Cross-scale intercomparison of climate change impacts simulated by regional and global hydrological models in eleven large river basins

F.F. Hattermann[1], V. Krysanova[1], S. Gosling[2], R. Dankers[8], P. Daggupati[19], Ch. Donnelly[6], M. Flörke[9], Sh. Huang[1], Yu. Motovilov[10] , Su Buda[16], T. Yang[17,18], C. Müller[1], G. Leng[11], Q. Tang[12], F.T. Portmann[13], S. Hagemann[7], D. Gerten[1], Y. Wada[3,4,5], Y. Masaki[14], T. Alemayehu[20], Y. Satoh[15], L. Samaniego[21]

1 Potsdam Institute for Climate Impact Research, Germany
2 School of Geography, University of Nottingham, UK
3 NASA Goddard Institute for Space Studies, USA
4 Center for Climate Systems Research, Columbia University, USA
5 Faculty of Geosciences, Utrecht University, The Netherlands
6 Swedens Meteorological and Hydrological Institute, Sweden
7 Max Planck Institute for Meteorology, Germany
8 Met Office, Exeter, UK
9 Center for Environmental Systems Research, University of Kassel, Germany
10 Water Problems Institute of Russian Academy of Science, Russia
11 Joint Global Change Research Institute, Pacific Northwest National Laboratory, USA
12 Key Laboratory of Water Cycle and Related Land Surface Processes, Institute of Geographic Sciences and Natural Resources Research, Chinese Ac. Sci., China
13 Johann Wolfgang Goethe-University Frankfurt am Main, Inst. of Phys. Geography, Germany
14 Center for Global Environmental Research, National Institute for Environmental Studies, Japan
15 International Inst. for Applied Systems Analysis, Austria
16 National Climate Center, China Meteorological Administration, China
17 Center for Global Change and Water Cycle, Hohai University, China
18 Xinjiang Institute of Ecology and Geography, Chinese Ac.Sci., China
19 Texas A&M University, USA
20 Vrije Universiteit Brussel, Belgium
21 Helmholtz Centre for Environmental Research, Germany

Abstract — Ideally, the results from models operating at different scales should agree in trend direction and magnitude of impacts under climate change. However, this implies that the sensitivity to climate variability and climate change is comparable for impact models designed for either scale. In this study, we compare hydrological changes simulated by 9 global and 9 regional hydrological models (HM) for 11 large river basins in all continents under reference and scenario conditions. The foci are on model validation runs, sensitivity of annual discharge to climate variability in the reference period, and sensitivity of the long-term average monthly seasonal dynamics to climate change. One major result is that the global models, mostly not calibrated against observations, often show a considerable bias in mean monthly discharge, whereas regional models show a much better reproduction of reference conditions. However, the sensitivity of the two HM ensembles to climate variability is in general similar. The simulated climate change impacts in terms of long-term average monthly dynamics evaluated for HM ensemble medians and spreads show that the medians are to a certain extent comparable in some cases, but with distinct differences in other cases, and the spreads related to global models are mostly notably larger. Summarizing, this implies that global HMs are useful tools when looking at large-scale impacts of climate change and variability, but whenever impacts for a specific river basin or region are of interest, e.g. for complex water management applications, the regional-scale models validated against observed discharge should be used.

Keywords — Hydrological impact models, global and regional scale, seasonal dynamics, ISI-MIP, WATCH, model inter-comparison.

---

1 Introduction

Climate change is a global phenomenon, but its impacts manifested at the regional scale (IPCC 2013). A global view on climate change impacts is important to quantify the aggregated effects, and developments at the global scale can influence driving forces in the region under study. The regional scale, on the other hand, is where most adaptation measures are planned and implemented and where interaction with affected stakeholders is most intense (Krysanova et al. 2005, Hattermann et al. 2011). Many insights into hydrological processes, impact pathways and adaptation options are only available at sufficient detail at the regional scale, but can be used to feedback into global assessments. As a result, both global and regional studies provide valuable information for decision-making and scientific understanding. The cross-scale interaction makes it important to bridge the scales in impact assessment and to compare the sensitivity of impact models of both scales to climate variability and change. Further, a comparison of regional and global hydrological models across a wide range of river basins provides a framework to test the consistency between the different scales of analysis and identifying a need for improvement.

Global hydrological models (Glob-HMs) are usually designed to supply consistent impact assessment for the continental and global scales. These models often compromise the model performance at the scale of individual catchments for the sake of overall model performance (Gosling and Arnell, 2011, Müller-Schmied et al. 2014). Regional or catchment-scale hydrological models (Cat-HMs) are typically more streamlined to the specific characteristics of the catchment under investigation, e.g. through local input data that better describe local conditions, calibration to observations and implementation of regionally important hydrological features such as wetland processes or water management (Koch et al. 2013, Hattermann et al. 2006). When looking at the specific model types and their inherent processes, there is no strict border between "purely" global and "purely" regional models. More and more hydrological features are implemented in global models, and model advancement and increase in computational power have led to the development that some global models are applied at the regional scale with higher resolution (e.g. WaterGAP3, Verzano 2009), while some regional models are applied at the continental scale (e.g. HYPE, Donnelly et al. 2015).

The way we distinguish the global and regional models in our study is that the former were applied for all continents with a spatial resolution of 0.5° without calibration (with the exception of WaterGAP2), while the regional models were applied for 11 large-scale river basins with a finer spatial resolution and were calibrated to observed discharge (see more details in Krysanova and Hattermann, this SI). In this study, we make use of global and regional HM output data uploaded in framework of the Inter-Sectoral Impact Model Intercomparison Project (ISI-MIP, Schellnhuber et al. 2014, Warszawski et al. 2014) (Table A1 I the Annex). ISI-MIP is a community-driven modelling effort bringing together impact modelers across sectors and scales to create consistent and comprehensive projections of impacts at different levels of global warming, based on the Representative Concentration Pathways (RCPs, van Vuuren et al. 2011) and Shared Socio-Economic Pathways (SSPs) scenarios (IPCC 2013).

In our study, we investigate the consistency of climate change impacts on the long-term average seasonal dynamics of discharge in 11 large-scale river basins (see Table A2 in Annex), covering the main climatic zones and hydrological regimes on all continents, using outputs of 9 Glob-HMs and 9 Cat-HMs. It was not possible to apply all regional models to all basins, because implementation of new model set-ups is work intensive and exceeded the capacity of the regional team.

