How well are we able to close the water budget at the global scale?

Fanny Lehmann[1], Bramha Dutt Vishwakarma[1,2], and Jonathan Bamber[1,3]

[1]School of Geographical Sciences, University of Bristol, United Kingdom
[2]Indian Institute of Science, Bangalore, India
[3]Technical University Munich, Germany

Correspondence: Fanny Lehmann (fanny.lehmann@bristol.ac.uk)

Abstract. The water budget equation describes the exchange of water between the land, ocean and atmosphere. Being able to adequately close the water budget gives confidence in our ability to model and/or observe the spatiotemporal variations in the water cycle and its components. Due to advances in observation techniques, satellite sensors, and modelling, a number of data products are available that represent the components of water budget both in space and time. Despite these advances, closure of the water budget at global scale has been elusive.

In this study, we attempt to close the global water budget using precipitation, evapotranspiration, and runoff data at the catchment scale. The large number of recent state-of-the-art datasets provides a new evaluation of well-used datasets. These estimates are compared to terrestrial water storage (TWS) changes as measured by the GRACE satellite mission. We investigated 189 river basins covering more than 90 % of the continental land area. TWS changes derived from the water balance equation were compared against GRACE data using two metrics: the Nash-Sutcliffe Efficiency (NSE) and cyclostationary NSE. These were used to assess the performance of more than 1600 combinations of the various datasets considered.

We found a positive NSE and cyclostationary NSE in 99% and 62% of the basins examined, respectively. This means that TWS changes reconstructed from the water balance equation were more accurate than the long-term (NSE) and monthly (cyclostationary NSE) mean of GRACE time series in the corresponding basins. By analyzing different combinations of the datasets that make up the water balance, we identified data products that performed well in certain regions based on, for example, climatic zone. We identified that some of the good results were obtained due to cancellation of errors in poor estimates of water budget components. Therefore, we used coefficients of variation to determine the relative quality of a data product, which helped us to identify bad combinations giving us good results. In general, water budget components from the ERA5 Land and the Catchment Land Surface Model (CLSM) performed better than other products for most climatic zones. Conversely, the latest version of the Catchment Land Surface Model, v2.2, performed poorly for evapotranspiration in snow-dominated catchments compared, for example, to its predecessor and other datasets available. Thus, the nature of the catchment dynamics and balance between components affects the optimum combination of datasets. For regional studies, the combination of datasets that provides the most realistic TWS for a basin will depend on its climatic conditions and factors that cannot be determined a-priori. We believe, the results of this study provide a roadmap for studying the water budget at catchment scale.

1 Introduction

A better understanding of hydrological processes at the catchment scale has been highlighted as one of the key challenges for hydrologists in the 21st century (Blöschl et al., 2019). One of the key processes is the terrestrial water cycle which can be described by the water balance equation,

$$\frac{dTWS}{dt} = P - ET - R. \tag{1}$$

This equation expresses the total amount of water gained by a river catchment in the form of precipitation (P) as a sum of, water returning back to the atmosphere through evapotranspiration (ET), water flowing out of the catchment in the form of runoff (R), and any changes in the terrestrial water storage (TWS). TWS is defined as the sum of water stored as snow, canopy, soil moisture, groundwater, and surface water (Scanlon et al., 2018). The water balance equation is a budget equation that follows the conservation of mass and it is an indispensable tool for validating our understanding of the catchment scale water cycle.

Several studies have used the water balance equation to explain hydro-climatic changes experienced in a river catchment (e.g., Landerer et al., 2010; Pan et al., 2012; Oliveira et al., 2014; Saemian et al., 2020), to validate modelled estimates of one component (e.g., Bhattarai et al., 2019; Long et al., 2015; Wan et al., 2015), or to estimate one component when others are known (Chen et al., 2020; Gao et al., 2010; Wang et al., 2014). It should be noted however that in these studies the accuracy of the result is limited by uncertainties associated ~~to~~ with individual components. For example, Sahoo et al. (2011) attempted to close the water balance equation for 10 large catchments and found that the imbalance error amounted to up to 25% of mean annual precipitation. Additionally, Zhang et al. (2018) highlighted the source of the imbalance error as being predominantly from stark disagreement between evapotranspiration estimates.

