Supplementary Sections and Text

Supplementary Section A. Case Studies: Effect of CaMa-Flood routing in different river basins

Here we examine closely on the hydrographs based on two GHMs, LPJmL and MATSIRO, which are representative of the large dispersion of performances revealed in Tables S5. This is done for three different river basins (Amazon, Mekong and Ob), which feature very different terrain and climate characteristics. For the two GHMs we evaluated the performance of their original discharge output and their runoff-driven discharge simulated by CaMa-Flood. Both GHMs employ a linear reservoir routing model with constant flow speed, but MATSIRO also explicitly simulates groundwater dynamics which can cause certain delay in the generation of subsurface runoff. Figures S1-S3 displays multi-year observed and simulated (by GHM and CaMa-Flood) daily discharges for the three basins. In all cases for these two GHMs, CaMa-Flood resulted in smaller amplitude and a delayed timing for peak discharge, likely due to its floodplain expansion mechanism. Note that this is not the case for the two GHMs using a routing scheme featuring a strong wetland mechanism (MPI-HM, ORCHIDEE); comparison for all nine GHMs at the Amazon basin is given in Figure S10.

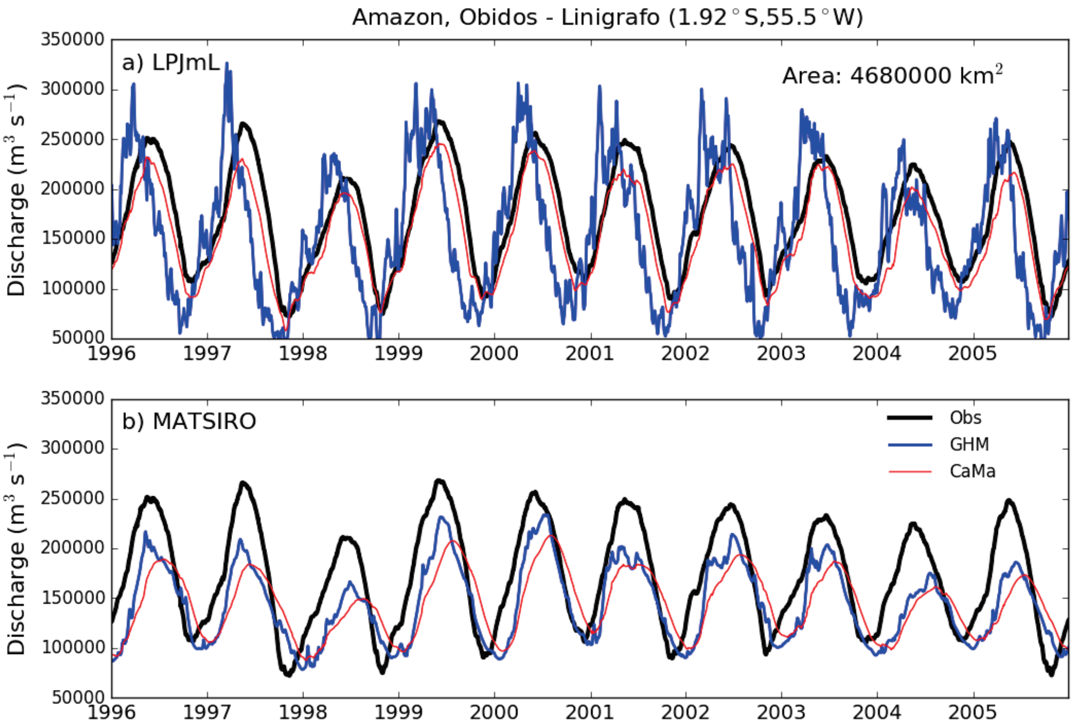

Figure S1. Observed (black), GHM simulated (blue) and CaMa-Flood (red) simulated daily river discharges at Obidos-Linigrafo, Amazon during 1995-2005, for a) LPJmL and b) MATSIRO.

In the case of Amazon, where the terrain is quite flat, the floodplain module in CaMa-Flood seems to be the main contribution to improved simulation for LPJmL, both in terms of the timing and amplitude of peak. For MATSIRO, however, its groundwater scheme substantially delays the timing of peak with certain reduction in the amplitude (Koirala *et al* 2014), CaMa-Flood tends to further amplify such delay mechanism, resulting in a relatively worse performance (Figure S1). This implies that the effect of floodplain dynamics may have exerted a similar buffering effect on river discharge as the groundwater scheme. A more realistic routing scheme should represent both mechanisms in order to avoid error overcompensations. In the case of Mekong where the terrain is relatively steep, LPJmL overestimates the amplitude of discharge, and features an earlier than observed flooding season; CaMa-Flood improves both aspects of the simulated discharge. Additionally, the high frequency variation in the original LPJmL discharge seems higher than observed, whereas such variance becomes lower than observed with CaMa-Flood. For MATSIRO, the GHM and CaMa-Flood simulated discharges are very similar, and the high frequency variation is better captured by the native routing in MATSIRO (Figure S2), which again could be due to the groundwater scheme that, in general, produces a smooth hydrograph compared to the models without groundwater representation. For the boreal Ob river basin, CaMa-Flood also significantly improved the amplitude and timing of LPJmL's discharge simulation. For MATSIRO, while the original amplitude is too large, the CaMa-Flood simulated amplitude is on the small side, although the magnitude of amplitude bias is reduced; the timing is not improved given that MATSIRO already simulates the timing of peak discharge well (Figure S3). In all three basins, the low flow simulations for LPJmL are improved with CaMa-Flood routing. Comparison for all nine GHMs at the Mekong and Ob basins are given in Figure S11 and S12.

