GRACEfully closing the water balance: a data-driven probabilistic approach applied to river basins in Iran

Gerrit Schoups[1] and Mohsen Nasseri[2]

[1]Department of Water Management
[2]School of Civil Engineering

November 24, 2022

Abstract

To fully benefit from remotely sensed observations of the terrestrial water cycle, bias and random errors in these datasets need to be quantified. This paper presents a Bayesian hierarchical model that fuses monthly water balance data and estimates the corresponding data errors and error-corrected water balance components (precipitation, evaporation, river discharge, and water storage). The model combines monthly basin-scale water balance constraints with probabilistic data error models for each water balance variable. Each data error model includes parameters that are in turn treated as unknown random variables to reflect uncertainty in the errors. Errors in precipitation and evaporation data are parameterized as a function of multiple data sources, while errors in GRACE storage observations are described by a noisy sine wave model with parameters controlling phase, amplitude and randomness of the sine wave. Error parameters and water balance variables are estimated using a combination of Markov Chain Monte Carlo sampling and iterative smoothing. Application to semi-arid river basins in Iran yields (i) significant reductions in evaporation uncertainty during water-stressed summers, (ii) basin-specific timing and amplitude corrections of the GRACE water storage dynamics, and (iii) posterior water balance estimates with average standard errors of 4-12 mm/month for water storage, 3.5-7 mm/month for precipitation, 2-6 mm/month for evaporation, and 0-2 mm/month for river discharge. The approach is readily extended to other datasets and other (gauged) basins around the world, possibly using customized data error models. The resulting error-filtered and bias-corrected water balance estimates can be used to evaluate hydrological models.

GRACEfully closing the water balance: a data-driven probabilistic approach applied to river basins in Iran

G. Schoups[1], M. Nasseri[2,3]

[1]Department of Water Management, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, The Netherlands
[2]School of Civil Engineering, College of Engineering, University of Tehran, Tehran, Iran
[3]Visiting Researcher, Department of Water Management, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, The Netherlands

Key Points:

- A Bayesian hierarchical model fuses water balance data containing unknown bias and random errors
- The model is solved using a combination of Markov Chain Monte Carlo sampling and iterative smoothing
- Computed posteriors provide hydrologically consistent data error and water balance estimates

Corresponding author: Gerrit Schoups, `g.h.w.schoups@tudelft.nl`

Abstract

To fully benefit from remotely sensed observations of the terrestrial water cycle, bias and random errors in these datasets need to be quantified. This paper presents a Bayesian hierarchical model that fuses monthly water balance data and estimates the corresponding data errors and error-corrected water balance components (precipitation, evaporation, river discharge, and water storage). The model combines monthly basin-scale water balance constraints with probabilistic data error models for each water balance variable. Each data error model includes parameters that are in turn treated as unknown random variables to reflect uncertainty in the errors. Errors in precipitation and evaporation data are parameterized as a function of multiple data sources, while errors in GRACE storage observations are described by a noisy sine wave model with parameters controlling phase, amplitude and randomness of the sine wave. Error parameters and water balance variables are estimated using a combination of Markov Chain Monte Carlo sampling and iterative smoothing. Application to semi-arid river basins in Iran yields (i) significant reductions in evaporation uncertainty during water-stressed summers, (ii) basin-specific timing and amplitude corrections of the GRACE water storage dynamics, and (iii) posterior water balance estimates with average standard errors of 4-12 mm/month for water storage, 3.5-7 mm/month for precipitation, 2-6 mm/month for evaporation, and 0-2 mm/month for river discharge. The approach is readily extended to other datasets and other (gauged) basins around the world, possibly using customized data error models. The resulting error-filtered and bias-corrected water balance estimates can be used to evaluate hydrological models.

1 Introduction

The increasing availability and accuracy of remote sensing data of the terrestrial water cycle holds great promise for calibration and validation of large-scale hydrological models. Several modeling studies have already taken advantage of these data for evaluating and constraining hydrological models, including water storage data from GRACE satellites (L. Zhang et al., 2017; Bai et al., 2018; Scanlon et al., 2018, 2019) and satellite-based evaporation data (Rientjes et al., 2013; Lopez et al., 2017; Odusanya et al., 2019; Jiang et al., 2020). A challenge with using remotely sensed data for model evaluation is that data errors need to be properly accounted for. Data errors are due to e.g. differences in scale, errors in the retrieval algorithms, and sensor insensitivities. However, without a reference "ground-truth" dataset, these errors are difficult to quantify, thereby undercutting the potential of remote sensing data for advancing large-scale hydrology. For example, ignoring or misrepresenting systematic data errors (bias) during calibration leads to biased parameter estimates and limits learning, especially when water balance data are hydrologically inconsistent, i.e. they do not close the water balance. Furthermore, proper characterization of random errors (noise) and information content of the data is important: underestimating or even ignoring data noise may lead to overfitting, while overestimating data noise limits learning by not fully exploiting information content of the data.

Processing and use of remotely sensed water balance data therefore requires (i) a methodology for estimating systematic and random errors in the data, and (ii) a methodology that corrects bias, filters out noise, and yields a hydrologically consistent set of water balance data that closes the water balance. These are of course well-known challenges, and the following paragraphs review some of the approaches that have been proposed in the literature to tackle error estimation and correction of water balance data.

A common approach for estimating bias and random data errors of individual water balance variables is to compare the data to a reference ground-truth dataset (Moreira et al., 2019). For example, satellite-based precipitation estimates are often evaluated by using rain gauge data as ground truth (Beck et al., 2017; Massari & Maggioni, 2020), while errors in evaporation data products have been estimated by comparing to ground-based measurements from eddy covariance flux towers (Chen et al., 2016; Yang et al., 2017) and soil moisture sensors (Martens et al., 2017). Another approach to error estimation is to create a reference dataset for the variable of interest by computing it as residual of the water balance, with all other water balance components assumed known. This approach has mainly been used for evaporation (Wan et al., 2015; Liu et al., 2016; Weerasinghe et al., 2019). Regardless of the approach used for creating the reference dataset, a conceptual drawback of the "ground-truth" approach is that the "true" values are never actually measured, since no dataset or estimate is completely error-free. For example, traditional ground observations, such as rain gauges, are limited in capturing variability across large areas, whereas remote sensing data suffer from uncertainties in converting electromagnetic signals into water balance variable estimates. Nevertheless, in practice the ground-truth approach may be justified as long as errors in the reference dataset are sufficiently small relative to the data errors being estimated (Massari & Maggioni, 2020).

Alternative error estimation techniques that do not assume a reference ground-truth dataset have also been developed. The main idea is to use an ensemble of (three or more) datasets of the same water balance variable, and either estimate errors based on variability across the ensemble (Tian & Peters-Lidard, 2010; Y. Zhang et al., 2018), or based on a triple collocation or three-cornered hat method, as has been applied to precipitation (Alemohammad et al., 2015; Massari et al., 2017) and evaporation (Long et al., 2014; Khan et al., 2018) error estimation.

A separate group of studies focuses on bringing together estimates of the different water balance variables and modifying the original estimates so as to close the water balance (Pan & Wood, 2006; Sahoo et al., 2011; Pan et al., 2012; Aires, 2014; Munier et al., 2014; Wang et al., 2015; Allam et al., 2016; Simons et al., 2016; Y. Zhang et al., 2016, 2018; Pellet et al., 2019; Hobeichi et al., 2020). In closing the water balance, variables with large errors are adjusted more than variables with small errors, a process that can be formalized by what Pan and Wood (2006) called a constrained Kalman filter. A crucial input of these water balance fusion studies is therefore specification of the magnitude of errors in each water balance variable. In existing water balance fusion studies, error estimates are typically fixed a priori based on expert judgment or on results from the error estimation techniques mentioned in the previous paragraphs. However, combining error estimates from different studies for water balance closure easily leads to inconsistencies, e.g. when error estimates of the different variables are based on conflicting underlying ground-truth assumptions, or on data from different regions. Furthermore, by fixing the data errors in advance, existing water balance fusion studies forego the opportunity to improve data error estimates: as we show in this paper, the idea of estimating errors by bringing together multi-source data, as used in triple collocation for a single variable, can also be applied to water balance fusion where data on the different water balance variables are combined.

The current paper builds on previous efforts and combines the error estimation and water balance fusion steps into a single methodology that removes the need for a reference ground-truth dataset. Instead, each water balance variable is assumed to be subject to unknown bias and random errors, and a single iterative approach is used to estimate an internally consistent set of data errors and water balance variables that close the water balance. The methodology relies on the formulation of a probabilistic model that combines monthly basin-scale water balance constraints with data error models for each water balance variable. The data error models relate observations to the underlying unknown true values and contain unknown parameters to account for uncertainty in the data errors. The overall probabilistic model takes the form of a Bayesian hierarchical model with two levels of uncertainty: unknown water balance variables are constrained by probability distributions with parameters that themselves are treated as unknown random variables with specified prior distributions. After conditioning on available water balance data, posteriors of all unknowns, i.e. error parameters and water balance variables, are computed using a combination of Markov Chain Monte Carlo sampling and an iterative form of (Kalman) smoothing. The posteriors automatically fuse all available information and yield best estimates with uncertainty for all water balance variables and error parameters. We note that (Kalman) smoothing, i.e. estimating water balance variables using data from the entire time-series, has not been used in previous water balance fusion studies, which have sometimes used additional postprocessing steps to remove high-frequency artefacts in the estimates (Munier et al., 2014).

The paper starts by introducing the river basins used in this study. Water balance data for these basins is used to motivate development of the probabilistic data error models in section 3. Section 4 details how the probabilistic water balance model is solved, i.e. how posteriors of interest are computed. Section 5 then presents results of applying the methodology to river basins in Iran, followed by an evaluation of different assumptions in the analysis (section 6) and a summary of the main findings.

2 Case study: river basins in Iran

Figure 1 shows locations of the Iranian river basins used in this study. The basins were selected for their availability of river discharge data, their relatively large size, and their geographical location across the country from west to east. Basin boundaries were identified by delineating the topographically upstream areas for each stream gauge providing river discharge data (Table 1). The endorheic Jazmoorian basin drains to an internal lake without natural outlet and hence does not have a stream gauge recording outflow. The basins range in size from 1,600 to 70,000 km^2 and are generally semi-arid or arid with potential evaporation equal to 1.4 to 5 times average precipitation. Consequently, runoff ratios (Q/P in Table 1) are small, mostly 0.1 or less, with the exception of the relatively steep mountainous Karoon basin. Surface and groundwater withdrawals for irrigation are common and tend to further reduce runoff ratios. All basins have pronounced seasonality in precipitation and runoff, with relatively wet winters and dry summers, translating into seasonal wetting and drying cycles.

The generally water-stressed nature and complex topography of the selected river basins, coupled with significant interventions in the natural water cycle in the form of dams, irrigation, and groundwater pumping, provide a good test-bed for the proposed water balance methodology.

