Improving the Representation of Long-Term Storage Variations With Conceptual Hydrological Models in Data-Scarce Regions

Hulsman, Petra; Hrachowitz, Markus; Savenije, Hubert H.G.

DOI
10.1029/2020WR028837

Publication date
2021

Document Version
Final published version

Published in
Water Resources Research

Citation (APA)
Hulsman, P., Hrachowitz, M., & Savenije, H. H. G. (2021). Improving the Representation of Long-Term Storage Variations With Conceptual Hydrological Models in Data-Scarce Regions. *Water Resources Research*, *57*(4), Article e2020WR028837. https://doi.org/10.1029/2020WR028837

Water Resources Research

RESEARCH ARTICLE
10.1029/2020WR028837

Special Section:
Advancing process representation in hydrologic models: Integrating new concepts, knowledge, and data

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

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:
P. Hulsman,
p.hulsman@tudelft.nl

Citation:
Hulsman, P., Hrachowitz, M., & Savenije, H. H. G. (2021). Improving the representation of long-term storage variations with conceptual hydrological models in data-scarce regions. *Water Resources Research*, *57*, e2020WR028837. https://doi.org/10.1029/2020WR028837

Received 18 SEP 2020
Accepted 1 MAR 2021

Improving the Representation of Long-Term Storage Variations With Conceptual Hydrological Models in Data-Scarce Regions

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

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

Abstract In the Luangwa basin in Zambia, long-term total water storage variations were observed with Gravity Recovery and Climate Experiment, but not reproduced by a standard conceptual hydrological model that encapsulates our current understanding of the dominant regional hydrological processes. The objective of this study was to identify potential processes underlying these low-frequency variations through combined data analysis and model hypothesis testing. First, we analyzed the effect of data uncertainty by contrasting observed storage variations with multi-annual estimates of precipitation and evaporation from multiple data sources. Second, we analyzed 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 analyzed 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 increasing or decreasing storage trends and multi-annual variabilities over ≥10 years. For example, decreasing storage trends were observed in Australia during the Millennium Drought in 1997–2010 (e.g., Chen et al., 2016; Leblanc et al., 2009; Zhao et al., 2017a), whereas both, increasing and decreasing long-term trends as well as multi-annual variabilities were observed in the United States (Boutt, 2017; Long et al., 2013), the La Plata basin in South America (Chen et al., 2010), China (Sun et al., 2018; Z. Zhang, Chao, et al., 2015), and different African river basins (Awange et al., 2016; Bonsor et al., 2018; Werth et al., 2017) which were attributed to rainfall variabilities, glacier melting, droughts or human activities such as groundwater abstractions and land cover changes.

These studies relied on satellite-based Gravity Recovery and Climate Experiment (GRACE) data, with the exception of Boutt (2017) who used in situ data.

However, many hydrological models fail to reproduce such observed long-term storage variations (Fowler et al., 2020; Scanlon et al., 2018; Winsemius et al., 2006). As highlighted by previous studies, these observed long-term storage trends and 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, unsuitable parameter values or any combination thereof (Fowler et al., 2018; Grigg & Hughes, 2018; Jing et al., 2019; Saft et al., 2016). 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 a few notable exceptions (e.g., Bouaziz et al., 2018; Goswami et al., 2007; Hrachowitz et al., 2014; Le Moine et al., 2007; Perrin et al., 2003; Samaniego et al., 2011), processes that could potentially allow long-term memory effects, such as groundwater export (Fowler et al., 2020; Istanbulluoglu et al., 2012), remain mostly unaccounted for in standard formulations of conceptual rainfall-runoff models (Bergström, 1992; Burnash et al., 1973; Euser et al., 2015; Fenicia et al., 2014; Liang et al., 1994; Willems, 2014). This leads to the situation that these models cannot sufficiently well 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 modeled annual minimum storage remained rather constant instead of showing a decreasing trend. The reason for this was that the modeled storage converged to or even reached zero toward 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 modeled 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 availability of high-quality ground observations. That is why in this context satellite observations may play an important role. For example, satellite-based 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 & Swenson, 2012; Swenson, 2012). For example, GRACE observations have been used for groundwater monitoring (Tangdamrongsub et al., 2018; J. Zhang et al., 2020), or drought analysis (Chao et al., 2016; Hulsman et al., 2021; Leblanc et al., 2009; van Dijk et al., 2013; Zhao et al., 2017b; D. Zhang, Zhang, et al., 2015).

While several previous studies focused on identifying long-term storage variations from (satellite-based) observations and potential drivers for these variations as well as on quantifying differences between observations and model results (e.g., Fowler et al., 2020; Jing et al., 2019; Joodaki et al., 2014; Leblanc et al., 2009; Meng et al., 2019; Scanlon et al., 2018), only very few studies attempted to modify a hydrological model to allow for meaningful representations of long-term storage variations. In one exception, Grigg and Hughes (2018) modified the GR4J rainfall-runoff model (Perrin et al., 2003) 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. Some other studies highlighted the value of incorporating groundwater import or export in hydrological models (e.g., Bouaziz et al., 2018; Hrachowitz et al., 2014; Le Moine et al., 2007). For example, Le Moine et al. (2007) analyzed 1040 French catchments and concluded inter-basin groundwater flow should be incorporated explicitly in hydrological models instead of correcting the rainfall or potential evaporation to close the water balance. Bouaziz et al. (2018) illustrated inter-catchment groundwater flow reached on average 10% of the precipitation in the Meuse river basin and should be accounted for in models to prevent overestimating the actual evaporation. However, these studies did not analyze the effect of such a process on the long-term variability in the total water storage. Other studies corrected for poor representations of long-term storage trends by assimilating total water storage anomaly observations according to GRACE into hydrological models (Khaki et al., 2018; Schumacher et al., 2018).

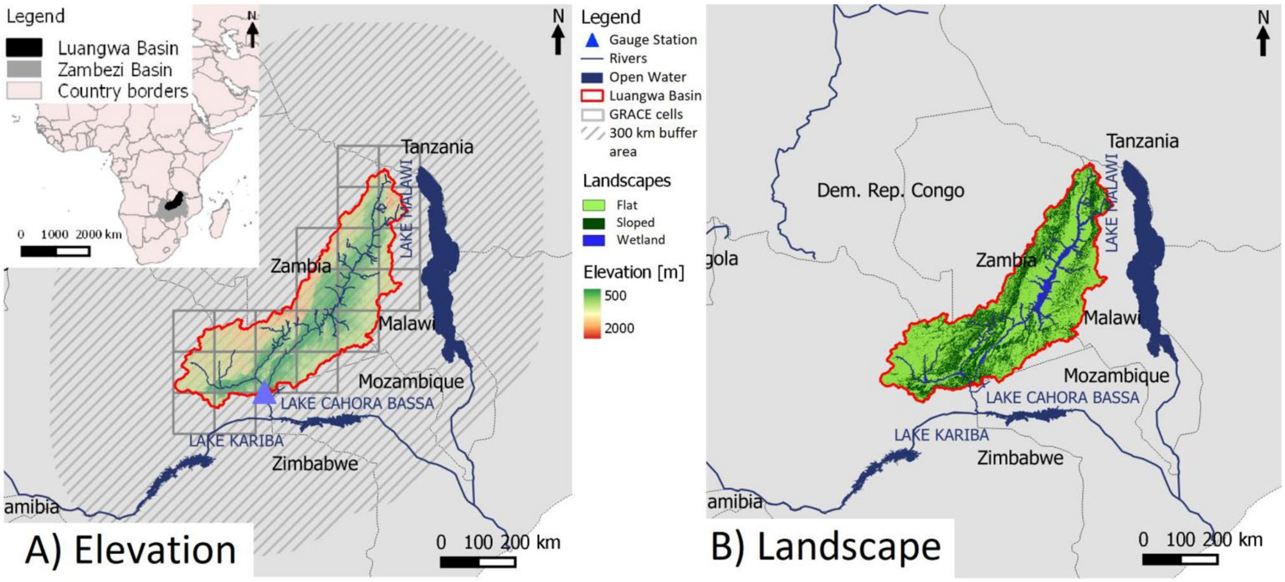

Figure 1. Map of the Luangwa River Basin in Zambia with (a) the elevation and (b) the main landscape types.

In this study, long-term storage variations were observed in the Luangwa river basin, but not reproduced by a distributed implementation of a conceptual model. This model was used in several previous studies to assess the potential of satellite-based altimetry, evaporation and total water storage anomaly observations for stepwise model development and spatial-temporal model calibration (Hulsman, Savenije, et al., 2020; Hulsman, Winsemius, et al., 2020). In these studies, the model successfully reproduced the dynamics of multiple hydrological variables in the Luangwa basin. 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 (Clark et al., 2011). In the spirit of Nearing et al. (2016) and Addor and Melsen (2019), we here more specifically tested the hypotheses that the frequently reported inability of conceptual hydrological models to reproduce observed long-term, low-frequency water storage variations is a result of the combined effects of (1) model forcing data that are incongruent with data of storage variations and (2) oversimplified representation of processes associated with basin-scale groundwater dynamics in such models and that (3) a careful choice of the data source and adaptation of groundwater-related model processes can significantly improve the representation of long-term storage variations in conceptual models.

2. Site Description

The Luangwa River is a 770 km long, mostly unregulated tributary of the Zambezi in Zambia (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 from 400 m up to 1,850 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 (1,555 mm yr^{-1}) exceeds the mean annual precipitation (970 mm yr^{-1}) (Hulsman, Winsemius, et al., 2020; The World Bank, 2010). The lithology is governed by intergranular/fractured siliciclastic sedimentary rocks in the center of the basin and weathered/fractured metamorphic rocks closer to the basin borders (Hartmann & Moosdorf, 2012; IGRAC & UNESCO-IHP, 2015; Ó Dochartaigh, 2019).

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 NDVI (Normalized Difference Vegetation Index). Temperature data from the CRU data set (Climatic Research Unit) Version 4.01 was used to estimate

Table 1
Data Used in This Study

	Time period	Temporal 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 & Gesch, 2011)
Precipitation	1998–2016	Daily	0.05°	CHIRPS	1,127 mm yr^{-1}	Version 2 (Funk et al., 2014)
	1998–2016	Daily	0.25°	TRMM	1,029 mm yr^{-1}	Version 3B42 (Huffman et al., 1995, 2007, 2014)
Evaporation	2009–2016	10 days	250 m	WaPOR	882 mm yr^{-1}	Version 1.1 (FAO, 2018; FAO & 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 (Allen et al., 2007; Bastiaanssen et al., 1998; Senay et al., 2007)
	2003–2016	Monthly	0.25°	GLEAM	751 mm yr^{-1}	Version 3.3b (Martens et al., 2017; Miralles et al., 2011)
	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
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/ (Landerer & Swenson, 2012; Swenson, 2012; Swenson & Wahr, 2006)
Altimetry	2002–2016	10 or 35 days	n/a	DAHITI	n/a	(Schwatke et al., 2015)
Discharge	2002–2016	Daily	n/a	n/a	138 mm yr^{-1}	WARMA

the daily potential evaporation with the Hargreaves (Hargreaves & Allen, 2003; Hargreaves & Samani, 1985) and Thornthwaite (Maes et al., 2019) methods. For this purpose, monthly temperature observations were interpolated to daily timescale using in situ observations at two locations (28° 30′ E, 14° 24′ S and 32° 35′ E, 13° 33′ S).

Processed GRACE observations generated by Centre for Space Research (CSR), GeoForschungsZentrum Potsdam (GFZ), and Jet Propulsion Laboratory (JPL) 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 (Landerer & Swenson, 2012; Swenson & Wahr, 2006; Wahr et al., 1998). 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 bodies, soil moisture and groundwater.

Altimetry data were extracted from the DAHITI website (https://dahiti.dgfi.tum.de/en/, Schwatke et al., 2015) for the Cahora Bassa reservoir, Kariba reservoir and Lake Malawi (Figure 1). In situ daily discharge data was available from the Great East Road Bridge gauging station at the basin outlet (30° 13′ E, 14° 58′ S; Figure 1) and was obtained from the Zambian Water Resources Management Authority (WARMA) in 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 arithmetic 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 and evaluate the model.

4. Approach

This study consisted of three steps. In the first step, we analyzed 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 artifact 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 analyzed, in a second step, four different combinations of data sources, that is, precipitation and potential evaporation, as input for a distributed implementation of a process-based 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 alternative and/or additional process representations, such as regional groundwater export, for their importance to meaningfully reproduce long-memory effects.

In general, long-term total water storage variations are a result of changes in precipitation, evaporation, discharge or any combination thereof (Equation 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 & Llorens, 2003; Hrachowitz et al., 2020; Li et al., 2017; Nijzink et al., 2016; Oguntunde et al., 2006; Saft et al., 2016; Warburton et al., 2012). In addition, long-term storage variations can be a result of slow inter-basin groundwater exchange (Bouaziz et al., 2018; Nelson & Mayo, 2014; Pellicer-Martínez & Martínez-Paz, 2014).

$$\frac{\mathrm{d}S}{\mathrm{d}t} = P - E - Q \tag{1}$$

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

4.1. Data Analysis

Long-term, basin-averaged satellite observations of precipitation from the CHIRPS and TRMM data products, actual evaporation from the WaPOR, SEBS, SSEBop, GLEAM and MOD16 products, potential evaporation according to the Hargreaves (Hargreaves & Allen, 2003; Hargreaves & Samani, 1985) and Thornthwaite (Maes et al., 2019) methods, respectively, as well as land-cover based on NDVI (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 visualized on monthly and/or annual timescale.

To assess the potential role of regional groundwater import to or export from the basin, the water balance was estimated using long-term 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 $\overline{P} - \overline{E} - \overline{Q} \neq 0$, can then be largely attributed to groundwater import/export. In case of groundwater export, the average annual loss term can then be estimated according to (e.g., Bouaziz et al., 2018):

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

where $\overline{Q_L}$ is annual mean groundwater export (mm yr^{-1}), $\overline{P}$ annual mean precipitation (mm yr^{-1}), $\overline{E}$ annual mean evaporation (mm yr^{-1}), and $\overline{Q}$ annual mean discharge (mm yr^{-1}).

