Text S1. Uncertainty of GRACE TWSA data

GRACE TWSA estimates for spatial units are affected by leakage errors that are caused by the need for spectral truncation and the need to filter the solutions, which, for averages of different spatial units, may lead to an under- or overestimation of TWSA, thus affecting model calibration using GRACE TWSA. Therefore, when consistently comparing simulated to GRACE TWSA, it is advised to filter the simulated grid cell data with the same filter that was used to process the GRACE data (Döll et al., 2014). However, given the large number of simulations required in ensemble-based calibration, this approach is computationally impractical. To roughly estimate the leakage effect, a re-scaling factor for GRACE TWSA was estimated for each CDA unit using Eq.1 of Swenson and Landerer (2012). The GRACE TWSA time series for CDA units can be multiplied with such a re-scaling factor to (ideally) reduce the leakage error and in this way make it better comparable to the simulated TWSA time series. First, the monthly time series of gridded TWSA as simulated by standard WaterGAP was filtered with the DDK3 filter, and then both the filtered and the unfiltered TWSA values were aggregated over all grid cells with a CDA unit. The re-scaling factor was then derived by minimizing the misfit between filtered and unfiltered TWSA time series through a simple least square regression. The re-scaling factors are between 1.00 and 1.03 for the CDA units MRB, Missouri and Upper MRB. They are 0.90 and 0.93 for the Ohio and Arkansas River basins, respectively, and 1.41 for the Lower MRB. As the re-scaling factors are close to 1 in all CDA units except the Lower MRB and we suspect that the large re-scaling for the MRB is due to an overestimation of the TWSA trend in the Lower MRB by WaterGAP, we did not apply re-scaling factors to GRACE TWSA.

The GRACE mission relies on accelerometers to measure non-gravitational forces. However, since August 2007, battery cell failures onboard the GRACE satellites led to increasing power supply problems, especially during orbital eclipses. As a result, the thermal control of the accelerometers was deactivated in April 2011 such that thermal variations would directly increase the measurement noise. To mitigate this problem, thermal variations and their impact on the GRACE instruments are modeled during the processing at TU Graz and the accelerometer data are calibrated (Klinger and Mayer-Gürr, 2016). This reduces the noise of the monthly gravity field solutions by an estimated 20-40% compared to solutions without accelerometer calibration (Klinger et al., 2016), but on balance, all GRACE solutions are deemed noisier from April 2011 onwards, the estimation of the noise floor is more uncertain, and the number of months without observations increases towards the end of the study period.