To our knowledge, this is one of the first comprehensive cross-scale inter-comparisons of multiple hydrological models considering river basins on *all* continents, although there have beencross-scale model intercomparisons involving fewer models and basins (Gosling et al. 2011, Piniewski et al. 2014). Responses to climate change in hydrological extremes of the same HMs are reported in another cross-scale paper by Gosling et al. 2015 (this SI). The most recent comparison of Glob-HMs was conducted within the framework of ISI-MIP and described by Schewe et al. 2014, Dankers et al. 2014, Prudhomme et al. 2014, Haddeland et al. 2014, Davie et al. 2013, Wada et al. 2013 and Portmann et al. 2014. Model intercomparisons for the regional scale are described in Breuer et al. 2009, Bosshard et al. 2013, Chen et al. 2013 and Vetter et al. 2014.

This article first presents a comparison of the model validation runs for the reference period 1971-2000, using re-analysis climate data from the WATCH project (Weedon et al. 2011) as driving data, and observed discharge. Secondly, the comparison is extended to impacts under climate change scenarios until 2099.

2 Methods, models, river basins and climate data

2.1 Models

In total, outputs from 9 Glob-HMs and 9 Cat-HMs are considered in this study. Annex Table A1 lists the models and references where more information on them can be found. While the global models consistently simulate hydrological processes and river routing with a spatial resolution of 0.5°, different approaches are used by the regional models: regular grids (e.g. VIC and WaterGAP3) and disaggregation schemes with subbasins and hydrological response units (SWIM, HYPE and SWAT). More information on basic processes represented in the models is given in Annex Table A3. All models simulate the full water cycle, with daily precipitation and temperature as main inputs, calculation of evapotranspiration, infiltration, generation of runoff, and application of a routing scheme to transfer the locally generated runoff along the river network to the outlet. Some of the models include more processes such as lake dampening of flow, regulation of flow, wetlands and more.

Table A2 illustrates which hydrological models were applied in which of the eleven river basins. While the Glob-HMs provided outputs for each river basin, only a subset of Cat-HMs was applied in most cases (minimum four in the Upper Yangtze and Darling, a maximum of nine in the Rhine), due to the workload associated with model set-ups and calibration in catchments. More information about the regional models, the calibration process and the validation results can be found in Krysanova and Hattermann and in Huang et al. (this SI).

The Glob-HMs are operated at the same spatial resolution as the provided climate data (0.5°), whereas further model-specific interpolation of climate data to the subbasin scale was necessary to run the regional models. In addition, some of the regional models corrected precipitation and temperature during interpolation taking into account elevation.

2.2 River basins

Eleven river basins were selected for this cross-scale comparison to cover the most important climate zones and hydrological regimes worldwide. The map in Figure A1 shows their location, and Table A2 summarizes some of their characteristics (Annex). Two of them are located in temperate climate (Upper Mississippi and Rhine), one in Mediterranean climate (Tagus), one in subarctic climate (Lena), four in monsoonal climate (Ganges, Upper Amazon, Upper Niger, Blue Nile), two in continental plateau climate (Upper Yellow and Upper Yangtze) and one in dry temperate climate (Darling). More information about these river basins is given in Krysanova and Hattermann (this SI). The upper parts of several basins (Mississippi, Amazon, Yangtze, Yellow, Niger and Blue Nile) were chosen because they have no or minor influence of human management, thus making it possible to compare close-to-natural discharge and avoid consideration of complex water management affecting river discharge.

2.3 Climate data

To obtain a coherent impact model intercomparison, the models are driven by climate forcing data from the same source and for the same periods. For the analysis of model performance under current conditions, all models were forced by global WATCH Forcing Data (WFD), daily 0.5 by 0.5 degree gridded meteorological data covering the period 1958-2001 (Weedon et al. 2011). The CMIP5 climate scenario data (Taylor et al. 2012) used in this study were provided by ISI-MIP. Five Earth System Models (HadGEM2-ES, IPSL-CM5A-LR, MIROC-ESM-CHEM, GFDL-ESM2M, NorESM1-M) which have been bias-corrected using a trend-preserving method(Hempel et al. 2013), were applied. In this study, only the high-end scenario RCP8.5 was used. In most cases, it can be shown that the selected GCMs cover well the spread of GCM uncertainty in the specific region. For more i.e. information about the climate scenario simulations for the individual river basins including statistics about projected climate see Krysanova and Hattermann (this SI).

Other important input data for hydrological models are soil, land cover, elevation and hydrological information such as the river network. In most cases, they were taken from globally available data sources (see Table 2 in Krysanova and Hattermann, this SI), and some already existing regional-scale models used different spatial data they were originally implemented with. Observed discharge time series for the considered gauges were provided by the Global Runoff Data Centre (GRDC 2013) or country specific agencies.

3 Results

3.1 Model performance during the reference period

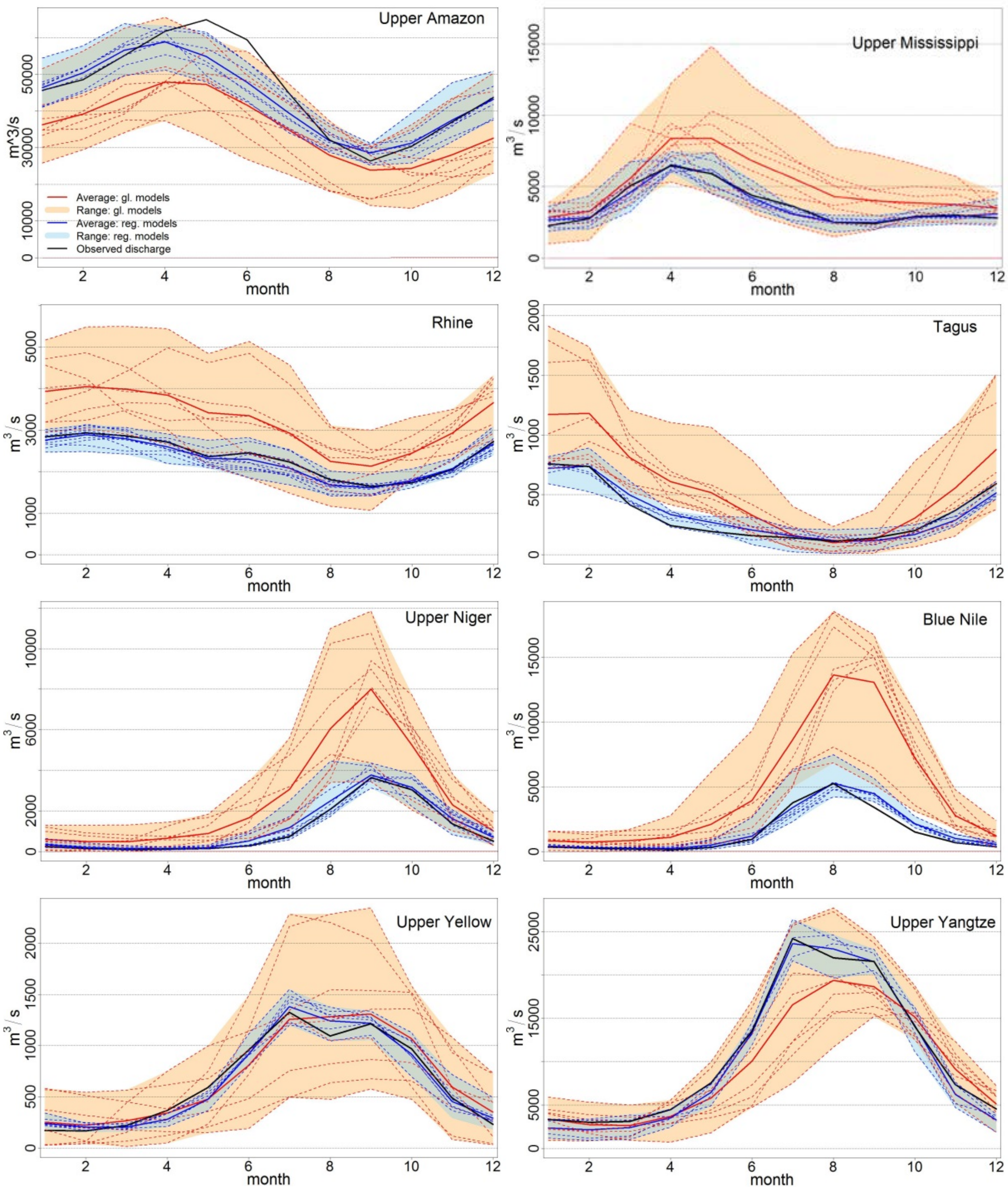
3.1.1 Comparison of simulated and observed discharges

