Learning from satellite observations: increased understanding of catchment processes through stepwise model improvement

Petra Hulsman[1], Hubert H.G. Savenije[1], Markus Hrachowitz[1]

[1]Water Resources Section, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Stevinweg 1, 2628 CN Delft, The Netherlands

Correspondence to: Petra Hulsman (p.hulsman@tudelft.nl)

Abstract. Satellite observations can provide valuable information for a better understanding of hydrological processes and thus serve as valuable tools for model structure development and improvement. While model calibration and evaluation has in recent years started to make increasing use of spatial, mostly remotely-sensed information, model structural development largely remains to rely on discharge observations at basin outlets only. Due to the ill-posed inverse nature and the related equifinality issues in the modelling process, this frequently results in poor representations of the spatiotemporal heterogeneity of system-internal processes, in particular for large river basins. The objective of this study is thus to explore the value of remotely-sensed, gridded data to improve our understanding of the processes underlying this heterogeneity and, as a consequence, their quantitative representation in models through a stepwise adaptation of model structures and parameters. For this purpose, a distributed, process-based hydrological model was developed for the study region, the poorly gauged Luangwa river basin. As a first step, this benchmark model was calibrated to discharge data only and, in a post-calibration evaluation procedure, tested for its ability to simultaneously reproduce (1) the basin-average temporal dynamics of remotely-sensed evaporation and total water storage anomalies, and (2) their temporally-averaged spatial patterns. This allowed the diagnosis of model structural deficiencies in reproducing these temporal dynamics and spatial patterns. Subsequently, the model structure was adapted in a step-wise procedure, testing five additional alternative process hypotheses that could potentially better describe the observed dynamics and pattern. These included, on the one hand, the addition and testing of alternative formulations of groundwater upwelling into wetlands as function of the water storage and, on the other hand, alternative spatial discretizations of the groundwater reservoir. Similar to the benchmark, each alternative model hypothesis was, in a next step, calibrated to discharge only and tested against its ability to reproduce the observed spatiotemporal pattern in evaporation and water storage anomalies. In a final step, all models were re-calibrated to discharge, evaporation and water storage anomalies simultaneously. The results indicated that (1) the benchmark model (Model A) could reasonably well reproduce the time series of observed discharge, basin-average evaporation and total water storage. In contrast, it poorly represented time series of evaporation in wetland dominated areas as well as the spatial pattern of evaporation and total water storage. (2) Step-wise adjusting the model structure (Models B – F) suggested that Model F, allowing for upwelling groundwater from a distributed representation of the groundwater reservoir and (3) simultaneously calibrating the model with respect to multiple variables, i.e. discharge, evaporation and total water storage anomalies, provided the best representation of all these variables with respect to their temporal dynamics and spatial patterns, except for the basin-average temporal dynamics in the total water storage anomalies. It was shown that satellite-based evaporation and total water storage anomaly data are not only valuable for multi-criteria calibration, but can play an important role in improving our understanding of hydrological processes through diagnosing model deficiencies and step-wise model structural improvement.

1. Introduction

Traditionally, discharge observations at basin outlets are used for hydrological model development and calibration, which can be a robust strategy in small watersheds with relatively uniform characteristics such as topography and land cover, but not for larger, heterogeneous basins (Blöschl and Sivapalan, 1995; Daggupati et al., 2015). As a result, temporal dynamics of discharge may be well reproduced. This however, does not ensure that the spatial patterns and temporal dynamics of model internal storage and flux variables provide a meaningful representation of their real pattern and dynamics (Beven, 2006b; Kirchner, 2006; Clark et al., 2008; Gupta et al., 2008; Hrachowitz et al., 2014; Garavaglia et al., 2017). Especially in large, poorly gauged basins this traditional model calibration and testing method is likely to result in a poor representation of spatial variability (Daggupati et al., 2015) due to equifinality and the related the boundary flux problem (Beven, 2006b).

An increasing number of satellite-based observations have become available over the last decade, giving us insight into a wide range of hydrology-relevant variables, including precipitation, total water storage anomalies, evaporation, surface soil moisture or river width (Xu et al., 2014; Jiang and Wang, 2019). These data are increasingly used as model forcing or for parameter selection and model calibration (e.g. Li et al., 2015; Mazzoleni et al., 2019; Tang et al., 2019).

Many studies used a single satellite product in the calibration procedure, some of them additionally using discharge data, others not. For instance, hydrological models have been calibrated with respect to evaporation (e.g. Immerzeel and Droogers, 2008; Winsemius et al., 2008; Vervoort et al., 2014; Bouaziz et al., 2018; Odusanya et al., 2019), water storage anomalies from GRACE (Gravity Recovery and Climate Experiment, Werth et al., 2009), river width (Revilla-Romero et al., 2015; Sun et al., 2018) or river altimetry (Getirana, 2010; Michailovsky et al., 2013; Sun et al., 2015; Hulsman et al., 2019).

Other studies simultaneously calibrated hydrological models with respect to multiple remotely-sensed variables, but only exploiting basin-average time series, without consideration for spatial patterns (e.g. Milzow et al., 2011; López et al., 2017; Kittel et al., 2018; Nijzink et al., 2018). On the other hand, some studies exclusively calibrated models to spatial patterns of observed variables (Stisen et al., 2011; Koch et al., 2016; Mendiguren et al., 2017; Demirel et al., 2018; Zink et al., 2018). As most satellite-based observations such as evaporation are not measured directly but are themselves a result of underlying models using satellite data as input (Xu et al., 2014), more focus has been recently placed on calibration to the relative spatial variability instead of using absolute magnitudes (Stisen et al., 2011; van Dijk and Renzullo, 2011; Dembélé et al., 2020).

To fully exploit the information content of satellite-based observations, simultaneous model calibration on both, temporal dynamics and spatial patterns of multiple variables has the potential to improve the representation of spatiotemporal variability and, linked to that, their underlying model internal processes and therefore the model realism (Rientjes et al., 2013; Rakovec et al., 2016; Herman et al., 2018). Strikingly, only a few studies so far used satellite-based observations to calibrate with respect to the temporal and spatial variation simultaneously (Rajib et al., 2018; Dembélé et al., 2020).

In general, most studies that made use of remotely-sensed data for model applications have exclusively addressed the problem of parameter selection and thus model calibration. However, as models are always abstract and simplified representations of reality, every model structure needs to be understood as a hypothesis to be tested (Clark et al., 2011; Fenicia et al., 2011; Hrachowitz and Clark, 2017). Yet, most studies on model structural improvement have so far only relied on spatially aggregated variables (Fenicia et al., 2008; Kavetski and Fenicia, 2011; Hrachowitz et al., 2014; Nijzink et al., 2016), while spatial data remain rarely used for that purpose (e.g. Fenicia et al., 2016; Roy et al., 2017).

The overall objective of this paper is therefore to explore the simultaneous use of spatial patterns and temporal dynamics of satellite-based evaporation and total water storage observations for a step-wise structural improvement and calibration of hydrological models for a large river system in a semi-arid, data scarce region. More specifically, we tested the research hypotheses that (1) spatial patterns and temporal dynamics in satellite-based evaporation and water storage anomaly data contain relevant information to diagnose and to iteratively improve on model structural deficiencies and that (2) these data, when simultaneously used with discharge data for calibration, do contain sufficient information for a more robust parameter selection.

2. Site description

The Luangwa River in Zambia is a large, mostly unregulated tributary of the Zambezi with a length of about 770 km (Figure 1). This poorly gauged river basin has an area of 159,000 km^2 which is mostly covered with deciduous forest, shrubs and savanna and where an elevation difference up to 1850 m can be found between the highlands and low lands along the river (The World Bank, 2010; Hulsman et al., 2019). In this semi-arid basin, the mean annual evaporation (1555 mm yr^{-1}) exceeds the mean annual precipitation (970 mm yr^{-1}).

The Luangwa River flows into the Zambezi upstream of the Cahora Bassa Dam which is used for hydropower production, and flood and drought protection. The operation of this dam is very difficult since there is only limited information available from the poorly gauged upstream tributaries (SADC, 2008; Schleiss and Matos, 2016). As a result, the local population has in the past suffered from severe floods and droughts (ZAMCOM et al., 2015; Schumann et al., 2016).

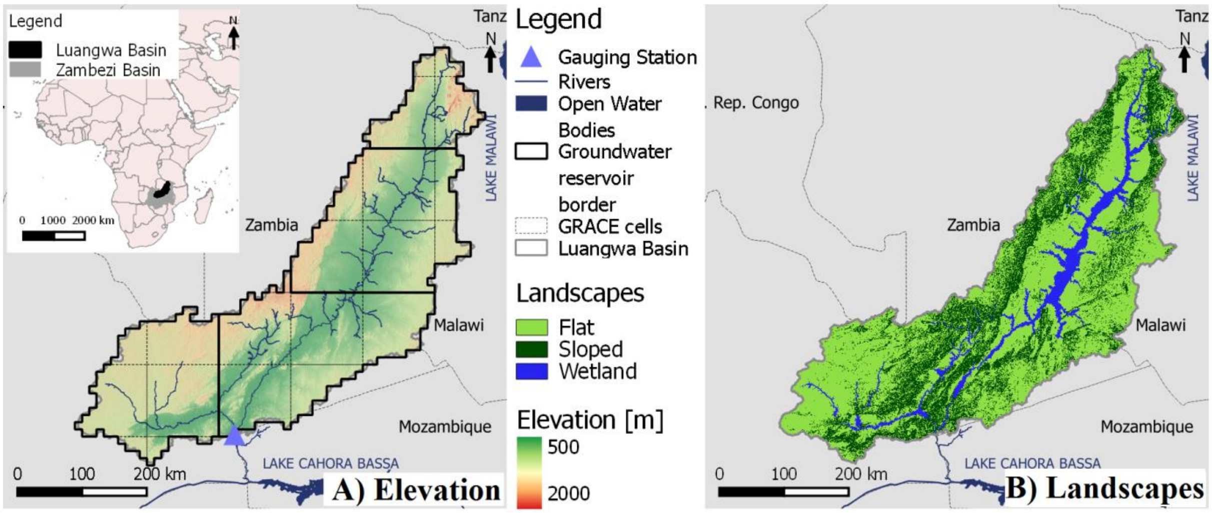

Figure 1: Map of the Luangwa River Basin in Zambia with A) the elevation, groundwater reservoir units at 0.1° resolution and 1° grid according to GRACE, and B) the main landscape types

2.1 Data availability

2.1.1 In-situ discharge observations

Historical daily in-situ discharge data was available from the Zambian Water Resources Management Authority at the Great East Road Bridge gauging station, located at 30^o 13’ E and 14^o 58’ S (Figure 1), for the time period 2004 to 2016 yet containing considerable gaps resulting in a temporal coverage of 31%.

2.1.2 Spatially gridded observation

Spatially gridded data were used for a topography-based landscape classification into hydrological response units (HRU, Savenije, 2010), as model forcing (precipitation and temperature) and for parameter selection (evaporation and total water storage, see Table 1).

More specifically, topography was extracted from GMTED with a spatial resolution of 0.002^o (Danielson and Gesch, 2011). Daily precipitation data was extracted from CHIRPS (Climate Hazards Group InfraRed Precipitation with Station) with a spatial resolution of 0.05^o. Monthly temperature data extracted from CRU at a spatial resolution of 1^o was used to estimate the potential evaporation applying the Hargreaves method (Hargreaves and Samani, 1985; Hargreaves and Allen, 2003). These monthly observations were interpolated to daily timescale using daily averaged in-situ temperature measured at two locations with the coordinates 28^o 30’ E, 14^o 24’ S and 32^o 35’ E, 13^o 33’ S. The satellite-based total evaporation data was extracted from WaPOR (Water Productivity Open-access portal, FAO, 2018) version 1.1 as it proved to perform well in African river basins (Weerasinghe et al., 2019). This product was available on 10-day temporal and 250 m spatial resolution. Satellite-based observations on the total water storage anomalies were extracted from the Gravity Recovery and Climate Experiment

(GRACE). With two identical GRACE satellites, the variations in the Earth's gravity field were measured to detect regional mass changes which are dominated by variations in the terrestrial water storage after having accounted for atmospheric and oceanic effects (Landerer and Swenson, 2012; Swenson, 2012). In this study, the long-term bias between the discharge, evaporation (WaPOR) and total water storage anomalies (GRACE) was corrected by multiplying the evaporation with a correction factor of 1.08 to close the long-term water balance.

The gridded information provided for the precipitation, temperature and evaporation were rescaled to the model resolution of 0.1˚. If the resolution of the satellite product was higher than 0.1˚, then the mean of all cells located within each model cell was used. Otherwise, each cell of the satellite product was divided into multiple cells such that the model resolution is obtained, retaining the original value. In contrast, the modelled total water storage was rescaled to 1°, the resolution of the GRACE data set, by taking the mean.

Table 1: Data used in this study

	Time period	Time Resolution	Spatial resolution	Product Name	Source
Digital elevation map	NA	NA	0.002°	GMTED	(Danielson and Gesch, 2011)
Precipitation	2002 – 2016	Daily	0.05°	CHIRPS	(Funk et al., 2014)
Temperature	2002 – 2016	Monthly	0.5°	CRU	(University of East Anglia Climatic Research Unit et al., 2017)
Evaporation	2009 – 2016	10-day	0.00223°	WaPOR	(FAO, 2018; FAO and IHE Delft, 2019)
Total water storage	2002 – 2016	Monthly	1°	GRACE	(Swenson and Wahr, 2006; Landerer and Swenson, 2012; Swenson, 2012)
Discharge (Luangwa Bridge gauging station)	2004 – 2016	Daily	NA	NA	WARMA

3. Modelling approach

A previously developed and tested (Hulsman et al., 2019) distributed, process-based hydrological model was implemented for the Luangwa Basin, see Section 3.1 for more information. This benchmark model (Model A) was calibrated with respect to discharge for the time period 2002 – 2012 and validated for the time period 2012 – 2016 with respect to discharge, evaporation and total water storage anomalies. Then, the model was calibrated with respect to all above variables, hence discharge, evaporation and total water storage anomalies simultaneously, for the time period 2002 – 2012 and validated with respect to the same variables for the time period 2012 – 2016. Model deficiencies were then diagnosed for this benchmark model (Model A) based on the results of both calibration strategies.

Next, model structure changes were applied creating Models B – D to improve the deficiencies found in Model A. These changes concerned the groundwater upwelling into the unsaturated zone as explained in Section 4.2. The same calibration and validation strategies as applied to Model A were applied to Models B – D. Model improvements were evaluated and further deficiencies were diagnosed for these models based on the calibration and validation results.

To improve the deficiencies diagnosed in Models B – D, further model structural changes, i.e. increased levels of spatial discretisation of the saturated zone as explained in Section 4.3, resulted in Models E and F. Similar to the previous models, the same calibration and validation strategies were applied to Models E and F, and model improvements and deficiencies were diagnosed based on the calibration and validation model performances.

The calculation of the model performance with respect to discharge, evaporation and total water storage are explained in Section 3.2. The calibration and validation procedures are described in Sections 3.3 and 3.4.