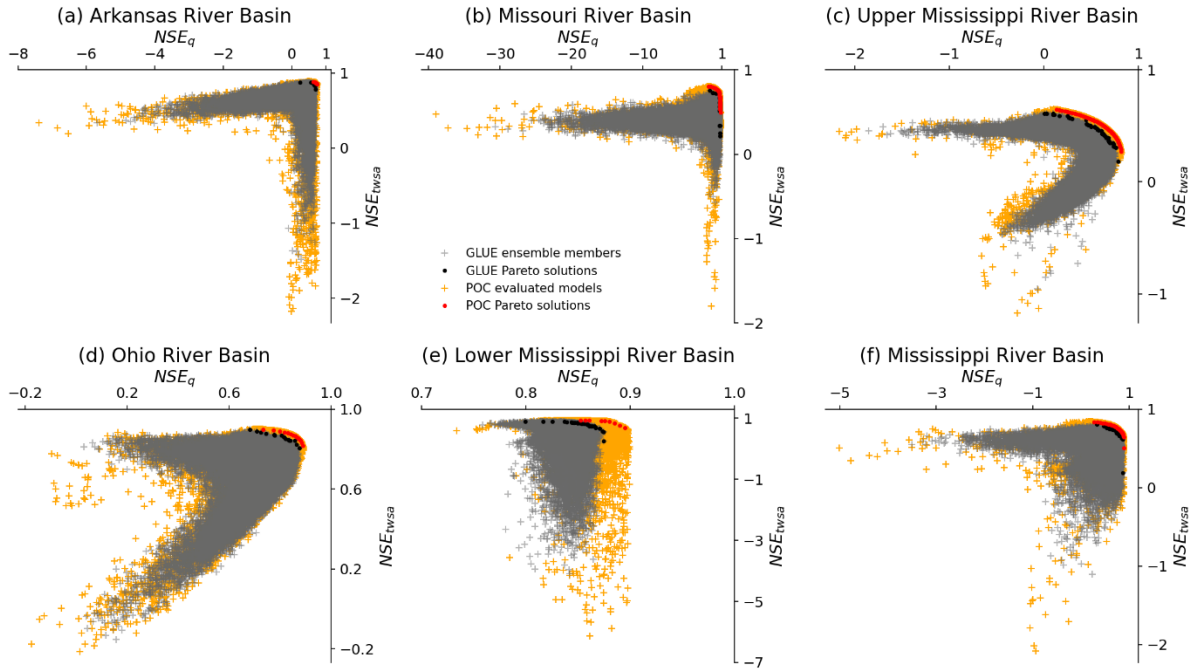

Figure S1. NSE_Q and NSE_{TWSA} of all 20,000 parameter sets derived by 1) a-priori assumptions on parameter uncertainty according to Table 2 in the case of GLUE, and 2) using an optimization algorithm in the case of POC. Solutions on the POC (red dots) and GLUE (black dots) Pareto front are indicated.

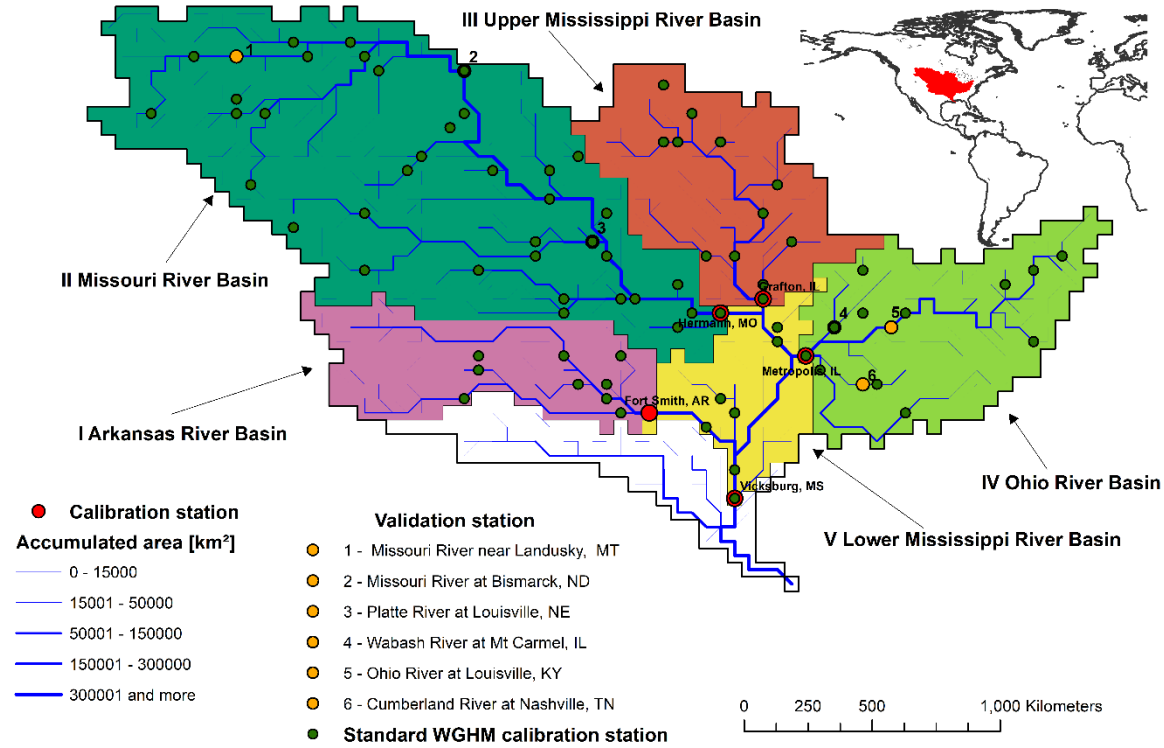

Figure S2. Streamflow stations used in the standard calibration of WGHM (in green), resulting in 77 spatial calibration units (CDA units), as well as the calibration and validation stations used in this paper.