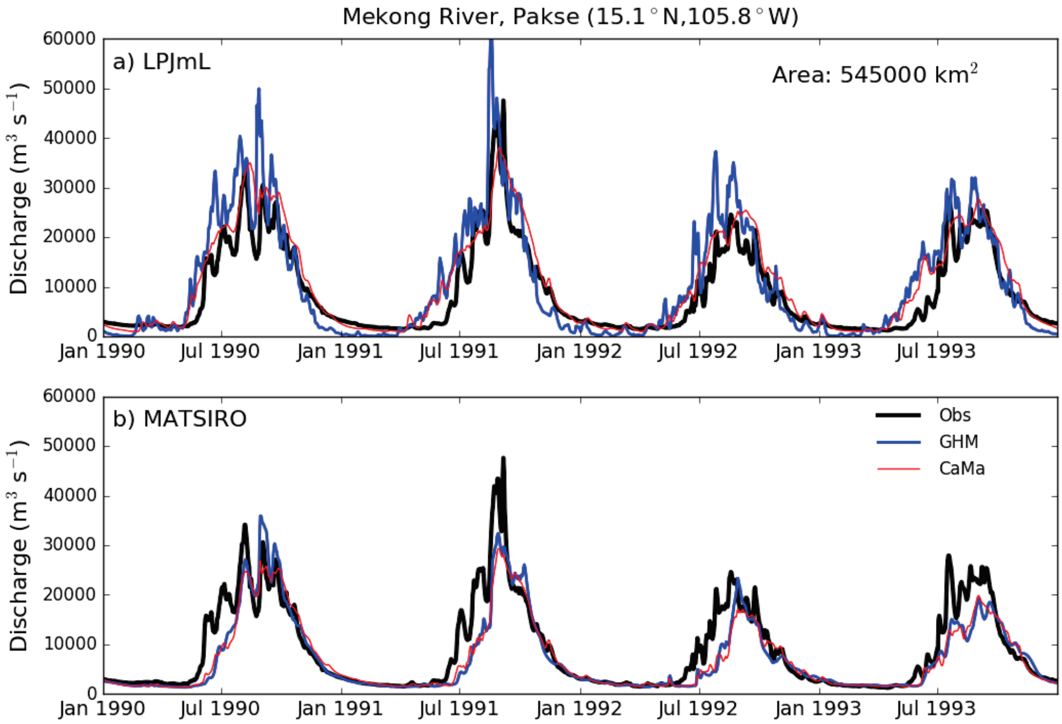

Figure S2. Same as Figure S1 but at Pakse, Mekong during 1990-1993.

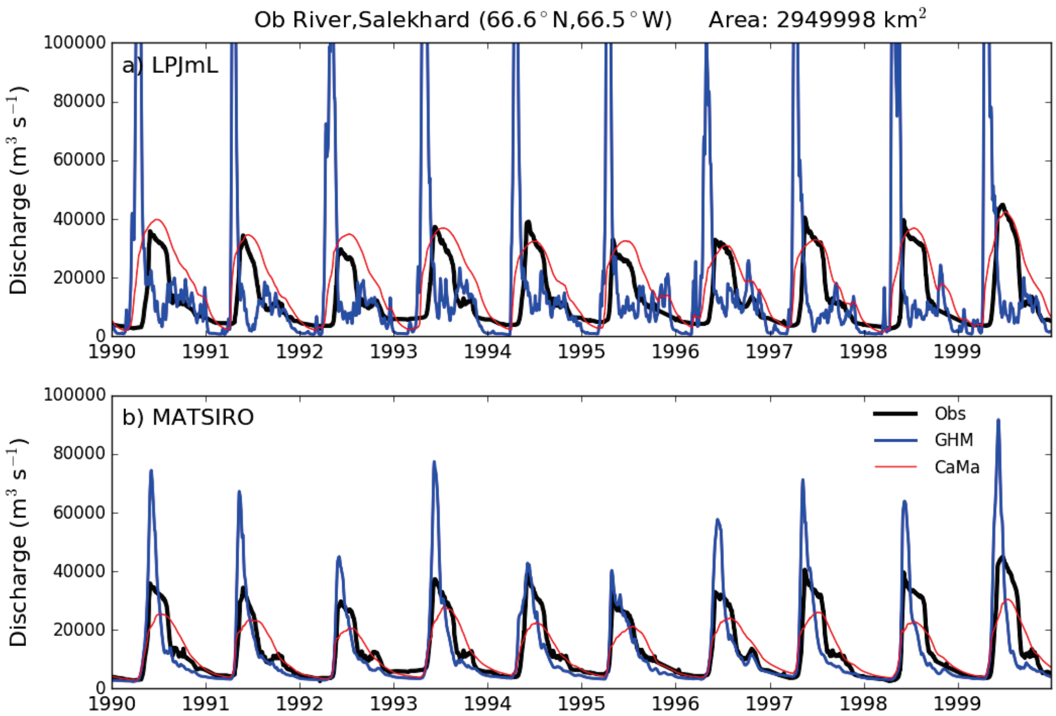

Figure S3. Same as Figure S1 and S2 but at Salekhard, Ob during 1990-1999.

Table S1 lists detailed performance statistics for the three case studies. With its native routing, MATSIRO outperforms LPJmL in all three basins. CaMa-Flood routing brings remarkable improvement on simulated discharge for LPJmL, especially for the Amazon and Ob river basins, where the terrain is relatively flat and floodplain mechanism may play an important role in regulating discharge.

Table S1. Performance of two selected models in simulating daily discharges in Amazon, Mekong and Ob river basins. Numbers in brackets are performance with CaMa-Flood routing. Statistics are based on all years in 1971-2010, where full-year observation data is available.

	Amazon		Mekong		Ob	
	LPJmL	**MATSIRO**	**LPJmL**	**MATSIRO**	**LPJmL**	**MATSIRO**
NSE	-0.29 (0.76)	0.46 (0.17)	0.65 (0.85)	0.78 (0.78)	-7.42 (0.43)	0.32 (0.74)
R	0.55 (0.96)	0.89 (0.69)	0.9 (0.94)	0.9 (0.91)	-0.1 (0.84)	0.79 (0.88)
BMEAN	-11% (-12%)	-17% (-16%)	15% (15%)	-20% (-20%)	34% (41%)	-3% (-2%)
BMAX	35% (-17%)	-15% (-27%)	16% (-24%)	-6% (-35%)	429% (-5%)	105% (-25%)

Supplementary Section B. Comparison to simulations including human impacts

Human hydraulic management through dams, reservoirs, and various water uses, has largely altered river discharge over many river basins across the world. The effects of these human interventions on the river flow are generally much smaller in the case of high flow than in the case of low flow, and their impact becomes even less important with increasing discharge (Veldkamp *et al* 2017). When we separately examined managed and near-natural stations, for peak discharge the results were similar (section 3.2; Table 2). Although dams and reservoirs are expected to largely reduce flood risks, it is possible that such protection is only limited to relatively small areas instead of at all sections of rivers due to financial/technical/environmental limitations/restrictions. Many of the flood-prone countries also have relatively low flood protection levels (Scussolini *et al* 2016), where human interventions are often not effective against large flood events. Additionally, the current representation of human management in GHMs still has much room for improvement. Therefore, even without considering human hydraulic management, CaMa-Flood routing might still simulate a more realistic river discharge than the the GHMs' native routing schemes, despite of their explicit consideration of human management.

Indeed, when we compare CaMa-Flood simulated discharge to the one from GHMs (using an ensemble of three GHMs: H08, LPJmL and WaterGAP2nc) accounting for time-varying human impacts (referred to as "VARSOC" in the ISIMIP2a protocol), we see similar level of improvement for the metrics related to peak discharge (BSTD, BMAX and BMYM) in most of the basins (Figure S4, Figure S13). In some cases (e.g., the Ganges basin in India) CaMa-Flood routing does lead to decreased performance in mean river discharge and NSE, for which human impacts are more important. This result confirms that human impacts as currently represented in GHMs have a limited effect on peak discharge at the global scale.