Obtaining high quality spatiotemporal estimates of components of the water balance is challenging due to a lack of global in situ measurement networks and political will to sustain any existing network. Therefore, the era of satellite remote sensing ~~offered~~ offers an excellent solution to monitoring the hydrosphere. With the help of dedicated satellite missions, we are able to measure variables that can be used to estimate water balance components. However monitoring TWS has been the most difficult part since it includes water on and below the surface of the Earth, and optical remote sensing can only offer information near the surface. This issue was solved by the launch of a satellite gravimetry mission from GFZ and NASA in 2002, also known as Gravity Recovery And Climate Experiment (GRACE) (Wahr et al., 1998; Tapley, 2004). This mission measures the temporal variations in the Earth's gravity field, which can then be related to water mass change on and below the surface of the Earth. GRACE provides the most accurate global estimations of TWS to date, which can be used in the water balance equation 1.

Another challenge concerns components like ET with a high spatial variability, which requires precise satellite estimates, not consistently available due to observational constraints (Fisher et al., 2017). Since ET accounts for up to 60% of precipitation in some regions, it is a crucial component of the water cycle (Oki and Kanae, 2006). It also constitutes the most significant uncertainties of the terrestrial water cycle components (Rodell et al., 2015). The water balance equation has been used to compensate for this lack of knowledge and increase our understanding of ET. Water-budget studies have generally found that ET inferred from the water balance equation agrees well with remote sensing estimates in terms of seasonal cycle but presents

larger inter-annual variability (Liu et al., 2016; Pascolini-Campbell et al., 2020; Swann and Koven, 2017) and larger magnitudes (Bhattarai et al., 2019; Long et al., 2014a; Wan et al., 2015).

60   Apart from ET, our ~~apprehension~~ knowledge of R also benefits from water budget estimations. Although river discharge can be measured by gauges, the spatio-temporal coverage of in situ measurements is limited due to a lack of ~~money~~ resources in some regions and political will to share data. Uncertainties and biases in P have been found to be the main drivers of the inaccuracy in budget inferred R (Sheffield et al., 2009; Oliveira et al., 2014; Sneeuw et al., 2014; Wang et al., 2014; Xie et al., 2019). Water budget studies using R as a reference variable also point out the difficulty to find datasets able to close the water

65   budget (Chen et al., 2020; Gao et al., 2010; Lorenz et al., 2014). Moreover, ET and R are strongly intertwined and accurate estimates of one cannot be achieved without a better constraint on the other (Armanios and Fisher, 2014; Lv et al., 2017; Penatti et al., 2015).

To improve the reliability of available data, the water budget can be used as a discriminating tool to assess the accuracy of various datasets. For this to be achieved, there is a need to first evaluate the water budget closure globally, including basins

70   of all sizes, and comparing as many state-of-the-art datasets as possible. This review is currently lacking because first, a majority of studies have concentrated only on a few selected basins with specific climatic conditions (e.g. the Amazon basin, ~~(Swann and Koven, 2017; Chen et al., 2020)~~Swann and Koven, 2017; Chen et al., 2020) or highly impacted by human activities (e.g. the Yellow river basin, ~~(Lv et al., 2017; Long et al., 2015)~~Lv et al., 2017; Long et al., 2015). Additionally, the studies which look at several basins worldwide have only evaluated sparsely distributed basins, which leaves entire zones without

75   analysis (Sahoo et al., 2011; Pan et al., 2012; Lorenz et al., 2014; Liu et al., 2016; Zhang et al., 2018). This has deprived hydrologists of a comprehensive global overview of the water budget.

Returning to the requirement for basins of all sizes, basins were also generally chosen to be quite large in the majority of studies. It is known that the accuracy of GRACE measurements is directly proportional to the size of the basin (Rodell and Famiglietti, 1999; Wahr et al., 2006; Vishwakarma et al., 2018), however the lower limit of $\sim 200,000\ km^2$ established by

80   Longuevergne et al. (2010) and which has long been used is no longer a requirement to retrieve GRACE signals. It has been shown that basins as small as $\sim 70,000\ km^2$ can be precisely recovered by GRACE measurements and that their size do not influence the closure of the water budget (Gao et al., 2010; Lorenz et al., 2014; Vishwakarma et al., 2018). They are therefore included in the current study.