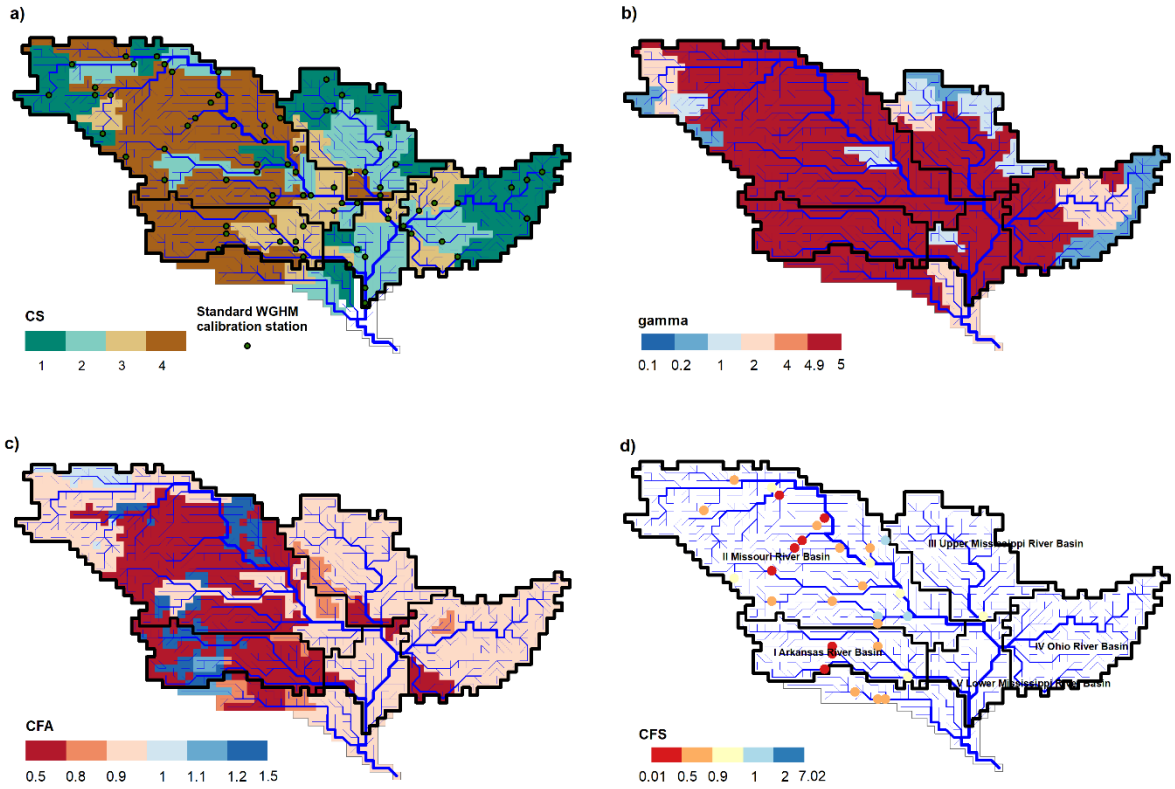

Figure S3. Calibration status (a) and values of calibration parameters SL-RC (b), areal correction factor (c), and station correction factor (d) obtained by the standard WaterGAP calibration for 77 CDA units. Calibration follows a four-step scheme with specific calibration status (CS): CS1: adjusting the basin-wide uniform parameter (Eq. (18)) in the range of [0.1-5.0] to match Q_{obs} within ± 1%. CS2: adjusting like in the case of CS1, but within a 10% uncertainty range (90-110% of observations). CS3: like CS2 but applying the areal correction factor CFA (adjusts runoff and, to conserve the mass balance, actual evapotranspiration of each grid cell within the range of [0.5-1.5]) to match Q_{obs} with 10% uncertainty. CS4: like CS3 but applying the station correction factor CFS (multiplies streamflow in the cell where the gauging station is located by an unconstrained factor) to match Q_{obs} with 10% uncertainty to avoid error propagation to the downstream basin. Different from this study, the maximum value of SL-RC in the standard calibration is 5.

