBOLS

%	Percent
°C	Degrees Celsius
Δ	Delta (change) symbol
∞	Infinity symbol
km	Kilometre(s)
km^2	Square kilometre(s)
m^3/s	Cubic metres per second
mm	Millimetre(s)

INTRODUCTION

Climate change has been one of the most severe challenges that humanity has been facing in recent decades. Since 1990, the Intergovernmental Panel on Climate Change (IPCC) has issued six assessment reports, which have documented the increasing certainty and magnitude of anthropogenic climate change and its consequences. The current Synthesis Report of the Sixth Assessment (AR6), published in 2023, confirms that rapid and worldwide alterations of the climate system have occurred over recent decades (IPCC, 2023). Adverse effects of climate change are, but not restricted to, global warming, increased sea levels, melting of glaciers, and even more and prolonged extreme events such as heavy precipitation, floods, droughts, and heatwaves (IPCC, 2014, 2023). Climate change severely threatens not only urban security, but also environmental stability and the long-term viability of economic and social development around the world. Socio-economic consequences range from reduced agricultural production and food security to displacement of communities and further pressure on public health systems (IPCC, 2023). Because of the scale and scope of these risks, adaptation and mitigation strategies to the present and future effects of climate change has become crucial.

Climate change risks can be reduced by implementing proper management strategies (Wilby & Dessai, 2010), which necessitate a thorough understanding of the magnitude and uncertainty of the projected changes. Reliable projections of future climate are therefore necessary to guide policymaking, enable risk management strategies, and influence sustainable development planning in the face of a changing climate (Knutti, Furrer, et al., 2010). Importantly, the effects of climate change are not uniform across the globe; they vary significantly by region due to differences in geography, climate systems, and socio-economic contexts (IPCC, 2021). As such, decision-making must be informed at national and local levels by regional considerations through regional-scale impact assessments, as opposed to global scale trends. Regional studies are required to inform adaptation planning, design robust infrastructure, and implement effective policies across sectors such as flood control, water resources management, and urban planning.

To evaluate the effects of climate change, researchers typically use mathematical models of the Earth's climate system referred to as General Circulation Models (GCMs) or Earth System models (ESMs) GCMs are highly advanced numerical models that incorporate interactions between the atmosphere, oceans, land surface, and cryosphere to project future and past climate conditions according to various greenhouse gas emission scenarios (Illangasingha et al., 2023). ESMs are advanced versions of GCMs that include additional components such as the carbon cycles and dynamic vegetation. Greenhouse gas emission scenarios, referred to as Representative Concentration Pathways (RCPs) or the more recent Shared Socioeconomic Pathways (SSPs), represent different trajectories of socio-economic development, energy use, and policy choices, each associated with a corresponding radiative forcing level (IPCC, 2021; Siabi et al., 2023). However, translating the outputs of GCMs into useful information for regional planning requires diligent assessment to overcome their shortfalls and associated uncertainties (Knutti, Furrer, et al., 2010).

In hydrological impact studies, the use of GCM outputs involves two key steps. The first is to take raw GCM data, which usually have coarse spatial resolution and contain systematic biases, and downscale and bias-correct them to provide site-specific climate data (Maraun et al., 2010; Menapace et al., 2025; Teutschbein & Seibert, 2012). This can be achieved using statistical techniques, which calibrate large-scale GCM outputs against local observations (Miller et al., 2025), or dynamical downscaling (Fallah et al., 2025), which uses Regional Climate Models (RCMs) nested within GCMs. In this approach, the GCM supplies large-scale atmospheric and oceanic boundary conditions, that are periodically updated, to drive the RCM. The RCM then simulates local processes like topographic effects and land–atmosphere interactions at a finer spatial resolution while remaining consistent with the global circulation. Second, the downscaled climate outputs, i.e., temperature and precipitation, are used to force hydrological models in order to simulate catchment-scale processes such as streamflow and evapotranspiration (e.g. Vano et al., 2014). The resulting hydrological projections can subsequently be used for the aim of informing real decisions, such as design of infrastructure (e.g., storm water systems, dams), water supply management, and flood risk assessment (Wilby & Dessai, 2010).

Climate modeling has progressed significantly in the last few decades, resulting in an outburst of GCMs that are presently available. The Coupled Model Intercomparison Project Phase 5 (CMIP5) and its successor, CMIP6, have been the key contributors to the increase in the number of climate models. CMIP5 featured simulations from 62 different models developed by 29 modeling centers worldwide (Taylor et al., 2012; *Why So Many Climate Models? | California Climate Commons*). CMIP6 expanded this further, involving around 100 models contributed by 44 institutions (Eyring et al., 2016; Hausfather, 2019). In addition to increasing the ensemble size, CMIP6 introduced major improvements such as enhanced spatial resolution, a broader range of greenhouse gas concentration scenarios (Shared Socioeconomic Pathways – SSPs), and more advanced representations of Earth system processes.

The growing diversity of climate models allows researchers to use large multi-model ensembles to better quantify uncertainty in future climate projections, account for structural differences among models, and identify robust signals of change (Lehner et al., 2020). However, from a practical standpoint, using all available models in hydrological impact assessments is infeasible. Running high-resolution downscaling and hydrological simulations across dozens of GCMs demands substantial computational resources, time, and data management capacity. Although modern hardware and storage solutions can accommodate large volumes of climate data, the bottleneck often lies in the workflow: repeatedly bias-correcting, forcing, calibrating, and analyzing hydrological simulations for dozens of GCMs requires substantial researcher time, model-specific tuning, and multi-stage post-processing. In operational and academic settings, the limiting factor is therefore less about raw computing power and more about managing, interpreting, and making decisions from a very large set of uncertain projections. Large ensembles can complicate impact assessment by producing an overwhelming range of possible outcomes, making it difficult for practitioners to extract actionable information. As a result, selecting a representative subset of credible models, or applying weighting strategies, remains a practical approach to maintaining interpretability and analytical tractability while still capturing the essential dimensions of uncertainty.

Since the predictions provided by these models are so critical, caution needs to be taken to ensure that the GCMs selected are representative and credible (Cannon, 2015; Dubrovsky et al., 2015). An ill-advised model selection can lead to biased estimates of the magnitude or timing of hydrological effects and, subsequently, to suboptimal planning measures. A selection of GCMs that is transparent and well-argued, therefore, together with careful downscaling and hydrological simulation, is necessary to provide useful and credible climate impact assessments (Chowdhury & Eslamian, 2014). Therefore, the selection of a representative subset of GCMs has become a key methodological issue in climate impact assessments (Chen et al., 2017; Knutti, Furrer, et al., 2010). The subset that captures the most significant dimensions of climate model uncertainty, but is small enough to manage, must be carefully selected (Herger et al., 2018; Wilcke & Bärring, 2016).

Given these constraints, the question of how to design GCM subsets that are both representative and computationally efficient remains central to the design of robust climate impact assessments. The selection of a smaller subset of GCMs from a large ensemble inevitably results in information loss. However, it is important to emphasize that subsetting does not reduce the actual uncertainty inherent in future climate projections, it only reduces the representation of that uncertainty (Wilcke & Bärring, 2016). In other words, fewer models mean fewer perspectives on possible climate futures, which can narrow the perceived range of outcomes without necessarily improving confidence in any one result. Therefore, the selected subset need to preserves the significant statistical and physical characteristics of the overall ensemble, such as central tendencies (means or medians), extremes, variability, and spatial coherence.

CHAPTER 1

LITERATURE REVIEW

1.1 Climate Change

Climate change is defined as the long-term changes in temperature, precipitation, and other weather patterns, and they can be caused by both human activity and natural processes like volcanic eruptions and variations in solar radiation (IPCC, 2021). However, the primary driver of the observed climate change since the late nineteenth century has been human activity. Global warming is one of the most prominent indications of this change. The global mean surface temperature has risen by roughly 1.09°C [0.95 to 1.20°C] in comparison to the pre-industrial baseline (1850–1900) (IPCC, 2023). Warming is not uniform across the globe, being stronger over land areas, approaching 2°C, and even more pronounced at higher latitudes. The main cause is the increase in atmospheric concentrations of carbon dioxide (CO_2) which trap outgoing longwave radiation and intensify the natural greenhouse effect (IPCC, 2023).

Greenhouse gases are emitted into the atmosphere from two primary sources. The first source is the natural systems, such as forest fires and volcanoes. It is noteworthy that the emissions and the sinks in the natural systems balance each other out, meaning that the greenhouse gases absorbed by the sinks, e.g., oceans, are of the same magnitude as the emission from sources, e.g., volcanoes (Yue & Gao, 2018). However, the addition of greenhouse gases from human activities, such as fossil fuel combustion, land-use change, agriculture, and industrial activities, interrupts the balance in the earth's system (IPCC, 2023).

With population growth and industrial advances in the last century, greenhouse gas emissions from human activities have risen to an unprecedented level in history. Representative Concentration Pathways (RCPs), defined as scenarios of different greenhouse gas concentration trajectories in the atmosphere, are used by climate modellers to assess the future of climate change. In addition, the most recent IPCC reports examine how socioeconomic

factors might evolve over the coming century (IPCC, 2023). Population, economic development, education, urbanization, and the pace of technological advancement are a few examples. These “Shared Socioeconomic Pathways” (SSPs) examine five potential futures and examine how various levels of climate change mitigation might be accomplished when the mitigation goals of RCPs are combined with the SSPs. According to the climate projections, even under an optimistic greenhouse gas emission scenario, the warming will exceed 1.5°C (SSP1-1.9) and 2°C (SSP1-2.6) (IPCC, 2023). Even with rigorous policies to reduce emissions, it will be challenging to control the warming without substantially lowering emissions in the upcoming decades (IPCC, 2023).

Another critical follow-up for climate change is the change in precipitation patterns. But, the distribution of changes in precipitation is more spatially variable. Over the last three decades, precipitation has generally increased in the higher latitudes of the northern hemisphere (north of 30°N) and the eastern part of North and South America. On the other hand, it has decreased in lower latitudes (30° to 10°) and in South Africa (IPCC, 2023). It is noteworthy that the results for other regions have more uncertainty, and the models do not provide consistent results. Overall, future changes are expected to amplify existing precipitation contrasts, with dry areas becoming drier and wet regions becoming wetter (Kundzewicz, 2008). It is also important to note that in arid regions, small absolute changes in precipitation may translate into large percentage changes because the baseline precipitation is low (IPCC, 2021).

1.2 Water Resources and Climate Change

The sustainable development of civilization has always been in debt to the availability of water resources since agriculture and food security are entirely dependent on accessible water. In addition, water is a crucial asset to industries, hydroelectricity, and environmental usage. The increased demand caused by the population increase puts extra pressure on this resource. In addition, climate change is affecting the distribution and adding uncertainty to future water distribution patterns.

The impact of climate change is not limited to warming temperatures and changes in precipitation. In addition to these effects, higher temperature increases the evaporation rate and results in surface drying, consequently raising the intensity and duration of drought (Trenberth, 2011). Because of the lack of long-term data for drought variables such as soil moisture, detection and attribution of drought is challenging (Easterling et al., 2017). Nonetheless, recent studies robustly show that climate change is already impacting droughts in several regions (Cook et al., 2018). For instance, Dai (2013) concluded that decreasing precipitation combined with increased evaporation will contribute to extreme drought events in many regions in the 21st century. Also, Woodhouse et al. (2016) concluded that the recent droughts in the Colorado river basin resulted from the recent warming.

The warming will also impact extreme precipitation, the primary flood generator in most regions. (Fischer & Knutti, 2016). Since the 1970s, the frequency of intense rainfall has increased, and more intense and prolonged droughts have been seen worldwide (IPCC, 2023). The capacity of the atmosphere to hold moisture is temperature-dependent and is governed by the Clausius–Clapeyron relationship, which suggests an increase of about 7% in atmospheric water-holding capacity per 1 °C of warming, provided moisture is available (Trenberth, 2011; Westra et al., 2014). This intrinsically means that a warmer atmosphere can retain more moisture, contributing to more extreme precipitation events (Trenberth, 2011). However, the link between warming and precipitation extremes is more complex. Global climate models project that average precipitation will increase more slowly, on the order of 2–3% per 1 °C of warming, due to energy balance constraints (Held & Soden, 2006; Allen & Ingram, 2002). Moreover, the magnitude of change varies regionally and depends on the intensity and duration of rainfall (Lenderink & Fowler, 2017; Martel et al., 2021; Westra et al., 2014).

The temperature rise will further alter the ratio of rain to snow precipitation, which has already been seen in many higher latitude regions, (e.g., Mote, 2003). Higher temperatures will result in reduced snowfall compared to an increase in rainfall. Warming will also cause earlier snowmelt, which will lead to lower water resources in summer (Trenberth, 2011). For instance,

in the western United States, among many other regions, the snowpack formed in winter provides freshwater as it melts in summer and spring. The lower snowfall rate, and earlier melt of the snowpack, will reduce the freshwater storage capacity. Furthermore, a higher rainfall rate is expected to increase the flood rate in winter and spring. Reduced natural freshwater and higher flood probability will challenge the managers' current flood control and reservoir management policies (Knowles et al., 2006).

In summary, climate change is changing the hydrological cycle, consequently impacting the quantity (Milly et al., 2005; Mourato et al., 2015) and quality (Whitehead et al., 2009) of water resources. Hydrologic processes, availability of water resources, agriculture, and hydroelectricity will be impacted by climate change (Christensen & Lettenmaier, 2007). In other words, the existing risks for humans and the ecosystem, such as floods and droughts, will be even worse with the changing climate (IPCC, 2023). However, the risks can be mitigated by adopting appropriate management strategies that require understanding the changes' magnitude and uncertainty. Hence, assessment of regional climate change impacts, especially on watershed hydrology, becomes very important.

1.3 General Circulation Models

General circulation models, also called global climate models (GCMs), and earth-system models (ESMs), which add the biogeochemical cycle, use mathematical equations that represent physical processes (conservation of mass, energy, momentum, etc.) to simulate the interaction between the atmosphere, land surface, oceans, and sea ice (Trzaska & Schnarr, 2014). These include, most notably, the Navier–Stokes equations for atmospheric and oceanic motion (conservation of momentum), along with equations representing the conservation of mass, energy, and the transfer of radiation and water vapor. Each equation is solved on vertical and horizontal grid cells and multiple layers in the atmosphere and ocean. GCMs generally have coarse resolutions (100km to 500km), because running them on smaller scales would be computationally too expensive (Wilby et al., 2009). Despite this

limitation, GCMs remain essential tools for analyzing current climate dynamics and projecting future changes under various greenhouse gas emission scenarios.

1.3.1 CMIP5 and CMIP6

The Coupled Model Intercomparison Project Phase 5 (CMIP5) has gathered coordinated simulations from different climate modeling groups to bridge the gap in understanding climate changes in the future and past. CMIP5 is a multimodel context of climate change and variability (Taylor et al., 2012). The long-term simulations in the CMIP5 span the range of the nineteenth century to the twenty-first century and build upon the previous CMIP projects. On the other hand, near-term projections are added to the CMIP5, which start from the observed state of the climate to assess the predictability of near-future climatic patterns. Near-term projections will help scientists identify the predictable variables and the corresponding time scales of predictability. Model robustness, initialization method, and data quality are the primary determiners of the prediction skill (Taylor et al., 2012).

CMIP ensembles are often described as “ensembles of opportunity,” meaning they are not designed according to a formal experimental plan but instead consist of all simulations voluntarily contributed by modelling groups worldwide. As a result, CMIP archives contain models with differing levels of complexity, varying numbers of realizations, shared code bases, and unequal institutional representation. This lack of experimental design has important implications for statistical inference: the ensemble does not constitute a random or balanced sample of all plausible climate models, and its spread cannot be interpreted as a probabilistic measure of uncertainty. Instead, it reflects the structural diversity, historical choices, and modelling philosophies of participating centres. Recognizing CMIP as an ensemble of opportunity is therefore essential when interpreting uncertainty, selecting subsets, or assigning weights, as these decisions must account for biases, interdependencies, and uneven sampling across the ensemble.

The more recent phase 6 (CMIP6) (Eyring et al., 2016) has been slowly taking the place of the CMIP5, whcih was widely used in the last decade (Hirabayashi et al., 2021; Martel et al., 2022; Zhang et al., 2023). The sixth IPCC assessment report (AR6) is based on brand-new and state-of-the-art CMIP6 models (IPCC, 2023). The CMIP6 models include several new and updated emission pathways that investigate a much wider range of potential future outcomes than were covered by CMIP5. While the IPCC fifth assessment report (AR5) included four RCPs that assessed various potential future greenhouse gas emissions, these scenarios have been updated to include various climate policies. The updated scenarios, namely SSP1-2.6, SSP2-4.5, SSP4-6.0, and SSP5-8.5, each produce levels of radiative forcing in 2100 that are comparable to those of their predecessor in AR5.

Climate sensitivity plays a central role in interpreting differences across climate models. While Equilibrium Climate Sensitivity (ECS) is the most widely used metric, defined as the long-term global temperature response to a doubling of CO_2 under equilibrium conditions, it does not capture transient behaviour. An alternative measure, the Transient Climate Response (TCR), represents the temperature change at the time of CO_2 doubling under a gradual 1% yr^{-1} increase scenario. Because TCR reflects near-term warming under non-equilibrium conditions, it is often more relevant for mid-century impact studies. Including both ECS and TCR allows for a more comprehensive assessment of how models differ in their response to radiative forcing and clarifies why “hot models” may diverge in both magnitude and timing of warming.

The CMIP6 offers temperature and precipitation projections with a smaller spread than those of the CMIP5, and except in mountainous areas, the CMIP6-driven hydrological projections produce a narrower range of future mean and high flow values (Martel et al., 2022). However, the CMIP6 includes a subset of “hot models” which predict much higher warmings than previously predicted by CMIP5 (e.g. Kreienkamp et al., 2020). The “hot models” exhibit greater ECS and TCR values (Flynn & 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).

Recently more research has been aimed at constraining the ECS based on historical and paleoclimatic data (Knutti, Rugenstein, et al., 2017; Sherwood et al., 2020) or the emergent constraints (Cox et al., 2018; Nijsse et al., 2020). For example, using multiple lines of evidence, Sherwood et al., (2020) concluded that the likely (with a 66% chance) ECS value is between 2.6°C and 4.1°C. Consequently, the most recent reports published by the IPCC have narrowed the range of likely ECS range to 2.5 and 4°C (IPCC, 2021).

1.4 Downscaling

The GCMs operate on a coarse scale relative to many small-scale phenomena, such as clouds or topography. Furthermore, outputs for each grid cell are homogeneous, meaning that each grid holds only one value for each output over each grid. In other words, small-scale phenomena cannot be modeled adequately with GCMs and are therefore simplified and parameterized. Yet, in many cases, the impact models require outputs on a smaller scale to accurately represent the events. Furthermore, the GCM outputs are often biased (Quintana Seguí et al., 2010). To solve this issue, downscaling methods have been proposed. Downscaling methods relate the coarse resolution outputs of GCMs to the local and regional scale events and derive more detailed information from GCMs (Hewitson & Crane, 1996). The outputs of downscaling methods can be used as inputs for impact models for climate change impact assessment studies and hydrological modeling. The two primary downscaling methods are dynamical and statistical downscaling methods, each with its advantages and disadvantages. The general limitations, theory, and practice of downscaling are now well described in the literature (Chokkavarapu & Mandla, 2019; Fowler et al., 2007).

1.4.1 Dynamical downscaling

Dynamical downscaling methods use regional climate models (RCMs) to derive local-scale data from large-scale GCM outputs. In principle, RCMs are similar to GCM models, but with a smaller scale (10-50 km). RCMs use the outputs of GCMs as the boundary conditions to model small-scale phenomena, which were simplified in GCMs, such as complex terrain

topography. In particular, in regions with complex topography, e.g., coastal areas, where GCMs cannot model regionally significant processes, the RCM projection becomes extremely valuable (Ekström et al., 2015). The main strength of dynamical downscaling is that RCMs provide physically consistent climate data while capturing intricate and nonlinear interactions within the climate system (Williams et al., 2010).

Dynamical downscaling methods are physically based yet computationally expensive. As such, RCMs are only used with a small selection of GCMs for each region (Ekström et al., 2015). RCM outputs strongly depend on the parent GCM (Chokkavarapu & Mandla, 2019). In this case, the selection of parent GCMs and their representation of the whole set of GCMs should be considered very carefully. Furthermore, RCM outputs also contain substantial biases compared to the observed historical data (Muerth et al., 2013). Some of these biases are inherited from the driving GCMs through boundary conditions, while others arise from the RCMs themselves due to their internal parameterizations, numerical schemes, or representation of local processes such as orography and land–atmosphere interactions. These biases may hinder impact assessment models from appropriate simulation of the processes. As such, a bias correction step is necessary before using the RCM outputs (Teutschbein & Seibert, 2012).

1.4.2 Statistical downscaling

Statistical downscaling methods develop an empirical relationship between GCM outputs and meteorological data at various scales such as station scale (Wilby et al., 1998). Fundamentally, statistical downscaling methods assume a stationary relationship between the predictor (GCM output) and the predictand (local climate information), which stays the same under the changing climate (Fowler et al., 2007; Gutmann et al., 2022). The statistical relationship is then used to interpolate the future GCM projections of the studied variable to the local scale.

Statistical methods are simpler to apply, straightforward, and have low computational costs (Nasseri et al., 2013). In addition, using these methods, one can derive information that

essentially is not available in RCMs (Fowler et al., 2007; Gutmann et al., 2022). Yet, the climatic region may affect downscaling skill (Chokkavarapu & Mandla, 2019). Moreover, for variables like precipitation, where much of the variability arises from sub-grid processes (e.g. Prein et al., 2017) poorly captured by GCMs (Chen & Zhang, 2021), statistical downscaling can be less effective (Chen & Zhang, 2021; Hernanz et al., 2022; Maraun et al., 2010).

1.5 Bias correction

GCM and RCM simulations have systematic biases compared to observed historical records (François et al., 2020). The biases stem from several sources, such as the intrinsic limitations in our understanding of the climate system and the imperfect parameterization of physical processes in the models (Chen et al., 2021). Raw outputs from climate models are rarely applied directly in impact assessment studies because they would provide unreliable results, considering that impact models, especially hydrological models, are sensitive to the quality of the data (Dinh & Aires, 2023; Potter et al., 2020). Bias correction is now a common post-processing technique used to solve the issue and correct important statistical characteristics, to make the outputs applicable in practical applications (Chen et al., 2021; Dinh & Aires, 2023; Gutmann et al., 2022; Kim et al., 2019).

Early bias correction techniques were designed to be simple and computationally efficient. For instance, the delta change method adjusts future climate projections with additive or multiplicative constant factors derived from the difference between modeled and observed means over a certain reference period (Räty et al., 2014; Teutschbein & Seibert, 2012). While effective at eliminating mean biases, such methods are not capable of addressing other aspects of the distribution, i.e., variability or extremes (Beyer et al., 2020). By contrast, advanced methods such as quantile mapping (QM) align the cumulative distribution functions (CDFs) of the modeled data with those of the record data CDFs, allowing adjustment of not only the mean but also the variance, skewness, and extremes (Beyer et al., 2020; Cannon et al., 2015). Because of its versatility and robustness, QM has become one of the most popular univariate bias correction methods. However, when individually applied to a set of variables such as

temperature and precipitation, QM may distort inter-variable relationships, and this distortion may compromise impact modeling, particularly in studies that depend on the co-variation of variables (Zscheischler et al., 2019; Zscheischler & Seneviratne, 2017). This issue is particularly significant in hydrological applications, where joint variability between variables governs key processes like evapotranspiration, snow accumulation, and melt (Cannon, 2018).

Multivariate bias correction (MBC) techniques have been developed to overcome this limitation, with the objective of correcting the individual distribution of climate variables and preserving their interdependencies (Cannon, 2018; Cannon et al., 2015). Recent MBC methods are generally categorized under three broad classes. Marginal-dependence techniques involve multivariate bias adjustment methods that correct marginal distributions and dependence relationships independently (Cannon, 2018; Vrac, 2018). Successive conditional techniques apply corrections sequentially, where each variable is adjusted using information from variables that have already been corrected earlier in the sequence (Bárdossy & Pegram, 2012; Dekens et al., 2017). Finally, all-in-one techniques attempt to correct marginal distributions and dependence structures simultaneously (Robin et al., 2019). While each approach has its merits, they also have limitations. Successive conditional methods, for instance, are order-sensitive and decline in performance as the number of variables increases. All-at-once methods provide powerful corrections, but at the expense of significantly higher computational demands (François et al., 2020).

Despite the advances in methodology, bias correction remains a controversial topic. At the heart of the debate is an assumption of bias stationarity, which assumes that biases observed during the historical record will remain unchanged in the future (Ehret et al., 2012). This is typically a questionable assumption under changing climate conditions, especially for extreme events and precipitation mechanisms, where biases may develop over time (Chen et al., 2015, 2021; Miao et al., 2016).

Yet, when applied carefully, bias correction methods can significantly enhance the value added to climate model output applied to impact studies (Chen et al., 2021; Maraun, 2016). They

have been shown to improve the representation of key hydrological parameters, such as streamflow, precipitation intensity, and temperature trends (Meyer et al., 2019; Teutschbein & Seibert, 2012; Worako et al., 2022).

1.6 Hydrological modelling

Hydrological modeling is the fundamental step in the climate-hydrology modeling chain and allows the transformation of meteorological inputs to estimates of streamflow, which are required to conduct impact studies. A hydrological model is a simplified representation of the real water cycle that enables simulation and understanding of runoff generation and streamflow dynamics on a variety of spatial and temporal scales (Devi et al., 2015). These models vary in complexity of structure, data requirements, spatial discretization, and representation of processes, and are usually categorized as empirical, conceptual, or physically based models (Jajarmizadeh et al., 2012; Pandi et al., 2021). Hydrological models can further be categorized as lumped or distributed, depending on how they represent spatial variability within a catchment.

Empirical models (data-driven models) are based only on observed input-output correlations and lack the description of the internal physical mechanisms in a catchment. Empirical models tend to employ statistical or machine learning techniques (e.g., artificial neural networks, fuzzy logic) to derive the relationships between meteorological inputs and streamflow (Hauswirth et al., 2021). Empirical models are characterized by simplicity, low computational cost, but by an absence of physical interpretability and limited extrapolation capability outside historical conditions (Abdulkareem et al., 2018; Devi et al., 2015).

Conceptual models simplify the catchment dynamics to connected reservoirs (e.g., soil water, snow, groundwater), regulated by empirical or semi-empirical equations. Conceptual models describe major hydrological processes using a limited number of parameters, which are typically calibrated from observations (Liu et al., 2019; Merz et al., 2009). Conceptual models trade physical realism with computational expense and are best suited for applications at the

basin scale where high-resolution spatial data may not be accessible (Biondi et al., 2012; Devi et al., 2015; Kavetski et al., 2006).

Physically based models attempt to simulate water flow by solving governing mass, momentum, and energy conservation equations (Newman et al., 2017; Paniconi & Putti, 2015). They require large spatial sets of soil, topography, land use, and meteorology data and are run in a distributed mode (Paniconi & Putti, 2015; Vieux et al., 2004). Physically based models offer detailed process representations at the expense of being computationally intensive and are subject to parameterization and input data uncertainty (Devi et al., 2015; Paniconi & Putti, 2015).

Models also vary in their spatial representation. Lumped models simulate the catchment as a homogeneous unit of space, averaging both input and output spatially (Seiller et al., 2012; Van Lanen et al., 2024). Such models are used most often when streamflow at the catchment outlet is most significant, and their simplicity makes them amenable to being used for large ensemble runs (Devi et al., 2015; Seiller et al., 2012).

Distributed models specifically explain spatial variation in inputs, parameters, and processes. Distributed models disaggregate the catchment into grid squares or hydrological response units (HRUs) to more accurately simulate local hydrological responses (Abbott & Refsgaard, 1996). Although potentially capable of producing simulations closer to reality, distributed models fail to consistently outperform lumped models due to increased complexity, higher data demands, and the equifinality of parameter estimation (Beven, 2001). Semi-distributed models maintain some spatial heterogeneity (e.g., by employing HRUs or sub-basins) without full spatial resolution.

Regardless of structural type, most hydrological models must be calibrated to reconcile simulations with observed streamflow (Bárdossy, 2007; Kavetski et al., 2006). Calibration involves the optimization of model parameters that minimize discrepancies between simulated and observed data, often measured by performance metrics like the Nash-Sutcliffe Efficiency

(NSE) (Nash & Sutcliffe, 1970) and the Kling-Gupta Efficiency (KGE) (Kling & Gupta, 2009) (Arsenault et al., 2014; S. K. Singh & Bárdossy, 2012). Calibration routines often employ optimization algorithms like Dynamically Dimensioned Search (DDS) (Tolson & Shoemaker, 2007) and the Shuffled Complex Evolution – University of Arizona (SCE-UA).

In recent years, deep learning methods have emerged as powerful alternatives to traditional hydrological models (Althoff et al., 2021; Zhong et al., 2023). Deep learning models learn parameters directly from large volumes of data, which allows them to capture complex and nonlinear relationships between climate drivers and streamflow (Tripathy & Mishra, 2024; Zhao et al., 2024). Among these, Long Short-Term Memory (LSTM) networks have become particularly popular because of their ability to handle sequential data and represent long-term dependencies in hydrological processes (Li et al., 2024; Zhong et al., 2023). Comparative studies have demonstrated that LSTMs frequently outperform both process-based models and conventional machine learning methods in large-sample hydrological forecasting (e.g. Kratzert et al., 2019). In the context of climate change impact studies, LSTM models have also been shown to outperform traditional hydrological models and to provide more robust streamflow projections with reduced climate sensitivity (e.g. Martel et al., 2025), highlighting their growing role in next-generation hydrological modeling.

No one model structure works best in all situations. Hydrological model selection must be tailored to the application and is a function of study intent, data availability, spatial and temporal resolution, and computational ability (Ghonchepour et al., 2021; Marshall et al., 2005; Nesru, 2023). Furthermore, different models can yield varying results when calibrated against the exact same data, and this contributes to impact studies' structural uncertainty (Beven, 2006). Due to the trade-offs between model complexity, precision, interpretability, and resource utilization, recent studies advocate the employment of several models or model ensembles (Huang et al., 2017; Velázquez et al., 2013; Wan et al., 2021). The employment of a multi-model strategy enhances the representation of uncertainty and also enhances confidence in simulated hydrological responses to climate change.

1.7 Uncertainty of Climate Change Impacts

Uncertainty is defined as incomplete information and knowledge, or lack of consensus on what we know and can know (IPCC, 2014). In climate change impact assessment studies, the primary sources of uncertainty include: 1) natural variability, 2) scenario uncertainty, 3) climate model uncertainty, 4) downscaling method, and 5) impact model uncertainty (Hawkins & Sutton, 2009; Poulin et al., 2011). Among these, numerous studies over the past two decades have shown that climate model and hydrological model uncertainty are the dominant contributors to overall uncertainty in climate change impact assessments (Brient, 2020; Deser et al., 2012; Hawkins & Sutton, 2009; Poulin et al., 2011; Vetter et al., 2017). Consequently, reducing uncertainty from climate models and hydrological models remains the most effective way to improve the reliability of future projections (Lorenz et al., 2018). The objective is not to eliminate uncertainty, an impossible and undesirable outcome given the inherently unpredictable nature of future climate and natural variability. Instead, the aim is to *better quantify and manage* the uncertainty that arises from avoidable or artificial sources, such as structural deficiencies, or methodological choices. Reducing these *avoidable* uncertainties improves the interpretability, credibility, and usefulness of future projections for decision-making (Lorenz et al., 2018).

1.7.1 Natural variability

Because of the natural processes, the atmosphere-ocean system fluctuates around its mean, causing daily and decadal variations. This variability may be due to internal reasons such as the El-Niño Southern Oscillation (ENSO) or external natural forcings outside the climate system, such as natural changes in radiative forcing (Deser et al., 2012). Large ensembles make it possible to estimate the magnitude of natural variability and distinguish its contribution from other uncertainty sources. However, because such fluctuations are inherently unpredictable beyond a few years to decades, this component of uncertainty cannot be reduced, even as climate models improve or greenhouse gas concentration pathways become better constrained (Deser et al., 2012).

1.7.2 Scenario Uncertainty

Changes in greenhouse gas concentration in the future are one of the primary sources of uncertainty in climate modeling (Kundzewicz et al., 2018; Vetter et al., 2017). Scenario uncertainty causes uncertainty in future radiative forcing and, hence, climate. Changes in population, economic activity, and climate policies are the main drivers of the uncertainty in greenhouse gas emissions. The IPCC has recognized different scenarios to account for future changes in greenhouse gas concentration, including low, intermediate, and high forcing scenarios (IPCC, 2023). Climate change impact studies done with different scenarios yield different results. The contribution of scenario uncertainty is of low importance in short time scales (less than 30 years); however, it gains more significance for longer lead times, depending on the study region (Hawkins & Sutton, 2009). Although scenario uncertainty cannot be eliminated, recent analyses suggest that it can be meaningfully constrained by observational evidence, socioeconomic trends, and policy modeling, which rule out some extreme high- or low-emission trajectories (Moore et al., 2022).

1.7.3 Climate Model Uncertainty

Climate models, and models in general, try to quantify natural phenomena using physical equations and through parameterization. Due to our incomplete understanding of nature, climate modeling groups use different simplifications and parameterization schemes. Hence, based on the parameterization scheme and model structure, each model may produce different results for the same input data (Knutti & Sedláček, 2013; Kundzewicz et al., 2018). Previous studies have shown that climate model uncertainty dominates the other components, such as downscaling methods and hydrological models (e.g., Chen et al., 2011; Joseph et al., 2018).

It is essential for climate change impact assessment studies to adequately characterize the climate model uncertainty given its dominant role (Merrifield et al., 2023; Qian et al., 2016). Therefore, a common strategy has been to construct *envelopes* of GCMs that span the range of

projected climate responses, providing an indication of the plausible bounds of future climate change (Crawford et al., 2019; Haughton et al., 2014; Maher et al., 2021; Merrifield et al., 2023; Sanderson & Knutti, 2012; Semenov & Stratonovitch, 2010). However, the number of available simulations has risen rapidly across successive CMIP phases, from about 25 GCMs in CMIP3 to more than 100 in CMIP6, often with multiple ensemble members per model (Eyring et al., 2016; Taylor et al., 2012). For impact studies, analyzing the full ensemble has become practically unfeasible. This necessitates the selection of GCM subsets, raising a critical methodological question: how representative are these subsets of the overall uncertainty space (Chen et al., 2016; Di Virgilio et al., 2022; Merrifield et al., 2023; Ruane & McDermid, 2017; H.-M. Wang et al., 2018).

Poorly designed selections risk underestimating or mischaracterizing the diversity of plausible futures, ultimately leading to biased or misleading conclusions in climate change impact assessments (Herger et al., 2018; Ito et al., 2020; Lutz et al., 2016). To address this, a range of ensemble design strategies has emerged: performance-based approaches that prioritize model skill, envelope methods that aim to capture the full spread of responses, and more recent weighting and optimization frameworks that explicitly balance performance, diversity, and independence (see section 1.8). Each method carries distinct advantages and limitations, but collectively they underscore a central point: because GCMs are typically the dominant source of uncertainty, the design of GCM subsets is one of the most consequential decisions in climate change impact assessments (Merrifield et al., 2023; Vano et al., 2015; H.-M. Wang et al., 2018).

1.7.4 Impact Model Uncertainty

Similar to climate models, the structure and parametrization of the impact models, e.g. hydrologic models, affect the results of climate change impact studies. For instance, Jiang et al., (2007) studied the hydrological model structure uncertainty using six conceptual rainfall-runoff models. They found that models which simulate the historical climate conditions similarly behave differently under future projections of climate change. In addition, Ludwig et

al., (2009) compared two physically based and one conceptual hydrological model to show that the complexity degree of the hydrologic models can impact the results. Furthermore, Poulin et al., (2011) concluded that the uncertainty of different model structures is more significant than different parameterizations of the models, suggesting the use of models of various complexity in climate change impact studies.

1.8 Climate Model Selection in Impact Assessment Studies

As noted earlier in Section 1.7.3, impact assessment studies cannot rely on the entire set of available GCMs. Nonetheless, a robust characterization of uncertainty remains essential. With the rapid expansion of climate simulations across successive CMIP phases, analyzing the full ensemble has become impractical, if not impossible, for most applications. This reality compels researchers to select a subset of GCMs, which in turn raises a critical question: to what extent do these subsets adequately represent the broader uncertainty space?

Despite the importance of this issue, there is still no consensus in the literature on selecting proper GCMs for impact assessment studies. Until recently, researchers often evaluated climate change's effects using only one climate model or a small number of different GCM scenarios (e.g., Ott et al., 2013). GCM scenarios were chosen based on arbitrary means or the researcher's subjective choice without standard criteria for selecting climate change scenarios. For instance, most climate impact assessments have been carried out using high-resolution GCM scenarios or a GCM developed by the country in question. The idea that one GCM scenario chosen now would perfectly reflect the future conditions decades in the future is not convincing (Lee & Kim, 2017). Instead, using multiple models from various institutions is a widely acknowledged method to grasp an understanding of the uncertainty of the outputs (Tebaldi & Knutti, 2007). Recently more objective methods have been developed to select GCMs for impact studies, yet these methods still have limitations and are susceptible to the subjective choices made by the researcher (e.g., Cannon, 2015; Mendlik & Gobiet, 2016).

1.8.1 Envelope-based GCM selection approach

As mentioned earlier, impact studies primarily use a top-down approach in which the downscaled GCM outputs are used with impact models to study the climate change impacts. These modeling steps have been called the impact modeling chain, where the uncertainty increases as multiple model choices must be made at each step of the study (Wilby & Dessai, 2010). Previous studies have shown that climate model uncertainty typically dominates the other components, such as downscaling methods and impact models (e.g., Chen et al., 2011). Since failing to account for the full range of uncertainty may result in substantially biased impact studies (Chen et al., 2011), ideally, any selected subset of GCMs should be as unbiased as possible relative to the statistical characteristics of the full ensemble, while still covering an adequate range of uncertainty.

To address this matter, an approach that researchers have turned to is the envelope-based GCM selection approach, which focuses on the selection of GCMs that span the range of future changes in climate signals, and clustering algorithms are typically used for this purpose (Houle et al., 2012; Mendlik & Gobiet, 2016; Ruane & McDermid, 2017; Wilcke & Bärring, 2016). For instance, Raju and Kumar (2016) used the K-means clustering technique to select GCM ensembles from 36 climate projections over India. The first limitation of this approach is that the used clustering algorithms are designed to maximize the explained variance of an ensemble, and are thus biased toward high-density areas of the climate space (Cannon, 2015; Seo et al., 2019). In the CMIP5 and CMIP6, for example, some models have contributed several realizations, some of which differ only in resolution, and some models are not entirely independent, sharing model components and development history (Knutti et al., 2013). The issue of model interdependency is discussed further in section 1.8.2. The clustering algorithms would favor the realizations that fall closer to each other in the climate space. Furthermore, the results are not ordered (the smaller subset results would not necessarily appear in larger subsets), meaning that increasing the size of the subset would not necessarily improve the coverage of the uncertainty (H.-M. Wang et al., 2018).

Cannon (2015) proposed using the KKZ method (Katsavounidis et al., 1994) to select subsets that cover the range of overall changes. The KKZ algorithm is a deterministic subset-selection approach that identifies a small but representative set of points from a larger dataset (Cannon, 2015). Multiple studies, including Ross and Najjar (2019) and Wang et al. (2018), showed that the KKZ method outperforms other clustering approaches with higher coverage of ensemble range and smaller subset size. However, the outlier simulations are more likely to be selected by the KKZ method as it is designed to choose GCMs lying on the edge of the ensemble (H.-M. Wang et al., 2018). The outliers may be the GCMs that cannot capture the climatic patterns realistically, and the selection of poor representations may reduce the credibility of the subset. Therefore, while the performance of all GCMs in representing key physical processes should ideally be evaluated prior to any subset-selection exercise, this assessment becomes particularly crucial for potential outliers, given their higher likelihood of being selected by the KKZ method. Also, the outlier assessment may help with the preselection of GCMs and improve the results, which is discussed in more detail in section 1.8.2. The main disadvantage of the envelope-based approach is that the models' performance over the historical period to simulate climate is not considered, and all available climate models are assumed to be equiprobable (Lutz et al., 2016). The assumption of equiprobability will also be further discussed in section 1.8.2.

A noteworthy advantage of the KKZ method is that the results improve by including more models in the subset. A subset with n+1 GCMs will either perform better or equally well as the subset with n models. Considering this fact, it is recommended to consider as many GCMs as possible in the subset to increase the chance of having adequate coverage of the uncertainty in the impact world (Chen et al., 2016).

The objective of using multiple climate simulations is to account for different sources of uncertainty in the projections (Wilcke & Bärring, 2016). However, regardless of the underlying approach, GCM selection studies have focused on the uncertainty in the climate world. Yet, there is no guarantee that the selected subset will cover the same range of uncertainty in the impact world. Impact models, particularly hydrologic models, are nonlinear models, and small

changes in the inputs may result in significant changes in the outputs (Muzik, 2001). It is still impossible to know a priori whether the selected subsets would cover the same range of uncertainty in the climate world.

In this regard, Chen et al. (2016) studied the transferability of GCM uncertainty to the hydrological impacts using KKZ and K-means clustering methods. They showed that the optimally selected GCMs in the climate world might not be optimal in the hydrology world. In another study, Wang et al. (2018) studied the transferability of the uncertainty between climate and hydrology worlds. They used the K-means clustering and the KKZ method with 31 climate variables to select GCMs from a pool of climate simulations. The results indicated that using a subset of 10 GCMs can cover an acceptable range of uncertainty for the hydrological variables studied over two watersheds. However, utilizing multiple climate variables for GCM selection may result in redundant information that does not have meaning and may reduce the performance of the subset selection (Seo et al., 2019). For example, the atmospheric pressure field outputs of GCMs can be included (or used solely) in the selection variables to select GCMs for flood assessment studies. The relationship between atmospheric pressure fields and floods is not as direct as between extreme precipitation and floods. Even random numbers can be used as selection criteria that logically provide no meaning and have no physical explanation.