3.1 Hydrological models

3.1.1 Benchmark model (Model A)

This model is a process-based hydrological model developed in a previous study by Hulsman et al. (2019) for the Luangwa basin. In this model, the water accounting was distributed by discretizing the basin and using spatially distributed forcing data while the same model structure and parameter set were used for the entire basin. Each 0.1° x 0.1° model cell was then further discretized into functionally distinct landscape classes, i.e. hydrological response units (HRU), inferred from topography (Figure 1B), but connected by a common groundwater component (Euser et al., 2015) following the FLEX-Topo modelling concept (Savenije, 2010) which was previously successfully applied in many different and climatically contrasting regions (Gao et al., 2014; Gharari et al., 2014; 2016; Nijzink et al., 2016). Here, the landscape was classified based on the local slope and “Height-above-the-nearest-drainage” (HAND, Rennó et al., 2008) into sloped areas (slope ≥ 4%), flat areas (slope < 4%, HAND ≥ 11 m) and wetlands (slope < 4%, HAND < 11 m). For this purpose, the drainage network was derived from a digital elevation map extracted from GMTED (Section 2.1.2) using a flow accumulation map after having burned-in a river network map extracted from OpenStreetMap (https://wiki.openstreetmap.org/wiki/Shapefiles) to obtain an as accurate as possible drainage network as done successfully in previous studies (Seyler et al., 2009). According to this classification, the wetland areas covered 8% of the basin, flat areas 64% and sloped areas 28% (Figure 1).

The model consisted of different storage components schematised as reservoirs representing interception and unsaturated storage, as well as a slow responding reservoir, representing the groundwater and a fast responding reservoir (Figure 2). The water balance for each reservoir and the associated constitutive equations are summarised in Table 2. The individual model structures of each parallel HRU were very similar. Functional differences between HRUs were thus mostly accounted for by different parameter sets. To allow the use of partly overlapping prior parameter distributions while maintaining relationships between parameters of individual HRUs that are consistent with our physical understanding of the system and to limit equifinality, model process constraints (Gharari et al., 2014; Hrachowitz et al., 2014) were applied for several parameters (Table 3). For instance, in the Luangwa Basin, the sloped areas are dominated by dense vegetation, suggesting higher interception capacities and larger storage capacities in the unsaturated zone compared to the remaining part of the basin. In addition, for each HRU the model structure was adjusted where necessary to include processes unique to that area. For instance, water percolates and recharges the groundwater system in sloped and flat areas whereas in wetlands this was assumed to be negligible due to groundwater tables that are very shallow and thus close to the surface.

The runoff was first calculated for each individual grid cell. A simple routing scheme based on the flow direction and constant flow velocity as calibration parameter was applied to estimate the flow at the outlet. In total, this model consisted of 16 calibration parameters with uniform prior distributions and constraints as summarized in Table 3.

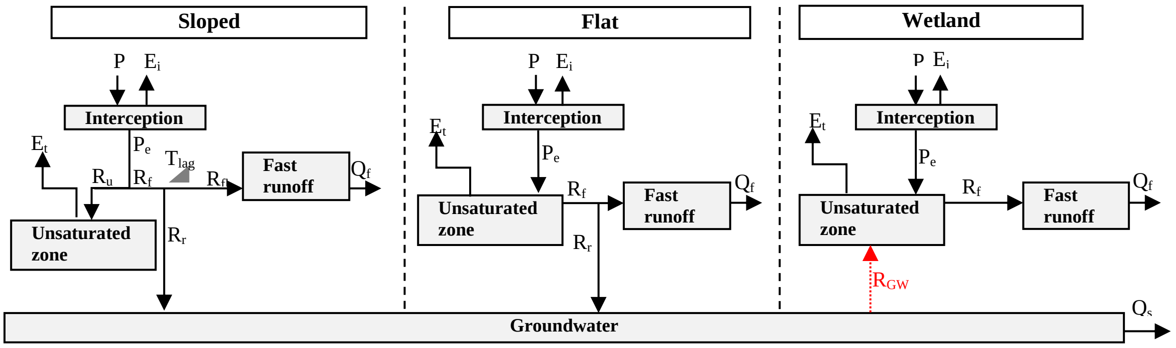

Figure 2: Schematisation of the model structure applied to each grid cell. Symbol explanation: precipitation (*P*), effective precipitation (*P*$_e$), interception evaporation (*E*$_i$), plant transpiration (*E*$_t$), infiltration into the unsaturated root zone (*R*$_u$), drainage to fast runoff component (*R*$_f$), delayed fast runoff (*R*$_{fl}$), lag time (*T*$_{lag}$), groundwater recharge (*R*$_r$), upwelling groundwater flux (*R*$_{GW}$), fast runoff (*Q*$_f$), groundwater/slow runoff (*Q*$_s$).

Table 2: Equations applied in the hydrological model. Fluxes [mm d^{-1}]: precipitation (P), effective precipitation (P_e), potential evaporation (E_p), interception evaporation (E_i), plant transpiration (E_t), infiltration into the unsaturated zone (R_u), drainage to fast runoff component (R_f), delayed fast runoff (R_{fl}), groundwater recharge (R_r for each relevant HRU and $R_{r,tot}$ combining all relevant HRUs), groundwater upwelling (R_{GW} for each relevant HRU and $R_{GW,tot}$ combining all relevant HRUs), fast runoff (Q_f for each HRU and $Q_{f,tot}$ combining all HRUs), groundwater/slow runoff (Q_s), total runoff (Q_m). Storages [mm]: storage in interception reservoir (S_i), storage in unsaturated root zone (S_u), storage in groundwater/slow reservoir (S_s), storage in fast reservoir (S_f). Parameters: interception capacity (I_{max}) [mm], maximum upwelling groundwater (C_{max}) [mm d^{-1}], maximum root zone storage capacity (S_{umax}) [mm], reference storage in the saturated zone ($S_{s,ref}$) [mm], splitter (W) [-], shape parameter (β) [-], transpiration coefficient (C_e) [-], time lag (T_{lag}) [d], exponent (γ) [-], reservoir time scales [d] of fast (K_f) and slow (K_s) reservoirs, areal weights for each grid cell (p_{HRU}) [-],time step (Δt) [d]. Model calibration parameters are shown in bold letters in the table below. The equations were applied to each hydrological response unit (HRU) unless indicated differently.

Reservoir system	Water balance equation	Equation	Process functions	Equation
Interception	$\frac{\Delta S_\mathrm{i}}{\Delta t} = P - P_\mathrm{e} - E_\mathrm{i}$	(1)	$E_\mathrm{i} = \min\left(E_\mathrm{p}, \min\left(P, \frac{\boldsymbol{I}_\mathbf{max}}{\Delta t}\right)\right)$	(2)
			$P_\mathrm{e} = P - E_\mathrm{i}$	(3)
Unsaturated zone	Sloped: $\frac{\Delta S_\mathrm{u}}{\Delta t} = R_\mathrm{u} - E_\mathrm{t}$	(4)	$E_\mathrm{t} = \min\left((E_\mathrm{p} - E_\mathrm{i}), \min\left(\frac{S_\mathrm{u}}{\Delta t}, (E_\mathrm{p} - E_\mathrm{i}) \cdot \frac{S_\mathrm{u}}{\boldsymbol{S}_\mathbf{u,max}} \cdot \frac{1}{\boldsymbol{C}_\mathbf{e}}\right)\right)$	(5)
	Flat: $\frac{\Delta S_\mathrm{u}}{\Delta t} = \mathrm{P_e} - E_\mathrm{t} - R_f$	(6)	Model A: $R_\mathrm{GW} = 0$	(7)
	Wetland: $\frac{\Delta S_\mathrm{u}}{\Delta t} = \mathrm{P_e} - E_\mathrm{t} - R_f + R_\mathrm{GW}$	(8)	Model B: $R_\mathrm{GW} = \min\left(\left(1 - \frac{S_\mathrm{u}}{\boldsymbol{S}_\mathbf{u,max}}\right) \cdot \boldsymbol{C}_\mathbf{max}, \frac{\frac{S_\mathrm{s}}{\Delta t}}{p_\mathrm{HRU}}\right)$	(9)
			Model C, E, F: $R_\mathrm{GW} = \min\left(\frac{\min(S_\mathrm{s}, \boldsymbol{S}_\mathbf{s,ref})}{\boldsymbol{S}_\mathbf{s,ref}} \cdot \boldsymbol{C}_\mathbf{max}, \frac{\frac{S_\mathrm{s}}{\Delta t}}{p_\mathrm{HRU}}\right)$	(10)
			Model D: $R_\mathrm{GW} = \min\left(\left(\frac{\min(S_\mathrm{s}, \boldsymbol{S}_\mathbf{s,ref})}{\boldsymbol{S}_\mathbf{s,ref}}\right)^{\boldsymbol{\gamma}} \cdot \boldsymbol{C}_\mathbf{max}, \frac{\frac{S_\mathrm{s}}{\Delta t}}{p_\mathrm{HRU}}\right)$	(11)
			if $S_\mathrm{u} + R_\mathrm{GW} \cdot \Delta t > S_\mathrm{u,max}$: $R_\mathrm{GW} = \frac{S_\mathrm{u,max} - S_\mathrm{u}}{\Delta t}$	(12)
			Sloped: $R_\mathrm{u} = (1 - C) \cdot P_\mathrm{e}$	(13)
			$C = 1 - \left(1 - \frac{S_\mathrm{u}}{\boldsymbol{S}_\mathbf{u,max}}\right)^{\boldsymbol{\beta}}$	(14)
Fast runoff	$\frac{\Delta S_\mathrm{f}}{\Delta t} = R_\mathrm{fl} - Q_\mathrm{f}$	(15)	$Q_\mathrm{f} = \frac{S_\mathrm{f}}{K_\mathrm{f}}$	(16)
			Flat/Wetland: $R_\mathrm{f} = \frac{\max(0, S_\mathrm{u} - \boldsymbol{S}_\mathbf{u,max})}{\Delta t}$	(17)
			$R_{fl} = R_f$	(18)
			Sloped: $R_\mathrm{f} = (1 - \boldsymbol{W}) \cdot C \cdot P_\mathrm{e}$	(19)
			$R_\mathrm{fl} = R_\mathrm{f} * \mathrm{f}(\boldsymbol{T}_\mathbf{lag})$	(20)
Groundwater	$\frac{\Delta S_\mathrm{s}}{\Delta t} = R_{\mathrm{r}_\mathrm{tot}} - R_{\mathrm{GW}_\mathrm{tot}} - Q_\mathrm{s}$	(21)	$R_\mathrm{r} = \boldsymbol{W} \cdot C \cdot P_e$	(22)
			$R_{\mathrm{r}_\mathrm{tot}} = \sum_\mathrm{HRU} p_\mathrm{HRU} \cdot R_\mathrm{r}$	(23)
			$R_{\mathrm{GW}_\mathrm{tot}} = \sum_\mathrm{HRU} p_\mathrm{HRU} \cdot R_\mathrm{GW}$	(24)
			$Q_\mathrm{s} = \frac{S_\mathrm{s}}{K_\mathrm{s}}$	(25)
Total runoff	$Q_\mathrm{m} = Q_\mathrm{s} + Q_{\mathrm{f}_\mathrm{tot}}$	(26)	$Q_{\mathrm{f}_\mathrm{tot}} = \sum_\mathrm{HRU} p_\mathrm{HRU} \cdot Q_\mathrm{f}$	(27)
Supporting literature	(Gao et al., 2014; Gharari et al., 2014; Euser et al., 2015; Hulsman et al., 2019)			

Table 3: Model parameter and ranges (Hulsman et al., 2019)

Landscape class	Parameter	min	max	Unit	Constraint	Comment
Entire basin	C_e	0	1	-		
	K_s	50	200	d		
	S_{sref}	100	500	mm		Only for Models C to F
Flat	I_{max}	0	5	mm d^{-1}		
	$S_{u,max}$	300	1000	mm		
	K_f	10	12	d		
	W	0.5	0.95	-		
Sloped	I_{max}	0	5	mm d^{-1}	$I_\mathrm{max,sloped} > I_\mathrm{max,flat}$	
	S_{umax}	300	1000	mm	$S_\mathrm{umax,sloped} > S_\mathrm{umax,flat}$	
	β	0	2	-		
	T_{lag}	1	5	d		
	K_f	10	12	d		
	W	0.5	0.95	-	$W_\mathrm{sloped} > W_\mathrm{flat}$	
Wetland	I_{max}	0	5	mm d^{-1}	$I_\mathrm{max,wetland} < I_\mathrm{max,sloped}$	
	S_{umax}	10	500	mm	$S_\mathrm{umax,wetland} < S_\mathrm{umax,sloped}$	
	K_f	10	12	d		
	C_{max}	0.1	5	mm d^{-1}		Only for Models B to F
	γ	0.01	0.5	-		Only for Model D
River profile	v	0.01	5.0	m s^{-1}		

3.1.2 First model adaptation: Adding groundwater upwelling (Models B – D)

Satellite-based evaporation and total water storage observations were used to evaluate the benchmark Model A with respect to the spatial and temporal variability visually and using model performance metrics as described in Section 3.2 to detect model deficiencies in these system-internal variables. The first model adaptation was applied to improve the hydrological model with respect to the deficiencies detected in Model A. Therefore, a detailed description of the reasoning behind the first model adaptation was explained in Section 4.2 after having described the deficiencies in Model A in Section 4.1.3.

In short, groundwater upwelling (R_{GW}) was added in wetland areas (see Figure 2). This upwelling groundwater was made (1) a linear function of the water content in the unsaturated reservoir (Model B, Eq.9 in Table 2), (2) a linear function of the water content in the slow responding reservoir (Model C, Eq.10) and (3) a non-linear function of the water content in the slow responding reservoir (Model D, Eq.11). As a result, upwelling water from the saturated zone feeds the unsaturated zone, controlled by the water content in the unsaturated (Model B) or in the saturated zone (Models C and D), and thus increasing the water availability for transpiration from the unsaturated zone in wetland areas. Compared to the benchmark Model A, Model B introduces one additional calibration parameter, Model C two and Model D three (Tables 2 and 3).

3.1.3 Second model adaptation: Discretizing the groundwater system (Models E – F)

Similar to the first model adaptation, the second model adaptation was developed to improve deficiencies detected in Models B – D. Therefore, a detailed description of the reasoning behind the second model adaptation was explained in Section 4.3 after having described the deficiencies in Models B – D in Section 4.2.3.

In short, the spatial resolution of the slow responding reservoir was gradually increased from lumped (Models A – D) to semi-distributed (Model E) and fully distributed (Model F). In Model E, the slow responding reservoir was divided into four units as visualised in Figure 1A, whereas in Model F it was further discretized into a grid of 10 x 10 km^2, equivalent to the remaining parts of the model. For both alternative formulations, Models E and F respectively, the slow reservoir timescales K_s remained constant throughout the basin to limit the number of calibration parameters. For both Models E and F, groundwater upwelling was included according to Eq.10 (Table 2), hence using Model C as basis, introducing two additional calibration parameters compared to the benchmark Model A (Tables 2 and 3).