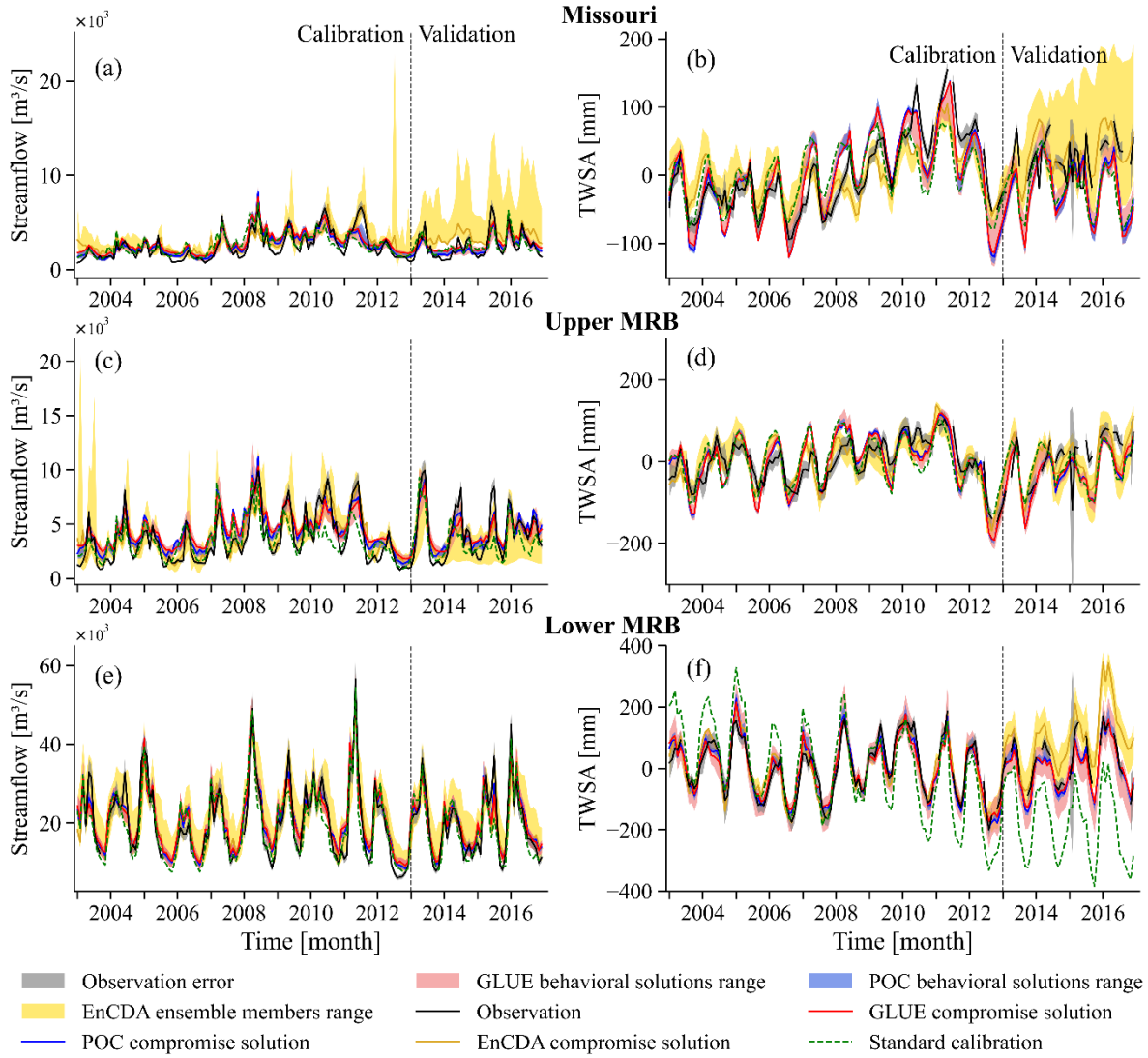

Figure S4. Monthly time series of simulated and observed Q (a, c, e) and TWSA (b, d, f) during calibration period 2003-2012 and validation period 2013-2016 for Missouri basin (a, b), Upper MRB (c, d) and Lower MRB (e, f). Observations and their assumed errors are shown together with simulated GLUE, POC, and EnCDA compromise solutions, with the range of GLUE and POC behavioral solutions (maximum and minimum monthly values of the behavioral solutions, Table 6) and the range of all 32 EnCDA ensemble members, as well as with the WaterGAP variant with standard calibration.