Regarding the number of datasets to be examined, each water budget study uses different datasets, some of which were

85   available only over a given continent or over short time periods. To the authors' best knowledge, Lorenz et al. (2014) conducted the study comparing the largest number of datasets by assessing more than 180 combinations of P, ET, and TWS datasets. However, many datasets have since improved, especially reanalyses such as Era-Interim (Dee et al., 2011) and MERRA Land (Reichle et al., 2011). It would be beneficial to provide an updated evaluation of those widely used datasets.

The aim of the current study is thus to provide a revised overview of the water budget closure on a global scale. Section 2

90   presents the study area covering all parts of the globe (excluding Greenland and Antarctica) and the datasets. Then, section 3 details the metrics used to evaluate the water budget closure as well as the selection process for the best combinations. Moreover, section 4 explains the results and discusses previous studies.

Figure 1. 189 basins larger than 63,000 km^2 with their corresponding climate zone

2 Data

2.1 Study area

We used the major river basins from the Global Runoff Data Centre (GRDC, 2020) to define the study area. Since the spatial resolution of GRACE products for hydrological applications is around 63,000 km^2 (Vishwakarma et al., 2018), catchments larger than this limit have been included in our analysis. Furthermore, these basins were assigned to a climate zone as defined by the Köppen-Geiger classification (Kottek et al., 2006). The 189 basins under study are depicted in Fig. 1 and their areas range from $\sim$ 65,600 km^2 to $\sim$ 5,965,900 km^2.

2.2 Datasets

We have used freely available global state-of-the-art datasets with a temporal resolution smaller than or equal to one month and coverage of at least 2003 to 2014. If necessary, data have been interpolated to 0.5° x 0.5° grids using bilinear interpolation to correspond with monthly TWS derived from GRACE satellite mission. In this study, GRACE mascon fields were obtained from the Jet Propulsion Laboratory (JPL) RL06 (Watkins et al., 2015; Wiese et al., 2018). Our results were also computed with mascons from the Center for Space Research (CSR) and they can be easily reproduced with the code we provided. Since this did not significantly changed our findings, we only ~~showed~~ show results using JPL mascons.

For other variables, daily data were aggregated to monthly values taking into account the number of days per month. Finally, gridded data were weighted by the area of each grid cell and then aggregated over a basin to obtain a time-series.

2.2.1 Precipitations datasets

Precipitation data were obtained from various sources that are summarised in Table S1. Three datasets rely only on rain-gauge measurements, namely the Climate Research Unit (CRU) which uses around 10,000 gauges (Harris et al., 2020), the Global Unified Gauge-Based Analysis of Daily Precipitation from the Climate Prediction Center (CPC) based on approximately 30,000 gauges (Chen and Xie, 2008), and the Global Precipitation Climatology Centre (GPCC) maintaining a database of around 67,000 gauges (Schneider et al., 2020). Surface observations are often used to calibrate satellite estimations or as input variables in reanalyses. Since the global coverage of rain gauges is not homogeneous, the quality of such products varies regionally, thus satellite-based products provide a good alternative.

Two satellite missions were specifically designed to measure precipitation. The Tropical Rainfall Measuring Mission (TRMM) operated from 1998 to 2015 and provided monthly estimations of precipitation over 50° N to 50° S. We used the TMPA 3B43 version that extends TRMM measurements until 2020 via calibration with other satellites (Huffman et al., 2007, 2010). The Global Precipitation Measurement mission (GPM) was built on TRMM findings since its launch in February 2014. This constellation of satellites is calibrated using previous satellites through the Integrated Multi-satellitE Retrievals for GPM (IMERG) to provide global coverage from 2000 onwards (Huffman et al., 2019). Finally, the Global Precipitation Climatology Project (GPCP) merges various satellite-based estimates with rain-gauge measurements from the GPCC (Adler et al., 2018). It provides a well-used and long dataset spanning from 1979 to the present.

