The benefits and trade-offs of multi-variable calibration of WGHM in the Ganges and Brahmaputra basins

H. M. Mehedi Hasan[1], Petra Döll[2,3], Seyed-Mohammad Hosseini-Moghari[2], Fabrice Papa[4], and Andreas Güntner[1,5]

[1] Helmholtz Centre Potsdam GFZ German Research Centre for Geosciences, Potsdam, Germany

[2] Institute of Physical Geography, Goethe University Frankfurt, Frankfurt am Main, Germany

[3] Senckenberg Leibniz Biodiversity and Climate Research Centre Frankfurt (SBiK-F), Frankfurt am Main, Germany

[4] University of Toulouse, LEGOS, IRD/CNES/CNRS/UPS, Toulouse, France

[5] University of Potsdam, Institute of Environmental Science and Geography, Potsdam, Germany

Correspondence to: H.M. Mehedi Hasan (mehedi.hasan@gfz-potsdam.de)

Abstract

While global hydrological models (GHMs) are affected by large uncertainties regarding model structure, forcing and calibration data, and parameters, observations of model output variables are rarely used to calibrate the model. Pareto dominance-based multi-objective calibration, often referred to as Pareto-Optimal Calibration (POC), may serve to estimate model parameter sets and analyse trade-offs among different objectives during calibration. Within a POC framework, we determined optimal parameter sets for the WaterGAP Global Hydrology Model (WGHM) in the two largest basins of the Indian subcontinent—the Ganges and the Brahmaputra, collectively supporting nearly 580 million inhabitants. The selected model parameters, determined through a multi-variable multi-signature sensitivity analysis, were estimated using up to four types of observations: in-situ streamflow (Q), GRACE and GRACE Follow-On terrestrial water storage anomalies (TWSA), LandFlux evapotranspiration (ET), and surface water storage anomalies (SWSA) derived from multi-satellite observations. While our sensitivity analysis assured that the model parameters that are most influential for the four variables were identified in a transparent and comprehensive way, the rather large number of calibration parameters, 10 for the Ganges and 16 for the Brahmaputra, had a negative impact on parameter identifiability during the calibration process. Calibration against observed Q resulted to be crucial for reasonable streamflow simulations, while additional calibration against TWSA was crucial for the Ganges basin and helpful for the Brahmaputra basin to obtain a reasonable

simulation of both Q and TWSA. Calibrating also against the other two observation types enhanced the overall model performance and enabled a more accurate representation of the water balance. We identified several trade-offs among the calibration objectives, with the nature of these trade-offs closely tied to the physiographic and hydrologic characteristics of the study basins. The trade-offs were particularly pronounced in the Ganges basin, in particular between Q and SWSA, as well as between Q and ET. When considering the observational uncertainty of the calibration data, model performance decreases in most cases. This indicates an overfitting to the singular observation time series by the calibration algorithm. We therefore propose a transparent algorithm to identify high-performing Pareto solutions under consideration of observational uncertainties of the calibration data. Recognizing these uncertainties, we anticipate that actual model performance may be lower in roughly 90% of cases.

1 Introduction

Global Hydrological Models (GHM), which quantify water fluxes and storage changes on the continents, are essential tools for understanding large-scale water dynamics (Grogan et al., 2022; Gudmundsson et al., 2012), for analysing the impact of humans on freshwater systems (Huang et al., 2015; Döll and Zhang, 2010), for developing scenarios of the future (Gu et al., 2022; Zheng et al., 2018; Giuntoli et al., 2015), and for supporting a sustainable water management (Ai and Hanasaki, 2023; Banda et al., 2022) in a globalized world. Even more than local to regional hydrological models, GHMs suffer from high predictive uncertainties which stem from input data and climate forcing uncertainties, incomplete knowledge about hydrological processes and their imprecise mathematical- description, unknown initial and boundary conditions, and uncertain parameters (Moges et al., 2021). It is state of the art to decrease predictive uncertainties of local to regional hydrological models by estimating model parameters through model calibration, i.e. by comparing the model output to observations and then identifying model parameters that lead to an optimal fit to observations. This is not yet general practice for global hydrological models (Yoshida et al., 2022).

There are many practical challenges of parameter calibration for GHMs. First, a GHM contains thousands of spatially distributed parameters. For example, the WaterGAP Global Hydrological Model (WGHM; Müller Schmied et al., 2021) has more than 30 parameters for each of its 68,420 0.5° grid cells. Second, commensurable observations of most hydrological variables are scarce. While streamflow observations aggregate over the upstream drainage areas, streamflow records are lacking in many regions of the Earth. Other observations, such as

groundwater recharge, are mostly point estimates that are difficult to relate to the behaviour in 0.5° grid cells that cover about 2000 km^2 (depending on the geographic latitude). In addition, observations may suffer from substantial uncertainties that are challenging to quantify (Di Baldassarre and Montanari, 2009). Third, despite the increase in computational power over the last few decades and the availability of high-capacity supercomputing facilities, the high runtimes of current GHMs do not allow most optimization algorithms to explore the high dimensional decision space (i.e., parameter space) in sufficient details, which results in premature termination in most cases before true convergence could be reached (Cheng et al., 2005). Due to these difficulties, GHMs are rarely calibrated at all or are calibrated against streamflow only. In its standard version, WGHM is calibrated in a simple manner against mean annual streamflow observed at more than 1300 gauging stations, by adjusting one to three parameters (Müller Schmied et al., 2021). For most GHMs, the estimation of distributed parameters is accomplished by transferring knowledge from gauged to ungauged basins through parameter regionalization (Beck et al., 2016; Hrachowitz et al., 2013). For example, Arheimer et al. (2020) used daily and monthly streamflow observations at more than 5000 gauging stations in a stepwise calibration approach for the GHM World-Wide HYPE (WWH; HYPE–Hydrological Predictions for the Environment), where process-specific parameters for representative catchments of different physiographic categories were calibrated in each step, and the parameters were transferred to similar catchments worldwide. Similarly, Beck et al. (2020) performed global-scale parameter regionalization for the hydrological model HVB using streamflow observations of over 4000 catchments.

The equifinality thesis proposed by Beven (1993) challenges the notion of a singular optimal model – whether in terms of structure, input, or parameters – particularly in the presence of multifaceted uncertainties. Instead, it suggests that there can be alternative models that exhibit comparable predictive capabilities while differing in their specific configurations. The fundamental causes of equifinality of model parameter sets are the uncertain model structure and inputs (e.g., climate or soil data) as well as the observations (and their errors) that are used to estimate model parameters or evaluate model outputs (Beven, 2006). It is common knowledge that different "optimal" parameter sets would be obtained from calibrations against observations of different periods or if a different model evaluation criterion were used, even though the calibration technique remains unchanged (Beven and Binley, 1992; Kirchner, 2006). Given all these uncertainties, a large number of model parameter sets is "optimal"; it is expected that the number of "optimal" parameter sets increases with the number of parameters that are adjusted. In addition, individual parameters can vary strongly among the "optimal" parameter

sets due to balancing effects among the parameters. For example, in a humid basin, a high soil water storage capacity may decrease streamflow while a low value for a parameter in the equation for potential evapotranspiration may increase it in a similar way; then, the values of each parameter in two equally "optimal" parameter sets can differ strongly, and an "optimal" parameter value cannot be identified. Nevertheless, non-identifiability can also arise when input parameters have little or no impact on the output variable of a model when compared to observations (Herrera et al., 2022). It is assumed that the identifiability of model parameters is enhanced by 1) adjusting only a small number of parameters, those to which model output is most sensitive, and 2) increasing the information content of observations, either by taking into account multiple characteristics (signatures) of the same observation time series or by using observations of more than one model output variable (Bai et al., 2018; Hosseini-Moghari et al., 2020). Jakeman and Hornberger (1993) demonstrated that conventional rainfall-runoff data provide sufficient information to constrain a simple hydrological model with a maximum of four free parameters. Gupta et al. (1998) recognized that parameter estimation for any hydrological model is inherently a multi-objective problem. Observations in addition to streamflow provide information on the behaviour of specific fluxes or storages and constrain parameters better than just streamflow observations.

The basis of any hydrological model is the water balance equation $P = ET + R + \Delta TWS$. That is, the only system input precipitation (P) has to be partitioned into evapotranspiration (ET), runoff (R) and terrestrial water storage change (ΔTWS) during a specific period. Clearly, the prediction accuracy of such a model may be significantly improved if the model could be constrained using observations of all three response variables of the water balance equation. Historically, the streamflow observations alone, i.e., aggregated and routed R from the upstream catchment area, have been used in most model calibration experiments. In the context of multi-objective calibration, Efstratiadis and Koutsoyiannis (2010) recommended a 1:5 or 1:6 ratio between the number of objectives and the number of calibration parameters to optimize parameter identifiability and to facilitate the search algorithm to find a robust solution of the given optimization problem. Developing criteria based on different features of the same observations could potentially increase the ratio of objectives to the number of parameters. However, this approach is not favoured, as the overall information content within any observation dataset is inherently limited. Thus, observations of multiple model output variables become essential to successful calibration (e.g., Denager et al., 2023). Advances in remote sensing technologies and the related generation of data products provide more large-scale information that often is the only source of observation for many data-scarce regions of the

world. The Gravity Recovery and Climate Experiment (GRACE) mission and its successor GRACE Follow-On, for example, provide global observations of terrestrial water storage anomaly (TWSA) starting from April 2002 onwards. Following these ideas of multi-variable parameter estimation, a more detailed calibration of six to eight parameters of WGHM was done by Werth and Güntner (2010) for 28 large basins worldwide using monthly time series of streamflow (Q) and terrestrial water storage anomaly (TWSA) observations following the multi-objective calibration methodology proposed by Werth et al. (2009). Hosseini-Moghari et al. (2020) performed a multi-objective calibration of WGHM parameters for the Lake Urmia basin using three observation variables, streamflow, TWSA, and groundwater storage, after adjusting the model input of human water use using observational data in the first step. More than three observation types have rarely been used for hydrological model calibrations (Meyer Oliveira et al., 2021).

Since TWSA from GRACE mission became available, TWSA observations have been added to in-situ streamflow observations as the measure of storage change (ΔS) in GHM calibration studies (Dembélé et al., 2020; Hosseini-Moghari et al., 2020; Demirel et al., 2019; Bai et al., 2018; Schumacher et al., 2018; Nijzink et al., 2018; Kittel et al., 2018; Rakovec et al., 2016; Milzow et al., 2011; Lo et al., 2010; Werth and Güntner, 2010; Werth et al., 2009). Döll et al. (2024) also used observations of Q and TWSA to calibrate WGHM alternatively for determining pareto-optimal parameter sets for the Mississippi basin as a whole or individually for each of five sub-basins. The whole-basin approach improved the fit to sub-basin observations in all sub-basins as compared to the uncalibrated model (with the exception of one sub-basin for Q). It did not degrade the fit to TWSA for three sub-basins compared to the computationally more demanding sub-basin approach but this was only the case in one sub-basin regarding Q. In contrast, only a few studies have attempted to incorporate global-scale ET products into hydrological model calibration, primarily because of their low reliability and high errors (Liu et al., 2022; Meyer Oliveira et al., 2021; Huang et al., 2020; Nijzink et al., 2018; López López et al., 2017). Demirel et al. (2018) demonstrated successful enhancement of spatial pattern performance in a distributed hydrological model through multi-objective calibration using discharge and remote-sensing-based ET observations. Additionally, Demirel et al. (2024) provide a discussion on the trade-offs between temporal and spatial pattern calibration of the same distributed model using discharge and ET observations. To the best of our knowledge, only a few studies have attempted to simultaneously use all three variables on the right-hand side of the water balance equation to condition a hydrological model (Yang et al., 2022; Dembélé et al., 2020; Livneh and Lettenmaier, 2012). While the study by Huang et

al. (2020) employed streamflow data for bias correction of the ET dataset, they utilized the bias-corrected ET and TWSA for parameterization of a hydrological model, aiming to establish a streamflow-independent calibration scheme. Similarly, in a study by Nijzink et al. (2018), in-situ streamflow observations were utilized to benchmark the performance of five hydrological models across 27 very small European catchments (area < 1600 km2). In that study, ten remote sensing data products, including TWSA and ET, were employed for model calibration, with the exclusion of streamflow observations. In their study, Meyer Oliveira et al. (2021) calibrated a hydrological model using several remote sensing products, including terrestrial water storage anomaly and evapotranspiration, while employing streamflow observations solely for benchmarking. Hulsman et al. (2021) utilized in-situ discharge, satellite-based evapotranspiration (ET), and GRACE Terrestrial Water Storage Anomaly (TWSA) data to calibrate a process-based distributed hydrological model in a large semi-arid basin in Africa, aiming to incrementally improve the process representation of the model. Also, Liu et al. (2022) calibrated 59 large basins worldwide using ET and TWSA observations, with streamflow observations exclusively utilized for validating the calibration results. Trautmann et al. (2018) and Trautmann et al. (2022) calibrated a global model at the scale of selected grid cells with TWSA, ET and a gridded runoff product, instead of using streamflow at the basin scale. For a slightly different purpose, Pellet et al. (2020) utilized observations of all the variables in the water balance equation to derive terrestrial water storage changes by reconstructing the water cycle in five southern Asian basins, including the Ganges and the Brahmaputra river basins.

The terms ‘multi-objective’ and ‘multi-variable’ are not always interchangeable, as multiple objectives can stem from the same variable and multiple variables can contribute to a single composite objective. We use these terms contextually based on their literal meanings. Our multi-objective calibration analyses involve multiple objectives and multiple variables, with one objective corresponding to each variable. A ‘signature’ of a data series consists of quantitative metrics or indices that describe its statistical or dynamic properties (McMillan, 2021). In this context, the term ‘multi-signature’ refers to a scenario where multiple quantitative properties of a data series are considered simultaneously.

In this study, we present a comprehensive multi-objective calibration framework for estimating optimal basin-specific parameter values for a global hydrological model by taking into account observations of multiple model output variables. The framework consists of 1) an approach for selecting model parameters that is based on a global sensitivity analysis and considers multiple signatures of each variable and 2) a multi-objective parameter optimization

that includes multiple variables. We apply the framework to WGHM and estimate, for the Ganges and the Brahmaputra basins of the Indian subcontinent, the most important model parameters using multi-variable multi-signature sensitivity analysis and multi-variable parameter optimization. We then analysed the calibration outcome to answer the following scientific questions.

- How does a multi-variable multi-signature sensitivity analysis enhance the identification of important model parameters?
- Does the inclusion of observations of multiple variables in model calibration increase parameter identifiability and thus reduce model equifinality?
- To what degree does the inclusion of TWSA, ET, and SWSA observations, in addition to Q observations, improve the simulation of important hydrologic variables by a GHM such as WGHM, and what is the value of streamflow observations?
- What is the impact of uncertainties on the calibration outcome? Can we integrate knowledge about observation uncertainties when selecting the so-called compromise solution?

2 Study area

The transboundary basins of the Ganges and Brahmaputra (Fig. 1) exert significant socio-economic, geo-political, and ecological influence in the region. These two basins are home to approximately 580 million human inhabitants and cover an aggregated area of 1.63 million square kilometres shared among India, China, Bangladesh, Nepal, and Bhutan (India-WRIS, 2014b, a; FAO, 2011). With a population density of 355 inhabitants per square kilometer and the necessity to irrigate crops outside the monsoon period, the basins and their inhabitants experience significant water stress. As a result of climate change and the rapid pace of economic growth aimed at lifting a large population out of poverty, the region's water scarcity is expected to intensify rapidly in the coming decades (Gain and Wada, 2014). The Ganges and Brahmaputra rivers collectively account for over 40% of the total freshwater discharge into the Bay of Bengal, which constitutes approximately 25% of the total freshwater inflow received by the Bay of Bengal (Papa et al., 2010). Streamflow in both rivers significantly influences delta formation, sediment deposition, and salinity dynamics in the coastal region (Becker et al., 2020; Akhil et al., 2014).

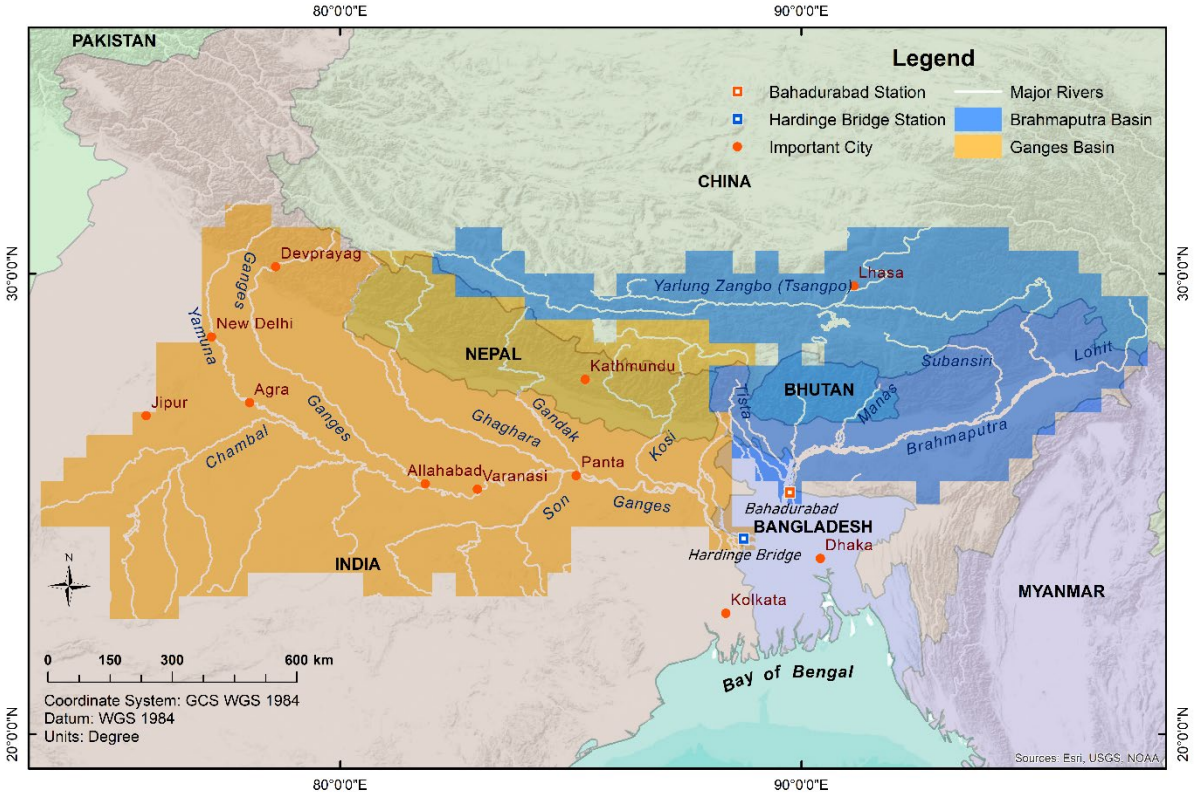

Figure 1: Spatial extent of the two calibration units: the Ganges and the Brahmaputra river basins, delineated as the upstream of the two streamflow gauging stations Hardinge Bridge (Ganges) and Bahadurabad (Brahmaputra)

In the current study, the Ganges and Brahmaputra basins were treated as two distinct calibration units, with calibration parameters adjusted uniformly within each unit. Drainage basins were defined as the upstream areas from the gauging stations at Hardinge Bridge and Bahadurabad, respectively, for the Ganges and Brahmaputra units (Fig. 1). This delineation was based on the drainage direction map DDM30 (Döll and Lehner, 2002). A detailed description of the basin's physiographic properties is provided in the supplementary materials in Sect. S1. Table 1 presents key characteristics of the two basins.