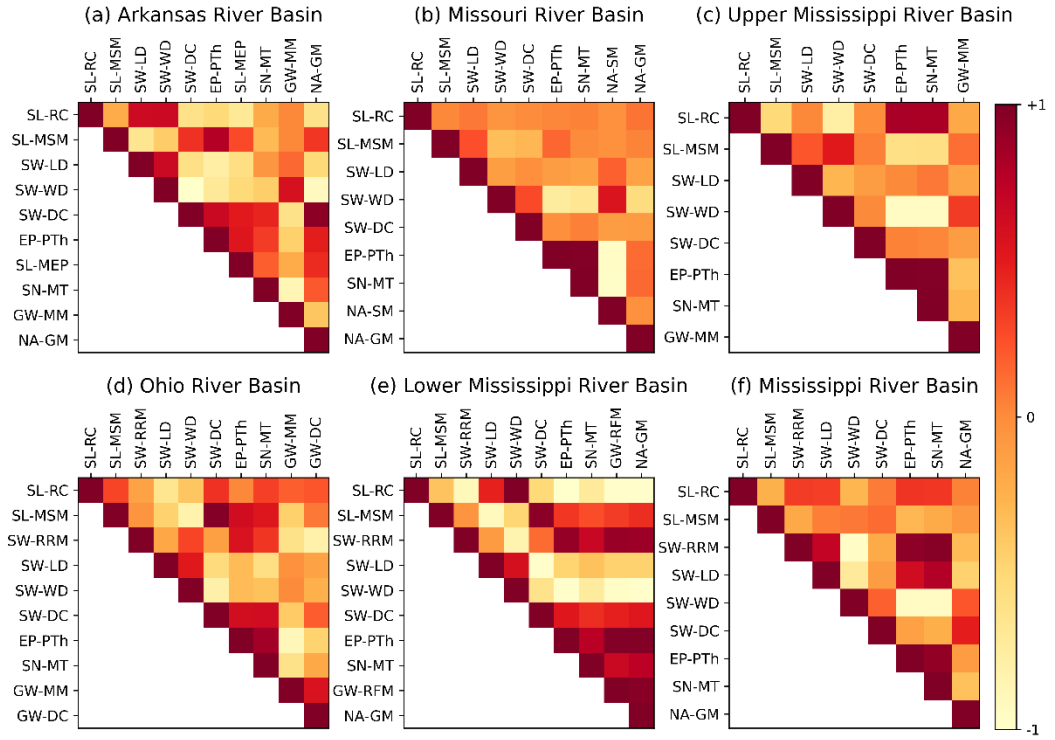

Figure S5. Correlation of calibration parameters in the ensemble of behavioral Pareto solutions derived by POC (see Table 3 for the number of ensemble members).

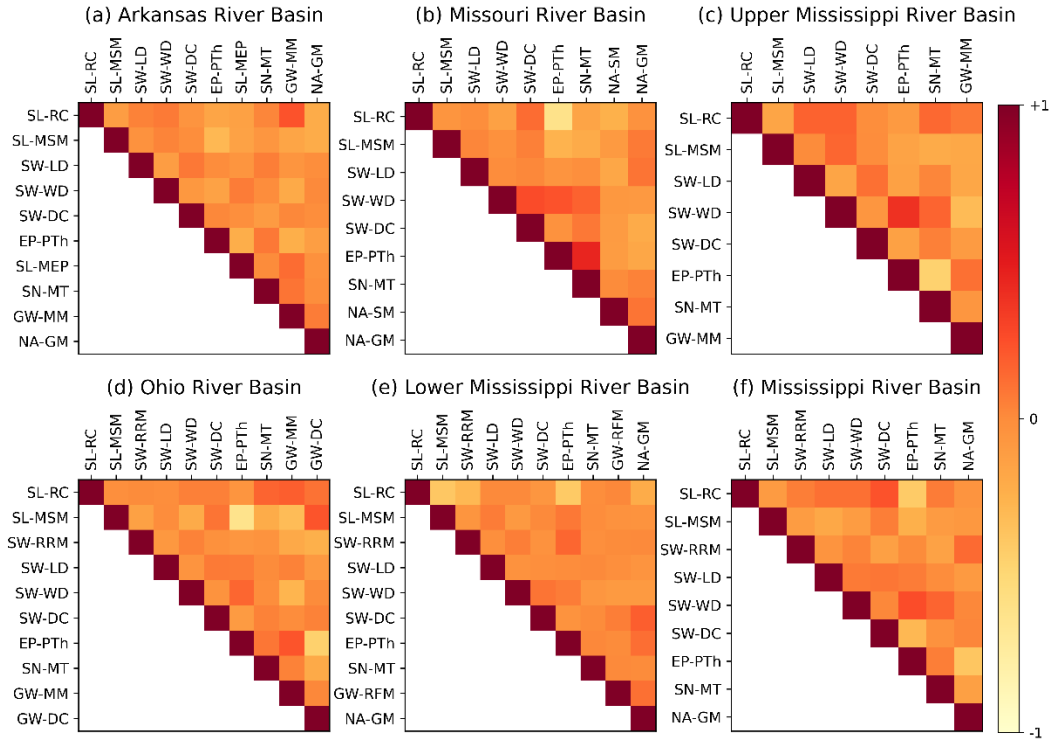

Figure S6. Correlation of calibration parameters in the ensemble of behavioral solutions derived by GLUE (see Table 3 for the number of ensemble members).

