AN ASSESSMENT OF SIMULATED RUNOFF FROM GLOBAL MODELS

by

IGNAZIO GIUNTOLI

A thesis submitted to the University of Birmingham for the degree of

DOCTOR OF PHILOSOPHY

School of Geography, Earth and Environmental Sciences
College of Life and Environmental Sciences
University of Birmingham

December 2016

Abstract

This thesis assesses long-term runoff projections form global multi-model ensembles used in hydrological impact studies. Firstly, the study investigates global-scale changes in frequency of high and low flow days towards the end of the current century compared to present day, quantifying the relative contribution to uncertainty from global climate (GCMs) and global impact models (GIMs). Results show increases in high flows for northern latitudes and in low flows for several hotspots worldwide. Overall, GCMs provide the largest uncertainty; but GIMs are the greatest source of uncertainty in snow-dominated regions. GIMs contribution to uncertainty is higher for low flows than for high flows. Secondly, the ability of a set of GIMs to reproduce observed runoff is evaluated at the regional scale. Results indicate that GIMs capture well overall trends in high, medium, and low flows, but differ from observations with respect to the timing of high and medium flows. In particular, GIMs that only include water balance tend to be closer to the observations than GIMs that also include the energy balance. Thirdly, the relative contribution to uncertainty from GCMs, GIMs, Representative Concentration Pathways (RCPs), and internal variability is quantified for transient runoff projections on different ensemble configurations. Over the USA, GCMs and GIMs are responsible for the largest uncertainty, followed by internal variability, which predominates in areas of topographic complexity like the Southwest USA. Culling least credible GIMs from the ensemble has a minor impact on low and medium flows uncertainty; but it has a substantial impact for high flows. Interestingly, regardless of the ensemble setup, RCPs always play a very small role in the uncertainty. In conclusion, efforts to improve (i.e. reduction of uncertainty) multi-model runoff projections from high to low flows should focus on GCMs and GIMs. In particular, GIMs should be evaluated in the region of study, so that ensembles are populated with models that exclude those GIMs that reproduce unrealistic runoff characteristics (e.g., disproportionate null runoff days in the year). This has the potential to yield greater confidence in future runoff projections.

Ma con quali vocavoli potrò io descrivere le nefande e spaventose inondazione, contro alla quale non vale alcuno umano riparo, ma colle gonfiate e superbe onde urina li alti monti, deriva le fortissime argine, disvelle le radicate piante, e colle rapaci onde intorbidate delle cultivate campagne portando con seco le intollerabili fatiche di miseri e stanchi agrecultori, lascia le valli denudate e vili per la lasciata povertà.

Leonardo

Codice Atlantico

Acknowledgements

I acknowledge the financial support from the U.K. Natural Environment Research Council and the School of Geography, Earth, and Environmental Sciences of the University of Birmingham.

I thank very much my supervisors, Christel Prudhomme and David Hannah, for their guidance and support throughout the PhD. I thank them for their encouragement and the helpful discussions, and for guaranteeing a prompt and continual support regardless of geography.

I thank Jean-Philippe Vidal from IRSTEA for hosting me in Lyon on a four-month placement in the fall of 2013, his advice and discussions helped to generate the research presented in Chapter 4. I thank Gabriele Villarini from IIHR University of Iowa for the opportunity to work on a six-month placement in Iowa City in 2014, his timely advice and continued collaboration and support while Iowa and thereafter have helped to generate the research presented in chapter 5 and 6.

I am grateful to the people I have met in the research centres that hosted me throughout the PhD. In particular, in the Centre for Ecology and Hydrology in Wallingford, Emma Robinson and Douglas Clark have been very helpful in the first year of the PhD while I was getting familiar with the global model world and the large data sets environment. I also thank Ilaria Prosdocimi and Marco Girardello for the stimulating discussions and precious suggestions on statistics and other research issues. Ilaria and Marco, together with my office mates Homan Wong and Richard Falk, have been wonderful colleagues to share the studious and hectic days at CEH.

I also thank David Lavers who helped liaising with Gabriele Villarini and his lab, and who has been very helpful for getting acquainted to the Iowan setting. Felipe Quintero and the black hole survivors group have been great companions in Iowa.

I am grateful to the providers of the data used in the thesis: the modelling groups that contributed to the ISI-MIP and WaterMIP projects; the USGS for the stream-flow data.

While in Florence, in the premises of the Italian National Library, Gian Piero Aielli and Matteo Borzoni have provided precious suggestions in statistics and research design.

Finally, the strong dose of optimism (nearing naivety) and the determination to obtain this title would have not sufficed without the precious support of my family and friends who greatly helped me through these years to stand some non ideal conditions (e.g., student status, mingy stipend) I was confronted with.

Table of contents

List of Figures

List of Tables

List of Abbreviations

ANOVA	Analysis Of Variance
AOGCM	Atmosphere-Ocean Global Circulation Model
CO_2	Carbon Dioxide
CRU	Climatic Research Unit
CV	Coefficient of Variation
DJF	December, January and February
ERA-40	ECMWF Re-analysis–40
ESM	Earth System Model
GCM	Global Climate Model
GHM	Global Hydrological Model
GIM	Global Impact Model
GPCC	Global Precipitation Climatology Centre
HDCN	Hydro-Climatic Data Network
IQR	Inter-Quartile Range
IPCC	Intergovernmental Panel on Climate Change
ISI-MIP	Inter-Sectoral Impact Model Intercomparison Project
JJA	June, July and August
LSM	Land Surface Model
MAM	March, April and May
MPI	Max-Planck Institut für Meteorologie
NOAA	National Oceanic and Atmospheric Administration
PDF	Probability Density Function
PE	Potential Evapotranspiration
RMSE	Root Mean Squared Error
S2N	Signal-to-Noise
SON	September, October and November
SST	Sea Surface Temperature
UEA	University of East Anglia
USGS	United States Geological Survey
WATCH	Integrated Project Water and Global Change
WFD	WATCH Forcing Data
WaterMIP	Water Model Intercomparison Project

1 Introduction

1.1 Background and rationale

"The flow and storage of water in the Earth's climate system are highly variable, but changes beyond those due to natural variability are expected by the end of the current century. In a warmer world, there will be net increases in rainfall, surface evaporation and plant transpiration. However, there will be substantial differences in the changes between locations. Some places will experience more precipitation and an accumulation of water on land. In others, the amount of water will decrease, due to regional drying and loss of snow and ice cover." (Collins et al. 2013 - IPCC WG1 5th AR Ch.12).

The hydrological cycle plays a vital role in the well-being of the human society and of the ecosystems. A changing climate and an increasing world population make the knowledge of future hydrology ever more valuable to inform adaptation strategies that will have to deal with unprecedented pressures on food/energy production and exposure to water related hazards (e.g., Lavell et al. 2012). Therefore, adaptation and mitigation decisions need information on potential future changes in the hydrological cycle, and more often at regional rather than global scales (Tebaldi and Smith, 2009). In this context, a valuable contribution for anticipating what the future holds is provided by multi-model ensembles (i.e.: global impact models (GIMs) fed by multiple global climate models (GCMs) simulations under different scenarios), which can simulate future hydrological variables of direct human relevance that account for changes in the climate. However, these global models suffer from uncertainties that originate from different sources (e.g. model structure, incomplete knowledge of the physical processes, external forcing, and initial boundary conditions). These sources of uncertainty are cascaded and amplified throughout the modelling

chain. In the case of extreme events (e.g., flood and drought occurrences), the uncertainty is even greater, as they are more difficult to simulate, partly because they are infrequent and thus poorly sampled, and partly because they tend to be more local and transient and thus more sensitive to differences in the definition of atmospheric quantities between the climate models and the observations (Stone and Knutti, 2010).

Therefore, key priorities for climate impact science are i) to assess hydrological projections from global models not only mean quantities (e.g., mean runoff), but also hydrological extremes; ii) to quantify uncertainties in these projections in an effort to communicate effectively how uncertain are the estimates in climate impact studies and to reduce overall uncertainty.

This thesis focuses on the assessment of runoff projections from global models. In particular, on sets of GIMs participating to multi-model ensemble experiments devoted to exploring the future water cycle: GIMs are used to assess future flows, and then are evaluated using observed data on different runoff characteristics, finally key contributors to uncertainty in future flows are quantified.

1.2 Research gaps and objectives

The literature review on simulated runoff from global models and its uncertainty (Chapter 2) identifies three research gaps for investigation:

1. There are few published studies that assess hydrological extremes jointly both at the global and regional scale using the latest CMIP5 GCMs and multiple Global Impact Models.

2. Multi-model ensembles of Global Impact Models fed by GCMs have proved useful tools for runoff simulation, but their ability to capture different characteristics of the terrestrial hydrology has not been tested outside of Europe.

3. There is a lack of knowledge in the dominant sources of uncertainty that are present in the modelling chain that affect the accuracy and the interpretation of the projections.

In light of these research gaps, the overarching aim of this thesis is to evaluate the ability of multi-model ensembles to simulate runoff and to quantify the uncertainties that lie in their projections. To address the aforementioned gaps, the objectives of this thesis are:

1) To assess the change in high and low flows at the global scale towards the end of the 21st century and the uncertainty share between GIMs and GCMs (Chapter 4).

2) To evaluate the ability of GIMs to simulate runoff in the control period (hindcast) using observed data as benchmark at the regional scale (Chapter 5).

3) To assess at the continental scale the contribution of the different sources of uncertainty (specifically, GCM, GIM, RCP, and internal variability) using transient runs spanning from the beginning to the end of the 21st century (Chapter 6).

1.3 Thesis structure

The thesis structure is depicted in Figure 1.1. Chapter 2 provides the literature review and identifies the research gaps presented in Section 1.2. Chapter 3 presents the research design, introducing the data and the methods used in the research. Chapter 4 investigates the changes in future runoff for high and low flows at the global scale and the uncertainty provided by GCMs and GIMs. Chapter 5 evaluates the ability of GIMs to simulate runoff at the regional scale. A quantification of the different sources of uncertainty over decadal transient hydrological indices is undertaken in Chapter 6. In Chapter 7 conclusions are drawn and suggestions are proposed for future research.

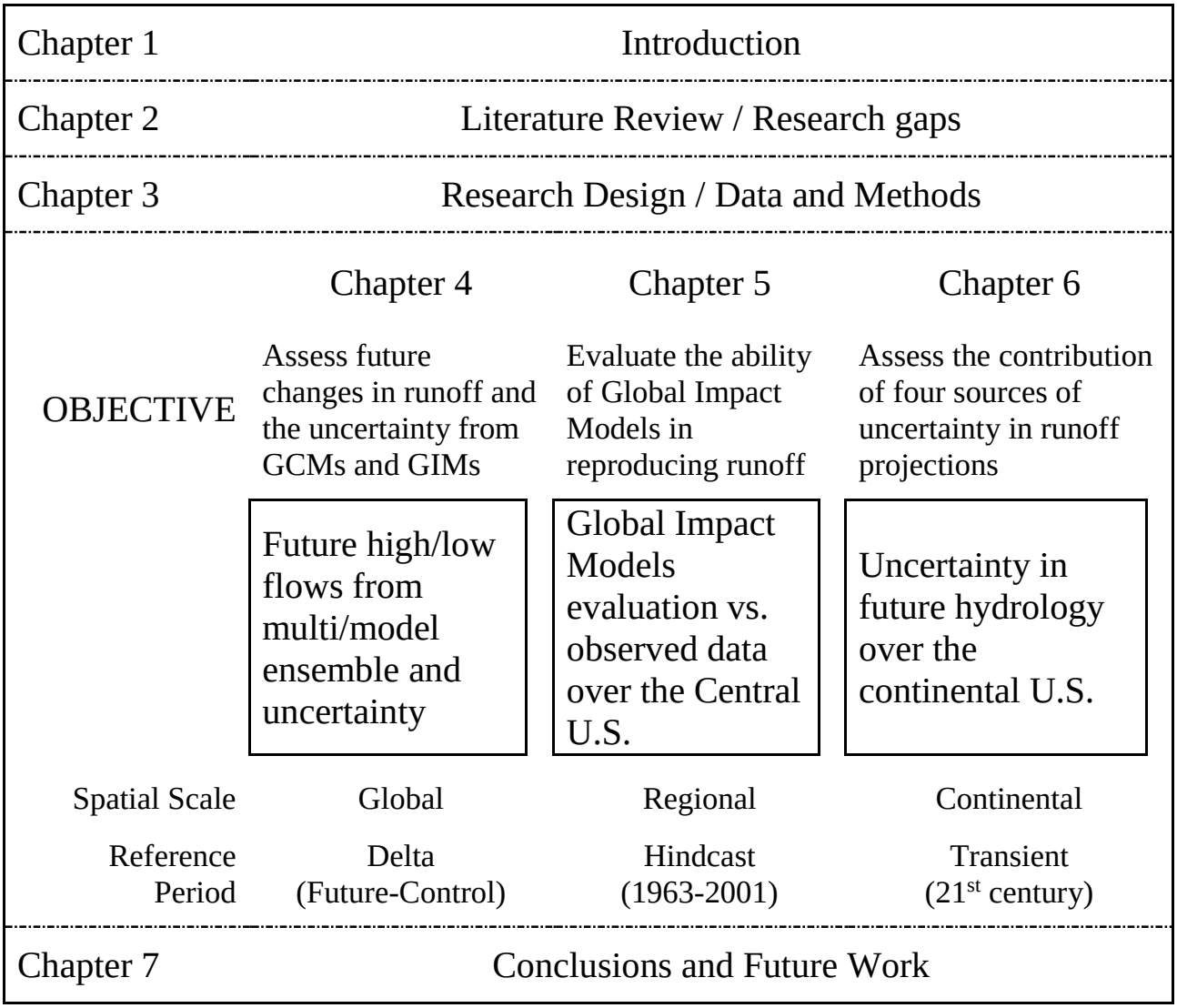


Figure 1.1 – Schematic diagram of the thesis (GCMs: Global Climate Models; GIMs: Global Impact Models).

1.4 Chapter summary

This chapter has provided a background and rationale to the research undertaken in the thesis on runoff simulation from global multi-model ensembles from the regional to the global scale. In Chapter 2 the literature review is presented and research gaps are identified; the thesis objectives are thus outlined to address these gaps in the following chapters.

This thesis is composed of papers published (two) or under review (one) in peer-reviewed journals. In particular, research from Chapter 4 corresponds to an *Earth System Dynamics* paper (APPENDIX V) on the projection of future global high and low flows and the uncertainty coming from GCMs and GIMs. The research in Chapter 5 on the evaluation of GIMs in reproducing observed streamflow was published in the *Journal of Geophysical Research* (APPENDIX VI). Research in Chapter 6 on the quantification of uncertainty sources in transient runoff is currently under review. Elements of these papers are part of Chapter 2 (Introduction section) and Chapter 3 (Data and Methods sections).

Specifically, the author contributions for the papers that compose the thesis are detailed below. For the 2015 *Earth System Dynamics* paper by Giuntoli, I., Vidal, J.-P., Prudhomme, C., Hannah, D.M., titled "Future hydrological extremes: the uncertainty from multiple global climate and global hydrological models". - I.G., J.-P., C.P., D.H. designed the study. - I.G. wrote computer code, performed analyses and prepared the manuscript with contributions from all co-authors. For the 2015 *Journal of Geophysical Research* paper by Giuntoli, I., Villarini, G., Prudhomme, C., Mallakpour, I., Hannah, D.M., titled "Evaluation of global impact models' ability to reproduce runoff characteristics over the central United States". - I.G., G.V., C.P., D.H. designed the study. - I.G. coded and performed analyses (I.M. extracted hydrological indices from observed data) with feedback from G.V., C.P., D.H. - I.G. prepared the manuscript with contributions from all co-authors. For the paper under revision by Giuntoli, I., Villarini, G., Prudhomme, C., Hannah, D.M., with prospective title "Uncertainties in projected runoff over the continental United States". - I.G. and G.V. designed the study. - I.G. coded, performed analyses and prepared the manuscript with feedback from all co-authors.

2 Literature Review and Research objectives

2.1 Introduction

In this chapter a literature review is undertaken to identify research gaps for investigation. The review consists of three main parts. Firstly, a summary is given on continental to global hydrology simulated by multi-model ensembles. Secondly, the importance of evaluating global impact models against observed data is reviewed. Thirdly, the interest of characterizing sources of uncertainty in space and time is discussed. Finally the research gaps identified are put in context with the objectives that the thesis aims to address.

2.2 Future global hydrology from multi-model ensembles

The ongoing intensification of the water cycle at the global scale is expected to continue in the coming decades (Huntington, 2006; Stott et al., 2010). Projected changes in climate variables from global climate models (GCMs) indicate an increase in the frequency of hydrological extremes (Kharin et al., 2013; Seneviratne et al., 2012; Sillmann et al., 2013; Tebaldi et al., 2006). These hydrological shifts go hand in hand with a growing world population that will become ever more vulnerable with respect to access to water and food, and resilience to natural hazards (Lavell et al., 2012). In this context, global multi-model ensembles yield a valuable opportunity for climate projections and impact assessments. In hydrology, multi-model ensemble experiments – consisting of global impact models (GIMs) fed by input forcing simulated by GCMs – can be used to project future changes in the water cycle and future hydrological extremes, using modelled variables such as runoff and soil moisture. In recent years, a number of studies have assessed the future changes in the global water cycle (e.g., Arnell, 2003; Hirabayashi et al., 2008; Milly et al., 2005; Nohara et al.,

2006; Sheffield and Wood, 2008). Although many of these studies have a representative number of GCMs in their ensembles, they rarely comprise more than one GIM – Table 2.1 lists and summarizes studies that have used runoff projections from ensembles comprising one GIM and one or more GCMs. This presents a limitation considering that GIMs provide more uncertainty than previously thought (Haddeland et al., 2011; Hagemann et al., 2013; Prudhomme et al., 2014; Schewe et al., 2014). In addition, the coarse temporal and spatial resolution of the climate signal used in these studies does not reflect well the potential changes in sub-monthly extreme events at the regional and local scale (Forzieri et al., 2014).

Table 2.1 – Overview of macro-scale (continental to global) studies using runoff projections to assess aspects of the future water cycle using multiple GCMs and GIMs, sorted by number of GCMs employed.

Study	Region	Number of GIMs	Number of GCMs	Runoff characteristic
(Hirabayashi et al., 2008)	Global	-	1	Drought and Flood
(Feyen and Dankers, 2009)	Europe	1	1	Drought
(Döll and Schmied, 2012)	Global	1	3	Drought to Flood
(Wanders et al., 2015)	Global	1	5	Drought
(Arnell, 2003)	Global	-	6	Drought to Flood
(Alfieri et al., 2015)	Europe	1	7	Flood
(Hirabayashi et al., 2013)	Global	1	11	Flood
(Milly et al., 2005)	Global	-	12	Mean flow
(Forzieri et al., 2014)	Europe	1	12	Drought
(Zhao and Dai, 2015)	Global	-	14	Drought
(Nohara et al., 2006)	Global	1	19	Mean flow
(Collins et al., 2013)*	Global	-	22	Mean flow
(Arnell and Gosling, 2013)	Global	1	21	Mean flow

*IPCC WG1 Fifth Assessment Report (Ch.12).

Recently, model inter-comparison projects like WaterMIP (Haddeland et al., 2011) and ISI-MIP (Warszawski et al., 2014) have made it possible to include multiple GCMs and GIMs in global impact studies at unprecedented temporal (up to daily) and spatial (0.5) resolution, thereby

providing frameworks for consistent assessments of the terrestrial water cycle. Table 2.2 lists and summarizes studies that have used ensembles with multiple GCMs and multiple GIMs.

In particular, the more recent ISI-MIP dataset has made available model runs at daily temporal scale, and use the most recent scenarios RCPs (representative concentration pathways, Moss et al. (2010); van Vuuren et al. (2011)), which have superseded the previous SRES scenarios (Nakicenovic et al., 2000) used in WaterMIP.

Table 2.2 – Overview of macro-scale (continental to global) studies using runoff projections to assess aspects of the future water cycle using multiple GCMs and GIMs, sorted by number of GIMs employed.

Study	Region	Number of GIMs	Number of GCMs	Runoff characteristic
(Hagemann et al., 2011)	Global	2	3	Mean flow
(Roudier et al., 2015)	Europe	3	5	Drought and Flood
(van Huijgevoort et al., 2014)	Global	5	3	Drought
(Prudhomme et al., 2014)	Global	7	5	Drought
(Wada et al., 2013)	Global	7	5	Irrigation water demand
(Hagemann et al., 2013)	Global	8	3	Mean flow
(Dankers et al., 2013)	Global	9	5	Flood
(Schewe et al., 2014)	Global	11	5	Water scarcity

The ISI-MIP data set has been used to assess future changes in runoff at global and regional scales. Dankers et al. (2013) explored changes in a 30-year return period of river flow showing that flood hazard is projected overall to increase globally, although not uniformly, and that decreases occur mainly in areas where the hydrograph is dominated by spring snowmelt. Schewe et al. (2014) assessed future water scarcity by analysing changes in mean annual runoff together with global population patterns, showing how the number of people living in water scarcity is projected to increase globally. Davie et al. (2013) investigated runoff changes across models by grouping GIMs into hydrological and biome (including CO_2 and vegetation dynamics) models, showing that while

both types agree on the sign of runoff change for most regions of the world (with contrasting exceptions like West Africa where biome models moisten and hydrological models dry), models accounting for varying CO_2 yield more runoff than those with constant CO_2. Prudhomme et al. (2014) examined the future frequency of droughts using a variable threshold method on daily runoff. They identified drought hotspots globally and observed, similarly to Davie et al. (2013), how biome models accounting for varying CO_2 concentrations tend to project more runoff with increasing CO_2, and generally higher increases and decreases of runoff, than the hydrological models. All of these studies emphasize how both GCM and GIM uncertainty contribute to the spread in projected changes in the hydrological cycle. Their findings highlight the importance of including different types of GIMs and GCMs for making comprehensive assessments of uncertainty in climate impact studies.

In this context, modelling-induced uncertainty (i.e. inter-model spread of GCMs and GIMs) has been expressed by looking at the variance across both types of models. For example, Schewe et al. (2014) and Dankers et al. (2013) used the ratio of the variances of GCM and GIM results (for GCMs: variance of the change across all GCMs for each GIM, then averaged over all of the GIMs; and vice versa for GIMs). Similarly, using WaterMIP data, Hagemann et al. (2013) expressed the spread due to the choice of model type using the standard deviation of GCMs and GIMs (for GCMs: the mean across all GIMs for each GCM, and standard deviation of the GCMs; and vice versa for GIMs). Prudhomme et al. (2014) omitted the partition into GCM/GIM and expressed the uncertainty through the signal-to-noise ratio (by grouping results per type of model) in order to infer which global model type in the ensemble brings about highest agreement.

The studies cited above have provided useful knowledge on climate change impacts on the water cycle using the ISI-MIP data set, however, a synthesis of future projections for high and low flows along with a consistent estimation of uncertainties from GCMs and GIMs is still missing.

The literature review of hydrological impact studies found that although multiple GCMs have been employed in ensembles to assess future hydrology, only recently have studies started to use more than one GIM in their ensembles. Moreover, aspects of future hydrology are studied at once, while it is essential to use the same dataset to concurrently explore multiple facets of runoff (low and high flows) and quantify the relative contribution to uncertainty from GCMs and GIMs.

2.3 Evaluation of Global Impact models

As noted in the previous Section 2.2, a valuable contribution in understanding the present and future hydrological processes in the context of climate impact studies is provided by Global Impact Models (GIMs), which allow simulation of the terrestrial water cycle at the global scale. Together with Global Circulation Models (GCMs), GIMs represent the physical processes in the atmosphere and land surface, and operate over relatively long time span (decades), at a coarse spatial resolution (typically 50-250 km), and time step from sub-daily to monthly. Broadly speaking, GIMs focus on simulating the land-surface whereas GCMs focus primarily on the atmosphere (although they generally include some sort of land-surface scheme, usually less sophisticated than that of the GIMs). Regarding the water cycle, the two model families meet at the land-surface/atmosphere interface, which represents the upper boundary for the GIMs and the lower boundary for the GCMs. Therefore, GCMs climate outputs often provide the basis for impact studies in which GIMs consider the interaction of the atmospheric and land-surface component of the water cycle (e.g., Mölders (2005)).

GIMs can be subdivided into two broad categories, which differ in the land-surface parameterizations: i) the Global Hydrological Models (GHMs) have the water budget and lateral

transfer of water as the main interest, requiring a partitioning of precipitation into evapotranspiration, infiltration, interception, storage and runoff to determine the water fluxes within the soil and the groundwater recharge – with lateral transfer over the surface but not at the subsurface; and ii) the Land Surface Models (LSMs) try additionally to close the energy budget and run at sub-daily time steps. With the aim to describe the vertical exchanges of heat, water, and sometimes carbon in considerable details, LSMs need a partitioning for precipitation between the aforementioned processes to determine the partitioning of radiative forcing between soil heat flux and the turbulent fluxes of sensible and latent heat (e.g., Mölders (2005)).

In the recent past, the hydrological impact research community has realized that the uncertainty associated with the GIMs (including model parameterization and structure) could be large and should not be neglected (Prudhomme and Davies, 2008). It has also been recognized that multi-model ensembles are much more robust tools to address the uncertainty associated with climate change impact than single models and hence should be used as much as possible in any climate change assessment work (e.g., Hagemann et al. (2013)). At the local/catchment scale, this is achieved through building hydrological catchment model ensembles (e.g., Smith et al. (2012)) from a wide range of models including simple lumped conceptual models to more complex physically-based distributed models (Beven, 2011). At continental to global scales, this relies on the GIMs, which are in turn, much more complex models that need a careful balance between accounting for the spatial heterogeneity of hydro-climatic processes and the computational burden associated with the multiplication of near-homogeneous areas. Also, differently from basin-scale hydrological models, which are routinely calibrated against observed river discharge, GIMs are usually not calibrated (Müller Schmied et al., 2014) and are instead tuned to set parameter values. For instance, for the MacPDM GIM, tuning involves tests of precipitation datasets and potential

evaporation calculations against long-term average runoff and long-term average within-year runoff patterns (Gosling and Arnell, 2011).

Following the climate community and programs like the Climate Model Intercomparison Project, e.g. phase five, CMIP5, Taylor et al. (2012), the hydrological community has started modelling experiments using different global impact models driven by the same climate forcing. The first such initiative was the WaterMIP project (Haddeland et al., 2011), since followed for example by the ISI-MIP project (Warszawski et al., 2014). As a result the scientific community has now easy access to many multi-impact model ensembles providing information on the possible projections in hydrological variables in the future for the world. Along with ease of access comes the danger of the data being used not appropriately, for example if some members of the ensemble are poor at reproducing some part of the hydrological processes, that could result in misleading interpretation of the projections if caution is not taken. This is because the global models used for experiments such as WaterMIP and ISI-MIP have generally been developed for different purposes – e.g.: water resource availability assessment (GWAVA, WaterGAP), carbon fluxes (LPJ), water and energy fluxes (JULES) – using different protocols for their parameterization and error-reduction, hence likely to have been tested differently for reproducing different processes. Moreover each model run can use a different set-up which is generally not fully published, and it is never guaranteed that the same set-up used to produce the result published in a paper have been used for another simulation. It might therefore not be appropriate to rely on previous assessment to evaluate the skill of a new ensemble. Furthermore, due to the scale and complexity of such global models, their parameterisation requires a long process, much more complex that that required for catchment models. In particular, comprehensive sensitivity testing of all parameters is a very ambitious task seldom undertaken by developers. While not all model codes are available to the research

community to use, it would require a huge (unrealistic) effort for someone not familiar with those models to undertake a uniform parameterisation testing for all global impact models together.

To improve our confidence in the GIMs, namely in climate impact studies, a necessary first step is the evaluation of the models' ability to reproduce the observational records. On this issue, Prudhomme et al. (2011) emphasized how an appraisal of the performance of large-scale models in replicating historical hydrological extremes is a necessary precursor to assessing the suitability of such models for projecting characteristics of hydrological extremes into the 21st century.

Model intercomparison frameworks like the aforementioned Water Model Intercomparison Project – WaterMIP, provide the opportunity to compare model simulations from a number of GIMs all driven with the same meteorological forcing: the WATCH Forcing Data – WFD (Weedon et al., 2011). The WaterMIP GIMs have been evaluated with respect to low, medium and high flow in a number of studies (Gudmundsson et al., 2012b; Haddeland et al., 2011; Prudhomme et al., 2011; Stahl et al., 2012; Tallaksen and Stahl, 2014; Van Loon et al., 2012) showing considerable variability in the magnitude and timing of the components of the hydrological cycle. Notably, all of these studies focused on Europe, despite the global coverage of the WaterMIP dataset. Little is known about the skill of these models in reproducing the hydrological processes for other regions of the world. Table 2.3 lists and summarizes studies that have assessed GIMs' simulated runoff against observed data.

The evaluation of the GIMs against observed data is not undertaken with the ISI-MIP runs because, with the forcing employed, these can reproduce the frequency but not the chronological occurrence of extreme events like the WaterMIP runs, which are forced with the WFD dataset and provide a valuable benchmark for the comparison.

Table 2.3 – Overview of macro-scale (continental to global) studies assessing simulated runoff from multiple GIMs, sorted by number of GIMs.

Study	Region	Number of GIMs	Forcing data	Reference period	Runoff characteristic
(Prudhomme et al., 2011)	Europe	3	WFD	1963-2001	Drought and Flood
(Wang et al., 2009)	Cont. U.S.A.	6	A2005	1920-2003	Drought
(Tallaksen and Stahl, 2014)	Europe	7	WFD	1963-2000	Drought
(Yang et al., 2015)	Global	7	TRENDY	1981-2010	Mean flow
(Stahl et al., 2012)	Europe	8	WFD	1963-2000	Drought to Flood
(Gudmundsson et al., 2012a)	Europe	9	WFD	1963-2000	Drought to Flood
(Gudmundsson et al., 2012b)	Europe	9	WFD	1963-2000	Mean flow
(van Huijgevoort et al., 2013)	Global	10	WFD	1963-2001	Drought
(van Loon et al., 2012)	Europe	10	WFD	1963-2000	Drought
(Haddeland et al., 2011)	Global	11	WFD	1985-1999	Mean flow

WFD = Watch Forcing Data. A2005 = same forcing as (Andreadis et al., 2005).

The use of multiple GIMs in hydrological impact studies should be accompanied by an evaluation of the models against observed data. This operation allows for an improved use/selection of GIMs in ensemble projects but is rarely undertaken as it is time consuming and as data is scarce in many areas of the world. The literature reviewed indicates that global data sets like WaterMIP, have not been evaluated outside of Europe in a systematic way.

2.4 Uncertainties in runoff projections

The utility of climate impact studies using runoff projections from multi-model ensemble experiments like WaterMIP (Haddeland et al., 2011) and ISI-MIP (Warszawski et al., 2014), is undermined by the large uncertainties that originate in the different components of the modelling chain. There is indeed a consensus on the growing need to well characterize uncertainty both to inform the selection/design of multi-model ensembles and to improve the components of the modelling chain (e.g., Northrop 2013). Uncertainty in climate projections comes from three main

sources: the models, the scenarios, and the internal variability (Hawkins and Sutton, 2009; Tebaldi and Knutti, 2007). Model uncertainty, or response uncertainty, results from different models yielding different responses to the same external forcing owing to differences in the physical and numerical formulations employed. Scenario uncertainty originates from the limited knowledge of the external factors that influence the climate system, for instance trajectories of greenhouse gases, land use change, ozone concentrations in the stratosphere. Internal variability is the natural variability of the climate system without external forcing, and it is caused by non-linear dynamical processes intrinsic to the atmosphere, the ocean, and the coupled ocean-atmosphere system (e.g., Deser et al. 2012b).

Dominant sources of uncertainty in climate projections depend on the variable of interest. Precipitation projections are generally dominated by global climate model (GCM) uncertainty and internal variability rather than scenarios (e.g., Hawkins and Sutton 2011; Deser et al. 2012b; Pendergrass et al. 2015). For runoff projections, while GCMs play a large role, the global impact models (GIMs) can outweigh the GCMs in the contribution to uncertainty, especially in those areas where storage-release processes (e.g., snow-ice) present a challenge (Giuntoli et al., 2015a; Hagemann et al., 2013; Tallaksen and Stahl, 2014; van Huijgevoort et al., 2013). In particular, the choice of RCP has a more systematic impact on temperature than on precipitation change. For instance, at the decadal scale: for temperature Hawkins and Sutton (2009) found that model uncertainty dominates before ~2040, after which scenario uncertainty becomes the leading source; for precipitation, in a later study Hawkins and Sutton (2011) found that internal variability dominates at the beginning and that after the first few decades model uncertainty becomes dominant. Little is known about the contribution to the uncertainty in runoff projections coming from internal variability and scenarios as well as the interplay with the other two sources (i.e., GCMs and GIMs) throughout the 21st century, although we can expect that the RCP has a marginal

role in uncertainty contribution as hinted by few recent studies (Arnell and Gosling, 2013; Orlowsky and Seneviratne, 2013; Tang and Lettenmaier, 2012; Wada et al., 2013).

If the uncertainty from model response (GCMs and GIMs) can potentially be reduced through improvement of the models, and if emission scenarios can be better constrained, internal variability is still unlikely to be reduced because of the inherently unpredictable nature of unforced climate fluctuations beyond a decade (Deser et al., 2012a). Because of internal variability, climate projections of different variables can be inherently uncertain in many parts of the world; with locations that are subject to large internal variability (e.g. the city of Seattle, U.S.A. in Deser et al. (2012a)) that are bound to be affected by an irreducible share of uncertainty in climate projections.

Uncertainty is generally characterized by partitioning the variance of the ensemble spread into different components using statistical frameworks like analysis of variance (ANOVA; e.g., Yip et al. 2011; Sansom et al. 2013). When runs with different initial conditions are unavailable (e.g. ISI-MIP daily runoff runs), internal variability can be sampled as a measure of the noise in the projections throughout the runs as in the framework proposed by Hawkins and Sutton (2009) and used in Orlowsky and Seneviratne (2013), among others.

Table 2.4 lists and summarizes studies that have used ensembles with multiple GCMs and multiple GIMs and have assessed uncertainty from these two sources (excluding Wada et al. (2013), who assessed uncertainty transiently throughout the 21st century on irrigated water demand having added the RCP source).

The studies presented in Table 2.4 show that the uncertainty in runoff projections has been assessed with regards to GCMs and GIMs on delta changes (future minus past) that refer to a time in the second half of the 21st century. Differently from the delta change approach, one can assess how the uncertainty evolves transiently throughout the 21st century (as in e.g., Hawkins and Sutton 2011; Hingray and Saïd 2014). In particular, a systematic uncertainty partition of future runoff

considering also RCPs and internal variability is still missing. Of particular interest is the analysis of high, medium and low flows jointly, showing how uncertainties differ across indices.

Table 2.4 – Overview of macro-scale (continental to global) studies assessing uncertainty in runoff projections from multiple GIMs and GCMs.

Study	Region	Uncertainty partition method	Number of sources	Period
(Roudier et al., 2015)	Europe	Interquartile range	2	Future-past
(Prudhomme et al., 2014)	Global	Signal-to-noise from GCM and GIM	2	Future-past
(Wada et al., 2013)	Global	Analysis of variance; fractional change	3	21st c. transient
(Hagemann et al., 2013)	Global	Spread owed to GCM, GIM, scenario	3	Future-past
(Dankers et al., 2013)	Global	Variance ratio GCM to GIM	2	Future-past
(Schewe et al., 2014)	Global	Variance ratio GCM to GIM	2	Future-past

The analyses presented in Chapter 4 and Chapter 5, have yielded an improved understanding of the GIMs in their ability to simulate runoff. This allows, in the third and final part of the thesis (Chapter 6), to examine how different ensemble configurations affect the uncertainty partition by: i) culling models (as discussed in Overland et al. (2011) and Thibeault and Seth (2014) among others) on the basis of credibility in medium/low flows representation and type (biome models, which include CO_2 and vegetation dynamics); ii) excluding intermediate RCPs (4.5 and 6.0). This is done to assess to what extent the uncertainty share changes using all of the available GIMs (e.g., Dankers et al. 2013, Schewe et al. 2014) or a subset (e.g., Prudhomme et al. 2014, Giuntoli et al. 2015a), and fewer RCPs with the aim to suggest a better use of resources making model runs and an improved choice of ensembles for future climate impact studies.

Any climate impact study is inherently affected by a degree of uncertainty that originates in the different components of the modelling chain (e.g., GCM, bias-correction, GIM, scenario, natural variability). A quantification of the uncertainty is essential for both improving the models and for providing useful information to end users on how large the uncertainty range is in the projections they are using for designing appropriate climate adaptation and mitigation policies. As hinted in few recent studies, RCP contribution to uncertainty in future water cycle is relatively small compared to e.g., GCMs and GIMs. It is important to undertake a study that explores multiple sources of uncertainty at once for different facets of the runoff spectrum.

2.5 Research objectives

With the aim to address the knowledge gaps identified in the previous sections, this thesis focuses on three research objectives.

2.5.1 Assessment of future changes in runoff from a multi-model ensemble globally

The first objective of the thesis is to assess future changes in runoff at the global scale using a global multi-model ensemble dataset. The use of both ends of the runoff spectrum (high and low flows) will allow for a systematic global assessment that is accompanied by an uncertainty analysis that will show which source is dominant between the GCMs and GIMs. This research objective is addressed in Chapter 4.

2.5.2 Evaluation of the ability of Global Impact Models in reproducing runoff in the Central United States

The second objective of the thesis is to evaluate how GIMs reproduce runoff characteristics over a data-rich region like the Central United States, over which gridded runoff data is compared to observed streamflow data. The use of metrics that span from high to medium and low flows ensures systematic evaluation of these models in their ability to reproduce the different facets of the runoff spectrum. This research is presented in Chapter 5.

2.5.3 Assessment of the contribution to uncertainty in runoff projections over the continental United States.

The third objective of the thesis is to carry out an uncertainty partition in runoff projections considering, in addition to GCMs and GIMs (as in objective 2.5.1), RCP scenarios and internal variability. Moreover, the partition is undertaken transiently throughout the 21st century in order to reveal how the relative contributions to uncertainty evolve over time. As in the previous objectives, the use of metrics that span from high to medium and low flows ensures systematic assessment of uncertainty over the different facets of the runoff spectrum. This research is presented in Chapter 6.

2.5.4 Chapter summary.

This chapter has presented a literature review of the research undertaken on: i) runoff projections from multi-model ensemble experiments; ii) the evaluation of GIMs in reproducing observed runoff; and iii) the partition of uncertainty in runoff projections. The knowledge gaps identified constitute the objectives of this thesis. The next chapter presents the research design and presents the data and the methods used in the thesis.

3 Research design, data and methods

3.1 Introduction

This chapter details the research design of the thesis providing guidance on how the chapters fit together. Thereafter the data and the generic statistical methods used in the analyses presented in chapters 4, 5, and 6 are described.

3.2 Research design

The research undertaken is presented in three chapters, as shown on the thesis schematic in Figure 1.1. The study areas range from the global (Chapter 4) to the regional (Chapter 5) to the continental (Chapter 6) scales, as shown in Figure 3.1.

Chapter 4 i) assesses changes in global high and low flows using an ensemble of multiple GIMs fed by multiple GCMs under RCP 8.5; ii) quantifies the uncertainty in the changes owed to GIMs and GCMs. This chapter thus addresses the research gap identified in Section 2.2 on the importance of using multiple GIMs/GCMs in hydrological impact studies exploring high and low flows jointly.

Following this work, Chapter 5 evaluates the ability of GIMs to reproduce runoff characteristics, from high to low flows. The Central U.S. was chosen as study region (Figure 3.1b) as it is a data rich region where no systematic GIMs runoff evaluation has been carried out to date with the WaterMIP dataset. This chapter provides valuable information on the ability of multiple GIMs in reproducing runoff characteristics against observed data, thus building confidence in the impact model runs used in Chapter 4 and later in Chapter 6. This chapter addresses the research gap identified in Section 2.3.

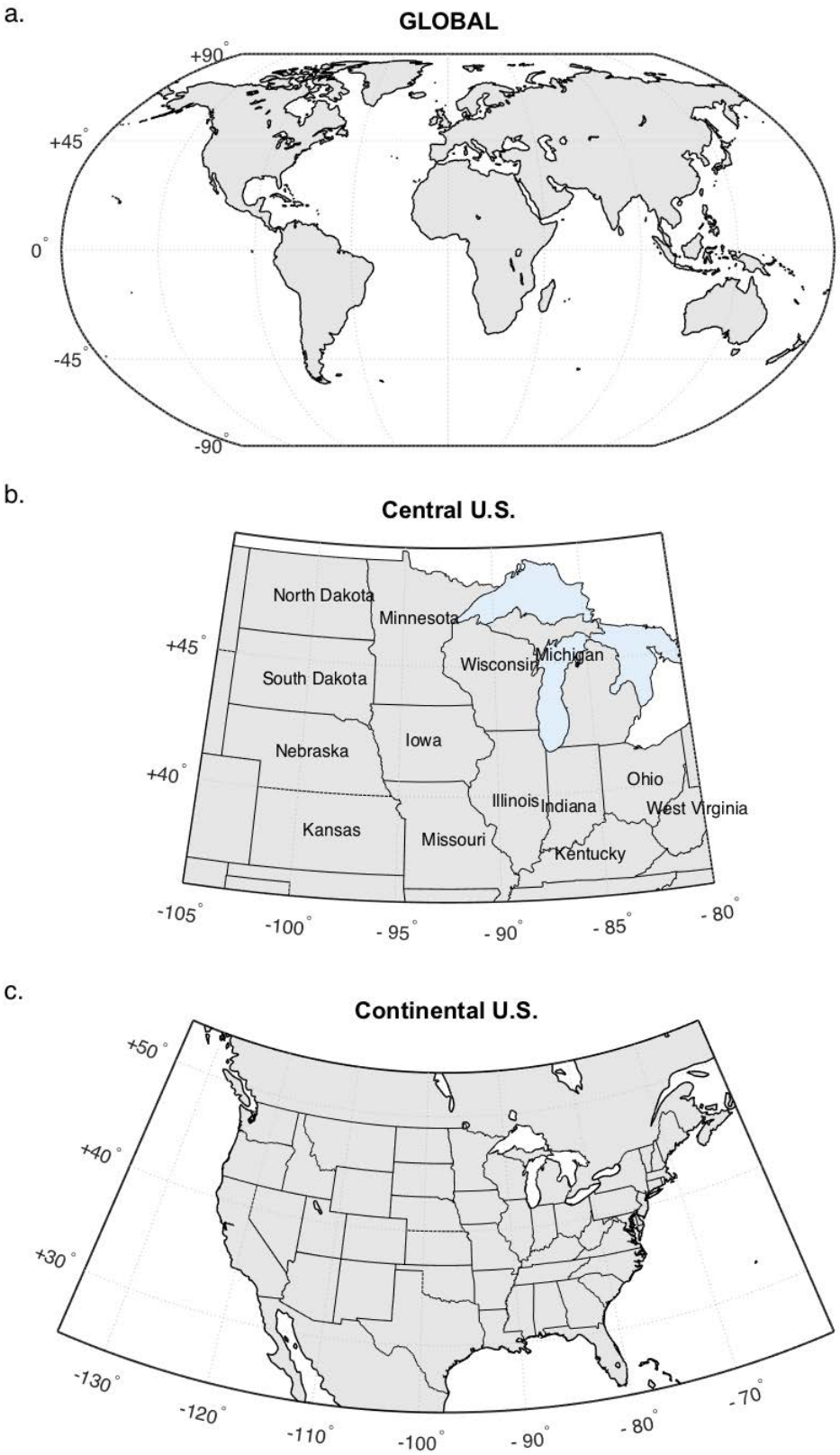


Figure 3.1 – Study areas for: a. changes in global future flows (Chapter 4); b. GIMs model evaluation (Chapter 5); c. Uncertainty partition in transient future flows (Chapter 6).

Chapter 6 brings further the research undertaken in Chapter 4 by zooming on the continental U.S. (which includes the focus study region of Chapter 5) with an assessment of the uncertainty owed to, in addition to GIMs and GCMs (quantified in Chapter 4), RCP scenarios and internal variability and going beyond the delta method with a year-by-year approach analysing transient runs throughout the 21st century. Finally, in Chapter 7 the main conclusions of the thesis are drawn and suggestions for future work are discussed.

3.3 Runoff data

The data used in this thesis consists of gridded simulated runoff from two multi-model ensembles, namely the ISI-MIP and the WaterMIP data sets (Figure 3.2). The first data set employs the most recent model setups and scenarios (RCPs) and it is used to assess future runoff firstly at the global and secondly at the continental scale, while the use of the earlier WaterMIP data set is limited, in this study, to the control period (i.e., hindcast) to evaluate GIMs' ability in reproducing runoff characteristic over a specific region using observed streamflow data as benchmark. Both ensembles do not account for anthropogenic influences (e.g. water abstraction, augmentation and

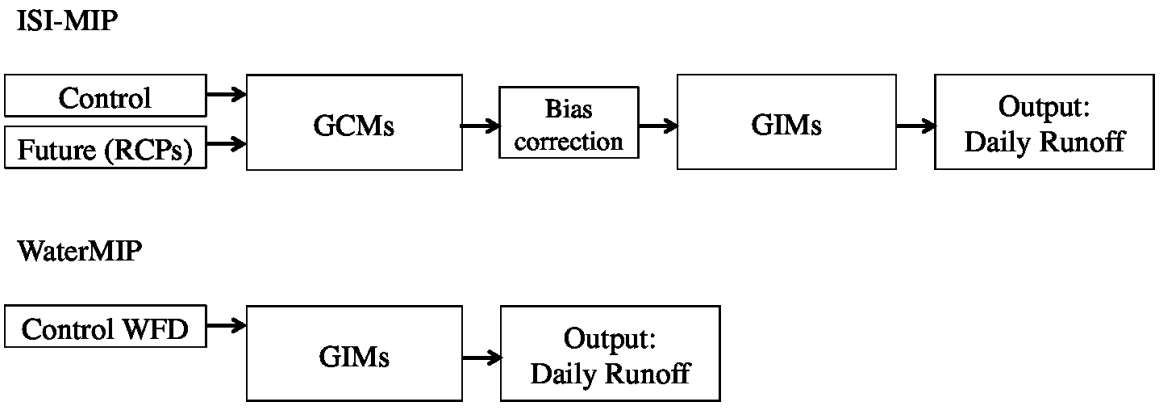


Figure 3.2 – Modelling chain corresponding to the two datasets used in the thesis.

artificial storage) or land-use changes. Descriptions of the simulated data sets and observed data are given in the following three sections.

3.3.1 The ISI-MIP data set

The daily unrouted runoff data used in Chapter 4 and Chapter 6 comes from the Inter-Sectorial Impact Model Intercomparison Project (ISI-MIP) (Warszawski et al., 2014), which comprises nine GIMs driven by five bias-corrected CMIP5 (fifth Coupled Model Inter-comparison Project; Taylor et al. (2012)) GCMs in their control (1971-2005) and future (2006-2099) period, under four RCP scenarios (RCPs 2.6, 4.5, 6.0 and 8.5).

3.3.1.1 Forcing data (GCMs)

The climate information used as input for the GIMs comes from five of the CMIP5 GCMs (HadGEM2-ES, IPSL-CM5A-LR, MIROC-ESM-CHEM, GFDL-ESM2M, NorESM1-M) covering the period from 1960 to 2099 (historical and all RCP scenarios). GCM selection was based on the availability, at the beginning of the ISI-MIP Project, of daily data for selected climate variables (see Hempel et al. (2013)) including average temperature and total precipitation. All five GCMs, whose characteristics and components are listed in Table A1.1 – APPENDIX I, consist of coupled Atmosphere-Ocean Global Circulation Model (AOGCM), which include earth system components like terrestrial and ocean carbon cycles. They can thus be called Earth System Models (ESMs). The inclusion of earth system components in an AOGCM allows for a consistent calculation of the impacts of the climate change on atmospheric composition or on ecosystems and it also allows for the incorporation of biogeochemical feedbacks: negative ones dampen the sensitivity of the climate to external forcing, while positive ones amplify the sensitivity (Collins et al., 2011). Adding Earth system components and processes increases the complexity of the model system and the model

spread in future projections. However, the larger spread should also better represent the true uncertainty of the future evolution of climate. In this thesis ESMs will henceforth be referred to as GCMs for consistency with the literature in the field of hydrological impact studies.

The GCM outputs have a spatial resolution ranging from 1.875°×1.25° for HadGEM2-ES to 2.8°×2.8° grid for MIROC-ESM-CHEM and have been bi-linearly interpolated to a common 0.5°×0.5° grid and bias-corrected towards the observation based WATCH Forcing Data (WFD) whose reference period is 1960-1999. The WFD combines the daily statistics of the ERA-40 dataset with the monthly characteristics of the Climate Research Unit (CRU TS2.1) data set and the Global Precipitation Climatology Centre (GPCC) version 4 data set, making it particularly suited as gridded observational dataset for bias-correcting global climate data (Weedon et al., 2011).