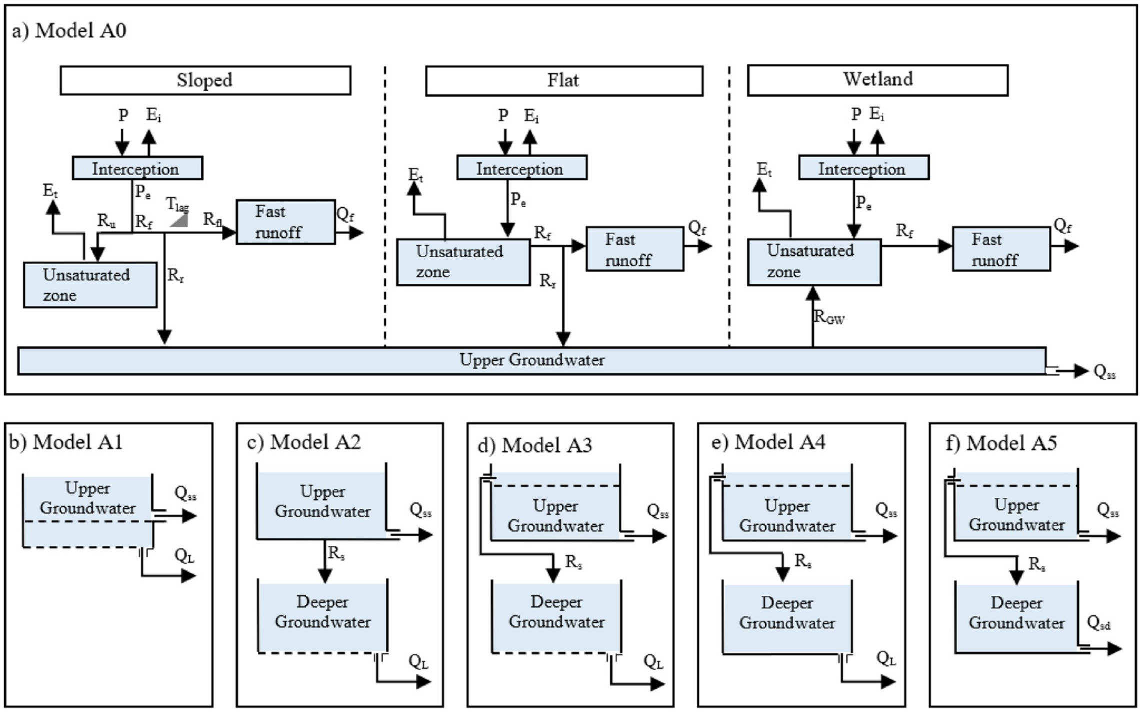

Figure 2. Schematization of the model structure applied to each grid cell for Models A0–A5. For Models A1–A5 (b–f), only the groundwater module is shown for brevity and clarity of the presentation, as the rest of the model structure remained the same. Abbreviations: precipitation (P), effective precipitation (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}).

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 stepwise developed and refined in previous studies (Hulsman, Savenije, et al., 2020; Hulsman, Winsemius, et al., 2020) following the FLEX-Topo modeling concept (Savenije, 2010). Each 0.25° × 0.25° model cell had the same model structure and parameter set, but was forced differently using spatially distributed precipitation and potential evaporation data (e.g., Euser et al., 2015). In addition, each cell was further discretized into functionally distinct landscape classes, that is, 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 lumped over the entire basin assuming a homogeneous groundwater system (Hulsman, Savenije, et al., 2020). The HRUs were classified based on the local slope and "Height-above-the-nearest-drainage" (HAND; Rennó et al., 2008) into sloped areas (slope ≥ 4%), flat areas (slope < 4%, HAND ≥ 11 m), and wetland areas (slope < 4%, HAND < 11 m). As a result, 68% of the basin was classified as flat, 28% as sloped, and 8% as wetlands (Figure 1b). This FLEX-Topo modeling concept was previously successfully applied in many different environments (Gao et al., 2014; Gharari et al., 2014; Hulsman, Winsemius, et al., 2020; Nijzink et al., 2016).

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 single 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 upwelling groundwater sustained the shallow groundwater tables and the unsaturated zone soil moisture content under the assumption that water is pushed upwards from the upland groundwater system into the wetlands due to the groundwater head difference between the uplands and wetlands (Hulsman, Savenije, et al., 2020).

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

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. The modeled total water storage anomaly was then calculated for each grid cell by taking the sum of all storage components, hence $S_{\mathrm{tot}} = S_{\mathrm{i,tot}} + S_{\mathrm{u,tot}} + S_{\mathrm{f,tot}} + S_{\mathrm{su}} + S_{\mathrm{sd}}$ (see Table 3 for an explanation of the abbreviations), and subtracting the 2004–2009 time-mean baseline similar to GRACE. The storages $S_{\mathrm{i,tot}}$, $S_{\mathrm{u,tot}}$ and $S_{\mathrm{f,tot}}$ are weighted averages from the storages $S_{\mathrm{i,HRU}}$, $S_{\mathrm{u,HRU}}$ and $S_{\mathrm{f,HRU}}$, respectively, in each HRU in a grid cell. This model consisted of 17 calibration parameters with uniform prior distributions and process constraints as summarized in Table 4 (Gharari et al., 2014). Parameter ranges were based on previous studies (e.g., Gao et al., 2014; Gharari et al., 2014; Wang-Erlandsson et al., 2016), their most extreme values (for example the splitter W) or selected based on trial and error such that different internal processes occur without introducing too much flexibility. In this benchmark model, the precipitation product CHIRPS was used and potential evaporation was estimated 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 were 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, that is, CHIRPS and TRMM, on the one hand and methods to estimate potential evaporation, that is, Hargreaves and Thornthwaite, on the other hand were tested in Models B0–D0 (Table 2).

4.2.3. Second Model Adaptation: Alternative Model Structure (Model A1–A5)

As second model adaptation, the model structure was changed to allow for additional alternative process formulations, representing deep groundwater flow or inter-basin groundwater export/import and to test their potential as relevant drivers 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, CHIRPS precipitation data and the Hargreaves method to estimate potential evaporation were used in these five model adaptations A1–A5.

More specifically, 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, a groundwater loss ($Q_{\mathrm{L};}$ Equation 36), which did not reach the river upstream of the gauging station, was introduced (Figures 2b and 3). In the spirit of model parsimony, Q_{L} was assumed to be constant, and thus independent of the water content in the Upper Groundwater reservoir to limit the number of calibration parameters in the absence of more detailed information. Thus, the Upper Groundwater reservoir (S_{su}) was formulated as a deficit store that can become negative and that loses water at a constant rate Q_{L}, expressed as a free calibration parameter. Note that, the shallow groundwater flow Q_{ss} only occurred when this storage was positive (if $S_{\mathrm{su}} > 0$, Equation 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., Bouaziz et al., 2018; Hrachowitz et al., 2014).

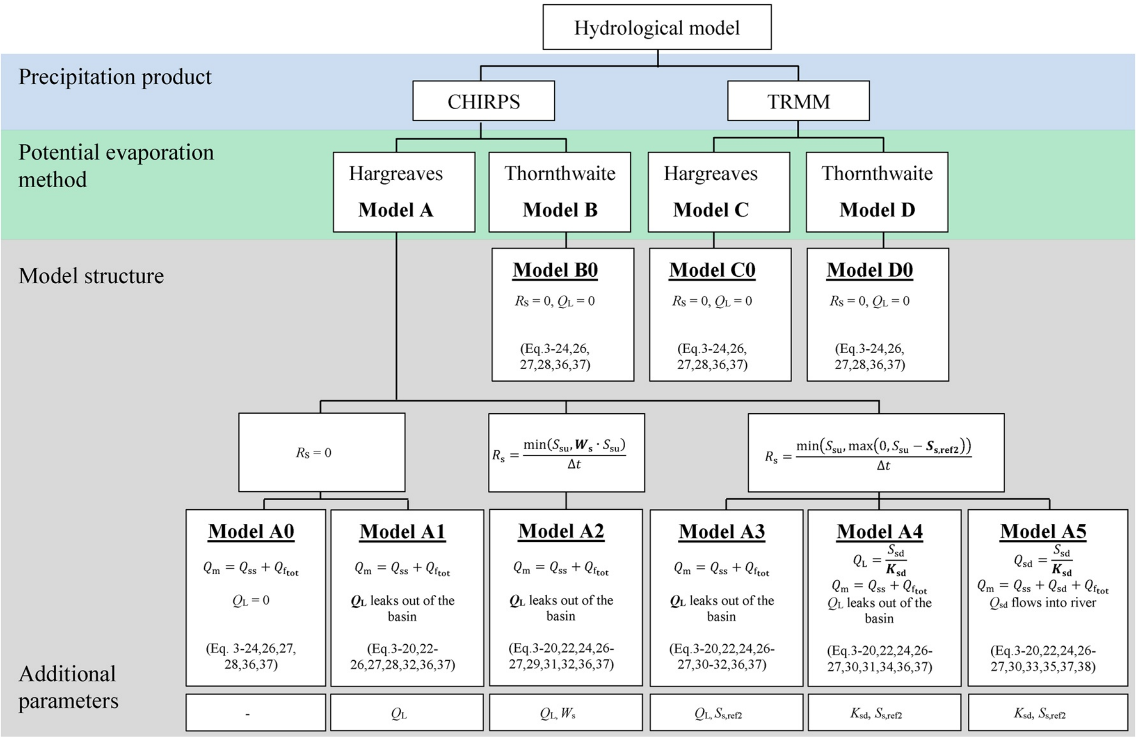
Hydrological model
Precipitation product
CHIRPS
TRMM
Potential evaporation method
Hargreaves
Model A
Thornthwaite
Model B
Hargreaves
Model C
Thornthwaite
Model D
Model structure
Model B0
$R_S = 0$, $Q_L = 0$
(Eq.3-24,26, 27,28,36,37)
Model C0
$R_S = 0$, $Q_L = 0$
(Eq.3-24,26, 27,28,36,37)
Model D0
$R_S = 0$, $Q_L = 0$
(Eq.3-24,26, 27,28,36,37)
$R_S = 0$
$R_s = \frac{\min(S_{su}, \boldsymbol{W_s} \cdot S_{su})}{\Delta t}$
$R_s = \frac{\min(S_{su}, \max(0, S_{su} - \boldsymbol{S_{s,ref2}}))}{\Delta t}$
Model A0
$Q_m = Q_{ss} + Q_{f_{tot}}$
$Q_L = 0$
(Eq. 3-24,26,27, 28,36,37)
Model A1
$Q_m = Q_{ss} + Q_{f_{tot}}$
Q_L leaks out of the basin
(Eq.3-20,22- 26,27,28,32,36,37)
Model A2
$Q_m = Q_{ss} + Q_{f_{tot}}$
Q_L leaks out of the basin
(Eq.3-20,22,24,26- 27,29,31,32,36,37)
Model A3
$Q_m = Q_{ss} + Q_{f_{tot}}$
Q_L leaks out of the basin
(Eq.3-20,22,24,26- 27,30-32,36,37)
Model A4
$Q_L = \frac{S_{sd}}{\boldsymbol{K_{sd}}}$
$Q_m = Q_{ss} + Q_{f_{tot}}$
Q_L leaks out of the basin
(Eq.3-20,22,24,26- 27,30,31,34,36,37)
Model A5
$Q_{sd} = \frac{S_{sd}}{\boldsymbol{K_{sd}}}$
$Q_m = Q_{ss} + Q_{sd} + Q_{f_{tot}}$
Q_{sd} flows into river
(Eq.3-20,22,24,26- 27,30,33,35,37,38)
Additional parameters
-
Q_L
Q_L, W_s
Q_L, $S_{s,ref2}$
K_{sd}, $S_{s,ref2}$
K_{sd}, $S_{s,ref2}$

Figure 3. Overview hydrological models.

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, Equation 29, Figures 2c and 3). From this Deeper Groundwater reservoir, constant groundwater loss (Q_L) subsequently leaked out of the basin equivalent 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 ($S_{s,ref2}$, Equation 30, Figures 2d and 3). From there constant groundwater loss (Q_L) then leaked out of the basin equivalent to Models A1 and A2.

With Model A4, it was tested whether temporally *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 (Equation 34). As in Models A1–A3, this groundwater loss (Q_L) did not reach the river.

With Model A5, it was tested whether temporally 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}, thereby contributing to the total river flow (Equation 38, Figures 2f and 3). Hence, only in Model A5 the additional contributions from a deep groundwater storage reached the gauged river system whereas in Models A1–A4 groundwater exclusively 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.