Table 1. River basin characteristics

ID	Basin	Stream gauge (°N, °E)	Area (km^2)	Elevation (m)	$\frac{E_p}{P}^*$	$\frac{Q}{P}^*$
1	Sepidrood	Gilvan (36.83, 49.02)	49246	332-3478	1.78	0.06
2	Karkheh	Abdolkhan (31.83, 48.36)	45497	36-3528	1.61	0.11
3	Karoon	Karoon-IV (32.25, 48.83)	32840	66-4199	1.36	0.38
4	Mond	Ghantareh (28.25, 51.87)	35397	68-3105	2.54	0.04
5	Jazmoorian	(endorheic)	70102	365-4226	5.04	0.00
6	Gorganrood	Bustan Dam (37.42, 55.41)	1620	85-1994	2.04	0.06

$^*P$, Q, and E_p are average precipitation, river discharge and potential evaporation

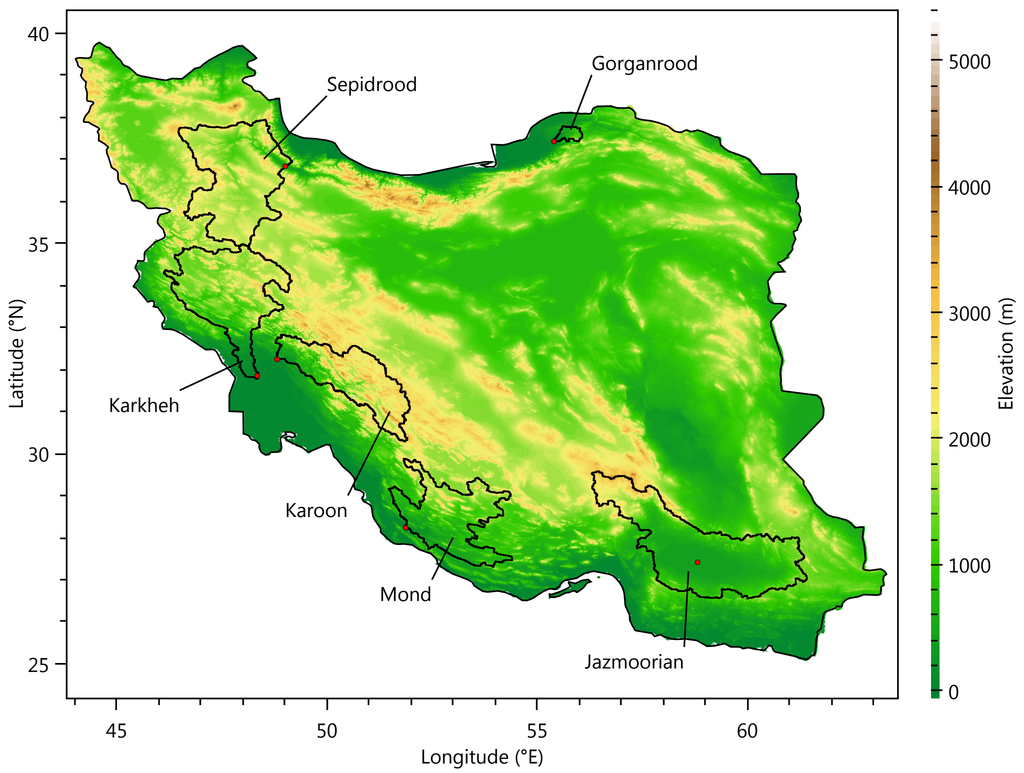

Figure 1. Topographic map of Iran with location of river basins and their outlets.

3 Probabilistic water balance model

Our interest is in estimating all terms in the monthly basin-scale water balance:

$$S_t = S_{t-1} + P_t - E_t - Q_t \tag{1}$$

where S_{t-1} and S_t are total water storage (surface and subsurface) in the basin at the start and end of month t, P_t and E_t are basin average precipitation and evaporation (including transpiration), and Q_t is river discharge at the basin outlet for month t. Each term is normalized by basin area and expressed in consistent water depth units (e.g. mm). Eq. 1 assumes negligible net lateral groundwater flow into or out of the basin. It also assumes no significant surface water flows crossing the basin boundary, except for river discharge at the basin outlet. Thus, upstream inflows and inter-basin water transfers are considered negligible, although intra-basin water transfers, e.g. via water diversions and groundwater pumping for irrigation, are captured by Eq. 1. Inter-basin water transfer is known to occur from the upstream part of Karoon basin (Fig. 1) into the semi-arid Zayanderood basin to the north; the transferred amount of water is however negligible compared to total runoff in Karoon basin (Abrishamchi & Tajrishy, 2005).

In principle, each term in Eq. 1 can be measured or estimated independently. However, bringing such independent estimates together does not typically lead to water balance closure, because all measurements and estimates are subject to systematic and random errors. Conceptually, it is then useful to distinguish between "true" and "observed" versions of each water balance variable: by definition, the true water balance variables close the water balance, and true and observed versions of each water balance variable are related via data error models that capture systematic and random deviations between observed and underlying true values.

Each data error model consists of parametric probabilistic relations between observed and true values, where parameters quantify the magnitude of systematic and random data errors. Since the magnitude of these errors is not known a priori, the parameters are themselves treated as random variables with specified prior distributions. The resulting model can hence be viewed as a Bayesian hierarchical model with two levels of uncertainty, i.e. one for error parameters and the other for water balance variables.

The monthly water balance data used here are summarized in Table 2. We follow previous water balance fusion studies and focus as much as possible on observational data instead of hydrological model outputs as source for the water balance data, thereby minimizing the impact of hydrological process assumptions. An exception is the GLEAM evaporation product, which internally relies on a soil water balance model. All data were spatially averaged across each basin to obtain monthly basin-scale data values. The following sections describe data sources and probabilistic data error models for each water balance variable (P, E, Q, S).

Table 2. Monthly water balance data

Variable	Symbol	Data source	Resolution	Reference
Precipitation	P_{obs1}	GPM IMERG Final V06B	0.1°	Huffman et al. (2019)
	P_{obs2}	CHIRPS v2.0	0.05°	Funk et al. (2014)
Evaporation	E_{obs1}	SSEBop v4	0.01°	Senay et al. (2020)
	E_{obs2}	GLEAM v3.3b	0.25°	Martens et al. (2017)
River discharge	Q_{obs}	Stream gauges	Basin	IWRMC (2020)
Storage	S_{obs}	GRACE JPL Mascon RL06v02	3°	Wiese et al. (2018)

3.1 Precipitation error model

The first dataset used is GPM IMERG (Table 2), which provides monthly precipitation values and associated standard errors. Monthly IMERG precipitation merges satellite-based estimates with the GPCC rain gauge dataset, while standard error estimates are based on the methodology of Huffman (1997). There is generally a good correspondence between IMERG and spatially interpolated rain-gauge precipitation for the basins studied here (Fig. 2, Fig. S1-S2), with the exception of Gorganrood basin. A recent evaluation of IMERG across Iran (Maghsood et al., 2020) reported small but systematic overestimation of monthly precipitation in dry regions and underestimation in the wettest parts of the country. To account for potential bias in IMERG, we included CHIRPS as a second precipitation dataset. In the semi-arid Mond basin for example (Fig. 2), CHIRPS tends to give lower precipitation than IMERG during the wet winter months.

The following error model was then used to relate observed and true precipitation:

$$m_{P,t} = (1 - w_P)P_{obs1,t} + w_P P_{obs2,t} \tag{2}$$

$$s_{P,t} = \max\left(\sigma_{P,t}, \frac{1}{2} r_P |P_{obs1,t} - P_{obs2,t}|\right) \tag{3}$$

$$P_t \sim \mathcal{N}(m_{P,t}, s^2_{P,t}) \tag{4}$$

$$P_t \geq 0 \tag{5}$$

The first equation models bias in the observations by describing prior mean precipitation $m_{P,t}$ in month t as a weighted average of IMERG ($P_{obs1,t}$) and CHIRPS ($P_{obs2,t}$) monthly basin precipitation. Parameter w_P represents the weight; since it is unknown a priori, it is given a quasi-uniform prior between 0 and 1 (specifically, a logit-normal

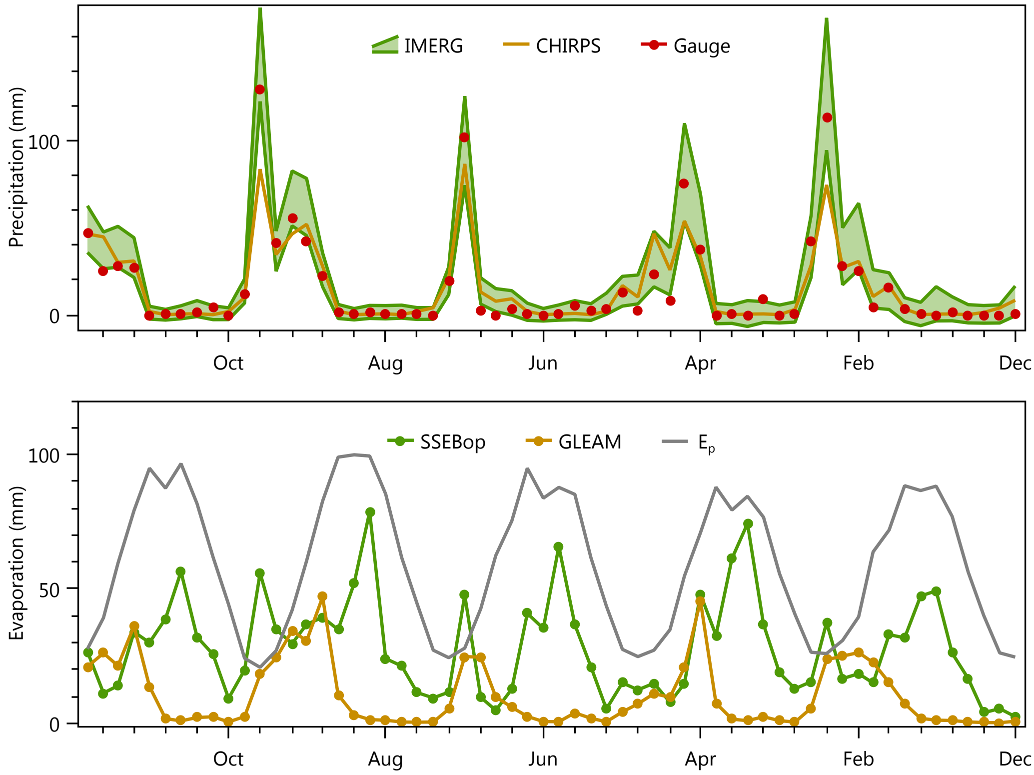

Figure 2. Monthly precipitation and evaporation data for Mond basin during 2006-2010. The IMERG data include standard errors and are plotted as 90% uncertainty bands. Spatially interpolated basin-average rain-gauge precipitation is included for comparison, but was not used in the model. Potential evaporation from the GLEAM dataset is shown as E_p.

prior with location parameter $\mu = 0$ and scale parameter $\sigma = 1.4$) to reflect prior uncertainty about the bias.

The second equation models random errors in the observations by describing prior standard deviation $s_{P,t}$ of precipitation in month t as the largest of either (i) the IMERG standard error $\sigma_{P,t}$, or (ii) the scaled absolute difference between the two precipitation datasets in each month, using r_P as the scaling parameter. The reasoning behind this is that large differences between the two datasets may not only indicate systematic but also significant random errors. Parameter r_P is given a quasi-uniform prior between 0 and 1 to reflect prior uncertainty about the relation between bias and random errors. In the limit when $r_P = 1$, the prior standard deviation is half the absolute difference between the two datasets. However, to avoid unrealistically small prior uncertainty in precipitation, e.g. when r_P is near 0 or the two datasets are in close agreement, the value of $s_{P,t}$ is not allowed to be less than the IMERG standard error $\sigma_{P,t}$. The latter is obtained by arithmetic averaging of the gridded "random error" variable in the IMERG dataset. This implicitly assumes that IMERG random errors are spatially perfectly correlated across the basin. As such, it provides a conservative estimate of the magnitude of basin-scale random errors, since averaging partially uncorrelated grid-scale random errors would result in some error cancellation and therefore smaller values for $\sigma_{P,t}$ at the basin scale.

Finally, the last two equations in the precipitation error model treat true precipitation P_t in month t as a random draw from a truncated normal distribution. Truncation at zero constrains precipitation to be non-negative.

3.2 Evaporation error model

To capture uncertainty and errors in evaporation, two different remote sensing evaporation products are used, i.e. GLEAM and SSEBop (Table 2). These datasets use different methods for estimating evaporation from remote sensing data. GLEAM uses Priestley-Taylor for potential evaporation and estimates actual evaporation as a function of microwave vegetation optical depth and soil moisture, in combination with a root-zone water balance. On the other hand, SSEBop uses Penman-Monteith for potential evaporation and estimates actual evaporation based on a surface energy balance and remotely sensed land surface temperature. For the basins studied in this paper, these two approaches translate into similar evaporation estimates under energy-limited conditions (wet winters), but significantly different evaporation estimates under water-limited conditions (dry summers). Figure 2 illustrates this for the Mond basin, with similar patterns observed in other basins (see Supporting Information): in the absence of significant rainfall during summer, GLEAM evaporation decreases to near-zero values, while SSEBop evaporation shows a peak in summer, suggesting water remains available to natural vegetation or crops (irrigation). These differences result in significant prior uncertainty in evaporation during summers.

A similar error model as for precipitation is adopted for evaporation:

$$
\begin{aligned}
m_{E,t} &= f_E\left[(1-w_E)E_{obs1,t} + w_E E_{obs2,t}\right] && (6)\\
s_{E,t} &= \max\left(0.1 m_{E,t}, \frac{1}{2} r_E |E_{obs1,t} - E_{obs2,t}|\right) && (7)\\
E_t &\sim \mathcal{N}(m_{E,t}, s_{E,t}^2) && (8)\\
E_t &\geq 0 && (9)
\end{aligned}
$$

Bias is modeled with two time-invariant parameters: w_E is a weight that interpolates between SSEBop $E_{obs1,t}$ and GLEAM $E_{obs2,t}$ evaporation, and f_E is an additional scaling factor that provides an additional degree of freedom to e.g. account for bias outside the range of the two datasets. Random errors are modeled using the same approach as for precipitation, with parameter r_E controlling to what extent prior uncertainty scales with the absolute difference between the two evaporation datasets. If difference between the two datasets is small, e.g. during energy-limited conditions in winter, a minimum relative error of 10% is assumed by setting $s_{E,t} = 0.1 m_{E,t}$. As with precipitation, true evaporation E_t in month t is treated as a random draw from a truncated normal distribution. Truncation at zero constrains evaporation to be non-negative.

