Why are long-term storage variations observed but not modelled in the Luangwa basin?

Petra Hulsman[1], Hubert H. G. Savenije[2], and Markus Hrachowitz[1]

[1]Delft University of Technology
[2]TU-Delft

November 24, 2022

Abstract

In the Luangwa basin, long-term total water storage variations were observed with GRACE, but not reproduced by a standard conceptual hydrological model that encapsulates our current understanding of the dominant regional hydrological processes. The objective of this paper was to identify potential processes underlying these low-frequency variations through combined data analysis and model hypothesis testing. First, we analysed the effect of data uncertainty by contrasting observed storage variations with multi-annual estimates of precipitation and evaporation from multiple data sources. Second, we analysed four different combinations of model forcing and evaluated their skill to reproduce the observed long-term storage variations. Third, we formulated alternative model hypotheses for groundwater export to potentially explain low-frequency storage variations. Overall, the results suggest that the initial model's inability to reproduce the observed low-frequency storage variations was partly due to the forcing data used and partly due to the missing representation of regional groundwater export. More specifically, the choice of data source affected the model's ability to reproduce annual maximum storage fluctuations, whereas the annual minima improved by adapting the model structure to allow for groundwater export from a deeper groundwater layer. This suggests that, in contrast to previous research, conceptual models can reproduce long-term storage fluctuations if a suitable model structure is used. Overall, the results highlight the value of alternative data sources and iterative testing of model structural hypotheses to improve runoff predictions in a poorly gauged basin leading to enhanced understanding of its hydrological processes.

Why are long-term storage variations observed but not modelled in the Luangwa basin?

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

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

Corresponding author: Petra Hulsman (p.hulsman@tudelft.nl)

Key Points:

- Observed long-term total water storage changes are not reproduced by a standard hydrological model in the Luangwa Basin
- Fluctuations in the annual maxima were improved by using alternative forcing data in the hydrological model
- Fluctuations in the annual minima were improved by introducing groundwater export from a deep groundwater layer

Abstract

In the Luangwa basin, long-term total water storage variations were observed with GRACE, but not reproduced by a standard conceptual hydrological model that encapsulates our current understanding of the dominant regional hydrological processes. The objective of this paper was to identify potential processes underlying these low-frequency variations through combined data analysis and model hypothesis testing. First, we analysed the effect of data uncertainty by contrasting observed storage variations with multi-annual estimates of precipitation and evaporation from multiple data sources. Second, we analysed four different combinations of model forcing and evaluated their skill to reproduce the observed long-term storage variations. Third, we formulated alternative model hypotheses for groundwater export to potentially explain low-frequency storage variations. Overall, the results suggest that the initial model's inability to reproduce the observed low-frequency storage variations was partly due to the forcing data used and partly due to the missing representation of regional groundwater export. More specifically, the choice of data source affected the model's ability to reproduce annual maximum storage fluctuations, whereas the annual minima improved by adapting the model structure to allow for groundwater export from a deeper groundwater layer. This suggests that, in contrast to previous research, conceptual models can reproduce long-term storage fluctuations if a suitable model structure is used. Overall, the results highlight the value of alternative data sources and iterative testing of model structural hypotheses to improve runoff predictions in a poorly gauged basin leading to enhanced understanding of its hydrological processes.

Plain Language Summary

According to satellite observations, the total amount of water stored on and below the land surface varied over the years in the Zambian Luangwa river basin. However, this variation was not well reproduced by existing rainfall-runoff models, resulting in inaccurate predictions of runoff and water availability. The goal of this study was to identify processes causing long-term fluctuations in the total water storage by using alternative data sources and by adjusting the model structure. First, we analysed whether similar long-term fluctuations existed in the climate using different satellite products. Second, we tested whether these fluctuations could be better represented using different data sources. Third, we tested whether they could be caused by inter-basin groundwater flow. We indeed showed that long-term storage fluctuations were better represented by alternative data sources and by incorporating groundwater loss from the basin, leading to more reliable runoff predictions in the poorly gauged Luangwa basin.

1 Introduction

Long-term and thus low-frequency total water storage variations have been observed in many regions world-wide (Long et al., 2017;Scanlon et al., 2018). This includes long-term storage variations in Australia during the Millennium Drought in 1997 – 2010 (e.g. Leblanc et al., 2009;Chen et al., 2016;Zhao et al., 2017a), in the United States (Long et al., 2013;Boutt, 2017), in the La Plata basin in South America (Chen et al., 2010), in China (Zhang et al., 2015b;Sun et al., 2018) and in different African river basins (Awange et al., 2016;Werth et al., 2017;Bonsor et al., 2018).

However, many hydrological models cannot reproduce these observed long-term storage variations (Winsemius et al., 2006;Scanlon et al., 2018;Fowler et al., 2020). As highlighted by previous studies, these observed long-term storage variations can be a result of climate variability, land-cover change, other human interventions or any combination thereof, while the inability of models to reproduce these variations can be a result of model structural deficiencies, poor parameterization, data errors, poor parameter values or any combination thereof (Saft et al., 2016;Fowler et al., 2018;Grigg and Hughes, 2018;Jing et al., 2019). For example, Bouaziz et al. (2020) showed that although a suite of different conceptual models could similarly well reproduce stream flow over almost two decades, they considerably varied in their skill to reproduce observed storage variations, which was attributed to deficiencies of different model architectures. With some exceptions (e.g. Perrin et al., 2003;Goswami et al., 2007;Le Moine et al., 2007;Samaniego et al., 2011;Hrachowitz et al., 2014;Bouaziz et al., 2018), processes that could potentially allow long-term memory effects, such as groundwater export, remain mostly unaccounted for in standard conceptual rainfall-runoff models (Burnash et al., 1973;Bergström, 1992;Liang et al., 1994;Fenicia et al., 2014;Willems, 2014;Euser et al., 2015). This leads to the situation that these models cannot capture long and slow processes dominating long-term storage variations, as convincingly demonstrated by Fowler et al. (2020). Their study, which focused on the Millennium Drought in Australia, illustrated that modelled annual minimum storage remained rather constant instead of showing a decreasing trend. The reason for this was that the modelled storage converged to or even reached zero towards the end of each dry season and hence could not decrease any further. Such an omission of processes that allow to account for long-term memory processes in rainfall-runoff models results in biased modelled discharge and impedes accurate estimations of water availability which is particularly crucial during extreme dry conditions (Saft et al., 2016).

In many river basins, detecting long-term storage variations and identifying their drivers is challenged by limited high-quality ground observations. That is why in this context satellite observations may play an important role. For example, satellite-based Gravity Recovery and Climate Experiment (GRACE) observations describe variations in the Earths' gravity field which can be used to detect regional mass changes that are dominated by variations in the

terrestrial water storage after removing atmospheric effects. In other words, GRACE observations, which are available on monthly timescale, provide valuable information on total water storage changes (Landerer and Swenson, 2012;Swenson, 2012). For example, GRACE observations have been used in the context of groundwater monitoring (Tangdamrongsub et al., 2018;Zhang et al., 2020), or drought analysis (Leblanc et al., 2009;van Dijk et al., 2013;Zhang et al., 2015a;Chao et al., 2016;Zhao et al., 2017b).

While several previous studies focused on identifying long-term storage variations in (satellite-based) observations, possible drivers for these variations, and differences between observations and model results (e.g. Leblanc et al., 2009;Joodaki et al., 2014;Scanlon et al., 2018;Jing et al., 2019;Meng et al., 2019;Fowler et al., 2020), only limited studies attempted to modify a hydrological model to allow for long-term storage variations. In one exception, Grigg and Hughes (2018) modified the GR4J rainfall-runoff model (Perrin et al., 2003) successfully to mimic long-term catchment memory effects. This was done by introducing a threshold in the storage reservoir such that percolation from this reservoir stopped when the storage was lower than the threshold while evaporation losses continued. Other studies improved the modelled long-term storage trends by assimilated total water storage observations according to GRACE into hydrological models (Khaki et al., 2018;Schumacher et al., 2018).

In this study, long-term storage variations were observed in the Luangwa river basin, but not reproduced by a standard implementation of a conceptual model. The objective of this paper was to identify potential and so far overlooked processes underlying these low-frequency variations in a combined data analysis and model hypothesis testing approach. More specifically, we here tested the hypotheses that the degree to which a conceptual hydrological model can reproduce observed long-term, low-frequency water storage variations depends (1) on the choice of the forcing data source used as input to the model and (2) on the incorporation of processes allowing long-term memory effects in the model.

2 Site description

The Luangwa River is a 770 km long tributary of the Zambezi in Zambia which is mostly unregulated (see Figure 1). Its 159,000 km^2 large basin area is poorly gauged and mostly covered with deciduous forests, shrubs and savanna. The elevation varies up to 1850 m between the low-lying areas around the river and the highlands. In this semi-arid area, there is a distinct wet season from October to April with heavy rains up to 100 mm $month^{-1}$. Nevertheless, the mean annual potential evaporation (1555 mm yr^{-1}) exceeds the mean annual precipitation (970 mm yr^{-1}) (The World Bank, 2010;Hulsman et al., 2020b).

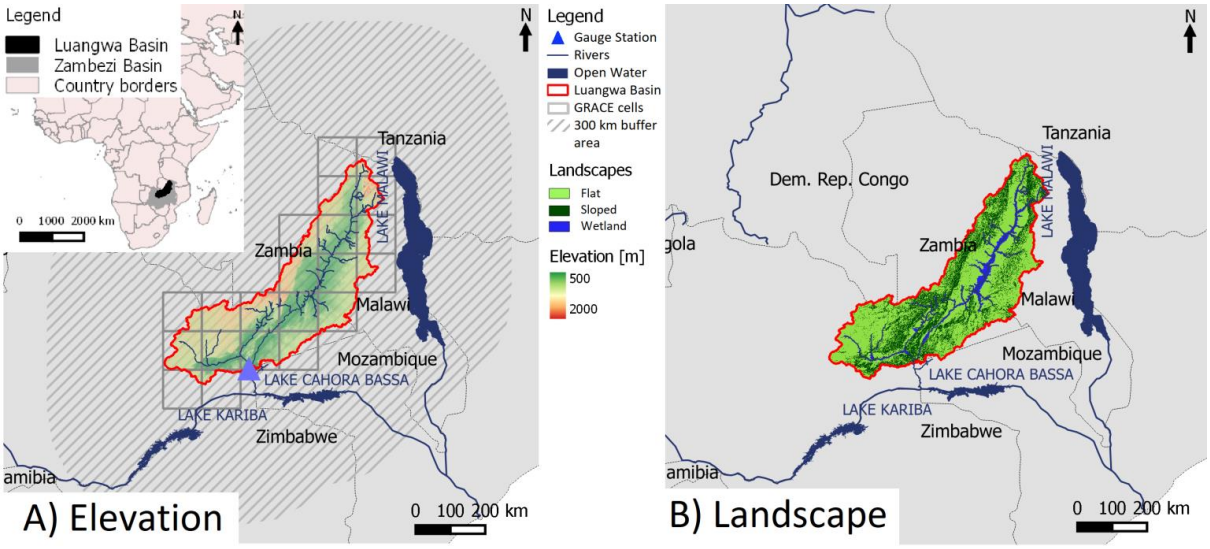


Figure 1. Map of the Luangwa River Basin in Zambia with a) the elevation, and b) the main landscape types

3 Data availability

In this study, hydro-meteorological data as shown in Table 1 were used. This included two satellite-based precipitation products (CHIRPS and TRMM) and five actual evaporation products (WaPOR, SEBS, SSEBop, GLEAM and MOD16). Land-cover changes were assessed using the NDVI (Normalized Difference Vegetation Index) and LAI (Leaf Area Index). Temperature data according to CRU (Climatic Research Unit) was used to estimate the potential evaporation with the Hargreaves (Hargreaves and Samani, 1985;Hargreaves and Allen, 2003) and Thornthwaite (Maes et al., 2019) method.

Processed GRACE (Gravity Recovery and Climate Experiment) observations generated by CSR (Centre for Space Research), GFZ (GeoForschungsZentrum Potsdam) and JPL (Jet Propulsion Laboratory) were obtained from the GRACE Tellus website (https://grace.jpl.nasa.gov/). This study used the average of these three sources which previously processed the raw data to remove atmospheric mass changes, systematic errors and noise, and to subtract the 2004 – 2009 time-mean baseline (Wahr et al., 1998;Swenson and Wahr, 2006;Landerer and Swenson, 2012). As a result, total water storage *anomalies* were available in equivalent water thickness. Total water storage anomaly observations include all terrestrial water storage components, hence water stored in the surface water, soil moisture and groundwater.

Altimetry data was extracted from the DAHITI website (https://dahiti.dgfi.tum.de/en/) for the Cahora Bassa reservoir, Kariba reservoir and Lake Malawi (Schwatke et al., 2015). In-situ

discharge data was used for the Great East Road Bridge gauging station at the basin outlet (30° 13' E and 14° 58' S) and was obtained from the Zambian Water Resources Management Authority (WARMA) for the time period 2002 to 2016 with a temporal coverage of 18%.

For the following data analysis, gridded observations were averaged for the entire basin, whereas for use in the distributed hydrological model, gridded observations were rescaled to the model resolution of 0.25° by (a) taking the mean of all cells located within a model cell if the resolution was smaller, or (b) dividing each cell into multiple cells if the resolution was larger. For the hydrological model, gridded observations were used for the topography to classify the landscape into hydrological response units (see Section 4.2.1), climate (precipitation and temperature) to force the model, and total water storage anomalies to calibrate/evaluate the model.

Table 1. Data used in this study

	Time period	Time resolution	Spatial Resolution	Product Name	Long-term annual mean	Source/Reference
Digital elevation map	n/a	n/a	0.02°	GMTED	n/a	GMTED2010 (Danielson and Gesch, 2011)
Precipitation	1998 – 2016	Daily	0.05°	CHIRPS	1127 mm yr^{-1}	Version 2 (Funk et al., 2014)
	1998 – 2016	Daily	0.25°	TRMM	1029 mm yr^{-1}	Version 3B42 (Huffman et al., 1995;Huffman et al., 2007;Huffman et al., 2014)
Evaporation	2009 – 2016	10 days	250 m	WaPOR	882 mm yr^{-1}	Version 1.1 (FAO, 2018;FAO and IHE Delft, 2019)
	2002 – 2013	Monthly	0.05°	SEBS	657 mm yr^{-1}	(Su, 2002)
	2003 – 2016	Monthly	0.01°	SSEBop	837 mm yr^{-1}	Version 4 (Bastiaanssen et al., 1998;Allen et al., 2007;Senay et al., 2007)
	2003 – 2016	Monthly	0.25°	GLEAM	751 mm yr^{-1}	Version 3.3b (Miralles et al., 2011;Martens et al., 2017)
	2002 – 2016	8 days	500 m	MOD16	793 mm yr^{-1}	MOD16A2 Version 6 (Running et al., 2017)
NDVI	2002 – 2016	8 days	30 m	NA	0.12	Derived from Landsat 7
LAI	2002 – 2016	Monthly	0.05°	NA	1.48	Version 5 (Claverie et al., 2014)
Temperature	2002 – 2016	Monthly	0.5°	CRU	22°	Time-series (TS) data version 4.01 (University of East Anglia Climatic Research Unit et al., 2017)
Total water Storage	2002 – 2016	Monthly	1°	GRACE	8.8 mm	Pre-processed by CSR & GFZ (Version RL05.DSTvSCS1409), and JPL (Version RL05_1.DSTvSCS1411) https://grace.jpl.nasa.gov/ (Swenson and Wahr, 2006;Landerer and Swenson, 2012;Swenson, 2012)
Altimetry	2002 – 2016	10 or 35 days	n/a	DAHITI	n/a	https://dahiti.dgfi.tum.de/en/ (Schwatke et al., 2015)
Discharge	2002 – 2016	Daily	n/a	n/a	138 mm yr^{-1}	WARMA

4 Approach

This study consisted of three steps. In the first step we analysed the effect of the choice of the data source used to explain observed total water storage variations to understand whether any of the data contain, in principle, sufficient information to at least broadly reflect the dynamics of storage variations. This was necessary to rule out that the model's inability to reproduce long-term storage variations is merely an artefact of unsuitable data. Thus, we investigated whether periods of high water storage anomalies roughly coincide with periods of high precipitation anomalies and/or low evaporation anomalies and vice versa. To do so, we contrasted long-term estimates of variables such as precipitation, potential and actual evaporation from multiple data sources with the observed water storage variations. This allowed a preliminary assessment of which data sources are more consistent with the observed low-frequency storage variations than others. Based on that, we then analysed, in a second step, four different combinations of data sources, i.e. precipitation and potential

evaporation, as input for a hydrological model and evaluated their respective effects to reproduce the observed long-term storage variations with the model. In a third step, we then iteratively formulated and tested several alternative model hypotheses, incorporating a model component, such as regional groundwater export, to account for long-memory effects.

In general, long-term total water storage variations are a result of changes in precipitation, evaporation, discharge or any combination thereof (Eq.1). While climate variability can cause long-term variations in precipitation and atmospheric water demand (i.e. potential evaporation), land-cover changes can affect the partitioning between evaporative fluxes and streamflow (Gallart and Llorens, 2003;Oguntunde et al., 2006;Warburton et al., 2012;Nijzink et al., 2016;Saft et al., 2016;Li et al., 2017;Hrachowitz et al., 2020). In addition, long-term storage variations can be a result of slow inter-basin groundwater exchange (Nelson and Mayo, 2014;Pellicer-Martínez and Martínez-Paz, 2014;Bouaziz et al., 2018).

$$\frac{dS}{dt} = P - E - Q \tag{1}$$

Where S is total water storage, P precipitation, E evaporation and Q discharge.