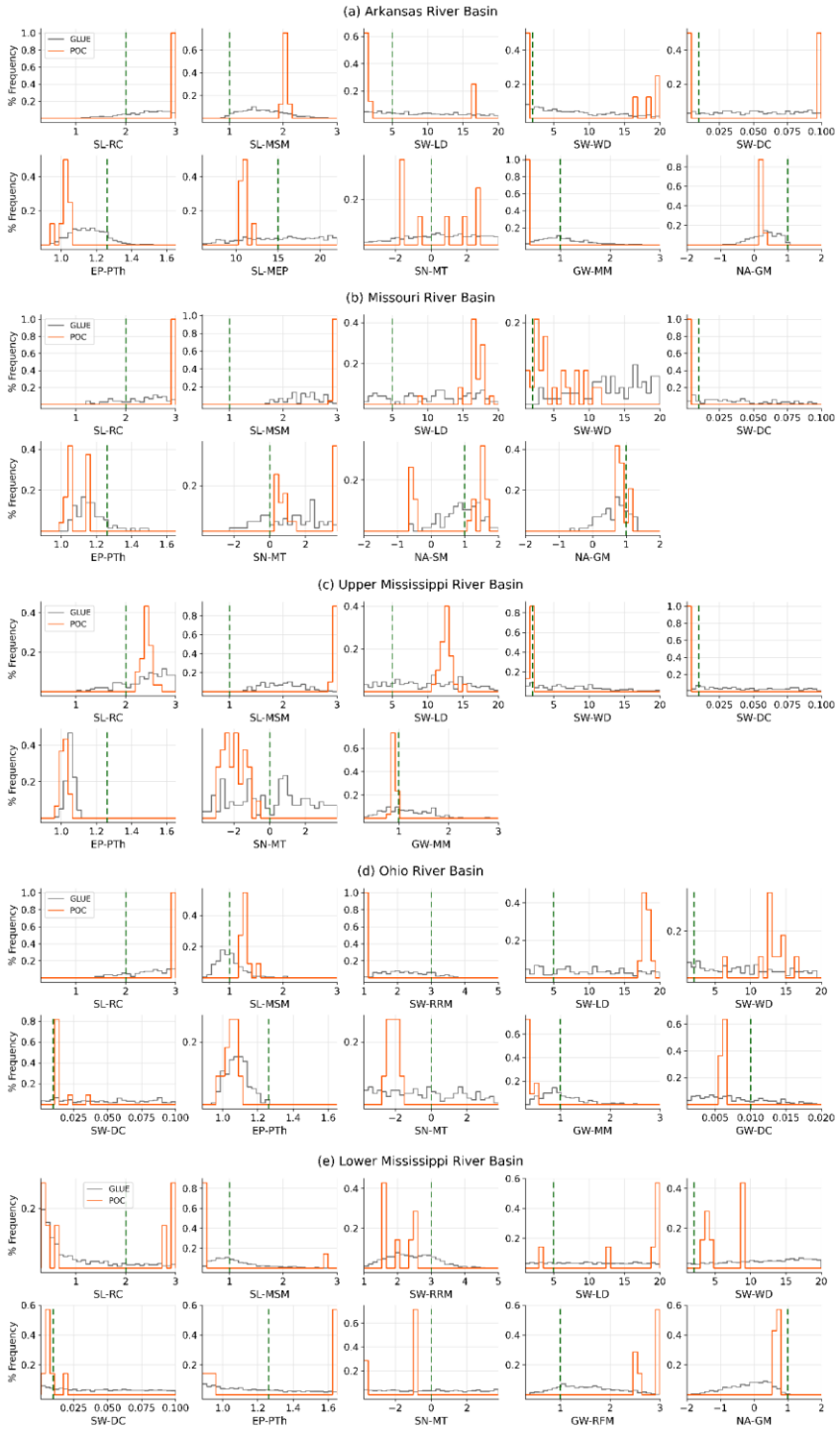

Figure S7. Histograms of parameter values in calibrated parameter sets for sub-basin CDA units Arkansas (a), Missouri (b), Upper MRB (c), Ohio (d) and Lower MRB (e). All behavioral parameter sets are considered for GLUE, while the smaller ensemble of behavioral Pareto-optimal parameter sets (neglecting observation errors) is shown for POC. The total number of parameters set for POC and GLUE is listed in Table 3. The y-axis shows the ratio of the number of parameter values in a class interval to the total number of parameter sets, while the x-axis provides the a-priori parameter range listed in Table 1. The green dashed line indicates the parameter values of the uncalibrated WaterGAP model.

Table S1. Comparison of mean annual precipitation in the CDA units for the calibration period 2003-2012 between GPCC-WFDEI used to drive WaterGAP and the high-resolution (4 km) PRISM* dataset for the USA [mm/yr]

CDA unit	GPCC-WFDEI	PRISM	Ratio PRISM/GPCC-WFDEI (potential P-PM)
I Arkansas	705	667	0.95
II Missouri	595	622	1.04
III Upper MRB	951	878	0.92
IV Ohio	1313	1242	0.95
V Lower MRB	1286	1254	0.97
MRB	839	829	0.99

*https://climatedataguide.ucar.edu/climate-data/prism-high-resolution-spatial-climate-data-united-states-maxmin-temp-dewpoint

Table S2. The most influential parameters for streamflow, TWSA, snow cover and local lake storage, covering together at least 50% of the total effect.