Since values of the error parameters are not known a priori, they are given vague prior distributions: quasi-uniform priors between 0 and 1 for w_E and r_E (specifically, flat logit-normal priors between 0 and 1 with location parameter $\mu = 0$ and scale parameter $\sigma = 1.4$), and a log-normal prior for f_E with mode at 1 (no bias) and a coefficient of variation CV of 50%.

3.3 River discharge error model

We assume the basin is gauged and a, possibly incomplete, record of measured monthly river discharge data Q_{obs} is available. A proportional error model is used to relate these data to underlying true discharge values Q:

$$m_{Q,t} = \mathcal{N}(Q_{obs,t}, v_{Q_{obs,t}}) \tag{10}$$
$$s_{Q,t} = a_Q Q_{obs,t} + b_Q \tag{11}$$
$$Q_t \sim \mathcal{N}(m_{Q,t}, s^2_{Q,t}) \tag{12}$$
$$Q_t \geq 0 \tag{13}$$

For months with observations, we set $v_{Q_{obs,t}} = 0$, so that the first equation becomes equivalent to $m_{Q,t} = Q_{obs,t}$, i.e. the mean of Q_t is equal to the (unbiased) observation for that month. For months with missing observations, $Q_{obs,t}$ and $v_{Q_{obs,t}}$ are set equal to the mean and variance of river discharge observed for that month across the entire observation record. This procedure works as long as only a few observations are missing. For the basins studied in this paper, Gorganrood basin has 1 month with missing data and Mond basin has 3 months with missing observations.

The magnitude of random observation errors is controlled by standard deviation $s_{Q,t}$, which is modeled as a linear function of the observed discharge for that month (or, the mean historical discharge for that month in case of a missing observation). This model assumes that observation errors increase linearly with discharge and includes two time-invariant parameters, a_Q and b_Q. Parameter a_Q is given a log-normal prior with mode at 0.1 (i.e. a relative error of 10%) and a small CV of 1%, while b_Q is given a log-normal prior with mode at 0.001 and also a CV of 1%. Sensitivity of the results to these assumed narrow priors will be evaluated in section 6.

As with precipitation and evaporation, monthly discharge Q_t is constrained to be non-negative.

3.4 Water storage error model

The JPL-mascon GRACE water storage data used here (see Table 2) consist of monthly total terrestrial water storage anomalies relative to the period 2004-2009 at a spatial resolution of 3°. The data come post-processed with the Coastline Resolution Improvement (CRI) filter of Wiese et al. (2016) to reduce leakage errors across land-ocean boundaries. Figure 3 shows measurement errors of the GRACE data across Iran.

Wiese et al. (2016) used simulations with the Community Land Model to downscale the coarse 3° storage data to a 0.5° global grid. Here, we use an alternative approach and instead downscale the data directly to the river basin of interest without using a hydrological model: first, the 3° data are weighted-area averaged over each river basin, and then an error model is specified to quantify systematic and random differences between the basin-averaged storage data and the true storage changes in the basin.

The monthly basin-scale data and true storages both typically have a seasonal cycle, but with possibly different amplitudes and phases, because the coarse-scale data are polluted by storage dynamics outside of the basin (”leakage”). This motivates the following noisy sine wave error model for quantifying differences between GRACE basin-scale water storages $S_{obs,t}$ and underlying true storages S_t:

$$m_{S,t} = S_t + A \sin\left(\omega(\frac{t}{12} - \delta)\right) \tag{14}$$
$$s_{S,t} = \sigma_S \tag{15}$$
$$S_{obs,t} \sim \mathcal{N}(m_{S,t}, s^2_{S,t}) \tag{16}$$

Here, A is amplitude (mm), ω is frequency (radians per year), and δ is phase (in years) of the errors. This model accounts for systematic differences in amplitude and phase between the observed and true values by means of time-invariant error parameters A and δ. Furthermore, time-invariant parameter σ_S quantifies magnitude of random errors in

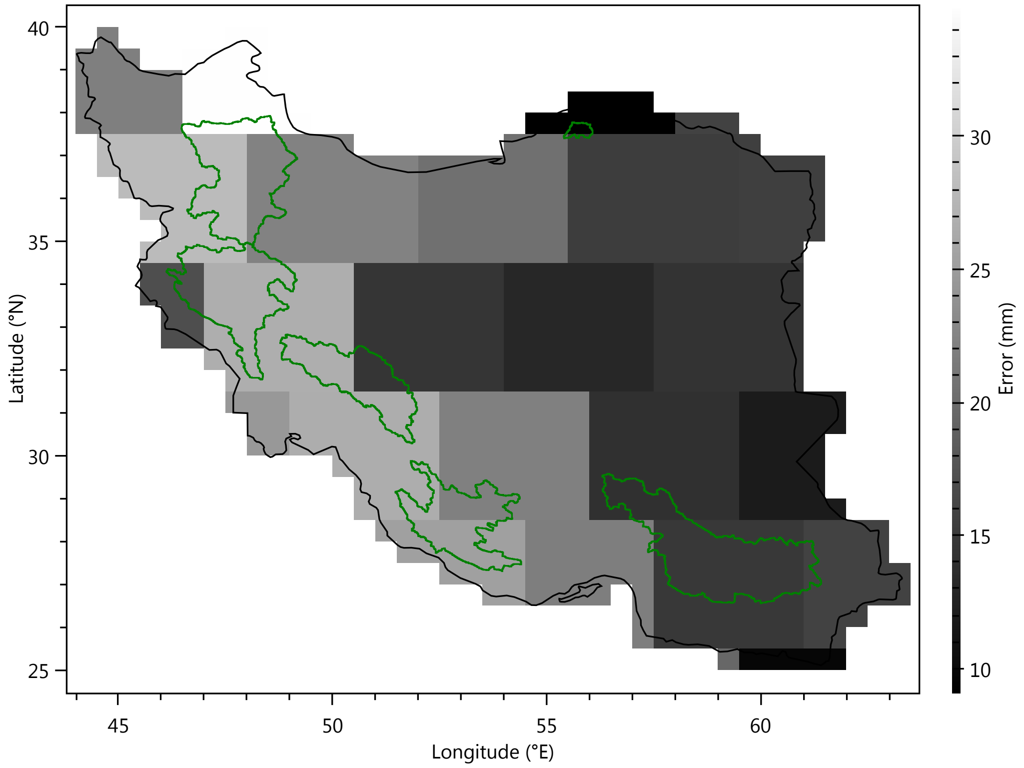

Figure 3. Time-averaged (2006-2015) measurement errors of the JPL GRACE data for each 3° mascon across Iran (based on the "uncertainty" variable in the JPL netcdf dataset). Errors tend to be smaller in arid parts of the country (east and central).

the basin-scale data, which may be caused by (i) inadequacies of the sine wave model and (ii) noise in the GRACE mascon inversion (Wiese et al., 2016), as shown by measurement errors in Fig. 3. We assume here σ_S is unknown and, in section 5, will compare its estimated value for each basin with the measurement errors in Fig. 3.

The value of ω is fixed at 2π radians per year, yielding a sine wave with a 12-month period, while A, σ_S, and δ are given vague priors to reflect prior uncertainty in the values of these parameters. Specifically, A is given a log-normal prior with mode at 30 mm and a CV of 200%, σ_S is given a log-normal prior with mode equal to 10 mm and a CV of 200%, and δ is given a flat logit-normal prior between 0 and 1 year with location parameter $\mu = 0$ and scale parameter $\sigma = 1.4$. Note that parameter δ represents phase of the errors; it should not be interpreted as phase difference between the observed and true signals. For example, if the observed and true signals are in phase, then δ will be equal to the shared phase of these signals, not equal to zero.

Note that the sine wave error model does not include a trend correction: it assumes that any long-term increasing or decreasing trend in the GRACE data is representative for water storage dynamics in the basin. If this assumption is invalid, then this may result in biased posterior estimates for precipitation and evaporation. However, this bias is likely to be relatively small, because water storage trends are sensitive to small changes in precipitation and evaporation. For example, a bias of 1 mm in monthly precipitation adds or removes 120 mm of water over a period of 10 years.

While the precipitation and evaporation error models rely on multiple datasets, the use of multiple GRACE solutions (e.g. the CSR mascon solution (Save, 2020) in addition to the JPL solution) is not expected to capture prior uncertainty caused by leakage or scaling errors, since the different solutions are generally limited by the same coarse spatial resolution of the GRACE observations. Therefore, the error model uses a single GRACE solution. Results in section 5 use the JPL data, while the effect of using the CSR data is evaluated in section 6.

4 Inference

The probabilistic water balance model described in the previous section defines a joint distribution over the data and all unknown variables, namely the 10 parameters (w_P, r_P, w_E, f_E, r_E, a_Q, b_Q, σ_S, A, δ) and the $4N$+1 monthly water balance variables (S_0, P_t, E_t, Q_t, S_t), where N is the number of months and S_0 is initial basin water storage at the start of the first month. This paper considers 10 years of data, so $N = 120$. Conceptually, we can write the joint distribution of the model as $p(\mathbf{x}, \boldsymbol{\theta}, \mathbf{S}_{obs})$, where $\mathbf{x}$ represents all $4N + 1$ water balance variables, $\boldsymbol{\theta}$ is the vector of 10 parameters, and $\mathbf{S}_{obs}$ represents the entire time-series of storage observations. Formally, this distribution depends on the input observations P_{obs}, E_{obs}, and Q_{obs}, but for notational simplicity this dependence is omitted here.

The goal is now to estimate posterior distributions for $\mathbf{x}$ and $\boldsymbol{\theta}$. The posteriors merge all available information and data, while accounting for all uncertainties in the model. We first describe the general form of the posteriors and then discuss the specific inference algorithm used.

4.1 Posterior distributions

The posterior for parameter vector $\boldsymbol{\theta}$ can be written as:

$$p(\boldsymbol{\theta}|\mathbf{S}_{obs}) \propto p(\boldsymbol{\theta})p(\mathbf{S}_{obs}|\boldsymbol{\theta}) \tag{17}$$

where $p(\boldsymbol{\theta})$ is the prior distribution for the parameters, and $p(\mathbf{S}_{obs}|\boldsymbol{\theta})$ is the likelihood. The prior is equal to the product of the individual parameter priors defined in the previous section. The likelihood on the other hand is obtained by computing the normalizing constant of the conditional water balance posterior $p(\mathbf{x}|\mathbf{S}_{obs}, \boldsymbol{\theta})$, as will be shown below.

The likelihood defines a scoring function for the parameters that quantifies how well storage predicted from the water balance matches the storage observations $\mathbf{S}_{obs}$. A good match can generally be achieved by picking bias parameters (f_E, w_P, etc) that move the storage predictions closer to the observations, and by making the noise parameters (r_E, σ_S, etc) as small as possible: this yields narrow predictive distributions centered on the observations, and thus large likelihood $p(\mathbf{S}_{obs}|\boldsymbol{\theta})$ for the parameters. However, since the error parameters are all time-invariant, such near-deterministic predictions generally cannot be achieved for all months simultaneously. Large likelihood is therefore achieved by setting the bias parameters to yield a good match on average across the entire time-series, and setting the noise parameters just large enough to "capture" all observations. Clearly, many error parameter combinations may yield large likelihood; this non-uniqueness is captured by characterizing the entire posterior distribution, rather than only determining the parameters with maximum likelihood or maximum posterior density. As described in the next section, the parameter posterior distribution is estimated using a Markov Chain Monte Carlo algorithm.

The joint posterior for all water balance variables $\mathbf{x}$ can be written as:

$$p(\mathbf{x}|\mathbf{S}_{obs}) = \int p(\mathbf{x}|\mathbf{S}_{obs}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{S}_{obs})d\boldsymbol{\theta} \tag{18}$$

where $p(\mathbf{x}|\mathbf{S}_{obs}, \boldsymbol{\theta}) = \frac{p(\mathbf{x}, \mathbf{S}_{obs}|\boldsymbol{\theta})}{p(\mathbf{S}_{obs}|\boldsymbol{\theta})}$ is the posterior distribution of $\mathbf{x}$, conditioned on specific values for the parameters. Note that the normalizing constant of this posterior is equal to the parameter likelihood function $p(\mathbf{S}_{obs}|\boldsymbol{\theta})$ in Eq. 17.