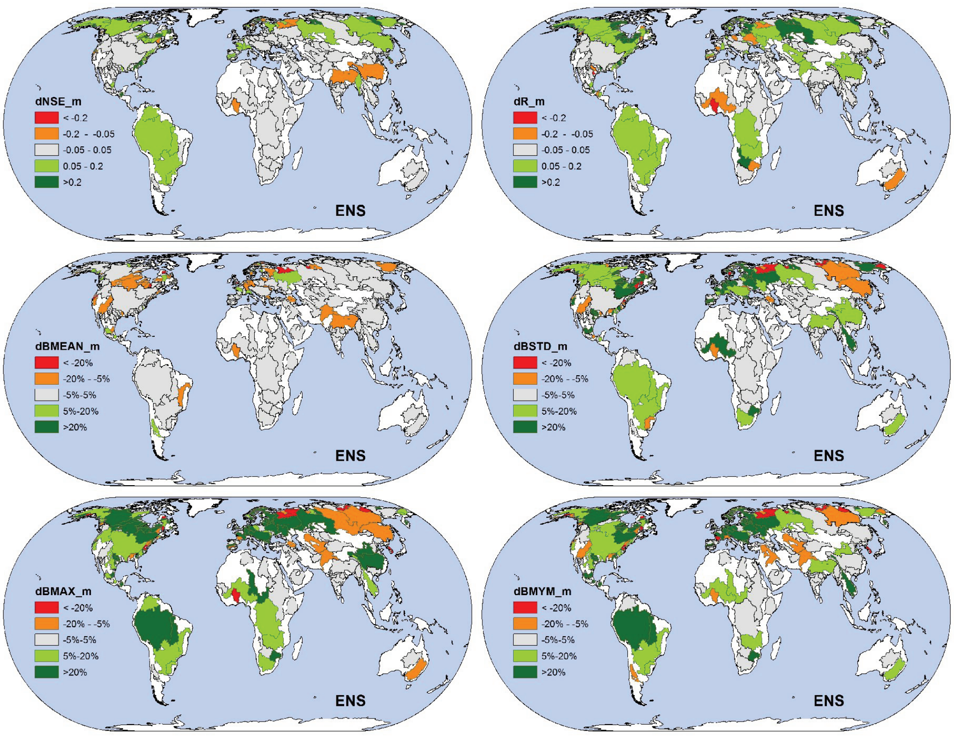

Figure S4. Ensemble mean performance differences between CaMa-Flood simulated discharge and GHM simulated discharge with time-varying human impacts (VARSOC) for three selected GHMs (H08, LPJmL and WaterGAP2nc) using daily GRDC observation as benchmark, all showing basin averages for the metrics.

Supplementary Text C

Similar to Döll et al. (2003), the corresponding grid cell to each GRDC station was determined according to each station's coordinate, if the difference in upstream area was within 5%; otherwise, the adjacent cell with minimum upstream area difference was selected. If the upstream area difference was greater than 30% for all surrounding cells, the station was excluded from further analyses. Cell locating was performed separately for DDM30 and CaMa-Flood's river network. After this procedure, visual inspection was performed with the aid of observed and simulated multi-year mean discharges to correct obvious mismatches in locating the cells. Around 5% of the cells located for CaMa-Flood were altered after this manual correction, mostly due to the location of wetland in CaMa-Flood's network and mainly in Boreal regions. Only a few cells had their location changed for the DDM30 network.

Considerable care and extensive manual correction was carried out in correctly locating the GRDC stations in the DDM30 and CaMa-Flood grids, respectively. A threshold of 30% or less upstream area difference for both grids was adopted, leading to a 5% (DDM30) or 3% (CaMa-Flood) difference in upstream area on average. While this relatively strict criterion reduced the number of stations in analyses, it was a worthwhile trade-off that minimizes the possibility of mis-locating and mitigates potential errors, so that possible mis-locating would likely be only shifting one cell upstream or downstream, where peak discharge is likely similar. However, it should be noted that CLM and PCR-GLOBWB deviated from using the provided DDM30 network such that it was necessary to perform re-location for them separately. About 40% (CLM) and 10% (PCR-GLOBWB) fewer grids meet the upstream area criteria and were included in analyses; therefore results regarding the two GHMs are less robust due to a smaller sample size. Nevertheless, the major findings in this study remain unchanged when excluding the two GHMs from the analyses.

Supplementary Figures and Tables

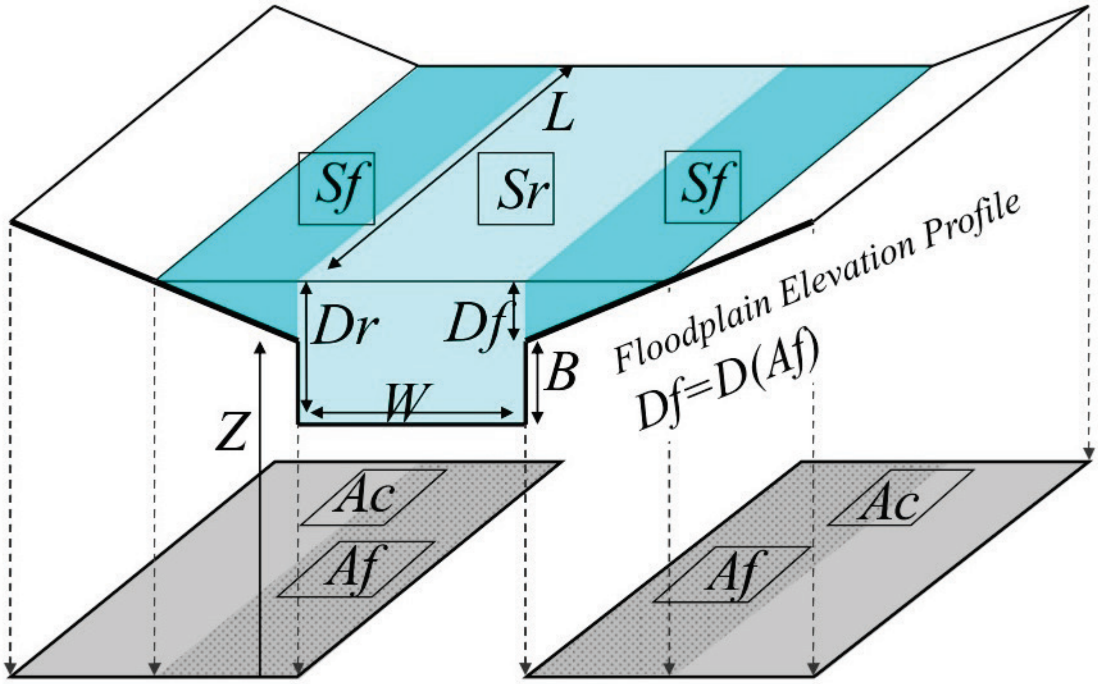

Figure S5. Illustration of a river channel reservoir and a floodplain reservoir defined in each grid in CaMa-Flood (Yamazaki, et al., 2011, Figure 1). L and W are channel length and width, B is bank height, Z is surface altitude, Ac and Af are unit catchment area and flooded area, Dr and Df are river and floodplain water depths, Sr and Sf are river channel and floodplain storages.