Apart from these, reanalyses products provide consistent estimations of precipitation, evapotranspiration, and runoff. ERA5-Land is a rerun of the land component from the ERA5 reanalysis developed by the European Centre for Medium-Range Weather Forecasts (ECMWF). Precipitation data are obtained from satellite measurements including but not restricted to TRMM and GPM results and are provided from 1981 onwards (Muñoz-Sabater, 2019). The Japanese 55-year Reanalysis (JRA55) also derives precipitation from satellite measurements with forecasts starting in 1958 (Kobayashi et al., 2015). Finally, the Modern-Era Retrospective Analysis for Research and Applications, version 2 (MERRA-2) uses two precipitation datasets from the CPC: the Global Unified Gauge-Based Analysis of Daily Precipitation described above and the Merged Analysis of Precipitation which combines gauge-based and satellite measurements (Reichle et al., 2017).

Finally, two additional datasets that combine rain-gauge observations, satellite measurements, and reanalyses: the Princeton Global Forcing dataset (PGF) and the Multi-Source Weighted Ensemble Precipitation (MSWEP), were used in this study. PGF was included as it is one of the forcing variables used in the Global Land Data Assimilation System (GLDAS) (Sheffield et al., 2006). Recently developed, MSWEP merges gauge observations (including GPCC), satellite measurements (including TRMM), and reanalyses (ERA-Interim and JRA55) (Beck et al., 2019).

Since there are large disagreements between different datasets, it is important to assess whether a dataset is in general agreement to others. By revealing datasets with significant bias, this method can limit the occurrence of error cancellation, which is a well-known problem in water budget studies (Sneeuw et al., 2014; Lorenz et al., 2014). We have used the coefficient of variation (CV) to evaluate various datasets of a water budget component in each basin. From a group of datasets, the CV is a time-series defined as the standard deviation divided by the mean. (A minimum value of 10 mm was enforced for the mean

to avoid high CVs during the dry season). The higher the CV, the greater the disagreement between datasets. Figure S1 shows the mean of CV time-series in each basin. Unsurprisingly, satellite datasets (TRMM, GPM, and GPCP) provide close results since they use similar measurements and are therefore not at all independent. Observations datasets (CPC, CRU, and GPCP) are more independent, which leads to higher CVs. However, apart from Australia where CRU led to precipitations consistently smaller than CPC and GPCC, there were no common patterns in the other regions. In addition, the major differences between reanalyses were found in Central Asia where MERRA2 gave much smaller precipitation values than ERA5 Land and JRA55. Interestingly, Fig. S1 also shows that the method used to create the dataset (i.e. rain gauge observations, satellite measurements, or reanalyses) is less ~~discriminant~~ relevant than differences within a method. The inter-category CV measuring differences between the mean of observations, satellite, and reanalyses datasets was found to be relatively low. The highest CVs were found in high latitude basins where reanalyses consistently led to larger precipitations whilst observations had the smallest precipitation values.

~~Coefficient of variation between different sets of precipitation datasets. Satellite: TRMM, GPM, and GPCP. Observations: CPC, CRU, and GPCC. Reanalyses: ERA5 Land, JRA55, and MERRA2. Inter-category: mean of satellite, observations, and reanalyses.~~

2.2.2 Evapotranspiration datasets

Evapotranspiration is the sum of evaporation from water surfaces and transpiration through vegetation. Datasets used in this study are listed in Table S2. One of the most accurate methods to estimate evapotranspiration is the Penman-Monteith equation (Penman, 1948; Monteith, 1965). The variables used in this equation are obtained from various land surface parametrizations and energy balance equations in reanalyses ERA5 Land and MERRA2, and in GLDAS land surface models (LSMs). We chose three variants of the GLDAS: the Variable Infiltration Capacity (VIC, Liang et al. (1994)), the Noah model (Chen et al., 1996; Koren et al., 1999; Ek et al., 2003), and the Catchment Land Surface Model (CLSM, Koster et al. (2000)). These LSMs are forced with different data depending on GLDAS version (Rodell et al., 2004). For example, PGF precipitation was used in version 2.0, GPCP precipitation in version 2.1, and ERA5 precipitation in version 2.2 coupled with GRACE data assimilation (for CLSM only, (Li et al., 2019)). MOD16 algorithm also uses the Penman-Monteith equation with measurements from the Moderate-Resolution Imaging Spectroradiometer (MODIS, NASA) (Mu et al., 2011).