3.2 Model performance metrics

3.2.1 Discharge

The model performance with respect to discharge was evaluated using eight distinct signatures simultaneously characterizing the observed discharge data (Euser et al., 2013; Hulsman et al., 2019). The model performance measure was based either on the Nash-Sutcliffe efficiency ($E_{NS,\theta}$, Eq.28 in Table 4) or the relative error ($E_{R,\theta}$, Eq.29) depending on the individual signature. The resulting performance metrics for the eight signatures then included the Nash-Sutcliffe efficiencies of the daily discharge time series ($E_{NS,Q}$), its logarithm ($E_{NS,logQ}$), the flow duration curve ($E_{NS,FDC}$), its logarithm ($E_{NS,logFDC}$) and of the autocorrelation function of daily flows ($E_{NS,AC}$) and the relative errors of the mean seasonal runoff coefficient during dry and wet periods ($E_{R,RCdry}$, $E_{R,RCwet}$) and of the rising limb density of the hydrograph ($E_{R,RLD}$). All these signatures were combined into an overall performance metric based on the Euclidian distance to the "perfect" model ($D_{E,Qcal}$, Eq.31). In absence of more information and to obtain balanced solutions, all individual performance metrics were equally weighted in Eq.31. Here, a $D_{E,Qcal} = 1$ indicates a perfect fit.

The discharge data availability was very limited during the validation time period (2012 – 2016). As a result, hydrological years were not fully captured resulting in incomplete information on the hydrologic signatures such as rising limb density or auto correlation function. That is why the overall model performance ($D_{E,Qval}$) was calculated using the signatures $E_{NS,Q}$, $E_{NS,logQ}$, $E_{NS,FDC}$ and $E_{NS,logFDC}$ excluding $E_{R,RCdry}$, $E_{R,RCwet}$, $E_{R,RLD}$ and $E_{NS,AC}$. It is therefore important to note that $D_{E,Qcal}$ cannot be meaningfully compared with $D_{\mathrm{E,Qval}}$. Instead, following the overall objective of the analysis, $D_{\mathrm{E,Qval}}$ of the different alternative model hypothesis were compared to evaluate the differences between the models.

Table 4: Overview of equations used to calculate model performance

Name	Objective Function	Equation	Variable explanation
Nash-Sutcliffe efficiency	$E_{\mathrm{NS},\theta} = 1 - \frac{\sum_t \left(\theta_{\mathrm{mod}}(t) - \theta_{\mathrm{obs}}(t)\right)^2}{\sum_t \left(\theta_{\mathrm{obs}}(t) - \overline{\theta_{\mathrm{obs}}}\right)^2}$	(28)	θ variable
Relative error	$E_{\mathrm{R},\theta} = 1 - \frac{\lvert\theta_{\mathrm{mod}} - \theta_{\mathrm{obs}}\rvert}{\theta_{\mathrm{obs}}}$	(29)	
Spatial efficiency metric	$E_{\mathrm{SP}} = \frac{1}{\mathrm{t_{max}}} \cdot \sum_t 1 - \sqrt{(\alpha - 1)^2 + (\beta - 1)^2 + (\gamma - 1)^2}$ With: $\alpha = \rho(\varphi_{\mathrm{obs}}, \varphi_{\mathrm{mod}})$ $\beta = \frac{\sigma_{\mathrm{obs}}/\mu_{\mathrm{obs}}}{\sigma_{\mathrm{mod}}/\mu_{\mathrm{mod}}}$ $\gamma = \left(\sum_{i=0}^{i=n} \min(K_i, L_i)\right) \cdot \left(\sum_{i=0}^{i=n} K_{\mathrm{i}}\right)^{-1}$	(30)	α Pearson correlation coefficient φ_{obs}, φ_{mod} observed/modelled map β coefficient of variation σ standard deviation μ mean γ fraction of histogram intersection between K and L K observed histogram L modelled histogram n = 100 bins t time step within the dry season with maximum t_{max}
Euclidian distance over multiple signatures	$D_{\mathrm{E,Q}} = 1 - \sqrt{\frac{1}{(N+M)}\left(\sum_n \left(1 - E_{\mathrm{NS},\theta_{\mathrm{n}}}\right)^2 + \sum_m \left(1 - E_{\mathrm{R},\theta_{\mathrm{m}}}\right)^2\right)}$	(31)	n signatures evaluated with Eq.28 with maximum N m signatures evaluated with Eq.29 with maximum M
Euclidian distance over multiple variables	$D_{\mathrm{E,ESQ}} = 1 - \sqrt{\frac{1}{N}\left(\sum_n (1 - E_{\mathrm{n}})^2\right)}$	(32)	n variables maximum N E_{n} model performance metric of variable n

3.2.2 Evaporation and total water storage

The model performance was also evaluated with respect to both the temporal dynamics and the spatial pattern of evaporation and total storage, respectively. For this purpose, satellite-based evaporation data (WaPOR) was used on 10-day time scale, and total water storage anomaly data (GRACE) on monthly time scale.

Temporal variation

To quantify the models' skill to reproduce the temporal dynamics of evaporation and total water storage anomalies, the respective Nash-Sutcliffe efficiencies (Eq. 28) were used as performance metrics. This performance metric was applied to assess the models' skill to reproduce the basin-average time series of evaporation and total water storage anomalies, i.e. $E_{\mathrm{NS,Basin,E}}$ and $E_{\mathrm{NS,Basin,S}}$, respectively. Similarly, the models' performance to mimic the dynamics of evaporation in all grid cells dominated by the wetland HRU was calculated with the Nash-Sutcliffe efficiency ($E_{\mathrm{NS,Wetland,E}}$). Grid cells were considered as wetland dominated if they were completely covered by wetlands, hence if $p_{\mathrm{HRU}} = 1$ with p_{HRU} the areal weight of wetland areas within that cell. With respect to evaporation, the flux was normalised first with Eq.33 to emphasize temporal variations rather than absolute values in an attempt to reduce bias related errors in the observation:

$$E_{\mathrm{normalised}} = \frac{E - E_{\mathrm{min}}}{E_{\mathrm{max}} - E_{\mathrm{min}}} \qquad (33)$$

Spatial variation

The model performance with respect to the spatial pattern of evaporation and total water storage anomalies was calculated with the spatial efficiency metrics $E_{\mathrm{SP,E}}$ and $E_{\mathrm{SP,S}}$ (Eq.30), respectively, which was successfully used in previous studies (Demirel et al., 2018; Koch et al., 2018). The spatial model performance was first calculated for each time step within the dry period which was in September/October and then averaged to obtain the overall model performance (E_{SP}, Eq.30). The spatial pattern was averaged over the dry season to minimize the effect of precipitation errors.

3.2.3 Multi-variable

The overall potential of the models to simultaneously reproduce the temporal dynamics as well as the spatial patterns of all observed variables, i.e. discharge, evaporation and total water storage anomalies, was tested with the overall model performance metric $D_{\mathrm{E,ESQ}}$. This metric was the Euclidian distance (Eq.32) of the following individual metrics: the temporal variation of the basin-average evaporation ($E_{\mathrm{NS,Basin,E}}$) and total water storage anomalies ($E_{\mathrm{NS,Basin,S}}$), spatial pattern of the evaporation ($E_{\mathrm{SP,E}}$) and total water storage anomalies ($E_{\mathrm{SP,S}}$) as well as discharge ($D_{\mathrm{E,Q}}$). See Table 5 for an overview of all model performance metrics used in this study.

Table 5: Overview of the applied model performance metrics

Data	**Temporal dynamics/ Spatial pattern**	**Performance metric**	**Symbol and equation nr.**	**Calibration/validation**
Discharge	Temporal dynamics	Euclidian distance over multiple signatures (combining $E_{\mathrm{NS,Q}}$, $E_{\mathrm{NS,logQ}}$, $E_{\mathrm{NS,FDC}}$, $E_{\mathrm{NS,logFDC}}$, $E_{\mathrm{NS,AC}}$, $E_{\mathrm{R,RCdry}}$, $E_{\mathrm{R,RCwet}}$ and $E_{\mathrm{R,RLD}}$)	$D_{\mathrm{E,Qcal}}$ (Eq.31)	Calibration (2002 – 2012)
	Temporal dynamics	Euclidian distance over multiple signatures (combining $E_{\mathrm{NS,Q}}$, $E_{\mathrm{NS,logQ}}$, $E_{\mathrm{NS,FDC}}$ and $E_{\mathrm{NS,logFDC}}$)	$D_{\mathrm{E,Qval}}$ (Eq.31)	Validation (2012 – 2016)
Evaporation	Temporal dynamics (basin-average)	Nash-Sutcliffe efficiency	$E_{\mathrm{NS,Basin,E}}$ (Eq.28)	Validation (2012 – 2016)
	Temporal dynamics (wetland areas)	Nash-Sutcliffe efficiency	$E_{\mathrm{NS,Wetland,E}}$ (Eq.28)	Validation (2012 – 2016)
	Spatial pattern	Spatial efficiency metric	$E_{\mathrm{SP,E}}$ (Eq.30)	Validation (2012 – 2016)
Total water storage anomalies	Temporal dynamics (basin-average)	Nash-Sutcliffe efficiency	$E_{\mathrm{NS,Basin,S}}$ (Eq.28)	Validation (2012 – 2016)
	Spatial pattern	Spatial efficiency metric	$E_{\mathrm{SP,S}}$ (Eq.30)	Validation (2012 – 2016)
Multi-variable (discharge, evaporation and total water storage anomalies)	Combination	Euclidian distance over multiple variables (combining $D_{\mathrm{E,Qcal}}$, $E_{\mathrm{NS,Basin,E}}$, $E_{\mathrm{SP,E}}$, $E_{\mathrm{NS,Basin,S}}$ and $E_{\mathrm{SP,S}}$)	$D_{\mathrm{E,ESQcal}}$ (Eq.32)	Calibration (2002 – 2012)
	Combination	Euclidian distance over multiple variables (combining $D_{\mathrm{E,Qval}}$, $E_{\mathrm{NS,Basin,E}}$, $E_{\mathrm{SP,E}}$, $E_{\mathrm{NS,Basin,S}}$ and $E_{\mathrm{SP,S}}$)	$D_{\mathrm{E,ESQval}}$ (Eq.32)	Validation (2012 – 2016)

3.3 Model calibration

In general, the model was calibrated by first running the model with $5 \cdot 10^4$ random parameter sets generated with a Monte-Carlo sampling strategy from uniform prior parameter distributions (Table 3). Then, the optimal and 5% best-performing parameter sets were selected according to the model performance metric as described in the previous section. The model was calibrated within the time period 2002 – 2012 with respect to 1) discharge ($D_{\mathrm{E,Qcal}}$) and 2) all variables simultaneously ($D_{\mathrm{E,ESQcal}}$). As the objective of this study was to explore the information content of multiple variables using multiple model evaluation criteria for step-wise model structure development and calibration, it was important to use the same parameter sets for all models as common starting point to rule out the effect of different parameter sets. This was efficiently possible with the

Monte-Carlo parameter sampling strategy, which, in addition also allowed a relatively straight-forward and intuitive interpretation and communication of the results.

3.4 Model validation

The model was validated with respect to discharge, evaporation and total water storage anomalies for the time period 2012 – 2016. During validation each variable was evaluated separately both temporally and spatially. This included the temporal variation of the basin-average evaporation ($E_{\mathrm{NS,Basin,E}}$) and total water storage anomalies ($E_{\mathrm{NS,Basin,S}}$), evaporation in wetland areas ($E_{\mathrm{NS,Wetland,E}}$), spatial pattern of the evaporation ($E_{\mathrm{SP,E}}$) and total water storage anomalies ($E_{\mathrm{SP,S}}$) as well as discharge ($D_{\mathrm{E,Qval}}$). In addition, the model was evaluated with respect to the overall performance ($D_{\mathrm{E,ESQval}}$). This was done for the solutions from both calibration strategies.

4. Model results

4.1 Benchmark model (Model A)

4.1.1 Discharge based calibration

For the benchmark model (Model A), the model performance of all model realizations following the first calibration strategy, i.e. calibrating to discharge, resulted in an optimum $D_{\mathrm{E,Qcal,opt}}$ = 0.76 and $D_{\mathrm{E,Qval}}$ = 0.37 during validation (Table 6, Figure 3). As shown in Figure 4, the main features of the hydrological response were captured reasonably well. However, particularly in the validation period, low flows were somewhat underestimated. Note that in 2013, the observed high flows were probably underestimated due to failures in the recording which resulted in a truncated top in the hydrograph and flat top in the flow duration curve during the validation time period (Figure 4) and which affect the validated model performance values ($D_{\mathrm{E,Qval}}$). The range in the calibrated model performance with respect to each discharge signature separately is visualised in Figure S1 in the supplementary material.

The basin-average evaporation ($E_{\mathrm{NS,\ Basin,E}}$ = 0.54) and total water storage anomalies ($E_{\mathrm{NS,\ Basin,S}}$ = 0.74) were in general also reproduced rather well (Figures S3 and S5). In contrast, the model failed to mimic the evaporation dynamics in wetland dominated areas as it decreased rapidly to zero in the dry season in contrast to the observations ($E_{\mathrm{NS,Wetland,E}}$ = 0.25, Figure 5). Similarly, the spatial variability in evaporation ($E_{\mathrm{SP,E}}$ = 0.17) and water storage anomalies ($E_{\mathrm{SP,S}}$ = -0.02) were poorly captured as several areas were over- or underestimated (Figures 6 and 7). Note that in both figures the normalised evaporation and total water storage anomalies were plotted applying Eq.33 to emphasize relative spatial differences rather than absolute values.

Table 6: Summary of model performance with respect to evaporation ($E_{\mathrm{NS,Basin,E}}$, $E_{\mathrm{NS,Wetland,E}}$ and $E_{\mathrm{SP,E}}$), total water storage anomalies ($E_{\mathrm{NS,Basin,S}}$ or $E_{\mathrm{SP,S}}$), discharge ($D_{\mathrm{E,Qcal}}$ and $D_{\mathrm{E,Qval}}$) and all variables combined ($D_{\mathrm{E,ESQval}}$): The parameter sets were selected based on discharge ($D_{\mathrm{E,Qcal}}$).