The bias-correction method applied to the GCM output first corrects the monthly mean data, and then the daily variability. More specifically, the climate variables' monthly variability and mean are corrected using a constant offset or multiplicative correction factor that corrects for long-term differences between the simulated and observed monthly mean data in the historical period. Then, the daily variability is modified about their monthly means to match the observed daily variability. This method is designed to preserve the long term absolute and relative trends in simulated data and it is presented in detail in (Hempel et al., 2013).

The bias-corrected daily runs from GCMs were used to force the nine GIMs to simulate runoff over the period 1971–2099.

3.3.1.2 Global Impact Models

The nine GIMs providing daily runoff output can be broadly grouped as hydrological models (H08, MPI-HM, PCRGlobWB, WBM, MacPDM, VIC, MATSIRO) and biome models (LPJmL and JULES; they are also hydrological models, but include effects of vegetation and CO_2

dynamics on runoff). Every GIM was run with every GCM under four RCPs. All simulations available form a sample of 156 runs in total (MacPDM, VIC, and MATSIRO lack RCPs 4.5 and 6.0 runs except when forced by the HadGEM2-ES GCM). All GIMs have a spatial resolution of 0.5°×0.5° degrees (except JULES, whose runs were regridded to 0.5° from 1.25°×1.875°) and vary in the parameterization and in the types of processes they represent (Table 3.1).

The GCM forcings employed vary across GIMs from precipitation and temperature (e.g., WBM), to additional variables like wind speed, air humidity, long (short) wave radiation flux and surface pressure (e.g., MacPDM). The method for potential evapotranspiration differs among the GIMs, as well as the schemes employed for runoff and snow representation. Moreover, H08, MATSIRO, and JULES solve also the energy balance at the land surface in addition to the water

Table 3.1– Main characteristics of the ISI-MIP GIMs with unrouted runoff daily runs (after Prudhomme et al. (2014)).

Model	Time Step	Meteorological forcing variables[a]	Energy Balance	Evapo-transpiration scheme[b]	Runoff Scheme[c]	Snow Scheme	Vege-tation dynamics	CO_2 effect
H08	Daily	R, S, T, W, Q, LW, SW, SP	Yes	Bulk formula	Saturation excess nonlinear	Energy Balance	No	No
MPI-HM	Daily	P, T, W, Q, LW, SP	No	Penman-Monteith	Saturation excess nonlinear	Degree Day	No	No
PCRGlobWB	Daily	P, T	No	Hamon	Infiltration & saturation excess, groundwater	Degree day	No	No
WBM	Daily	P, T	No	Hamon	Saturation excess	Empirical T &P formula	No	No
MacPDM	Daily	P, T, W, Q, LW_{net}, SW, SP	No	Penman-Monteith	Saturation excess nonlinear	Degree day	No	No
VIC	Daily 3h sn	P, T_{max}, T_{min}, W, Q, LW, SW,SP	No	Penman-Monteith	Saturation excess/ beta function	Energy balance	No	No
MATSIRO	1 h	R, S, T, W, Q, LW, SW, SP	Yes	Bulk formula	Infiltration & saturation excess, groundwater	Energy balance	No	constant 345ppm
LPJmL	Daily	P, T, LW_{net}, SW	No	Priestley-Taylor	Saturation excess	Degree day	Yes	Varying
JULES	1 h	R, S, T, W, Q, LW, SW, SP	Yes	Penman-Monteith	Infiltration & saturation excess, groundwater	Energy balance	Yes	Varying

[a] R = rainfall rate; S = snowfall rate; P = precipitation (rain or snow distinguished in the model); T = air temperature; W = wind speed; Q = specific humidity; LW = longwave radiation flux (downward); LWnet = longwave radiation flux (net); SW = shortwave radiation flux (downward); and SP = surface pressure. [b] Bulk formula: Bulk transfer coefficients are used when calculating the turbulent heat fluxes. [c] Beta function: Runoff is a nonlinear function of soil moisture.

balance (for this they can be called LSMs, i.e., Land Surface Models); however the most compelling structural feature in runoff generation is found in LPJmL and JULES, with the representation of the effects of CO_2 on stomatal opening and the inclusion of a dynamic vegetation model that allows vegetation to grow in response to its environment (Prudhomme et al., 2014). In particular, with increased CO_2, biome models have shown to yield increased runoff compared to the other GIMs (e.g., Wada et al. (2013); Prudhomme et al. (2014)), and this is the result of the combination of two main opposing effects of elevated CO_2 on evapotranspiration and runoff: on the one hand, CO_2 may increase plant productivity and consequently evapotranspiration from the canopy, thus decreasing runoff; on the other hand, CO_2 may inhibit evapotranspiration by reducing stomatal conductance at the leaf level, leading to increased runoff (Davie et al., 2013).

3.3.2 The WaterMIP data set

The daily unrouted runoff data used in Chapter 5 comes from the WaterMIP Project. This data set comprises nine GIMs that, within a model evaluation effort, are analysed in the control period in which they are driven by the Watch Forcing Data (WFD) (Weedon et al., 2011). GIMs runs in this dataset provide a valuable basis for evaluation as the forcing employed reproduces the frequency and the chronology of occurrence of hydrological events, while for the ISI-MIP runs only the frequency can be reproduced.

3.3.2.1 Forcing data (WFD)

The WFD consists of gridded sub-daily meteorological forcing data derived from the surface variables of the ERA-40 reanalysis product (Uppala et al., 2005) for the period 1958-2001, but from reordered ERA-40 data for the period 1901-1957 (Weedon et al., 2011). More specifically, the WFD has been produced by combining i) the Climatic Research Unit's monthly observations of

temperature, wet days and cloud cover; ii) the GPCCv4 monthly precipitation observations; and iii) the ERA-40 reanalysis products (with the addition of corrections for seasonal and decadal varying atmospheric aerosols needed to adjust the solar radiation components) (Weedon et al., 2010). The spatial resolution is 0.5° (~50 km) x 0.5° grid for a total of 67420 land cells globally. The variables included are: air temperature, surface pressure, specific humidity, wind speed, downwards long-wave (infra-red) radiation flux – provided at 6-hour intervals; downwards short-wave (solar) radiation flux, rainfall and snowfall rates – provided at 3-hour intervals.

Among the different variables constituting the WFD, the reliability of precipitation data can be crucial in the context of extreme events (e.g., floods). The ERA-40 reanalysis, on which WFD precipitation is based, has shown acceptable skills in the simulation of the variability of extreme precipitation in the cold season, although this does not imply confidence in the actual amounts of precipitation, which are considerably underestimated in the case of extreme events (Zolina et al., 2004). The ERA-40 precipitation has been corrected for the WFD in order to improve credibility for hydrological modelling at least at the sub-monthly, if not weekly or daily scale (Weedon et al., 2010). In particular, precipitation was corrected using the CRU number of wet days, thus imposing an average number of precipitation days; then the monthly precipitation was corrected using the GPCCv4 data product, integrating with the CRU monthly totals to produce alternative rainfall and snowfall products; and finally, monthly precipitation totals were also corrected for gauge "undercatch" via the gridded average catch ratios of Adam and Lettenmaier (2003). These corrections have brought about improvements in the quality of precipitation. Weedon et al. (2010) report important disparities between ERA-40 monthly precipitation totals and both CRU and GPCP totals especially in tropical latitudes. Similarly, Weedon et al. (2011) describe how the wet-day correction employed for the WFD guarantees a spatial continuity and coherence of significant frontal precipitation across grid cells, emphasizing how large scale hydrological modelling remains

meaningful at the daily scale. The same authors show, in a comparison with FLUXNET data (www.fluxnet.ornl.gov/fluxnet/) at selected sites, a close correspondence between field- measured and the adjusted reanalysis data for all variables, including precipitation. Among the WFD precipitation issues Weedon et al. (2010) report the following ones: i) presence of outliers in precipitation rates in a few isolated places where exceptionally extreme precipitation rates were created, especially near the boundaries of the Inter-tropical Convergence Zone (e.g. Northern India); ii) underestimation of orographic effects as no attempt was made to adjust precipitation rates to allow for the effects of orography.

3.3.2.2 Global Impact Models

The WaterMIP GIMs dataset comprises both land surface (LSMs) and global hydrological models (GHMs). As mentioned in the previous chapter (Section 2.3), the key difference between these two types of models is whether they solve at the land surface both the water and the energy balances (LSMs) or only the water balance (GHMs). Table 3.2 provides a brief overview of these models, which vary in structure and parameterization (for a comprehensive description of the characteristics see Haddeland et al. (2011)). Five out of nine GIMs consist of antecedent versions of the GIMs presented in the previous section (ISI-MIP dataset) and listed in Table 3.1, namely: LPJmL, MPI-HM, MacPDM, JULES, and MATSIRO.

All of the global models were run over the period 1963-2001 (except GWAVA: 1963-2000) at a spatial resolution of 0.5° degrees and forced by the same meteorological input data (WFD). The models vary substantially in the parameterizations of evaporation and runoff, and do not all use the same input variables or model time steps (in particular, all GHMs are run at a daily time step whereas LSMs are run at a sub-hourly time step). As noted in Chapter 2 (Section 2.3), in contrast to

basin-scale hydrological models, which are routinely calibrated against observed river discharge, GIMs are usually not calibrated (Müller Schmied et al., 2014).

Table 3.2 – Main characteristics of the WaterMIP GIMs with unrouted runoff daily runs (after Haddeland et al. (2011)).

Model	Time Step	Meteorological forcing variables[a]	Energy Balance	Evapotranspiration scheme[b]	Runoff Scheme[c]	Snow scheme
GHMs						
WaterGAP	Daily	P, T, LWnet, SW	No	Priestley-Taylor	Beta function	Degree-day
LPJmL	Daily	P, T, LWnet, SW	No	Priestley-Taylor	Saturation excess	Degree-day
MPI-HM	Daily	P, T	No	Thornthwaite	Saturation excess/ Beta function	Degree-day
GWAVA	Daily	P, T, W, Q, LWnet, SW, SP	No	Penman-Monteith	Saturation excess/ Beta function	Degree-day
MacPDM	Daily	P, T, W, Q, LWnet, SW	No	Penman-Monteith	Saturation excess/ Beta function	Degree-day
LSMs						
HTESSEL	1 h	R, S, T, W, Q, LW, SW, SP	Yes	Penman-Monteith	Infiltration excess/ Darcy	Energy balance
JULES	1 h	R, S, T, W, Q, LW, SW, SP	Yes	Penman-Monteith	Infiltration excess/ Darcy	Energy balance
MATSIRO	1 h	R, S, T, W, Q, LW, SW, SP	Yes	Bulk formula	Infiltration and saturation excess	Energy balance
Orchidee	15 min	R, S, T, W, Q, LW, SW, SP	Yes	Bulk formula	Saturation excess	Energy balance

[a] R = rainfall rate; S = snowfall rate; P = precipitation (rain or snow distinguished in the model); T = air temperature; W = wind speed; Q = specific humidity; LW = longwave radiation flux (downward); LWnet = longwave radiation flux (net); SW = shortwave radiation flux (downward); and SP = surface pressure. [b] Bulk formula: Bulk transfer coefficients are used when calculating the turbulent heat fluxes. [c] Beta function: Runoff is a nonlinear function of soil moisture.

With the exception of WaterGAP, none of the models used in this dataset were calibrated specifically for the WaterMIP experiment, although they may have been calibrated for previous studies (Haddeland et al., 2011). The GIMs use their default soil and vegetation information derived from mapped land properties (e.g. soil texture and vegetation density) (Gudmundsson et al., 2012b), and no attempt was made to standardize these parameters (Haddeland et al., 2011). WaterGAP

underwent a limited calibration procedure using local measured streamflow data (for details see Hunger and Döll (2008)).

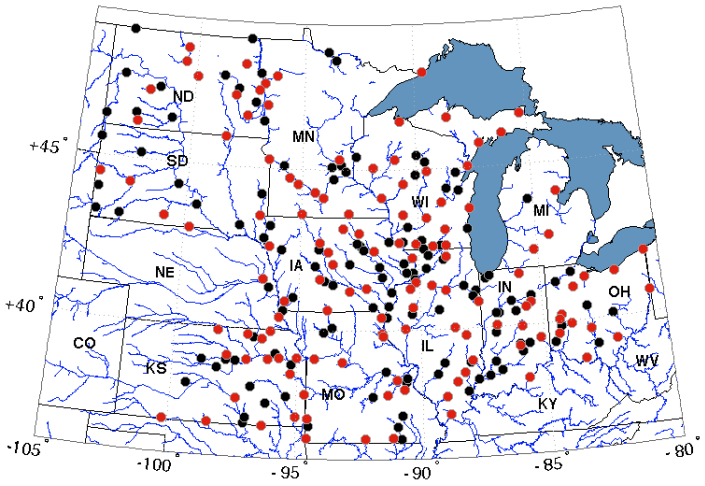
3.3.3 Observed runoff data

The reference dataset for evaluating simulated runoff consists of observed discharge data covering the 1963-2001 period over the selected study area (Figure 3.1b). A total of 252 gauging stations (Figure 3.3 and Table 3.3) were selected after intersecting ~700 gauges available over the study area with the Hydro-Climatic Data Network (HCDN) because no land use changes or water management interventions are accounted for in the modelled data. The HCDN dataset was introduced in 1992 and updated in 2011 (Whitfield et al., 2012) as a subset of U.S. Geological Survey (USGS) streamflow gauging stations with historical streamflow data responsive to climatic variations, so relatively free of anthropogenic influences such as dam impoundment, regulation and wide-scale urbanization (although minor impacts may still be present, e.g. land use change).

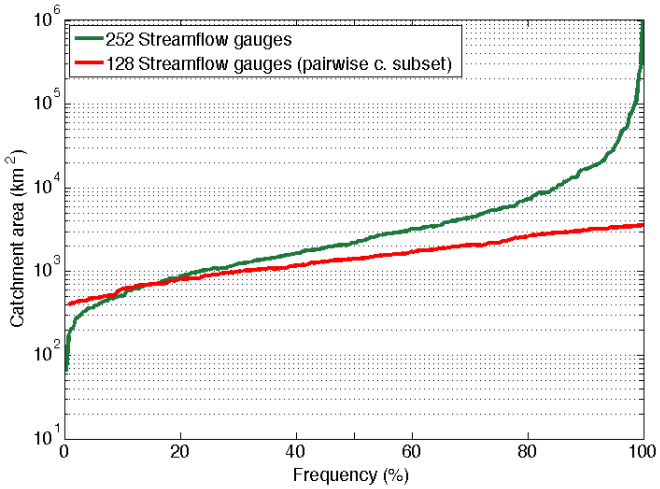
The size of the catchments of the 252 gauges varies, with drainage areas ranging from 64 to 1,350,000 km^2, with a majority (80%) with area up to 7000 km^2 (see Figure 3.3b, while the catchment boundaries are shown in Figure 3.3c).

For an in-depth comparison gauge-grid-cell, a subset of 128 gauges (shown in red in Figure 3.3a-c) have been selected within the 400 to 3500 km^2 catchment area range in order to match more closely the size of the model grid cells whose area ranges from approximately 2500 km^2 at 36°N to 2000 km^2 at 49.5°N depending on the latitude.

a.



b.



c.

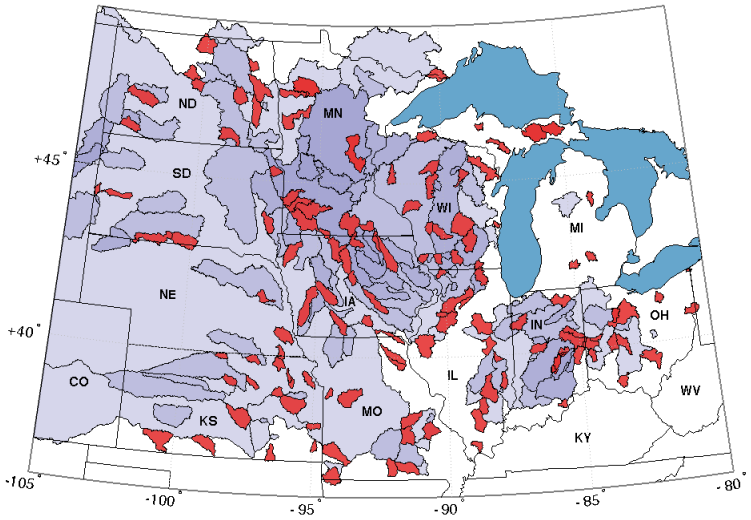


Figure 3.3 – The 252 streamflow gauges used in this study: a) their location (in red the 128 subset used for pairwise comparison, in black the remainder of the gauges, in blue the river network); b) the distribution of their catchment areas (green the whole 252 set, in red the 128 subset); c) Catchment boundaries relative to the streamflow gauges (in red the 128 subset, in purple the remainder of the gauges).

Table 3.3 – List of the 252 stream-gauges used for the GIMs' evaluation (*128 pairwise comparison subset).

	USGS str.flow gauge	Catch. area [Km^2]		USGS str.flow gauge	Catch. area [Km^2]		USGS str.flow gauge	Catch. area [Km^2]		USGS str.flow gauge	Catch. area [Km^2]
*1	3109500	1285	*64	4198000	3240	127	5444000	378	*190	6471200	1971
2	3144000	363	*65	4201500	692	128	5446500	24732	191	6478500	53501
*3	3159500	2442	*66	4213000	453	*129	5447500	2598	192	6480000	8645
4	3219500	1469	67	5053000	5387	130	5451500	3968	*193	6481500	1606
*5	3230500	1383	*68	5054500	1098	131	5454500	8472	194	6483500	4123
6	3234500	13289	69	5056000	5361	*132	5457000	1033	195	6485500	20407
*7	3237500	1002	70	5057000	16757	*133	5458000	793	*196	6600500	2295
*8	3262000	666	*71	5057200	1790	134	5458500	4302	197	6606600	6475
*9	3265000	1303	*72	5059700	2183	*135	5459500	1362	*198	6800000	953
10	3266000	1683	73	5060500	300	136	5462000	4522	199	6800500	17871
*11	3269500	1269	*74	5062000	2525	137	5464000	13328	*200	6808500	3434
*12	3272000	712	*75	5066500	3116	138	5464500	16861	201	6809500	2315
13	3274000	9402	*76	5069000	1088	139	5465500	32375	202	6810000	7268
*14	3275000	1352	*77	5078500	3574	*140	5466000	401	*203	6811500	2051
*15	3294000	490	78	5082500	77959	*141	5470000	816	*204	6814000	715
16	3324000	681	*79	5084000	1181	*142	5472500	1891	*205	6853800	588
*17	3324300	1101	80	5100000	8832	143	5474000	11168	206	6864500	19632
*18	3326500	1766	81	5113600	4636	144	5474500	308209	207	6867000	3890
19	3328500	2044	*82	5120500	1826	*145	5476000	3237	208	6869500	7304
20	3333600	64	*83	5123400	3004	*146	5479000	3388	*209	6876700	995
21	3334500	629	84	5131500	4351	147	5480500	10852	210	6876900	17534
22	3335500	18821	85	5133500	50246	*148	5481000	2186	*211	6878000	777
*23	3339500	1318	86	5280000	6838	149	5482500	4193	*212	6884200	891
24	3340800	360	*87	5286000	3522	*150	5484000	2574	213	6884400	8679
*25	3343400	482	88	5288500	49469	151	5484500	8912	*214	6885500	1062
26	3345500	3926	*89	5291000	1031	152	5486490	1268	*215	6888500	824
*27	3346000	824	90	5304500	4869	*153	5489000	969	216	6889200	386
*28	3351500	438	*91	5313500	1725	154	5490500	36358	*217	6889500	751
29	3360500	12142	*92	5316500	1629	*155	5495000	1036	218	6891500	1101
30	3362000	277	*93	5317000	3367	*156	5500000	1606	*219	6892000	1052
*31	3362500	1228	*94	5320500	2875	*157	5501000	917	*220	6894000	477
32	3363500	785	95	5330000	41958	158	5520500	5941	221	6897500	5827
33	3364000	4421	96	5331000	95312	*159	5525000	1777	*222	6898000	1816
34	3373500	12761	97	5340500	16162	160	5526000	5416	223	6899500	4455
35	3374000	28814	*98	5362000	1492	161	5527500	13338	*224	6908000	2901
36	3377500	74164	*99	5368000	1083	*162	5555300	3240	*225	6913500	3237
*37	3379500	2929	*100	5379500	1665	*163	5556500	508	*226	6917000	813
*38	3380500	1202	*101	5381000	1940	164	5567500	1987	227	6933500	7356
39	3381500	8034	102	5393500	211	*165	5569500	2776	228	6934500	1353269
*40	3612000	632	103	5394500	477	166	5570000	4237	*229	7013000	2023
*41	4010500	1577	*104	5397500	971	*167	5572000	1424	*230	7016500	2093
*42	4027000	1546	105	5399500	580	*168	5585000	3349	*231	7018100	1904
*43	4040500	443	*106	5405000	1577	169	5592500	5025	232	7018500	2375
*44	4045500	2046	107	5407000	26936	170	5593000	7042	233	7019000	9811
*45	4056500	2849	*108	5408000	689	*171	5597000	2051	*234	7057500	1453
*46	4059500	1165	109	5412500	4002	172	6334500	5113	235	7061500	1254
*47	4069500	2797	*110	5413500	697	173	6335500	12018	236	7067000	4318
48	4071000	1826	111	5414000	368	174	6337000	21523	237	7068000	5278
*49	4073500	3471	112	5418500	4022	*175	6339500	3186	*238	7071500	2054
50	4079000	5853	*113	5419000	640	176	6340500	5802	239	7141200	5563
51	4084500	15566	114	5420500	221703	177	6350000	1502	*240	7144200	3437
*52	4086000	1083	*115	5421000	2714	*178	6352000	1432	241	7146500	113216
53	4087000	1803	116	5422000	6050	179	6354000	10619	242	7147800	4869
54	4093000	321	*117	5426000	1974	180	6359500	6724	*243	7149000	2339
55	4094000	171	118	5430500	8651	181	6395000	18658	*244	7157500	2997
*56	4100500	1538	*119	5431486	515	182	6409000	205	245	7167500	334
*57	4105000	624	*120	5432500	707	*183	6425500	1422	*246	7172000	1153
*58	4112500	919	121	5433000	572	*184	6431500	427	247	7180500	285
59	4121500	3711	122	5434500	2678	185	6441500	8151	248	7183000	9643
*60	4142000	829	123	5435500	3434	186	6446000	5594	*249	7184000	510
*61	4189000	896	*124	5436500	1355	*187	6449500	2675	*250	7186000	3015
62	4191500	6004	*125	5438500	1393	188	6452000	25693	251	7187000	1106
63	4193500	16395	126	5440000	2846	*189	6464500	2924	*252	7189000	2204

3.4 Overview of the statistical methods

This section introduces the statistical methods used in the thesis. Details of the methods used in this research are given in the respective chapters.

3.4.1 Hydrological indices

3.4.1.1 Changes in the frequency of high and low flows.

In Chapter 4, delta indices are used to express changes in the frequency of high and low flows. The indices extraction consists of two steps: i) time series of days classified as high and low flows are extracted from daily total runoff record; ii) high and low flow indices (i.e. change in frequency of high/flow flows) are calculated (future minus historical period).

In the first step, in order to quantify high and low flow inter-annual variability, daily binary series (zero or one) are extracted for every land grid cell: high flow days, HFD; and low flows days, LFD. The series extraction uses daily varying threshold curves obtained from the daily runoff series for the historical period (1972–2005), which are then applied to the historical period and future projections to identify days above (for HFD) or below (for LFD) threshold, as in e.g. for low flows Prudhomme et al. (2014). High flows are characterized by the 95^{th} percentile (Q_{95} – runoff equalled or exceeded 5 % of the time) and low flows by the 10^{th} percentile (Q_{10} – runoff equalled or exceeded 90 % of the time). For HFD, a value of 1 (high flow) is assigned to each cell if the cell's runoff exceeds the Q_{95} value, otherwise a value of 0 (no high flow) is assigned. For LFD, a value of 1 (low flows) is assigned to each cell if the cell's runoff is below the Q_{10} value, otherwise a value of 0 (no low flow) is assigned. Explanatory graphics for both threshold and binary series extraction are shown in Figure 3.4.

In the second step, the indices expressing the change in the frequency (in %) of future high (HFI) and low (LFI) flows are then calculated as follows: for each ensemble member HFI (LFI) is equal to the difference between the frequency (in %) of high (low) flows days (100× mean of HFD (LFD)) from the future (e.g., 2066–2099) and historical period (e.g., 1972–2005), for the whole year or per season.

It is worth describing in more detail the first step, i.e. the development of the threshold and binary series extraction of HFD and LFD, undertaken after participating to the study on future droughts of Prudhomme et al. (2014) to overcome a glitch in their drought index extraction.

The threshold curves are obtained by linearly interpolating percentiles calculated over fixed 5-day windows (e.g. 1–5 December, 6–10 December, and so forth, i.e. 73 for the whole year) of the historical period runoff (i.e. December 1971 to December 2005), having considered the hydrological year from December to November. In general, the identification of high and low flows at the global scale imposes the selection of a universal threshold level serving many hydrological regimes and climate regions at once (thereby pooling events that may not always be extreme) and it is based on physical processes: low flows are generally characterized by a slower onset, and a longer duration, and high flows by a sudden onset, and a shorter duration. Accordingly, high and low flows are not necessarily symmetric with respect to the median flow (Q_{50}). For low flows in particular, the choice of Q_{10} comes from seeking a sufficiently low quantile without compromising the analysis, as quantiles lower than 10% become intractable for the large presence of zero pools in some time series. This is in agreement with e.g. Gudmundsson et al. (2012a) who showed how the performance of a similar set of WaterMIP global models decreased systematically from high Q_{95} to low Q_5 runoff percentile over Europe. The choice of a fixed 5-day time window with interpolation was preferred over the 30-day moving average used in e.g. Prudhomme et al. (2014) because the

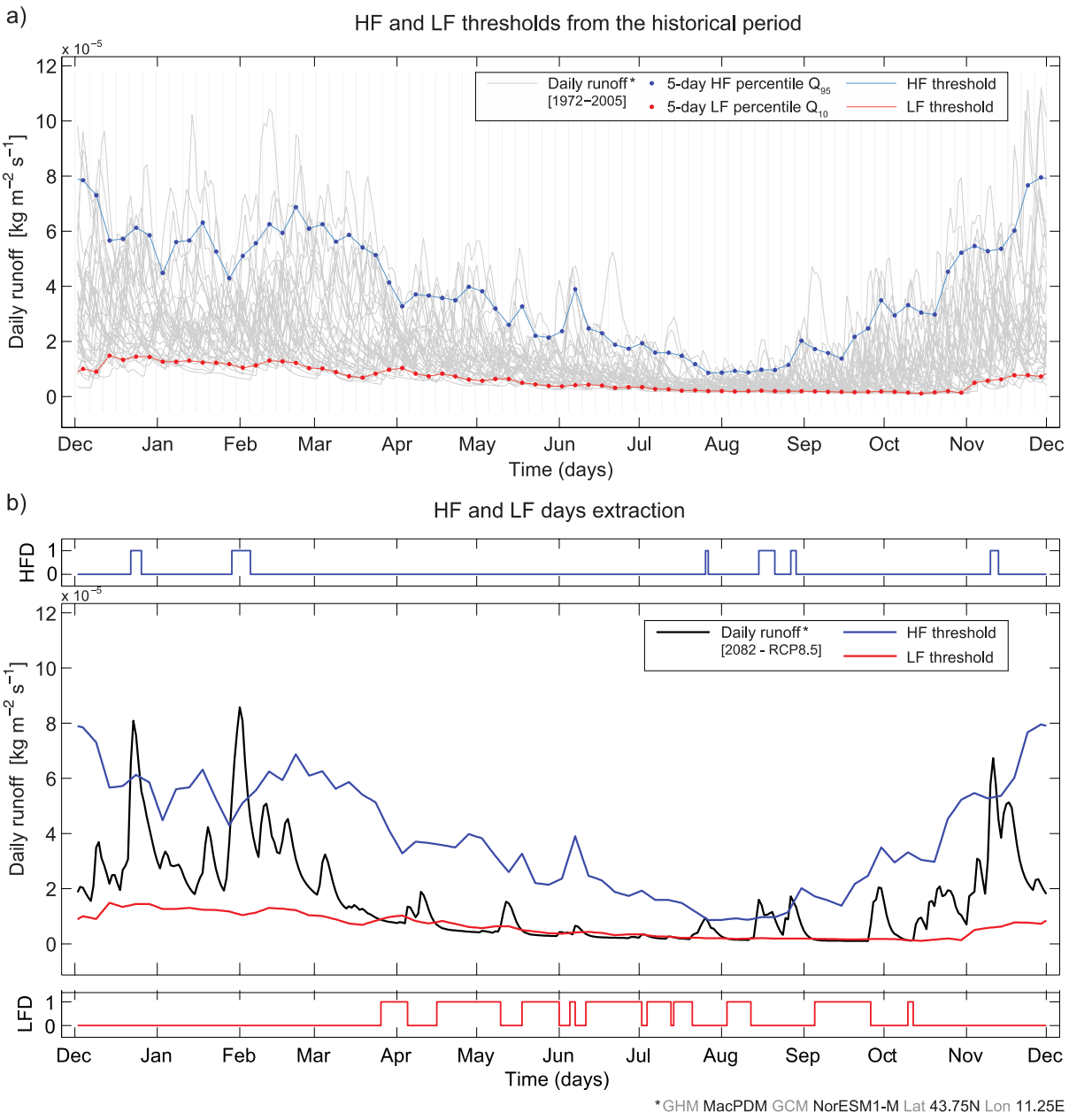


Figure 3.4 – Schematic of HFD and LFD extraction (days under high and low flows): a) daily varying threshold curves for HF and LF from 5-day percentiles calculated over the historical period; b) High and low flow days extraction for a given year.

latter had shown some limitations with regards to the low flow quantile extraction. The effect of levelling out over 30 days could lead to lower values than expected in the control period (10% by design but). The choice of 5-day was made after testing 1, 5, 11, 31-day windows at nine locations

world-wide (considering, for each location, a cluster of 5 contiguous grid cells) as 5-day proved to be closest to the 10% (Figure 3.5). In addition, favouring the use of the same framework for high and low flows, the 5-day window was considered to be most appropriate for the identification of both types of events. The choice of a linear interpolation was preferred over the moving window approach to minimize dependence (i.e. inertia) within quantile estimates with the following rationale:

i) moving average aims to smooth out wiggles for a less spiky identification of hydrological events like droughts that could result in erratic threshold crossings, thereby pooling several times over the same event; however, its quantile estimates use the same information from neighbouring days (as many as the time window), resulting in a quantile series holding a correlation that is higher the longer the time window, potentially leading to inadvertent effects of large inertia during the extraction of the hydrological index.

ii) In our case, as we count high (low) flow days (as opposed to single events), smoothing the threshold is unnecessary.

iii) A 1-day window would ensure a series of independent quantile estimates, but the computation over 34 points (i.e. 34 years of the control period) was considered insufficient for quantile estimation.

iv) Seeking a representative number of points for quantile extraction (170, i.e. 5 days × 34 years), a choice was made to compute the quantile by extracting a point every 5 days and extrapolating values for intermediate days to the next 5-day point; as a result threshold values were obtained with a nonrecursive use of data, thereby minimizing dependence.

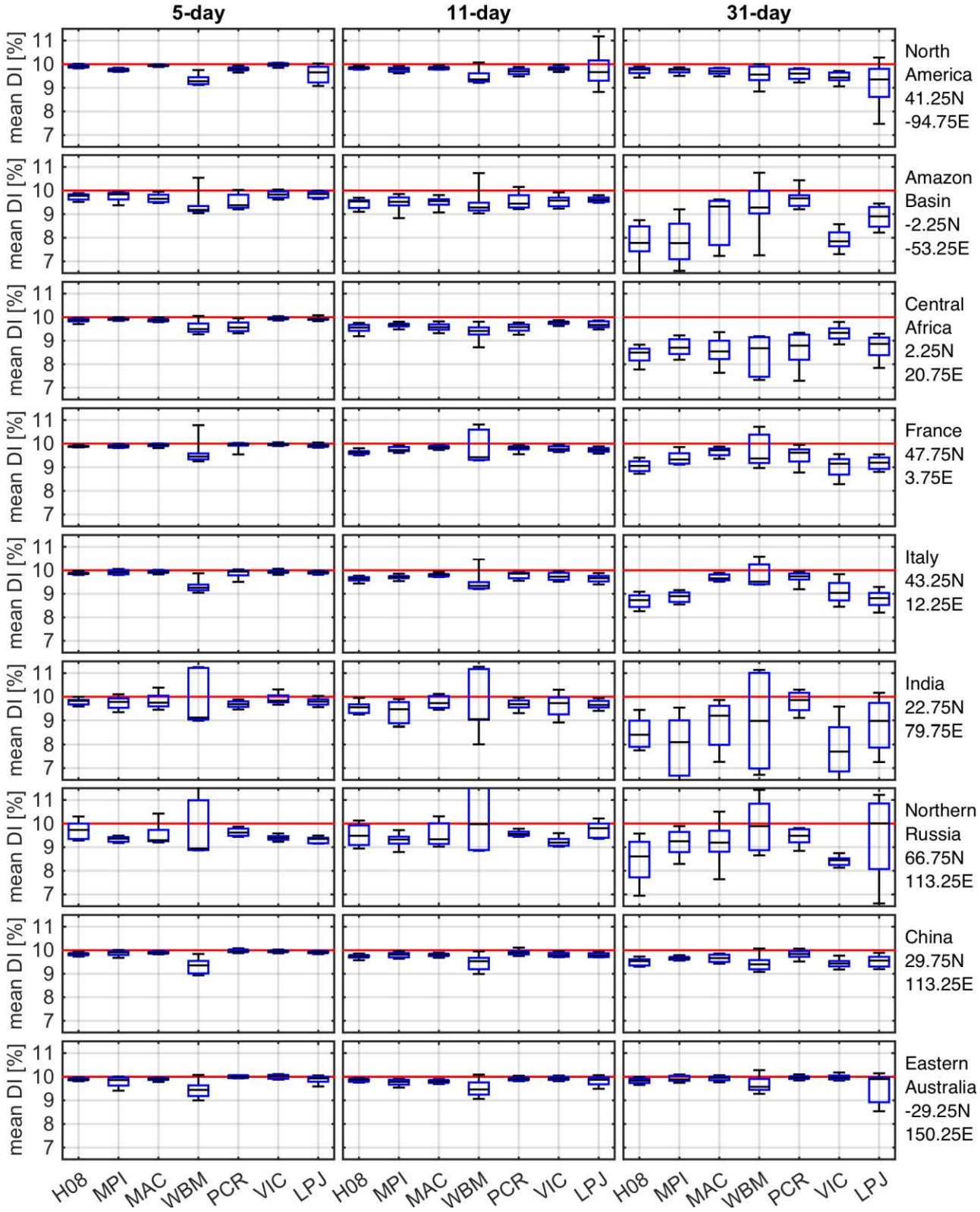


Figure 3.5 – Sensitivity analysis for low flow threshold estimation using 5-, 11-, 31-day windows for 5 contiguous grid cells from 9 selected locations worldwide over the reference period 1972-2005. Boxplots are per GIMs and comprise four GCMs for 5 grid cells (20 points). On the y-axis: the closest to 10% the better.

The index extraction described above is not applicable when the runoff is very low, i.e. when long periods of the year have the same value. Therefore, with reference to the control period (1972–2005), grid cells showing little or no seasonal change in daily runoff were screened out (details on the masking procedure are provided in APPENDIX II).

3.4.1.2 Annual block indices

Hydrological indices extracted annually from daily runoff series are used in Chapter 5 for the model evaluation (over the Central U.S.A.) and in Chapter 6 for the transient uncertainty partition (over the continental U.S.A.). In particular, in order to analyse changes in runoff over different parts of the flow regime (including high, medium, and low flows), the following three magnitude indices are considered: 1) annual maximum flow (AMax: a record of the largest daily discharge value for every year), 2) annual median flow (AMed: a record of the median daily discharge value for every year), and 3) annual minimum flow (AMin: a record of the smallest daily discharge value for every year). For the model evaluation analysis, in addition to these three indices, three timing indices were used to gain a basic understanding of whether the models are able to capture the timing of flooding, medium flow, and drought: 1) annual maximum date (AMaxDate: the day of the year in which the largest daily discharge value occurs for every year); 2) medium flow date (V50Date: the day of the year by which half of the annual total discharge volume has occurred); V50Date follows the concept of "center of mass" timing proposed by Stewart et al. (2005), also used, for instance, in Moore et al. (2007); and 3) drought start date (VDef10Date: the day of the year by which 10% of the annual volume deficit has occurred). The threshold used to define the VDef10Date corresponds to the 20^{th} quantile of the time series; following the center of mass concept over the volume deficit (as, for instance, in Giuntoli et al. (2013), which provide a

schematic of the index), the drought starts on the day the 10% of the annual volume deficit has occurred.

3.4.2 Multi-model ensemble metrics

The different model runs make up a multi-model ensemble spread, whose strength is to express the range of plausible future outcomes. In order to use this information, both the mean (arithmetic average) or the median (the value separating the data sample in halves) can be used, although the latter – used in this study – is usually favoured as it is more robust and less influenced by potential outliers (Roudier et al., 2015). In order to express the degree of agreement of multiple model outputs the Signal-to-Noise ratio, S2N, is used and it is obtained by dividing the median of the ensemble indices by the inter-quartile range (75^{th} percentile minus 25^{th} percentile). The higher the S2N, the higher the agreement of the model members in the signal, as in Prudhomme et al. (2014).

3.4.3 Uncertainty partition

The uncertainty assessed in the thesis is reflected by the spread of the flow indices. In Chapter 4, while assessing changes in high and low flows globally, the uncertainty is quantified with regards to the choice of GCM or GHM. Whereas in Chapter 6, while analysing decadal high, medium, low flows in the continental U.S.A., the uncertainty is quantified with regards to the choice of GCM, GHM, RCP scenario and the contribution of the internal variability.

Analysis of variance frameworks have become common in hydroclimatology for quantifying sources of uncertainty in climate projections. In the Chapter 4, a 2-way ANOVA is used to quantify GCM and GHM uncertainty under one RCP scenario (8.5), while in Chapter 6, the methodology proposed by Hawkins and Sutton (2009) is used to quantify relative contributions of also RCP and

internal variability, provided that the latter (theoretically not samplable in our dataset owing to the unavailability of run replicates) is sampled using the distance from the mean of each source and is considered and assumed constant throughout the 21st century.

3.5 Chapter summary

This chapter has presented the research design of the thesis. After illustrating the study areas corresponding to the three main analyses, the data employed and the generic statistical methods used are presented. The methods are detailed in each of the chapters. The next chapter assesses future changes in high and low runoff globally and assesses the contribution to uncertainty form GCMs and GIMs.

4 Changes in future hydrological extremes and the uncertainty from global impact and global climate models

4.1 Abstract

Projections of changes in the hydrological cycle from global impact models (GIMs) driven by global climate models (GCMs) are critical for understanding future occurrence of hydrological extremes. However, uncertainties remain large and need to be better assessed. In particular, recent studies have pointed to a considerable contribution of GIMs that can equal or outweigh the contribution of GCMs to uncertainty in hydrological projections. Using six GIMs and five GCMs from the ISI-MIP multi-model ensemble, this chapter aims: (i) to assess future changes in the frequency of both high and low flows at the global scale using control and future (RCP8.5) simulations by the 2080s, and (ii) to quantify, for both ends of the runoff spectrum, GCMs and GIMs contributions to uncertainty using a two-way ANOVA. Increases are found in high flows for northern latitudes and in low flows for several hotspots. Globally, the largest source of uncertainty is associated with GCMs, but GIMs are the greatest source in snow-dominated regions. More specifically, results vary depending on the runoff metric, the temporal (annual and seasonal) and regional scale of analysis. For instance, uncertainty contribution from GIMs is higher for low flows than it is for high flows, partly owing to the different processes driving the onset of the two phenomena (e.g. the more direct effect of the GCMs' precipitation variability on high flows). This study provides a comprehensive synthesis of where future hydrological extremes are projected to increase and where the ensemble spread is owed to either GCMs or GIMs. Finally, our results underline the need for improvements in modelling snowmelt and runoff processes to project future

hydrological extremes and the importance of using multiple GCMs and GIMs to encompass the uncertainty range provided by these two sources.

4.2 Introduction

The use of global multi-model ensembles in climate impact studies has been discussed in Chapter 2 (Section 2.2). The paucity of studies considering a representative group of GIMs has been highlighted together with the interest to assess two extremes at once and to quantify uncertainty coming from GIMs and GCMs using a formal statistical framework.

The present study builds on the work on low flows of Prudhomme et al. (2014), but introduces several new aspects. Firstly, low flows (Q_{10}) are now analysed using an improved index extraction. The variable threshold method used in Prudhomme et al. (2014) has been revisited to overcome a limitation of the 30-day moving window for which grid cells were assigned lower threshold values than the theoretical threshold assigned (Q_{10}) (i.e. a tendency to capture fewer occurrences, an effect perhaps attributable to GIMs' slow emptying of reservoirs during the recession phase). A shorter 5-day fixed time window eliminates this effect. Note that, in order to gather further data for the estimate of the quantile flow, the period of analysis was increased from 30 to 34 years, starting 4 years earlier (1972 for control and 2066 for future). Secondly, we now analyse high flows (Q_{95}), with the same method used for low flows (5-day fixed-window variable-threshold method). Dankers et al. (2013), who also analysed high flows, have focused on a different metric (annual extreme monthly flood peak with 30-year return level), as their aim was to describe changes in flood hazard, while our focus is on change in frequency of high flow days. In our study high and low flows are hence identified jointly with the same ensemble of five GCMs and six GIMs. While comprising the same number of GCMs, the ensemble used by Prudhomme et al. (2014) uses one additional GIM (JULES) and Dankers et al. (2013) uses three additional GIMs

(JULES, LPJmL, MATSIRO). We did not use these additional GIMs as they showed large areas with long pools of zero values hindering the index extraction, making them unsuitable for our analysis, especially for the low flows; additionally, JULES was run at a coarser resolution (1.25–1.875 vs. 0.5–0.5) that would potentially influence the uncertainty analysis. Thirdly, we assess systematically the relative contribution of GIMs and GCMs to uncertainty using an analysis of variance (ANOVA) framework as in e.g., Yip et al. (2011) and Sansom et al. (2013). This uncertainty assessment moves beyond the signal-to-noise ratio by Prudhomme et al. (2014), as the quantification of each source (GCM/GIM) to total uncertainty allows us to describe the spatial variability of the contributions grid cell per grid cell. While Dankers et al. (2013) and Schewe et al. (2014) partition GCM/GIM uncertainty using ratios between the variances, our ANOVA approach adds the contribution of the error (or residual) to the partition of the variance along with *post hoc* testing on the residuals for model adequacy. We thus describe how high and low flows and inherent uncertainty vary at the seasonal and spatial scale, identifying areas where we have more confidence in the climate or in the hydrology (i.e. uncertainty is owed to GCMs or GIMs). Finally, to understand how the variance of the changes differs regionally, we carry out analysis at the regional scale expressing the ANOVA sum-of-squares of each source using homogeneous geo-climate regions (Köppen–Geiger). This allows for an improved understanding of how the climate and hydrological processes drive uncertainty for both runoff ends.

By comparing an ensemble of GCMs (5) and GIMs (6) for future projections (2066–2099) against the historical period (1972–2005), this study aims (i) to assess future high and low flows changes at global and annual and seasonal scales, and (ii) to quantify the uncertainty attributable to GIMs and GCMs using ANOVA. In the next section, the data set and the different steps of the methodology are detailed. The results of projected hydrological extremes and respective uncertainty

are presented in Section 4.4 before discussing the important and wider implications of this research in the fifth and final section.

4.3 Data and Methods

The data set used herein – introduced in Section 3.3.1 – comes from the Inter-Sectorial Impact Model Intercomparison Project (ISI-MIP) (Warszawski et al., 2014) and consists of daily total unrouted runoff at a spatial resolution of 0.5 degrees from an ensemble of six GIMs forced with five CMIP5 GCMs' bias-corrected climate (Hempel et al., 2013) for the historical (1972–2005) and future (2066–2099) periods under the RCP8.5 scenario. The six GIMs are: H08, MPI-HM, MacPDM, VIC, WBM, PCRGlobWB (see Table 3.1 in Section 3.3.1.2 for a summary of the main characteristics), and the five GCMs are: HadGEM2-ES, IPSL-CM5A-LR, MIROC-ESM-CHEM, GFDL-ESM2M, NorESM1-M (see Section 3.3.1.1 – refer to Warszawski et al. (2014), for further details on the models and to www.isi-mip.org to access the simulation protocol). It should be noted that the selection of GIMs was dictated by temporal (daily runoff) resolution and time series tractability: models with lengthy pools of runoff equal to zero over large portions of the globe imposing constraints to the index extraction were not included (this aspect has been described further at the end of Section 3.4.1.1). The selected model combinations form an ensemble of 30 experiments, each consisting of a historical and future period; none of the GIMs include varying CO_2.

The analytical framework employed was composed of three steps: (i) high and low flow indices (i.e. change in frequency of high/flow flows) were calculated (future minus historical period) and mapped; (ii) ANOVA was carried out on the high and low flow indices considering GCMs and GIMs as factors; and (iii) the dominant uncertainty factors were explored for high and low flows across different climate regions based on the Köppen–Geiger classification.

As detailed in Section 3.4.1.1, the extraction of high and low flow indices consists of extracting daily binary series (zero or one) for every land grid cells using daily varying threshold curves obtained from the historical period (1972–2005), which are then applied to the historical period and future projections to identify days above (for HFD) or below thresholds (for LFD). High flows are characterized by the 95th percentile (Q_{95} – runoff equalled or exceeded 5% of the time) and low flows by the 10th percentile (Q_{10} – runoff equalled or exceeded 90% of the time).

Grid cells showing little or no seasonal change in the daily runoff of the control period (1972–2005) were screened-out and represented in grey on the maps. These screened-out grid cells are often located in arid or frozen regions where there is little or no runoff during long periods of the year and so the index extraction becomes intractable due to the presence of repeated zero values in the series. The grid cell masking procedure is detailed in APPENDIX II, while Table 4.1 shows

Table 4.1 – Percentage of available land grid cells after masking per GIM-GCM model combination.

			GCM				
			HadGEM	IPSL	MIROC	GFDL	NorESM
GIM	H08	Q_{10}	99.97	99.82	99.96	99.96	99.95
		Q_{95}	99.98	99.98	99.98	99.99	99.98
	MPIHM	Q_{10}	89.85	89.14	89.69	89.68	89.68
		Q_{95}	92.75	92.24	92.52	93.17	93.08
	MacPDM	Q_{10}	100	100	100	100	100
		Q_{95}	100	100	100	100	100
	VIC	Q_{10}	96.25	96.25	96.47	96.59	96.39
		Q_{95}	99.48	97.72	99.41	99.20	99.36
	WBM	Q_{10}	96.19	96.29	95.72	96.02	96.27
		Q_{95}	97.38	97.97	96.81	97.75	97.58
	PCRGlobWB	Q_{10}	90.91	91.17	90.39	91.26	90.71
		Q_{95}	92.92	92.84	92.16	93.16	92.79
	JULES*	Q_{10}	64.07	64.05	65.45	66.06	66.59
		Q_{95}	84.71	89.16	91.39	89.57	91.06
	LPJmL*	Q_{10}	26.97	25.07	25.95	26.12	26.89
		Q_{95}	70.22	67.27	69.76	68.50	69.72
	MATSIRO*	Q_{10}	25.73	23.27	29.60	25.39	27.70
		Q_{95}	64.56	61.26	67.15	69.10	67.42

*Models not included in the ensemble.

percentages of available land grid cells after screening for the different GCM–GIM combinations and runoff percentile, which led to the choice of the first six models (notice the low percentage values for LPJmL and MATSIRO especially).

Once binary series are obtained, the indices expressing the change in the frequency (in %) of future high (HFI) and low (LFI) flows can be calculated as follows: for each ensemble member HFI (LFI) is equal to the difference between the frequency (in %) of high (low) flows days (100× mean of HFD (LFD)) from the future (2066–2099) and historical period (1972–2005), for the whole year and per season (DJF and JJA). Both HFI and LFI are composed of 30 series (i.e. six GIMs fed by five GCMs each). The agreement in the change across ensemble members is expressed by the signal-to-noise ratio, S2N, calculated by dividing the median of the ensemble flow indices (HFI and LFI) by the inter-quartile range (75th percentile minus 25th percentile). The higher the S2N, the higher the members' agreement in the signal; assuming signal greater than noise if S2N > 1.