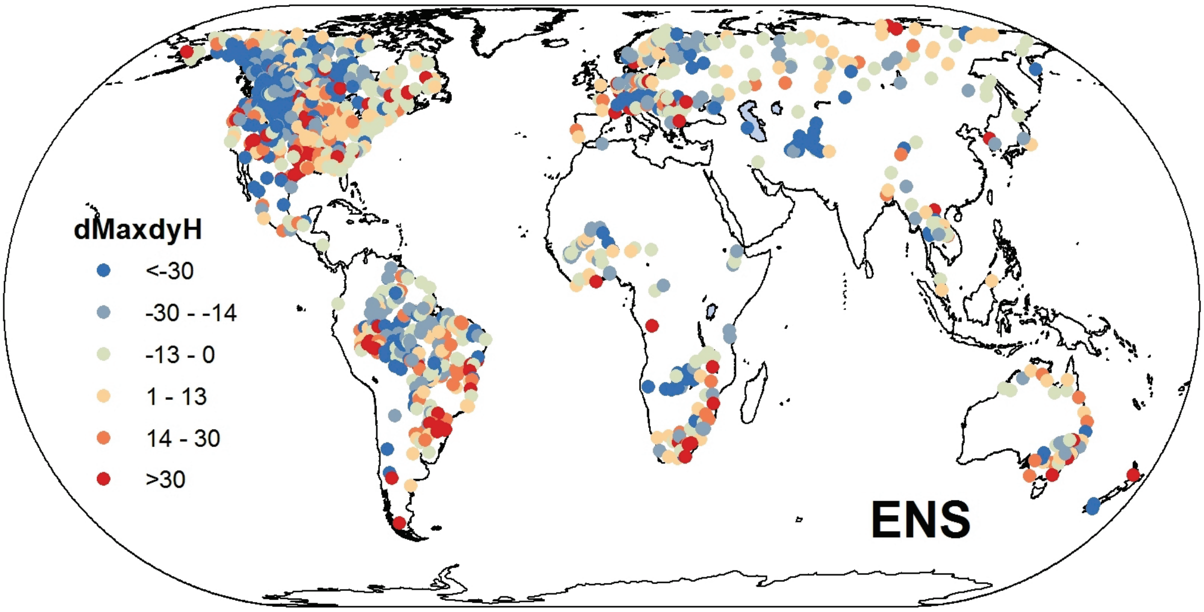

Figure S6. Multi-model ensemble mean changes in timing of climatological daily maximum discharge simulated by GHMs compared to observation. Note the time periods for mean daily hydrograph could be different as observation could be shorter than the 1971-2010 period at many stations. A positive value indicates max discharge occurring later than observation.

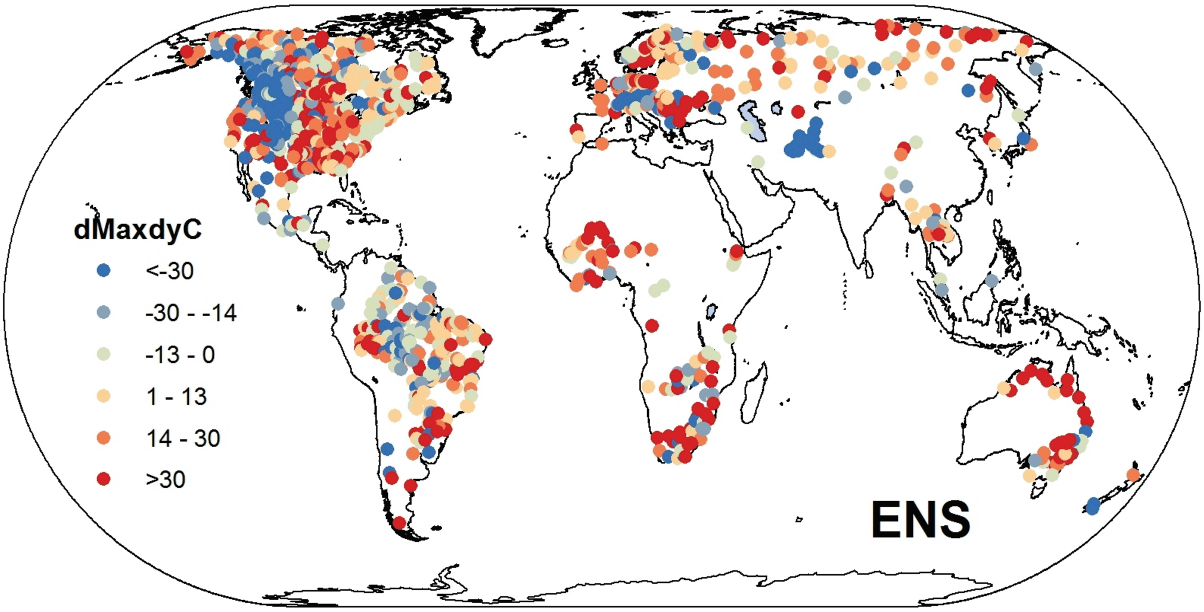

Figure S7. Multi-model ensemble mean changes in timing of climatological daily maximum discharge simulated by CaMa-Flood compared to observation. Note the time periods for mean daily hydrograph could be different as observation could be shorter than the 1971-2010 period at many stations. A positive value indicates max discharge occurring later than observation.