4.1 Data analysis

Long-term, basin-averaged satellite observations of the precipitation according to CHIRPS and TRMM, actual evaporation according to WaPOR, SEBS, SSEBop, GLEAM and MOD16, potential evaporation according to the Hargreaves (Hargreaves and Samani, 1985;Hargreaves and Allen, 2003) and Thornthwaite (Maes et al., 2019) methods, respectively, and land-cover based on the NDVI and LAI (Table 1) were contrasted with and compared to the water storage variations estimated by GRACE. For each of these data sources, the temporal variability was visualised on monthly and/or annual timescale.

To assess the potential role of regional groundwater import to or export from the basin, the long-term water balance was estimated using the average annual precipitation, evaporation and discharge from the different satellite products. Assuming negligible long-term storage changes and data uncertainties, surpluses or deficits in the long-term water balance, hence if $\bar{P} - \bar{E} - \bar{Q} \neq 0$, are then the result of groundwater import/export. In case of groundwater export, the average annual leaking flow can then be estimated according to (e.g. Bouaziz et al., 2018):

$$\overline{Q_L} = \bar{P} - \bar{E} - \bar{Q} \tag{2}$$

Where $\overline{Q_L}$ is annual mean groundwater export [mm yr^{-1}], $\bar{P}$ annual mean precipitation [mm yr^{-1}], $\bar{E}$ annual mean evaporation [mm yr^{-1}] and $\bar{Q}$ annual mean discharge [mm yr^{-1}].

4.2 Hydrological models

4.2.1 Benchmark model (Model A0)

The process-based distributed hydrological model used in this study for the Luangwa basin was step-wise developed and refined in previous studies (Hulsman et al., 2020a;2020b) following the FLEX-Topo modelling concept (Savenije, 2010). Each 0.25° x 0.25° model cell had the same model structure and parameter set, but was forced differently using spatially distributed forcing data with respect to the precipitation and potential evaporation (e.g. Euser et al., 2015). In addition, each cell was further discretized into functionally distinct landscape classes, i.e. hydrological response units (HRUs) based on the topography (Nijzink et al., 2016). All HRUs within a cell were connected through a common groundwater component (Figure 2a). This groundwater reservoir was also lumped over the entire basin assuming a homogeneous groundwater system (Hulsman et al., 2020a). The landscape was classified based on the local slope and "Height-above-the-nearest-drainage" (HAND; Rennó et al., 2008) into sloped areas (slope $\geq$ 4%), flat areas (slope < 4%, HAND $\geq$ 11 m) and wetland areas (slope < 4%, HAND < 11 m). As a result, 68% of the basin was classified as flat areas, 28% as sloped areas and 8% as wetlands (Figure 1b). This FLEX-Topo modelling concept was applied successfully in previous studies (Gao et al., 2014;Gharari et al., 2014;Hulsman et al., 2020b).

As illustrated in Figure 2a, the hydrological model consisted of multiple storage components representing the interception storage, unsaturated root-zone storage, as well as fast and slow responding storages. Each storage component was schematized as reservoir with corresponding water balance and constitutive equations as shown in Table 3. As the dominant processes and thus the associated model structures of the three individual HRUs were very similar to each other, the major differences between the HRUs were accounted for by different parameter values. Model process constraints were applied as shown in Table 4 to allow partly overlapping prior parameter distributions with relationships consistent with our physical understanding of the system (Gharari et al., 2014;Hrachowitz et al., 2014), and to limit equifinality (Beven, 2006). For example in the Luangwa basin, higher interception evaporation and larger root-zone storage capacities were expected in the densely vegetated, forest dominated sloped areas compared to the flat, grass- and shrub-land dominated areas and wetlands. Processes unique to a HRU were incorporated by adjusting the model structure where necessary. In sloped and flat areas for example, the groundwater system was recharged by downward infiltration whereas in wetlands this flow was assumed to be negligible due to shallow groundwater tables. Rather, water was assumed to be pushed upwards from the groundwater system into the unsaturated root-zone due to the groundwater head difference between the upland and wetland (Hulsman et al., 2020a).

After having calculated the runoff for each grid cell, the total flow at the outlet was estimated by applying a simple routing scheme based on the flow distance to the outlet and a constant, calibrated flow velocity. This model consisted of 17 calibration parameters with uniform prior distributions and constraints as summarized in Table 4. In this benchmark model, the precipitation product CHIRPS was used and potential evaporation was calculated with the Hargreaves method (see Table 2).

4.2.2 First model adaptation: Alternative forcing data (Models B0 – D0)

As first model adaptation, the forcing data was changed to assess the role of data uncertainty for the model's ability to reproduce the observed long-term storage variations and to test whether some combinations of data sources allow model results to be more consistent with the observed storage variations than others. Starting with Model A0 as benchmark, different combinations of precipitation products, i.e. CHIRPS and TRMM, on the one hand and methods to estimate potential evaporation, i.e. Hargreaves and Thornthwaite, on the other hand were tested in Models B0 – D0 (Table 2).

Table 2. Overview of model combinations

	Precipitation product	**Potential evaporation method**
Model A0	CHIRPS	Hargreaves
Model B0	CHIRPS	Thornthwaite
Model C0	TRMM	Hargreaves
Model D0	TRMM	Thornthwaite

4.2.3 Second model adaptation: Alternative model structure (Model A1– A5)

As second model adaptation, the model structure was changed to test whether deep groundwater flow or inter-basin groundwater export/import was a relevant driver for the observed long-term storage variations. In this study, a distinction was made between shallow groundwater flow (Q_{ss}), deep groundwater flow (Q_{sd}) and groundwater loss (Q_L). While the shallow and deep groundwater flow reached the river, the groundwater loss (Q_L) leaked out of the Luangwa basin and potentially reached the Zambezi river further downstream. Based on benchmark Model A0, hence using CHIRPS for precipitation and the Hargreaves method to estimate potential evaporation, the model structure was modified to introduce long-term storage memory.

With Model A1, it was tested whether only groundwater export, hence groundwater leaking out of the Luangwa basin, was a dominant driver for the long-term storage variations. In this model, groundwater loss (Q_L) was introduced (Figures 2b and 3) which did not reach the river (Eq.36) and, in the spirit of model parsimony, was assumed to be constant, regardless of the water content in the Upper Groundwater reservoir to limit the number of calibration

parameters as no additional information was available. Thus, the Upper Groundwater reservoir (S_{su}) was formulated as a deficit store that can become negative. However, the shallow groundwater flow Q_{ss} only occurred when this storage was positive (if $S_{su} > 0$, Eq.27). Such a formulation allowed groundwater to keep on draining, and thus groundwater levels falling, even if discharge in the river ceased during dry periods (e.g. Hrachowitz et al., 2014;Bouaziz et al., 2018).

With Model A2, it was tested whether constant groundwater export from a second, Deeper Groundwater reservoir can explain the observed long-term storage variations. In this model, groundwater seeped from the Upper Groundwater reservoir into a Deeper Groundwater reservoir as fraction of the water content in the Upper Groundwater reservoir (R_s, Eq.29, Figures 2c and 3). From this Deeper Groundwater reservoir, constant groundwater loss (Q_L) leaked out of the basin similar to Model A1.

With Model A3, it was tested whether constant groundwater export from the Deeper Groundwater reservoir recharged only during wet seasons, was the main driver for long-term storage variations. In this model, groundwater only seeped into the Deeper Groundwater reservoir when the groundwater level in the Upper Groundwater reservoir exceeded a reference level ($S_{s,ref2}$, Eq.30, Figures 2d and 3). From there constant groundwater loss (Q_L) leaked out of the basin similar to Models A1 and A2.

With Model A4, it was tested whether *variable* groundwater export from the Deeper Groundwater reservoir recharged only during wet seasons, was the main driver for long-term storage variations. In this model, the groundwater loss (Q_L, Figures 2e and 3) was a function of the water content in the Deeper Groundwater reservoir (Eq.34). This groundwater loss (Q_L) did not reach the river similar to Models A1 – A3.

With Model A5, it was tested whether variable groundwater *flow* from the Deeper Groundwater reservoir recharged only during wet seasons, was the main driver for long-term storage variations. In this model, the groundwater drained from the Deeper reservoir into the river as Q_{sd} contributing to the total river flow (Eq.38, Figures 2f and 3). Hence, only in Model A5 deep groundwater reached the gauged river system whereas in Models A1 – A4 groundwater leaked out of the basin.

Figure 3 gives an overview of all alternative model hypotheses tested in this study. The relevant model equations are given in Table 3 and the corresponding prior parameter distributions in Table 4.

4.2.4 Third model adaptation: Alternative forcing data and model structure

As third model adaptation, the forcing and the model structure were changed simultaneously. For this purpose, the best performing model based on the results of the first model adaptation, i.e. changing the forcing data (Models A0 – D0) and the second model adaptation, i.e. changing the model structure (Models A0 – A5) were combined. For example, if Models D0

and A4 performed best, respectively, then the combined Model D4 using the forcing data applied in Model D0 and the model structure of Model A4 was tested. To ensure a robust representation of both, discharge and total water storage, the above model selection was based on the combined performance metrics for both variables. We explicitly acknowledge the possibility of this not being the combination that most reliably reflects real world processes. However, exhaustively testing all possible combinations goes beyond our computational capacity.

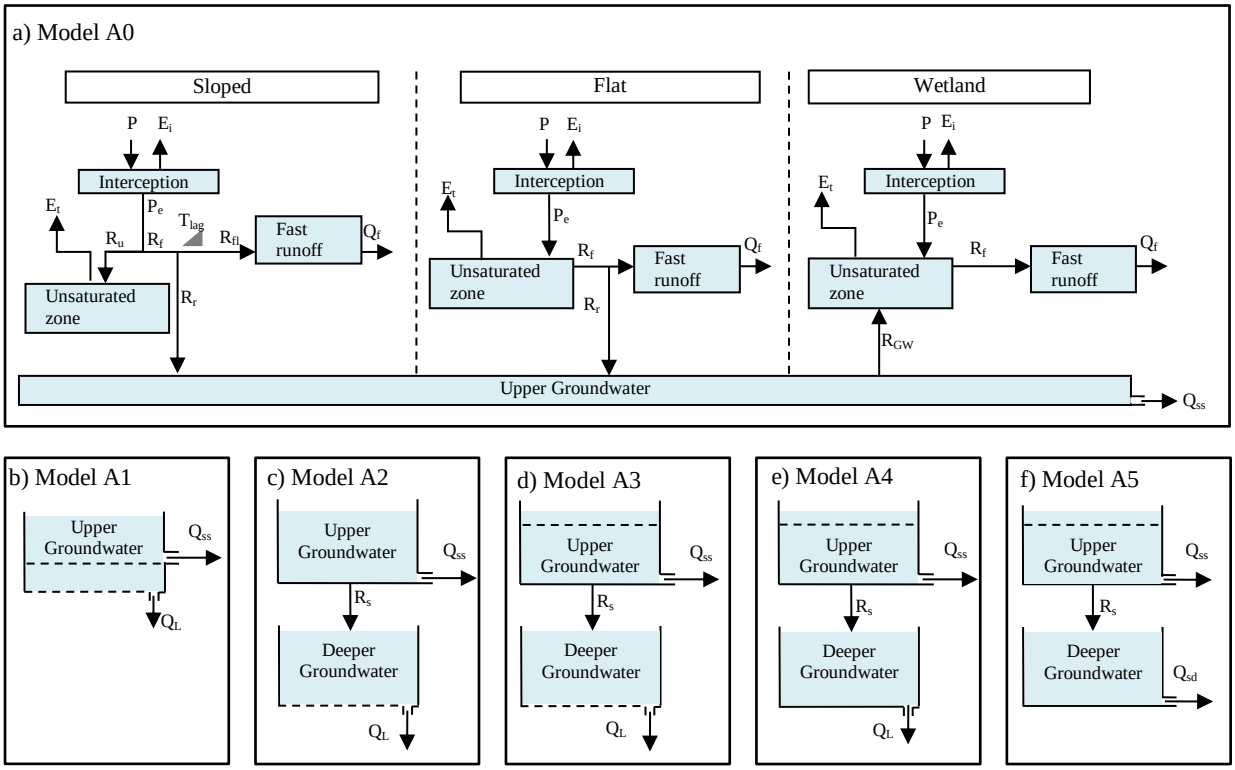


Figure 2. Schematisation of the model structure applied to each grid cell for Models A0 – A5. For Models A1 – A5 (b – f), only the groundwater module is shown for brevity and clarity of the presentation, as the rest of the model structure remained the same. Abbreviations: precipitation (P), effective 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), groundwater upwelling (R_{GW}), fast runoff (Q_f), groundwater recharge into Deeper Groundwater reservoir (R_s), shallow groundwater flow (Q_{ss}), groundwater loss (Q_L) and deep groundwater flow (Q_{sd}).