Table 1. Key characteristics of the study basins Granges and Brahmaputra

	Ganges	Brahmaputra
Area [km^2]	1.09 million [b]	543400 [e]
Population [millions]	448 [b, c]	130 [b, d]
Annual precipitation [mm/yr]	760-2290	1347 [a]; 2371 [d], 2143 [f]
Mean summer[1] temperature [°C]	28.7[2] [c]	33.85[2] [d]
Mean winter[1] temperature [°C]	19.6[2] [c]	25.5[2] [d]
Mean annual streamflow [m^3/s]	11300 [e]	20000 [a, e, h]

[1] Apr-Oct considered summer months and Nov-Mar considered winter months

[2] Calculated with data from 1969 to 2004

Sources: (a) Immerzeel (2008) (b) FAO (2011), (c) India-WRIS (2014b), (d) India-WRIS (2014a) (e) Masood et al. (2015), (f) Khan et al. (2015), (g) Ray et al. (2015), (h) Wang et al. (2023)

3 Data and methods

3.1 WaterGAP Global Hydrological Model (WGHM) and forcing data

WaterGAP Global Hydrological Model (WGHM; Müller Schmied et al., 2021, 2014) simulates the continental water cycle to estimate water storage dynamics in ten different storage compartments and water fluxes (ET and streamflow) for all continents (except Antarctica) at 0.5^0 spatial and daily temporal resolution. In this study, we consider the sum of water storage in lakes, wetlands, man-made reservoirs, and rivers as surface water storage (SWS), and the sum of SWS, canopy, snow, soil, and groundwater storage as the terrestrial water storage (TWS). Glacier dynamics could not be taken into account in the WGHM version that was available for this study. For some storage compartments such as lakes and groundwater, WGHM does not simulate absolute values of storage but only storage anomalies such that SWSA and TWSA with respect to a temporal mean over a reference period are analysed, consistent with observations of TWSA and SWSA. The conceptual framework of the model is based on solving the vertical water balance of precipitation, snow accumulation and melt, interception by the vegetation canopy, evapotranspiration, soil water storage and groundwater recharge, and the lateral water movement of generated surface runoff and groundwater outflow through the surface water bodies until it reaches to the ocean or inland sinks. The vertical water balance and the horizontal water movement depend on various geomorphological and physiographic characteristics including soil storage capacity, land cover specific interception capacity and root depth, area of surface water bodies and drainage directions. WGHM account for the impact of man-made reservoirs and human water use on water flows and storages. It is driven by potential net abstractions from groundwater and surface water bodies that are computed by other modules of WaterGAP. For a detailed description of WGHM, see (Müller Schmied et al., 2021).

The WGHM model in its standard version is calibrated for one parameter (the runoff coefficient, SL-RC) against river discharge observations of 1319 gauging stations worldwide such that the simulated long-term mean annual river discharge of the corresponding river basin is within a 10% error range of the observed mean. Upon failure of the above calibration target, two additional correction factors (i.e., the areal correction factor – CFA, and the station correction factor – CFS) are introduced in the standard model version for synthetic runoff

adjustment (Müller Schmied et al., 2021). To suppress these corrections for the calibration experiments in this study, we set both correction factors to 1.0 in all cells. A total of 24 model parameters, including the runoff coefficient SL-RC, were considered in this study (Table 2). The spatial distribution of parameter values is according to one of the following schemes: (U) Uniform parameter value in all 0.5° cells of a river basin; (S) parameter values are specific to sub-areas of the river basin, e.g., in the case of the Priestley-Taylor coefficients, all cells in the arid or humid part of the river basin have the same parameter value, respectively; (M) multiplier parameters are uniform throughout the river basin but multiply the spatially distributed cell-specific values of their base parameter. For example, a value of 1.5 of the river roughness coefficient multiplier (SW-RRM) parameter increases the cell-specific roughness coefficient values by 50% in all cells in a basin. Out of the 24 parameters, two multipliers – the net radiation multiplier (EP-NM) and the precipitation multiplier (P-PM) alter the climate input variables radiation and precipitation, respectively. They were excluded from the sensitivity analysis because of their predominating influence on the target model variables which masks the relative importance of the rest of the parameters. Nevertheless, P-PM was selected as an additional calibration parameter because precipitation forcing data, in contrast to radiation data, contain high uncertainties and biases which need to be corrected during model calibration, if possible. Recently, Goteti & Famiglietti (2024) pointed out the underestimation of precipitation in data sets of India that need to be corrected (here by P-PM) to avoid non-physical or process-based compensation by calibration of other parameters.

Table 2. WGHM parameters with a-priori parameter ranges; spatial scheme U (uniform), M (multiplier), S (sub-area specific) (see text for details). The parameters are categorized according to the storage compartments or processes that they directly affect. P: precipitation, EP: potential evapotranspiration, CA: canopy, SN: snow, SL: soil, SW: surface water, GW: groundwater, NA: net abstraction of water by human

Compartment	Parameter name [units if not unitless]	Acronym	Spatial scheme	Range	Standard WGHM value
P	Precipitation multiplier	P-PM	M	0.5-2	1
EP	Net radiation multiplier	EP-NM	M	0.5-2	1
	Priestley-Taylor coefficient (humid)	EP-PTh	S	0.885-1.65	1.26
	Priestley-Taylor coefficient (semi-arid/arid)	EP-PTa	S	1.365-2.115	1.74

CA	Max. canopy water height [mm]	CA-MC	U	0.1-1.4	0.3
	LAI Multiplier	CA-LAIM	M	0.2-2.5	1
SN	Snow-freeze temperature [°C]	SN-FT	U	-1-3	0
	Snow-melt temperature [°C]	SN-MT	U	-3.75-3.75	0
	Degree-day factor multiplier	SN-DM	M	0.5-2	1
	Temperature gradient [°C/m]	SN-TG	U	0.001-0.01	0.006
SL	Runoff coefficient	SL-RC	U	0.3-3	Variable[1]
	S_{max}[2] multiplier	SL-MSM	M	0.5-3	1
	Maximum EP (mm/d)	SL-MEP	U	6-22	15
SW	River roughness coefficient multiplier	SW-RRM	M	1-5	3
	Active lake depth [m]	SW-LD	U	1-20	5
	Active wetland depth [m]	SW-WD	U	1-20	2
	SW discharge coefficient [1/d]	SW-DC	U	0.001-0.1	0.01
	ET reduction factor multiplier	SW-ERM	M	0.33-1.5	1
GW	GW recharge factor multiplier	GW-RFM	M	0.3-3	1
	Max. GW recharge multiplier	GW-MM	M	0.3-3	1
	Critical precipitation for GW recharge (arid/semi-arid) [mm/d]	GW-CP	S	2.5-20	12.5
	GW discharge coefficient [1/d]	GW-DC	U	0.001-0.02	0.01
NA	Net SW abstraction multiplier	NA-SM	M	-2-2	1
	Net GW abstraction multiplier	NA-GM	M	-2-2	1

[1] Spatially variable among grid-cells

[2] Maximum soil water storage in the effective root zone

WGHM is driven by a climate forcing dataset which is a homogenized combination of WFD (WATCH Forcing Data based on ERA40; Weedon et al., 2011) for 1901-1978 and WFDE5 (WATCH Forcing Data methodology applied to ERA5 reanalysis data; Cucchi et al., 2020) for 1979-2019, with precipitation data being bias corrected using monthly precipitation from GPCC (Global Precipitation Climatology Centre) according to Schneider et al. (2015). The climate forcing dataset includes precipitation, air temperature, downward shortwave radiation and downward longwave radiation.

3.2 Observations

3.2.1 Surface water storage anomaly (SWSA)

Based on multi-satellite observations of surface water extent and water level, Salameh et al. (2017) produced a 15-years data set of SWSA for the Ganges and the Brahmaputra basins by analysing pixel-wise hypsographic curves that represent area-volume relationships. A detailed description of the method can be found in Papa et al. (2013) and Papa and Frappart (2021). Two different Global Digital Elevation Models (GDEM) were used: (i) ASTER (Advance Spaceborne Thermal Emission and Reflection Radiometer) GDEM and (ii) HyMAP (Hydrological Modelling and Analysis Platform) based on SRTM30 (Shuttle Radar Topography Mission) GDEM. Thus, two SWSA observation products were produced for the period 1993-2007 on an equal area (773 km^2) grid of 0.25^0 resolution at the equator. We used the basin-scale monthly mean values of the two products in our analysis. As we considered the river basin area upstream of the last gauging station only, our SWSA basin-scale value for Ganges and Brahmaputra are substantially smaller than those presented in Salameh et al. (2017).

The uncertainties in data products like the one of Salameh et al. (2017) is difficult to assess. Nevertheless, we provide a maximum error estimate from other similar SWS products combining GIEMS and radar altimetry. Frappart et al. (2012) estimated SWS uncertainty of 23% over the Amazon and Papa et al. (2015) estimated the uncertainty to be 24% over the Ganges-Brahmaputra. Based on these two similar estimates we used an error estimate of 25% for the basin-average monthly SWS data in our study.

3.2.2 Actual evapotranspiration (ET)

We used the benchmark ET product LandFlux-EVAL of Mueller et al. (2013) which is a merged synthesis of available global ET products covering observation-based estimations, estimations from several Land Surface Models (LSM), and from atmospheric reanalyses. Many studies have used or compared this product in recent years (Lienert and Joos, 2018; Nanteza et al., 2016; Orth and Seneviratne, 2015; Tsarouchi et al., 2014; Liu et al., 2014). Here, we used the LandFlux-EVAL product that was merged from 14 data sets for the period 1989 to 2005. We used the ensemble mean as monthly ET observation. The standard deviation (σ with N = 14) of the mean is also provided in the dataset as an estimate of the monthly observation error; we used 2-σ range as the uncertainty of the ET observations in our analysis. The errors are provided

in absolute terms with the unit of mm water equivalent for each month, which equivalenting relative terms corresponds to 25% in the Ganges basin and 24% in the Brahmaputra basin.

Table 3. Availability of observation variables and their error estimates

Observation variables	Unit	Period	Estimated error	Source
Streamflow (Q) at the outlet of the basin	[m^3s^{-1}]	1980-2012	20% of monthly value	Masood et al. (2015)
Basin average of total water storage anomaly (TWSA)	[mm]	2003-2012	Propagated 2-σ GRACE errors for each month	Personal communication[1]
Basin average of evapotranspiration (ET)	[mm]	1989-2005	2-σ monthly error	Mueller et al. (2013)
Basin average of surface water storage anomaly (SWSA)	[km^3]	1993-2007	25% of monthly value	Salameh et al. (2017)

[1] Personal communication with the Astronomical, Physical and Mathematical Geodesy Group, Institute of Geodesy and Geoinformation, University of Bonn, Germany

3.2.3 Terrestrial water storage anomalies (TWSA)

The TWSA dataset is based on Level-2 data (Spherical Harmonic Coefficient, SHC) of the GRACE and GRACE Follow-On of TU Graz monthly solutions (Mayer-Gürr et al., 2018) up to degree and order 96, and by applying the anisotropic DDK3 filter (Kusche et al., 2009) to correct the degree-related correlated noise. Further corrections were necessary to eliminate errors related to low-degree effects such as glacial isostatic adjustment (Gerdener et al., 2020). The residual geoid changes due to two large earthquake signals with magnitude over 9.0 (West Coast of Northern Sumatra, Indonesia on 24 Dec 2004 and near the East Coast of Honshu, off Tohoku, Japan on 11 Mar 2011) were removed according to the estimated values following Einarsson et al. (2010). The TWSA data were aggregated to area-average monthly time-series from 2003-2019 for the two study basins. The anomalies were computed by using the mean of the period 2003-2009 as the reference mean.

2-sigma errors based on the full variance-covariance matrix of the TU Graz data, which accounts for orbital effects and the meridional behaviour of errors, were propagated to estimate the uncertainty of the TWSA data. The resultant time series of propagated errors is used to bracket the monthly uncertainty of TWSA observations.

3.2.4 Streamflow (Q)

We use monthly river discharge from 1980-2012 at the Hardinge Bridge and the Bahadurabad gauging stations (Figure 1) derived by Masood et al. (2015), using daily water level observations acquired from the Hydrology Department of the Bangladesh Water Development Board (BWDB) and rating curves developed by the Institute of Water Modelling, Bangladesh (IWM). We assume an error of monthly discharge values of 20% following McMillan et al. (2012) who compared reported streamflow uncertainties and concluded that the uncertainty varies between 10-20% for medium to high in-bank flows, 50-100% for low flows, and over 40% for out of bank flows. Considering the large average streamflow in the two study basins (11,300 m^3s^{-1} in the Ganges and 20,000 m^3s^{-1} in the Brahmaputra basin), we took a pessimistic range of error of 20% which aligns with the estimate of Sir William Halocrow and Partners Ltd. (1991 cited in Mirza, 2003) who reported that uncertainty of streamflow could reach 20% in those stations due to the method of velocity measurement from non-anchored boats and inaccurate measurement of depths of current meters.

3.2.5 Water balance closure of observations

To avoid an ill-posed calibration problem in particular when using all major terms of the water balance equation as forcing (P) or calibration variables (Q, ET, ΔTWS), it is important to check to which extent the water balance is closed in the observations. For this purpose, we calculated the water balance (P – ET – Q – ΔTWS) of the observation data using annual mean values over the available data period. ΔS from GRACE TWSA was computed as the difference between the December values of consecutive years. Observation gaps were filled for all variables through linear regressions for individual months, accounting for seasonality and trends. It is worth noting that the water balance of WGHM is closed at all time intervals.

The mean values of annual precipitation, streamflow, ET, and ΔS for the Ganges basin used in this study as observation data are 1119 mm, 402 mm, 621 mm, and -14 mm, respectively. This results in an annual mean non-closure of the water balance of +109 mm. While reconciling the water budget for the entire India, Narasimhan (2008) argued that ET estimation in India is significantly underestimated. However, in a recent study, Kushwaha et al. (2021) estimated ET for the Ganges basin to be in the range of 511-622 mm yr-1, which closely aligns with our observed mean. In the Brahmaputra basin, the discrepancy is even more pronounced (-436.6 mm). The mean values of P, Q, ET, and ΔS from GRACE for the Brahmaputra basin are 1490 mm, 1361 mm, 564 mm, and -11 mm, respectively. In the case of the Brahmaputra basin, input precipitation is significantly underestimated (Schneider et al., 2017; Michailovsky et al., 2013),

a phenomenon attributed to convective rainfall (Bookhagen and Burbank, 2006) resulting from pronounced differences in basin topology. In their study, Schneider et al. (2017) used a scaling factor of 1.4 to correct the input precipitation and achieve a reasonable water balance closure, while Michailovsky et al. (2013) applied a factor of 1.25 to scale the TRMM 3B42 precipitation data for the Brahmaputra basin in their work. This also underscores the importance of using the WGHM parameter P-PM, a multiplicative factor to adjust the precipitation amount of the forcing data, in the calibration experiments.

It is important to note that mismatches in water balance can also occur in data-rich regions. For example, Rakovec et al. (2016) reported water balance closure errors (P – Q – ET) in some European basins, ranging from -200 mm to 100 mm per year for most of the 179 considered basins.

3.3 Sensitivity analysis

We used the multi-start perturbation Sensitivity Analysis (SA) method of Morris (1991), also known as Elementary Effect Test (EET). Elementary Effect (EE) is expressed as the derivative of a response variable with respect to change in a parameter. The EET method measures the sensitivity to a parameter as the average of elementary effects (EE) at many locations of the parameter space. The sensitivity index (SI) of i^{th} parameter ($i \in \{1,2, \dots m\}$; m denotes total number of parameters) is calculated as:

$$SI_i = \frac{1}{r}\sum_{j=1}^{r}\frac{f(\theta_{per}) - f(\theta_{ref})}{\Delta_i^j} \times C_i$$

$$= \frac{1}{r}\sum_{j=1}^{r}\frac{f(\theta_1, \theta_2, \dots \theta_i + \Delta_i^j, \dots \theta_m) - f(\theta_1, \theta_2, \dots \theta_i, \dots \theta_m)}{\Delta_i^j} \times C_i \quad (1)$$

where, r is the total number of EEs at random locations of the parameter space; Θ_{ref} and Θ_{per} ($\Theta_{ref} \in \mathbf{R}^m$; $\Theta_{per} \in \mathbf{R}^m$) are respectively a reference parameter set and a perturbed parameter set where only i^{th} parameter being perturbed from the reference parameter set; Δ_i^j ($j \in \{1,2, \dots r\}$) is the amount of change in i^{th} parameter at j^{th} location ($j \in \{1,2, \dots r\}$); $f(\Theta)$ is the model response of parameter set Θ ($\Theta \in \mathbf{R}^m$); and C_i is the scaling factor of i^{th} parameter. The scaling factors (C_i) correspond to the range of the respective parameter values (Table 2) and facilitate inter-parameter comparisons in parameter ranking, for instance, as the parameters mostly have differing units and ranges. While the Morris method does not explicitly show interaction terms, it produces a variance term for the elementary effect that accounts for parameter interactions

and the functional non-linearity of the model response. We computed the standard error of the sensitivity index from this variance term and used it for parameter selection (Algorithm 4 in section S2 of the supplementary material).

As a measure of change in the model response (i.e., $f(\Theta_{per}) - f(\Theta_{ref})$), we used the Root Mean Squared Deviation (RMSD) between simulated values of a response variable using a reference parameter set (Θ_{ref}) and simulated values of that variable using the perturbed parameter set (Θ_{per}). The sensitivity index of the EET method averages out the local influences by taking samples from many locations in the parameter space, making it a global sensitivity analysis method (Pianosi et al., 2016). The method is computationally inexpensive and recommended for ranking and screening purposes by Pianosi et al. (2016). 1000 random reference samples were taken using Latin Hypercube Sampling (LHS) which were then perturbed one-at-a-time based on radial-design (Campolongo et al., 2011). We used SAFE MATLAB toolbox developed by Pianosi et al. (2015) for sampling and later computing the sensitivity index. We included 22 WGHM model parameters during the SA from Table 2 (excluding EP-NM and P-PM as stated earlier). During the SA, a total of 23,000 samples were analysed for each of the river basins. Model simulations were conducted for the period 1990-2019, with the spin-up period from 1985 to 1989. The initial year of the spin-up was run five times to allow water storages to reach an equilibrium state.

Parameter sensitivity differs among the response variables and their statistics, i.e., hydrological signatures. To identify parameters that are important for characterizing different features of the target response variables, i.e., those against which the model will be calibrated, we performed a multi-variable multi-signature sensitivity analysis on the four variables with available observations (Q, ET, TWSA, SWSA), considering four signatures, 1) the continuous monthly time series (MTS), 2) the 'climatology' or seasonality, i.e., the 12 mean monthly values, averaged over the study period (MM), 3), the time series of annual means (ATS), and 4) the time series of the seasonal amplitudes computed as the difference between the largest and the smallest monthly values of a year (SNA). The sensitivity indices for each signature were computed separately. We observed that the sensitivities of the four response variables to the individual parameters as well as the share of cumulative effect of top-ranking parameters to the "total effect" (sum of sensitivity indices for all parameters) vary considerably among the response variables (Figure 2). Thus, we decided to select, for each response variable, those top-ranking parameters that together contribute at least 50% of the combined total effect. Application of this threshold ensures that (i) only the most influential parameters for a given

signature of a given variable are selected; and (ii) the total number of selected parameters does not become very large.