Although initially developed to be unsupervised, objective methods of GCM selection, the evidence indicates that identifying key climate indices correlated with the impact variable under study is key to subset selection, with the KKZ method performing poorly when unrelated indices are used (Seo et al., 2019). In other words, instead of using similar climate variables to select GCMs in flood and drought-related impact assessment studies, climate variables that relate specifically to floods and droughts must be identified separately and used as the selection criteria for the subset of climate models when using the KKZ method. Seo et al., (2019) concluded that before selecting the GCM subsets, careful identification of the most important climatic indicators for the studied impact (e.g., floods) and the region is necessary. A robust understanding of the region's climatic system and the key physical processes would guide the

choice of climatic variables (Seo et al., 2019). Further research is still required to evaluate the KKZ method and the necessary climatic indicators' efficiency in various climatic regions with different hydrological regimes.

1.8.2 Weighting Multimodel Ensembles

Another complication in GCM selection is that there is no consensus on how to combine the GCMs in multimodel ensembles (MME). Typically, the "model democracy" approach is used, in which all the models are considered plausible and equiprobable (e.g., Collins et al., 2013). "Model democracy" is based on the fact that all models have strengths and limitations and that their performance varies regionally but is globally similar (Chen et al., 2017), and a binary set of weights is assigned to the models (unit weight for selected GCMs and zero for others). It has been shown that the average of equally-weighted projections outperforms every single model in simulating the mean climatic patterns (Reichler & Kim, 2008). However, this method is arguably a suboptimal way of utilizing the available information (Knutti, Sedláček, et al., 2017).

An equally-weighted average implies that the simulations in the ensemble are independent; however, this might not always be the case (Sanderson et al., 2017). For instance, CMIP5 and CMIP6 contain multiple simulations from the same research group, which may only differ in resolution. Some simulations share parts of code or parameterization schemes and certainly share model-developing expertise (Eyring et al., 2019; Knutti, 2010). The number of independent models in an ensemble such as CMIP5 may be significantly lower than the actual number of models (Caldwell et al., 2014). The interdependent simulations, at worst, bias the results towards repeated simulations and, at best, add little information to the ensemble (Knutti, Sedláček, et al., 2017; H.-M. Wang et al., 2019). Ideally, a subset of climate simulation should account for the inter-dependency of simulations, however, identifying and accounting for interdependence is a difficult task and is not straightforward, and even the definition of dependence is a subjective matter depending on the problem at hand (Herger et al., 2018; Knutti, Abramowitz, et al., 2010).

Recently, efforts have been made to account for model dependence (e.g., Knutti et al., 2017); for instance, Herger et al. (2018) accounted for the interdependency of models by comparing the correlated biases, arguing that errors in an independent ensemble would be random and cancel each other out. Nonetheless, these approaches are practically sensitive to every component of the issue, including the chosen measure, variable, analysis period, and dataset (Eyring et al., 2019). Although a complicated task affiliated with multiple subjective and problem-specific choices regarding how to define dependence or evaluate it, accounting for model interdependency of GCMs increases the reliability of the ensemble (Herger et al., 2018).

Furthermore, the performance of GCMs in reproducing climatic patterns depends on the location and variable under study (Abramowitz et al., 2019), and in regions where some models are more reliable than others, model democracy might not be the ideal choice (Knutti et al., 2013; Lorenz et al., 2018). Consequently, another common approach for subset selection has been based on GCM performance over the historical period (Ahmadalipour et al., 2017; Ahmed et al., 2019; Evans et al., 2013; Hamed et al., 2022; Hassan et al., 2020; Salehie et al., 2023). To this end, GCMs are evaluated on representing climatic patterns of the recent past based on climate metrics defined by modelers (H.-M. Wang et al., 2018). For instance, Raju and Kumar (2015) ranked 11 GCMs based on their skill in simulating recent past precipitation and temperature patterns. However, the selected subset was completely different from the suggested subset by Raju and Kumar (2014) because of the different performance evaluation metrics.

Although there is no consensus in the literature on suitable performance metrics, the definition of performance measures is straightforward (e.g., the bias between simulated and observed precipitation); the challenge is how to translate them into a measure of model quality and then to model weight (Knutti, Sedláček, et al., 2017). A quality index assesses the model's suitability for a particular purpose by subjectively aggregating numerous indicators necessary for an application (Knutti, Abramowitz, et al., 2010). For example, multiple climatic variables (e.g., precipitation and temperature) may impact an environmental impact under study (e.g., mean

streamflow). Yet, the relative importance of these variables may not be equal and challenging to compute (H.-M. Wang et al., 2019).

There have been some attempts to select GCMs with unequal weights based on dependence and performance (e.g., Lorenz et al., 2018; Sanderson et al., 2017). For instance, Chen et al. (2017) studied the effect of unequal weight assignment to temperature and precipitation on the hydrology of a Canadian watershed. They concluded that weight assignment to GCMs may not significantly improve the ensemble's performance regarding streamflow. It is worth mentioning that the assigned weights are calculated in the climate world based on reproducing climate variables such as temperature or precipitation. These weights may not be optimal in the impact world since the relation is nonlinear and complex, yet it is impossible to know a priori which GCMs perform best in the impact world (Chen et al., 2017; H.-M. Wang et al., 2019). In addition, Wang et al. (2019) concluded that bias-corrected GCMs assigned with equal weights have the same capability as the weighted raw GCM data. However, the question remains as to whether weighing climate simulation would impact different hydrological variables in various climatic regions, which requires further research.

The quality indices must be defined based on the studied impact and region of the study, and general skill scores may not be adequate to evaluate the model's performance (Jagannathan et al., 2020). Physical understanding of the region’s climatic system and the dominating processes can help in the choice of climatic variables used as the performance metric. It is also possible to analyze multiple variables to determine which ones are the most crucial in the study region (Wenzel et al., 2016). Although expert judgment is inevitably involved in this step, transparency can be maintained by clearly documenting the criteria and physical rationale used. It is essential to combine numerous metrics to avoid overconfidence in the subset, but using a large set of metrics will reduce the impact of weighting (Borodina et al., 2017; Lorenz et al., 2018). The optimal number of the metrics is still not known, but including the most relevant ones must be the priority.

The GCMs that perform well over the observation period may not have realistic future projections. The climate and GCM performance are non-stationary, and the GCM performance may change for future projections (Hui et al., 2019). Nevertheless, the reliability of poor-quality models is under question (Lorenz et al., 2018). One of the approaches to address this issue is the use of emergent constraints, which are empirical relationships between observed behavior and model projections that have a physical explanation (Borodina et al., 2017; Eyring et al., 2019). With the use of observational data and by accounting for data uncertainty, the empirical relationship can become an emergent constraint that narrows down the range of plausible projections by picking a set of models compatible with the observations (Brient, 2020). For instance, based on evidence from paleoclimate, surface temperature, ocean heat content, and physical process models, Hausfather et al. (2022) concluded that some of the models in the CMIP6 archive are too sensitive to greenhouse gas emissions and that projected temperatures are too hot calling them "hot models," which are recommended to be excluded entirely from the impact assessment studies.

In this regard, Shiogama et al., (2021) presented a subset selection method in which the first step of model selection was to screen out the hot models. On the other hand, Palmer et al., (2022) showed that models with higher sensitivity better represent some of the key climatic processes over Europe. Although they were unable to provide a robust physical explanation for their findings, it is still noteworthy 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 another layer of complexity to climate model selection for regional impact studies. Removing GCMs that fail to adequately represent key physical processes in the past climatic patterns of the study region improves the ensemble's reliability (Klein & Hall, 2015). As Sanderson et al. (2017) noted, "a climate model is fit for the purpose if it can adequately represent the response of relevant physical processes in the required range of boundary conditions." However, further research is required to assess the impact of, and to further justify, dismissing outlier climate simulations in hydrological climate change impact studies.

It is worth mentioning that, compared to the non-weighted subset, where the spread is not a measure of uncertainty, the spread of the weighted multimodel mean can be regarded as a measure of uncertainty given everything that we know (Lorenz et al., 2018). Weighted approaches increase our confidence in impact assessment studies using multimodel ensembles. Nonetheless, it is recommended that modelers who use model-weighting approaches present the unweighted and the weighted results and analyze the sensitivity of the results under various weighting strategies, performance, or quality indices (Wootten et al., 2022). In addition, performance or quality indices must be justified based on physical reasoning, discussed, and analyzed to prevent overconfidence in results (Knutti, Sedláček, et al., 2017; Lorenz et al., 2018; Weigel et al., 2010).

A combination of envelope-based and performance-based approaches has been recommended to ensure that the subset covers multiple future projections and includes adequately performing models (Lutz et al., 2016; McSweeney et al., 2015). For instance, Lee & Kim (2017) used K-means clustering to group the GCMs based on their statistical characteristics and then selected a representative of each cluster based on the calculated skill score. Yet, the selected subset is sensitive to the selection sequence, meaning that if the performance-based selection is the second step, the selected GCMs may not be the best-performing ones overall (Lutz et al., 2016). Nonetheless, the selection sequence is not fixed, and the modelers must choose how to implement this approach based on the study's objective. As these approaches are, in some sense, similar to the past performance approach, they may intrinsically have the same weaknesses (H.-M. Wang et al., 2018).

1.9 Research Objectives

Given this context, the overall aim of this thesis is to comparatively examine and evaluate GCM selection and weighting approaches for hydrological impact studies. This study does not seek to introduce one “optimal” approach that can be applied in all regions and sectors. Rather, it aims to explain how different subsetting and weighting approaches affect future streamflow predictions in different North American catchments. By doing so, the thesis provides the trade-

offs embodied in different ensemble design choices and facilitates stronger, more transparent practices in regional-scale climate change impact assessments. The specific research objectives of this study are:

1. To quantify the impact of GCM sub-selection based on climate indices on uncertainty transferability to hydrological projections.

2. To investigate the influence of high-sensitivity climate models on the translation of climate signals into hydrological responses, and to understand the mechanisms through which these models shape future streamflow projections.

3. To compare the effects of different GCM weighting schemes, through the use of a pseudo-reality framework, on the uncertainty of future streamflow projections.

This study contributes to the growing literature emphasizing the need for transparent, reproducible, and context-dependent ensemble design in climate impact research. By bringing into focus the applied significance of model selection and weighting choices, it offers methodological findings and actionable recommendations, as much for hydrology-focused climate adaptation planning as for climate modeling and impact research more broadly.

This dissertation follows a manuscript-based format composed of three research articles that collectively investigate how climate model ensemble design influences hydrological impact assessments. Although each article is self-contained, they are intentionally ordered to build a coherent methodological progression that addresses the thesis objectives. Chapter 2 examines how uncertainty from the “climate-model world” propagates into the “hydrological-model world.” It evaluates whether reduced subsets of GCMs, selected using climate indices, can preserve the hydrological uncertainty captured by the full ensemble. This establishes the foundation for understanding uncertainty transferability and the challenges of ensemble reduction.

Chapter 3 investigates whether excluding high-ECS climate models alters projected streamflow responses across North America. By examining how a specific structural property of climate models shapes hydrological impacts, it provides insight into the consequences of model exclusion and the importance of regional process evaluation. Chapter 4 evaluates whether unequal weighting schemes improve streamflow projections relative to the commonly used model-democracy approach. Using a pseudo-reality framework, it tests the performance of multiple weighting methods and assesses how model credibility and dependence influence impact outcomes. This extends the previous chapters by exploring the role of model importance, not just model inclusion.

Together, these three articles form a unified investigation of climate-model ensemble design, from subsetting, to selective exclusion, to weighting, and how these methodological choices shape hydrological projections across more than 3,000 North American catchments. The final chapter synthesizes the cross-cutting insights, connects the findings back to the research objectives, highlights limitations, and outlines promising directions for future work.

CHAPTER 2

USING A REDUCED CLIMATE MODEL ENSEMBLE WHICH PRESERVES FUTURE STREAMFLOW UNCERTAINTY

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

[a] Hydrology, Climate and Climate Change Laboratory, École de Technologie Supérieure, 1100 Notre-Dame Street West, Montreal, Quebec, Canada H3C 1K3

Article submitted to *Climate Dynamics*, September 2025

Abstract

Climate change impact studies often use ensembles of climate projections from General Circulation Models (GCMs). These ensembles generate a distribution of future impacts which often dominates other uncertainty sources. While using all available GCMs was once considered ideal for representing uncertainty, the rapid growth in projections makes this approach unfeasible, necessitating representative subsets instead. Understanding how climate-model uncertainty from these subsets propagates through hydrological impact assessments is essential for robust adaptation planning. Using 3,540 catchments across Canada and the contiguous United States, we evaluate whether the representativity of GCM subset uncertainty is preserved once climate projections are transferred to hydrological projections. We drive three lumped hydrological models (GR4J+CemaNeige, HMETS, HSAMI) with 20 CMIP5 GCMs under RCP8.5, bias-corrected with TSQM and MBCn, and decompose variance in projected changes for mean, high, and low flows. We find that GCMs dominate uncertainty for mean and high flows, whereas hydrological model structure dominates low-flows. We then test GCM sub-selection using K-means and the deterministic KKZ algorithm within multivariate spaces defined by climate indices. KKZ consistently preserves ensemble spread better than K-means. Crucially, index choice matters: small, physically meaningful pairs tailored to the target hydrologic metric (e.g., PRCPTOT with ΔT for mean flows; Rx1day with wet-day frequency for high flows) outperform larger index sets. Across most catchments, five well-chosen GCMs reproduce most of the full-ensemble spread for mean and low flows, while

high-flows require larger subsets. Results provide practical guidance for designing compact, representative GCM ensembles that retain key uncertainties in hydrological applications.

2.1 Introduction

Understanding future hydrological regimes in a warmer climate is essential for developing adaptation strategies to meet growing demands on food production and increased risks from water-related hazards due to climate change and population growth (Lavell et al., 2012; Mahadevan et al., 2024; Smirnov et al., 2016). To this end, climate change impact studies targeting the water cycle are now routinely performed to guide decision makers into choosing the best possible adaptation measures. These impact studies typically use ensembles of projections from General Circulation Models (GCMs) and hydrological models to project future hydrological regimes (e.g. Feng & Beighley, 2019). However, each component within the modeling chain introduces uncertainties that must be carefully studied and accounted for (Ashraf Vaghefi et al., 2019; Senatore et al., 2022; H. Wang et al., 2020). A proper characterization of uncertainty related to the various components of the hydroclimatic modeling chain is critical for impact studies (Clark et al., 2016; Giuntoli et al., 2018). Of all uncertainties present in the hydroclimatic modeling chain, the uncertainty related to the choice of climate models is often dominant. However, whether subsets chosen in the climate domain remain representative after being processed through bias correction and hydrological models is still unclear. This raises the broader problem of uncertainty "transferability" between climate and impact domains.

Climate model projection uncertainty arises primarily from three factors: internal variability, emission scenarios, and model uncertainty (Deser et al., 2012; Tebaldi & Knutti, 2007). Model uncertainty arises because different models employ distinct physical and numerical formulations, resulting in varied responses to identical external forcing. Scenario uncertainty stems from limited knowledge about external factors influencing the climate system, such as trends in greenhouse gas emissions, land-use changes, and stratospheric ozone concentrations. Internal variability is due to the nonlinear dynamic processes within the atmosphere, ocean,

and the coupled ocean-atmosphere system, reflecting the climate system's inherent fluctuations independent of external forces (Deser et al., 2012). The internal variability related uncertainty is irreducible, but can be evaluated by running a climate model multiple times with slightly different initial conditions (Deser et al., 2012; Leduc et al., 2019).

Impact assessment studies typically involve multiple steps in a modeling chain starting with a greenhouse gases emission scenario and ending with an impact model such as a hydrological model (Chen et al., 2011; H. Wang et al., 2020). Each of these steps adds uncertainty along the way. Several studies have shown that GCM projections are, in most cases, the main source of uncertainty in climate change impact assessments (Chen et al., 2011; Giuntoli et al., 2018; Hodson et al., 2013; H. Wang et al., 2020). Ideally, capturing the full range of uncertainty would involve using all available GCMs, allowing for a comprehensive spectrum of potential future scenarios generated by multiple models. However, as the number of GCMs has increased from 25 simulations in CMIP3 to 61 in CMIP5 (Taylor et al., 2012) and over 100 in CMIP6 (Eyring et al., 2016), incorporating them all in impact studies has become impractical. The plethora of existing model choices poses a challenge for researchers and decision-makers in selecting the most appropriate models for their assessments.

An ideal subset of GCMs should both accurately reconstruct historical climate patterns and represent a range of potential future scenarios (Vano et al., 2015). Accordingly, one approach to selecting GCMs has been to evaluate their historical performance, considering models that closely replicate past observations as optimal (Gleckler et al., 2008; Palmer et al., 2022; Parding et al., 2020; Perkins et al., 2007; Rupp et al., 2013). However, even high-performing models during the reference period do not necessarily guarantee the most reliable future projections. Many researchers have therefore explored strategies to select climate change scenarios that minimize the number of required scenarios while effectively capturing a broad spectrum of potential inter-model variability (Cannon, 2015; Mendlik & Gobiet, 2016). Other researchers have tried to improve their subset by first eliminating GCMs with the weakest performance over the reference period (Dubrovsky et al., 2015; George & Athira, 2022; Lutz et al., 2016; McSweeney et al., 2015; Prein et al., 2019; Ruane & McDermid, 2017; Shiogama

et al., 2021). While excluding GCMs that fail to capture key processes in the study region may enhance accuracy, this approach is still limited in guaranteeing future reliability (Palmer et al., 2022). The question of whether selection should include all available GCM simulations, even those with subpar historical performance, remains under debate (Rahimpour Asenjan et al., 2023).

To capture inter-model variability, earlier studies predominantly employed clustering-based techniques, such as hierarchical (e.g. Mendlik & Gobiet, 2016; Wilcke & Bärring, 2016) and k-means clustering (e.g. Casajus et al., 2016; Houle et al., 2012). However, the clustering approaches often select scenarios that represent only the central trend rather than the full spectrum of an ensemble's variability, and they do not allow for scenarios to be arranged by priority (Cannon, 2015). To address these limitations, Cannon (2015) proposed the Katsavounidis–Kuo–Zhang (KKZ) algorithm (Katsavounidis et al., 1994), which selects a subset of GCMs that more effectively captures a wide range of variability. The KKZ algorithm operates recursively, selecting members that thoroughly span multivariate space, thereby offering a more effective method for preserving the full inter-model variability than traditional clustering approaches. Previous studies demonstrated that the KKZ algorithm better retains the comprehensive variability of the ensemble (Cannon, 2015; Chen et al., 2016; Golian & Murphy, 2021; Ross & Najjar, 2019).

Previous studies have primarily focused on the “climate world,” conducting analyses within the domain of climate models (e.g. Sung et al., 2019). While these methods effectively select climate simulations that capture the uncertainty inherent in climate models, they do not guarantee that the selected subset will remain optimal when applied to the impact domain. This limitation becomes particularly evident after processes such as downscaling, bias correction, or passing through the non-linear filters of impact models. For instance, hydrological models exhibit highly non-linear responses to even minor variations in temperature and precipitation. A limited number of previous studies have explored the transferability of GCM uncertainty to hydrological impacts using various methods. For example, Chen et al. (2016) assessed transferability over a Canadian watershed using two climate variables and found limited

transferability of the uncertainty. Wang et al. (2018) extended this work by incorporating multiple climate variables, concluding that a selection of ten GCMs could adequately represent uncertainty in both climate and hydrology domains. Seo et al., (2019) and Seo and Kim, (2018) extended this line of research by employing the KKZ algorithm with climate extreme indices and emphasized that the choice of indices should reflect the hydrological extremes being projected. Nonetheless, their analysis was restricted to a limited number of basins and did not address how the spatial scale of climate indices may influence hydrological outcomes.

Building on this context, the aim of our study is to investigate the transferability of uncertainty from the climate world to the hydrologic domain when GCMs are selected based on climatic variables. As Seo et al., (2019) and Seo and Kim, (2018) demonstrated, identifying key climatic indices is a crucial step before selecting a representative subset of GCMs. This approach minimizes the need to account for a broad range of climate indicators, allowing for a tailored selection process based on the specific dependency of each hydrologic variable. In this study, various combinations of climate indices are compared to identify the most effective set of climate variables for GCM selection in North American catchments.

This paper evaluates climate model sampling methods for preserving uncertainty across diverse climatic and hydrologic regimes. Specifically, it investigates whether including multiple indices closely related to the hydrologic variable under study improves uncertainty preservation and examines how narrowing the ensemble of GCMs impacts this preservation. Additionally, it assesses the performance of extreme indices compared to classic indices in North American catchments. By addressing these objectives, our study contributes to advancing methodological foundations and provides guidance for designing representative, yet computationally feasible, GCM ensembles for hydrological applications.

2.2 Methods

2.2.1 Study area and data

The meteorological and streamflow data for 3540 catchments spanning Canada and the contiguous United States were used in this study. Data was extracted from the NAC^{2}H dataset (Arsenault, Brissette, Chen, et al., 2020). The NAC^{2}H dataset integrates 2,842 U.S. catchments from the National Hydrography Dataset (U.S. Geological Survey, 2019) with 698 additional catchments from both the CANOPEX database (Arsenault et al., 2016) and the U.S. Geological Survey's National Water Information System (NWIS) (U.S. Geological Survey, 2016). The meteorological data are obtained from the Livneh gridded database for U.S. catchments (Livneh et al., 2015) and CANOPEX for Canadian catchments. The CANOPEX dataset includes 10 km resolution gridded climate data from Natural Resources Canada (Hutchinson et al., 2009), while the Livneh dataset provides 6 km resolution gridded meteorological data for the United States. Daily precipitation (mm/day) and minimum and maximum temperatures (°C) are included in the dataset, which are used as input for hydrological model calibration and as reference data for climate model bias correction.

The reference period selected for this study is 1971–2000, while the future period for analysis is 2070–2099. Catchments with drainage areas smaller than 300 km² were excluded to avoid challenges associated with daily-scale hydrological modeling. Figure 2.1 illustrates the spatial distribution of catchments across multiple regions, capturing multiple hydrological and climatic conditions.

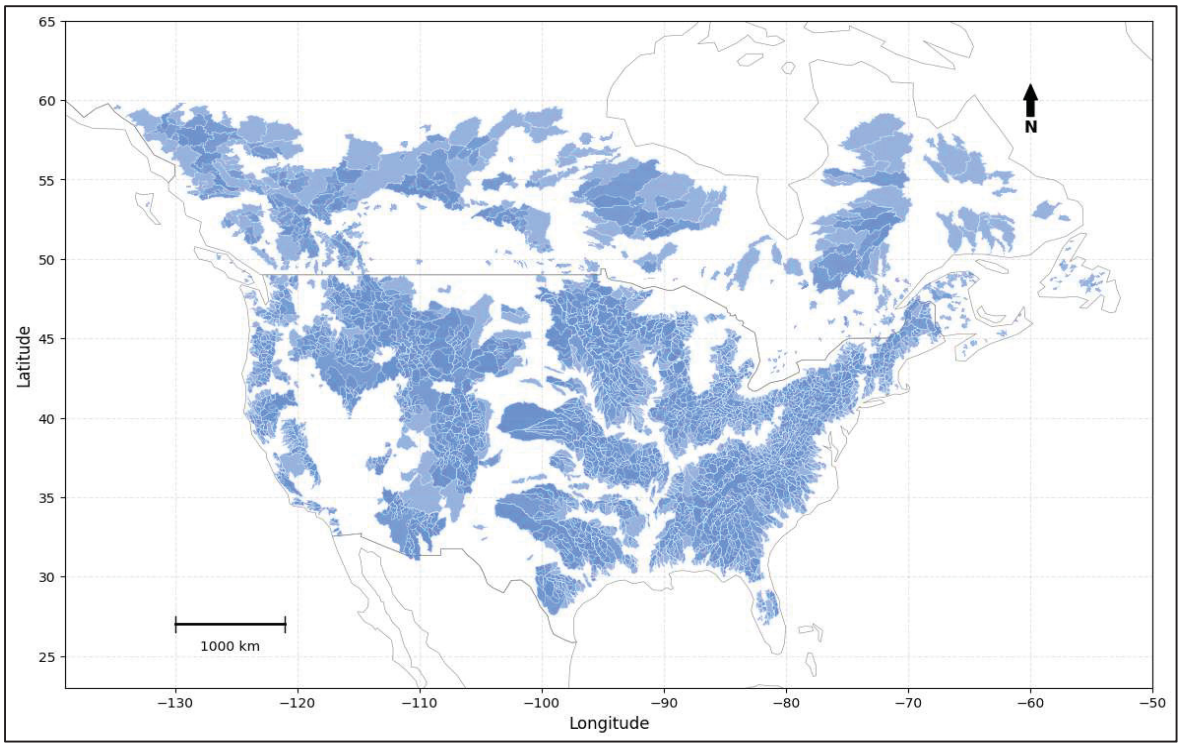

Figure 2.1 Geographic distribution of the 3,540 catchments in the dataset. The figure includes nested basins, with smaller catchments overlaid atop larger ones

2.2.2 Experimental Setup

Following a top-down hydroclimatic modeling approach, similar to Rahimpour Asenjan et al., (2023) and Arsenault et al. (2020), the RCP 8.5 scenario was selected as it represents a high-emission scenario with substantial warming potential, allowing for the assessment of extreme hydrological changes under a worst-case climate change trajectory. While this scenario has been increasingly considered as overly pessimistic (e.g. Hausfather & Peters, 2020), it has the advantage of limiting the impact of internal variability (irreducible uncertainty) on the interpretation of future impacts. Twenty (20) GCMs from the CMIP5 archive were used, which provide a diverse ensemble for hydrological simulations and impact analysis. A complete list of the selected GCMs is provided in Table 2.1.

Table 2.1 The 20 GCMs employed in this study

Climate Models	Resolution (Lon × Lat)	Modeling center
ACCESS1-0	1.875° × 1.25°	CSIRO
ACCESS1-3	1.875° × 1.25°	CSIRO
BCC-CSM1-1	2.8° × 2.8°	BCC
BCC-CSM1-1-m	1.125° × 1.125°	BCC
BNU-ESM	2.8° × 2.8°	GCESS
CanESM2	2.8° × 2.8°	CCCma
CMCC-CMS	1.875° × 1.875°	CMCC
CSIRO-Mk3-6-0	10.8° × 1.8°	CSIRO
FGOALS-g2	1.875° × 1.25°	CESS
GFDL-ESM 2G	2.5° × 2.0°	NOAA-GFDL
GFDL-ESM 2M	2.5° × 2.0°	NOAA-GFDL
GISS-E2-R	2.5° × 2.0°	NOAA-GISS
Inmcm4	2.0° × 1.5°	INM
IPSL-CM5A-LR	3.75° × 1.9°	IPSL
IPSL-CM5A-MR	2.5° × 1.25°	IPSL
IPSL-CM5B-LR	3.75° × 1.9°	IPSL
MIROC5	1.4° × 1.4°	MIROC
MIROC-ESM	2.8° × 2.8°	MIROC
MIROC-ESM-CHEM	2.8° × 2.8°	MIROC
MRI-CGCM3	1.1° × 1.1°	MRI

To correct the systematic biases in raw GCM simulations the Two-Stage Quantile Mapping (TSQM) and Multivariate Bias Correction (MBCn) (Cannon, 2018) methods were used. TSQM is a two-step quantile mapping approach (Guo et al., 2019) developed to enhance bias

correction by preserving the relationships between climate variables. TSQM first corrects the marginal distributions of precipitation and temperature using a univariate quantile-mapping approach, and then restores their dependence structure through a distribution-free shuffling step. MBCn, on the other hand, performs a fully multivariate correction by rotating the variables into an independent coordinate system, applying quantile mapping in that space, and then rotating them back, thereby adjusting both marginal distributions and the full multivariate dependence structure simultaneously. In addition to correcting biases in temperature and precipitation distributions, the two multivariate bias correction methods also maintain the inter-variable relationships between temperature and precipitation (Arsenault, Brissette, Chen, et al., 2020; H. Wang et al., 2020).

In this study, three lumped hydrological models namely GR4J, HMETS and HSAMI were employed. GR4J (Génie Rural à 4 Paramètres Journalier) is a four-parameter conceptual model (Perrin et al., 2003), which is paired with the CemaNeige snow module (Oudin et al., 2005) to account for snow processes and improve performance in snow-dominated catchments, since it lacks a built-in snow routine. On the other hand, the 21-parameter HMETS (Hydrological Model of the École de Technologie Supérieure) model was created especially for cold climates (Martel et al., 2017). Ten parameters are dedicated to snow accumulation and melt processes, and PET is calculated internally using the Oudin formulation. HSAMI is a 23-parameter model that has a similar structure to HMETS but employs a different snow routine and an empirical PET formulation (Poulin et al., 2011). The selected hydrological models were chosen to represent different conceptual structures, allowing for an assessment of structural uncertainty in hydrological simulations (Poulin et al., 2011).

In this study, we did not perform new hydrological model calibrations; instead, we relied on the pre-calibrated simulations provided in the NAC^{2}H dataset. In NAC^{2}H, the hydrological models were calibrated using the CMAES optimization algorithm (Hansen et al., 2003), with the Kling-Gupta Efficiency (KGE) metric as the objective function (Arsenault et al., 2014; Gupta et al., 2009). Each calibration was repeated 15,000 times. The meteorological data of the reference period (1971-2000) was used for calibration, with the first two years allocated

for model warm-up and the remaining 28 years for calibration. Many catchments achieved KGE values above 0.7, demonstrating the effectiveness of hydrological models in simulating streamflow. However, modeling remains challenging in regions such as the Midwestern U.S. and the Canadian Prairies, where lumped models often struggle to capture hydrological processes accurately (Newman et al., 2015).

2.2.3 Uncertainty Decomposition

The study consists of 20 GCMs, 3 hydrological models (HMs), and 2 bias correction (BC) methods, resulting in 120 total combinations. To attribute variance in projected streamflow changes to these three sources of uncertainty, an Analysis of Variance (ANOVA) was applied. ANOVA partitions the overall variance of a projected hydrological variable into components attributable to each factor, thereby quantifying the distinct contribution of GCMs, HMs, and BC methods (Giuntoli, Vidal, et al., 2015; Meresa et al., 2022; S. Zhang et al., 2024). Because our models are deterministic, we obtain only one simulation for each GCM–HM–BC combination. In classical ANOVA, interaction effects can only be estimated if multiple independent values (replicates) exist for each treatment, allowing variability due to interactions to be separated from random noise. Since no such replicates exist in our dataset, interaction effects cannot be distinguished and are absorbed into the residual error. The ANOVA is therefore simplified to a first-order variance decomposition with one case per treatment and no interaction terms ($\alpha\beta_{ij} = 0$) (Giuntoli, Vidal, et al., 2015). The overall variance in hydrological responses is divided between the separate components using equation (2.1)

$$Y_{ijk} = \mu + G_i + H_j + B_k + \varepsilon_{ijk} \tag{2.1}$$

Where Y_{ijk} is the simulated streamflow response under the i-th GCM, j-th hydrological model, and k-th bias correction method, μ represents the ensemble mean across all model simulations, G_i captures the effect of the i-th GCM, H_j accounts for the effect of the j-th hydrological model, B_k represents the effect of the k-th bias correction method, ε_{ijk} is the residual error. The total variance is then decomposed to each factor using the sum of squares method.

$$SS_X = \frac{\sum \left(X - \overline{X}\right)^2}{\sum \left(Y - \overline{Y}\right)^2} \quad (2.2)$$

where SS_X represents the sum of squares for factor X (e.g., climate models, hydrological models, or bias correction methods), and Y denotes the total response variable representing the simulated streamflow projections. Here, total variance is used as a measure of overall uncertainty, aligning with methodologies applied in previous studies (Meresa et al., 2022; Sansom et al., 2013; H.-M. Wang et al., 2018).

2.2.4 GCM subset selection

GCM selection is typically based on a set of *n* climate variables (e.g. annual mean precipitation and temperature in its simplest form with *n*=2). These variables define an n-dimensional climate space**,** where each GCM is represented as a point according to its projected changes. In order to cover the uncertainty of a large ensemble of GCMs, subsets of *N* GCMs ($2 \leq N < 20$) are chosen using either K-means clustering or the KKZ method (see details below). The selection process relies on selected climate variables (see next section) that are used to assess similarity (or differences) between GCMs. Subset selection is therefore dependent on both the approach (K-means vs KZZ) as well as on the climate variables (and scaling of said variables) used to measure similarity between the various GCMs. In order to assess the ability of those subsets at representing explained variance, all possible combinations of *n* GCMs will also be computed. This exhaustive benchmark allows us to determine whether structured selection methods preserve ensemble variability more effectively than random sub-selection. Figure 2.2 schematizes the approach used for testing GCM subset selection approaches.

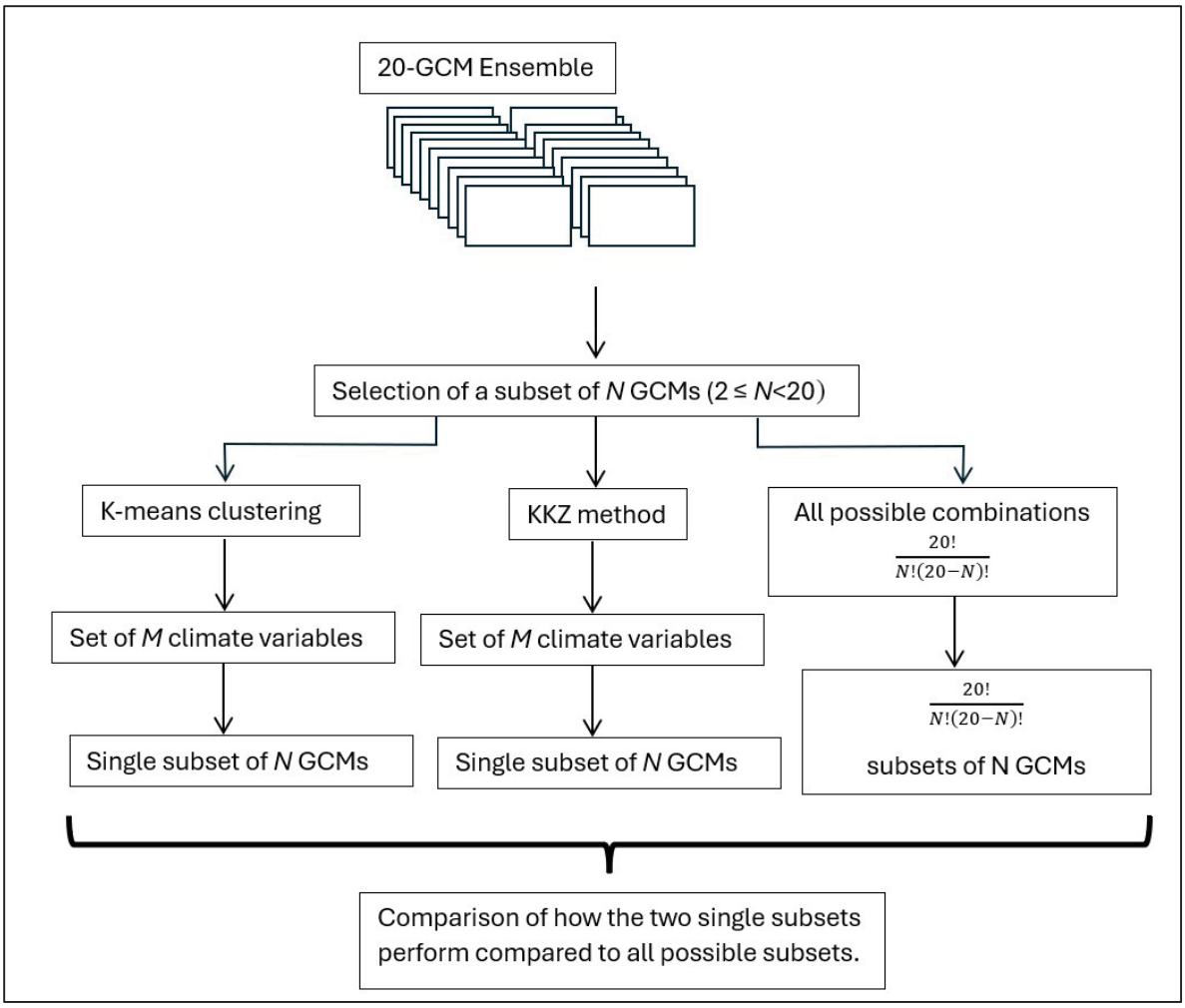

Figure 2.2 Methodological framework for testing GCM subset selection

2.2.5 Climate variables

Subsets of climate models are typically selected using future climate signals, often using a reduced 2-D space consisting of mean temperature change (ΔT) and relative precipitation change (ΔP/P), however, in practice any number of climate variables can be used. To characterize climate change signals relevant to hydrological impacts, 25 climate indices were computed, primarily based on those defined by the Expert Team on Climate Change Detection and Indices (ETCCDI). These indices capture extremes in temperature and precipitation, and were further complemented by climate variables used in Wang et al. (2018) and Seo et al. (2018, 2019) to ensure comprehensive coverage of relevant climate characteristics.

Table 2.2 List of climate variables. All indices represent the change between historical and future periods. Indices marked with (%) are expressed as ratios (future/historic)

Index	Description	Change
FD	No. of frost days (Tmin < 0°C)	
SU	No. of summer days (Tmax > 25°C)	
ID	No. of icing days (Tmax < 0°C)	
TR	No. of tropical nights (Tmin > 20°C)	
TNn	Annual minimum of daily minimum temperature	
TNx	Annual maximum of daily minimum temperature	
TXx	Annual maximum of daily maximum temperature	
TXn	Annual minimum of daily maximum temperature	
DTR	Change in diurnal temperature range	
ΔT	Change in annual mean temperature	
WSDI	Warm spell duration index	
CSDI	Cold spell duration index	
PRCPTOT	Total annual precipitation	
SDII	Simple precipitation intensity index	%
Rx1day	Annual maximum 1-day precipitation	%
Rx3day	Annual maximum consecutive 3-day precipitation	%
Rx5day	Annual maximum consecutive 5-day precipitation	%
R10mm	No. of wet days with ≥10mm precipitation	%
R20mm	No. of wet days with ≥20mm precipitation	%
R1mm	No. of wet days with ≥1mm precipitation	%
CDD	Maximum number of consecutive dry days	
CWD	Maximum number of consecutive wet days	
Rn30day	Annual min consecutive 30-day precipitation	
R95pTOT	Precipitation from days >95th percentile	%
R99pTOT	Precipitation from days >99th percentile	%

Each index was computed by comparing values from the reference period to those of the future projection period, and subsequently normalized to make the indices directly comparable. Precipitation-related indices were expressed as relative changes, while temperature-based and duration-related metrics were represented as absolute changes. This set of standardized climate indicators formed the basis for GCM selection using clustering and sampling methods such as KKZ and K-means. Table 2.2 provides a summary of the climate variables used in this study.

2.2.6 K-Means clustering

K-means clustering is a widely used unsupervised learning technique that partitions the set of climate simulations into a specified number of clusters, aiming to minimize the within-cluster sum of squared errors (SSE) (Hartigan & Wong, 1979). Each cluster is identified by a centroid, which is the average position of all simulations allocated to that cluster. The SSE is calculated as the Euclidean distance between each simulation and its centroid. Simulations closest to these centroids are selected to form a representative subset, a strategy commonly applied in climate modeling studies to reduce ensemble size while preserving variability (Logan et al., 2011; Cannon, 2015; Houle et al., 2012).

A key limitation of K-means clustering is its sensitivity to initial centroid placement, which can strongly influence the final clustering results. To mitigate this issue, the clustering was run 10,000 times with various initializations, and only the solution that produced the lowest SSE was retained. However, a major drawback of this approach is that subset selection is not hierarchical, meaning the simulations chosen in a smaller subset may not necessarily be included in a larger subset. Because of this lack of ordering, it is less flexible and less appropriate for applications in which users must dynamically modify subset sizes in accordance with certain specifications.

2.2.7 KKZ method

Originally developed for initializing centroids in k-means clustering, the Katsavounidis-Kuo-Zhang (KKZ) method (Katsavounidis et al., 1994) is a deterministic algorithm later adapted by Cannon (2015) for selecting representative climate model simulations. Unlike stochastic clustering techniques, KKZ is a deterministic procedure that systematically identifies a subset of models that optimally capture the variability within an ensemble. This approach is particularly valuable in climate studies, where it is essential to represent a wide range of climate projections. By ensuring that the selected models are evenly distributed across the multivariate space, KKZ provides a more comprehensive representation of future climate scenarios.

Unlike random sampling or conventional clustering, which may favor models concentrated in high-density regions, KKZ prioritizes models that span the full range of climate variability. This ensures that the selected ensemble reflects the entire spectrum of climate conditions, making it a reliable method for scenario selection in climate impact assessments. The KKZ algorithm follows a structured, step-by-step approach to ensure optimal selection of models that best represent variability within an ensemble:

1. Select the first model: Identify the model closest to the ensemble centroid, determined by the lowest SSE across all variables.

2. Select the second model: Choose the model farthest from the first selection based on Euclidean distance in the multivariate space.

3. For subsequent selections (starting from the third model onward):

    1. Compute the Euclidean distance between each remaining model and all previously selected models.
    2. Retain only the minimum distance for each remaining model, ensuring it is evaluated based on its closest selected counterpart.

3. Select the model with the maximum minimum distance, ensuring that each new selection maximizes diversity within the ensemble.

4. Repeat Step 3 until the desired number of models has been selected.

By employing this method, KKZ ensures that the selected models effectively capture the full range of variability within the dataset. However, its deterministic nature makes it susceptible to outliers, as extreme values may be preferentially chosen. Despite this limitation, KKZ remains a valuable tool for selecting climate scenarios in a systematic, reproducible, and unbiased manner.

2.3 Results

Figure 2.3 illustrates the contribution of different sources to future mean flow uncertainty using the full ensemble of 20 GCMs. Uncertainty is categorized into HMs, BCs, and GCMs, with colors representing the percentage contribution to total uncertainty. For mean flow, GCMs are identified as the dominant source of uncertainty across most of North America, explaining over 80% of the variance in many regions. In contrast, HMs generally contribute less to mean flow uncertainty, except in certain central regions where they struggle with modeling accuracy, resulting in a higher contribution to the overall uncertainty. BCs, while contributing modestly, are less influential compared to GCMs and HMs in shaping the uncertainty associated with mean flows. These patterns underscore the critical role of GCMs in driving uncertainty in future projections of mean flows across North America.