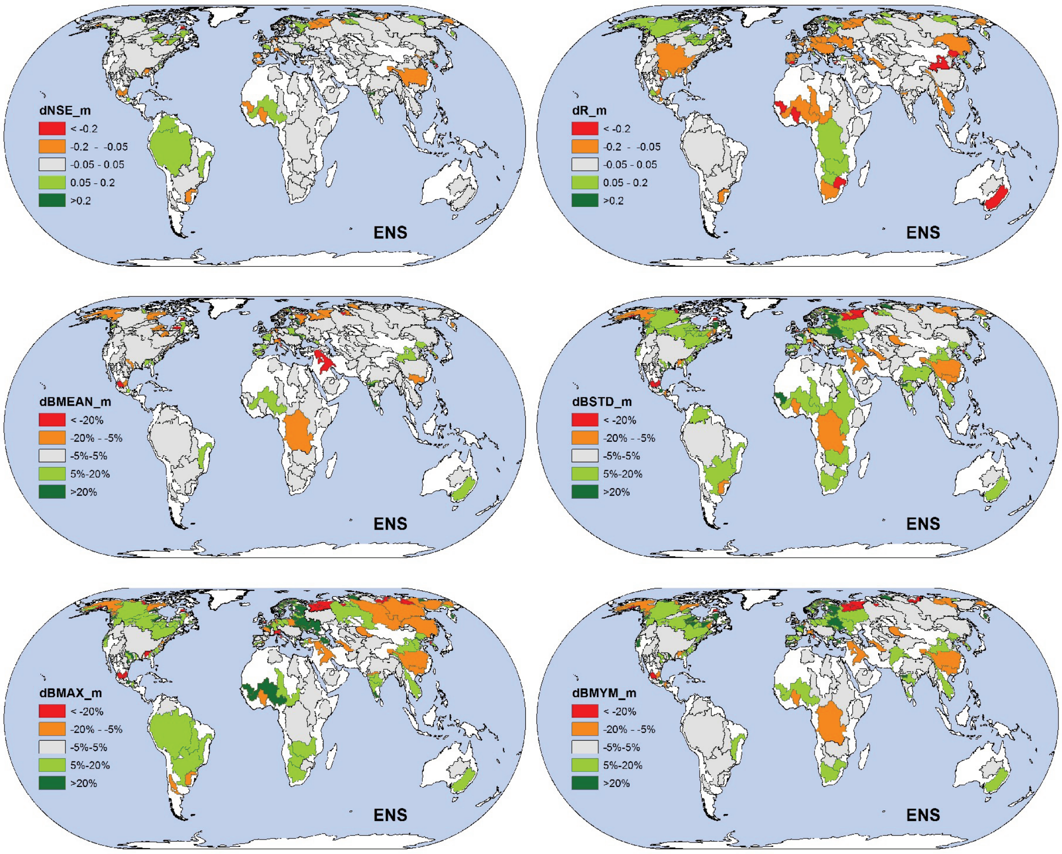

Figure S8. Multi-model ensemble mean performance differences compared to monthly GRDC data, all shown as basin averages (denoted by _m). Grey colour shows differences <5% in basin-averaged performance metrics. Green colours show basins where a discharge metrics is improved with CaMa-Flood compared to native GHM routing.

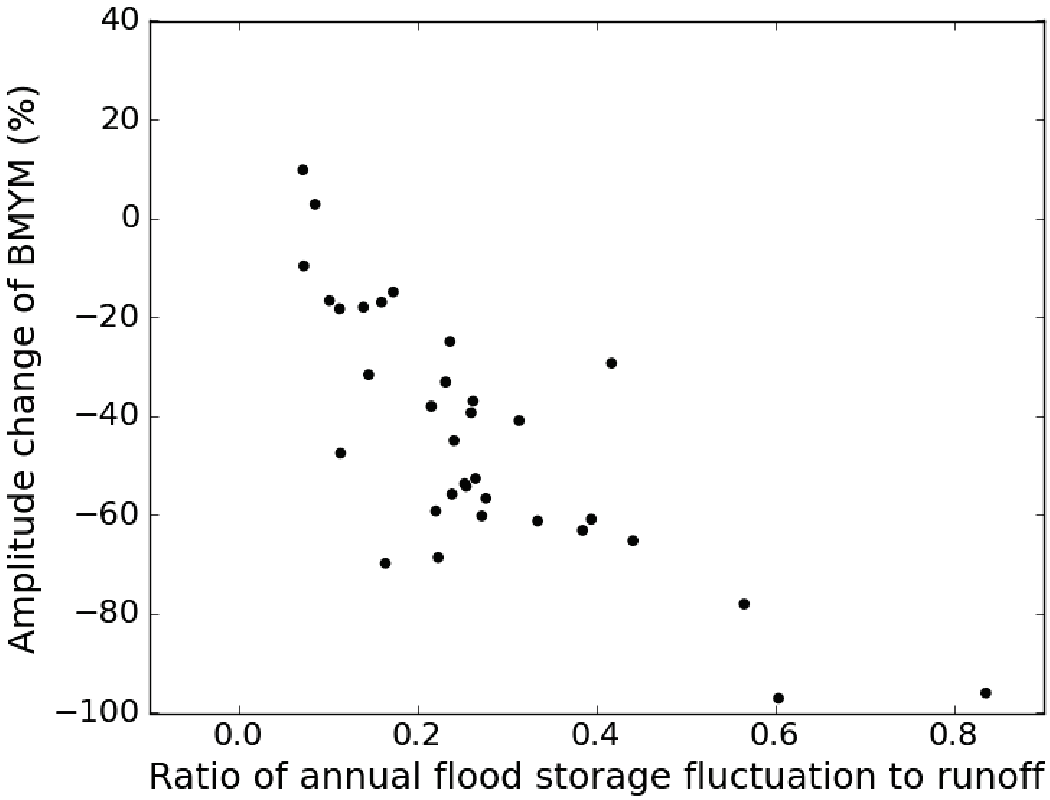

Figure S9. Relationship between ratio of annual basin floodplain storage fluctuation to runoff and amplitude change of daily peak discharge at basin outlet, averaged over the 1971-2010 period. Each dot represents multi-model ensemble median (DBH, H08, LPJmL, MATSIRO, WaterGAP2nc) for one of 34 selected large basins (area >100, 000 km^2) worldwide.