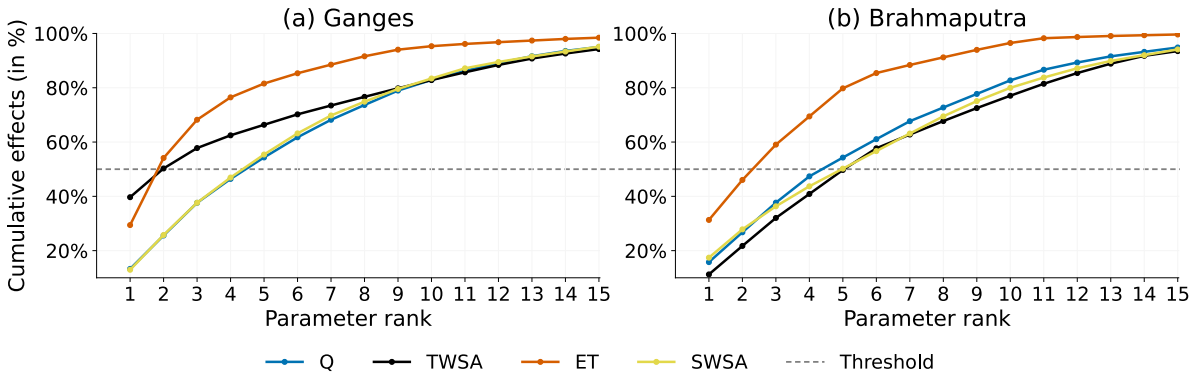

Figure 2. Cumulative effect in percent of the "total effect" (sum of all effects of all parameters) of top-ranked parameters up to different cut-off ranks. The function of the cumulative effect and the cut-off level differ among the four response variables Q, TWSA, ET, and SWSA in the two basins – Ganges (a) and Brahmaputra (b). Sensitivity to parameters in this example was for the monthly time-series (MTS) of the four target response variables. The grey line indicates the cut-off threshold at 50% of the "total effect" that must be surpassed by the top-ranked parameters for each variable and for each signature. For example, the cumulative effect of the three highest-ranked parameters on TWSA in the Ganges basin accounts for 58% of the total combined effects of all parameters on this variable, while the cumulative effect of the top three parameters on Q is only 38% of the total combined effects of all parameters in the basin.

3.4 Calibration

We used the Borg Multi-Objective Evolutionary Algorithm (Borg-MOEA; Hadka and Reed, 2013) to identify non-dominated Pareto-optimal parameter sets of WGHM against one to a maximum of four objectives. A parameter set is considered non-dominated if it outperforms all other competing sets in at least one objective.

The Borg-MOEA has been successfully used for hydrological model calibration in many studies (Fernandez-Palomino et al., 2020; Chilkoti et al., 2018) because of its superior performance over many state-of-the-art multi-objective algorithms (Reed et al., 2013; Hadka and Reed, 2013). The critical features of the Borg-MOEA include amalgamation of multiple (six) search operators and strategies from benchmark optimization algorithms (e.g., NSGA-II of Deb et al. (2002), ε-MOEA of Deb et al. (2005), ε-NSGA-II of Kollat and Reed (2006), GDE3 of Kukkonen and Lampinen (2005), and others), an auto-adaptive recombination mechanism for search operators based on operators' success rates of producing non-dominated solutions over time, a restart mechanism upon detection of a search stagnation, and straight-forward adaptation of the algorithm in parallel computation framework (Reed and Hadka, 2014). Except for the initial population size, all algorithmic parameters were kept to their

recommended values of Hadka and Reed (2013). The ε-precision level for all objectives was set to 0.005 to obtain a detailed Pareto front consisting of a high number of solutions. We deployed a master-slave parallel implementation of Borg-MOEA and ran the algorithm on 401 nodes of a cluster machine operating under Scientific Linux 7 environment. Related to this configuration, the initial population size 400 was used which was equal to the number of slave processes.

The objectives are to maximize values of the Nash-Sutcliffe Efficiency (NSE; Eq. 2; Nash and Sutcliffe, 1970) of streamflow (NSE_Q), terrestrial water storage anomaly (NSE_{TWSA}), evapotranspiration (NSE_{ET}), and surface water storage anomaly (NSE_{SWSA}).

$$NSE = 1 - \frac{\sum_{t=1}^{T}(sim_{(t)} - obs_{(t)})^2}{\sum_{t=1}^{T}(sim_{(t)} - \mu_{\text{obs}})^2} = 1 - \frac{MSE}{\sigma_{obs}^2} \quad (2)$$

where $sim_{(t)}$ and $obs_{(t)}$ are the simulated and observed monthly values at time step t respectively, and μ_{obs} is the mean of the observations; σ_{obs} is the standard deviation of observations and MSE is the Mean Squared Error. NSE serves as a good indicator for inter-basin comparison since it normalizes the MSE by the observed variance (Livneh and Lettenmaier, 2012). While four signatures (MTS, ATS, MM, and SATS) analysed by the SA, we restricted the parameter estimation to using the monthly time series data (MTS) as observations, which has the least aggregation of temporal information among the four. However, the calibration parameters were selected considering the sensitivities of all four signatures (Sect. 4.1).

Fifteen calibration experiments were carried out for each of the two basins, covering all possible combinations of objectives. Each experiment was repeated eight times with different initial populations generated by varying random seeds. The experiments and their objective(s) are listed in Table 4. The maximum number of model runs was limited to 20,000, which proved sufficient for approximating the Pareto front (PF), representing the frontier formed by the set of non-dominated parameter sets. This adequacy is evident from the relatively small difference in PFs between 10,000 and 20,000 model runs (as shown in Figure 3a for a two-objective case), and the stabilization of the mean objective value of the compromise solution occurring well before reaching 20,000 runs in most experiments (Figure 3b).

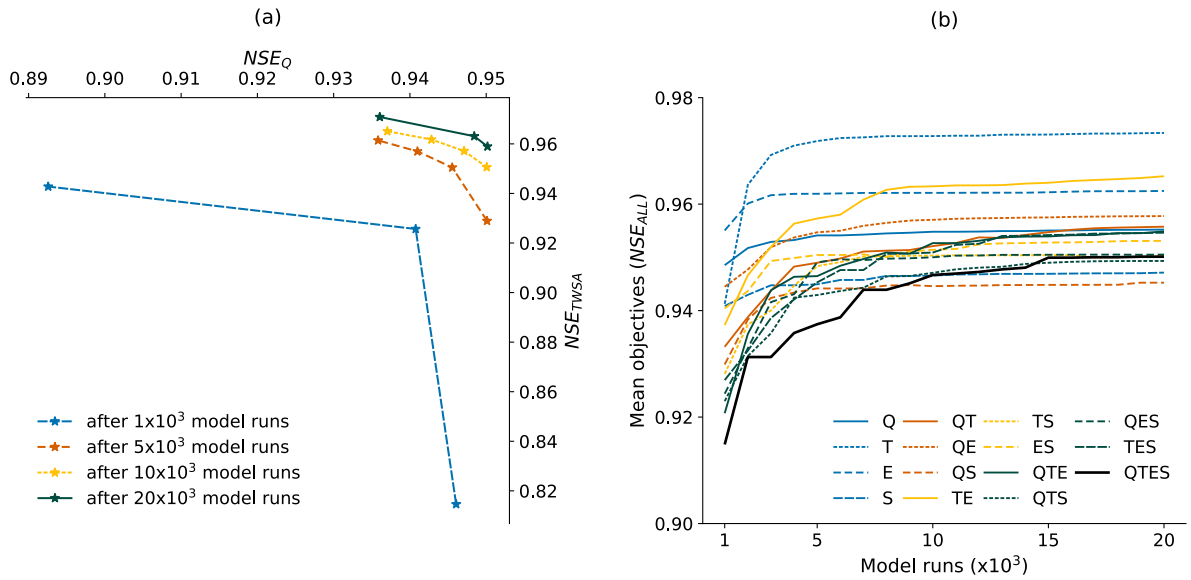

Figure 3. Convergence of single and multi-objective calibrations of the Brahmaputra basin. (a) Pareto fronts of a 2-objective calibration experiment (Experiment-QT with objectives NSE_Q and NSE_{TWSA}) after one, five, ten, and twenty thousand model evaluations. (b) the mean objective value (NSE) of the compromise solution of all calibration experiments as a function of no. of model evaluations. The Pareto fronts and the compromise solutions have been determined after merging all solutions in all replications.

The "compromise" solution or parameter set is said to have the "best" overall performance among the non-dominated solutions, and it is determined by finding the solution with the lowest Euclidian distance (ED, Eq. (3)) in the objective space from the point of theoretical best values of the objectives, known as the "utopia" point. Separate compromise solutions were determined for each replication of an experiment, and an "overall compromise solution", which is the result of merging all solutions from the eight replications, was also determined.

$$ED = \sqrt{\sum_{i=1}^{n}(U_i - O_i)^2} \tag{3}$$

where n is the number of objectives, and U_i and O_i represent respectively the best value of the i^{th} objective and the i^{th} object of a solution parameter set.

The observation datasets available for this study cover different periods (Table 3). The only overlapping period of all four observables are the three years of 2003 to 2005. We considered this period insufficiently short for calibration. Thus, we used observations in partly non-overlapping periods for model calibration while still trying to include overlapping data sets as far as possible. The calibration period was set to 1980-2009. Like in the sensitivity analysis, the model run started 5 years before the start of the calibration period; additionally, the first

year was repeated 5 times. The initial 5 years (1975-1979) were considered as the spin-up period.

3.5 Validation

The validation period was set to 2010-2012, while using the same start time (year 1975) of model runs as for calibration. Validation covered Q and TWSA only as no observations for validating ET and SWSA were available for this period. For validation, we also used several performance metrics including Root Mean Squared Error (RMSE), Mean Absolute Deviation (MAD), Pearson correlation coefficient (r), and the Nash and Sutcliffe efficiency (NSE) for the four signatures MTS, MM, ATS, and SNA. Furthermore, a thorough visual inspection of the simulation results was also performed.

3.6 Uncertainty estimation

To account for the uncertainty of the observation data in the calibration results in its entirety, we would need to repeat the calibration multiple times with alternative realizations of the observation time series. This is not feasible given its high computational demand. Alternatively, Werth and Güntner (2010) defined an error ellipse around the compromise solution after the regular calibration against the original observation time series. For defining the length of one axis of the ellipse, they generated 5000 perturbed observation time series of the observable according to its assumed error characteristics and calculated the performance indices of the simulated time series for each of them. The axis length was then determined by the standard deviation of the performance indices. We recognized that the method of Werth and Güntner (2010) does not consider the uncertainty of the compromise solution itself given the fact that a different solution will probably result as the compromise solution if the parameter search by the algorithm starts from a different starting location, a different initial population is used, or simply a different realization of the observational time series was used in the calibration. With the change of the compromise solution, the axis lengths of the error ellipse are also expected to change. In our analysis, we employed a Monte Carlo process to generate 1000 realizations of observation time series for each variable, taking into account the given uncertainty range. Subsequently, we computed the objective values (NSE) for all variables using the 1000 observation time series for each variable, separately for each of the eight compromise solutions (with one compromise solution per replication). We established thresholds for each objective to extract high-performing solutions from the combined set of solutions across all replications and referred to them as “acceptable” Pareto solutions, accounting for observation uncertainty. These solutions can be viewed as “equivalent” to the compromise solution in the context of

uncertain observations. By applying thresholds to subset solutions, we effectively delineate a hyperrectangle in the objective space, which is conceptually similar to the error ellipsoid used by Werth and Güntner (2010).

4. Results and discussion

4.1 Parameter importance

The sensitivity to parameters varies among the response variables in the two river basins and in many cases, it also varies among the different signatures of a response variable (Figure 4). The response variables, especially streamflow (Q) and TWSA, represent an aggregate response of many complex processes over various temporal and spatial scales. Thus, they are often sensitive to parameters associated with many storage compartments or processes (Table 2) such as ET, soil (SL), surface water (SW), groundwater (GW), snow (SN) (predominantly in the Brahmaputra basin), and net abstraction (NA) by human water use (mainly in the Ganges basin). In addition to the SW parameters, SWSA is highly sensitive to one soil parameter (SL-MSM), a few groundwater-related parameters (GW-RFM, GW-MM and GW-DC) with varying importance depending on the considered signature, two snow parameters (SN-MT, SN-TG) and the ET parameter EP-PTh only in the Brahmaputra basin. ET is computed as the sum of evaporation and transpiration from canopy, snow, soil, and surface water bodies. However, the soil component dominates total ET, and ET is highly sensitive to the parameter SL-MSM which governs the soil water storage capacity. ET is also sensitive to the snow melt (SN-SM) in the Brahmaputra basin as apparently sublimation in the basin contributes substantially to total ET. Apart from these storage parameters, the EP-PTh which scales potential ET in the humid zone highly influences the simulated actual ET.

Several parameters influence most or all response variables across various signatures. However, certain parameters affect only one or two signatures of the response variables. For instance, the Runoff Coefficient (SL-RC) – which is one of the parameters considered in the standard WGHM calibration – significantly influences monthly means (MM) of ET in the Ganges basin and MTS of streamflow. Similarly, the snow melt temperature (SN-MT) is important for some cases in snow-dominated catchments in the Brahmaputra basin. These parameters may also affect other response variables and signatures to some extent but do not meet the defined threshold for calibration selection (Figure 4) The relative contributions of all parameters to all response variables and signatures are presented in Table S2 and Table S3.

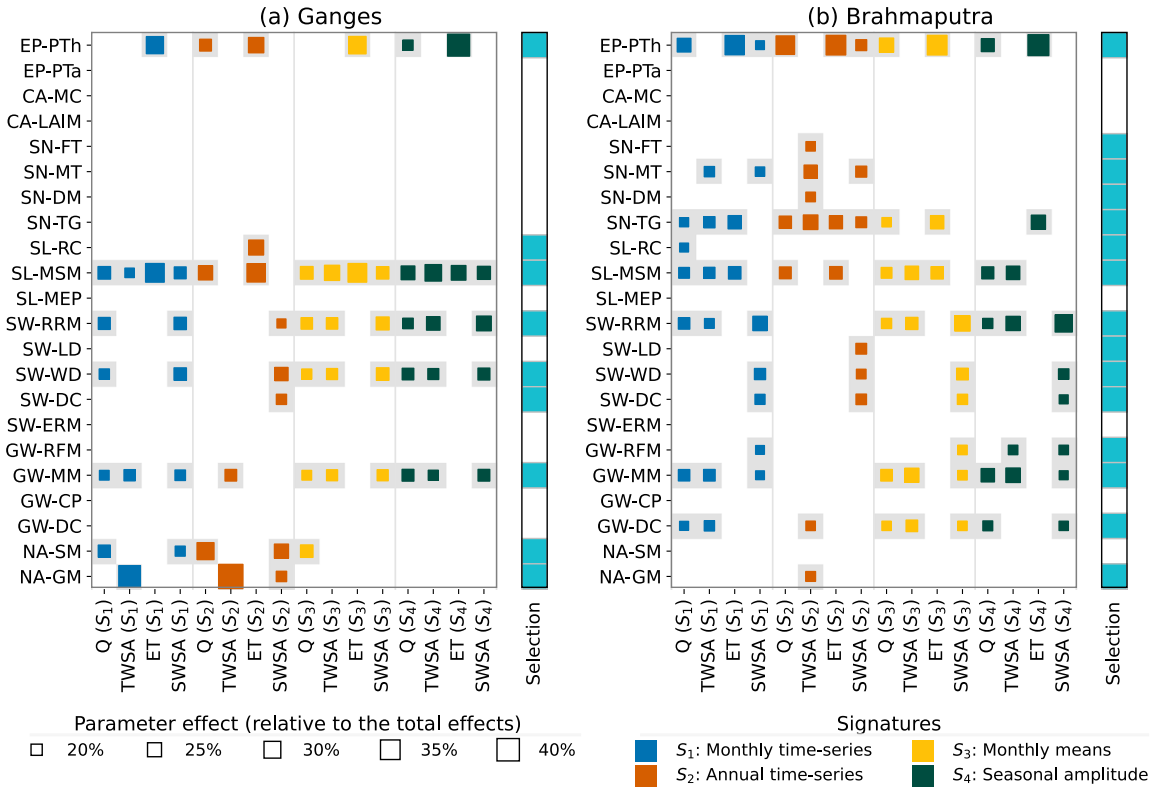

Figure 4: Most influential parameters for the Ganges (a) and the Brahmaputra (b) river basin based on the sensitivity of four signatures S_1 to S_4 - monthly time-series (MTS; blue), monthly means (MM; gold), annual time series (ATS; orange), and seasonal amplitude (SNA; darkgreen) of simulated Q, TWSA, ET, and SWSA. The size of each box represents the effect of a parameter relative to the "total effect", i.e. the sum of the effects of all parameters. The final set of calibration parameters with a significant impact on any signature of the four variables is shown on the right of each plot (cyan boxes). For parameter abbreviations, see Table 2.

Based solely on sensitivity to streamflow and TWSA, we identified seven influential parameters in the Ganges basin and twelve influential parameters in the Brahmaputra basin (Figure 4). Additionally, three SW parameters in the Brahmaputra basin and one in the Ganges basin were selected due to their significant impact on SWSA. Furthermore, one additional parameter in the Ganges basin was found to be sufficiently influential on ET and was included as a calibration parameter. After including the P-PM parameter for both basins, we selected 10 WGHM parameters for calibration in the Ganges basin and 16 parameters in the Brahmaputra basin.

The use of multiple signatures from various variables ensures that key parameters governing all critical hydrological processes in the model are identified. For instance, if only one signature were considered, 5-9 parameters in the Ganges basin and 9-12 parameters in the Brahmaputra basin would have been selected for calibration. Similarly, if parameter sensitivity was assessed based on Q or TWSA only, influential parameters governing other important

observables could have been overlooked. However, the method of parameter selection is not without challenges. Parameters with significant impacts may be excluded if the cut-off threshold (e.g., 50% of the total effect, as used in this study) is surpassed by only a few top-ranked parameters. For example, in the Brahmaputra basin, despite contributing a substantial 12% to the total impact, the parameter SL-RC deemed non-influential for ET according to this threshold (Table S3). Raising the threshold would result in the selection of a larger number of parameters, potentially leading to an unnecessary expansion of the decision space. This could increase computational demands and exacerbate issues of equifinality (Sect. 4.2.3).

4.2 Model Calibration

Calibration experiments with all 15 possible combinations of the four objectives (NSE_Q, NSE_{TWSA}, NSE_{ET}, and NSE_{SWSA}) were carried out for the two study basins. Furthermore, each of the experiments was repeated 8 times with random seeds, resulting a total number of 240 calibrations. Overall, the study involved the evaluation of 4.8 million samples, requiring approximately 3.2[1] million CPU hours of model run time.

Table 4: Configuration of the 15 calibration experiments, with the observed variable(s), number of objectives, number of replications, minimum and maximum number of non-dominated (Pareto-optimal) solutions (i.e. estimated parameter sets) obtained among the 8 replications, and the total number of non-dominated solutions over all 8 replications.