The validation of the hydrological models was done for the gauging stations listed in Table A2 for the reference period 1971-2000. All Cat-HMs were calibrated against observed discharge and afterwards validated in a split sample mode, i.e. validating the model using discharge data of a time period different from calibration, normally with an 8-10 year period for calibration, depending on data availability. The Glob-HMs were not calibrated, except WaterGAP2 (which was calibrated against long-term average monthly discharge for a number of gauges worldwide) (see Krysanova and Hattermann, this SI).

Figure 1 visualizes the long-term average monthly seasonal dynamics of discharge for 1971-2001 simulated by Cat-HMs and Glob-HMs at the downstream gauges of the eleven basins, and Table 1 provides quantitative assessment. In general, Cat-HMs reproduce the observed long-term average seasonal dynamics of discharge well, with narrow ranges of uncertainty. This is partly so because the minimizing volume error is generally a calibration target of regional models. Results of the Glob-HMs in most cases show much higher uncertainty ranges in terms of deviation from the mean, and often a considerable bias towards observed data, mostly too high discharge, e.g. for the rivers Rhine, Tagus, Upper Mississippi, Upper Niger, Blue Nile, Ganges and Darling. In these cases, evapotranspiration is underestimated. The best performance of the mean of the nine Glob-HM results is for the Upper Yellow, followed by the Upper Yangtze, and Ganges (Figure 1).

The Darling is an extreme case, with a strong overestimation of the long-term average seasonal dynamics by Glob-HMs, while the Cat-HMs perform better but not as well as for the other basins (see also Table 1). A possible reason for the poor results in the Darling and in other arid and semi-arid climates may be the low runoff coefficient (i.e. the fraction of precipitation that reaches the basin outlet) because even a small underestimation of evapotranspiration (or overestimation of precipitation in the forcing) may lead to large overestimation of river discharge. Also, lots of unregulated and regulated water abstractions are reported for the Darling, including water harvesting (Kingsford 2000, Thoms and Sheldon 2000), which were not considered in the modelling

In the Upper Amazon, all models underestimate discharge in May and June, due to underestimation of precipitation in the rainy season in the driving WATCH ERA-40 data (Strauch et al., this SI). In the Lena, the inclusion of frozen soil is very likely to influence river discharge as it results in a higher runoff peak in the spring (Haddeland et al. 2011), a process considered only in the ECOMAG model and, by a static permafrost mask, in MPI-HM.



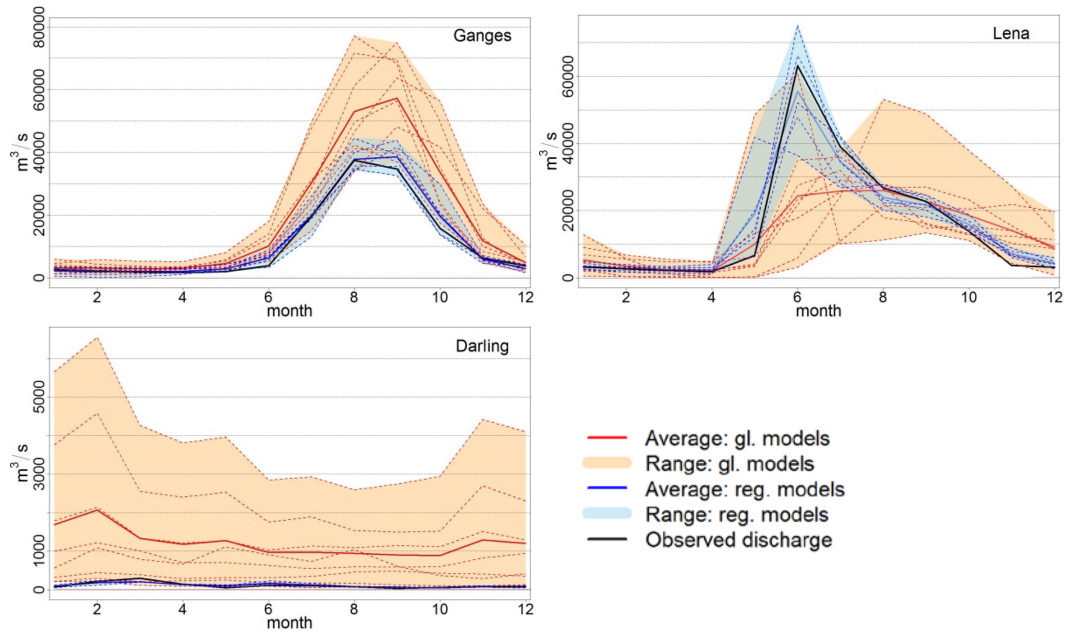


Figure 1: Comparison of the observed and simulated long-term average monthly seasonal dynamics of river discharge for 1971-2000 as modelled by the Cat-HMs and Glob-HMs in the selected 11 river basins.