One of the main drawbacks of the Penman-Monteith equation is the reliance on a large number of parameters such as vegetation characteristics, air temperature, wind, vapour pressure, etc. Since these parameters can be difficult to assess accurately, alternative approaches have been developed. For example, the Global Land Evaporation Amsterdam Model (GLEAM) uses an equation involving fewer parameters, the Priestley-Taylor equation (Martens et al., 2017; Miralles et al., 2011). Another method relies on the energy budget to compute the fraction of energy leading to water vaporization, as done in the Simplified Surface Energy Balance for operational applications (SSEBop), (Senay et al., 2013). Finally, algorithms also take advantage of the FLUXNET network of eddy-covariance towers measuring evapotranspiration. To this extent, the machine learning FLUXCOM algorithm (Jung et al., 2019) extends the methodology of the well-used Multi-Tree Ensemble (Jung et al., 2009) by exploiting relationships between meteorological variables and latent heat flux measured by eddy-covariance towers.

Similar to precipitation, Fig. S2 shows the coefficient of variation for different categories of evapotranspiration datasets. CVs were relatively low between the mean of all categories, as was found for precipitation. The largest differences between reanalyses were also found in Central Asia with MERRA2 predicting lower evapotranspiration. In addition, it is striking to see the large CVs among land surface models (CLSM, Noah, and VIC with versions 2.0 and 2.1). In this category, there were consistent patterns across all basins with VIC tending to underestimate ET while CLSM provided slightly larger values. The CVs were especially large in high-latitude basins due to low ET in the cold season. Moreover, in Fig. S2 we see that the differences between remote sensing datasets (FLUXCOM, GLEAM, MOD16, and SSEBop) are not spatially consistent. In Australia, MOD16 led to significantly lower ET, especially during the hot season (October to February). In South Africa, differences were constant all year long with MOD16 being lower while FLUXCOM was rather high. We do not comment on CVs in hot deserts (Sahara, Arabian peninsula, and Central Asia) because FLUXCOM and MOD16 are not available in non-vegetated land areas.

Coefficient of variation between different sets of evapotranspiration datasets. Remote sensing: FLUXCOM, GLEAM, MOD16, and SSEBop. Land surface models: CLSM, Noah, and VIC with versions 2.0 and 2.1. Reanalyses: ERA5 Land, JRA55, and MERRA2. Inter-category: mean of remote sensing, LSMs, and reanalyses.

2.2.3 Runoff datasets

Runoff is computed in LSMs as the excess water not evaporated from soils. This water infiltrates through the soil to the lowest layers without communicating with adjacent grid cells. All LSMs presented above provide runoff estimates that were included in this study. River discharge measurements are also available from gauge records but they are not temporally consistent and heterogeneously distributed along rivers. To exploit these in situ measurements without the need to pre-process themacross the study period. In addition, discharge areas from the gauge stations with the longest records do not necessarily match the area of GRDC basins that we selected. Therefore, we decided to use only spatially and temporally consistent datasets by excluding gauge records from our analyses. However, we used the recently developed machine learning GRUN dataset which provides runoff values at 0.5° x 0.5° spatial resolution from 1902 to 2014 (Ghiggi et al., 2019). This algorithm was trained with precipitation, temperature, and runoff measurements and validated against independent river discharge observations from the GRDC.

As for precipitation and evapotranspiration, Fig. S3 shows the coefficients of variation. CVs were generally higher for runoff than that for evapotranspiration and precipitation. Even though it reflects high uncertainties in runoff values, this should play a relatively smaller role in the water balance because the runoff is the smallest water cycle component. In Fig. S3, the inter-category CVs were computed between GRUN, the mean of LSMs, and the mean of reanalyses. The general observations are complementary to those made about evapotranspiration. VIC generally led to the highest values among all datasets. Reanalyses tended to be lower, along with CLSM. Finally, compared with the mean across all datasets, GRUN was relatively close in general (not shown). The largest differences were found in Australia and Central Africa where GRUN was lower, and in Central Asia where it led to higher values.