					No. of non-dominated solutions			
					Ganges		Brahmaputra	
Experiment	Name	Observed variable(s)	No. of Objs.[a]	Repli-cations	Min/Max	Total	Min/Max	Total
(1)	Q	Q	1	8	1/1	8	1/1	8
(2)	T	TWSA	1	8	1/1	8	1/1	8
(3)	E	ET	1	8	1/1	8	1/1	8
(4)	S	SWSA	1	8	1/1	8	1/1	8
(5)	QT	Q, TWSA	2	8	3/5	34	2/3	20
(6)	QE	Q, ET	2	8	5/6	45	1/1	8
(7)	QS	Q, SWSA	2	8	10/13	94	1/2	15
(8)	TE	TWSA, ET	2	8	6/7	49	1/2	10
(9)	TS	TWSA, SWSA	2	8	14/17	123	1/4	18

[1] The execution time for a single run of the WGHM model was approximately 40 minutes but exhibited significant variations depending on the specific CPU used and the concurrent I/O traffic on the cluster machine during the model run

(10)	ES	ET, SWSA	2	8	7/9	63	1/2	10
(11)	QTE	Q, TWSA, ET	3	8	106/146	1006	6/11	60
(12)	QTS	Q, TWSA, SWSA	3	8	402/472	3526	3/11	63
(13)	QES	Q, ET, SWSA	3	8	546/637	4712	2/16	30
(14)	TES	TWSA, ET, SWSA	3	8	65/84	616	3/12	63
(15)	QTES	Q, TWSA, ET, SWSA	4	8	1031/1155	8705	20/77	339

[a] NSE used as calibration objective

4.2.1 Added value of multi-objective calibration and trade-offs among objectives

A high cardinality, i.e., a high number of solutions in the non-dominated Pareto solution set, was obtained in most multi-objective calibrations (Table 4). The cardinality depends on the shape of the Pareto frontier (PF) and the allowed crowding distance, which was constant (0.005) for all objectives in all experiments. A wider PF resulting in high cardinality reflects a high trade-off between the objectives. The high cardinality observed in the Ganges experiments indicates marked trade-offs among objectives, especially between NSE_Q and NSE_{SWSA}, as well as between NSE_{SWSA} and NSE_{TWSA}. This observation is further supported by the experiments involving solely those objective pairs, which yielded a larger number of solutions. The 3-objective calibration TES has a lower cardinality than the other 3-objective cases, which indicates the simulation of Q is in a rather strong conflict with the simulation of the three other variables. This is supported by the poor fits to streamflow observations of the TES calibration variant for both basins (Table 5 and Table 6). As expected, the 4-objective calibration produced the highest number of non-dominated solutions.

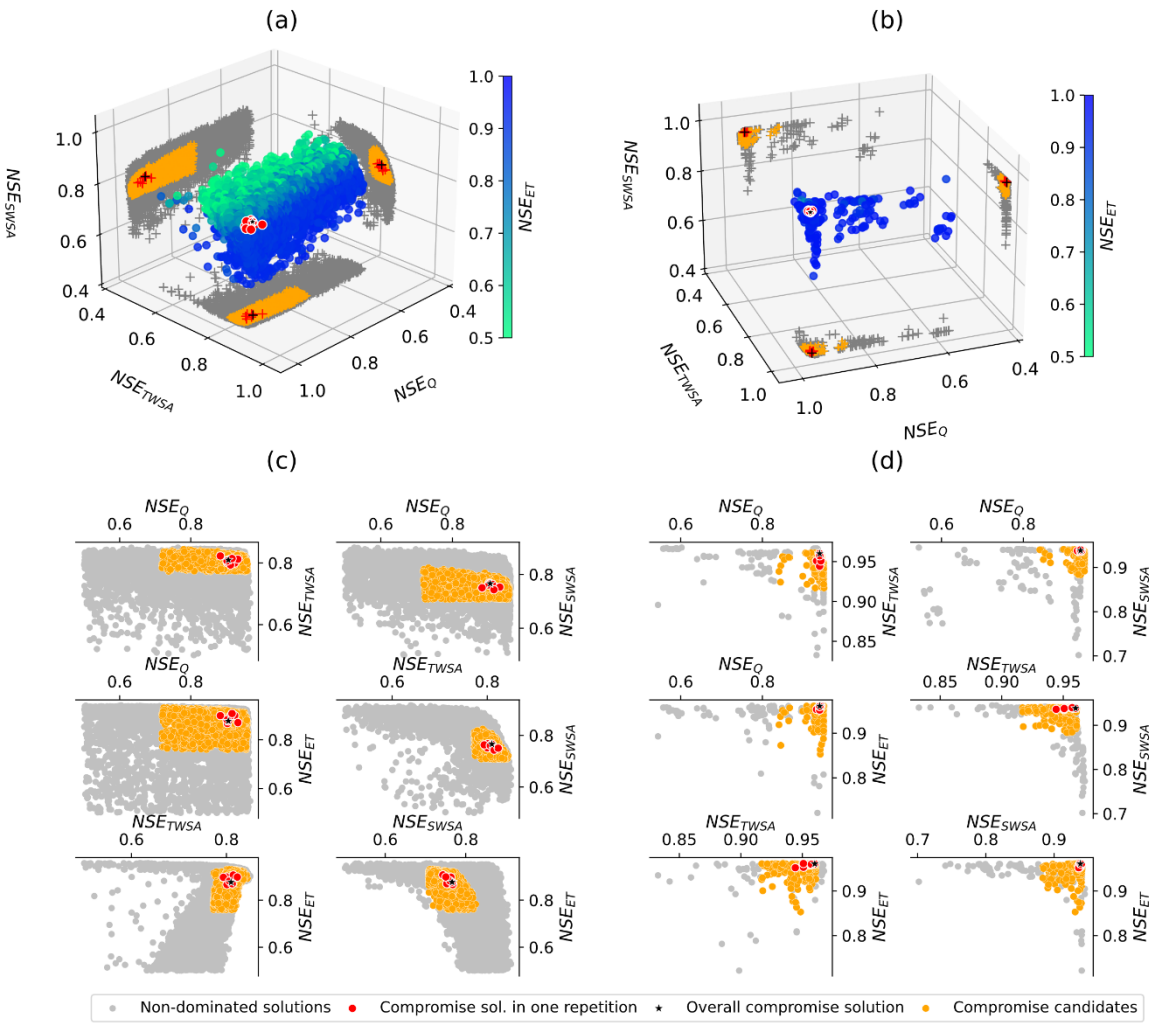

Figure 5: The Pareto front of 4-objective calibration experiments of the Ganges basin and for both the Brahmaputra basin. The 3-D view of the 4-D PF of Ganges (a) and Brahmaputra (b), while the 4th dimension is colour coded. Only solutions having NSE greater than 0.5 are shown. The 2-D projection is shown with crosses. The bottom row shows the 2-D view of each pair of objectives for the Ganges (c) and the Brahmaputra (d) basin. All non-dominated solutions are shown in grey, the compromise solution of each replication in red, the overall compromise solution in black, and the candidate compromise solutions in orange.

The single-objective calibration experiments obtained the best NSE values for the specific objective under consideration (Table 5 and Table 6). The mean NSE of all four objectives ($\mu_{NSE,ALL}$) was used as a simple indicator of the overall performance of an experiment. In multi-objective calibrations, although the objective values for each individual objective decrease slightly, the overall performance tends to increase when more objectives are included. In the Brahmaputra basin, the highest $\mu_{NSE,ALL}$ increases from 0.84 for single-objective calibration to 0.90 for 2-objective calibrations, to 0.93 for 3-objective calibration and to 0.95 for 4-objective calibration. In the Ganges basin, the highest $\mu_{NSE,\ ALL}$ is slightly smaller in 3-objective calibrations than in 2-objective calibrations. Nevertheless, the 4-objective calibration

experiments achieved the highest overall performances in the basin. In their study, Livneh and Lettenmaier (2012) demonstrated that the overall performance of the calibrated model improved with the inclusion of ET and TWSA observations in addition to streamflow observations. In contrast, Mei et al. (2023) observed a reduction in the overall performance of a three-objective calibration, including observations of Q, soil moisture, and ET, when compared to single and two-objective calibrations. This reduction was attributed to suspected model structural errors and/or erroneous observations.

Different from the Brahmaputra, calibration against only Q in the Ganges basin resulted in worse fits to all three other variables as compared to the uncalibrated model version. Multi-variable calibration, however, works best if streamflow observations are included. Excluding NSE_Q as an objective in any calibration resulted in significantly poorer performance in streamflow simulation (Table 5 and Table 6). The importance of streamflow observations in model calibration is well documented in the literature, with a particular focus on multi-variable calibration scenarios (Dembélé et al., 2020; Livneh and Lettenmaier, 2012). Liu et al. (2022) reported that calibrating model with ET and TWSA observations can occasionally produce reasonable streamflow simulations in certain basin. In their study, Livneh and Lettenmaier (2012) concluded that calibrating the model with either ET or TWSA alone was insufficient to achieve good performance in streamflow simulation. We also discovered that calibrating with these two variables resulted in high NSE_Q (> 0.8) only in a few replications in the Brahmaputra basin.

Table 5: Mean and standard deviation of model performance indicator NSE for the compromise solutions (N = 8) of the calibration experiments in the Ganges river basin during calibration period. The WGHM model was rerun using parameters from the compromise solutions to compute NSEs of all variables. The $\mu_{NSE, ALL}$ represents the mean NSE across all objectives over all eight compromise solutions per experiment. The highest NSE for each objective is highlighted using bold face, also the highlighted mean across objectives ($\mu_{NSE, ALL}$) show the highest value in each group (2-objective, 3-objective, and 4-objective). The objective obtained in the standard calibration and in the uncalibrated model is also shown.

	Mean ± Std. Deviation				
Experiment	NSE_Q	NSE_{TWSA}	NSE_{ET}	NSE_{SWSA}	$\mu_{NSE, ALL}$
Q	**0.97± 0.002**	-4.07± 3.572	0.54± 0.135	0.56± 0.088	-0.5
T	0.70± 0.030	**0.85± 0.001**	0.89± 0.007	0.63± 0.021	**0.77**
E	-14.83± 0.739	-2.56± 6.952	**0.96± 0.000**	-4.39± 2.476	-5.21
S	-2.40± 1.453	0.57± 0.095	0.07± 0.064	**0.92± 0.001**	-0.21
QT	0.95± 0.003	0.84± 0.001	0.87± 0.001	0.64± 0.008	**0.83**
QE	0.96± 0.001	-5.93± 0.588	0.93± 0.001	-0.09± 0.092	-1.03
QS	0.94± 0.003	-23.73± 5.807	-0.35± 0.035	0.89± 0.002	-5.56
TE	0.52± 0.026	0.85± 0.000	0.93± 0.001	0.61± 0.016	0.73

TS	-2.64± 0.775	0.81± 0.001	0.88± 0.025	0.88± 0.004	-0.02
ES	-6.24± 0.206	0.66± 0.004	0.94± 0.003	0.89± 0.001	-0.94
QTE	0.94± 0.004	0.83± 0.002	0.92± 0.003	0.60± 0.028	**0.82**
QTS	0.93± 0.005	0.79± 0.008	0.63± 0.057	0.80± 0.008	0.79
QES	0.92± 0.013	-13.91± 6.842	0.87± 0.010	0.77± 0.007	-2.83
TES	-3.42± 0.106	0.80± 0.002	0.93± 0.005	0.87± 0.002	-0.2
QTES	0.91± 0.015	0.81± 0.009	0.89± 0.015	0.76± 0.009	**0.84**
Std. Calibration[a]	0.96	0.80	-0.39	0.70	0.52
Uncalibrated[b]	0.84	0.80	0.55	0.76	**0.74**

[a] SL-RC and two correction factors are calibrated by adjusting mean annual streamflow was calibrated against observed values (Müller-Schmied et al., 2021).
[b] SL-RC is set to the default 2.0 and correction factors were set to 1

Table 6: Mean and standard deviation of model performance for the compromise solutions (N = 8) of the calibration experiments of the Brahmaputra basin. $\mu_{NSE,\ ALL}$ represents the mean across the four objectives. The highest objective values in all experiments and in each group are highlighted. Objectives of the standard calibration and uncalibrated model is also shown.

	Mean ± Std. Deviation				
Experiment	NSE_Q	NSE_{TWSA}	NSE_{ET}	NSE_{SWSA}	$\mu_{NSE,\ ALL}$
Q	**0.95± 0.001**	0.74± 0.052	0.79± 0.208	0.86± 0.043	**0.84**
T	0.23± 0.978	**0.97± 0.004**	0.72± 0.335	0.70± 0.166	0.66
E	0.01± 0.589	0.79± 0.108	**0.96± 0.002**	0.66± 0.169	0.61
S	-0.19± 0.709	0.77± 0.065	0.74± 0.178	**0.95± 0.002**	0.57
QT	0.94± 0.003	0.96± 0.003	0.73± 0.221	0.79± 0.073	0.86
QE	0.95± 0.001	0.83± 0.027	0.96± 0.004	0.85± 0.037	**0.9**
QS	0.95± 0.002	0.73± 0.095	0.69± 0.209	0.94± 0.001	0.83
TE	-0.19± 0.965	0.96± 0.003	0.96± 0.005	0.83± 0.024	0.64
TS	-0.12± 0.698	0.96± 0.005	0.71± 0.339	0.94± 0.002	0.62
ES	-0.32± 0.606	0.86± 0.044	0.96± 0.004	0.94± 0.001	0.61
QTE	0.94± 0.003	0.96± 0.003	0.96± 0.001	0.84± 0.065	**0.93**
QTS	0.94± 0.002	0.96± 0.003	0.71± 0.268	0.94± 0.002	0.89
QES	0.95± 0.002	0.87± 0.020	0.96± 0.003	0.94± 0.001	**0.93**
TES	-0.14± 0.590	0.95± 0.003	0.96± 0.002	0.94± 0.003	0.68
QTES	0.94± 0.003	0.95± 0.005	0.96± 0.004	0.94± 0.001	**0.95**
Std. Calibration	0.90	0.77	0.26	0.64	0.64
Uncalibrated	0.72	0.81	0.68	0.57	**0.70**

In comparison to the standard calibration, the 4-objective calibration resulted in better performance in the Brahmaputra for all four response variables, and in all variables except streamflow in the Ganges basin, where the standard calibration leads to a very high NSE_Q of 0.96 (Table 6). As the streamflow simulation for the single objective calibration with Q only is better in the two basins than the standard calibration, this suggests that the slight decrease of streamflow performance in 4-objective calibration in the Ganges basin is due to some trade-offs among the objectives. The improvement in ET by the 4-objective calibration was much

higher than the improvement in other variables which underpins the need to include ET as a calibration variable. This is corroborated by the observation that the standard calibration procedure of Müller Schmied et al. (2021) with Q only degrades the ET simulation even in comparison to the uncalibrated WGHM. This, in fact, contradicts the conclusion of the study by Nijzink et al. (2018), in which they analysed the potential of several remote sensing products to constrain hydrological models and calibrated five hydrological models for 27 small catchments in Europe. They concluded that remote sensing-based ET observations were less effective at adequately constraining the posterior parameter distribution compared to other observations such as soil moisture, TWSA, and snow. One probable cause could be the fact that the catchment size in that study was too small (< 1600 km^2) for the ET products to be effective; in our study, the catchment size is significantly larger.

As mentioned above, multi-objective calibration enhances the overall model performance at the expense of a slight decrease in the performance of individual variables, which is common and often expected. Many studies have reported a reduction in the performance of streamflow simulation when the model is calibrated with streamflow and TWSA observations, as compared to models calibrated solely with streamflow data (Li et al., 2018; Bai et al., 2018; Yassin et al., 2017; Rakovec et al., 2016; Livneh and Lettenmaier, 2012). The trade-offs among other variables are not well-documented in the literature. Mei et al. (2023) compiled a list of the previous studies that incorporated streamflow observations and observations of some additional variables in model calibration. They documented changes in performance in four target variables—streamflow, ET, soil moisture, and TWS—as a result of incorporating additional variables in those studies. In addition to the trade-offs between streamflow and TWSA, we also observed substantial trade-offs between NSE_Q and NSE_{SWSA}, between NSE_{TWSA} and NSE_{SWSA}, and NSE_{SWSA} and NSE_{ET} (Figure 5). The trade-offs among the objectives behave differently in the two basins as the shape of the Pareto Front (PF) of non-dominated solutions differs significantly between the basins. In general, PFs of the Ganges experiments have a smooth curvature with extended spread near the theoretical optimum of the objectives, while the Pareto fronts in the Brahmaputra basin are mostly very steep resembling right angles (Figure 5). Due to the conflicts among objectives, the number of non-dominated solutions in the Ganges basin became much larger than in the Brahmaputra where the trade-offs are much smaller.

The substantial variations of the performance values of the compromise solutions across the calibration repetitions, which can be regarded as the “uncertainty” of the calibration method itself, further complicates the assessment of trade-offs among objectives. Bai et al. (2018) observed inconsistent conclusions in the literature regarding the impact on streamflow simulations when incorporating GRACE data for model calibration in addition to streamflow observations. Some studies reported a "positive" impact, while others reported a "negative" impact. We argue that the source of such inconsistency could be attributed to (i) the failure to account for the uncertainty in the calibration method, (ii) the lack of convergence to the Pareto front, and (iii) an ill-posed problem formulation resulting from the choice of an inappropriate model, non-identifiable parameters, or inadequate data. Thus, the uncertainty of the calibration outcome should be considered whenever possible when discussing trade-offs. While we consider the impact of observational uncertainty in the next chapter of this study, we found here that the uncertainty stemming from the calibration method differed significantly between the two basins. In the Ganges basin, the highest uncertainties were observed in single-objective calibration cases involving only NSE_Q and only NSE_{ET}. Among the 3-objective calibrations, the highest level of uncertainty was observed in calibrations without NSE_{TWSA} and without NSE_Q. High variations were observed in those objectives that were not used in the calibration. In the Brahmaputra basin, among the single objective calibration cases, calibrations with only NSE_{TWSA} and with only NSE_{ET} exhibited the highest level of variation. When one object is omitted from calibration, the calibrations without NSE_Q and without NSE_{ET} generated the highest degree of uncertainty in the objectives NSE_Q and NSE_{ET}, respectively. Probably the most important calibration cases for trade-off analysis are the bivariate cases with two objectives, which in our case exhibit an insignificant level of uncertainty resulting from the calibration method itself.

In the Ganges, calibrating against ET leads to negative NSE values for TWSA and SWSA. This conflict between ET and TWSA results in the only positive NSE_Q value that was obtained in two-objective calibrations without Q. In a single-objective calibration against ET, the calibration algorithm aims at keeping storage as high as possible to ensure that there is enough water for evaporation and there is no penalty for overestimating storage. However, including TWSA forces the algorithm to release some of the storage to achieve a good fit for TWSA, which leads to a better simulation of Q compared to a single calibration against ET. Calibration against ET and SWSA does not improve Q in the Ganges, suggesting that the adjustment of TWSA is likely related to soil storage. In the Brahmaputra basin, the trade-off

between ET and TWSA is very small and the two-objective calibration against ET and TWSA does not improve Q.