Table 1: Performance of the global and regional models considering reproduction of the long-term average seasonal dynamics (monthly values) in the period 1971-2000 using WATCH data as climate input for 11 river basins. Indicators are the correlation coefficient r between simulated and observed monthly seasonal dynamics (r), and percent bias in standard deviation (Δσ) (see Equation 1 in Annex), both averaged over all models in columns 2, 3, 7 and 8, and d-factor as a measure of uncertainty in columns 6 and 11 (Abbapour et al., 2007). The percentage share of models with a moderate fit of r > 0.8 and Δσ < ±30% is shown in columns 4, 5, 9 and 10. Usually the thresholds r ≥ 0.9 and Δσ < ±15 % denote a good performance (Huang et al., this SI). The high average fit (r ≥ 0.9, Δσ < ±15%, d-factor<1) is indicated by shading.

	Cat-HMs					***Glob-HMs***				
Basin	*Average dynamics: corr. coef. r*	*Average dynamics: bias in STD Δσ*	*Share of models with r>0.8, in %*	*Share of models with Δσ:< 30, in %*	*d-factor*	*Average dynamics: corr. coef. r*	*Average dynamics: bias in STD Δσ*	*Share of models with r>0.8, in %*	*Share of models with Δσ:< 30, in %*	*d-factor*
Rhine	0.95*	1.9	100	78	1.08	0.87	68	88	50	4.60
Tagus	0.96	-5.4	100	60	0.75	0.91	67	100	50	2.86

Niger, Kouli-koro	0.96	7.3	100	100	0.72	0.89	116	75	13	2.59
Bl.Nile, El Deim	0.97	4.5	100	83	0.65	0.93	187	100	13	3.23
Lena	0.92	-10.6	80	100	0.51	0.61	66	38	0	1.23
U. Yellow	0.97	4.5	100	100	0.55	0.89	7.2	75	25	2.34
U. Yangtze	0.99	7.2	100	100	0.37	0.90	-16	88	50	0.99
Ganges	0.98	7.4	100	100	0.45	0.95	60	100	38	1.32
Darling	0.83	-29.5	50	50	0.68	0.34	431	0	38	47.2
U. Mississippi	0.92	2.0	88	88	1.13	0.80	59	50	25	3.61
U. Amazon	0.90	-16.5	83	100	0.89	0.87	-25	100	50	1.99

*with shading: r>0.9, Δσ< 15%, d-factor<1

Table 1 provides quantitative results of the model comparison shown in Figure 1. It summarizes the validation results in terms of two criteria of fit applied to the average seasonal dynamics from two model sets and to separate models: shown are the correlation coefficient (r) between the simulated and observed mean annual cycles of the years 1971-2000, and bias in standard deviation (Δσ). In addition the d-factor is added as a measure of uncertainty.

According to these thresholds, high correlation was found for 10 basins (all except the Darling) for means of Cat-HMs, but for only 4 out of 11 basins for means of Glob-HMs, and low bias in standard deviation was found in 9 cases for means of Cat-HMs, but only in one case for means of Glob-HMs. In addition, shares of regional and global models fulfilling the moderate thresholds of $r > 0.8$ and $\Delta\sigma < \pm 30$ % are given in Table 1. The values of d-factor below 1 denoting a low uncertainty related to observations (see Abbaspour et al., 2007) were found in 9 basins with Cat-HMs, but only in one case with Glob-HMs.

3.1.2 Sensitivity of modelled river discharge to climate variability

We investigated the sensitivity of discharge simulated by the Glob-HMs and Cat-HMs to climate variability by calculating the anomalies of annual precipitation and annual discharge for the reference period 1971-2000 and fitting outputs from the two model sets to a nonlinear regression (Figure 2). The anomalies are defined as the differences between the annual values for each year and the long-term average annual values over the period 1971-2000.

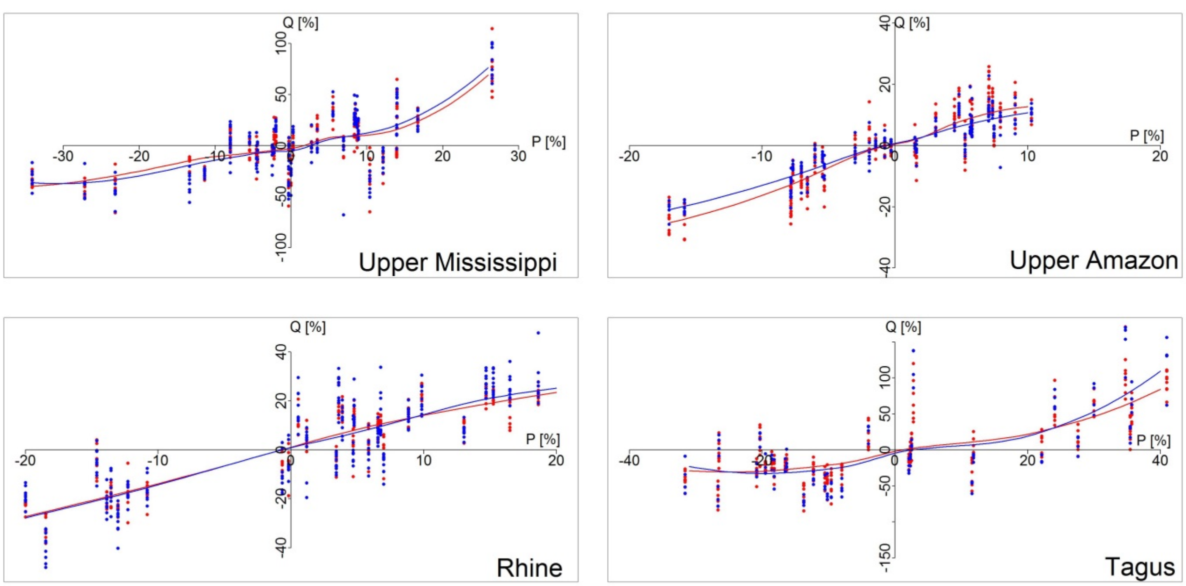
The lowest variability in precipitation was found for the Upper Amazon and Blue Nile basins with anomalies ranging from -20 % to +10. The largest variability in precipitation was found for the Darling and Tagus basins with annual precipitation anomalies ranging from -40 % to +40 %. The latter two are the driest regions considered, and they are consequently also the basins that show the highest variability in discharge (from less than -80 % to more than 150 % in the case of Tagus and from -100 % to more than

300 % in the case of Darling), proving their high vulnerability to climate variability. The lowest variability in discharge was found for the Upper Amazon and the Upper Yangtze (between -30 % to +30 %).