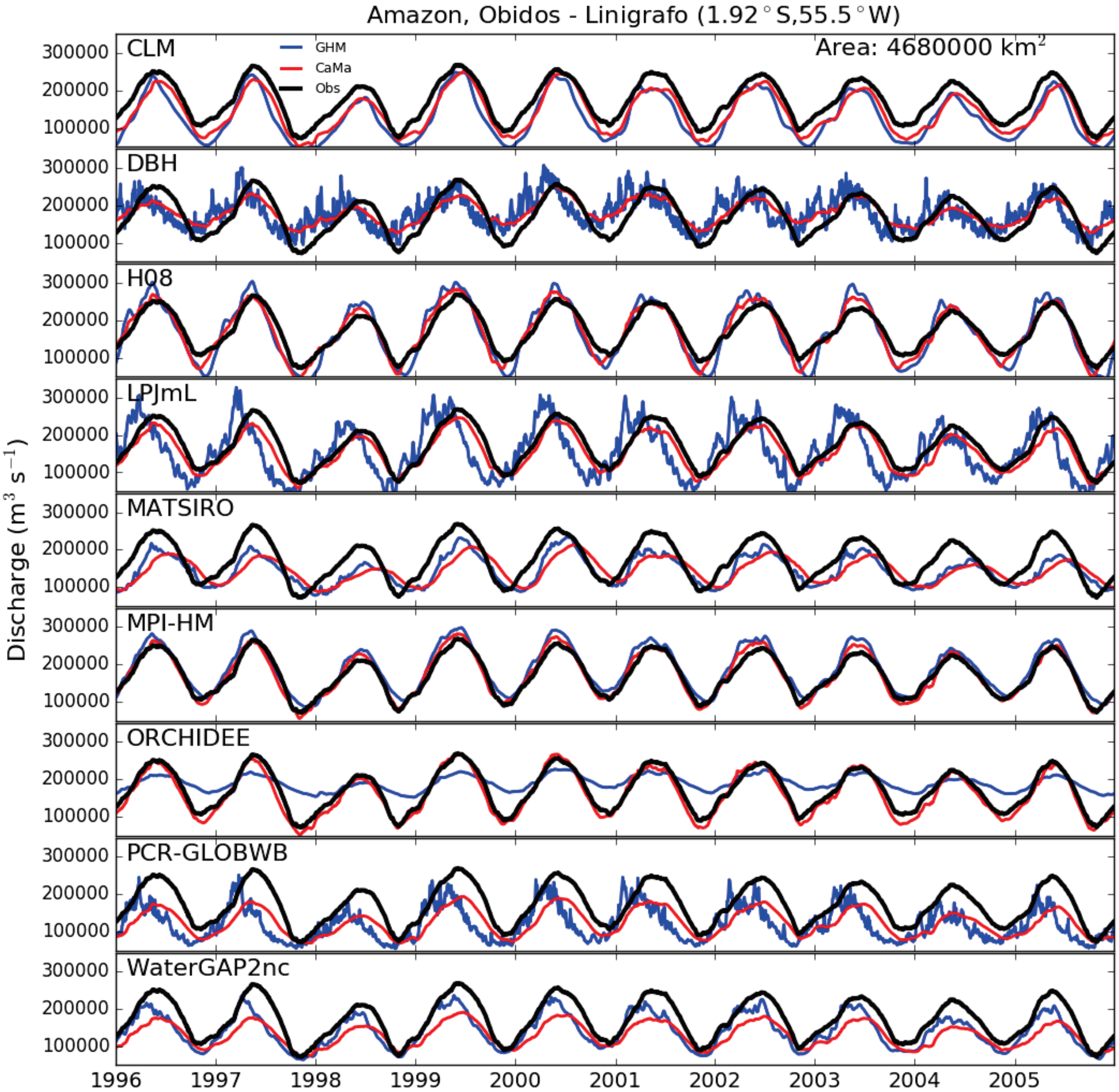

Figure S10. Observed (black), GHM (blue) and CaMa-Flood (red) simulated daily river discharges at Amazon, Obidos-Linigrafo during 1996-2005, for the nine GHMs.

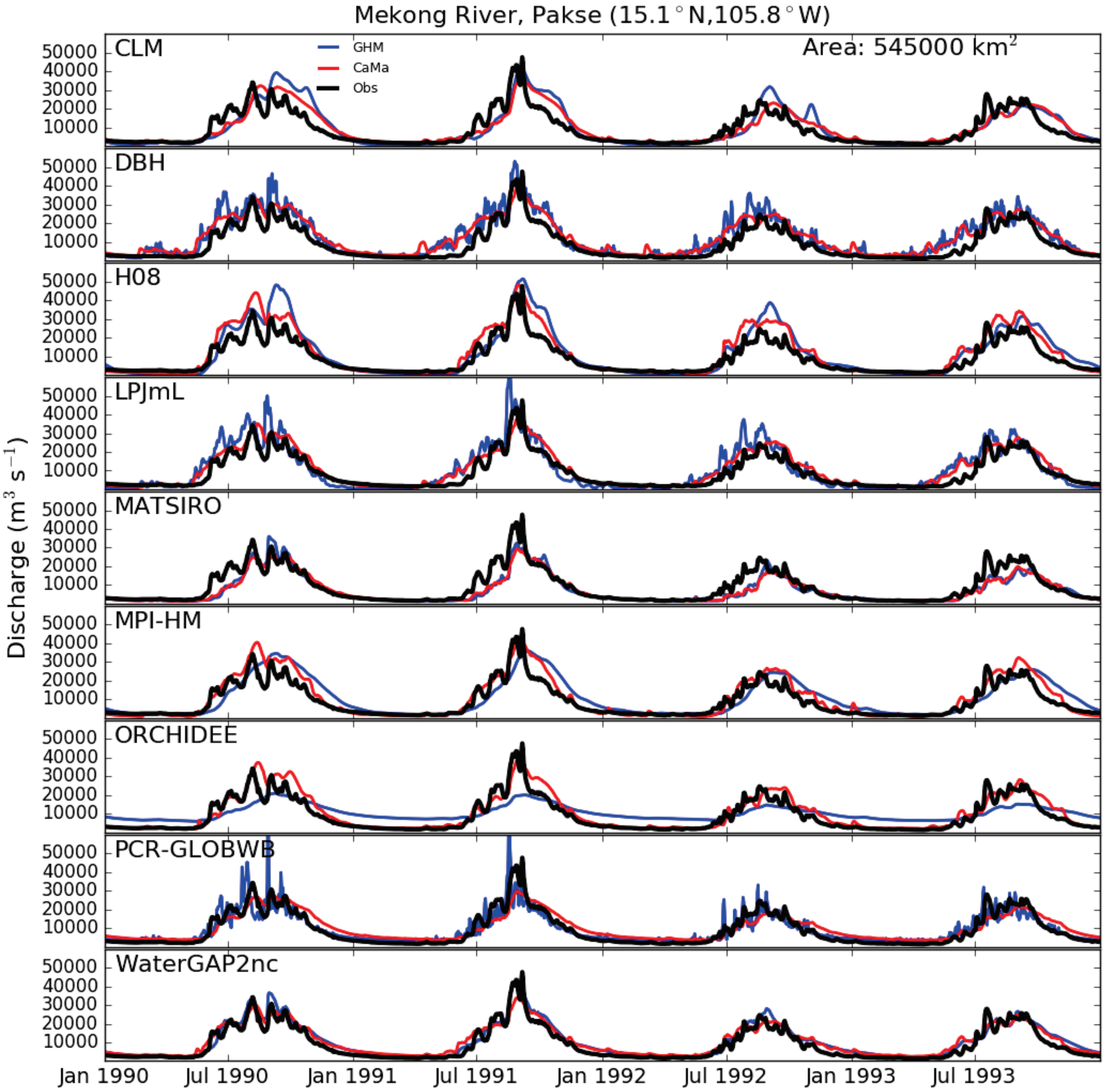

Figure S11. Same as Figure S10 but at Pakse, Mekong during 1990-1993.