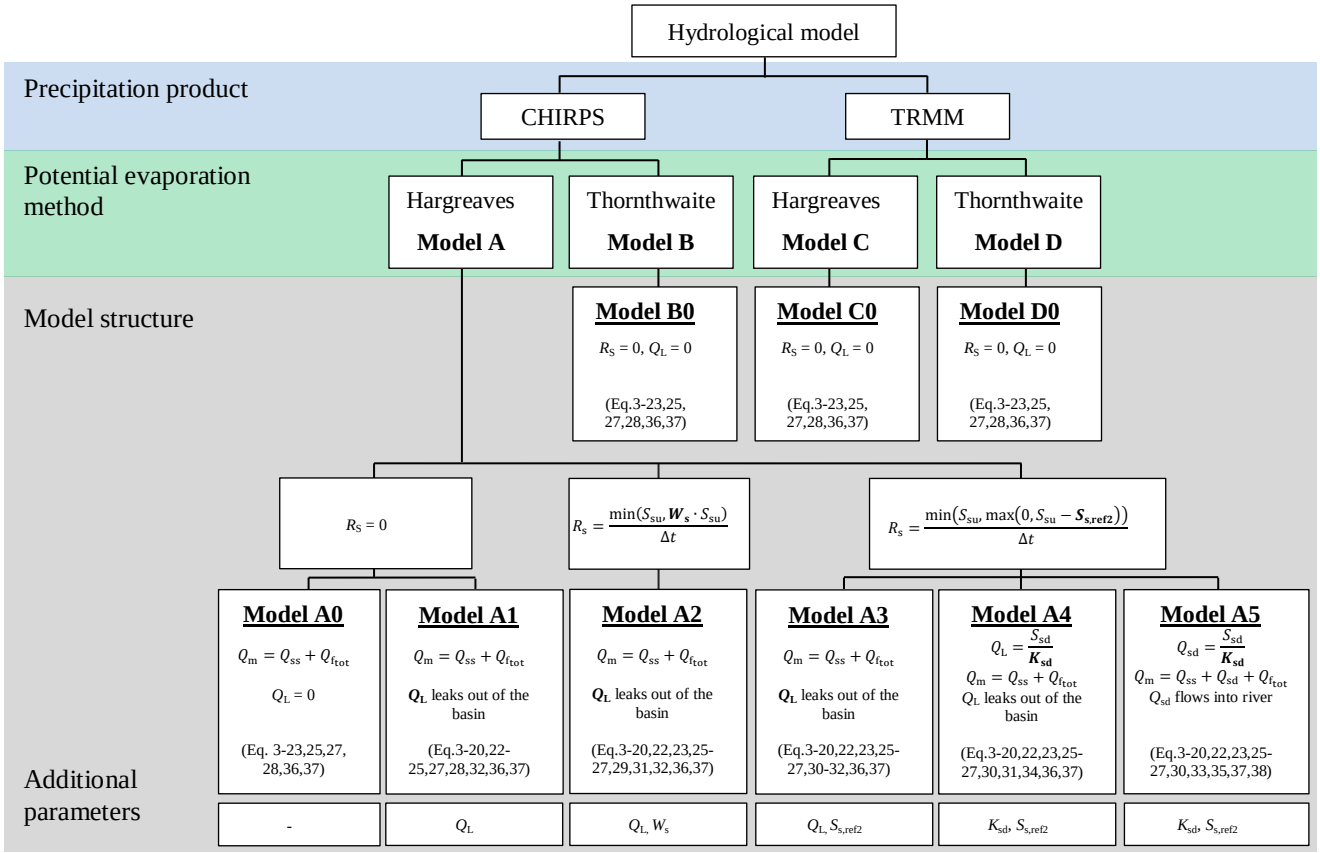


Figure 3. Overview hydrological models

Table 3. Equations applied in the hydrological model

Reservoir system	Water balance equations	Eq.	Process functions	Eq.
Interception	$\frac{\Delta S_\mathrm{i}}{\Delta t} = P - P_\mathrm{e} - E_\mathrm{i} \approx 0$	(3)	$E_\mathrm{i} = \min\left(E_\mathrm{p}, \min\left(P, \frac{\boldsymbol{I_{max}}}{\Delta t}\right)\right)$	(4)
			$P_\mathrm{e} = P - E_\mathrm{i}$	(5)
Unsaturated Root-zone	Sloped: $\frac{\Delta S_\mathrm{u}}{\Delta t} = R_\mathrm{u} - E_\mathrm{t}$	(6)	$E_\mathrm{t} = \min\left(\left(E_\mathrm{p} - E_\mathrm{i}\right), \min\left(\frac{S_\mathrm{u}}{\Delta t}, \left(E_\mathrm{p} - E_\mathrm{i}\right) \cdot \frac{S_\mathrm{u}}{\boldsymbol{S_{u,max}}} \cdot \frac{1}{\boldsymbol{C_e}}\right)\right)$	(7)
	Flat: $\frac{\Delta S_\mathrm{u}}{\Delta t} = \mathrm{P_e} - E_\mathrm{t} - R_f$	(8)	$R_\mathrm{GW} = \min\left(\frac{\min(S_\mathrm{su}, \boldsymbol{S_{s,ref1}})}{\boldsymbol{S_{s,ref1}}} \cdot \boldsymbol{C_{max}}, \frac{\frac{S_\mathrm{su}}{\Delta t}}{p_\mathrm{HRU}}\right)$	(9)
	Wetland: $\frac{\Delta S_\mathrm{u}}{\Delta t} = \mathrm{P_e} - E_\mathrm{t} - R_f + R_\mathrm{GW}$	(10)	if $S_\mathrm{u} + R_\mathrm{GW} \cdot \Delta t > \boldsymbol{S_{u,max}}$: $R_\mathrm{GW} = \frac{\boldsymbol{S_{u,max}} - S_\mathrm{u}}{\Delta t}$	(11)
			Sloped: $R_\mathrm{u} = (1 - C) \cdot P_\mathrm{e}$	(12)
			$C = 1 - \left(1 - \frac{S_\mathrm{u}}{\boldsymbol{S_{u,max}}}\right)^{\boldsymbol{\beta}}$	(13)
Fast runoff	$\frac{\Delta S_\mathrm{f}}{\Delta t} = R_\mathrm{fl} - Q_\mathrm{f}$	(14)	$Q_\mathrm{f} = \frac{S_\mathrm{f}}{\boldsymbol{K_f}}$	(15)
			Sloped: $R_\mathrm{f} = C \cdot P_\mathrm{e}$	(16)
			$R_\mathrm{fl} = (1 - \boldsymbol{W}) \cdot R_\mathrm{f} * \mathrm{f}(\boldsymbol{T_{lag}})$	(17)
			Flat/Wetland: $R_\mathrm{f} = \frac{\max(0, S_\mathrm{u} - \boldsymbol{S_{u,max}})}{\Delta t}$	(18)
			Flat: $R_\mathrm{fl} = (1 - \boldsymbol{W}) \cdot R_\mathrm{f}$	(19)
			Wetland: $R_\mathrm{fl} = R_f$	(20)
Upper Groundwater	$\frac{\Delta S_\mathrm{su}}{\Delta t} = R_{\mathrm{r_{tot}}} - R_{\mathrm{GW_{tot}}} - Q_\mathrm{ss}$	(21)	$R_\mathrm{r} = \boldsymbol{W} \cdot R_\mathrm{f}$	(22)
			$R_{\mathrm{r_{tot}}} = \sum_\mathrm{HRU} p_\mathrm{HRU} \cdot R_\mathrm{r}$	(23)
	$\frac{\Delta S_\mathrm{su}}{\Delta t} = R_{\mathrm{r_{tot}}} - R_{\mathrm{GW_{tot}}} - Q_\mathrm{ss} - Q_\mathrm{L}$	(24)	$R_{\mathrm{GW_{tot}}} = \sum_\mathrm{HRU} p_\mathrm{HRU} \cdot R_\mathrm{GW}$	(25)
	$\frac{\Delta S_\mathrm{su}}{\Delta t} = R_{\mathrm{r_{tot}}} - R_{\mathrm{GW_{tot}}} - Q_\mathrm{ss} - R_\mathrm{s}$	(26)	$Q_\mathrm{ss} = \frac{\max(0, S_\mathrm{su})}{\boldsymbol{K_s}}$	(27)
			$R_\mathrm{s} = 0$	(28)
			$R_\mathrm{s} = \frac{\boldsymbol{W_s} \cdot S_\mathrm{su}}{\Delta t}$	(29)
			$R_\mathrm{s} = \frac{\min(S_\mathrm{su}, \max(0, S_\mathrm{su} - \boldsymbol{S_{s,ref2}}))}{\Delta t}$	(30)
Deeper Groundwater	$\frac{\Delta S_\mathrm{sd}}{\Delta t} = R_\mathrm{s} - Q_\mathrm{L}$	(31)	$\boldsymbol{Q_L}$ = const.	(32)
	$\frac{\Delta S_\mathrm{sd}}{\Delta t} = R_\mathrm{s} - Q_\mathrm{sd}$	(33)	$Q_\mathrm{L} = \frac{S_\mathrm{sd}}{\boldsymbol{K_{sd}}}$	(34)
			$Q_\mathrm{sd} = \frac{S_\mathrm{sd}}{\boldsymbol{K_{sd}}}$	(35)
Total runoff	$Q_\mathrm{m} = Q_{\mathrm{f_{tot}}} + Q_\mathrm{ss}$	(36)	$Q_{\mathrm{f_{tot}}} = \sum_\mathrm{HRU} p_\mathrm{HRU} \cdot Q_\mathrm{f}$	(37)
	$Q_\mathrm{m} = Q_{\mathrm{f_{tot}}} + Q_\mathrm{ss} + Q_\mathrm{sd}$	(38)		

Note. 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), groundwater recharge into Deeper Groundwater reservoir (R_s), shallow groundwater flow (Q_{ss}), deep groundwater flow (Q_{sd}), groundwater loss (Q_L), total runoff (Q_m). Storages [mm]: storage in interception reservoir (S_i), storage in unsaturated root zone (S_u), storage in upper/deeper groundwater reservoir (S_{su}, S_{sd}), storage in fast reservoir (S_f). Calibration parameters (shown in bold): interception capacity (I_{max}) [mm], maximum upwelling groundwater (C_{max}) [mm d^{-1}], maximum root zone storage capacity (S_{umax}) [mm], splitter (W) [-], shape parameter (β) [-], transpiration coefficient (C_e) [-], time lag (T_{lag}) [d], reservoir timescales [d] of fast (K_f) and slow (K_s, K_{sd}) reservoirs, reference groundwater level ($S_{s,ref1}$, $S_{s,ref2}$) [mm], groundwater splitter (W_s) [-]. Remaining parameters: areal weights for each grid cell (p_{HRU}) [-], time step (Δt) [d]. The equations were applied to each hydrological response unit (HRU) unless indicated differently.

Table 4. Model parameters and prior distributions

Landscape class	Parameter	min	max	Unit	Constraint	Comment
Entire basin	C_e	0	1	-		All models
	K_s	90	110	d		All models
	$S_{sref,1}$	1	50	mm		All models
	Q_L	0	0.5	mm d^{-1}		Models A1, A2, A3
	K_{sd}	100	2500	d		Models A4, A5
	$S_{sref,2}$	1	50	mm		Models A3, A4, A5
	W_s	0	1	-		Model A2
Flat	I_{max}	0	5	mm d^{-1}		All models
	$S_{u,max}$	10	800	mm		All models
	K_f	10	12	d		All models
	W	0.01	1	-		All models
Sloped	I_{max}	0	5	mm d^{-1}	$I_{max,sloped} > I_{max,flat}$	All models
	S_{umax}	10	800	mm	$S_{umax,sloped} > S_{umax,flat}$	All models
	β	0	2	-		All models
	T_{lag}	1	5	d		All models
	K_f	10	12	d		All models
	W	0.01	1	-	$W_{sloped} > W_{flat}$	
Wetland	I_{max}	0	5	mm d^{-1}	$I_{max,wetland} < I_{max,sloped}$	All models
	S_{umax}	10	400	mm	$S_{umax,wetland} < S_{umax,sloped}$	All models
	K_f	10	12	d		All models
	C_{max}	0.01	5	mm d^{-1}		All models
River profile	v	0.01	5	m s^{-1}		All models

4.3 Model performance metrics

The model performance was evaluated with respect to discharge and basin-average total water storage anomalies. With respect to discharge, eight hydrological signatures were evaluated simultaneously using the Nash-Sutcliffe efficiency ($E_{NS,\theta}$, Eq.39) or relative error ($E_{R,\theta}$, Eq.40), depending on the signature. The individual performance metrics included the Nash-Sutcliffe efficiency of the daily flow time-series ($E_{NS,Q}$) and its logarithm ($E_{NS,logQ}$), of the flow duration curve ($E_{NS,FDC}$) and its logarithm ($E_{NS,logFDC}$), and of the autocorrelation function of the daily flows ($E_{NS,AC}$). In addition the relative error of the mean seasonal runoff during dry and wet periods ($E_{R,RCdry}$, $E_{R,RCwet}$), and the rising limb density of the hydrograph ($E_{R,RLD}$) (Euser et al., 2013) were used. These signatures were combined, assuming equals weights, using the Euclidian distance ($D_{E,Q}$, Eq.41) with $D_{E,Q}$ = 1 corresponding to the "perfect" model.

The model performance with respect to the basin-average total water storage anomalies was evaluated with the Euclidian distance ($D_{E,S}$, Eq.41) of the Nash-Sutcliffe efficiencies on monthly ($E_{NS,S,monthly}$) and annual ($E_{NS,S,annual}$) timescale. On annual timescale, the Nash-

Sutcliffe efficiency was calculated for the annual minima and maxima separately which were then averaged to obtain $E_{NS,S,annual}$. The annual time-series were normalised by dividing it with the maximum range in the observed annual minima or maxima total water storage respectively. With this performance measure for the total water storage, more emphasis could be given to annual variations rather than to seasonal variations only.

The combined model performance with respect to discharge and total water storage anomalies ($D_{E,QS}$) was calculated with the Euclidian distance (Eq.41) using $D_{E,Q}$ for the discharge and $D_{E,S}$ for the total water storage. This performance measure was used to select the best performing models representing both the discharge and the total storage as good as possible.

Table 5. Overview of equations used to calculate the model performance

Name	Objective Function	Equation	Variable explanation
Nash-Sutcliffe efficiency	$E_{\mathrm{NS},\theta} = 1 - \frac{\sum_t (\theta_{\mathrm{mod}}(t) - \theta_{\mathrm{obs}}(t))^2}{\sum_t (\theta_{\mathrm{obs}}(t) - \overline{\theta_{\mathrm{obs}}})^2}$	(39)	θ variable
Relative error	$E_{\mathrm{R},\theta} = 1 - \frac{\lvert\theta_{\mathrm{mod}} - \theta_{\mathrm{obs}}\rvert}{\theta_{\mathrm{obs}}}$	(40)	
Euclidian distance over multiple variables	$D_{\mathrm{E}} = 1 - \sqrt{\frac{1}{N}(\sum_n (1 - E_{\mathrm{n}})^2)}$	(41)	E_{n} model performance metric of variable n

4.4 Parameter selection procedure

The hydrological model was calibrated by running the model with 10^5 random parameter sets generated with a Monte-Carlo sampling strategy with uniform prior parameter distributions. Then, following two different strategies, the optimal parameter set was selected according to the model performance metrics as previously described with respect to 1) discharge ($D_{E,Q}$) and 2) discharge combined with total water storage ($D_{E,QS}$). The 5% best-performing parameter sets with respect to $D_{E,Q}$ or $D_{E,QS}$ were considered as feasible. The feasible parameter sets were used to evaluate the model performance with respect to discharge and total water storage anomalies individually and combined. The model was run for the time period 1995 – 2016 and calibrated/evaluated for the time period 2002 – 2016 using the first seven years as warm-up period. The entire time period (2002 – 2016) was used to estimate the model performance with respect to discharge and total water storage to capture the long-term variability as good as possible.

In addition, the predictive strength of the benchmark Model A0 and the best performing model hypothesis (i.e. third model adaptation; Section 4.2.4) were compared by calibrating

both models with respect to discharge and total water storage simultaneously ($D_{E,QS}$) for the time period 2002 – 2012, and post-calibration evaluating the models with respect to total water storage for the time period 2012 – 2016. Due to the limited data availability in 2012 – 2016, the model could not be evaluated with respect to discharge.

5 Results

5.1 Data analysis

5.1.1 GRACE total water storage anomalies

In the Luangwa basin, the total water storage anomalies varied both seasonally and in the long-term (for example Figure 4a). The seasonal variation, hence the difference between the annual maximum and minimum, remained rather similar throughout the years (on average 225 mm). However, the annual minima, mean and maxima changed over the years indicating relatively dry conditions in the Luangwa basin for example during the 2005 – 2007 period and wetter conditions in the 2009 – 2011 period. The annual minima varied between -164 mm in 2016 and -67 mm in 2009, while the annual maxima varied between 75 mm in 2016 and 183 mm in 2010. Also the annual mean varied over the years between -46 mm in 2006 and 48 mm in 2010. This study focused on annual minima/maxima separately instead of the annual mean to distinguish processes dominant in wet seasons influencing the annual maxima and dry seasons affecting the annual minima.

One possibility is that these variations were a result of uncertainties in GRACE observations as the Luangwa basin is relatively small (150,000 km^2) relative to the resolution of GRACE. Previous studies estimated errors in GRACE observations to be about 20 mm for areas of around 63,000 km^2 (Landerer and Swenson, 2012;Vishwakarma et al., 2018). But similar long-term variations were also observed for the entire Zambezi basin (Figure 4b), which is considerably larger (1,390,000 km^2) and where the maximum variation (194 mm) was an order of magnitude larger than the average uncertainty error of 20 mm.

In addition, long-term variations in large open water bodies could influence the GRACE signal. In this study, multiple open water bodies were within a radius of 300 km of the Luangwa Basin (Figure 1A) which typically is the distance used for data smoothing when processing GRACE data (Landerer and Swenson, 2012;Blazquez et al., 2018). The area of these open water bodies were 2% of the Luangwa basin for the Cahora Bassa reservoir, 4% for the Kariba reservoir and 20% for Lake Malawi. As no long-term variations were observed in the altimetry observations for the Cahora Bassa reservoir (Figure S1 in the Supplementary Material) and since this reservoir had a small area compared to the Luangwa basin, the effect of this reservoir was assumed to be negligible. For the Kariba reservoir (Figure 4c) and Lake Malawi (Figure 4f), long-term variations were observed in the altimetry data, but with a low

temporal correlation with the total water storage as shown in Figure S2 in the Supplementary Material. For the Zambezi basin where similar long-term storage variations were observed (Figure 4b), these three open water bodies covered together 2.7% of the basin. This was considered to be too small to have a significant effect. That is why it is plausible to assume that these long-term storage variations were not dominated by uncertainties in the GRACE observations.

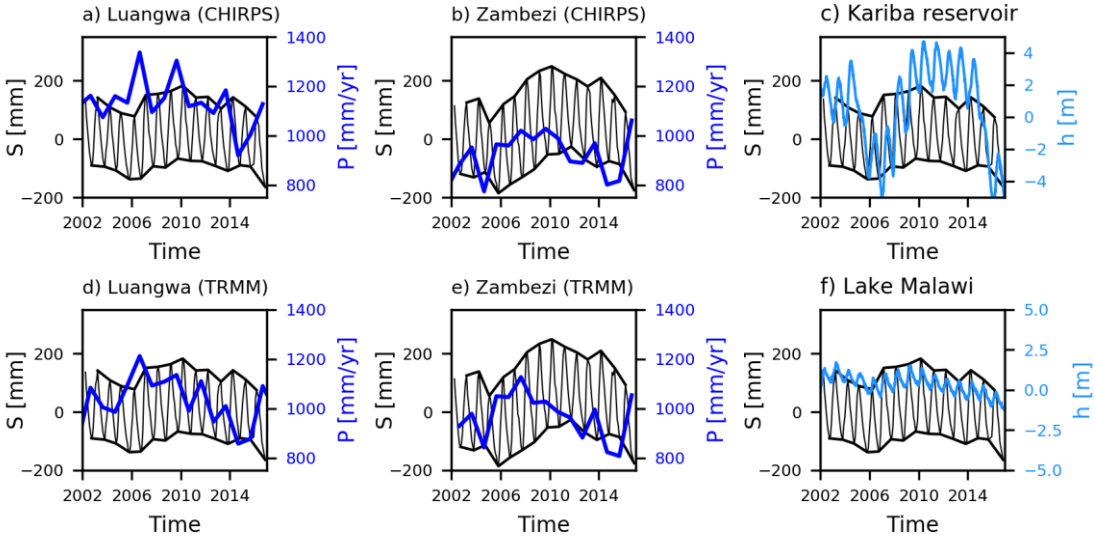


Figure 4. Basin-average total water storage (black) and annual rainfall (dark blue) according to CHIRPS (a and b) and TRMM (d and e) for the Luangwa (a and d) and Zambezi (b and e) river basin, or altimetry observations (light blue) at c) Kariba reservoir and f) Lake Malawi.