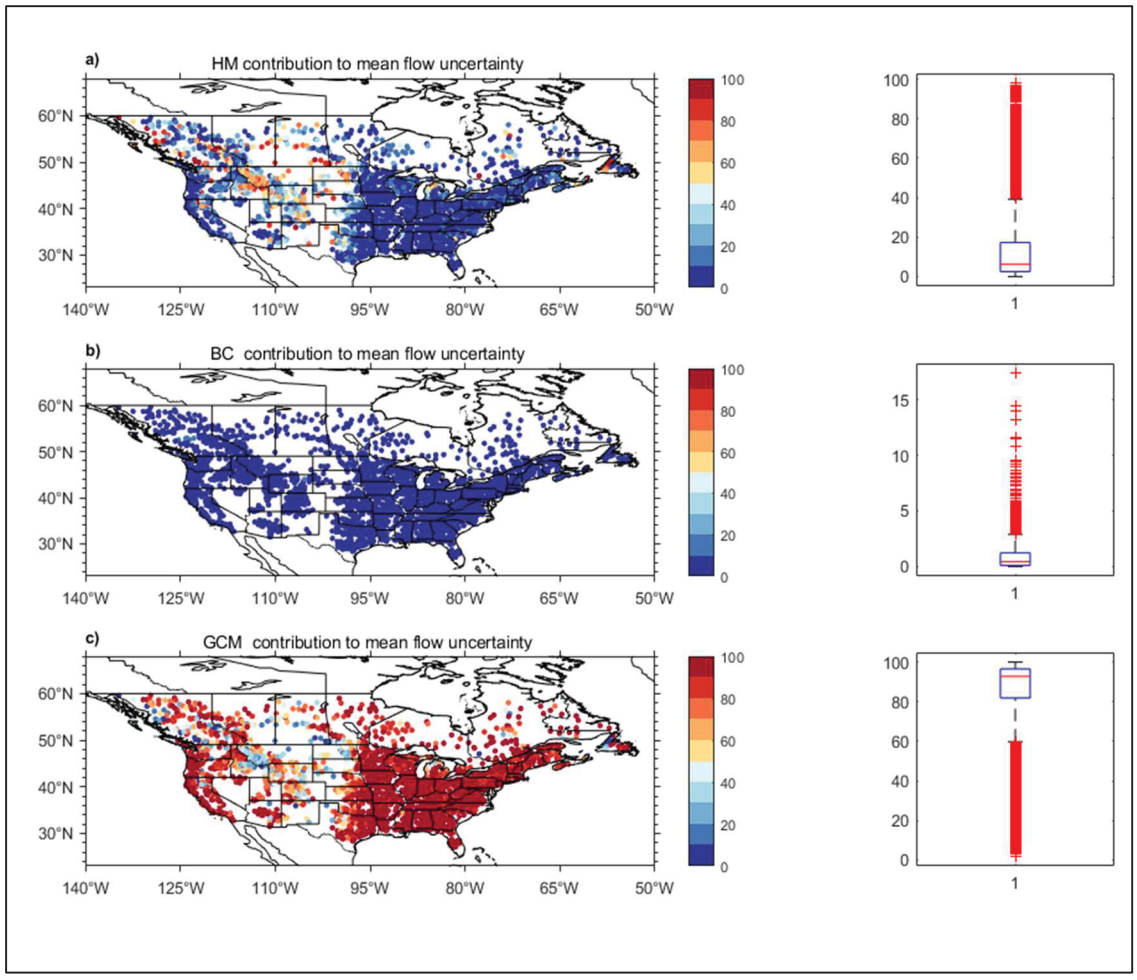

Figure 2.3 Contribution to future mean flow uncertainty, separated by (a) Hydrological Model (HM), (b) Bias Correction (BC), and (c) General Circulation Model (GCM). Colors represent the percentage contribution to total uncertainty

In the same format as Figure 2.3, Figure 2.4 presents uncertainty partitioning for low-flows. HMs are identified as the largest source of uncertainty for low flows, with a median contribution of 82%. This high contribution reflects the challenges HMs face in accurately simulating low flows, which amplifies their impact on overall uncertainty. In comparison, BC and GCM contributions to low flow uncertainty are significantly lower, with little regional variation observed.

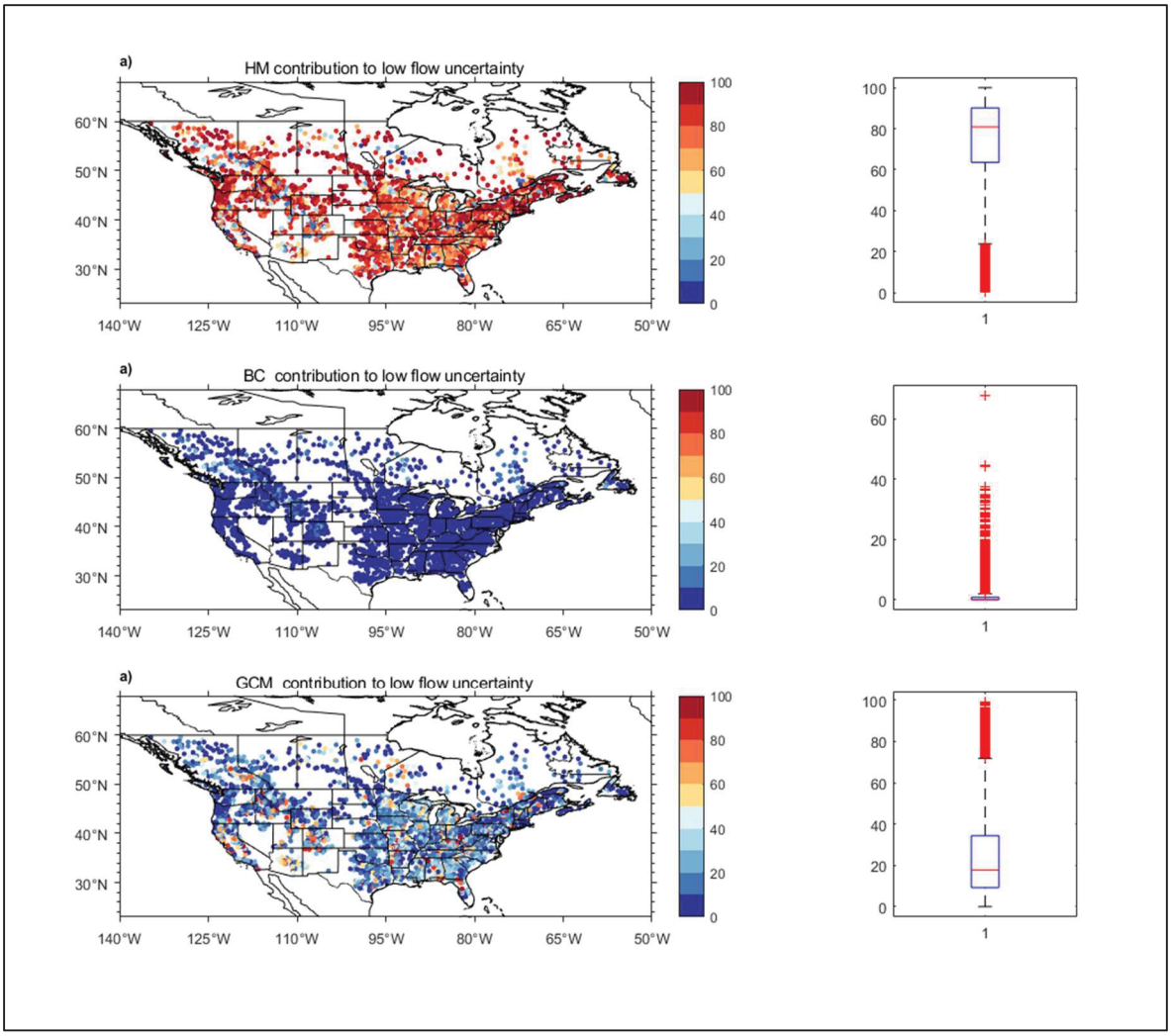

Figure 2.4 Contribution to future low flow uncertainty, separated by (a) Hydrological Model (HM), (b) Bias Correction (BC), and (c) General Circulation Model (GCM). Colors represent the percentage contribution to total uncertainty

Similarly, Figure 2.5 illustrates the contributions to uncertainty for future high flows. GCMs emerge as the dominant source of uncertainty for high flows, with a median contribution of 60%. HMs contribute 40% to high flow uncertainty, performing better in modeling high flows compared to low flows. However, HMs exhibit a significant regional pattern, with a notable contribution to uncertainty in northern and western regions. These regions are characterized by the presence of snow. In contrast, BC methods contribute minimally to high flow uncertainty, having a less significant impact compared to GCMs and HMs.

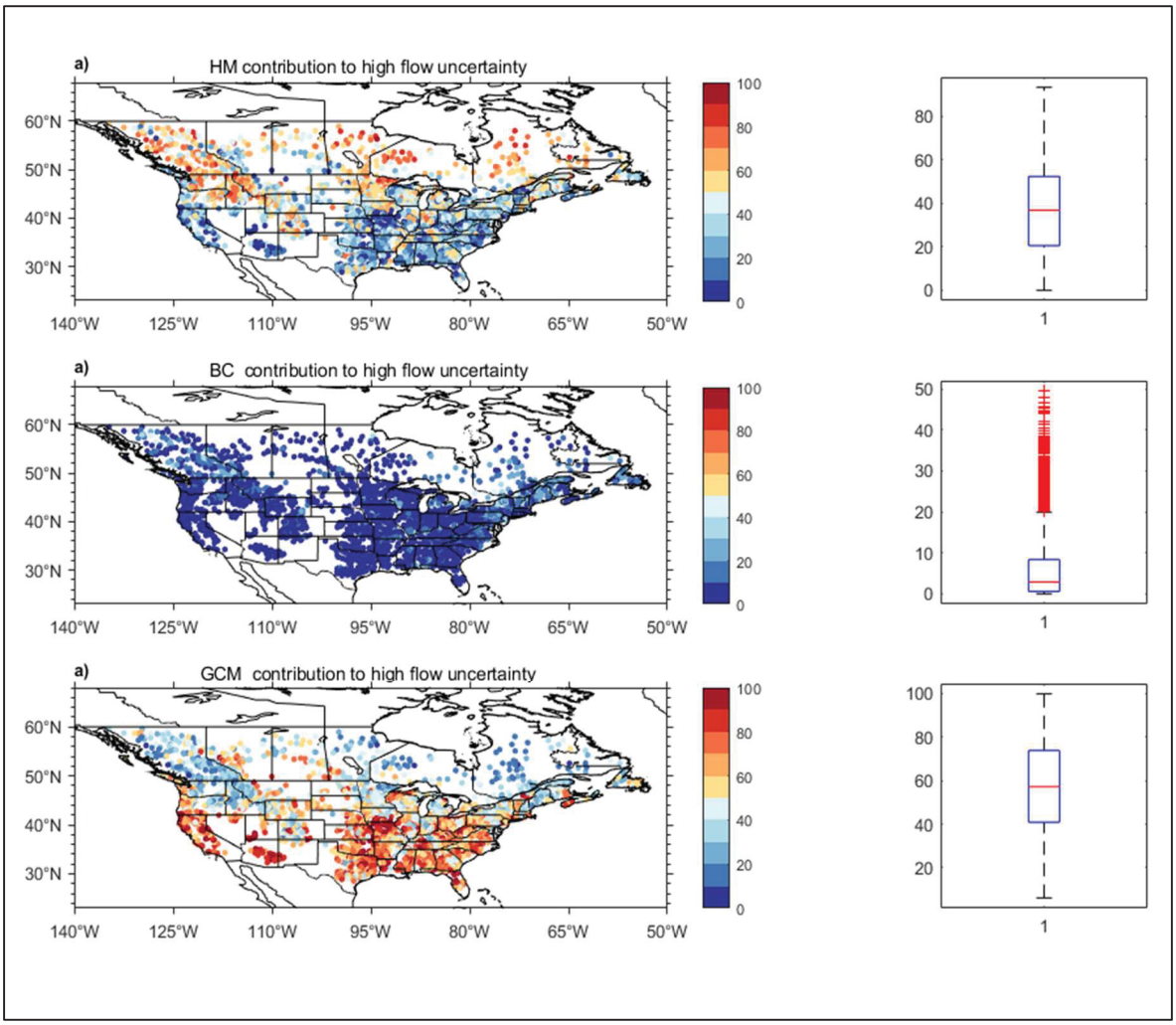

Figure 2.5 Contribution to future high flow uncertainty, separated by (a) Hydrological Model (HM), (b) Bias Correction (BC), and (c) General Circulation Model (GCM). Colors represent the percentage contribution to total uncertainty

The main objective of the paper is to investigate if a smaller sample of General Circulation Models (GCMs) can preserve the variance as the full GCM ensemble. The impact of sampling GCMs on their contribution to overall streamflow variance for mean, high, and low flows is shown in Figure 2.6. The graph depicts the distribution of GCM contributions to uncertainty, obtained from all possible combinations of 5, 10, and 15 GCMs, for one representative catchment, though the same analysis was performed for all catchments, with additional examples provided in the Supplementary Material. However, the key features of the graph are

consistent across most catchments, even though the total GCM contribution varies depending on catchment, as seen in Figures 2.3, 2.4 and 2.5. The red line on the right represents the contribution to variance of the original 20-member ensemble.

It is apparent that when using a random sample of 5 GCMs, the likelihood of over or underestimating future uncertainty is highest, with values ranging from 6% and 94% for mean flow. A low value would arise from selecting 5 GCMs that give similar future mean flow projections, whereas a large value would result from the selection of 5 highly diverse GCMs in terms of future mean flow projections. Increasing the number of GCMs from 5 to 10 and 15 significantly decreases the potential for over or underestimation of the original ensemble variance. However, significant deviations can still arise depending on the choice of GCM within the reduced sample. Nevertheless, selecting a larger number of GCMs improves the probability of accurately representing the uncertainty of the original GCM ensemble. The behavior for high and low flows is similar, although a smaller proportion of the overall variance can be attributed to GCMs for these two metrics. Importantly, Figure 2.6 shows that random sampling of GCMs is a strategy susceptible to result in large errors in the representation of uncertainties, unless a large number of GCMs is included. This is why GCM selection strategies have typically been favored over random sampling. Figure 2.6 also shows that there is a much higher risk of underestimating GCM variance than overestimating it. In addition, the magnitude of the underestimation can be a lot more severe than that resulting from an overestimation.

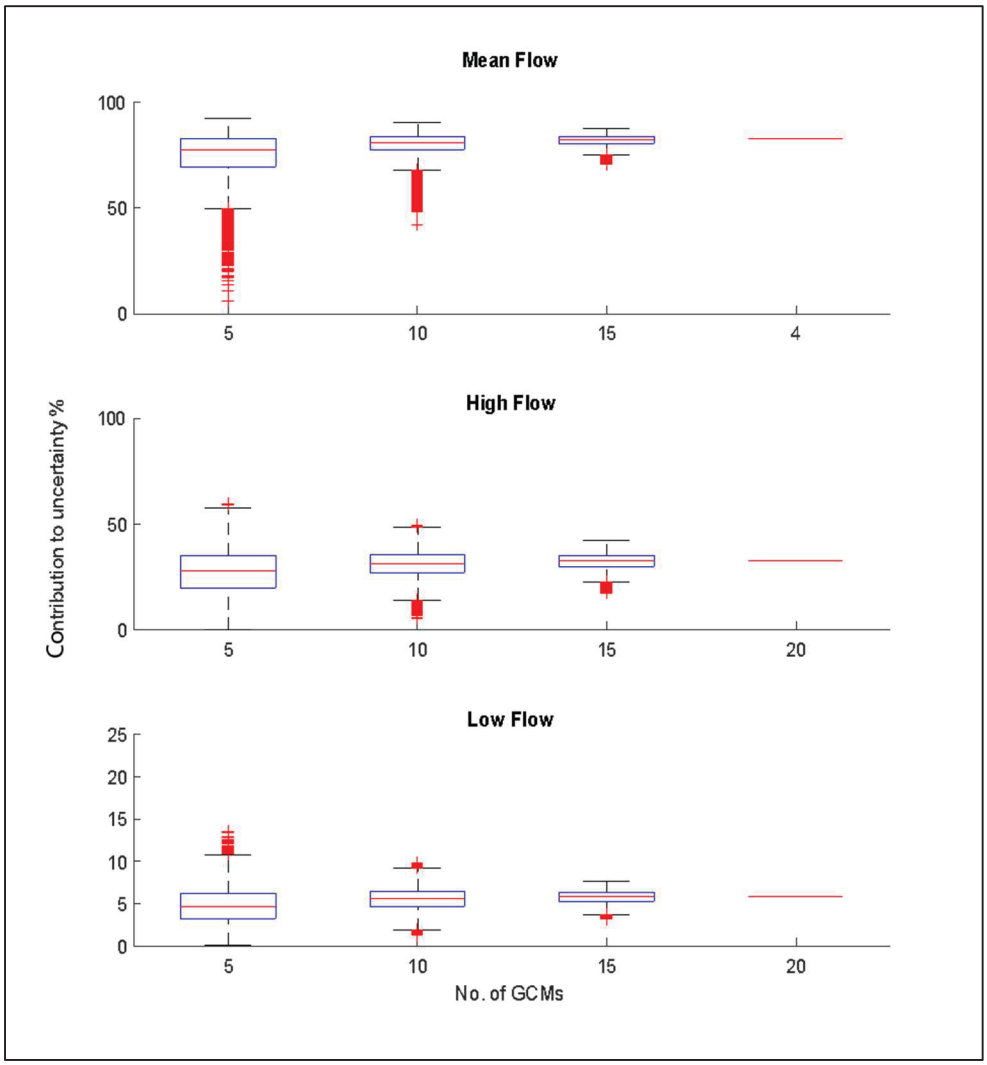

Figure 2.6 Boxplot of GCM contribution to uncertainty for all possible combinations of 5, 10 and 15 GCMS from the original ensemble of 20. These results are for one typical catchment

Figure 2.7 presents the results of the different GCM selection strategies discussed above. For each selection strategy, it shows the boxplot of the ratio of spread coverage (RSC) computed at each catchment, when using a subset of 5 GCMs. Each boxplot is therefore composed of 3540 catchment values. In the Figures below, the RSC is defined as the ratio of the variable's range in the subset (from each selection strategy) to the variable's range in the full 20-member ensemble. A RSC equal to 1 implies that the subset fully reproduces the spread of the full ensemble, while values less than 1 indicate that the subset underestimates the spread of climate

model uncertainty. While we acknowledge the limitations of the RSC metric, for example, a subset could technically achieve RSC = 1 by including only the models at the extremes, it nonetheless provides a simple, transparent representation of how much of the original ensemble spread is retained, rather than a complete measure of uncertainty itself. We use RSC as a practical diagnostic tool to compare subset performance, while recognizing that true uncertainty is multidimensional and cannot be fully captured by a single metric.

The figure compares the two chosen GCM selection methods, KKZ and K-means. For both methods, GCM similarity is evaluated using 5 different strategies: 1- using all of the climate variables listed in Table 2 (“All”), 2- using all climate variables but excluding highly correlated climate variables (“Low Corr”), 3- using the pair of climate variables suggested by Seo et al. (2019) which link specific climate indices to different hydrologic regimes; annual total precipitation (PRCPTOT) and mean annual temperature (ΔT) for mean flows, Rx5day and Rx3day (see table 2.2) precipitation for high flows, and diurnal temperature range (DTR) combined with Rn30day for low flows (“Seo”), 4- selecting the best-performing pair that maximizes the median RSC across all catchments (“Best”), and 5-choosing a random pair of climate variables (“Random”). It’s important to state the ‘best’ performing pair was chosen by optimizing the median RSC across all catchments and is NOT catchment-specific. This approach will therefore yield the best median result, but may not be the best one on all catchments.

For mean flows, the best-performing climate variables (the one resulting in the best RSC) was PRCPtot and precipitation above the 95th percentile (R95pTOT). For high flows, the best pair was Rx1day and R1mm. For low flows, the best indices were change in annual mean temperature (ΔT) and PRCPtot. For any given catchment, the choice of a single random subset is not representative (a subset could be good, bad, or average in terms of RSC), but when aggregated over 3,540 catchments (as shown in Figure 2.7, for example), it provides representative expected results from this strategy.

Our findings show that the KKZ method consistently outperformed K-means for selecting GCMs that better preserve the range of variability (the boxplots on the left hand side of Figure 2.7, are systematically better than the ones on the right hand side). The Seo climate variables were close to that of the best performing ones for mean flows, but significantly worse for low and high flow. For low flows, GCM selection proved to be less critical, as the contribution of GCMs to overall uncertainty is relatively small. For low flow, all selection methods give somewhat similar results with exception of the ‘best’ approach which clearly outperforms the others. Overall results show that a subset of 5 properly selected GCMs allows to preserve most of the variability of the full ensemble of 20 GCMs, as the median RSC exceeds 0.9 for the ‘best’ method. it should be noted that even though the ‘best’ method outperforms the other approaches when looking at Figure 2.7, this may not be the case on all catchments. In fact, in some catchments, the ‘Seo’ approach outperforms the ‘best’ pair.

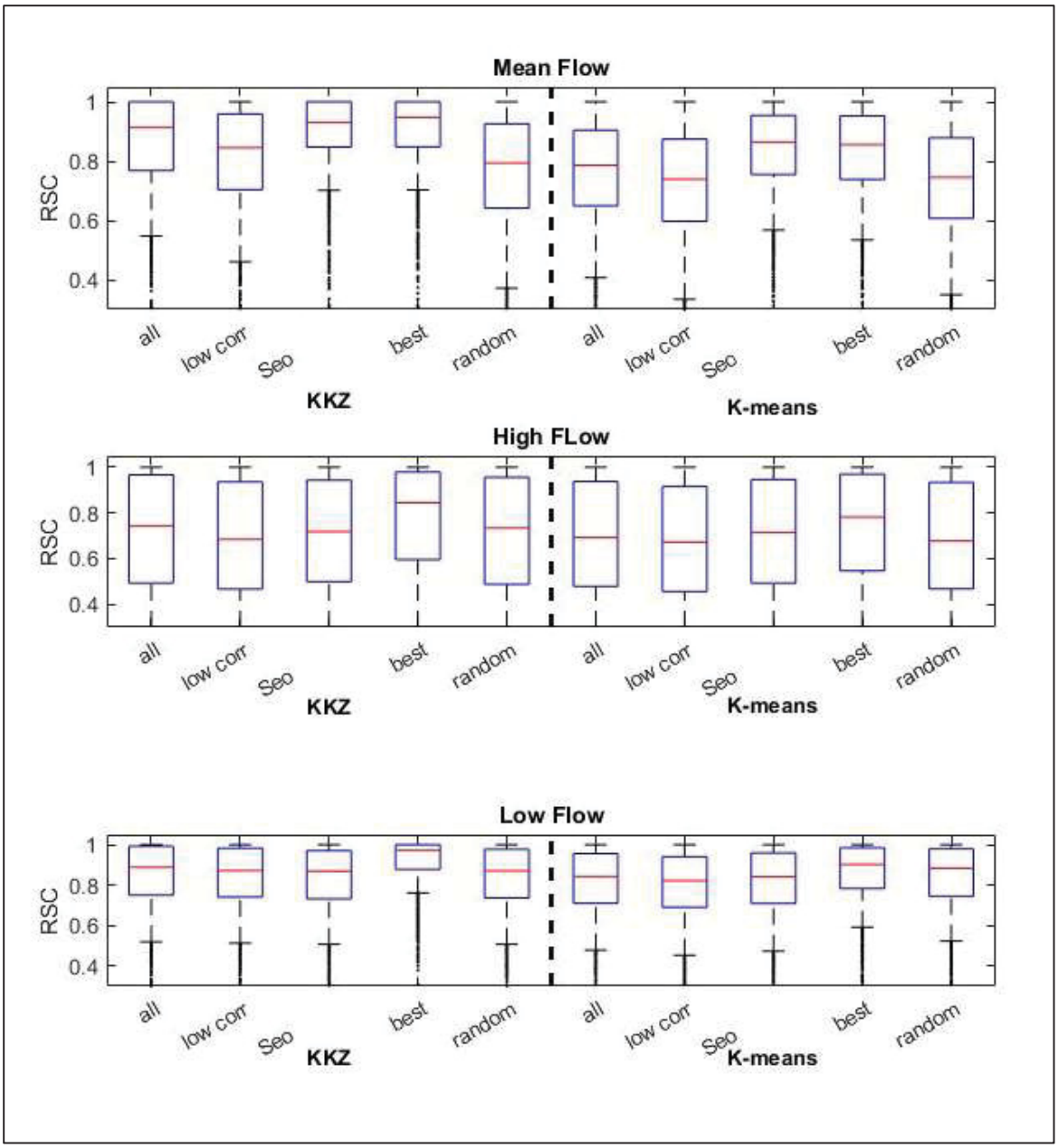

Figure 2.7 Boxplot of RSC across 3,540 catchments. The RSC is defined as the ratio of the range of selected GCMs to the range of all GCMs. The X-axis labels correspond to different selection methods: the first five boxplots represent the KKZ method, and the next five represent the K-means method. These results are for subsets of 5 GCMs.

Figure 2.8 presents a scatter plot comparing the RSC when using the KKZ method with “Seo” and “best” pairs of indices. The results are once again obtained from a subset of 5 GCMs. As expected, across all three flow indices, the ‘best’ indices give improved results with RSC

consistently closer to one (warmer colors). The difference is however small for mean flow, as both approaches provide similar results. Differences are however much larger for high and low flow, suggesting that the selection of appropriate climate variables is more complex. This is perhaps not surprising since the meteorological, climatological and physical processes leading to low flows and high flows are more complex than the ones leading to mean annual flow. For mean flow, mean annual precipitation and a temperature related index (e.g. mean annual temperature, annual PET, aridity index) are natural choices.

No clear spatial patterns are present in Figure 2.8, suggesting that local catchment characteristics or flow regime dynamics may play a larger role in performance. Indeed, in some catchments, the “Seo” indices outperform the “best” pairs, underscoring the context-dependent nature of index selection.

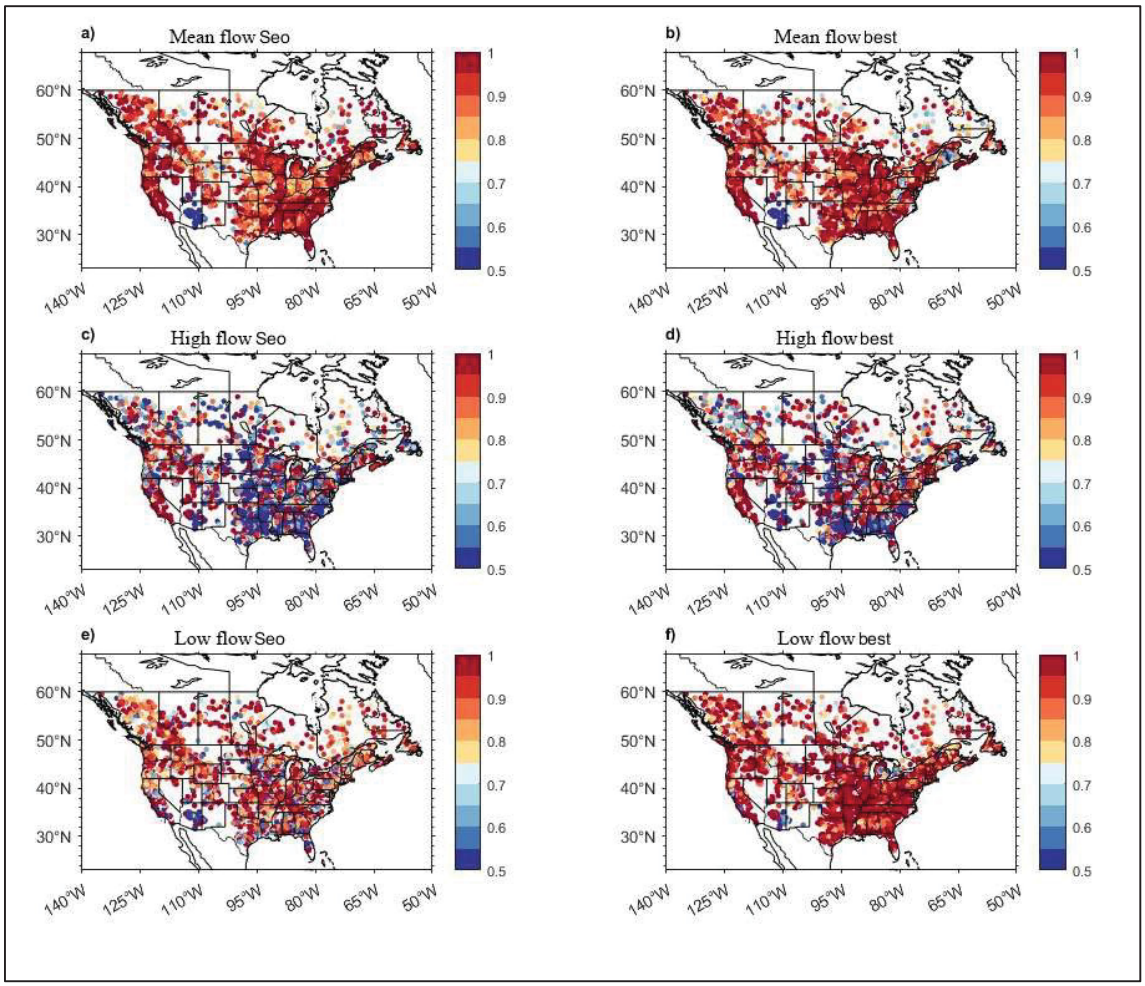

Figure 2.8 Map showing the RSC covered when selecting five GCMs (one combination per catchment) using the KKZ method with (a, c, e) Seo indices and (b, d, f) best indices across 3,540 catchments.

Results for Figure 2.7 and 2.8 were detailed for subsets of 5 GCMs. To look into the impact of subset size, Figure 2.9 presents the median RSC for low, high, and mean flows, for subsets of 1 to 19 GCMs out of the original ensemble of 20. To simplify the Figure only two selections methods are shown: (1) the KKZ method using the "best" pair of climate variables (solid lines in Figure 2.9), and (2) KKZ method using the "Seo" pair of climate variables (dashed lines in Figure 2.9). These two choices represent the best-performing approaches based on the preceding results. The plotted values represent the median RSC from the 3,540 catchments, providing a generalized view of subset performance. While the median offers a generalized view, the full distributions shown in earlier figures illustrate the underlying spatial variability.

It is important to note that RSC is only one representation of uncertainty. Although range-based measures are intuitive and easy to interpret, they do not capture all aspects of uncertainty (e.g., variance or distributional shape). Here, RSC is used as a practical and interpretable metric for large-sample comparison, while acknowledging its limitations.

The “best” variable pair consistently outperforms the “Seo” one. This is by design, since the “best” pair was optimized to have the best median RSC. The difference between the two is smallest for mean flows and largest for low flows as was also shown in Figure 2.7. All results eventually converge toward a RSC of 1. However, the rate of convergence depends on the chosen streamflow metric. Convergence is particularly slow for the high flows with respectively 8 (best) and 11 (Seo) GCMs needed to reach a RSC of 95%, compared with 4 (best) and 9 (Seo) for low flows and 6 (best) and 7 (Seo) for mean flow. The large difference between these two options for high and low flow indicate that results are sensitive to the choice of climate variable used for GCM selection. In those two cases, 10 GCMs need to be selected so that differences become much smaller. For mean flow, which is less sensitive to the choice of climate variables, only 5 GCMs are needed to reach comparable results.

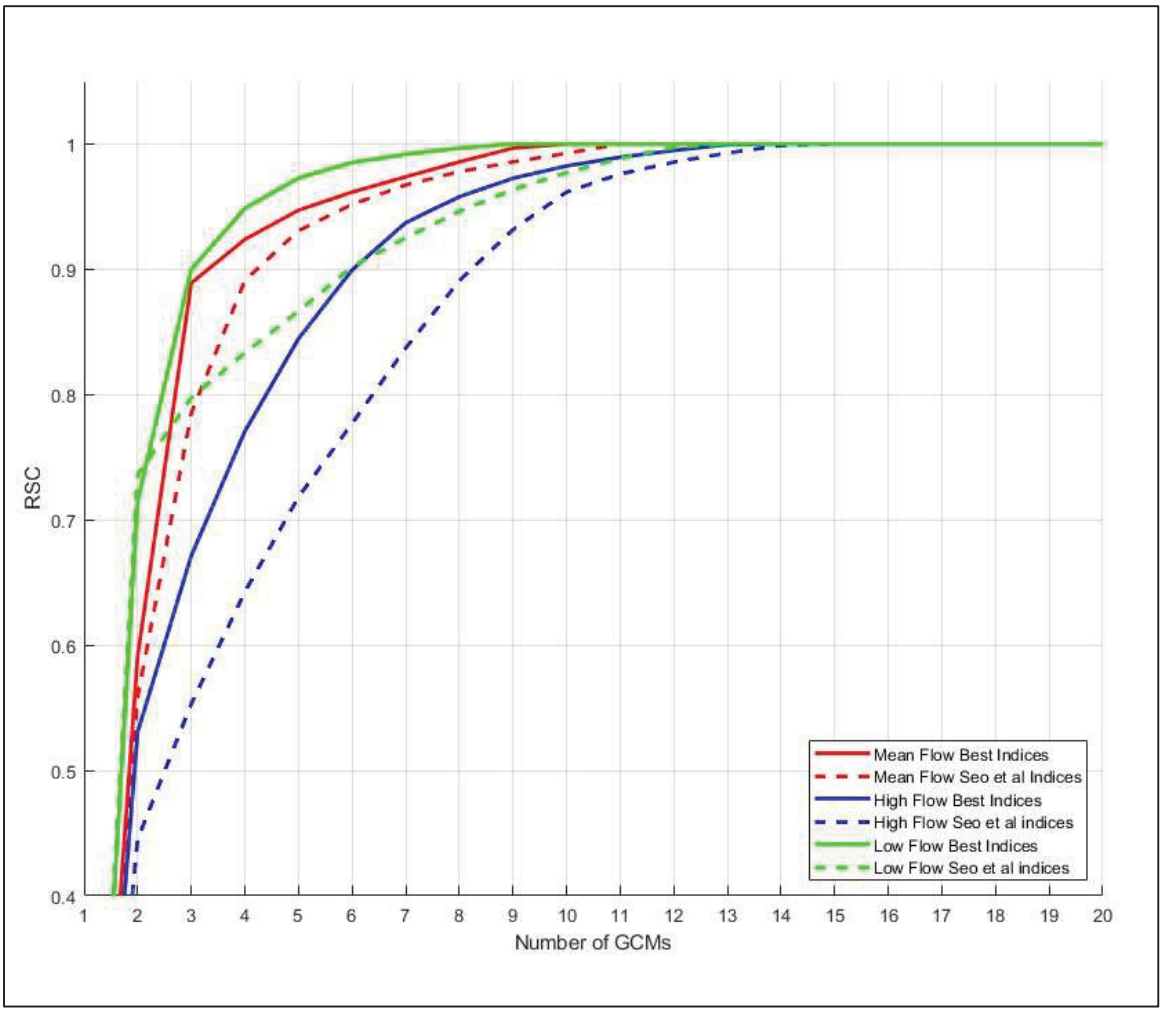

Figure 2.9 RSC for low, high, and mean flows. GCM subsets are selected using the KKZ method with "Seo" and "best" indices. The values shown represent the median RSC across 3,540 catchments.

2.4 Discussion

Hydrological impact studies often use ensemble approaches to capture uncertainty across the modeling chain. Since climate models contribute significantly to this uncertainty, selecting a representative subset is essential but nontrivial. The main objective of this study was to investigate if the uncertainty from the climate modeling domain can be effectively transferred to the hydrological domain. GCM subset selection was performed using KKZ and K-means

algorithms, both of which rely on a multivariate space defined by climate variables to capture the differences among GCM outputs. A key objective was to identify the most effective climate indices for capturing GCM-driven variability and translating it into hydrological impact projections.

The findings show that carefully selected climate variables can significantly improve the transferability of variability from the climate modeling domain to hydrological impact assessments. The choice of indices used to define the multivariate space plays a critical role in determining how well the selected GCMs preserve the relevant variability for hydrological impacts. In many cases, a well-chosen pair of climate indices is sufficient to achieve strong performance. Adding more variables does not necessarily improve results and can even degrade performance by introducing redundancy, particularly when the added indices are highly correlated. For example, in the case of high flows, a randomly selected pair of indices performed similarly to the full set of climatic indices, which contain ETCCDI indices as well, suggesting that increasing the number of variables does not guarantee better transferability.

A pairwise comparison of climate indices was conducted to identify the best combinations for transferring variability from climate simulations to hydrological responses. For mean flows, the indices suggested by Seo et al. (2019) (e.g., annual total precipitation and mean temperature) performed almost as well as the best-performing pairs. This is likely because mean flows are largely influenced by long-term averages in temperature and precipitation, which these commonly used indices capture effectively. Although a comprehensive list of precomputed indices is available through ETCCDI, reducing this list by removing highly correlated indices did not improve performance, for mean flows, using the full set of indices generally yielded better results. This shows that indiscriminate elimination of the correlated climate variables does not necessarily yield improvements. On the other hand, a well-chosen small subset of indices (e.g., two variables) can perform nearly as well as the best possible combinations as shown above, underscoring the importance of informed selection rather than simple reduction.

However, for high flows and low flows, the performance differences between the suggested indices by Seo et al., the full set of ETCCDI indices, the reduced set of low-correlated indices, and even randomly selected pairs were relatively small. The underlying reason remains uncertain, but this may indicate that these flow regimes are influenced by more complex or indirect climate drivers, which are not easily captured by a limited set of general-purpose indices. For high flows, the most effective combination was Rx1day and the number of wet days, while for low flows, the optimal pair was the change in annual mean temperature (ΔT) and PRCPtot. In both cases, the optimal sets make a lot of hydrological sense. This shows that rather than relying solely on standardized indices such as those from the ETCCDI or default pairs like temperature and precipitation changes a more effective strategy may be to prioritize climate variables that are directly and physically linked to the specific hydrologic metric of interest. Such a targeted approach can enhance the precision and relevance of GCM subset selection, ultimately leading to more robust and meaningful hydrological impact assessments.

It's also noteworthy that the best-performing climate indices varied across catchments. The "best indices" identified in this study reflect those with the best median performance across all catchments; however, different pairs of indices may perform better in specific locations. For example, the previously identified Rx1day predictor for high flows may be a great one for medium size catchments with a response time close to one day, but may fail for larger catchments who may be more sensitive to long wet periods, or for catchments whose spring snowmelt is the main driver of high flows. Crucially, there is no way to know in advance which indices will perform best for a given catchment. Even catchments with similar characteristics can yield different results, indicating that the relationship between climate indices and hydrological responses is not easily predictable. Stepwise regression analysis revealed no consistent relationship between catchment characteristics and the most effective indices, highlighting the highly site-specific nature of GCM sub-selection. This variability makes it difficult to generalize best practices across regions. While tailored index selection can improve performance, it limits scalability for regional or continental-scale studies. As such, GCM selection should align with the specific objectives of the impact assessment and the hydrological variables of interest. Although it is feasible to determine optimal indices at the

catchment scale, these combinations often lack broad applicability, reinforcing the need for flexible, context-sensitive selection strategies. To offer more reliable recommendations, regional-scale (e.g., sub-national or basin-level) studies, perhaps guided by expert judgement on local hydrological processes, should be done to assist in identifying index combinations that balance transferability and performance.

The KKZ method generally outperforms K-means in preserving the spread of projections, as it tends to select models located near the boundaries of the multivariate space. That said, KKZ's tendency to prioritize outlier models can pose challenges, particularly when such models are unavailable or excluded due to poor historical performance. In these situations, the representativeness and reliability of the resulting subset may be reduced. To mitigate this issue, some studies recommend pre-screening GCMs and excluding those with inadequate performance before applying subset selection techniques. This ensures that only credible models are considered, improving the robustness of the selected ensemble. However, it is important to recognize that no single GCM selection method performs optimally across all catchments or hydrological metrics. The effectiveness of each method is context-dependent and should be aligned with the specific goals and characteristics of the study.

The number of GCMs needed to adequately capture uncertainty also varied by flow regime. Mean and low flows were often well represented with just five GCMs, whereas high flows required larger subsets to preserve the full range of projected variability. This can be attributed to the heightened sensitivity of high flows to extreme precipitation events, which are less consistently simulated across GCMs. Consequently, studies focused on high-flow metrics may need to retain a broader ensemble to avoid underestimating future risk.

Strategically selecting a representative GCM subset offers a practical pathway to reduce computational burden without sacrificing uncertainty representation. This is particularly valuable for regional-scale assessments or large-sample hydrology studies where running full GCM ensembles across many catchments is often infeasible. Our results support the notion

that well-designed GCM subsets can serve as viable alternatives to full ensembles in hydrological impact assessments.

Climate variable selection should be tailored to the specific flow regime and objectives of the impact study. A one-size-fits-all approach may overlook important aspects of hydrological behavior. The choice of climate variables used in GCM selection should reflect the physical processes relevant to the hydrologic variable of interest, be it rainfall intensity for floods or temperature trends for drought.

While subset selection is a simplification of the full ensemble approach, it remains a viable and often necessary alternative in data- and computation-intensive applications. By aligning selection strategies with study objectives and hydrological contexts, researchers can achieve reliable, efficient assessments that still account for the key sources of uncertainty.

2.5 Conclusion

This study used two GCM subset selection techniques (KKZ and K-means) and numerous climate indices to examine the transferability of climate model uncertainty to hydrological impact estimates over 3,540 North American catchments. The following is a summary of the key findings.

- K-means clustering and KKZ clustering are effective tools for choosing subsets of GCMs that cover the spectrum of climate projections. However, when taking the transferability of uncertainty to the hydrological world into account, KKZ mostly outperformed K-means.

- No single subset of GCMs can simultaneously preserve uncertainty across multiple hydrological metrics. The optimal subset depends on the variable of interest: the models that best capture mean flows are not the same as those that best capture high or low

flows. Consequently, to achieve a high transferability of uncertainty, carefully selected climatic indices are crucial. For mean flows, indices such as mean temperature and annual precipitation (Seo et al., 2019) are almost as effective as the best performing pairs of indices. For both high and low flows, higher performance is obtained when using indices that are more closely associated with these extremes (e.g., Rx1day, R1mm, PRCPTOT, ΔT).