In this study, the uncertainty is reflected by the spread of the flow indices due to the choice of GCM or GIM. To quantify the individual contribution of GCMs and GIMs to total uncertainty, a 2-factor ANOVA was carried out on the flow indices HFI and LFI for each grid cell. For this data set, model runs had no replicates, therefore the ANOVA model considers one case per treatment (Neter et al. 1999, Chap. 21), so no interactions $\alpha\beta_{ij} = 0$) and fixed factors levels (n=1):

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \text{ , (1)}$$

where Y_{ij} is the mean change for GCM_i and GIM_j , μ is a constant (the overall mean of all ensemble members), α_i is the main effect for GCM at the ith level, β_j is the main effect for GIM at the jth level, ε_{ij} is the residual $\approx N(0, \sigma^2)$iid. Thus, the variance is partitioned into two factors, GCMs and GIMs, plus the residuals. The results, expressed in terms of sum of squares, are used to quantify the factors' contributions to the total variance, here considered as uncertainty as in e.g. Sansom et al.

(2013). ANOVA models are reasonably robust against certain types of departures from the model (e.g. error terms not being exactly normally distributed). Nonetheless, the suitability of the ANOVA model with the data at hand should be checked for serious departures from the conditions assumed by the model by looking at the residuals (Neter et al. 1999, Chap. 18) and testing their normality (e.g. Lilliefors test) and constancy of variance (e.g. Hartley test). Unsatisfactory results would require remedial measures like data transformation or a modification of the model. To understand how variance differs between climate regions, the ANOVA sum of squares for all model combinations are shown per Köppen–Geiger class. We used the Köppen–Geiger data classification[1] based on the present day as proposed by Kottek et al. (2006). A total of 15 (out of 31) regions are considered leaving out under-represented regions with too few grid cells (<1000). The number of grid cells required to retain a region was set to 1000 so that regions excluded would not account for more than 10% of global unmasked grid cells (precisely 9.1% for both indices).

4.4 Results

Annual mean changes and associated S2N across all GIMs and GCMs are shown for HFI and LFI in Figure 4.1a and Figure 4.1b. For high and low flow indices, the mean changes vary spatially and in magnitude (Figure 4.2) but they are positive generally. This means increases in number of days with (i) high flows, mostly over high northern latitudes; and (ii) low flows, spread over all latitudes with hotspots in southern Europe, south-western and mid Latin America, south-eastern USA and south-eastern Canada, lower parts of Central Africa, north/north-eastern China, and south-western Australia. Regions screened-out represent 14 and 18 % of land for HFI and LFI, respectively. The S2N shows model agreement generally over the same regions for both indices (e.g. southern Europe, south-western and mid Latin America, south-eastern US). However,

[1] Map link: http://koeppen-geiger.vu-wien.ac.at/pdf/kottek_et_al_2006_A4.pdf.

model agreement is found for HFI – but not for LFI – over Alaska, eastern Canada, and north-western and eastern Russia. In some regions increases are not associated with a strong S2N (e.g. for high flows over western China and the Horn of Africa).

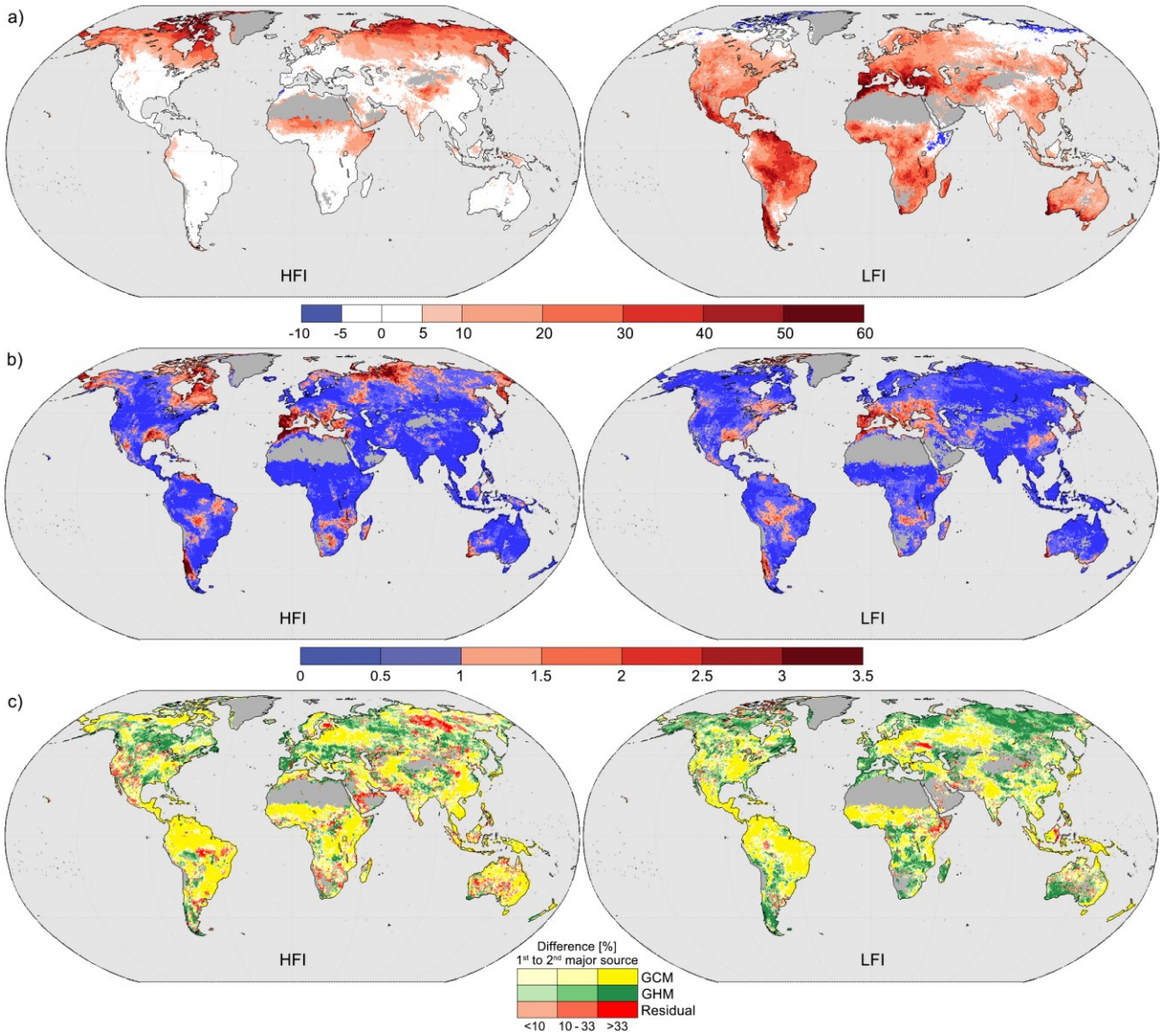


Figure 4.1 – Change in the frequency (in %) of days under high (left) and low (right) flow conditions for the period 2066–2099 relative to 1972–2005, based on a multi-model ensemble (MME) experiment under RCP8.5 from five GCMs and six GIMs: (a) MME mean change and associated (b) signal-to-noise ratio; (c) Proportion of variance per factor for the MME mean change: GCM (yellow), GIM (green), Residual (red). The colour of the dominant factor is given.

Mean changes and S2N for boreal winter (DJF) and summer (JJA), in Figure 4.3 and Figure 4.4 respectively, show an increased intensity with very similar spatial patterns to their annual counterparts in DJF for the high flows and in JJA for low flows. Conversely, high flows in JJA show virtually no change, while low flows in DJF show decreases at high northern latitudes with high model agreement and increases elsewhere with smaller model agreement (S2N). This can be seen also in Figure 4.2: the PDF (i.e. the density of the mean change percentage) stretches towards higher mean changes for high flows in DJF and for low flows in JJA. Global results are dominated

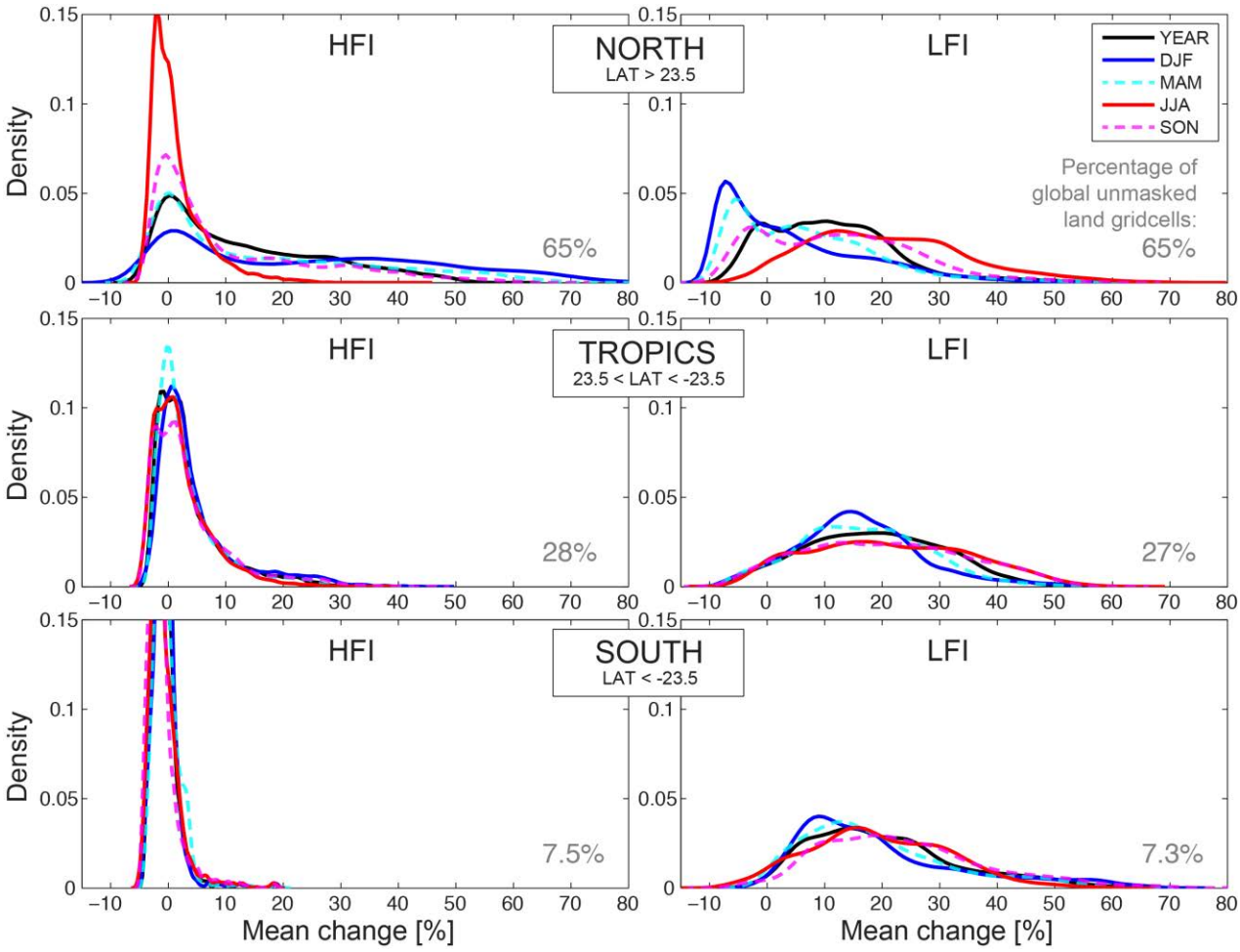


Figure 4.2 – PDFs of mean changes in high (HFI) and low (LFI) flows, annually and per season (DJF and JJA) for North, Tropics, and South latitude bands. Based on a multi-model ensemble (MME) experiment under RCP8.5 from five GCMs and six GIMs.

by boreal seasonality (high flow changes dominant in DJF, and low flow changes dominant in JJA) as the majority of global land cells 65% (of unmasked land) are located north of latitude 23.5. The remainder of the land cells (35%) are located within the Tropics and south latitude bands, and depict weak changes for high flows in all seasons, and increased changes for low flows in all seasons, though JJA's are more marked.

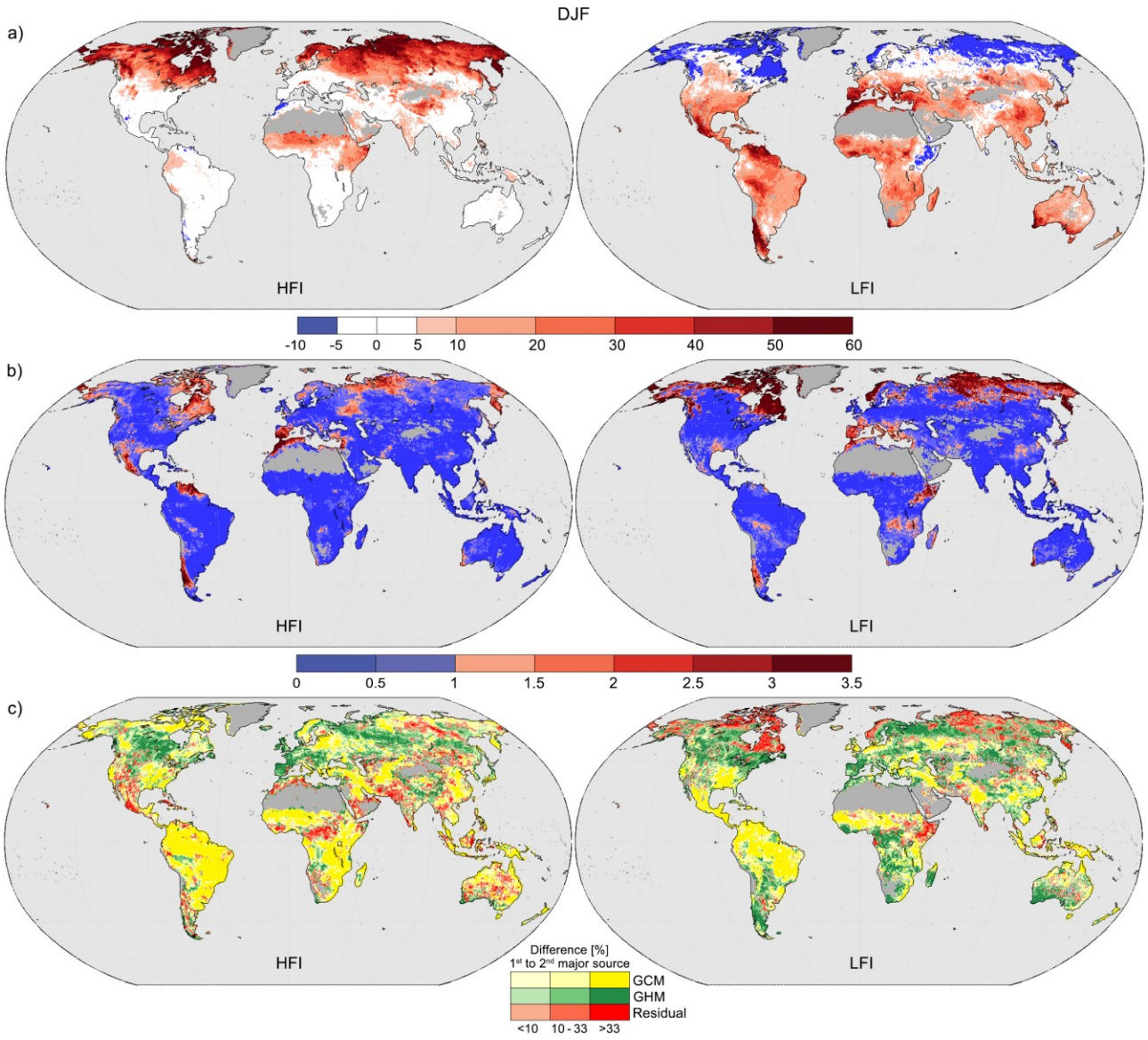


Figure 4.3 – As Figure 4.1, for the season DJF.

The results of the ANOVA across the 30 members of HFI and LFI are shown in Figure 4.1c; they are expressed, for each factor, as the proportion of sum of squares divided by the total sum of squares (refer to APPENDIX II for residuals testing for model adequacy). For the high flows, the variance is explained mostly by the GCMs (yellow, 47 % of unmasked land, Figure 4.1c), although

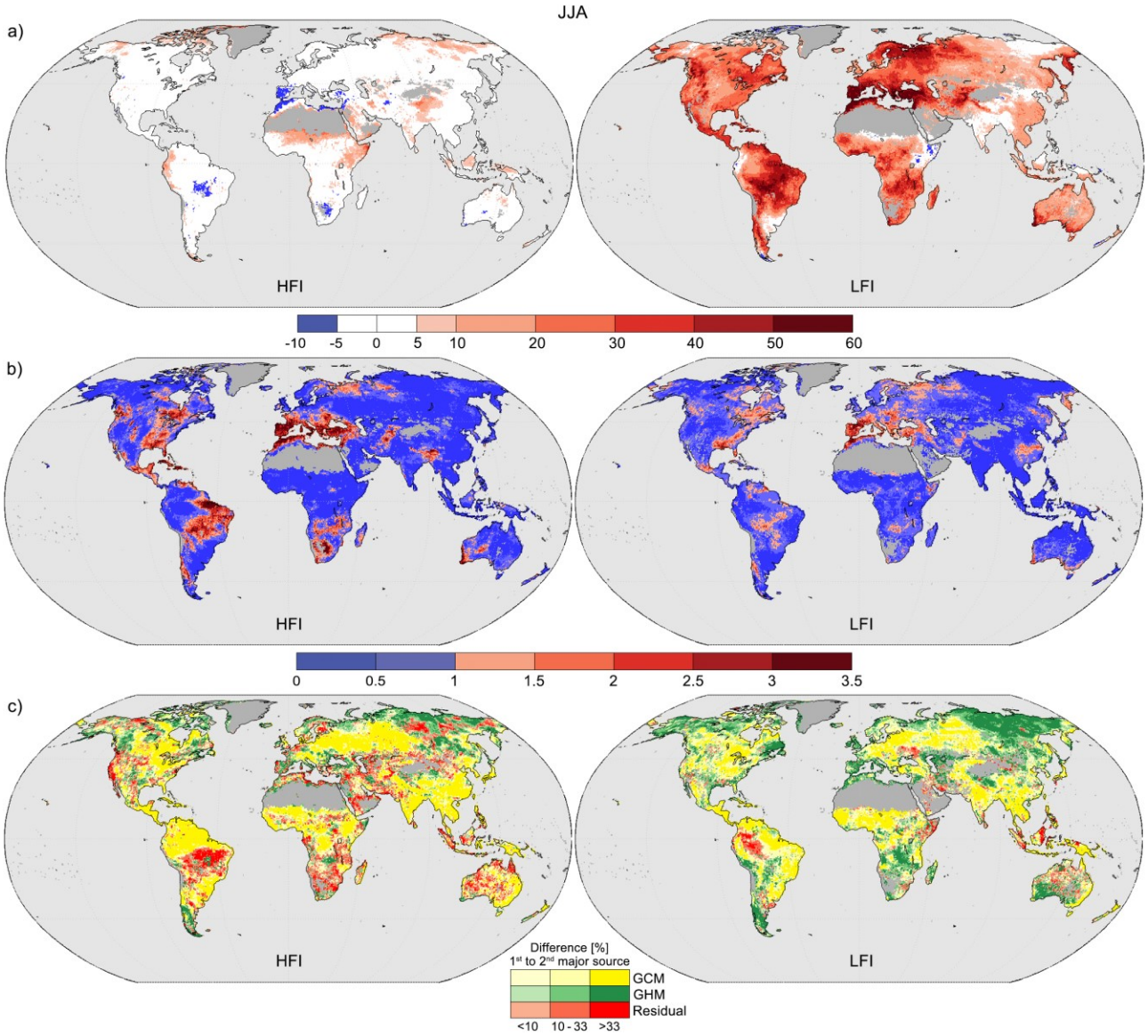


Figure 4.4 – As Figure 4.1, for the season JJA.

the GIMs are the major factor over western Europe and central Canada (green, 28% of unmasked land, Figure 4.1c). For low flows, the proportions change: the GCMs (43%) remain the major contributors over the globe, but GIMs (35%) increase to a relative influence closer to the GCMs, and become the major factor in some northern (e.g. north-eastern Russia) and southern (e.g. southern Africa, south-western Australia) regions. Seasonal results (Figure 4.3c and Figure 4.4c) are very similar to their annual counterparts in the case of high flows in DJF and low flows in JJA, whereas for high flows in JJA and for low flows in DJF higher residual rates (i.e. decreased overall GIM and GCM contributions) are found, perhaps owing to fewer events occurring in these seasons for both low and high flow indices.

To capture better the spatial distribution of the major sources of uncertainty, ANOVA results are aggregated by climatic homogeneous regions based on the climatological Köppen-Geiger classification. Scatterplots in Figure 4.5 show the proportions of sums of squares of GIMs (y-axis)

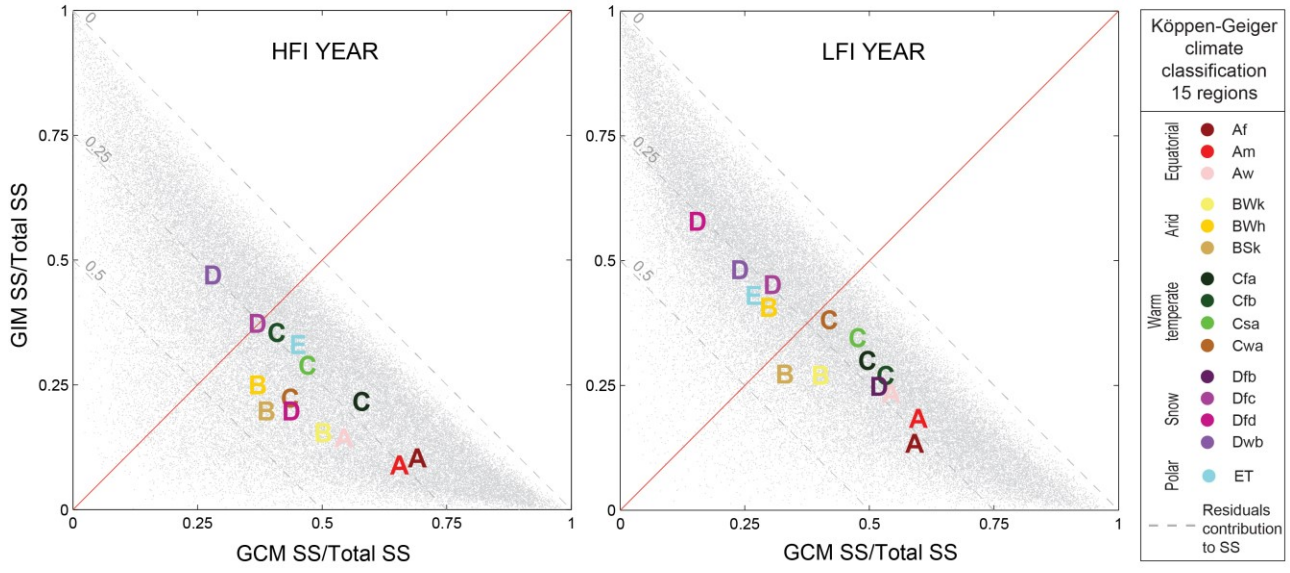


Figure 4.5 – ANOVA sum of squares (SS) of the two factors (GIM y axis; GCM x axis) divided by the total sum of squares (TSS) for all grid cells as grey dots; and for each Köppen–Geiger climate region (15 most represented), as region letters shown at the medians of the region's GCM SS/TSS as x-coord and of the region's GIM SS/TSS as y-coord.

vs. GCMs (x-axis); medians for each climatic region are shown as their class letter and summarize the prominent factor of uncertainty. For both high and low flows calculated over the year and seasonally, uncertainty in equatorial regions (A) is dominated by GCMs (median closest to the x- axis); while in snow-dominated climate (D) it is dominated by GIMs (median closest to the y-axis).

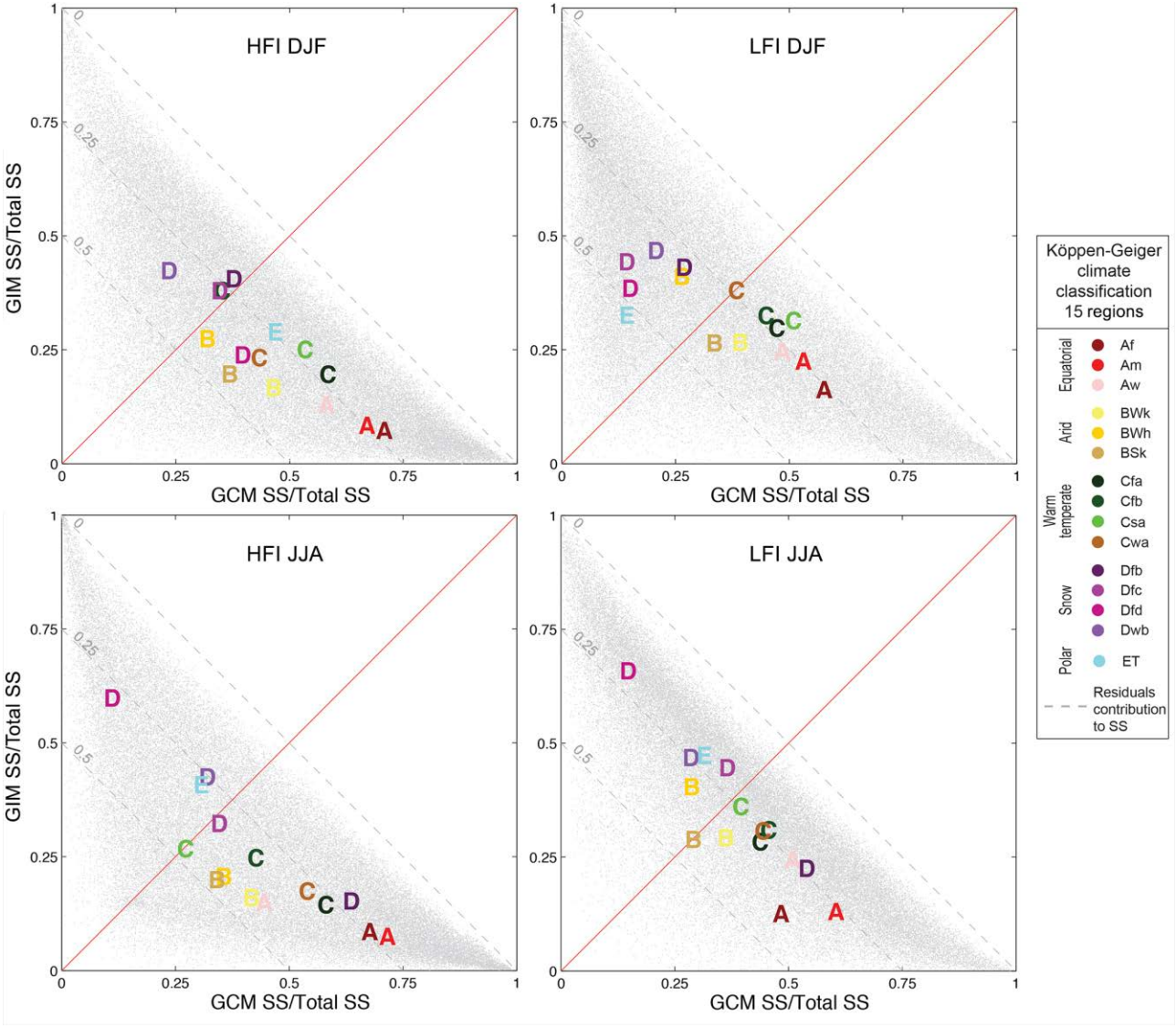


Figure 4.6 – As Figure 4.5, for the seasons DJF (top) and JJA (bottom).

In warm temperate regions (C), uncertainty is slightly higher for GCMs than GIMs. In arid regions (B), the variance is not well explained by either GCMs or GIMs (median farthest from 1; i.e. residuals explain most of the variance), suggesting that reproducing hydroclimatology over these regions represents a challenge for both GCMs and GIMs.

The ANOVA results for the whole year and those for winter and summer seasons (DJF and JJA shown in Figure 4.3c and Figure 4.4c) are quantified further in Table 4.2. This table provides a breakdown with both the regional and global results expressed for mean changes, S2N and percentage of sum of squares per factor at the annual and seasonal (DJF and JJA) scale. Looking jointly at the annual and seasonal results in Table 4.2, it is clear that the widespread dominance of the GCMs' contribution to uncertainty is outweighed by the GIMs in the snow- and ice-dominated regions (D). This pattern is visible also on the scatterplots (Figure 4.5 and Figure 4.6) with the GIM uncertainty-dominated regions (near the y-axis) often populated by D regions for both HFI and LFI (although to a lesser extent for the former).

Table 4.2 – Summary of mean changes, signal-to-noise S2N, and sources of variance for high and low flows at the annual and seasonal (DJF, JJA) scale, and at the global and climate region scale. The first source of variance is shown in grey, the second in light grey.

			YEAR					DJF					JJA					
Koppen-Geigen Region*		Area [Km²]	Mean[1] change [%]	Signal[2] to noise adim.	Source of variance GCM [%]	GIM [%]	Resid. [%]	Mean change [%]	Signal to noise adim.	Source of variance GCM [%]	GIM [%]	Resid. [%]	Mean change [%]	Signal to noise adim.	Source of variance GCM [%]	GIM [%]	Resid. [%]	
Equatorial																		
1	Af	2468	3.1	0.5	65.5	14.2	20.3	3.4	0.5	64.4	11.4	24.3	3.2	0.6	64.0	12.4	23.6	
2	Am	1836	1.4	0.5	61.6	15.5	22.9	2.8	0.4	60.2	14.8	25.0	0.7	0.9	65.9	11.0	23.1	
3	Aw	6017	1.6	0.6	52.7	20.4	27.0	2.9	0.4	56.3	18.3	25.5	0.5	0.9	47.0	19.7	33.3	
Arid																		
4	Bwk	3095	4.1	0.5	49.7	19.8	30.5	5.5	0.4	47.1	20.7	32.3	2.3	0.7	42.9	20.3	36.8	
5	Bwh	3139	1.2	0.6	39.3	28.5	32.2	3.1	0.6	35.6	29.6	34.8	0.1	0.7	37.7	25.3	37.1	
6	BSk	4255	6.4	0.4	40.5	25.1	34.4	7.0	0.4	39.6	23.9	36.4	5.1	0.5	36.2	25.5	38.3	
Warm temp.																		
7	Cfa	2955	-0.6	0.6	56.1	26.8	17.1	-0.5	0.4	57.1	22.9	20.0	-0.8	0.8	55.4	20.1	24.5	
8	Cfb	2360	0.4	0.9	45.0	35.9	19.0	1.8	0.6	40.5	37.7	21.9	-0.5	1.0	45.4	29.1	25.5	
9	Csa	1099	-1.2	1.5	45.3	32.8	21.9	-0.8	1.3	50.8	30.2	19.0	-2.0	1.9	31.0	30.5	38.5	
10	Cwa	1504	0.9	0.6	43.8	28.0	28.2	0.9	0.4	45.6	27.6	26.8	1.1	0.8	49.8	23.9	26.3	
Snow																		
11	Dfb	4459	5.4	0.7	42.7	37.1	20.2	17.5	0.7	38.2	40.3	21.5	-1.1	0.8	57.3	21.6	21.1	
12	Dfc	11008	18.6	1.2	38.0	38.8	23.2	37.5	1.0	36.0	39.7	24.2	0.2	0.7	39.7	34.2	26.2	
13	Dfd	1405	27.4	1.1	42.7	22.4	34.9	39.8	0.9	38.1	26.2	35.6	7.4	0.4	13.6	55.6	30.8	
14	Dwb	1311	6.7	0.5	29.1	46.1	24.8	11.0	0.4	25.7	41.9	32.3	1.4	0.3	35.0	41.9	23.1	
Polar																		
15	ET	5937	26.3	1.3	42.9	36.1	20.9	40.6	1.2	43.7	32.2	24.1	5.5	0.5	34.3	41.0	24.7	
-	Global	128.9M	6.5	0.7	46.5	28.0	25.5	11.8	0.6	45.5	27.5	27	1.3	0.8	45.0	25.4	29.7	
Equatorial																		
1	Af	2463	14.6	0.5	58.7	18.1	23.1	12.7	0.3	56.5	20.4	23.1	15.4	0.5	50.1	17.1	32.8	
2	Am	1834	23.7	0.7	57.0	25.2	17.8	19.7	0.4	50.5	29.1	20.4	27.1	0.7	59.5	18.1	22.4	
3	Aw	5997	21.7	0.7	52.4	28.5	19.0	18.4	0.5	48.2	29.8	22.0	25.7	0.6	50.8	28.3	20.9	
Arid																		
4	Bwk	2927	15.6	0.5	41.4	31.0	27.6	14.3	0.5	40.0	31.1	28.9	17.1	0.4	38.4	32.1	29.5	
5	Bwh	2821	20.2	0.6	32.1	42.6	25.3	18.4	0.5	29.2	42.8	28.0	22.2	0.6	30.3	42.9	26.7	
6	BSk	2693	14.3	0.6	35.9	32.1	32.0	13.5	0.6	35.0	31.2	33.7	15.1	0.5	33.8	33.3	33.0	
Warm temp.																		
7	Cfa	2950	18.2	1.0	49.1	32.9	18.1	17.6	0.7	47.5	32.4	20.0	19.1	0.9	44.1	30.9	25.0	
8	Cfb	2358	20.2	1.1	51.7	32.3	16.0	15.2	0.8	43.7	36.4	19.9	24.4	1.0	46.2	33.6	20.1	
9	Csa	1096	35.7	1.4	47.0	37.6	15.5	31.0	1.3	48.7	35.8	15.5	41.9	1.4	41.0	37.7	21.3	
10	Cwa	1500	18.5	0.8	42.0	39.4	18.5	18.1	0.7	39.5	39.8	20.7	18.4	0.7	44.7	34.1	21.2	
Snow																		
11	Dfb	4440	15.8	0.8	50.6	28.1	21.3	4.1	0.5	29.8	43.5	26.7	26.3	0.9	52.4	26.0	21.6	
12	Dfc	10920	8.7	0.5	33.6	44.8	21.7	-2.0	1.5	17.4	45.1	37.5	25.0	0.8	38.9	43.1	18.1	
13	Dfd	1402	-2.5	0.7	15.3	59.3	25.4	-5.7	2.3	16.8	40.1	43.1	4.4	0.2	14.5	66.4	19.1	
14	Dwb	1306	9.5	0.3	26.8	48.5	24.7	9.9	0.3	23.3	47.3	29.4	11.4	0.5	31.7	46.7	21.5	
Polar																		
15	ET	5650	3.4	0.5	29.8	45.0	25.2	-1.7	2.1	20.2	37.9	41.9	14.3	0.5	35.2	46.4	18.3	
-	Global	122M	16.1	0.7	43.1	34.8	22.1	11.8	0.8	36.6	35.9	27.6	21.5	0.7	42.5	34.2	23.3	

HIGH FLOWS

LOW FLOWS

[1] Mean change weighted over gridcells surface areas

[2] Signal-to-noise weighted over gridcells surface areas

Source of variance 1st 2nd

4.5 Discussions and conclusions

Using six global impact models (GIMs) fed by five global climate models (GCMs) under the RCP8.5 scenario, this study aimed to assess future high and low flow changes globally by the 2080s, and to quantify the uncertainty attributable to GIMs and GCMs. We decided to focus solely on the uncertainty coming from GIMs and GCMs using as many ensemble members (from the ISI-MIP project data set) as possible under the RCP8.5, in which change signals are expected to be larger (i.e. emissions continue to rise leading to global radiative forcing levels of 8.5 W/m^2 by the end of the 21^{st} century). The hydrological simulations used in this study do not account for anthropogenic influences (e.g. water abstraction, augmentation and artificial storage) or land-use changes.

High and low flow changes in the future (2066–2099) relative to the control period (1972–2005) exhibit a number of robust large-scale features. Increases in high flow days were found at northern latitudes, with a strong signal over eastern Canada, Scandinavia, north-western Russia, and around the Bering Sea (eastern Russia and Alaska). Increases in low flow days were found in southern Europe, south-western and central Latin America, south-eastern USA, more southerly parts of Central Africa, and south-western Australia. These patterns are largely consistent with the few other studies carried out on runoff at the global scale with several GIM–GCM combinations: e.g. for high flows (Hirabayashi et al., 2013), low flows (Prudhomme et al., 2014; van Huijgevoort et al., 2014) and for mean flows (Davie et al., 2013; Hagemann et al., 2013; Schewe et al., 2014). More specifically, the comparison of flood hazard patterns by Dankers et al. (2013) with the changes in the occurrence of high flow days from our study reveals some similarities, mostly northern North America and Northern Asia, while in some regions like north-eastern Europe

patterns are opposite. Low flow patterns are similar to Prudhomme et al. (2014) although they find a weaker S2N.

In this study we provide for the first time a comprehensive assessment of both ends of the runoff spectrum at the same time using the same data set globally. Moreover, we undertake a consistent partition of uncertainty via ANOVA for both high and low flows, showing that GCMs provide the largest uncertainty, although the GIM contribution can be substantial in particular regions. The results from our ANOVA framework are consistent with other global studies based on the ratios between the variances (or standard deviations) of ensemble members averaged per type of model (Dankers et al., 2013; Hagemann et al., 2013; Schewe et al., 2014). In particular, uncertainty results that Dankers et al. (2013) expressed with GCM/GIM variance are in agreement with our findings for high flows in the Southern Hemisphere, mainly driven by GCM uncertainty, whereas there is less agreement for the Northern Hemisphere (in North America, Central Canada is GCM-driven uncertainty, whereas it is GIM driven in our results). Uncertainty results for low flows from Prudhomme et al. (2014), expressed as S2N ratio, are not directly comparable, but as will be discussed later, the inclusion of the JULES GIM in their ensemble has pointed to lower model agreement (i.e. increased uncertainty).

At the regional level, the uncertainty partition enables us to delineate in which climate region each factor (GCMs or GIMs) provides the largest uncertainty at the annual and seasonal scales. Notably, for snow- and ice-dominated polar regions, and for arid zones, GIMs bring about the largest portion of uncertainty, especially for low flows. This is likely to reflect uncertainty in the way the hydrological storage–release processes can modify the climate signal, particularly where these storage components are relatively large or water residence times high – hence the importance of considering several GIMs in studying changes in high and low flows. GCM and GIM uncertainty shares are similar for HFI and LFI globally, although the spatial patterns differ slightly (e.g. north-

eastern Russia, south-western Australia and Alaska are GCM driven in HFI, and GIM driven in LFI). This could reflect different dominant processes for high and low flow generation, with high flow events mainly driven by precipitation inputs or snow/ice-melt (i.e. atmospheric-driven processes); whereas low flows event develop over longer durations and are influenced more by land-surface processes like evaporation, infiltration and storage, which are simulated by the GIMs, each one with its own scheme and parameterization: e.g. for evapotranspiration, Penman–Monteith, Hamon (Haddeland et al. 2011; and Table 3.2). Haddeland et al. (2011) have identified in the snow scheme employed by different GIMs a major source of difference between the model runoff simulations, and recent studies at global (e.g. Hagemann et al. 2013) and regional scale (e.g. Jung et al. 2012) hint at an increase in uncertainty in snow-dominated regions. Our study shows that in snow-dominated and arid regions GIM uncertainty equals or outweighs GCM uncertainty for both high and low flows, highlighting the importance of comprising balanced sets of both global hydrological and climate models to encompass the overall uncertainty in these regions.

To put the current study in context and to provide suggestions for further studies, it is worth making a few considerations on the hydrological index extraction and clarify a few aspects of the uncertainty partition concerning the method and the data set we used.

The identification of high and low flows over long time series, and particularly over climate projections, is nontrivial. As an illustration, van Huijgevoort et al. (2014) in their multi-model ensemble study on droughts report that applying the threshold level method to the future period using a threshold derived from the control period can lead to spurious pooling of drought events. They suggest that future changes could be accounted for by linking the drought threshold to adaptation scenarios like Vidal et al. (2012) did over France. Wanders et al. (2015) used a transient threshold level method for a moving reference period, in order to reflect the changes in hydrological regime over time, finding that the nontransient threshold method projected larger shares of areas in

drought (except in snow-dominated regions). For our study, the threshold was calculated over the control period, as changes in future extremes with respect to present day were sought. In general, the selection of threshold approach should consider that if, on the one hand, a consistent pooling of extreme events may be hampered by incremental shifts or shape changes of the hydrograph throughout the future; on the other hand, when assessing the changes in frequency with respect to the present, information on the present used for comparison is lost when the threshold adapts throughout the projections.

The model runs used in this study have no replicates; therefore, our ANOVA partition set-up poses some limitations as it assumes that the factors do not interact (no degrees of freedom are available for the estimation of the experimental error). However, interactions between the factors may indeed be present and, as pointed out by Bosshard et al. (2013), these interactions may represent uncertainty contributions that do not behave linearly: e.g. a snowmelt bias of a GIM may depend on the temperature projection of the driving GCM that could lead to a nonlinear response in the simulated runoff. This could in part explain the high rate of residuals' contribution seen in some grid cells for which potential interactions hinder the ANOVA to properly disclose the factors main effects. To avoid this drawback multiple model runs would be necessary.

Bias correction and CO_2 and vegetation dynamics represent other sources of uncertainty that were not accounted for in this study, though their influence should be further investigated in future works. Bias correction is commonly used to overcome bias inconsistencies between GCMs and GIMs in climate impact studies; however, this technique alters the model output by e.g. reducing the inter-GCM variability and potentially their contribution to total uncertainty in climate projections (Dankers et al., 2013; Wada et al., 2013), and it is argued that its use is not always justified (Ehret et al., 2012). Hagemann et al. (2011) even found that uncertainty due to bias-correction can be of the same order of magnitude as that related to the choice of GCM or GIM.

As Huber et al. (2014) points out, findings on relative contributions of GCMs and GIMs to total impact uncertainty would need to stand the test of using non bias-corrected runs, but runs that have not been bias corrected (with a method designed to preserve the long-term trends in temperature and precipitation projections, Hempel et al. 2013) are unavailable within ISI-MIP or with the same GCM/GIM combinations.

As mentioned in the Introduction, biome models have shown a larger spread than GIMs without varying CO_2 and vegetation dynamics processes, and it is argued that, due to the additional processes that they simulate, the inclusion of biome models in multi-model ensemble studies is important to capture a comprehensive range of uncertainty (Davie et al., 2013; Prudhomme et al., 2014). Within our study specifically, biome models with runs at daily resolution were JULES and LPJmL. These models were excluded primarily for intractability in low flow analysis. Therefore, uncertainty from varying CO_2 is not sampled and could suggest overconfidence (or bias) in favour of non-biome GIMs, which simulate less runoff than biome models. During our exploratory analysis we actually included JULES in the ensemble and found that the uncertainty was driven towards the GIM source (in agreement with Prudhomme et al. 2014, who found higher S2N, i.e. stronger agreement between the models, when considering the ensemble without JULES). However, the inclusion of models in the ensemble must be compatible with the applicability of the method, and the biome models available through ISI-MIP proved to hamper the global comparison assessment for the heavy masking over large areas with zero-rich time series. As shown in Table 4.2, low flow index extraction was vetoed over large areas of the globe, ultimately leaving 61 and 20 % of land cells for JULES and LPJmL respectively (note that the masking is formed by superimposing masking from each GIM–GCM combination). Also, JULES' coarser resolution (7558 vs. 67420 total land grid cells for JULES and the other GIMs respectively, i.e. a ratio of 1 to 9 cells) may contribute to more uncertainty, although lower-resolution runs would be necessary to

assess such contribution. Index extraction for high flows proved more favourable, but we adopted the pragmatic approach of using the largest possible ensemble of models common to both high and low flows. We are aware that the inclusion of multiple models is not sufficient to fully scope model uncertainty due to resolution and structural errors that are common across models and place a limit to the confidence we obtain from robustness (Knutti, 2010). However, our results demonstrated that, even excluding biome models and other model structure differences in the ISI-MIP ensemble, large uncertainty in the signal of changes in high and low flows is attributable to GIMs and not only on GCMs.

Were biome models' shortcomings not present, their inclusion in our ensemble would have required a modification of our uncertainty partition strategy because the presence of outliers (likely introduced by biome models) would limit our ANOVA model (whose assumptions include no or minimal presence of outliers). For their distinct behaviour from the other GIMs, biome models could be considered as a factor level in a two-way ANOVA framework with unequal sample sizes (Neter et al. 1999, Chap. 23), i.e. the spread of future hydrological extremes would be examined as the function of factor 1 – the type of hydrological model (level 1: six GIMs; level 2: two biome models) and factor 2 – the GCMs.

Finally, the focus of our uncertainty analysis was on GCMs and GIMs, therefore the effect of emission scenarios (RCPs) was neglected. The few studies that have considered this aspect hint at a relatively small role of emission scenarios (Hagemann et al., 2013; Wada et al., 2013) all throughout the 21st century when compared to GCMs and GIMs, which play a stronger role in uncertainty contribution over most of the globe.

To conclude, knowledge of the dominant source of uncertainty in climate-to-hydrology signal is critical to modellers for improving modelling of the terrestrial water cycle and to scientists for putting together targeted multi-model ensembles for climate impact studies. In addition to GIMs

and GCMs, further work is needed to assess the degree to which internal variability, bias correction, biome models (i.e. GIMs that simulate vegetation dynamics and varying CO_2) and emission scenarios contribute to total uncertainty.

5 Evaluation of global impact models' ability to reproduce runoff characteristics over the central United States

5.1 Abstract

The central United States experiences a wide array of hydrological extremes, with the 1993, 2008, 2013, and 2014 flooding events and the 1988 and 2012 droughts representing some of the most recent extremes, and is an area where water availability is critical for agricultural production. This study aims to evaluate the ability of a set of global impact models (GIMs) from the WaterMIP project to reproduce the regional hydrology of the central United States for the period 1963-2001. Hydrological indices describing annual daily maximum, medium and minimum flow and their timing are extracted from both modelled daily runoff data by nine GIMs and from observed daily streamflow measured at 252 river gauges. We compare trend patterns for these indices, and their ability to capture runoff volume differences for the 1988 drought and 1993 flood. In addition, we use a subset of 128 gauges and corresponding grid cells to perform a detailed evaluation of the models on a gauge-to-grid-cell basis. Results indicate that these GIMs capture the overall trends in high, medium, and low flows well. However, the models differ from observations with respect to the timing of high and medium flows. More specifically, GIMs that only include water balance tend to be closer to the observations than GIMs that also include the energy balance. In general, as it would be expected, the performance of the GIMs is the best when describing medium flows, as opposed to the two ends of the runoff spectrum. With regards to low flows, some of the GIMs having considerably large pools of zeros or low values in their time series, undermining their ability in capturing low flow characteristics and weakening the ensemble's output. Overall, this study

provides a valuable examination of the capability of GIMs to reproduce observed regional hydrology over a range of quantities for the central United States.

5.2 Introduction

Model intercomparison frameworks like WaterMIP, provide the opportunity to compare model simulations from a number of GIMs all driven with the same meteorological forcing: the WATCH Forcing Data – WFD (Weedon et al., 2011). As noted in Chapter 2 (Section 2.3), the WaterMIP GIMs have been evaluated with respect to low, medium and high flow in a number of studies (Gudmundsson et al., 2012b; Haddeland et al., 2011; Prudhomme et al., 2011; Stahl et al., 2012; Tallaksen and Stahl, 2014; Van Loon et al., 2012) showing considerable variability in the magnitude and timing of the components of the hydrological cycle. Notably, despite the WaterMIP global coverage, all of these studies focused on Europe, and – apart from two studies (Haddeland et al., 2011; van Huijgevoort et al., 2014) – little is known about the skill of these models in reproducing the hydrological processes for other regions of the world. In this study, we address this gap in our knowledge by aiming to examine the capability of nine GIMs to reproduce key features of the hydrological regime, including high, medium and low flow over the central United States (defined as the region between 36°N to 49.5°N and -105°E to -80°E): a region that experiences a wide array of hydrological extremes, with the 1993, 2008, 2013 and 2014 flooding events and the 1988 and 2012 droughts representing some of the most recent extremes, and where water availability is critical for agricultural production.

5.3 Data and Methods

In this study a first level of analysis uses a larger streamflow dataset to verify whether the models are able to capture overall trend patterns of regional hydrology and two specific extreme

events (1988 drought and 1993 flood), and a second level uses a smaller set of gauges (whose catchment have comparable size with the grid cells) to evaluate model performance matching observed and modelled data at the gauge-grid-cell scale. This framework was chosen to ensure a first level of analysis with a sufficient number of streamflow gauges for spatial representativeness in the trend (Section 5.3.4) and extreme events (Section 5.3.5) comparison and a robust second level of analysis on carefully selected pairs (Section 5.3.6).