5.1.2 Precipitation

Alternatively, long-term variations in the total water storage can be caused by changes in the precipitation. In the Luangwa basin, the annual observed precipitation volumes varied over the years, depending on the data source, from 920 mm to 1337 mm (CHIRPS) and from 858 mm to 1213 mm (TRMM), as shown in Figures 4a) and d). In general, precipitation anomalies preceded storage variations by roughly 1 – 3 years. According to CHIRPS (Figure 4a), the rainfall volumes peaked in 2006 and 2009 with a significant decrease in 2008 – 2009 and 2014. While the increased rainfall volumes in 2006 and 2009 could explain the increased total water storage anomalies between 2008 and 2010, the significantly decreased rainfall volumes in 2008 – 2009 did not correspond to the long-term total water storage pattern. The correlation between the annual rainfall volumes according to CHIRPS and the annual maximum total water storage showed a R^2 = 0.10 without taking any time shift into account and reached up to R^2 = 0.29 with a two year time shift.

According to TRMM, the annual rainfall volumes decreased in 2004 – 2005 which could explain the decreased lower total water storage in 2006. This was followed by several wet years with a maximum rainfall volume of 1213 mm in 2006 which could explain the increased total water storage starting in 2007. The annual rainfall volumes decreased significantly in 2014 – 2015 as low as 858 mm which corresponded to the decreased total water storage in 2016. The correlation between the annual rainfall volumes according to TRMM and the annual maximum total water storage reached R^2 = 0.28 without taking any time shift into account and reached up to R^2 = 0.34 with a two year time shift.

This difference between CHIRPS and TRMM illustrated the high sensitivity of the annual rainfall volumes to the underlying processing techniques (Cohen Liechti et al., 2012;Thiemig et al., 2012;Le Coz and van de Giesen, 2019;Mazzoleni et al., 2019). Strikingly, for the entire Zambezi river basin the annual variability in the precipitation according to both CHIRPS and TRMM show a similar pattern compared to each other and to the storage variations. The annual rainfall volumes decreased in 2004 followed by low total water storages in 2006, after which both the rainfall and total water storage increased with a maximum in 2009 (CHIRPS), 2007 (TRMM) and 2010 (GRACE). These observations suggest that long-term variations in precipitation alone already contain considerable information to potentially explain much of the observed long-term storage variations.

5.1.3 Potential and actual Evaporation

The two different methods to estimate potential evaporation and its variations over the study time period, gave dramatically different results. While the Hargreaves method suggested a long-term mean annual E_P = 1565 mm yr^{-1} (Figure 5a), Thornthwaite estimated long-term mean E_P = 1904 mm yr^{-1} (Figure 5b). Major long-term variations in E_P were only observed for estimates based on the Thornthwaite method (Figure 5b), but with a different pattern compared to the total water storage resulting in low correlation coefficients when focusing on the annual mean variations (R^2 = 0.02). In contrast, no discernible long-term fluctuations were observed when applying the Hargreaves method (R^2 = 0.03). As the potential evaporation did change over the years according to the Thornthwaite method, it is possible this was one of the reasons why the modelled total water storage did not capture any long-term variations when using the Hargreaves method for the potential evaporation.

Analysis of the actual evaporation did not reveal any systematic long-term patterns that could clearly explain observed variations in the total water storage for most of the satellite products used in this study (Figure S3 in the Supplementary Material). In general, the magnitudes and long-term fluctuations varied for each satellite product as a result of different underlying assumptions and input data which could influence whether or not long-term fluctuations are visible. This resulted in a range of R^2 = 0.02 – 0.17 with respect to the annual minima for all satellite products used in this study except for SSEBop which showed the highest R^2 = 0.37

and where the evaporation increased between 2006 and 2010 similar to the storage (Figures 5c and S3 in the Supplementary Material). Note, that the observed annual minimum storage increase of 67 mm over three years (2006 – 2009), which in fact is an accumulated difference arising from the combined history of inputs and outputs over that period, can result among others from a mean deviation of only 0.06 mm d^{-1} in the evaporation, which is by far within the uncertainty range of many satellite-based evaporation products (Long et al., 2014;Westerhoff, 2015). Hence, evaporation can potentially be one of the drivers for the observed long-term storage fluctuations, but additional in-depth analyses is necessary to substantiate this hypothesis which was outside the scope of this study due to the limited ground observations available.

Overall, long-term variations in potential and actual evaporation, according to most satellite products used here, exhibited less direct correspondence with water storage variations, which was likely a consequence of the subtle spatially varying interactions between water supply and atmospheric water demand in this largely water limited environment. Thus, while actual evaporation is largely controlled by water supply in hillslope regions, it is to a higher degree dominated by variations in atmospheric water demand in wetland areas, where sufficient water supply is sustained by shallow groundwater throughout most of the year. On the basin average, these processes can, to some degree, cancel each other out and thus prevent the development of a clear long-term signal. Based on the above analysis it therefore remains difficult to meaningfully assess the uncertainty of the different analysed evaporation products.

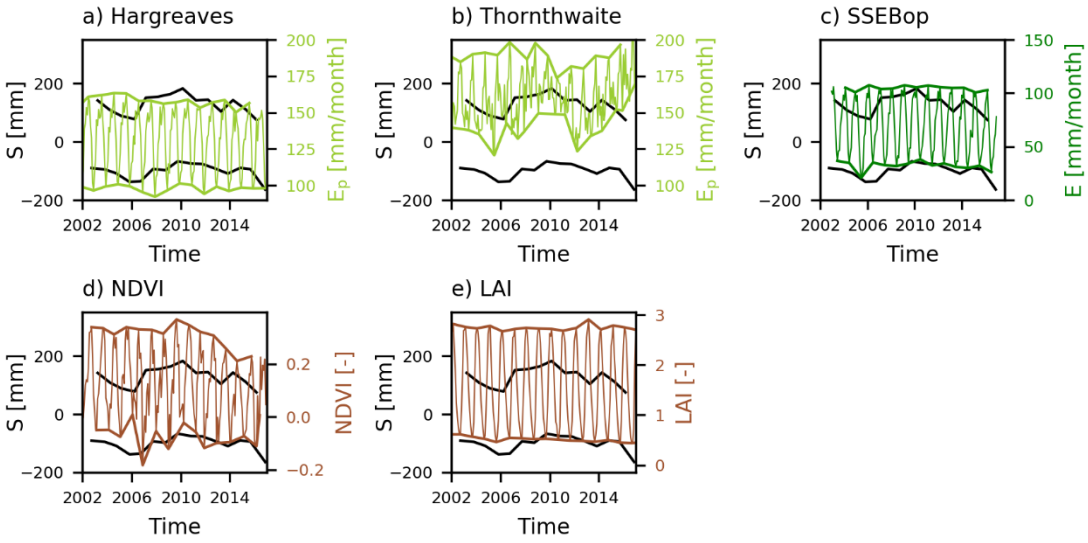


Figure 5. Basin-average total water storage (black) with respect to the annual minima/maxima combined with basin-average a) monthly potential evaporation according to Hargreaves (light green) and b) Thornthwaite (light green), c) monthly actual evaporation according to SSEBop (dark green), d) NDVI (brown), and e) LAI (brown) including the annual minima/maxima of the respective variables.

5.1.4 Land-cover

Affecting the magnitudes of transpiration, land-cover changes could also be one of the drivers for the observed annual storage variations. In the Luangwa basin, deforestation, forest recovery and agricultural expansion have occurred in the past (Handavu et al., 2019;Phiri et al., 2019b, a). However, inspections of time-series of LAI and NDVI (Figure 5) did not reveal any significant long-term variations directly corresponding with water storage variations over the 2002 – 2016 period. While LAI did not exhibit any significant long-term variation, NDVI showed some fluctuations, including a considerable decrease after 2010, which, however, did not directly correspond with the observed water storage variations. This resulted in low correlations between the annual mean total water storage and LAI (R^2 = 0.003) and NDVI (R^2 = 0.04). It was therefore assumed that land use change did not play a major role for the observed long-term storage variations.

5.1.5 Overall water balance

Another potential reason for the observed long-term storage variations can be regional, inter-basin groundwater exchange. For example, groundwater may leak out of the Luangwa basin below the river, thus never contributing to the (river) flow at the basin outlet, and into the Zambezi river basin further downstream eventually draining into that river or potentially even directly into the sea. Given the available observations, this would result in a water balance surplus for the Luangwa basin. Depending on the rainfall and evaporation products used, the water balance surplus in the Luangwa basin for the study period ranged between 9 and 332 mm yr^{-1} (Table 1). This suggested that even in the likely presence of data uncertainty, groundwater export may occur at least to some degree in the study region. Assuming an inter-basin export of $\overline{Q_L}$ = 332 mm yr^{-1}, discharge would be considerably overestimated as compared to actual discharge observations (Figure 6). To remain within the ranges spanned by multiple analytical solutions for water partitioning in the Budyko space (dark grey area in Figure 6; Gerrits et al., 2009), groundwater export should not exceed $\overline{Q_L}$ = 143 mm yr^{-1}, which corresponds to a mean daily flow of $\overline{Q_L}$ = 0.39 mm d^{-1} or ~13% of the annual rainfall. Therefore, based on the water balance, a plausible estimate for groundwater export of $\overline{Q_L}$ = 0 – 0.39 mm d^{-1} is in the following assumed for the study basin.

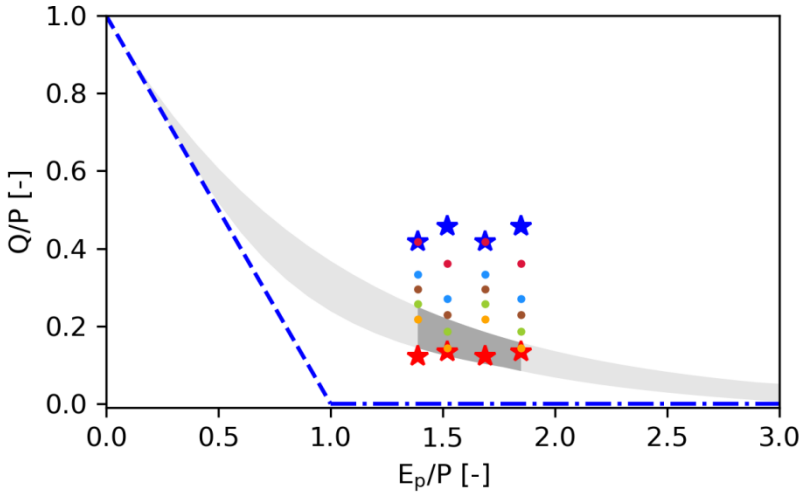


Figure 6. Runoff coefficient (Q/P) as a function of the dryness index (E_p/P) where Q is discharge, P precipitation, and E_p potential evaporation. The blue dashed line indicates the energy limit and the blue horizontal dash-dotted line the water limit. The grey area indicates envelope of analytical solutions according to Schreiber (1904), Ol'dekop (1911), Turc (1953), Pike (1964) and Budyko (1974). The dryness index was estimated using CHIRPS or TRMM for the precipitation and the Hargreaves method ($\overline{E_P}$ = 1565 mm yr^{-1}) or Thornthwaite ($\overline{E_P}$ = 1904 mm yr^{-1}) for the potential evaporation. The runoff coefficient was estimated with the same precipitation products and 1) recorded discharge without groundwater exchange (red stars), 2) estimated discharge including groundwater exchange ($\bar{Q} + \overline{Q_L} = \bar{P} - \bar{E}$, Eq.(2) using the same precipitation products and SEBS (red dots), GLEAM (blue dots), MOD16 (brown dots), SSEBop (green dots) and WaPOR (orange dots) for the evaporation resulting in $\overline{Q_L}$ = 9 – 332 mm y^{-1} depending on the chosen satellite products, and 3) sum of recorded discharge and maximum groundwater export ($\overline{Q_L}$ = 332 mm yr^{-1}, blue stars). To remain within the Budyko space (dark grey area), the groundwater exchange should range between $\overline{Q_L}$ = -51 – 143 mm yr^{-1} depending on the satellite products used. See Table 1 for the corresponding long-term values of the individual fluxes.

5.2 Hydrological models

5.2.1 Benchmark model (Model A0)

Following the first strategy, i.e. calibrating with respect to discharge, the benchmark Model A0 captured the discharge well (Figures 7a and b) with an optimum model performance of $D_{E,Q,opt}$ = 0.85 (Table 6, Figure 8a). The modelled flow dynamics such as the timings of the wet and dry season were broadly consistent with the observations (Figure 7a), but the high flows were slightly underestimated and low flows somewhat overestimated (Figure 7b). In contrast, and in spite of its general ability to reproduce discharge, the model could only poorly reproduce the time-series of monthly and annual total water storage anomalies with $D_{E,S}$ = -14 (Table 6, Figure 8a). On monthly timescale, the general seasonal fluctuations were

modelled well with respect to the timings of the wet and dry season (Figure S6a in the Supplementary Material). However, the annual maxima were significantly overestimated and the annual minima underestimated (Figure 7c). In addition, the modelled total water storage did not reflect any fluctuations in the annual minima in contrast to the observations (Figure 7e, R^2 = 0.07), whereas the modelled annual maxima varied throughout the years, but with a different pattern compared to the observations (Figure 7d, R^2 = 0.20). As a result, the overall model performance with respect to discharge and total water storage $D_{E,QS}$ = -9.6 remained poor.

Following the second strategy, i.e. calibration with respect to discharge and total water storage simultaneously, the ability of the model to reproduce flow decreased significantly to $D_{E,Q}$ = -0.23 (Table 6, Figure 8b). While the general flow dynamics were modelled well (Figure 9a), the flows were continuously overestimated (Figure 10a). In contrast, the modelled monthly and annual total water storage time-series improved ($D_{E,S}$ = -0.11). The modelled total water storage mimicked the seasonal variations in the observation better (Figure S7a in the Supplementary Material), but with slight differences in the storage decrease during the dry seasons. The magnitudes of the annual maxima and minima corresponded better with the observations (Figure 10b) and the fluctuations in the annual maxima improved slightly (Figure 10c, R^2 = 0.31). However, the modelled storage did not reflect any fluctuations in the annual minima (Figure 10d, R^2 = 0.06). Hence, the overall model performance $D_{E,QS}$ = -0.17 improved, but remained poor. Even when calibrating with respect to total water storage only, the annual minima did not reflect any fluctuations (Figure S8 in the Supplementary Material, R^2 = 0.08).

As a result, with this benchmark Model A0 the flows were modelled well as also the seasonal fluctuations in the total water storage. However, the long-term variations in the total water storage with respect to the annual maxima were poorly modelled and with respect to the annual minima completely missed.

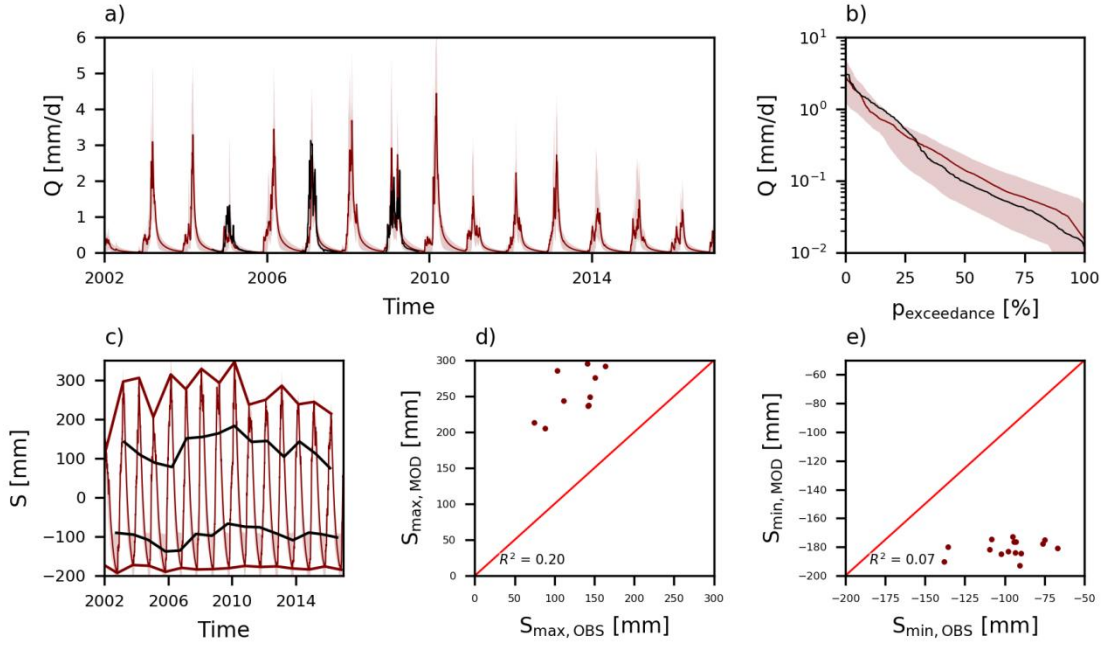


Figure 7. Range of model solutions for Model A0 for calibration strategy 1 with respect to a) hydrograph, b) flow duration curve, c) total water storage time-series, d) annual maximum total water storage, and e) annual minimum total water storage. In a) to c), the black line indicates the recorded data, the coloured line the solution with the highest calibration objective function with respect to discharge ($D_{E,Q}$) and the shaded area the envelope of the solutions retained as feasible. In d) and e), the recorded data are plotted on the horizontal axis and on the vertical axis the model solution with the highest calibration objective function with respect to discharge ($D_{E,Q}$). The red line indicates the 1:1 line.