- Expanding the selection space with additional indices does not necessarily improve transferability and may instead introduce redundancy, especially when variables are not directly linked to the hydrological and climatological dynamics of the system. In many cases, small, well-chosen subsets of indices outperform the full ETCCDI set.

- The hydrologic metric taken into consideration determines how many GCMs are required. While high flow typically requires larger subsets to preserve uncertainty due to the increased sensitivity to extremes, mean and low flows are generally well represented with five GCMs.

- The best-performing indices varied across catchments, with no consistent relationship to catchment characteristics. This demonstrates how GCM sub-selection is site-specific, making it challenging to generalize best practices across regions. Finding index combinations that strike a balance between performance and transferability may consequently need regional or basin-scale research, possibly assisted by expert judgement.

In conclusion, the study shows that well-crafted subsets of GCMs, chosen with suitable indices and strong sampling techniques, can retain a significant amount of the uncertainty from complete ensembles, providing a computationally viable and reliable substitute for large-sample hydrological impact analyses.

2.6 Code and data availability

The processed data and the used codes are available via contacting the authors.

2.7 Author Contribution

M.R.A. and F.B. designed the study, carried out the analyses, and wrote the first draft. R.A. and F.B. supervised the research and contributed to the interpretation of results. J.L.M. contributed to improving the manuscript through multiple revisions. All authors revised the manuscript and approved the final version.

2.8 Acknowledgments

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

2.9 Supplementary material

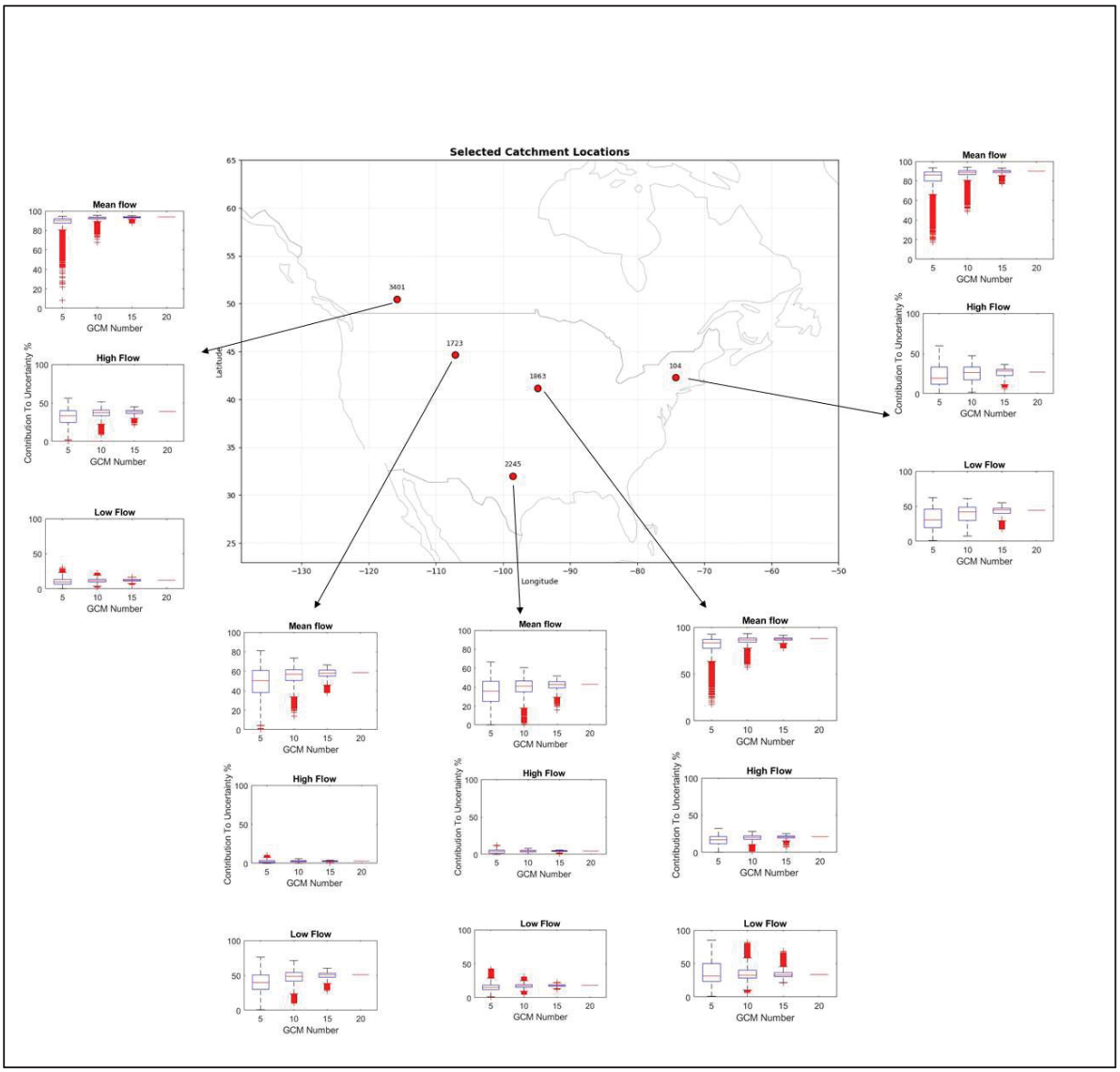

Figure S2.10 Boxplots of GCM contribution to total uncertainty in streamflow projections, calculated across all possible combinations of 5, 10, and 15 GCMs (out of a 20-member ensemble). Results are shown for five additional representative catchments, with the location of each catchment indicated on the accompanying map (red markers).

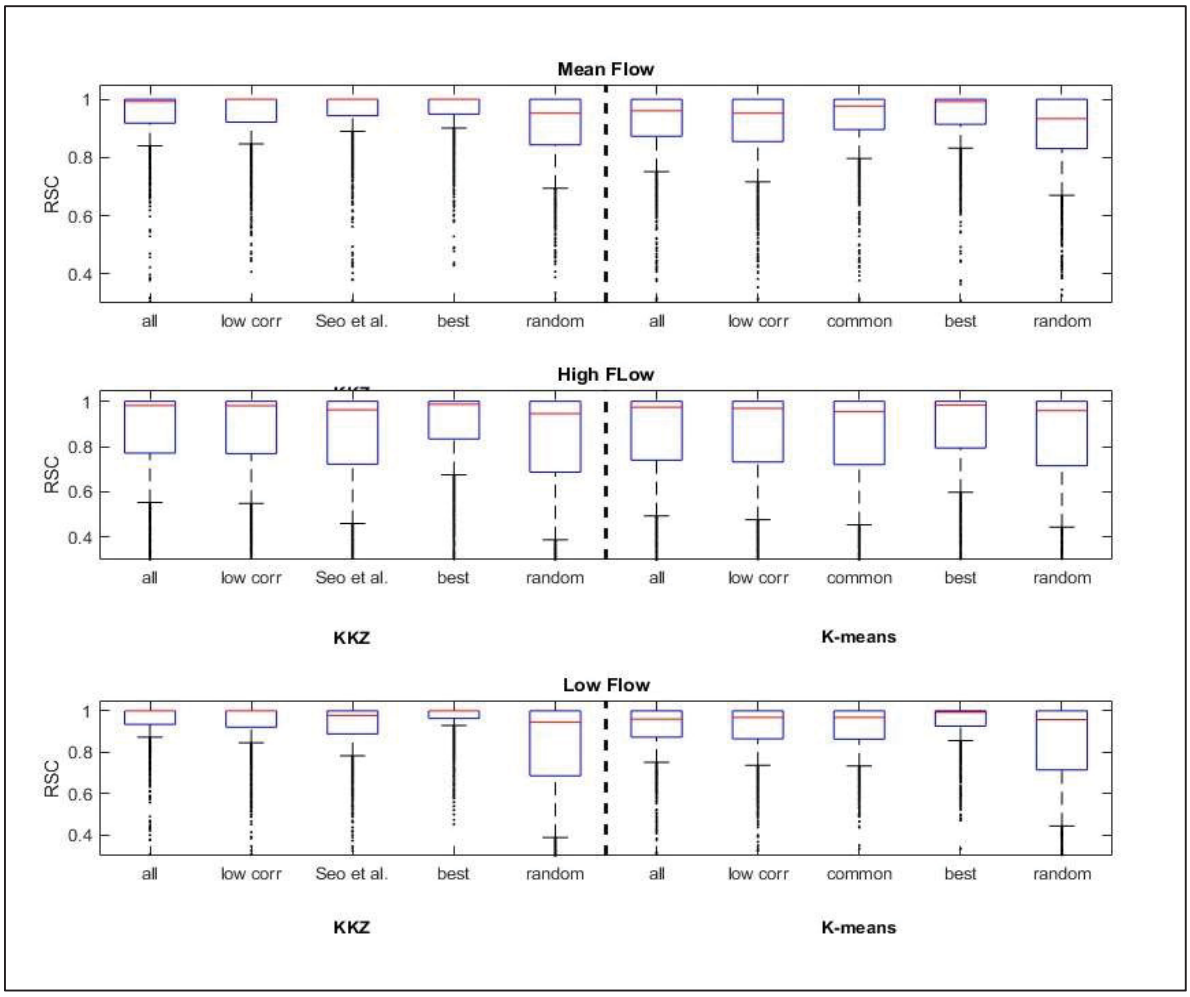

Figure 2.11 Same as Figure 2.7, but for subsets of 10 GCMs.

CHAPTER 3

UNDERSTANDING THE INFLUENCE OF 'HOT' MODELS IN CLIMATE IMPACT STUDIES: A HYDROLOGICAL PERSPECTIVE

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

[a] Hydrology, Climate and Climate Change Laboratory, École de Technologie Supérieure, 1100 Notre-Dame Street West, Montreal, Quebec, Canada H3C 1K3

Article published in *Hydrology and Earth System Sciences*, December 2023

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 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 hot models be removed from impact studies to avoid skewing impact results toward unlikely futures. Indeed, the inclusion of these models in impact studies carries a significant risk of overestimating the impact of climate change.

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 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 global outlier climate models do not necessarily provide regional 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.

3.1 Introduction

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 (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; Y. Zhang et al., 2023).

The concept of "model democracy" has been widely used in impact studies (e.g. Collins et al., 2013; IPCC, 2014) despite criticism (Knutti, 2010). 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 & Kim, 2008). However, this approach may be less effective for CMIP6 ensemble as the validity of some simulations is under question (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

sensitivity (ECS), a measure of the steady-state temperature increase in the event of doubled carbon dioxide (CO_2) concentrations in the atmosphere (Flynn & 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).

However, a plethora of evidence based on observations and our understanding of planetary physics indicate that we can confidently restrict the likely range of future warming trend 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, Rugenstein, et al., 2017; Sherwood et al., 2020) or emergent constraints (Cox et al., 2018; Nijsse et al., 2020; Shiogama, Watanabe, et al., 2022). For example, Sherwood et al. (2020) used multiple lines of evidence and concluded that the likely (with a 66% chance) ECS 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). It should be noted that the uncertainty surrounding the cooling impact (both direct and indirect) of aerosols on radiative forcing poses challenges in constraining future warming estimates (Bellouin et al., 2020; Forster et al., 2013; Smith et al., 2021). In essence, the current historical measurements do not provide a clear understanding of whether we are in a scenario of high sensitivity, fast-warming, accompanied by strong contemporary aerosol cooling, or if the situation is the opposite.

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). Incorporating climate models with high sensitivity into impact studies may potentially lead to an overestimation of the overall economic consequences arising from future climate changes (Shiogama, Takakura, et al., 2022). 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.

3.2 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, Brissette, Martel, 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 successfully calibrating hydrological models and bias-correcting climate models. Overall, 3107 catchments were retained.

Table 3-1 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: 5.62), NESM3 (ECS: 4.68), IPSL-CM6A-LR (ECS: 4.52), 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.

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

Table 3.1 The 19 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
CANESM5	5.62[1]
NESM3	4.68[1]
IPSL-CM6A-LR	4.52[1]
EC-Earth3-Veg	4.3[1]
EC-Earth3	4.2[1]
ACCESS-ESM1-5	3.88[2]
GFDL-CM4_gr1	3.89[2]
GFDL-CM4_gr2	3.89[2]
MRI-ESM2-0	3.14[1]
MPI-ESM1-2-LR	3.02[2]
BCC-CSM2-MR	3.01[1]
MPI-ESM1-2-HR	2.98[2]
FGOALS-g3	2.87[2]
GFDL-ESM4	2.62[1]
NorESM2-LM	2.60[1]
MIROC6	2.57[1]
NorESM2-MM	2.49[2]
INM-CM5-0	1.92[1]
INM-CM4-8	1.83[1]

Figure 3.1 illustrates the methodological framework for each study catchment. Precipitation and temperature data are first extracted from 19 CMIP6 climate models under the SSP8.5 scenario for both the reference and future periods. Using precipitation and temperature from the ERA5 reanalysis over the reference period, climate data is then bias-corrected using the

MBCn method. These bias-corrected climate scenarios are subsequently employed as inputs for a calibrated hydrological model to compute streamflows. These computed streamflows are then used to examine the impact of including (or not including) 'hot' models in the impact study, using a set of defined metrics. Further details are provided below.

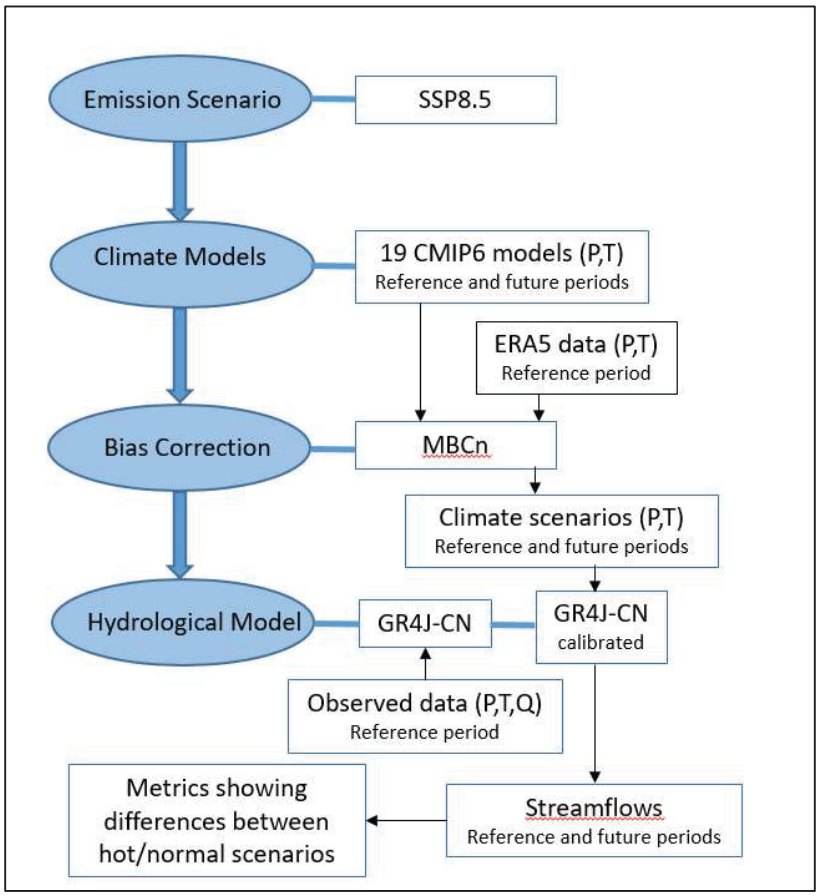

Figure 3.1 Methodological framework performed for each of the study catchments

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 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. (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 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 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). The calibration was performed using the Kling-Gupta Efficiency (KGE) metric. The KGE metric (Gupta et al., 2009) directly combines the bias, ratio of variance, and correlation into a single metric. It provides a more robust and refined assessment of model performance when calibrating hydrological models, addressing the drawbacks of the Nash-Sutcliffe Efficiency metric (NSE, Nash & Sutcliffe, 1970) (Knoben et al., 2019). Figure 3.2 presents the location of the 3107 retained catchments, each having a KGE calibration value above 0.5.

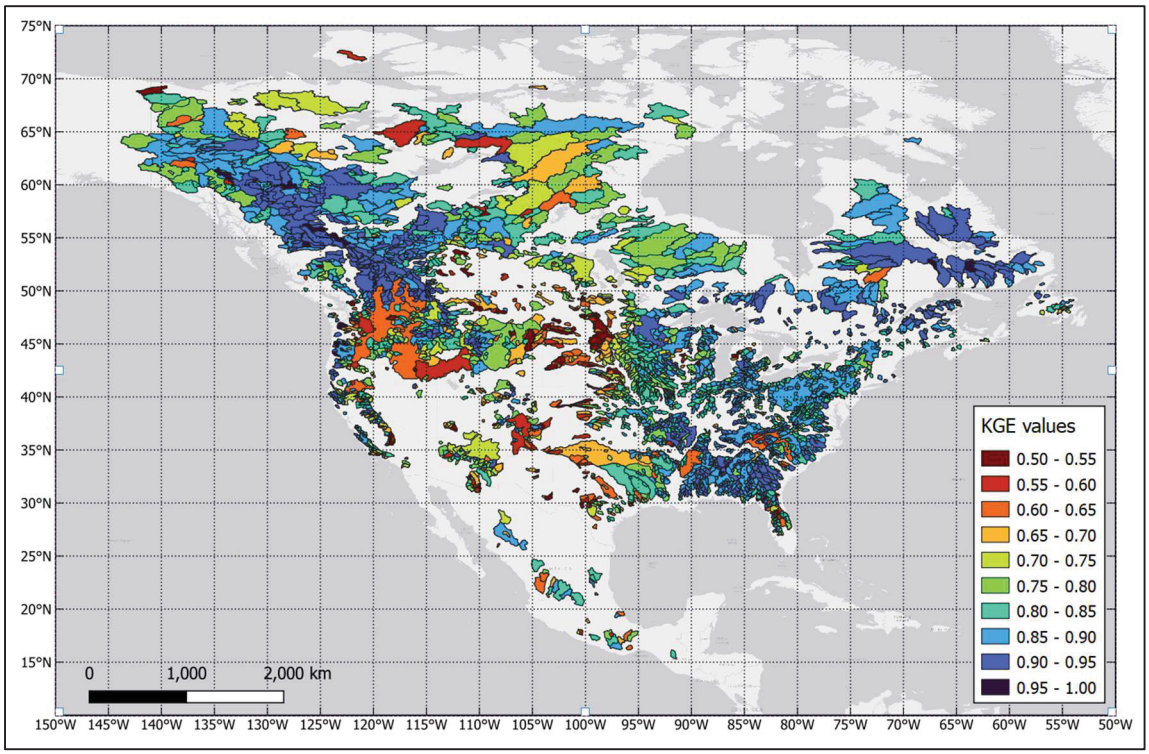

Figure 3.2 Study catchment location. The color scale corresponds to the hydrological model KGE calibration score over the reference period. Only catchments with available data, KGE values higher than 0.5 and area larger than 500 km^2 were selected

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 (Q_{mean}) and high (Q_{max}) and low (Q_{min}) flows:

- Q_{mean}: obtained by averaging daily streamflow over the 30-year period.
- Q_{max}: obtained by averaging the 30 annual maximum simulated streamflows.
- Q_{min}: 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 3.3 provides a schematic representation illustrating how the three dispersion metrics are interpreted in this study. It serves as a guide for understanding the spread (or uncertainty) of future streamflow projections. For the three streamflow metrics, 19 values from the original ensemble and 14 from the reduced 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.

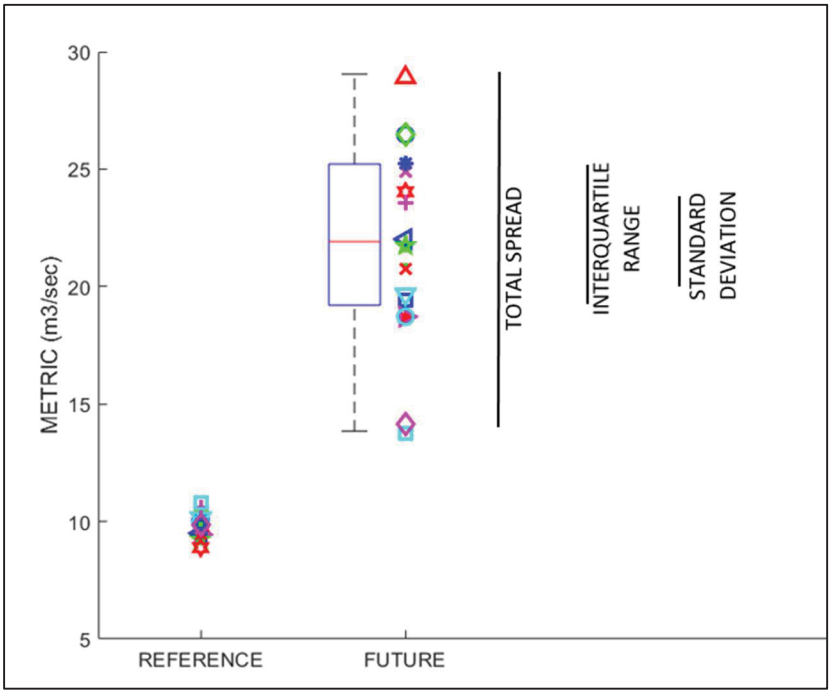

Figure 3.3 Representation of the dispersion metrics used in this paper. Each marker represents one of the 19 climate models. METRIC will either be Q_{mean}, Q_{max} or Q_{min}, all having units of m^3/sec

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

$$TS = metric_{max} - metric_{min} \tag{3.1}$$

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 3.3.

$$IQR = Q_{75} - Q_{25} \tag{3.2}$$

Finally, the standard deviation (σ) is the standard mathematical measure of dispersion. In the case of a normal distribution, 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 m^3/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:

$$TS_{nd} = \frac{TS_{14}}{TS_{19}} \tag{3.3}$$

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:

$$IQR_{nd} = \frac{IQR_{14}}{IQR_{19}} \tag{3.4}$$

However, in this case, the potential values vary in the 0 to ∞ range. More practically, a value below 1 indicates that the IQR has 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.

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

$$\sigma_{nd} = \frac{\sigma_{14}}{\sigma_{19}} \tag{3.5}$$

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 ∞).

3.3 Results

Figure 3.4-a 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 2.5th and 97.5th quantile of the distribution, providing a 95% coverage of the dataset. Quantiles below 2.5 and above 97.5 are shown as dots. Results indicate that the distribution of projected temperature increases 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).

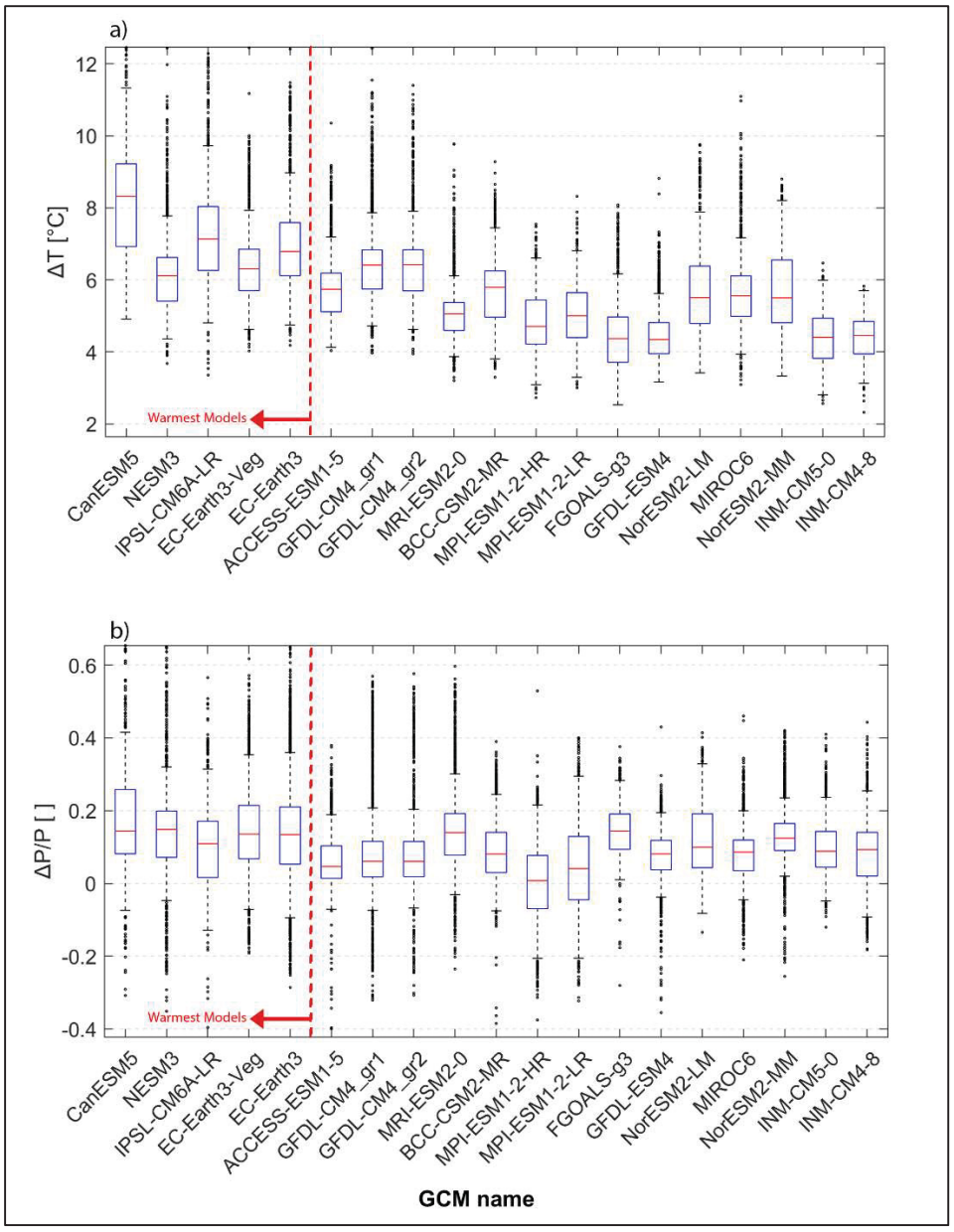

Figure 3.4 a) Distribution of projected temperature increase (ΔT) and b) projected relative annual precipitation increase (ΔP/P) for the 19 CMIP6 selected model for the 2070-2099 future period, compared to the 1971-2000 reference period. Each boxplot represents the distribution of projected increases 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. The boxplot whiskers correspond to the 2.5th and 97.5th quantiles and a few catchments that were beyond the Y-axis limits are not shown

Figure 3.4-b 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 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.

In order to show regional patterns related to Figure 3.4, Figure 3.5 displays the mean ΔT (3-5a) and mean ΔP/P (3-5b) ratios between hot models and normal models. For temperature a red color indicates that hot models are warmer than the other models on average. For precipitation, blue colors highlight increased precipitation in the hot models compared to the normal models. Overall, the hot global models exhibit a systematically larger temperature increase over the entire study domain. The hot models mostly exhibit increased precipitation compared to the normal models. However, the west coast of the U.S., as well as some catchments in the southwestern U.S., exhibit a decrease in precipitation according to the hot models. These observations underscore the regional variability in temperature and precipitation patterns when comparing hot and normal models.

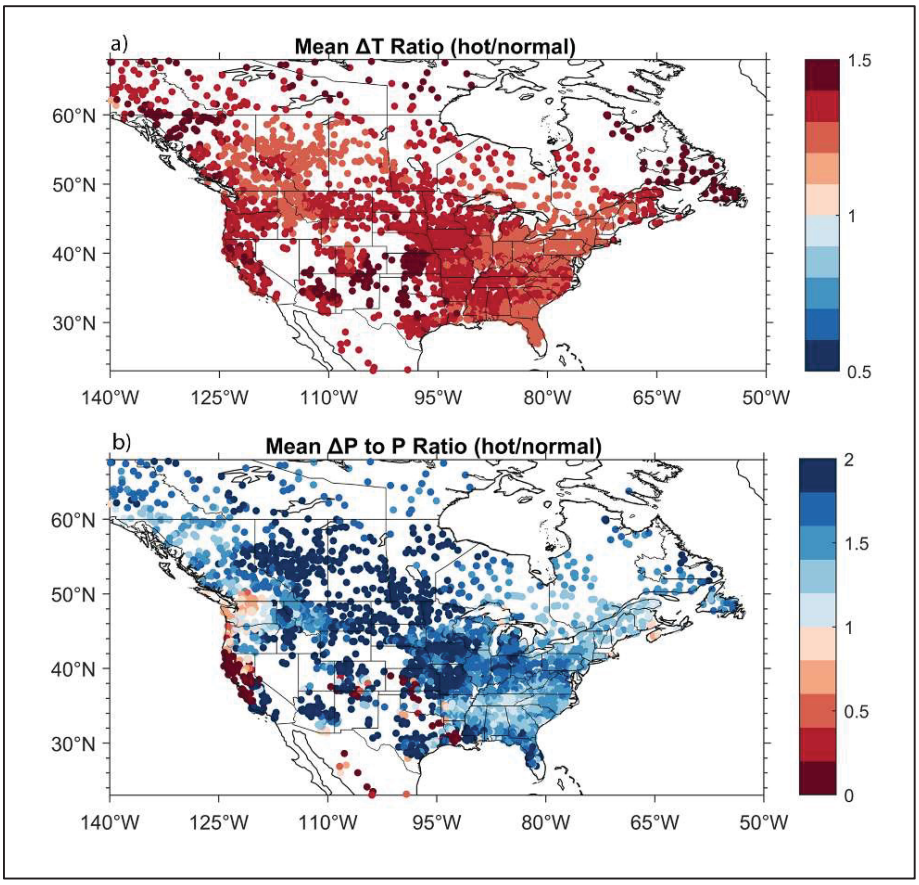

Figure 3.5 Mean ΔT (a) and ΔP/P (b) ratios (hot models to normal models). For ΔT, a red color indicates that hot models, on average, are warmer than their normal (non-hot) counterparts. For ΔP/P, a blue color shows that hot models are wetter than their normal (non-hot) counterparts. The graphs represent the differences computed between the future and reference periods

Figure 3.6 presents the ratio of mean projected streamflow changes (hot models/normal models) for Q_{mean}, Q_{max} and Q_{min}. A blue color indicates larger projected streamflows by the 'hot' models. Results show spatial patterns which differ depending on the streamflow metrics. Hot models project higher mean flows over most of the study domain, except in the south-west regions, where increased evapotranspiration nullifies potential increases in precipitation. For Q_{max}, increases are mostly localized in the Eastern US, whereas Q_{min} are widely increasing in Canada and mostly decreasing in the US.

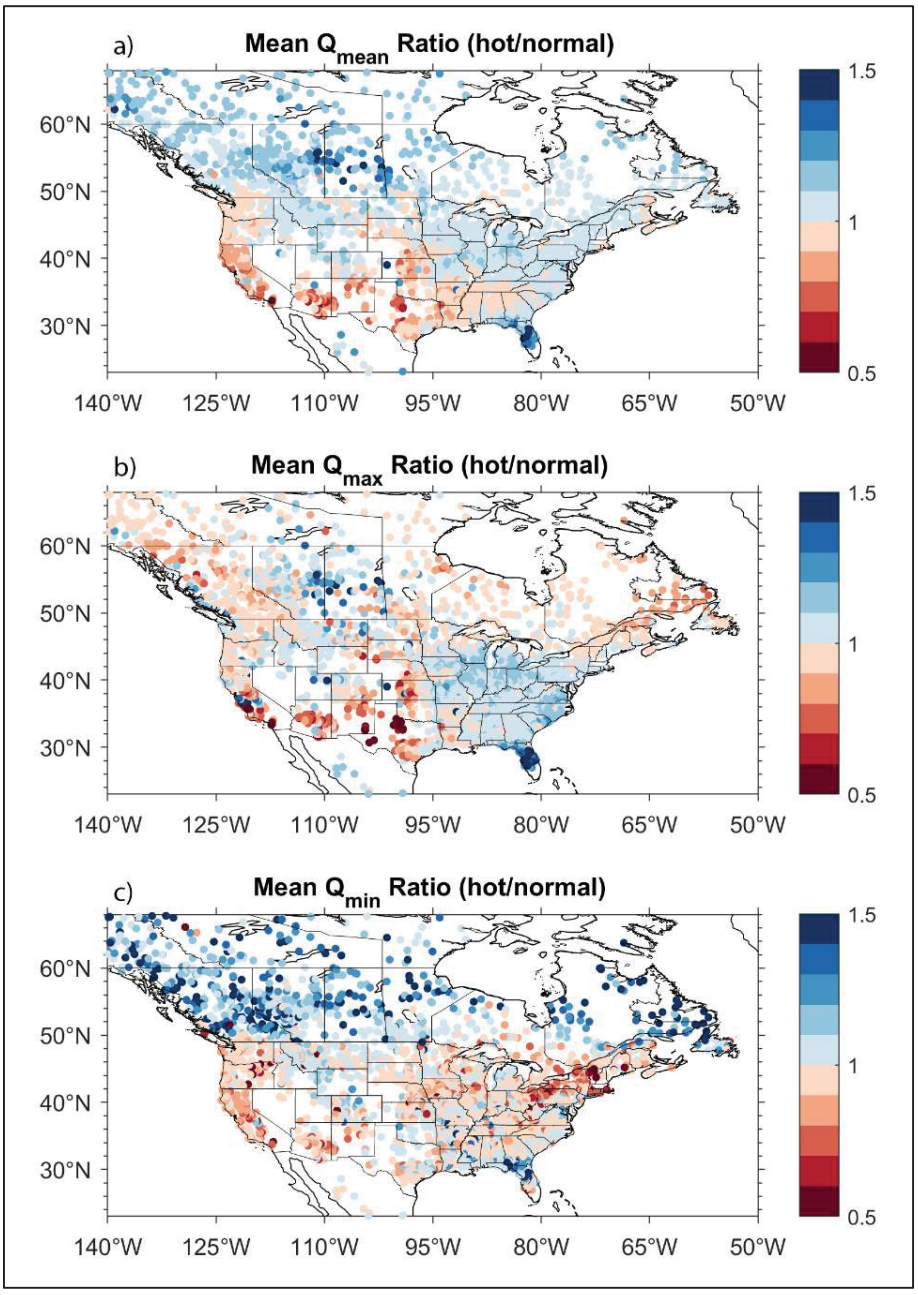

Figure 3.6 Ratio of mean projected changes: 'hot' divided by normal models. a): Q_{mean}; b) Q_{min}; c): Q_{max}). A blue color shows that hot project larger streamflows than their normal (non-hot) counterparts

Figure 3.7 presents the TS_{nd} 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 TS 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 TS 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 TS 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 TS for Q_{mean}, 53.3% for Q_{max} and 61.7% for Q_{min}.

The data from Figure 3.7 are shown in the form of boxplots in the left side of each panel to better illustrate the range of TS reduction. It shows that the median TS_{nd} 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 TS (43%, 46.7%, and 38.3% respectively). However, there is a significant reduction in TS observed in many catchments, and this decrease is strongly dependent on the geographical location of the catchments. Additionally, it can be seen that removing the hot models has a greater impact on Q_{min} than on the other two metrics.

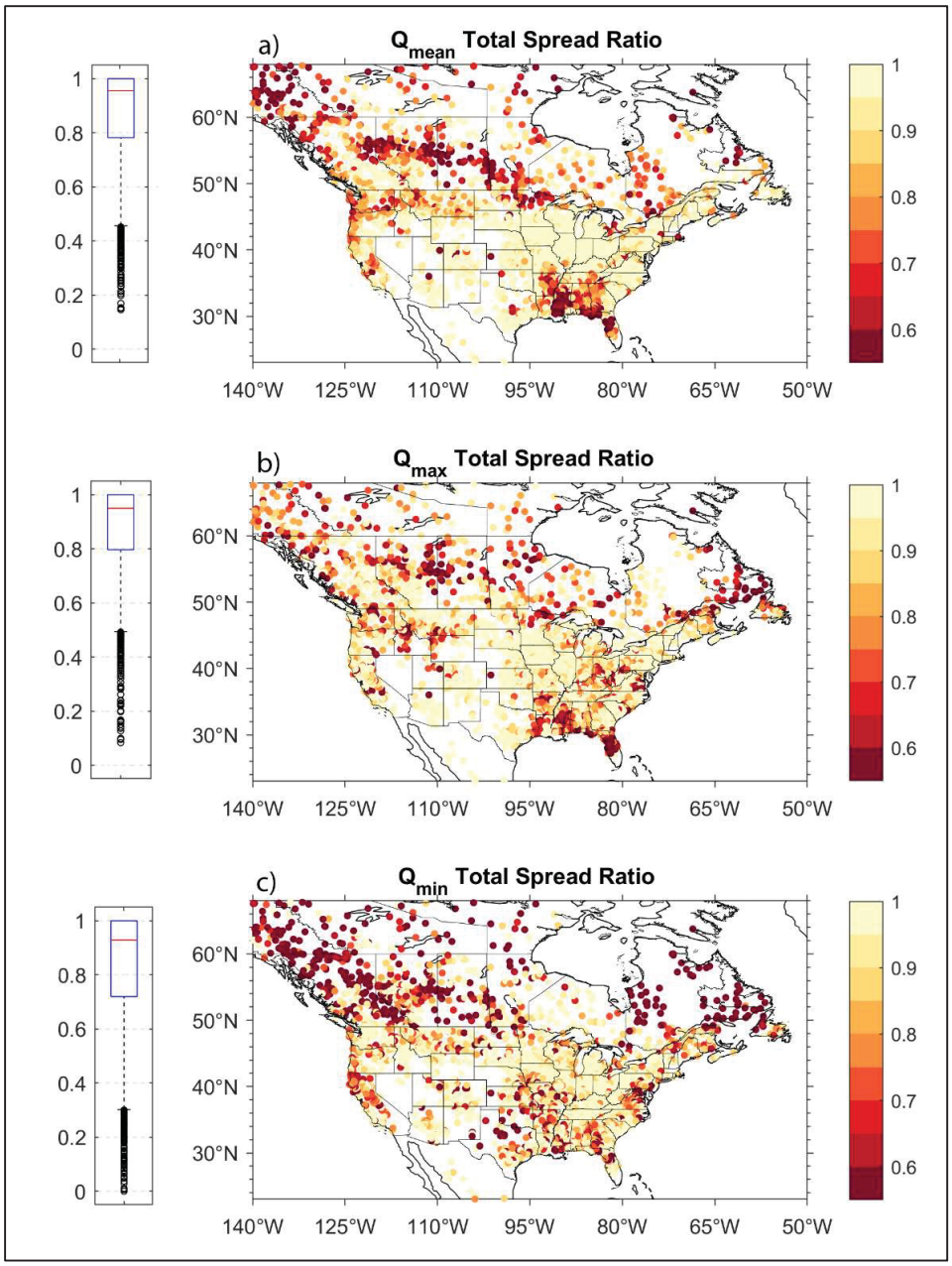

Figure 3.7 Total spread ratio ($TS_{nd} = \frac{TS_{14}}{TS_{19}}$) for Q_{mean} (a), Q_{max} (b), and Q_{min} (c) resulting from the removal of the five hot models. Boxplots are shown in the left

The TS_{nd} is heavily impacted by outliers and may not accurately represent the overall spread of models. Figure 3.8 presents the σ_{nd} for the three streamflow metrics. A red color (σ_{nd}> 1) indicates that the model spread has increased following the removal of the hot models whereas

a blue color (σ_{nd}< 1) corresponds to a decrease. Results indicate that removing the hot models consistently reduces σ_{nd} 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 outlier models in the north central, north-east, and southwest of the US results in an increase in σ_{nd} for both Q_{mean} and Q_{max}. Overall, as shown in the boxplots of Figure 3.8, removing the hot models likely reduces the spread in roughly two-thirds of catchments, while one-third see an increase. These values are larger than those obtained for TS. The Trends seen in IQR_{nd} is also very similar to that of σ_{nd} (see figures S3.10 and S3.11).

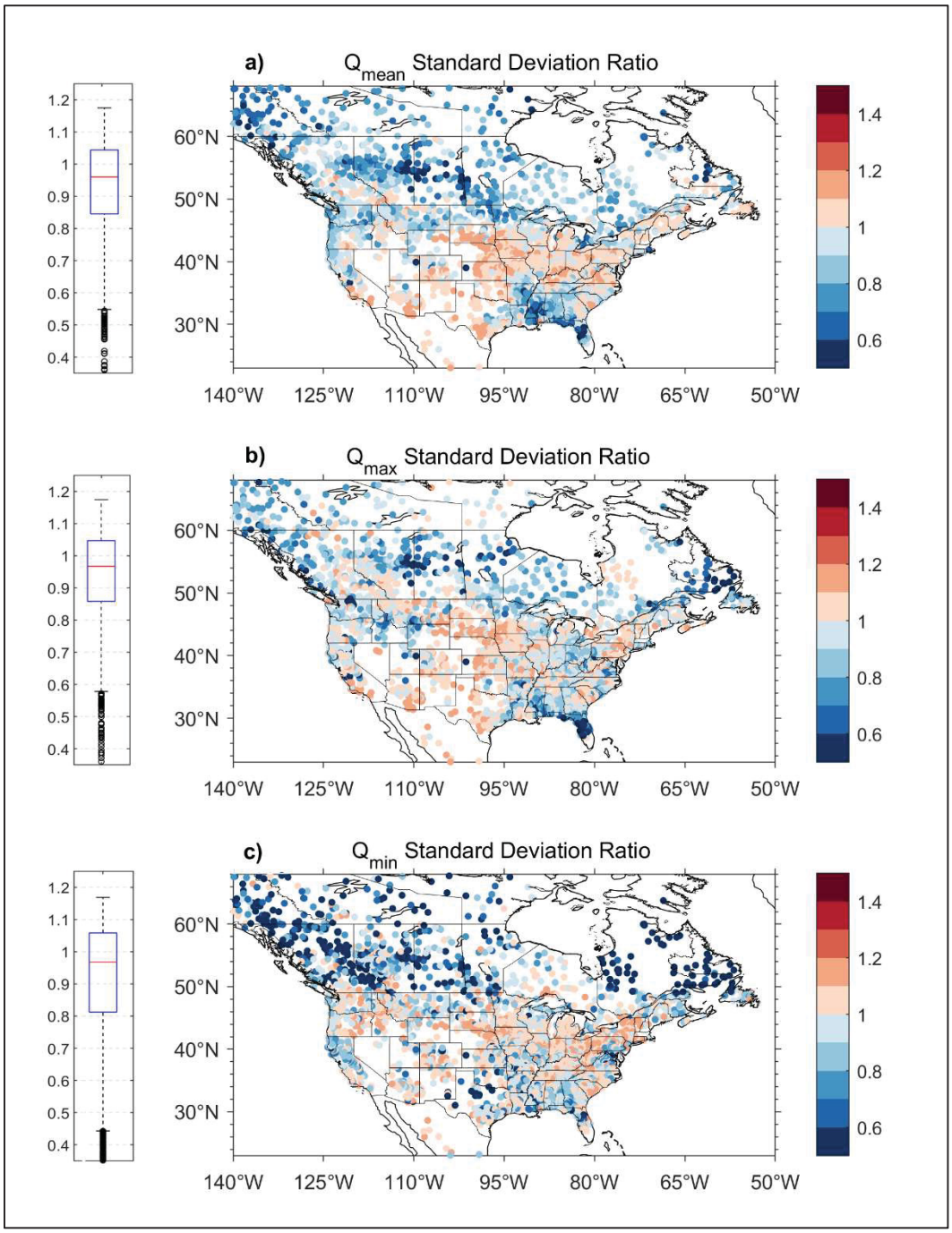

Figure 3.8 Standard deviation ratio ($\sigma_{nd} = \frac{\sigma_{14}}{\sigma_{19}}$) for Q_{mean} (a), Q_{max} (b), and Q_{min} (c) resulting from the removal of the five hot models. Boxplots are shown in the left side of each panel

3.4 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 uncertainty in climate impact studies to develop effective adaptation measures. Incorrectly representing uncertainty can lead to poor adaptation.