4.2.2 Impact of observation uncertainty on the calibration outcome

The uncertainties associated with individual data points in the observation time series alter the values of the performance criteria. We conducted an assessment of this effect by perturbing all observations using their respective uncertainties through a Monte Carlo simulation, resulting in 1000 perturbed time series for each variable. Subsequently, we calculated the objective (NSE) values for the compromise solutions across all eight replications of the 4-objective calibration. The largest deviations of NSE from the reference values, i.e., the values of the compromise solutions with the original observation time series, were found in streamflow for both the Ganges and the Brahmaputra basins (on average 0.26 and 0.12 respectively), followed by ET (0.12 and 0.10) (Figure 6-b, c). Low deviation in the range of 0.05-0.07 was observed for both storage variables (TWSA and SWSA) for which the means are always zero in both the simulation values and observation data. The changes of the objective function values with the perturbed time series may result in different Pareto solutions. For this reason, we propose a mechanism to objectively identify a group of solutions that could be considered alternatives to the compromise solution.

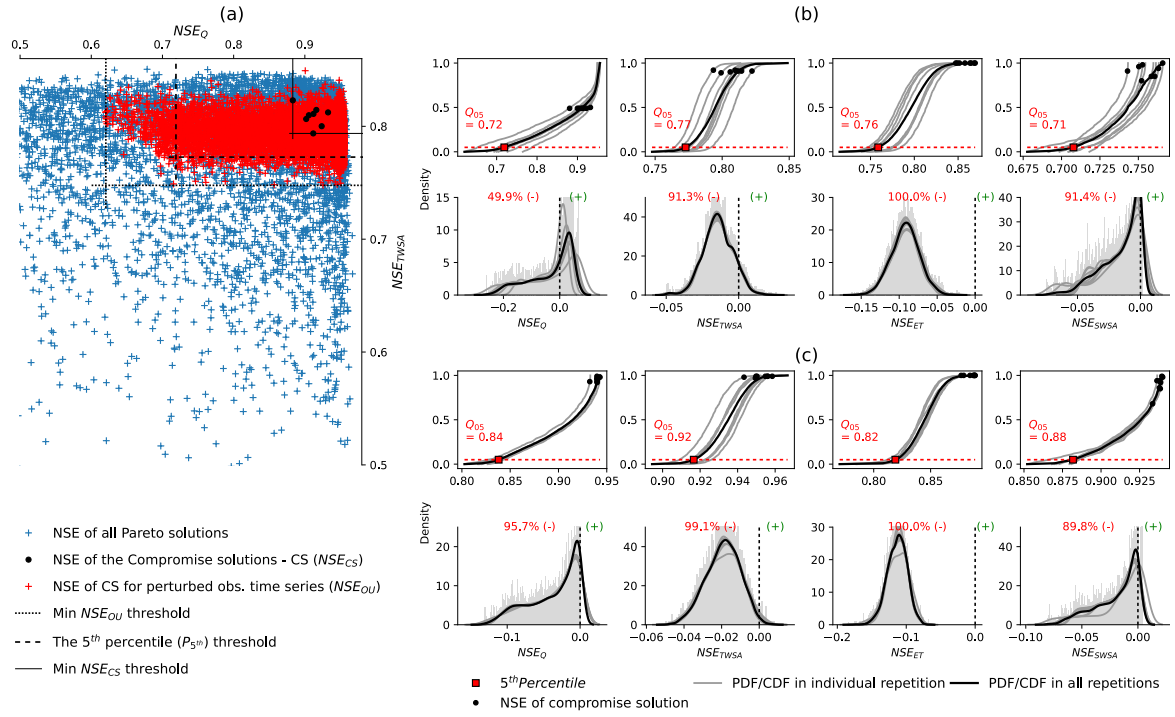

Figure 6: Uncertainty in NSE of the compromise solutions (CS) in the repeated experiments of 4-objective calibration, obtained by propagation of observational uncertainties into objectives. (a) Scatter plot of objective values (NSE) of (i) all Pareto

solutions, (ii) the compromise solutions (NSE_{CS}) and (iii) the compromise solutions computed for 1000 perturbed observation time series (NSE_{OU}) for the Ganges basin. Three thresholds (M2, M3, M4) are visualized as options to delineate the space of compromise solutions under consideration of observational uncertainties (see text for details). Density functions (CDFs and PDFs) of NSE_{OU} in the Ganges basin (b) and in the Brahmaputra basin (c). PDFs represent the deviations $NSE_{OU} - NSE_{CS}$. The black dashed vertical lines in the density plots delineate the zones of NSE decrease and NSE increase. Densities of NSE_{OU} for each compromise solution of a single repetition are plotted in grey and the density function of NSE_{OU} of all compromise solutions are in black. The black dots show the objective values of the compromise solutions.

We tested several objective thresholding methods to delineate the space of solutions that can be considered equivalent to the compromise solution in view of the observation uncertainties. In the following, we name them "acceptable Pareto solutions considering observation uncertainties". The minimum NSE value of all compromise solutions (Min NSE_{CS}) in the repeated experiments is discarded as a threshold (M1) as it represents the uncertainty due to the random start of the parameter search in the calibration algorithm only, but not the observation uncertainties themselves (Table 7). The second threshold (M2) is computed by subtracting the standard deviation of all the objective values obtained with perturbed observation time-series (Std. dev. NSE_{OU}) from Min NSE_{CS}. The third threshold (M3) is computed by subtracting the mean absolute deviation of objective values with perturbed time series from the reference values (MAD $NSE_{OU, REF}$; the objectives of compromise solutions are used as reference values), the threshold is set at Min NSE_{CS} – MAD $NSE_{OU, REF}$. Threshold M4 is the 5th percentile of NSE_{OU} which ensures 95% of NSE_{OU} remain above the threshold. For the most unrestricted option (M5) the minimum value of NSE_{OU} is taken as the threshold.

Table 7: Metrics related to the spread of the objectives of the compromise solutions (CS) for two sources of uncertainties – observation uncertainty and uncertainty due to the calibration method, i.e., random starting population used during the parameter search. Observation uncertainty is propagated by perturbation of the observation with a Monte-Carlo process within the estimated uncertainty bound of an observable and then computing objectives for those perturbed observation time series (NSE_{OU}) (N = 1000).

		Ganges				Brahmaputra			
Source of uncertainty	Metric	Q	TWSA	ET	SWSA	Q	TWSA	ET	SWSA
Optimization algorithm	Min NSE_{CS}	0.883	0.794	0.870	0.743	0.932	0.944	0.952	0.934
	Range NSE_{CS}	0.050	0.029	0.037	0.023	0.011	0.016	0.009	0.006
Observations	Min NSE_{OU}	0.620	0.748	0.715	0.678	0.803	0.894	0.770	0.852
	The 5th Percentile of NSE_{OU}	0.719	0.773	0.759	0.708	0.838	0.917	0.819	0.882
	Std. dev NSE_{OU}	0.080	0.013	0.022	0.017	0.034	0.010	0.015	0.019
	MAD of $NSE_{OU, REF}$	0.063	0.015	0.092	0.016	0.036	0.019	0.113	0.019

In some instances, we found that the standard deviation of NSE_{OU} (Std. dev. NSE_{OU}) and MAD NSE_{OU} are smaller than the range of objectives in the compromise solutions of the repeated experiments. For this reason, we argue M2 and M3 are incapable of distinguishing the objective uncertainty attributed to the observational uncertainty from other sources of uncertainty like the randomness in the calibration method itself. Thus, we rejected them as appropriate thresholds for identifying candidates for the compromise solution, although they could find a reasonable number of good solutions (Table S5). Worth noting, Werth and Güntner (2010) used a similar strategy like M2 for identifying the uncertainty in objectives. We also rejected the least restrictive threshold M5 as a singular extreme low value of an objective can extremely limit the efficacy of the threshold. On the contrary, the threshold M4 holds a balance between restrictedness and efficacy. While it excludes the poor extremes, 95% of good objective values are kept in the final set. Using M4, we obtained over 1400 solutions (16% of the total number) in the Ganges basin and 221 solutions (65%) in the Brahmaputra basin having model performance the threshold. Overall, the performance of the "acceptable" Pareto solutions considering observation uncertainties is seen generally higher with smaller dispersion in the Brahmaputra basin than those of the Ganges basin.

Any perturbation of the ET observation time series within its uncertainty ranges leads to a lower NSE than the reference value in both basins (Figure 6). In all cases, NSE_{OU} is worse than the objectives of the compromise solutions. A similar performance decrease was observed in NSE_{OU} for all variables, except for streamflow in the Ganges basin where only about half of the uncertainty-perturbed time series lead to a decrease in NSE_{OU} while the rest causes NSE_{OU} to increase. The aforementioned indicates that during calibration the parameters are so finely tuned to the (undisturbed) observation time series that any modification of the time series leads to a deterioration of the objective values. On the one hand, this corroborates the strength of the optimization algorithm. On the other hand, it clearly indicates overfitting of the parameters. This raises questions about their usability in scenarios where variations in observations are anticipated such as model predictions in a different time period or when extrapolating parameters for uncalibrated basins (parameter regionalization). The reason why streamflow in the Ganges behaves differently in this regard is not very clear; one probable cause could be that during the choice of the compromise solutions, most high-performing solutions for the streamflow variable were rejected due to low performance in other variables (the maximum NSE_Q in all solutions is 0.97 but max. NSE_Q in the compromise solutions is 0.93).

It is also noteworthy to observe that the shape of the density function for changes in objectives (Figure 6 b, c) is closely associated with the error structure of observations. An average percentage bias was considered as error of the streamflow and SWSA observables, whereas for TWSA and ET absolute errors were assumed. When converted to the percentage error, it was observed that the TWSA and ET observation error has a sinusoidal seasonal structure. In contrast, the constant percent bias in streamflow and SWSA causes high errors in the monsoon season in the perturbed observations and added only very small bias in the dry winter season. This ultimately causes a left skewness to the distribution of deviations of NSE_{OU} for streamflow and SWSA (Figure 6 b,c).

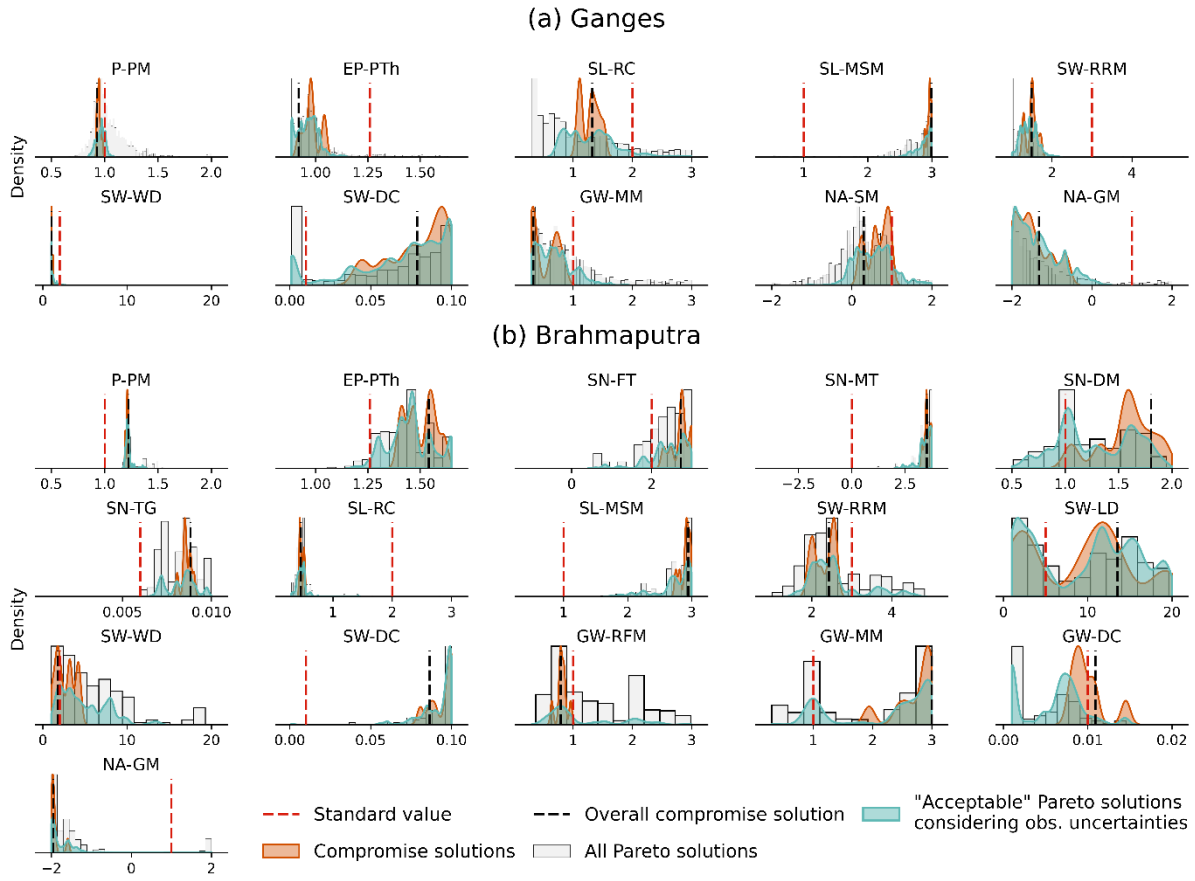

Figure 7: The distribution of parameters of the compromise solutions in the 8 repeated 4-objective calibration experiments (darkorange), and the "acceptable" Pareto solutions considering observation uncertainties (cyan), and all Pareto solutions in all replications (grey). The overall compromise solution (black dashed vertical line) represents the compromise solution among all solutions in all replications. The parameter value of the standard WGHM is shown with the red dashed vertical line.

When comparing the parameter values of the eight compromise solutions (group 1) to those of the "acceptable" Pareto solutions considering observation uncertainties (group 2) (Figure 7), the parameter distributions of these two groups are very similar in most cases, although the total number of solutions in group 2 is very high. Mostly we observed flattening

and widening of the density curves for the parameter values in the solutions of group 2. The peak of the density curves of two groups matches most of the time with few exceptions with slight horizontal shifts. The overall compromise solution, which is the compromise solution among all outcomes of all replications, does however not always coincide with the peak of the density distributions. The standard WGHM parameter values significantly differ from the calibrated values for most parameters in the two basins.

The impact of observation uncertainty is often overlooked in hydrological model calibrations, largely due to the lack of strategies for incorporating this uncertainty into the calibration framework. These uncertainties can severely limit the applicability of calibrated models, as they are often overfitted to mean or single observation values. Even small fluctuations in the observations can degrade the performance of the model, potentially disqualifying it as a non-dominated solution. This, in turn, reduces the reliability of what is considered a compromise solution. Conversely, identifying solutions similar to the compromise solution that account for observation uncertainty, while maintaining an acceptable level of performance, could enhance the reliability of the calibration outcomes. However, scrutinizing these "acceptable solutions considering observation uncertainties", comes with challenges. The influence of observational uncertainties varies across variables and depends on the objective functions used in calibration as well as the nature of the data uncertainties.

4.2.3 Parameter identifiability and equifinality

Figure 8 illustrates that the 4-objective calibration effectively reduced the parameter space substantially in the two basins for most parameters, particularly when comparing the compromise solutions across repeated experiments. Even when we consider the acceptable Pareto solutions considering observation uncertainties, a significant reduction in the a-priori parameter range was achieved for most cases, except two parameters in the Ganges and five in Brahmaputra basin. This already indicates that a good level of identifiability has been achieved. A parameter is considered identifiable for a given set of observations if the true value of the parameter can be inferred with confidence (Wu et al., 2019). The degree of identifiability is usually measured by the posterior standard deviation for individual parameters and posterior covariance matrix for multiple parameters (Wu et al., 2019; Arendt et al., 2012a, b). Cibin et al. (2010) determined parameter identifiability in the Soil and Water Assessment Tool (SWAT) model through visual inspection of scatter plots of model parameters against their corresponding performance metric, considering a parameter as identifiable if a distinct performance metric maximum was observable in the scatter plot. In the absence of a posterior distribution, we measured the degree of identifiability as the ratio of the parameter range in the

compromise solutions of the eight replications to the a-priori parameter range (Table S8 and Table S9).

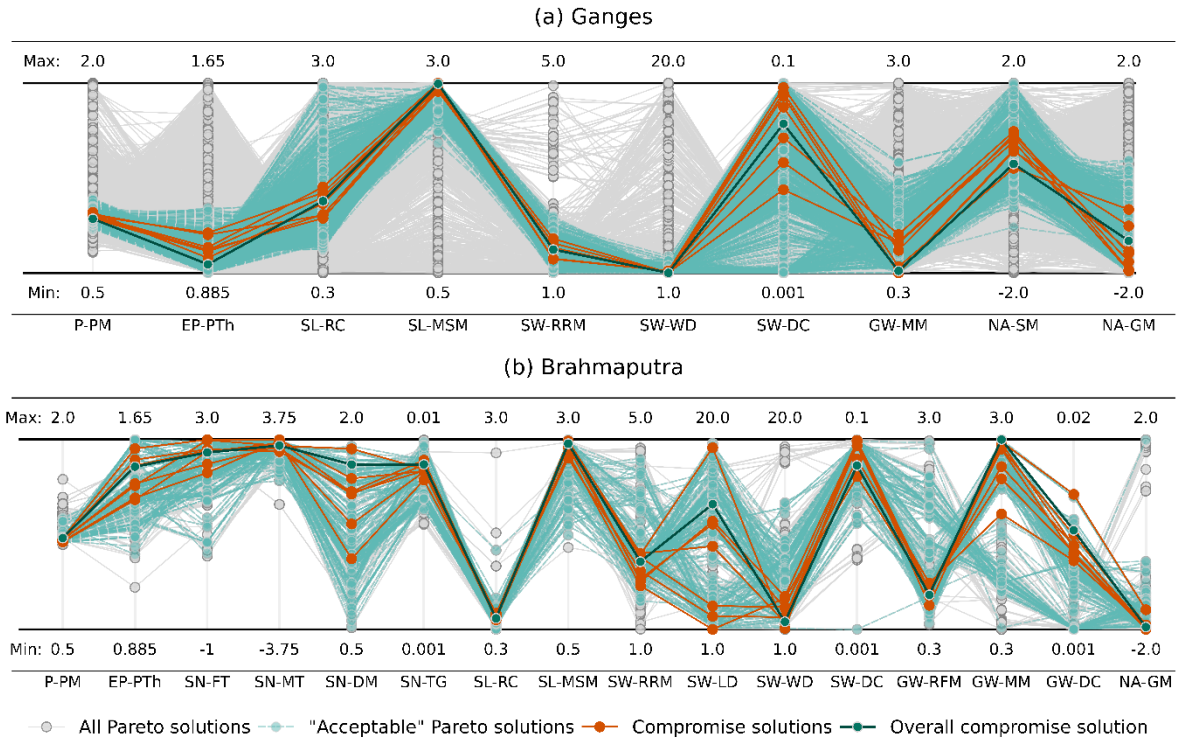

Figure 8: Parallel coordinate plot of all Pareto solutions (grey), the compromise parameter sets (N=8, darkorange), the overall compromise solution (solid cyan), and the set of "acceptable Pareto solutions considering observation uncertainty" (dark green) of the 4-objective calibration in the Ganges basin (a) and the Brahmaputra (b) basin

Due to the fewer parameters involved in the Ganges calibration experiments, better parameter identifiability is observed within the basin compared to experiments in the Brahmaputra basin. We investigated how individual observations influence parameter identifiability during calibration and explored the impact of sensitivity on parameter identifiability. The least satisfactory result was obtained for the calibration with Q only where the ranges of only two parameters (P-PM and SL-RC) in the compromise solutions are less than 15% of the a-priori range. Five parameters (P-PM, EP-PTh, SL-MSM, SL-RC, and SW-RRM) are better constrained by the calibration with TWSA alone. Two sets of six parameters are best constrained by the ET and the SWSA variables, respectively. For ET, they are P-PM, EP-PTh, SL-MSM, SL-RC, GW-MM, and SW-DC; for SWSA the parameters are SL-MSM, SL-RC, SW-RRM, SW-WD, NA-GM, and NA-SM. Compared to the sensitivity indices (mean EET, Table S2), the parameters that are better identified in the SWSA-only calibration are those with the highest sensitivity for this variable. Wu et al. (2019) demonstrated that identifiability is largely related to the sensitivity or significance of the calibration parameters with respect to

response variables. For ET, however, SW-DC and GW-MM are well constrained by the calibration but are not among the influential parameters for any signature of ET in the SA. For TWSA and Q the relationships between parameter identifiability based on the range ratio and the most sensitive signatures from SA are more diverse. In their study, Soares and Calijuri (2021) also observed a clear disparity between the results of their identifiability analysis and sensitivity analysis, although the majority of the results in the two analyses were similar. One should keep in mind, though, that the objective function used in calibration is only one of the signatures that was used to measure sensitivity in the SA. Nevertheless, we usually observe high correlations among the objectives and the parameters for at least one of the variables (Table S4).