5.2.2 First model adaptation: Alternative forcing data (Models B0 – D0)

Following the first calibration strategy, Models B0 – D0, using different combinations of input data sources, represented the discharge in general well with $D_{E,Q}$ = 0.85 – 0.92 (Table 6, Figure 8a). All models reproduced the overall flow dynamics and magnitudes well (Figures S4 and S5a in the Supplementary Material), especially Models C0 ($D_{E,Q}$ = 0.91) and D0 ($D_{E,Q}$ = 0.92). The monthly and annual total water storage remained poorly modelled for all models with $D_{E,S}$ = -3.4 – -0.48 (Table 6, Figure 8a). On monthly timescale, the general seasonal fluctuations were modelled well with slight differences mostly in the storage decrease during dry seasons (Figure S6 in the Supplementary Material). The magnitudes of the modelled annual minima corresponded well with the observation for all models, but the annual maxima were overestimated for Models B0 and C0, whereas this improved the most for Model D0 (Figure S5b in the Supplementary Material). In addition, the annual minimum storage did not exhibit any of the observed long-term variations in any of the models (R^2 = 0.02 – 0.10,

Figure S5c – d in the Supplementary Material), whereas the fluctuations in the annual maxima improved the most for Model D0 (R^2 = 0.35). As a result, the overall model performance with respect to discharge and total water storage improved the most for Model D0 with $D_{E,QS}$ = -0.05 (Table 6, Figure 8a) which remained poor.

Following the second calibration strategy, the modelled flow improved for all Models B0 – D0 to $D_{E,Q}$ = 0.32 – 0.83 compared to the benchmark Model A0 (Table 6, Figure 8b). The general flow dynamics were represented well for all models (Figure 9), but the flow magnitudes were only captured well for Models C0 and D0 (Figure 10a). While Models A0 and B0 significantly overestimated the flows continuously, Model C0 only slightly overestimated the flows continuously and Model D0 only slightly underestimated the medium to low flows (Figure 10a). As a result, Model D0 had the highest model performance with respect to discharge with $D_{E,Q}$ = 0.83 (Table 6, Figure 8b). Also the modelled monthly and annual total water storage improved for Models B0 – D0 with $D_{E,S}$ = 0.00 – 0.34 compared to the benchmark Model A0 (Table 6, Figure 8b). On monthly timescale, the general seasonal variations were captured well for all models, but with slight differences in the storage decrease during dry seasons (Figure S7 in the Supplementary Material). The magnitudes of the annual minima and maxima corresponded well with the observations for all models (Figure 10b), whereas the fluctuations in the annual maxima only improved for Model D0 with R^2 = 0.39 (Figure 10c). On the other hand, the annual minima remained close to constant for all models (R^2 = 0.00 – 0.03; Figure 10d). The overall model performance with respect to discharge and total water storage improved the most for Model D0 with $D_{E,QS}$ = 0.52 (Table 6, Figure 8b).

As a result, the ability of the model to reproduce long-term variations of the total water storage during the wet seasons, i.e. the annual maxima, was considerably influenced by the choice of precipitation data source and the method to estimate potential evaporation. In contrast, the modelled dry season storage, i.e. annual minima, did not reflect the observed pattern for any combination of data sources but remained rather stable. Overall, the combination of TRMM with the Thornthwaite method (Model D0) here produced model results that were most consistent simultaneously with observed discharge and the observed total water storage variations. This suggests that the choice of data source can explain some of the inability of the model to reproduce long-term water storage variations.

Table 6. Model performance with respect to discharge (D_E), total water storage ($D_{E,S}$) and both combined ($D_{E,QS}$) including their 5/95% percentile ranges of the feasible parameter sets for Models A0 – D4.

	Strategy 1: Discharge calibration (D_E)			**Strategy 2: Multi-variable calibration ($D_{E,QS}$)**		
	$D_{E,Q}$ ($D_{E,Q,5/95\%}$)	$D_{E,S}$ ($D_{E,S,5/95\%}$)	$D_{E,QS}$ ($D_{E,QS,5/95\%}$)	$D_{E,Q}$ ($D_{E,Q,5/95\%}$)	$D_{E,S}$ ($D_{E,S,5/95\%}$)	$D_{E,QS}$ ($D_{E,QS,5/95\%}$)
Model A0	0.85 (0.70 – 0.81)	-14 (-18 – -5.5)	-9.6 (-12 – -3.6)	-0.23 (-0.71 – -0.06)	-0.11 (-0.80 – -0.10)	-0.17 (-0.52 – -0.31)
Model B0	0.85 (0.72 – 0.81)	-3.4 (-9.2 – -1.7)	-2.1 (-6.2 – -0.94)	0.32 (-0.14 – 0.49)	0.00 (-0.65 – -0.09)	0.14 (-0.25 – 0.01)
Model C0	0.91 (0.80 – 0.88)	-0.85 (-4.5 – -0.34)	-0.31 (-2.9 – 0.05)	0.64 (0.26 – 0.72)	0.22 (-0.13 – 0.19)	0.39 (0.16 – 0.31)
Model D0	0.92 (0.84 – 0.90)	-0.48 (-2.2 – 0.21)	-0.05 (-1.3 – 0.43)	0.83 (0.56 – 0.88)	0.34 (0.09 – 0.28)	0.52 (0.34 – 0.46)
Model A1	0.84 (-0.13 – 0.71)	-15 (-15 – -0.87)	-11 (-10 – -0.51)	-0.20 (-1.1 – 0.07)	0.05 (-1.4 – -0.15)	-0.08 (-0.90 – -0.35)
Model A2	0.82 (-5.1 – 0.51)	-1066 (-813 – -3.4)	-753 (-575 – -3.3)	-0.24 (-11 – -1.0)	-0.47 (-7.5 – -0.68)	-0.36 (-7.6 – -3.3)
Model A3	0.87 (0.73 – 0.83)	-425 (-1133 – -11)	-300 (-801 – -7.2)	0.28 (-1.2 – 0.49)	-0.45 (-3.9 – -0.66)	-0.14 (-2.6 – -0.53)
Model A4	0.87 (0.73 – 0.83)	-9.8 (-27 – -3.6)	-6.7 (-19 – -2.3)	0.54 (-0.42 – 0.50)	0.16 (-0.64 – 0.11)	0.32 (-0.31 – 0.12)
Model A5	0.84 (0.68 – 0.79)	-13 (-18 – -5.3)	-9.0 (-12 – -3.5)	-0.31 (-0.72 – 0.03)	0.23 (-0.73 – 0.08)	-0.07 (-0.46 – -0.20)
Model D4	0.93 (0.85 – 0.91)	0.31 (-6.9 – 0.29)	0.51 (-4.6 – 0.49)	0.85 (0.61 – 0.89)	0.50 (0.11 – 0.37)	0.63 (0.35 – 0.53)

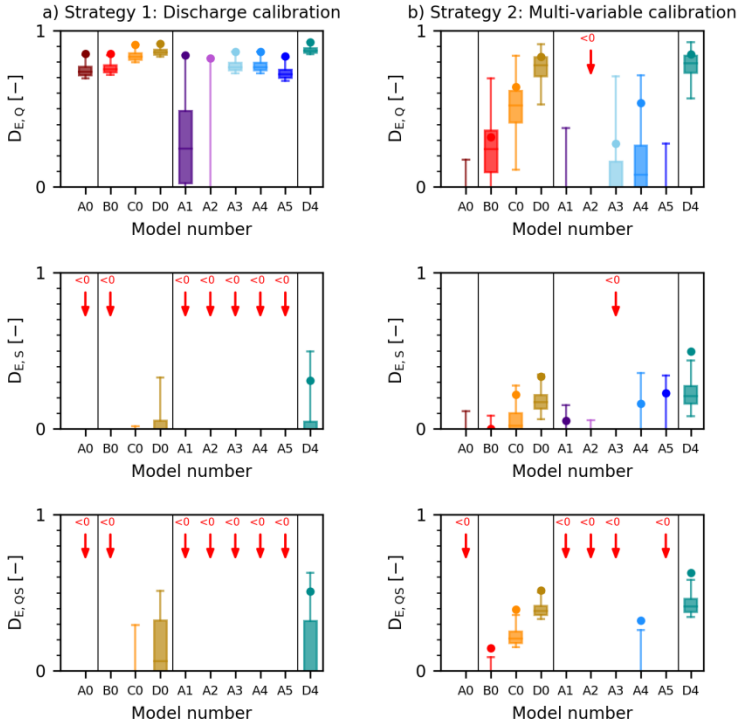


Figure 8. Model performance for Models A0 – D4 with respect to discharge ($D_{E,Q}$), total water storage anomalies ($D_{E,S}$) and both combined ($D_{E,QS}$). The model is calibrated with respect to a) discharge or b) both variables simultaneously. The dots represent the model performance using the "optimal" parameter set and the boxplot the range of the best 5% solutions according to $D_{E,Q}$ or $D_{E,QS}$. A red arrow was added if all solutions are below zero.

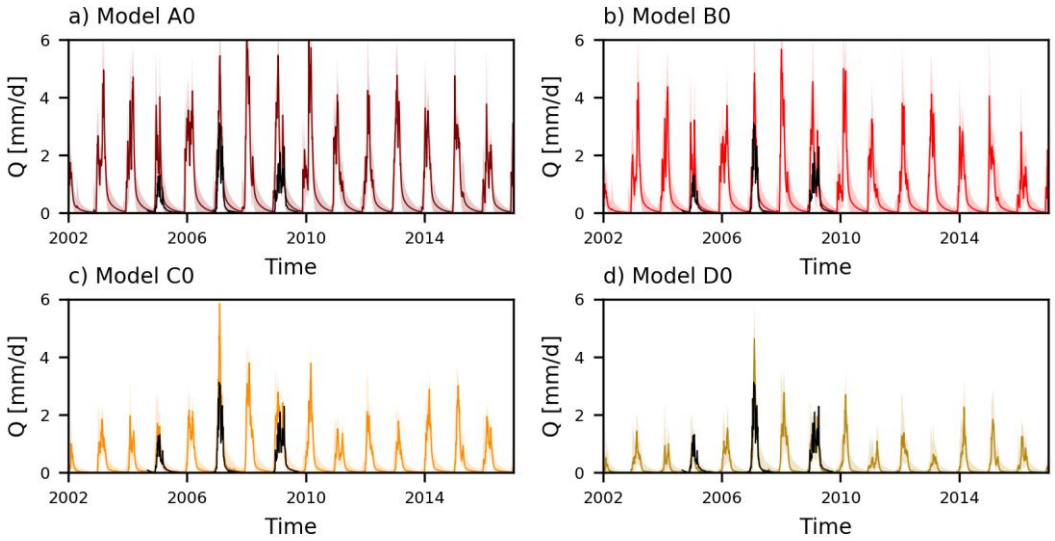
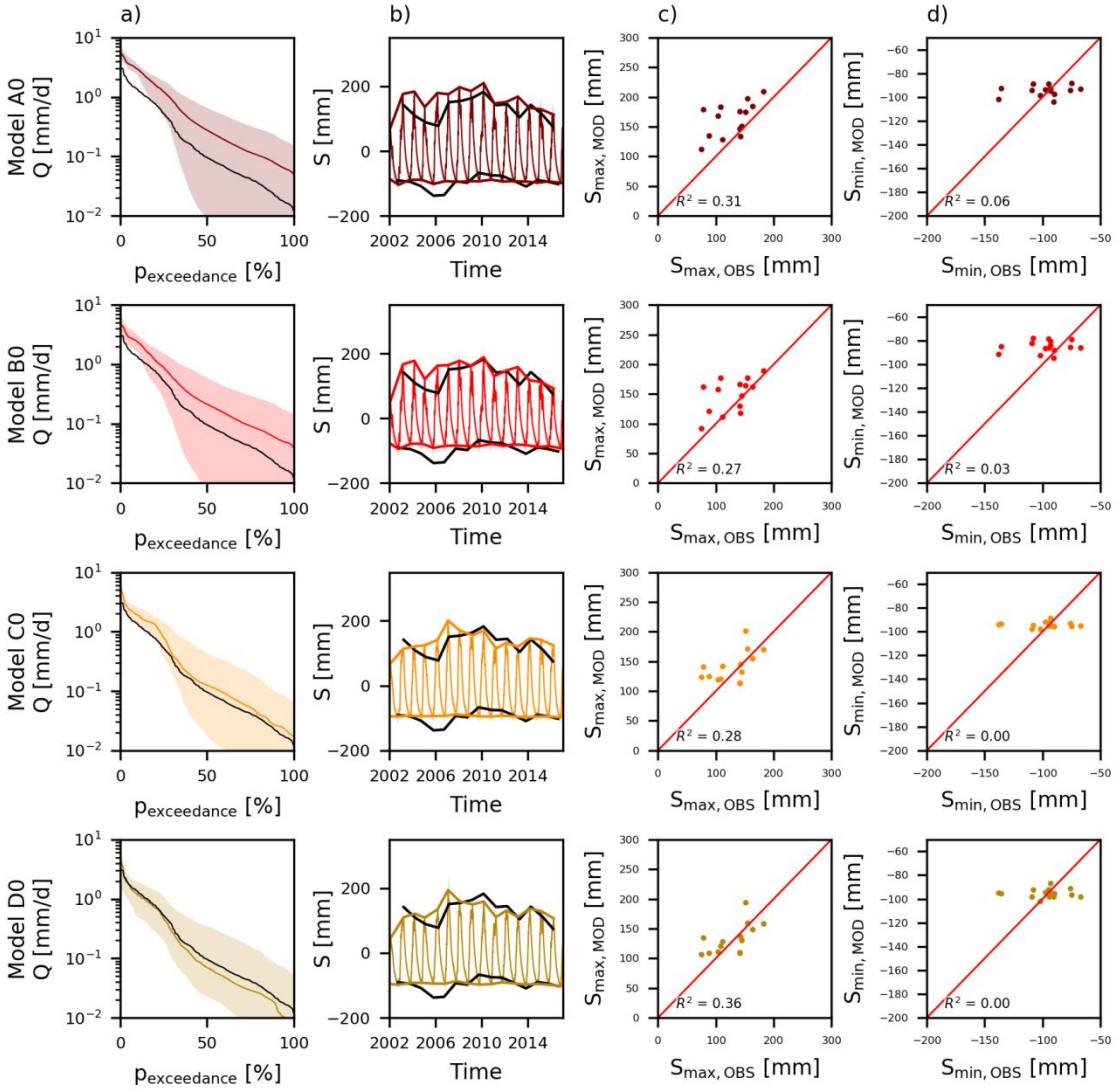


Figure 9. Range of model solutions for Models A0 – D0 for calibration strategy 2 with respect to discharge (hydrograph). The black line indicates the recorded data, the coloured line the solution with the highest calibration objective function with respect to discharge and total water storage ($D_{E,QS}$) and the shaded area the envelope of the solutions retained as feasible.



655 **Figure 10.** Range of model solutions for Models A0 – D0 for calibration strategy 2 with respect to a) flow duration curve, b) total water storage time-series, c) annual maximum total water storage, d) annual minimum total water storage. In a) – b), the black line indicates the recorded data, the coloured line the solution with the highest calibration objective function with respect to discharge and total water storage ($D_{E,QS}$) and the shaded area the envelope of the solutions retained as feasible. In c) – d), the recorded data are plotted on the horizontal
660 axis and on the vertical axis the model solution with the highest calibration objective function with respect to discharge and total water storage ($D_{E,QS}$). The red line indicates the 1:1 line.

5.2.3 Second model adaptation: Alternative model structure (Model A1 – A5)

Following the first calibration strategy, all Models A1 – A5 reproduced the discharge well with $D_{E,Q}$ = 0.82 – 0.87 (Table 6, Figure 8a). All models captured the general flow dynamics and magnitudes (Figures S9 and S10a in the Supplementary Material). The monthly and annual total water storage time-series was modelled very poorly for all models ($D_{E,S}$ = -1066 – -9.8, Table 6, Figure 8a). While Models A1 and A5 consistently over- or underestimated the storage with little resemblance in the fluctuations of the annual maxima (R^2 = 0.19 – 0.22) and minima (R^2 = 0.08 – 0.16), Models A2 and A3 substantially overestimated the long-term variations (R^2 = 0.00 – 0.11, Figures S10 and S11 in the Supplementary Material). Also in Model A4, the storage was over- or underestimated, but the long-term variations improved with respect to the annual maxima (R^2 = 0.56) and minima (R^2 = 0.27). As a result, the overall model performance with respect to discharge and total water storage simultaneously improved the most for Model A4 with $D_{E,QS}$ = 0.32 (Table 6, Figure 8a).

Following the second calibration strategy, the modelled discharge improved considerably for Models A3 ($D_{E,Q}$ = 0.28) and A4 ($D_{E,Q}$ = 0.54) compared to the benchmark Model A0, but was poorly represented for the remaining models with $D_{E,Q}$ = -0.31 – -0.20 (Table 6, Figure 8b). The general flow dynamics were reproduced well for Models A1 – A4 (Figure 11), albeit with slight differences in the timing of the wet season and dry season recession, whereas Model A5 poorly represented the recession during dry seasons. In addition, the flows were significantly over- or underestimated with Models A1 – A3 and A5 (Figure 12a), whereas Model A4 only slightly overestimated the high flows and underestimated the low flows. The monthly variations in the total water storage were captured well for all models with some differences in the storage decrease during dry seasons especially for Model A2 (Figure S12 in the Supplementary Material). While the magnitudes of the annual maxima and minima were captured well for all models (Figure 12b), the annual fluctuations improved the most Model A5 with respect to the annual maxima (R^2 = 0.51, Figure 12c) and for Models A2 and A5 with respect to the annual minima (R^2 = 0.23, Figure 12d). When considering both the monthly and annual fluctuations and magnitudes, Models A4 ($D_{E,S}$ = 0.16) and A5 ($D_{E,S}$ = 0.23) improved the most (Table 6, Figure 8b).