The rationale behind this choice is that model evaluations must deal with a misalignment between modelled and observational data: as pointed out by other authors (e.g., Gudmundsson et al., 2012b), large-scale hydrological models are not designed to model runoff at the catchment scale and interpreting localized model performance by comparing it with observed data may yield misleading results. Modelled data are systematically distributed in grid cells over the study region at a given spatial resolution, while the observational records do not have the same homogeneous coverage. Also stream gauges provide an integrated measurement over a catchment (e.g., Hannah et al. (2011)), while the runoff information provided by the models represents values uniformly distributed over grid cells.

5.3.1 Simulated data

We use daily total (surface plus subsurface) unrouted runoff outputs from nine GIMs created as part of the WaterMIP project. WaterMIP, comprises both land surface models (LSMs), which solve at the land surface both the water and the energy balances, and global hydrology models (GHMs), which only solve the water balance. These models vary in structure and parameterization (see Section 3.3.2.2 and Table 3.2 for an overview of the models). All of the global models were run over the period 1963-2001 (except GWAVA: 1963-2000) at a spatial resolution of 0.5 decimal degrees and forced by the same meteorological input data (WFD).

The models vary substantially in the parameterizations of evaporation and runoff, and do not all use the same input variables or model time steps (in particular, all GHMs are run at a daily time step whereas LSMs are run at a sub-hourly time step).

None of the models were calibrated specifically for the WaterMIP experiment except WaterGAP, which underwent a limited calibration procedure using local measured streamflow data (Hunger and Döll (2008)).

5.3.2 Observations

The observed data, described in Section 3.3.3, consists of daily discharge data covering the 1963-2001 period from 252 stream gauging stations (shown in Figure 3.3 and listed in Table 3.3). The size of the gauging stations' catchments ranges from 64 to 1,350,000 km^2, with a majority (80%) with area up to 7000 km^2. Because no land use changes or water management interventions are accounted for in the modelled data, the 252 gauges were selected from the Hydro-Climatic Data Network (HCDN).

5.3.3 Hydrological indices

We aim to analyse changes in discharge over different parts of the flow regime (including high, medium and low flows). The central United States is a region marked by a high flow season mostly from April to July (e.g., Villarini et al. (2011)) and a low flow season usually from September to February. We focus on different hydrological indices extracted from daily discharge time series over the period 1963-2001 (except for GWAVA, for which data were available for 1963-2000) for both observed (252 gauges) and modelled (1350 grid cells) data. The hydrological year is January-December for high and medium flows indices, and April-March for low flows indices. We use (introduced in Section 3.4.1.20) three magnitude:

i) Annual Maximum Flow (AMax)

ii) Annual Median Flow (AMed)

iii) Annual Minimum Flow (AMin)

and three timing indices:

i) Annual Maximum Date (AMaxDate)

ii) Medium flows Date (V50Date)

iii) Drought Start Date (VDef10Date).

The latter timing index poses some limitations in the presence zero (or very low values) rich time series for which the index cannot be extracted or there are too few threshold crossings over the time series to provide useful information. Therefore, if the index has insufficient non-zero values (at least 25 over 38) it is screened out (shown in grey on the maps). In this regard, it is worth noting that other studies have highlighted how low flow tractability can be problematic for GIMs. For instance, Gudmundsson et al. (2012a) found that the performance of this same set of GIMs decreased systematically from high (Q_{95}) to low (Q_5) runoff percentiles over Europe. The ensemble median of the GIMs, calculated as the median of the single GIMs' indices series, was added to complement the results and assess whether its results are more satisfactory than for any of the GIMs.

5.3.4 Trend patterns in hydrological indices

A first step in our evaluation is geared towards the assessment of the skill of the GIMs in reproducing regional patterns of changes in the selected metrics, as well as their temporal evolutions. We examine temporal changes in discharge using the Mann-Kendall test (among others, consult Helsel and Hirsch (1992) for a description of this test). This is a non-parametric test (it does not require any distributional assumption) that allows the detection of monotonic patterns in the record of interest.

5.3.5 The 1988 drought and the 1993 flood

We selected two major hydrological extremes that occurred during the time of analysis (1963-2001), namely the 1988 summer drought, which affected most of the conterminous U.S., and the 1993 summer flood, which affected the U.S. Midwest particularly. Both events have developed over a time span of approximately three months, from June to August, as reported by the NOAA Billion-Dollar Weather and Climate Disasters[2]. We thus assessed how well the GIMs captured these events by considering the mean summer runoff volumes (from June 1st to August 31st) of the year in which the event occurred, and compared them to the mean summer runoff volumes over the whole time series. These differences are quantified using the following coefficient of variation (e.g. for the 1988 drought):

$$CV = Q_{JJA[88]} - Q_{JJA[63\text{-}01]} \,/\, \sigma(Q_{JJA[63\text{-}01]})$$

We thus map this quantity to show whether the models indicate negative (positive) balances for drought (flood). In addition, we express the exceedance probability (p) by ranking the years based on their summer runoff volumes, and compute the plotting position of the particular year event (1988, or 1993) with reference to the whole time series:

$$p = m \,/\, n+1$$

where m, is the rank position, and n is the number of years in record.

5.3.6 Modelled – Observed pairwise comparison

We carry out a pairwise comparison between observed and modelled discharge using a subset of 128 non-nested gauges, which were selected within the 400 to 3500 km^2 catchment area range (Figure 3.3) – while the size of the model grid cells ranges depending on the latitude from approximately 2500 km^2 at 36°N to 2000 km^2 at 49.5°N. The selection of the pairs was carried out

[2] https://www.ncdc.noaa.gov/billions/events

on a GIS using the streamgauges' catchment boundaries obtained from the National Weather Service[3]: the grid cell corresponding to a given catchment was selected on the basis of centroid proximity. Priority was given to larger catchments (i.e. with area closer to the grid cells) and, in case of more catchments overlapping over the same grid cell, the one that shared the majority of the area was selected. Because of the different units used for modelled and observed data – except for the timing indices (expressed in number of days from beginning of hydrological year) – the indices series for the observed streamflow data were converted from cubic feet per second to mm of runoff per unit area per second.

The comparison is carried out first on the timing indices assessing the monthly frequency of occurrence; this is followed by analyses on all of the index series using three performance metrics: Pearson correlation coefficient, computed to assess the similarity of the indices series across pairs, with optimal value R=1; the relative difference in standard deviation, computed to compare the amplitude of observed and modelled indices data, with optimal value $\Delta\sigma=0$; Root Mean Squared Error, computed to express the magnitude of the difference between observed and modelled indices series, with optimal value RMSE=0.

5.4 Results

5.4.1 Trend patterns

Results related to the temporal change in AMax (Figure 5.1), AMed (Figure 5.2), and AMin (Figure 5.3) are presented through maps showing the sign and significance of the results of the Mann-Kendall test. Note that grid cells were greyed out when the total runoff was negative. These negative values can be achieved if, for instance, there is high evaporation and no sufficient

[3] http://www.nws.noaa.gov/geodata/catalog/hydro/html/basins.htm

precipitation to generate runoff, as seen for the WaterGAP and JULES (Döll and Schmied, 2012; Williams and Clark, 2014). Grid cells were also greyed out when runoff was unavailable – for the Great Lakes (WaterGAP, LPJmL, MPI-HM, GWAVA), or when the hydrological index tested had null variance (e.g. all annual minima equal to zero). Also note that there are very few streamflow gauges in the south-western part of the study region. While there are a number of USGS stream gaging stations, a very small number are included in the HDCN, mostly because of large water withdrawal for agriculture (e.g., Rasmussen and Perry (2001)).

The 1-day annual maximum index based on the observations (Figure 5.1, top-left panel) shows a weak tendency towards increasing trends over most of the region, although the trends are generally not significant at the 0.1 significance level. These results are consistent with what discussed in the literature (e.g., Hirsch and Ryberg (2012); Peterson et al. (2013); Villarini et al. (2011); Vogel et al. (2011); Mallakpour and Villarini (2015)) where there is not a very strong indication of changes in extreme discharge over this area, but more of a tendency towards increasing trends. For the GIMs, MacPDM depicts a rather muted signal with virtually no significant trends over the entire region. In comparison, the remaining models show stronger patterns of change. In particular, WaterGAP, MPI-HM, and MATSIRO yield spatial patterns that more closely resemble the observations, with an even stronger signal of change than observed. Most of the models indicate a decreasing trend in northern Minnesota that could not be compared with the observations due to the lack of stream gaging stations in the area. The lack of observational records holds true for the area including Nebraska and Kansas, for which the models suggest increasing trends in annual maximum daily discharge. The models GWAVA, HTESSEL, JULES, and Orchidee show a generally noisier signal with both positive and negative trends over the region of study.

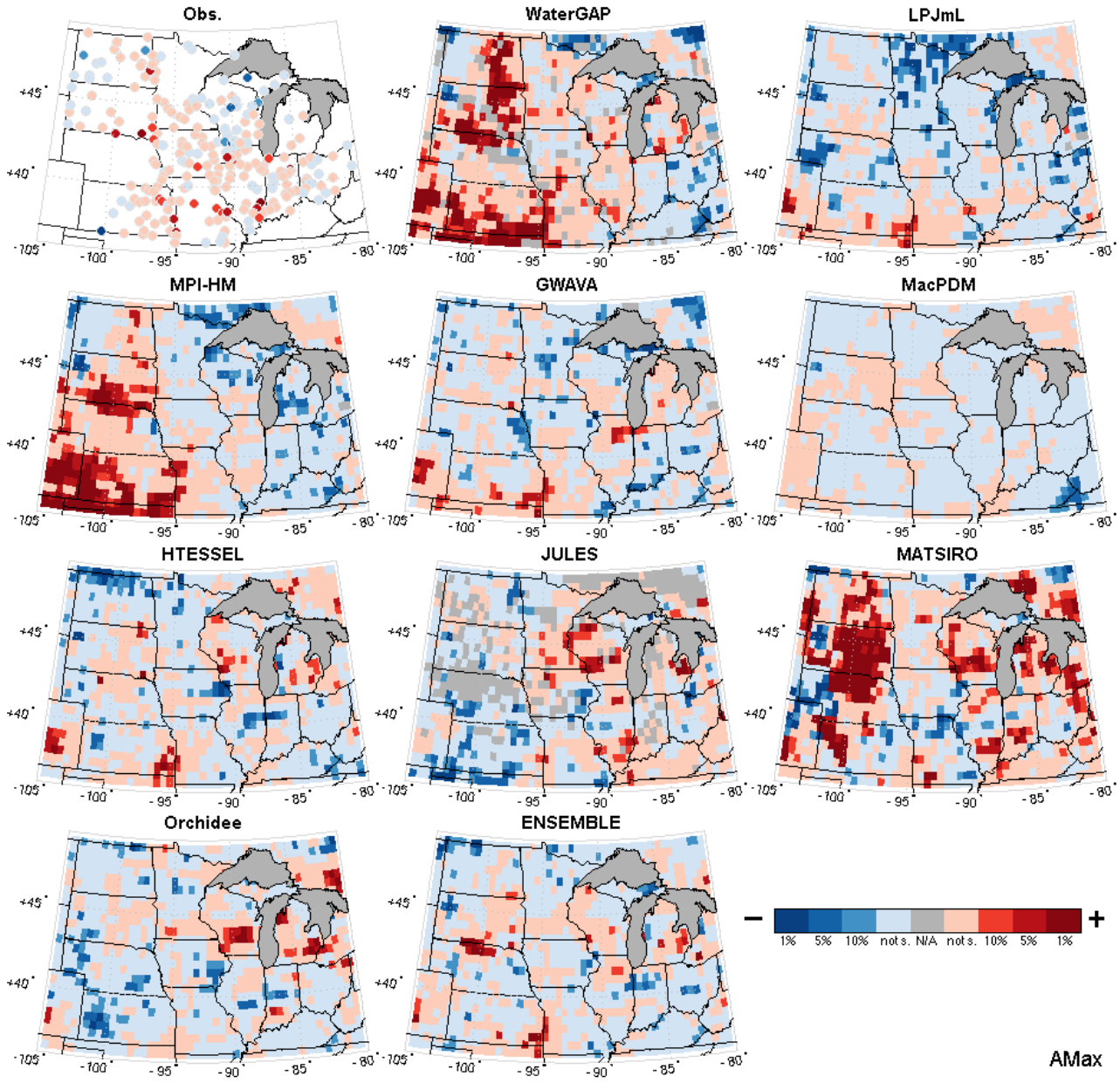


Figure 5.1 – Trends in the annual maximum flow for observed data (top left), the nine GIMs and their ensemble median. Negative trends are shown in blue and positive trends in red, with three levels of significance (1, 5, 10%) from pale (not significant) to dark (significant at the 1% level).

Trends in median (Figure 5.2) and minimum (Figure 5.3) discharge show a much clearer pattern than for the annual maximum daily series. These results are consistent with published work (e.g., Douglas et al. (2000); Lins and Slack (1999, 2005); McCabe and Wolock (2002)), in which most of the statistically significant increasing trends were detected for low to moderate quantiles,

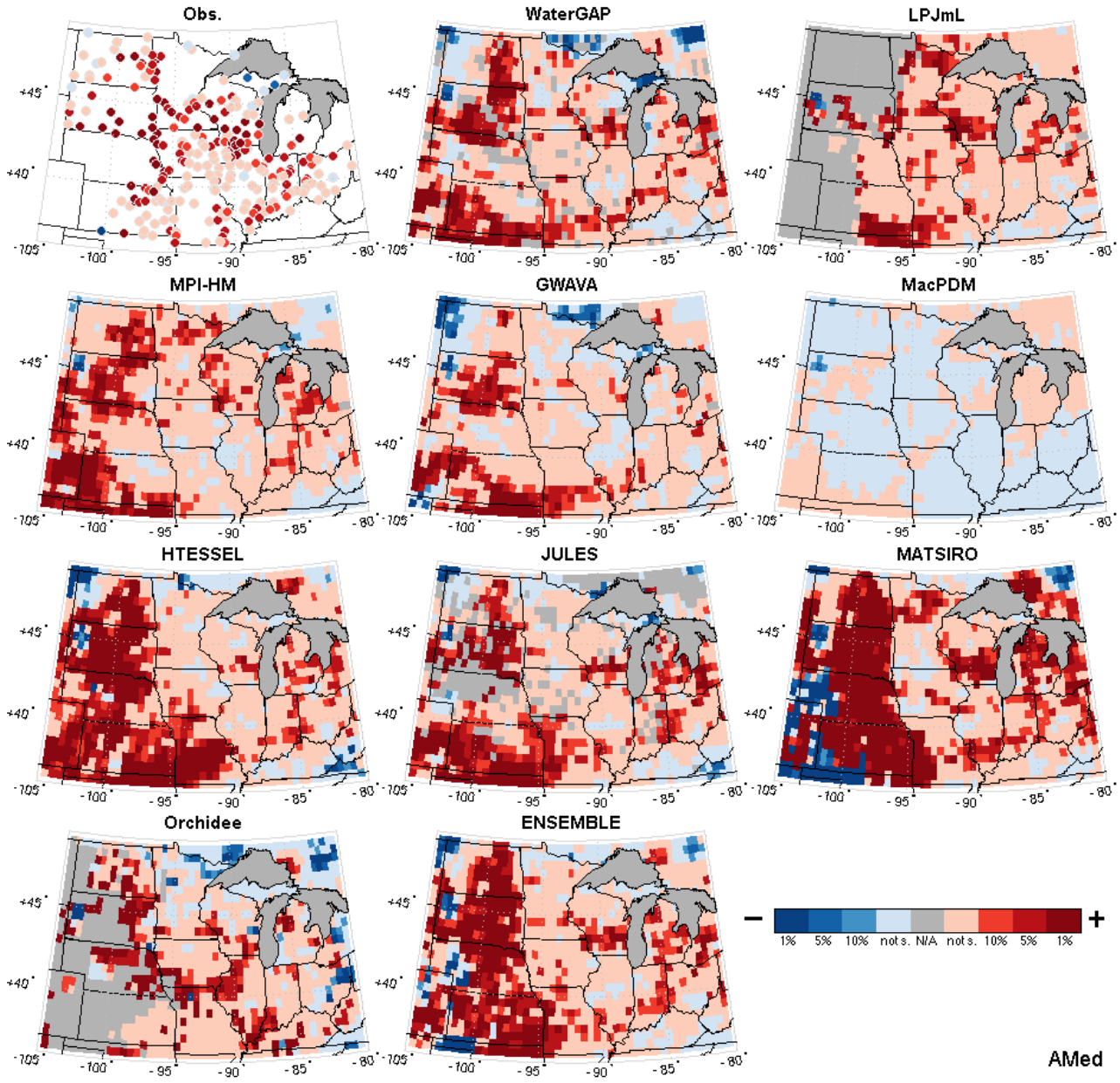


Figure 5.2 – Same as Figure 5.1 for annual median flow.

and much fewer when dealing with annual maximum discharge.

Trends in observed annual minimum indicate strong and highly significant (p-values generally < 0.01, i.e. 1%) increasing trends over most of the region, with the exception of the south-eastern part of the domain (weaker signal). Overall, the models capture well this increasing

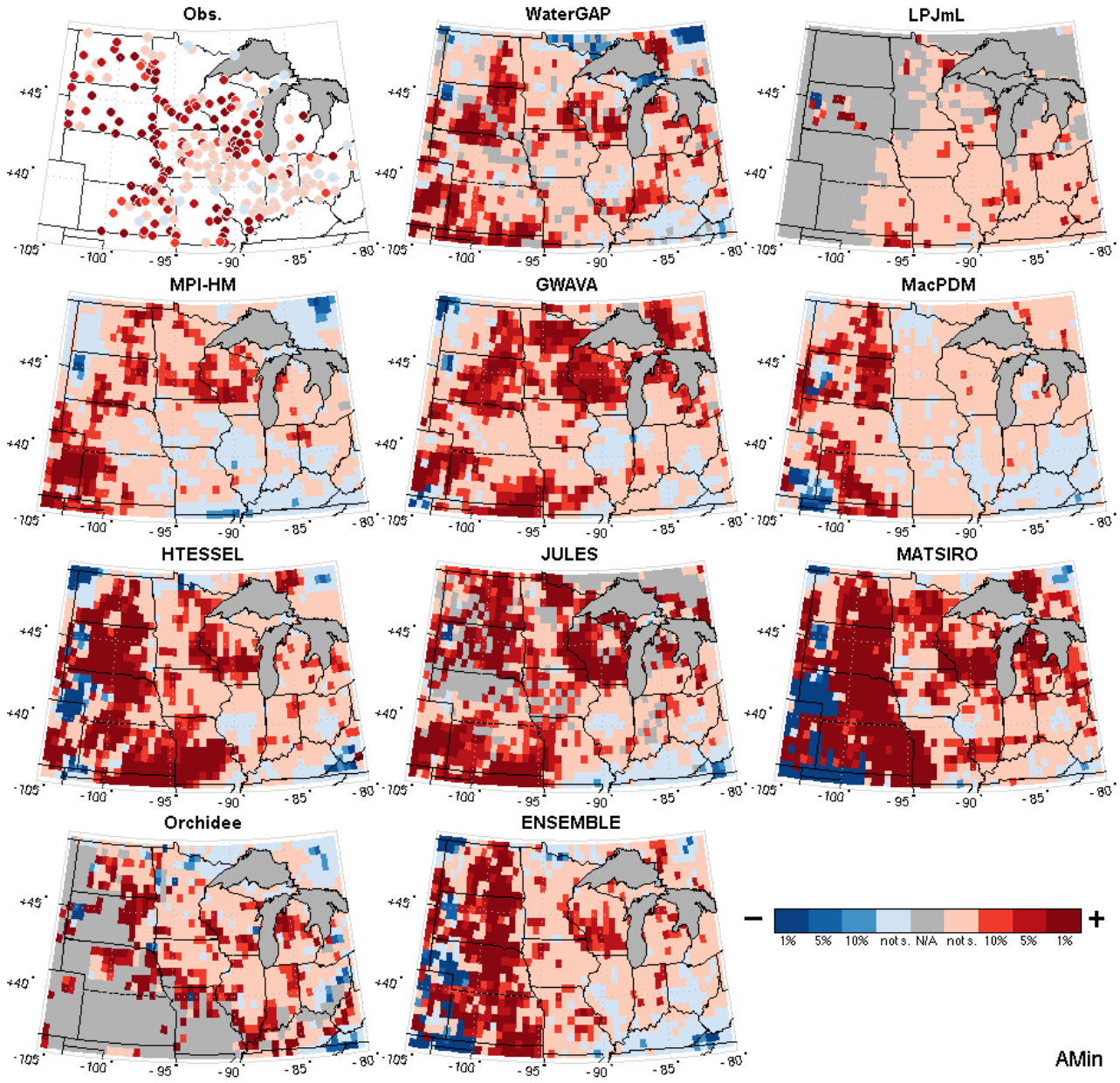


Figure 5.3 – Same as Figure 5.1 for annual minimum flow.

pattern. In particular, the LSMs (HTESSEL, JULES, MATSIRO) show strong increasing trends that are also detected, although not as strongly, in the GHMs (WaterGAP, GWAVA, MPI-HM) and to a lesser extent in MacPDM (positive significant detections are limited to the western part of the domain). The models LPJmL and Orchidee have a substantial number of grid cells screened out (grey), where the annual minimum is equal to zero over the 38 years considered. This behaviour

results in a large part of the pixels being removed from the analysis in south-west for Orchidee, and in the west and the north-east for LPJmL. In the unmasked areas, Orchidee reproduces well the spatial signal patterns with positive trends, whereas LPJmL shows no significant detections (this is also the case for MacPDM over the same area). Thus, LPJmL and MacPDM do not seem to capture the overall trend in runoff Annual Minima as well as the other GIMs.

Trends in median flow (Figure 5.2) are broadly similar to those for the annual minimum flow, with most GIMs capturing the observed overall increasing signal. In contrast with the other GIMs, MacPDM has virtually no significant trends. Although less than for the AMin, LPJmL and Orchidee have grid cells screened out even for the median flow. This is rather surprising because it indicates that at least half of the days every year have daily discharge equal to zero. At this stage, it is unclear what the issues with these two models are, although this issue was also noted by Gudmundsson et al. (2012a) where the two GIMs have constant low values of interannual variability at low percentiles (i.e. Q_5, Q_{25}), and by Prudhomme et al. (2014) where LPJmL displays a similar behaviour in the runs of the ISI-MIP experiment.

We focused also on the timing of high (AMaxDate; Figure 5.4), medium (V50Date; Figure 5.5), and low flows (Vdef10Date; Figure 5.6) to aid inference of the discharge-generating processes over this region. The observations do not point to a change in the seasonality of high flow or medium discharge, with no statistically significant (at the 0.1 level) trends. The lack of a clear spatial pattern and significant trends in the date of annual maxima is reproduced by most GIMs (WaterGAP, GWAVA, HTESSEL, JULES, MATSIRO, and Orchidee; Figure 5.4). However, decreasing trends are simulated in the north/ north-east part of the region by three of the GHMs (LPJmL, MPI-HM, and MacPDM) and by the ensemble median.

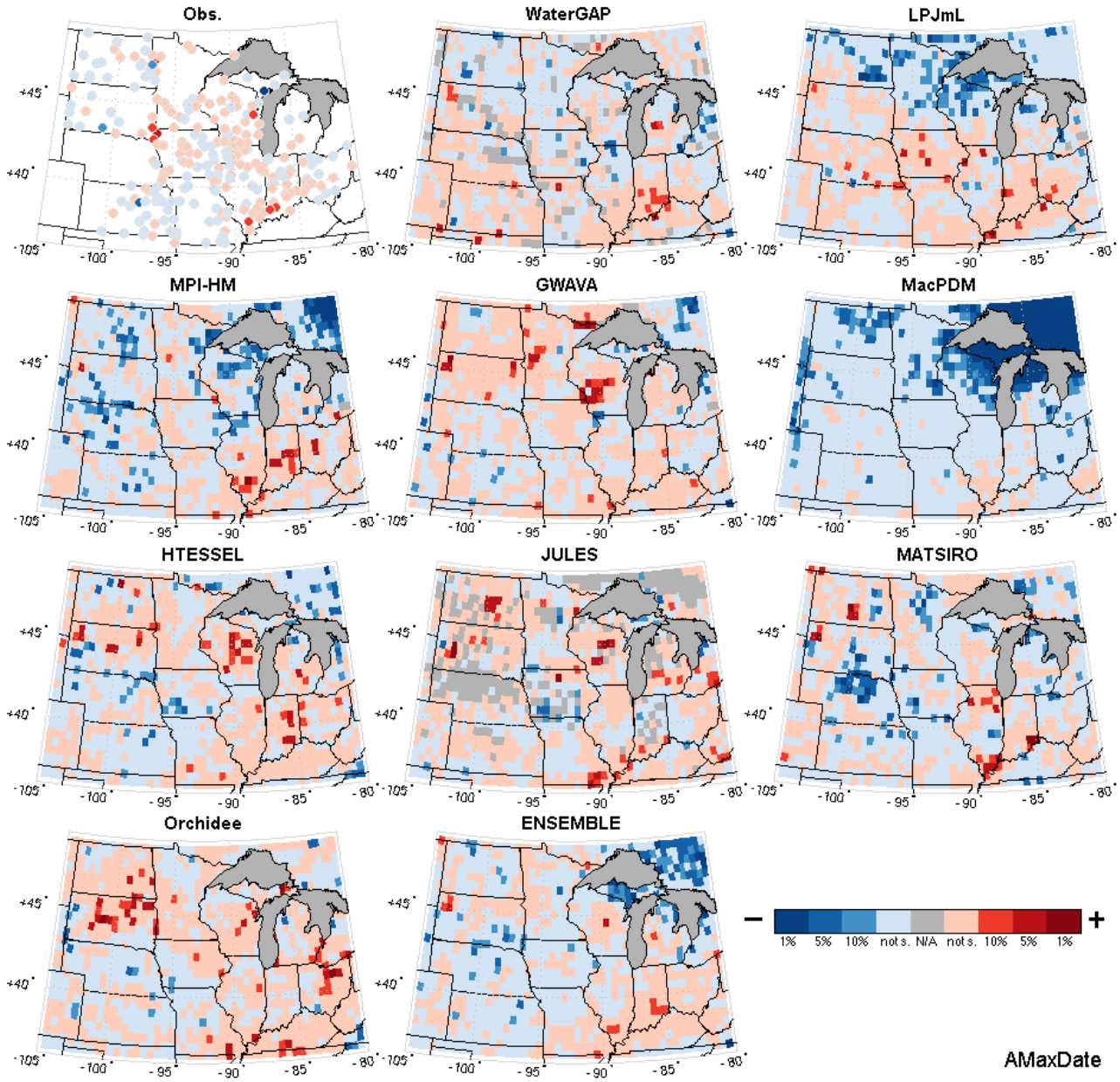


Figure 5.4 – Same as Figure 5.1 for annual maximum flow date (positive trends indicate events occurring later, negative trends earlier).

Results for the AMaxDate index would indicate an earlier occurrence of annual peaks, potentially linked to an earlier melting of the snowpack. While this finding would be consistent with increasing temperatures (e.g., Villarini et al. (2013)), it is not picked up in observational records at the 0.1 significance level. The medium flow date (Figure 5.5) shows very few trend detections for both the observed and the GIMs (including the ensemble median). A few models show areas with

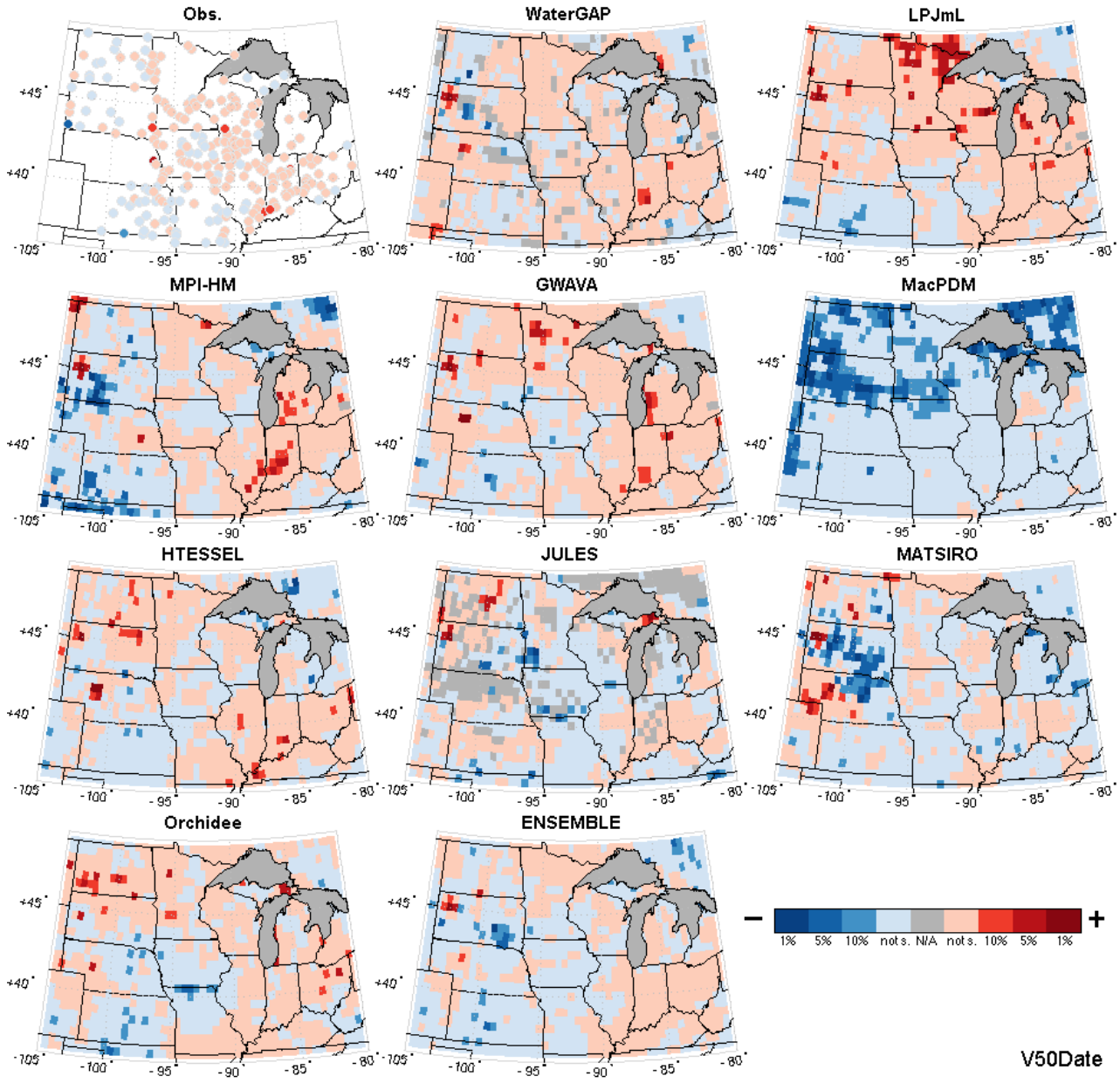


Figure 5.5 – Same as Figure 5.1 for annual medium flow date (positive trends indicate events occurring later, negative trends earlier).

decreasing trends – as seen for the maximum flow date – especially in the north (MacPDM) and to a lesser degree in the west (MATSIRO, MPI-HM); while LPJmL shows an increasing trend in the north. Except for the marked decreasing pattern of MacPDM, the few hotspots seen in the other models are small and point to scarce detections and no clear overall pattern.

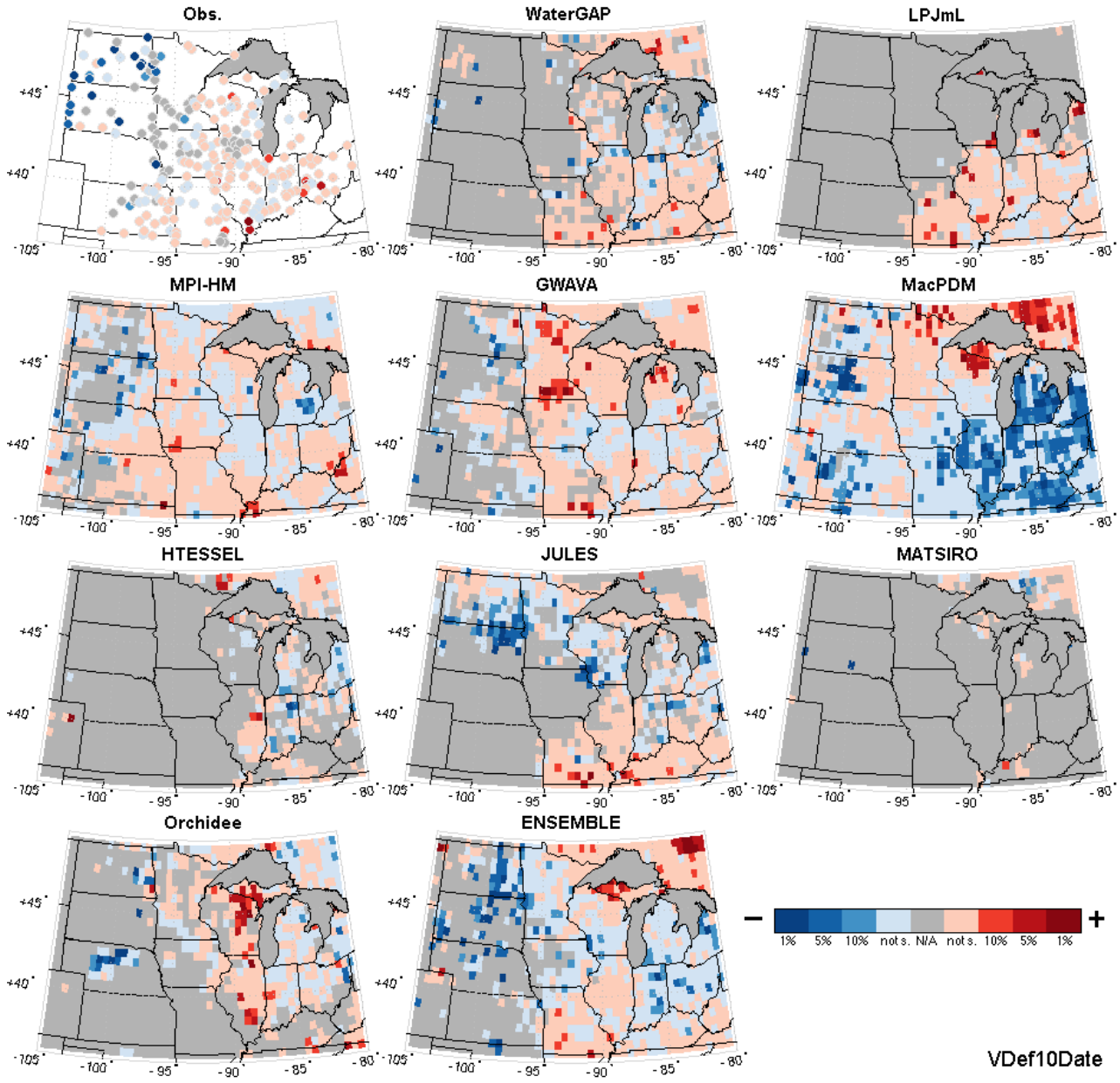


Figure 5.6 – Same as Figure 5.1 for drought start, namely the annual volume deficit 10% date (positive trends indicate events occurring later, negative trends earlier).

The drought start for observed data (expressed as volume deficit date; Vdef10Date) shows a few decreasing trends in the northwest (mostly North and South Dakota) and very few increasing trends in the south-eastern part of the domain. This would hint at an earlier onset of the drought start in the northwest. The masking applied to the GIMs depends on whether the grid cells had sufficient

non-zero values in the index (<25). In spite of the considerable masking, most models seem to match the weak pattern in the trends detected on the observed data, though MacPDM shows marked decreases in the southeast and an increase in the northeast. Finally, JULES and the Ensemble median seem to capture well the light decreasing pattern present in the observations in the northwest part of the study region.

5.4.2 The 1988 drought and the 1993 flood

All GIMs and the Ensemble, show good agreement with the observed data in capturing both the 1988 drought (Figure 5.7) and the 1993 flood (Figure 5.8). While the pattern is more evenly distributed for the 1988 drought, the 1993 flood appears intensified with a patch spanning from the southwest (Kansas) to the north east (Wisconsin) of the domain. The intensity of the variations is different for the two events, CVs vary mostly between 0 and -2 for the 1988 drought, and between 0 and 5 for the 1993 flood. This indicates that the 1993 summer flood volumes have a more pronounced departure from the whole period's summer volumes than the 1988 summer drought does. This is to be expected and can be explained by the more erratic nature of the flood runoff volumes compared to slower onset and development of the drought ones (whose values, differently from the flood, are bound to zero). The good performance in capturing these two events is confirmed by the exceedance probability maps (Figure A3.1 and Figure A3.2 in APPENDIX III), where, as expected, low probabilities result for the 1993 mean summer runoff and vice versa for the 1988. While all GIMs tend to capture the mean runoff differences with similar intensity and spatial pattern, MacPDM appears to capture the spatial pattern equally well, but with a weaker intensity with regards to the 1993 flood (Figure 5.8).

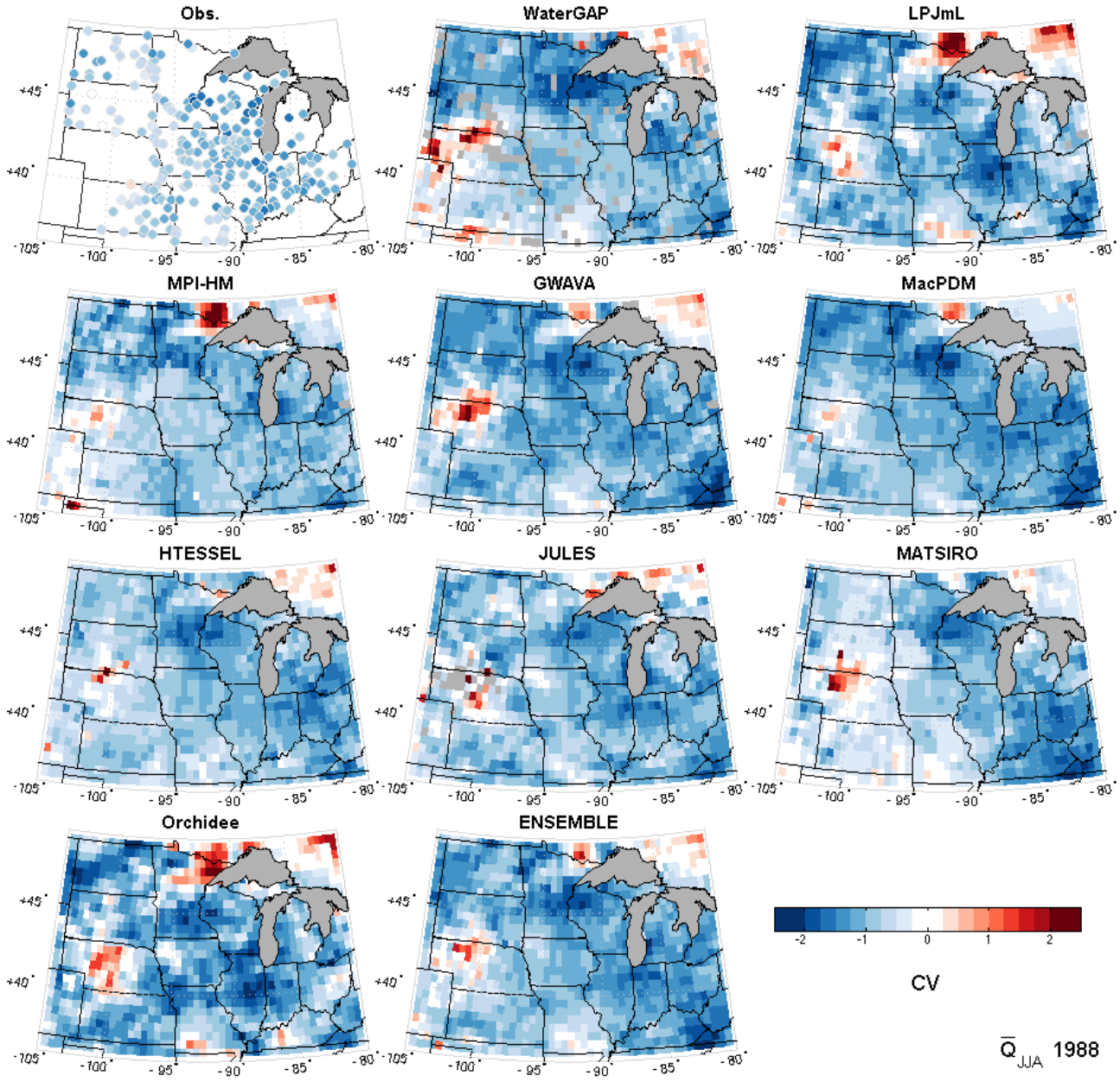


Figure 5.7 – 1988 drought Coefficient of Variation for observed data (top left), the nine GIMs and their ensemble median. Negative CVs are shown in blue and positive CVs in red (negative CVs indicate 1988 summer mean runoff smaller than mean 1963-2001 summer mean runoff).

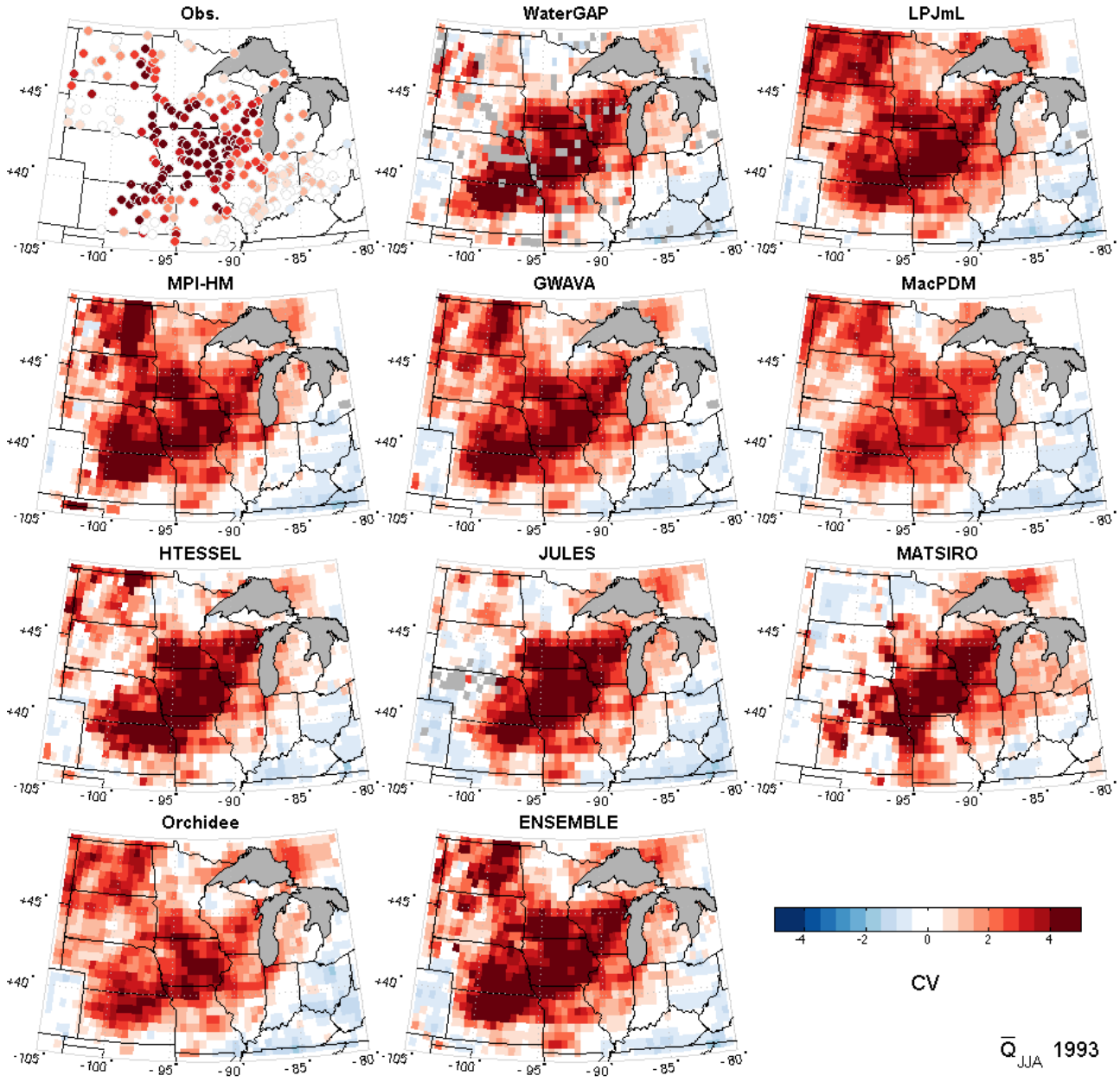


Figure 5.8 – Same as Figure 5.7 for 1993 flood (positive CVs indicate 1993 summer mean runoff larger than mean 1963-2001 summer mean runoff).

5.4.3 Modelled – Observed pairwise comparison

After considering the whole domain for the examination of trends of magnitude and timing indices, and the consideration of two particularly extreme events, we focus on a subset of stations to

examine whether the models are able to capture the seasonality in these quantities. More specifically, we focus on 128 grid cells selected to correspond to 128 streamflow gauges.