	Calibration (2002 – 2012)	Validation (2012 – 2016)						
	$D_{\mathrm{E,Qcal,opt}}$ ($D_{\mathrm{E,\ Qcal\ ,5/95\%}}$)	$D_{\mathrm{E,Qval}}$ ($D_{\mathrm{E,\ Qval\ ,5/95\%}}$)	$E_{\mathrm{NS,Basin,E}}$ ($E_{\mathrm{NS,Basin,E,5/95}}$)	$E_{\mathrm{NS,wetland,E}}$ ($E_{\mathrm{NS,wetland,E,5/95}}$)	$E_{\mathrm{SP,E}}$ ($E_{\mathrm{SP,E,5/95}}$)	$E_{\mathrm{NS,Basin,S}}$ ($E_{\mathrm{NS,Basin,S,5/95}}$)	$E_{\mathrm{SP,S}}$ ($E_{\mathrm{SP,S,5/95}}$)	$D_{\mathrm{E,ESQval}}$ ($D_{\mathrm{E,ESQval,5/95}}$)
A	0.76 (0.54 – 0.68)	0.37 (0.26 – 0.85)	0.54 (0.34 – 0.57)	0.25 (-0.14 – 0.58)	0.17 (-0.37 – 0.04)	0.74 (0.62 – 0.80)	-0.02 (-0.23 – 0.03)	0.30 (0.12 – 0.29)
B	0.75 (0.36 – 0.60)	0.08 (-3.9 – 0.78)	0.46 (0.34 – 0.63)	0.29 (0.09 – 0.65)	0.12 (-0.68 – 0.12)	0.69 (0.61 – 0.82)	-0.07 (-0.20 – 0.08)	0.21 (-1.3 – 0.27)
C	0.79 (0.58 – 0.70)	0.81 (0.27 – 0.85)	0.50 (0.34 – 0.58)	0.19 (-0.01 – 0.57)	0.10 (-0.39 – 0.06)	0.76 (0.62 – 0.81)	-0.08 (-0.23 – 0.04)	0.32 (0.12 – 0.30)
D	0.77 (0.53 – 0.68)	-1.7 (-2.4 – 0.84)	0.36 (0.33 – 0.60)	0.41 (0.11 – 0.62)	-0.04 (-0.57 – 0.10)	0.63 (0.61 – 0.82)	-0.17 (-0.22 – 0.06)	-0.41 (-0.72 – 0.28)
E	0.78 (0.58 – 0.70)	0.81 (0.27 – 0.85)	0.50 (0.34 – 0.58)	0.07 (-0.04 – 0.59)	0.05 (-0.39 – 0.05)	0.77 (0.62 – 0.81)	-0.08 (-0.23 – 0.04)	0.30 (0.12 – 0.29)
F	0.91 (0.86 – 0.89)	0.52 (0.12 – 0.74)	0.61 (0.45 – 0.63)	0.56 (-0.08 – 0.61)	-0.03 (-0.49 – 0.19)	0.66 (0.44 – 0.71)	0.08 (-0.07 – 0.13)	0.31 (0.12 – 0.34)

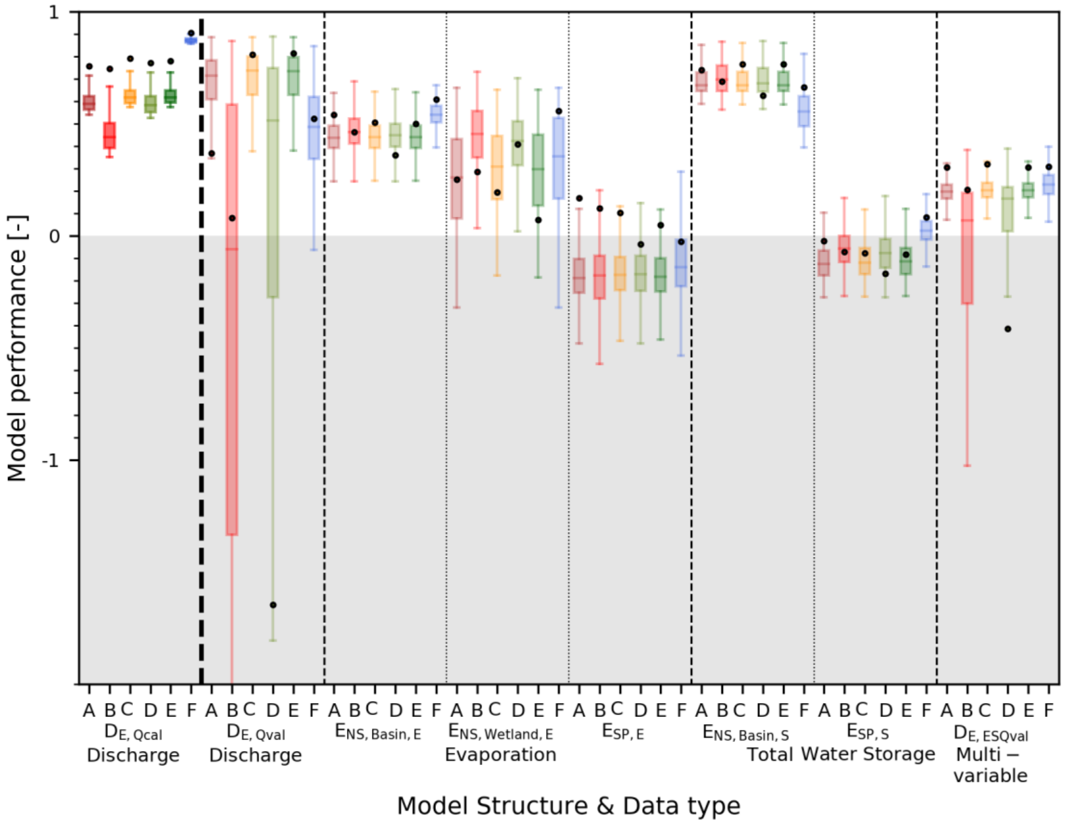

Figure 3: Model performance with respect to discharge, evaporation and storage for all models. The model is calibrated to discharge ($D_{E,Qcal}$, darker boxplots in the first column) and validated to the discharge, evaporation and storage (lighter boxplots). The dots represent the model performance using the "optimal" parameter set and the boxplot the range of the best 5% solutions both according to discharge ($D_{E,Qcal}$). The following performance metrics were used: 1) discharge using the overall model performance metric ($D_{E,Qcal}$ for calibration and $D_{E,Qval}$ for validation), 2) evaporation temporally basin-average ($E_{NS,Basin,E}$), 3) evaporation temporally wetland areas only ($E_{NS,Wetland,E}$), 4) evaporation spatially ($E_{SP,E}$), 5) storage temporally basin-average ($E_{NS,Basin,S}$), 6) storage spatially ($E_{SP,S}$), and 7) the combination of evaporation, storage and discharge (combined metric $D_{E,ESQval}$).

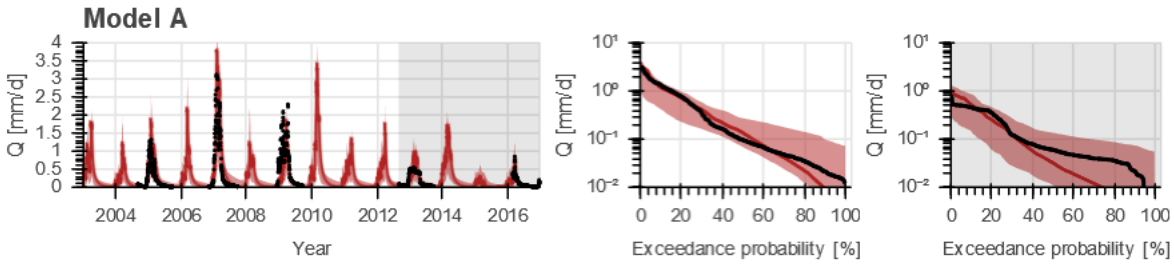

Figure 4: Range of model solutions for Model A. The left panel shows the hydrograph and the right panel the flow duration curve of the recorded (black) and modelled discharge: the line indicates the solution with the highest calibration objective function with respect to discharge ($D_{E,Qcal}$) and the shaded area the envelope of the solutions retained as feasible. The data in the white area were used for calibration and the grey shaded area for validation.

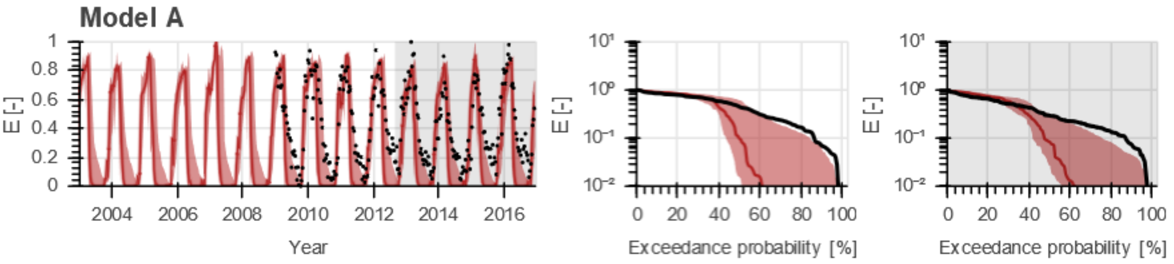

Figure 5: Range of model solutions for Model A. The left panel shows the time series and the right panel the duration curve of the recorded (black) and modelled normalised evaporation for wetland dominated areas: the line indicates the solution with the highest calibration objective function with respect to discharge ($D_{E,Qcal}$) and the shaded area the envelope of the solutions retained as feasible. The data in the grey shaded area were used for validation.

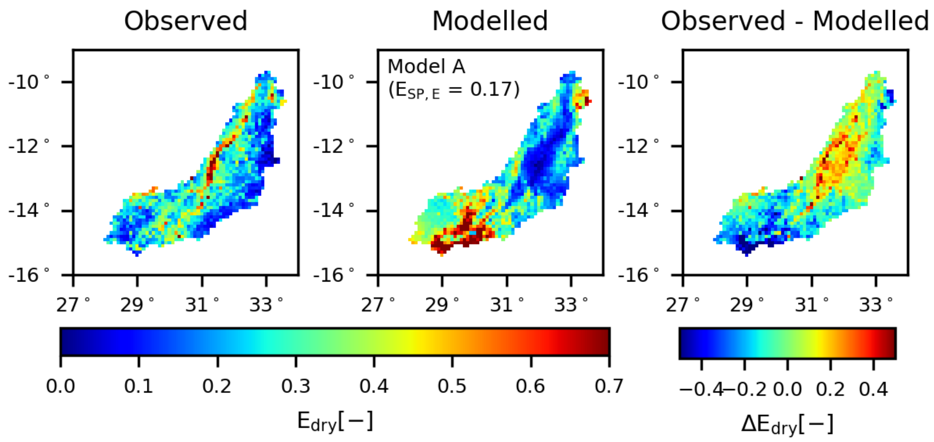

Figure 6: Spatial variability of the normalised total evaporation for Model A averaged over all days within the dry season. The left panel shows the observation according to WaPOR data, the middle panel the model result using the "optimal" parameter set with respect to discharge ($D_{E,Qcal}$), and the right panel the difference between the observation and model.

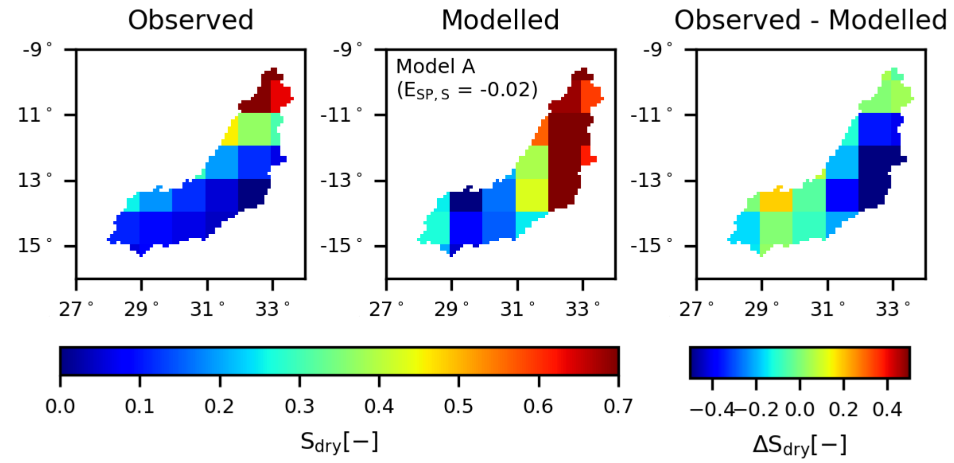

Figure 7: Spatial variability of the normalised total water storage anomalies for Model A averaged over all days within the dry season. The left panel shows the observation according to GRACE data, the middle panel the model result using the "optimal" parameter set with respect to discharge ($D_{E,Qcal}$), and the right panel the difference between the observation and model.

4.1.2 Multi-variable calibration

Calibrating with respect to multiple variables simultaneously in the second calibration strategy, resulted in a reduced model skill to simultaneously reproduce all flow signatures in the validation period with $D_{E,Qval}$ = 0.07 (Table 7, Figures 8 and 9). Compared to the first calibration strategy, the simulated evaporation did not change significantly with respect to the temporal dynamics ($E_{NS,Wetland,E}$ = 0.27, $E_{NS,Basin,E}$ = 0.57) and spatial pattern ($E_{SP,E}$ = -0.18). Evaporation from wetland dominated areas remained underestimated in the dry season (Figure 10) and large areas in the basin were still under- or overestimated (Figure 11). The reproduction of the total water storage anomalies decreased though, mostly with respect to the spatial pattern ($E_{SP,S}$ = -0.14, Figure 12). On the other hand, when looking at the 5/95th percentile range instead of the "optimal" parameter set, then an improvement was observed in the spatial pattern in evaporation ($E_{SP,E,5/95}$ =-0.10 – 0.22) and in total water storage ($E_{SP,S,5/95}$ = -0.17 – 0.08, compare Tables 6 and 7).

Table 7: Summary of the model performance with respect to evaporation ($E_{NS,Basin,E}$, $E_{NS,Wetland,E}$ and $E_{SP,E}$), total water storage anomalies ($E_{NS,Basin,S}$ or $E_{SP,S}$), discharge ($D_{E,Qval}$) and all variables combined ($D_{E,ESQval}$): Parameter sets selected based on multiple variables simultaneously ($D_{E,ESQcal}$).