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)	$E_{\mathrm{i}} = \min\left(E_{\mathrm{p}}, \min\left(P, \frac{\boldsymbol{I}_{\mathbf{max}}}{\Delta t}\right)\right)$	(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)	$E_{\mathrm{t}} = \min\left(\left(E_{\mathrm{p}} - E_{\mathrm{i}}\right), \min\left(\frac{S_{\mathrm{u}}}{\Delta t}, \left(E_{\mathrm{p}} - E_{\mathrm{i}}\right) \cdot \frac{S_{\mathrm{u}}}{\boldsymbol{S}_{\mathbf{u,max}}} \cdot \frac{1}{\boldsymbol{C}_{\mathbf{e}}}\right)\right)$	(7)
	Flat: $\frac{\Delta S_{\mathrm{u}}}{\Delta t} = P_{\mathrm{e}} - E_{\mathrm{t}} - R_{f}$	(8)	$R_{\mathrm{GW}} = \min\left(\frac{\min\left(S_{\mathrm{su}}, \boldsymbol{S}_{\mathbf{s,ref1}}\right)}{\boldsymbol{S}_{\mathbf{s,ref1}}} \cdot \boldsymbol{C}_{\mathbf{max}}, \frac{\frac{S_{\mathrm{su}}}{\Delta t}}{p_{\mathrm{HRU}}}\right)$	(9)
	Wetland: $\frac{\Delta S_{\mathrm{u}}}{\Delta t} = P_{\mathrm{e}} - E_{\mathrm{t}} - R_{f} + R_{\mathrm{GW}}$	(10)	if $S_{\mathrm{u}} + R_{\mathrm{GW}} \cdot \Delta t > \boldsymbol{S}_{\mathbf{u,max}} : R_{\mathrm{GW}} = \frac{\boldsymbol{S}_{\mathbf{u,max}} - S_{\mathrm{u}}}{\Delta t}$	(11)
			Sloped: $R_{\mathrm{u}} = \left(1 - C\right) \cdot P_{\mathrm{e}}$	(12)
			$C = 1 - \left(1 - \frac{S_{\mathrm{u}}}{\boldsymbol{S}_{\mathbf{u,max}}}\right)^{\beta}$	(13)
Fast runoff	$\frac{\Delta S_{\mathrm{f}}}{\Delta t} = R_{\mathrm{fl}} - Q_{\mathrm{f}}$	(14)	$Q_{\mathrm{f}} = \frac{S_{\mathrm{f}}}{\boldsymbol{K}_{\mathbf{f}}}$	(15)
			Sloped: $R_{\mathrm{f}} = C \cdot P_{\mathrm{e}}$	(16)
			$R_{\mathrm{fl}} = \left(1 - \boldsymbol{W}\right) \cdot R_{\mathrm{f}} * \mathrm{f}\left(\boldsymbol{T}_{\mathbf{lag}}\right)$	(17)
			Flat/Wetland: $R_{\mathrm{f}} = \frac{\max\left(0, S_{\mathrm{u}} - \boldsymbol{S}_{\mathbf{u,max}}\right)}{\Delta t}$	(18)
			Flat: $R_{\mathrm{fl}} = \left(1 - \boldsymbol{W}\right) \cdot R_{\mathrm{f}}$	(19)
			Wetland: $R_{\mathrm{fl}} = R_{f}$	(20)
Upper Groundwater	$\frac{\Delta S_{\mathrm{su}}}{\Delta t} = R_{\mathrm{r_{tot}}} - R_{\mathrm{GW_{tot}}} - Q_{\mathrm{ss}}$ (A0)	(21)	$R_{\mathrm{r}} = \boldsymbol{W} \cdot R_{\mathrm{f}}$	(22)
	$\frac{\Delta S_{\mathrm{su}}}{\Delta t} = R_{\mathrm{r_{tot}}} - R_{\mathrm{GW_{tot}}} - Q_{\mathrm{ss}} - Q_{\mathrm{L}}$ (A1)	(23)	$R_{\mathrm{r_{tot}}} = \sum_{\mathrm{HRU}} p_{\mathrm{HRU}} \cdot R_{\mathrm{r}}$	(24)
	$\frac{\Delta S_{\mathrm{su}}}{\Delta t} = R_{\mathrm{r_{tot}}} - R_{\mathrm{GW_{tot}}} - Q_{\mathrm{ss}} - R_{\mathrm{s}}$ (A2–A5)	(25)	$R_{\mathrm{GW_{tot}}} = \sum_{\mathrm{HRU}} p_{\mathrm{HRU}} \cdot R_{\mathrm{GW}}$	(26)
			$Q_{\mathrm{ss}} = \frac{\max\left(0, S_{\mathrm{su}}\right)}{\boldsymbol{K}_{\mathbf{s}}}$	(27)
			$R_{\mathrm{s}} = 0$ (A0–A1)	(28)
			$R_{\mathrm{s}} = \frac{\boldsymbol{W}_{\mathbf{s}} \cdot S_{\mathrm{su}}}{\Delta t}$ (A2)	(29)
			$R_{\mathrm{s}} = \frac{\min\left(S_{\mathrm{su}}, \max\left(0, S_{\mathrm{su}} - \boldsymbol{S}_{\mathbf{s,ref2}}\right)\right)}{\Delta t}$ (A3–A5)	(30)

Table 3
Continued

Reservoir system	Water balance equations	Eq.	Process functions	Eq.
Deeper Groundwater	$\frac{\Delta S_{\mathrm{sd}}}{\Delta t} = R_{\mathrm{s}} - Q_{\mathrm{L}}$ (A2–A4)	(31)	$\boldsymbol{Q}_{\mathbf{L}} = \text{const.}$ (A1–A3)	(32)
	$\frac{\Delta S_{\mathrm{sd}}}{\Delta t} = R_{\mathrm{s}} - Q_{\mathrm{sd}}$ (A5)	(33)	$Q_{\mathrm{L}} = \frac{S_{\mathrm{sd}}}{\boldsymbol{K}_{\mathbf{sd}}}$ (A4)	(34)
			$Q_{\mathrm{sd}} = \frac{S_{\mathrm{sd}}}{\boldsymbol{K}_{\mathbf{sd}}}$ (A5)	(35)
Total runoff	$Q_{\mathrm{m}} = Q_{\mathrm{f_{tot}}} + Q_{\mathrm{ss}}$ (A0–A4)	(36)	$Q_{\mathrm{f_{tot}}} = \sum_{\mathrm{HRU}} p_{\mathrm{HRU}} \cdot Q_{\mathrm{f}}$	(37)
	$Q_{\mathrm{m}} = Q_{\mathrm{f_{tot}}} + Q_{\mathrm{ss}} + Q_{\mathrm{sd}}$ (A5)	(38)		

Notes. 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 $R_{\mathrm{r,tot}}$ combining all relevant HRUs), groundwater upwelling (R_{GW} for each relevant HRU and $R_{\mathrm{GW,tot}}$ 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 ($S_{\mathrm{s,ref1}}$, $S_{\mathrm{s,ref2}}$) (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) and each model (A0–A5) unless indicated differently.

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, that is, changing the forcing data (Models A0–D0) and the second model adaptation, that is, 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 anomalies, 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.

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 ($E_{\mathrm{NS},\theta}$, Equation 39 in Table 5 or relative error ($E_{\mathrm{R},\theta}$, Equation 40), depending on the signature. The individual performance metrics included the Nash-Sutcliffe efficiency of the daily flow time-series ($E_{\mathrm{NS,Q}}$) and its logarithm ($E_{\mathrm{NS,logQ}}$), of the flow duration curve ($E_{\mathrm{NS,FDC}}$) and its logarithm ($E_{\mathrm{NS,logFDC}}$), and of the autocorrelation function of the daily flows ($E_{\mathrm{NS,AC}}$). In addition, the relative error of the mean seasonal runoff coefficients during dry and wet periods ($E_{\mathrm{R,RCdry}}$, $E_{\mathrm{R,RCwet}}$), and the rising limb density of the hydrograph ($E_{\mathrm{R,RLD}}$) (Euser et al., 2013) were used. These signatures were combined, assuming equals weights, using the Euclidian distance ($D_{\mathrm{E,Q}}$, Equation 41) with $D_{\mathrm{E,Q}} = 1$ corresponding to the "perfect" model.

The model performance with respect to the basin-average total water storage anomalies was evaluated with the Euclidian distance ($D_{\mathrm{E,S}}$, Equation 41) of the Nash-Sutcliffe efficiencies on monthly ($E_{\mathrm{NS,S,monthly}}$) and annual ($E_{\mathrm{NS,S,annual}}$) timescale. On annual timescale, the Nash-Sutcliffe efficiency was calculated for the annual minima and maxima separately which were then averaged to obtain $E_{\mathrm{NS,S,annual}}$. The annual time-series were normalized by dividing it with the maximum range in the observed annual minima or maxima total water

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	–	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	–	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	$I_{max,sloped} > I_{max,flat}$	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	$I_{max,wetland} < I_{max, sloped}$	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	–	All models
River profile	v	0.01	5	$m\ s^{-1}$	–	All models

storage anomalies, respectively. With this performance measure for the total water storage anomalies, 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 (Equation 41) using $D_{E,Q}$ for the discharge and $D_{E,S}$ for the total water storage anomalies. This performance measure was used to select the best performing models representing both the discharge and the total storage as good as possible.

4.4. Parameter Selection Procedure

Each hydrological model (A0–D0 and A1–A5) 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 anomalies ($D_{E,QS}$). 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 ($D_{E,Q}$), total water storage anomalies ($D_{E,S}$) and both simultaneously ($D_{E,QS}$). The model was run for the time period 1995–2016 and calibrated/evaluated for the time period 2002–2016 using the first 7 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 anomalies to capture the long-term variability in an efficient way.

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 anomalies simultaneously ($D_{\mathrm{E,QS}}$) for the time period 2002–2012, and post-calibration evaluating the models with respect to total water storage anomalies 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. Similarly, 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 & 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 (Blazquez et al., 2018; Landerer & Swenson, 2012). The area of these open water bodies was 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 anomalies 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.

Furthermore, GRACE observations used in this study were an average of products generated by CSR, GFZ, and JPL as explained in Section 3. The individual products showed similar long-term and seasonal patterns for the 2002–2015 time-period as shown in Figure S3a in the Supplementary Material. The standard deviation between the three solutions reached up to 11.7 mm for the annual minima and 19.1 mm for the annual maxima which is significantly smaller compared to the multi-annual variations. In 2016, the standard deviation increased to 73 mm for the annual maximum when considering the mascon (mass concentration block) solution according to JPL too. That is why it is plausible to assume that these long-term storage variations were not dominated by uncertainties in the GRACE observations for the 2002–2015 time-period.

5.1.2. Precipitation

Alternatively, long-term variations in the total water storage can be caused by changes in precipitation. In the Luangwa basin, the annual observed precipitation volumes varied over the years, depending on the data source, from 920 to 1,337 mm (CHIRPS) and from 858 to 1,213 mm (TRMM), as shown in Figures 4a and 4d.

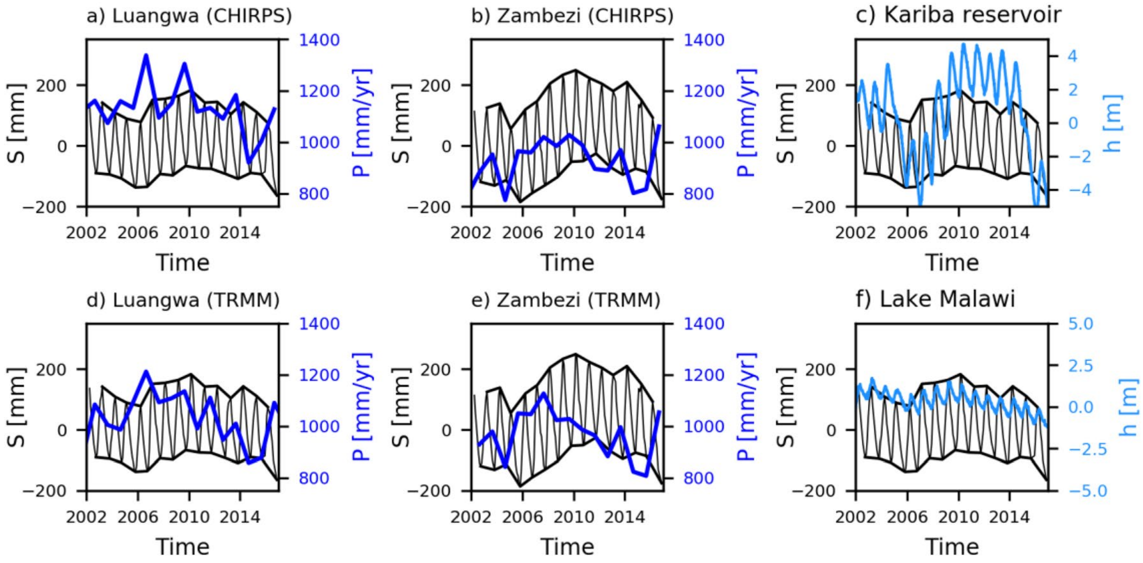

Figure 4. Basin-average total water storage anomalies according to GRACE (black) and annual rainfall (dark blue) according to CHIRPS (a and b) and TRMM (d and e) for the Luangwa (a and d) and Zambezi (b and e) river basin, or altimetry observations (light blue) at (c) Kariba reservoir and (f) Lake Malawi.