CDA Unit	Streamflow	TWSA	Snow cover	Local lake storage
I Arkansas	SL-RC, SL-MSM, EP-PTh, SL-MEP, GW-MM	SL-RC, SL-MSM, NA-GM	SN-MT	SW-LD, SW-DC
II Missouri	SL-RC, SL-MSM, EP-PTh, SN-MT, NA-SM	SL-RC, SL-MSM, SW-WD, EP-PTh, NA-GM	SN-MT	SW-LD, SW-DC, NA-SM
III Upper MRB	SL-RC, SL-MSM, EP-PTh, SN-MT, GW-MM	SL-RC, SL-MSM, SW-WD, SW-DC, EP-PTh	SN-MT	SW-LD, SW-DC
IV Ohio	SL-RC, SL-MSM, SW-RRM, EP-PTh, GW-MM	SL-RC, SL-MSM, EP-PTh, GW-DC	SN-MT	SW-LD, SW-DC
V Lower MRB	SL-RC, SL-MSM, SW-RRM, EP-PTh, SN-MT	SL-MSM, GW-RFM, NA-GM	SN-MT	SW-LD, SW-DC
MRB	SL-RC, SL-MSM, SW-RRM, EP-PTh	SL-RC, SL-MSM, EP-PTh, NA-GM	SN-MT	SW-LD, SW-DC

Note that although SW-WD was not selected in unit I, IV, V, MRB, we decided to select the parameter for all units due to effect on groundwater recharge from surface water bodies

Table S3. Comparison of model performance in the five sub-basins of the MRB between the calibration of MRB as a whole (CDA unit VI) and calibration of the individual sub-basins (CDA units I – V). Model performance is indicated by NSE_Q and NSE_{TWSA} during the validation period 2013-2016, as achieved by the compromise solutions of the three calibration approaches POC, GLUE, and EnCDA. The values in parenthesis in the line "EnCDA compromise" are NSE_{TWSA} values that are computed after normalizing TWSA during the validation period by the mean TWSA of the validation period.