5.4.3.1 Timing of annual high, medium and low flows

For the entire region, the monthly frequency of occurrence of annual maxima, annual medium flows, and annual drought start are shown as boxplots in Figure 5.9, Figure 5.10, and Figure 5.11, while Figure 5.12 quantifies the differences in the median and the interquartile range of the models from the observations. It is worth clarifying that the boxplots summarize the results grouping outcomes from different regions and on a limited number of grid cells (128 of 1350). The observed annual maxima (Figure 5.9) occur mostly from March to June, with the highest frequency in April. This pattern is reproduced by the GIMs, but specific behaviours emerge depending on the nature of the model (LSMs versus GHMs). The GHMs tend to show a seasonality characterized by an enhanced frequency of occurrence of annual maxima about 1-2 months earlier than the observations, with medians that are closer overall to observed data. The LSMs, on the other hand, tend to exhibit a delayed seasonality (1-2 months later) and to show an overall greater discrepancy from observations. This pattern is shown very clearly in Figure 5.12 (top panels), where the GHMs (WaterGAP, LPJmL, MPI-HM) tend to overestimate count rates in AMaxDate occurrences from December to March, and to underestimate them from April to September. Opposite to this pattern, the LSMs (JULES, MATSIRO, Orchidee) tend to underestimate count rates from February to April and to overestimate them from June to September. The spread (quantified in terms of interquartile range IQR) of the LSMs is higher when there is an overestimation of the count rates and lower in the case of underestimation, whereas the spread of the GHMs is generally closer to the observational one throughout the year. Between these two marked behaviours lay GWAVA, MacPDM (GHMs), and HTESSEL (LSM), which show the smallest differences from the

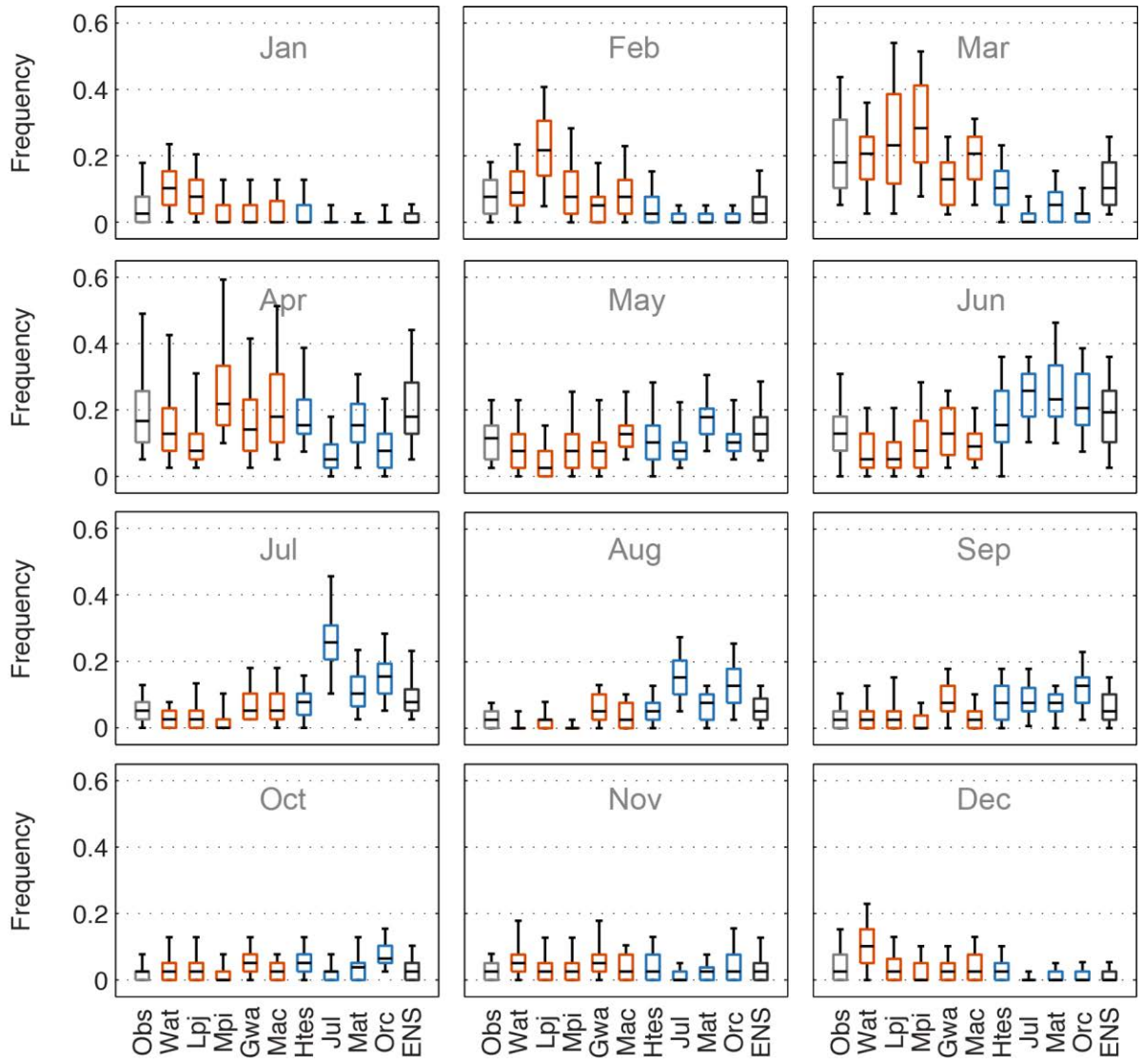


Figure 5.9 – Frequency of occurrence of annual maximum flow per month for 128 gauges and coresponding grid cells (bar: median, box: interquartile range, whiskers: 10th and 90th percentiles). In light grey the observed records, in orange the GHMs, in blue the LSMs, in dark grey the ensemble median.

observations both in the median and IQR. The observed data indicate that the V50Dates occur from March to June with the highest counts in June (Figure 5.10). Few or no events are counted from September through February, and this is captured unanimously by all the models. For March to August, GHMs tend to capture better the timing of the medium flows than the LSMs, although there are some discrepancies among these models. More specifically, WaterGAP, GWAVA and MacPDM underestimate the count rates in V50Date occurrences during March and April, while

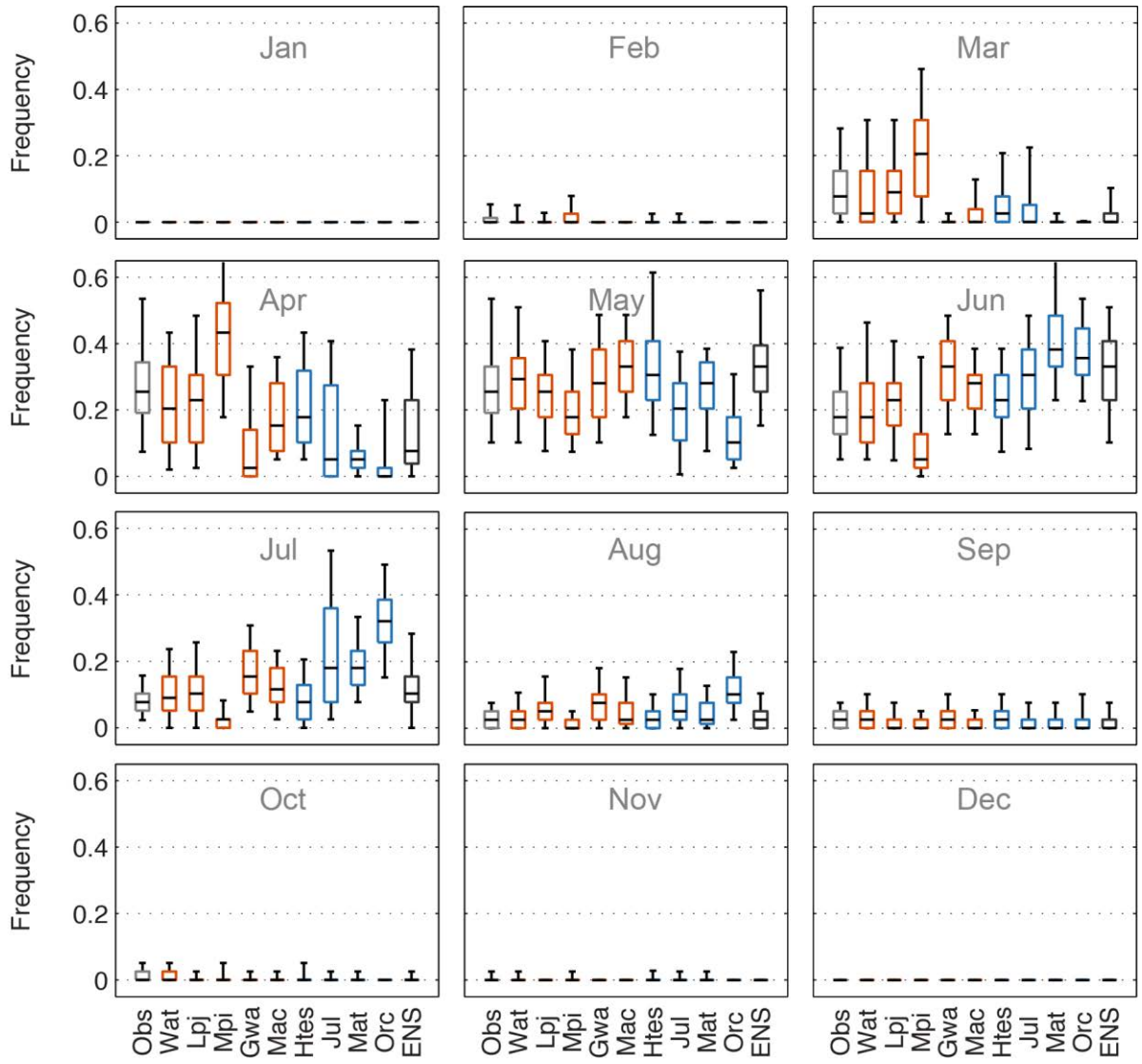


Figure 5.10 – Same as Figure 5.9 for annual medium flow.

GWAVA and MacPDM also overestimate from May to August; LPJmL and MPI-HM underestimate them in late spring. With the exception of HTESSEL, which captures rather well the timing throughout the entire year, for the LSMs there is a marked underestimation during the spring (March to May) and an overestimation in the summer (June to August). The LSMs are strikingly not in line with the observations, and they appear to be out of phase with a lag of 1-2 months.

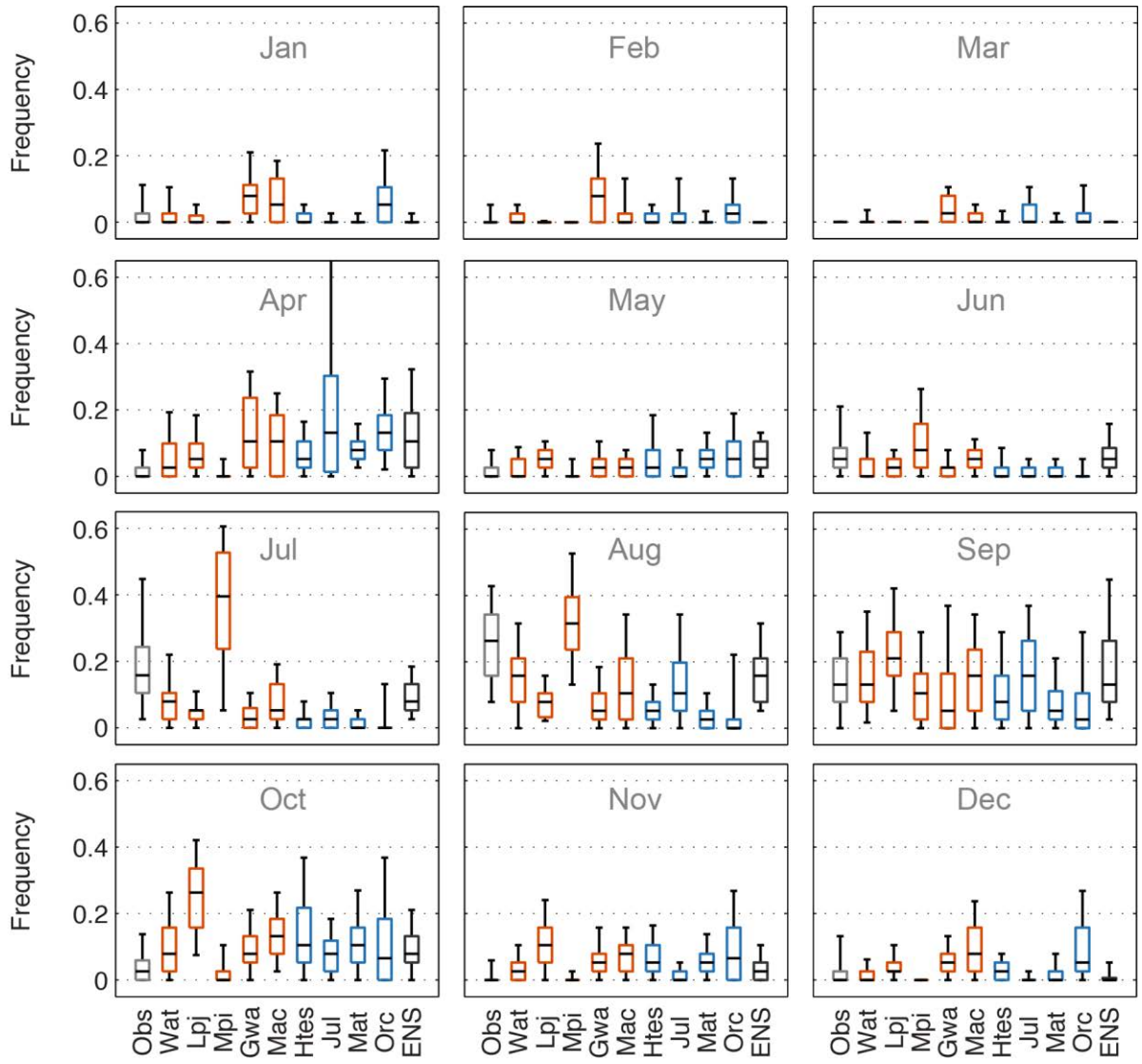


Figure 5.11 – Same as Figure 5.9 for annual drought start.

Figure 5.12 (middle panels), shows this phase shift for which the largest differences in the median and the IQR appear for the LSMs JULES, MATSIRO, and Orchidee and to a smaller extent for the GHMs MPI-HM and GWAVA. The drought starts (Vdef10date) in observed data show few occurrences in the spring (April-May) and an increasing frequency in the summer, peaking in August and decreasing in early fall (September-October), with virtually no occurrences in the winter from November to March (Figure 5.11). The GIMs ability to reproduce ground observations is weak, highlighting the difficulty in capturing the timing of low flows with respect to high and

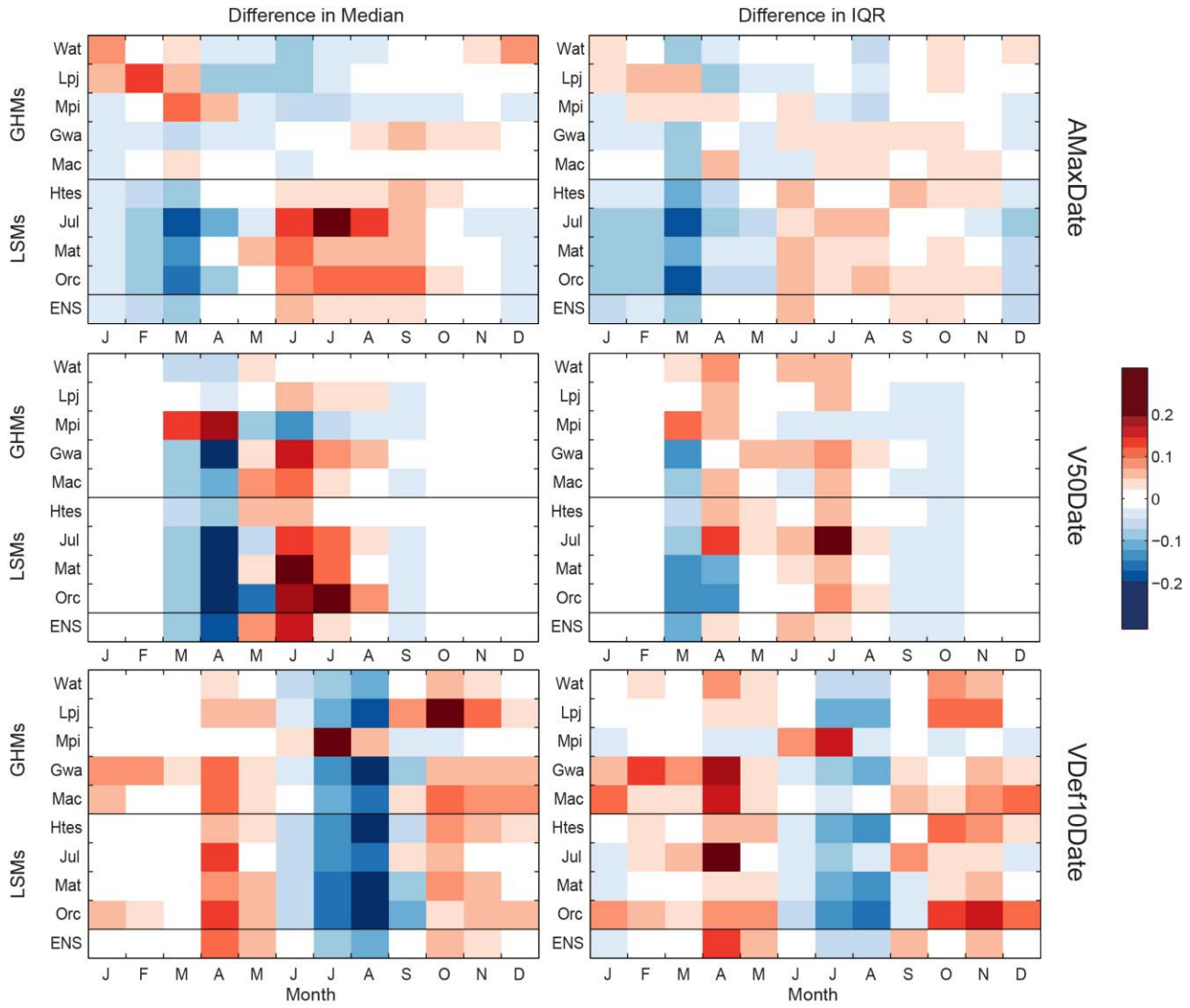


Figure 5.12 – Occurrence of annual maximum (top panels), annual medium flow (middle panels), and annual drought start (bottom panels) events per month (as seen in Figure 5.9-Figure 5.11): difference in median (left) and the interquartile range IQR (right) of the models from the observations – red, overestimation; blue, underestimation.

medium flows. For instance, for the two previous indices (Figure 5.9; Figure 5.10), months with no occurrences were broadly well reproduced by the majority of the GIMs, while for the drought start some GIMs show considerable frequencies, especially in the winter as opposed to the frequencies of the observations that are near zero; there is also a less pronounced homogeneous response per type of GIM seen thus far. With the exception of MPI-HM, which seems to follow the most closely the

observed results, all of the other GIMs show noticeable fewer counts in the summer when counts are high. The situation changes in September, when GIMs' counts increase and tend to decrease in the fall at a much slower rate than the observed data. This lag seems to indicate that GIMs tend to capture the drought onset later in the year, with approximately a 1-2 month delay. In addition, there are higher frequencies in winter and spring. This is visible in Figure 5.12 (lower panels), where there is clear marked underestimation of the drought start in the summer (especially July and August) and an overestimation in spring and fall.

5.4.3.2 An assessment of the GIMs' performance

Figure 5.13 summarizes the results of the performance achieved by the GIMs in the pairwise comparison for the hydrological indices from the streamflow gauges and from the corresponding grid cell. The first index, AMax, depicts a performance that is fairly homogeneous across the GIMs. The main differences are for the Pearson coefficient (R), according to which GHMs perform slightly better than the LSMs. For the annual median discharge performances improve in all metrics compared to the AMax: the GIMs' correlation to the observed data improves noticeably, with R values closer to 1; the $\Delta\sigma$ are closer to zero and their spreads decrease; the RMSE values show that GIMs are closer to the observations. The other end of the hydrological regime, the annual minima, seems to perform better in the RMSE and R correlation than the annual maximum, but results within models in the $\Delta\sigma$ can differ considerably in the spread. The results for the annual median discharge have less pronounced variability in $\Delta\sigma$. This can be due to the description of the central part of the hydrological regime, as opposed to intrinsically more erratic nature at the tails (AMax, AMin). Similarly, lower values of correlation (R) of AMax compared to Amin may be partly owed to AMax's more erratic behaviour, while AMin is bounded below at zero. It should be noted that

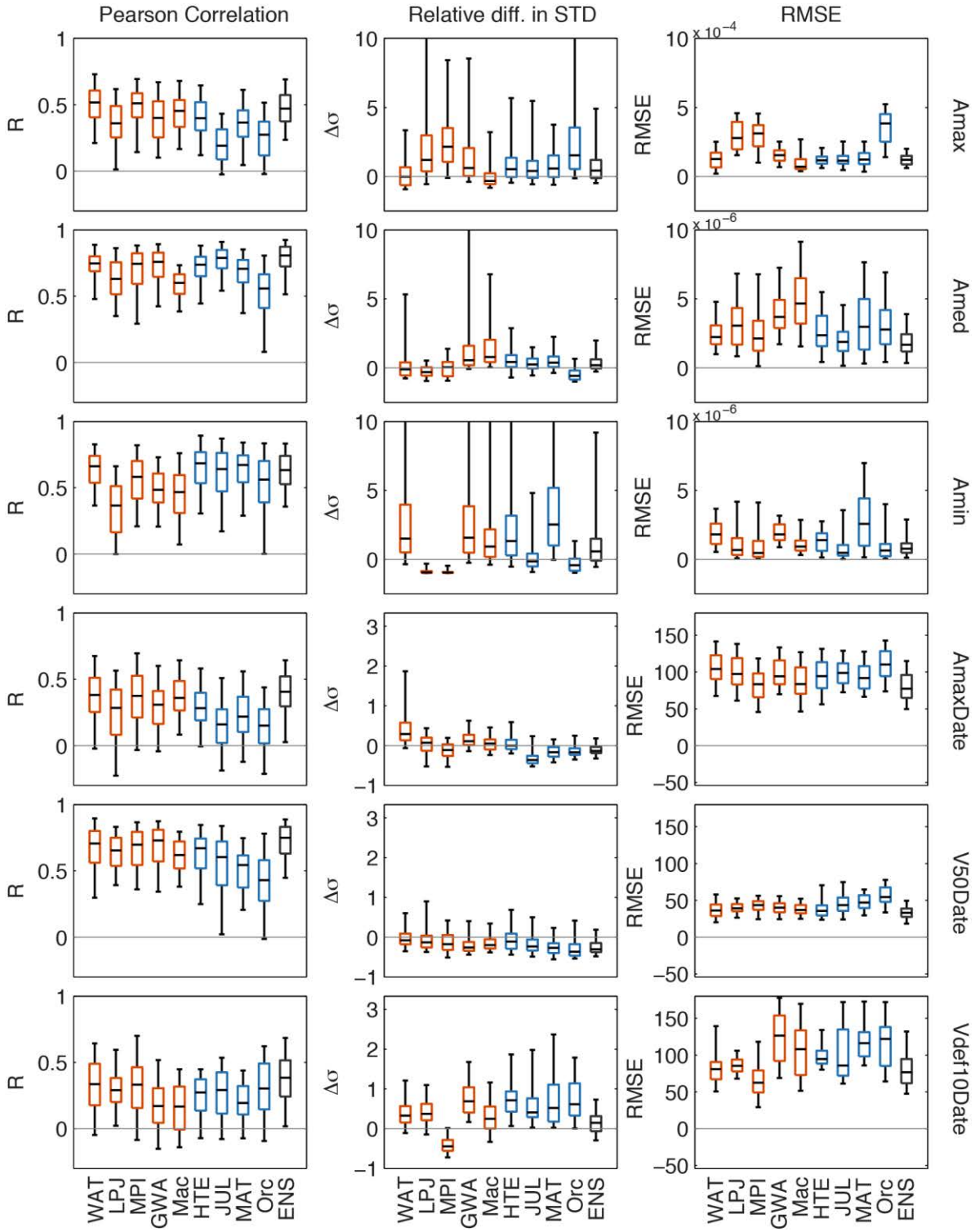


Figure 5.13 – Performance metrics (in column: Pearson R correlation coefficient, relative difference in standard deviation Δσ, RMSE) on the pairwise comparison observed-modelled (128 points) for the six hydrological indices (in row). For the boxplots: bar, median; box, interquartile range; whiskers, 10th and 90th percentiles. Note that the vertical scales are different for Δσ (middle column) and RMSE (right column).

the index series comparison modelled-to-observed is based on approximately 39 points (1963-2001) for high and median flows, and on 38 (1963-2000 as the hydrological year starts in April) for low flows. Also, while computing the metrics, a year with a missing value found in one of the two series is excluded from both series.

The following three rows in Figure 5.13 describe the timing of high, medium and low flows indices expressed in number of days from the beginning of the hydrological year. In general, similar to the previous three indices, the Ensemble median seems to outperform individual GIMs, and the medium flow (V50Date) is the index that is the closest to the observations. Focusing on the correlation coefficient, the second best index is the annual maximum flow (AMaxDate) with the GHMs performing better than the LSMs, followed by the annual drought start (Vdef10Date). Similar to V50Date, the $\Delta\sigma$ nears zero for most of the GIMs for AMaxDate. This is not true for the Vdef10Date, which have higher values and larger spreads. The RMSE stays below 50 for V50Date, and around 100 for AMaxDate and Vdef10Date, though the latter shows stronger variations from GIM to GIM, including in the spread. It should be noted that results for Vdef10Date tend to include fewer than 128 pairs because the presence of zeros in the index series (the threshold was not always crossed) affecting the pairwise comparison: series with less than 25 values different from zeros were excluded. The GIMs using fewer pairs are LPJmL and MATSIRO (47 pairs), followed by HTESSEL (71), Orchidee (78), with the remainder of the GIMs having between 109 and 124 pairs.

5.5 Discussion and conclusions

The aim of this paper was to assess how well the regional hydrology of the central United States (based on observations at 252 reference gauges from 1963-2001) was reproduced by a set of nine global impact models from the WaterMIP Project and their ensemble medians. The focus was

on the examination of a number of discharge indices related to high, medium and low flows, as well as the seasonality and timing of the flow regime.

In our model-observation comparison, there are few elements that we need to keep in mind when interpreting these results. The spatial resolutions of the models and observed records used as reference are not the same. The models do share a historical forcing (the Watch Forcing Data: WFD) that has been provided globally and whose quality can vary depending on the region, as noted in Section 3.3.2.1. However, our study region lacks high elevation features, which typically have a negative effect on the quality of the forcing, and, more importantly, the scale at which we operate for the trend detection is sufficiently large to allow for a comprehensive comparison of the patterns, while for the pairwise comparison analysis, the observed dataset is reduced using only catchment of comparable size with the grid cell.

To date, the WaterMIP GIMs have been used in other studies (e.g., Gudmundsson et al. (2012b); Prudhomme et al. (2011); Stahl et al. (2012)) comparing their control period with observed data over parts of Europe. A general conclusion was that the models tend to capture the interannual variability of high, medium and low flows well. All of these studies show that simulated runoff can vary substantially depending on the GIM, as every model has different characteristics in the way it simulates the different components of the water cycle. The type of flow (high or low) also plays a role: Gudmundsson et al. (2012a) show that for low runoff percentiles the performance of the models decreases, reflecting the uncertainty associated with the representation of the hydrological processes (e.g. the depletion of soil moisture storage). The same authors confirm the results by Haddeland et al. (2011) on MATSIRO's propensity to predict less seasonal variation in runoff than the other models. This is owed to a deep groundwater reservoir that buffers the timing of runoff, in turn leading to an underestimation of the magnitudes and to delays of the high flows peaks. Moreover, Prudhomme et al. (2011) focused on three WaterMIP GIMs and showed that WaterGAP

is the model that best reproduces the regional characteristics of high and low flow events in Europe, while JULES and MPI-HM tend to have a slow and fast responding runoff respectively. Tallaksen and Stahl (2014) focused on droughts (using seven WaterMIP GIMs) and also suggested that WaterGAP and GWAVA are better at capturing hydrological droughts over Europe.

The findings outlined above are generally consistent with our study: most of these GIMs are able to reproduce the spatial trends in the observational records over the central United States. However, a new element in our results is the clear dichotomy between LSMs and GHMs, which is reflected in the ability of each model to capture the timing of maximum and medium flows. The LSMs are less capable of capturing the timing exhibited in the observed data than the GHMs. For the annual maximum flow, GHMs tend to overestimate frequencies in the winter and to underestimate them during spring and summer, while the opposite is true and more marked for the LSMs. For medium flow, a strong underestimation of the frequencies is shown for the LSMs in the spring and an overestimation in the summer, while the GHMs are closer to the observations and show a less marked behaviour in general. Though less marked, indications of similar behaviour can be found in the works by Haddeland et al. (2011) and Gudmundsson et al. (2012b). Over basins with a climate comparable to our study region (i.e., Northern Europe), Haddeland et al. (2011) showed that peaks occur earlier for GHMs than LSMs and linked this behaviour to the snow scheme employed: the energy balance approach used by LSMs predicts reduced snow water equivalent (SWE) values, leading to lower winter and spring runoff volumes than predicted by the degree-day approach used by GHMs. The snowy winters in the northern part of the central United States may explain the clear shift in the timing of high and medium flows yielded by GHMs and LSMs. It should be noted that, as shown in Table 3.2, energy balance models (LSMs) comprise more forcing variables than degree-day models (GHMs), and are thus prone to additional associated errors.

The timing of low flows depicts a less marked behaviour in terms of the type of GIM as seen for medium and high flows and also a poorer ability in capturing the frequencies of occurrence. In particular, GIMs' counts of drought start occur sporadically during seasons for which the observations display no counts. More importantly, during the summer, when observed data frequencies are high, GIMs tend to a generalized underestimation of the occurrences, and to an overestimation in the fall, when the results based on observations tend to decrease while the GIMs continue to have fairly higher rates. It is worth noting that the identification of drought start can be cumbersome when dealing with zero/very small values rich time series, by which some GIMs (e.g. LPJmL, MATSIRO, Orchidee) are particularly affected (and to a lesser extent some streamflow gauges in part of the study domain). The problem is present even when choosing large thresholds quantiles, because for those grid cells/gauges whose runoff tends to plateau over most of the year and have an isolated very large peak, the threshold crossing may not occur every year (i.e. metric is not computed). This underlines the aforementioned increased difficulty of the GIMs to describe the lower tail of the runoff spectrum and the interest for future research in considering alternative low flow timing approaches (e.g., van Huijgevoort et al. (2012)).

The 1988 drought and the 1993 flood events were overall well captured by all the GIMs, with runoff variations compared to the observed data of comparable spatial pattern and intensity. This result provides insightful confidence on the capability of these models to simulate single specific multi-month events on both ends of the runoff spectrum.

To complement our evaluation of the GIMS, we carried out an in-depth pairwise comparison between observations and model outputs using a subset of streamflow gauges and corresponding grid cells. The GIMs' performance was assessed on all hydrological indices through a number of performance metrics. Results from this assessment indicate a better performance of the GIMs in describing the medium flow and its timing compared to the annual maximum and minimum flows.

This could be expected as it reflects the increased difficulty of the GIMs in describing extreme events whose occurrence is more erratic (especially high flows) and whose onset is harder to capture (especially low flows) considering the uncertainties that are cascaded across the different model components, and the limited knowledge of the world. In general the ensemble median proved to perform better and to be more stable than any of the GIMs individually as seen in other previous studies. This is consistent with Stahl et al. (2012) who used the same dataset over Europe. They found both a better performance of the ensemble mean over each GIM, and a decreasing agreement between observed and modelled trends as they moved from annual mean runoff to the tails of the distribution. They also found the widest spread among models for low flows trends, in the same way the performance metrics of our low flow indices were more variable than medium and high flows. Tallaksen and Stahl (2014) also revealed considerable model dispersion in simulating temporal and spatial persistence of drought. They warned about the importance of validating GIMs specifically for hydrological drought when analysing drought characteristics from a limited number of models. Generally, this is valid for all hydrological studies that involve the use of GIMs: the validation of their performance in either high, medium or low flows is key depending on the flow of interest. However, this is particularly relevant for low flows, because GIMs tend to provide larger uncertainties (i.e. inter-model spread) than the other flow types (high and medium) due to their high sensitivity to model structure and parameterization (Wang et al., 2009).

Multimodel studies like WaterMIP comprise many participating GIMs, each of them developed using different conceptual approaches. This make it difficult to identify the reasons for different model behaviour and more generally to attribute model error. For instance, conducting parameter sensitivity on an ensemble of GIMs is theoretically possible, but unrealistic in practice, as it would require full control over each model. Similarly, the effect of calibration on model output is rarely quantified for large-scale models, which rarely undergo calibration as the traditional

catchment models do. The study by Müller Schmied et al. (2014) provides some insights in this regard as it uses the only WaterMIP calibrated model WaterGAP in different configurations to investigate the sensitivity of simulated freshwater fluxes and storages to five major sources of uncertainty: climate forcing, land cover input, model structure/refinement, human water use and calibration against observed mean river discharge. They find that the largest impacts on freshwater fluxes and water storages came from calibration and model structure (e.g. modelling groundwater depletion), and to a lesser extent to alternative climate forcings, and land cover data, whose effects tend to compensate and cancel each other out. In a study on the MacPDM model Gosling and Arnell (2011) present a sensitivity analysis and report that simulated runoff is more sensitive to the choice of method to calculate PE (having tested Penman-Mointeith and Priestley-Taylor) than to perturbations in soil moisture capacity and field capacity for each specific vegetation type. In particular, they suggest that regional projections from GIMs are likely to be conditional upon the PE method applied, because each method may be more reliable in dry rather than in wet regions. For instance, for much of the United States, the Priestley-Taylor is associated with positive runoff anomalies compared the Penman-Monteith (used in our study), and the situation is reversed for wetter regions. The same authors also report that MacPDM, when running with monthly input data (in our study, however forcing data from WaterMIP is provided at daily time step), produces a negative runoff bias in several regions of the world where day-to-day variability in relative humidity is high, and attribute this bias to difficulties of this GIM in disaggregating monthly relative humidity into daily data.

These results represent a key step toward an improved understanding of the ability of the models to reproduce the hydrologic processes and their temporal changes over the central United States. In particular, this study provides a benchmark for the application of data from intercomparison experiments that make use of this type of GIMs. Building confidence in the

models' ability to capture the overall temporal trends and the timing of the hydrology at the regional scale is of great importance for the climate impact studies that will follow, in light of the large socio-economic impacts of too little or too much water will have over this region in a warmer climate.

6 Uncertainties in projected runoff over the continental United States

6.1 Abstract

Projections of runoff from global multi-model ensembles provide a valuable basis for the estimation of future hydrological extremes. However, projections suffer from uncertainty that originates from different error sources along the modelling chain. Hydrological impact studies have generally partitioned these error sources into global impact (GIMs) and global climate (GCMs) models uncertainties, neglecting other sources, including scenarios and internal variability. Using a set of GIMs driven by GCMs under different representative concentration pathways (RCPs) from the ISI-MIP Project, this study aims to partition the uncertainty of future flows coming from GIMs, GCMs, RCPs, and internal variability over the United States. We focus on annual maximum, median, and minimum runoff, analysed decadally over the 21st century. Results show that GCMs and GIMs are responsible for the largest fraction of uncertainty over most of the study area, followed by internal variability and to a smaller extent RCPs. In order to investigate the influence of the ensemble setup on uncertainty, three ensemble configurations are also studied using fewer GIMs (culling least credible GIMs in low/medium flows representation and GIMs accounting for vegetation and CO_2 dynamics), and excluding intermediate RCPs. Overall, the use of fewer GIMs has a minor impact on uncertainty for low and medium flows (except for the north-eastern U.S.), but a substantial impact for high flows. Regardless of the exclusion of intermediate pathways, RCPs still play a very small role, suggesting that improvement of GCMs and GIMs and more informed ensemble selections a can lead to a reduction of projected uncertainties, rather than a better constraint in the path we may be on.

6.2 Introduction

As highlighted in Chapter 2, runoff projections from multi-model ensemble experiments (e.g., WaterMIP, ISI-MIP) are increasingly used in climate impact studies, but their utility is undermined by the large uncertainties that originate in the different components of the modelling chain. Therefore, uncertainties need to be quantified in order to: i) acknowledge which components provide the largest uncertainties and ii) attempt to reduce the uncertainty (not for e.g., internal variability).

In order to better constrain future projections of runoff, it is important to quantify major sources of uncertainty not only at one extreme (e.g., flood or drought), but across the runoff spectrum. To this aim, this study examines the partition of uncertainty in annual maximum, medium, and minimum flows from the ISI-MIP runoff projections. This multimodel ensemble has been studied globally for low (Giuntoli et al., 2015a; Prudhomme et al., 2014; Schewe et al., 2014), medium (Davie et al., 2013), and high (Dankers et al., 2013; Giuntoli et al., 2015a) flows, comparing runoff metrics from future and past periods. We focus on the continental United States using a fractional change approach to assess how the uncertainty evolves transiently throughout the 21st century (e.g., Hawkins and Sutton 2011; Hingray and Saïd 2014). In particular, we analyse high, medium and low flows jointly, showing how uncertainties differ across indices; moreover, in addition to GCMs and GIMs (as done in the aforementioned studies), we also include the contribution of RCPs and internal variability.

We also investigate whether the ensemble configuration influences the partition of uncertainty by: i) culling models (as discussed in Overland et al. (2011), Thibeault and Seth (2014), and van Huijgevoort et al. (2014) among others) on the basis of credibility in medium/low flows representation and type (biome models, including CO_2 and vegetation dynamics); ii) excluding

intermediate RCPs (4.5 and 6.0). This is done to assess to what extent the uncertainty shares change using all of the available GIMs (e.g., Dankers et al. 2013, Schewe et al. 2014) or a subset (e.g., Prudhomme et al. 2014, Giuntoli et al. 2015a), and fewer RCPs with the aim to suggest a better use of resources making model runs and an improved choice of ensembles for future climate impact studies.

After describing the data and methods in Section 6.3 we examine how uncertainty evolves over time using transient runs focusing on the continental United States and the full ensemble of five GCMs and nine GIMs. In Section 6.4.2 we investigate the effect of culling models on the basis of credibility and type. Finally, the effect of considering fewer RCPs is investigated. Section 6.5 provides a discussion of the results, followed by summary and conclusions.

6.3 Data and Methods

6.3.1 Simulated runoff

We use simulations of daily unrouted runoff from the ISI-MIP project comprising nine GIMs driven by five bias-corrected CMIP5 (fifth Coupled Model Inter-comparison Project; Taylor et al. 2012) GCMs in their control (1971-2005) and future (2006-2099) period, under four RCP scenarios (RCPs 2.6, 4.5, 6.0 and 8.5). All GIMs have a spatial resolution of 0.5°×0.5° degrees (except JULES, whose runs were regridded to 0.5° from 1.25°×1.875°) and vary in the parameterization and in the types of processes they represent (Table 3.1). As noted in Chapter 3 (Section 3.3.1.2), only two of the GIMs (LPJmL and JULES), so-called biome models, represent the effects of CO_2 on stomatal opening. This feature have shown that, with increased CO_2, biome models yield higher runoff increases and decreases compared to the other GIMs. As noted by other authors, the inclusion of a diverse set of GIMs is important to sample the GIM uncertainty range

(Davie et al., 2013; Prudhomme et al., 2014). However, even though biome models are pioneering the inclusion of an important process affecting runoff projections – absent in the other GIMs – they have shown unrealistic representations of runoff over some areas of the globe (Giuntoli et al., 2015a, 2015b; Prudhomme et al., 2014); this is particularly true for LPJmL, which is primarily a dynamic global vegetation model, and was therefore designed to model global vegetation chiefly, rather than surface hydrology. For this reason, we decided to analyse the ensemble uncertainty considering all GIMs as well as culling the ensemble leaving out the GIMs that have shown unrealistic runoff characteristics and the biome GIMs, so to determine the extent of their influence on uncertainty.

6.3.2 Hydrological indices

We aim to partition uncertainty in runoff over the high-, median- and low-flow regime focusing on three hydrological indices for the period 2006-2099 over the conterminous United States (approximately 5700 land cells) – defined as the area bounded in longitude by -135° and -60° E, and in latitude by 54° and 24° N. To this aim, over the period 1971-2099 we extracted the following three indices (introduced in Section 3.4.1.2) from the daily runoff projections:

i) Annual Maximum (AMax)

ii) Annual Median (AMed)

iii) Annual Minimum (AMin)

We use these indices decadally, as for temperature and precipitation in e.g., Hawkins and Sutton (2009) and Pendergrass et al. (2015), thereby reducing the noise in the signal of the ensemble spread and the contribution of internal variability to uncertainty. As discussed in Hawkins and Sutton (2009), internal variability would be a dominant component at the annual scale, with its contribution decreasing for increasing spatial and temporal aggregation scales. Hence, we run a

10-yr moving average on the annual indices obtaining decadal AMax, AMed, AMin series for the period 1975-2095 (later denoted by *X*).

6.3.3 Partition of uncertainty

We followed the statistical framework proposed by Hawkins and Sutton (2009) to quantify four sources of uncertainty: GCMs, GIMs, RCPs, and internal variability (henceforth called IVar). The uncertainty is expressed by the variance of multimodel anomalies averaged for each source (except for IVar). Prior to computing the variance for each source, each decadally averaged projection *X* is fitted with a robust locally weighted regression function - loess (Cleveland, 1979) with a 2nd degree polynomial model and a time window of 20 years, represented by *x*. Next, the temporal mean, *i*, calculated over the common reference period 1976–2005 (estimated from the loess-smoothed series, *x*) is removed. Therefore, every projection *X*, for each GCM *c*, GIM *h*, RCP *s*, and year *t*, can be written as:

$$X_{c,h,s,t} = x_{c,h,s,t} + i_{c,h,s} + \varepsilon_{c,h,s,t} \quad [1]$$

The residuals, ε, resulting from the difference between the decadally averaged and the loessed projections, are used to express the uncertainty from internal variability. IVar is thus estimated independently of RCPs and lead time and is defined as the variance of the residuals from the fits across RCPs and time:

$$IVar = \sum_c \sum_h var_{s,t} (\varepsilon_{c,h,s,t}) \quad [2]$$

In this study, GCMs and GIMs are assumed to be independent and, differently from Hawkins and Sutton (2009), received the same weight (as in Villarini and Vecchi (2012)). The GCM uncertainty, denoted by *C*, is estimated from the variance in the mean of the different GCM prediction fits (five):

$$C(t) = 1/N_s N_h \bullet \sum_s \sum_h var_c (x_{c,h,s,t}) \quad [3]$$

where N_s is the number of RCPs (four), and N_h is the number of GIMs (nine). Similarly, the GIM uncertainty, denoted by *H*, is estimated from the variance in the mean of the different GIM prediction fits (nine):

$$H(t) = 1/N_s N_c \bullet \sum_s \sum_c var_h (x_{c,h,s,t}) \quad [4]$$

where N_c is the number of GCMs, therefore the mean is computed over 20 values (four RCPs × five GCMs). The RCP uncertainty, denoted by *S*, is the variance of the multimodel means for the four RCP scenarios:

$$S(t) = var_s (\sum_c \sum_h x_{c,h,s,t}) \quad [5]$$

Finally, the total variance, *T* (representing the total uncertainty) at time *t* can be written as:

$$T(t) = C(t) + H(t) + S(t) + IVar \quad [6]$$

To represent how the different sources contribute to the total uncertainty, we map the fraction of total variance corresponding to each source.

We express transient uncertainty over the entire domain on a grid-cell basis, as well as over regions characterized by aggregating grid cells into nine climate homogeneous areas (http://www.ncdc.noaa.gov/monitoring-references/maps/us-climate-regions.php).

For a selected grid cell in the northeast, Figure 6.1 shows the overall spread of the transient runs from 2006 to 2095 and the corresponding fractional variance (uncertainty): AMax shows a large share of uncertainty from GCMs at the beginning of the run that decreases rapidly in the first decade and continues to decrease until the end of the run in favour of the GIMs; on the other hand, when focusing on AMin, after a balanced start with GIMs and GCMs accounting for almost 80% of the uncertainty, the share of GIM rises quickly and explains (around 2020) virtually all of the variance in the projections. Similarly to AMax, AMed depicts a declining GCM contribution, although the RCP uncertainty share is generally larger. For these three indices all throughout the run, RCP and the IVar keep a constant and marginal contribution to total variance with respect to

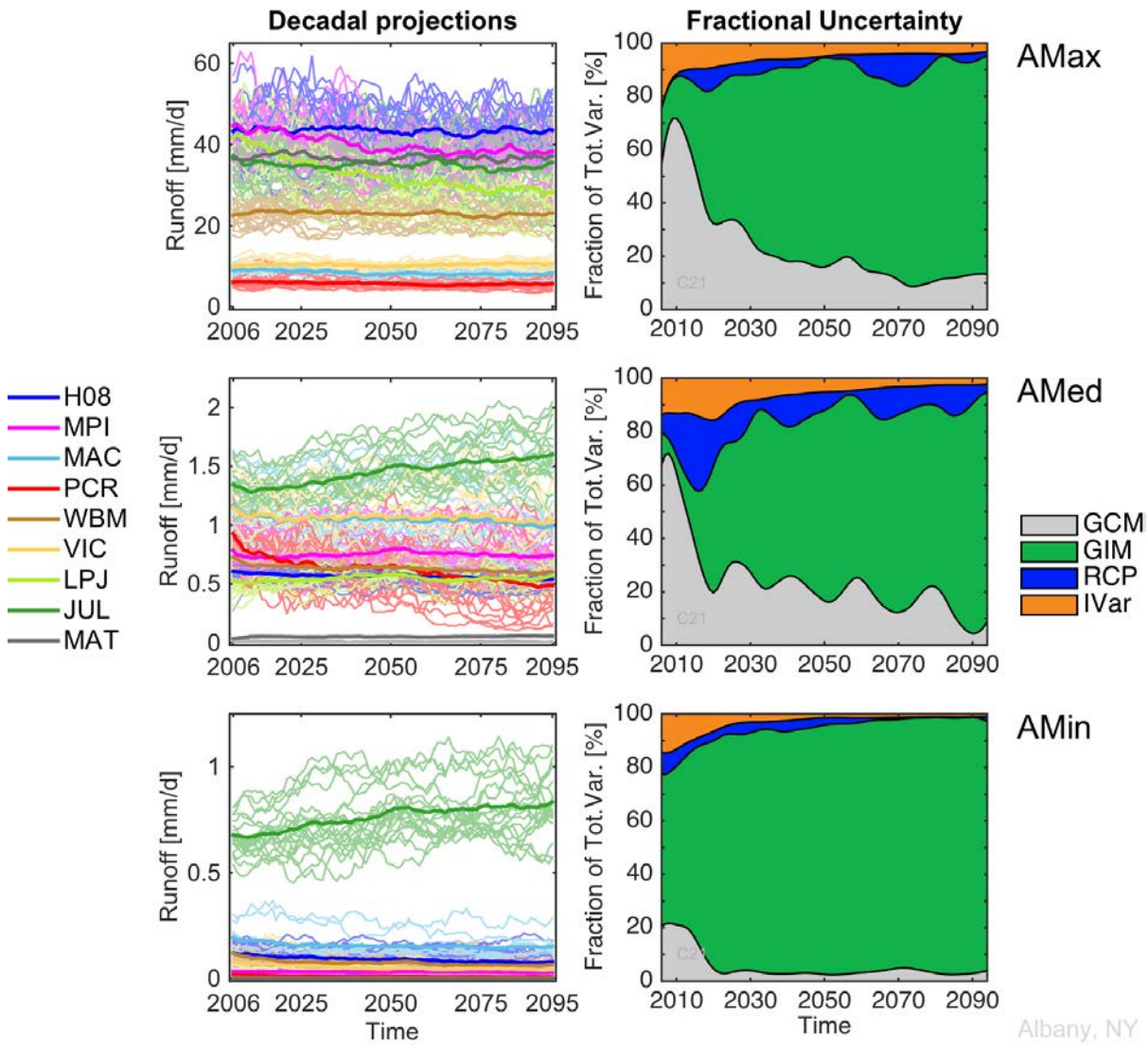


Figure 6.1 – Decadally averaged AMax (top panels), AMed (mid panels), and AMin (lower panels) projections (left) coloured according to GIM (thick lines are averages per GIM) and corresponding fractional uncertainty (right column), for a selected grid cell (42.7° N -73.9° E; Albany, NY).

the GIM and GCM contributions. The differences in the magnitude of the projections across GIMs can be considerable and this explains how they dominate future uncertainty.

It should be noted that the loess window of 20-yr was chosen after testing 10, 20, and 30-yr windows, for which results do not differ greatly. In particular, increasing the year window from 10 to 30 led to a general smoothing of the fluctuations in the fractional uncertainty contributions, with a moderate increase of the IVar contribution (at the expenses of GCM and GIM contributions

primarily) as a result of greater residual values from the stronger smoothing fit on the decadal indices series. Importantly, the overall results for the three windows tested provide similar results across the domain of study (see, e.g., in APPENDIX IV, Figure A4.1, based on the same grid cell of Figure 6.1, and Figure A4.2, for a contrasting grid cell within the study region).

6.3.4 Ensemble configurations

The uncertainty partition was carried out over the entire ensemble and over selected subsets obtained by culling specific GIMs and ultimately intermediate RCPs (Table 6.1). This is done to improve our understanding of the effect of a reduced ensemble on the results. In particular, a multi-model ensemble can be used in its entirety or in part, depending on the availability of the runs or on the focus of the study (e.g. often projections are used under one or two RCPs scenarios rather than all four); therefore it is important to find out to what extent the use of different ensemble

Table 6.1 - The ensemble combinations presented in this study comprise five GCMs (HadGEM2-ES, IPSL-CM5A-LR, MIROC-ESM-CHEM, GFDL-ESM2M, NorESM1-M), four RCPs (2.6, 4.5, 6.0, 8.5), and six to nine GIMs; "(–)" not included.

	GIM	Runs	Ensemble combination			
			oE	clcE	clcbE	clcirE §
	H08	20				
	MPI-HM	20				
	PCRGlobWB	20				
	WBM	20				
Missing RCP4.5 RCP6.0	MacPDM*	12				
	VIC*	12				
	MATSIRO*	12		–	–	–
Biome Models	LPJmL	20		–	–	–
	JULES	20			–	
	Total Runs		156	124	104	70

Low credibility AMed AMin

§ RCPs 4.5 and 6.0 excluded.

* GIMs lacking runs from intermediate RCPs (4.5, 6.0) for all GCMs except HadGEM2-ES.

configurations can impact the uncertainty for each of the indices selected.

6.3.4.1 Ensemble of opportunity – oE

We first examine the uncertainty in the general case of using the ensemble of opportunity, i.e., all runs available (156), henceforth called oE (Table 6.1 – first column). This is the general case that ensures the sampling of the largest spread provided by the ISI-MIP dataset, comprising 156 runs. For instance, all nine GIMs with daily runoff were used in Dankers et al. (2013) to assess the risk of future floods. Results for the oE are presented in Section 6.4.1.

6.3.4.2 Culled low credibility GIMs Ensemble – clcE

In addition to the general case, we have analysed subsets of the oE to investigate the effect of the exclusion of the GIMs showing limitations in simulating realistic low and medium runoff on the uncertainty. Indeed, two GIMs have shown a disproportionate number of days with null runoff (Giuntoli et al., 2015a, 2015b), notably in the western United States (Figure 6.2). We define low credibility GIMs as those with more than 40% of land grid cells in the domain of study with more than 40% of days in the year with null runoff on average, calculated across the five GCMs in the control period 1971-2005. Consequently, MATSIRO and LPJmL are identified as less credible, with 50% and 53% of land grid cells with more than 40% with null runoff days, respectively. JULES is retained with 27% of land with more than 40% of null runoff days on average, with grid cells located mostly in the arid U.S. southwest. The remaining GIMs have negligible percentages of null runoff days as shown in Figure 6.2 (in Figure A4.3, APPENDIX IV, results are shown for Europe for comparison). By analysing the ensemble excluding MATSIRO and LPJmL, named *clcE*, (Section 6.4.2.1), we address the question of what the effect of using low credibility GIMs is on the

uncertainty, compared to the whole ensemble. This GIM configuration has been used for the study of future droughts in Prudhomme et al. (2014).