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 anomalies 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 total water storage anomalies in 2006. This was followed by several wet years with a maximum rainfall volume of 1,213 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 anomalies 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; Le Coz & van de Giesen, 2019; Mazzoleni et al., 2019; Thiemig et al., 2012). 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 anomalies 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 $E_P = 1{,}565$ mm yr^{-1} (Figure 5a), Thornthwaite estimated long-term mean $E_P = 1{,}904$ 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 anomalies resulting in weak correlations with respect to the annual minimum ($R^2 = 0.03$) and maximum variations ($R_2 = 0.34$). In contrast, no discernible long-term fluctuations were observed when applying the Hargreaves method ($R^2 = 0.00$ and

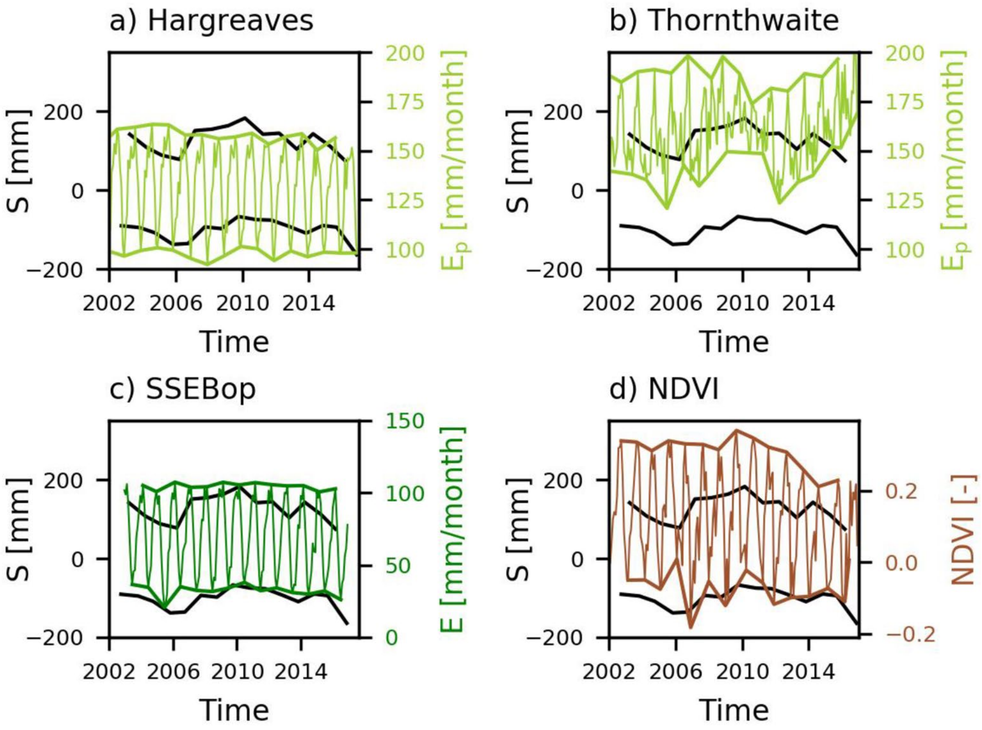

Figure 5. Basin-average total water storage anomalies according to GRACE (black) with respect to the annual minima/maxima combined with basin-average (a) monthly potential evaporation according to Hargreaves (light green) and (b) Thornthwaite (light green), (c) monthly actual evaporation according to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

$R^2 = 0.12$ with respect to the annual minima and maxima, respectively). 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 modeled total water storage anomalies 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 S4 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 $R^2 = 0.00$–0.13 with respect to the annual maxima and $R^2 = 0.02$–0.17 with respect to the annual minima for all satellite products used in this study except for SSEBop which showed the highest $R^2 = 0.37$ (Figure 5c and Figure S4 in the Supplementary Material). Note, that the observed annual minimum storage increase of 67 mm over 3 years (2006–2009), which in fact is an accumulated difference arising from the combined history of inputs and outputs over that period, can result from a mean daily deviation of only 0.06 mm d^{-1} in 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 only very limited direct correspondence with water storage variations, which was likely a consequence of the subtle and 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 analyzed evaporation products.

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., 2019a, 2019b). However, inspections of NDVI time-series (Figure 5) did not reveal any significant long-term variations directly corresponding with water storage variations over the 2002–2016 period. 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 minimum/maximum total water storage anomalies and NDVI ($R^2 = 0.01$ and $R^2 = 0.06$, respectively). 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 gray 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 range of groundwater export of $\overline{Q_L} = 0.02–0.39$ mm d^{-1} is in the following assumed for the study basin.

5.2. Hydrological Models

5.2.1. Benchmark Model (Model A0)

Following the first calibration strategy, that is, calibrating with respect to discharge, the benchmark Model A0 captured the discharge well (Figures 7a and 7b) with an optimum model performance of $D_{\mathrm{E,Q,opt}} = 0.85$ (Table 6, Figure 8a). The modeled 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 $D_{\mathrm{E,S}} = -14$ (Table 6, Figure 8a). On the monthly timescale, the general seasonal storage fluctuations were modeled well with respect to the timings of the wet and dry season (Figure S7a in the Supplementary Material). However, the annual storage maxima were significantly overestimated, and the annual minima underestimated (Figure 7c). In addition, the modeled total water storage anomalies did not reflect any fluctuations in the annual minima in contrast to the observations (Figure 7e, $R^2 = 0.07$), whereas the modeled 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 anomalies $D_{\mathrm{E,QS}} = -9.6$ remained poor.

Following the second calibration strategy, that is, calibration with respect to discharge and total water storage anomalies simultaneously, the ability of the model to reproduce flow decreased significantly to $D_{\mathrm{E,Q}} = -0.23$ (Table 6, Figure 8b). While the general flow dynamics were modeled well (Figure S8a in the Supplementary Material), the flows were systematically overestimated (Figure 9a). In contrast, the modeled monthly and annual total water storage anomaly time-series improved ($D_{\mathrm{E,S}} = -0.11$). The modeled total water storage anomalies mimicked the seasonal variations in the observation better (Figure S9a in

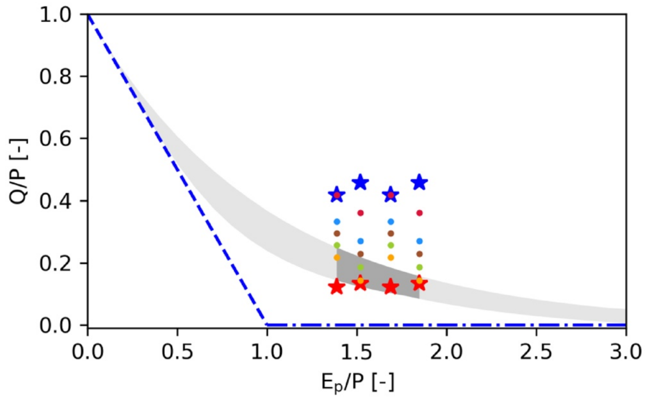

Figure 6. Runoff coefficient (Q/P) as a function of the dryness index (E_p/P) where Q is discharge, P precipitation, and E_p potential evaporation. The blue dashed line indicates the energy limit and the blue horizontal dash-dotted line the water limit. The gray area indicates envelope of analytical 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 ($\overline{E_P}$ = 1,565 mm yr^{-1}) or Thornthwaite ($\overline{E_P}$ = 1,904 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 ($\overline{Q} + \overline{Q_L} = \overline{P} - \overline{E}$, Equation 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 gray area), the groundwater exchange should range between $\overline{Q_L}$ = 9–143 mm yr^{-1} depending on the satellite products used. See Table 1 for the corresponding long-term values of the individual fluxes.

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 9b) and the fluctuations in the annual maxima improved slightly (Figure 9c, R^2 = 0.31). However, the modeled storage did not reflect any fluctuations in the annual minima (Figure 9d, 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 anomalies only, the annual minima did not reflect any fluctuations (Figure S10 in the Supplementary Material, R^2 = 0.08).

As a result, this benchmark Model A0 reproduced the flows well only with calibration strategy 1, while the seasonal fluctuations in the total water storage were better reproduced with calibration strategy 2. However, the long-term variations in the total water storage anomalies with respect to the annual maxima were poorly modeled and with respect to the annual minima completely missed for both calibration strategies.

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 S5 and S6a 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 anomalies remained poorly modeled for all models with $D_{E,S}$ = −3.4 to −0.48 (Table 6, Figure 8a). On monthly timescale, the general seasonal fluctuations were modeled well with slight differences mostly in the storage decrease during dry seasons (Figure S7 in the Supplementary Material). The magnitudes of the modeled 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 S6b 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, Figures S6c–S6d 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 anomalies 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 modeled 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 S8 in the Supplementary Material), but the flow magnitudes were only captured well for Models C0 and D0 (Figure 9a). While Models A0 and B0 significantly overestimated the flows continuously, Model C0 only slightly overestimated the flows and Model D0 only slightly underestimated the medium to low flows (Figure 9a). 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 modeled monthly and annual total water storage anomalies 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 S9 in the Supplementary Material). The magnitudes of the annual minima and maxima corresponded well with the observations for all models (Figure 9b), whereas the fluctuations in the annual maxima only improved for Model D0 with R^2 = 0.39 (Figure 9c). On the other hand, the annual minima remained close to constant for all models

Table 5
Overview of Equations Used to Calculate the Model Performance

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

(R^2 = 0.00–0.03; Figure 9d). The overall model performance with respect to discharge and total water storage anomalies improved the most for Model D0 with $D_{\mathrm{E,QS}} = 0.52$ (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, that is, the annual maxima, was considerably influenced by the choice of precipitation data source and the method to estimate potential evaporation. In contrast, the modeled dry season storage, that is, 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 already explain a significant part of the inability of the model to reproduce long-term water storage variations.

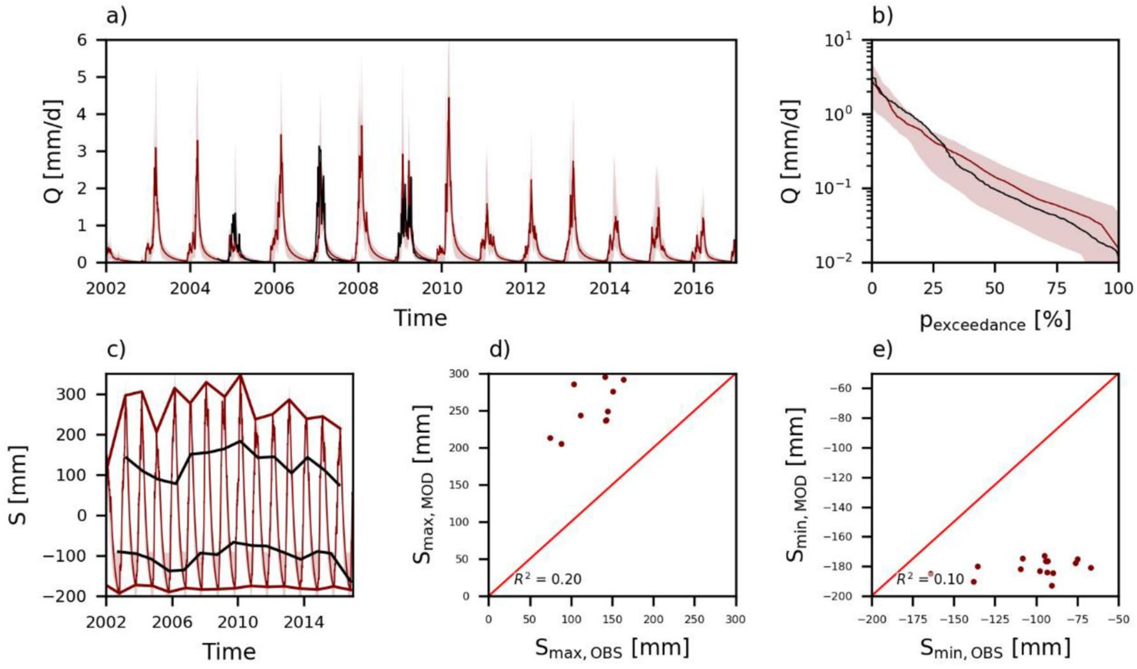

Figure 7. Range of model solutions for Model A0 for calibration strategy 1 with respect to (a) hydrograph, (b) flow duration curve, (c) total water storage anomaly time-series, (d) annual maximum total water storage anomalies, and (e) annual minimum total water storage anomalies. In (a–c), the black line indicates the recorded data, the colored line the solution with the highest calibration objective function with respect to discharge ($D_{\mathrm{E,Q}}$) 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 ($D_{\mathrm{E,Q}}$). The red line indicates the 1:1 line and R^2 is the correlation with respect to a fitted regression line.

Table 6

Model Performance With Respect to Discharge (D_E), Total Water Storage Anomalies ($D_{E,S}$) 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 to −5.5)	−9.6 (−12 to −3.6)	−0.23 (−0.71 to −0.06)	−0.11 (−0.80 to −0.10)	−0.17 (−0.52 to −0.31)
Model B0	0.85 (0.72–0.81)	−3.4 (−9.2 to −1.7)	−2.1 (−6.2 to −0.94)	0.32 (−0.14–0.49)	0.00 (−0.65 to −0.09)	0.14 (−0.25–0.01)
Model C0	0.91 (0.80–0.88)	−0.85 (−4.5 to −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 to −0.87)	−11 (−10 to −0.51)	−0.20 (−1.1–0.07)	0.05 (−1.4 to −0.15)	−0.08 (−0.90 to −0.35)
Model A2	0.82 (−5.1–0.51)	−1,066 (−813 to −3.4)	−753 (−575 to −3.3)	−0.24 (−11 to −1.0)	−0.47 (−7.5 to −0.68)	−0.36 (−7.6 to −3.3)
Model A3	0.87 (0.73–0.83)	−425 (−1,133 to −11)	−300 (−801 to −7.2)	0.28 (−1.2–0.49)	−0.45 (−3.9 to −0.66)	−0.14 (−2.6 to −0.53)
Model A4	0.87 (0.73–0.83)	−9.8 (−27 to −3.6)	−6.7 (−19 to −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 to −5.3)	−9.0 (−12 to −3.5)	−0.31 (−0.72–0.03)	0.23 (−0.73–0.08)	−0.07 (−0.46 to −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)

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 S11 and S12a in the Supplementary Material). The monthly and annual total water storage anomaly time-series was modeled very poorly for all models ($D_{E,S} = -1{,}066$ to -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 S12 and S13 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 anomalies simultaneously improved the most for Model A4 with $D_{E,QS} = 0.32$ (Table 6, Figure 8a).