With the increased future temperatures, an intensification in the hydrological cycle is expected. However, it does not guarantee an automatic increase in water flow rates. This is because the rise in average temperature can also have a considerable impact on evapotranspiration. The outcome of these two factors working together is complex and varies based on the geographical location and primary climate zones. The research paper indicates that regions characterized as 'hot' tend to be associated with increased precipitation, further complicating the relationship between temperature and water flow.

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 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. Shiogama, Watanabe, et al., (2022) also concluded that the inclusion of hot models leads to an overestimation of annual mean precipitation increases in Alaska, Canada, and the western United States, where there is a substantial decrease in the variability of streamflow metrics.

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 standard deviation ratio. This was repeated for all catchments and the aggregated results are shown in Figure 3.9.

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

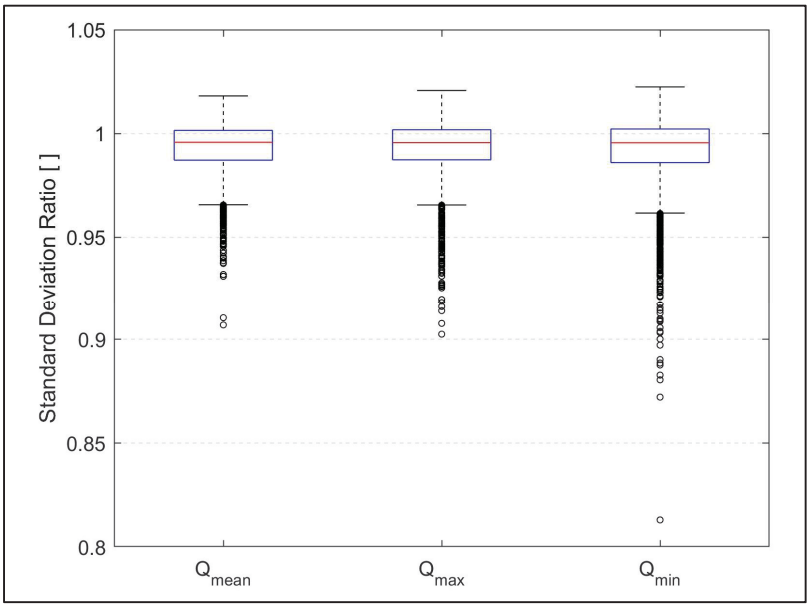

Figure 3.9 Boxplots of the average standard deviation ratio for Q_{mean}, Q_{max}, and Q_{min} resulting from the removal of 5 random models, after sampling 100 random combinations of 5 models

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

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

The regional impact of model importance is also compared (see figures S3.12 and S3.13 supporting information), which demonstrate the total spread ratio resulting from removing a single climate model and creating an 18-member ensemble. CanESM5 (Figure S3.12) and NESM3 (Figure S3.11) have the highest global sensitivity in this study. Removing CanESM5 leads to a clear reduction of total spread in Alaska and Yukon (for Q_{mean} and Q_{min}) and in the Southeast USA for Q_{max}, 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, as the high ECS value of NESM3 does not automatically translate into a correspondingly higher level of regional warming (see also Figure 3.4), demonstrating that it is not an outlier in most regions. This underscores the strong regional differences among globally identified hot models.

The only uncertainty in this study is that originating from GCM/EMSs. As stated earlier, in most impact studies, additional 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 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 *a posteriori* the impact of including hot models or not. If a selection of a subset of climate models is necessary (whether due to computational constraints or to avoid redundant or poorly performing 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.

3.5 Conclusion

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).

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 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.

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 selection of model subsets for impact studies.

3.6 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 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.

3.7 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 provided editorial feedback on the paper's early drafts.

3.8 Acknowledgments

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC; grant no. RGPIN-2020-07242). We extend our gratitude to the editor, Efrat Morin, and the anonymous reviewers for their constructive comments that helped improve the quality of this paper.

3.9 Supplementary material

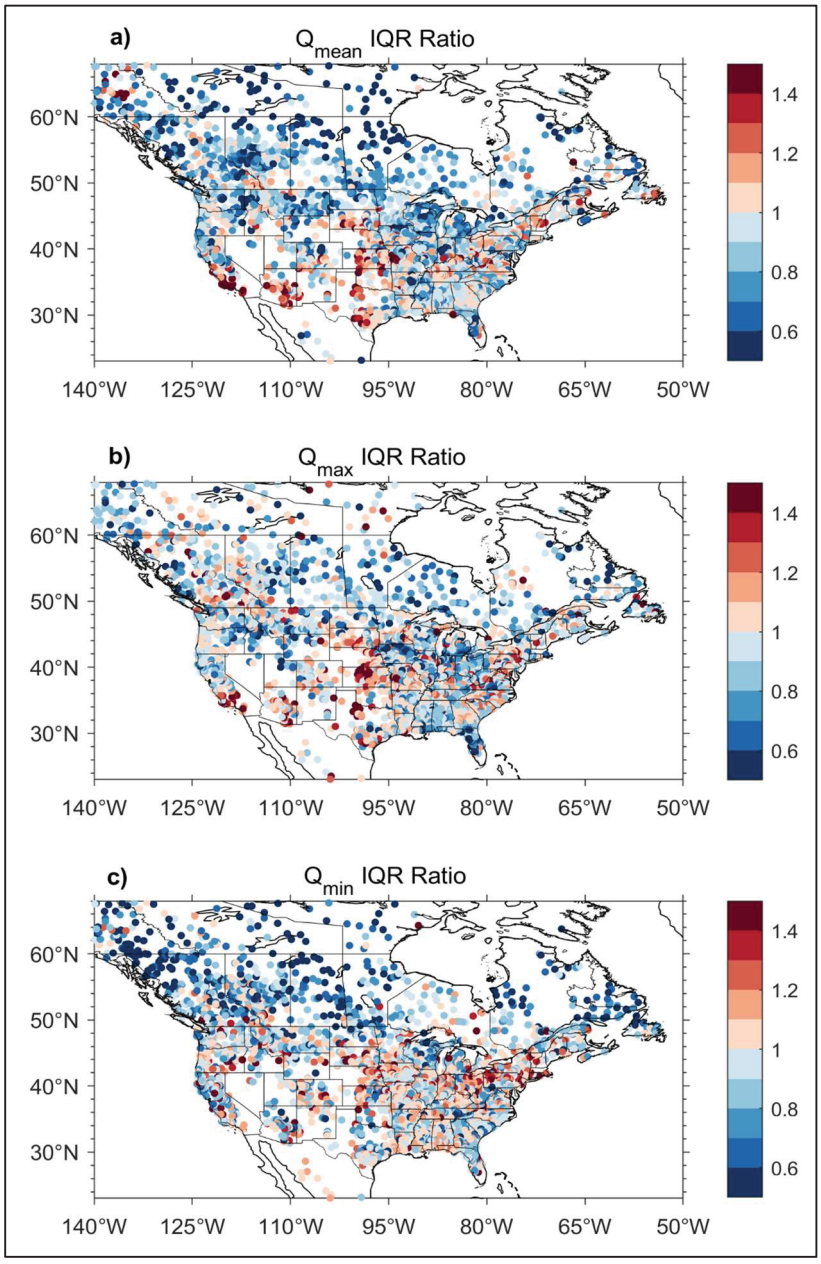

Figure S3.10 Change in the IQR ratio for Qmean (a), Q_{max} (b), and Q_{min} (c) resulting from the removal of the five hot models

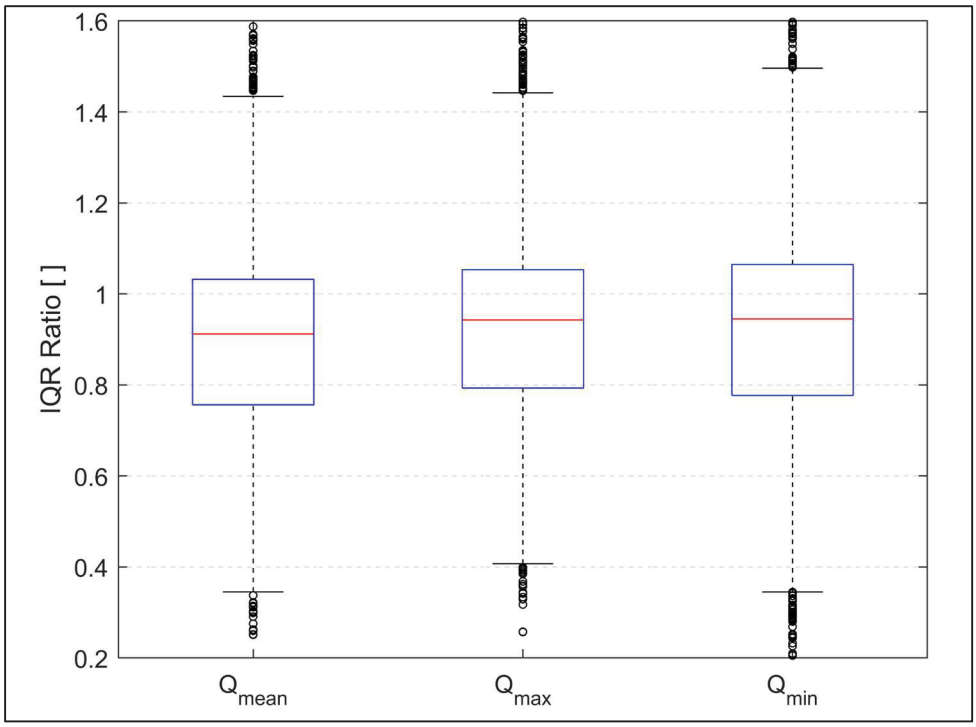

Figure S3.11 Boxplots of change in the interquartile range ratio ($\boldsymbol{IQR_{nd}}$) for Qmean, Qmax and Qmin resulting from the removal of the 5 hot models. A few outliers are beyond the Y-axis limits

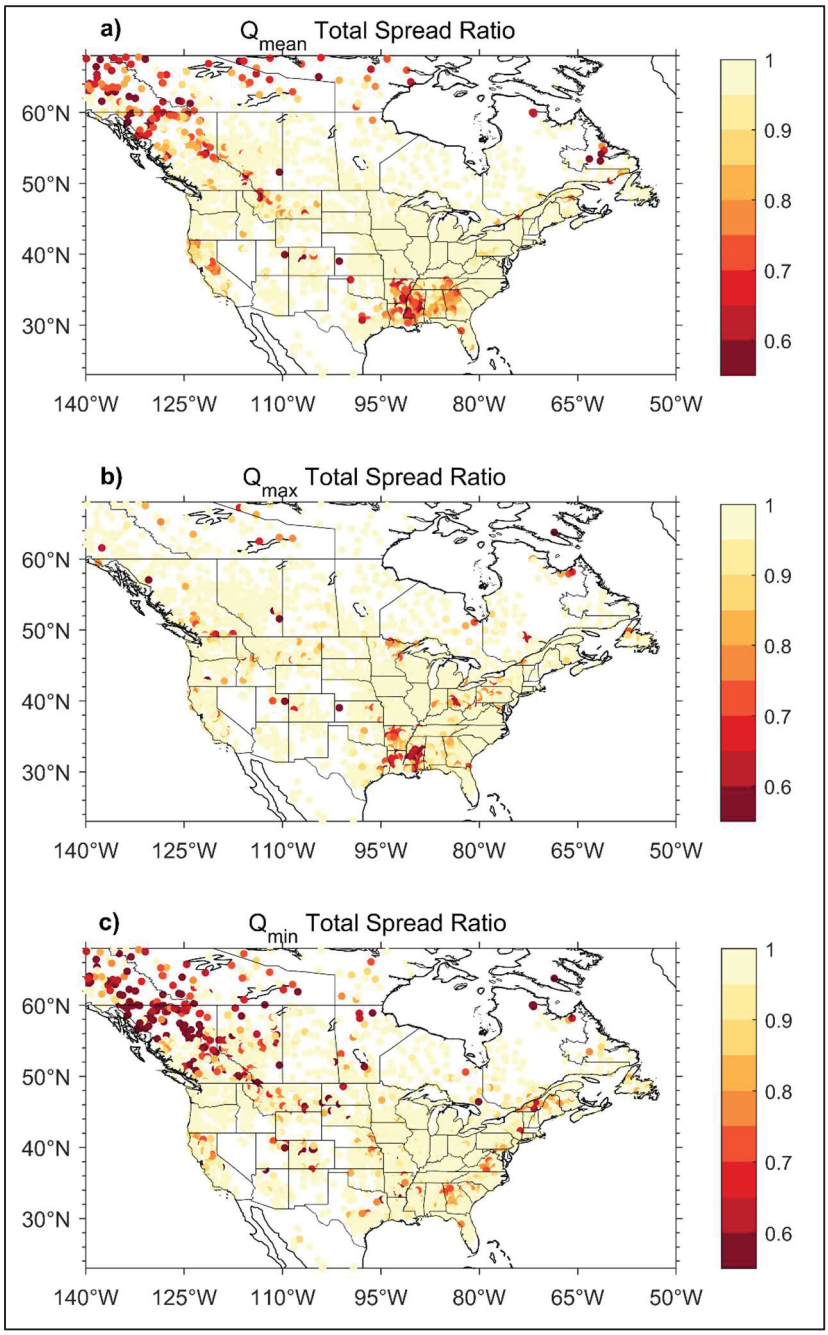

Figure S3.12 Total spread ratio for Q_{mean} (a), Q_{max} (b), and Q_{min} (c) resulting from the removal of a single climate model (CanESM5)

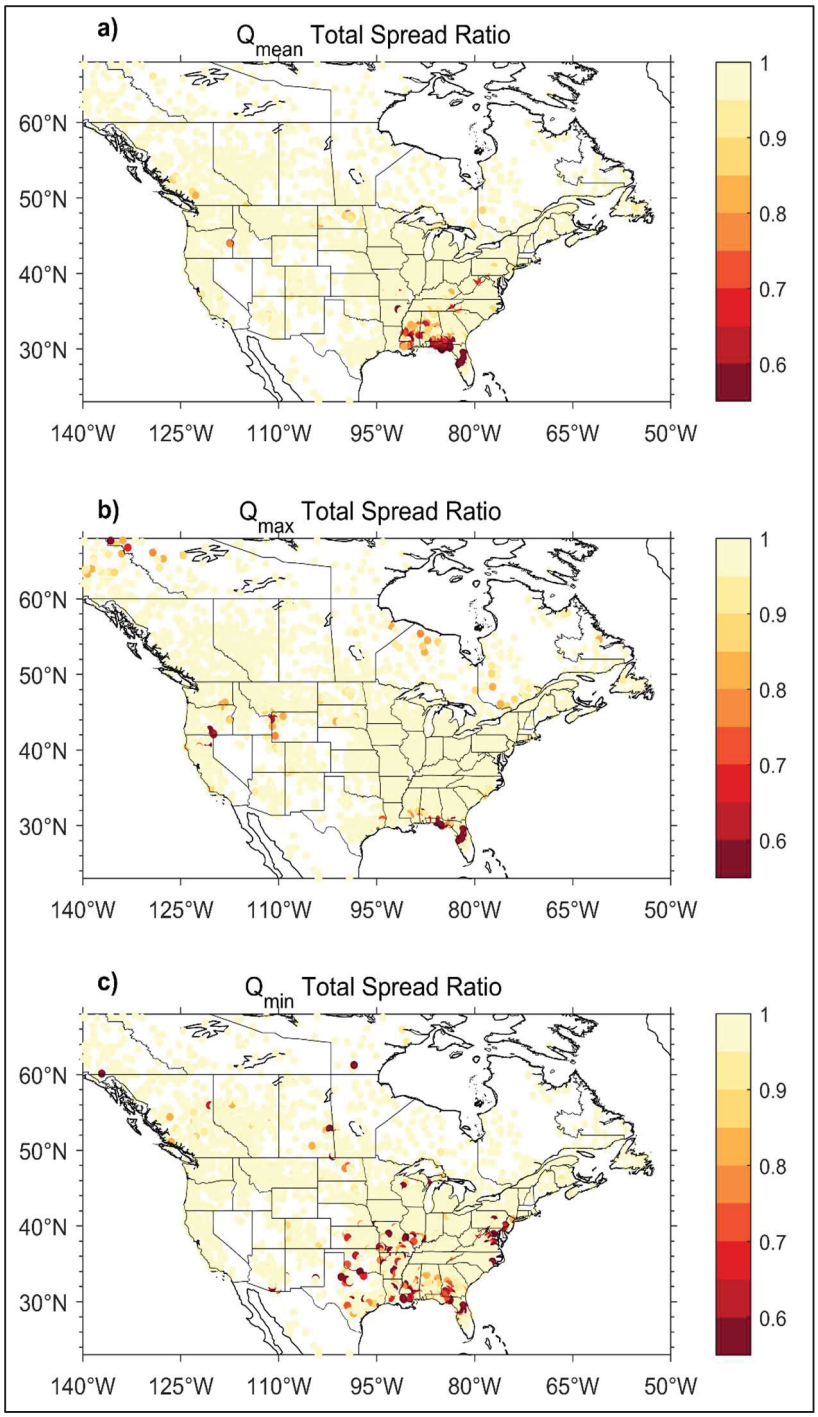

Figure S3.13 Total spread ratio for Q_{mean} (a), Q_{max} (b), and Q_{min} (c) resulting from the removal of a single climate model (NESM3)

CHAPTER 4

ASSESSING THE HYDROLOGICAL IMPACT SENSITIVITY TO CLIMATE MODEL WEIGHTING STRATEGIES

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

[a] Hydrology, Climate and Climate Change Laboratory, École de Technologie Supérieure, 1100 Notre-Dame Street West, Montreal, Quebec, Canada H3C 1K3

Article submitted to *Hydrological Sciences Journal*, September 2025

Abstract

Climate change impact studies use 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 remains unclear due to the absence of future observational data.

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. 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 improved 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 emphasizes the need for careful approaches and further research to manage uncertainties in climate change impact studies.

4.1 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. While 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), it is essential for capturing the spectrum of plausible future conditions (Hallegatte, 2009). However, this is compounded by numerous 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). A common approach for presenting results from such multi-model ensembles is by providing a best estimate along with an uncertainty range or a probabilistic distribution (Brunner et al., 2020). 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”, which assumes all models are equally capable of simulating past and future climates (Chen et al., 2017; Knutti, 2010).

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 perform worse than others (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, resulting in more confidence in the projections compared to simple averaging (e.g. Lorenz et al., 2018; Palmer et al., 2023; Yuan et al., 2020). Other studies 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 the ideal set of weights for climate simulations that considers interdependence is challenging and somewhat subjective (Herger et al., 2018), with a risk of information loss due to inappropriate weighting (Weigel et al., 2010).

In hydrological impact studies, a common method to weight or select GCMs assesses their capability to effectively depict historical climate conditions such as 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 affects streamflow projections derived from GCMs (e.g., Chen et al., 2017; Kolusu et al., 2021).

Recent impact assessment studies have utilized streamflow values to weigh simulations (Castaneda-Gonzalez et al., 2023; Dong et al., 2021; Giuntoli 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 when using raw GCM outputs to simulate streamflows, applying streamflow-based weighting schemes enhanced the reproduction of observed mean hydrographs better than using weights based on climate variables. However, the impact of weighting diminished once bias correction was applied to the GCM outputs.

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 (e.g., Chen et al., 2017). To further assess the suitability of these weights for a specific application, a calibration–validation framework can be employed, in which historical data is divided into two sets: one for calibrating the weights and the other for testing the sub-ensemble's performance (e.g. Bishop & Abramowitz, 2013). This approach is limited for the majority of regional to global climate applications due to the lack of high-quality observational data, which, depending on the region and variable, usually spans no more than 60 years (Abramowitz et al., 2019). Another limitation is the absence of future observational data, which makes it impossible to directly evaluate model performance in future scenarios. Thus, most studies on the efficiency of climate model weighting for future streamflow projections focus on whether unequal weights produce different future projections (e.g. Lorenz et al., 2018). While improving the skill scores during the reference period is important (Eyring et al., 2019), comprehensive out-of-sample testing is crucial to validate weighting methods for future projection periods (Abramowitz et al., 2019; Herger et al., 2018). However, only few studies have explored how well these schemes perform in future scenarios for their intended application by using pseudo-reality testing (Abramowitz et al., 2019; Abramowitz & Bishop, 2015; Bishop & Abramowitz, 2013; Brunner et al., 2020; Herger et al., 2018; Knutti et al., 2017; Sanderson et al., 2017; Shin et al., 2020).

In this context, our study aims to address a critical gap by investigating how different climate model weighting strategies influence hydrological impact assessments, specifically in the context of future projections where observational data is unavailable. 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 (Abramowitz et al., 2019; Brunner et al., 2020; Herger et al., 2018; Shin et al., 2020). The remaining models are then calibrated to the pseudo-reality during the reference period, after which the ensemble's performance is evaluated for future conditions using the known projections of the "truth" member as a benchmark. By doing so, we can provide a more robust validation framework for evaluating weighting schemes in future hydrological projections, which is not possible with real-world data alone (Herger et al., 2018; Knutti et al., 2017). By comparing different weighting schemes against this pseudo-reality, researchers can infer their effectiveness for future projections (Chen et al., 2020; Hernanz et al., 2022; Mendoza Paz & Willems, 2023). This study contributes to the ongoing debate on model weighting effectiveness by offering a thorough evaluation of multiple weighting schemes, including equal and random weighting as benchmarks. By conducting multiple iterations of the pseudo-reality method across various climate variables and geographic regions, we aim to gain a nuanced understanding of the sensitivity of these schemes. Ultimately, our goal is to provide valuable insights into how climate model weighting influences hydrological impact assessments, helping to better inform adaptation and mitigation strategies.

4.2 Materials and Methods

4.2.1 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 to avoid flashy

catchments due to the daily scale of data, and at least 10 years of data availability, as dictated by the requirements of the hydrological models used. While no upper area limit was imposed, 97.5% of the selected catchments had areas smaller than 11,000 km², with the largest being 650,000 km². This is comparable to the range suggested by Giuntoli et al. (2015) where catchments are selected to be of comparable size to the grid cell resolution of the global models, and our results show no significant effect of catchment size on the regional consistency of the hydrological simulations. The spatial distribution of these catchments is illustrated in Figure 4.1. Additionally, the meteorological data required for our study were obtained from the ERA5 reanalysis dataset. While ERA5 performance varies by variable and region, it has been shown to provide reliable precipitation estimates in extratropical regions, which include much of our study area (Lavers et al., 2022). This dataset has been demonstrated to perform as good as using observational data in hydrological modelling over most of the USA, without the problems related to missing data, thus ensuring complete temporal coverage (Tarek et al., 2020).

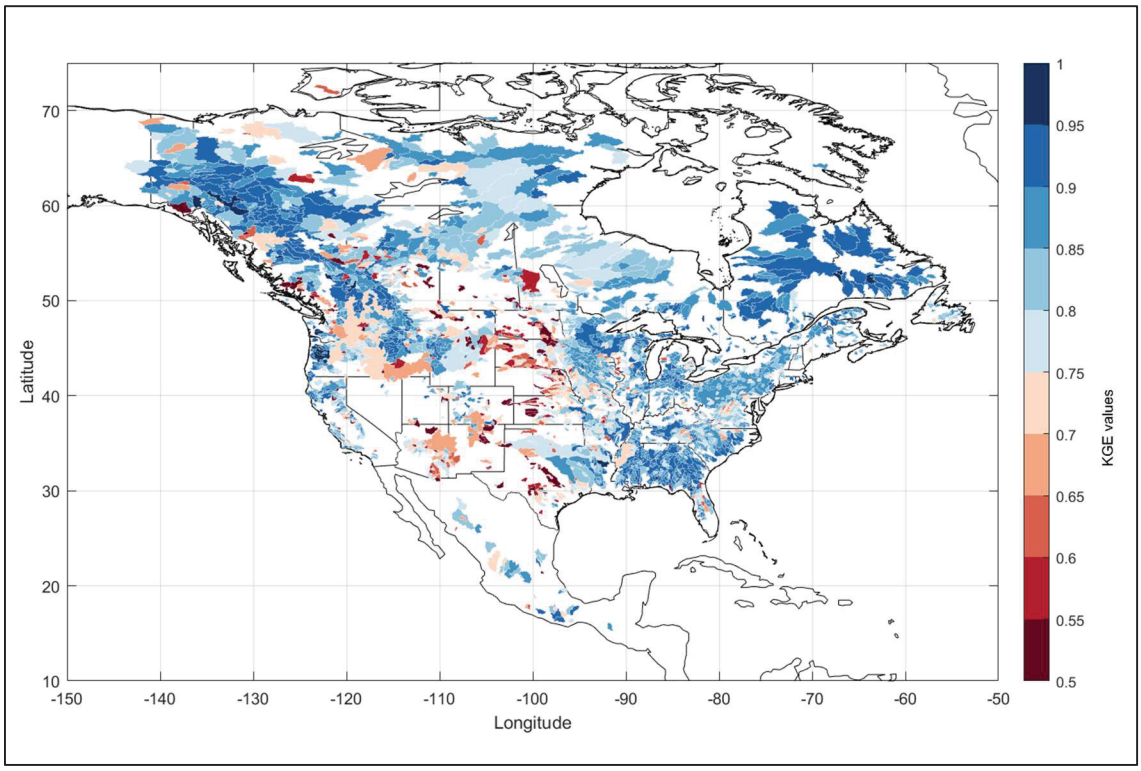

Figure 4.1 Map of the 3,107 catchments used in this study. The color code represents the hydrological model Kling–Gupta efficiency (KGE) calibration score over the reference period. In the case of nested catchments, the smaller ones were plotted on top of larger catchments.

4.2.2 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 4.2 displays the projected changes in temperature (difference between the

future and reference periods) and precipitation (change ratio, calculated as (future P - reference P) / reference P) 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. It has been suggested that GCMs with higher ECS values may present less realistic or less probable future scenarios (Hausfather et al., 2022). Consequently, the exclusion (Rahimpour Asenjan et al., 2023) or down-weighting (Massoud et al., 2023) of these models could be considered, making ECS a critical factor in the weighting of models.

Table 4.1 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). The models are listed by their ECS values

GCM	ECS	Modeling Center
CanESM5	5.62[1]	CCCma
NESM3	4.68[1]	NUIST
IPSL-CM6A-LR	4.52[1]	IPSL
EC-Earth3-Veg	4.3[1]	EC-Earth-Consortium
EC-Earth3-CC	4.23[2]	EC-Earth-Consortium
EC-Earth3	4.2[1]	EC-Earth-Consortium
EC-Earth3-Veg-LR	4.2[2]	EC-Earth-Consortium
GFDL-CM4_gr1	3.89[2]	NOAA-GFDL
GFDL-CM4_gr2	3.89[2]	NOAA-GFDL
ACCESS-ESM1-5	3.88[1]	CSIRO
KIOST-ESM	3.36[2]	KIOST
MRI-ESM2-0	3.14[1]	MRI
MPI-ESM1-2-HR	3.02[2]	DKRZ
BCC-CSM2-MR	3.01[1]	BCC
MPI-ESM1-2-LR	2.98[2]	MPI-M
FGOALS-g3	2.87[2]	CAS

Table 4.2 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). The models are listed by their ECS values (continued)

GCM	ECS	Modeling Center
GFDL-ESM4	2.62[1]	NOAA-GFDL
NorESM2-LM	2.60[1]	NCC
MIROC6	2.57[1]	MIROC
NorESM2-MM	2.49[2]	NCC
INM-CM5-0	1.92[1]	INM
INM-CM4-8	1.83[1]	INM

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, with bias correction performed exclusively using the pseudo-reality GCM data. In this approach, pseudo-reality is treated as the true climate condition, and the GCM data is corrected based on this assumed truth. Typically, bias correction relies on observed real-world data (e.g. Mendez et al., 2020), but in this case, pseudo-reality was used to establish a hypothetical baseline (Hui et al., 2019; Maraun, 2012; Schmith et al., 2021). This allows us to apply bias correction in a controlled environment where the "true" future climate is also known, ensuring that the corrected GCM data aligns more closely with these hypothetical true conditions (Chen et al., 2020; Schmith et al., 2021). Subsequently, both the raw and bias-corrected climate data were used as inputs for a pre-calibrated hydrological model, which then generated streamflow simulations.

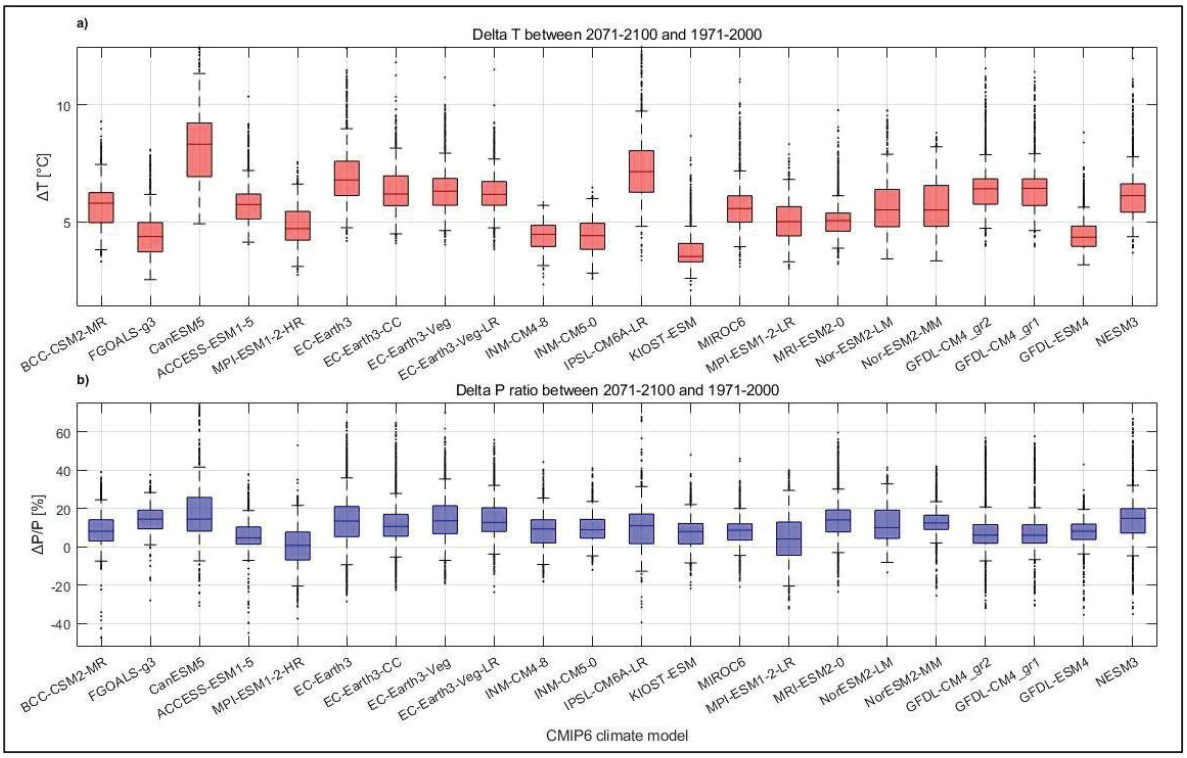

Figure 4.2 Projected temperature (a) and precipitation (b) changes between the reference (1971-2000) and future (2071-2100) periods over all 3,107 catchments for all 22 GCMs

The HMETS conceptual lumped rainfall–runoff model was used for simulating streamflow (Martel et al., 2017). The HMETS model operates on a daily time scale, with both inputs and outputs at this temporal resolution, and 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. The studied watersheds have a minimum KGE value of 0.5, indicating a satisfactory performance of the hydrological model (see Figure 4.1).

4.2.3 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 4.2. These methods were selected based on recent literature to ensure a comprehensive evaluation of different criteria, including GCM performance, model independence, and the inclusion of random and equal weighting for comparison. These methods are described below.

Table 4.3 Weighing methods used in this study

Method	Description	References
RAC	Evaluating how closely models match pseudo reality 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
Equal	Weights are assigned equally	
Random	Weights are assigned randomly to models for benchmarking and comparison	

4.2.3.1 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 pseudo reality and simulated series, with the maximum correlation (r_0) set to 1 for this study. To apply this analysis, we aggregate the daily data to a monthly temporal scale. This aggregation is necessary because the RAC method is designed to capture larger-scale seasonal and annual patterns. Additionally, the parameter $\sigma = \sigma_s / \sigma_o$ represents the ratio between the standard deviations of the monthly simulated series and the monthly pseudo reality series. The RAC method aims to quantify the degree of resemblance between the simulated and pseudo reality annual cycles (Wang et al., 2019).

$$RAC_i = \frac{4\,(1+r)^4}{(\sigma\ + 1/\sigma)^2(1+r_0)^4} \tag{4.1}$$

4.2.3.2 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 the pseudo reality in the reference period, 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). The convergence criterion assumes that models that closely follow the collective behavior of the ensemble are more reliable. However, we recognize that this assumption may overlook the potential value of outlier models, which could offer important information in certain cases.

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 4.2. 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 $\varepsilon/|\beta_i|$ or $\varepsilon/|D_i|$ 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.

$$REA_i = \{[\frac{\epsilon}{abs\ (\beta_i)}]^m \times [\frac{\epsilon}{abs\ (D_i)}]^n\}^{1/mn} \tag{4.2}$$

4.2.3.3 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 pseudo reality data receive higher weights, thus having a greater influence within the ensemble. The weights, $W_{m,skill}(i)$, are calculated according to equation 4.3 (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 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.

$$W_{m,skill}(i) = e^{-(\frac{\delta_i}{D_q})^2} \tag{4.3}$$

4.2.3.4 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.

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

4.2.4 Experiment Design

The main methodological steps are depicted in Figure 4.3. Specifically, Figure 4.3-a illustrates the steps for evaluating the performance of each weighting method for both future precipitation and temperature, while Figure 4.3-b 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. Using a larger number of simulations allows us to better differentiate between structural differences and internal variability, an issue that earlier studies with fewer simulations struggled to address (Deser et al., 2020). The steps in Figure 4.3 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.

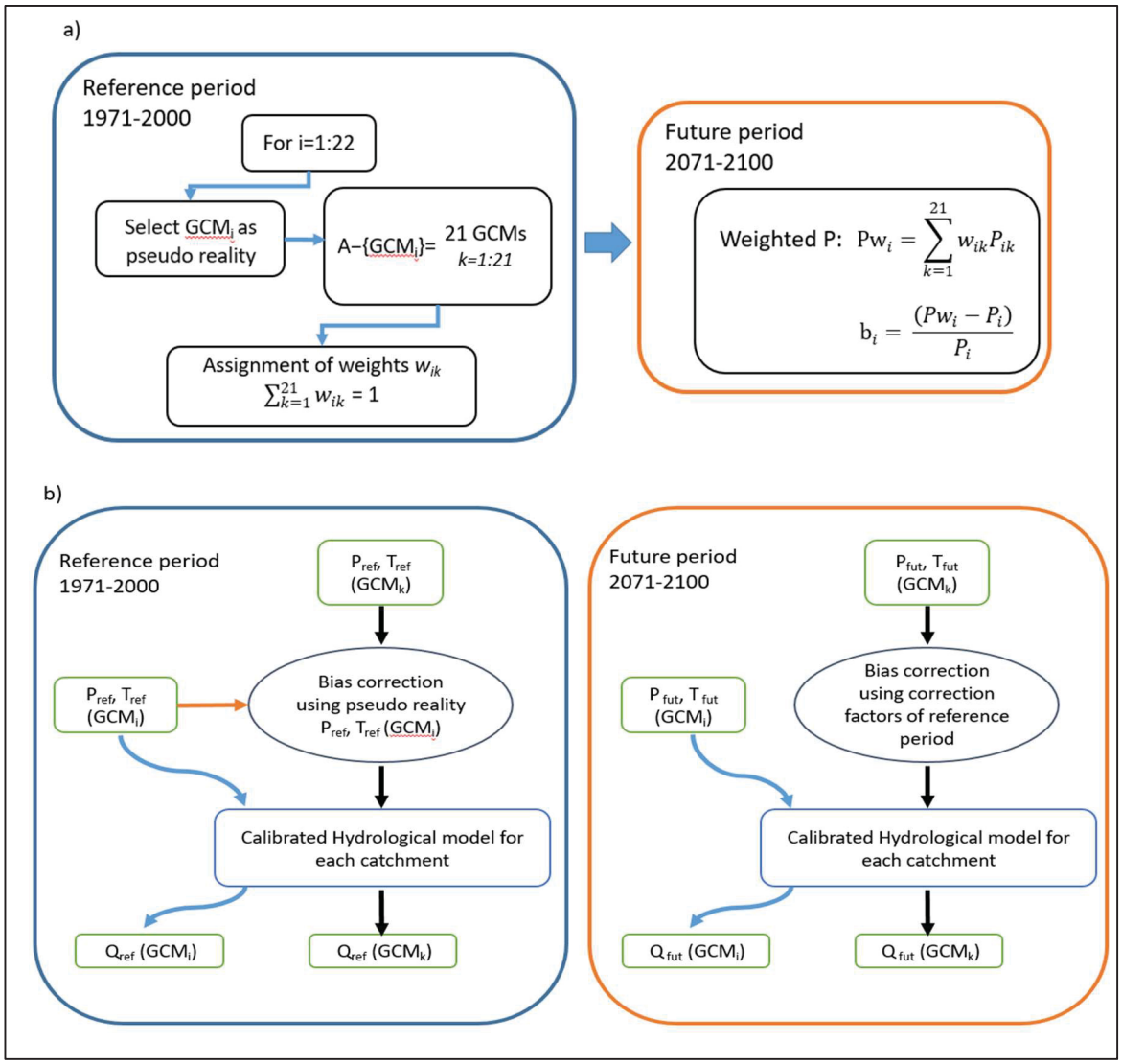

Figure 4.3 Main methodological steps (a) for the evaluation of the performance of each weighing method for precipitation (shown as P) and temperature (not shown). A = {GCM1, GCM2, GCM3, …, GCM22}, and bias = median {b1, b2, b3, …, b22}. Additional methodological (b) steps for the evaluation of the performance of each weighing method for streamflow metrics

To assess the impact of weighting on streamflow values, it is necessary to include an additional step of bias correction, as detailed in Figure 4.3-b. 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). Due to the inherent limitations of climate models, using uncorrected simulations often leads to systematic discrepancies when compared to projections that have undergone bias correction (Dinh & Aires, 2023; Ehret et al., 2012; Räty et al., 2014). However, it is important to acknowledge that bias correction introduces additional uncertainty into the modeling process, as it may reduce inter-GCM variability and obscure some of the original characteristics of the climate models (Ehret et al., 2012).

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. This adjustment allows for the computation of streamflow values using the bias-corrected P and T with the calibrated hydrological model. In this study, we matched the GCM meteorological forcing to the lumped hydrological model by using the mean of all included grid points within the watershed or, if none were available, the closest grid point to ensure accurate simulations. It is important to note that the hydrological model is calibrated using observation data. While the absolute performance of the hydrological model is important, our primary focus remains on effectively representing the key underlying hydrological processes. As long as these processes are reasonably represented, the hydrologic model's absolute performance may not be of critical concern. The input data for the hydrological model comprises GCM data, which has been bias-corrected against the pseudo-reality, which serves as the hypothetical truth in our study. Crucially, the pseudo-reality is not intended for comparison with real-world observations; instead, it acts as a controlled framework to evaluate different climate model weighting strategies in future projections where observational data is unavailable. The comparison is performed on climatological statistics (e.g., interannual means,

long-term distributions) rather than on day-to-day correspondence, since individual daily sequences from the GCMs do not, and are not expected to, match the pseudo-reality.

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. Since, after bias correction, the streamflow characteristics of the 21 GCMs should align closely with the pseudo-reality GCM, the different weighting methods are not expected to result in a weighted average that significantly deviates from the pseudo-reality. However, 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 4.3-b.

4.3 Results

4.3.1 Climate Variable Sensitivity to Weighting Methods

Figure 4.4 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 4.3. A bias of 0 for Equal Weighting (Figure 4.4-

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. The metric used in Figures 4.4b-f is the difference in the absolute values of bias, meaning that the absolute values of the equal weighting method biases and those of the tested methods are first computed before taking the difference. This means the initial direction of the bias (positive or negative) is not considered in this calculation. This approach helps us discern the deviation of each method from Equal Weighting, aiding in understanding their relative effectiveness. Supplementary material Figures S4.12 and S4.13 show the median bias for each method.