	Calibration (2002 – 2012)	Validation (2012 – 2016)						
	$D_{E,ESQcal,opt}$ ($D_{E,ESQcal,5/95}$)	$D_{E,Qval}$ ($D_{E, Qval,5/95\%}$)	$E_{NS,Basin,E}$ ($E_{NS,Basin,E,5/95}$)	$E_{NS,wetland,E}$ ($E_{NS,wetland,E,5/95}$)	$E_{SP,E}$ ($E_{SP,E,5/95}$)	$E_{NS,Basin,S}$ ($E_{NS,Basin,S,5/95}$)	$E_{SP,S}$ ($E_{SP,S,5/95}$)	$D_{E,ESQval}$ ($D_{E,ESQval,5/95}$)
A	0.42 (0.28 – 0.36)	0.07 (-1.4 – 0.80)	0.57 (0.37 – 0.60)	0.27 (-0.05 – 0.61)	0.18 (-0.10 – 0.22)	0.72 (0.60 – 0.77)	-0.14 (-0.17 – 0.08)	0.21 (-0.25 – 0.32)
B	0.40 (0.23 – 0.33)	0.46 (-4.2 – 0.70)	0.55 (0.39 – 0.63)	0.56 (0.04 – 0.64)	0.16 (-0.14 – 0.25)	0.73 (0.61 – 0.79)	-0.16 (-0.17 – 0.09)	0.28 (-1.4 – 0.29)
C	0.44 (0.29 – 0.37)	0.61 (-1.6 – 0.79)	0.48 (0.37 – 0.61)	0.51 (0.08 – 0.60)	0.19 (-0.07 – 0.25)	0.70 (0.60 – 0.77)	-0.03 (-0.16 – 0.09)	0.33 (-0.31 – 0.33)
D	0.43 (0.27 – 0.36)	-0.08 (-3.5 – 0.75)	0.51 (0.38 – 0.62)	0.59 (0.06 – 0.61)	0.24 (-0.09 – 0.26)	0.69 (0.60 – 0.78)	-0.04 (-0.16 – 0.09)	0.21 (-1.1 – 0.32)
E	0.43 (0.29 – 0.36)	0.30 (-1.6 – 0.79)	0.43 (0.38 – 0.61)	0.30 (0.03 – 0.61)	0.17 (-0.08 – 0.25)	0.64 (0.60 – 0.77)	-0.02 (-0.16 – 0.10)	0.27 (-0.31 – 0.32)
F	0.52 (0.39 – 0.45)	0.51 (-0.24 – 0.81)	0.56 (0.45 – 0.63)	0.45 (0.01 – 0.63)	0.23 (0.08 – 0.27)	0.63 (0.53 – 0.73)	0.09 (-0.10 – 0.13)	0.37 (0.15 – 0.38)

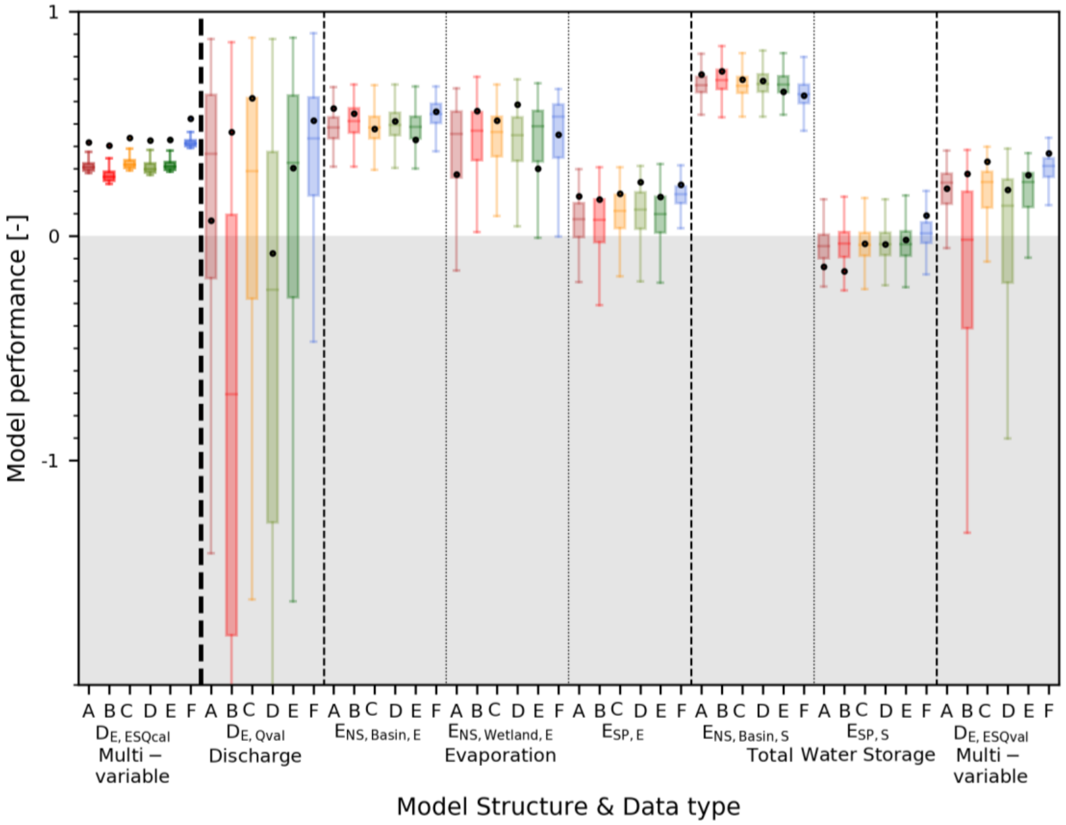

Figure 8: Model performance with respect to discharge, evaporation and storage for all models. The model is calibrated to multiple variables simultaneously ($D_{E,ESQcal}$, darker boxplots in the first column) and evaluated with respect to each flux individually (lighter boxplots). The dots represent the model performance using the "optimal" parameter set and the boxplot the range of the best 5% solutions both according to $D_{E,ESQcal}$. The following performance metrics were used: 1) discharge using the overall model performance metric ($D_{E,Qval}$), 2) evaporation temporally basin-average ($E_{NS,Basin,E}$), 3) evaporation temporally wetland areas only ($E_{NS,Wetland,E}$), 4) evaporation spatially ($E_{SP,E}$), 5) storage temporally basin-average ($E_{NS,Basin,S}$), 6) storage spatially ($E_{SP,S}$), and 7) the combination of evaporation, storage and discharge (combined metric $D_{E,ESQval}$).

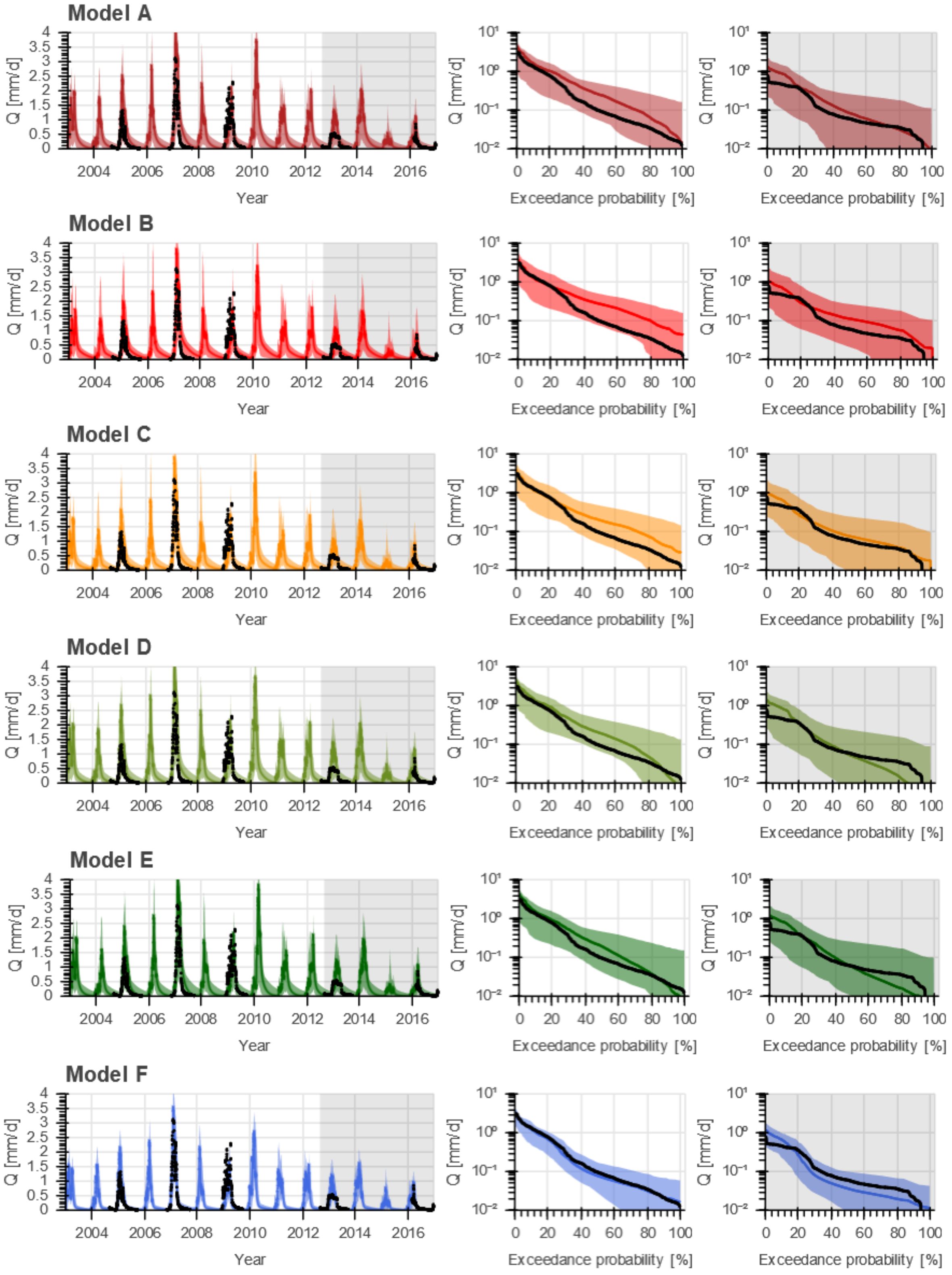

Figure 9: Range of model solutions for Models A to F. The left panel shows the hydrograph and the right panel the flow duration curve of the recorded (black) and modelled discharge: the line indicates the solution with the highest calibration objective function with respect to multiple variables ($D_{E,ESQcal}$) and the shaded area the envelope of the solutions retained as feasible. The data in the grey shaded area were used for validation.

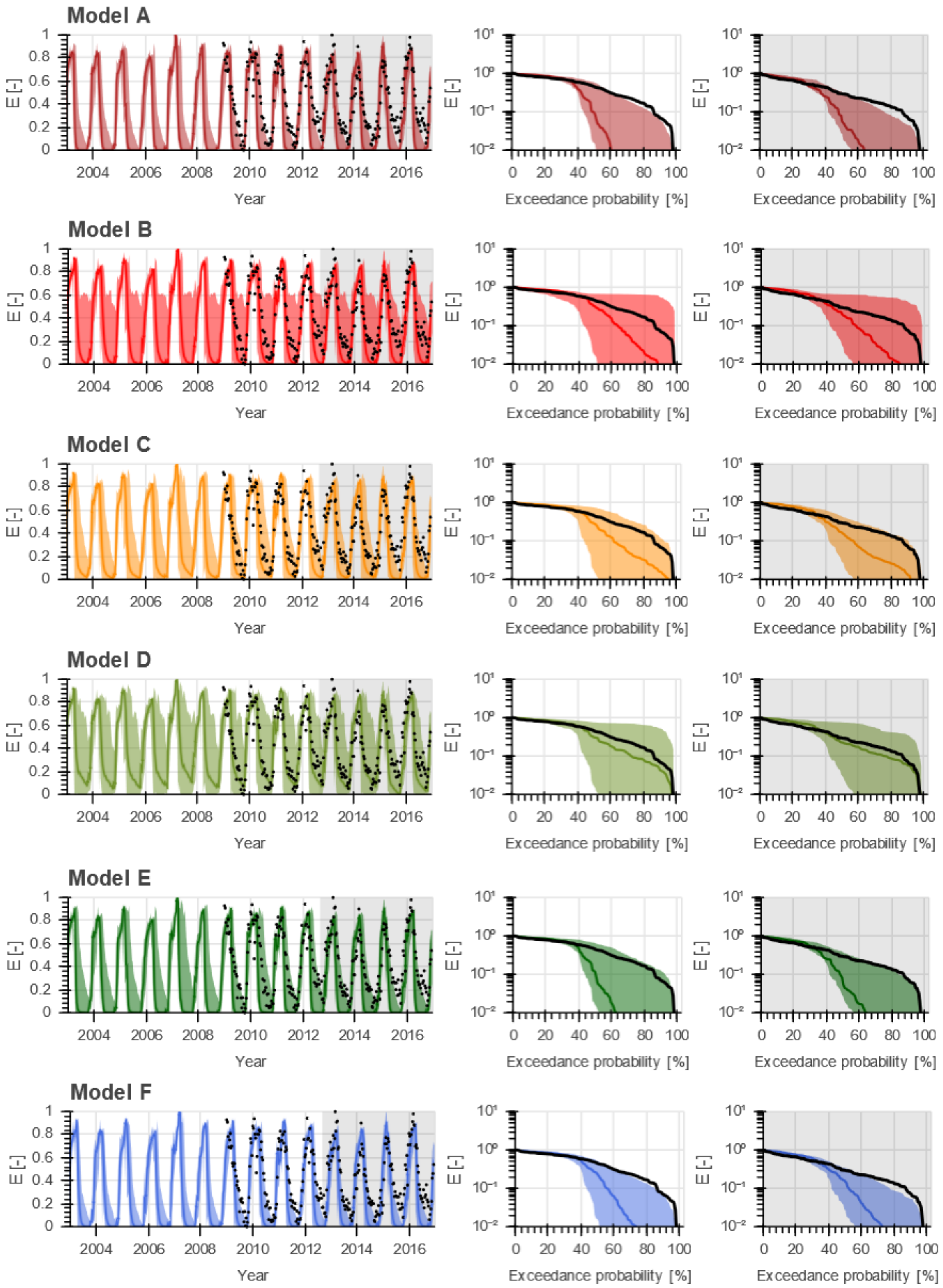

Figure 10: Range of model solutions for Models A to F. The left panel shows the time series and the right panel the duration curve of the recorded (black) and modelled normalised evaporation for wetland dominated areas: the line indicates the solution with the highest calibration objective function with respect to multiple variables ($D_{\mathrm{E,ESQcal}}$) and the shaded area the envelope of the solutions retained as feasible. The data in the grey shaded area were used for validation.

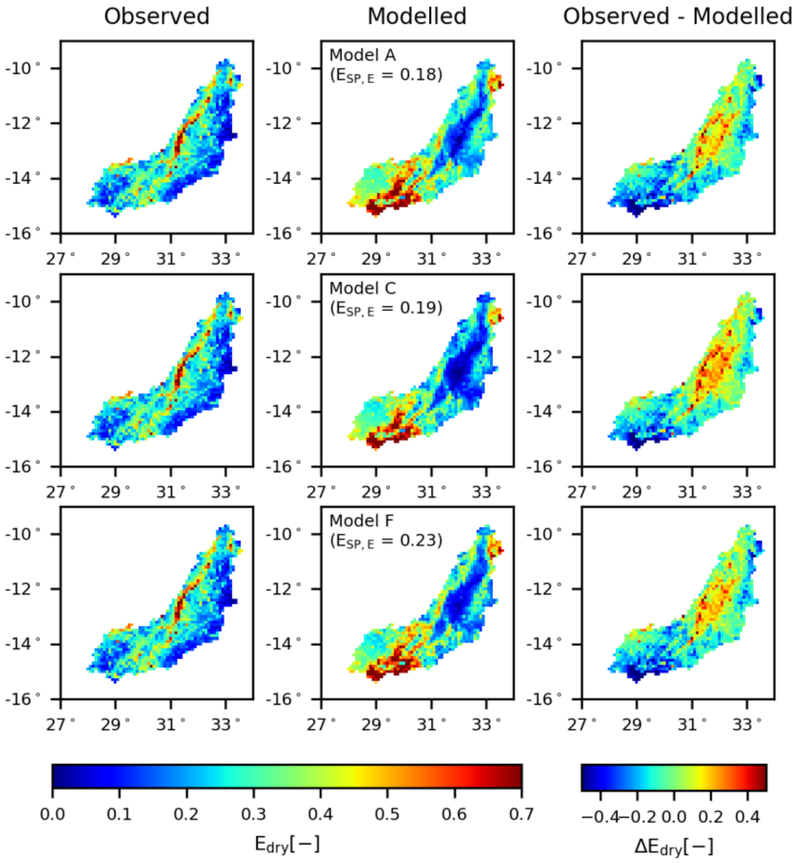

Figure 11: Spatial variability of the normalised total evaporation for Models A, C and F averaged over all days within the dry season. The left panel shows the observation according to WaPOR data, the middle panel the model result using the "optimal" parameter set with respect to multiple variables ($D_{E,ESQcal}$), and the right panel the difference between the observation and model.