Following the second calibration strategy, the modeled 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 $D_{E,Q} = -0.31$ to -0.20 (Table 6, Figure 8b). The general flow dynamics were reproduced well for Models A1–A4 (Figure S14 in the Supplementary Material), 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 10a), whereas Model A4 only slightly overestimated the high flows and underestimated the low flows. The monthly variations in the total water storage anomalies were captured well for all models with some differences in the storage decrease during dry seasons especially for Model A2 (Figure S15 in the Supplementary Material). While the magnitudes of the annual maxima and minima were captured well for all models (Figure 10b), the annual fluctuations improved the most Model A5 with respect to the annual maxima ($R^2 = 0.51$, Figure 10c) and for Models A2 and A5 with respect to the annual minima ($R^2 = 0.23$, Figure 10d). 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, that is, annual minima and maxima, was significantly influenced by the model structure. The modeled annual and monthly total water storage anomalies improved the most for Models A4 and A5 (Table 6, Figure 8b) where a Deeper Groundwater reservoir was incorporated with groundwater loss and/or 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 10a). Therefore, when considering the overall model performance with respect to discharge and total water storage anomalies 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

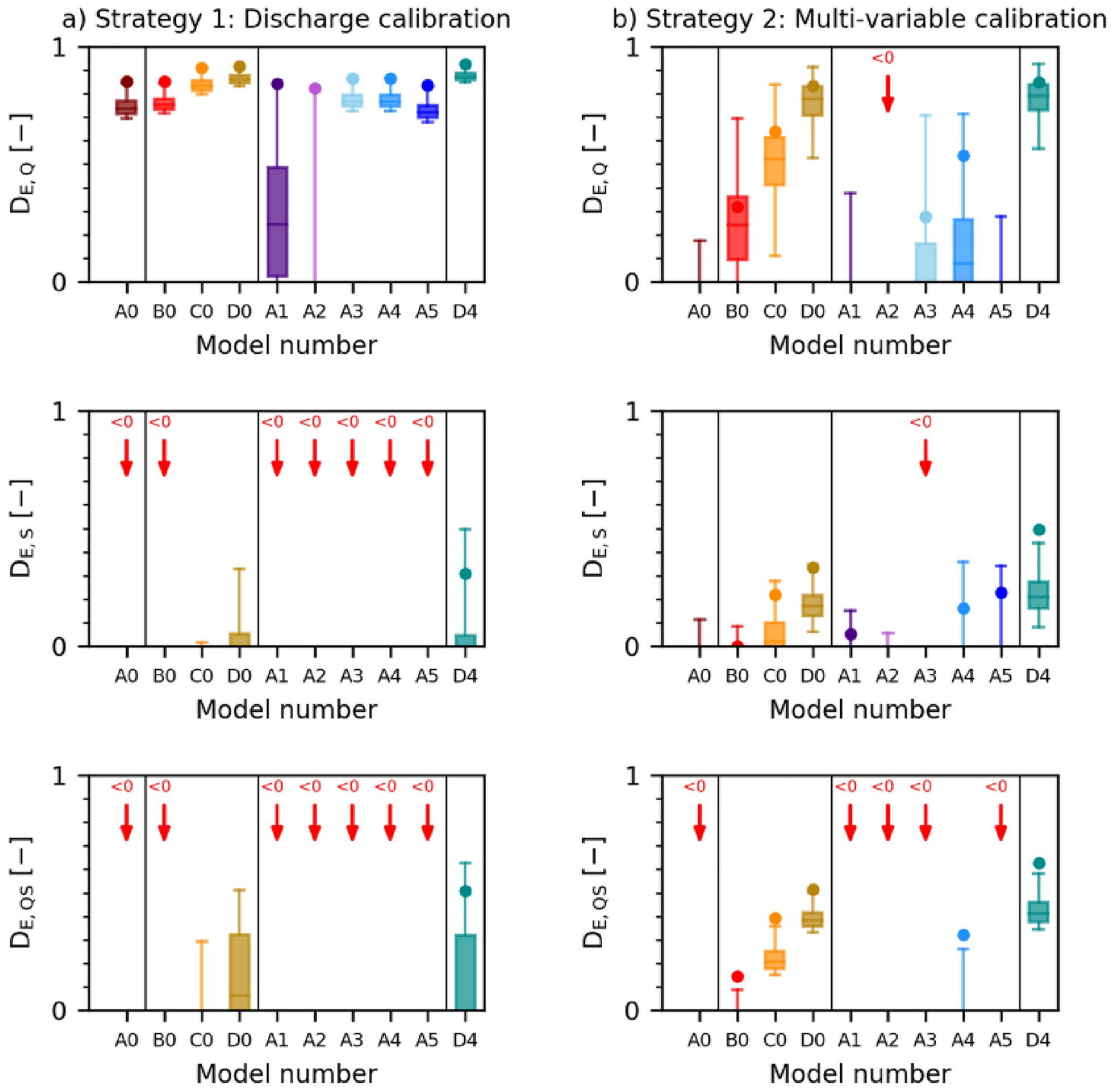

Figure 8. Model performance for Models A0–D4 with respect to discharge ($D_{E,Q}$), total water storage anomalies ($D_{E,S}$) and both combined ($D_{E,QS}$). The model is calibrated with respect to (a) discharge or (b) both variables simultaneously. The dots represent the model performance using the "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.

annual total water storage anomaly magnitudes and fluctuations, albeit with a slight overestimation of the annual minima and maxima in 2004–2006 (Figure 10b). These results provide evidence that the model hypotheses A0–A3 as well as A5 generate hydrological response patterns that are less consistent with the available data than those of Model A4. It is thus not implausible to reject hypotheses A0–A3 and A5 and to assume that long-term storage fluctuations are potentially the result of groundwater export according to the loss rate Q_L from the Deeper Groundwater reservoir (Model A4).

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 from 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

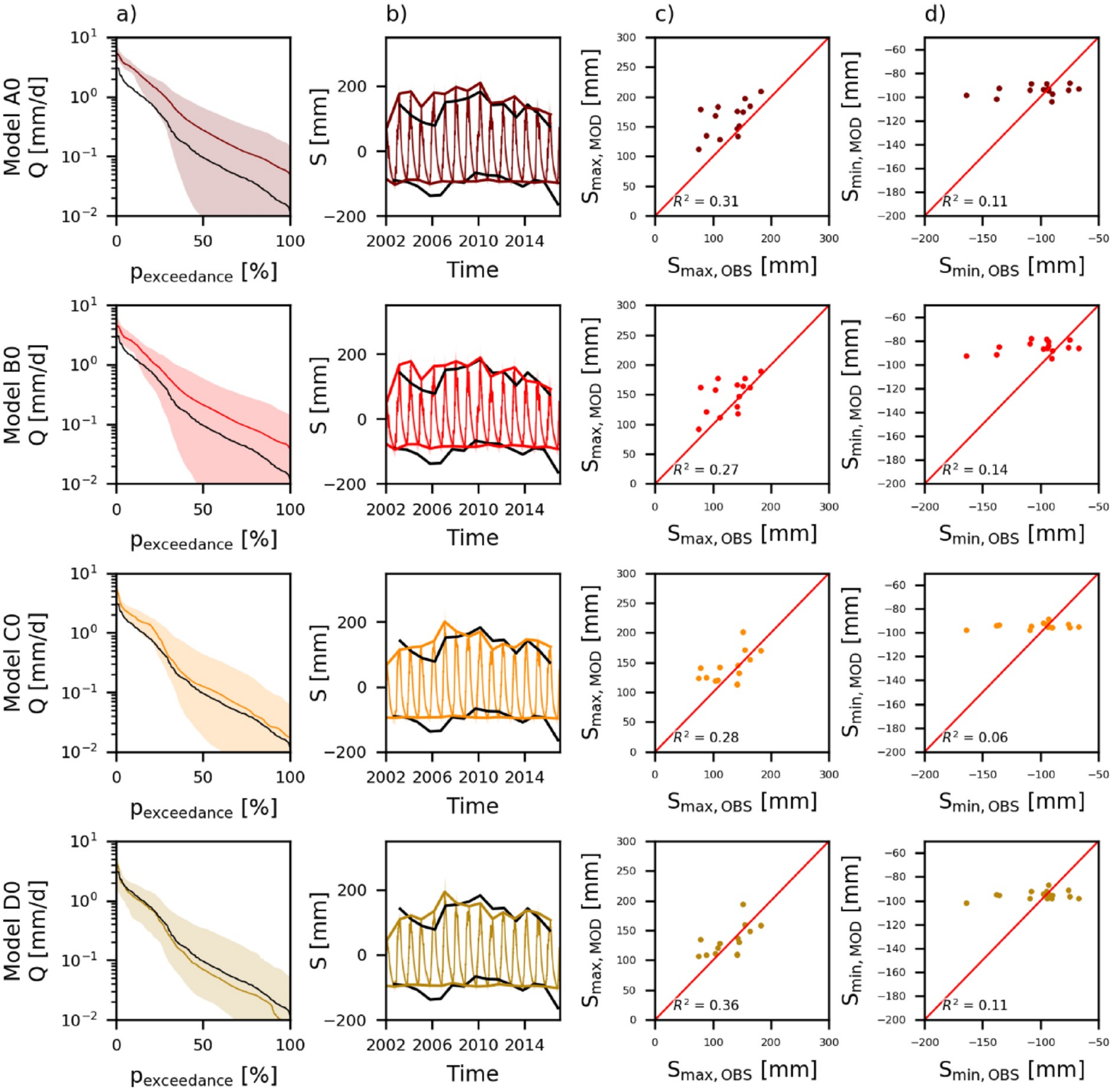

Figure 9. Range of model solutions for Models A0–D0 for calibration strategy 2 with respect to (a) flow duration curve, (b) total water storage anomaly time-series, (c) annual maximum total water storage anomalies, (d) annual minimum total water storage anomalies. In (a and b), the black line indicates the recorded data, the colored line the solution with the highest calibration objective function with respect to discharge and total water storage anomalies ($D_{E,QS}$) and the shaded area the envelope of the solutions retained as feasible. In (c and 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 anomalies ($D_{E,QS}$). The red line indicates the 1:1 line and R^2 is the correlation with respect to a fitted regression line.

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 S16a 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 (Figures S16a and S16b in the Supplementary Material). The monthly and annual total water storage anomalies improved significantly to $D_{E,S}$ = 0.31 (Table 6, Figure 8a). The modeled monthly storage variations were broadly consistent with the observation (Figure S17 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 stor-

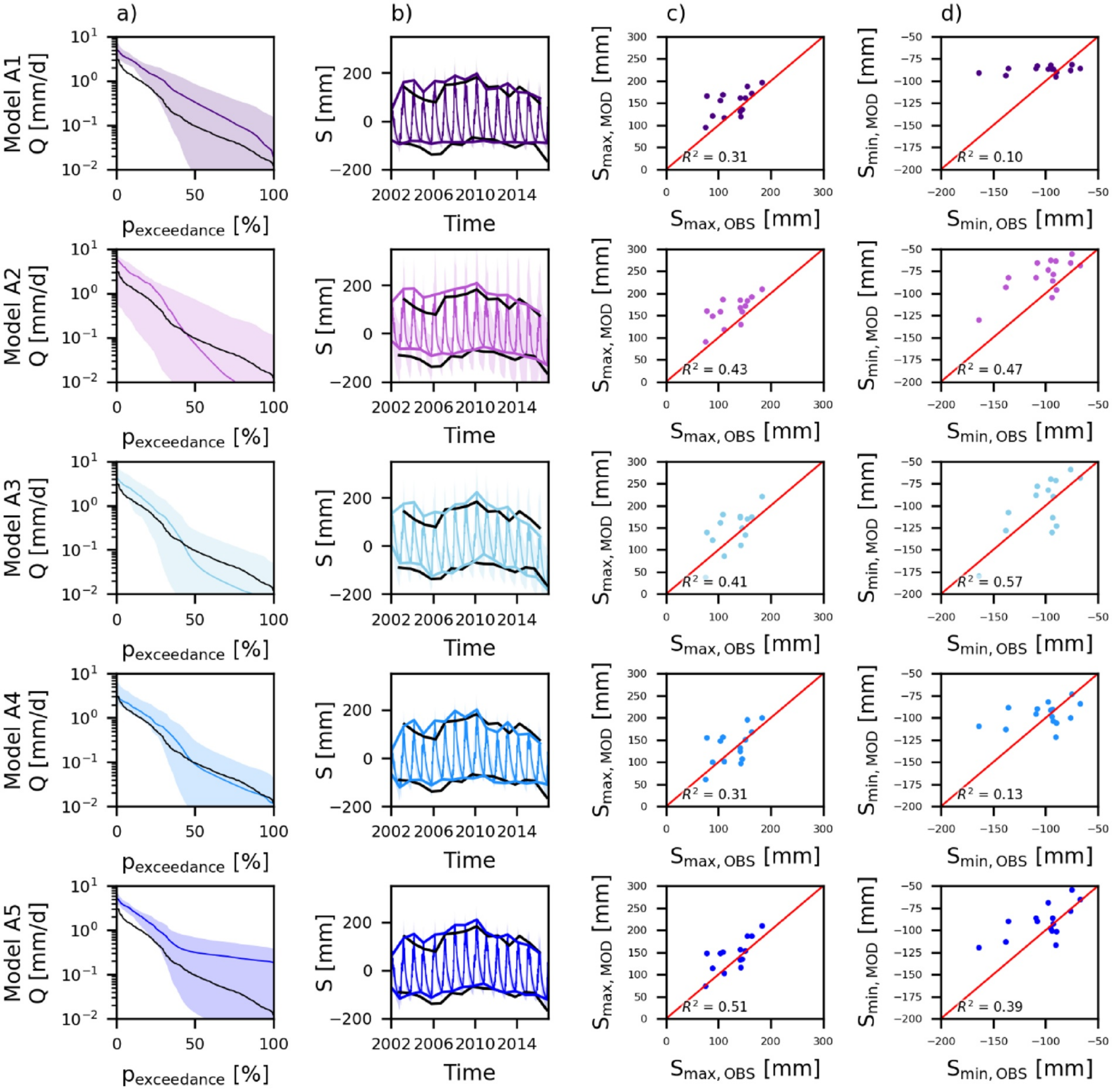

Figure 10. Range of model solutions for Models A1–A5 for calibration strategy 2 with respect to (a) flow duration curve, (b) total water storage anomaly time-series, (c) annual maximum total water storage anomalies, and (d) annual minimum total water storage anomalies. In (a and b), the black line indicates the recorded data, the colored line the solution with the highest calibration objective function with respect to discharge and total water storage anomalies ($D_{E,QS}$) and the shaded area the envelope of the solutions retained as feasible. In (c and 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 anomalies ($D_{E,QS}$). The red line indicates the 1:1 line and R^2 is the correlation with respect to a fitted regression line.