Instead of the joint posterior in Eq. 18, we are interested in marginal posterior distributions $p(x|\mathbf{S}_{obs})$ over individual water balance variables x, where x is a scalar variable equal to one of (S_0, P_t, E_t, Q_t, S_t). For example, if x corresponds to S_t, then we aim to compute the posterior distribution for S_t based on all observations before, on, and after time t. Such posterior distributions can be computed, as in Eq. 18, by averaging conditional posterior distributions $p(x|\mathbf{S}_{obs}, \boldsymbol{\theta})$ over the parameter posterior distribution $p(\boldsymbol{\theta}|\mathbf{S}_{obs})$. An efficient way of computing all conditional posteriors $p(x|\mathbf{S}_{obs}, \boldsymbol{\theta})$ is to use a smoothing algorithm, such as a Kalman smoother, as discussed next. Incidentally, a smoothing algorithm also computes normalizing constant $p(\mathbf{S}_{obs}|\boldsymbol{\theta})$ of $p(\mathbf{x}|\mathbf{S}_{obs}, \boldsymbol{\theta})$, which is used to compute the likelihood in Eq. 17, without explicitly constructing the $(4N+1)$-dimensional joint water balance posterior.

4.2 Algorithm

Following the discussion in the previous section, posterior distributions are computed using a double-loop algorithm that combines Markov Chain Monte Carlo (MCMC) sampling for the parameter posteriors with Expectation Propagation (EP) (Minka, 2001), an iterative smoothing algorithm, for the water balance posteriors. Essentially, the MCMC algorithm forms an outer loop that iteratively proposes and accepts/rejects new parameter values, while the EP algorithm forms an inner loop that iteratively computes (i) the (unnormalized) posterior density, Eq. 17, of parameter values proposed by the MCMC algorithm, and (ii) conditional water balance posteriors $p(x|\mathbf{S}_{obs}, \boldsymbol{\theta})$ for specific parameter vectors sampled by the MCMC algorithm.

For linear-Gaussian models, the EP algorithm is equivalent to a Kalman smoother for S_t, and computes exact Gaussian water balance posteriors via a single forward-backward pass through the time series, with the backward pass also updating the P_t, E_t and Q_t posteriors (see Appendix B). The forward-backward pass ensures that water balance posteriors are estimated using data from the entire time-series. Given values for the error parameters, the probabilistic water balance model in this paper consists of a linear transition model at each time step (i.e. water balance equation, Eq. 1) with Gaussian storage observations. However, as discussed in the previous section, the model also uses physical non-negativity constraints for each P_t, E_t, and Q_t. These constraints render the input distributions and water balance posteriors non-Gaussian. The EP algorithm used here approximates the exact non-Gaussian water balance posteriors with Gaussian distributions that have the same moments (mean and variance) as the exact posteriors. This strategy is called moment-matching. Since moment-matching is applied to the posterior, not the prior, approximations made in one month affect approximations in other months and the algorithm is iterative: instead of a single forward-backward pass, multiple forward-backward passes are used, where each pass further refines the approximations until convergence, i.e. until there is no more change in the approximate posteriors.

We implement the probabilistic water balance model in C# using the open-source probabilistic programming library Infer.NET (Minka et al., 2018). The resulting model code (see Fig. A1) uses the Infer.NET modeling API to implement the model equations listed in the previous section. This code is then automatically translated by the Infer.NET compiler into code for running inference, i.e. for computing the water balance posteriors with EP.

The MCMC algorithm used in this paper is a single-chain version of the differential-evolution MCMC algorithm of ter Braak and Vrugt (2008). The algorithm iteratively proposes new parameter vectors and evaluates their posterior density, Eq. 17, by calling the EP inference code. The latter computation is done in Infer.NET by placing the entire model inside a stochastic if-block and using EP to compute the posterior odds of being inside vs outside the block, i.e. of the model being "true".

Finally, since the EP algorithm only computes conditional water balance posteriors (conditioned on specific parameter values), a post-processing step is used that averages computed water balance posteriors over the MCMC sampled parameter sets, as in Eq. 18. That way, the final water balance posteriors account for posterior uncertainty in the data error parameters. For example, if $p(x|\mathbf{S}_{obs}, \boldsymbol{\theta})$ represents the (Gaussian) posterior for variable x (e.g. E_t), conditioned on data $\mathbf{S}_{obs}$ and on parameter vector $\boldsymbol{\theta}$, then the final marginal posterior $p(x|\mathbf{S}_{obs})$ is computed from n posterior parameter samples $\boldsymbol{\theta}_i$ as:

$$p(x|\mathbf{S}_{obs}) = \int p(x|\mathbf{S}_{obs}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{S}_{obs}) d\boldsymbol{\theta} \approx \frac{1}{n} \sum_{i=1}^{n} p(x|\mathbf{S}_{obs}, \boldsymbol{\theta}_i) \tag{19}$$

As such, each marginal water balance posterior is strictly speaking a (Gaussian) mixture distribution, although empirically it turns out to be well approximated by a single Gaussian distribution using moment matching. While this last approximation is not strictly necessary, it avoids storing the entire Monte Carlo mixture (for each water balance variable and each month).

5 Results

First, detailed results are presented for one of the basins (Mond), followed by a summary of results for all basins. Detailed results for all basins are available in the Supporting Information.

5.1 Mond basin

Mond basin is one of the drier basins in this study (Table 1). Water balance posteriors for Mond basin are shown in Fig. 4, and error parameter posteriors are shown in Fig. 5. In Fig. 4, inferred precipitation tends to more closely follow the CHIRPS data than the IMERG data, especially during the wet winter months, with IMERG apparently overestimating precipitation. This is reflected in the inferred value for parameter w_P (last row in Fig. 4), which is shifted towards 1, indicating greater weight on CHIRPS than on IMERG for this basin. The wide posterior for noise scaling parameter r_P indicates that this parameter does not play an important role here, and the posterior uncertainty in precipitation is not markedly different from the prior uncertainty shown in Fig. 2.

In contrast, posterior uncertainty in evaporation is significantly smaller than its prior uncertainty, as shown by the posterior uncertainty bands in Fig. 4 (second row) and posterior values of $r_E < 0.5$, indicating that random errors in evaporation are smaller than the absolute difference between the SSEBop and GLEAM data. Estimated evaporation lies more or less right between the two datasets, with an estimated w_E value around 0.5 (equal weights) and no additional bias (f_E around 1). Posterior uncertainty increases during dry summers when differences between the two datasets are largest.

River discharge in this basin is an order of magnitude smaller than the other water balance variables. With the assumed 10% relative error, this results in small posterior uncertainty that closely follows prior uncertainty (third row in Fig. 4). Note however the significant increase in discharge uncertainty at the end of the time series: no river discharge observations are available in the basin for the last three months of 2015, and historical discharge variability is instead used as prior for these months, as discussed in section 4. The larger posterior uncertainty in discharge for these months does not appear to affect uncertainty in the other water balance components. This will be further explored in section 6.

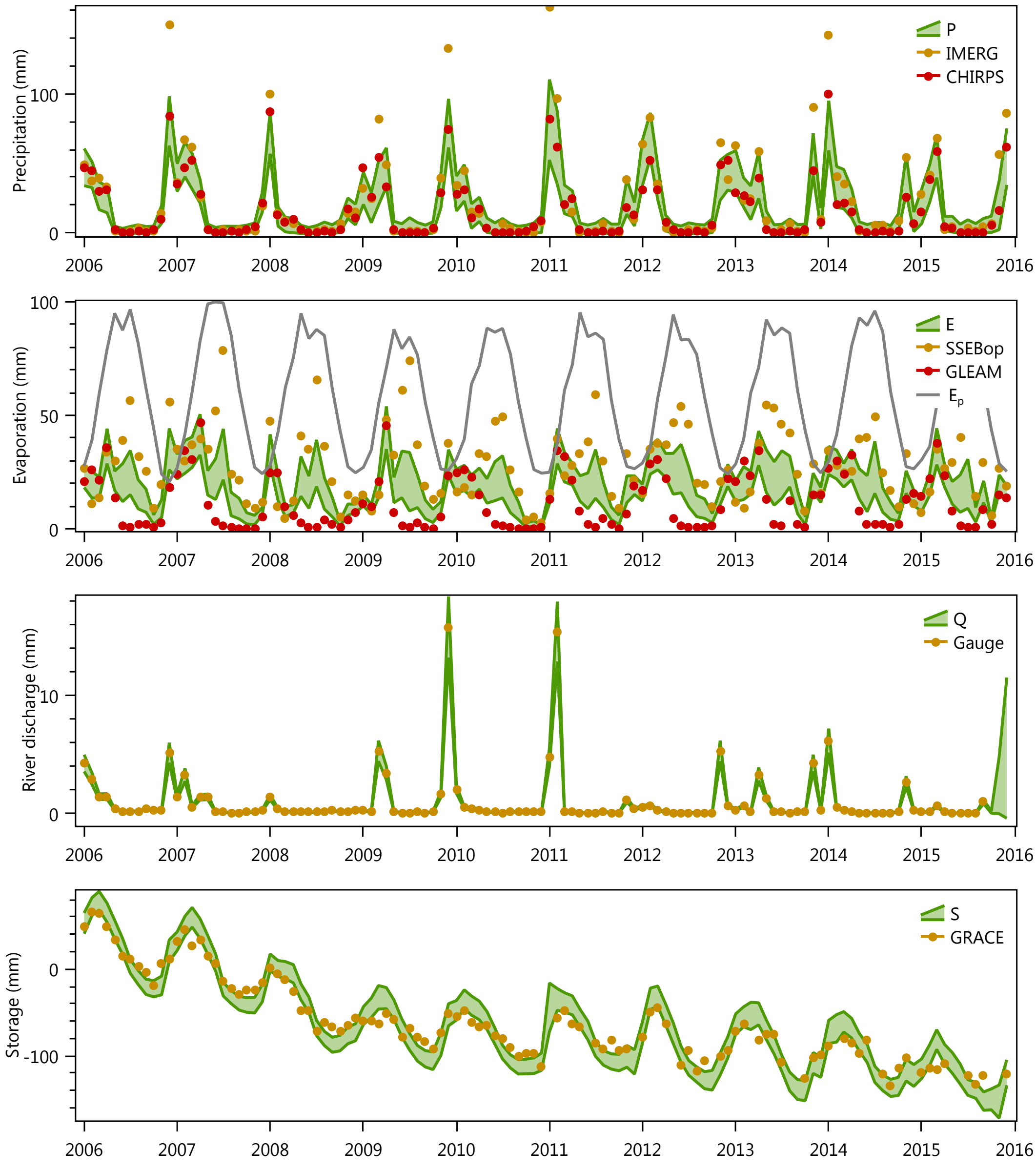

Figure 4. Monthly water balance estimates for Mond basin, shown as 90% posterior uncertainty bands. Each year label indicates start of the year (January). All values are in mm/month.

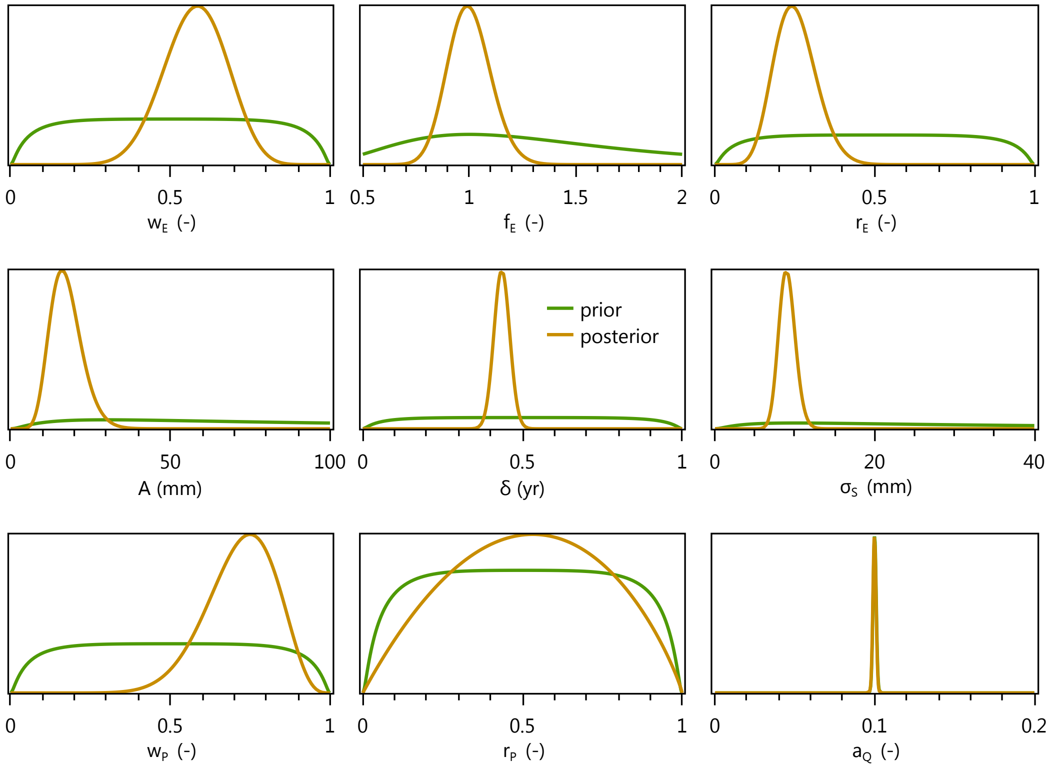

Figure 5. Normalized prior and posterior densities of error parameters for Mond basin.