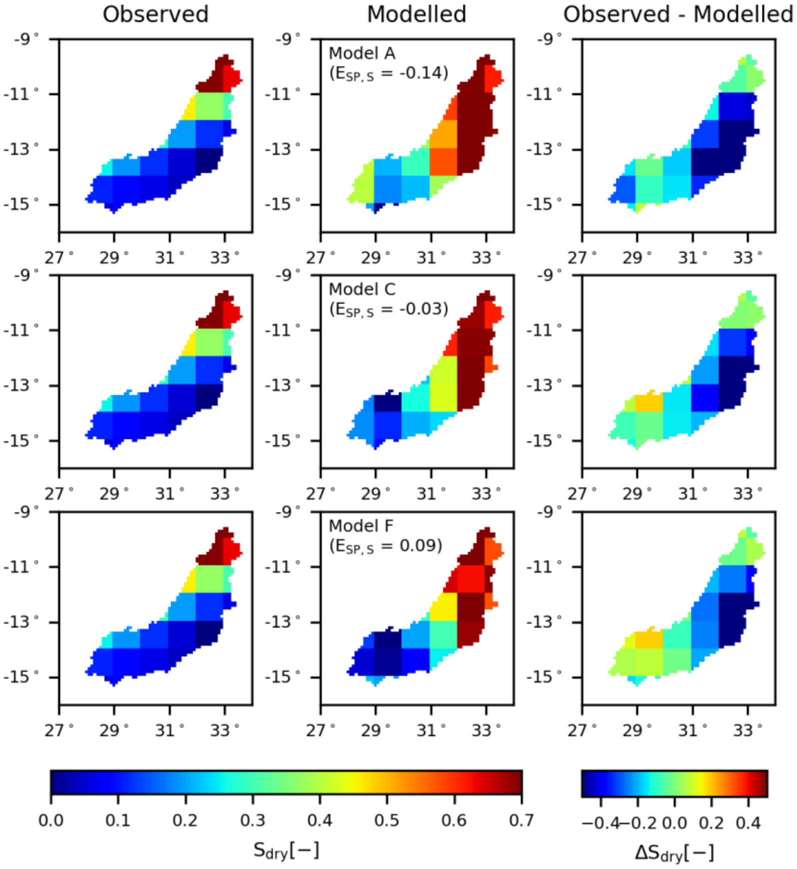

Figure 12: Spatial variability of the normalised total water storage anomalies for Models A, C and F averaged over all days within the dry season. The left panel shows the observation according to GRACE data, the middle panel the model result using the "optimal" parameter set with respect to multiple variables ($D_{E,ESQcal}$), and the right panel the difference between the observation and model.

4.1.3 Model deficiencies

Regardless of the calibration strategy, the benchmark model failed in particular to adequately reproduce evaporation dynamics in wetland dominated areas. During the dry seasons, the modelled evaporation decreased rapidly to zero in contrast to the observations (Figures 5 and 10). Partly as a consequence of that, the spatial pattern of evaporation was captured poorly as illustrated in Figures 6 and 11. Apart from the wetlands, the modelled average dry season evaporation was also extremely low in the centre of the basin which did not correspond with the satellite observations. At the same time, the evaporation was significantly overestimated in the southern part of the basin. Also the spatial pattern in total water storage anomalies were poorly represented since the model significantly overestimated storage anomalies in large parts of the basin (Figures 7 and 12). Note, overestimations in specific regions do not necessarily mean the actual (non-normalised) model values were also higher compared to the observation, but it does mean the model results in this cell/region were high relative to the remainder of the basin compared to observations. This was the case for the evaporation and total water storage even though it was negative during dry seasons (compare Figures S8 and S9 in the supplementary material).

4.2 First model adaptation: Adding groundwater upwelling (Models B, C and D)

In the benchmark model (Model A), there was no groundwater upwelling into the wetlands and floodplains around the river channels, similar to many distributed conceptual hydrological models (eg. Samaniego et al., 2010; Bieger et al., 2017). However, according to field and satellite-based observations, wetland areas remain moist at the end of the dry season while the remaining areas of the basin become very dry. Given the low elevation of these wetlands above rivers, it is plausible to assume that groundwater from higher parts of the catchment is pushed up into the unsaturated root zone of these wetlands. As a result, water deficits in the unsaturated zone are partly replenished by upwelling groundwater. It thereby can sustain relatively elevated levels of moisture, available for plant transpiration long into the dry season.

To improve the representation of evaporation in the model, the process of upwelling groundwater (R_{GW}) was added to the model. In principle, it was assumed that the upwelling groundwater is regulated by the head difference between upland groundwater and the groundwater in the wetland. As this information was not available, due to the lack of continuous gradients in the type of model used (Hrachowitz and Clark, 2017), this was done in a simplified way. In three alternative formulations of this hypothesis, the upwelling groundwater was made (1) a linear function of the water content in the unsaturated reservoir (Model B, Eq.9), (2) a linear function of the water content in the slow responding reservoir (Model C, Eq.10) and (3) a non-linear function of the water content in the slow responding reservoir (Model D, Eq.11). In other words, in Model B the groundwater upwelling was driven by the water deficit in the unsaturated zone, hence the lower the water content in the unsaturated zone, the higher the groundwater upwelling. In Models C and D, the groundwater upwelling was driven by the water content in the slow responding reservoir, the groundwater system, such that the higher the water content in the slow responding reservoir, the higher the groundwater upwelling. As a result of the non-linear relation between the groundwater upwelling and the water content in the slow responding reservoir in Model D, the groundwater upwelling increased the most under dry conditions and less under wet conditions. In Models B – D, the groundwater upwelling flowed into the unsaturated zone until it was saturated, hence until its maximum $S_{u,max}$ was reached (Eq.12). Model B required one additional calibration parameter, Model C two and Model D three (Tables 2 and 3).

4.2.1 Discharge based calibration

Following the first calibration strategy, the performances of Models B – D with respect to discharge did not improve significantly for the calibration period ($D_{E,Qcal}$ = 0.75 – 0.79) compared to Model A, regardless of the model (Table 6, Figures 3 and S2). For the validation period, Models B and D experienced a pronounced reduction of their ability to adequately reproduce the discharge signatures with $D_{E,Qval}$ = 0.08 and -1.7, respectively, since the flows were mostly underestimated (Figure S2). On the other hand, Model C showed significant improvements with $D_{E,Qval}$ = 0.81. With respect to the evaporation from wetland dominated areas, the largest improvements were found for Model D ($E_{\text{NS,Wetland,E}}$ = 0.41) where the evaporation did not drop rapidly to zero anymore even though it was still significantly underestimated in the dry season (Figure S4). But this came at the cost of decreased simulations of all remaining variables (Table 6, Figure 3), hence the discharge, basin-average evaporation and total water storage and their spatial patterns (Figures S2 – S7). For example Figure S6 illustrates the poorly simulated temporally-averaged dry season evaporation for Model D which was higher in wetland areas (centre of the basin) compared to the surrounding areas which was not observed in the satellite based observations. For Models B and C, the model performances with respect to the remaining variables remained comparable to Model A or even decreased as can be seen in Table 6 and Figure 3. As a result, when considering all variables simultaneously, Model C performed the best with $D_{\text{E,ESQval}}$ = 0.32.

4.2.2 Multi-variable calibration

Following the second calibration strategy, Model C experienced the largest increases compared to Model A in its ability to describe features of discharge with $D_{\text{E,Qval}}$ = 0.61, while Model D decreased the most to $D_{\text{E,Qval}}$ = -0.08 with the high flows being overestimated and low flows underestimated (Table 7, Figures 8 and 9). With this calibration strategy, large improvements were observed in the reproduction of the evaporation from wetland dominated areas for all three Models B – D, especially for Model D with $E_{\text{NS,Wetland,E}}$ = 0.59 where the evaporation was simulated well even during the dry season as it did not decrease rapidly to zero in the dry season compared to Model A (Figure 10). For Models C and D, the spatial pattern in evaporation and total water storage anomalies improved, albeit moderately (Table 7) as large areas were still under- or overestimated (Figures S12 and S13), whereas it decreased slightly for Model B. For all Models B – D, the basin-average temporal dynamics in evaporation and total water storage anomalies remained similar or decreased slightly (Table 7, Figures S10 and S11). Overall, when considering the model performance with respect to all variables simultaneously, Model C showed the highest performances with $D_{\text{E,ESQval}}$ = 0.33.

4.2.3 Model deficiencies

According to the results, the representation of evaporation strongly benefitted from including upwelling groundwater as function of the water content in the slow responding reservoir (Eq.10, Model C) especially for the second calibration strategy. The incorporation of this flux resulted in increased levels of water supply to the unsaturated zone of wetlands to sustain higher levels of transpiration throughout the dry periods (Figure 10). But even though the evaporation increased during dry periods, it was still underestimated especially towards the end of the dry season due to too large groundwater upwelling depleting the slow responding reservoir. The major weakness of the model remained its very limited ability to represent the spatial pattern in evaporation as there were several local clusters of considerable mismatches, both over- and underestimating observed evaporation. This was clearly visible for example in the centre and southern part of the basin (Figure 11). Also the spatial pattern in the total water storage anomalies remained poorly represented, in spite of some improvements compared to Model A, as they were considerably overestimated in the northern parts of the basin (Figure 12). This could be a result of deficiencies in the hydrological models or in the satellite-based observations.

4.3 Second model adaptation: Discretizing the groundwater system (Models E and F)

In all above models, the groundwater layer was simulated as a single lumped reservoir assuming equal groundwater availability throughout the entire basin. As groundwater processes can occur on relatively large spatial scales, this assumption may be valid for small- or mesoscale catchments, but not necessarily for larger basins such as the Luangwa basin. This may partly be responsible for the deficiency of all above models to meaningfully reproduce the spatial pattern of the total water storage. Taking Model C as a basis for further model adaptations, two more alternative model hypothesis were formulated. In both models the slow responding reservoir, representing the groundwater, was spatially discretized. For Model E, the reservoir was split into four units with an area of 15,396 – 47,239 km^2 each containing four to six different GRACE cells (see Figure 1A). In contrast, Model F was formulated with a completely distributed slow reservoir at the resolution of the remaining parts of the model, i.e. 10 x 10 km^2. In Models E and F, the slow reservoir timescales K_s remained constant throughout the basin to limit the number of calibration parameters. Models E and F did not require additional calibration parameters. See Tables 2 and 3 for the corresponding model equations and calibration parameter ranges.

4.3.1 Discharge based calibration

Following the first calibration strategy, the calibrated and validated model performance with respect to discharge did not change significantly for Model E compared to Model C. For Model F on the other hand, the calibrated model performance increased to $D_{\mathrm{E,Qcal}}$ = 0.91 (Table 6, Figures 3 and S2), but during validation it decreased to $D_{\mathrm{E,Qval}}$ = 0.52 compared to Model C as a result of overestimated high flows (Figure S2). In other words, the discharge simulation was only affected when applying a fully distributed groundwater system (Model F). Also the simulated dynamics of the evaporation improved for Model F, especially for wetland dominated areas ($E_{\mathrm{NS,Wetland,E}}$ = 0.56, Table 6) even though it remained significantly underestimated during the dry season (Figure S4). But for both models, no improvements in the spatial pattern of evaporation can be observed with $E_{\mathrm{SP,E}}$ = 0.05 and -0.03 for Models E and F, respectively. As shown in Figure S6, for Model E and F the temporally-averaged dry season evaporation was very low in the centre of the basin compared to the remaining part of the basin in contrast to the satellite-based observations. The spatial pattern of total water storage anomalies were at least slightly better mimicked by Model F with $E_{\mathrm{SP,S}}$ = 0.08 (Figure S7), which, in turn, came at the price of a poorer reproduction of the temporal dynamics of the basin-averaged total water storage anomalies ($E_{\mathrm{NS,Basin,S}}$ = 0.66, Figure S5).

4.3.2 Multi-variable calibration

Including multiple variables in the calibration process did not improve the representation of the hydrological response with respect to discharge for Models E and F compared to Model C with $D_{\mathrm{E,Qval}}$ = 0.30 and 0.51, respectively (Table 7, Figures 8 and 9). For both models, the flows were underestimated during low flows and overestimated during high flows (Figure 9). Also the evaporation from wetland dominated areas did not improve for both models as it decreased rapidly in the dry season (Figure 10). On the other hand, the spatial pattern in the evaporation was slightly better mimicked for Model F ($E_{\mathrm{SP,E}}$ = 0.23), but still at low performance levels similar to Models A – D with large areas still being under- or overestimated (Figure S12). Slight improvements could be observed though for the representation of spatial pattern in total water storage in Models F ($E_{\mathrm{SP,S}}$ = 0.09, Figure S13), albeit modestly. Overall, when considering the model performance with respect to all variables simultaneously, Model F showed the highest performances with $D_{\mathrm{E,ESQval}}$ = 0.37.

4.3.3 Model deficiencies

Applying the second calibration strategy, Model F poorly reproduced the evaporation from wetlands (Figure 10) since the water availability for evaporation decreased rapidly in the dry season due to the limited water availability in the slow responding reservoir. This was a direct result of the limited connectivity in the distributed groundwater system within the basin and very likely points to the presence of contiguous groundwater systems extending beyond the modelling resolution that sustain dry season evaporation in wetlands. Strikingly, discretizing the groundwater basin only had limited effects on the spatial pattern in evaporation and total water storage anomalies. Despite their limited improvements, they remained poorly captured as several local clusters were over- and underestimated (Figures 11 and 12).

5. Discussion

As illustrated in the previous sections, satellite-based evaporation and storage anomaly data were used in an attempt to (1) iteratively improve a benchmark model structure and 2) identify parameter sets with which the model can simultaneously reproduce the temporal dynamics as well as the spatial patterns of multiple flux and storage variables.

The results suggested that among the tested models, Models C and F provided the overall best representation of the hydrological processes in the Luangwa basin, following the first and second calibration strategy respectively. The addition of upwelling groundwater alone (Model C) significantly improved the discharge simulations during validation regardless of the calibration strategy and the simulation of evaporation from wetland areas following the second calibration strategy. Discretizing the slow responding reservoir (Model F) reached reasonable overall performance levels, i.e. $D_{\mathrm{E,ESQval}}$, when calibrating on discharge and its signatures only (Figure 3), with improved simulations of evaporation from wetland areas. But calibrating on multiple variables proved instrumental as it allowed to improve the spatial pattern of the evaporation compared to calibrating with respect to discharge (Figures 11 and S12 in the supplementary material), while maintaining high levels for the other performance criteria (Figure 8). In general, it could also be observed that a further discretization of the model lead to a better representation of the system especially with respect to the spatial patterns. Nevertheless, while the model structure and calibration strategy did influence the spatial pattern in the evaporation (Figures S6 and S12 in the supplementary material) and total water storage anomalies (Figures S7 and S13 in the supplementary material), none of the tested models could adequately reproduce the observed spatial pattern which could be a result of model deficiencies or uncertainties in the satellite-based spatial patterns.