age were modeled well for the time period 2010–2016, whereas before 2010 the storage was overestimated (Figure S16c in the Supplementary Material). Also the fluctuations in the annual maximum storage were modeled well with $R^2 = 0.48$ (Figure S16d in the Supplementary Material), but the annual minima remained to be poorly captured ($R^2 = 0.19$, Figure S16e 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 modeled well (Figure 11a), 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 11a), low flows were slightly underestimated (Figure 11b). The monthly and annual total water storage anomaly time-series improved considerably to

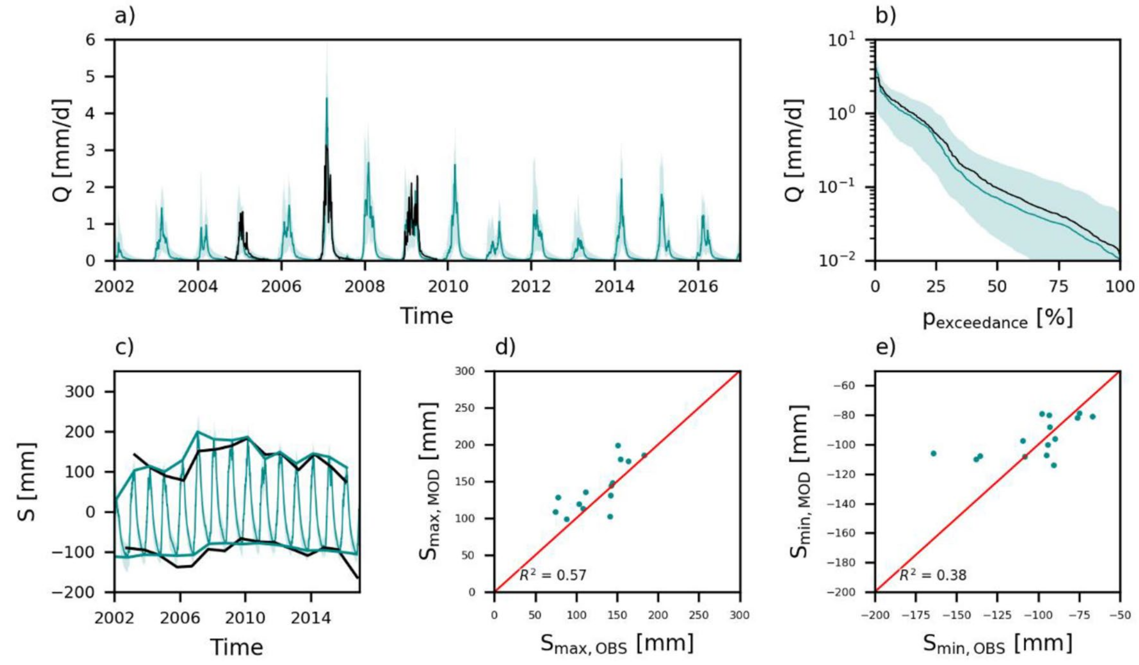

Figure 11. Range of model solutions for Model D4 for calibration strategy 2 with respect to (a) hydrograph, (b) flow duration curve, (c) total water storage anomaly time-series, (d) annual maximum total water storage anomalies, and (e) annual minimum total water storage anomalies. In (a–c), the black line indicates the recorded data, the colored line the solution with the highest calibration objective function with respect to discharge and total water storage anomalies ($D_{E,QS}$) 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 anomalies ($D_{E,QS}$). The red line indicates the 1:1 line and R^2 is the correlation with respect to a fitted regression line.

$D_{E,S} = 0.50$ (Table 6, Figure 8b). With this model and this calibration strategy, the monthly variations were captured well (Figure S18 in the Supplementary Material), as also magnitudes and fluctuations in the annual maxima ($R^2 = 0.57$, Figures 11c and 11d) and minima ($R^2 = 0.41$, Figures 11c and 11e). The overall model performance increased to $D_{E,QS} = 0.63$ which was better than all other alternative model hypotheses (Table 6, Figure 8b).

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 anomalies 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 anomalies for the time period 2012–2016 (see Section 4.4). While the general flow dynamics were modeled well for both models (Figure S19 in the Supplementary Material), the magnitudes improved significantly for Model D4 as the flows were only slightly underestimated during medium flows (Figure 12a). For the calibration period, the modeled 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 anomaly 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 S20 in the Supplementary Material). While the magnitudes of the annual minima/maxima were captured well for both models (Figure 12b), long-term fluctuations improved for Model D4 with respect to the annual maxima ($R^2 = 0.57$, Figure 12c) and minima ($R^2 = 0.44$, Figure 12d). In both cases, R^2 was calculated for the calibration time-period 2002–2012 since merely four to five points were available during 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 anomalies improved to $D_{E,S} = -1.0$ (Table 7) which remained negative due to the low model performance metrics with respect to the annual

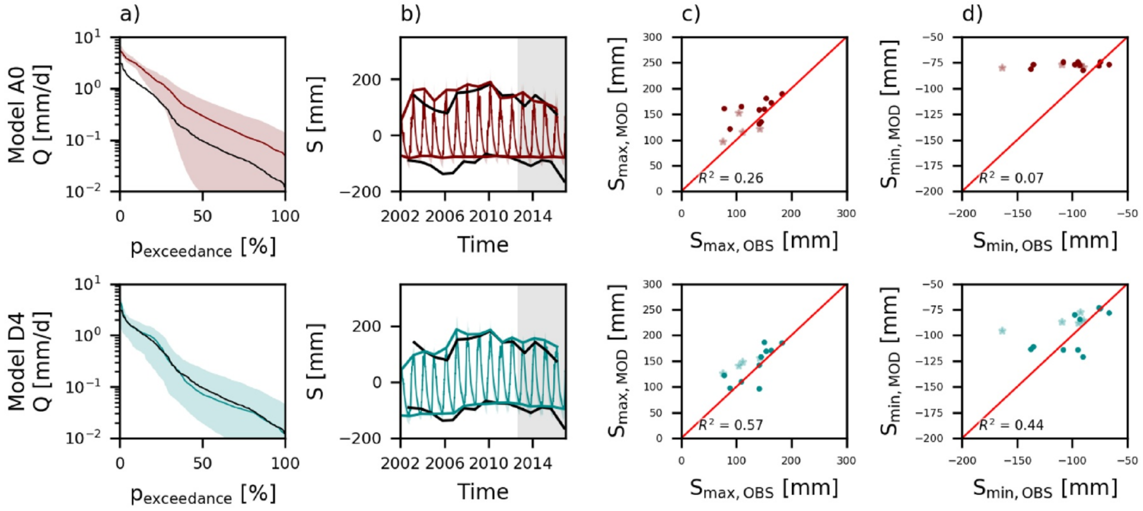

Figure 12. Range of model solutions for Models A0 and D4 for calibration strategy 2 with respect to (a) flow duration curve, (b) total water storage anomaly time-series, (c) annual maximum total water storage anomalies, and (d) annual minimum total water storage anomalies. In (a and b), the black line indicates the recorded data, the colored line the solution with the highest calibration objective function with respect to discharge and total water storage anomalies ($D_{E,QS}$) and the shaded area the envelope of the solutions retained as feasible. The white area was used for calibration (2002–2012) and the gray area for evaluation (2012–2016). In (c and 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 anomalies ($D_{E,QS}$). 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 and R^2 is the correlation with respect to a fitted regression line.

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 (Equation 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 anomalies and discharge time-series. This Model D4 allowed for a better representation of the discharge and better prediction of the total water storage anomalies 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).

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 modeled annual maximum storage fluctuations were to a large extent controlled by the choice of forcing data, whereas modeled annual minima were influenced by pro-

Table 7
Model Performance With Respect to Total Water Storage Anomalies and Discharge ($D_{E,QS}$), and Total Water Storage Anomalies ($D_{E,S}$) 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
	$D_{E,QS}$ ($D_{E,QS,5/95\%}$)	$D_{E,Q}$ ($D_{E,Q,5/95\%}$)	$D_{E,S}$ ($D_{E,S,5/95\%}$)	$D_{E,S}$ ($D_{E,S,5/95\%}$)
Model A0	−0.29 (−0.71 to −0.10)	−0.13 (−0.76 to −0.11)	−0.21 (−0.51 to −0.33)	−2.7 (−6.2 to −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)

cesses generating long-term memory effects which are missing in the original benchmark Model A0 and in many other common hydrological models (cf. Fowler et al., 2020). More specifically, the representation of monthly and annual total water storage fluctuations improved when using TRMM precipitation data, the Thornthwaite method to estimate potential evaporation and allowing for groundwater export via a loss from a deeper groundwater layer (Model D4). In 2005–2007, most models poorly reproduced the annual maximum/minimum total water storage anomalies. This could be further improved in future studies by testing alternative forcing data sources, model formulations for groundwater loss and calibration strategies. Depending on the model, the modeled total water storage anomalies improved considerably when calibrating only with respect to this variable, but at the cost of decreased discharge performances (Figure S21 in the Supplementary Material).

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 anomalies in the calibration procedure only improved the modeled 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, that is, 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. The average groundwater loss/flow reached up to 0.68 mm d^{-1} for Models A1–A5 and D4 when considering the optimal and feasible parameter sets. However, in many commonly used hydrological models such processes allowing long-term memory effects are missing (e.g., Bergström, 1992; Fenicia et al., 2014; Liang et al., 1994) resulting in biased predictions of discharge and storage which is especially crucial during extreme dry conditions (Fowler et al., 2020; Saft et al., 2016).

Furthermore, this study showed that simple process formulations allowing for long-term memory effects can be readily 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 ($D_{\mathrm{E,S}} < 0$, 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 Section 5.2.4. In addition, the above model modifications also improved the model's skill to reproduce observed discharge time-series such that the general dynamics and magnitudes were represented better with Model D4 (Figure 11) compared to the benchmark Model A0 (Figure 7). Overall, the results suggest that the model hypotheses A0–D0 as well as A1–A5 can be rejected in favor of hypothesis D4. This underlines the crucial role of model hypothesis testing for improving the simultaneous representation of multiple variables in models and thereby providing evidence that the process representation in D4 is likely a more consistent representation of real-world processes than the other hypotheses tested here (Beven, 2018; Clark et al., 2011).

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 (Fowler et al., 2018; Jing et al., 2019; Saft et al., 2016; Scanlon et al., 2018; Winsemius et al., 2006). 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, iterative data and model selection procedure, the representation of long-term storage variations in a conceptual model can 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, that is, 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 *per se*. Instead, our results provide 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).

The (satellite) observations used in this study are prone to data uncertainties. Uncertainties in GRACE observations are a result of data (post-) processing which includes data smoothening with a radius of 300 km (Landerer & Swenson, 2012). This results in signal leakage between neighboring cells of 1° and in grid cells which are not completely independent from each other. In addition, data gaps occur in GRACE observations due to instrument issues, calibration campaigns and battery management activities every 6 months since 2011 (Figure S3b in the Supplementary Material). Uncertainties in precipitation data can be considerable, in particular for extreme events on small scale or in mountainous regions (Beck et al., 2020; Hrachowitz & Weiler, 2011; Kimani et al., 2017; Le Coz & van de Giesen, 2019). As shown in Figure 4 for CHIRPS and TRMM, different methods and input data underlying satellite-based precipitation products affect the long-term patterns and hence the modeled long-term total water storage variations. In addition, bias errors in the precipitation affect the estimated long-term average groundwater export based on the water balance (Liu et al., 2020). Similarly, satellite-based evaporation data are a result of models with uncertainties in the input data, parameterization or model conceptualization (K. Zhang et al., 2016). This affects the monthly values and the long-term patterns as shown in Figure S4 in the Supplementary Material. In Section 5.1.3, it was illustrated that daily deviations within the uncertainty range of many evaporation satellite products can result in a considerable storage change of over multiple years. In addition, uncertainties in the evaporation affect the long-term water balance closure and hence also the estimated long-term groundwater export/import. Uncertainties in the potential evaporation are a result of the underlying equations and input data. This study compared the Hargreaves and Thornthwaite methods which both use temperature data to estimate potential evaporation but with different equations (Hargreaves & Allen, 2003; Hargreaves & Samani, 1985; Maes et al., 2019). This resulted in different monthly values and long-term variations as shown in Figure 5. Uncertainties in the potential evaporation affect the modeled actual evaporation especially during wet seasons when the total evaporation is limited by the energy, whereas during dry seasons the total evaporation is limited by the water availability. Discharge uncertainties are a result of rating curve uncertainties (Domeneghetti et al., 2012; McMillan & Westerberg, 2015; Westerberg et al., 2011) and limited data availability. Due to the limited data availability in this study, it was not possible to validate these observations with field measurements to estimate the magnitude of the uncertainties.

According to the International Groundwater Resources Assessment Centre (IGRAC), there are two aquifers which are shared by the Luangwa basin and neighboring basins (Figure S22 in the Supplementary Material). These aquifers are located in the South upstream of the gauge station and in the East at the border with Malawi and mostly consist of alluvial sediments/sands and fractured crystalline - metamorphic basement rocks, respectively (IGRAC & UNESCO-IHP, 2015; TWAP, 2015a, 2015b). Both aquifers were identified in previous studies as part of an effort to identify transboundary aquifers world-wide to support transboundary aquifer management activities. Additional studies are needed to characterize these aquifers more detailed and to analyze whether there are additional aquifers shared by Luangwa and surrounding river basins within Zambia. According to Fraser et al. (2020), who identified transboundary aquifers in Malawi based on lithology, hydrogeology, groundwater levels and literature, water flows from Zambia to Malawi in the aquifer at the eastern border of the Luangwa basin. This supports our findings with respect to the significance of potential groundwater leakage in the Luangwa basin.