Relatively low variability in annual precipitation appears also in the basins of the Upper Niger, Upper Yellow, Upper Yangtze and Lena. Variability in discharge is low in the Upper Yangtze, while the Upper Mississippi shows relatively high variability in discharge. The correlation of changes in precipitation to changes in discharge has mostly close-to-linear character, only the Lena, Upper Mississippi, Tagus and Darling show more nonlinear responses (Figure 2). A positive anomaly in precipitation greater than 10 % usually produces a positive anomaly in discharge, but a smaller increase in rainfall may be associated with a decrease in discharge in single model runs (e.g. when in the specific application evapotranspiration increases more than precipitation).

The coefficient of determination R^2 of the fitted curves (see Table A4) is high for the Ganges, U. Niger, U. Amazon, Rhine and Blue Nile (both model types) in connection with their mostly high precipitation and runoff coefficients, and much lower for the Tagus, U. Mississippi, U. Yangtze and Darling. In general, there is no clear and distinct relation to the runoff coefficient, but interesting is that the single R^2 values of the two model sets are comparable.

A robust conclusion that can be drawn from Figure 2 is that no systematic differences in Glob-HM and Cat-HM sensitivities to climate variability can be observed, only the Darling River (where also the bias in discharge is highest for both model ensembles), as well as the Upper Niger and Blue Nile rivers show larger deviations.



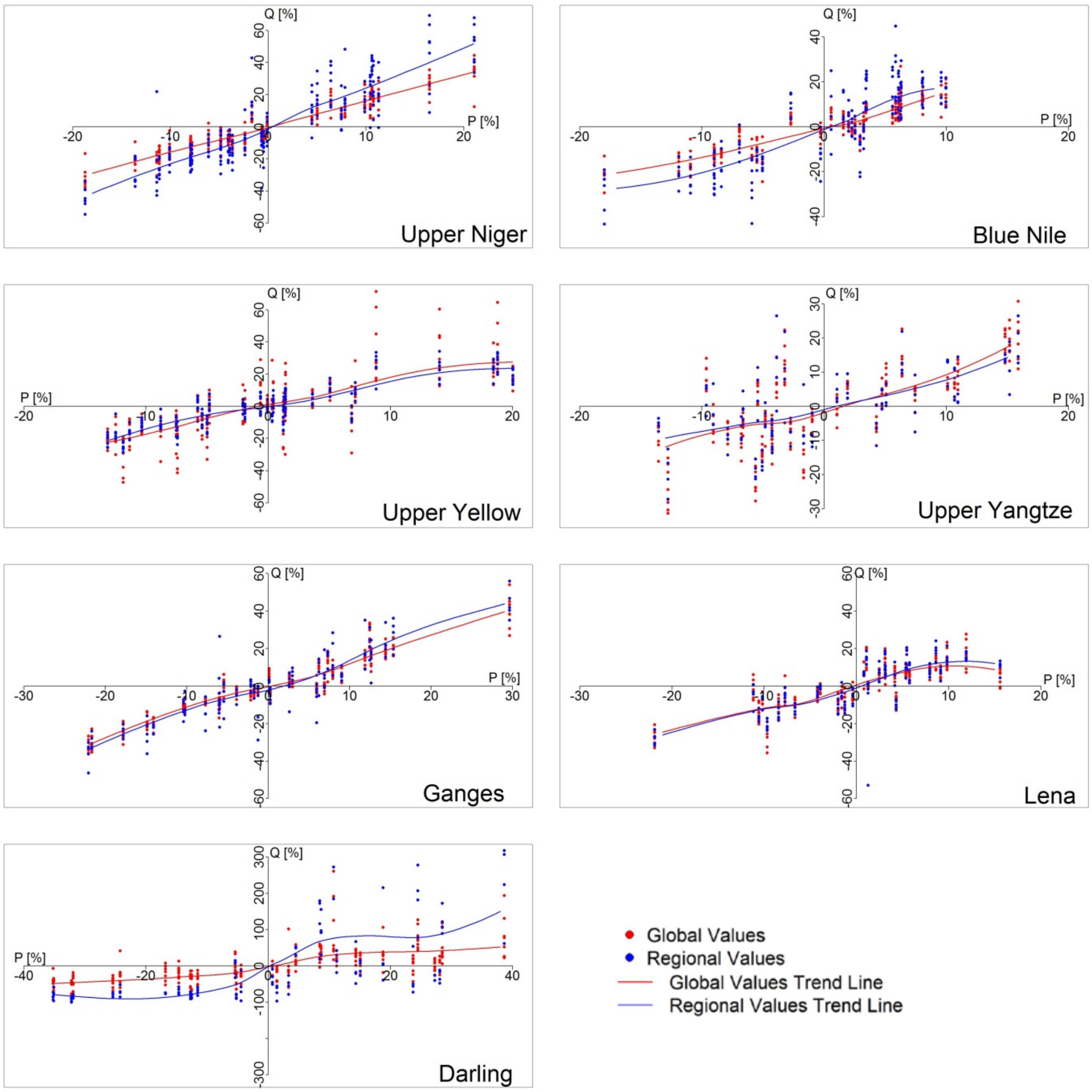


Figure 2: Sensitivity of annual discharge simulated by Glob-HMs and Cat-HMs to annual variability in precipitation for the 11 basins: anomalies in discharge (y-axis) versus anomalies in precipitation (x-axis) and the period 1971-2000 in percent. The lines were calculated using the LOESS technique, a nonparametric regression method that combines multiple regression models in a k-nearest-neighbor-based meta-model.