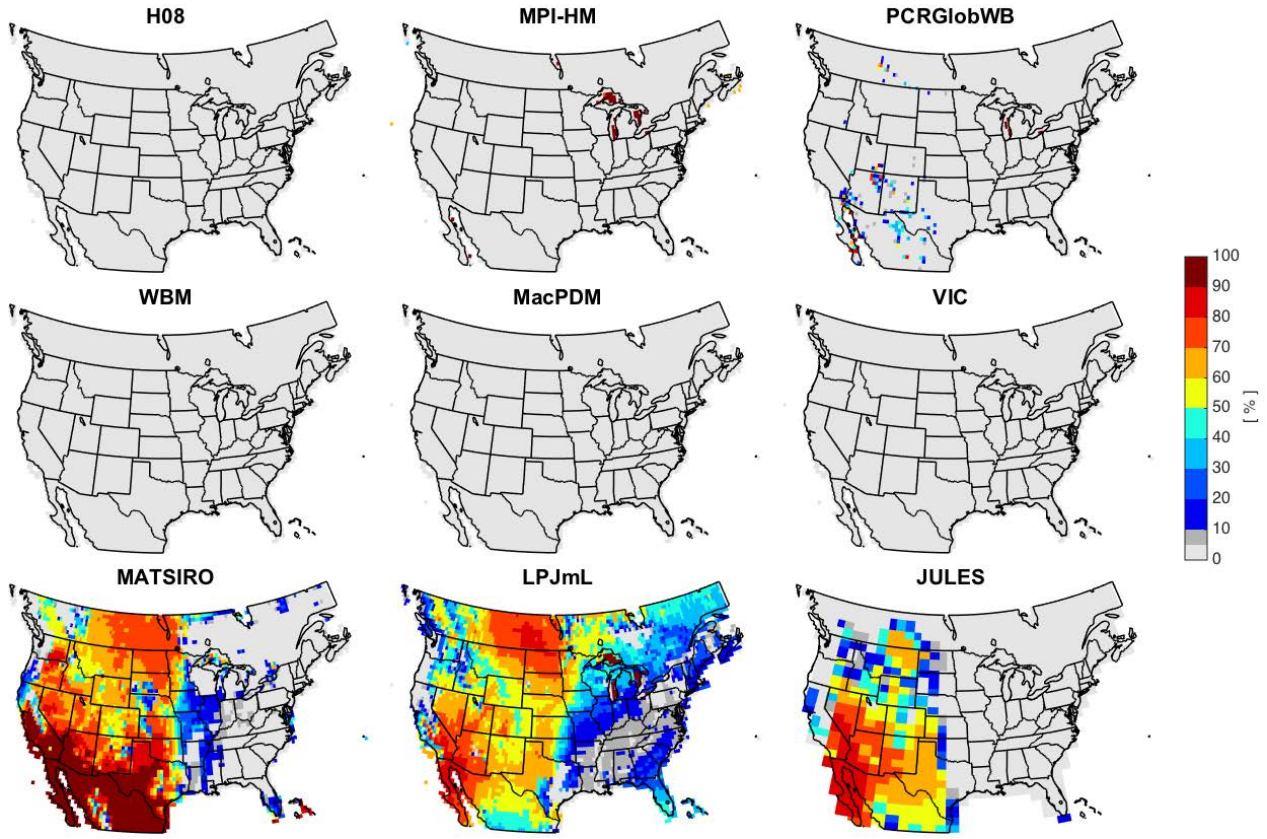


Figure 6.2 - Days in the year [%] with runoff equal to zero per GIM (averaged over the five GCMs) for the historical period (1971-2005).

6.3.4.3 Culled low credibility and biome GIMs Ensemble – clcbE

We also analyse the effect of excluding the so-called biome GIMs (Section 6.4.2.2). These GIMs account for vegetation dynamics and have a behaviour distinct from the other GIMs as noted in the introduction. For instance, this GIM configuration has been used for the study of future floods and droughts by Giuntoli et al. (2015a).

6.3.4.4 Culled low credibility GIMs Ensemble – clcirE

We finally examine the effect of the exclusion of intermediate RCP runs from the clcE (Section 6.4.2.3) to test whether the use of the lower (2.6) and upper (8.5) boundaries alone brings about the same results as using all RCPs.

6.4 Results

6.4.1 Ensemble of opportunity - oE

We present the partition of uncertainty into four sources for decadally averaged transient runs of AMax, AMed, AMin for all available runs of the ensemble. We map the relative contributions to the total uncertainty over the whole domain at five time slices from 2010 to 2090 (Figure 6.3a-Figure 6.4a-Figure 6.5a). Throughout the 21st century, the uncertainty in AMax (Figure 6.3a) is explained mostly by the GCMs and GIMs (first two columns), with the latter increasing their contribution to the total uncertainty from the west of the domain spreading to the northwest (around 2030) and to the upper Midwest and north-eastern United States; this leads to a reduced importance of the GCMs over the 21st century, especially for the northern half of the domain. The share of IVar is large at the beginning of the century over the domain but declines steadily with the exception of the south and southeast, with a contribution to the total variance (15-30%) comparable to that of the GIMs. In contrast and similar to what highlighted in Figure 6.1, the contribution to the total uncertainty from RCPs is very limited compared to GIMs and GCMs, with shares ranging from 5% to 20% of the total variance. In particular, both RCPs and IVar show the smallest contributions in the northern part of the domain, although RCPs have larger contributions in the east, while IVar in the south (see mean relative contributions per region in Figure A4.4, APPENDIX IV). Figure 6.4a summarizes the results for AMed and depicts a picture similar to that

of AMax. GCMs and GIMs are by far the major contributors, with GIMs increasing their contribution to uncertainty over the domain during the 21st century. This is particularly true in the northern and eastern part of the continental United States, even though regional patterns are different and shares are more balanced between the two sources, without the clear predominance of GIM contribution as seen, for instance, in the western United States for AMax. Also, GCMs show larger contributions to the total variance both at the beginning of the run and, in general, in the western and north-western United States compared to AMax (see Figure A4.5). Interestingly, in the U.S. Southeast the RCP contribution to uncertainty is the largest thus far, with a share of around 20%, making it comparable to that of the GCMs. Conversely, in the U.S. West and Southwest, IVar has high shares from as much as 40% in 2006 to 15-20% in 2095. This is due to the large fluctuations in the projections that make up large deviations from the smooth fit, thereby making the contribution of IVar to uncertainty substantial (as seen for the Las Vegas grid cell in Figure A4.2).

For AMin (Figure 6.5a) the similar contribution to the uncertainty by GCMs and GIMs seen for AMax and AMed only holds at the beginning of the 21st century, in which IVar, especially through the mountainous north-south chain of the Rockies has a substantial contribution to uncertainty. GIMs increase steadily their share throughout the period, leaving the remainder of the contributions to GCMs and IVar. In particular, the GCMs lose their share to the advantage of GIMs in the east, while in the west IVar shows considerable percentages of about 40%, which slowly reduce to 15-20% at the end of the run. The strong role that IVar plays in the southwest (in the area spanning from the lower Rocky Mountains to the arid areas of the south) can partly be attributed to the difficulty of both GCMs and GIMs in simulating runoff in mountainous and arid areas, where indeed AMin time series generally suffer from poor simulations with anomalous erratic departures from zero or very low values. With the exception of the U.S. Southeast around the years 2030-2040 (~20%), RCPs contribute only marginally to the total variance (Figure A4.6).

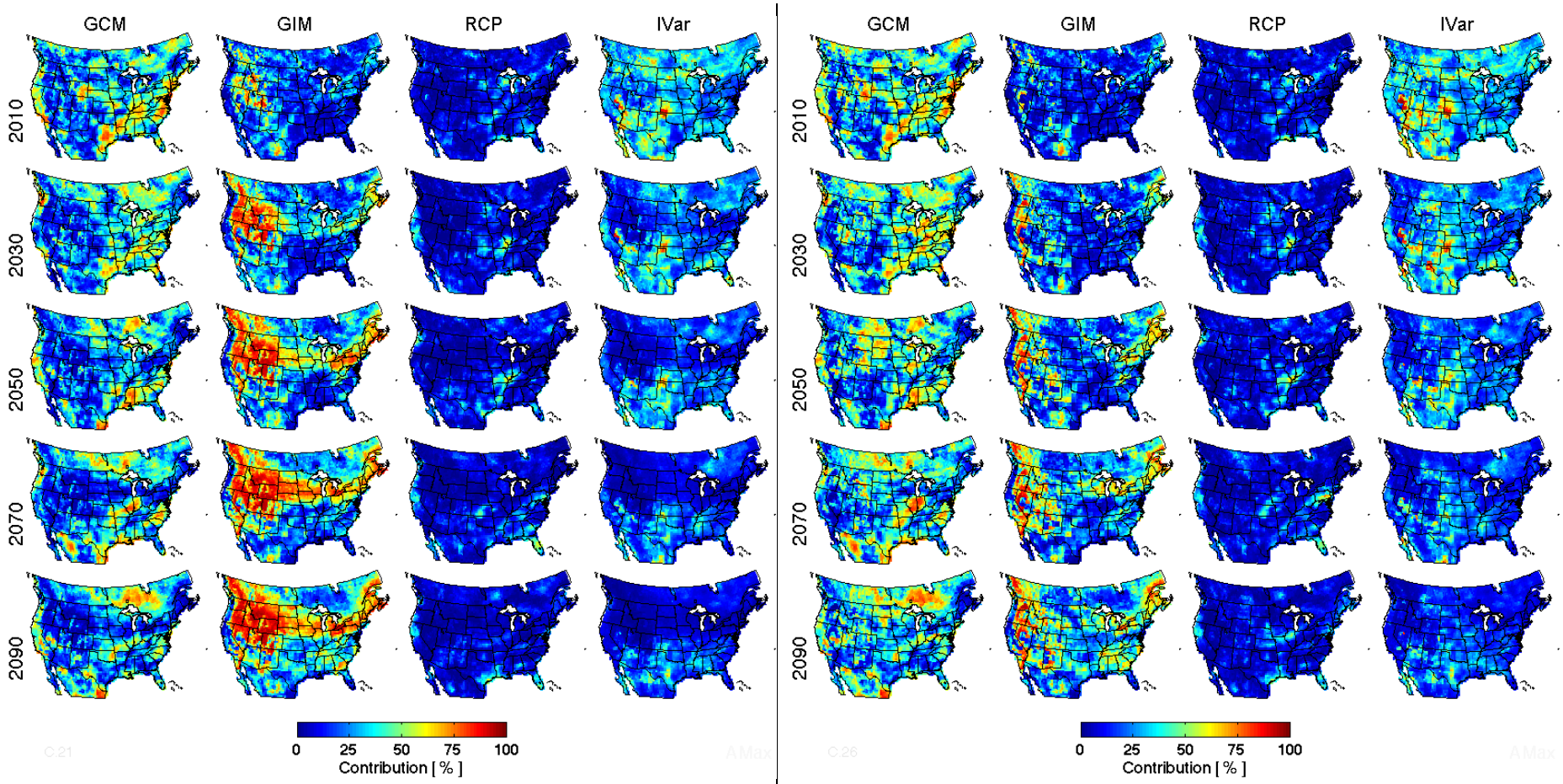


Figure 6.3 - Relative contribution (on 100%) to uncertainty in AMax of (from left to right): GCMs, GIMs, RCPs, and Internal variability. Results are shown for different time slices (in row): 2010, 2030, 2050, 2070, 2090. On the left: a) the whole ensemble (oE); on the right: b) the ensemble excluding low credibility GIMs (clcE); (refer to Table 6.1).

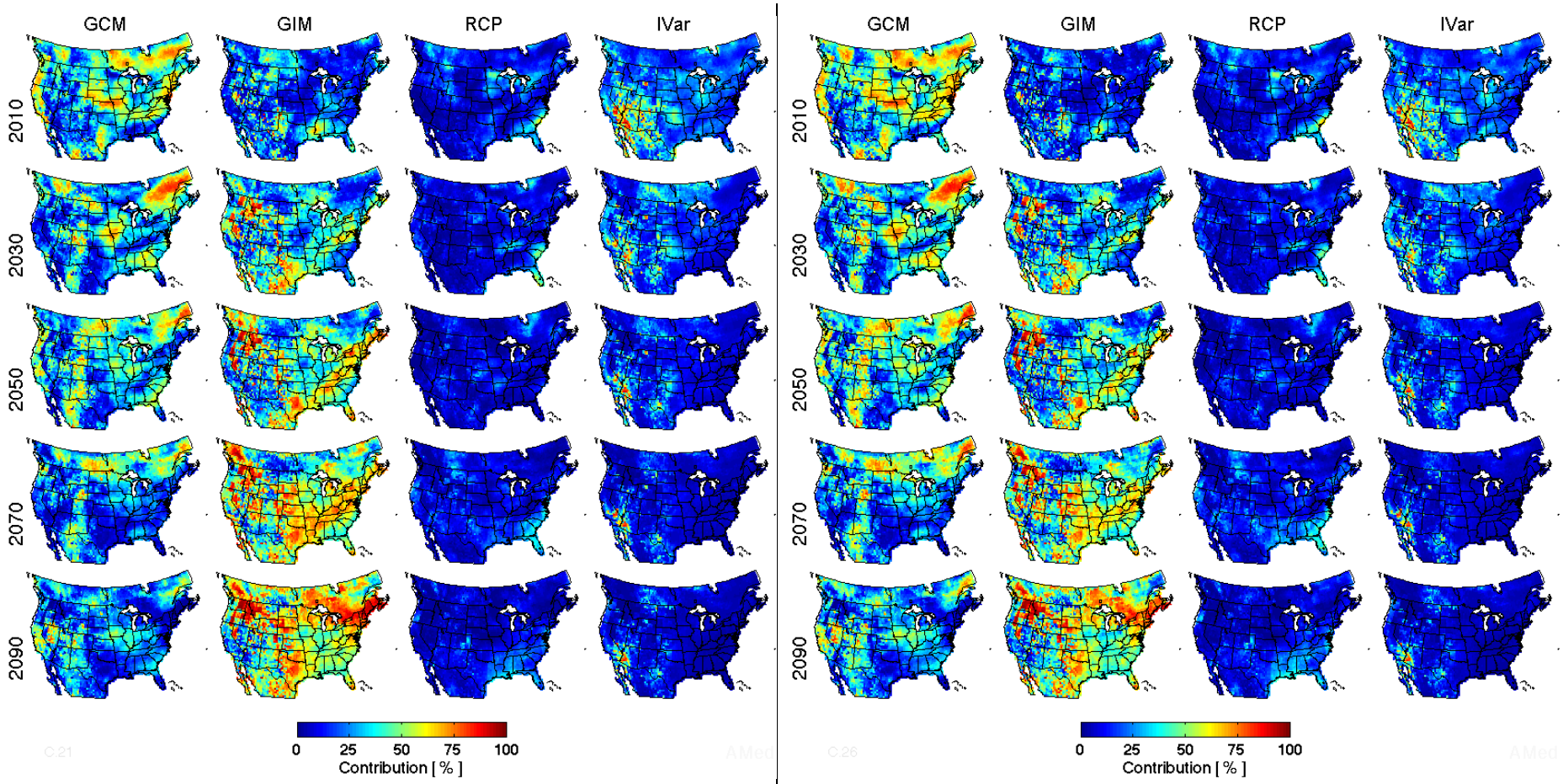


Figure 6.4 – Same as Figure 6.3 but for AMed. Results are shown: a) for the whole ensemble (oE); b) for the ensemble excluding low credibility GIMs (clcE).

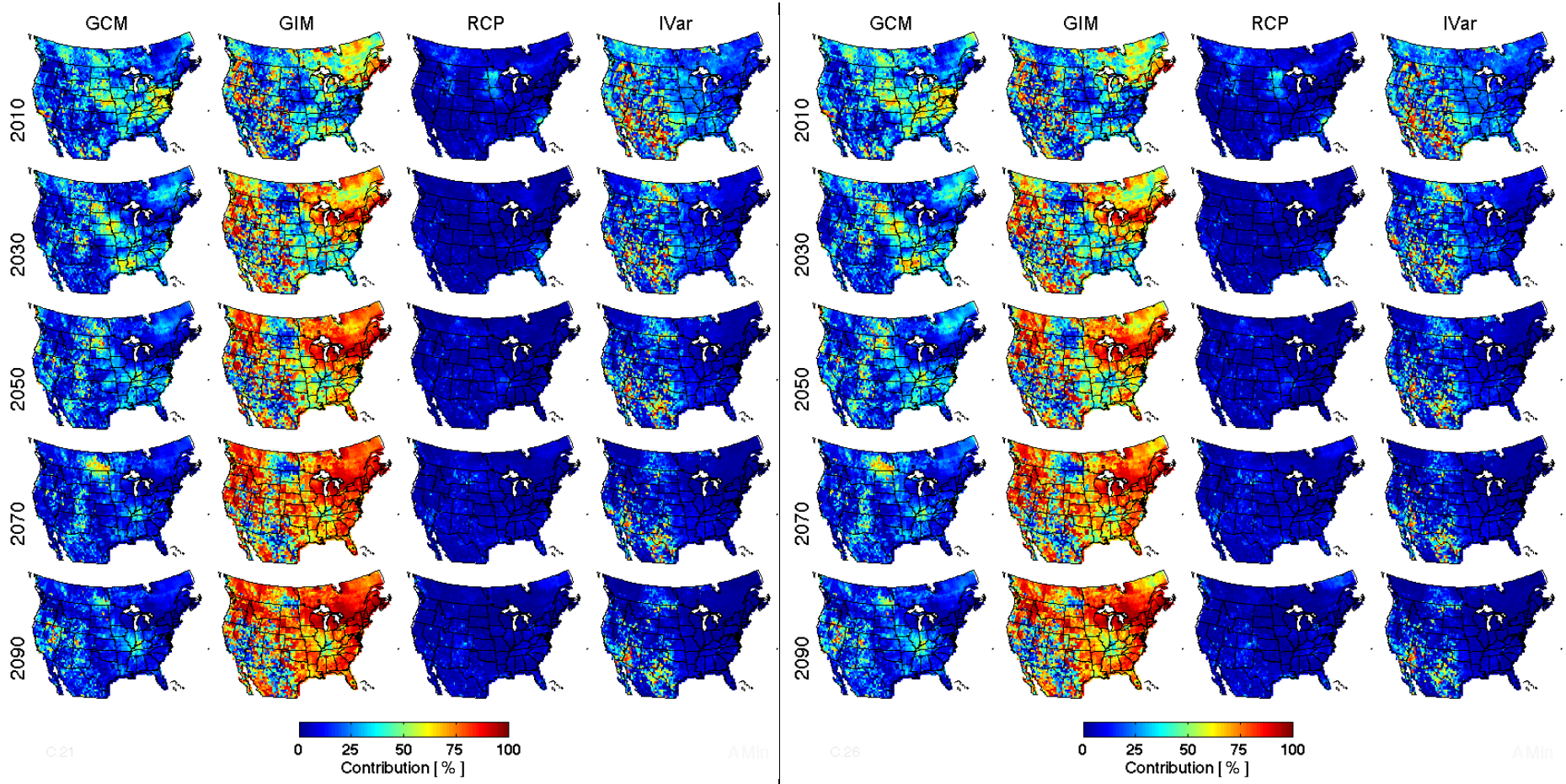


Figure 6.5 – Same as Figure 6.3 but for AMin. Results are shown: a) for the whole ensemble (oE); b) for the ensemble excluding low credibility GIMs (clcE).

The relative contributions to the uncertainty described above refer to a measure of total variance on the ensemble spread. As expected, total variance increases with time to various degrees depending on region and on runoff index. Plots of regional averages of total variance throughout the 21st century are shown in Figure 6.6 (see Figure A4.7-Figure A4.9 for a breakdown of transient total variance per index). Orders of magnitude are closer for AMin and AMed than for the much higher AMax. All three indices show a steady increase across the period, with the highest values and the largest increases in the U.S. Northeast and, in contrast, the smallest values and the weakest

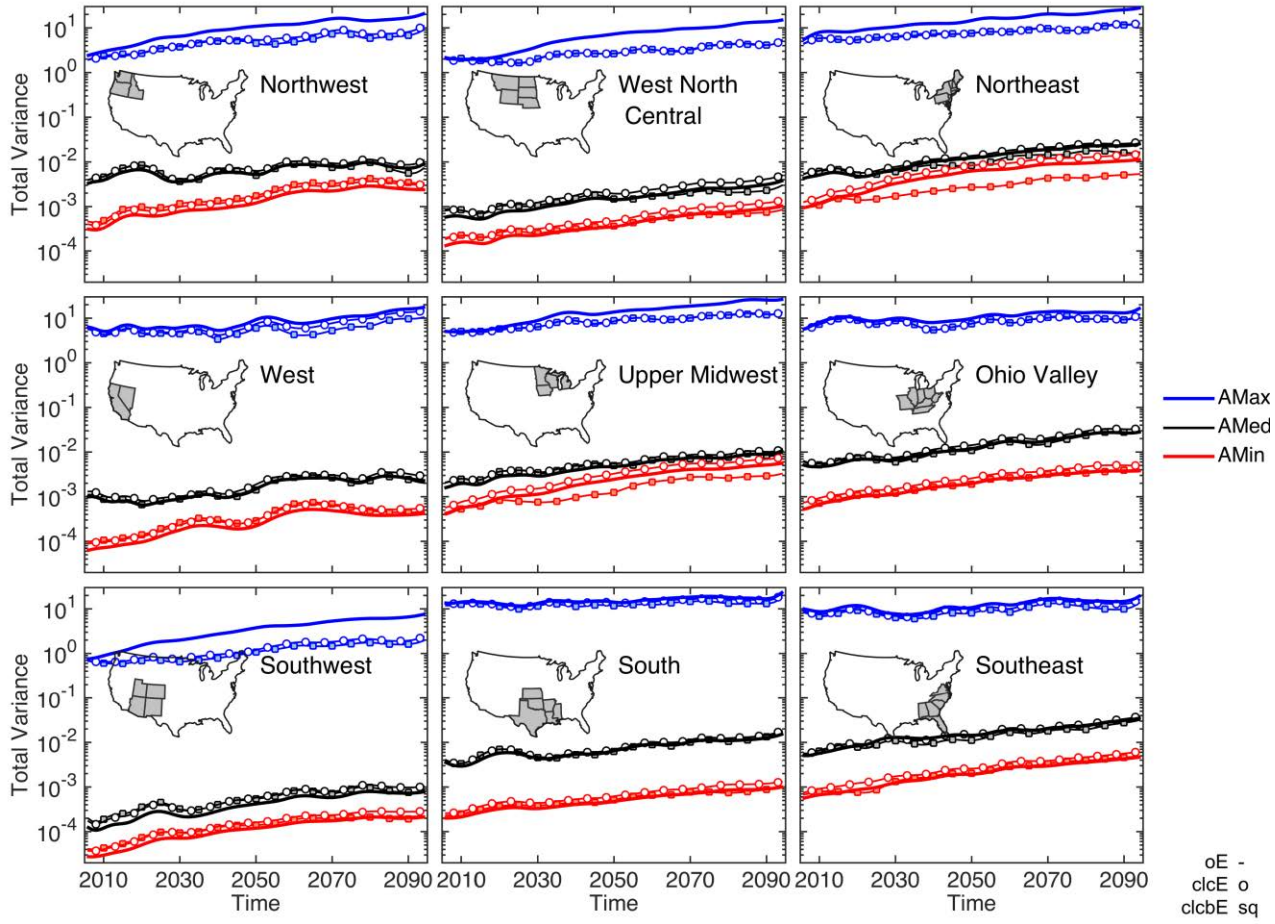


Figure 6.6 - Increase of total variance per region for AMax, AMed, AMin (in mm/day). Thick line: whole ensemble (oE); circle-marked line: low-credibility GIMs excluded (clcE); square-marked line: low-credibility and biome GIMs excluded (clcbE).

increases in the U.S. Southwest. In the south-western United States in particular, very low variances translate into a noisy pattern in the uncertainty partition. The AMax shows the highest increases in the U.S. Northwest, Northeast and Upper Midwest regions. AMed shows the strongest increases in the eastern part of the domain, while AMin shows rates of increase in the eastern part and virtually no increase in the western United States, where values remain very small over the 21st century.

6.4.2 Culled Ensembles

We show results for subsets of the oE by culling i) less-credible GIMs on medium and low flows – clcE; and ii) biome GIMs – clcbE. This is done to examine the effect of the use of different subsets from the oE on the uncertainty partition (Table 6.1 shows the combinations presented in the study). Finally, the effect of excluding intermediate RCPs (4.5, 6.0) – clcirE is investigated.

6.4.2.1 Less-credible GIMs culled Ensemble – clcE

The culling of the oE by excluding two (MATSIRO and LPJmL) out of nine GIMs has an effect on both relative contributions and total variance. For AMax in particular, patterns are markedly different: the GIM source ceases to dominate over most of the domain (Figure 6.3b), the strong contribution in the northern half is indeed limited to the west and northeast of the domain in favour of the uncertainties in the GCMs. Compared to the oE, the lower variance of the clcE is reflected in lower total variance in all regions, particularly those in the northern ones (Figure A4.7). When we focus on AMed (Figure 6.4b), GCM and RCP contributions grow from 5% to 10% over the entire domain, with a reduction in GIM contribution. There are large fractions of GCM uncertainty at the beginning of the run over most of the domain that decline in favour of the GIM uncertainty; GIM uncertainty explains most of the uncertainty at the end of the run, with the exception of the western region. The total variance for clcE is close and only slightly higher to that

of oE (Figure A4.8), suggesting that MATSIRO and LPJmL projections lie within the range of the other GIMs, and do not increase the AMed spread.

Similar to AMed, AMin (Figure 6.5b) has close shares of fractional variance to the oE (Figure 6.5a) all throughout the run. In particular, fractions remain the same, with the exception of a shift of 5-10% in favour of the GCMs, together with a concurrent decrease for GIMs over the entire domain. Total variance increases are in line with those of oE in all regions except for the north-eastern United States (Figure A4.9).

6.4.2.2 Less-credible and biome GIMs culled Ensemble – clcbE

If, in addition to MATSIRO and LPJmL, we exclude the JULES, the resulting ensemble (clcbE) depicts marked differences especially for AMax. It should be noted that the LPJmL and JULES account for vegetation dynamics and varying CO_2 (absent in the other GIMs) and tend to project higher runoff, depending on the region (e.g., not in the southwest).

Consistent with what discussed for clcE, when we focus on AMax the clcbE (Figure A4.10a) depicts even lower fractions of GIM uncertainty, in favor of the GCM and IVar. This is especially true in the southern and south-western parts of the study region. Both clcE and clcbE have in common the exclusion of LPJmL, which is responsible for bringing about high variance in the northern half of the domain from the northwest to the east. Substantial changes in the proportion of total variance result from the exclusion of the three GIMs: the high uncertainty in the northwest and north seen with the oE for the GIMs is no longer present, and GCMs become the dominant source over the entire domain during the whole period. Except for the south and southeast where the GIMs contribution was already small (~15-30%), GCM and IVar increase their shares. This shows that this type of GIMs bring about greater variance to the ensemble for the peak flows, with the exception of the southern/south-eastern United States. It is worth noting that, over the regions

covering the southern half of the United States, the RCP contribution to uncertainty is similar to that of GIMs (~10-20%) and sometimes greater, reaching 30% in the South and Southeast. This suggests that low credibility and biome GIMs can be considered as outliers in the GIMs spread; therefore, their exclusion facilitates an emerging signal of the RCP in these regions, although the GCMs and IVar are still the dominant contributors to uncertainty.

Changes in proportions from oE to clcE and clcbE for AMax are not as pronounced as those obtained for AMed (Figure A4.11a) and AMin (Figure A4.12a) for which we see a clear shift from the GIMs' to the GCMs contribution. Namely, for AMed there is a marked gain in GCM contribution, especially in the northern regions. Interestingly, the exclusion of JULES increases the RCP relative variance: it starts with similar percentages at the beginning of the run (~10% for all regions), increases considerably up to approximately 35%, essentially equalling and almost outweighing those of GIMs and GCMs in the southern and eastern of the domain.

The influence of biome GIMs on AMin's total variance (Figure A4.9) is fairly weak from the west to the south of the study region, where these models tend to simulate zero runoff over extended periods of time, leading to lower variance when included in the oE. Elsewhere (Figure A4.12a), GCMs increase their share by about 10%. As seen for AMed, RCP's contribution increases (especially in the eastern part of the domain) to equal the GCMs' contribution; this holds even though GIMs still remain the dominant source of uncertainty.

The inclusion of biome models in the ensemble brings about larger shares of GIM uncertainty. In particular, the marked difference is the larger and increasing GIM share in the southern half of the domain when JULES is present, owing to the fact that MATSIRO has considerably lower values not only in AMed and AMin (for which annual zero values are common), but also for AMax.

6.4.2.3 Intermediate RCPs culled Ensemble – clcirE

We tested the influence on the model spread due to the use of only the lower (2.6) and higher (8.5) RCPs, as runs for all scenarios are not always available for climate impact studies (in our dataset three GIMs lack intermediate RCPs 4.5 and 6.0). For AMax (Figure A4.10b), compared to the ensemble with all RCPs (Figure 6.3b), apart from a slight reduction in IVar and a slight increase of the GCM variance, the exclusion of intermediate RCPs is barely noticeable on the partition of uncertainty. For AMed (Figure A4.11b – versus all RCPs in Figure 6.4b), the RCP fraction of total variance has higher values in the eastern United States when compared to the clcE, especially in the south and southeast U.S. after 2050. For AMin (Figure A4.12b), IVar proportions decline in the southern half of the domain (as seen for AMax), and RCP proportions increase at the end of the run especially in the east (as seen for AMed). In essence, all indices show proportions similar to those of clcE, but these proportion tend to systematically fluctuate throughout the period, perhaps owing to the use of fewer projections (i.e., decreased information), which is also reflected on the total variances (Figure A4.7-Figure A4.9). In the absence of the two intermediate RCPs, weak changes occur in the uncertainty shares along the different sources, but this does not lead to a change in the overall results; as seen for the previous ensemble setups, this is also consistent with the fact that RCPs account for very little uncertainty in runoff projections.

6.5 Discussion

GCMs and GIMs are responsible for most of the uncertainty in runoff projections over the United States. This is consistent with Wada et al. (2013), who analysed projections of irrigated water demand.

If we consider the physical processes pertaining the study region, Villarini (2016) gives an overview of the meteorological patterns that influence flood seasonality in the United States: the North American monsoon and the North Pacific tropical cyclones in Arizona and New Mexico; snowmelt in the north-central United States; thunderstorms and mesoscale convective systems towards the Gulf of Mexico; North Atlantic tropical cyclones in Florida; extratropical cyclones and atmospheric rivers in the central and western United States; snowmelt, rain on snow and rain on frozen ground in the north-eastern U.S.; tropical and extratropical systems along the U.S. East Coast. Berghuijs et al. (2016) report that the variability of annual flows is much larger for the more arid catchments of the central U.S. and that snow controls the flood response in the Rockies and in the northern states. Moreover, evaporation-controlled soil moisture plays a dominant role over most of the United States.

We can summarize the main contributing factors affecting the uncertainty in: snow-melt, storage-release processes and soil moisture accounting for the GIMs, potentially affecting the AMin; precipitation variability, intensity and spatial/temporal distribution for the GCMs, potentially affecting AMax (for its sudden onset and short duration); uncertainty in CO2 emissions and in the way a warmer world changes the climate for RCPs; and the natural variability in the climate and the hydrology for IVar.

As described in Chapter 3, large differences in parameterization and model structure exist across GIMs. Notably, GIMs accounting for vegetation and CO_2 dynamics tend to produce increased (reduced) runoff in the east (southwest) of the United States, while two GIMs have shown unrealistic runoff representation over the west-southwest. Given the variety of GCMs and GIMs forming the ensemble of projections, we have limited scope to reconcile how the main physical processes play out in the uncertainty contributions and their evolution in time. Clearly, orography

(i.e. mountainous areas) represents a challenge for both GCMs and GIMs, and this may explain the higher fractional contribution of internal variability over the Rocky Mountains. Another difficult setting both in terms of GIMs and GCMs is represented by arid regions (e.g., U.S. southwest); GIMs have also difficulties in the cold regions (e.g., U.S. Midwest or the northern Great Plains), where snow and ice dynamics play a role in the lag at which runoff is generated from melting. Generally, GIM uncertainty is amplified in those areas where storage-release processes play a role in the reproduction of runoff.

In addition to the structural model differences of GIMs and GCMs, more uncertainty lies in the grid transformation from the coarser resolution of the GCMs to the finer resolution of the GIMs, and in the bias-correction of the climate variables. Even though the bias-correction may affect the results of the different sources of uncertainty by reducing inter-GCM and RCP variability (Hempel et al., 2013), we are unable to quantify these impacts because of the lack of non-bias-corrected runs in our dataset. Hempel et al. (2013) also note that, while trends and long-term means are well represented, the ISI-MIP bias-correction approach can present limitations with regards to the adjustment of variability; this is because it corrects daily data about the monthly mean, but it neglects the variability at other time scales (e.g., weekly and monthly variability of precipitation), which can affect the representation of droughts and floods.

Similarly, we do not know to what extent the coarser grid resolution of JULES impacts uncertainty, as no simulations at other scales are available; Prudhomme et al. (2014) note that results based on JULES are generally not very sensitive to the size of the grid cells, at least for modest changes in resolution and for regionally or globally averaged statistics. However, it is difficult to extend these findings to our work.

We have tested to what extent the uncertainty contributions change using the whole ensemble or subsets of it as; this is of relevance to those who may conduct a study with an approach of including all runs, as well as excluding some of the runs deemed less credible or that simply poorly adapt to the methodology. Appraisal of model credibility is arbitrary and much depends on the variable of interest and the metric used. Within our dataset, the five GCMs available have been considered to represent CMIP5 GCMs reasonably well, although McSweeney and Jones (2016) argue that twice as many CMIP5 GCMs would be necessary to adequately represent the range of temperature and precipitation changes globally. This suggests that the GCM uncertainty could be larger if more GCMs were available within ISI-MIP.

While there has been a consensus on using all GCMs, the other studies using this dataset have occasionally excluded some of the nine GIMs from the ensemble. This choice was generally dictated by tractability of the data with respect to the metric used rather than by model selection based on some sort of evaluation or credibility criterion. For instance, with the ISI-MIP runoff dataset at the daily scale, Dankers et al. (2013) have employed all nine GIMs available when working on high flows; on the other hand, Prudhomme et al. (2014) excluded LPJmL and MATSIRO whose runs would poorly adapt to the threshold level method used to quantify drought changes; finally Giuntoli et al. (2015a) excluded LPJmL, MATSIRO and JULES for the joint assessment of high and low flows, with the purpose of keeping a consistent GIM sets for both flow quantiles. One merit of our study was to reveal the difference it makes to use the whole ensemble or to exclude the GIMs that have progressively been culled from the ensemble in the aforementioned studies. This was possible thanks to the indices we employed, for which the presence of long lower tails (zero rich time series) does not pose any type of limitation.

This study clearly shows that the contribution of RCPs to uncertainty is small compared to the other sources. This was already suggested by previous studies, including Wada et al. (2013) Orlowsky and Seneviratne (2013). However, we have shown how this weak contribution holds for runoff indices going from one end (AMin) to the other (AMax) of the runoff distribution, regardless of time and of whether less-credible and biome GIMs are excluded from the analysis.

Finally, within the uncertainty quantification, the sampling of internal variability bears the brunt of lacking model run replicates; therefore, with the method we employed, instead of calculating uncertainty from internal variability via, for instance the variance among replicates, we assess it on the basis of the distance of the departures from to the smooth fit and we assume it to be constant throughout the period of analysis. However, were run replicates available, and therefore e.g., a 4-way ANOVA feasible, we could expect results to change quantitatively, but we have no reason to expect them to change qualitatively.

Moreover, in the interpretation of the uncertainty results, one should bear in mind that fewer models from a culled ensemble do not necessarily lead to a smaller spread and to a lower uncertainty (i.e. uncertainty can increase as a result of computing the variance on the same range of values but with fewer projections). This explains the higher total variances for the smaller sample analysed excluding RCPs that lie between the upper and lower radiative forcing levels (RCPs 2.6 and 8.5). Similarly for AMin, the presence in the ensemble of GIMs producing very low values (e.g., LPJmL and MATSIRO) tends to lower the variance as these GIMs add projections with smaller deviations from the mean.

6.6 Conclusions

The aim of this paper was to partition the uncertainty of ISI-MIP runoff projections throughout the 21st century over the continental United States. The contributions of GCMs, GIMs,

RCPs and Internal Variability to the uncertainties in projected runoff vary regionally and through time depending on the runoff index (annual minimum, median and maximum runoff).

Regardless of the runoff index, GCMs and GIMs account for the largest uncertainty in decadal runoff projections, followed by IVar and ultimately by RCPs, whose contribution is smallest. In particular, uncertainty in AMax is increasingly dominated by GIMs in the western and north-eastern United States, while in the east the GCMs' contributions decrease their share over the 21st century. Although regional patterns differ, similar to Amax, the uncertainty in AMed sees an increase in GIM uncertainty with a concurrent decrease of GCMs' contribution over the 21st century. For AMin, uncertainty is mainly governed by GIMs, the predominant source over the continental United States, in particular towards the end of the 21st century.

After considering the entire ensemble, we have focused on subsets of it, considering fewer GIMs on the basis of their credibility and consideration of vegetation and CO_2 dynamics. We found that the use of the full ensemble or the subsets tested, delivers essentially the same results in the case of AMin and AMed. Conversely, GIMs' uncertainty is markedly lower when fewer projections are used for AMax, in particular when biome models are excluded. This is due to an attenuation of the higher peaks introduced by the biome GIMs (LPJmL and JULES), which simulate increased runoff (thus variance), especially in the north-eastern United States.

As hinted by previous studies at the regional (Giuntoli et al., 2015b; Gudmundsson et al., 2012b) and global (Giuntoli et al., 2015a; Prudhomme et al., 2014) scale, some GIMs have difficulties in simulating runoff, especially low flows. Over our study area, this difficulty is most evident in the U.S. south-southwest, and it explains the higher fractions of variance explained by IVar, a source that is otherwise generally small. Interestingly, the oE, which includes GIMs with lower credibility in medium and low flows, has a lower degree of uncertainty than the clcE. As

mention in the previous section, this is due to the inclusion of projections that have low variance (for being near zero most of the time), producing the apparent beneficial effect of reducing the variance (uncertainty), while it is just an artefact.

Finally, we have tested the influence of excluding two intermediate RCPs, as it can occur when modelling centres contribute to some scenarios and not others to reduce the burden of producing simulations. We found that the use of only the higher (RCP 8.5) and lower (RCP 2.6) scenarios slightly increases the variance (as expected excluding projections from within the sample) and intensifies fluctuations of the different sources in time, but it returns essentially the same results of using all RCPs. Overall, the variety of GIMs used makes it difficult to identify key drivers responsible for how the uncertainty is partitioned and evolves over time across these models.

To conclude, the quest for reducing uncertainty, either via model culling or choosing impact models that are deemed best or more credible, is of little relevance for medium and low flows over the United States, for which uncertainty remains largely the same over time regardless of the selected ensemble. Overall, promising margins for reducing uncertainties will come from improved constraints on the GCMs and GIMs, rather than scenarios. Moreover, the limited impact of internal variability indicates that the majority of the uncertainties are indeed predictable, and not chaotic and irreducible.

7 Conclusions and future work

7.1 Introduction

The overarching aim of the thesis was to assess future runoff from multi-model ensembles and to quantify the uncertainties that lie in the projections. Particular attention was given to GIMs, directly responsible to generate runoff output at the land surface. In light of the knowledge gaps identified in Chapter 2, the objectives of this thesis were:

1) To estimate the changes in high and low flows projections at the global scale towards the end of the 21st century (~2080) providing an assessment of the uncertainty from GIMs and GCMs. This was presented in Chapter 4.

2) To evaluate the ability of GIMs in simulating runoff characteristics in the control period (hindcast) using observed data at the regional scale. This was presented in Chapter 5.

3) To assess the contribution of the different sources of uncertainty (specifically, GCM, GIM, RCP, and internal variability) from the beginning to the end of the 21st century using different ensemble configurations. This was the topic of Chapter 6.

This final chapter summarizes the results on the research undertaken and presents areas for future research.

7.2 Key research findings

Before summarizing conclusions for each objective, the thesis's major research findings are presented as follows:

1) Uncertainty in runoff projections from global impact models is large. The inclusion of a representative number of GIMs in climate impact studies is not sufficient to encompass the overall

uncertainty in future projections from this source. The GIMs need improvement in describing the hydrological processes over land and should be routinely evaluated against observations for avoiding the selection of GIMs that are poorly consistent and for identifying weaknesses and modules that need improvement (e.g., parameterization, processing of evapotranspiration, representation of snow-ice and storage and release processes).

2) Multi-model ensembles tend to project medium flows with a lower uncertainty than at the extremes (high and low flows). In particular, a few GIMs have shown poor realistic representations of runoff (low flows), especially in arid and mountainous areas (e.g., U.S. arid southwest and Rocky Mountains) in which runoff equals zero during large portions of the year. This can affect the uncertainty estimation, indicating to a lower variance (which is the metric that expresses uncertainty) that hints to an erroneously apparent decreased uncertainty.

3) Regardless of considering high, medium, or low flows, RCP scenarios account for negligible shares of uncertainty throughout the 21st century compared to GCMs, GIMs, and, in arid or mountainous areas especially, internal variability.

7.3 Synthesis

7.3.1 Changes in future hydrological extremes and GCM/GIM uncertainty (Chapter 4)

Addressing objective 1, this chapter assessed changes in the frequency of high and low flow days at the end of the 21st century (2066-2099) compared to present day (1972-2005) quantifying the uncertainty from GCMs and GIMs using an ANOVA framework. These changes have shown robust large-scale features: for high flow days, increases were found at northern latitudes with strong signal over eastern Canada, Scandinavia, north-western Russia; for low flow days, increases

were found in several hotspots e.g., southern Europe, south-western and central Latin America, south-eastern USA. As noted in Chapter 4, these patterns are largely consistent with the literature, although, in general, the task of comparing results is non-trivial because the ensembles can vary significantly across experiments in the forcings, the model structures, and the spatial and temporal scales. The identification of both high and low flows at once using the same approach (5-day variable threshold method) had the merit of revealing the different spatial patterns and model agreement depending on which flow is considered. Moreover, the inability to extract the low flows quantile from the control period (1972-2005) of two of the GIMs in particular (LPJmL and MATSIRO), has shed light on the unrealistic simulation of this flow (see, e.g., percentages of vetoed cells per quantile flow in Table 4.1). Without this step, these models would have normally been included in a high flows analysis, hence, the importance to consider multiple aspects of the runoff spectrum when using multi-models ensembles. It is in fact assumed that models should not be limited to satisfactorily simulate one aspect of runoff, but possibly the whole spectrum, although bearing in mind that global models are generally more capable of simulating mean quantities rather than the extremes (Stone and Knutti, 2010). The quantification of the relative uncertainty from GCMs and GIMs has shown how GIMs' contribution can be substantial, especially for low flows, and how it is predominant in snow- and ice-dominated and arid zones. This was attributed to the way the hydrological storage-release processes can modify the climate signal, particularly where storage components are large or water residence times high. The differences in the relative contributions obtained in high and low flows is reflected by the different processes that govern their generation (e.g., mainly precipitation or snow- ice-melt for high flows; evaporation or infiltration and storage for low flows), and provides insightful information on the ensemble uncertainty depending on the flow that is analysed.

7.3.2 Evaluation of GIMs ability in reproducing runoff over the Central U.S. (Chapter 5)

Objective 2 was addressed in Chapter 5. A comparison of simulated runoff with observed streamflow was undertaken considering indices of magnitude and timing for high, medium, and low flows. The comparison consisted in a first level of analysis, in which streamflow gauges were compared to grid cells testing temporal trends, and a second more in-depth level of analysis, in which a pair-wise comparison was undertaken using fewer gauges – those with catchment of size comparable to that of the grid cells. Most GIMs were able to reproduce spatial trends in the observed data over the Central United States. An interesting element was the similarity in the results in relation to the GIM type for several indices. Generally, GIMs closing the water balance (GHMs) were closer to the observations than GIMs that also close the energy balance (LSMs). Depending on the index, the tendency to be out of sync with the observations was generally clustered per GIM type, as in the over- or under- estimation in the frequencies of high and medium flows. This was at least partly attributed to the snow scheme employed: the energy balance for LSMs and the degree-day for GHMs. This dichotomy between GIM types appeared less marked for low flows, for which all GIMs have shown a poorer ability in capturing frequencies of occurrence generally (occurring mostly in the fall for the models and in the summer for the observed data). The difficulty of the models in describing low flows (noted in Chapter 4 and pointed out by e.g., Gudmundsson et al., (2012a)), was further highlighted upon the extraction of the low flow timing index, which was hindered by the abnormal presence of zero-rich time series in some of the GIMs (LPJmL, MATSIRO, and Orchidee, especially). The gauge-to-grid-cell pairwise comparison indicated a better performance of the GIMs in describing the medium flows and its timing compared to the annual maximum and minimum flows, thus confirming the increased difficulty of the GIMs in capturing the ends of the runoff spectrum. However, results also provided confidence on the

capability of the GIMs in simulating specific multimonth extreme events as all models captured successfully intensity and spatial extent of the 1988 drought and the 1993 flood.

As in multi-model ensemble evaluations, the reasons behind the different model behaviour are extremely difficult to track down and disentangle. Indeed, each GIM is unique, but may share few to many characteristics with its peer models (meteorological forcing variables, evapotranspiration scheme) and therefore, it becomes intractable to attribute e.g., a delayed runoff peak to one or more characteristic without being able to perform sensitivity analyses on each of the GIMs employed.

Updated versions of the GIMs assessed in this chapter (or similar ones), have been employed in the ISI-MIP project runs used in Chapter 4 and Chapter 6. Therefore, the overall good performance of the WaterMIP GIMs in describing regional scale hydrology has helped building confidence for using these models in hydrological impact studies.

7.3.3 Uncertainties in projected runoff over the continental U.S. (Chapter 6)

The research on the uncertainty from GCMs and GIMs (Chapter 4) was taken further in Chapter 6 to determine the contribution of also RCPs and internal variability, thus completing objective 3. Uncertainty was partitioned transiently throughout the 21st century (as opposed to ~2080 in Chapter 4) and fractional uncertainties in decadal maximum, median, and minimum flows were thus described over time. The continental USA was chosen as study region as it encompasses a diverse range of morphology and climates, and it includes the region in which the GIMs have been evaluated (in Chapter 5). Regardless of the flow, internal variability predominates over areas of topographic complexity (U.S. southwest, the Rocky Mountains), while GCMs and GIMs account for the largest uncertainty, followed by RCPs, whose contribution is smallest. Uncertainty shares

vary depending on the index, with GCMs generally contributing the most in maximum and median flows while GIMs contributed the most for minimum flows. The exclusion (culling) from the entire ensemble of GIMs on the basis of their credibility and consideration of vegetation and CO_2 dynamics has delivered the same results essentially in the case of minimum and medium flows, whereas for maximum flows GIMs' uncertainty was markedly lower when biome models were excluded. It can be concluded that for maximum flows the biome models (LPJmL and JULES) contribute to a larger spread (especially in north-eastern United States) that yields a more comprehensive sampling of plausible futures (i.e. uncertainty). On the other hand, the difficulty of GIMs in simulating low flows, most evident in the south-southwest U.S. (and documented throughout this thesis), is not reflected in the uncertainty shares, which are lower when the GIMs that simulate low flows poorly (e.g., MATSIRO) are included. This effect, due to the lower variance produced when unrealistic zero-rich time series are included, highlights the importance of performing model evaluation in the region of study to avoid interpretations of uncertainty that are biased by low credibility runs in the ensemble. Overall, provided that uncertainty from internal variability is inherently irreducible, promising margins for reducing uncertainty in runoff projections will come from improvements in GCMs and GIMs, rather than scenarios, whose contribution proved minimal in all indices and ensemble setups.