In addition, Tóth (1963) illustrated groundwater flow occurs on local or regional scale depending on the topography such that local groundwater flow generally occurs in regions with large local relief and shallow aquifers, whereas regional groundwater flow generally occurs in regions with large regional relief and deep aquifers. In the Luangwa river basin, elevation differences are less pronounced near the eastern and southern border compared to the North and West (Figure S23 in the Supplementary Material) which supports the possibility of regional groundwater flow in the East and South. Schaller and Fan (2009) illustrated regional groundwater flow often occur in arid regions as the groundwater level often remains below the local topography. This further supports the possibility of regional groundwater flow in the semi-arid Luangwa basin. According to Condon et al. (2020), conceptual hydrological models typically focus on the shallow groundwater system assuming the deep groundwater system is negligible even though deep flow paths can be relevant depending on the river basin and research question. This results in for example open water balances and flawed conceptual models (Condon et al., 2020). Our study illustrated that the addition of a Deeper Groundwater Reservoir with groundwater loss has the potential to substantially improve the modeled discharge and long-term total water storage anomalies in conceptual hydrological models.

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 (Hobbins et al., 2008; Huang et al., 2015; Roderick & Farquhar, 2005; 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 (Bai et al., 2019; Feng et al., 2019; Goroshi et al., 2017; Wang et al., 2018). As shown in a previous study by Hulsman, Savenije, et al. (2020) and in Figure S24 in the Supplementary Material, the basin-averaged evaporation was modeled well and incorporating this flux in the calibration procedure did not improve the modeled long-term storage variabilities. That is why, more in-depth analyses on the occurrence of long-term variations in the actual and potential evaporation, their main drivers and their effect on long-term storage fluctuations testing different model hypotheses is recommended. This was outside the scope of this study due to the limited data availability.

Furthermore, the calibration approach used in this study allowed to analyze the influence of different model adjustments on the behavioral parameter sets. We assumed that if a model adjustment is relevant, a clear improvement in the distribution of performances for the parameter sets retained as feasible should be visible. Other calibration schemes are recommended when attempting to identify the mathematically optimal parameter set. However, a solution that may mathematically be the best fit, is unlikely the most plausible representation of the real-world system given the many sources of uncertainty in the modeling process especially in data scarce regions (e.g., Beven, 2006). Sensitivity analysis of the parameters related to the groundwater loss indicated that these parameters influenced the modeled total water storage significantly with some exceptions, whereas the impact on the discharge was mostly limited (Figure S25). We recommend using additional information sources to constrain these parameters, which was outside the scope of this study due to limited data availability. As this study focused on a necessarily limited number of model hypotheses, it should be noted that additional alternative model hypotheses clearly may lead to similar model improvements. This also includes alternative hypotheses with respect to the conceptualization of the groundwater system (e.g., de Graaf et al., 2015; Reinecke et al., 2019; Stoelzle et al., 2015). The results of our study therefore need to be understood in that context. Model hypothesis D4 allowed the rejection of all other hypotheses tested here, yet it may in the future be rejected itself in favor of another hypothesis.

7. Conclusion

In the Luangwa basin, long-term total water storage variations were observed with GRACE, but not reproduced by a previously developed 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. The data analysis results revealed different long-term patterns in the precipitation, potential and actual evaporation depending on the satellite product which could partly explain the observed long-term storage variations. The results of the model hypotheses testing 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 maximum total water storage anomalies. In contrast, the incorporation of a process representing regional groundwater export from a deep groundwater layer significantly improved the model's ability to reproduce the observed long-term 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 toward more reliable predictions of hydrological models.

Data Availability Statement

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 publicly available online databases as described in Section 3 and Table 1.

Acknowledgment
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.

References

Addor, N., & Melsen, L. A. (2019). Legacy, rather than adequacy, drives the selection of hydrological models. *Water Resources Research*, *55*, 378–390. https://doi.org/10.1029/2018wr022958

Allen, R. G., Tasumi, M., & Trezza, R. (2007). 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)

Awange, J. L., Khandu, M., Schumacher, M., Forootan, E., & Heck, B. (2016). 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

Bai, M., Shen, B., Song, X., Mo, S., Huang, L., & Quan, Q. (2019). 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

Bastiaanssen, W. G. M., Menenti, M., Feddes, R. A., & Holtslag, A. A. M. (1998). A remote sensing surface energy balance algorithm for land (SEBAL). 1. Formulation. *Journal of Hydrology*, *212–213*, 198–212. https://doi.org/10.1016/s0022-1694(98)00253-4

Beck, H. E., Wood, E. F., McVicar, T. R., Zambrano-Bigiarini, M., Alvarez-Garreton, C., Baez-Villanueva, O. M., et al. (2020). Bias correction of global high-resolution precipitation climatologies using streamflow observations from 9372 catchments. *Journal of Climate*, *33*, 1299–1315. https://doi.org/10.1175/JCLI-D-19-033210.1175/jcli-d-19-0332.1

Bergström, S. (1992). The HBV model—Its structure and applications (32). Sweden: SMHI Norrköping.

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

Beven, K. J. (2018). 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

Blazquez, A., Meyssignac, B., Lemoine, J., Berthier, E., Ribes, A., & Cazenave, A. (2018). 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. https://doi.org/10.1093/gji/ggy293

Bonsor, H. C., Shamsudduha, M., Marchant, B. P., MacDonald, A. M., & Taylor, R. G. (2018). 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

Bouaziz, L. J. E., Thirel, G., de Boer-Euser, T., Melsen, L. A., Buitink, J., Brauer, C. C., et al. (2020). Behind the scenes of streamflow model performance. *Hydrology and Earth System Sciences Discussions*, 1–38. https://doi.org/10.5194/hess-2020-176

Bouaziz, L. J. E., Weerts, A., Schellekens, J., Sprokkereef, E., Stam, J., Savenije, H., & Hrachowitz, M. (2018). Redressing the balance: Quantifying net intercatchment groundwater flows. *Hydrology and Earth System Sciences*, *22*, 6415–6434. https://doi.org/10.5194/hess-22-6415-2018

Boutt, D. F. (2017). 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

Budyko, M. I. (1974). *Climate and life* (p. 508). New York, NY: Academic Press.

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

Chao, N., Wang, Z., Jiang, W., & Chao, D. (2016). 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

Chen, J. L., Wilson, C. R., Tapley, B. D., Longuevergne, L., Yang, Z. L., & Scanlon, B. R. (2010). Recent La Plata basin drought conditions observed by satellite gravimetry. *Journal of Geophysical Research*, *115*, D22108. https://doi.org/10.1029/2010jd014689

Chen, J. L., Wilson, C. R., Tapley, B. D., Scanlon, B., & Güntner, A. (2016). 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

Clark, M. P., Kavetski, D., & Fenicia, F. (2011). Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resources Research*, *47*, W09301. https://doi.org/10.1029/2010wr009827

Cohen Liechti, T., Matos, J. P., Boillat, J.-L., & Schleiss, A. J. (2012). Comparison and evaluation of satellite derived precipitation products for hydrological modeling of the Zambezi River Basin. *Hydrology and Earth System Sciences*, *16*, 489–500. https://doi.org/10.5194/hess-16-489-2012

Condon, L. E., Markovich, K. H., Kelleher, C. A., McDonnell, J. J., Ferguson, G., & McIntosh, J. C. (2020). Where is the bottom of a watershed? *Water Resources Research*, *56*, e2019WR026010. https://doi.org/10.1029/2019wr026010

Danielson, J. J., & Gesch, D. B. (2011). *Global multi-resolution terrain elevation data 2010 (GMTED2010)* (Open-file report 2011-1073). Reston, VA: U.S. Geological Survey. https://doi.org/10.3133/ofr20111073

de Graaf, I. E. M., Sutanudjaja, E. H., van Beek, L. P. H., & Bierkens, M. F. P. (2015). A high-resolution global-scale groundwater model. *Hydrology and Earth System Sciences*, *19*, 823–837. https://doi.org/10.5194/hess-19-823-2015

Domeneghetti, A., Castellarin, A., & Brath, A. (2012). Assessing rating-curve uncertainty and its effects on hydraulic model calibration. *Hydrology and Earth System Sciences*, *16*, 1191–1202. https://doi.org/10.5194/hess-16-1191-2012

Euser, T., Hrachowitz, M., Winsemius, H. C., & Savenije, H. H. G. (2015). 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

Euser, T., Winsemius, H. C., Hrachowitz, M., Fenicia, F., Uhlenbrook, S., & Savenije, H. H. G. (2013). 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

FAO. (2018). *WaPOR database methodology: Level 1* (Remote sensing for water productivity technical report: Methodology series, p. 72). Rome: FAO. Retrieved from http://www.fao.org/3/I7315EN/i7315en.pdf

FAO, IHE Delft. (2019). *WaPOR quality assessment* (Technical report on the data quality of the WaPOR FAO database version 1.0, p. 134). Rome: FAO and IHE Delft. Retrieved from http://www.fao.org/3/ca4895en/CA4895EN.pdf

Feng, T., Su, T., Zhi, R., Tu, G., & Ji, F. (2019). 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

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

Fowler, K., Coxon, G., Freer, J., Peel, M., Wagener, T., Western, A., et al. (2018). Simulating runoff under changing climatic conditions: A framework for model improvement. *Water Resources Research*, *54*, 9812–9832. https://doi.org/10.1029/2018wr023989

Fowler, K., Knoben, W., Peel, M., Peterson, T., Ryu, D., Saft, M., et al. (2020). 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

Fraser, C. M., Kalin, R. M., Kanjaye, M., & Uka, Z. (2020). A national border-based assessment of Malawi's transboundary aquifer units: Towards achieving sustainable development goal 6.5.2. *Journal of Hydrology: Regional Studies*, *31*, 100726. https://doi.org/10.1016/j.ejrh.2020.100726

Funk, C. C., Peterson, P. J., Landsfeld, M. F., Pedreros, D. H., Verdin, J. P., Rowland, J. D., et al. (2014). A quasi-global precipitation time series for drought monitoring (Data Series 832, 4). South Dakota: U.S. Geological Survey. https://doi.org/10.3133/ds832

Gallart, F., & Llorens, P. (2003). Catchment management under environmental change: Impact of land cover change on water resources. *Water International*, *28*, 334–340. https://doi.org/10.1080/02508060308691707

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

Gerrits, A. M. J., Savenije, H. H. G., Veling, E. J. M., & Pfister, L. (2009). 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

Gharari, S., Hrachowitz, M., Fenicia, F., Gao, H., & Savenije, H. H. G. (2014). Using expert knowledge to increase realism in environmental system models can dramatically reduce the need for calibration. *Hydrology and Earth System Sciences*, *18*, 4839–4859. https://doi.org/10.5194/hess-18-4839-2014

Goroshi, S., Pradhan, R., Singh, R. P., Singh, K. K., & Parihar, J. S. (2017). 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

Goswami, M., O'Connor, K. M., & Bhattarai, K. P. (2007). 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

Grigg, A. H., & Hughes, J. D. (2018). 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

Handavu, F., Chirwa, P. W. C., & Syampungani, S. (2019). 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

Hargreaves, G. H., & Allen, R. G. (2003). 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)

Hargreaves, G. H., & Samani, Z. A. (1985). Reference crop evapotranspiration from temperature. *Applied Engineering in Agriculture*, *1*, 96–99. https://doi.org/10.13031/2013.26773

Hartmann, J., & Moosdorf, N. (2012). The new global lithological map database GLiM: A representation of rock properties at the Earth surface. *Geochemistry, Geophysics, Geosystems*, *13*. https://doi.org/10.1029/2012gc004370

Hobbins, M. T., Dai, A., Roderick, M. L., & Farquhar, G. D. (2008). Revisiting the parameterization of potential evaporation as a driver of long-term water balance trends. *Geophysical Research Letters*, *35*. https://doi.org/10.1029/2008gl033840

Hrachowitz, M., Fovet, O., Ruiz, L., Euser, T., Gharari, S., Nijzink, R., et al. (2014). 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

Hrachowitz, M., Stockinger, M., Coenders-Gerrits, M., van der Ent, R., Bogena, H., Lücke, A., & Stumpp, C. (2020). Deforestation reduces the vegetation-accessible water storage in the unsaturated soil and affects catchment travel time distributions and young water fractions. *Hydrology and Earth System Sciences Discussions*, 1–43. https://doi.org/10.5194/hess-2020-293

Hrachowitz, M., & Weiler, M. (2011). Uncertainty of precipitation estimates caused by sparse gauging networks in a small, mountainous watershed. *Journal of Hydrologic Engineering*, *16*, 460–471. https://doi.org/10.1061/10.1061/(asce)he.1943-5584

Huang, H., Han, Y., Cao, M., Song, J., Xiao, H., & Cheng, W. (2015). Spatiotemporal characteristics of evapotranspiration paradox and impact factors in China in the period of 1960–2013. *Advances in Meteorology*, 519207. https://doi.org/10.1155/2015/519207

Huffman, G. J., Adler, R. F., Rudolf, B., Schneider, U., & Keehn, P. R. (1995). 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

Huffman, G. J., Bolvin, D. T., Nelkin, E. J., Wolff, D. B., Adler, R. F., Gu, G., et al. (2007). 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

Huffman, G. J., Stocker, E. F., Bolvin, D. T., & Nelkin, E. J. (2014). *TRMM 3B43V7 data sets*. Greenbelt, MD: Goddard Earth Sciences Data and Information Services Center (GES DISC). https://doi.org/10.5067/TRMM/TMPA/MONTH/7

Hulsman, P., Savenije, H. H. G., & Hrachowitz, M. (2020). Learning from satellite observations: increased understanding of catchment processes through stepwise model improvement. *Hydrology and Earth System Sciences Discussions*, 1–26. https://doi.org/10.5194/hess-2020-191

Hulsman, P., Savenije, H. H. G., & Hrachowitz, M. (2021). Satellite-based drought analysis in the Zambezi River Basin: Was the 2019 drought the most extreme in several decades as locally perceived? *Journal of Hydrology: Regional Studies*, *34*, 100789. https://doi.org/10.1016/j.ejrh.2021.100789