As a result, the model's ability to reproduce the long-term total water storage variations during dry and wet seasons, i.e. annual minima and maxima, was significantly influenced by the model structure. The modelled annual and monthly total water storage improved the most for Models A4 and A5 (Table 6, Figure 8b) where a Deeper Groundwater reservoir was incorporated with groundwater loss/flow as function of the water content in the Deeper Groundwater reservoir. However, Model A5 only poorly captured the discharge ($D_{E,Q}$ = -0.31, Figure 12a). Therefore, when considering the overall model performance with respect to discharge and total water storage simultaneously ($D_{E,QS}$), Model A4 performed the best with $D_{E,QS}$ = 0.32 (Table 6, Figure 8b). This model captured the flows well as also the monthly and

annual total water storage magnitudes and fluctuations, albeit with a slight overestimation of the annual minima and maxima in 2004 – 2006 (Figure 12b). These results indicated long-term storage fluctuations were most likely a result of groundwater loss from the Deeper Groundwater reservoir (Model A4).

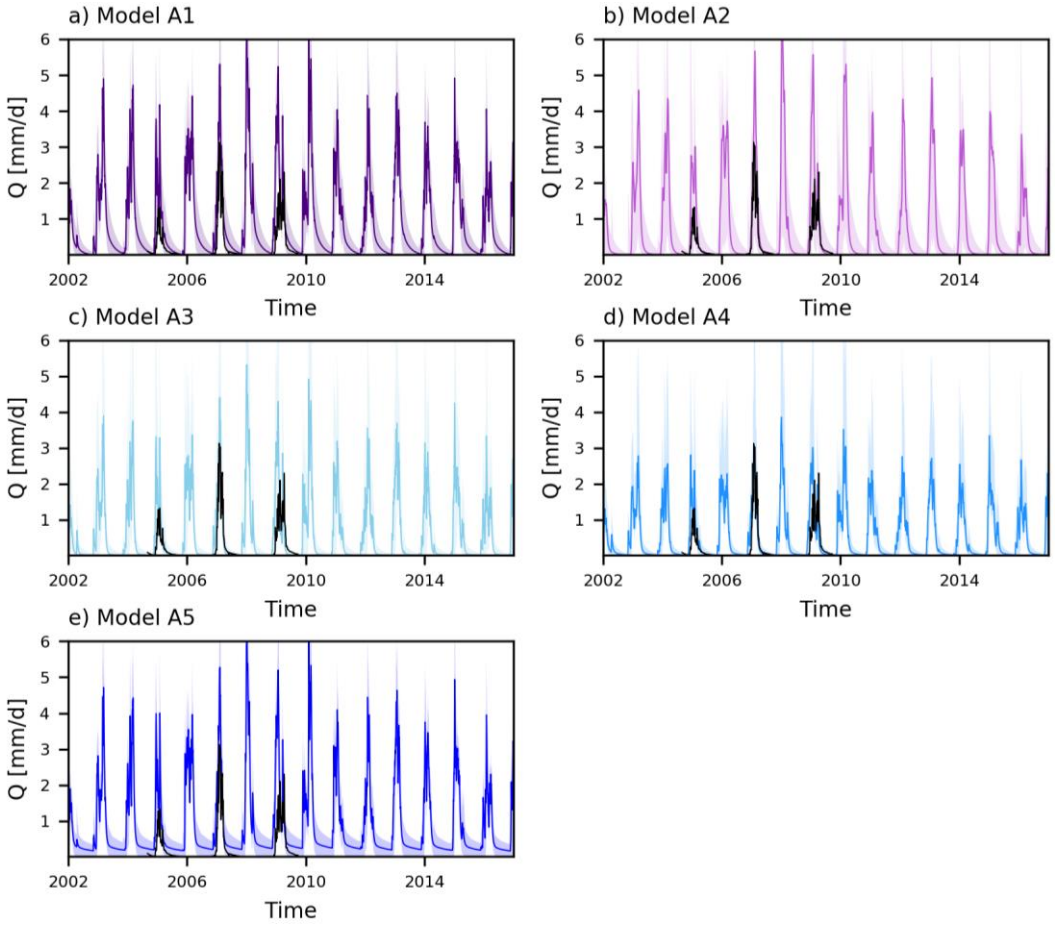


Figure 11. Range of model solutions for Models A1 – A5 for calibration strategy 2 with respect to discharge (hydrograph). The black line indicates the recorded data, the coloured line the solution with the highest calibration objective function with respect to discharge and total water storage ($D_{E,QS}$) and the shaded area the envelope of the solutions retained as feasible.

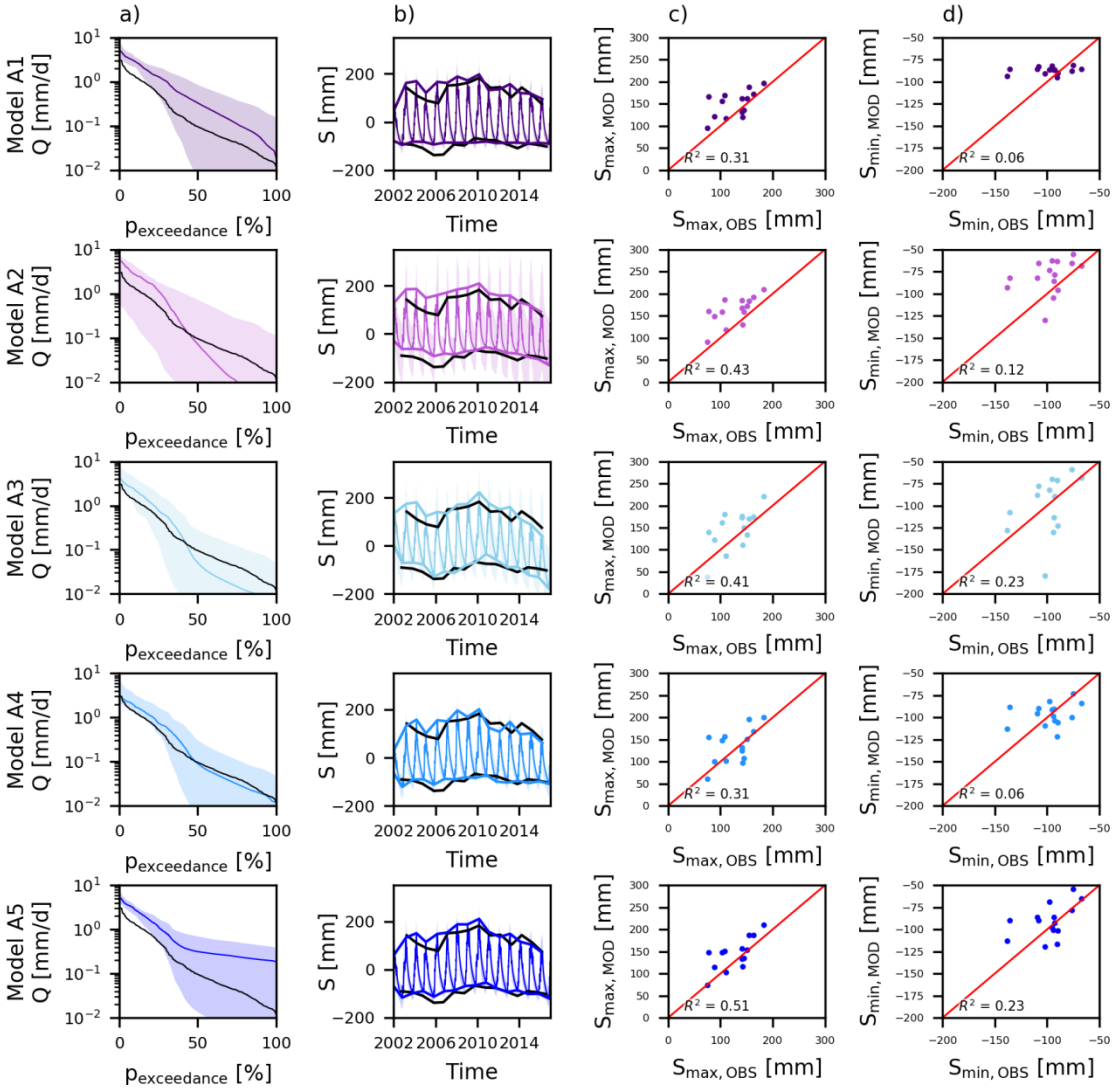


Figure 12. Range of model solutions for Models A1 – A5 for calibration strategy 2 with respect to a) flow duration curve, b) total water storage time-series, c) annual maximum total water storage, and d) annual minimum total water storage. In a) – b), the black line indicates the recorded data, the coloured line the solution with the highest calibration objective function with respect to discharge and total water storage ($D_{E,QS}$) and the shaded area the envelope of the solutions retained as feasible. In c) – d), the recorded data are plotted on the horizontal axis and on the vertical axis the model solution with the highest calibration objective function with respect to discharge and total water storage ($D_{E,QS}$). The red line indicates the 1:1 line.

5.2.4 Third model adaptation: Alternative forcing data and model structure

According to the first model adaptation (comparing Models A0 – D0), Model D0 performed the best using precipitation data according to TRMM and estimating the potential evaporation with the Thornthwaite method. According to the second model adaptation (comparing Models A0 – A5), Model A4 performed the best featuring a Deeper Groundwater reservoir which was only recharged during the wet season and from where groundwater leaked out of the basin (Figures 2 and 3). In this section, both models D0 and A4 were combined into Model D4 where we used TRMM as data source for precipitation, the Thornthwaite method to estimate potential evaporation and the model structure associated with Model A4.

Following the first calibration strategy, this model reproduced the discharge well (Figure S13a in the Supplementary Material) with $D_{E,Q}$ = 0.93 which was better than all other alternative model hypotheses (Table 6, Figure 8a). Both, the general flow dynamics and magnitudes were captured well with this model (Figure S13a, b in the Supplementary Material). The monthly and annual total water storage improved significantly to $D_{E,S}$ = 0.31 (Table 6, Figure 8a). The modelled monthly storage variations were broadly consistent with the observation (Figure S14 in the Supplementary Material), albeit with differences in the decrease during dry seasons and with high parameter uncertainty. The magnitudes of the annual minimum and maximum storage were modelled well for the time period 2010 – 2016, whereas before 2010 the storage was overestimated (Figure S13c in the Supplementary Material). Also the fluctuations in the annual maximum storage were modelled well with R^2 = 0.48 (Figure S13d in the Supplementary Material), but the annual minima were captured poorly (R^2 = 0.19, Figure S13e in the Supplementary Material). The overall model performance increased to $D_{E,QS}$ = 0.51 which was better than all other alternative model hypotheses (Table 6, Figure 8a).

Following the second calibration strategy, the discharge was modelled well (Figure 13a), albeit with a slight decrease in the model performance ($D_{E,Q}$ = 0.85) compared to the first calibration strategy (Table 6, Figure 8b). While the flow dynamics were captured well (Figure 13a), low flows were slightly underestimated (Figure 13b). The monthly and annual total water storage time-series improved considerably to $D_{E,S}$ = 0.50 (Table 6, Figure 8b). With this model and this calibration strategy, the monthly variations were captured well (Figure S15 in the Supplementary Material), as also magnitudes and fluctuations in the annual maxima (R^2 = 0.57, Figure 13c,d) and minima (R^2 = 0.41, Figure 13c,e). The overall model performance increased to $D_{E,QS}$ = 0.63 which was better than all other alternative model hypotheses (Table 6, Figure 8b).

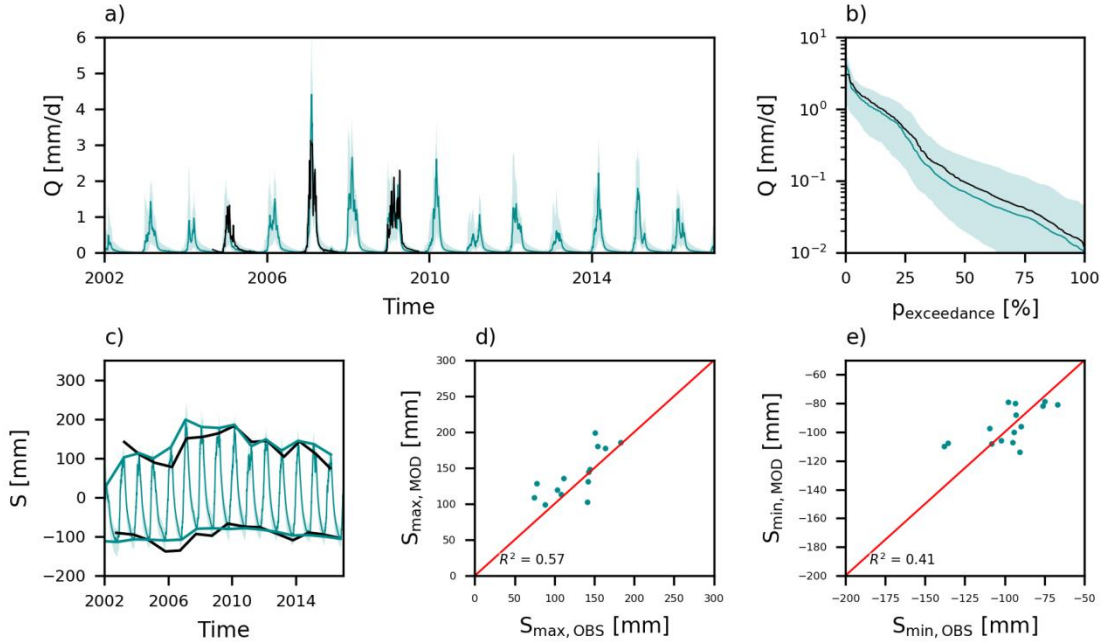


Figure 13. Range of model solutions for Model D4 for calibration strategy 2 with respect to a) hydrograph, b) flow duration curve, c) total water storage time-series, d) annual maximum total water storage, and e) annual minimum total water storage. In a) to c), the black line indicates the recorded data, the coloured line the solution with the highest calibration objective function with respect to discharge and total water storage ($D_{E,QS}$) and the shaded area the envelope of the solutions retained as feasible. In d) and e), the recorded data are plotted on the horizontal axis and on the vertical axis the model solution with the highest calibration objective function with respect to discharge and total water storage ($D_{E,QS}$). The red line indicates the 1:1 line.

In a last step, the predictive strength of Model D4 was compared to that of the benchmark Model A0. For this purpose, both models were calibrated with respect to discharge and total water storage simultaneously (calibration strategy 2) for the time period 2002 – 2012, and post-calibration evaluated due to the lack of flow data only with respect to total water storage for the time period 2012 – 2016 (see Section 4.4). While the general flow dynamics were modelled well for both models (Figure S16 in the Supplementary Material), the magnitudes improved significantly for Model D4 as the flows were only slightly underestimated during medium flows (Figure 14a). Hence, the modelled flow improved from $D_{E,Q}$ = -0.13 for Model A0 to $D_{E,Q}$ = 0.51 for Model D4 (Table 7). Also the monthly and annual total water storage time-series improved for Model D4 to $D_{E,S}$ = 0.63. On monthly timescale, Model D4 captured the seasonal variations better with considerable improvements in the storage decrease during dry seasons (Figure S17 in the Supplementary Material). While the magnitudes of the annual minima/maxima were captured well for both models (Figure 14b), long-term fluctuations improved for Model D4 with respect to the annual maxima (R^2 = 0.57, Figure 14c) and

minima (R^2 = 0.44, Figure 14d) where R^2 corresponded to the calibration time-period 2002 – 2012 as merely four to five points were available for the evaluation time-period 2012 – 2016. With Model D4, the annual minimum and maximum storage increased before 2010 after which it decreased similar to the observations and in contrast to the benchmark Model A0. However, the annual minimum/maximum storage were frequently overestimated except in 2002 – 2004 when it was underestimated. During the evaluation time-period 2012 – 2016, the model performance with respect to the monthly and annual total water storage improved to $D_{E,S}$ = -1.0 (Table 7) which remained negative due to the low model performance metrics with respect to the annual minima/maxima ($E_{NS,S,annual}$, Section 4.3). In this short time-period, the difference between the observed time-series and its mean was significantly lower compared to a longer time-period such as 2002 – 2012 resulting in a low denominator and hence a low Nash-Sutcliffe efficiency (Eq.39).

Overall, the results suggest that the model's ability to simultaneously reproduce both the observed discharge *and* long-term and seasonal total water storage variations was considerably influenced by both, the choice of forcing data and model structure, respectively. Overall, the combination of TRMM data for precipitation, the Thornthwaite method for potential evaporation and the model structure associated with Model A4 here produced model results most consistent with the observed total water storage and discharge time-series. This Model D4 allowed for a better representation of the discharge and better prediction of the total water storage with respect to the seasonal and long-term fluctuations. The forcing data mostly controlled the model's ability to mimic annual storage maxima, whereas the annual storage minima improved the most when incorporating groundwater loss from the Deeper Groundwater reservoir (Model A4 and D4).