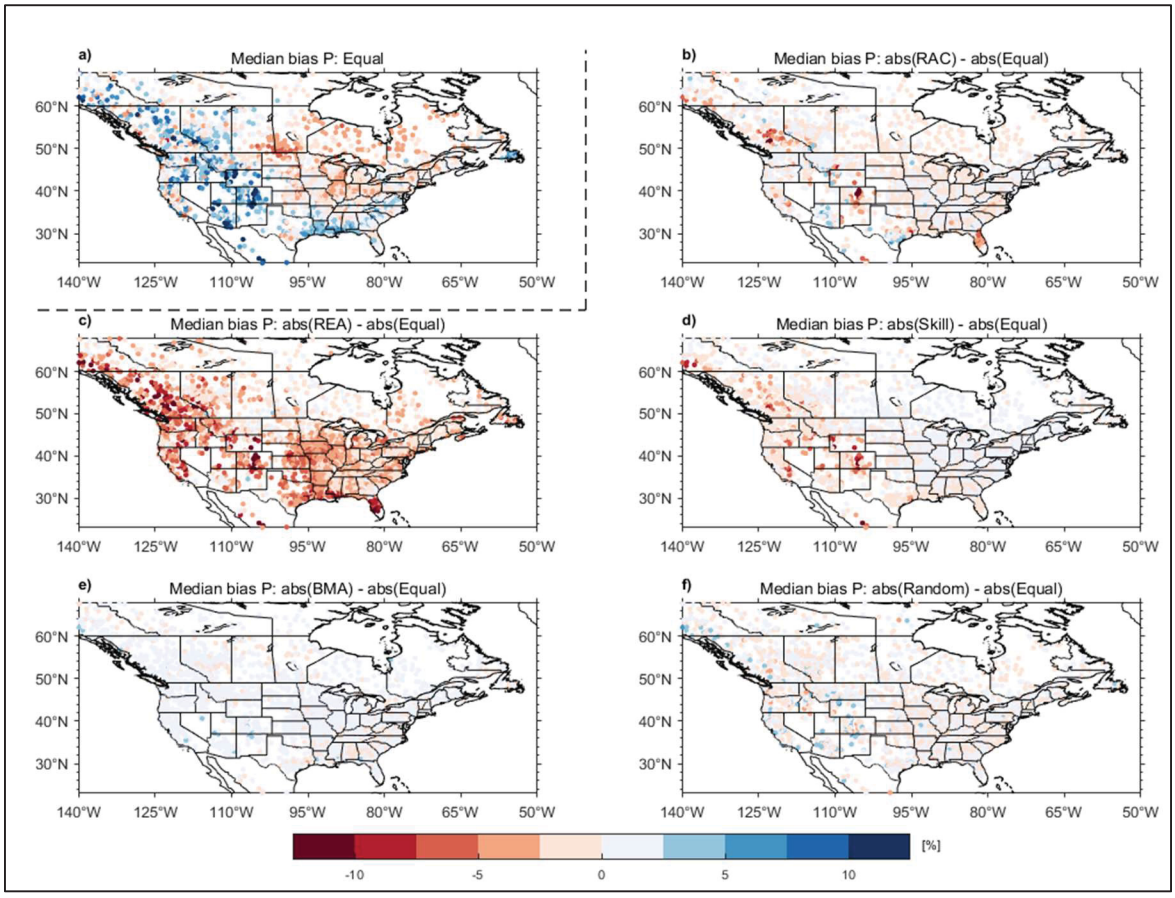

Figure 4.4 Difference in median absolute precipitation (prcptot) bias across all catchments for the future period (2071-2100). Equal weighting (a) is presented as the actual bias value, while the biases from all other methods (b-f) are expressed as differences between the

absolute values of the tested method bias and the absolute value of the equal weighting method 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. Overall, these findings emphasize the need to account for regional variations when evaluating the effectiveness of different weighting schemes.

Looking at the median precipitation with equal weights (Figure 4.4-a), the western and southeastern catchments display positive biases, while other regions exhibit negative biases. This pattern suggests regional differences in model behavior. For instance, areas with negative biases (in red) are predominantly continental climates (Dfa, Dfb, Dfc, Cfa in the Köppen classification), while regions with positive biases tend to have maritime or mountainous climates. Similar discrepancies were observed in ERA5 precipitation biases (Tarek et al., 2020), where precipitation was negatively biased in these same zones. This could point to potential shortcomings in the climate models' ability to accurately capture certain climatic conditions, particularly in regions influenced by maritime or orographic effects. A full analysis of bias distribution across all 22 models may reveal if specific GCMs disproportionately affect the median bias, but this is beyond the scope of the paper and will be left for future regional studies.

Figure 4.5 presents results for mean annual temperature (tas) using the same format as Figure 4.4. 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 4.5, the overall median absolute biases remain small, always less than 0.25 (less than 25% of the original value) for all catchments.

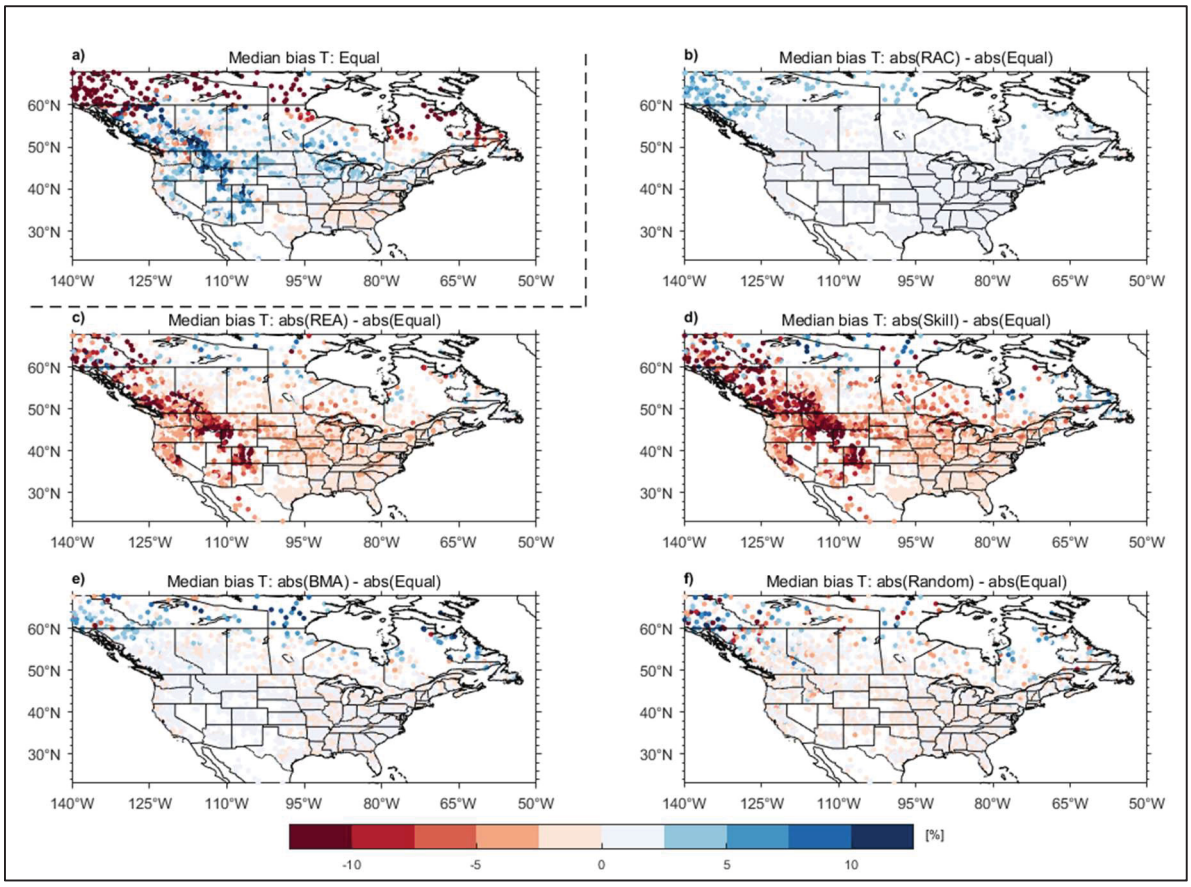

Figure 4.5 Same as Figure 4.4, but for mean annual temperature (tas)

Figure 4.6 presents the median bias for mean annual streamflow (Qm) in the same format as Figures 4.4 and 4.5. 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 of the maximum and minimum annual discharge values, as shown in supplementary material Figures S4.14 and S4.15.

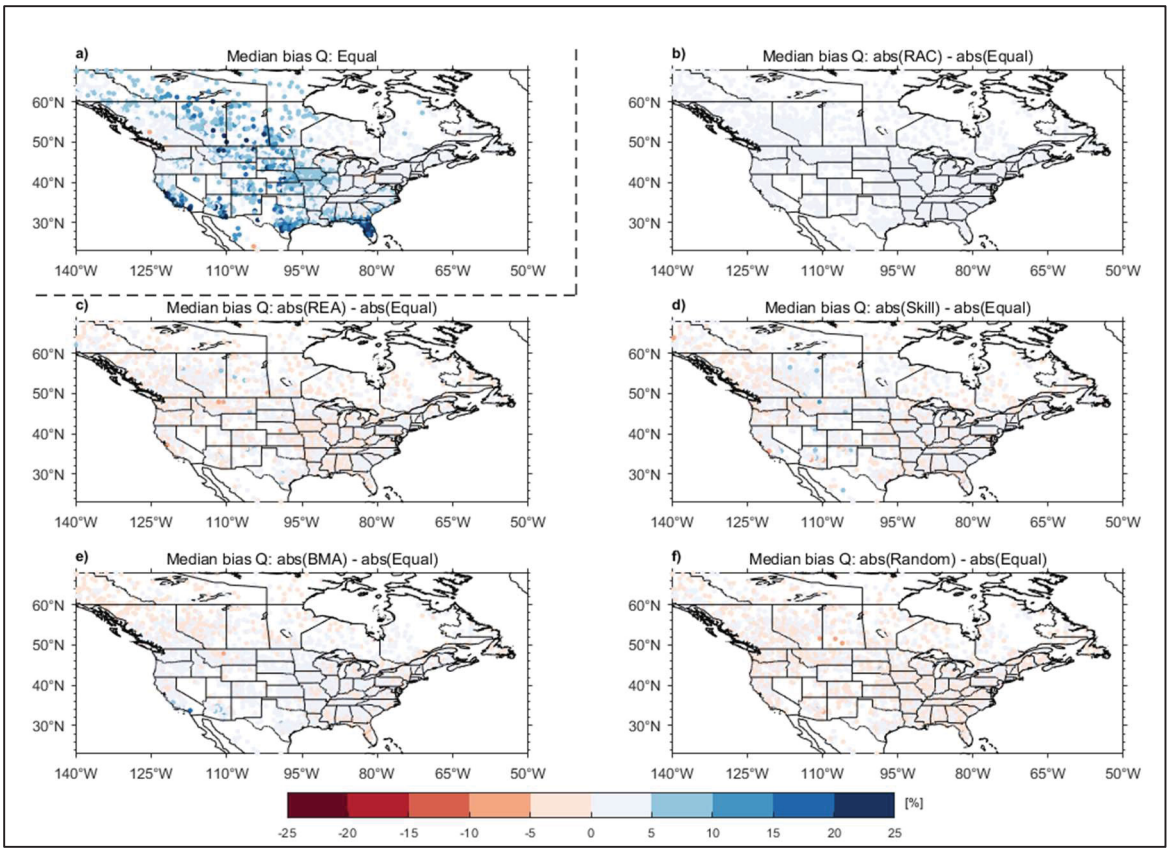

Figure 4.6 Same as Figure 4.4, but for mean annual streamflow (Q_m)

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 of each weighting method, Figure 4.7 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 4.6, 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 S4.14 and S4.15). 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.

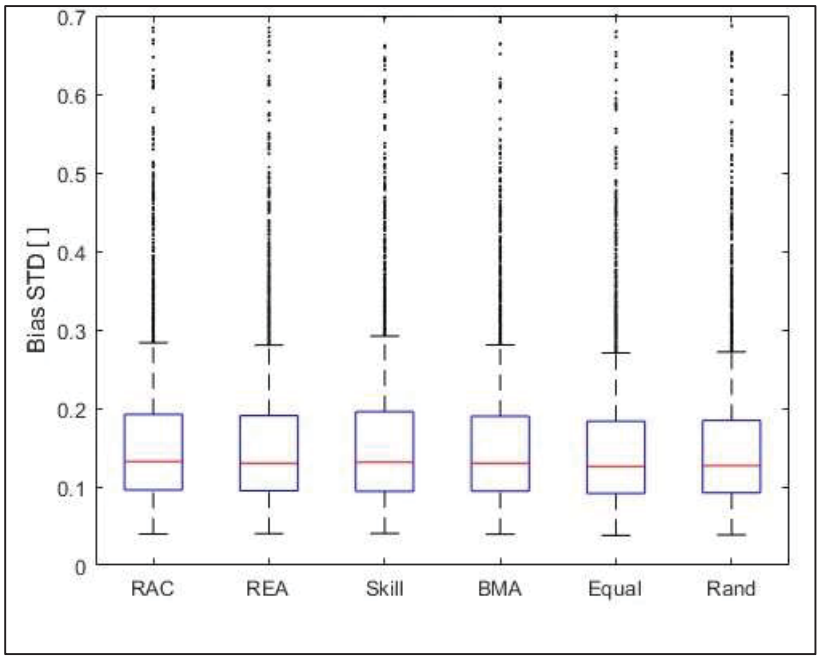

Figure 4.7 Boxplot for the standard deviation of the distribution of the 22 bias values of mean annual streamflow (Q_m)

Results from Figures 4.6 and 4.7 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 4.4 and 4.5).

4.3.2 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:

1. Weighting based on climate variables: In this approach, model weights were derived from the raw (non–bias-corrected) precipitation and temperature data, as in Figure 4.6. These weights were then used to combine the corresponding uncorrected streamflow simulations.

2. Weighting based on streamflow simulations: In this approach, the weights were computed directly from streamflow outputs simulated with raw precipitation and temperature over the reference period, rather than from the climate variables themselves.

This distinction allows assessing whether deriving weights from meteorological variables or from hydrological responses leads to different outcomes when no bias correction is applied.

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 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 4.8 (for the first approach) and Figure 4.9 (for the second approach), both of which illustrate the median bias for the mean annual streamflow discharge in the same format as Figure 4.6. 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.

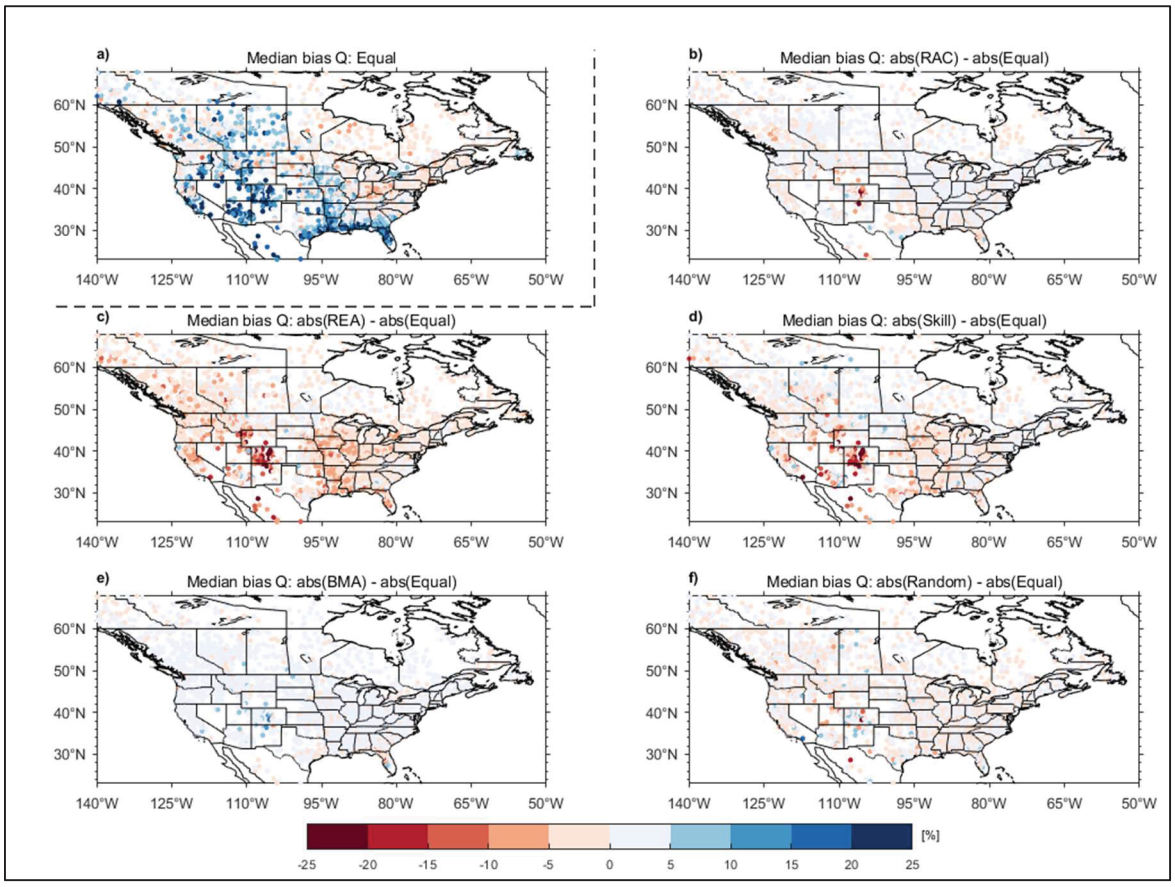

Figure 4.8 Same as Figure 4.4, but for mean annual streamflow (Q_m) and using the first approach.

In both experiments, the biases are considerably larger than those observed in Figure 4.6. 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.

A slight improvement is observed when weighting is based on streamflow performance rather than precipitation and temperature. 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.

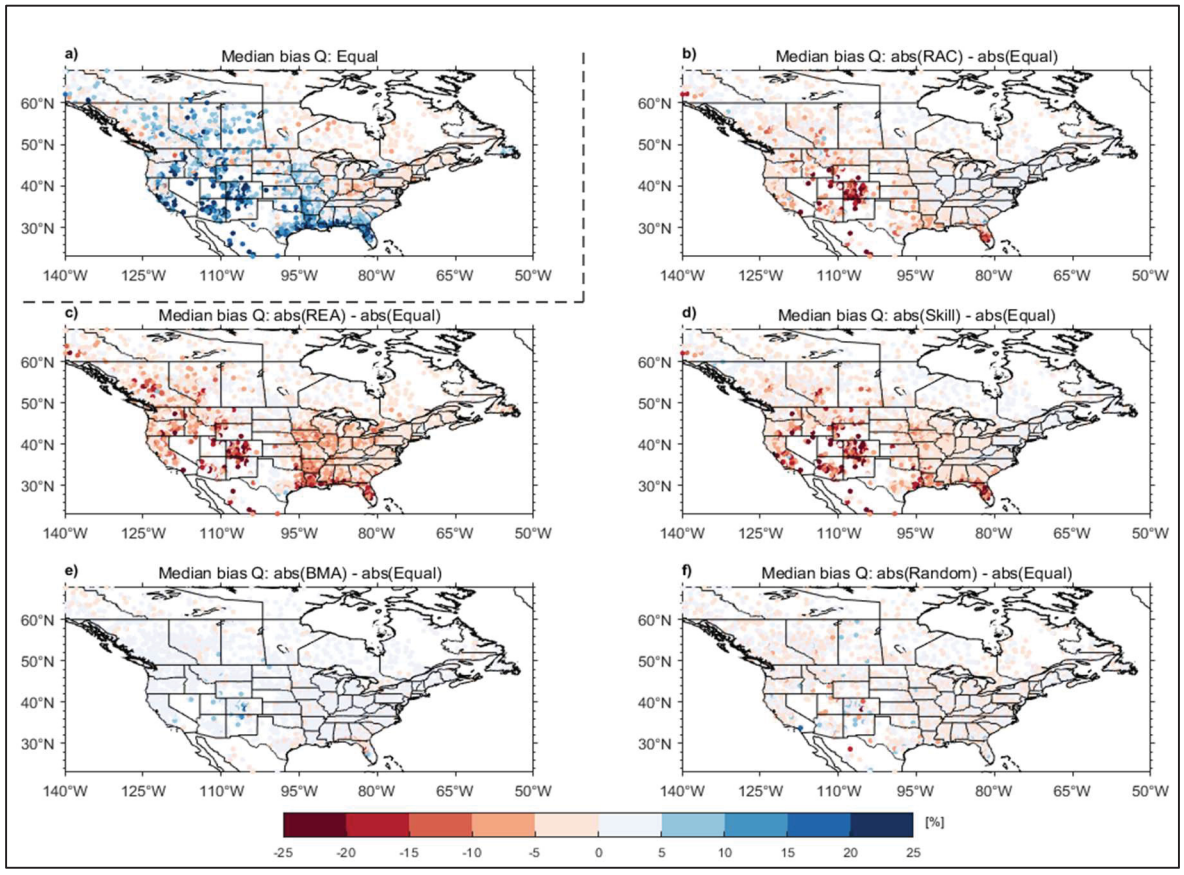

Figure 4.9 Same as Figure 4.4, but for mean annual streamflow (Qm) and using the second approach

4.4 Discussion

4.4.1 Evaluation of Weighting Methods in Hydrological Impact Studies

Assigning weights to climate model projections can be subjective and introduces additional uncertainty into impact analysis, making the selection of an appropriate weighting method a challenging task (Knutti, Furrer, et al., 2010). There is considerable debate over the best

approach to weighing climate models in impact assessment studies. A key concern is that weights are often based on past performance, which may not translate to future conditions (Hui et al., 2019). Model weighting is inherently complex and requires a comprehensive assessment of the uncertainties involved (Abramowitz et al., 2019; Brunner et al., 2020). Moreover, performance metrics are subjective and vary depending on the parameters chosen for evaluation (Sanderson et al., 2015). It is worth noting that relying solely on outputs like temperature and precipitation for weighting may fail to capture the intricate relationship between climate variables and hydrological responses, potentially limiting the models' effectiveness in representing hydrological changes (Wang et al., 2019; Wootten et al., 2023).

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 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.

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 4.10. This observation serves to underscore the effectiveness of the REA method.

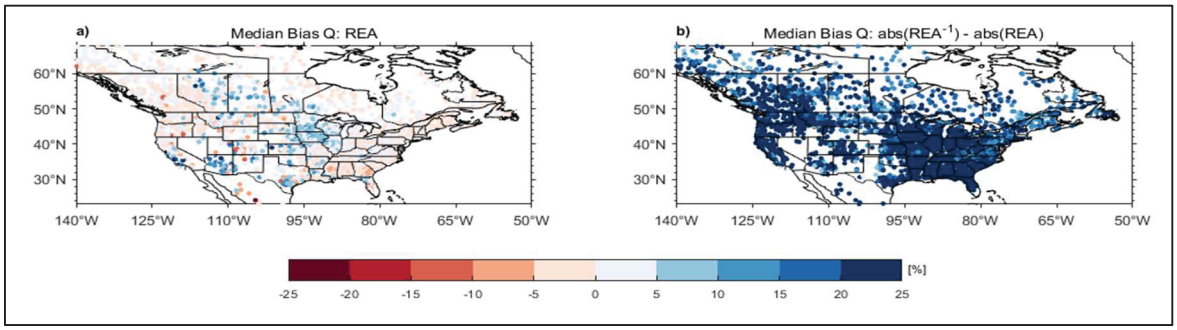

Figure 4.10 Same as figure 4.9 with a) REA and b) inverted REA weights

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 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).

4.4.2 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 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 weighting schemes, as well as the availability of reliable observational data, as noted by Singh & AchutaRao (2020). These considerations are crucial for improving the robustness of future change estimates and the uncertainties associated with them. 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.

4.4.3 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 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 4.11, demonstrate that using random weights in this smaller ensemble performed worse than equal weighting, as shown by the darker blue colors compared to Figure 4.9-b. This supports the previously mentioned hypothesis.

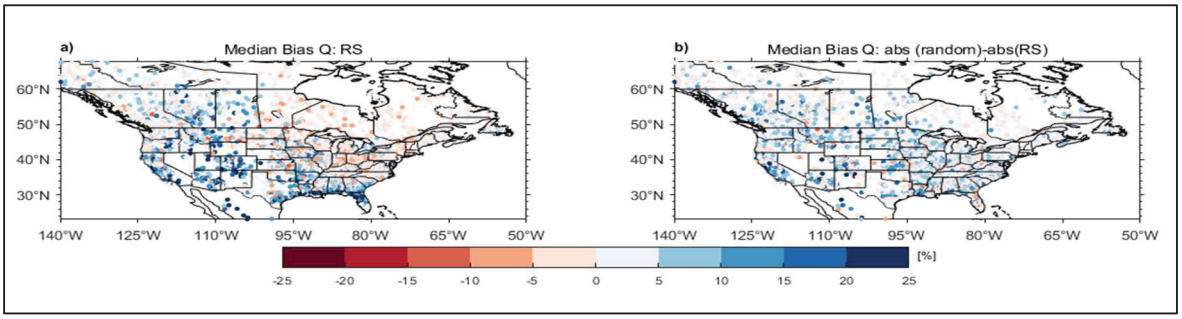

Figure 4.11 Similar to figure 4.9, comparing two scenarios: a) Using 7 randomly selected and equally weighted GCMs, and b) the difference in median streamflow bias when using 7 randomly selected GCMs with random and equal weights.

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 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.

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

In this study, we used each of the 22 GCMs as the pseudo-reality target in turn, an important methodological step to account for the potential impact of selecting a model with either low or high sensitivity. As a result, the median findings provide a robust estimate of the expected performance of each weighting method. The underlying hypothesis of using pseudo-reality is that if a weighting scheme can accurately replicate the pseudo-reality scenario, it is likely to be effective in projecting future climate impacts. By utilizing multiple pseudo-reality scenarios, we simulate the range of uncertainties inherent in climate projections, helping to identify weighting schemes that are consistently reliable across different conditions.

However, the use of pseudo-reality comes with several limitations. First, pseudo-reality is a hypothetical construct, and while it mimics future climate conditions, it does not represent

actual observations. Without future observational data, it is impossible to verify how well the pseudo-reality reflects real-world climate outcomes, introducing further uncertainty, especially when making long-term projections. Another limitation is that, as with any method operating in a 'climate model world,' the model-as-truth approach may oversimplify complex real-world processes, potentially overlooking important factors that influence climate impacts. Pseudo-reality may also fail to fully capture real-world extremes, resulting in an incomplete assessment of model performance in predicting such events. Moreover, without future observations, the results remain theoretical despite the methodological advantages. Despite these limitations, pseudo-reality remains a valuable tool for evaluating model weighting strategies when applied cautiously and in conjunction with other methods. It provides important insights into model performance and uncertainty, helping to enhance the robustness of climate projections across diverse scenarios.

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 4.3-b 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). The two lumped conceptual hydrological used, while effective in simulating general hydrological processes, may not fully capture the complexity of spatially distributed processes or account for the detailed physical interactions at the sub-basin scale. Therefore, future work could benefit from incorporating more diverse hydrological models, including physically-based or distributed models, to provide a broader evaluation of the impacts of climate model weighting on hydrological simulations.

4.5 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.

4.6 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 used codes are available via contacting the authors.

4.7 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.

4.8 Acknowledgments

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

4.9 Supplementary material

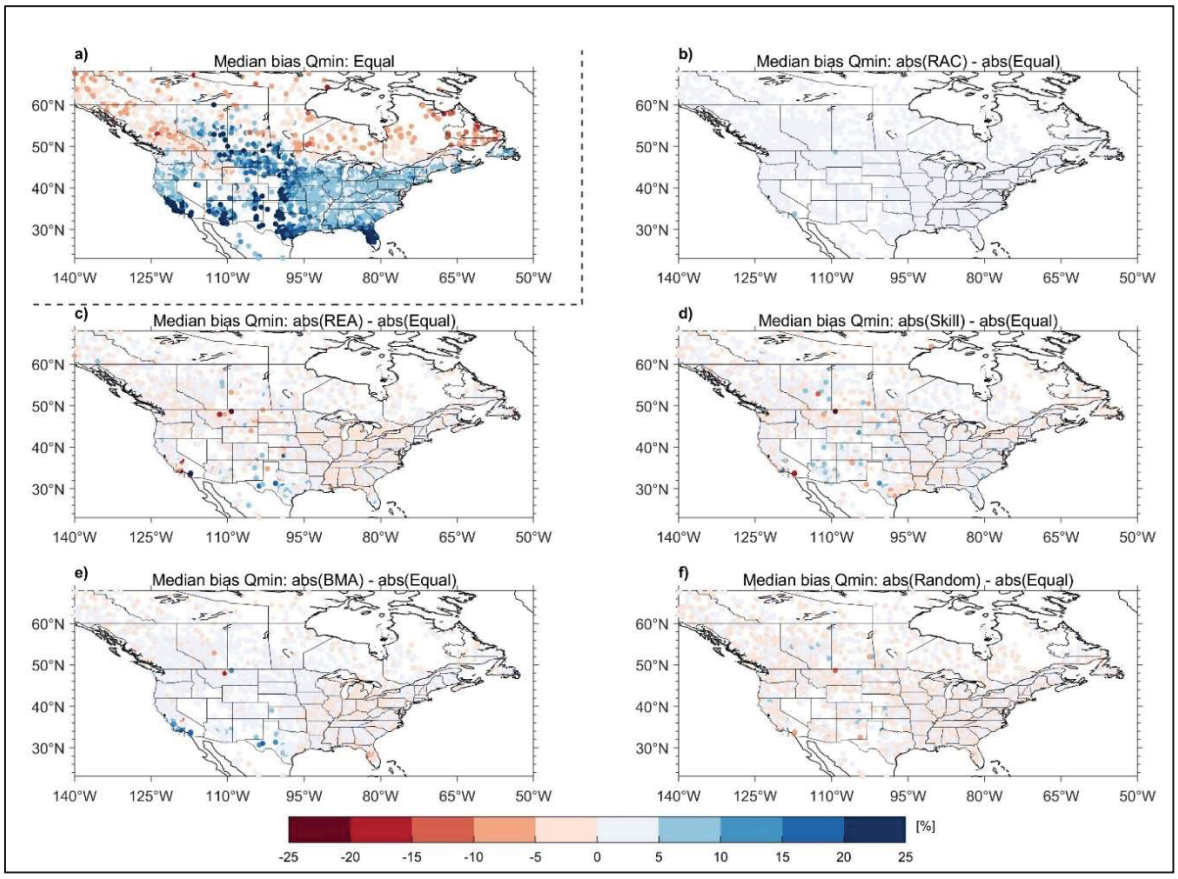

Figure S4.12 Similar to Figure 4.4 but median bias is plotted.

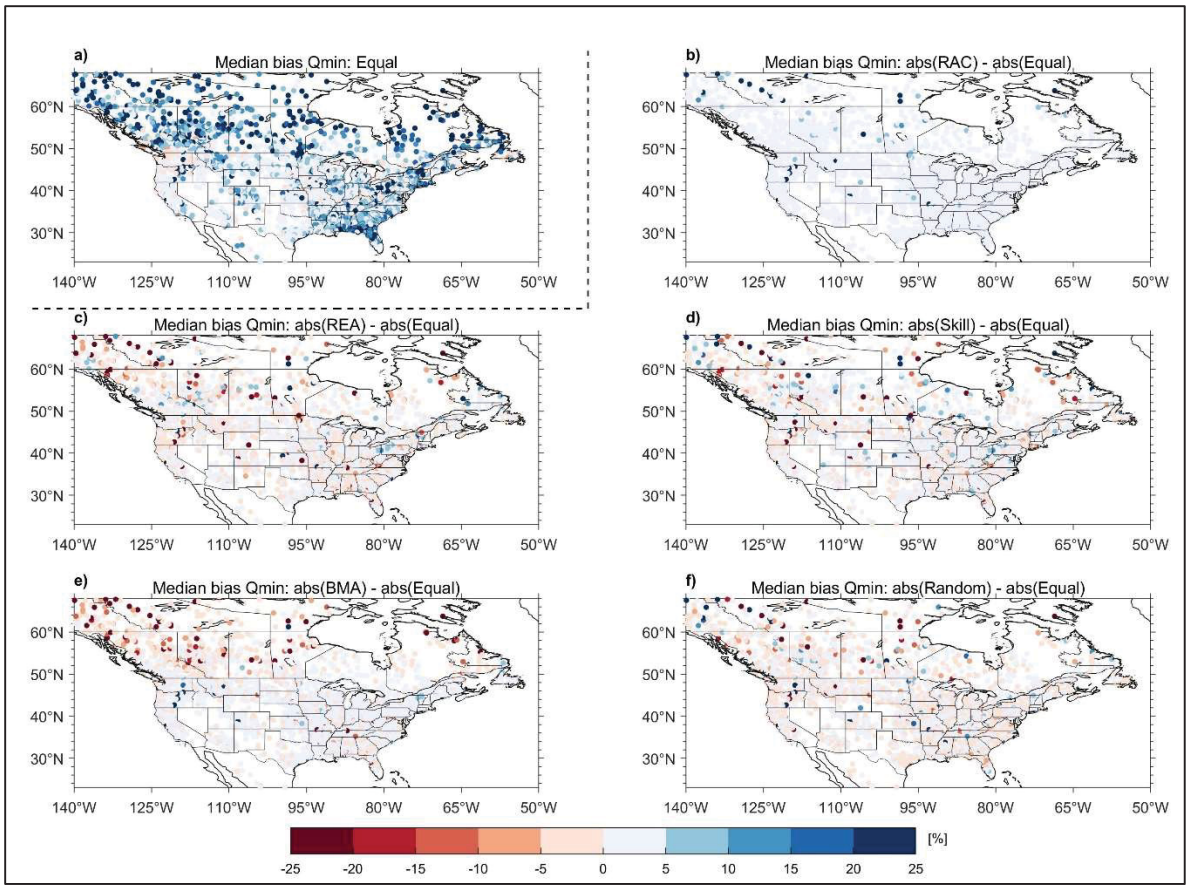

Figure S4.13 Similar to Figure 4.5, but bias is plotted as negative and positive values

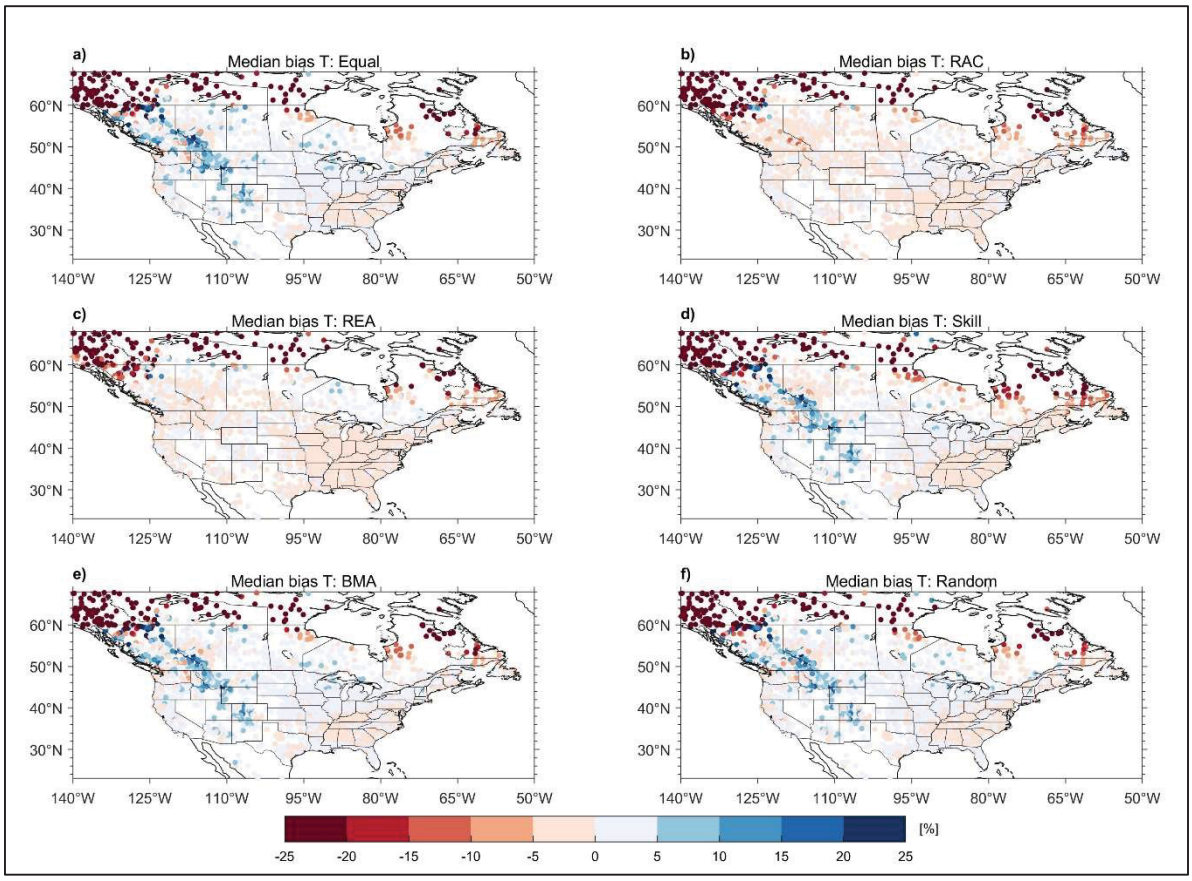

Figure S4.14 Similar to Figure 4.6 but for minimum streamflow (Q_{min}).

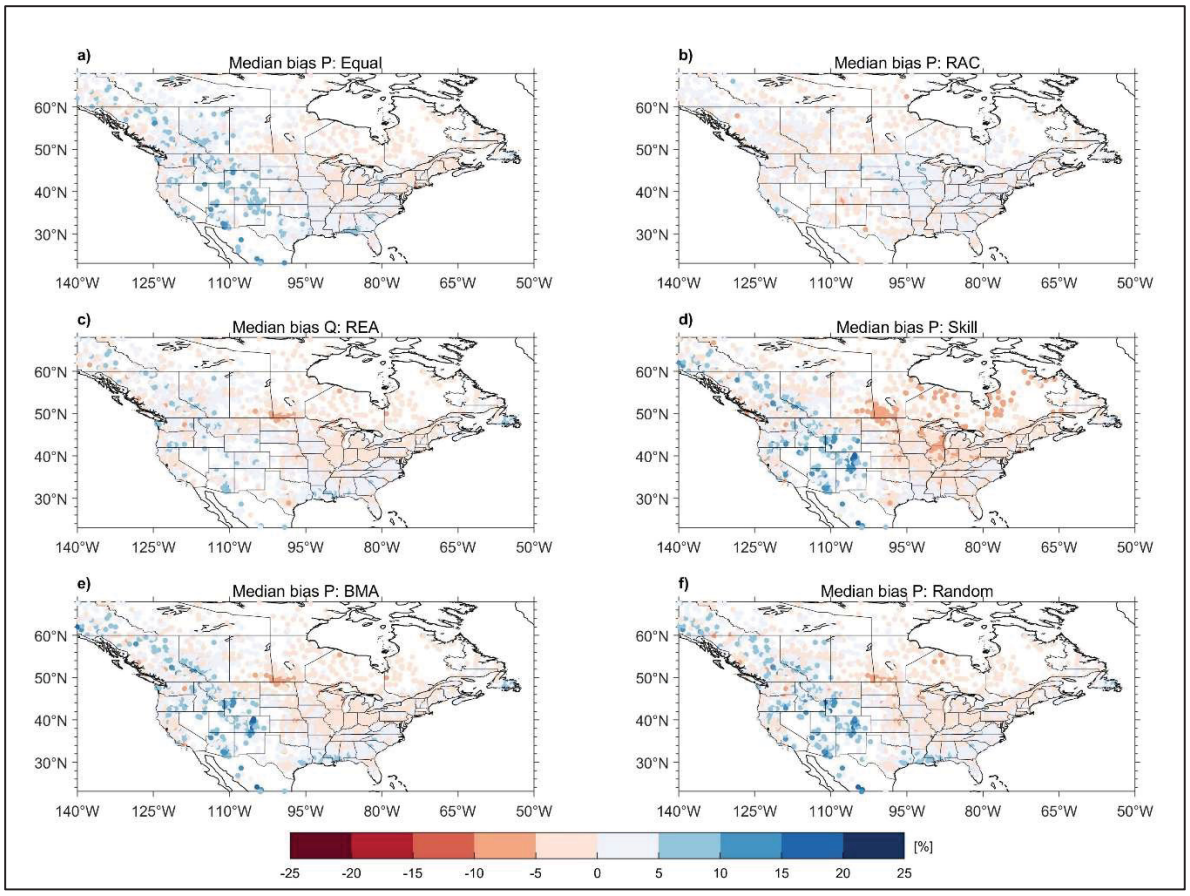

Figure S4.15 Similar to Figure 4.6 but for maximum streamflow (Q_{max}).

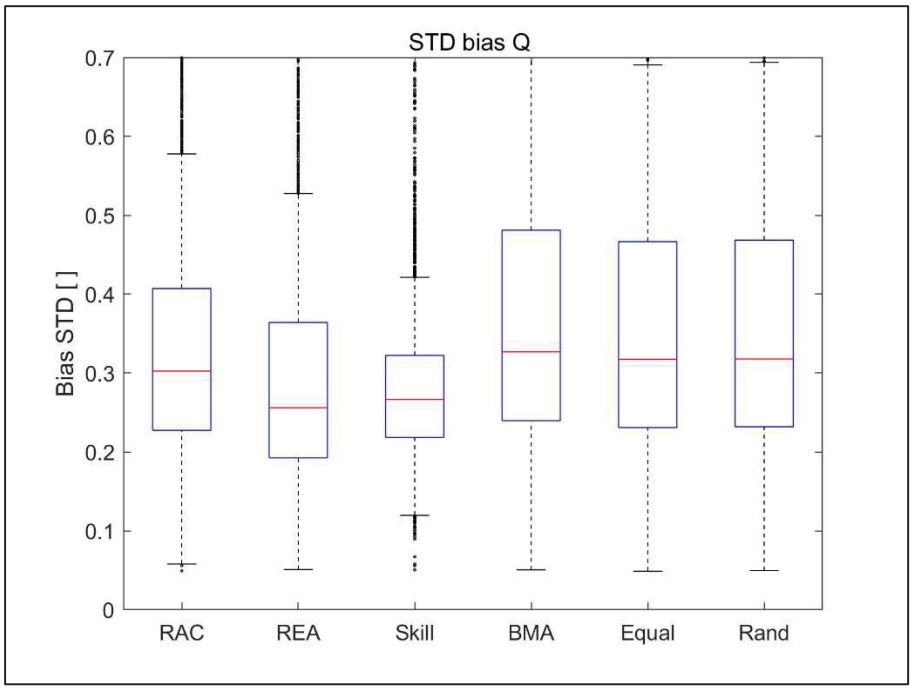

Figure S4.16 Standard deviation of streamflow bias from weighting applied to streamflow simulated with raw precipitation and temperature (no bias correction), using the HMETS hydrological model.