	NSE_Q/NSE_{TWSA}					
	Arkansas	Missouri	Upper MRB	Ohio	Lower MRB	MRB
POC: whole basin calibration	0.47/0.22	0.63/-0.92	0.71/-0.16	0.80/0.77	0.85/0.60	0.85/0.31
POC: sub-basin calibration	0.59/-0.03	0.72/-2.77	0.79/-0.05	0.85/0.75	0.83/0.78	0.83/0.02[1]
GLUE: whole basin calibration	0.49/0.06	0.67/-0.99	0.68/-0.28	0.83/0.67	0.84/0.61	0.84/0.11
GLUE: sub-basin calibration	0.61/0.66	0.68/-3.45	0.74/0.02	0.86/0.72	0.80/0.76	0.80/-0.28[1]
EnCDA: whole basin calibration	0.40/0.65 (-0.25)	-1.08/ -0.26 (-0.43)	0.19/-0.28 (-0.36)	0.50/0.34 (0.25)	0.61/0.47 (0.46)	0.61/-1.00 (-1.72)
EnCDA: sub-basin calibration	0.07/0.11 (-3.99)	0.02/-0.30 (-0.30)	0.68/-0.07 (-0.07)	0.74/0.20 (-2.60)	0.76/0.43 (-0.66)	0.76/-1.15 (-5.86)
Standard calibration	0.44/-0.85	0.60/-3.70	0.47/-0.40	0.85/0.62	0.76/-6.24	0.76/-2.38

[1] based on Q at Vicksburg and TSWA averaged over the whole MRB computed by a WaterGAP run, in which the calibration parameters in the five sub-basins (CDA units I-V) were set to their respective compromise solution values.

Table S4. Comparison of model performance at the six streamflow validation stations in the Missouri and Ohio sub-basins of the MRB (Fig. 2) between the calibration of MRB as a whole (CDA unit VI) or calibration of the individual sub-basins (CDA units I–V). Model performance is indicated by NSE_Q and the three KGE components during the validation period 2013-2016 as achieved by the compromise solutions of the three calibration approaches POC, GLUE, and EnCDA. The best-performing calibration variant for each station is shown in bold if NSE>0. In addition, the performance of the standard and uncalibrated WaterGAP model is shown.

	NSE_Q/CC/RBias/RVar					
	Missouri near Landusky	Missouri at Bismarck[1]	Platte at Louisville[1]	Wabash at Mt Carmel[1]	Ohio at Louisville	Cumberland at Nashville
POC: whole basin calibration	-1.33/0.65/ 0.57/1.68	-0.13/0.42/ 0.88/0.55	0.36/0.78/ **1.12**/**1.11**	0.67/**0.86**/ 1.15/0.78	0.78/0.92/ 1.09/0.64	0.55/0.89/ 1.29/0.58
POC: sub-basin calibration	-2.15/0.77/ 0.45/2.38	-3.27/0.42/ 0.58/0.81	0.10/0.78/ 0.58/1.32	0.44/0.76/ 1.14/0.94	**0.90**/**0.95**/ 1.04/0.89	0.65/0.91/ 1.26/0.77
GLUE: whole basin calibration	-0.67/0.81/ 0.64/1.88	-0.01/0.41/ 0.93/0.76	0.29/0.88/ 1.15/1.32	**0.70**/**0.86**/ 1.11/0.87	0.80/0.90/ **1.02**/0.79	0.68/0.90/ 1.22/0.64
GLUE: sub-basin calibration	-0.80/0.79/ 0.62/1.84	-0.83/0.37/ 0.78/0.73	0.32/0.78/ 0.68/1.13	0.48/0.76/ 1.15/0.85	0.85/0.93/ **1.02**/0.83	0.66/**0.**91/ 1.25/0.68
EnCDA: whole basin calibration	-0.67/0.40/ 0.73/1.13	-7.14/0.37 /1.56/0.93	-4.57/0.58/ 2.23/0.59	0.63/0.87/ 0.91/0.60	0.26/0.81/ 0.73/0.51	0.56/0.85/ 0.92/0.51
EnCDA: sub-basin calibration	-0.70/0.26/ 1.16/0.86	-9.93/0.47/ 1.68/0.91	-2.69/0.67/ 1.91/0.80	0.66/0.83/ 0.91/1.04	0.43/0.80/ 0.73/**0.94**	0.78/0.90/ **1.03**/0.74
Standard calibration	-0.15/0.73/ 1.00/1.57	-0.75/0.55/ 1.23/0.62	**0.49**/**0.87**/ 0.75/1.55	0.54/0.79/ **0.99**/1.08	0.74/0.86/ 0.98/0.82	**0.81**/0.91/ 1.07/**0.78**
Uncalibrated	-0.09/0.78 /0.95/1.68	-7.52/0.56/ 1.50/1.49	-6.01/0.82/ 1.89/1.54	0.58/0.82/ 1.11/**0.95**	0.71/0.85/ 0.92/0.89	0.78/**0.93**/ 1.15/0.68

[1]Calibration station of standard calibration