In 2- or more-objective calibrations in the Ganges basin, the degree of parameter identifiability varies with the participating variables. Interestingly, parameters are best identifiable in the experiment ES in which the range of all parameters after calibration is less than 2.2% of the a-priori range except one for which the range ratio is only 11%. In the case of the 4-objective calibration, most parameter range ratios are below or around 20%, only two parameters have a range ratio of more than 30%. Arendt et al. (2012a) demonstrated that employing multiple responses, which exhibit mutual dependencies on a common set of parameters, can enhance the identifiability of those parameters. However, if the dependencies among themselves exhibit inverse relationships, parameter identifiability may worsen, as observed in the case of the 4-objective calibration.

In the Brahmaputra basin, four parameters (P-PM, SN-MT, SN-TG, and SL-RC) are constrained well (i.e., they have low coverage of their a-priori range in the compromise solution sets) with the variable Q, two parameters (SN-MT and SL-MSM) by the ET variable, and two (SN-TG and SW-RRM) by the SWSA observations. However, TWSA seems to have no discernible control on any of the parameters in the basin. Parameters with a strong inverse correlation, such as GW-MM and GW-RFM (correlation of -0.90, Table S7) leading to reduced identifiability in both the parameters in Brahmaputra basin (Figure 8), may become better identifiable if one parameter is omitted from the calibration process.

Non-uniqueness, or equifinality, arises from the fact that a singular set of parameters is incapable of generating a unique set of model responses. This occurs because there are numerous pathways to achieve the same target (Beven and Binley, 1992), or the capacity for uniqueness is compromised, whether observation data is absent or present, often as a consequence of summarizing the response variable (Wagener et al., 2003). In the presence of

input and data uncertainties, the degree of equifinality increases, even if the structural discrepancies within the model remain unchanged. As the non-uniqueness problem intensifies, it is expected that non-identifiability also rises. If, however, the equifinality arises from an inverse correlation among parameters, a parameter can still be identifiable if the opposing parameter is omitted. For example, GW-MM exhibits a strong negative correlation (-0.90, Table S7) with the parameter GW-RFM among the "acceptable" Pareto solutions considering observation uncertainties, which reduced identifiability in both the parameters in Brahmaputra basin (Figure 8). If one of them is omitted from the calibration process, the other may become identifiable.

The analysis of parameter identifiability does not provide clear evidence that including additional observables in the calibration process necessarily enhances identifiability, particularly in the presence of strong correlations among parameters. On the contrary, interactions among highly sensitive parameters must be carefully considered when selecting parameters for calibration, especially in multi-variable calibration scenarios. However, a strong relationship was observed between parameter identifiability and the number of parameters being optimized. Specifically, in almost all single and multi-objective calibrations within the Ganges basin, the parameter ranges were significantly reduced compared to the calibration experiments conducted in the Brahmaputra basin (Tables S8 and S9).

4.2.4 Validation

We compare the simulations of the four variables of the overall compromise solution and the ensemble of the acceptable Pareto solutions considering observation uncertainties (Sect. 4.2.2) of the 4-objective calibration experiment with the observations. Except in one to two months during monsoon, the simulation with the overall compromise solution overestimates monthly streamflow in the Ganges basin (Figure 9). This led to an overestimation of monthly means for all months. The annual streamflow amplitudes are also overestimated in all years except one (2002). In comparison, the standard WGHM calibration with only streamflow observation resulted in a better fit in streamflow simulations with a lower magnitude of under- and overestimations of monthly means, annual amplitudes, and annual means. Surprisingly, the uncalibrated model simulates the streamflow in the basin better than the compromise solutions in all signatures except in the seasonal amplitude (Table 8). The uncalibrated model mostly performed well to represent the low-flows.

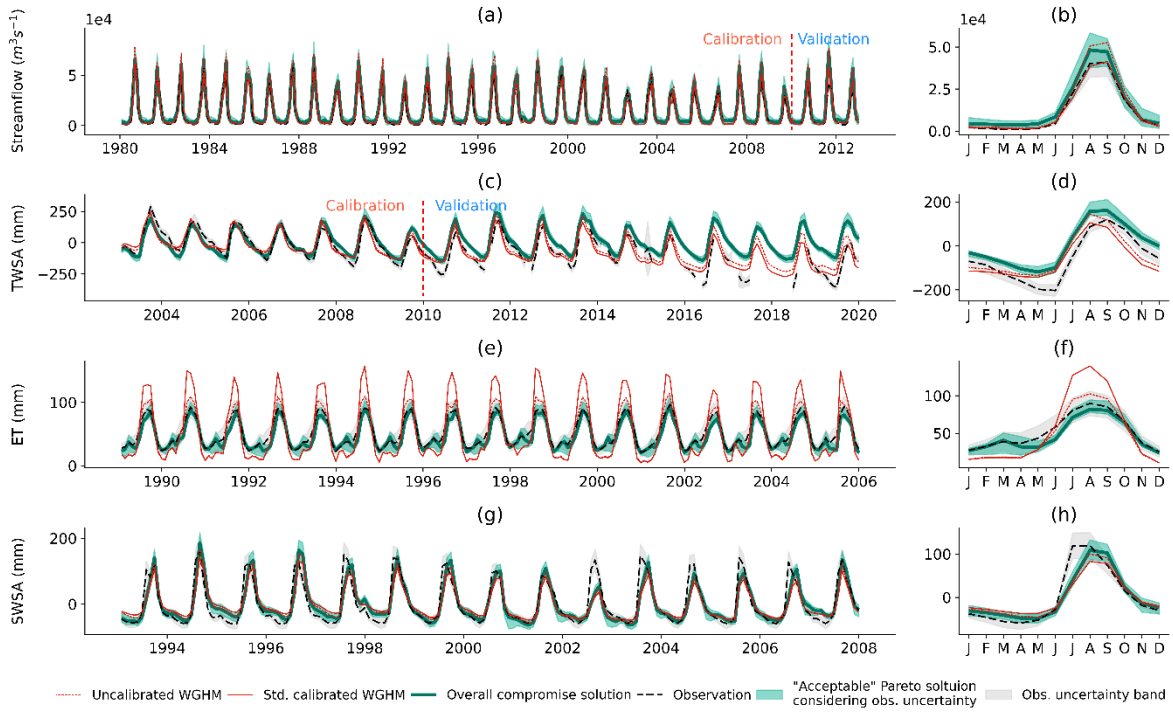

Figure 9: Simulations of streamflow (a, b), TWSA (c, d), ET (e, f), and SWSA (g, h) in the Ganges basin with the overall compromise solution of the 4-objective calibration. The simulation results with the acceptable Pareto solutions considering observation uncertainties (darkgreen), with uncalibrated WGHM (red dotted line), and with the calibrated model with standard approach (red solid) are also shown; monthly time series (a, c, e, g) and monthly mean values (b, d, f, h). The mean was computed over the entire period with available observations.

In the Brahmaputra basin, the simulation with the compromise solution underestimates streamflow for the monthly time series and annual averages (Figure 10a). However, except for few high rainfall years, the seasonal amplitude was mostly overestimated (Figure 10b). In comparison to the model that was calibrated with the standard approach, the compromise solution in this basin performed better with lower mean absolute deviation in all aspects of streamflow simulation (Table 8).

Table 8: Comparing the overall compromise solution of the 4-objective calibration, standard calibration, and uncalibrated model across four signatures: monthly time series (MTS), monthly means (MM), annual time series (ATS), and seasonal amplitude (SNA). The mean absolute deviation (MAD) was computed by comparing observations and simulations across the entire observation period, expressed as a percentage of the observation mean. Observation gaps were filled using linear regressions ($y = \beta_0 + \beta_1 x$) for individual months, accounting for seasonality and trends.

		Mean Absolute Deviation (MAD) as fractions of observation mean							
		Ganges				Brahmaputra			
	Sig	Obs. Mean[a,b]	Comp. Sol.	Std. calibration	Un-calibrated	Obs. Mean	Comp. Sol.	Std. calibration	Un-calibrated
Q	MTS	11855.8	30.6%	16.7%	28.2%	22140.2	14.3%	19.4%	34.7%

	MM		26.8%	7.5%	24.7%		7.5%	13.6%	34.6%
	ATS		26.8%	8.1%	24.7%		7.0%	7.0%	34.6%
	SNA	43274.4	20.4%	11.0%	34.0%	48345.7	13.9%	17.1%	16.9%
TWSA	SNA	323.4	11.7%	19.4%	13.4%	270.0	7.3%	28.9%	25.0%
ET	MTS	51.7	11.3%	41.9%	26.5%	46.7	6.1%	33.3%	17.9%
	MM		10.2%	39.5%	25.2%		3.2%	33.3%	17.0%
	ATS		9.7%	5.1%	13.0%		1.9%	33.3%	9.6%
	SNA	65.0	6.9%	105.3%	48.6%	52.7	8.0%	7.6%	45.1%
SWSA	SNA	187.8	15.4%	31.2%	19.0%	245.6	5.8%	55.8%	61.2%

[a] units for Q is [m^3s^{-1}] and [mm] for all other variables
[b] means of seasonal amplitude in [mm $year^{-1}$]

Negative trends of TWSA are clearly visible in observations of both the Ganges and the Brahmaputra basins (Figure 9c and Figure 10c). Because we used only few years of TWSA data in calibration (2003-2009) and within this period the trend was not very obvious, all calibrations fail to represent the TWSA decrease in the two basins. Nevertheless, beyond the calibration period, the seasonality and peaks are correctly represented (with high correlation coefficients between observation and simulation in 2010-2019, Table 9) by the simulations with the compromise solutions of the two basins. Also, the mean deviation of TWSA is 11.7% and only 7.3% of mean annual amplitude in the Ganges and the Brahmaputra basin respectively. The model's inability to produce the negative TWSA trends causes an overestimation of TWSA in the later years of the simulation time series. The performance of the calibrated model with standard calibration in simulating the TWSA is worse than that of the compromise solution and even worse than the uncalibrated model.

ET simulation with the compromise solution in the Ganges basin mostly underestimates the observed values except for a few winter months. It should be noted that the interannual variability of amplitude is very small in the ET observation in the two basins (13.6 mm in the Ganges and 11mm in the Brahmaputra). The seasonal amplitude is underestimated by the compromise solution in most years in the Ganges basin; whereas, slightly overestimated in the Brahmaputra basin in most years. The monthly means in the Ganges is underestimated for all months except March. In the Brahmaputra basin the monthly means match well with the observations with only 3% of mean absolute deviation; the annual mean in the basin also has a very small deviation from the observations.

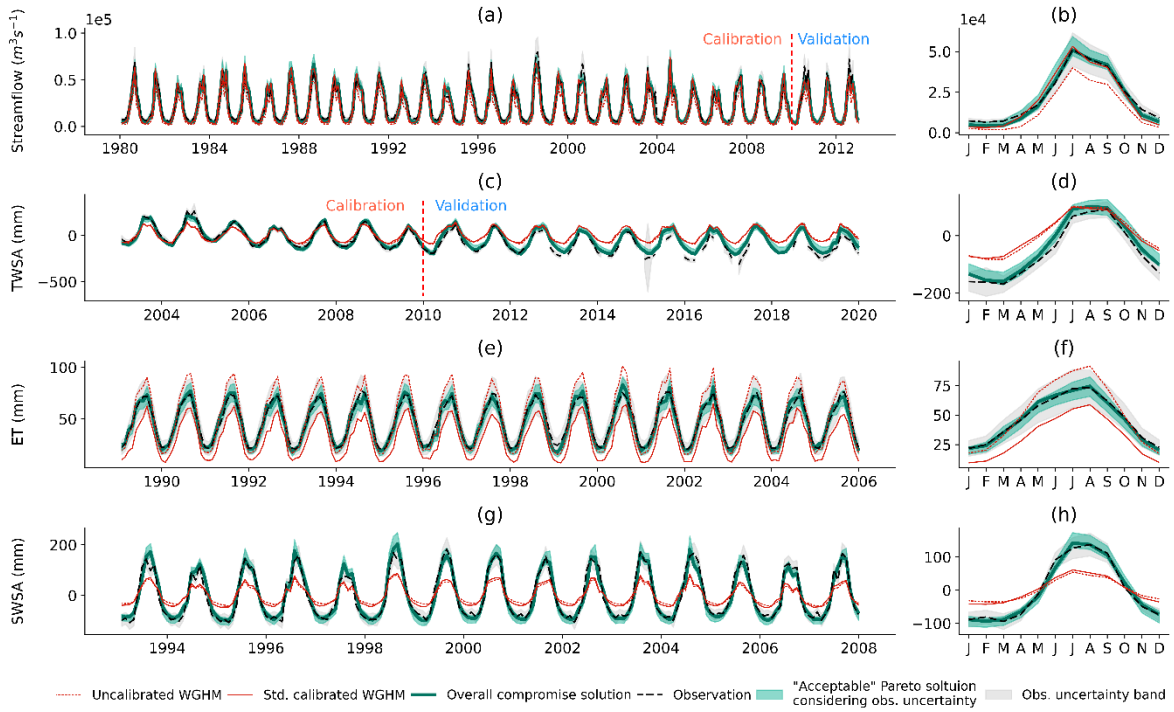

Figure 10: Simulations of streamflow (a, b), TWSA (c, d), ET (e, f), and SWSA (g, h) in the Brahmaputra basin with different solutions. See description of Figure 9.

For SWSA, the simulation with the compromise solutions is limited in accurately reproducing the mean seasonality of the observed SWSA: the seasonal peak of the simulated SWSA is on average one month delayed relative to the observations. The earlier increase of SWSA by one month as compared to the simulations could be explained by the detection of rice paddies or wet soil signal by the satellite method (Papa et al., 2006) which are not captured by the model.

The seasonal amplitude is slightly underestimated by the simulations, especially in the later observational years (2002-2007) (Figure 9h). In contrast, in the Brahmaputra basin, the compromise solution simulates the SWSA dynamics very well (Figure 10g, h).

The simulation results with the ensemble of compromise solutions obtained by consideration of observation uncertainties (group 2) usually show similar dynamics to the compromise solution itself. The uncertainty of the group 2 simulation results, i.e., the bandwidth around the compromise solution, is smaller than the observation uncertainty bandwidth (Table S13) except for streamflow and SWSA simulations in the Ganges basin. The bandwidth of ET in the group 2 solutions is the lowest (32.2% and 23.8% of the observed mean in the Ganges and Brahmaputra basins respectively in comparison to observation average uncertainty band width of 53.4% and 50.5% in the two basins). In the Brahmaputra basin, the monthly streamflow simulations with the group 2 solutions fall mostly within the observation while in

the Ganges basin the monthly streamflow peaks are mostly overestimated and pass beyond the upper limit of observation uncertainty. Although the group 2 simulation bandwidth was smaller for TWSA for the two basins than the observation uncertainty bands, the simulation bandwidth follows the observation uncertainty band only in the Brahmaputra basin. In the Ganges basin, with the exception of a few years, the group 2 simulation uncertainty band misses the observation uncertainty band even in the calibration period. The group 2 uncertainty of ET simulations falls within the observation uncertainty band in both basins. For SWSA, due to the better representation of its seasonality in Brahmaputra, also the uncertainty of SWSA simulations is better covered by the observation uncertainty than in the Ganges basin.

As discussed before, due to the unavailability of observation data in ET and SWSA in the validation period, model validation was only possible for the Q and TWSA simulations of the compromise solution (Table 9). The simulation error (RMSE) in the validation period increased by factors of 2 to 5 relative to the calibration in most cases, indicating strong degradation of model performance in the validation period. The performance in monthly values (MTS) and mean monthly values (MM) is slightly better than the annual means (ATS) and seasonal amplitudes (SNA) in both basins because the calibration objectives were monthly time series. The performance with respect to these last two signatures is worse also in the calibration period. However, all the signatures stay with high correlations in the validation period which implies the timing and the seasonal dynamics were well captured by the simulations with the compromise solutions. The validation metrics should be carefully interpreted as the amount of validation data is small.

Table 9: Performance metrics of the overall compromise solution of the 4-objective calibration of the Ganges and Brahmaputra basin for all observables and for four signatures: monthly time-series (MTS), annual average time-series (ATS), monthly mean (MM), and seasonal amplitude (SNA). Calibration period is until 2009 from the start of observation availability and validation period starts from 2010 till the last observation year

NSE/Correlation[a] (r)/RMSE[b]

	Calibration [variable year – 2009]				Validation [2010 – variable year]			
	MTS	ATS	MM	SNA	MTS	ATS	MM	SNA
	Ganges							
Q	0.91/0.98/4978.1	0.72/0.98/3378.6	0.95/1.00/3407.6	-0.01/0.86/8930.3	0.55/0.96/9560.1	-4.68/0.99/5826.1	0.66/0.97/7704.2	-17.99/0.38/21303.2
TWSA	0.81/0.92/47.0	0.15/0.61/42.1	0.95/0.99/21.1	0.54/0.90/38.2	0.20/0.88/114.4	-1.13/0.82/114.8	0.18/0.97/98.6	0.22/0.87/56.6

ET	0.88/0.96/8.0	-8.34/0.83/5.2	0.90/0.97/7.2	-1.32/0.49/5.3				
SWSA	0.77/0.88/31.5	-6.83/0.15/10.2	0.84/0.92/25.6	-2.72/0.24/35.3				
	Brahmaputra							
Q	0.94/0.98/4038.8	0.56/0.82/1828.5	0.98/1.00/2042.4	0.56/0.80/7411.8	0.85/0.97/7323.5	-0.28/0.92/4942.3	0.87/0.99/6109.9	-1.95/0.79/13054.9
TWSA	0.96/0.98/20.7	0.94/0.99/10.2	0.99/1.00/9.3	0.78/0.92/22.4	0.72/0.96/54.5	-0.58/0.92/54.3	0.78/1.00/45.6	0.55/0.85/26.0
ET	0.96/0.98/3.7	-0.98/0.18/1.1	0.99/1.00/1.7	-3.63/-0.09/5.5				
SWSA	0.94/0.97/22.3	0.43/0.66/8.8	0.98/0.99/11.0	0.50/0.74/18.3				

[a] Pearson correlation coefficient
[b] unit of RMSE for Q is m3s-1 and mm for the rest of the variables

5 Conclusions

In this study, we introduced a multi-objective calibration framework for estimating basin-specific optimal parameter sets for large-scale hydrological models. The framework can make use of observations for multiple model output variables as well as for multiple signatures of each variable. Applying this approach to the Ganges and Brahmaputra basins with the global hydrological model WGHM, we analysed the impacts, benefits and challenges of multi-variable multi-signature sensitivity analysis and multi-variable calibration.