A potential reason for the models' problems to meaningfully describe the spatial pattern of the evaporation was in this study the use of the same parameters within a specific HRU in different model grid cells as also observed in previous studies (Stisen et al., 2018). As a result, the simulated spatial pattern was strongly influenced by the catchment classification method into distinct HRUs. In this study, the catchment was classified merely on the basis of topography into flat, sloped and wetland areas, whereas ecosystem diversity could also be considered as an additional layer in the classification. The poor representation of the spatial pattern in total water storage was also partly linked to that. Spatially distributing calibration parameters could improve the modelled spatial pattern assuming there is sufficient data available to meaningfully constrain the increased number of calibration parameters and thus to avoid elevated equifinality. In a preliminary test, the maximum interception storage (I_{max}) was spatially distributed using a linear transfer function with LAI (leaf area index) data similar to previous studies (Samaniego et al., 2010; Kumar et al., 2013) and using Model F as basis. This did not result in obvious improvements as shown in Figure S14 in the supplementary material. It was considered outside the scope of this study to analyse additional parameter distribution strategies with the limited data availability in this study region.

Another likely reason for the poorly modelled spatial pattern is the absence of lateral exchange of sub-surface water between model grid cells in the tested models, as contiguous groundwater bodies of varying but unknown spatial scale will shape water transfer through the landscape in the real world which remain unaccounted for in the model. Lateral exchange fluxes are as any flux driven by continuous gradients and resistances. However, conceptual-type models, such as the one used in this study, only mimic gradients within grid cells, but not between grid cells. As a result, the head difference between neighbouring cells remains unknown which entails that the direction and magnitude of lateral exchange between cells is unknown. Consequently, these fluxes can only be expressed on basis of free calibration parameters. However, in this data-scarce region it will not be possible to test whether the additional calibration parameters and the associated exchange fluxes are physically plausible. These unspecified boundary fluxes across grid cells are at the core of the closure problem (Beven, 2006a) and touch on the limits of what can be done in hydrology with our current observational technology and the available data. Therefore, adding lateral exchange flow to the model was considered outside the scope of this study.

In addition, each of the applied data sources have their own uncertainties and bias. These include uncertainties in observed discharge due to rating curve uncertainties (Westerberg et al., 2011; Domeneghetti et al., 2012; Tomkins, 2014) and limited data availability, in precipitation data, often as a result of poorly capturing mountainous regions or extreme events on small scales (Hrachowitz and Weiler, 2011; Kimani et al., 2017; Dinku et al., 2018; Le Coz and van de Giesen, 2019), in estimates of total water storage anomalies as a result of data (post-) processing including data smoothing using a radius of for example 300 km affecting the spatial variability on basin scale (Landerer and Swenson, 2012; Blazquez et al., 2018) and in evaporation data due to model, input data and parameter estimation uncertainties (Zhang et al., 2016). In general satellite products are a result of models that are prone to uncertainties related to the input data or model conceptualisation. Uncertainties in for example the spatial pattern of the precipitation affect the spatial pattern of the evaporation considerably as shown in Figure S15 in the supplementary material. In the ideal situation, the data would be validated with field measurements to assess the error magnitude. However, this was not possible due to data limitations. To compensate for bias errors in the satellite-based evaporation and to allow more reliable comparisons with model results, the satellite-based evaporation was adjusted with a correction factor of 1.08 (Section 2.1.2). Correcting the precipitation in a similar manner instead of the evaporation did not significantly affect the model results since normalised values were used for model calibration and evaluation (Figure S16 in the supplementary material).

The results in this study were sensitive to the choice of performance metrics with respect to the individual variables (discharge, evaporation and total water storage) and all variables combined. For instance, the overall model performance measure $D_{\mathrm{E,ESQval}}$ (Eq.32) was strongly influenced by the validated discharge model performance $D_{\mathrm{E,Qval}}$ due to its large range and variation between models compared to the remaining variables where the range was smaller and similar for all models (Figure 8). As a result, the overall model performance measure might not reflect each variable equally well which affected the choice of best performing model. However, this did not cause the poorly reproduced spatial pattern in the evaporation as it remained poorly modelled also when calibrating only with respect to that variable ($E_{\mathrm{SP,E}}$, Figure S17 in the supplementary material). In addition, the histogram component (γ) in the Spatial efficiency metric (E_{SP}, Eq.30) becomes less meaningful for very coarse resolutions when the river basin consists of only a few grid cells as was the case for GRACE. It would be interesting to examine the different components in E_{SP} more detailed in future studies to assess the overall suitability of this metric to identify feasible parameter sets across different spatial scales.

Reflecting the results of previous studies, this study found that calibrating to multiple variables including the spatial patterns improved the simulation of the evaporation and storage with some trade-off in the discharge simulation depending on the model structure (Stisen et al., 2011; Rientjes et al., 2013; Demirel et al., 2018; Herman et al., 2018; 2018; Dembélé et al., 2020). But in contrast and additional to previous studies, this study also provided an example, illustrating that spatial data, here evaporation and total water storage, can contain relevant information to diagnose model deficiencies and to therefore enable step-wise model structural improvement. Previous studies have largely relied on discharge observations to improve model structures (Hrachowitz et al., 2014; Fenicia et al., 2016) and only few studies used satellite data (Roy et al., 2017) even though it provides valuable information on the internal processes temporally and spatially which is not available with discharge data alone (Daggupati et al., 2015; Rakovec et al., 2016). Roy et al. (2017) observed that the simulated evaporation according to the spatially lumped model HYMOD (HYdrological MODel) rapidly dropped to zero in contrast to the satellite product GLEAM (Global Land Evaporation Amsterdam Model) in the Nyangores river basin in Kenya. They improved this simulated evaporation while maintaining good discharge performances by modifying the corresponding equation in HYMOD such that it was a function of the soil moisture.

While here we focussed on upwelling groundwater and spatial discretization, a promising avenue for future studies may be to evaluate the incorporation of simple formulations of subsurface exchange fluxes between model grid cells. Similarly, a further discretization of HRUs into different land cover and ecosystem types may be worthwhile. In addition, a systematic sensitivity analysis is recommended to explore the influence of individual factors such as model structure and parameters on the spatial and temporal variability of different variables and to further improve the representation of the hydrological processes.

6. Conclusion

The objective of this paper was to explore the added value of satellite-based evaporation and total water storage anomaly data to increase the understanding of hydrological processes through step-wise model structure improvement and model calibration for large river systems in a semi-arid, data scarce region. For this purpose, a distributed process-based hydrological model with sub-grid process heterogeneity for the Luangwa River basin was developed and iteratively adjusted. The results suggested that (1) the benchmark model (Model A) calibrated with respect to discharge reproduced observed discharge but also basin-average evaporation and total water storage anomalies rather well, while poorly capturing the evaporation for wetland dominated areas as well as the spatial pattern of evaporation and total water storage anomalies. Testing five further alternative model structures (Models B – F), it was found that (2) among the tested model hypotheses Model F, allowing for upwelling groundwater from a distributed representation of the groundwater reservoir and (3) simultaneously calibrating this model with respect to multiple variables, i.e. discharge, evaporation and total water storage anomalies, resulted in marked improvements of the model performance, providing the best simultaneous representation of all these variables with respect to their temporal dynamics and spatial pattern. As a result of the limited data availability and model hypotheses tested in this study, it should be noted that Model F allowed the rejection of alternative hypotheses tested here but may be rejected in future studies in favour of another hypothesis. However, this study illustrated ~~Overall, it was shown~~ that satellite-based evaporation and total water storage anomaly data are not only valuable for multi-criteria calibration, but can play an important role in improving our understanding of hydrological processes through diagnosing model deficiencies and step-wise model structural improvement.

Abbreviations

CHIRPS	Climate Hazards Group InfraRed Precipitation with Station
CMRSET	CSIRO MODIS Reflectance Scaling EvapoTranspiration
CRU	Climatic Research Unit
CSIRO	Commonwealth Scientific and Industrial Research Organisation
FAO	Food and Agriculture Organization
GEOS	Goddard Earth Observing System Model
GMTED	Global Multi-resolution Terrain Elevation Data
GRACE	Gravity Recovery and Climate Experiment
HRU	Hydrological Response Unit
MERRA	Modern-Era Retrospective analysis for Research and Applications
MODIS	Moderate Resolution Imaging Spectroradiometer
NDVI	Normalized Difference Vegetation Index
SSEBop	operational Simplified Surface Energy Balance
WaPOR	Water Productivity Open Access Portal

References

Beven, K.: Searching for the Holy Grail of scientific hydrology: Qt=(S, R, Δt)A as closure, Hydrol. Earth Syst. Sci., 10, 609-618, 10.5194/hess-10-609-2006, 2006a.

Beven, K. J.: A manifesto for the equifinality thesis, Journal of Hydrology, 320, 18-36, https://doi.org/10.1016/j.jhydrol.2005.07.007, 2006b.

Bieger, K., Arnold, J. G., Rathjens, H., White, M. J., Bosch, D. D., Allen, P. M., Volk, M., and Srinivasan, R.: Introduction to SWAT+, A Completely Restructured Version of the Soil and Water Assessment Tool, JAWRA Journal of the American Water Resources Association, 53, 115-130, 10.1111/1752-1688.12482, 2017.

Blazquez, A., Meyssignac, B., Lemoine, J. M., Berthier, E., Ribes, A., and Cazenave, A.: Exploring the uncertainty in GRACE estimates of the mass redistributions at the Earth surface: implications for the global water and sea level budgets, Geophysical Journal International, 215, 415-430, 10.1093/gji/ggy293, 2018.

Blöschl, G., and Sivapalan, M.: Scale issues in hydrological modelling: A review, Hydrological Processes, 9, 251-290, 10.1002/hyp.3360090305, 1995.

Bouaziz, L. J. E., Weerts, A., Schellekens, J., Sprokkereef, E., Stam, J., Savenije, H., and Hrachowitz, M.: Redressing the balance: quantifying net intercatchment groundwater flows, Hydrol. Earth Syst. Sci., 22, 6415-6434, https://doi.org/10.5194/hess-22-6415-2018, 2018.

Clark, M. P., Rupp, D. E., Woods, R. A., Zheng, X., Ibbitt, R. P., Slater, A. G., Schmidt, J., and Uddstrom, M. J.: Hydrological data assimilation with the ensemble Kalman filter: Use of streamflow observations to update states in a distributed hydrological model, Advances in Water Resources, 31, 1309-1324, https://doi.org/10.1016/j.advwatres.2008.06.005, 2008.

Clark, M. P., Kavetski, D., and Fenicia, F.: Pursuing the method of multiple working hypotheses for hydrological modeling, Water Resources Research, 47, W09301, https://doi.org/10.1029/2010WR009827, 2011.

Daggupati, P., Yen, H., White, M. J., Srinivasan, R., Arnold, J. G., Keitzer, C. S., and Sowa, S. P.: Impact of model development, calibration and validation decisions on hydrological simulations in West Lake Erie Basin, Hydrological Processes, 29, 5307-5320, 10.1002/hyp.10536, 2015.

Danielson, J. J., and Gesch, D. B.: Global multi-resolution terrain elevation data 2010 (GMTED2010), in: Open-File Report 2011-1073, U.S. Geological Survey, Reston, Virginia, 2011.

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 Resources Research, 56, e2019WR026085, 10.1029/2019WR026085, 2020.

Demirel, M., Mai, J., Mendiguren González, 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, 2018.

Dinku, T., Funk, C., Peterson, P., Maidment, R., Tadesse, T., Gadain, H., and Ceccato, P.: Validation of the CHIRPS satellite rainfall estimates over eastern Africa, Quarterly Journal of the Royal Meteorological Society, 144, 292-312, 10.1002/qj.3244, 2018.

Domeneghetti, A., Castellarin, A., and Brath, A.: Assessing rating-curve uncertainty and its effects on hydraulic model calibration, Hydrology and Earth System Sciences, 16, 1191-1202, 10.5194/hess-16-1191-2012, 2012.

Euser, T., Winsemius, H. C., Hrachowitz, M., Fenicia, F., Uhlenbrook, S., and Savenije, H. H. G.: A framework to assess the realism of model structures using hydrological signatures, Hydrology and Earth System Sciences, 17, 1893-1912, https://doi.org/10.5194/hess-17-1893-2013, 2013.

Euser, T., Hrachowitz, M., Winsemius, H. C., and Savenije, H. H. G.: The effect of forcing and landscape distribution on performance and consistency of model structures, Hydrological Processes, 29, 3727-3743, https://doi.org/10.1002/hyp.10445, 2015.

FAO: WaPOR Database Methodology: Level 1. Remote Sensing for Water Productivity Technical Report: Methodology Series, in, FAO, Rome, 72, 2018.

FAO and IHE Delft: WaPOR quality assessment. Technical report on the data quality of the WaPOR FAO database version 1.0, in, FAO and IHE Delft, Rome, 134, 2019.

Fenicia, F., McDonnell, J. J., and Savenije, H. H. G.: Learning from model improvement: On the contribution of complementary data to process understanding, Water Resources Research, 44, 10.1029/2007WR006386, 2008.

Fenicia, F., Kavetski, D., and Savenije, H. H. G.: Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development, Water Resources Research, 47, 10.1029/2010WR010174, 2011.

Fenicia, F., Kavetski, D., Savenije, H. H. G., and Pfister, L.: From spatially variable streamflow to distributed hydrological models: Analysis of key modeling decisions, Water Resources Research, 52, 954-989, 10.1002/2015WR017398, 2016.

Funk, C. C., Peterson, P. J., Landsfeld, M. F., Pedreros, D. H., Verdin, J. P., Rowland, J. D., Romero, B. E., Husak, G. J., Michaelsen, J. C., and Verdin, A. P.: A quasi-global precipitation time series for drought monitoring, in: Data Series 832, U.S. Geological Survey, South Dakota, 4, 2014.

Gao, H., Hrachowitz, M., Fenicia, F., Gharari, S., and Savenije, H. H. G.: Testing the realism of a topography-driven model (FLEX-Topo) in the nested catchments of the Upper Heihe, China, Hydrol. Earth Syst. Sci., 18, 1895-1915, https://doi.org/10.5194/hess-18-1895-2014, 2014.

Gao, H., Hrachowitz, M., Sriwongsitanon, N., Fenicia, F., Gharari, S., and Savenije, H. H. G.: Accounting for the influence of vegetation and landscape improves model transferability in a tropical savannah region, Water Resources Research, 52, 7999-8022, 10.1002/2016WR019574, 2016.

Garavaglia, F., Le Lay, M., Gottardi, F., Garçon, R., Gailhard, J., Paquet, E., and Mathevet, T.: Impact of model structure on flow simulation and hydrological realism: from a lumped to a semi-distributed approach, Hydrol. Earth Syst. Sci., 21, 3937-3952, 10.5194/hess-21-3937-2017, 2017.

Getirana, A. C. V.: Integrating spatial altimetry data into the automatic calibration of hydrological models, Journal of Hydrology, 387, 244-255, http://dx.doi.org/10.1016/j.jhydrol.2010.04.013, 2010.