3.2 Climate change impacts on seasonal flows

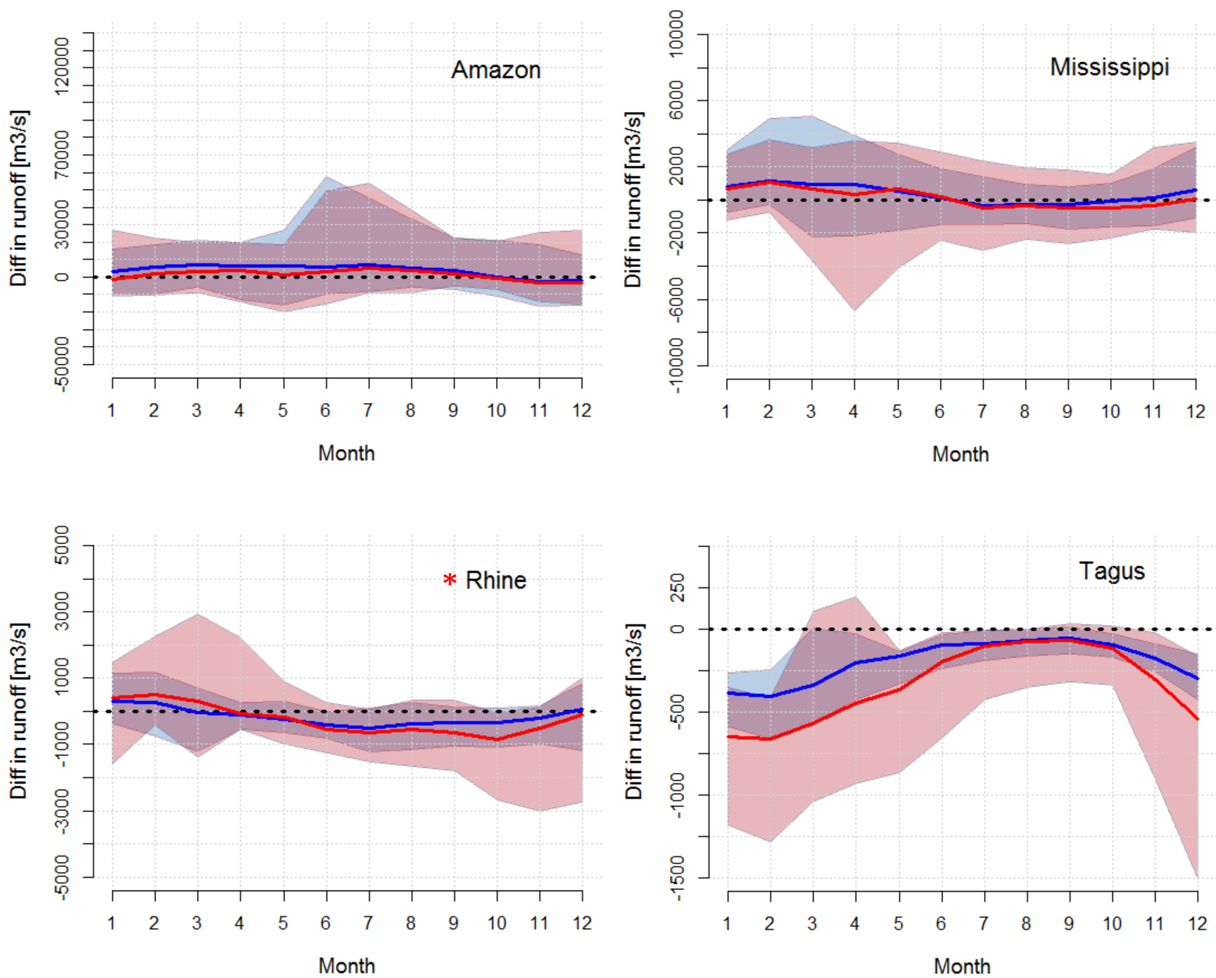
Comparison of the climate change impacts simulated by Glob-HMs and Cat-HMs was done for the high-end scenario RCP8.5 by comparing the *differences in long-term average monthly discharges* between the periods 2071-2099 and 1971-2000 in *terms of medians and uncertainty ranges from two HM sets* (Figures 3, A3 and Table 2).

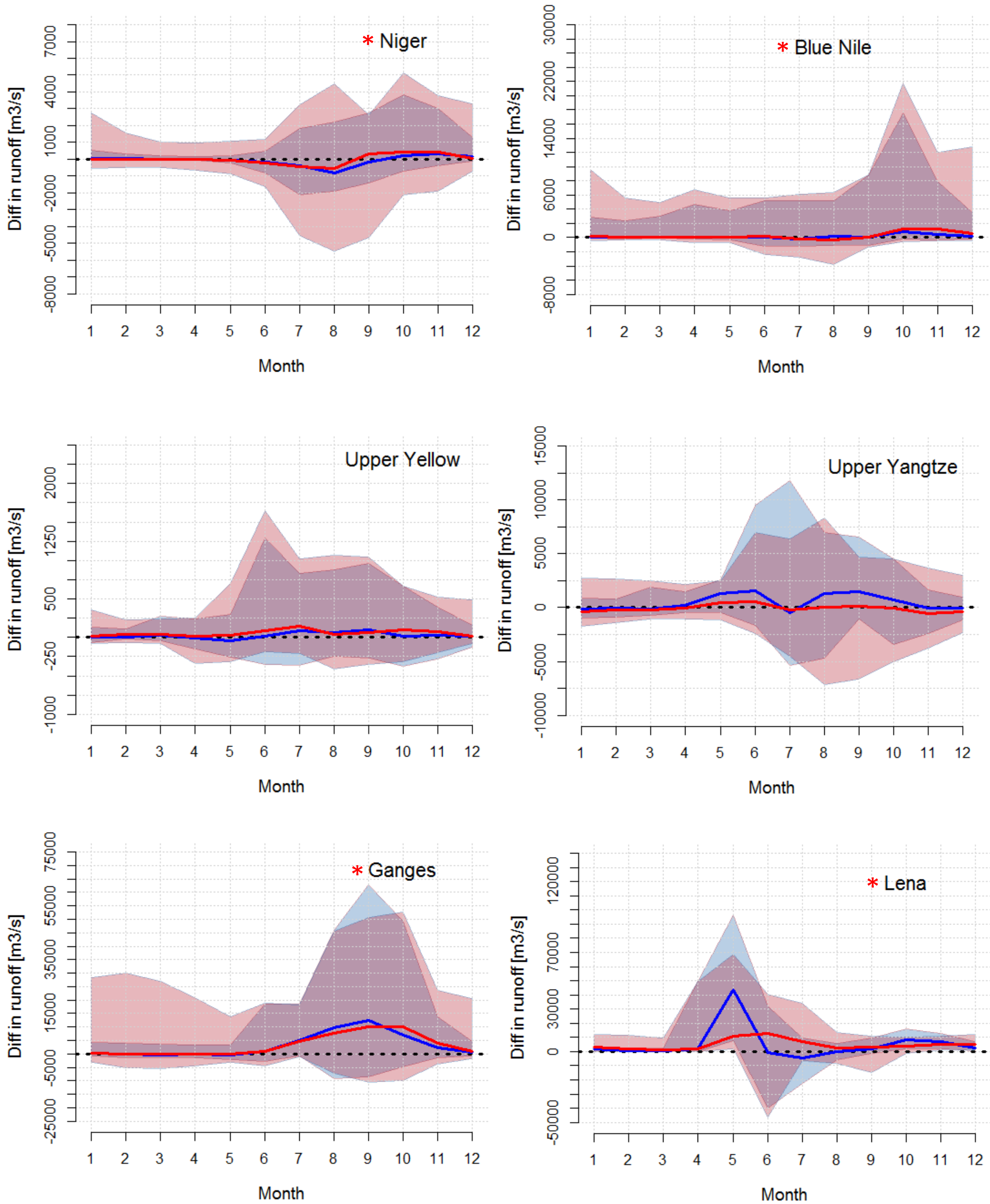
While temperature increases in all basins under scenario conditions, trends in precipitation are diverse (Krysanova and Hattermann, this SI). In general, the rivers showing the strongest overall decrease in mean seasonal discharge are the Tagus, Rhine and Darling, whereas increases are most pronounced for the Ganges, Lena and U. Amazon. The changes in medians without uncertainty ranges are shown additionally in Figure A3 in the Annex.