The multi-variable multi-signature sensitivity analysis facilitates the identification of important parameters that would remain unidentified if not all variables or signatures were considered. Due to the different hydrological characteristics of the modelling units to be calibrated, the sensitivity analysis has to be carried out individually for each unit, resulting in identification of different influential parameters and in different numbers of parameters to be calibrated. The proposed parameter selection method is based on the relative impact of individual parameters compared to all parameters. The method can be adjusted with respect to the impact thresholds and by weighting variables and signatures depending on the modelling purpose and is thus an approach that can be used in a flexible way in other studies.

The results of this study show that parameter identifiability is inversely related to the number of parameters that were selected for calibration. Although a reasonably good level of parameter identifiability in the multi-variable calibrations was achieved, the results do not provide evidence that using multiple observational variables generally enhances parameter identifiability. Certain combinations of observations used for calibration resulted in a high parameter identifiability, e.g., calibration with ET and SWSA or with TWSA, ET, and SWSA in the Ganges basin but this is not the case of the Brahmaputra basin. While in the Ganges basin,

all 2-variable calibrations lead to a better identifiability than all 1-variable calibrations, calibration against Q only results in a better identifiability than all the 2-variable calibration except calibration with Q and SWSAöThus it depends on the basin as well as on the selected calibration parameters, to what degree the inclusion of observations of multiple variable in the model calibration increase parameter identifiability.

Including additional observations in the calibration consistently improved the overall model performance. The highest overall performance is achieved by the calibration that takes into account all four output variables. The value of calibrating with Q and TWSA observations for the overall model performance was higher than that of ET and SWSA observations. The degree of improvements depends on basin characteristics as well as on the trade-offs and interactions among the objectives. This, in turn, depends on the capability of the model to represent the relevant hydrological processes in the basin. In line with Döll et al. (2024), we found that using streamflow observations in the calibration is essential for achieving good streamflow simulations, which are the primary target for most hydrological model applications. In contrast, good simulation results for TWSA could also be achieved in the Brahmaputra basin even when TWSA was not used for calibration but Q and ET.

In this study, we considered two sources of uncertainty in the calibration process: (i) those arising from the search algorithm used to identify the non-dominated Pareto-optimal parameter sets and (ii) those stemming from observational errors. As the random seeds used in the BORG algorithm lead to non-negligible variations in the calibration results and model performances, in particular for the variables that were not used for calibration, a sufficient number of replications of the calibration runs with different initial parameter sets is vital. The results show that a large part of the variations of "optimal" parameter sets can be attributed to observational uncertainties, a factor often overlooked in calibration exercises. We demonstrated that in the presence of observational uncertainty, relying solely on a 'best solution' or a compromise solution can become unreliable, leading to decreased overall efficiency. To address this challenge, we propose a method to select an ensemble of 'acceptable' solutions from the Pareto solutions derived by the search algorithm, taking into account uncertainties in the observation data used for calibration.

The multi-variable multi-signature parameter selection and calibration methodology presented in this study is suggested for other calibration studies with GHMs or other large-scale hydrological models for all large river basins of the globe where diverse observations of model output variables are available. While the methodology also allows for considering the effect of observational uncertainties on the multi-criterial calibration results, it is imperative to further

explore how accounting for observation uncertainties can enhance the robustness of calibration outcomes. Developing uncertainty-based performance metrics would represent a significant advancement in this direction. In regions with limited data availability, leveraging remote sensing-based streamflow observations such as HydroSAT (http://hydrosat.gis.uni-stuttgart.de) or SWOT can provide new insights, complementing TWSA data from GRACE, GRACE-FO, and GRACE-C (GRACE-Continuity). Given the availability of numerous contemporary ET products, future calibration efforts should explore the benefits of considering these ET data sources.

Code availability: The WaterGAP code is accessible at https://github.com/HydrologyFrankfurt/WaterGAP2

Data availability. All optimal parameter sets obtained for the two basins together with the resulting performance metrics are compiled in an Excel file included in the supplement. The supplement (MS Excel file) is available online at URL???

Supplement. The supplement related to this article (MS Word file) is available online at URL???

Author contributions. HMMH and AG designed the study, with contributions from PD, SMHM, and FP. HMMH performed SA. Calibration and data analyses was performed by HMMH and SMHM. HMMH produced the graphics. HMMH wrote the original draft of the manuscript. All authors contributed to the final draft.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgments. This study was enabled by the financial support of the German Research Foundation for the research unit “Understanding the global freshwater system by combining geodetic and remote sensing information with modelling using a calibration/data assimilation approach (GlobalCDA)”. The authors thank Felix Theodor Portmann for first analyses and discussions, Kerstin Schulze, University of Bonn, Germany for processing the GRACE TWSA data, and Muhammad Masood of Bangladesh Water Development Board (BWDB), Dhaka, Bangladesh for providing streamflow observations.

Review statement. This paper was edited by xxx and reviewed by xxx anonymous referees.

References

Ai, Z. and Hanasaki, N.: Simulation of crop yield using the global hydrological model H08 (crp.v1), Geosci Model Dev, 16, 3275–3290, https://doi.org/10.5194/gmd-16-3275-2023, 2023.

Akhil, V. P., Durand, F., Lengaigne, M., Vialard, J., Keerthi, M. G., Gopalakrishna, V. V, Deltel, C., Papa, F., and de Boyer Montégut, C.: A modeling study of the processes of surface salinity seasonal cycle in the Bay of Bengal, J Geophys Res Oceans, 119, 3926–3947, https://doi.org/10.1002/2013JC009632, 2014.

Arendt, P. D., Apley, D. W., Chen, W., Lamb, D., and Gorsich, D.: Improving Identifiability in Model Calibration Using Multiple Responses, Journal of Mechanical Design, 134, 100909, https://doi.org/10.1115/1.4007573, 2012a.

Arendt, P. D., Apley, D. W., and Chen, W.: Quantification of Model Uncertainty: Calibration, Model Discrepancy, and Identifiability, Journal of Mechanical Design, 134, 100908, https://doi.org/10.1115/1.4007390, 2012b.

Arheimer, B., Pimentel, R., Isberg, K., Crochemore, L., Andersson, J. C. M., Hasan, A., and Pineda, L.: Global catchment modelling using World-Wide HYPE (WWH), open data, and stepwise parameter estimation, Hydrol Earth Syst Sci, 24, 535–559, https://doi.org/10.5194/hess-24-535-2020, 2020.

Bai, P., Liu, X., and Liu, C.: Improving hydrological simulations by incorporating GRACE data for model calibration, J Hydrol (Amst), 557, 291–304, https://doi.org/10.1016/j.jhydrol.2017.12.025, 2018.

Di Baldassarre, G. and Montanari, A.: Uncertainty in river discharge observations: a quantitative analysis, Hydrol Earth Syst Sci, 13, 913–921, https://doi.org/10.5194/hess-13-913-2009, 2009.

Banda, V. D., Dzwairo, R. B., Singh, S. K., and Kanyerere, T.: Hydrological Modelling and Climate Adaptation under Changing Climate: A Review with a Focus in Sub-Saharan Africa, Water (Basel), 14, https://doi.org/10.3390/w14244031, 2022.

Beck, H. E., van Dijk, A. I. J. M., de Roo, A., Miralles, D. G., McVicar, T. R., Schellekens, J., and Bruijnzeel, L. A.: Global-scale regionalization of hydrologic model parameters, Water Resour Res, 52, 3599–3622, https://doi.org/10.1002/2015WR018247, 2016.

Beck, H. E., Pan, M., Lin, P., Seibert, J., van Dijk, A. I. J. M., and Wood, E. F.: Global Fully Distributed Parameter Regionalization Based on Observed Streamflow From 4,229 Headwater Catchments, Journal of Geophysical Research: Atmospheres, 125, e2019JD031485, https://doi.org/10.1029/2019JD031485, 2020.

Becker, M., Papa, F., Karpytchev, M., Delebecque, C., Krien, Y., Khan, J. U., Ballu, V., Durand, F., Cozannet, G. Le, Islam, A. K. M. S., Calmant, S., and Shum, C. K.: Water level changes, subsidence, and sea level rise in the Ganges–Brahmaputra–Meghna delta, Proceedings of the National Academy of Sciences, 117, 1867–1876, https://doi.org/10.1073/pnas.1912921117, 2020.

Beven, K.: Prophecy, reality and uncertainty in distributed hydrological modelling, Adv Water Resour, 16, 41–51, https://doi.org/10.1016/0309-1708(93)90028-E, 1993.

Beven, K.: A manifesto for the equifinality thesis, J Hydrol (Amst), 320, 18–36, https://doi.org/10.1016/j.jhydrol.2005.07.007, 2006.

Beven, K. and Binley, A.: The future of distributed models: Model calibration and uncertainty prediction, Hydrol Process, 6, 279–298, https://doi.org/10.1002/hyp.3360060305, 1992.

Bookhagen, B. and Burbank, D. W.: Topography, relief, and TRMM-derived rainfall variations along the Himalaya, Geophys Res Lett, 33, https://doi.org/10.1029/2006GL026037, 2006.

Campolongo, F., Saltelli, A., and Cariboni, J.: From screening to quantitative sensitivity analysis. A unified approach, Comput Phys Commun, 182, 978–988, https://doi.org/10.1016/j.cpc.2010.12.039, 2011.

Cheng, C. T., Wu, X. Y., and Chau, K. W. : Multiple criteria rainfall–runoff model calibration using a parallel genetic algorithm in a cluster of computers / Calage multi-critères en modélisation pluie–débit par un algorithme génétique parallèle mis en œuvre par une grappe d'ordinateurs. Hydrological Sciences Journal, 50(6), 1087. https://doi.org/10.1623/hysj.2005.50.6.1069, 2005

Chilkoti, V., Bolisetti, T., and Balachandar, R.: Multi-objective autocalibration of SWAT model for improved low flow performance for a small snowfed catchment, Hydrological Sciences Journal, 63, 1482–1501, https://doi.org/10.1080/02626667.2018.1505047, 2018.

Cibin, R., Sudheer, K. P., and Chaubey, I.: Sensitivity and identifiability of stream flow generation parameters of the SWAT model, Hydrol Process, 24, 1133–1148, https://doi.org/10.1002/hyp.7568, 2010.

Cucchi, M., Weedon, G. P., Amici, A., Bellouin, N., Lange, S., Müller Schmied, H., Hersbach, H., and Buontempo, C.: WFDE5: bias-adjusted ERA5 reanalysis data for impact studies, Earth Syst Sci Data, 12, 2097–2120, https://doi.org/10.5194/essd-12-2097-2020, 2020.

Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II, IEEE Transactions on Evolutionary Computation, 6, 182–197, https://doi.org/10.1109/4235.996017, 2002.

Deb, K., Mohan, M., and Mishra, S.: Evaluating the ϵ-Domination Based Multi-Objective Evolutionary Algorithm for a Quick Computation of Pareto-Optimal Solutions, Evol Comput, 13, 501–525, https://doi.org/10.1162/106365605774666895, 2005.

Dembélé, M., Hrachowitz, M., Savenije, H. H. G., Mariéthoz, G., and Schaefli, B.: Improving the Predictive Skill of a Distributed Hydrological Model by Calibration on Spatial Patterns With Multiple Satellite Data Sets, Water Resour Res, 56, e2019WR026085, https://doi.org/10.1029/2019WR026085, 2020.

Demirel, M. C., Koch, J., Rakovec, O., Kumar, R., Mai, J., Müller, S., Thober, S., Samaniego, L., and Stisen, S.: Tradeoffs Between Temporal and Spatial Pattern Calibration and Their Impacts on Robustness and Transferability of Hydrologic Model Parameters to Ungauged Basins, Water Resources Research, 60, e2022WR034193, https://doi.org/10.1029/2022WR034193, 2024.

Demirel, M. C., Mai, J., Mendiguren, G., Koch, J., Samaniego, L., and Stisen, S.: Combining satellite data and appropriate objective functions for improved spatial pattern performance of a distributed hydrologic model, Hydrol. Earth Syst. Sci., 22, 1299–1315, https://doi.org/10.5194/hess-22-1299-2018, 2018.

Demirel, M. C., Özen, A., Orta, S., Toker, E., Demir, H. K., Ekmekcioğlu, Ö., Tayşi, H., Eruçar, S., Sağ, A. B., Sarı, Ö., Tuncer, E., Hancı, H., Özcan, T. İ., Erdem, H., Koşucu, M. M., Başakın, E. E., Ahmed, K., Anwar, A., Avcuoğlu, M. B., Vanlı, Ö., Stisen, S., and Booij, M. J.: Additional Value of Using Satellite-Based Soil Moisture and Two Sources of Groundwater Data for Hydrological Model Calibration, Water (Basel), 11, https://doi.org/10.3390/w11102083, 2019.

Denager, T., Sonnenborg, T. O., Looms, M. C., Bogena, H., and Jensen, K. H.: Point-scale multi-objective calibration of the Community Land Model (version~5.0) using in~situ observations of water and energy fluxes and variables, Hydrol Earth Syst Sci, 27, 2827–2845, https://doi.org/10.5194/hess-27-2827-2023, 2023.

Döll, P., Hasan, H. M. M., Schulze, K., Gerdener, H., Börger, L., Shadkam, S., Ackermann, S., Hosseini-Moghari, S.-M., Müller Schmied, H., Güntner, A., Kusche, J. (2024): Leveraging multi-variable observations to reduce and quantify the output uncertainty of a global hydrological model: evaluation of three ensemble-based approaches for the Mississippi River basin. Hydrol. Earth Syst. Sci., 28, 2259–2295. https://doi.org/10.5194/hess-28-2259-2024.

Döll, P. and Lehner, B.: Validation of a new global 30-min drainage direction map, J Hydrol (Amst), 258, 214–231, https://doi.org/10.1016/S0022-1694(01)00565-0, 2002.

Döll, P. and Zhang, J.: Impact of climate change on freshwater ecosystems: a global-scale analysis of ecologically relevant river flow alterations, Hydrol Earth Syst Sci, 14, 783–799, https://doi.org/10.5194/hess-14-783-2010, 2010.

Efstratiadis, A. and Koutsoyiannis, D.: One decade of multi-objective calibration approaches in hydrological modelling: a review, Hydrological Sciences Journal, 55, 58–78, https://doi.org/10.1080/02626660903526292, 2010.

Einarsson, I., Hoechner, A., Wang, R., and Kusche, J.: Gravity changes due to the Sumatra-Andaman and Nias earthquakes as detected by the GRACE satellites: a reexamination, Geophys J Int, 183, 733–747, https://doi.org/10.1111/j.1365-246X.2010.04756.x, 2010.

FAO: AQUASTAT Transboundary River Basins – Ganges-Brahmaputra-Meghna River Basin, Rome, Italy, 2011.

Fernandez-Palomino, C. A., Hattermann, F. F., Krysanova, V., Vega-Jácome, F., and Bronstert, A.: Towards a more consistent eco-hydrological modelling through multi-objective calibration: a case study in the Andean Vilcanota River basin, Peru, Hydrological Sciences Journal, 0, 1–16, https://doi.org/10.1080/02626667.2020.1846740, 2020.

Frappart, F., Papa, F., Silva, J. S. da, Ramillien, G., Prigent, C., Seyler, F., and Calmant, S.: Surface freshwater storage and dynamics in the Amazon basin during the 2005 exceptional drought, Environmental Research Letters, 7, 44010, https://doi.org/10.1088/1748-9326/7/4/044010, 2012.

Gain, A. K. and Wada, Y.: Assessment of Future Water Scarcity at Different Spatial and Temporal Scales of the Brahmaputra River Basin, Water Resources Management, 28, 999–1012, https://doi.org/10.1007/s11269-014-0530-5, 2014.

Gerdener, H., Engels, O., and Kusche, J.: A framework for deriving drought indicators from the Gravity Recovery and Climate Experiment (GRACE), Hydrol Earth Syst Sci, 24, 227–248, https://doi.org/10.5194/hess-24-227-2020, 2020.

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

Goteti, G. and Famiglietti, J.: Extent of gross underestimation of precipitation in India, Hydrol. Earth Syst. Sci. Discuss. [preprint], https://doi.org/10.5194/hess-2024-18, in review, 2024.

Gu, L., Chen, J., Yin, J., Slater, L. J., Wang, H.-M., Guo, Q., Feng, M., Qin, H., and Zhao, T.: Global Increases in Compound Flood-Hot Extreme Hazards Under Climate Warming, Geophys Res Lett, 49, e2022GL097726, https://doi.org/10.1029/2022GL097726, 2022.

Gupta, H. V., Sorooshian, S., and Yapo, P. O.: Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information, Water Resour Res, 34, 751–763, https://doi.org/10.1029/97WR03495, 1998.

Hadka, D. and Reed, P.: Borg: An Auto-Adaptive Many-Objective Evolutionary Computing Framework, Evol Comput, 21, 231–259, https://doi.org/10.1162/EVCO_a_00075, 2013.

Herrera, P. A., Marazuela, M. A., and Hofmann, T.: Parameter estimation and uncertainty analysis in hydrological modeling, WIREs Water, 9, e1569, https://doi.org/10.1002/wat2.1569, 2022.

Hornberger, G. M. and Spear, R. C.: Approach to the preliminary analysis of environmental systems, J. Environ. Manage.; (United States), 12:1, 1981.

Hosseini-Moghari, S.-M., Araghinejad, S., Tourian, M. J., Ebrahimi, K., and Döll, P.: Quantifying the impacts of human water use and climate variations on recent drying of Lake Urmia basin: the value of different sets of spaceborne and in situ data for calibrating a global hydrological model, Hydrol Earth Syst Sci, 24, 1939–1956, https://doi.org/10.5194/hess-24-1939-2020, 2020.

Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., Arheimer, B., Blume, T., Clark, M. P., Ehret, U., Fenicia, F., Freer, J. E., Gelfan, A., Gupta, H. V, Hughes, D. A., Hut, R. W., Montanari, A., Pande, S., Tetzlaff, D., Troch, P. A., Uhlenbrook, S., Wagener, T., Winsemius, H. C., Woods, R. A., Zehe, E., and Cudennec, C.: A decade of Predictions in Ungauged Basins (PUB)—a review, Hydrological Sciences Journal, 58, 1198–1255, https://doi.org/10.1080/02626667.2013.803183, 2013.

Huang, Q., Qin, G., Zhang, Y., Tang, Q., Liu, C., Xia, J., Chiew, F. H. S., and Post, D.: Using Remote Sensing Data-Based Hydrological Model Calibrations for Predicting Runoff in Ungauged or Poorly Gauged Catchments, Water Resour Res, 56, e2020WR028205, https://doi.org/10.1029/2020WR028205, 2020.

Huang, Y., Salama, Mhd. S., Krol, M. S., Su, Z., Hoekstra, A. Y., Zeng, Y., and Zhou, Y.: Estimation of human-induced changes in terrestrial water storage through integration of GRACE satellite detection and hydrological modeling: A case study of the Yangtze River basin, Water Resour Res, 51, 8494–8516, https://doi.org/10.1002/2015WR016923, 2015.