Gharari, S., Hrachowitz, M., Fenicia, F., Gao, H., and Savenije, H. H. G.: Using expert knowledge to increase realism in environmental system models can dramatically reduce the need for calibration, Hydrol. Earth Syst. Sci., 18, 4839-4859, https://doi.org/10.5194/hess-18-4839-2014, 2014.

Gupta, H. V., Wagener, T., and Liu, Y.: Reconciling theory with observations: elements of a diagnostic approach to model evaluation, Hydrological Processes, 22, 3802-3813, 10.1002/hyp.6989, 2008.

Hargreaves, G. H., and Samani, Z. A.: Reference Crop Evapotranspiration from Temperature, Applied Engineering in Agriculture, 1, 96-99, https://doi.org/10.13031/2013.26773, 1985.

Hargreaves, G. H., and Allen, R. G.: History and evaluation of hargreaves evapotranspiration equation, Journal of Irrigation and Drainage Engineering, 129, 53-63, https://doi.org/10.1061/(ASCE)0733-9437(2003)129:1(53), 2003.

Herman, M. R., Nejadhashemi, A. P., Abouali, M., Hernandez-Suarez, J. S., Daneshvar, F., Zhang, Z., Anderson, M. C., Sadeghi, A. M., Hain, C. R., and Sharifi, A.: Evaluating the role of evapotranspiration remote sensing data in improving hydrological modeling predictability, Journal of Hydrology, 556, 39-49, https://doi.org/10.1016/j.jhydrol.2017.11.009, 2018.

Hrachowitz, M., and Weiler, M.: Uncertainty of Precipitation Estimates Caused by Sparse Gauging Networks in a Small, Mountainous Watershed, Journal of Hydrologic Engineering, 16, 460-471, 10.1061/(ASCE)HE.1943-5584.0000331, 2011.

Hrachowitz, M., Fovet, O., Ruiz, L., Euser, T., Gharari, S., Nijzink, R., Freer, J., Savenije, H. H. G., and Gascuel-Odoux, C.: Process consistency in models: The importance of system signatures, expert knowledge, and process complexity, Water Resources Research, 50, 7445-7469, https://doi.org/10.1002/2014WR015484, 2014.

Hrachowitz, M., and Clark, M. P.: HESS Opinions: The complementary merits of competing modelling philosophies in hydrology, Hydrol. Earth Syst. Sci., 21, 3953-3973, 10.5194/hess-21-3953-2017, 2017.

Hulsman, P., Winsemius, H. C., Michailovsky, C., Savenije, H. H. G., and Hrachowitz, M.: Using altimetry observations combined with GRACE to select parameter sets of a hydrological model in data scarce regions, Hydrol. Earth Syst. Sci. Discuss., 2019, 1-35, 10.5194/hess-2019-346, 2019.

Immerzeel, W. W., and Droogers, P.: Calibration of a distributed hydrological model based on satellite evapotranspiration, Journal of Hydrology, 349, 411-424, http://dx.doi.org/10.1016/j.jhydrol.2007.11.017, 2008.

Jiang, D., and Wang, K.: The role of satellite-based remote sensing in improving simulated streamflow: A review, Water (Switzerland), 11, 10.3390/w11081615, 2019.

Kavetski, D., and Fenicia, F.: Elements of a flexible approach for conceptual hydrological modeling: 2. Application and experimental insights, Water Resources Research, 47, 10.1029/2011WR010748, 2011.

Kimani, W. M., Hoedjes, C. B. J., and Su, Z.: An Assessment of Satellite-Derived Rainfall Products Relative to Ground Observations over East Africa, Remote Sensing, 9, 10.3390/rs9050430, 2017.

Kirchner, J. W.: Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology, Water Resources Research, 42, 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, 10.5194/hess-22-1453-2018, 2018.

Koch, J., Siemann, A., Stisen, S., and Sheffield, J.: Spatial validation of large-scale land surface models against monthly land surface temperature patterns using innovative performance metrics, Journal of Geophysical Research: Atmospheres, 121, 5430-5452, 10.1002/2015JD024482, 2016.

Koch, J., Demirel, M. C., and Stisen, S.: The SPAtial EFficiency metric (SPAEF): Multiple-component evaluation of spatial patterns for optimization of hydrological models, Geoscientific Model Development, 11, 1873-1886, 10.5194/gmd-11-1873-2018, 2018.

Kumar, R., Samaniego, L., and Attinger, S.: Implications of distributed hydrologic model parameterization on water fluxes at multiple scales and locations, Water Resources Research, 49, 360-379, 10.1029/2012WR012195, 2013.

Landerer, F. W., and Swenson, S. C.: Accuracy of scaled GRACE terrestrial water storage estimates, Water Resources Research, 48, W04531, https://doi.org/10.1029/2011WR011453, 2012.

Le Coz, C., and van de Giesen, N.: Comparison of rainfall products over sub-Sahara Africa, Journal of Hydrometeorology, 21, 553-596, https://doi.org/10.1175/JHM-D-18-0256.1, 2019.

Li, Z., Yang, D., Gao, B., Jiao, Y., Hong, Y., and Xu, T.: Multiscale hydrologic applications of the latest satellite precipitation products in the Yangtze river basin using a distributed hydrologic model, Journal of Hydrometeorology, 16, 407-426, 10.1175/JHM-D-14-0105.1, 2015.

López, P. L., 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, Hydrology and Earth System Sciences, 21, 3125-3144, 10.5194/hess-21-3125-2017, 2017.

Mazzoleni, M., Brandimarte, L., and Amaranto, A.: Evaluating precipitation datasets for large-scale distributed hydrological modelling, Journal of Hydrology, 578, 124076, https://doi.org/10.1016/j.jhydrol.2019.124076, 2019.

Mendiguren, G., Koch, J., and Stisen, S.: Spatial pattern evaluation of a calibrated national hydrological model – a remote sensing based diagnostic approach, Hydrol. Earth Syst. Sci. Discuss., 2017, 1-28, 10.5194/hess-2017-233, 2017.

Michailovsky, C. I., Milzow, C., and Bauer-Gottwein, P.: Assimilation of radar altimetry to a routing model of the Brahmaputra River, Water Resources Research, 49, 4807-4816, 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, 10.5194/hess-15-1729-2011, 2011.

Nijzink, R. C., Samaniego, L., Mai, J., Kumar, R., Thober, S., Zink, M., Schäfer, D., Savenije, H. H. G., and Hrachowitz, M.: The importance of topography-controlled sub-grid process heterogeneity and semi-quantitative prior constraints in distributed hydrological models, Hydrol. Earth Syst. Sci., 20, 1151-1176, https://doi.org/10.5194/hess-20-1151-2016, 2016.

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 Resources Research, 54, 8332-8362, 10.1029/2017WR021895, 2018.

Odusanya, A. E., Mehdi, B., Schürz, C., Oke, A. O., Awokola, O. S., Awomeso, J. A., Adejuwon, J. O., and Schulz, K.: Multi-site calibration and validation of SWAT with satellite-based evapotranspiration in a data-sparse catchment in southwestern Nigeria, Hydrology and Earth System Sciences, 23, 1113-1144, 10.5194/hess-23-1113-2019, 2019.

Rajib, A., Evenson, G. R., Golden, H. E., and Lane, C. R.: Hydrologic model predictability improves with spatially explicit calibration using remotely sensed evapotranspiration and biophysical parameters, Journal of Hydrology, 567, 668-683, 10.1016/j.jhydrol.2018.10.024, 2018.

Rakovec, O., Kumar, R., Attinger, S., and Samaniego, L.: Improving the realism of hydrologic model functioning through multivariate parameter estimation, Water Resources Research, 52, 7779-7792, 10.1002/2016WR019430, 2016.

Rennó, C. D., Nobre, A. D., Cuartas, L. A., Soares, J. V., Hodnett, M. G., Tomasella, J., and Waterloo, M. J.: HAND, a new terrain descriptor using SRTM-DEM: Mapping terra-firme rainforest environments in Amazonia, Remote Sensing of Environment, 112, 3469-3481, https://doi.org/10.1016/j.rse.2008.03.018, 2008.

Revilla-Romero, B., Beck, H. E., Burek, P., Salamon, P., de Roo, A., and Thielen, J.: Filling the gaps: Calibrating a rainfall-runoff model using satellite-derived surface water extent, Remote Sensing of Environment, 171, 118-131, http://dx.doi.org/10.1016/j.rse.2015.10.022, 2015.

Rientjes, T. H. M., Muthuwatta, L. P., Bos, M. G., Booij, M. J., and Bhatti, H. A.: Multi-variable calibration of a semi-distributed hydrological model using streamflow data and satellite-based evapotranspiration, Journal of Hydrology, 505, 276-290, https://doi.org/10.1016/j.jhydrol.2013.10.006, 2013.

Roy, T., Gupta, H. V., Serrat-Capdevila, A., and Valdes, J. B.: Using satellite-based evapotranspiration estimates to improve the structure of a simple conceptual rainfall–runoff model, Hydrol. Earth Syst. Sci., 21, 879-896, 10.5194/hess-21-879-2017, 2017.

SADC: Integrated Water Resources Management Strategy and Implementation Plan for the Zambezi River Basin, Euroconsult Mott MacDonald, 2008.

Samaniego, L., Kumar, R., and Attinger, S.: Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale, Water Resources Research, 46, 10.1029/2008WR007327, 2010.

Savenije, H. H. G.: Topography driven conceptual modelling (FLEX-Topo), Hydrol. Earth Syst. Sci., 14, 2681-2692, https://doi.org/10.5194/hess-14-2681-2010, 2010.

Schleiss, A. J., and Matos, J. P.: Chapter 98: Zambezi River Basin, in: Chow's Handbook of Applied Hydrology, edited by: Singh, V. P., McGraw-Hill Education - Europe, United States, 2016.

Schumann, G., Kirschbaum, D., Anderson, E., and Rashid, K.: Role of Earth Observation Data in Disaster Response and Recovery: From Science to Capacity Building, in: Earth Science Satellite Applications edited by: Hossain, F., Springer International Publishing, Seattle, USA, 2016.

Seyler, F., Muller, F., Cochonneau, G., Guimarães, L., and Guyot, J. L.: Watershed delineation for the Amazon sub-basin system using GTOPO30 DEM and a drainage network extracted from JERS SAR images, Hydrological Processes, 23, 3173-3185, 10.1002/hyp.7397, 2009.

Stisen, S., McCabe, M. F., Refsgaard, J. C., Lerer, S., and Butts, M. B.: Model parameter analysis using remotely sensed pattern information in a multi-constraint framework, Journal of Hydrology, 409, 337-349, http://dx.doi.org/10.1016/j.jhydrol.2011.08.030, 2011.

Stisen, S., Koch, J., Sonnenborg, T. O., Refsgaard, J. C., Bircher, S., Ringgaard, R., and Jensen, K. H.: Moving beyond run-off calibration—Multivariable optimization of a surface–subsurface–atmosphere model, Hydrological Processes, 32, 2654-2668, 10.1002/hyp.13177, 2018.

Sun, W., Song, H., Cheng, T., and Yu, J.: Calibration of hydrological models using TOPEX/Poseidon radar altimetry observations, Proc. IAHS, 368, 3-8, 10.5194/piahs-368-3-2015, 2015.

Sun, W., Fan, J., Wang, G., Ishidaira, H., Bastola, S., Yu, J., Fu, Y. H., Kiem, A. S., Zuo, D., and Xu, Z.: Calibrating a hydrological model in a regional river of the Qinghai–Tibet plateau using river water width determined from high spatial resolution satellite images, Remote Sensing of Environment, 214, 100-114, https://doi.org/10.1016/j.rse.2018.05.020, 2018.

Swenson, S. C., and Wahr, J.: Post-processing removal of correlated errors in GRACE data, Geophys. Res. Lett., 33, L08402, https://doi.org/10.1029/2005GL025285, 2006.

Swenson, S. C.: GRACE monthly land water mass grids NETCDF RELEASE 5.0, in, PO.DAAC, CA, USA, 2012.

Tang, X., Zhang, J., Gao, C., Ruben, G. B., and Wang, G.: Assessing the uncertainties of four precipitation products for SWAT modeling in Mekong River Basin, Remote Sensing, 11, 10.3390/rs11030304, 2019.

The World Bank: The Zambezi River Basin: A Multi-Sector Investment Opportunities Analysis, in: Volume 3 State of the Basin, The International Bank for Reconstruction and Development, The World Bank, Washington DC, 2010.

Tomkins, K. M.: Uncertainty in streamflow rating curves: Methods, controls and consequences, Hydrological Processes, 28, 464-481, 10.1002/hyp.9567, 2014.

University of East Anglia Climatic Research Unit, Harris, I. C., and Jones, P. D.: CRU TS4.01: Climatic Research Unit (CRU) Time-Series (TS) version 4.01 of high-resolution gridded data of month-by-month variation in climate (Jan. 1901- Dec. 2016), in, Centre for Environmental Data Analysis,, 2017.

van Dijk, A. I. J. M., and Renzullo, L. J.: Water resource monitoring systems and the role of satellite observations, Hydrol. Earth Syst. Sci., 15, 39-55, 10.5194/hess-15-39-2011, 2011.

Vervoort, R. W., Miechels, S. F., van Ogtrop, F. F., and Guillaume, J. H. A.: Remotely sensed evapotranspiration to calibrate a lumped conceptual model: Pitfalls and opportunities, Journal of Hydrology, 519, Part D, 3223-3236, http://dx.doi.org/10.1016/j.jhydrol.2014.10.034, 2014.

Weerasinghe, I., van Griensven, A., Bastiaanssen, W., Mul, M., and Jia, L.: Can we trust remote sensing ET products over Africa?, Hydrol. Earth Syst. Sci. Discuss., 2019, 1-27, 10.5194/hess-2019-233, 2019.

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

Westerberg, I., Guerrero, J. L., Seibert, J., Beven, K. J., and Halldin, S.: Stage-discharge uncertainty derived with a non-stationary rating curve in the Choluteca River, Honduras, Hydrological Processes, 25, 603-613, 10.1002/hyp.7848, 2011.

Winsemius, H. C., Savenije, H. H. G., and Bastiaanssen, W. G. M.: Constraining model parameters on remotely sensed evaporation: justification for distribution in ungauged basins?, Hydrol. Earth Syst. Sci., 12, 1403-1413, 10.5194/hess-12-1403-2008, 2008.

Xu, X., Li, J., and Tolson, B. A.: Progress in integrating remote sensing data and hydrologic modeling, Progress in Physical Geography, 38, 464-498, 10.1177/0309133314536583, 2014.

ZAMCOM, SADC, and SARDC: Zambezi Environment Outlook 2015, Harare, Gaborone, 2015.

Zhang, K., Kimball, J. S., and Running, S. W.: A review of remote sensing based actual evapotranspiration estimation, Wiley Interdisciplinary Reviews: Water, 3, 834-853, 10.1002/wat2.1168, 2016.

Zink, M., Mai, J., Cuntz, M., and Samaniego, L.: Conditioning a Hydrologic Model Using Patterns of Remotely Sensed Land Surface Temperature, Water Resources Research, 54, 2976-2998, 10.1002/2017WR021346, 2018.