As one can see in Figure 3, similar to the model validation against observed discharge (Figure 1), the Glob-HMs mostly span much wider ranges, especially for the Tagus, U. Niger and Darling. The medians of the simulated changes of the two model ensembles are comparablefor the Lena and Ganges, but differ significantly in most of other cases (Figure A3). In some cases, for example the Mississippi, U. Niger and U. Yangtze Rivers, the uncertainty in changes from the Glob-HMs is very large compared to the average changes making it difficult to draw conclusions regarding the projected direction of changes and the comparability of both the two data sets. Therefore, a more formal analysis of similarity of the long-term average discharges from two ensemble results (Figure 3) was done using the non-parametric Wilcoxon signed-rank test, with a confidence level of 95% and in the two-sided mode. The hypothesis of similarity of the population mean ranks (i.e. of the signals of change) was confirmed in five cases (Rhine, Niger, Ganges, Mississippi and Lena) by this test.

In addition, the change signals in terms of means and medians (presented in Figure 3) as well as spreads and spreads related to means were estimated (Table 2, columns 2 - 9) and analyzed. The last two columns provide a qualitative estimation of similarity. As we see from this table, the means and medians are well comparable for the Ganges and Lena (though the shapes of seasonal dynamics for the Lena are different, Figure 3), and the differences are not large for the Rhine and Blue Nile. For the remaining seven basins differences are higher than 70%, and in three cases they are very high (U. Niger, U. Yangtze and Darling). The spreads from Glob-HM simulations are higher than those from Cat-HMs in 10 cases of 11. The spreads are well comparable in four cases: for the Lena, U. Amazon and two Chinese basins, and in four cases, the spreads from Glob-HMs are moderately (33 - 79%) larger. In three cases (Tagus, U. Niger and Darling) the spreads from Glob-HMs are more than doubled compared to spreads from Cat-HMs.

It is important to mention that the large uncertainty ranges in Figure 3 are the combined effects of global climate model and hydrological model uncertainty. The uncertainty related only to HMs can be seen in Figure A4, where results driven only by one climate model, GFDL, are presented as partial results from Figure 3. The evaluation of means, medians and spreads for this figure is included in Table A5, confirming that the uncertainty (spread related to mean) corresponding to Glob-HMs is significantly larger than that corresponding to Cat-HMs in most cases.




* Niger
* Blue Nile
Upper Yellow
Upper Yangtze
* Ganges
* Lena
Diff in runoff [m3/s]
Month

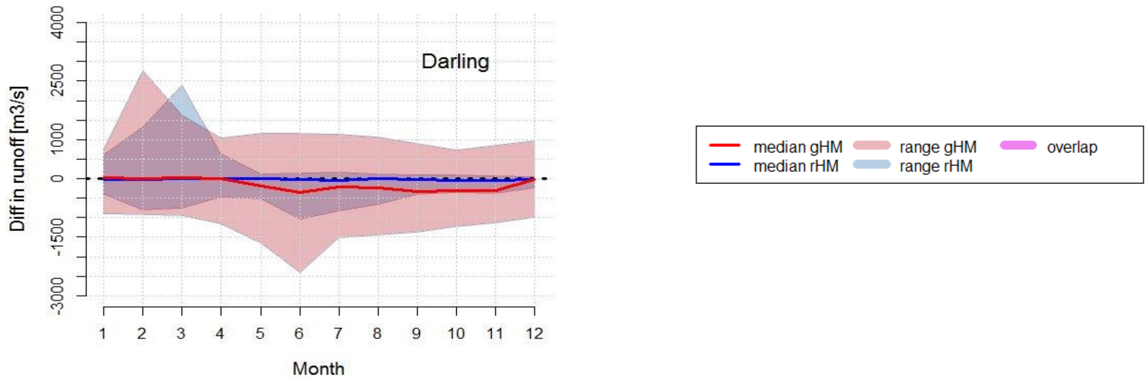


Figure 3: Comparison of climate change impacts on the long-term average monthly discharge modelled by the Glob-HMs and by Cat-HMs driven by 5 GCMs (scenario RCP8.5) for the period 2071-2099 compared to the reference period 1971-2000. Red stars indicate that the medians are not distinguishable with the confidence level of 95 % following the two-sided Wilcoxon signed-rank test.

Table 2. Comparison of differences in seasonal dynamics of discharge between end of the century and reference period (RCP 8.5) simulated by Glob-HMs and Cat-HMs (as in Figure 3) in terms of annual means, medians and spreads. The change signals (columns 2, 3, 6, 7 are calculated by averaging 12 values of the long-term mean or median values). The signs in last two columns: ++ similar (differ < 25%); + quite similar (differ < 40%); +/- moderately different (difference 40-70%); --- different (difference > 70%.

	Global Models				Regional Models				Comparison of means, medians and spreads				
	change signal (seas. mean)	change signal (seas. median)	average spread	av. spread / mean	change signal (seas. mean)	change signal (seas. median)	average spread	av. spread / mean	difference in means (abs. values, without sign)	difference in medians (abs. values, without sign)	spread(Glob-HM)/ spread(Cat-HM) (in %)	similarity of means & medians	similarity of spreads
Rhine	-278	-237	2067	7.4	-169	-164	1154	6.8	Glob-HM: 64% higher (neg)	Glob-HM: 45% higher (neg)	179	+/-	---
Tagus	-366	-341	652	1.8	-205	-196	244	1.2	Glob-HM: 79% higher (neg)	Glob-HM: 74% lower (neg)	267	---	---
U. Niger	25	-5	3841	153	137	-66	1746	12.7	Cat-HM: 5.5 times larger (pos)	Cat-HM: 13 times larger (neg)	220	---	---
Blue Nile	1131	274	7763	6.9	838	159	5413	6.4	Glob-HM: 35% higher (pos)	Glob-HM: 72% higher (pos)	143	+/-	+/-
Ganges	5101	3206	21789	4.3	4161	3132	16373	3.9	Glob-HM: 23% higher (pos)	Glob-HM: 2% higher (pos)	133	++	+
Lena	5923	4934	24365	4.1	6211	5239	21492	3.4	Cat-HM: 5% higher (pos)	Cat-HM: 6% higher (pos)	113	++	++
U. Yangtze	9	-73	5861	651	813	487	5253	6.4	Cat-HM: 116 times higher	Cat-HM and Glob-HM: diff. signs	112	---	++
U. Yellow	88	53	648	7.4	61	19	681	11.1	Glob-HM: 44% higher (pos)	Glob-HM: 2.8 times larger (pos)	95	---	++
Darling	-196	-161	2084	10.7	-45	-21	854	18.9	Glob-HM: 4.4 times larger (neg)	Glob-HM: 7.7 times larger (neg)	244	---	---
U. Mississippi	25	126	4859	194	405	353	3324	8.2	Cat-HM: 16 times larger (pos)	Cat-HM: 180% higher (pos)	146	---	+/-
U. Amazon	2928	1311	33640	11.5	5271	3849	32051	6.1	Cat-HM: 80% higher (pos)	Cat-HM: 2.9 times larger	105	---	++