7.4 Future directions

Hydrological impact studies make use of multi-model ensembles to which different modelling groups contribute form across the world. Scientists and modellers work on common frameworks to produce model output using high performance computers, and the resources required for letting this happen are considerable, as it is considerable the amount of data that is produced for each experiment. For instance, the ISI-MIP daily unrouted runoff runs alone – considering all

GCMs, all GIMs, and all scenarios available – amount to ~1.2 Terabytes of storage memory. If the modelling effort is huge and time consuming, the data retrieval and analysis is also very demanding. As a result, it is hard to exploit the data and to identify model errors and biases in a comprehensive way. As discussed in Chapter 5, the risk is then to generate a plethora of studies that describe a more or less certain future for a given variable, time-window, and metric, without being able to fully discern model errors and deficiencies. This is particularly difficult to do at the global scale, at which many hydrology relevant processes are levelled out (e.g., precipitations). Clearly, the long-term goal is to enhance global models to include more detailed processes (e.g., vegetation dynamics, water abstraction from human activities) and to run them at higher spatial and temporal resolution. However, there is an immediate demand for evaluating these models at the regional scale (where observed data is available) in order to ensure reliability of simulations and to provide insight on how the processes develop over land and how the different variables are intertwined. This calls for an effort to develop standardized tools and assessment strategies that will facilitate comparisons, increase robustness in the results and help in the selection of models, discounting those that are found unsatisfactory in the representation of key processes (McSweeney et al., 2015). For instance, in the context of assessing changes in future flood risk, as in e.g., Dankers et al. (2013), fitting a generalized extreme value distribution (GEV) (e.g., Coles (2001)) to historical and future annual peak flows could be undertaken at the regional scale, benchmarking modelled historical peak distributions against peak distribution from observed data, retaining only consistent model runs for fitting peak flows in the future.

This evaluation effort is also valid for GCMs, which feed the GIMs. For instance, McSweeney et al. (2015) have provided a valuable evaluation of GCMs in capturing the range in surface temperature and precipitation over selected regions of the world, attesting which models are

'implausible', thus suggesting to avoid the poorest ones in the selection of models suitable for climate change impact assessments. In a later study (McSweeney and Jones, 2016), the same authors have shed light on the representativeness of the five (out of the ~36 available) CMIP5 GCMs used for the ISI-MIP project runs, suggesting that they sufficiently express the uncertainty range in future climate impact, but that at least 13 GCMs would be required to capture 80% of the range in >75% of regions and seasons. The fact that they also emphasize how the selection of regionally optimized subsets of models (i.e. fewer models that perform well in a given region) can capture the uncertainty range equally well, suggests that climate impact studies, that are often carried out at the regional or continental scale, should be undertaken with an informed selection of GCMs and GIMs.

In addition to model evaluation, further investigations are needed to better characterize the uncertainty in hydrological projections. This thesis has highlighted that an improvement in global models is needed to reduce overall uncertainty, as the contribution from GCMs and GIMs – regardless of the flow type – is very large and that the role of RCPs is negligible. In addition to assess uncertainty in other regions of the world, additional sources neglected in this study should be quantified, like bias-correction, which requires the use of both bias-corrected and non-corrected runs, as noted in Chapter 4. The influence of internal variability on uncertainty should also be further investigated through the use of run replicates (i.e. runs with different initial conditions, unavailable within ISI-MIP at this stage), as opposed to a measure of the distance of the projections from their mean (as done in Chapter 6).

Discarding 'implausible' models and characterizing uncertainty from the aforementioned sources can indeed yield greater confidence in the projected climate.

Finally, promising prospects for improved climate impact studies lie in the assessment of GIMs output variables (e.g., temperature, precipitation, runoff) whose joint analysis can help both to identify biases in the models and to better quantify the uncertainty in the projections, as it has been done in the literature, especially for the GCMs (e.g., Tebaldi and Sansó (2009)).

REFERENCES

Adam, J.C., Lettenmaier, D.P., 2003. Adjustment of global gridded precipitation for systematic bias. J. Geophys. Res. Atmos. 108, n/a–n/a. doi:10.1029/2002JD002499

Alfieri, L., Burek, P., Feyen, L., Forzieri, G., 2015. Global warming increases the frequency of river floods in Europe. Hydrol. Earth Syst. Sci. 19, 2247–2260. doi:10.5194/hess-19-2247-2015

Andreadis, K.M., Clark, E.A., Wood, A.W., Hamlet, A.F., Lettenmaier, D.P., 2005. Twentieth-Century Drought in the Conterminous United States. J. Hydrometeorol. 6, 985–1001. doi:10.1175/JHM450.1

Arnell, N.W., 2003. Effects of IPCC SRES* emissions scenarios on river runoff: a global perspective. Hydrol. Earth Syst. Sci. 7, 619–641. doi:10.5194/hess-7-619-2003

Arnell, N.W., Gosling, S.N., 2013. The impacts of climate change on river flow regimes at the global scale. J. Hydrol. 486, 351–364. doi:10.1016/j.jhydrol.2013.02.010

Bentsen, M., Bethke, I., Debernard, J.B., Iversen, T., Kirkevåg, A., Seland, Ø., Drange, H., Roelandt, C., Seierstad, I. a., Hoose, C., Kristjánsson, J.E., 2012. The Norwegian Earth System Model, NorESM1-M – Part 1: Description and basic evaluation. Geosci. Model Dev. Discuss. 5, 2843–2931. doi:10.5194/gmdd-5-2843-2012

Berghuijs, W.R., Woods, R.A., Hutton, C.J., Sivapalan, M., 2016. Dominant flood generating mechanisms across the United States. Geophys. Res. Lett. 1–9. doi:10.1002/2016GL068070

Beven, K., 2011. Rainfall-runoff modelling: the primer. John Wiley & Sons.

Bosshard, T., Carambia, M., Goergen, K., Kotlarski, S., Krahe, P., Zappa, M., Schär, C., 2013. Quantifying uncertainty sources in an ensemble of hydrological climate-impact projections. Water Resour. Res. 49, n/a–n/a. doi:10.1029/2011WR011533

Cleveland, W.S., 1979. Robust Locally Weighted Regression and Smoothing Scatterplots. J. Am. Stat. Assoc. 74, 829. doi:10.2307/2286407

Coles, S., 2001. An introduction to statistical modeling of extreme values, Springer Series in Statistics. doi:10.1007/978-1-4471-3675-0

Collins, M., Knutti, R., Arblaster, J., Dufresne, J.-L., Fichefet, T., Friedlingstein, P., Gao, X., Gutowski, W.J., Johns, T., Krinner, G., Shongwe, M., Tebaldi, C., Weaver, A.J., Wehner, M., 2013. Long-term Climate Change: Projections, Commitments and Irreversibility, in: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. pp. 1029–1136. doi:10.1017/CBO9781107415324.015

Collins, W.J., Bellouin, N., Doutriaux-Boucher, M., Gedney, N., Halloran, P., Hinton, T., Hughes, J., Jones, C.D., Joshi, M., Liddicoat, S., Martin, G., O'Connor, F., Rae, J., Senior, C., Sitch, S., Totterdell, I., Wiltshire, A., Woodward, S., 2011. Development and evaluation of an Earth-System model – HadGEM2. Geosci. Model Dev. 4, 1051–1075. doi:10.5194/gmd-4-1051-2011

Dankers, R., Arnell, N.W., Clark, D.B., Falloon, P.D., Fekete, B.M., Gosling, S.N., Heinke, J., Kim, H., Masaki, Y., Satoh, Y., Stacke, T., Wada, Y., Wisser, D., 2013. First look at changes in flood hazard in the Inter-Sectoral Impact Model Intercomparison Project ensemble. Proc. Natl. Acad. Sci. U. S. A. 1–5. doi:10.1073/pnas.1302078110

Davie, J.C.S., Falloon, P.D., Kahana, R., Dankers, R., Betts, R., Portmann, F.T., Wisser, D., Clark, D.B., Ito, A., Masaki, Y., Nishina, K., Fekete, B., Tessler, Z., Wada, Y., Liu, X., Tang, Q., Hagemann, S., Stacke, T., Pavlick, R., Schaphoff, S., Gosling, S.N., Franssen, W., Arnell, N., 2013. Comparing projections of future changes in runoff from hydrological and biome models

in ISI-MIP. Earth Syst. Dyn. 4, 359–374. doi:10.5194/esd-4-359-2013

Deser, C., Knutti, R., Solomon, S., Phillips, A.S., 2012a. Communication of the role of natural variability in future North American climate. Nat. Clim. Chang. 2, 775–779. doi:10.1038/nclimate1562

Deser, C., Phillips, A., Bourdette, V., Teng, H., 2012b. Uncertainty in climate change projections: the role of internal variability. Clim. Dyn. 38, 527–546. doi:10.1007/s00382-010-0977-x

Döll, P., Schmied, H.M., 2012. How is the impact of climate change on river flow regimes related to the impact on mean annual runoff? A global-scale analysis. Environ. Res. Lett. 7, 014037. doi:10.1088/1748-9326/7/1/014037

Douglas, E.M., Vogel, R.M., Kroll, C.N., 2000. Trends in floods and low flows in the United States: impact of spatial correlation. J. Hydrol. 240, 90–105. doi:10.1016/S0022-1694(00)00336-X

Dufresne, J.-L., Foujols, M.-A., Denvil, S., Caubel, A., Marti, O., Aumont, O., Balkanski, Y., Bekki, S., Bellenger, H., Benshila, R., Bony, S., Bopp, L., Braconnot, P., Brockmann, P., Cadule, P., Cheruy, F., Codron, F., Cozic, A., Cugnet, D., de Noblet, N., Duvel, J.-P., Ethé, C., Fairhead, L., Fichefet, T., Flavoni, S., Friedlingstein, P., Grandpeix, J.-Y., Guez, L., Guilyardi, E., Hauglustaine, D., Hourdin, F., Idelkadi, A., Ghattas, J., Joussaume, S., Kageyama, M., Krinner, G., Labetoulle, S., Lahellec, A., Lefebvre, M.-P., Lefevre, F., Levy, C., Li, Z.X., Lloyd, J., Lott, F., Madec, G., Mancip, M., Marchand, M., Masson, S., Meurdesoif, Y., Mignot, J., Musat, I., Parouty, S., Polcher, J., Rio, C., Schulz, M., Swingedouw, D., Szopa, S., Talandier, C., Terray, P., Viovy, N., Vuichard, N., 2013. Climate change projections using the IPSL-CM5 Earth System Model: from CMIP3 to CMIP5. Clim. Dyn. 40, 2123–2165. doi:10.1007/s00382-012-1636-1

Dunne, J.P., John, J.G., Adcroft, A.J., Griffies, S.M., Hallberg, R.W., Shevliakova, E., Stouffer, R.J., Cooke, W., Dunne, K.A., Harrison, M.J., Krasting, J.P., Malyshev, S.L., Milly, P.C.D., Phillipps, P.J., Sentman, L.T., Samuels, B.L., Spelman, M.J., Winton, M., Wittenberg, A.T., Zadeh, N., 2012. GFDL's ESM2 Global Coupled Climate–Carbon Earth System Models. Part I: Physical Formulation and Baseline Simulation Characteristics. J. Clim. 25, 6646–6665. doi:10.1175/JCLI-D-11-00560.1

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?" Hydrol. Earth Syst. Sci. 16, 3391–3404. doi:10.5194/hess-16-3391-2012

Feyen, L., Dankers, R., 2009. Impact of global warming on streamflow drought in Europe. J. Geophys. Res. 114, D17116. doi:10.1029/2008JD011438

Forzieri, G., Feyen, L., Rojas, R., Flörke, M., Wimmer, F., Bianchi, A., 2014. Ensemble projections of future streamflow droughts in Europe. Hydrol. Earth Syst. Sci. 18, 85–108. doi:10.5194/hess-18-85-2014

Giuntoli, I., Renard, B., Vidal, J.-P., Bard, A., 2013. Low flows in France and their relationship to large-scale climate indices. J. Hydrol. 482, 105–118. doi:10.1016/j.jhydrol.2012.12.038

Giuntoli, I., Vidal, J.-P., Prudhomme, C., Hannah, D.M., 2015a. Future hydrological extremes: the uncertainty from multiple global climate and global hydrological models. Earth Syst. Dyn. 6, 267–285. doi:10.5194/esd-6-267-2015

Giuntoli, I., Villarini, G., Prudhomme, C., Mallakpour, I., Hannah, D.M., 2015b. Evaluation of global impact models' ability to reproduce runoff characteristics over the central United States. J. Geophys. Res. Atmos. 120, 9138–9159. doi:10.1002/2015JD023401

Gosling, S.N., Arnell, N.W., 2011. Simulating current global river runoff with a global hydrological

model: model revisions, validation, and sensitivity analysis. Hydrol. Process. 25, 1129–1145. doi:10.1002/hyp.7727

Gudmundsson, L., Tallaksen, L.M., Stahl, K., Clark, D.B., Dumont, E., Hagemann, S., Bertrand, N., Gerten, D., Heinke, J., Hanasaki, N., Voss, F., Koirala, S., 2012a. Comparing Large-Scale Hydrological Model Simulations to Observed Runoff Percentiles in Europe. J. Hydrometeorol. 13, 604–620. doi:10.1175/JHM-D-11-083.1

Gudmundsson, L., Wagener, T., Tallaksen, L.M., Engeland, K., 2012b. Evaluation of nine large-scale hydrological models with respect to the seasonal runoff climatology in Europe. Water Resour. Res. 48, W11504. doi:10.1029/2011WR010911

Haddeland, I., Clark, D.B., Franssen, W., Ludwig, F., Voß, F., Arnell, N.W., Bertrand, N., Best, M., Folwell, S., Gerten, D., Gomes, S., Gosling, S.N., Hagemann, S., Hanasaki, N., Harding, R., Heinke, J., Kabat, P., Koirala, S., Oki, T., Polcher, J., Stacke, T., Viterbo, P., Weedon, G.P., Yeh, P., 2011. Multimodel Estimate of the Global Terrestrial Water Balance: Setup and First Results. J. Hydrometeorol. 12, 869–884. doi:10.1175/2011JHM1324.1

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

Hagemann, S., Chen, C., Haerter, J.O., Heinke, J., Gerten, D., Piani, C., 2011. Impact of a Statistical Bias Correction on the Projected Hydrological Changes Obtained from Three GCMs and Two Hydrology Models. J. Hydrometeorol. 12, 556–578. doi:10.1175/2011JHM1336.1

Hannah, D.M., Demuth, S., van Lanen, H.A.J., Looser, U., Prudhomme, C., Rees, G., Stahl, K., Tallaksen, L.M., 2011. Large-scale river flow archives: importance, current status and future

needs. Hydrol. Process. 25, 1191–1200. doi:10.1002/hyp.7794

Hawkins, E., Sutton, R., 2011. The potential to narrow uncertainty in projections of regional precipitation change. Clim. Dyn. 37, 407–418. doi:10.1007/s00382-010-0810-6

Hawkins, E., Sutton, R., 2009. The Potential to Narrow Uncertainty in Regional Climate Predictions. Bull. Am. Meteorol. Soc. 90, 1095–1107. doi:10.1175/2009BAMS2607.1

Helsel, D.R., Hirsch, R.M., 1992. Statistical methods in water resources. Elsevier 49.

Hempel, S., Frieler, K., Warszawski, L., Schewe, J., Piontek, F., 2013. A trend-preserving bias correction – the ISI-MIP approach. Earth Syst. Dyn. 4, 219–236. doi:10.5194/esd-4-219-2013

Hingray, B., Saïd, M., 2014. Partitioning Internal Variability and Model Uncertainty Components in a Multimember Multimodel Ensemble of Climate Projections. J. Clim. 27, 6779–6798. doi:10.1175/JCLI-D-13-00629.1

Hirabayashi, Y., Kanae, S., Emori, S., Oki, T., Kimoto, M., 2008. Global projections of changing risks of floods and droughts in a changing climate. Hydrol. Sci. J. 53, 754–772. doi:10.1623/hysj.53.4.754

Hirabayashi, Y., Mahendran, R., Koirala, S., Konoshima, L., Yamazaki, D., Watanabe, S., Kim, H., Kanae, S., 2013. Global flood risk under climate change. Nat. Clim. Chang. 3, 816–821. doi:10.1038/nclimate1911

Hirsch, R.M., Ryberg, K.R., 2012. Has the magnitude of floods across the USA changed with global CO 2 levels? Hydrol. Sci. J. 57, 1–9. doi:10.1080/02626667.2011.621895

Huber, V., Schellnhuber, H.J., Arnell, N.W., Frieler, K., Friend, A.D., Gerten, D., Haddeland, I., Kabat, P., Lotze-Campen, H., Lucht, W., Parry, M., Piontek, F., Rosenzweig, C., Schewe, J., Warszawski, L., 2014. Climate impact research: beyond patchwork. Earth Syst. Dyn. 5, 399–408. doi:10.5194/esd-5-399-2014

Hunger, M., Döll, P., 2008. Value of river discharge data for global-scale hydrological modeling. Hydrol. Earth Syst. Sci. … 12, 841–861.

Huntington, T., 2006. Evidence for intensification of the global water cycle: Review and synthesis. J. Hydrol. 319, 83–95. doi:10.1016/j.jhydrol.2005.07.003

Jones, C.D., Hughes, J.K., Bellouin, N., Hardiman, S.C., Jones, G.S., Knight, J., Liddicoat, S., O'Connor, F.M., Andres, R.J., Bell, C., Boo, K.-O., Bozzo, A., Butchart, N., Cadule, P., Corbin, K.D., Doutriaux-Boucher, M., Friedlingstein, P., Gornall, J., Gray, L., Halloran, P.R., Hurtt, G., Ingram, W.J., Lamarque, J.-F., Law, R.M., Meinshausen, M., Osprey, S., Palin, E.J., Parsons Chini, L., Raddatz, T., Sanderson, M.G., Sellar, A.A., Schurer, A., Valdes, P., Wood, N., Woodward, S., Yoshioka, M., Zerroukat, M., 2011. The HadGEM2-ES implementation of CMIP5 centennial simulations. Geosci. Model Dev. 4, 543–570. doi:10.5194/gmd-4-543-2011

Jung, I.-W., Moradkhani, H., Chang, H., 2012. Uncertainty assessment of climate change impacts for hydrologically distinct river basins. J. Hydrol. 466-467, 73–87. doi:10.1016/j.jhydrol.2012.08.002

Kharin, V. V., Zwiers, F.W., Zhang, X., Wehner, M., 2013. Changes in temperature and precipitation extremes in the CMIP5 ensemble. Clim. Change 119, 345–357. doi:10.1007/s10584-013-0705-8

Kirkevåg, a., Iversen, T., Seland, Ø., Hoose, C., Kristjánsson, J.E., Struthers, H., Ekman, a. M.L., Ghan, S., Griesfeller, J., Nilsson, E.D., Schulz, M., 2012. Aerosol-climate interactions in the Norwegian Earth System Model – NorESM. Geosci. Model Dev. Discuss. 5, 2599–2685. doi:10.5194/gmdd-5-2599-2012

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

Kottek, M., Grieser, J., Beck, C., Rudolf, B., Rubel, F., 2006. World Map of the Köppen-Geiger climate classification updated. Meteorol. Zeitschrift 15, 259–263. doi:10.1127/0941-2948/2006/0130

Lavell, A., Oppenheimer, M., Diop, C., Hess, J., Lempert, R., Li, J., Muir-Wood, R., Myeong, S., 2012. Climate change: new dimensions in disaster risk, exposure, vulnerability, and resilience, in: Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. pp. 25–64. doi:10.1017/CBO9781139177245.004

Lins, H.F., Slack, J.R., 2005. Seasonal and Regional Characteristics of U.S. Streamflow Trends in the United States from 1940 to 1999. Phys. Geogr. 26, 489–501. doi:10.2747/0272-3646.26.6.489

Lins, H.F., Slack, J.R., 1999. Streamflow trends in the United States. Geophys. Res. Lett. 26, 227–230. doi:10.1029/1998GL900291

Mallakpour, I., Villarini, G., 2015. The changing nature of flooding across the central United States. Nat. Clim. Chang. 1–5. doi:10.1038/nclimate2516

Maloney, E.D., Camargo, S.J., Chang, E., Colle, B., Fu, R., Geil, K.L., Hu, Q., Jiang, X., Johnson, N., Karnauskas, K.B., Kinter, J., Kirtman, B., Kumar, S., Langenbrunner, B., Lombardo, K., Long, L.N., Mariotti, A., Meyerson, J.E., Mo, K.C., Neelin, J.D., Pan, Z., Seager, R., Serra, Y., Seth, A., Sheffield, J., Stroeve, J., Thibeault, J., Xie, S.-P., Wang, C., Wyman, B., Zhao, M., 2014. North American Climate in CMIP5 Experiments: Part III: Assessment of Twenty-First-Century Projections*. J. Clim. 27, 2230–2270. doi:10.1175/JCLI-D-13-00273.1

McCabe, G.J., Wolock, D.M., 2002. A step increase in streamflow in the conterminous United States. Geophys. Res. Lett. 29, 2185. doi:10.1029/2002GL015999

McSweeney, C.F., Jones, R.G., 2016. How representative is the spread of climate projections from

the 5 CMIP5 GCMs used in ISI-MIP? Clim. Serv. 1, 24–29. doi:10.1016/j.cliser.2016.02.001

McSweeney, C.F., Jones, R.G., Lee, R.W., Rowell, D.P., 2015. Selecting CMIP5 GCMs for downscaling over multiple regions. Clim. Dyn. 44, 3237–3260. doi:10.1007/s00382-014-2418-8

Meinshausen, M., Smith, S.J., Calvin, K., Daniel, J.S., Kainuma, M.L.T., Lamarque, J.-F., Matsumoto, K., Montzka, S.A., Raper, S.C.B., Riahi, K., Thomson, A., Velders, G.J.M., van Vuuren, D.P.P., 2011. The RCP greenhouse gas concentrations and their extensions from 1765 to 2300. Clim. Change 109, 213–241. doi:10.1007/s10584-011-0156-z

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, 347–50. doi:10.1038/nature04312

Mölders, N., 2005. Feedbacks at the hydro-meteorological interface, in: Bronstert, A., Carrera, J., Kabat, P., Lütkemeier, S. (Ed.), Coupled Models for the Hydrological Cycle - Integrating Atmosphere, Biosphere, and Pedosphere. Springer, pp. 192–208.

Moore, J.N., Harper, J.T., Greenwood, M.C., 2007. Significance of trends toward earlier snowmelt runoff, Columbia and Missouri Basin headwaters, western United States. Geophys. Res. Lett. 34. doi:10.1029/2007GL031022

Moss, R.H., Edmonds, J. a, Hibbard, K. a, Manning, M.R., Rose, S.K., van Vuuren, D.P., Carter, T.R., Emori, S., Kainuma, M., Kram, T., Meehl, G. a, Mitchell, J.F.B., Nakicenovic, N., Riahi, K., Smith, S.J., Stouffer, R.J., Thomson, A.M., Weyant, J.P., Wilbanks, T.J., 2010. The next generation of scenarios for climate change research and assessment. Nature 463, 747–756. doi:10.1038/nature08823

Müller Schmied, H., Eisner, S., Franz, D., Wattenbach, M., Portmann, F.T., Flörke, M., Döll, P., 2014. Sensitivity of simulated global-scale freshwater fluxes and storages to input data,

hydrological model structure, human water use and calibration. Hydrol. Earth Syst. Sci. 18, 3511–3538. doi:10.5194/hess-18-3511-2014

Nakicenovic, N., Alcamo, J., Davis, G., de Vries, B., Fenhann, J., Gaffin, S., Gregory, K., Grübler, A., Yong Jung, T., Kram, T., Lebre La Rovere, E., Michaelis, L., Mori, S., Morita, T., Pepper, W., Pitcher, H., Price, L., Riahi, K., Roehrl, A., Rogner, H.-H., Sankovski, A., Schlesinger, M., Shukla, P., Smith, S., Swart, R., van Rooijen, S., Victor, N., Dadi, Z., 2000. IPCC Special Report on Emissions Scenarios (SRES), Special Report on Emissions Scenarios: A Special Report of Working Group III of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, U.K.

Neter, J., Kutner, M.H., Nachtsheim, C.J., Wasserman, W., 1999. Applied Linear Statistical Models, Irwin series in statistics. McGraw-Hill, New York.

Nohara, D., Kitoh, A., Hosaka, M., Oki, T., 2006. Impact of Climate Change on River Discharge Projected by Multimodel Ensemble. J. Hydrometeorol. 7, 1076–1089.

Northrop, P.J., 2013. Comments on "A Simple, Coherent Framework for Partitioning Uncertainty in Climate Predictions." J. Clim. 26, 4375–4376. doi:10.1175/JCLI-D-12-00527.1

Orlowsky, B., Seneviratne, S.I., 2013. Elusive drought: uncertainty in observed trends and short- and long-term CMIP5 projections. Hydrol. Earth Syst. Sci. 17, 1765–1781. doi:10.5194/hess-17-1765-2013

Overland, J.E., Wang, M., Bond, N.A., Walsh, J.E., Kattsov, V.M., Chapman, W.L., 2011. Considerations in the Selection of Global Climate Models for Regional Climate Projections: The Arctic as a Case Study*. J. Clim. 24, 1583–1597. doi:10.1175/2010JCLI3462.1

Pendergrass, A.G., Lehner, F., Sanderson, B.M., Xu, Y., 2015. Does extreme precipitation intensity depend on the emissions scenario? Geophys. Res. Lett. 42, 8767–8774.

doi:10.1002/2015GL065854

Peterson, T.C., Heim, R.R., Hirsch, R., Kaiser, D.P., Brooks, H., Diffenbaugh, N.S., Dole, R.M., Giovannettone, J.P., Guirguis, K., Karl, T.R., Katz, R.W., Kunkel, K., Lettenmaier, D., McCabe, G.J., Paciorek, C.J., Ryberg, K.R., Schubert, S., Silva, V.B.S., Stewart, B.C., Vecchia, A. V., Villarini, G., Vose, R.S., Walsh, J., Wehner, M., Wolock, D., Wolter, K., Woodhouse, C. a., Wuebbles, D., 2013. Monitoring and Understanding Changes in Heat Waves, Cold Waves, Floods, and Droughts in the United States: State of Knowledge. Bull. Am. Meteorol. Soc. 94, 821–834. doi:10.1175/BAMS-D-12-00066.1

Prudhomme, C., Davies, H., 2008. Assessing uncertainties in climate change impact analyses on the river flow regimes in the UK. Part 2: future climate. Clim. Change 93, 197–222. doi:10.1007/s10584-008-9461-6

Prudhomme, C., Giuntoli, I., Robinson, E.L., Clark, D.B., Arnell, N.W., Dankers, R., Fekete, B.M., Franssen, W., Gerten, D., Gosling, S.N., Hagemann, S., Hannah, D.M., Kim, H., Masaki, Y., Satoh, Y., Stacke, T., Wada, Y., Wisser, D., 2014. Hydrological droughts in the 21st century, hotspots and uncertainties from a global multimodel ensemble experiment. Proc. Natl. Acad. Sci. U. S. A. 111, 3262–7. doi:10.1073/pnas.1222473110

Prudhomme, C., Parry, S., Hannaford, J., Clark, D.B., Hagemann, S., Voss, F., 2011. How Well Do Large-Scale Models Reproduce Regional Hydrological Extremes in Europe? J. Hydrometeorol. 12, 1181–1204. doi:10.1175/2011JHM1387.1

Rasmussen, T.J., Perry, C.A., 2001. Trends in Peak Flows of Selected Streams in Kansas. U.S. Geological Survey Water-Resources Investigations Report 01-4203.

Roudier, P., Andersson, J.C.M., Donnelly, C., Feyen, L., Greuell, W., Ludwig, F., 2015. Projections of future floods and hydrological droughts in Europe under a +2°C global warming. Clim.

Change. doi:10.1007/s10584-015-1570-4

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. J. Clim. 26, 4017–4037. doi:10.1175/JCLI-D-12-00462.1

Schewe, J., Heinke, J., Gerten, D., Haddeland, I., Arnell, N.W., Clark, D.B., Dankers, R., Eisner, S., Fekete, B.M., Colón-González, F.J., Gosling, S.N., Kim, H., Liu, X., Masaki, Y., Portmann, F.T., Satoh, Y., Stacke, T., Tang, Q., Wada, Y., Wisser, D., Albrecht, T., Frieler, K., Piontek, F., Warszawski, L., Kabat, P., 2014. Multimodel assessment of water scarcity under climate change. Proc. Natl. Acad. Sci. 111, 3245–3250. doi:10.1073/pnas.1222460110

Seneviratne, S., Nicholls, N., Easterling, D., Goodess, C., Kanae, S., Kossin, J., Luo, Y., Marengo, J., McInnes, K., Rahimi, M., Reichstein, M., Sorteberg, A., Vera, C., Zhang, X., 2012. Changes in climate extremes and their impacts on the natural physical environment. Manag. Risk Extrem. Events Disasters to Adv. Clim. Chang. Adapt. A Spec. Rep. Work. Groups I II IPCC, Annex IIanaging Risks Extrem. Events Disasters to Adv. Clim. Chang. Adapt. 109–230.

Sheffield, J., Wood, E.F., 2008. Projected changes in drought occurrence under future global warming from multi-model, multi-scenario, IPCC AR4 simulations. Clim. Dyn. 31, 79–105. doi:10.1007/s00382-007-0340-z

Sillmann, J., Kharin, V. V., Zwiers, F.W., Zhang, X., Bronaugh, D., 2013. Climate extremes indices in the CMIP5 multimodel ensemble: Part 2. Future climate projections. J. Geophys. Res. Atmos. 118, 2473–2493. doi:10.1002/jgrd.50188

Smith, M.B., Koren, V., Zhang, Z., Zhang, Y., Reed, S.M., Cui, Z., Moreda, F., Cosgrove, B. a., Mizukami, N., Anderson, E. a., 2012. Results of the DMIP 2 Oklahoma experiments. J. Hydrol. 418-419, 17–48. doi:10.1016/j.jhydrol.2011.08.056

Stahl, K., Tallaksen, L.M., Hannaford, J., van Lanen, H. a. J., 2012. Filling the white space on maps of European runoff trends: estimates from a multi-model ensemble. Hydrol. Earth Syst. Sci. 16, 2035–2047. doi:10.5194/hess-16-2035-2012

Stewart, I.T., Cayan, D.R., Dettinger, M.D., 2005. Changes toward Earlier Streamflow Timing across Western North America. J. Clim. 18, 1136–1155. doi:10.1175/JCLI3321.1

Stone, D.A., Knutti, R., 2010. Weather and Climate, in: Fung, F., Lopez, A., New, M. (Eds.), Modelling the Impact of Climate Change on Water Resources. John Wiley & Sons, Ltd, Chichester, UK, pp. 4–33. doi:10.1002/9781444324921.ch2

Stott, P.A., Gillett, N.P., Hegerl, G.C., Karoly, D.J., Stone, D. a., Zhang, X., Zwiers, F., 2010. Detection and attribution of climate change: a regional perspective. Wiley Interdiscip. Rev. Clim. Chang. 1, 192–211. doi:10.1002/wcc.34

Tallaksen, L.M., Stahl, K., 2014. Spatial and temporal patterns of large-scale droughts in Europe: Model dispersion and performance. Geophys. Res. Lett. 41, 429–434. doi:10.1002/2013GL058573

Tang, Q., Lettenmaier, D.P., 2012. 21st century runoff sensitivities of major global river basins. Geophys. Res. Lett. 39, n/a–n/a. doi:10.1029/2011GL050834

Taylor, I.H., Burke, E., McColl, L., Falloon, P., Harris, G.R., McNeall, D., 2012. Contributions to uncertainty in projections of future drought under climate change scenarios. Hydrol. Earth Syst. Sci. Discuss. 9, 12613–12653. doi:10.5194/hessd-9-12613-2012

Taylor, K.E., Stouffer, R.J., Meehl, G. a., 2012. An Overview of CMIP5 and the Experiment Design. Bull. Am. Meteorol. Soc. 93, 485–498. doi:10.1175/BAMS-D-11-00094.1

Tebaldi, C., Hayhoe, K., Arblaster, J.M., Meehl, G.A., 2006. Going to the Extremes. Clim. Change 79, 185–211. doi:10.1007/s10584-006-9051-4

Tebaldi, C., Knutti, R., 2007. The use of the multi-model ensemble in probabilistic climate projections. Philos. Trans. A. Math. Phys. Eng. Sci. 365, 2053–75. doi:10.1098/rsta.2007.2076

Tebaldi, C., Sansó, B., 2009. Joint projections of temperature and precipitation change from multiple climate models: a hierarchical Bayesian approach. J. R. Stat. Soc. Ser. A 172, 83–106.

Tebaldi, C., Smith, R.L., 2009. Characterizing the Uncertainty of Climate Change Projections Using Hierarchical Models, in: O'Hagan, A., West, M. (Eds.), The Handbook of Applied Bayesian Analysis. Oxford University Press, pp. 545–594.

Thibeault, J.M., Seth, A., 2014. A Framework for Evaluating Model Credibility for Warm-Season Precipitation in Northeastern North America: A Case Study of CMIP5 Simulations and Projections. J. Clim. 27, 493–510. doi:10.1175/JCLI-D-12-00846.1

Uppala, S.M., Kallberg, P.W., Simmons, a. J., Andrae, U., Bechtold, V.D.C., Fiorino, M., Gibson, J.K., Haseler, J., Hernandez, A., Kelly, G. a., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R.P., Andersson, E., Arpe, K., Balmaseda, M. a., Beljaars, a. C.M., Berg, L. Van De, Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hólm, E., Hoskins, B.J., Isaksen, L., Janssen, P. a. E.M., Jenne, R., Mcnally, a. P., Mahfouf, J.-F., Morcrette, J.-J., Rayner, N. a., Saunders, R.W., Simon, P., Sterl, A., Trenberth, K.E., Untch, A., Vasiljevic, D., Viterbo, P., Woollen, J., 2005. The ERA-40 re-analysis. Q. J. R. Meteorol. Soc. 131, 2961–3012. doi:10.1256/qj.04.176

van Huijgevoort, M.H.J., Hazenberg, P., van Lanen, H.A.J., Teuling, A.J., Clark, D.B., Folwell, S., Gosling, S.N., Hanasaki, N., Heinke, J., Koirala, S., Stacke, T., Voss, F., Sheffield, J., Uijlenhoet, R., 2013. Global Multimodel Analysis of Drought in Runoff for the Second Half of the Twentieth Century. J. Hydrometeorol. 14, 1535–1552. doi:10.1175/JHM-D-12-0186.1

van Huijgevoort, M.H.J., Hazenberg, P., van Lanen, H.A.J., Uijlenhoet, R., 2012. A generic method

for hydrological drought identification across different climate regions. Hydrol. Earth Syst. Sci. 16, 2437–2451. doi:10.5194/hess-16-2437-2012

van Huijgevoort, M.H.J., van Lanen, H.A.J., Teuling, A.J., Uijlenhoet, R., 2014. Identification of changes in hydrological drought characteristics from a multi-GCM driven ensemble constrained by observed discharge. J. Hydrol. 512, 421–434. doi:10.1016/j.jhydrol.2014.02.060

van Loon, A.F., van Huijgevoort, M.H.J., van Lanen, H.A.J., 2012. Evaluation of drought propagation in an ensemble mean of large-scale hydrological models. Hydrol. Earth Syst. Sci. 16, 4057–4078. doi:10.5194/hess-16-4057-2012

van Vuuren, D.P., Edmonds, J., Kainuma, M., Riahi, K., Thomson, A., Hibbard, K., Hurtt, G.C., Kram, T., Krey, V., Lamarque, J.-F., Masui, T., Meinshausen, M., Nakicenovic, N., Smith, S.J., Rose, S.K., 2011. The representative concentration pathways: an overview. Clim. Change 109, 5–31. doi:10.1007/s10584-011-0148-z

Vidal, J.-P., Martin, E., Kitova, N., Najac, J., Soubeyroux, J.-M., 2012. Evolution of spatio-temporal drought characteristics: validation, projections and effect of adaptation scenarios. Hydrol. Earth Syst. Sci. 16, 2935–2955. doi:10.5194/hess-16-2935-2012

Villarini, G., 2016. On the seasonality of flooding across the continental United States. Adv. Water Resour. 87, 80–91. doi:10.1016/j.advwatres.2015.11.009

Villarini, G., Smith, J.A., Baeck, M.L., Krajewski, W.F., 2011. Examining Flood Frequency Distributions in the Midwest U.S. JAWRA J. Am. Water Resour. Assoc. 47, 447–463. doi:10.1111/j.1752-1688.2011.00540.x

Villarini, G., Smith, J.A., Vecchi, G.A., 2013. Changing Frequency of Heavy Rainfall over the Central United States. J. Clim. 26, 351–357. doi:10.1175/JCLI-D-12-00043.1

Villarini, G., Vecchi, G.A., 2012. Twenty-first-century projections of North Atlantic tropical storms from CMIP5 models. Nat. Clim. Chang. 2, 604–607. doi:10.1038/nclimate1530

Vogel, R.M., Yaindl, C., Walter, M., 2011. Nonstationarity: Flood Magnification and Recurrence Reduction Factors in the United States. JAWRA J. Am. Water Resour. Assoc. 47, 464–474. doi:10.1111/j.1752-1688.2011.00541.x

Wada, Y., Wisser, D., Eisner, S., Flörke, M., Gerten, D., Haddeland, I., Hanasaki, N., Masaki, Y., Portmann, F.T., Stacke, T., Tessler, Z., Schewe, J., 2013. Multimodel projections and uncertainties of irrigation water demand under climate change. Geophys. Res. Lett. 40, 4626–4632. doi:10.1002/grl.50686

Wanders, N., Wada, Y., Van Lanen, H.A.J., 2015. Global hydrological droughts in the 21st century under a changing hydrological regime. Earth Syst. Dyn. 6, 1–15. doi:10.5194/esd-6-1-2015

Wang, A., Bohn, T.J., Mahanama, S.P., Koster, R.D., Lettenmaier, D.P., 2009. Multimodel Ensemble Reconstruction of Drought over the Continental United States. J. Clim. 22, 2694–2712. doi:10.1175/2008JCLI2586.1

Warszawski, L., Frieler, K., Huber, V., Piontek, F., Serdeczny, O., Schewe, J., 2014. The Inter-Sectoral Impact Model Intercomparison Project (ISI-MIP): project framework. Proc. Natl. Acad. Sci. U. S. A. 111, 3228–32. doi:10.1073/pnas.1312330110

Watanabe, S., Hajima, T., Sudo, K., Nagashima, T., Takemura, T., Okajima, H., Nozawa, T., Kawase, H., Abe, M., Yokohata, T., Ise, T., Sato, H., Kato, E., Takata, K., Emori, S., Kawamiya, M., 2011. MIROC-ESM: model description and basic results of CMIP5-20c3m experiments. Geosci. Model Dev. Discuss. 4, 1063–1128. doi:10.5194/gmdd-4-1063-2011

Weedon, G.P., Gomes, S., Adam, J.C., Bellouin, N., Viterbo, P., Bellouin, N., Boucher, O., Best, M., 2010. The Watch Forcing Data 1958-2001: a Meteorological Forcing Dataset for Land

Surface-and Hydrological-Models. Watch Tech. Rep. 22 41p.

Weedon, G.P., Gomes, S., Viterbo, P., Shuttleworth, W.J., Blyth, E., Österle, H., Adam, J.C., Bellouin, N., Boucher, O., Best, M., 2011. Creation of the WATCH Forcing Data and Its Use to Assess Global and Regional Reference Crop Evaporation over Land during the Twentieth Century. J. Hydrometeorol. 12, 823–848. doi:10.1175/2011JHM1369.1

Whitfield, P.H., Burn, D.H., Hannaford, J., Higgins, H., Hodgkins, G. a., Marsh, T., Looser, U., 2012. Reference hydrologic networks I. The status and potential future directions of national reference hydrologic networks for detecting trends. Hydrol. Sci. J. 57, 1562–1579. doi:10.1080/02626667.2012.728706

Williams, K., Clark, D.B., 2014. Hadley Centre Technical Note 96 Disaggregation of daily data in JULES. Exeter, UK.

Yang, H., Piao, S., Zeng, Z., Ciais, P., Yin, Y., Friedlingstein, P., Sitch, S., Ahlström, A., Guimberteau, M., Huntingford, C., Levis, S., Levy, P.E., Huang, M., Li, Y., Li, X., Lomas, M.R., Peylin, P., Poulter, B., Viovy, N., Zaehle, S., Zeng, N., Zhao, F., Wang, L., 2015. Multicriteria evaluation of discharge simulation in Dynamic Global Vegetation Models. J. Geophys. Res. Atmos. 120, 7488–7505. doi:10.1002/2015JD023129

Yip, S., Ferro, C. a. T., Stephenson, D.B., Hawkins, E., 2011. A Simple, Coherent Framework for Partitioning Uncertainty in Climate Predictions. J. Clim. 24, 4634–4643. doi:10.1175/2011JCLI4085.1

Zhao, T., Dai, A., 2015. The Magnitude and Causes of Global Drought Changes in the Twenty-First Century under a Low–Moderate Emissions Scenario. J. Clim. 28, 4490–4512. doi:10.1175/JCLI-D-14-00363.1

Zolina, O., Kapala, A., Simmer, C., Gulev, S.K., 2004. Analysis of extreme precipitation over

Europe from different reanalyses: a comparative assessment. Glob. Planet. Change 44, 129–161. doi:10.1016/j.gloplacha.2004.06.009

APPENDIX I

Table A1.1 – CMIP5 GCM models used in ISI-MIP and their attributes (adapted from Maloney et al. (2014)).

Model	Expanded name	Centre	Atmospheric horizontal resolution (°lon×°lat)	N. of model levels	Reference Models' main components
HadGEM2-ES	Hadley Centre Global Environment Model, version 2– Earth System	Met Office Hadley Centre, United Kingdom	1.875 × 1.25	60	References: Collins et al. (2011); Jones et al. (2011) HadGEM2-ES is a coupled AOGCM, which also represents interactive land and ocean carbon cycles and dynamic vegetation. HadGEM2-ES is based on the following components: Met Office Unified Model for the ocean atmosphere; TRIFFID for the dynamic global vegetation; diat-HadOCC for the ocean biology and carbonate chemistry; UKCA for the chemistry and aerosols. The model is 30min (atmosphere and land) and 1 h (ocean).
IPSL-CM5A-LR	Institut Pierre-Simon Laplace Coupled (Earth System) Model, version 5, coupled with NEMO, low resolution	Institut Pierre-Simon Laplace, France	3.75 × 1.875	39	Reference: Dufresne et al. (2013) IPSL-CM5A-LR is a coupled ocean-atmosphere general circulation model coupled to a carbon model composed of a biogeochemical model (PISCES) for the ocean part and of a dynamic global vegetation model (ORCHIDEE) for the terrestrial part. IPSL-CM5 is based on the following model components: LMDZ atmospheric model; NEMO ocean model, including sea ice and marine biogeochemistry; ORCHIDEE model of continental surfaces including carbon cycle; INCA model of chemistry and aerosols; LMDz-REPROBUS coupled chemistry-climate model.
MIROC-ESM-CHEM	Model for Interdisciplinary Research on Climate, Earth System Model, Chemistry Coupled	Japan Agency for Marine-Earth Science and Technology, Atmosphere and Ocean Research Institute, and National Institute for Env. Studies, Japan.	2.8 × 2.8	80	Reference: Watanabe et al. (2011) MIROC-ESM-CHEM is an Earth System Model based on the MIROC-AGCM atmospheric general circulation model, which includes an on-line aerosol component (SPRINTARS 5), an ocean GCM with sea-ice component (COCO 3.4), and a land surface model (MATSIRO). A flux coupler couples the components. MIROC-ESM further includes an atmospheric chemistry component (CHASER 4.1), a nutrient-phytoplankton-zooplankton-detritus (NPZD) type ocean ecosystem component, and a terrestrial ecosystem component dealing with dynamic vegetation (SEIB-DGVM).
GFDL-ESM2M	Geophysical Fluid Dynamics Laboratory Earth System Model with Modular Ocean Model 4 (MOM4) component.	NOAA/Geophysical Fluid Dynamics Laboratory, United States	2.5 × 2.0	24	References: Dunne et al. (2012) GFDL-ESM2M is global coupled carbon–climate Earth System Model. The atmosphere and sea ice components are based on Climate Model CM2.1, while the Land Model (LM3.0) component represents land, water, energy, and carbon cycles. The model time step is 30min (atmosphere, land and sea-ice) and 2 h (ocean, atmosphere).
NorESM1-M	Norwegian Earth System Model, version 1 (intermediate resolution) and with carbon cycle	Norwegian Climate Center, Norway	2.5 × 1.9	26	References: Bentsen et al. (2012); Kirkevåg et al. (2012) NorESM1-M is an Earth System Model based on the NCAR CCSM4 climate model. The atmospheric module is modified with the CAM4-Oslo model chemistry-aerosol-cloud-radiation interaction schemes. NorESM1-M uses the biogeochemical ocean module (HAMOCC) adapted to an isopycnic ocean model (MICOM), the sea ice CICE4 model, the CLM4 land model with with carbon-nitrogen cycle enabled. The components are coupled with CPL7 (used in CCSM4).

APPENDIX II

Appendix to Chapter 4.

Masking of grid cells – With reference to the control period (1972–2005), grid cells showing little or no seasonal change in daily runoff were screened out using the 5-day percentiles series that form the threshold curves (i.e. one mask for HF and one for LF) following these rules: (i) percentiles are equal to zero for more than one-third of the year; (ii) standard deviation of percentiles of first and/or second half-year equals zero; (iii) annual percentiles Q_{10} and Q_{95} series are equal.

Tests on ANOVA's residuals – To verify whether the ANOVA model assumptions hold, statistical tests were performed on the ANOVA residuals. For every unmasked grid cell, for both HFI and LFI, residuals were assessed as follows: we tested (i) normality with the Lilliefors test; and then, for grid cells for which the null hypothesis (that the residuals' vector comes from a distribution in the normal family) was not rejected, we tested (ii) constancy of variance with the Hartley test. Results for the annual and seasonal ANOVAs show that HFI has higher rates of residuals for which the hypotheses of normality and constancy of variance were rejected compared to the LFI. For the year, the percentages of unmasked grid cells not meeting the residuals requirements were: HFI 22 % not normal, 15 % no constant variance, for a total of 37 % globally; LFI 12 % not normal, 15 % no constant variance, for a total of 27 % globally. JJA and DJF have the lowest proportions of residuals' requirements not met for HFI and LFI respectively. We also applied the ANOVA on HFI and LFI transformed via the normal-score method (seeking normality

of the data); this showed lower percentages of cells not satisfying the ANOVA assumptions of normality and constant variance (HFI: 7.5 and 11 %; and LFI: 7 and 12 % respectively) for a total of 19 % globally. It should be noted that the residuals' contribution to uncertainty tends to be lower for the transformed data (e.g. grid cells with residuals' dominated uncertainty decreased by 6 % for HFI and 1 % for LFI). Because the partition of uncertainty between GCMs and GHMs are similar from both ANOVA applied to raw and transformed data sets, and because the areas of non-satisfaction of normality are not located where the residuals dominate the uncertainty, we discussed results obtained from the raw, non-transformed data.

APPENDIX III

Appendix to Chapter 5.