The last row of Fig. 4 shows that the inferred water storage dynamics largely follow the GRACE observations, with a small increase in seasonal amplitude in the posteriors compared to the data. The corresponding inferred storage error parameters are shown in the second row of Fig. 5. All three parameters (A, δ, σ_S) have well defined posterior distributions compared to their vague priors. Residual noise in the data, after making amplitude (A) and phase adjustments (δ), is relatively small as indicated by an inferred value for σ_S of around 10 mm. Note that inferred posteriors for months with missing GRACE observations (e.g. May-June 2015, October-November 2015) do not markedly differ from months with observations. This is because error parameter values learned from months with data are shared across all months, and because smoothing infers posteriors using data from all months. A more dramatic example of this effect will be seen in section 6.

5.2 Other basins

The Supporting Information contains posterior plots for all other basins, similar to the ones for Mond basin shown above. Here, we highlight the main findings from these results. In terms of water storage posteriors, the basins can roughly be divided into basins without a significant change in amplitude or phase between the estimated posteriors and the GRACE data (Mond, Karoon, Karkheh), basins with only a change in phase (Sepidrood), basins with only a change in amplitude (Jazmoorian), and basins with both a change in amplitude and phase (Gorganrood).

Figure 6 illustrates this for the Sepidrood and Gorganrood basins. In both basins the inferred storage dynamics (posteriors shown in green) are shifted earlier in time than the corresponding GRACE observations. Apparently, the observed GRACE dynamics do not fit with the other water balance observations in terms of water balance closure. Interestingly, both basins are in the north of the country where the large footprint of the GRACE observations (Fig. 3) is possibly affected by the Caspian Sea to the north, which is not included in the Coastline Resolution Improvement (CRI) filter of the JPL GRACE dataset. The sine wave error model appears to restore the underlying water storage dynamics, including an increase in amplitude for the relatively small Gorganrood basin. The increase in amplitude can be explained by the strong spatial smoothing inherent in the coarse-scale GRACE data, which tends to be more severe in smaller basins.

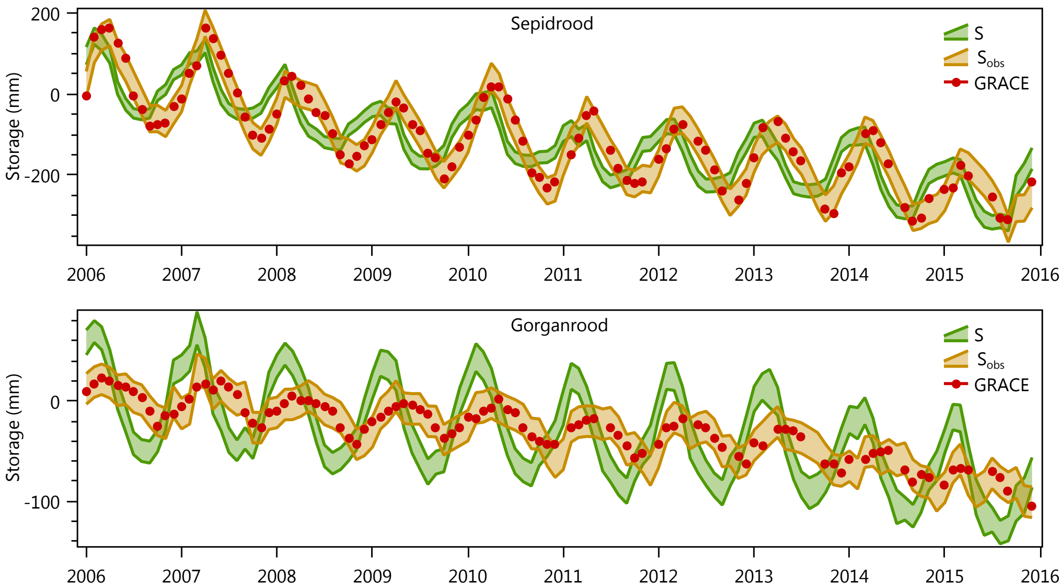

Figure 6. 90% uncertainty bands of storage posteriors (S) and GRACE posterior predictive distributions (S_{obs}), along with GRACE data, for Sepidrood and Gorganrood basins.

Fig. 6 also shows posterior predictive distributions for the GRACE observations (S_{obs}), conditioned on the posterior mean of the true water storage (S). These plots illustrate validity of the proposed sine wave model, since the original GRACE observations fall within the posterior predictive distributions obtained by taking the inferred posterior mean of S_t in each month and applying the noisy sine wave model to generate a predictive distribution for the corresponding observation $S_{obs,t}$. This however does not mean that the probabilistic water balance model is generally suitable for making water balance predictions, as will be illustrated in section 6.

Error parameter posterior distributions for all basins are shown in Fig. 7. The third row in this figure shows that for most basins IMERG fits better with the other water balance data than does CHIRPS, since inferred values for w_P are mostly less than 0.5 (more weight on IMERG). Mond basin is the exception, with $w_P > 0.5$, as discussed above. The insensitivity of parameter r_P that was already observed in Mond basin, also occurs in two other basins (Sepidrood and Karkheh), while in the three other basins r_P does matter and tends toward a value of 1.

The three evaporation error parameters are mostly well identified (first row in Fig. 7). In most basins, more weight is given to the GLEAM dataset ($w_E > 0.5$), with the exception of the wettest basin (Karoon), where SSEBop provides a better fit. However, in all basins a weighted average of the two datasets is preferred to using either dataset alone. Inferred values for bias parameter f_E range between 0.5 and 1.5, with the largest values for Karkheh and Sepidrood basins. While a multiplicative bias of 1.5 may seems excessive, the inferred evaporation posteriors remain at or below potential evaporation (see Supporting Information), even though potential evaporation was not used in the model. Finally, the reduction in prior evaporation uncertainty found in Mond basin also occurs in other basins, as evidenced by inferred values for r_E below 0.5, with the exception of Karkheh and Sepidrood basins, where prior evaporation uncertainty is less pronounced than in the other basins.

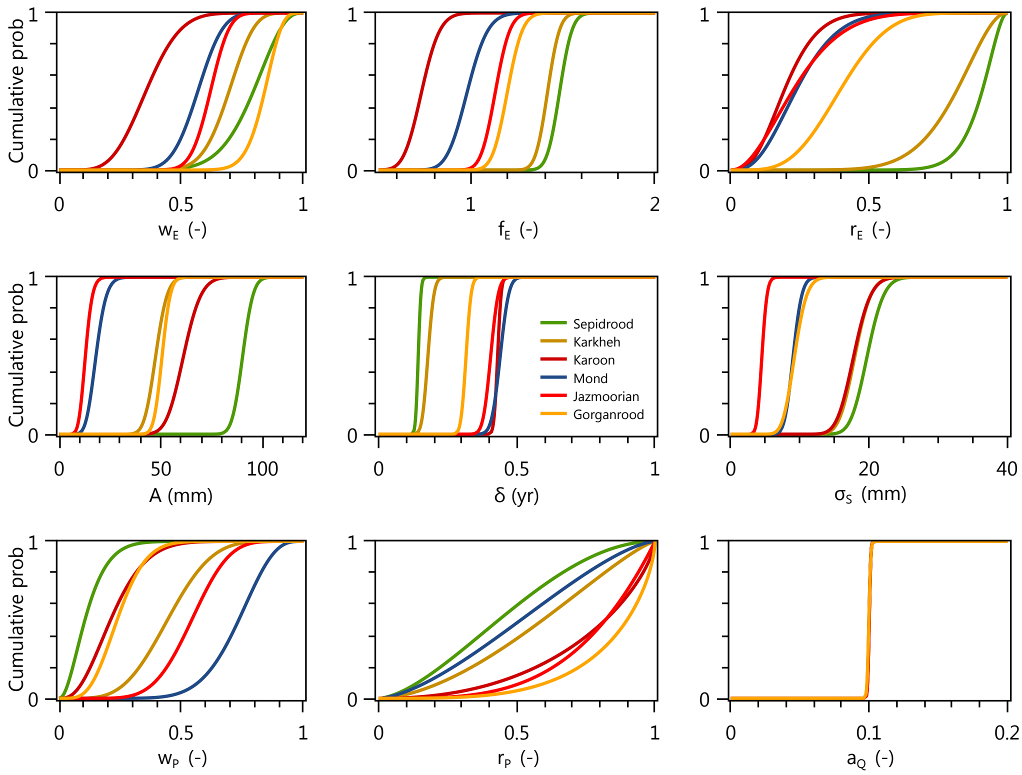

Figure 7. Posterior error parameter distributions for all basins.

The storage error parameters (second row in Fig. 7) are also well identified in all basins. Standard deviation σ_S of random errors in the GRACE observations, after amplitude and phase corrections, is 10 mm or less for the drier basins in the east (Mond, Jazmoorian, Gorganrood) and 15-20 mm for the wetter basins in the west (Sepidrood, Karkheh, Karoon). As shown in Fig. 8, the inferred posterior mean values for σ_S closely follow a similar west-to-east decreasing trend as the JPL-mascon GRACE measurement errors, with an increase in inferred noise for the smaller Gorganrood basin. These results suggest that the sine wave model adequately captured and corrected systematic errors in the GRACE data due to a mismatch in scale, yielding random errors similar to and even smaller than the reported GRACE measurement errors.

Finally, Table 3 summarizes and compares posterior standard deviations for the different water balance variables. The table includes results for a second scenario with vague prior on a_Q, which is further discussed in section 6. Results in this table show that posterior uncertainty, in terms of posterior standard deviation, decreases from water storage (4-12 mm/month), to precipitation (3.5-7 mm/month), to evaporation (2-6 mm/month), and to discharge (0-2 mm/month). The small posterior uncertainty in river discharge is a direct consequence of the assumed 10% error and the generally small discharge val-

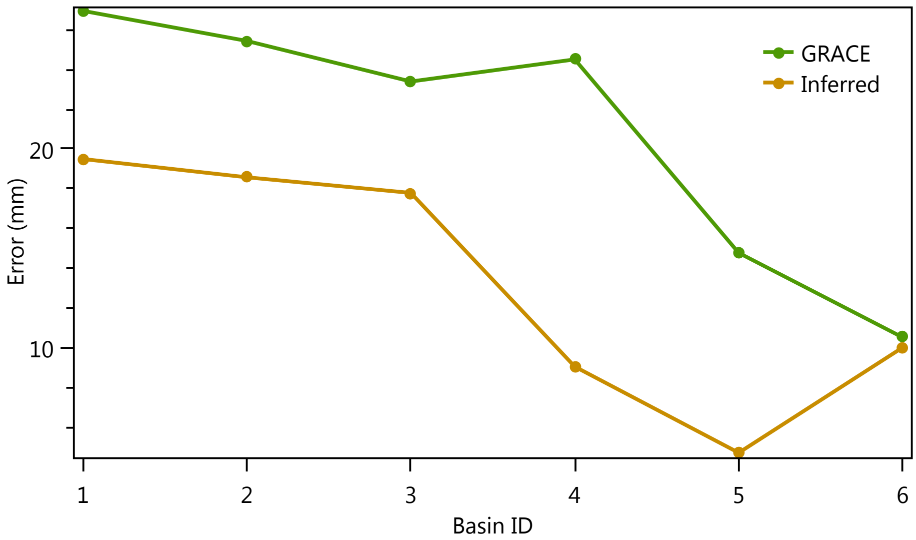

Figure 8. Posterior mean of σ_S compared to mascon-scale standard error of the JPL GRACE observations.

ues in the semi-arid basins studied here. At the extreme end, the endorheic Jazmoorian basin has no outflow, and thus zero discharge and error.