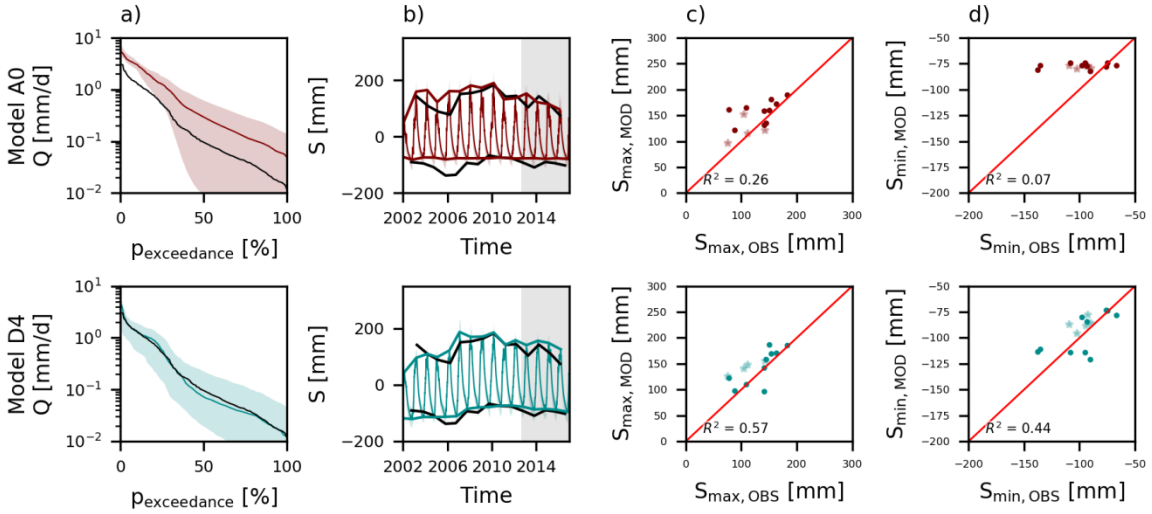


Figure 14. Range of model solutions for Models A0 and D4 for calibration strategy 2 with respect to a) flow duration curve, b) total water storage time-series, c) annual maximum total water storage, and d) annual minimum total water storage. In a) – b), the black line indicates the recorded data, the coloured line the solution with the highest calibration objective function with respect to discharge and total water storage ($D_{E,QS}$) and the shaded area the envelope of the solutions retained as feasible. The white area was used for calibration (2002 – 2012) and the grey area for evaluation (2012 – 2016). In c) – d), the recorded data are plotted on the horizontal axis and on the vertical axis the model solution with the highest calibration objective function with respect to discharge and total water storage ($D_{E,QS}$). The darker dots correspond to the 2002 – 2012 time-period and was used to calculate R^2, whereas the lighter stars correspond to the 2012 – 2016 time-period. The red line indicates the 1:1 line.

Table 7. Model performance with respect to total water storage and discharge ($D_{E,QS}$), and total water storage ($D_{E,S}$) including their 5/95% percentile ranges of the feasible parameter sets for Models A0 and D4 calibrated with respect to $D_{E,QS}$ for the time period 2002 – 2012.

	2002 – 2012			**2012 – 2016**
	$\boldsymbol{D_{E,QS}}$ ($\boldsymbol{D_{E,QS,5/95\%}}$)	$\boldsymbol{D_{E,Q}}$ ($\boldsymbol{D_{E,Q,5/95\%}}$)	$\boldsymbol{D_{E,S}}$ ($\boldsymbol{D_{E,S,5/95\%}}$)	$\boldsymbol{D_{E,S}}$ ($\boldsymbol{D_{E,S,5/95\%}}$)
Model A0	-0.29 (-0.71 – -0.10)	-0.13 (-0.76 – -0.11)	-0.21 (-0.51 – -0.33)	-2.7 (-6.2 – -0.70)
Model D4	0.83 (0.62 – 0.89)	0.51 (0.08 – 0.37)	0.63 (0.33 – 0.53)	-1.0 (-3.3 – 0.43)

6 Discussion

In this study, we identified plausible drivers for the observed long-term total water storage variations in the Luangwa Basin. The results indicated modelled annual maximum storage

fluctuations were to a large extent controlled by the choice of forcing data, whereas modelled annual minima were influenced by processes allowing long-term memory effects which were missing in the original benchmark Model A0. More specifically, the representation of monthly and annual total water storage fluctuations improved when using TRMM for the precipitation, the Thornthwaite method to estimate potential evaporation and incorporating groundwater loss from a deeper groundwater layer (Model D4).

The results demonstrated that models that can adequately reproduce discharge do not necessarily reproduce storage well which was also observed by Bouaziz et al. (2020). In this study, the benchmark Model A0 reproduced the general dynamics and magnitudes of the discharge well but did not reproduce the observed storage magnitudes nor the long-term storage fluctuations. Incorporating the total water storage in the calibration procedure only improved the modelled storage magnitudes, but not the long-term fluctuations. While alternative forcing data sources improved the representation of the annual maximum storage fluctuations, the storage conditions during dry seasons, i.e. annual minima, remained poorly represented (Models A0 – D0) and only improved after modifying the model structure (Model D4). These results suggested that groundwater loss from the Luangwa basin played an important role to explain long-term annual storage variations. However, in many commonly used hydrological models such processes allowing long-term memory effects are missing (e.g. Bergström, 1992;Liang et al., 1994;Fenicia et al., 2014) resulting in biased predictions of discharge and storage which is especially crucial during extreme dry conditions (Saft et al., 2016;Fowler et al., 2020).

Furthermore, this study showed that processes allowing for long-term memory effects can be incorporated in conceptual hydrological models. In this study, several model hypotheses were tested to assess which processes most likely dominated long-term memory effects in the Luangwa basin (Models A1 – A5). The results suggested long-term storage variations were a result of groundwater loss from a deeper groundwater layer which was only recharged during wet seasons (Model D4). With this model, the storage prediction substantially improved compared to the benchmark Model A0, yet remained at a modest level ($D_{E,S} < 0$, Table 7) most likely due to the chosen model performance metric and the limited number of data points for the evaluation when considering annual minima/maxima for the time-period 2012 – 2016 as explained in the previous section. In addition, these modifications also improved the modelled discharge time-series such that the general dynamics and magnitudes were represented better with Model D4 (Figure 13) compared to the benchmark Model A0 (Figure 7). Therefore, model hypothesis testing played a crucial role in improving the representation of real world processes to reproduce multiple variables simultaneously (Clark et al., 2011;Beven, 2018).

Previous studies highlighted the inability of many conceptual models to reproduce long-term storage variations and attributed this to data errors, poor parameterization, model structural

deficiencies or a combination thereof (Winsemius et al., 2006;Saft et al., 2016;Fowler et al., 2018;Scanlon et al., 2018;Jing et al., 2019). Fowler et al. (2020) recently demonstrated that commonly used conceptual hydrological models cannot reproduce long-term storage variations as they lack long-term memory processes and hence should not be used for discharge predictions in for example drying climates. However, here we could show that following a careful data and model selection procedure, the representation of long-term storage variations in a conceptual model could be considerably improved. This further implies that although many typical implementations of hydrological models indeed cannot reproduce long-term storage changes, in particular with respect to annual fluctuations in dry season conditions, i.e. annual minima, as shown by Fowler et al. (2020) and here with Models A0 – D0, this inability is not an inherent property of conceptual models. Instead, our results provide some evidence that this inability can, at least to some degree, be overcome when adopting a systematic procedure to test alternative model hypotheses and thus to improve the representation of real world processes (here: Models A1 – A5).

For future studies, it will be interesting to explore the effects of evaporation on long-term storage fluctuations in a more detailed analysis. Our results suggest that long-term fluctuations in the potential evaporation can occur depending on the chosen estimation method (Roderick and Farquhar, 2005;Hobbins et al., 2008;Huang et al., 2015;Xu et al., 2018). It would therefore be interesting to look into alternative, potentially more accurate estimation methods. In addition, long-term fluctuations in the actual evaporation were observed depending on the satellite product due to the different underlying assumptions and input data (Goroshi et al., 2017;Wang et al., 2018;Bai et al., 2019;Feng et al., 2019). That is why, more in-depth analyses on the effect of evaporation on long-term storage fluctuations is recommended which was outside the scope of this study due to the limited data availability.

7 Conclusion

In the Luangwa basin, long-term total water storage variations were observed with GRACE, but not reproduced by the existing process-based hydrological model that encapsulates our current understanding of the dominant regional hydrological processes. The objective of this paper was to identify so far overlooked processes underlying these low-frequency variations in a combined data analysis and model hypothesis testing approach. Overall, the results suggest that the initial model's inability to reproduce the observed low-frequency storage variations was a combined effect of the data source used to run the model and the missing representation of regional groundwater export. More specifically, it was shown that a different choice of the model input data source produced model results that are more consistent with observed fluctuations in long-term annual total water storage maxima. In contrast, the incorporation of a process representing regional groundwater export from a deep

groundwater layer improved the model's ability to reproduce the observed variations in the annual minimum storage. The results highlighted the combined value of alternative data sources and iterative hypothesis testing to improve our understanding of hydrological processes, their quantitative description in models and eventually towards more reliable predictions of hydrological models.

Acknowledgement

This research is supported by the TU Delft | Global Initiative, a program of the Delft University of Technology to boost Science and Technology for Global Development. Discharge data for the study region were made available by WARMA (Water Resources Management Authority in Zambia) and can be accessed upon request at WARMA. Satellite observations were obtained from publically available online databases as described in Section 3.

Abbreviations

CHIRPS	Climate Hazards Group InfraRed Precipitation with Station data
CRU	Climatic Research Unit
GLEAM	Global Land Evaporation Amsterdam Model
GRACE	Gravity Recovery and Climate Experiment
LAI	Leaf Area Index
MOD16	MODIS Global Evapotranspiration Project
NDVI	Normalized Difference Vegetation Index
SEBS	Surface Energy Balance System
SSEBop	operational Simplified Surface Energy Balance
TRMM	Tropical Rainfall Measuring Mission
WaPOR	Water Productivity Open-access portal
WARMA	Zambian Water Resources Management Authority

References

Allen, R. G., Tasumi, M., and Trezza, R.: Satellite-Based Energy Balance for Mapping Evapotranspiration with Internalized Calibration (METRIC)—Model, Journal of Irrigation and Drainage Engineering, 133, 380-394, https://doi.org/10.1061/(ASCE)0733-9437(2007)133:4(380), 2007.

Awange, J. L., Khandu, Schumacher, M., Forootan, E., and Heck, B.: Exploring hydro-meteorological drought patterns over the Greater Horn of Africa (1979–2014) using remote sensing and reanalysis products, Advances in Water Resources, 94, 45-59, https://doi.org/10.1016/j.advwatres.2016.04.005, 2016.

Bai, M., Shen, B., Song, X., Mo, S., Huang, L., and Quan, Q.: Multi-Temporal Variabilities of Evapotranspiration Rates and Their Associations with Climate Change and Vegetation Greening in the Gan River Basin, China, Water, 11, 2568, https://doi.org/10.3390/w11122568, 2019.

Bastiaanssen, W. G. M., Menenti, M., Feddes, R. A., and Holtslag, A. A. M.: A remote sensing surface energy balance algorithm for land (SEBAL). 1. Formulation, Journal of Hydrology, 212–213, 198-212, http://doi.org/10.1016/S0022-1694(98)00253-4, 1998.

Bergström, S.: The HBV model – its structure and applications, in, SMHI Norrköping, Sweden, 32, 1992.

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, 2006.

Beven, K. J.: On hypothesis testing in hydrology: Why falsification of models is still a really good idea, WIREs Water, 5, e1278, https://doi.org/10.1002/wat2.1278, 2018.

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.

Bonsor, H. C., Shamsudduha, M., Marchant, B. P., MacDonald, A. M., and Taylor, R. G.: Seasonal and Decadal Groundwater Changes in African Sedimentary Aquifers Estimated Using GRACE Products and LSMs, Remote Sensing, 10, 904, https://doi.org/10.3390/rs10060904, 2018.

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.

Bouaziz, L. J. E., Steele-Dunne, S. C., Schellekens, J., Weerts, A. H., Stam, J., Sprokkereef, E., Winsemius, H. H. C., Savenije, H. H. G., and Hrachowitz, M.: Improved Understanding of the Link Between Catchment-Scale Vegetation Accessible Storage and Satellite-Derived Soil Water Index, Water Resources Research, 56, e2019WR026365, https://doi.org/10.1029/2019WR026365, 2020.

Boutt, D. F.: Assessing hydrogeologic controls on dynamic groundwater storage using long-term instrumental records of water table levels, Hydrological Processes, 31, 1479-1497, https://doi.org/10.1002/hyp.11119, 2017.

Budyko, M. I.: Climate and Life, Academic Press, New York, 508 pp., 1974.

Burnash, R. J. C., Ferral, R. L., and McGuire, R. A.: A generalized streamflow simulation system: conceptual modeling for digital computers, in, US Department of Commerce, National Weather Service and State of California, Department of Water Resources, California, 1973.

Chao, N., Wang, Z., Jiang, W., and Chao, D.: A quantitative approach for hydrological drought characterization in southwestern China using GRACE, Hydrogeology Journal, 24, 893-903, https://doi.org/10.1007/s10040-015-1362-y, 2016.

Chen, J. L., Wilson, C. R., Tapley, B. D., Longuevergne, L., Yang, Z. L., and Scanlon, B. R.: Recent La Plata basin drought conditions observed by satellite gravimetry, Journal of Geophysical Research: Atmospheres, 115, D22108, https://doi.org/10.1029/2010JD014689, 2010.

Chen, J. L., Wilson, C. R., Tapley, B. D., Scanlon, B., and Güntner, A.: Long-term groundwater storage change in Victoria, Australia from satellite gravity and in situ observations, Global and Planetary Change, 139, 56-65, https://doi.org/10.1016/j.gloplacha.2016.01.002, 2016.

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.

Claverie, M., Vermote, E., and NOAA CDR Program: NOAA Climate Data Record (CDR) of Leaf Area Index (LAI) and Fraction of Absorbed Photosynthetically Active Radiation (FAPAR), Version 5, in, NOAA National Climatic Data Center, 2014.

Cohen Liechti, T., Matos, J. P., Boillat, J. L., and Schleiss, A. J.: Comparison and evaluation of satellite derived precipitation products for hydrological modeling of the Zambezi River Basin, Hydrol. Earth Syst. Sci., 16, 489-500, https://doi.org/10.5194/hess-16-489-2012, 2012.

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.

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.

Feng, T., Su, T., Zhi, R., Tu, G., and Ji, F.: Assessment of actual evapotranspiration variability over global land derived from seven reanalysis datasets, International Journal of Climatology, 39, 2919-2932, https://doi.org/10.1002/joc.5992, 2019.

Fenicia, F., Kavetski, D., Savenije, H. H. G., Clark, M. P., Schoups, G., Pfister, L., and Freer, J.: Catchment properties, function, and conceptual model representation: is there a correspondence?, Hydrological Processes, 28, 2451-2467, https://doi.org/10.1002/hyp.9726, 2014.

Fowler, K., Coxon, G., Freer, J., Peel, M., Wagener, T., Western, A., Woods, R., and Zhang, L.: Simulating Runoff Under Changing Climatic Conditions: A Framework for Model Improvement, Water Resources Research, 54, 9812-9832, https://doi.org/10.1029/2018WR023989, 2018.

Fowler, K., Knoben, W., Peel, M., Peterson, T., Ryu, D., Saft, M., Seo, K.-W., and Western, A.: Many Commonly Used Rainfall-Runoff Models Lack Long, Slow Dynamics: Implications for Runoff Projections, Water Resources Research, 56, e2019WR025286, https://doi.org/10.1029/2019WR025286, 2020.

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.

Gallart, F., and Llorens, P.: Catchment Management under Environmental Change: Impact of Land Cover Change on Water Resources, Water International, 28, 334-340, https://doi.org/10.1080/02508060308691707, 2003.

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.

Gerrits, A. M. J., Savenije, H. H. G., Veling, E. J. M., and Pfister, L.: Analytical derivation of the Budyko curve based on rainfall characteristics and a simple evaporation model, Water Resources Research, 45, W04403, https://doi.org/10.1029/2008WR007308, 2009.

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.

Goroshi, S., Pradhan, R., Singh, R. P., Singh, K. K., and Parihar, J. S.: Trend analysis of evapotranspiration over India: Observed from long-term satellite measurements, Journal of Earth System Science, 126, 113, https://doi.org/10.1007/s12040-017-0891-2, 2017.

Goswami, M., O'Connor, K. M., and Bhattarai, K. P.: Development of regionalisation procedures using a multi-model approach for flow simulation in an ungauged catchment, Journal of Hydrology, 333, 517-531, https://doi.org/10.1016/j.jhydrol.2006.09.018, 2007.

Grigg, A. H., and Hughes, J. D.: Nonstationarity driven by multidecadal change in catchment groundwater storage: A test of modifications to a common rainfall–run-off model, Hydrological Processes, 32, 3675-3688, https://doi.org/10.1002/hyp.13282, 2018.

Handavu, F., Chirwa, P. W. C., and Syampungani, S.: Socio-economic factors influencing land-use and land-cover changes in the miombo woodlands of the Copperbelt province in Zambia, Forest Policy and Economics, 100, 75-94, https://doi.org/10.1016/j.forpol.2018.10.010, 2019.