4 Discussion

The results suggest that the model sensitivity of Cat-HMs and Glob-HMs to climate variability is comparable in most cases (Figure 2). When looking at the spreads of the annual discharges for each model set in Figure 3, both have similar ranges, indicating that the model sensitivity of both ensembles to climate variability is comparable and not altered by calibration of the Cat-HMs.

The comparable model sensitivity is in contrast to the fact that the long-term average seasonal dynamics of discharge simulated by the Glob-HMs often show large biases compared to the observed values when driven with WATCH data for the reference period 1971-2000 (Figure 1, Table 1). For a more detailed look and statistical analysis of the behavior of the single models in different basins see Huang et al., Eisner et al. and Vetter et al., all in this special issue.

When looking at climate change impacts, the highly aggregated outputs such as the long-term monthly averages of the two model sets show visually comparable shapes for many of the 11 catchments (Figure 3) with large uncertainty bounds stemming from GCMs and HMs. The changes in simulated seasonal discharge in general are mainly the result of changes in precipitation (where the input is the same for both model sets) and changes in evapotranspiration, where models from both scales often use the same equations to calculate potential evapotranspiration (see Table A3). Therefore, it is not very surprising that the aggregated long-term average monthly impacts are comparable. However, a more formal statistical analysis of similarity of the long-term average discharges (signals of change in discharge and empirical distributions) from the two ensembles confirms the hypothesis of similarity by the Wilcoxon test only in five cases of eleven (Figure 3). Also the analysis of differences in means, medians and spreads (Table 2) reveals many differences between of two HM ensembles.

While the focus here was on absolute changes in discharge, for many applications it might be sufficient to evaluate relative changes only (Schewe et al. 2014), or in the case of floods or droughts to use extreme value statistics (Feyen et al. 2012, Hattermann et al. 2014, Gosling et al. 2015 (this SI)). Figure A2 (Annex) shows the relative changes in discharge under climate change for three basins where the absolute results of the two model sets showed stronger differences, the Mississippi, Yangtze and Darling. Especially for the Darling and Mississippi, the similarity of results from the two model ensembles strongly increases.

The results presented here generally support those ones from an earlier multi-scale hydrological model intercomparison (Gosling et al. 2011), which showed that Glob-HMs can be useful tools for understanding catchment-scale hydrological responses to climate change, if *mean impacts* on annual flows, sign of change, or the seasonal cycle are of interest. However, the fact that ensemble medians of both model sets tend to be comparable while single models (especially the global ones) often generate high uncertainty ranges proves that there is a real benefit in using a multi-model ensemble, as also reported in previous studies (e.g. Hagemann et al. 2013). In addition, this allows the model-related uncertainty to be quantified.

Under climate scenario conditions, the range of uncertainty in the Glob-HM results is mostly higher than that in the Cat-HM results, and in some cases much higher (e.g. in the Tagus, U. Niger and Darling). The larger uncertainty in absolute values is probably the result of the large biases during the reference period (Darling). Most practitioners would certainly prefer a lower uncertainty in scenario results, while it might be of interest in some cases to screen a larger range of uncertainty, for example when planning sensitive infrastructure in riverine areas. Generally, models overestimating runoff by far during the reference period will likely do so also under climate change conditions.

In most cases, when simulated water components are used in subsequent management applications, accuracy of the data is important, for example in the case of water availability per capita, hydropower production, flood protection and crop production. In these cases, data of uncalibrated models should be used with care. For water resources applications, changes in many components of the water cycle also within the catchment may be equally important, and in this case a multi-site and multi-criteria validation is necessary (Hattermann et al. 2005). However, in some cases, the good model performance we observed for the Cat-HMs could be a sign of over-calibration, e.g. where hydrological processes are influenced by management which was not included.

Calibration of hydrological models is complex and the stability of calibrated parameters over time (and into the future) may be questionable and is under discussion (Merz et al. 2011). However, the fact that a model can respond to the climatic variability within the historical period lends some more trust to the projections using this specific model.

5 Conclusions

Our study is, to our knowledge, one of the first comprehensive cross-scale hydrological model intercomparisons, applying 9 global and 9 regional hydrological models in 11 large scale river basins. Some of the results were as to be expected: Glob-HMs, mostly uncalibrated, often show a large bias in the long-term average seasonal discharge when results are compared against observations, although they do in many cases reproduce the intra-annual variability well. More surprising is the fact that the sensitivity of models of both scales to climate variability (evaluated for model ensembles) is quite similar in most basins. The simulated climate change impacts in terms of long-term average monthly dynamics evaluated for HM ensemble medians and spreads show that the medians are to a certain extent comparable in some basins – but with distinct differences in others, and the spreads related to global models are mostly notably larger. The hypothesis of similarity of the long-term average signals of change from the two ensembles is confirmed by Wilcoxon tests only in five cases out of eleven.

This study was limited to analysis of river discharge at the outlet of large scale river basins, an indicator for changes in the water balance of large regions. In future studies, it would be good to have a more balanced number of models from the global and regional scales. In follow-up investigations, more attention should be given to improving performance of global models, including spatially-distributed calibration of regional models, analysis of other components of the water cycle, and also to other sources of uncertainty in scenario analysis, such as the emission scenarios and the driving global climate models.

6 Acknowledgements

This work has been conducted under the framework of ISI-MIP (funded by the German Federal Ministry of Education and Research (BMBF) with project funding reference number 01LS1201A). Responsibility for the content of this publication lies with the authors. We acknowledge the World Climate Research Programme's Working Group on Coupled Modeling, which is responsible for CMIP, and we thank the respective climate modelling groups for producing and making available their model output. We also acknowledge the support of the Global Runoff Data Center and of the Technology Development Fund (S-10) of the Ministry of the Environment, Japan.

7 References

(See additional file)