Hulsman, P., Savenije, H. H. G., and Hrachowitz, M.: Learning from satellite observations: increased understanding of catchment processes through stepwise model improvement, Hydrol. Earth Syst. Sci., 25, 957–982, https://doi.org/10.5194/hess-25-957-2021, 2021.

Immerzeel, W.: Historical trends and future predictions of climate variability in the Brahmaputra basin, International Journal of Climatology, 28, 243–254, https://doi.org/10.1002/joc.1528, 2008.

India-WRIS: Basin Reports - Brahmaputra Basin, New Delhi - 110066, INDIA, 2014a.

India-WRIS: Basin Reports - Ganga Basin, New Delhi - 110066, INDIA, 2014b.

Jakeman, A. J. and Hornberger, G. M.: How much complexity is warranted in a rainfall-runoff model?, Water Resour Res, 29, 2637–2649, https://doi.org/10.1029/93WR00877, 1993.

Khan, A. A., Pant, N. C., Goswami, A., Lal, R., and Joshi, R.: Critical Evaluation and Assessment of Average Annual Precipitation in The Indus, The Ganges and The Brahmaputra Basins, Northern India, in: Dynamics of Climate Change and Water Resources of Northwestern Himalaya, 67–84, 2015.

Kirchner, J. W.: Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology, Water Resour Res, 42, https://doi.org/10.1029/2005WR004362, 2006.

Kittel, C. M. M., Nielsen, K., Tøttrup, C., and Bauer-Gottwein, P.: Informing a hydrological model of the Ogooué with multi-mission remote sensing data, Hydrol Earth Syst Sci, 22, 1453–1472, https://doi.org/10.5194/hess-22-1453-2018, 2018.

Kollat, J. B. and Reed, P. M.: Comparing state-of-the-art evolutionary multi-objective algorithms for long-term groundwater monitoring design, Adv Water Resour, 29, 792–807, https://doi.org/10.1016/j.advwatres.2005.07.010, 2006.

Kukkonen, S. and Lampinen, J.: GDE3: the third evolution step of generalized differential evolution, in: 2005 IEEE Congress on Evolutionary Computation, 443-450 Vol.1, https://doi.org/10.1109/CEC.2005.1554717, 2005.

Kusche, J., Schmidt, R., Petrovic, S., and Rietbroek, R.: Decorrelated GRACE time-variable gravity solutions by GFZ, and their validation using a hydrological model, J Geod, 83, 903–913, https://doi.org/10.1007/s00190-009-0308-3, 2009.

Kushwaha, A. P., Tiwari, A. D., Dangar, S., Shah, H., Mahto, S. S., and Mishra, V.: Multimodel assessment of water budget in Indian sub-continental river basins, J Hydrol (Amst), 603, 126977, https://doi.org/10.1016/j.jhydrol.2021.126977, 2021.

Li, Y., Grimaldi, S., Pauwels, V. R. N., and Walker, J. P.: Hydrologic model calibration using remotely sensed soil moisture and discharge measurements: The impact on predictions at gauged and ungauged locations, J Hydrol (Amst), 557, 897–909, https://doi.org/10.1016/j.jhydrol.2018.01.013, 2018.

Lienert, S. and Joos, F.: A Bayesian ensemble data assimilation to constrain model parameters and land-use carbon emissions, Biogeosciences, 15, 2909–2930, https://doi.org/10.5194/bg-15-2909-2018, 2018.

Liu, X., Yang, K., Ferreira, V. G., and Bai, P.: Hydrologic Model Calibration With Remote Sensing Data Products in Global Large Basins, Water Resour Res, 58, e2022WR032929, https://doi.org/10.1029/2022WR032929, 2022.

Liu, Y., Zhuang, Q., Pan, Z., Miralles, D., Tchebakova, N., Kicklighter, D., Chen, J., Sirin, A., He, Y., Zhou, G., and Melillo, J.: Response of evapotranspiration and water

availability to the changing climate in Northern Eurasia, Clim Change, 126, 413–427, https://doi.org/10.1007/s10584-014-1234-9, 2014.

Livneh, B. and Lettenmaier, D. P.: Multi-criteria parameter estimation for the Unified Land Model, Hydrol Earth Syst Sci, 16, 3029–3048, https://doi.org/10.5194/hess-16-3029-2012, 2012.

Lo, M.-H., Famiglietti, J. S., Yeh, P. J.-F., and Syed, T. H.: Improving parameter estimation and water table depth simulation in a land surface model using GRACE water storage and estimated base flow data, Water Resour Res, 46, https://doi.org/10.1029/2009WR007855, 2010.

López López, P., Sutanudjaja, E. H., Schellekens, J., Sterk, G., and Bierkens, M. F. P.: Calibration of a large-scale hydrological model using satellite-based soil moisture and evapotranspiration products, Hydrol Earth Syst Sci, 21, 3125–3144, https://doi.org/10.5194/hess-21-3125-2017, 2017.

Masood, M., Yeh, P. J.-F., Hanasaki, N., and Takeuchi, K.: Model study of the impacts of future climate change on the hydrology of Ganges-Brahmaputra-Meghna basin, Hydrol Earth Syst Sci, 19, 747–770, https://doi.org/10.5194/hess-19-747-2015, 2015.

Mayer-Gürr, T., Behzadpour, S., Ellmer, M., Klinger, B., Kvas, A., Strasser, S., and Zehentner, N.: ITSG-Grace2018: The new GRACE time series from TU Graz. Abstract from GRACE / GRACE-FO Science Team Meeting 2018, Potsdam, Germany, 2018.

McMillan, H. K.: A review of hydrologic signatures and their applications, WIREs Water, 8, e1499, https://doi.org/10.1002/wat2.1499, 2021.

McMillan, H., Krueger, T., and Freer, J.: Benchmarking observational uncertainties for hydrology: rainfall, river discharge and water quality, Hydrol Process, 26, 4078–4111, https://doi.org/10.1002/hyp.9384, 2012.

Mei, Y., Mai, J., Do, H. X., Gronewold, A., Reeves, H., Eberts, S., Niswonger, R., Regan, R. S., and Hunt, R. J.: Can Hydrological Models Benefit From Using Global Soil Moisture, Evapotranspiration, and Runoff Products as Calibration Targets?, Water Resour Res, 59, e2022WR032064, https://doi.org/10.1029/2022WR032064, 2023.

Meyer Oliveira, A., Fleischmann, A. S., and Paiva, R. C. D.: On the contribution of remote sensing-based calibration to model hydrological and hydraulic processes in tropical regions, J Hydrol (Amst), 126184, https://doi.org/10.1016/j.jhydrol.2021.126184, 2021.

Michailovsky, C. I., Milzow, C., and Bauer-Gottwein, P.: Assimilation of radar altimetry to a routing model of the Brahmaputra River, Water Resour Res, 49, 4807–4816, https://doi.org/10.1002/wrcr.20345, 2013.

Milzow, C., Krogh, P. E., and Bauer-Gottwein, P.: Combining satellite radar altimetry, SAR surface soil moisture and GRACE total storage changes for hydrological model calibration in a large poorly gauged catchment, Hydrol. Earth Syst. Sci., 15, 1729–1743, https://doi.org/10.5194/hess-15-1729-2011, 2011.

Mirza, M. M. Q.: The Choice of Stage-Discharge Relationship for the Ganges and Brahmaputra Rivers in Bangladesh, Hydrology Research, 34, 321–342, https://doi.org/10.2166/nh.2003.0010, 2003.

Moges, E., Demissie, Y., Larsen, L., and Yassin, F.: Review: Sources of Hydrological Model Uncertainties and Advances in Their Analysis, Water (Basel), 13, https://doi.org/10.3390/w13010028, 2021.

Morris, M. D.: Factorial Sampling Plans for Preliminary Computational Experiments, Technometrics, 33, 161–174, https://doi.org/10.1080/00401706.1991.10484804, 1991.

Mueller, B., Hirschi, M., Jimenez, C., Ciais, P., Dirmeyer, P. A., Dolman, A. J., Fisher, J. B., Jung, M., Ludwig, F., Maignan, F., Miralles, D. G., McCabe, M. F., Reichstein, M., Sheffield, J., Wang, K., Wood, E. F., Zhang, Y., and Seneviratne, S. I.: Benchmark products for land evapotranspiration: LandFlux-EVAL multi-data set synthesis, Hydrol. Earth Syst. Sci., 17, 3707–3720, https://doi.org/10.5194/hess-17-3707-2013, 2013.

Müller Schmied, H., Eisner, S., Franz, D., Wattenbach, M., Portmann, F. T., Flörke, M., and Döll, P.: 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, https://doi.org/10.5194/hess-18-3511-2014, 2014.

Müller Schmied, H., Cáceres, D., Eisner, S., Flörke, M., Herbert, C., Niemann, C., Peiris, T. A., Popat, E., Portmann, F. T., Reinecke, R., Schumacher, M., Shadkam, S., Telteu, C.-E., Trautmann, T., and Döll, P.: The global water resources and use model WaterGAP v2.2d: model description and evaluation, Geosci Model Dev, 14, 1037–1079, https://doi.org/10.5194/gmd-14-1037-2021, 2021.

Nanteza, J., de Linage, C. R., Thomas, B. F., and Famiglietti, J. S.: Monitoring groundwater storage changes in complex basement aquifers: An evaluation of the GRACE

satellites over East Africa, Water Resour Res, 52, 9542–9564, https://doi.org/10.1002/2016WR018846, 2016.

Narasimhan, T. N.: A note on India's water budget and evapotranspiration, Journal of Earth System Science, 117, 237–240, https://doi.org/10.1007/s12040-008-0028-8, 2008.

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

Nijzink, R. C., Almeida, S., Pechlivanidis, I. G., Capell, R., Gustafssons, D., Arheimer, B., Parajka, J., Freer, J., Han, D., Wagener, T., van Nooijen, R. R. P., Savenije, H. H. G., and Hrachowitz, M.: Constraining Conceptual Hydrological Models With Multiple Information Sources, Water Resour Res, 54, 8332–8362, https://doi.org/10.1029/2017WR021895, 2018.

Orth, R. and Seneviratne, S. I.: Introduction of a simple-model-based land surface dataset for Europe, Environmental Research Letters, 10, 44012, https://doi.org/10.1088/1748-9326/10/4/044012, 2015.

Papa, F. and Frappart, F.: Surface Water Storage in Rivers and Wetlands Derived from Satellite Observations: A Review of Current Advances and Future Opportunities for Hydrological Sciences, Remote Sens (Basel), 13, https://doi.org/10.3390/rs13204162, 2021.

Papa, F., Prigent, C., Durand, F., and Rossow, W. B.: Wetland dynamics using a suite of satellite observations: A case study of application and evaluation for the Indian Subcontinent, Geophys Res Lett, 33, https://doi.org/10.1029/2006GL025767, 2006.

Papa, F., Durand, F., Rossow, W. B., Rahman, A., and Bala, S. K.: Satellite altimeter-derived monthly discharge of the Ganga-Brahmaputra River and its seasonal to interannual variations from 1993 to 2008, J Geophys Res Oceans, 115, https://doi.org/10.1029/2009JC006075, 2010.

Papa, F., Frappart, F., Güntner, A., Prigent, C., Aires, F., Getirana, A. C. V, and Maurer, R.: Surface freshwater storage and variability in the Amazon basin from multi-satellite observations, 1993–2007, Journal of Geophysical Research: Atmospheres, 118, 2013JD020500, https://doi.org/10.1002/2013JD020500, 2013.

Papa, F., Frappart, F., Malbeteau, Y., Shamsudduha, M., Vuruputur, V., Sekhar, M., Ramillien, G., Prigent, C., Aires, F., Pandey, R. K., Bala, S., and Calmant, S.: Satellite-derived

surface and sub-surface water storage in the Ganges–Brahmaputra River Basin, J Hydrol Reg Stud, 4, 15–35, https://doi.org/10.1016/j.ejrh.2015.03.004, 2015.

Pellet, V., Aires, F., Papa, F., Munier, S., and Decharme, B.: Long-term total water storage change from a Satellite Water Cycle reconstruction over large southern Asian basins, Hydrol Earth Syst Sci, 24, 3033–3055, https://doi.org/10.5194/hess-24-3033-2020, 2020.

Pianosi, F., Sarrazin, F., and Wagener, T.: A Matlab toolbox for Global Sensitivity Analysis, Environmental Modelling & Software, 70, 80–85, https://doi.org/10.1016/j.envsoft.2015.04.009, 2015.

Pianosi, F., Beven, K., Freer, J., Hall, J. W., Rougier, J., Stephenson, D. B., and Wagener, T.: Sensitivity analysis of environmental models: A systematic review with practical workflow, Environmental Modelling & Software, 79, 214–232, https://doi.org/10.1016/j.envsoft.2016.02.008, 2016.

Rakovec, O., Kumar, R., Attinger, S., and Samaniego, L.: Improving the realism of hydrologic model functioning through multivariate parameter estimation, Water Resour Res, 52, 7779–7792, https://doi.org/10.1002/2016WR019430, 2016.

Ray, P. A., Yang, Y.-C. E., Wi, S., Khalil, A., Chatikavanij, V., and Brown, C.: Room for improvement: Hydroclimatic challenges to poverty-reducing development of the Brahmaputra River basin, Environ Sci Policy, 54, 64–80, https://doi.org/10.1016/j.envsci.2015.06.015, 2015.

Reed, P. M. and Hadka, D.: Evolving many-objective water management to exploit exascale computing, Water Resour Res, 50, 8367–8373, https://doi.org/10.1002/2014WR015976, 2014.

Reed, P. M., Hadka, D., Herman, J. D., Kasprzyk, J. R., and Kollat, J. B.: Evolutionary multiobjective optimization in water resources: The past, present, and future, Adv Water Resour, 51, 438–456, https://doi.org/10.1016/j.advwatres.2012.01.005, 2013.

Salameh, E., Frappart, F., Papa, F., Güntner, A., Venugopal, V., Getirana, A., Prigent, C., Aires, F., Labat, D., and Laignel, B.: Fifteen Years (1993–2007) of Surface Freshwater Storage Variability in the Ganges-Brahmaputra River Basin Using Multi-Satellite Observations, Water (Basel), 9, https://doi.org/10.3390/w9040245, 2017.

Schneider, R., Godiksen, P. N., Villadsen, H., Madsen, H., and Bauer-Gottwein, P.: Application of CryoSat-2 altimetry data for river analysis and modelling, Hydrol Earth Syst Sci, 21, 751–764, https://doi.org/10.5194/hess-21-751-2017, 2017.

Schneider, U., Becker A., Finger P., Meyer-Christoffer A., Rudolf B., and Ziese, M.: GPCC full data monthly productversion 7.0 at 0.5: Monthly land-surface precipitationfrom rain-gauges built on GTS-based and historic data, https://doi.org/10.5676/DWD_GPCC/FD_M_V7_050, 2015.

Schumacher, M., Forootan, E., van Dijk, A. I. J. M., Schmied, H. M., Crosbie, R. S., Kusche, J., and Döll, P.: Improving drought simulations within the Murray-Darling Basin by combined calibration/assimilation of GRACE data into the WaterGAP Global Hydrology Model, Remote Sens Environ, 204, 212–228, https://doi.org/10.1016/j.rse.2017.10.029, 2018.

Sir William Halocrow and Partners Ltd. (1991). River Training Studies of the Brahmaputra River. First Interim Report, Annex 1, Part 4: Analysis of Discharge Measurements. Sir William Halocrow and Partners Ltd., Dhaka. Cited in Mirza, M. M. Q. (2003). The choice of stage-discharge relationship for the Ganges and Brahmaputra rivers in Bangladesh. Hydrology Research, 34(4), 321-342.

Soares, L. M. V and Calijuri, M. C.: Sensitivity and identifiability analyses of parameters for water quality modeling of subtropical reservoirs, Ecol Modell, 458, 109720, https://doi.org/10.1016/j.ecolmodel.2021.109720, 2021.

Trautmann, T., Koirala, S., Carvalhais, N., Eicker, A., Fink, M., Niemann, C., and Jung, M.: Understanding terrestrial water storage variations in northern latitudes across scales, Hydrol. Earth Syst. Sci., 22, 4061–4082, https://doi.org/10.5194/hess-22-4061-2018, 2018.

Trautmann, T., Koirala, S., Carvalhais, N., Güntner, A., and Jung, M.: The importance of vegetation in understanding terrestrial water storage variations, Hydrol. Earth Syst. Sci., 26, 1089–1109, https://doi.org/10.5194/hess-26-1089-2022, 2022.

Tsarouchi, G. M., Buytaert, W., and Mijic, A.: Coupling a land-surface model with a crop growth model to improve ET flux estimations in the Upper Ganges basin, India, Hydrol Earth Syst Sci, 18, 4223–4238, https://doi.org/10.5194/hess-18-4223-2014, 2014.

Wagener, T., McIntyre, N., Lees, M. J., Wheater, H. S., and Gupta, H. V: Towards reduced uncertainty in conceptual rainfall-runoff modelling: dynamic identifiability analysis, Hydrol Process, 17, 455–476, https://doi.org/10.1002/hyp.1135, 2003.

Wang, J., Wei, J., Shan, W., and Zhao, J.: Modeling the water-energy-food-environment nexus and transboundary cooperation opportunity in the Brahmaputra River Basin, J Hydrol Reg Stud, 49, 101497, https://doi.org/10.1016/j.ejrh.2023.101497, 2023.

Weedon, G. P., Gomes, S., Viterbo, P., Shuttleworth, W. J., Blyth, E., Österle, H., Adam, J. C., Bellouin, N., Boucher, O., and Best, M.: 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, https://doi.org/10.1175/2011JHM1369.1, 2011.

Werth, S. and Güntner, A.: Calibration analysis for water storage variability of the global hydrological model WGHM, Hydrol. Earth Syst. Sci., 14, 59–78, https://doi.org/10.5194/hess-14-59-2010, 2010.

Werth, S., Güntner, A., Petrovic, S., and Schmidt, R.: Integration of GRACE mass variations into a global hydrological model, Earth Planet Sci Lett, 277, 166–173, https://doi.org/10.1016/j.epsl.2008.10.021, 2009.

Wu, X., Shirvan, K., and Kozlowski, T.: Demonstration of the relationship between sensitivity and identifiability for inverse uncertainty quantification, J Comput Phys, 396, 12–30, https://doi.org/10.1016/j.jcp.2019.06.032, 2019.

Yang, D., Xu, X., and Scanlon, B. R.: Multisource remote sensing data facilitate ecohydrological simulations without runoff calibration, Hydrol Process, 36, e14773, https://doi.org/10.1002/hyp.14773, 2022.

Yassin, F., Razavi, S., Wheater, H., Sapriza-Azuri, G., Davison, B., and Pietroniro, A.: Enhanced identification of a hydrologic model using streamflow and satellite water storage data: A multicriteria sensitivity analysis and optimization approach, Hydrol Process, 31, 3320–3333, https://doi.org/10.1002/hyp.11267, 2017.

Yoshida, T., Hanasaki, N., Nishina, K., Boulange, J., Okada, M., and Troch, P. A.: Inference of Parameters for a Global Hydrological Model: Identifiability and Predictive Uncertainties of Climate-Based Parameters, Water Resour Res, 58, e2021WR030660, https://doi.org/10.1029/2021WR030660, 2022.

Zheng, H., Chiew, F. H. S., Charles, S., and Podger, G.: Future climate and runoff projections across South Asia from CMIP5 global climate models and hydrological modelling, J Hydrol Reg Stud, 18, 92–109, https://doi.org/10.1016/j.ejrh.2018.06.004, 2018.