The reported posterior standard deviations result from the fusion of all water balance data. For example, the posterior of S_t in a particular month t results from the fusion of three noisy information streams: the GRACE observation for that month (if not missing), the water balance constraint for month t, and the water balance constraint for month $t+1$, for which S_t provides the initial storage. Combination of these three information streams results in a posterior that is narrower than any of the individual streams, with each stream or distribution more or less constraining the final posterior estimate of S_t. A similar process happens when inferring the other water balance variables (P_t, E_t, Q_t), although for those variables only two information streams are involved (one from the prior, and the other from the water balance of month t).

Table 3. Average posterior standard deviation (mm/month) of each water balance variable for two cases: (i) relative river discharge error a_Q fixed at 0.1 (10%) and (ii) a vague lognormal prior for a_Q with mode at 0.1 and CV equal to 0.9.

	$a_Q = 0.1$				Vague prior on a_Q			
Basin	P	E	Q	S	P	E	Q	S
Sepidrood	6.0	5.1	0.2	10.1	6.0	5.1	0.4	10.1
Karkheh	6.1	6.1	0.4	11.2	6.2	6.1	1.0	11.1
Karoon	6.9	4.8	1.7	11.3	6.8	4.6	5.6	11.7
Mond	4.7	3.5	0.1	6.7	4.7	3.6	0.3	6.8
Jazmoorian	3.5	1.9	0.0	4.1	3.5	1.9	0.0	4.0
Gorganrood	6.7	4.9	0.2	8.3	6.8	5.0	0.4	8.3

6 Discussion

This section evaluates how results are affected when changing some of the data and assumptions of the probabilistic water balance model.

6.1 Sensitivity to assumed river discharge errors

Results in the previous section were based on a narrow prior for the relative error a_Q of monthly river discharge data centered on 0.1 (10%). To test sensitivity of the results to this choice, an alternative vague lognormal prior for a_Q was used, i.e. one with mode at 0.1 and with a coefficient of variation of 0.9. Table 3 shows that this change increases the posterior standard deviation of monthly river discharge, but has otherwise little effect on posterior uncertainty of the other water balance variables. The largest absolute increase in posterior standard deviation of Q is observed for Karoon basin, which is the wettest basin included in the analysis. In fact, for Karoon basin, the posterior standard deviation of river discharge becomes larger than that of evaporation (Table 3).

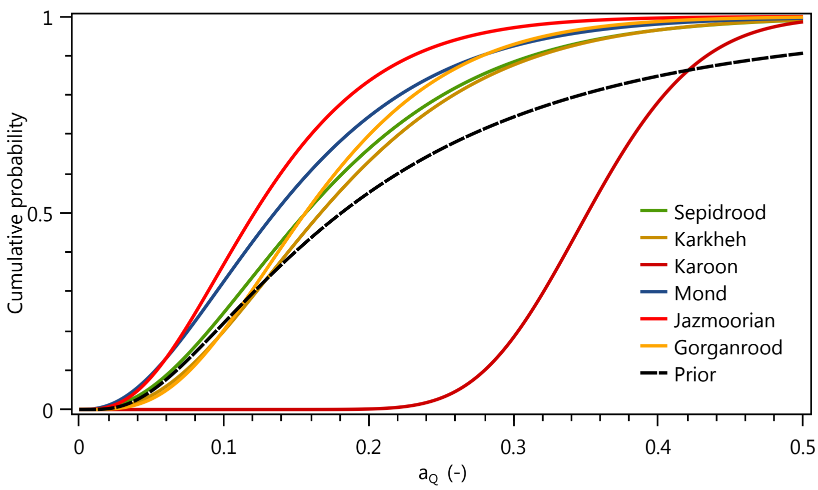

Figure 9. Posterior distributions (cdf) for a_Q when using a vague prior (dashed) for a_Q.

When using a vague prior, posterior distributions for relative error a_Q in Fig. 9 show that the posteriors are generally close to the prior. Most basins show a slight contraction of the posterior relative to the prior toward smaller relative errors, with the exception of Karoon basin, where the posterior moves to larger, likely unrealistic, values for a_Q around 0.3-0.4. These large values suggest that uncertainty in river discharge increases to compensate for errors somewhere else in the water balance. Due to the small magnitude of river discharge relative to the other water balance terms, a large relative error is needed to get a sizeable effect.

These results indicate that, for the semi-arid basins studied here, the value of a_Q cannot reliably be estimated from water balance data, and instead river discharge errors should be estimated independently, e.g. using a formal rating curve error analysis (Horner et al., 2018; Kiang et al., 2018). The value of a_Q can then be fixed a priori, or given a narrow prior, based on the independent estimate. On the other hand, accurate estimates of a_Q are only relevant for estimating uncertainty of the river discharge data. For the goal of estimating the other water balance variables, approximate estimates of a_Q suffice, at least when river discharge is the smallest term in the water balance.

6.2 Effect of missing GRACE observations

Results in section 5 already showed that missing GRACE observations do not significantly affect the inferred posteriors. Sharing of error parameters across the entire time-series, combined with fusion of all data via smoothing, allows the model to fill in occasional gaps in the data record. It is however instructive to evaluate a few more drastic scenarios of missing GRACE observations to gain additional insight into the predictive capabilities and limitations of the probabilistic water balance model.

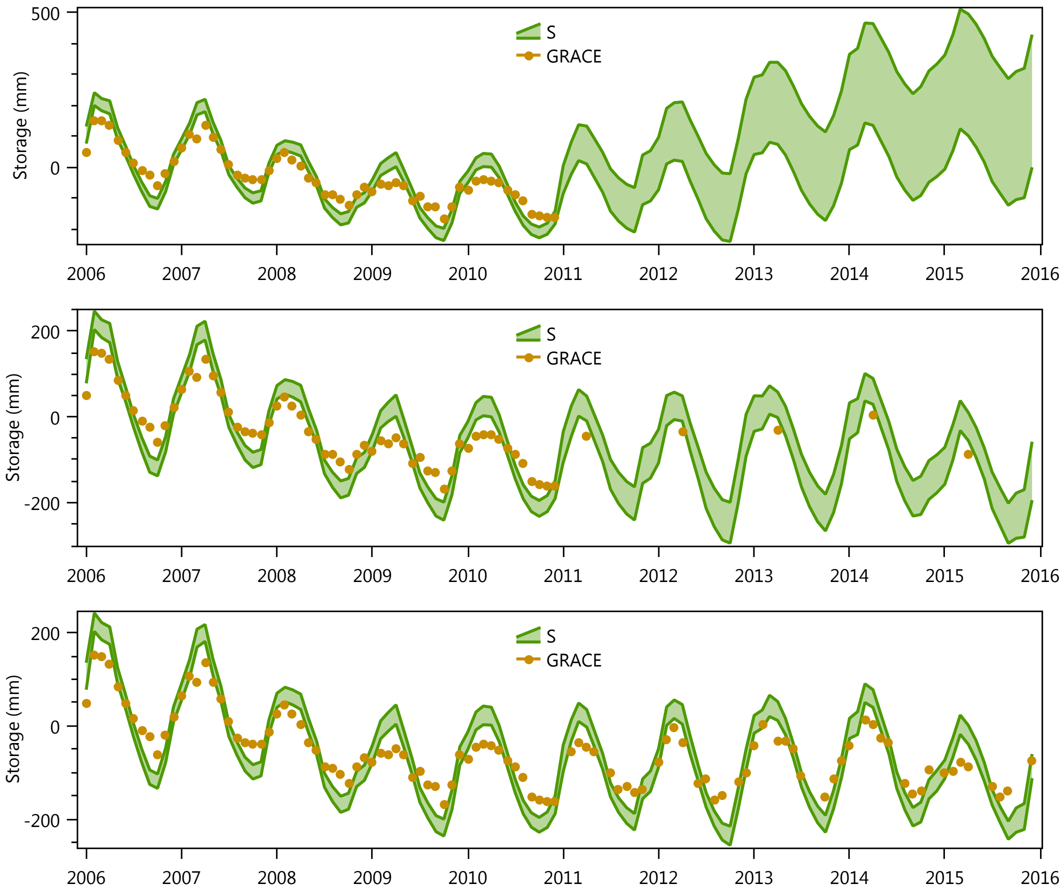

Figure 10. Storage posteriors for Karoon basin for three scenarios of missing GRACE observations: (i) no GRACE observations in the last 5 years, (ii) one GRACE observation per year in the last 5 years, (iii) using all available observations.

Two fictitious scenarios are evaluated. The first scenario assumes that all GRACE observations after 2010 are missing; the first five years provide a complete data record to learn the model error parameters, which are then applied to infer and predict storage posteriors in the next five years. Fig. 10 shows that in the absence of constraining GRACE observations in the second part of the period, posterior uncertainty grows over time, and an increasing trend in storage is (wrongly) predicted. In the second scenario, which assumes a single annual observation is available after 2010, this trend is removed and posterior uncertainty is smaller, although it remains larger than when the full GRACE observation record is used.

These results illustrate that the model is less suitable for long-range predictions without storage observations: uncertainties quickly accumulate, and small imbalances between precipitation and evaporation easily lead to erroneous trend predictions. On the other hand, the model works well for interpolating and filling in gaps when observations are occasionally missing.

6.3 Using a different GRACE solution

The results in this paper are based on the JPL-mascon GRACE data. The model can also use other GRACE solutions by simply replacing S_{obs} in the model by the relevant dataset. Fig. 11 compares inferred posterior distributions for σ_S when using the CSR mascon solution instead of the JPL mascon solution. For the basins studied in this paper, the JPL data consistently yield smaller noise, i.e. smaller posterior values for σ_S. This indicates that the JPL data provide a better fit with the other monthly water balance data used in this study.

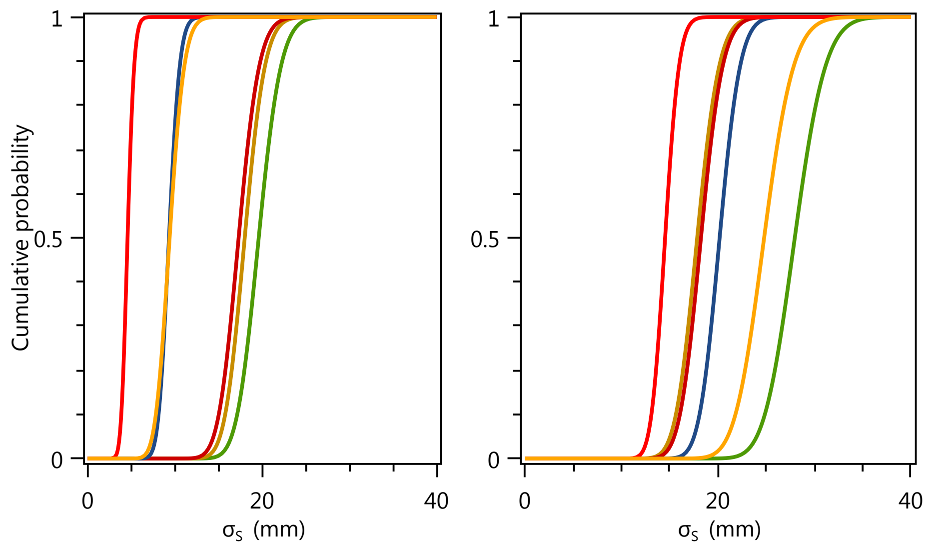

Figure 11. Posterior distributions (cdf) of σ_S for two different GRACE mascon solutions: JPL (left) and CSR (right).

6.4 Effect of positivity constraints

As described in section 4, the model includes positivity constraints on water balance variables P, E, and Q, since these variables cannot physically be negative. To what extent do these constraints affect the inferred posteriors? This can be assessed by removing the positivity constraints from the model, which is achieved by commenting out the three `Variable.ConstrainPositive` statements in Fig. A1) and recomputing the posteriors. Conditional on the model parameters, the model now only contains Gaussian and linear relations. As such, inference does not require any iteration and a single forward-backward pass over the monthly time-series is sufficient to compute all water balance posteriors. The Infer.NET compiler in fact automatically detects this and, in the absence of positivity constraints, generates inference code that is equivalent to a Kalman smoother.