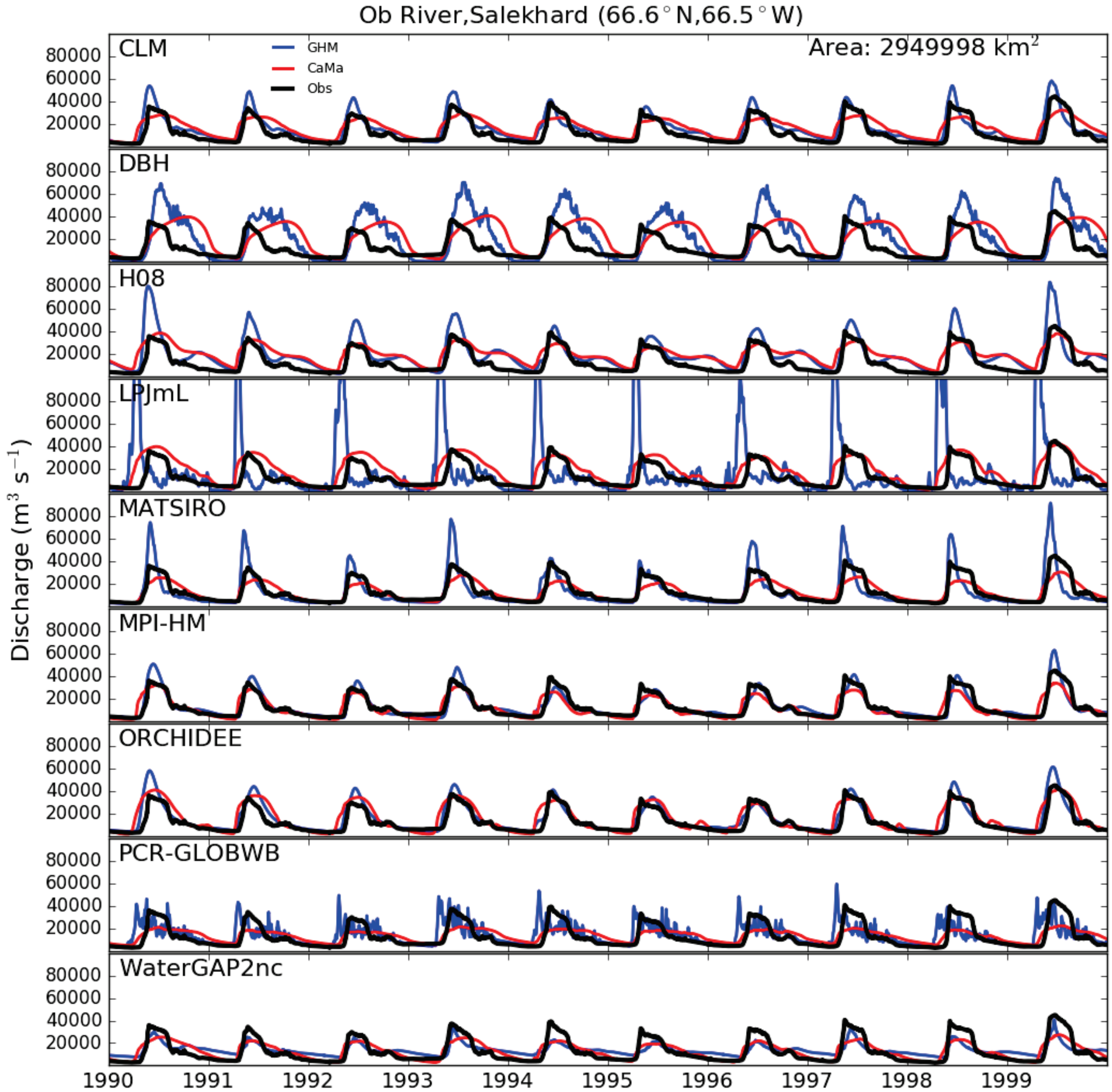

Figure S12. Same as Figure S10 and S11 but at Salekhard, Ob during 1990-1999.

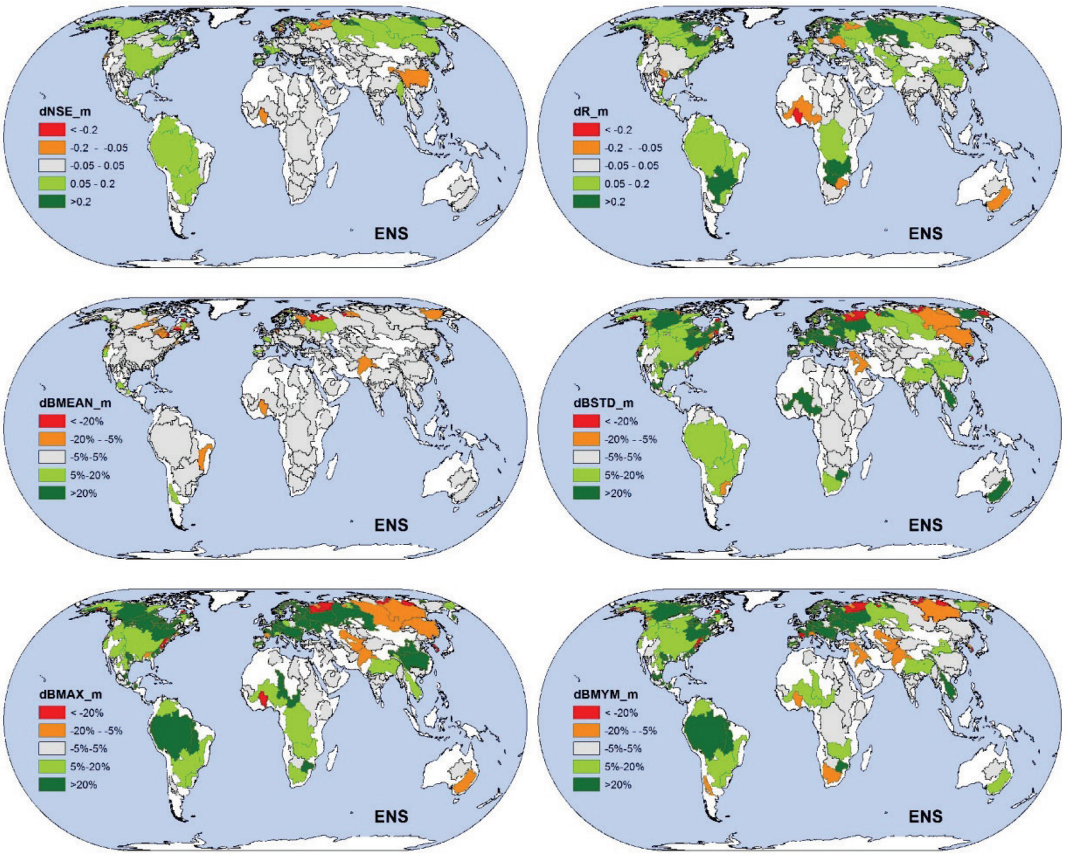

Figure S13. Ensemble mean performance differences between CaMa-Flood simulated discharge and GHM simulated discharge with naturalized run for three selected GHMs (H08, LPJmL and WaterGAP2nc) using daily GRDC observation as a benchmark (similar to Figure 4 except that only three models are used for the ensemble mean, in order to compare with Figure S4), all showing basin averages for the metrics.