Hulsman, P., Winsemius, H. C., Michailovsky, C. I., Savenije, H. H. G., & Hrachowitz, M. (2020). Using altimetry observations combined with GRACE to select parameter sets of a hydrological model in a data-scarce region. *Hydrology and Earth System Sciences*, *24*, 3331–3359. https://doi.org/10.5194/hess-24-3331-2020

IGRAC (International Groundwater Resources Assessment Centre), & UNESCO-IHP (UNESCO International Hydrological Programme). (2015). *Transboundary aquifers of the world [map], revised 2020, scale 1:50 000 000*. Delft, The Netherlands. Retrieved from https://www.un-igrac.org/resource/transboundary-aquifers-world-map-2015

Istanbulluoglu, E., Wang, T., Wright, O. M., & Lenters, J. D. (2012). Interpretation of hydrologic trends from a water balance perspective: The role of groundwater storage in the Budyko hypothesis. *Water Resources Research*, *48*. https://doi.org/10.1029/2010wr010100

Jing, W., Yao, L., Zhao, X., Zhang, P., Liu, Y., Xia, X., et al. (2019). 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

Joodaki, G., Wahr, J., & Swenson, S. (2014). 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

Khaki, M., Forootan, E., Kuhn, M., Awange, J., van Dijk, A. I. J. M., Schumacher, M., & Sharifi, M. A. (2018). 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

Kimani, W. M., Hoedjes, C. B. J., & Su, Z. (2017). An assessment of satellite-derived rainfall products relative to ground observations over East Africa. *Remote Sensing*, *9*. https://doi.org/10.3390/rs9050430

Landerer, F. W., & Swenson, S. C. (2012). Accuracy of scaled GRACE terrestrial water storage estimates. *Water Resources Research*, *48*, W04531. https://doi.org/10.1029/2011wr011453

Leblanc, M. J., Tregoning, P., Ramillien, G., Tweed, S. O., & Fakes, A. (2009). 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

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

Le Moine, N., Andréassian, V., Perrin, C., & Michel, C. (2007). 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

Liang, X., Lettenmaier, D. P., Wood, E. F., & Burges, S. J. (1994). A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *Journal of Geophysical Research*, *99*, 14415–14428. https://doi.org/10.1029/94jd00483

Li, C., Wu, P., Li, X., Zhou, T., Sun, S., Wang, Y., et al. (2017). Spatial and temporal evolution of climatic factors and its impacts on potential evapotranspiration in Loess Plateau of Northern Shaanxi, China. *The Science of the Total Environment*, *589*, 165–172. https://doi.org/10.1016/j.scitotenv.2017.02.122

Liu, Y., Wagener, T., Beck, H. E., & Hartmann, A. (2020). What is the hydrologically effective area of a catchment? *Environmental Research Letters*, *15*, 104024. https://doi.org/10.1088/1748-9326/aba7e5

Long, D., Longuevergne, L., & Scanlon, B. R. (2014). Uncertainty in evapotranspiration from land surface modeling, remote sensing, and GRACE satellites. *Water Resources Research*, *50*, 1131–1151. https://doi.org/10.1002/2013wr014581

Long, D., Pan, Y., Zhou, J., Chen, Y., Hou, X., Hong, Y., et al. (2017). 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

Long, D., Scanlon, B. R., Longuevergne, L., Sun, A. Y., Fernando, D. N., & Save, H. (2013). 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

Maes, W. H., Gentine, P., Verhoest, N. E. C., & Miralles, D. G. (2019). Potential evaporation at eddy-covariance sites across the globe. *Hydrology and Earth System Sciences*, *23*, 925–948. https://doi.org/10.5194/hess-23-925-2019

Martens, B., Miralles, D. G., Lievens, H., van der Schalie, R., de Jeu, R. A. M., Fernández-Prieto, D., et al. (2017). GLEAM v3: Satellite-based land evaporation and root-zone soil moisture. *Geoscientific Model Development*, *10*, 1903–1925. https://doi.org/10.5194/gmd-10-1903-2017

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

McMillan, H. K., & Westerberg, I. K. (2015). Rating curve estimation under epistemic uncertainty. *Hydrological Processes*, *29*, 1873–1882. https://doi.org/10.1002/hyp.10419

Meng, F., Su, F., Li, Y., & Tong, K. (2019). 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

Miralles, D. G., Holmes, T. R. H., De Jeu, R. A. M., Gash, J. H., Meesters, A. G. C. A., & Dolman, A. J. (2011). Global land-surface evaporation estimated from satellite-based observations. *Hydrology and Earth System Sciences*, *15*, 453–469. https://doi.org/10.5194/hess-15-453-2011

Nearing, G. S., Tian, Y., Gupta, H. V., Clark, M. P., Harrison, K. W., & Weijs, S. V. (2016). A philosophical basis for hydrological uncertainty. *Hydrological Sciences Journal*, *61*, 1666–1678. https://doi.org/10.1080/02626667.2016.1183009

Nelson, S. T., & Mayo, A. L. (2014). 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

Nijzink, R. C., Samaniego, L., Mai, J., Kumar, R., Thober, S., Zink, M., et al. (2016). The importance of topography-controlled sub-grid process heterogeneity and semi-quantitative prior constraints in distributed hydrological models. *Hydrology and Earth System Sciences*, *20*, 1151–1176. https://doi.org/10.5194/hess-20-1151-2016

Ó Dochartaigh, B. E. (2019). *User guide: Africa groundwater atlas country hydrogeology maps, version 1.1* (Open report OR/19/035, 21). Nottingham, UK: British Geological Survey. Retrieved from http://nora.nerc.ac.uk/id/eprint/523272/

Oguntunde, P. G., Friesen, J., van de Giesen, N., & Savenije, H. H. G. (2006). 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

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

Pellicer-Martínez, F., & Martínez-Paz, J. M. (2014). 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

Perrin, C., Michel, C., & Andréassian, V. (2003). Improvement of a parsimonious model for streamflow simulation. *Journal of Hydrology*, *279*, 275–289. https://doi.org/10.1016/s0022-1694(03)00225-7

Phiri, D., Morgenroth, J., & Xu, C. (2019a). 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

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

Pike, J. G. (1964). 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

Reinecke, R., Foglia, L., Mehl, S., Trautmann, T., Cáceres, D., & Döll, P. (2019). Challenges in developing a global gradient-based groundwater model (G3M v1.0) for the integration into a global hydrological model. *Geoscientific Model Development*, *12*, 2401–2418. https://doi.org/10.5194/gmd-12-2401-2019

Rennó, C. D., Nobre, A. D., Cuartas, L. A., Soares, J. V., Hodnett, M. G., Tomasella, J., & Waterloo, M. J. (2008). 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

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

Running, S., Mu, Q., & Zhao, M. (2017). *MOD16A2 MODIS/Terra net evapotranspiration 8-day L4 global 500m SIN grid V006*: NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/MODIS/MOD16A2.006

Saft, M., Peel, M. C., Western, A. W., Perraud, J. M., & Zhang, L. (2016). Bias in streamflow projections due to climate-induced shifts in catchment response. *Geophysical Research Letters*, *43*, 1574–1581. https://doi.org/10.1002/2015gl067326

Samaniego, L., Kumar, R., & Jackisch, C. (2011). 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

Savenije, H. H. G. (2010). HESS opinions "topography driven conceptual modelling (FLEX-Topo)". *Hydrology and Earth System Sciences*, *14*, 2681–2692. https://doi.org/10.5194/hess-14-2681-2010

Scanlon, B. R., Zhang, Z., Save, H., Sun, A. Y., Müller Schmied, H., van Beek, L. P. H., et al. (2018). Global models underestimate large decadal declining and rising water storage trends relative to GRACE satellite data. *Proceedings of the National Academy of Sciences of the United States of America*, *115*, E1080. https://doi.org/10.1073/pnas.1704665115

Schaller, M. F., & Fan, Y. (2009). River basins as groundwater exporters and importers: Implications for water cycle and climate modeling. *Journal of Geophysical Research*, *114*. https://doi.org/10.1029/2008jd010636

Schreiber, P. (1904). Über die Beziehungen zwischen dem Niederschlag und der Wasserführung der Flüsse in Mitteleuropa. *Zeitschrift für Meteorologie*, *21*, 441–452.

Schumacher, M., Forootan, E., van Dijk, A. I. J. M., Müller Schmied, H., Crosbie, R. S., Kusche, J., & Döll, P. (2018). 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

Schwatke, C., Dettmering, D., Bosch, W., & Seitz, F. (2015). DAHITI—An innovative approach for estimating water level time series over inland waters using multi-mission satellite altimetry. *Hydrology and Earth System Sciences*, *19*, 4345–4364. https://doi.org/10.5194/hess-19-4345-2015

Senay, G. B., Budde, M., Verdin, J. P., & Melesse, A. M. (2007). A coupled remote sensing and simplified surface energy balance approach to estimate actual evapotranspiration from irrigated fields. *Sensors*, *7*, 979–1000. https://doi.org/10.3390/s7060979

Stoelzle, M., Weiler, M., Stahl, K., Morhard, A., & Schuetz, T. (2015). Is there a superior conceptual groundwater model structure for baseflow simulation? *Hydrological Processes*, *29*, 1301–1313. https://doi.org/10.1002/hyp.10251

Sun, Z., Zhu, X., Pan, Y., Zhang, J., & Liu, X. (2018). Drought evaluation using the GRACE terrestrial water storage deficit over the Yangtze River Basin, China. *The Science of the Total Environment*, *634*, 727–738. https://doi.org/10.1016/j.scitotenv.2018.03.292

Su, Z. (2002). The Surface Energy Balance System (SEBS) for estimation of turbulent heat fluxes. *Hydrology and Earth System Sciences*, *6*, 85–100. https://doi.org/10.5194/hess-6-85-2002

Swenson, S. C. (2012). *GRACE monthly land water mass grids NETCDF RELEASE 5.0*. CA: PO.DAAC. https://doi.org/10.5067/TELND-NC005

Swenson, S. C., & Wahr, J. (2006). Post-processing removal of correlated errors in GRACE data. *Geophysical Research Letters*, *33*, L08402. https://doi.org/10.1029/2005GL025285

Tangdamrongsub, N., Han, S.-C., Tian, S., Müller Schmied, H., Sutanudjaja, E. H., Ran, J., & Feng, W. (2018). 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

The World Bank. (2010). The Zambezi River Basin: A multi-sector investment opportunities analysis (3). Washington DC: State of the Basin, The International Bank for Reconstruction and Development, The World Bank.

Thiemig, V., Rojas, R., Zambrano-Bigiarini, M., Levizzani, V., & De Roo, A. (2012). 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

Tóth, J. (1963). A theoretical analysis of groundwater flow in small drainage basins. *Journal of Geophysical Research*, *68*, 4795–4812. https://doi.org/10.1029/JZ068i016p04795

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

TWAP. (2015a). *Information sheet: AF18—Arangua alluvial*. Retrieved from https://services.geodan.nl/public/document/AGRC0001XXXX/api/data/AGRC0001XXXX/mim/AF18_20150911.pdf_fpg6vamr2

TWAP. (2015b). *Information sheet: AF24—Weathered basement*. Retrieved from https://services.geodan.nl/public/document/AGRC0001XXXX/api/data/AGRC0001XXXX/mim/AF24_20150911.pdf_h6ojjjme3

University of East Anglia Climatic Research Unit, Harris, I. C., & Jones, P. D. (2017). *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*: Centre for Environmental Data Analysis. https://doi.org/10.5285/58a8802721c94c66ae45c3baa4d814d0

van Dijk, A. I. J. M., Beck, H. E., Crosbie, R. S., de Jeu, R. A. M., Liu, Y. Y., Podger, G. M., et al. (2013). 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

Vishwakarma, B., Devaraju, B., & Sneeuw, N. (2018). What is the spatial resolution of grace satellite products for hydrology? *Remote Sensing*, *10*, 852. https://doi.org/10.3390/rs10060852

Wahr, J., Molenaar, M., & Bryan, F. (1998). Time variability of the Earth's gravity field: Hydrological and oceanic effects and their possible detection using GRACE. *Journal of Geophysical Research*, *103*, 30205–30229. https://doi.org/10.1029/98jb02844

Wang-Erlandsson, L., Bastiaanssen, W. G. M., Gao, H., Jägermeyr, J., Senay, G. B., van Dijk, A. I. J. M., et al. (2016). Global root zone storage capacity from satellite-based evaporation. *Hydrology and Earth System Sciences*, *20*, 1459–1481. https://doi.org/10.5194/hess-20-1459-2016

Wang, W., Li, J., Yu, Z., Ding, Y., Xing, W., & Lu, W. (2018). 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

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

Werth, S., White, D., & Bliss, D. W. (2017). 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

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

Westerhoff, R. S. (2015). 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

Willems, P. (2014). 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

Winsemius, H. C., Savenije, H. H. G., van de Giesen, N. C., van den Hurk, B. J. J. M., Zapreeva, E. A., & Klees, R. (2006). 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

Xu, S., Yu, Z., Yang, C., Ji, X., & Zhang, K. (2018). 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

Zhang, D., Zhang, Q., Werner, A. D., & Liu, X. (2015). 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

Zhang, J., Liu, K., & Wang, M. (2020). 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

Zhang, K., Kimball, J. S., & Running, S. W. (2016). A review of remote sensing based actual evapotranspiration estimation. *Wiley Interdisciplinary Reviews: Water*, *3*, 834–853. https://doi.org/10.1002/wat2.1168

Zhang, Z., Chao, B. F., Chen, J., & Wilson, C. R. (2015). 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

Zhao, M., Geruo, A., Velicogna, I., & Kimball, J. S. (2017a). 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

Zhao, M., Geruo, A., Velicogna, I., & Kimball, J. S. (2017b). 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