Fig. 12 shows that constraining the water balance variables to be positive results in smaller posterior uncertainty when the unconstrained posterior extends into the neg-

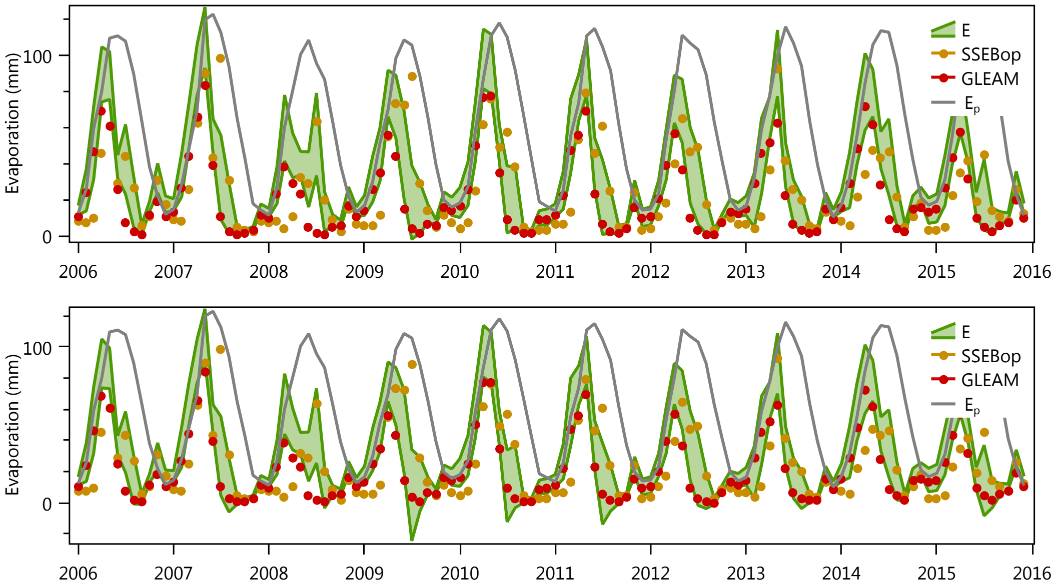

Figure 12. Posterior 90% uncertainty bands for monthly evaporation in Karkheh basin with (top) and without (bottom) positivity constraints in the model.

ative domain. In this case (Karkheh basin), the unconstrained evaporation posterior has a negative tail whenever there is a large difference between the two evaporation datasets (e.g. summer 2009), because then the (prior) uncertainty is large. However, overall for the basins analyzed here, the effect of the positivity constraints is fairly limited and does not significantly change the results. This is also why the number of EP iterations to achieve convergence is small (we used 3 iterations); the studied problems are only mildly non-Gaussian. However, the positivity constraints do maintain physically realistic posteriors and thus are useful for general applicability of the model.

7 Conclusions

The paper presents a probabilistic model to estimate monthly basin-scale precipitation, evaporation, terrestrial water storage and river discharge based on independent observations of each water balance term and monthly water balance constraints. The main contribution compared to previous water balance fusion studies is that data errors are not fixed a priori but are treated as unknown random variables that are estimated from the data. This results in a data fusion approach that combines data error and water balance estimation into a single coherent methodology.

The approach is based on formulating a Bayesian hierarchical model that ties together all data, water balance variables and data error parameters, followed by computing posteriors of all unknown parameters and water balance variables in the model. The model combines monthly basin-scale water balance constraints with data error models for each water balance variable (precipitation, evaporation, river discharge, water storage) that account for random and systematic data errors.

Specifically, bias in precipitation and evaporation data is modeled as a weighted average of two different datasets (IMERG and CHIRPS for precipitation, and SSEBop and GLEAM for evaporation), where the weight is treated as an unknown parameter. For evaporation, a second unknown bias parameter is included for additional flexibility in modeling bias. Random errors in precipitation and evaporation are modeled as a function of differences between the two respective datasets, with unknown parameters controlling magnitude of the random errors. The JPL-mascon GRACE data are used as basin-scale water storage observations. Measurement and scaling errors in the GRACE data are described by a noisy sine-wave error model, with amplitude, phase and noise of the sine wave controlled by unknown parameters. Finally, monthly river discharge data are taken from river gauging stations, with random errors described by a relative error parameter.

The resulting probabilistic model is solved for the unknown water balance variables and data error parameters using Markov Chain Monte Carlo sampling (for the parameters) in combination with an iterative smoothing algorithm (for the water balance variables) that maintains non-negativity of the water balance variables. Computed posteriors provide (i) hydrologically consistent, error-filtered and bias-corrected water balance estimates, and (ii) statistically consistent, basin-specific error estimates of the water balance data.

Application to semi-arid river basins in Iran illustrates usefulness of the approach. First, computed evaporation posteriors achieve significant reductions in prior evaporation uncertainty during water-stressed summers. Other studies have also reported reductions in errors by combining multiple evaporation products (Mueller et al., 2011; Hobeichi et al., 2018). Second, the approach leads to basin-specific phase and amplitude corrections of the GRACE data, and is able to extract the underlying water storage dynamics. Third, by fusing all water balance data, posterior water balance estimates are obtained with time-averaged standard errors of 4-12 mm/month for water storage, 3.5-7 mm/month for precipitation, 2-6 mm/month for evaporation, and 0-2 mm/month for river discharge. Data error parameters are generally well identified, with the exception of relative error of the river discharge data, which is best estimated using an independent rating curve analysis. This lack of sensitivity however also means that the other water balance estimates are not strongly affected by the assumed discharge errors, and an approximate estimate suffices as long as river discharge is the smallest term in the water balance, as is the case for the semi-arid basins studied here.

The proposed methodology is data-driven in that no hydrological process assumptions are made beyond the monthly water balance constraints. As such, the water balance posteriors can be used for independent evaluation and calibration of monthly water balance models. Nevertheless, an interesting extension could be to embed the data errors models used here into a monthly water balance model, and perform joint estimation of all error and hydrological parameters. Another modification would be to consider spatially distributed error models, e.g. using land cover specific error models for evaporation and elevation or temperature specific error models for precipitation, and sharing these parameters across multiple basins to ensure identifiability.

The approach can also be extended to other datasets and other (gauged) basins around the world, possibly using tailor-made data error models. Modifications may be warranted to describe data errors in different climates and landscapes, e.g. in snow-dominated basins, where satellite data may underestimate snow accumulation. A benefit in this respect is that the model is implemented in a general-purpose and extensible probabilistic programming tool (Infer.NET) that separates model assumptions from inference (model solving): when the individual data error models are modified, inference code is automatically generated to compute posteriors for the new model.

Appendix A Implementation of the probabilistic water balance model in Infer.NET

Figure A1 shows how the probabilistic water balance model in section 3 translates directly into a probabilistic program implemented with the Infer.NET modeling API. The Infer.NET compiler automatically translates the model code into an iterative smoothing algorithm for computing water balance posteriors using Expectation Propagation (EP). The complete code is at `http://doi.org/10.5281/zenodo.4116451`.

```
// Time loop
using (var time = Variable.ForEach(timeInterval))
{
    var t = time.Index;

    // P
    var mP = (1 - wP) * PObs1[t] + wP * PObs2[t];
    var sP = Variable.Max(PStd[t], rP * 0.5 * Abs(PObs1[t] - PObs2[t]));
    P[t] = Variable.GaussianFromMeanAndVariance(mP, sP * sP);
    Variable.ConstrainPositive(P[t]);

    // E
    var mE = fE * ((1 - wE) * EObs1[t] + wE * EObs2[t]);
    var sE = Variable.Max(0.1 * mE, rE * 0.5 * Abs(EObs1[t] - EObs2[t]));
    E[t] = Variable.GaussianFromMeanAndVariance(mE, sE * sE);
    Variable.ConstrainPositive(E[t]);

    // Q
    var mQ = Variable.GaussianFromMeanAndVariance(QObs[t], QObsVar[t]);
    var sQ = aQ * QObs[t] + bQ;
    Q[t] = Variable.GaussianFromMeanAndVariance(mQ, sQ * sQ);
    Variable.ConstrainPositive(Q[t]);

    // S: water balance
    using (Variable.If(t == 0))
    {
        S[t] = S0 + P[t] - E[t] - Q[t];
    }
    using (Variable.If(t > 0))
    {
        S[t] = S[t - 1] + P[t] - E[t] - Q[t];
    }

    // SObs
    var missingSObs = IsNaN(SObs[t]);
    using (Variable.IfNot(missingSObs))
    {
        const double omega = 2 * Math.PI;
        var mS = S[t] + A * Sin(omega * (Variable.Double(t) / 12 - Delta));
        var sS = SStd;
        SObs[t] = Variable.GaussianFromMeanAndVariance(mS, sS * sS);
    }
}
```

Figure A1. Implementation of the probabilistic water balance model using the Infer.NET probabilistic programming API in C#.

Appendix B Details of EP

Here, we give details of how Expectation Propagation (EP) computes conditional water balance posteriors. EP uses "messages", i.e. Gaussian distributions in this case, to propagate uncertainty through the model. If we write the water balance at each time as $S = S_0 + P - E - Q$ (omitting time index for simplicity), then the forward message (Gaussian distribution) to S is computed by propagating Gaussian distributions for the inputs (S_0, P, E, Q) through the water balance:

$$\text{forward message to } S = \mathcal{N}(S|m_{S_0} + m_P - m_E - m_Q, v_{S_0} + v_P + v_E + v_Q) \tag{B1}$$

where m_x and v_x represent mean and variance of input x. Mean and variance of P, E, and Q are given by the model priors described in section 3, modified for truncation at zero, see below. Mean and variance of previous storage S_0 is given by multiplying two Gaussian distributions: the forward message that was sent to S_0 in the previous time step and the Gaussian likelihood of a GRACE observation, if any. Mean and variance of the resulting Gaussian message (distribution) is given by the general Gaussian multiplication formula:

$$\begin{aligned} \mathcal{N}(x|m_1, v_1)\mathcal{N}(x|m_2, v_2) &\propto \mathcal{N}(x|m, v) && \text{(B2)} \\ m &= w_2 m_1 + w_1 m_2 && \text{(B3)} \\ v &= w_2 v_1 = w_1 v_2 && \text{(B4)} \end{aligned}$$

where $w_1 = \frac{v_1}{v_1+v_2}$, $w_2 = \frac{v_2}{v_1+v_2}$, and x in this case would be S_0. This formula is the scalar version of the Kalman filter update equation. Forward messages are computed by a forward pass through the entire time series.

Likewise, backward messages represent (Gaussian) distributions that propagate uncertainty through the model in backward direction. They are computed by a backward pass through the entire time series, analogous to a Kalman smoother. The backward message (Gaussian distribution) to S_0 is computed by propagating Gaussian distributions for the inputs (P, E, Q) and for S through the water balance back to S_0:

$$\text{backward message to } S_0 = \mathcal{N}(S_0|m_S - m_P + m_E + m_Q, v_S + v_P + v_E + v_Q) \tag{B5}$$

where mean m_S and variance v_S of the backward message from S are obtained by multiplying the backward message to S (computed in previous step of backward pass) with the Gaussian likelihood of a GRACE observation, if any, using the same Gaussian multiplication formula given above. The posterior for each S (or S_0) is obtained by multiplying the forward and backward message it receives, as well as a GRACE likelihood message, if any.

Backward messages to the inputs are computed in a similar way:

$$\begin{aligned} &\text{backward message to } P = \mathcal{N}(P|m_S - m_{S_0} + m_E + m_Q, v_{S_0} + v_S + v_E + v_Q) && \text{(B6)} \\ &\text{backward message to } E = \mathcal{N}(E|m_{S_0} - m_S + m_P - m_Q, v_{S_0} + v_S + v_P + v_Q) && \text{(B7)} \\ &\text{backward message to } Q = \mathcal{N}(Q|m_{S_0} - m_S + m_P - m_E, v_{S_0} + v_S + v_P + v_E) && \text{(B8)} \end{aligned}$$

These backward messages correspond to what Pan and Wood (2006) call a "constrained Kalman filter". The product of these backward messages and the corresponding priors gives the posterior for each input. However, since P, E, and Q are constrained to be positive, the actual posteriors are truncated Gaussians. Moments of each truncated posterior are given by:

$$\mathbb{E}[x^n] = Z^{-1} \int_0^\infty x^n p(x) b(x) dx \tag{B9}$$

where x is P, E, or Q, $n = 1, 2$, $p(x)$ is the unconstrained Gaussian prior of x, $b(x)$ is the backward message to x (Eq. B6-B8), and $Z = \int_0^\infty p(x)b(x)dx$. The posterior is then approximated by a Gaussian with mean equal to $\mathbb{E}[x]$ and variance equal to $\mathbb{E}[x^2] - \mathbb{E}[x]^2$. Finally, using a Gaussian division formula analogous to the Gaussian multiplication formula given earlier, the input messages used in Eq. B1 and B5 are computed by dividing the approximate Gaussian posterior by the corresponding backward message $b(x)$.