Table S2: Main characteristics of the GHMs as used in this study

Model name	Water and energy budgets					Routing			References
	Energy balance	**Soil scheme**	**Evaporation scheme**	**Runoff scheme**	**Snow scheme**	**Routing scheme**	**Flow velocity**	**Floodplain scheme**	
CLM	Yes	10 soil layers up to 3.8m	Modified Penman-Monteith	Saturation and infiltration excess	Degree day	Linear reservoir	0.35m/s	No	(Leng *et al* 2015)
DBH	Yes	Three soil layers with varied depth up to 1.5-2m	Energy balance	Infiltration excess	Energy balance	Linear reservoir	Variable based on topographic gradient	No	(Tang *et al* 2007, 2008)
H08	Yes	One soil layer with a depth of 1m	Bulk formula	Saturation excess, non-linear	Energy balance	TRIP (Oki and Sud 1998, linear reservoir)	0.5m/s	No	(Hanasaki *et al* 2008b, 2008a)
LPJmL	No	Five layers of 20, 30, 50, 100 and 100 cm thickness	Priestley-Taylor	Saturation excess	Degree-day	Continuity equation derived from linear reservoir model	1 m/s	No	(Rost *et al* 2008, von Bloh *et al* 2010)
MATSIRO	Yes	12 fully resolved layers (5cm, 20cm, 75cm, and nine next layers of 1m) and a 90m groundwater layer	Bulk formula	Overland flow, infiltration excess, saturation excess, groundwater.	Energy balance	TRIP (Oki and Sud 1998, linear reservoir)	0.5m/s	No	(Takata *et al* 2003, Pokhrel *et al* 2012, 2015)
MPI-HM	No	prescribed by the plant routing depth	Penman-Monteith	Saturation excess, non-linear	Degree-day	Linear reservoir	Variable, based on Manning-Strickler	Yes	(Hagemann and Dümenil Gates 2003, Stacke and Hagemann 2012)
PCR-GLOBWB	No	Variable up to 1.5 m soil layers and 50 m groundwater layer	Hamon	Saturation Excess Beta Function	Degree Day	Travel time routing (characteristic distance)	Variable based on channel dimensions and gradient with Manning's Equation	No	(Wada *et al* 2010, van Beek *et al* 2011, Wada *et al* 2011)
ORCHIDEE	Yes	11 layers in a 2 m soil	Bulk formula	Infiltration excess	Energy balance	Same as MPI-HM*	Same as MPI-HM*	Same as MPI-HM*	(Guimberteau *et al* 2014)

WaterGAP2	No	One soil layer, varying depth in dependence on land cover type (0.1 to 4 m)	Priestley Taylor with two alpha factors depending on the aridity of the grid cell	Beta function, saturation excess	Degree Day	Linear reservoir	Variable, based on Manning-Strickler (details see Verzano et al. 2012)	No	(Müller Schmied *et al* 2014, 2016)

*ORCHIDEE's discharge is post-processed using the same MPI-HD model from MPI-HM for ISIMIP submission.

Table S3. Percentage of land area showing a considerably better(left)/worse(right) performance in their basin-average representation of R (over 0.05 difference) with CaMa-Flood routing compared to the GHMs' native routing schemes; using all studied stations, managed stations only, and (near-)natural stations only.

	All Stations (%)	Managed Stations only (%)	Natural Stations only (%)
CLM	25 / **29**	**35** / 31	27 / **31**
DBH	29 / **38**	33 / **35**	30 / **33**
H08	**34** / 22	**23** / 14	**42** / 22
LPJmL	**90** / 1	**84** / 9	**94** / 1
MATSIRO	21 / **32**	24 / **40**	24 / **34**
MPI-HM	3 / **43**	17 / **54**	4 / **38**
ORCHIDEE	**40** / 23	**34** / 22	**42** / 19
PCR-GLOBWB	**45** / 24	**56** / 22	**48** / 23
WaterGAP2nc	33 / **34**	27 / **39**	**33** / 26
WaterGAP2	20 / **46**	16 / **35**	21 / **40**
ENS*	**49** / 17	**46** / 17	**55** / 15

*Note that the ensemble (ENS uses the uncalibrated (WaterGAP2nc) instead of calibrated version of WaterGAP2.

Table S4. **Similar to Table S3, but for NSE.**

	All Stations (%)	Managed Stations only (%)	Natural Stations only (%)
CLM	**24** / 19	9 / **11**	**28** / 18
DBH	**23** / 7	**7** / 4	**23** / 6
H08	**42** / 7	**24** / 0	**36** / 9
LPJmL	**60** / 0	**28** / 0	**67** / 0
MATSIRO	**25** / 24	14 / **24**	20 **/ 23**
MPI-HM	4 / **53**	6 / **35**	6 / **50**
ORCHIDEE	14 / **28**	2 / **13**	19 **/ 30**
PCR-GLOBWB	**32** / 11	**18** / 5	**35** / 12
WaterGAP2nc	**28** / 13	**18** / 7	**27** / 26
WaterGAP2	22 / **52**	15 / **39**	25 **/ 52**
ENS	**24** / 3	**14** / 8	**26** / 4

*Note that the ensemble (ENS) uses the uncalibrated (WaterGAP2nc) instead of calibrated version of WaterGAP2.

Table S5. Land area-based mean performance of the individual GHMs with CaMa-Flood/GHM simulated daily discharge compared to GRDC observations. Percent biases are weighted averages of the absolute value, regardless of over- or under-estimation. Numbers in bold indicate better agreement with observations.

	NSE	R	PBSTD (%)	PBMAX (%)	PBMYM (%)
CLM	**0.16** / 0.15	0.54 **/ 0.56**	38 / 38	51 / **48**	41 / 41
DBH	**0.09** / 0.07	0.49 / 0.49	**44** / 70	**44** / 63	**46** / 73
H08	**0.14** / 0.10	**0.56** / 0.54	**47** / 59	**52** / 62	**50** / 59
LPJmL	**0.14** / 0.01	**0.55** / 0.27	**43** / 75	**44** / 84	**47** / 85
MATSIRO	0.17 **/ 0.18**	0.51 / **0.55**	37 / **34**	46 / **45**	41 / **38**
MPI-HM	0.16 / **0.23**	0.58 / **0.66**	43 / **35**	43 / **37**	43 / **36**
ORCHIDEE	0.11 / **0.13**	**0.49** / 0.47	41 / **37**	47 / **36**	44 / **36**
PCR-GLOBWB	**0.11** / 0.09	**0.51** / 0.47	**40** / 47	**49** / 66	**43** / 64
WaterGAP2nc	**0.17** / 0.16	**0.61** / 0.60	**41** / 49	**47** / 56	**46** / 56
WaterGAP2	0.24 **/ 0.30**	0.56 / **0.61**	32 / **27**	42 / **38**	39 / **34**