CHAPTER 5

GENERAL DISCUSSION

This thesis explored the challenges and implications associated with climate model selection and their combination methods in hydrological climate change impact assessments. Through three complementary studies, it investigated different aspects of uncertainty of future streamflow projections over a comprehensive sample of North American catchments. As a whole, these studies provide an integrative description of how methodological decisions regarding model selection, weighting, and bias correction influence hydrological projections. This chapter discusses the key findings, reflects on their implications, and offers recommendations for future research and practice. It is noteworthy to mention that the objective of this thesis is not to point out a single method or a single set of models that is practical for all impact studies, as, arguably, such a method or subset has not been developed yet (Kolusu et al., 2021), but to provide overall strategies that can be used to guide the choice of models.

5.1 Ensemble Design and Uncertainty Transfer in Hydrological Impact Modeling

In climate change impact assessment studies, multi-model ensembles have become a standard approach to account for the uncertainties inherent in projecting future climate (Bellucci et al., 2015; Grose et al., 2023; Maher et al., 2021; Semenov & Stratonovitch, 2010; Tebaldi & Knutti, 2007). By combining outputs from multiple GCMs, multi-model ensembles are able to capture a broad spectrum of potential futures while enabling the assessment of uncertainty contributions from individual modeling components.

Of the many links in the chain of climate-impact modeling; starting with GCMs, through bias correction, and hydrological modeling; GCMs have most frequently been cited as the main source of uncertainty for variables such as flood extremes and flood risk (Gao et al., 2020; Shen et al., 2018; H. Wang et al., 2020). The results presented in this thesis also showcase the central role of GCMs in introducing uncertainty into hydrological projections. However, their

relative contribution to uncertainty is not consistent through different regions and hydrological flow regimes. This observation aligns with previous research that has shown that uncertainty contributions of GCMs differ with the dominant hydroclimatic process of a basin, seasonality, and structural characteristics of the hydrological model employed (e.g. Castaneda-Gonzalez et al., 2022; Meresa et al., 2021). For instance, our analysis revealed that hydrological models made larger contributions to uncertainty in dry regions, and for low-flow and high-flow regimes. This larger contribution is presumably because the models have a limited capacity to represent conditions in arid areas and during peak events effectively. On the other hand, Troin et al. (2022) showed that in cold and snow-dominated catchments, uncertainty was largely caused by hydrological models, mainly due to the fact that it was challenging for the models to adequately simulate snow accumulation and melt processes.

The thesis further examined how uncertainty originating from climate models transfers into hydrological projections. While larger ensembles naturally represent a wider diversity of possible futures, they also increase the computational, analytical, and interpretative burden—particularly when downscaling, bias correction, and hydrological modeling must be applied consistently across thousands of catchments. Even with modern computational resources, the practical challenge often lies not only in running the simulations, but in managing, comparing, and interpreting the resulting volume of data in a transparent and reproducible way. This work demonstrated that by appropriate design, reduced ensembles can still preserve a large part of the original hydrological uncertainty envelope. In particular, the KKZ algorithm performed well in preserving spread while reducing ensemble size, offering a good compromise between representational accuracy and computational efficiency.

Uncertainty transfer from climate world to hydrological responses is, however, non-uniform and nonlinear. For instance, small changes in precipitation projections, which is often the most uncertain climate projection, can result in considerably large differences in simulated streamflow, particularly during low-flow and high-flow conditions. These nonlinear responses critically depend on the structure of the hydrological model, the characteristics of the catchment, and the design of the climate model ensemble. As a result, variability in climate

space does not translate directly into variability in impact space, underscoring the importance of incorporating hydrological sensitivities into ensemble construction.

This leads to a broader methodological insight that ensemble selection methods cannot be based solely on climate performance metrics (e.g., temperature or precipitation bias or RMSE). Otherwise, they will potentially overlook the behavior of impact models in simulating the variables of concern, such as streamflow. For example, a moderately warm-biased GCM can project winter precipitation as rain instead of snow and produce greatly underestimated spring runoff even if other "acceptable" temperature performance is realized. Therefore, it is necessary that impact-based criteria be applied during ensemble selection in order to remain applicable in the final ensemble for the intended application, including hydrological sensitivity, seasonal character, and representation of regional processes.

No single selection or weighting strategy performs optimally across all variables, regions, or hydrological metrics, indeed, this is one of the central findings of this thesis. While model selection and weighting are both feasible and widely used, they are inherently constrained by trade-offs and methodological choices. Ultimately, ensemble design is not only a technical process but an intellectual bridge between climate modelling and impact assessment. Subset selection strategies involve trade-offs and choices regarding what model properties or performance statistics are more important. Rather than searching for an optimal "best" subset, a better approach is to have a robust and logical set of choices that are appropriate to the context of the research. Through the development of methods that simultaneously respect uncertainty structure, computational efficiency, and end-use relevance, we can progress toward more credible and actionable hydrological projections in a changing world. The thesis thus makes the case for the strategic, contextually aware ensemble design that considers the nonlinearity of interactions between climate and hydrology and prioritizes impact-relevance over purely climate-space performance.

5.2 Trade-offs in Excluding Climate Models

Model evaluation is fundamental in GCM selection, as the ability to realistically reproduce past climates is generally viewed as a prerequisite for confidence in future projections, though on its own, it does not provide sufficient grounds for such confidence (Nguyen et al., 2024). Model performance can be evaluated using different metrics at both the global (e.g. Donat et al., 2023; Ridder et al., 2021) and regional sclaes (e.g. Di Virgilio et al., 2022; Palmer et al., 2023). Yet the lack of standard evaluation metrics between studies makes direct comparison of results or consistent tracking of model performance between regions difficult.

One of the most controversial discussions in climate impact studies, particularly in hydrology, is whether to filter out climate models with high ECS (Hausfather et al., 2022; Ribes et al., 2021). Removing models that project extreme global warming can be justified on the basis of physical credibility, with specific concerns that these models fall outside projected warming ranges according to current observations and theoretical limits (Knutti, Rugenstein, et al., 2017). Our results nevertheless highlight that such exclusion decisions are far from easy to make, particularly when moving from the global climate modeling to regional hydrological impact assessments. While the exclusion of high-ECS models reduced the spread of streamflow projections in the majority of the catchments, it counterintuitively resulted in increased uncertainty in others, a slightly more than one-third of the catchments examined experienced this uncertainty rise after model exclusion. Our results point out that globally extreme climate models will not necessarily be outliers for hydrological simulations. In fact, in certain situations, e.g., snow-dominated basins or dry regions, high-ECS models can generate hydrological responses that appear more plausible when compared with regional hydroclimatic behavior and pseudo-reality experiments.

Model exclusion on the basis of global metrics such as ECS alone runs the danger of overlooking crucial hydrological behaviour. High-ECS models can be implausible at the global scale but still have valuable insights to offer at the regional scale, particularly in stress-testing contexts or in examining worst-case futures. Model exclusion inevitably results in information loss, particularly regarding the uncertainty in the projections. It should be noted that by reducing the ensemble, only information about the uncertainty is reduced, not the uncertainty

itself (Wilcke & Bärring, 2016). Although subsetting GCMs unavoidably results in loss of information, it is critical to identify and maintain the most relevant information for the study's objective. While physical and statistical relationships between data and projections could help keep as much information as possible, at the end of the day, this is a subjective choice best determined in the discussion between researchers and stakeholders.

To this end, the thesis recommends exclusion decisions to be made cautiously with justification, taking into account the study's specific hydrological objective, regional sensitivities, and risk management needs (Palmer et al., 2023; Ribes et al., 2021; Shiogama et al., 2024). Rather than simply focusing on global plausibility, hydrological modellers are to consider if the inclusion of such models helps to inform plausible extreme futures or worst-case scenarios that are beneficial to adaptation planning.

5.3 Evaluating the Utility of Model Weighting in Hydrological Impact Studies

Climate model weighting strategies aim to constrain the uncertainty of ensemble projections by assigning more weights to the models that better capture key climatic processes in the historical period. While this approach appears logically sound, this thesis's findings and broader literature suggest that the effectiveness of model weighting is conditional. Its success strongly depends on methodological decisions, the modeling task context, and the nature of the downstream impact models (H.-M. Wang et al., 2019; Wootten et al., 2022).

Through a pseudo-reality framework, this thesis explored model weighting to assess the performance of various weighting approaches on streamflow projections. The results confirmed that, if uncorrected (raw) climate model outputs are used, unequal weighting can improve hydrological projection skill. The findings support earlier studies, such as Knutti et al., 2017, who suggest unweighted means can be misleading if ensemble members have substantially varying qualities or are non-independent. Weighting, in these cases, is a remedial action, lessening the effect of poor or too-similar models (Lorenz et al., 2018; Merrifield et al., 2020, 2023). However, in bias-corrected projections, the added value of any weighting

approach became insignificant with little to no difference observed between weighted and unweighted ensemble outputs. These findings suggest that weighting loses its value when used alongside preprocessing steps (e.g., bias correction) that already align model outputs with observations.

To this is the added complexity that the performance of weighting schemes is heavily dependent on the variable of interest, geographic area, and performance measure (Palmer et al., 2023; Wootten et al., 2022). The use of weights without regard to these factors can reduce, rather than enhance, projection robustness. Our analysis also revealed hydrological-based weighting to outperform climate-only approaches for projecting streamflow, highlighting the importance of weighting criteria to be aligned to the specific aims of an impact study. Even so, these performance-weighted weights were of limited value when bias correction was applied.

In particular, the relationship between climate drivers and hydrological outputs is often nonlinear and site-dependent. Therefore, evaluating weighting schemes solely in “climate space”, i.e., based on a model’s skill in simulating temperature or precipitation, can be misleading. To ensure models contribute meaningfully to realistic impact estimates, performance must also be assessed in “impact space,” where the end-use variable (e.g., streamflow) is directly simulated. Applying climate-based weights without verifying their validity in impact models can reduce, rather than enhance, the credibility of projections. While impact-based weighting remains underdeveloped, it offers a promising direction, especially when improving projections of hydrologically significant outcomes is the goal.

These findings collectively suggest a pragmatic modeling philosophy: model democracy, or equal weighting, remains a robust and justifiable strategy for many hydrological applications, particularly after bias correction. However, model democracy should not be mistaken for indiscriminate inclusion; strategic model exclusion, based on regional hydrological performance or physical realism, remains a valuable means to improve ensemble reliability.

5.4 Limitations and Recommendations

5.4.1 Emission Scenario

This study relied only on a single high-emission greenhouse gas scenario, RCP8.5, and its new equivalent SSP5-8.5. While in hindsight these scenarios are seen as pessimistic, their use remains legitimate in climate effect studies. Even though the overall conclusions of this study will most likely concur with general trends reported in the literature and will not significantly differ under other emission pathways, addition of additional scenarios, can be beneficial to position the results in a wider socio-economic context. This addition would make the findings more robust as well as express the set of plausible climatic futures.

5.4.2 Bias Correction

Advanced multivariate bias correction techniques were employed, chosen for their ability to preserve inter-variable relationships and reduce systematic error. These techniques have been demonstrated recently in the literature to perform well. However, alternative bias correction strategies, particularly designed for extreme event, seasonality, or temporal ordering, were not explored. These techniques, taken into account in future work, may strengthen hydro-climatic extreme representation and further improve impact projection credibility.

Similar to other bias correction strategies, the employed techniques in this study assume that the statistical relationships between observed and modeled data within the calibration period are also applicable in the future (i.e., stationarity). However, in cases of strong climate change, this assumption may not be valid, especially in extreme cases or unprecedented climates, potentially introducing non-negligible biases in impact estimates.

5.4.3 Hydrological Modeling

Hydrological simulations were carried out with lumped conceptual models, chosen for their computational efficiency in large-sample applications. Even though the choice is appropriate for the scale of our study, it inherently constrains the spatial representation of key hydrological processes, such as groundwater flow, snowpack heterogeneity, and land surface heterogeneity. Conceptual models, including GR4J, often exhibit reduced transferability when applied to climate states that differ substantially from the calibration period, reflecting well-documented challenges in representing hydrologic non-stationarity (Harvey et al., 2024; Saavedra et al., 2022; Stephens et al., 2019). Their fixed parameter sets and simplified treatment of snow, vegetation, and evapotranspiration processes limit their ability to capture evolving hydroclimatological dynamics under climate change. Nonetheless, conceptual models provide a practical and transparent framework for large-ensemble climate impact experiments, serving as a useful baseline from which more process-rich modelling strategies can evolve.

Future work would benefit from incorporating models with more explicit representations of these changing processes or models that allow climate-dependent parameters or structural flexibility. Semi-distributed or fully distributed models, physically based frameworks, or data-driven architectures such as Long Short-Term Memory (LSTM) networks may offer improved robustness when extrapolating beyond historical conditions. Machine-learning approaches, particularly deep learning, have shown strong performance in large-sample hydrology and potential advantages in ungauged basins, though challenges remain regarding interpretability and generalizability. Recent studies also suggest that LSTMs may provide more stable projections under climate change (e.g. Martel et al., 2025).

The hydrologic model was run on a daily time step, which, although common in literature, limits the analysis of short-duration extreme events such as flash flood or sub-daily drought dynamics. Events that evolve at sub-daily timescales are not fully resolved, potentially underestimating impacts under more intense precipitation scenarios. To partly mitigate this limitation, we excluded very small catchments from the analysis, using catchment area as a proxy for hydrological response time. The assumption is that smaller basins tend to respond more rapidly to precipitation inputs and are thus more sensitive to sub-daily variability. By

filtering out these basins, we reduced the risk of misrepresenting sub-daily processes within a daily time-step framework. While this does not fully resolve the scale mismatch, it provides a pragmatic balance between computational feasibility and representativeness of hydrological responses across the domain.

Finally, the subset of GCMs identified as most representative in this study is conditional on the hydrological model structure and the climate forcing used. More complex or spatially explicit hydrological models, or dynamically downscaled RCM inputs, may respond differently to climatic drivers, potentially altering which GCMs span the relevant hydrological uncertainty. This reinforces a central conclusion of the thesis: model selection is inherently context-specific, and no universal subset can be expected to perform optimally across modeling frameworks.

5.4.4 Physically Informed Model Selection

Finally, future work should also explore physically informed model selection approaches. While this thesis focuses on statistical and performance-based criteria, an additional avenue involves evaluating GCMs based on their ability to represent key regional hydroclimatological processes (e.g., snow accumulation and melt, atmospheric rivers, monsoon dynamics, soil moisture feedbacks). Incorporating such process-based diagnostics could help identify models that are not only statistically adequate but also physically credible for the region of interest, thereby improving the interpretability and robustness of hydrological impact assessments.

CONCLUSION

This thesis explored the influence of ensemble design decisions such as climate model sub-selection and weighting on hydrological impact assessments over North American catchments. In three connected studies, this research offers important insights on the impact of methodological decisions on projected streamflow, highlighting that choices made in the design of climate model ensembles are not only technical decisions but have significant implications for the credibility, interpretability, and usefulness of hydrological impact assessments.

The first study demonstrated that sub-selection of climate models by using informed sampling techniques, such as the KKZ algorithm using key climate indices, is able to preserve much of the uncertainty captured by full GCM ensembles. However, the transfer of uncertainty from the climate domain to the hydrologic domain is nonlinear and catchment-dependent. While ensemble reduction offers computational advantages, its implications extend further, shaping the extent to which the model ensemble captures the plausible range of future hydrological conditions.

The second study evaluated the effect of excluding high ECS models on variability in future streamflow. Even though these models produce higher spread in global climate projections, their impact on hydrological response in their respective regions is less critical. Even though their removal yielded reduced projection spread in Canada, Alaska, the southeastern United States, and the Pacific coast, small changes or even greater variability was seen in other locations. This kind of spatial heterogeneity means hot models are not hydrologically outliers in and of themselves. Their generally wetter projections may well be compatible with potential futures for some locations within some catchments, and a one-model-fits-all model exclusion approach becomes more challenging. The implications emphasize the need to choose models not merely on the basis of global climate metrics like ECS but also on local hydrological relevance.

Finally, the value of climate model weighting schemes was studied using a pseudo-reality experiment in which all GCMs were run as the “true” future and the performance of six weighting schemes were compared. Model weighting improved temperature and precipitation projections, but these improvements did not always lead to improved streamflow projections, particularly when bias correction had been done. when bias correction was applied, no weighting scheme outperformed the baseline equal-weighting case. These findings suggest that model democracy, or equal weighting, remains a robust and justifiable strategy for many hydrological applications, particularly after bias correction.

The thesis proposes cautious, context-driven approaches to ensemble sub-selection and weighting. Instead of rigorously applying global performance metrics or statistical abstractions, researchers would be well advised to consider regional hydroclimatic relevance, the nature of the impact variable being simulated, and the nature of the modeling pipeline (e.g., bias correction). Although equal weighting and model democracy remain justified, especially following bias correction, strategic model exclusion and performance-weighting are useful under specific circumstances, provided they are applied judiciously and with full understanding of trade-offs. In the absence of a universally applicable best practice, an open and adaptable modeling philosophy is proposed: exploit a large ensemble where possible to allow for uncertainty, meticulously estimate the impact of omission, and employ weighting cautiously, being aware when and where it adds value. It is this pragmatic approach that serves to advance the evolution of reproducible, reliable, and policy-relevant hydrological climate impact assessments.

BIBLIOGRAPHY

Abbott, M. B., & Refsgaard, J. C. (Eds). (1996). *Distributed Hydrological Modelling* (Vol. 22). Springer Netherlands. https://doi.org/10.1007/978-94-009-0257-2

Abdulkareem, J. H., Pradhan, B., Sulaiman, W. N. A., & Jamil, N. R. (2018). Review of studies on hydrological modelling in Malaysia. *Modeling Earth Systems and Environment*, *4*(4), 1577–1605. https://doi.org/10.1007/s40808-018-0509-y

Abramowitz, G., & Bishop, C. H. (2015). *Climate Model Dependence and the Ensemble Dependence Transformation of CMIP Projections*. https://doi.org/10.1175/JCLI-D-14-00364.1

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

Ahmadalipour, A., Rana, A., Moradkhani, H., & Sharma, A. (2017). Multi-criteria evaluation of CMIP5 GCMs for climate change impact analysis. *Theoretical and Applied Climatology*, *128*(1–2), 71–87. https://doi.org/10.1007/s00704-015-1695-4

Ahmed, K., Sachindra, D. A., Shahid, S., Demirel, M. C., & Chung, E.-S. (2019). Selection of multi-model ensemble of general circulation models for the simulation of precipitation and maximum and minimum temperature based on spatial assessment metrics. *Hydrology and Earth System Sciences*, *23*(11), 4803–4824. https://doi.org/10.5194/hess-23-4803-2019

Althoff, D., Rodrigues, L. N., & Silva, D. D. da. (2021). Addressing hydrological modeling in watersheds under land cover change with deep learning. *Advances in Water Resources*, *154*, 103965. https://doi.org/10.1016/j.advwatres.2021.103965

Arsenault, R., Bazile, R., Ouellet Dallaire, C., & Brissette, F. (2016). CANOPEX: A Canadian hydrometeorological watershed database. *Hydrological Processes*, *30*(15), 2734–2736. https://doi.org/10.1002/hyp.10880

Arsenault, R., Brissette, F., Chen, J., Guo, Q., & Dallaire, G. (2020). NAC2H: The North American Climate Change and Hydroclimatology Data Set. *Water Resources Research*, *56*(8), e2020WR027097. 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. (2020). 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

Arsenault, R., Poulin, A., Côté, P., & Brissette, F. (2014). Comparison of Stochastic Optimization Algorithms in Hydrological Model Calibration. *Journal of Hydrologic Engineering*, *19*(7), 1374–1384. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000938

Ashraf Vaghefi, S., Iravani, M., Sauchyn, D., Andreichuk, Y., Goss, G., & Faramarzi, M. (2019). Regionalization and parameterization of a hydrologic model significantly affect the cascade of uncertainty in climate-impact projections. *Climate Dynamics*, *53*(5), 2861–2886. https://doi.org/10.1007/s00382-019-04664-w

Bárdossy, A. (2007). Calibration of hydrological model parameters for ungauged catchments. *Hydrology and Earth System Sciences*, *11*(2), 703–710. https://doi.org/10.5194/hess-11-703-2007

Bárdossy, A., & Pegram, G. (2012). Multiscale spatial recorrelation of RCM precipitation to produce unbiased climate change scenarios over large areas and small. *Water Resources Research*, *48*(9), 2011WR011524. https://doi.org/10.1029/2011WR011524

Bellouin, N., Quaas, J., Gryspeerdt, E., Kinne, S., Stier, P., Watson-Parris, D., Boucher, O., Carslaw, K. S., Christensen, M., Daniau, A.-L., Dufresne, J.-L., Feingold, G., Fiedler, S., Forster, P., Gettelman, A., Haywood, J. M., Lohmann, U., Malavelle, F., Mauritsen, T., … Stevens, B. (2020). Bounding Global Aerosol Radiative Forcing of Climate Change. *Reviews of Geophysics*, *58*(1), e2019RG000660. https://doi.org/10.1029/2019RG000660

Bellucci, A., Haarsma, R., Gualdi, S., Athanasiadis, P. J., Caian, M., Cassou, C., Fernandez, E., Germe, A., Jungclaus, J., Kröger, J., Matei, D., Müller, W., Pohlmann, H., Salas Y Melia, D., Sanchez, E., Smith, D., Terray, L., Wyser, K., & Yang, S. (2015). An assessment of a multi-model ensemble of decadal climate predictions. *Climate Dynamics*, *44*(9–10), 2787–2806. https://doi.org/10.1007/s00382-014-2164-y

Beven*, K. (2001). How far can we go in distributed hydrological modelling? *Hydrology and Earth System Sciences*, *5*(1), 1–12. https://doi.org/10.5194/hess-5-1-2001

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology*, *320*(1), 18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007

Beyer, R., Krapp, M., & Manica, A. (2020). An empirical evaluation of bias correction methods for palaeoclimate simulations. *Climate of the Past*, *16*(4), 1493–1508. https://doi.org/10.5194/cp-16-1493-2020

Biondi, D., Freni, G., Iacobellis, V., Mascaro, G., & Montanari, A. (2012). Validation of hydrological models: Conceptual basis, methodological approaches and a proposal for a code of practice. *Physics and Chemistry of the Earth, Parts A/B/C*, *42–44*, 70–76. https://doi.org/10.1016/j.pce.2011.07.037

Bishop, C. H., & Abramowitz, G. (2013). Climate model dependence and the replicate Earth paradigm. *Climate Dynamics*, *41*(3–4), 885–900. https://doi.org/10.1007/s00382-012-1610-y

Borodina, A., Fischer, E. M., & Knutti, R. (2017). Potential to Constrain Projections of Hot Temperature Extremes. *Journal of Climate*, *30*(24), 9949–9964.

Brient, F. (2020). Reducing Uncertainties in Climate Projections with Emergent Constraints: Concepts, Examples and Prospects. *Advances in Atmospheric Sciences*, *37*(1), 1–15. https://doi.org/10.1007/s00376-019-9140-8

Brunner, L., Pendergrass, A. G., Lehner, F., Merrifield, A. L., Lorenz, R., & Knutti, R. (2020). Reduced global warming from CMIP6 projections when weighting models by performance and independence. *Earth System Dynamics*, *11*(4), 995–1012. https://doi.org/10.5194/esd-11-995-2020

Caldwell, P. M., Bretherton, C. S., Zelinka, M. D., Klein, S. A., Santer, B. D., & Sanderson, B. M. (2014). Statistical significance of climate sensitivity predictors obtained by data mining. *Geophysical Research Letters*, *41*(5), 1803–1808. https://doi.org/10.1002/2014GL059205

Cannon, A. J. (2015). Selecting GCM Scenarios that Span the Range of Changes in a Multimodel Ensemble: Application to CMIP5 Climate Extremes Indices. *Journal of Climate*, *28*(3), 1260–1267. https://doi.org/10.1175/JCLI-D-14-00636.1

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

Cannon, A. J., Sobie, S. R., & Murdock, T. Q. (2015). *Bias Correction of GCM Precipitation by Quantile Mapping: How Well Do Methods Preserve Changes in Quantiles and Extremes?* https://doi.org/10.1175/JCLI-D-14-00754.1

Casajus, N., Périé, C., Logan, T., Lambert, M.-C., Blois, S. de, & Berteaux, D. (2016). An Objective Approach to Select Climate Scenarios when Projecting Species Distribution under Climate Change. *PLOS ONE*, *11*(3), e0152495. https://doi.org/10.1371/journal.pone.0152495

Castaneda-Gonzalez, M., Poulin, A., Romero-Lopez, R., Turcotte, R., & Chaumont, D. (2022). Uncertainty sources in flood projections over contrasting hydrometeorological regimes. *Hydrological Sciences Journal*, *67*(15), 2232–2253. https://doi.org/10.1080/02626667.2022.2137415

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. (2015). Assessing the limits of bias-correcting climate model outputs for climate change impact studies. *Journal of Geophysical Research: Atmospheres*, *120*(3), 1123–1136. https://doi.org/10.1002/2014JD022635

Chen, J., Brissette, F. P., & Lucas-Picher, P. (2016). Transferability of optimally-selected climate models in the quantification of climate change impacts on hydrology. *Climate Dynamics*, *47*(9–10), 3359–3372. https://doi.org/10.1007/s00382-016-3030-x

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

Chen, J., Brissette, F. P., Poulin, A., & Leconte, R. (2011). Overall uncertainty study of the hydrological impacts of climate change for a Canadian watershed: OVERALL

UNCERTAINTY OF CLIMATE CHANGE IMPACTS ON HYDROLOGY. *Water Resources Research*, *47*(12). https://doi.org/10.1029/2011WR010602

Chen, J., Li, C., Brissette, F. P., Chen, H., Wang, M., & Essou, G. R. C. (2018). 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

Chen, J., & Zhang, X. J. (2021). Challenges and potential solutions in statistical downscaling of precipitation. *Climatic Change*, *165*(3), 63. https://doi.org/10.1007/s10584-021-03083-3

Chokkavarapu, N., & Mandla, V. R. (2019). Comparative study of GCMs, RCMs, downscaling and hydrological models: A review toward future climate change impact estimation. *SN Applied Sciences*, *1*(12), 1698. https://doi.org/10.1007/s42452-019-1764-x

Chowdhury, R. K., & Eslamian, S. (2014). Climate Change and Hydrologic Modeling. In *Handbook of Engineering Hydrology*. CRC Press.

Christensen, N. S., & Lettenmaier, D. P. (2007). A multimodel ensemble approach to assessment of climate change impacts on the hydrology and water resources of the Colorado River Basin. *Hydrol. Earth Syst. Sci.*, 18.

Clark, M. P., Wilby, R. L., Gutmann, E. D., Vano, J. A., Gangopadhyay, S., Wood, A. W., Fowler, H. J., Prudhomme, C., Arnold, J. R., & Brekke, L. D. (2016). Characterizing Uncertainty of the Hydrologic Impacts of Climate Change. *Current Climate Change Reports*, *2*(2), 55–64. https://doi.org/10.1007/s40641-016-0034-x

Collins, M., Knutti, R., Arblaster, J., Dufresne, J.-L., Fichefet, T., Gao, X., Jr, W. J. G., Johns, T., Krinner, G., Shongwe, M., Weaver, A. J., Wehner, M., Allen, M. R., Andrews, T., Beyerle, U., Bitz, C. M., Bony, S., Booth, B. B. B., Brooks, H. E., … Tett, S. (2013). *Long-term Climate Change: Projections, Commitments and Irreversibility*. 108.

Cook, B. I., Mankin, J. S., & Anchukaitis, K. J. (2018). Climate Change and Drought: From Past to Future. *Current Climate Change Reports*, *4*(2), 164–179. https://doi.org/10.1007/s40641-018-0093-2

Cox, P. M., Huntingford, C., & Williamson, M. S. (2018). Emergent constraint on equilibrium climate sensitivity from global temperature variability. *Nature*, *553*(7688), Article 7688. https://doi.org/10.1038/nature25450

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

Dai, A. (2013). Increasing drought under global warming in observations and models. *Nature Climate Change*, *3*(1), Article 1. https://doi.org/10.1038/nclimate1633

Dekens, L., Parey, S., Grandjacques, M., & Dacunha-Castelle, D. (2017). Multivariate distribution correction of climate model outputs: A generalization of quantile mapping approaches. *Environmetrics*, *28*(6), e2454. https://doi.org/10.1002/env.2454

Deser, C., Lehner, F., Rodgers, K. B., Ault, T., Delworth, T. L., DiNezio, P. N., Fiore, A., Frankignoul, C., Fyfe, J. C., Horton, D. E., Kay, J. E., Knutti, R., Lovenduski, N. S., Marotzke, J., McKinnon, K. A., Minobe, S., Randerson, J., Screen, J. A., Simpson, I. R., & Ting, M. (2020). Insights from Earth system model initial-condition large ensembles and future prospects. *Nature Climate Change*, *10*(4), 277–286. https://doi.org/10.1038/s41558-020-0731-2

Deser, C., Phillips, A., Bourdette, V., & Teng, H. (2012). Uncertainty in climate change projections: The role of internal variability. *Climate Dynamics*, *38*(3), 527–546. https://doi.org/10.1007/s00382-010-0977-x

Devi, G. K., Ganasri, B. P., & Dwarakish, G. S. (2015). A Review on Hydrological Models. *Aquatic Procedia*, *4*, 1001–1007. https://doi.org/10.1016/j.aqpro.2015.02.126

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

Donat, M. G., Delgado-Torres, C., De Luca, P., Mahmood, R., Ortega, P., & Doblas-Reyes, F. J. (2023). How Credibly Do CMIP6 Simulations Capture Historical Mean and Extreme Precipitation Changes? *Geophysical Research Letters*, *50*(14), e2022GL102466. https://doi.org/10.1029/2022GL102466

Dos Santos, F. M., De Oliveira, R. P., & Mauad, F. F. (2018). Lumped versus Distributed Hydrological Modeling of the Jacaré-Guaçu Basin, Brazil. *Journal of Environmental Engineering*, *144*(8). https://doi.org/10.1061/(asce)ee.1943-7870.0001397

Dubrovsky, M., Trnka, M., Holman, I. P., Svobodova, E., & Harrison, P. A. (2015). Developing a reduced-form ensemble of climate change scenarios for Europe and its application to selected impact indicators. *Climatic Change*, *128*(3), 169–186. https://doi.org/10.1007/s10584-014-1297-7

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

Ehret, U., Zehe, E., Wulfmeyer, V., Warrach-Sagi, K., & Liebert, J. (2012). HESS Opinions 'Should we apply bias correction to global and regional climate model data?' *Hydrology and Earth System Sciences*, *16*(9), 3391–3404. https://doi.org/10.5194/hess-16-3391-2012

Ekström, M., Grose, M. R., & Whetton, P. H. (2015). An appraisal of downscaling methods used in climate change research. *WIREs Climate Change*, *6*(3), 301–319. https://doi.org/10.1002/wcc.339

Evans, J. P., Ji, F., Abramowitz, G., & Ekström, M. (2013). Optimally choosing small ensemble members to produce robust climate simulations. *Environmental Research Letters*, *8*(4), 044050. https://doi.org/10.1088/1748-9326/8/4/044050

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

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., … 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

Fallah, B., Rostami, M., Russo, E., Harder, P., Menz, C., Hoffmann, P., Didovets, I., & Hattermann, F. F. (2025). Climate model downscaling in central Asia: A dynamical

and a neural network approach. *Geoscientific Model Development*, *18*(1), 161–180. https://doi.org/10.5194/gmd-18-161-2025

Feng, D., & Beighley, E. (2019). *Identifying uncertainties in simulated streamflow from hydrologic model components for climate change impact assessments*. https://doi.org/10.5194/hess-2019-328

Fischer, E. M., & Knutti, R. (2016). Observed heavy precipitation increase confirms theory and early models. *Nature Climate Change*, *6*(11), 986–991. https://doi.org/10.1038/nclimate3110

Flynn, C. M., & Mauritsen, T. (2020). On the climate sensitivity and historical warming evolution in recent coupled model ensembles. *Atmospheric Chemistry and Physics*, *20*(13), 7829–7842. https://doi.org/10.5194/acp-20-7829-2020

Forster, P. M., Andrews, T., Good, P., Gregory, J. M., Jackson, L. S., & Zelinka, M. (2013). Evaluating adjusted forcing and model spread for historical and future scenarios in the CMIP5 generation of climate models. *Journal of Geophysical Research: Atmospheres*, *118*(3), 1139–1150. https://doi.org/10.1002/jgrd.50174

Fowler, H. J., Blenkinsop, S., & Tebaldi, C. (2007). Linking climate change modelling to impacts studies: Recent advances in downscaling techniques for hydrological modelling: ADVANCES IN DOWNSCALING TECHNIQUES FOR HYDROLOGICAL MODELLING. *International Journal of Climatology*, *27*(12), 1547–1578. https://doi.org/10.1002/joc.1556

François, B., Vrac, M., Cannon, A. J., Robin, Y., & Allard, D. (2020). Multivariate bias corrections of climate simulations: Which benefits for which losses? *Earth System Dynamics*, *11*(2), 537–562. https://doi.org/10.5194/esd-11-537-2020

Gao, C., Booij, M. J., & Xu, Y.-P. (2020). Assessment of extreme flows and uncertainty under climate change: Disentangling the uncertainty contribution of representative concentration pathways, global climate models and internal climate variability. *Hydrology and Earth System Sciences*, *24*(6), 3251–3269. https://doi.org/10.5194/hess-24-3251-2020

George, J., & Athira, P. (2022). Process informed selection of climate models for climate change impact assessment in the Western Coast of India. *Theoretical and Applied Climatology*, *150*(1), 805–828. https://doi.org/10.1007/s00704-022-04197-z

Ghonchepour, D., Sadoddin, A., Bahremand, A., Croke, B., Jakeman, A., & Salmanmahiny, A. (2021). A methodological framework for the hydrological model selection process in water resource management projects. *Natural Resource Modeling*, *34*(3), e12326. https://doi.org/10.1111/nrm.12326

Giuntoli, I., Prosdocimi, I., & Hannah, D. M. (2021). Going Beyond the Ensemble Mean: Assessment of Future Floods From Global Multi-Models. *Water Resources Research*, *57*(3), e2020WR027897. https://doi.org/10.1029/2020WR027897

Giuntoli, I., Vidal, J.-P., Prudhomme, C., & Hannah, D. M. (2015). Future hydrological extremes: The uncertainty from multiple global climate and global hydrological models. *Earth System Dynamics*, *6*(1), 267–285. https://doi.org/10.5194/esd-6-267-2015

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

Giuntoli, I., Villarini, G., Prudhomme, C., Mallakpour, I., & Hannah, D. M. (2015). Evaluation of global impact models' ability to reproduce runoff characteristics over the central United States. *Journal of Geophysical Research: Atmospheres*, *120*(18), 9138–9159. https://doi.org/10.1002/2015JD023401

Gleckler, P. J., Taylor, K. E., & Doutriaux, C. (2008). Performance metrics for climate models. *Journal of Geophysical Research: Atmospheres*, *113*(D6). https://doi.org/10.1029/2007JD008972

Golian, S., & Murphy, C. (2021). Evaluation of Sub-Selection Methods for Assessing Climate Change Impacts on Low-Flow and Hydrological Drought Conditions. *Water Resources Management*, *35*(1), 113–133. https://doi.org/10.1007/s11269-020-02714-1

Grose, M. R., Narsey, S., Trancoso, R., Mackallah, C., Delage, F., Dowdy, A., Di Virgilio, G., Watterson, I., Dobrohotoff, P., Rashid, H. A., Rauniyar, S., Henley, B., Thatcher, M., Syktus, J., Abramowitz, G., Evans, J. P., Su, C.-H., & Takbash, A. (2023). A CMIP6-based multi-model downscaling ensemble to underpin climate change services in Australia. *Climate Services*, *30*, 100368. https://doi.org/10.1016/j.cliser.2023.100368

Guo, Q., Chen, J., Zhang, X., Shen, M., Chen, H., & Guo, S. (2019). A new two-stage multivariate quantile mapping method for bias correcting climate model outputs. *Climate Dynamics*, *53*(5–6), 3603–3623. https://doi.org/10.1007/s00382-019-04729-w

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Gutmann, E. D., Hamman, J. J., Clark, M. P., Eidhammer, T., Wood, A. W., & Arnold, J. R. (2022). *En-GARD: A Statistical Downscaling Framework to Produce and Test Large Ensembles of Climate Projections*. https://doi.org/10.1175/JHM-D-21-0142.1

Hallegatte, S. (2009). Strategies to adapt to an uncertain climate change. *Global Environmental Change*, *19*(2), 240–247. https://doi.org/10.1016/j.gloenvcha.2008.12.003

Hamed, M. M., Nashwan, M. S., & Shahid, S. (2022). A novel selection method of CMIP6 GCMs for robust climate projection. *International Journal of Climatology*, *42*(8), 4258–4272. https://doi.org/10.1002/joc.7461

Hansen, N., Müller, S. D., & Koumoutsakos, P. (2003). Reducing the Time Complexity of the Derandomized Evolution Strategy with Covariance Matrix Adaptation (CMA-ES). *Evolutionary Computation*, *11*(1), 1–18. https://doi.org/10.1162/106365603321828970

Harvey, N., Razavi, S., & Bilish, S. (2024). Review of hydrological modelling in the Australian Alps: From rainfall-runoff to physically based models. *Australasian Journal of Water Resources*, *28*(2), 208–224. https://doi.org/10.1080/13241583.2024.2343453

Hassan, I., Kalin, R. M., White, C. J., & Aladejana, J. A. (2020). Selection of CMIP5 GCM Ensemble for the Projection of Spatio-Temporal Changes in Precipitation and Temperature over the Niger Delta, Nigeria. *Water*, *12*(2), 385. https://doi.org/10.3390/w12020385

Haughton, N., Abramowitz, G., Pitman, A., & Phipps, S. J. (2014). On the generation of climate model ensembles. *Climate Dynamics*, *43*(7), 2297–2308. https://doi.org/10.1007/s00382-014-2054-3

Hausfather, Z. (2019). *CMIP6: The next generation of climate models explained*. Carbon Brief.