This creates a mutual dependence that is solved by iteration: repeat forward and backward passes over the entire time-series until the approximate posteriors don't change anymore.

Acknowledgments

This work was supported by Dutch Research Council (NWO) Visitor Travel Grant no. 040.11.731. Data and code used in this study are available at `http://doi.org/10.5281/zenodo.4116451`.

References

Abrishamchi, A., & Tajrishy, M. (2005). Interbasin water transfers in Iran. In *Water conservation, reuse, and recycling: Proceedings of an Iranian-American workshop.* National Academies Press.

Aires, F. (2014). Combining datasets of satellite-retrieved products. Part I: Methodology and water budget closure. *Journal of Hydrometeorology*, *15*(4), 1677–1691.

Alemohammad, S., McColl, K., Konings, A., Entekhabi, D., & Stoffelen, A. (2015). Characterization of precipitation product errors across the US using multiplicative triple collocation. *Hydrology and Earth System Sciences*, *12*(2).

Allam, M. M., Jain Figueroa, A., McLaughlin, D. B., & Eltahir, E. A. (2016). Estimation of evaporation over the upper Blue Nile basin by combining observations from satellites and river flow gauges. *Water Resources Research*, *52*(2), 644–659.

Bai, P., Liu, X., & Liu, C. (2018). Improving hydrological simulations by incorporating GRACE data for model calibration. *Journal of Hydrology*, *557*, 291–304.

Beck, H. E., Vergopolan, N., Pan, M., Levizzani, V., Van Dijk, A. I., Weedon, G. P., ... Wood, E. F. (2017). Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling. *Hydrology and Earth System Sciences*, *21*(12), 6201–6217.

Chen, M., Senay, G. B., Singh, R. K., & Verdin, J. P. (2016). Uncertainty analysis of the Operational Simplified Surface Energy Balance (SSEBop) model at multiple flux tower sites. *Journal of Hydrology*, *536*, 384–399.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Rowland, J., ... Verdin, A. (2014). A quasi-global precipitation time series for drought monitoring. *U.S. Geological Survey Data Series 832*. `data.chc.ucsb.edu/products/CHIRPS-2.0`.

Hobeichi, S., Abramowitz, G., Contractor, S., & Evans, J. (2020). Evaluating precipitation datasets using surface water and energy budget closure. *Journal of Hydrometeorology*, *21*(5), 989–1009.

Hobeichi, S., Abramowitz, G., Evans, J., & Ukkola, A. (2018). Derived optimal linear combination evapotranspiration (dolce): a global gridded synthesis et estimate. *Hydrology and Earth System Sciences (Online)*, *22*(2).

Horner, I., Renard, B., Le Coz, J., Branger, F., McMillan, H., & Pierrefeu, G. (2018). Impact of stage measurement errors on streamflow uncertainty. *Water Resources Research*, *54*(3), 1952–1976.

Huffman, G. (1997). Estimates of root-mean-square random error for finite samples of estimated precipitation. *Journal of Applied Meteorology*, *36*(9), 1191–1201.

Huffman, G., Stocker, E., Bolvin, D., Nelkin, E., & Tan, J. (2019). GPM IMERG Final Precipitation L3 1 month 0.1 degree x 0.1 degree V06. *Greenbelt, MD, Goddard Earth Sciences Data and Information Services Center (GES DISC).* `gpm.nasa.gov/data/directory`. doi: 10.5067/GPM/IMERG/3B-MONTH/06

IWRMC. (2020). *Iran Water Resources Management Company.* `http://wrs.wrm`

.ir/amar/login.asp.

Jiang, L., Wu, H., Tao, J., Kimball, J. S., Alfieri, L., & Chen, X. (2020). Satellite-based evapotranspiration in hydrological model calibration. *Remote Sensing*, *12*(3), 428.

Khan, M. S., Liaqat, U. W., Baik, J., & Choi, M. (2018). Stand-alone uncertainty characterization of GLEAM, GLDAS and MOD16 evapotranspiration products using an extended triple collocation approach. *Agricultural and Forest Meteorology*, *252*, 256–268.

Kiang, J. E., Gazoorian, C., McMillan, H., Coxon, G., Le Coz, J., Westerberg, I. K., … others (2018). A comparison of methods for streamflow uncertainty estimation. *Water Resources Research*, *54*(10), 7149–7176.

Liu, W., Wang, L., Zhou, J., Li, Y., Sun, F., Fu, G., … Sang, Y.-F. (2016). A worldwide evaluation of basin-scale evapotranspiration estimates against the water balance method. *Journal of Hydrology*, *538*, 82–95.

Long, D., Longuevergne, L., & Scanlon, B. R. (2014). Uncertainty in evapotranspiration from land surface modeling, remote sensing, and GRACE satellites. *Water Resources Research*, *50*(2), 1131–1151.

Lopez, P. L., Sutanudjaja, E. H., Schellekens, J., Sterk, G., & Bierkens, M. F. (2017). Calibration of a large-scale hydrological model using satellite-based soil moisture and evapotranspiration products. *Hydrology and Earth System Sciences*, *21*(6), 3125–3144.

Maghsood, F. F., Hashemi, H., Hosseini, S. H., & Berndtsson, R. (2020). Ground validation of GPM IMERG precipitation products over Iran. *Remote Sensing*, *12*(1), 48.

Martens, B., Gonzalez Miralles, D., Lievens, H., Van Der Schalie, R., De Jeu, R. A., Fernández-Prieto, D., … Verhoest, N. (2017). GLEAM v3: satellite-based land evaporation and root-zone soil moisture. *Geoscientific Model Development*, *10*(5), 1903–1925.

Massari, C., Crow, W., & Brocca, L. (2017). An assessment of the performance of global rainfall estimates without ground-based observations. *Hydrology and Earth System Sciences*, *21*(9), 4347.

Massari, C., & Maggioni, V. (2020). Error and uncertainty characterization. In *Satellite precipitation measurement* (pp. 515–532). Springer.

Minka, T. (2001). *A family of algorithms for approximate Bayesian inference* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Minka, T., Winn, J., Guiver, J., Zaykov, Y., Fabian, D., & Bronskill, J. (2018). *Infer.net 0.3.* (Microsoft Research Cambridge. http://dotnet.github.io/infer)

Moreira, A. A., Ruhoff, A. L., Roberti, D. R., de Arruda Souza, V., da Rocha, H. R., & de Paiva, R. C. D. (2019). Assessment of terrestrial water balance using remote sensing data in South America. *Journal of Hydrology*, *575*, 131–147.

Mueller, B., Seneviratne, S. I., Jimenez, C., Corti, T., Hirschi, M., Balsamo, G., … others (2011). Evaluation of global observations-based evapotranspiration datasets and ipcc ar4 simulations. *Geophysical Research Letters*, *38*(6).

Munier, S., Aires, F., Schlaffer, S., Prigent, C., Papa, F., Maisongrande, P., & Pan, M. (2014). Combining data sets of satellite-retrieved products for basin-scale water balance study: 2. Evaluation on the Mississippi basin and closure correction model. *Journal of Geophysical Research: Atmospheres*, *119*(21), 12–100.

Odusanya, A. E., Mehdi, B., Schürz, C., Oke, A. O., Awokola, O. S., Awomeso, J. A., … Schulz, K. (2019). Multi-site calibration and validation of SWAT with satellite-based evapotranspiration in a data-sparse catchment in southwestern Nigeria. *Hydrology and Earth System Sciences*, *23*(2).

Pan, M., Sahoo, A. K., Troy, T. J., Vinukollu, R. K., Sheffield, J., & Wood, E. F. (2012). Multisource estimation of long-term terrestrial water budget for major global river basins. *Journal of Climate*, *25*(9), 3191–3206.

Pan, M., & Wood, E. F. (2006). Data assimilation for estimating the terrestrial water budget using a constrained ensemble Kalman filter. *Journal of Hydrometeorology*, *7*(3), 534–547.

Pellet, V., Aires, F., Munier, S., Jordà, G., Prieto, D., Dorigo, W., ... Brocca, L. (2019). Integrating multiple satellite observations into a coherent dataset to monitor the full water cycle-application to the Mediterranean region. *Hydrology and Earth System Sciences*, *23*(1), 465–491.

Rientjes, T., Muthuwatta, L. P., Bos, M., Booij, M. J., & Bhatti, H. (2013). Multi-variable calibration of a semi-distributed hydrological model using streamflow data and satellite-based evapotranspiration. *Journal of hydrology*, *505*, 276–290.

Sahoo, A. K., Pan, M., Troy, T. J., Vinukollu, R. K., Sheffield, J., & Wood, E. F. (2011). Reconciling the global terrestrial water budget using satellite remote sensing. *Remote Sensing of Environment*, *115*(8), 1850–1865.

Save, H. (2020). CSR GRACE and GRACE-FO RL06 Mascon Solutions v02. *doi: 10.15781/cgq9-nh24*.

Scanlon, B., Zhang, Z., Rateb, A., Sun, A., Wiese, D., Save, H., ... others (2019). Tracking seasonal fluctuations in land water storage using global models and GRACE satellites. *Geophysical Research Letters*, *46*(10), 5254–5264.

Scanlon, B., Zhang, Z., Save, H., Sun, A. Y., Schmied, H. M., Van Beek, L. P., ... others (2018). Global models underestimate large decadal declining and rising water storage trends relative to GRACE satellite data. *Proceedings of the National Academy of Sciences*, *115*(6), E1080–E1089.

Senay, G. B., Kagone, S., & Velpuri, N. M. (2020). Operational global actual evapotranspiration: Development, evaluation and dissemination. *Sensors*, *20*(7), 1915. `earlywarning.usgs.gov/fews/product/460`.

Simons, G., Bastiaanssen, W., Ngô, L. A., Hain, C. R., Anderson, M., & Senay, G. (2016). Integrating global satellite-derived data products as a pre-analysis for hydrological modelling studies: A case study for the Red River basin. *Remote Sensing*, *8*(4), 279.

ter Braak, C. J., & Vrugt, J. A. (2008). Differential evolution Markov chain with snooker updater and fewer chains. *Statistics and Computing*, *18*(4), 435–446.

Tian, Y., & Peters-Lidard, C. D. (2010). A global map of uncertainties in satellite-based precipitation measurements. *Geophysical Research Letters*, *37*(24).

Wan, Z., Zhang, K., Xue, X., Hong, Z., Hong, Y., & Gourley, J. J. (2015). Water balance-based actual evapotranspiration reconstruction from ground and satellite observations over the conterminous United States. *Water Resources Research*, *51*(8), 6485–6499.

Wang, S., Huang, J., Yang, D., Pavlic, G., & Li, J. (2015). Long-term water budget imbalances and error sources for cold region drainage basins. *Hydrological Processes*, *29*(9), 2125–2136.

Weerasinghe, I., Griensven, A. v., Bastiaanssen, W., Mul, M., & Jia, L. (2019). Can we trust remote sensing ET products over Africa? *Hydrology and Earth System Sciences Discussions*, 1–27.

Wiese, D., Landerer, F. W., & Watkins, M. M. (2016). Quantifying and reducing leakage errors in the JPL RL05M GRACE mascon solution. *Water Resources Research*, *52*(9), 7490–7502.

Wiese, D., Yuan, D., Boening, C., Landerer, F., & Watkins, M. (2018). JPL GRACE mascon ocean, ice, and hydrology equivalent water height, Release 06, Coastal Resolution Improvement (cri) filtered version 1.0. ver. 1.0. *PO.DAAC, CA, USA*. `grace.jpl.nasa.gov/data`.

Yang, X., Yong, B., Ren, L., Zhang, Y., & Long, D. (2017). Multi-scale validation of GLEAM evapotranspiration products over China via ChinaFLUX ET measurements. *International Journal of Remote Sensing*, *38*(20), 5688–5709.

Zhang, L., Dobslaw, H., Stacke, T., Güntner, A., Dill, R., & Thomas, M. (2017).

Validation of terrestrial water storage variations as simulated by different global numerical models with GRACE satellite observations. *Hydrology and Earth System Sciences*, *21*(2).

Zhang, Y., Pan, M., Sheffield, J., Siemann, A. L., Fisher, C. K., Liang, M., … others (2018). A Climate Data Record (CDR) for the global terrestrial water budget: 1984–2010. *Hydrology and Earth System Sciences*, *22*(PNNL-SA-129750).

Zhang, Y., Pan, M., & Wood, E. F. (2016). On creating global gridded terrestrial water budget estimates from satellite remote sensing. *Surveys in Geophysics*, *37*(2), 249–268.