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.

Hobbins, M. T., Dai, A., Roderick, M. L., and Farquhar, G. D.: Revisiting the parameterization of potential evaporation as a driver of long-term water balance trends, Geophysical Research Letters, 35, L12403, https://doi.org/10.1029/2008GL033840, 2008.

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., Stockinger, M., Coenders-Gerrits, M., van der Ent, R., Bogena, H., Lücke, A., and Stumpp, C.: Deforestation reduces the vegetation-accessible water storage in the unsaturated soil and affects catchment travel time distributions and young water fractions, Hydrol. Earth Syst. Sci. Discuss., 2020, 1-43, https://doi.org/10.5194/hess-2020-293, 2020.

Huang, H., Han, Y., Cao, M., Song, J., Xiao, H., and Cheng, W.: Spatiotemporal Characteristics of Evapotranspiration Paradox and Impact Factors in China in the Period of 1960–2013, Advances in Meteorology, 2015, 519207, https://doi.org/10.1155/2015/519207, 2015.

Huffman, G. J., Adler, R. F., Rudolf, B., Schneider, U., and Keehn, P. R.: Global Precipitation Estimates Based on a Technique for Combining Satellite-Based Estimates, Rain Gauge Analysis, and NWP Model Precipitation Information, Journal of Climate, 8, 1284-1295, https://doi.org/10.1175/1520-0442(1995)008<1284:GPEBOA>2.0.CO;2, 1995.

Huffman, G. J., Bolvin, D. T., Nelkin, E. J., Wolff, D. B., Adler, R. F., Gu, G., Hong, Y., Bowman, K. P., and Stocker, E. F.: The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-Global, Multiyear, Combined-Sensor Precipitation Estimates at Fine Scales, Journal of Hydrometeorology, 8, 38-55, https://doi.org/10.1175/JHM560.1, 2007.

Huffman, G. J., Stocker, E. F., Bolvin, D. T., and Nelkin, E. J.: TRMM 3B43V7 Data Sets, in, Goddard Earth Sciences Data and Information Services Center (GES DISC), Greenbelt, MD, USA, 2014.

Hulsman, P., Savenije, H. H. G., and Hrachowitz, M.: Learning from satellite observations: increased understanding of catchment processes through stepwise model improvement, Hydrol. Earth Syst. Sci. Discuss., 2020, 1-26, https://doi.org/10.5194/hess-2020-191, 2020a.

Hulsman, P., Winsemius, H. C., Michailovsky, C. I., Savenije, H. H. G., and Hrachowitz, M.: Using altimetry observations combined with GRACE to select parameter sets of a hydrological model in a data-scarce region, Hydrol. Earth Syst. Sci., 24, 3331-3359, https://doi.org/10.5194/hess-24-3331-2020, 2020b.

Jing, W., Yao, L., Zhao, X., Zhang, P., Liu, Y., Xia, X., Song, J., Yang, J., Li, Y., and Zhou, C.: Understanding Terrestrial Water Storage Declining Trends in the Yellow River Basin, Journal of Geophysical Research: Atmospheres, 124, 12963-12984, https://doi.org/10.1029/2019JD031432, 2019.

Joodaki, G., Wahr, J., and Swenson, S.: Estimating the human contribution to groundwater depletion in the Middle East, from GRACE data, land surface models, and well observations, Water Resources Research, 50, 2679-2692, https://doi.org/10.1002/2013WR014633, 2014.

Khaki, M., Forootan, E., Kuhn, M., Awange, J., van Dijk, A. I. J. M., Schumacher, M., and Sharifi, M. A.: Determining water storage depletion within Iran by assimilating GRACE data into the W3RA hydrological model, Advances in Water Resources, 114, 1-18, https://doi.org/10.1016/j.advwatres.2018.02.008, 2018.

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.

Le Moine, N., Andréassian, V., Perrin, C., and Michel, C.: How can rainfall-runoff models handle intercatchment groundwater flows? Theoretical study based on 1040 French catchments, Water Resources Research, 43, W06428, https://doi.org/10.1029/2006WR005608, 2007.

Leblanc, M. J., Tregoning, P., Ramillien, G., Tweed, S. O., and Fakes, A.: Basin-scale, integrated observations of the early 21st century multiyear drought in southeast Australia, Water Resources Research, 45, W04408, https://doi.org/10.1029/2008WR007333, 2009.

Li, C., Wu, P. T., Li, X. L., Zhou, T. W., Sun, S. K., Wang, Y. B., Luan, X. B., and Yu, X.: Spatial and temporal evolution of climatic factors and its impacts on potential evapotranspiration in Loess Plateau of Northern Shaanxi, China, Science of The Total Environment, 589, 165-172, https://doi.org/10.1016/j.scitotenv.2017.02.122, 2017.

Liang, X., Lettenmaier, D. P., Wood, E. F., and Burges, S. J.: A simple hydrologically based model of land surface water and energy fluxes for general circulation models, Journal of Geophysical Research: Atmospheres, 99, 14415-14428, https://doi.org/10.1029/94JD00483, 1994.

Long, D., Scanlon, B. R., Longuevergne, L., Sun, A. Y., Fernando, D. N., and Save, H.: GRACE satellite monitoring of large depletion in water storage in response to the 2011 drought in Texas, Geophysical Research Letters, 40, 3395-3401, https://doi.org/10.1002/grl.50655, 2013.

Long, D., Longuevergne, L., and Scanlon, B. R.: Uncertainty in evapotranspiration from land surface modeling, remote sensing, and GRACE satellites, Water Resources Research, 50, 1131-1151, https://doi.org/10.1002/2013WR014581, 2014.

Long, D., Pan, Y., Zhou, J., Chen, Y., Hou, X., Hong, Y., Scanlon, B. R., and Longuevergne, L.: Global analysis of spatiotemporal variability in merged total water storage changes using multiple GRACE products and global hydrological models, Remote Sensing of Environment, 192, 198-216, https://doi.org/10.1016/j.rse.2017.02.011, 2017.

Maes, W. H., Gentine, P., Verhoest, N. E. C., and Miralles, D. G.: Potential evaporation at eddy-covariance sites across the globe, Hydrol. Earth Syst. Sci., 23, 925-948, https://doi.org/10.5194/hess-23-925-2019, 2019.

Martens, B., Miralles, D. G., Lievens, H., van der Schalie, R., de Jeu, R. A. M., Fernández-Prieto, D., Beck, H. E., Dorigo, W. A., and Verhoest, N. E. C.: GLEAM v3: satellite-based land evaporation and root-zone soil moisture, Geosci. Model Dev., 10, 1903-1925, https://doi.org/10.5194/gmd-10-1903-2017, 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.

Meng, F., Su, F., Li, Y., and Tong, K.: Changes in Terrestrial Water Storage During 2003–2014 and Possible Causes in Tibetan Plateau, Journal of Geophysical Research: Atmospheres, 124, 2909-2931, https://doi.org/10.1029/2018JD029552, 2019.

Miralles, D. G., Holmes, T. R. H., De Jeu, R. A. M., Gash, J. H., Meesters, A. G. C. A., and Dolman, A. J.: Global land-surface evaporation estimated from satellite-based observations, Hydrol. Earth Syst. Sci., 15, 453-469, https://doi.org/10.5194/hess-15-453-2011, 2011.

Nelson, S. T., and Mayo, A. L.: The role of interbasin groundwater transfers in geologically complex terranes, demonstrated by the Great Basin in the western United States, Hydrogeology Journal, 22, 807-828, https://doi.org/10.1007/s10040-014-1104-6, 2014.

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.

Oguntunde, P. G., Friesen, J., van de Giesen, N., and Savenije, H. H. G.: Hydroclimatology of the Volta River Basin in West Africa: Trends and variability from 1901 to 2002, Physics and Chemistry of the Earth, Parts A/B/C, 31, 1180-1188, https://doi.org/10.1016/j.pce.2006.02.062, 2006.

Ol'dekop, E. M.: On evaporation from the surface of river basins, Transactions on Meteorological Observations, 4, 1911.

Pellicer-Martínez, F., and Martínez-Paz, J. M.: Assessment of interbasin groundwater flows between catchments using a semi-distributed water balance model, Journal of Hydrology, 519, 1848-1858, https://doi.org/10.1016/j.jhydrol.2014.09.067, 2014.

Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation, Journal of Hydrology, 279, 275-289, https://doi.org/10.1016/S0022-1694(03)00225-7, 2003.

Phiri, D., Morgenroth, J., and Xu, C.: Four decades of land cover and forest connectivity study in Zambia—An object-based image analysis approach, International Journal of Applied Earth Observation and Geoinformation, 79, 97-109, https://doi.org/10.1016/j.jag.2019.03.001, 2019a.

Phiri, D., Morgenroth, J., and Xu, C.: Long-term land cover change in Zambia: An assessment of driving factors, Science of The Total Environment, 697, 134206, https://doi.org/10.1016/j.scitotenv.2019.134206, 2019b.

Pike, J. G.: The estimation of annual run-off from meteorological data in a tropical climate, Journal of Hydrology, 2, 116-123, https://doi.org/10.1016/0022-1694(64)90022-8, 1964.

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.

Roderick, M. L., and Farquhar, G. D.: Changes in New Zealand pan evaporation since the 1970s, International Journal of Climatology, 25, 2031-2039, https://doi.org/10.1002/joc.1262, 2005.

Running, S., Mu, Q., and Zhao, M.: MOD16A2 MODIS/Terra Net Evapotranspiration 8-Day L4 Global 500m SIN Grid V006, in, NASA EOSDIS Land Processes DAAC, 2017.

Saft, M., Peel, M. C., Western, A. W., Perraud, J.-M., and Zhang, L.: Bias in streamflow projections due to climate-induced shifts in catchment response, Geophysical Research Letters, 43, 1574-1581, https://doi.org/10.1002/2015GL067326, 2016.

Samaniego, L., Kumar, R., and Jackisch, C.: Predictions in a data-sparse region using a regionalized grid-based hydrologic model driven by remotely sensed data, Hydrology Research, 42, 338-355, https://doi.org/10.2166/nh.2011.156, 2011.

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.

Scanlon, B. R., Zhang, Z., Save, H., Sun, A. Y., Müller Schmied, H., van Beek, L. P. H., Wiese, D. N., Wada, Y., Long, D., Reedy, R. C., Longuevergne, L., Döll, P., and Bierkens, M. F. P.: Global models underestimate large decadal declining and rising water storage trends relative to GRACE satellite data, Proceedings of the National Academy of Sciences, 115, E1080, https://doi.org/10.1073/pnas.1704665115, 2018.

Schreiber, P.: Über die Beziehungen zwischen dem Niederschlag und der Wasserführung der Flüsse in Mitteleuropa, Z. Meteorol, 21, 441-452, 1904.

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

Schwatke, C., Dettmering, D., Bosch, W., and Seitz, F.: DAHITI – an innovative approach for estimating water level time series over inland waters using multi-mission satellite altimetry, Hydrol. Earth Syst. Sci., 19, 4345-4364, https://doi.org/10.5194/hess-19-4345-2015, 2015.

Senay, G. B., Budde, M., Verdin, J. P., and Melesse, A. M.: A Coupled Remote Sensing and Simplified Surface Energy Balance Approach to Estimate Actual Evapotranspiration from Irrigated Fields, Sensors (Basel), 7, 979-1000, https://doi.org/10.3390/s7060979 2007.

Su, Z.: The Surface Energy Balance System (SEBS) for estimation of turbulent heat fluxes, Hydrol. Earth Syst. Sci., 6, 85-100, https://doi.org/10.5194/hess-6-85-2002, 2002.

Sun, Z., Zhu, X., Pan, Y., Zhang, J., and Liu, X.: Drought evaluation using the GRACE terrestrial water storage deficit over the Yangtze River Basin, China, Science of The Total Environment, 634, 727-738, https://doi.org/10.1016/j.scitotenv.2018.03.292, 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.

Tangdamrongsub, N., Han, S.-C., Tian, S., Müller Schmied, H., Sutanudjaja, E. H., Ran, J., and Feng, W.: Evaluation of Groundwater Storage Variations Estimated from GRACE Data Assimilation and State-of-the-Art Land Surface Models in Australia and the North China Plain, Remote Sensing, 10, 483, https://doi.org/10.3390/rs10030483, 2018.

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.

Thiemig, V., Rojas, R., Zambrano-Bigiarini, M., Levizzani, V., and De Roo, A.: Validation of satellite-based precipitation products over sparsely Gauged African River basins, Journal of Hydrometeorology, 13, 1760-1783, https://doi.org/10.1175/JHM-D-12-032.1, 2012.

Turc, L.: Le bilan d'eau des sols: relations entre les précipitations, l'évaporation et l'écoulement, in, Institut national de la recherche agronomique, Paris, 1953.

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., Beck, H. E., Crosbie, R. S., de Jeu, R. A. M., Liu, Y. Y., Podger, G. M., Timbal, B., and Viney, N. R.: The Millennium Drought in southeast Australia (2001–2009): Natural and human causes and implications for water resources, ecosystems, economy, and society, Water Resources Research, 49, 1040-1057, https://doi.org/10.1002/wrcr.20123, 2013.

Vishwakarma, D. B., Devaraju, B., and Sneeuw, N.: What Is the Spatial Resolution of grace Satellite Products for Hydrology?, Remote Sensing, 10, 852, https://doi.org/10.3390/rs10060852, 2018.

Wahr, J., Molenaar, M., and Bryan, F.: Time variability of the Earth's gravity field: Hydrological and oceanic effects and their possible detection using GRACE, Journal of Geophysical Research: Solid Earth, 103, 30205-30229, https://doi.org/10.1029/98JB02844, 1998.

Wang, W., Li, J., Yu, Z., Ding, Y., Xing, W., and Lu, W.: Satellite retrieval of actual evapotranspiration in the Tibetan Plateau: Components partitioning, multidecadal trends and dominated factors identifying, Journal of Hydrology, 559, 471-485, https://doi.org/10.1016/j.jhydrol.2018.02.065, 2018.

Warburton, M. L., Schulze, R. E., and Jewitt, G. P. W.: Hydrological impacts of land use change in three diverse South African catchments, Journal of Hydrology, 414–415, 118-135, http://dx.doi.org/10.1016/j.jhydrol.2011.10.028, 2012.

Werth, S., White, D., and Bliss, D. W.: GRACE Detected Rise of Groundwater in the Sahelian Niger River Basin, Journal of Geophysical Research: Solid Earth, 122, 10459-10477, https://doi.org/10.1002/2017JB014845, 2017.

Westerhoff, R. S.: Using uncertainty of Penman and Penman–Monteith methods in combined satellite and ground-based evapotranspiration estimates, Remote Sensing of Environment, 169, 102-112, https://doi.org/10.1016/j.rse.2015.07.021, 2015.

Willems, P.: Parsimonious rainfall–runoff model construction supported by time series processing and validation of hydrological extremes – Part 1: Step-wise model-structure identification and calibration approach, Journal of Hydrology, 510, 578-590, https://doi.org/10.1016/j.jhydrol.2014.01.017, 2014.

Winsemius, H. C., Savenije, H. H. G., van de Giesen, N. C., van den Hurk, B. J. J. M., Zapreeva, E. A., and Klees, R.: Assessment of Gravity Recovery and Climate Experiment (GRACE) temporal signature over the upper Zambezi, Water Resources Research, 42, W12201, https://doi.org/10.1029/2006WR005192, 2006.

Xu, S., Yu, Z., Yang, C., Ji, X., and Zhang, K.: Trends in evapotranspiration and their responses to climate change and vegetation greening over the upper reaches of the Yellow River Basin, Agricultural and Forest Meteorology, 263, 118-129, https://doi.org/10.1016/j.agrformet.2018.08.010, 2018.

Zhang, D., Zhang, Q., Werner, A. D., and Liu, X.: GRACE-Based Hydrological Drought Evaluation of the Yangtze River Basin, China, Journal of Hydrometeorology, 17, 811-828, https://doi.org/10.1175/JHM-D-15-0084.1, 2015a.

Zhang, J., Liu, K., and Wang, M.: Seasonal and Interannual Variations in China's Groundwater Based on GRACE Data and Multisource Hydrological Models, Remote Sensing, 12, 845, https://doi.org/10.3390/rs12050845, 2020.

Zhang, Z., Chao, B. F., Chen, J., and Wilson, C. R.: Terrestrial water storage anomalies of Yangtze River Basin droughts observed by GRACE and connections with ENSO, Global and Planetary Change, 126, 35-45, https://doi.org/10.1016/j.gloplacha.2015.01.002, 2015b.

Zhao, M., A, G., Velicogna, I., and Kimball, J. S.: A Global Gridded Dataset of GRACE Drought Severity Index for 2002–14: Comparison with PDSI and SPEI and a Case Study of the Australia Millennium Drought, Journal of Hydrometeorology, 18, 2117-2129, https://doi.org/10.1175/JHM-D-16-0182.1, 2017a.

Zhao, M., A, G., Velicogna, I., and Kimball, J. S.: Satellite Observations of Regional Drought Severity in the Continental United States Using GRACE-Based Terrestrial Water Storage Changes, Journal of Climate, 30, 6297-6308, https://doi.org/10.1175/JCLI-D-16-0458.1, 2017b.