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

Hausfather, Z., & Peters, G. P. (2020). Emissions – the 'business as usual' story is misleading. *Nature*, *577*(7792), 618–620. https://doi.org/10.1038/d41586-020-00177-3

Hauswirth, S. M., Bierkens, M. F. P., Beijk, V., & Wanders, N. (2021). The potential of data driven approaches for quantifying hydrological extremes. *Advances in Water Resources*, *155*, 104017. https://doi.org/10.1016/j.advwatres.2021.104017

Hawkins, E., & Sutton, R. (2009). The Potential to Narrow Uncertainty in Regional Climate Predictions. *Bulletin of the American Meteorological Society*, *90*(8), 1095–1108. https://doi.org/10.1175/2009BAMS2607.1

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

Hewitson, B., & Crane, R. (1996). Climate downscaling: Techniques and application. *Climate Research*, *7*, 85–95. https://doi.org/10.3354/cr007085

Hirabayashi, Y., Tanoue, M., Sasaki, O., Zhou, X., & Yamazaki, D. (2021). Global exposure to flooding from the new CMIP6 climate model projections. *Scientific Reports*, *11*(1), Article 1. https://doi.org/10.1038/s41598-021-83279-w

Hodson, D. L. R., Keeley, S. P. E., West, A., Ridley, J., Hawkins, E., & Hewitt, H. T. (2013). Identifying uncertainties in Arctic climate change projections. *Climate Dynamics*, *40*(11), 2849–2865. https://doi.org/10.1007/s00382-012-1512-z

Houle, D., Bouffard, A., Duchesne, L., Logan, T., & Harvey, R. (2012). Projections of Future Soil Temperature and Water Content for Three Southern Quebec Forested Sites. *Journal of Climate*, *25*(21), 7690–7701. https://doi.org/10.1175/JCLI-D-11-00440.1

Huang, S., Kumar, R., Flörke, M., Yang, T., Hundecha, Y., Kraft, P., Gao, C., Gelfan, A., Liersch, S., Lobanova, A., Strauch, M., van Ogtrop, F., Reinhardt, J., Haberlandt, U., & Krysanova, V. (2017). Evaluation of an ensemble of regional hydrological models in 12 large-scale river basins worldwide. *Climatic Change*, *141*(3), 381–397. https://doi.org/10.1007/s10584-016-1841-8

Hui, Y., Chen, J., Xu, C.-Y., Xiong, L., & Chen, H. (2019). Bias nonstationarity of global climate model outputs: The role of internal climate variability and climate model sensitivity. *International Journal of Climatology*, *39*(4), 2278–2294. https://doi.org/10.1002/joc.5950

Illangasingha, S., Koike, T., Rasmy, M., Tamakawa, K., Matsuki, H., & Selvarajah, H. (2023). A holistic approach for using global climate model (GCM) outputs in decision making. *Journal of Hydrology*, *626*, 130213. https://doi.org/10.1016/j.jhydrol.2023.130213

IPCC. (2014). *Intergovernmental Panel on Climate Change. Fifth Assessment Report: Climate Change 2014 Synthesis Report.*

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

IPCC. (2023). *IPCC, 2023: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland.* (First). Intergovernmental Panel on Climate Change (IPCC). https://doi.org/10.59327/IPCC/AR6-9789291691647

Ito, R., Shiogama, H., Nakaegawa, T., & Takayabu, I. (2020). Uncertainties in climate change projections covered by the ISIMIP and CORDEX model subsets from CMIP5. *Geoscientific Model Development*, *13*(3), 859–872. https://doi.org/10.5194/gmd-13-859-2020

Jagannathan, K., Jones, A. D., & Kerr, A. C. (2020). Implications of climate model selection for projections of decision-relevant metrics: A case study of chill hours in California. *Climate Services*, *18*, 100154. https://doi.org/10.1016/j.cliser.2020.100154

Jajarmizadeh, M., Harun, S., & Salarpour, M. (2012). A Review on Theoretical Consideration and Types of Models in Hydrology. *Journal of Environmental Science and Technology*, *5*(5), 249–261. https://doi.org/10.3923/jest.2012.249.261

Jiang, T., Chen, Y. D., Xu, C., Chen, X., Chen, X., & Singh, V. P. (2007). Comparison of hydrological impacts of climate change simulated by six hydrological models in the Dongjiang Basin, South China. *Journal of Hydrology*, *336*(3–4), 316–333. https://doi.org/10.1016/j.jhydrol.2007.01.010

Joseph, J., Ghosh, S., Pathak, A., & Sahai, A. K. (2018). Hydrologic impacts of climate change: Comparisons between hydrological parameter uncertainty and climate model

uncertainty. *Journal of Hydrology*, *566*, 1–22. https://doi.org/10.1016/j.jhydrol.2018.08.080

Karlsson, I. B., Sonnenborg, T. O., Refsgaard, J. C., Trolle, D., Børgesen, C. D., Olesen, J. E., Jeppesen, E., & Jensen, K. H. (2016). 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

Katsavounidis, I., Jay Kuo, C.-C., & Zhang, Z. (1994). A new initialization technique for generalized Lloyd iteration. *IEEE Signal Processing Letters*, *1*(10), 144–146. IEEE Signal Processing Letters. https://doi.org/10.1109/97.329844

Kavetski, D., Kuczera, G., & Franks, S. W. (2006). Calibration of conceptual hydrological models revisited: 1. Overcoming numerical artefacts. *Journal of Hydrology*, *320*(1), 173–186. https://doi.org/10.1016/j.jhydrol.2005.07.012

Kim, D.-I., Kwon, H.-H., & Han, D. (2019). Bias correction of daily precipitation over South Korea from the long-term reanalysis using a composite Gamma-Pareto distribution approach. *Hydrology Research*, *50*(4), 1138–1161. https://doi.org/10.2166/nh.2019.127

Klein, S. A., & Hall, A. (2015). Emergent Constraints for Cloud Feedbacks. *Current Climate Change Reports*, *1*(4), 276–287. https://doi.org/10.1007/s40641-015-0027-1

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

Knoben, W. J. M., Freer, J. E., & Woods, R. A. (2019). *Technical note: Inherent benchmark or not? Comparing Nash-Sutcliffe and Kling-Gupta efficiency scores* [Preprint]. Catchment hydrology/Modelling approaches. https://doi.org/10.5194/hess-2019-327

Knowles, N., Dettinger, M. D., & Cayan, D. R. (2006). Trends in Snowfall versus Rainfall in the Western United States. *Journal of Climate*, *19*(18), 4545–4559. https://doi.org/10.1175/JCLI3850.1

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., Furrer, R., Tebaldi, C., Cermak, J., & Meehl, G. A. (2010). *Challenges in Combining Projections from Multiple Climate Models*. https://doi.org/10.1175/2009JCLI3361.1

Knutti, R., Masson, D., & Gettelman, A. (2013). Climate model genealogy: Generation CMIP5 and how we got there: CLIMATE MODEL GENEALOGY. *Geophysical Research Letters*, *40*(6), 1194–1199. https://doi.org/10.1002/grl.50256

Knutti, R., Rugenstein, M. A. A., & Hegerl, G. C. (2017). Beyond equilibrium climate sensitivity. *Nature Geoscience*, *10*(10), 727–736. https://doi.org/10.1038/ngeo3017

Knutti, R., & Sedláček, J. (2013). Robustness and uncertainties in the new CMIP5 climate model projections. *Nature Climate Change*, *3*(4), 369–373. https://doi.org/10.1038/nclimate1716

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

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019). Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning. *Water Resources Research*, *55*(12), 11344–11354. https://doi.org/10.1029/2019WR026065

Kreienkamp, F., Lorenz, P., & Geiger, T. (2020). Statistically Downscaled CMIP6 Projections Show Stronger Warming for Germany. *Atmosphere*, *11*(11), Article 11. https://doi.org/10.3390/atmos11111245

Kundzewicz, Z. W. (2008). Climate change impacts on the hydrological cycle. *Ecohydrology & Hydrobiology*, *8*(2–4), 195–203. https://doi.org/10.2478/v10104-009-0015-y

Kundzewicz, Z. W., Krysanova, V., Benestad, R. E., Hov, Ø., Piniewski, M., & Otto, I. M. (2018). Uncertainty in climate change impacts on water resources. *Environmental Science & Policy*, *79*, 1–8. https://doi.org/10.1016/j.envsci.2017.10.008

Lavell, A., Oppenheimer, M., Diop, C., Hess, J., Lempert, R., Li, J., Muir-Wood, R., Myeong, S., Moser, S., Takeuchi, K., Cardona, O.-D., Hallegatte, S., Lemos, M., Little, C., Lotsch, A., & Weber, E. (2012). Climate Change: New Dimensions in Disaster Risk, Exposure, Vulnerability, and Resilience. In C. B. Field, V. Barros, T. F. Stocker, & Q. Dahe (Eds), *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation* (1st edn, pp. 25–64). Cambridge University Press. https://doi.org/10.1017/CBO9781139177245.004

Lavers, D. A., Simmons, A., Vamborg, F., & Rodwell, M. J. (2022). An evaluation of ERA5 precipitation for climate monitoring. *Quarterly Journal of the Royal Meteorological Society*, *148*(748), 3152–3165. https://doi.org/10.1002/qj.4351

Leduc, M., Mailhot, A., Frigon, A., Martel, J.-L., Ludwig, R., Brietzke, G. B., Giguère, M., Brissette, F., Turcotte, R., Braun, M., & Scinocca, J. (2019). *The ClimEx Project: A 50-Member Ensemble of Climate Change Projections at 12-km Resolution over Europe and Northeastern North America with the Canadian Regional Climate Model (CRCM5)*. https://doi.org/10.1175/JAMC-D-18-0021.1

Lee, J.-K., & Kim, Y.-O. (2017). Selection of representative GCM scenarios preserving uncertainties. *Journal of Water and Climate Change*, *8*(4), 641–651. https://doi.org/10.2166/wcc.2017.101

Lehner, F., Deser, C., Maher, N., Marotzke, J., Fischer, E. M., Brunner, L., Knutti, R., & Hawkins, E. (2020). Partitioning climate projection uncertainty with multiple large ensembles and CMIP5/6. *Earth System Dynamics*, *11*(2), 491–508. https://doi.org/10.5194/esd-11-491-2020

Lenderink, G., & Fowler, H. J. (2017). Understanding rainfall extremes. *Nature Climate Change*, *7*(6), 391–393. https://doi.org/10.1038/nclimate3305

Li, H., Zhang, C., Chu, W., Shen, D., & Li, R. (2024). A process-driven deep learning hydrological model for daily rainfall-runoff simulation. *Journal of Hydrology*, *637*, 131434. https://doi.org/10.1016/j.jhydrol.2024.131434

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

Liu, Z., Wang, Y., Xu, Z., & Duan, Q. (2019). Conceptual Hydrological Models. In *Handbook of Hydrometeorological Ensemble Forecasting* (pp. 389–411). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-39925-1_22

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

Ludwig, R., May, I., Turcotte, R., Vescovi, L., Braun, M., Cyr, J.-F., Fortin, L.-G., Chaumont, D., Biner, S., Chartier, I., Caya, D., & Mauser, W. (2009). The role of hydrological model complexity and uncertainty in climate change impact assessment. *Advances in Geosciences*, *21*, 63–71. https://doi.org/10.5194/adgeo-21-63-2009

Lutz, A. F., ter Maat, H. W., Biemans, H., Shrestha, A. B., Wester, P., & Immerzeel, W. W. (2016). Selecting representative climate models for climate change impact studies: An advanced envelope-based selection approach: ADVANCED ENVELOPE-BASED CLIMATE MODEL SELECTION APPROACH. *International Journal of Climatology*, *36*(12), 3988–4005. https://doi.org/10.1002/joc.4608

Mahadevan, M., Noel, J. K., Umesh, M., Santhosh, A. S., & Suresh, S. (2024). Climate Change Impact on Water Resources, Food Production and Agricultural Practices. In P. Singh & N. Yadav (Eds), *The Climate-Health-Sustainability Nexus: Understanding the Interconnected Impact on Populations and the Environment* (pp. 207–229). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-56564-9_9

Maher, N., Milinski, S., & Ludwig, R. (2021). Large ensemble climate model simulations: Introduction, overview, and future prospects for utilising multiple types of large ensemble. *Earth System Dynamics*, *12*(2), 401–418. https://doi.org/10.5194/esd-12-401-2021

Maraun, D. (2012). Nonstationarities of regional climate model biases in European seasonal mean temperature and precipitation sums. *Geophysical Research Letters*, *39*(6). https://doi.org/10.1029/2012GL051210

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

Maraun, D., Wetterhall, F., Ireson, A. M., Chandler, R. E., Kendon, E. J., Widmann, M., Brienen, S., Rust, H. W., Sauter, T., Themeßl, M., Venema, V. K. C., Chun, K. P., Goodess, C. M., Jones, R. G., Onof, C., Vrac, M., & Thiele-Eich, I. (2010). Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Reviews of Geophysics*, *48*(3). https://doi.org/10.1029/2009RG000314

Marshall, L., Nott, D., & Sharma, A. (2005). Hydrological model selection: A Bayesian alternative. *Water Resources Research*, *41*(10). https://doi.org/10.1029/2004WR003719

Martel, J.-L., Brissette, F., Arsenault, R., Turcotte, R., Castañeda-Gonzalez, M., Armstrong, W., Mailhot, E., Pelletier-Dumont, J., Rondeau-Genesse, G., & Caron, L.-P. (2025). Assessing the adequacy of traditional hydrological models for climate change impact studies: A case for long short-term memory (LSTM) neural networks. *Hydrology and Earth System Sciences*, *29*(13), 2811–2836. https://doi.org/10.5194/hess-29-2811-2025

Martel, J.-L., Brissette, F. P., Lucas-Picher, P., Troin, M., & Arsenault, R. (2021). Climate Change and Rainfall Intensity–Duration–Frequency Curves: Overview of Science and Guidelines for Adaptation. *Journal of Hydrologic Engineering*, *26*(10), 03121001. https://doi.org/10.1061/(ASCE)HE.1943-5584.0002122

Martel, J.-L., Brissette, F., Troin, M., Arsenault, R., Chen, J., Su, T., & Lucas-Picher, P. (2022). CMIP5 and CMIP6 Model Projection Comparison for Hydrological Impacts Over North America. *Geophysical Research Letters*, *49*(15), e2022GL098364. https://doi.org/10.1029/2022GL098364

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.

McSweeney, C. F., Jones, R. G., Lee, R. W., & Rowell, D. P. (2015). Selecting CMIP5 GCMs for downscaling over multiple regions. *Climate Dynamics*, *44*(11–12), 3237–3260. https://doi.org/10.1007/s00382-014-2418-8

Meehl, G. A., Covey, C., Delworth, T., Latif, M., McAvaney, B., Mitchell, J. F. B., Stouffer, R. J., & Taylor, K. E. (2007). THE WCRP CMIP3 Multimodel Dataset: A New Era in

Climate Change Research. *Bulletin of the American Meteorological Society*, *88*(9), 1383–1394. https://doi.org/10.1175/BAMS-88-9-1383

Menapace, A., Dhawan, P., Dalla Torre, D., Kaffas, K., Crespi, A., Larcher, M., Righetti, M., & Cannon, A. J. (2025). Review of bias correction methods for climate model outputs in hydrology. *Journal of Hydrology*, *660*, 133213. https://doi.org/10.1016/j.jhydrol.2025.133213

Mendez, M., Maathuis, B., Hein-Griggs, D., & Alvarado-Gamboa, L.-F. (2020). Performance Evaluation of Bias Correction Methods for Climate Change Monthly Precipitation Projections over Costa Rica. *Water*, *12*(2), Article 2. https://doi.org/10.3390/w12020482

Mendlik, T., & Gobiet, A. (2016). Selecting climate simulations for impact studies based on multivariate patterns of climate change. *Climatic Change*, *135*(3–4), 381–393. https://doi.org/10.1007/s10584-015-1582-0

Meresa, H., Murphy, C., Fealy, R., & Golian, S. (2021). Uncertainties and their interaction in flood hazard assessment with climate change. *Hydrology and Earth System Sciences*, *25*(9), 5237–5257. https://doi.org/10.5194/hess-25-5237-2021

Meresa, H., Tischbein, B., & Mekonnen, T. (2022). Climate change impact on extreme precipitation and peak flood magnitude and frequency: Observations from CMIP6 and hydrological models. *Natural Hazards*, *111*(3), 2649–2679. https://doi.org/10.1007/s11069-021-05152-3

Merrifield, A. L., Brunner, L., Lorenz, R., Humphrey, V., & Knutti, R. (2023). Climate model Selection by Independence, Performance, and Spread (ClimSIPS v1.0.1) for regional applications. *Geoscientific Model Development*, *16*(16), 4715–4747. https://doi.org/10.5194/gmd-16-4715-2023

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

Merz, R., Parajka, J., & Blöschl, G. (2009). Scale effects in conceptual hydrological modeling. *Water Resources Research*, *45*(9). https://doi.org/10.1029/2009WR007872

Meyer, J., Kohn, I., Stahl, K., Hakala, K., Seibert, J., & Cannon, A. J. (2019). Effects of univariate and multivariate bias correction on hydrological impact projections in alpine

catchments. *Hydrology and Earth System Sciences*, *23*(3), 1339–1354. https://doi.org/10.5194/hess-23-1339-2019

Miao, C., Su, L., Sun, Q., & Duan, Q. (2016). A nonstationary bias-correction technique to remove bias in GCM simulations. *Journal of Geophysical Research: Atmospheres*, *121*(10), 5718–5735. https://doi.org/10.1002/2015JD024159

Miara, A., Macknick, J. E., Vörösmarty, C. J., Tidwell, V. C., Newmark, R., & Fekete, B. (2017). Climate and water resource change impacts and adaptation potential for US power supply. *Nature Climate Change*, *7*(11), 793–798. https://doi.org/10.1038/nclimate3417

Miller, S., Ormaza-Zulueta, N., Koppa, N., & Dancer, A. (2025). Statistical downscaling differences strongly alter projected climate damages. *Communications Earth & Environment*, *6*(1), 1–10. https://doi.org/10.1038/s43247-025-02134-2

Milly, P. C. D., Dunne, K. A., & Vecchia, A. V. (2005). Global pattern of trends in streamflow and water availability in a changing climate. *Nature*, *438*(7066), Article 7066. https://doi.org/10.1038/nature04312

Moore, F. C., Lacasse, K., Mach, K. J., Shin, Y. A., Gross, L. J., & Beckage, B. (2022). Determinants of emissions pathways in the coupled climate–social system. *Nature*, *603*(7899), 103–111. https://doi.org/10.1038/s41586-022-04423-8

Mote, P. W. (2003). Trends in snow water equivalent in the Pacific Northwest and their climatic causes: TRENDS IN SNOW WATER EQUIVALENT. *Geophysical Research Letters*, *30*(12). https://doi.org/10.1029/2003GL017258

Mourato, S., Moreira, M., & Corte-Real, J. (2015). Water Resources Impact Assessment Under Climate Change Scenarios in Mediterranean Watersheds. *Water Resources Management*, *29*(7), 2377–2391. https://doi.org/10.1007/s11269-015-0947-5

Muerth, M. J., Gauvin St-Denis, B., Ricard, S., Velázquez, J. A., Schmid, J., Minville, M., Caya, D., Chaumont, D., Ludwig, R., & Turcotte, R. (2013). On the need for bias correction in regional climate scenarios to assess climate change impacts on river runoff. *Hydrology and Earth System Sciences*, *17*(3), 1189–1204. https://doi.org/10.5194/hess-17-1189-2013

Muzik, I. (2001). Sensitivity of Hydrologic Systems to Climate Change. *Canadian Water Resources Journal*, *26*(2), 233–252. https://doi.org/10.4296/cwrj2602233

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Nasseri, M., Tavakol-Davani, H., & Zahraie, B. (2013). Performance assessment of different data mining methods in statistical downscaling of daily precipitation. *Journal of Hydrology*, *492*, 1–14. https://doi.org/10.1016/j.jhydrol.2013.04.017

Nesru, M. (2023). A review of model selection for hydrological studies. *Arabian Journal of Geosciences*, *16*(2), 102. https://doi.org/10.1007/s12517-023-11194-7

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R. J., Blodgett, D., Brekke, L., Arnold, J. R., Hopson, T., & Duan, Q. (2015). Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: Data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, *19*(1), 209–223. https://doi.org/10.5194/hess-19-209-2015

Newman, A. J., Mizukami, N., Clark, M. P., Wood, A. W., Nijssen, B., & Nearing, G. (2017). Benchmarking of a Physically Based Hydrologic Model. *Journal of Hydrometeorology*, *18*(8), 2215–2225. https://doi.org/10.1175/JHM-D-16-0284.1

Nguyen, P. L., Alexander, L. V., Thatcher, M. J., Truong, S. C. H., Isphording, R. N., & McGregor, J. L. (2024). Selecting CMIP6 global climate models (GCMs) for Coordinated Regional Climate Downscaling Experiment (CORDEX) dynamical downscaling over Southeast Asia using a standardised benchmarking framework. *Geoscientific Model Development*, *17*(19), 7285–7315. https://doi.org/10.5194/gmd-17-7285-2024

Nijsse, F. J. M. M., Cox, P. M., & Williamson, M. S. (2020). Emergent constraints on transient climate response (TCR) and equilibrium climate sensitivity (ECS) from historical warming in CMIP5 and CMIP6 models. *Earth System Dynamics*, *11*(3), 737–750. https://doi.org/10.5194/esd-11-737-2020

Ott, I., Duethmann, D., Liebert, J., Berg, P., Feldmann, H., Ihringer, J., Kunstmann, H., Merz, B., Schaedler, G., & Wagner, S. (2013). High-Resolution Climate Change Impact Analysis on Medium-Sized River Catchments in Germany: An Ensemble Assessment. *Journal of Hydrometeorology*, *14*(4), 1175–1193. https://doi.org/10.1175/JHM-D-12-091.1

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., & Loumagne, C. (2005). Which potential evapotranspiration input for a lumped rainfall–runoff model?: Part 2—Towards a simple and efficient potential evapotranspiration model for rainfall–runoff modelling. *Journal of Hydrology*, *303*(1), 290–306. https://doi.org/10.1016/j.jhydrol.2004.08.026

Palmer, T. E., McSweeney, C. F., Booth, B. B. B., Priestley, M. D. K., Davini, P., Brunner, L., Borchert, L., & Menary, M. B. (2022). *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

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

Pandi, D., Kothandaraman, S., & Kuppusamy, M. (2021). Hydrological models: A review. *International Journal of Hydrology Science and Technology*, *12*(3), 223–242. https://doi.org/10.1504/IJHST.2021.117540

Paniconi, C., & Putti, M. (2015). Physically based modeling in catchment hydrology at 50: Survey and outlook. *Water Resources Research*, *51*(9), 7090–7129. https://doi.org/10.1002/2015WR017780

Parding, K. M., Dobler, A., McSweeney, C. F., Landgren, O. A., Benestad, R., Erlandsen, H. B., Mezghani, A., Gregow, H., Räty, O., Viktor, E., El Zohbi, J., Christensen, O. B., & Loukos, H. (2020). GCMeval – An interactive tool for evaluation and selection of climate model ensembles. *Climate Services*, *18*, 100167. https://doi.org/10.1016/j.cliser.2020.100167

Perkins, S. E., Pitman, A. J., Holbrook, N. J., & McAneney, J. (2007). *Evaluation of the AR4 Climate Models' Simulated Daily Maximum Temperature, Minimum Temperature, and Precipitation over Australia Using Probability Density Functions*. https://doi.org/10.1175/JCLI4253.1

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

Potter, N. J., Chiew, F. H. S., Charles, S. P., Fu, G., Zheng, H., & Zhang, L. (2020). Bias in dynamically downscaled rainfall characteristics for hydroclimatic projections.

Hydrology and Earth System Sciences, *24*(6), 2963–2979. https://doi.org/10.5194/hess-24-2963-2020

Poulin, A., Brissette, F., Leconte, R., Arsenault, R., & Malo, J.-S. (2011). Uncertainty of hydrological modelling in climate change impact studies in a Canadian, snow-dominated river basin. *Journal of Hydrology*, *409*(3–4), 626–636. https://doi.org/10.1016/j.jhydrol.2011.08.057

Prajapati, S., Sabokruhie, P., Brinkmann, M., & Lindenschmidt, K.-E. (2023). 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

Prein, A. F., Bukovsky, M. S., Mearns, L. O., Bruyère, C. L., & Done, J. M. (2019). Simulating North American Weather Types With Regional Climate Models. *Frontiers in Environmental Science*, *7*. https://doi.org/10.3389/fenvs.2019.00036

Prein, A. F., Rasmussen, R., & Stephens, G. (2017). *Challenges and Advances in Convection-Permitting Climate Modeling*. https://doi.org/10.1175/BAMS-D-16-0263.1

Qian, Y., Jackson, C., Giorgi, F., Booth, B., Duan, Q., Forest, C., Higdon, D., Hou, Z. J., & Huerta, G. (2016). *Uncertainty Quantification in Climate Modeling and Projection*. https://doi.org/10.1175/BAMS-D-15-00297.1

Quintana Seguí, P., Ribes, A., Martin, E., Habets, F., & Boé, J. (2010). Comparison of three downscaling methods in simulating the impact of climate change on the hydrology of Mediterranean basins. *Journal of Hydrology*, *383*(1–2), 111–124. https://doi.org/10.1016/j.jhydrol.2009.09.050

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

Raju, K. S., & Kumar, D. N. (2014). Ranking of global climate models for India using multicriterion analysis. *Climate Research*, *60*(2), 103–117. https://doi.org/10.3354/cr01222

Räty, O., Räisänen, J., & Ylhäisi, J. S. (2014). Evaluation of delta change and bias correction methods for future daily precipitation: Intermodel cross-validation using ENSEMBLES simulations. *Climate Dynamics*, *42*(9), 2287–2303. https://doi.org/10.1007/s00382-014-2130-8

Reed, S., Koren, V., Smith, M., Zhang, Z., Moreda, F., Seo, D.-J., & Dmip Participants, A. (2004). Overall distributed model intercomparison project results. *Journal of Hydrology*, *298*(1–4), 27–60. https://doi.org/10.1016/j.jhydrol.2004.03.031

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

Ribes, A., Qasmi, S., & Gillett, N. P. (2021). Making climate projections conditional on historical observations. *Science Advances*, *7*(4), eabc0671. https://doi.org/10.1126/sciadv.abc0671

Riboust, P., Thirel, G., Moine, N. L., & Ribstein, P. (2019). Revisiting a Simple Degree-Day Model for Integrating Satellite Data: Implementation of Swe-Sca Hystereses. *Journal of Hydrology and Hydromechanics*, *67*(1), 70–81. https://doi.org/10.2478/johh-2018-0004

Ridder, N. N., Pitman, A. J., & Ukkola, A. M. (2021). Do CMIP6 Climate Models Simulate Global or Regional Compound Events Skillfully? *Geophysical Research Letters*, *48*(2), e2020GL091152. https://doi.org/10.1029/2020GL091152

Robin, Y., Vrac, M., Naveau, P., & Yiou, P. (2019). Multivariate stochastic bias corrections with optimal transport. *Hydrology and Earth System Sciences*, *23*(2), 773–786. https://doi.org/10.5194/hess-23-773-2019

Ross, A. C., & Najjar, R. G. (2019). Evaluation of methods for selecting climate models to simulate future hydrological change. *Climatic Change*, *157*(3–4), 407–428. https://doi.org/10.1007/s10584-019-02512-8

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

Rupp, D. E., Abatzoglou, J. T., Hegewisch, K. C., & Mote, P. W. (2013). Evaluation of CMIP5 20th century climate simulations for the Pacific Northwest USA. *Journal of Geophysical Research: Atmospheres*, *118*(19), 10,884-10,906. https://doi.org/10.1002/jgrd.50843

Saavedra, D., Mendoza, P. A., Addor, N., Llauca, H., & Vargas, X. (2022). A multi-objective approach to select hydrological models and constrain structural uncertainties for climate impact assessments. *Hydrological Processes*, *36*(1), e14446. https://doi.org/10.1002/hyp.14446

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

Salehie, O., Hamed, M. M., Ismail, T. bin, Tam, T. H., & Shahid, S. (2023). Selection of CMIP6 GCM with projection of climate over the Amu Darya River Basin. *Theoretical and Applied Climatology*, *151*(3), 1185–1203. https://doi.org/10.1007/s00704-022-04332-w

Sanderson, B. M., & Knutti, R. (2012). On the interpretation of constrained climate model ensembles. *Geophysical Research Letters*, *39*(16). https://doi.org/10.1029/2012GL052665

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

Sansom, P. G., Stephenson, D. B., Ferro, C. A. T., Zappa, G., & Shaffrey, L. (2013). *Simple Uncertainty Frameworks for Selecting Weighting Schemes and Interpreting Multimodel Ensemble Climate Change Experiments*. https://doi.org/10.1175/JCLI-D-12-00462.1

Schmith, T., Thejll, P., Berg, P., Boberg, F., Christensen, O. B., Christiansen, B., Christensen, J. H., Madsen, M. S., & Steger, C. (2021). Identifying robust bias adjustment methods for European extreme precipitation in a multi-model pseudo-reality setting. *Hydrology and Earth System Sciences*, *25*(1), 273–290. https://doi.org/10.5194/hess-25-273-2021

Seiller, G., Anctil, F., & Perrin, C. (2012). Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions. *Hydrology and Earth System Sciences*, *16*(4), 1171–1189. https://doi.org/10.5194/hess-16-1171-2012

Semenov, M., & Stratonovitch, P. (2010). Use of multi-model ensembles from global climate models for assessment of climate change impacts. *Climate Research*, *41*, 1–14. https://doi.org/10.3354/cr00836

Senatore, A., Fuoco, D., Maiolo, M., Mendicino, G., Smiatek, G., & Kunstmann, H. (2022). Evaluating the uncertainty of climate model structure and bias correction on the hydrological impact of projected climate change in a Mediterranean catchment. *Journal of Hydrology: Regional Studies*, *42*, 101120. https://doi.org/10.1016/j.ejrh.2022.101120

Seo, S. B., & Kim, Y.-O. (2018). Impact of Spatial Aggregation Level of Climate Indicators on a National-Level Selection for Representative Climate Change Scenarios. *Sustainability*, *10*(7), 2409. https://doi.org/10.3390/su10072409

Seo, S. B., Kim, Y.-O., Kim, Y., & Eum, H.-I. (2019). Selecting climate change scenarios for regional hydrologic impact studies based on climate extremes indices. *Climate Dynamics*, *52*(3–4), 1595–1611. https://doi.org/10.1007/s00382-018-4210-7

Shen, M., Chen, J., Zhuan, M., Chen, H., Xu, C.-Y., & Xiong, L. (2018). Estimating uncertainty and its temporal variation related to global climate models in quantifying climate change impacts on hydrology. *Journal of Hydrology*, *556*, 10–24. https://doi.org/10.1016/j.jhydrol.2017.11.004

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 der Heydt, A. S., Knutti, R., Mauritsen, T., … Zelinka, M. D. (2020). An Assessment of Earth's Climate Sensitivity Using Multiple Lines of Evidence. *Reviews of Geophysics*, *58*(4), e2019RG000678. https://doi.org/10.1029/2019RG000678

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

Shiogama, H., Hayashi, M., Hirota, N., & Ogura, T. (2024). Emergent Constraints on Future Changes in Several Climate Variables and Extreme Indices from Global to Regional Scales. *SOLA*, *20*(0), 122–129. https://doi.org/10.2151/sola.2024-017

Shiogama, H., Ishizaki, N. N., Hanasaki, N., Takahashi, K., Emori, S., Ito, R., Nakaegawa, T., Takayabu, I., Hijioka, Y., Takayabu, Y. N., & Shibuya, R. (2021). Selecting CMIP6-

Based Future Climate Scenarios for Impact and Adaptation Studies. *SOLA*, *17*(0), 57–62. https://doi.org/10.2151/sola.2021-009

Shiogama, H., Takakura, J., & Takahashi, K. (2022). Uncertainty constraints on economic impact assessments of climate change simulated by an impact emulator. *Environmental Research Letters*, *17*(12), 124028. https://doi.org/10.1088/1748-9326/aca68d

Shiogama, H., Watanabe, M., Kim, H., & Hirota, N. (2022). Emergent constraints on future precipitation changes. *Nature*, *602*(7898), 612–616. https://doi.org/10.1038/s41586-021-04310-8

Siabi, E. K., Awafo, E. A., Kabo-bah, A. T., Derkyi, N. S. A., Akpoti, K., Mortey, E. M., & Yazdanie, M. (2023). Assessment of Shared Socioeconomic Pathway (SSP) climate scenarios and its impacts on the Greater Accra region. *Urban Climate*, *49*, 101432. https://doi.org/10.1016/j.uclim.2023.101432

Singh, S. K., & Bárdossy, A. (2012). Calibration of hydrological models on hydrologically unusual events. *Advances in Water Resources*, *38*, 81–91. https://doi.org/10.1016/j.advwatres.2011.12.006

Smirnov, O., Zhang, M., Xiao, T., Orbell, J., Lobben, A., & Gordon, J. (2016). The relative importance of climate change and population growth for exposure to future extreme droughts. *Climatic Change*, *138*(1), 41–53. https://doi.org/10.1007/s10584-016-1716-z

Smith, C. J., Harris, G. R., Palmer, M. D., Bellouin, N., Collins, W., Myhre, G., Schulz, M., Golaz, J. -C., Ringer, M., Storelvmo, T., & Forster, P. M. (2021). Energy Budget Constraints on the Time History of Aerosol Forcing and Climate Sensitivity. *Journal of Geophysical Research: Atmospheres*, *126*(13). https://doi.org/10.1029/2020JD033622

Srinivasa Raju, K., & Nagesh Kumar, D. (2015). Ranking general circulation models for India using TOPSIS. *Journal of Water and Climate Change*, *6*(2), 288–299. https://doi.org/10.2166/wcc.2014.074

Srinivasa Raju, K., & Nagesh Kumar, D. (2016). Selection of global climate models for India using cluster analysis. *Journal of Water and Climate Change*, *7*(4), 764–774. https://doi.org/10.2166/wcc.2016.112

Stephens, C. M., Marshall, L. A., & Johnson, F. M. (2019). Investigating strategies to improve hydrologic model performance in a changing climate. *Journal of Hydrology*, *579*, 124219. https://doi.org/10.1016/j.jhydrol.2019.124219

Su, T., Chen, J., Cannon, A. J., Xie, P., & Guo, Q. (2020). Multi-site bias correction of climate model outputs for hydro-meteorological impact studies: An application over a watershed in China. *Hydrological Processes*, *34*(11), 2575–2598. https://doi.org/10.1002/hyp.13750

Sung, J. H., Kwon, M., Jeon, J.-J., & Seo, S. B. (2019). A Projection of Extreme Precipitation Based on a Selection of CMIP5 GCMs over North Korea. *Sustainability*, *11*(7), 1976. https://doi.org/10.3390/su11071976

Tabari, H. (2020). Climate change impact on flood and extreme precipitation increases with water availability. *Scientific Reports*, *10*(1), Article 1. https://doi.org/10.1038/s41598-020-70816-2

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

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An Overview of CMIP5 and the Experiment Design. *Bulletin of the American Meteorological Society*, *93*(4), 485–498. https://doi.org/10.1175/BAMS-D-11-00094.1

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

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

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

Tolson, B. A., & Shoemaker, C. A. (2007). Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resources Research*, *43*(1). https://doi.org/10.1029/2005WR004723

Trenberth, K. (2011). Changes in precipitation with climate change. *Climate Research*, *47*(1), 123–138. https://doi.org/10.3354/cr00953

Tripathy, K. P., & Mishra, A. K. (2024). Deep learning in hydrology and water resources disciplines: Concepts, methods, applications, and research directions. *Journal of Hydrology*, *628*, 130458. https://doi.org/10.1016/j.jhydrol.2023.130458

Troin, M., Martel, J.-L., Arsenault, R., & Brissette, F. (2022). Large-sample study of uncertainty of hydrological model components over North America. *Journal of Hydrology*, *609*, 127766. https://doi.org/10.1016/j.jhydrol.2022.127766

Trzaska, S., & Schnarr, E. (2014). *A Review of Downscaling Methods for Climate Change Projections*.

U.S. Geological Survey. (2016). *U.S. Geological Survey. (2016). National Water Information System data available on the World Wide Web (USGS Water Data for the Nation). Retrieved from http://waterdata.usgs.gov/nwis/. Doi:10.5066/F7P55KJN*. https://www.usgs.gov/national-hydrography/access-national-hydrography-products

U.S. Geological Survey. National Hydrography Dataset. Retrieved from https://www.usgs.gov/core-science-systems/ngp/nationalhydrography/access-national-hydrography-products. (2019). https://www.usgs.gov/national-hydrography/access-national-hydrography-products

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

Van Lanen, H. A. J., Van Loon, A. F., Wanders, N., & Prudhomme, C. (2024). Process-based modelling. In *Hydrological Drought* (pp. 427–476). Elsevier. https://doi.org/10.1016/B978-0-12-819082-1.00019-9

Vano, J. A., Kim, J. B., Rupp, D. E., & Mote, P. W. (2015). Selecting climate change scenarios using impact-relevant sensitivities. *Geophysical Research Letters*, *42*(13), 5516–5525. https://doi.org/10.1002/2015GL063208

Vano, J. A., Udall, B., Cayan, D. R., Overpeck, J. T., Brekke, L. D., Das, T., Hartmann, H. C., Hidalgo, H. G., Hoerling, M., McCabe, G. J., Morino, K., Webb, R. S., Werner, K., & Lettenmaier, D. P. (2014). *Understanding Uncertainties in Future Colorado River Streamflow*. https://doi.org/10.1175/BAMS-D-12-00228.1

Velázquez, J. A., Schmid, J., Ricard, S., Muerth, M. J., Gauvin St-Denis, B., Minville, M., Chaumont, D., Caya, D., Ludwig, R., & Turcotte, R. (2013). An ensemble approach to assess hydrological models' contribution to uncertainties in the analysis of climate change impact on water resources. *Hydrology and Earth System Sciences*, *17*(2), 565–578. https://doi.org/10.5194/hess-17-565-2013

Vetter, T., Reinhardt, J., Flörke, M., van Griensven, A., Hattermann, F., Huang, S., Koch, H., Pechlivanidis, I. G., Plötner, S., Seidou, O., Su, B., Vervoort, R. W., & Krysanova, V. (2017). Evaluation of sources of uncertainty in projected hydrological changes under climate change in 12 large-scale river basins. *Climatic Change*, *141*(3), 419–433. https://doi.org/10.1007/s10584-016-1794-y

Vieux, B. E., Cui, Z., & Gaur, A. (2004). Evaluation of a physics-based distributed hydrologic model for flood forecasting. *Journal of Hydrology*, *298*(1–4), 155–177. https://doi.org/10.1016/j.jhydrol.2004.03.035

Vrac, M. (2018). Multivariate bias adjustment of high-dimensional climate simulations: The Rank Resampling for Distributions and Dependences (R^2 D^2) bias correction. *Hydrology and Earth System Sciences*, *22*(6), 3175–3196. https://doi.org/10.5194/hess-22-3175-2018

Wan, Y., Chen, J., Xu, C.-Y., Xie, P., Qi, W., Li, D., & Zhang, S. (2021). Performance dependence of multi-model combination methods on hydrological model calibration strategy and ensemble size. *Journal of Hydrology*, *603*, 127065. https://doi.org/10.1016/j.jhydrol.2021.127065

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., Cannon, A. J., Xu, C.-Y., & Chen, H. (2018). Transferability of climate simulation uncertainty to hydrological impacts. *Hydrology and Earth System Sciences*, *22*(7), 3739–3759. https://doi.org/10.5194/hess-22-3739-2018

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/10.1175/2010JCLI3594.1

Wenzel, S., Eyring, V., Gerber, E. P., & Karpechko, A. Y. (2016). Constraining Future Summer Austral Jet Stream Positions in the CMIP5 Ensemble by Process-Oriented Multiple Diagnostic Regression. *Journal of Climate*, *29*(2), 673–687. https://doi.org/10.1175/JCLI-D-15-0412.1

Westra, S., Fowler, H. J., Evans, J. P., Alexander, L. V., Berg, P., Johnson, F., Kendon, E. J., Lenderink, G., & Roberts, N. M. (2014). Future changes to the intensity and frequency of short-duration extreme rainfall. *Reviews of Geophysics*, *52*(3), 522–555. https://doi.org/10.1002/2014RG000464

Whitehead, P. G., Wilby, R. L., Battarbee, R. W., Kernan, M., & Wade, A. J. (2009). A review of the potential impacts of climate change on surface water quality. *Hydrological Sciences Journal*, *54*(1), 101–123. https://doi.org/10.1623/hysj.54.1.101

Why So Many Climate Models? | California Climate Commons. (n.d.). Retrieved 13 July 2025, from http://climate.calcommons.org/article/why-so-many-climate-models

Wilby, R. L., & Dessai, S. (2010). Robust adaptation to climate change. *Weather*, *65*(7), 180–185. https://doi.org/10.1002/wea.543

Wilby, R. L., & Harris, I. (2006). A framework for assessing uncertainties in climate change impacts: Low-flow scenarios for the River Thames, UK. *Water Resources Research*, *42*(2). https://doi.org/10.1029/2005wr004065

Wilby, R. L., Troni, J., Biot, Y., Tedd, L., Hewitson, B. C., Smith, D. M., & Sutton, R. T. (2009). A review of climate risk information for adaptation and development planning. *International Journal of Climatology*, *29*(9), 1193–1215. https://doi.org/10.1002/joc.1839

Wilby, R. L., Wigley, T. M. L., Conway, D., Jones, P. D., Hewitson, B. C., Main, J., & Wilks, D. S. (1998). Statistical downscaling of general circulation model output: A comparison of methods. *Water Resources Research*, *34*(11), 2995–3008. https://doi.org/10.1029/98WR02577

Wilcke, R. A. I., & Bärring, L. (2016). Selecting regional climate scenarios for impact modelling studies. *Environmental Modelling & Software*, *78*, 191–201. https://doi.org/10.1016/j.envsoft.2016.01.002

Williams, M. W., Erickson, T. A., & Petrzelka, J. L. (2010). Visualizing meltwater flow through snow at the centimetre-to-metre scale using a snow guillotine. *Hydrological Processes*, *24*(15), 2098–2110. https://doi.org/10.1002/hyp.7630

Woodhouse, C. A., Pederson, G. T., Morino, K., McAfee, S. A., & McCabe, G. J. (2016). Increasing influence of air temperature on upper Colorado River streamflow. *Geophysical Research Letters*, *43*(5), 2174–2181. https://doi.org/10.1002/2015GL067613

Wootten, A., Massoud, E., Waliser, D., & Lee, H. (2022). *To weight or not to weight: Assessing sensitivities of climate model weighting to multiple methods, variables, and domains* [Preprint]. Dynamics of the Earth system: models. https://doi.org/10.5194/esd-2022-15

Worako, A. W., Haile, A. T., & Taye, M. T. (2022). Implication of bias correction on climate change impact projection of surface water resources in the Gidabo Sub-basin, southern Ethiopia. *Papers Published in Journals (Open Access)*, 13(5):2070-2088.

Yue, X.-L., & Gao, Q.-X. (2018). Contributions of natural systems and human activity to greenhouse gas emissions. *Advances in Climate Change Research*, *9*(4), 243–252. https://doi.org/10.1016/j.accre.2018.12.003

Zelinka, M. D., Myers, T. A., McCoy, D. T., Po-Chedley, S., Caldwell, P. M., Ceppi, P., Klein, S. A., & Taylor, K. E. (2020). Causes of Higher Climate Sensitivity in CMIP6 Models. *Geophysical Research Letters*, *47*(1), e2019GL085782. https://doi.org/10.1029/2019GL085782

Zhang, S., Zhou, Z., Peng, P., & Xu, C. (2024). A New Framework for Estimating and Decomposing the Uncertainty of Climate Projections. *Journal of Climate*, *37*(2), 365–384. https://doi.org/10.1175/JCLI-D-23-0064.1

Zhang, Y., Liu, H., Qi, J., Feng, P., Zhang, X., Liu, D. L., Marek, G. W., Srinivasan, R., & Chen, Y. (2023). 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. *Science of The Total Environment*, *857*, 159482. https://doi.org/10.1016/j.scitotenv.2022.159482

Zhao, X., Wang, H., Bai, M., Xu, Y., Dong, S., Rao, H., & Ming, W. (2024). A Comprehensive Review of Methods for Hydrological Forecasting Based on Deep Learning. *Water*, *16*(10), 1407. https://doi.org/10.3390/w16101407

Zhong, L., Lei, H., & Gao, B. (2023). Developing a Physics-Informed Deep Learning Model to Simulate Runoff Response to Climate Change in Alpine Catchments. *Water Resources Research*, *59*(6), e2022WR034118. https://doi.org/10.1029/2022WR034118

Zscheischler, J., Fischer, E. M., & Lange, S. (2019). The effect of univariate bias adjustment on multivariate hazard estimates. *Earth System Dynamics*, *10*(1), 31–43. https://doi.org/10.5194/esd-10-31-2019

Zscheischler, J., & Seneviratne, S. I. (2017). Dependence of drivers affects risks associated with compound events. *Science Advances*, *3*(6), e1700263. https://doi.org/10.1126/sciadv.1700263