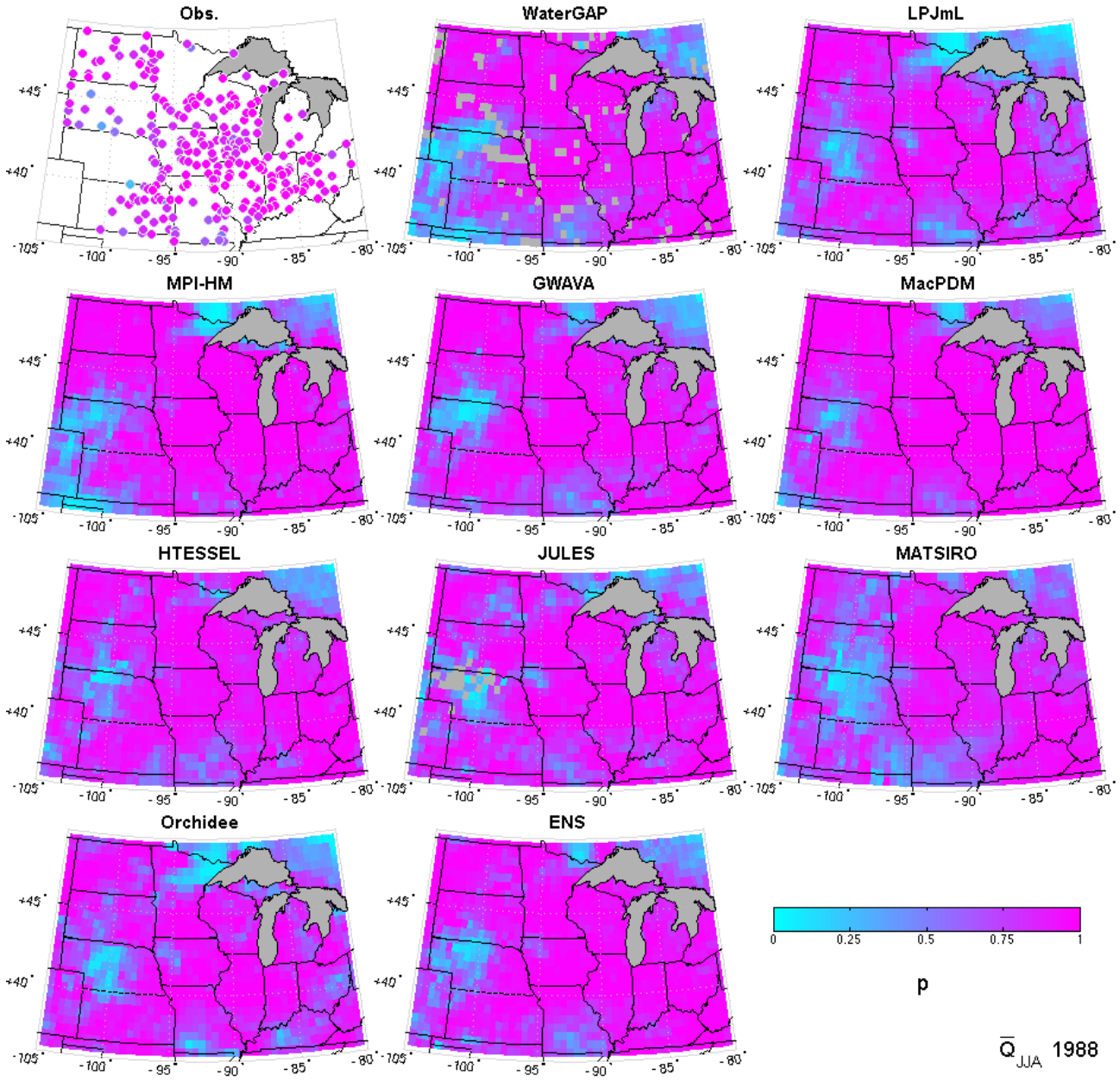


Figure A3.1 – Estimated exceedance probabilities [p] of mean summer (JJA) runoff for the 1988 drought.

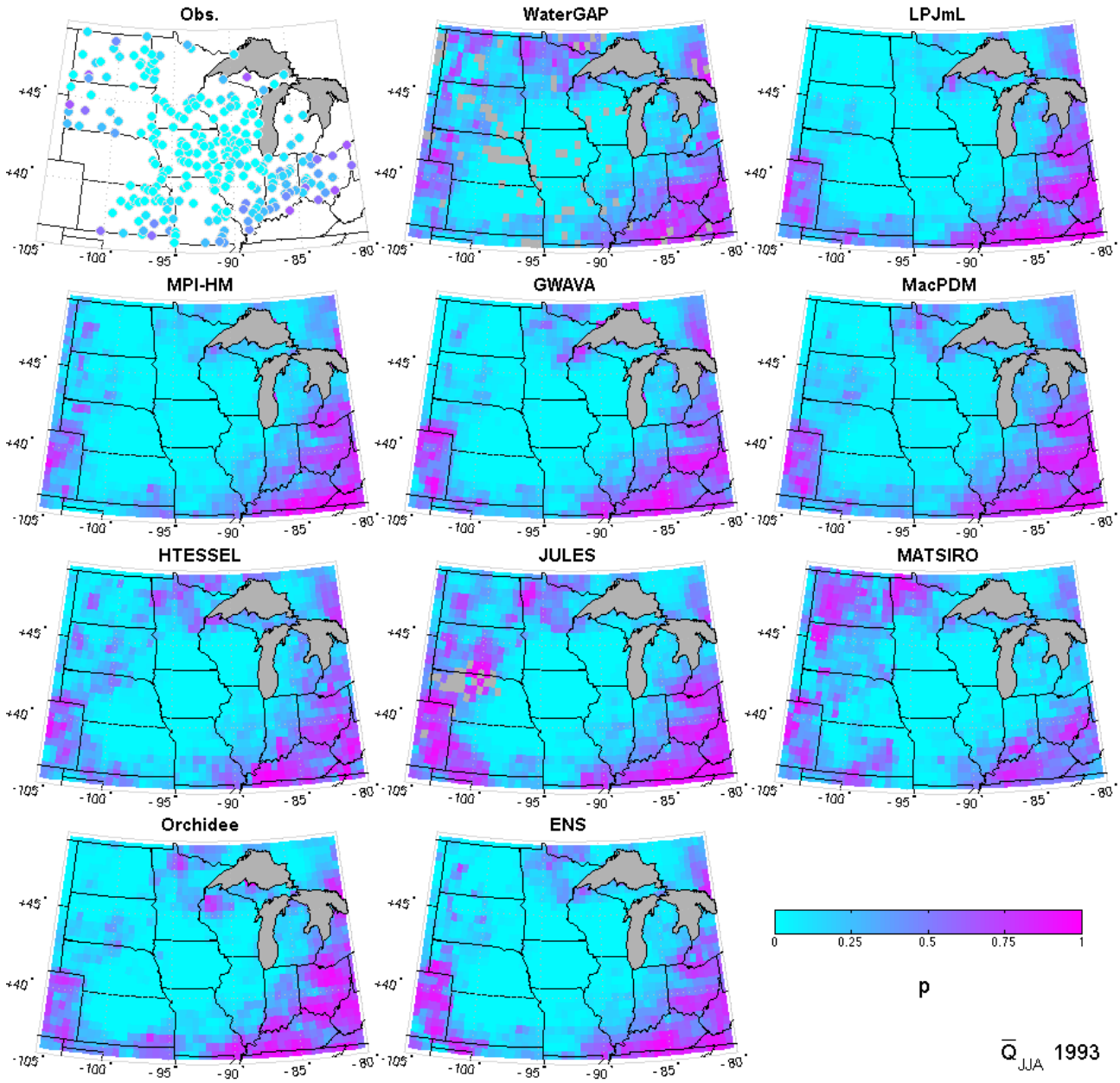


Figure A3.2– Same as Figure A3.1 for the 1993 flood.

APPENDIX IV

Appendix to Chapter 6.

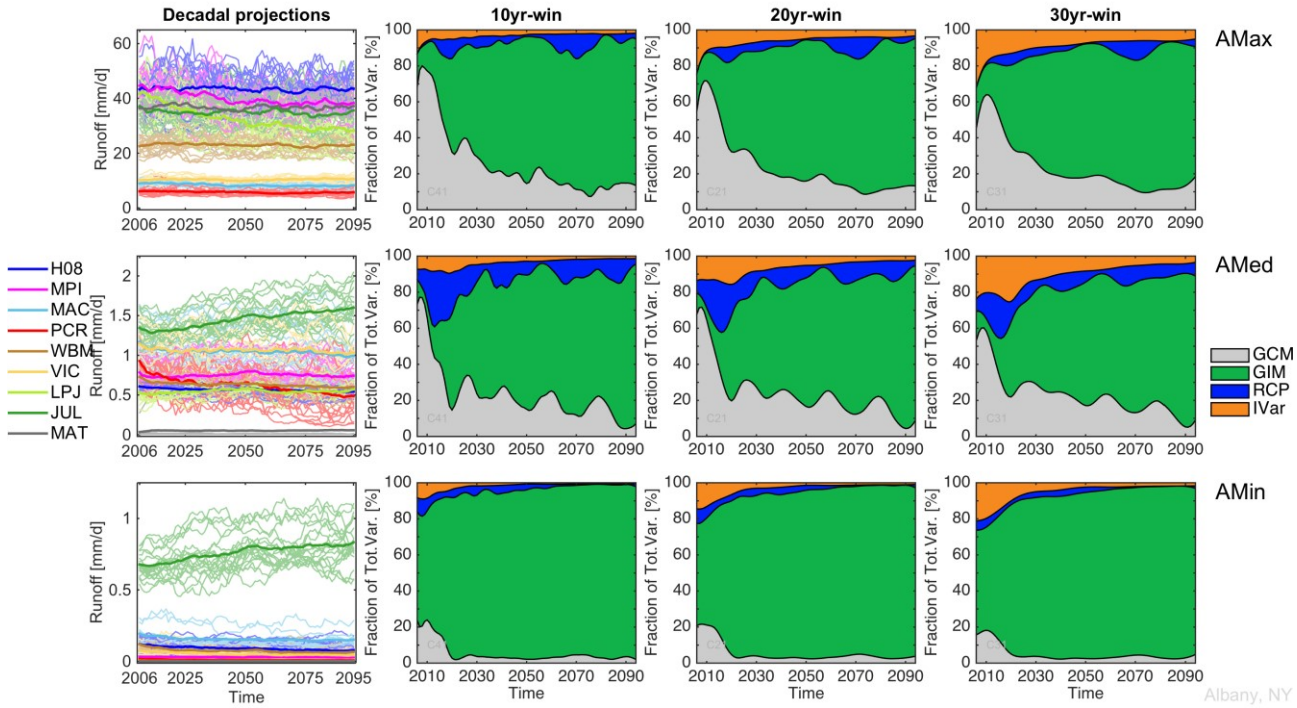


Figure A4.1 – For a selected grid cell (42.7° N -73.9° E; Albany, NY). On the first column: decadally averaged AMax, AMed, and AMin projections (top, mid, and lower panels respectively) colored according to GIM (thick lines are averages per GIM); second to fourth columns: corresponding fractional uncertainty using a loess window of 10, 20 (chosen for the study), and 30 years respectively.

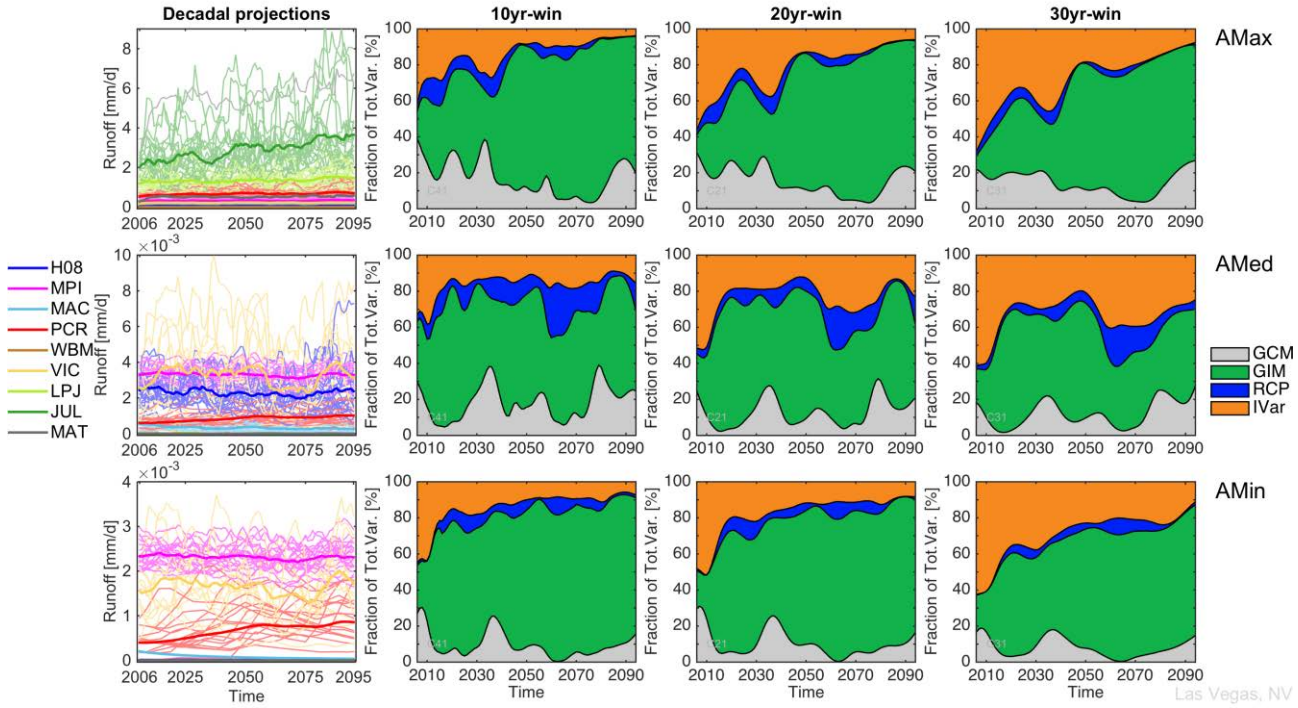


Figure A4.2 – Same as Figure A4.1 but for a contrasting grid cell (36.1° N -115.3° E; Las Vegas, NV).

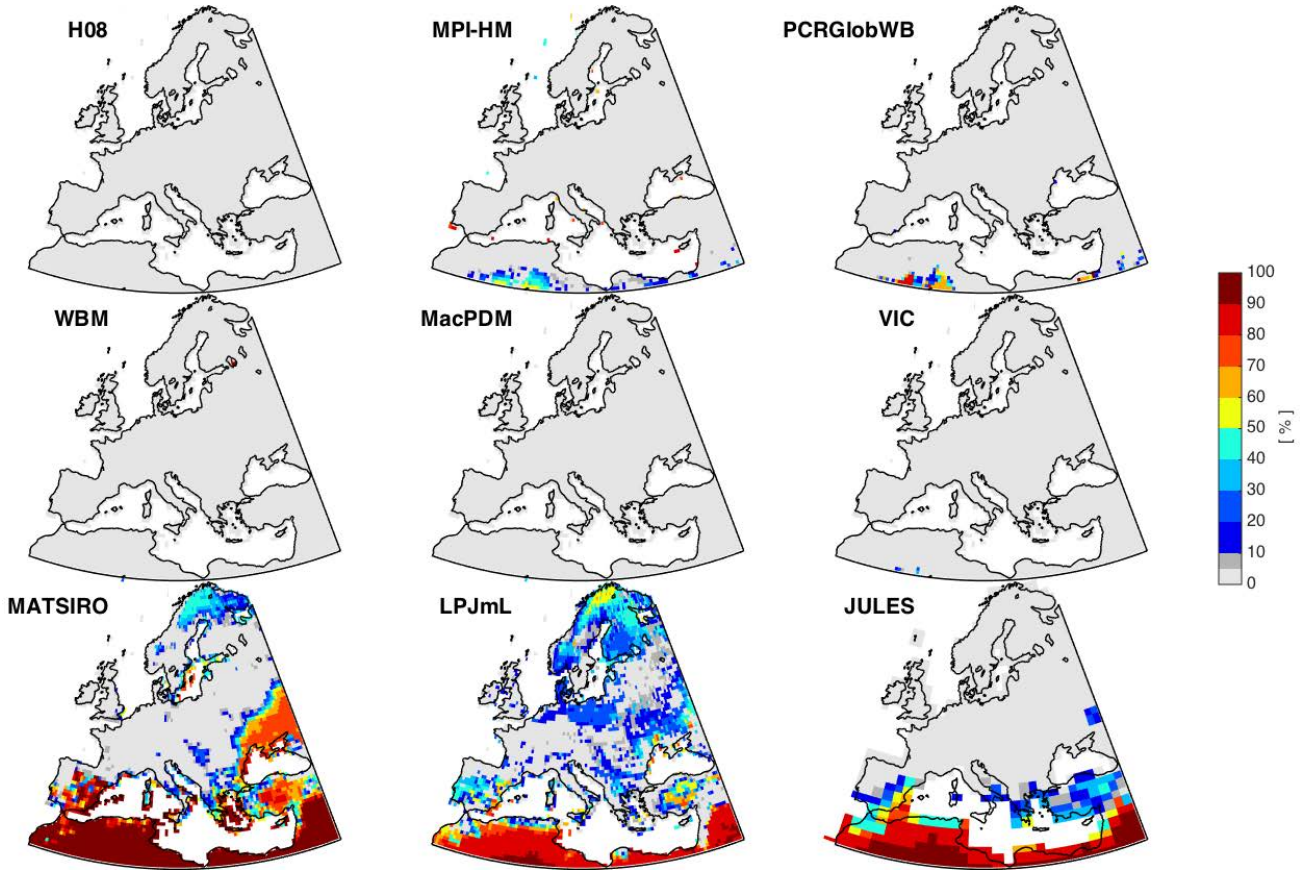


Figure A4.3 – Same as Figure 6.2 but for Europe.

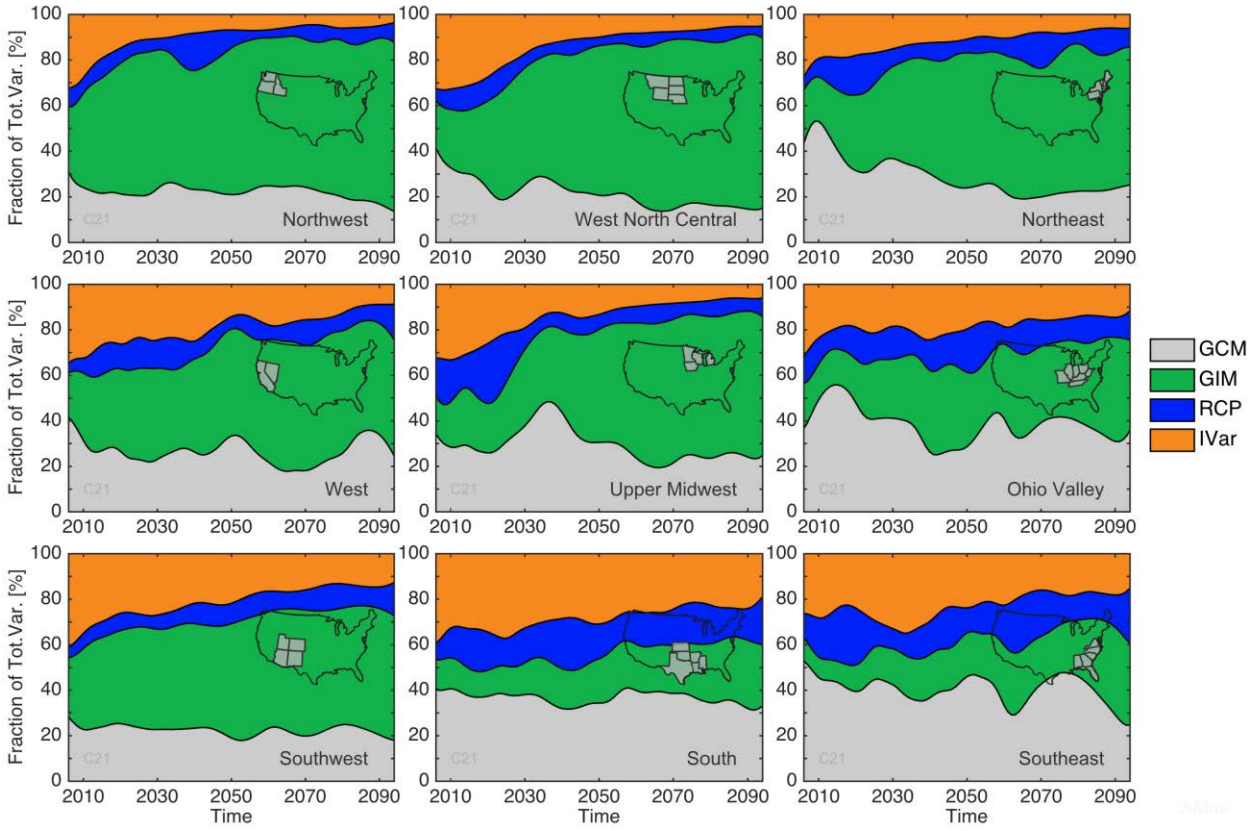


Figure A4.4 – Relative contributions to uncertainty in AMax: GCMs, GIMs, RCPs, and IVar for oE.

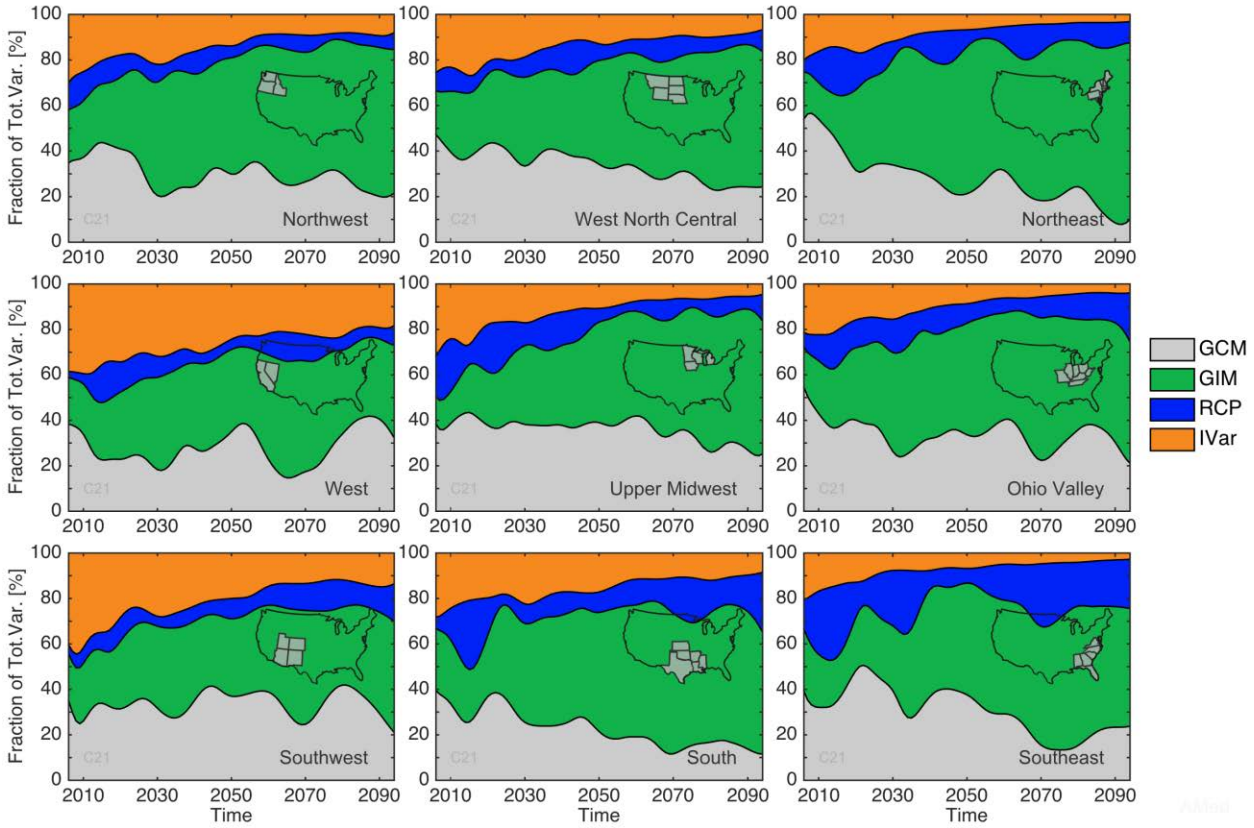


Figure A4.5– Same as Figure A4.4 but for AMed.

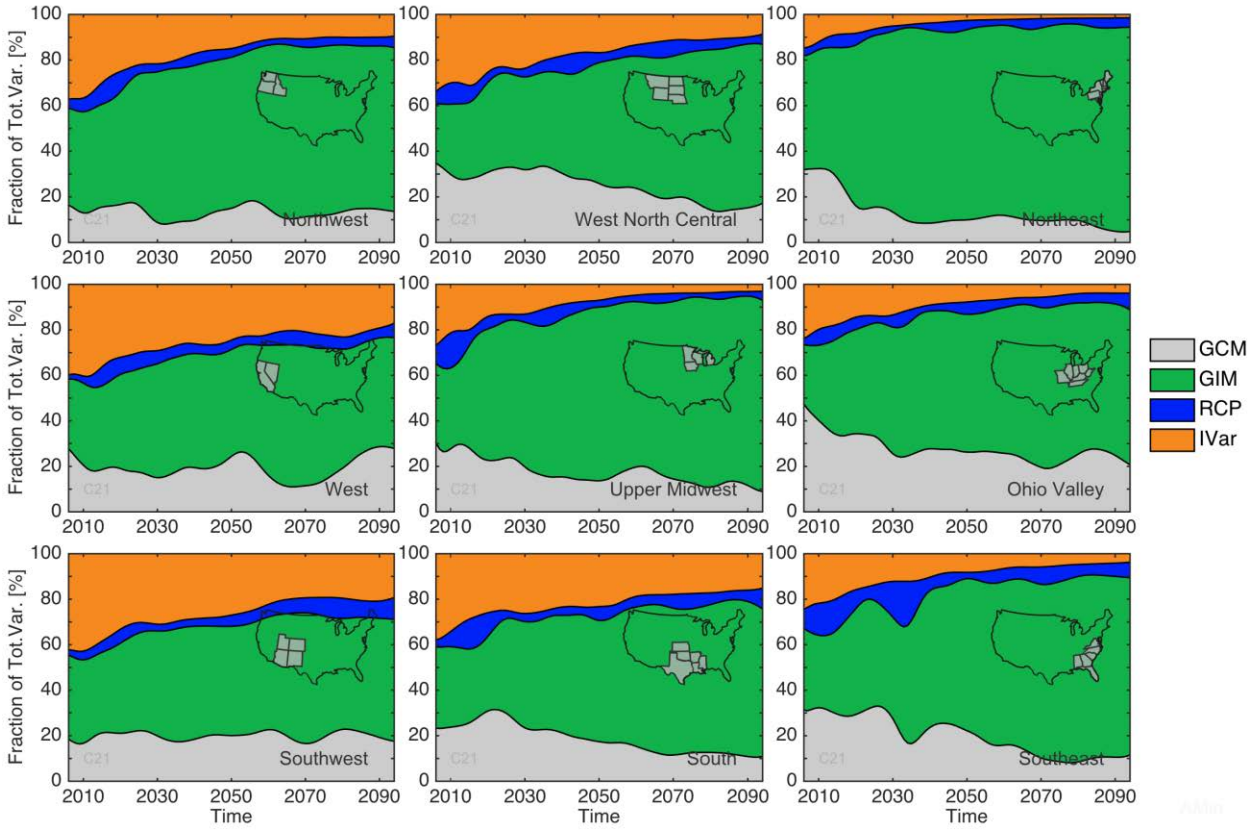


Figure A4.6 – Same as Figure A4.4 but for AMin.

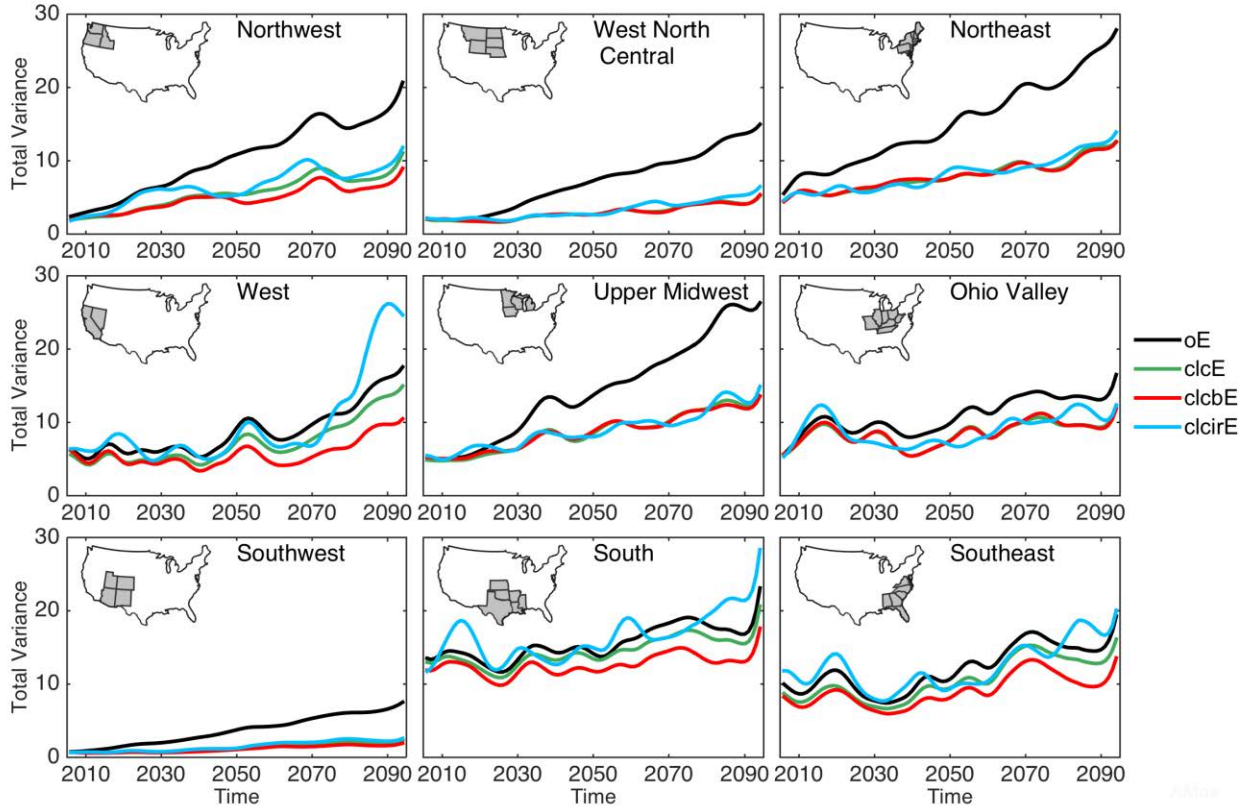


Figure A4.7 – Total variance increase in AMax per ensemble combination.

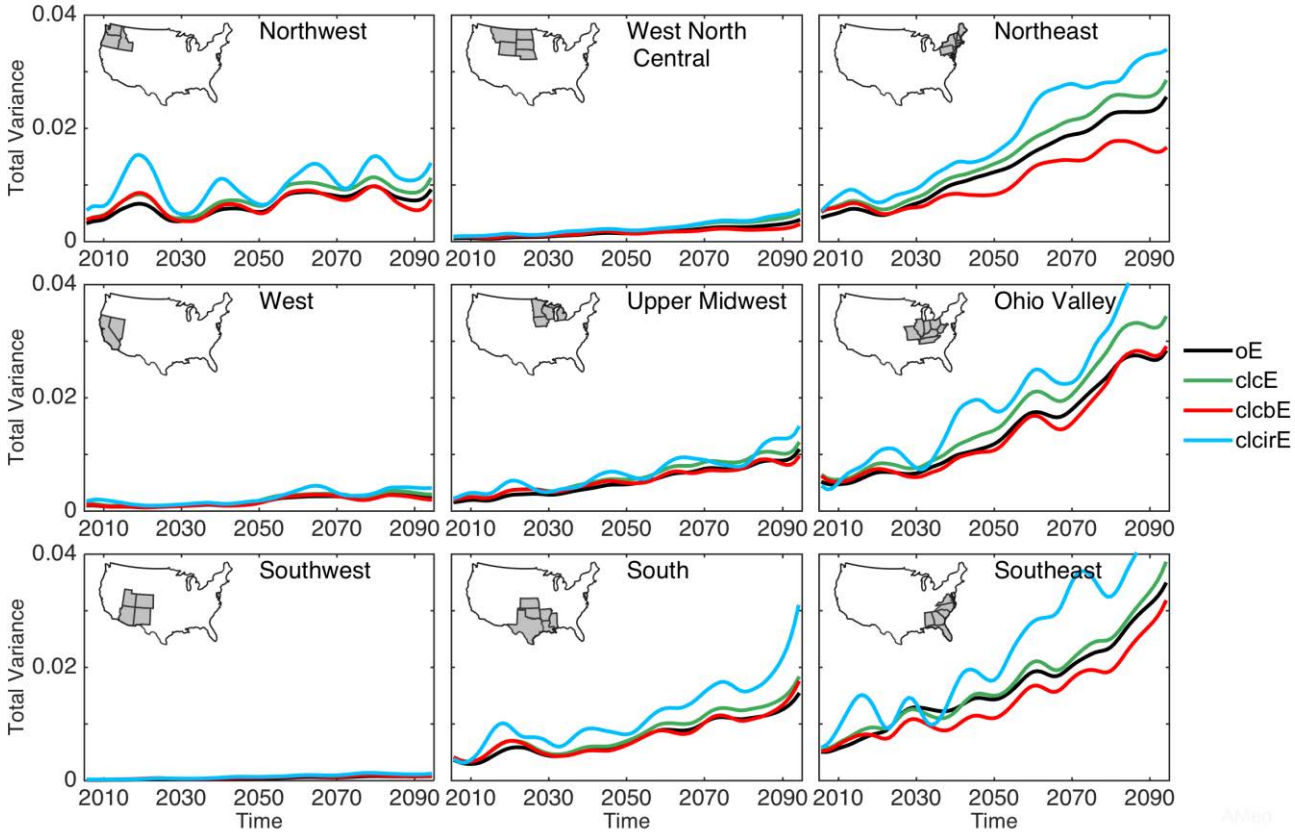


Figure A4.8 – Total variance increase in AMed per ensemble combination.

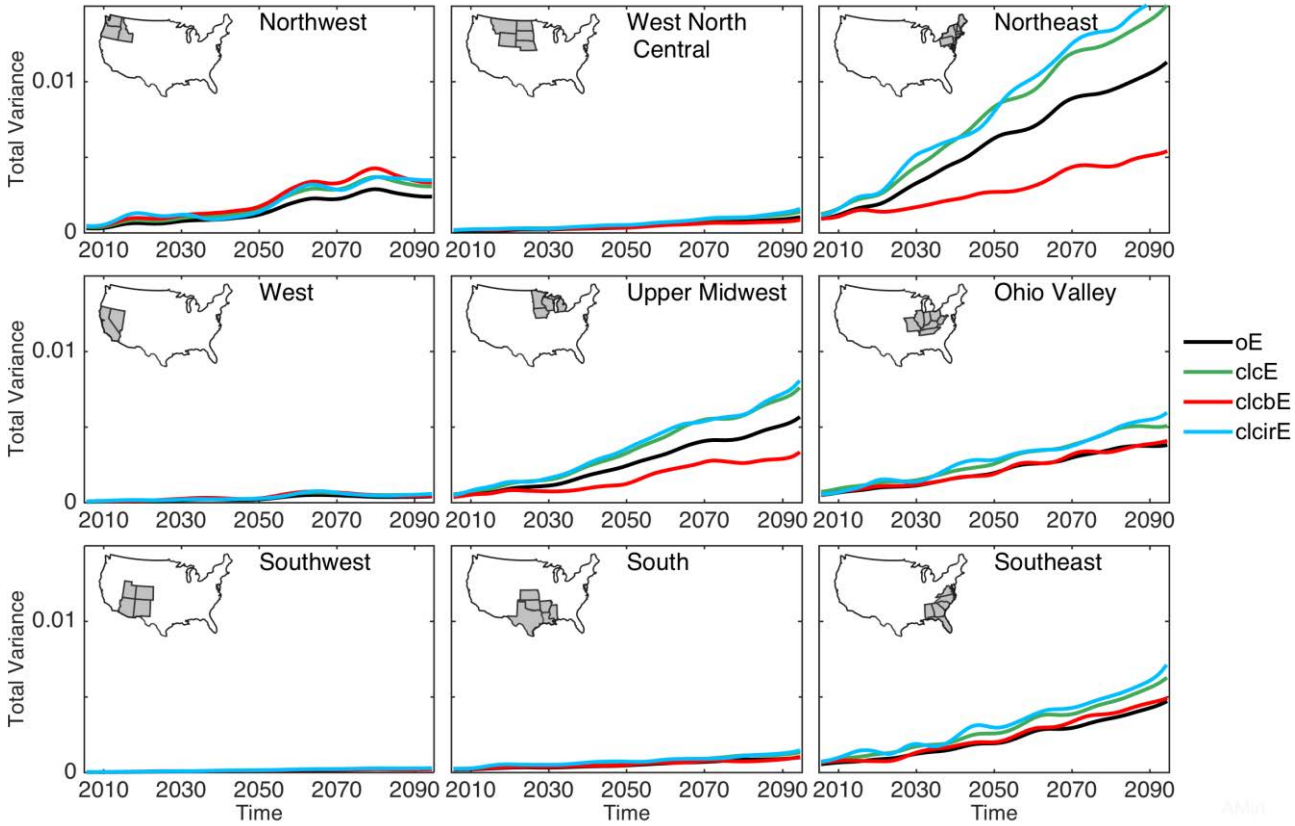


Figure A4.9 – Total variance increase in AMin per ensemble combination.

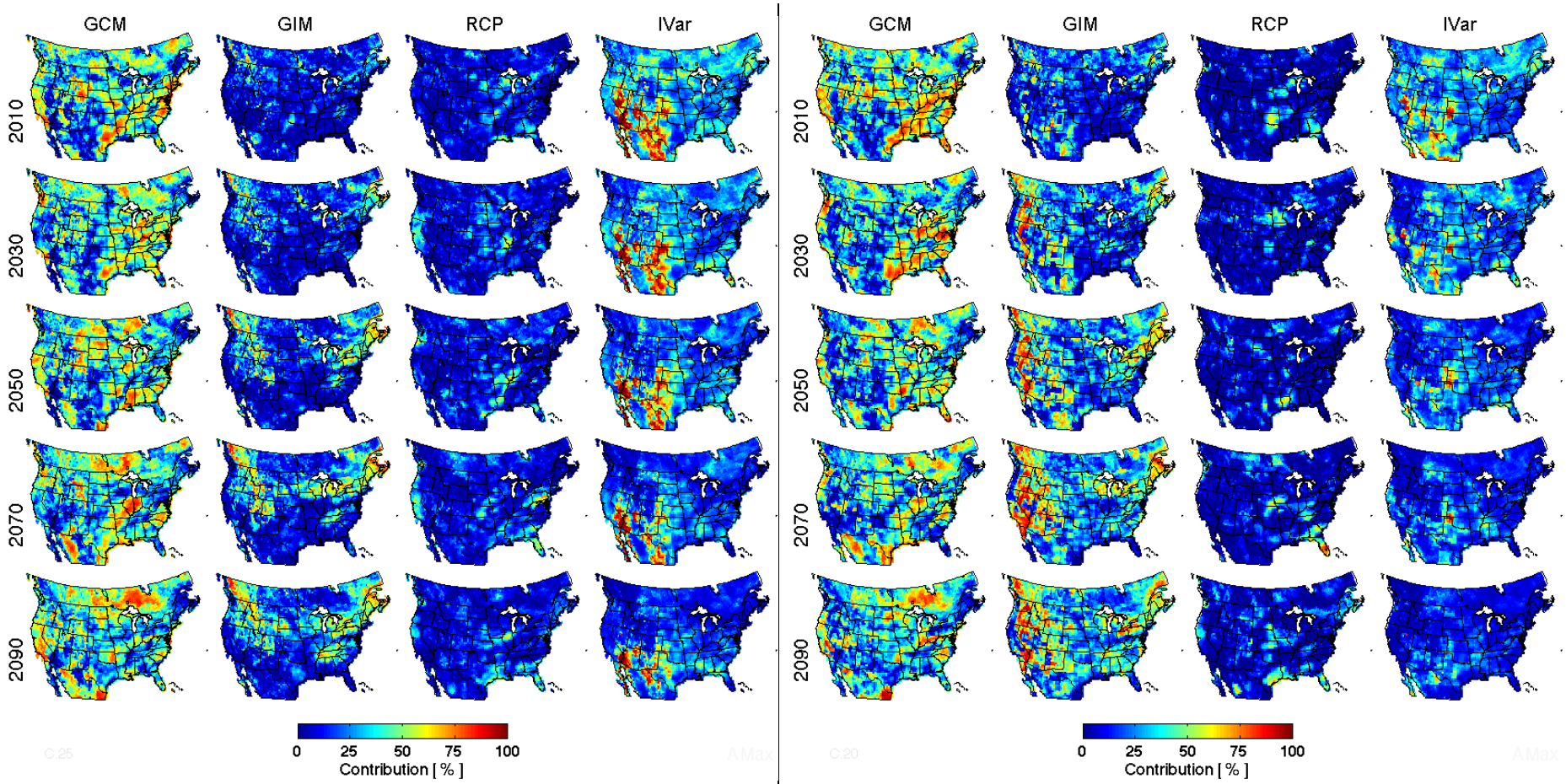
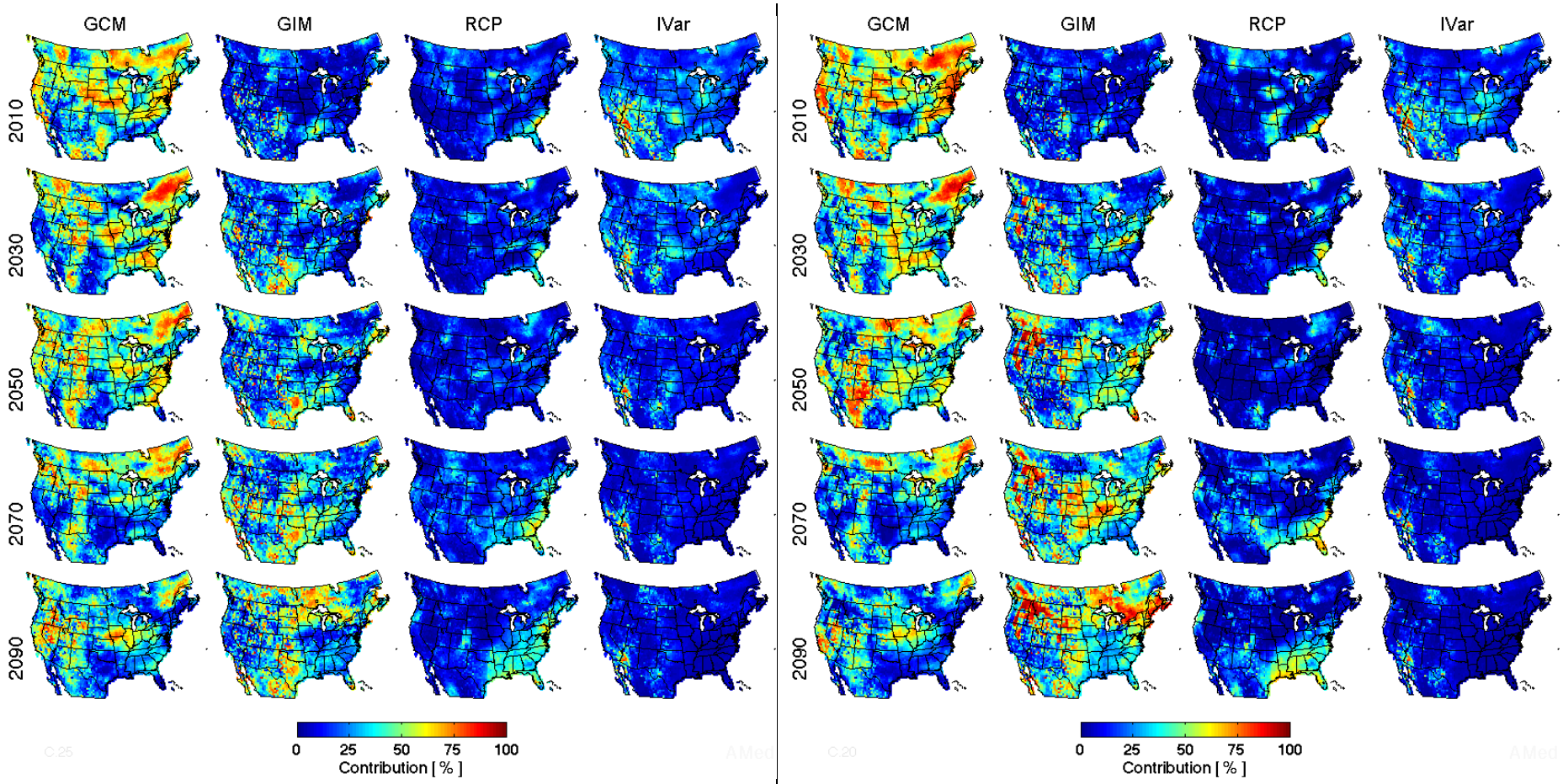


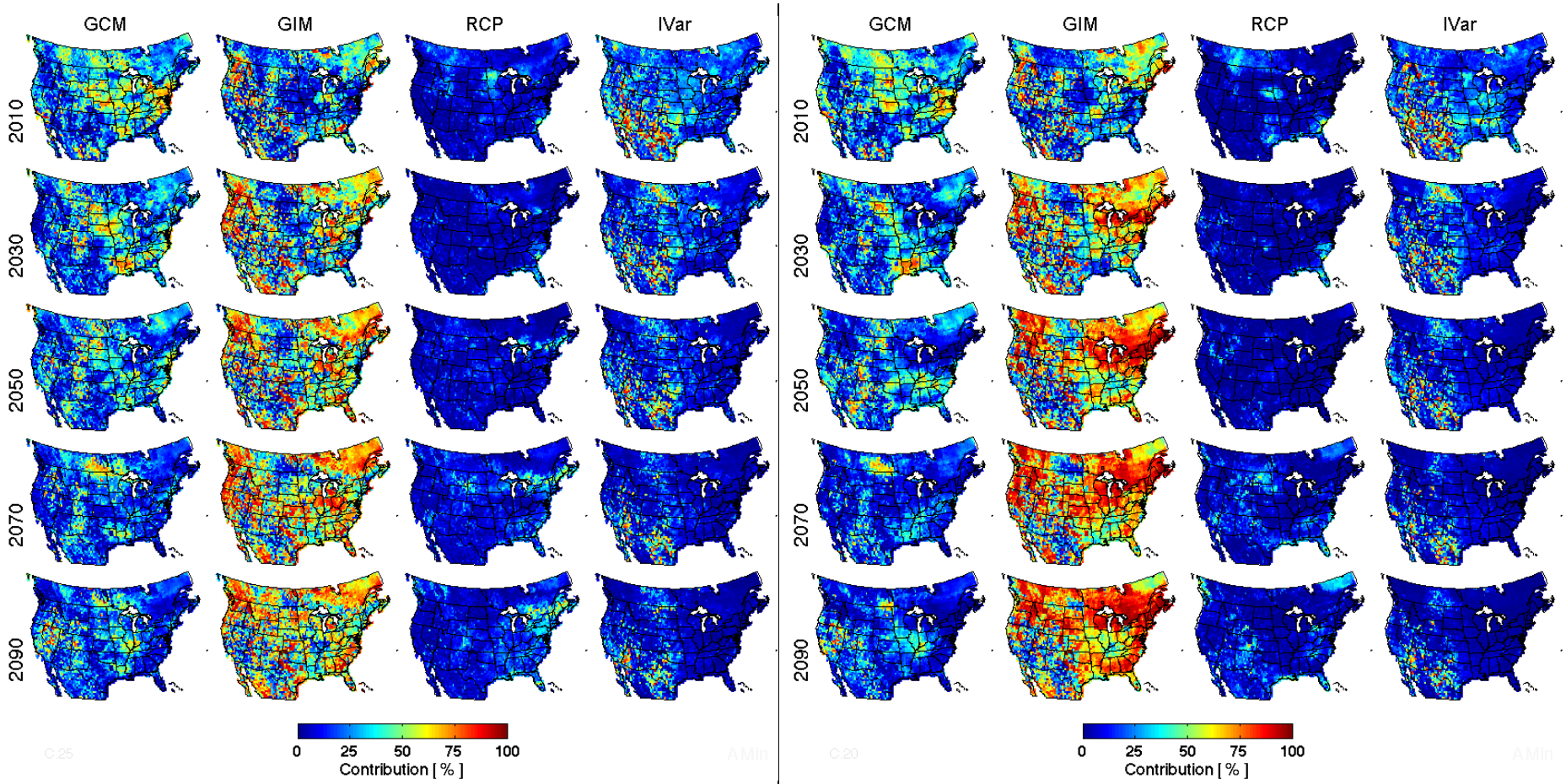
Figure A4.10 - Relative contribution (on 100%) to uncertainty in AMax of (from left to right): GCMs, GIMs, RCPs, and Internal variability. Results are shown for different time slices (in row): 2010, 2030, 2050, 2070, 2090. On the left: a) the ensemble excluding low credibility and biome GIMs (clcbE); on the right: b) ensemble excluding low credibility GIMs and RCP 4.5 and 6.0 (clcirE); (refer to Table Table 6.1).



a. [clcbE]

b. [clcirE]

Figure A4.11 - Same as Figure A4.10 but for AMed. Results are shown: a) for the ensemble excluding low credibility and biome GIMs (clcbE); b) ensemble excluding low credibility GIMs and RCP 4.5 and 6.0 (clcirE).



a. [clcbE]

b. [clcirE]

Figure A4.12 - Same as Figure A4.10 but for AMin. Results are shown: a) for the ensemble excluding low credibility and biome GIMs (clcbE); b) ensemble excluding low credibility GIMs and RCP 4.5 and 6.0 (clcirE).

APPENDIX V

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 Syst. Dyn. 6, 267–285. doi:10.5194/esd-6-267-2015

APPENDIX VI

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. J. Geophys. Res. Atmos. 120, 9138–9159. doi:10.1002/2015JD023401