3   Methods

3.1   Water budget reconstruction

GRACE mascon fields were used to compute time-series of TWS anomalies relative to the mean between 2004 and 2009.

215   Since equation 1 involves the variation of TWS over a time period, which is called Terrestrial Water Storage Change (TWSC). To obtain TWSC from TWS anomalies, the time ~~derivation~~ derivative was computed with centered finite difference (as in e.g., ~~Long et al. (2014b)~~ Long et al. (2014a) or Pascolini-Campbell et al. (2020))

$$TWSC(t) = \frac{TWS(t+1) - TWS(t-1)}{2\Delta t}, \tag{2}$$

where Δt equals 1 month and $t-1$, t, $t+1$ are three consecutive months. Missing monthly values were filled with cubic

220   interpolation. In order to match the temporal shift induced by the central difference, time-series of P, ET, and R also needed to be time-filtered by equation 3 (Landerer et al., 2010)

$$\tilde{X}(t) = \frac{1}{4}X(t-1) + \frac{1}{2}X(t) + \frac{1}{4}X(t+1), \tag{3}$$

where X denotes either P, ET, or R. All variables referred to hereafter are filtered variables but are denoted without the tilde notation for the sake of clarity.

225   Each triplet of datasets $(data_P, data_{ET}, data_R)$ was called a *combination* and led to a *budget reconstruction* of TWSC computed with equation 1: $TWSC_{budget}(t) = P_{data_P}(t) - ET_{data_{ET}}(t) - R_{data_R}(t)$. This reconstruction was compared with the derivatives obtained from equation 2 and denoted $TWSC_{GRACE}(t)$. Since we used 11 precipitation, 14 evapotranspiration, and 11 runoff datasets, we finally evaluated 1694 combinations.

3.2   Metrics

230   Differences between two time-series are commonly evaluated with the Root Mean Square Deviation (RMSD)

$$RMSD = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(TWSC_{budget}(t) - TWSC_{GRACE}(t))^2}, \tag{4}$$

The main drawback of the RMSD is that it is not normalized *i.e.* basins with large TWSC tend to have larger RMSD. A very common normalization is the Nash-Sutcliffe Efficiency (NSE) introduced by Nash and Sutcliffe (1970) to evaluate modeled runoff compared to observations

235    $$NSE = 1 - \frac{\frac{1}{T}\sum_{t=1}^{T}(TWSC_{budget}(t) - TWSC_{GRACE}(t))^2}{\frac{1}{T}\sum_{t=1}^{T}(TWSC_{GRACE}(t) - \overline{TWSC_{GRACE}})^2} = 1 - \frac{RMSD^2}{\delta_{cst}^2}, \tag{5}$$

where $\overline{TWSC_{GRACE}} = \frac{1}{T}\sum_{t=1}^{T} TWSC_{GRACE}(t)$ is the long-term mean of TWSC and $\overline{err_{cst}}$ δ_{cst} is the deviation of monthly values from the long-term mean. In our case, any positive value of the NSE means that the budget reconstruction of $TWSC_{GRACE}$ is a better approximation than the long-term mean. The maximum value of 1 describes a perfect reconstruction and a negative value denotes a poor performance. One major advantage of the NSE is that it requires both phase agreement (usually assessed with the correlation coefficient) and a small long-term mean error (evaluated with the bias, or percentage bias) to yield high values (Lorenz et al., 2014).

However, although several attempts have been made to associate positive NSE values to a performance (e.g. Henriksen et al., 2003; Samuelsen et al., 2015), it is known that this index suffers from several weaknesses, for example, a high positive NSE can be obtained with a poor time-series if the time-series ~~have~~ has a large variance (Jain and Sudheer, 2008). In the context of the current study, basins with large seasonal variations of TWSC, especially tropical basins, are more likely to exhibit a NSE close ~~from~~ to 1 even though the budget reconstruction presents substantial errors.

To overcome this issue, it has been proposed to compare the budget reconstruction to the mean monthly value of TWSC instead of comparing it to the constant long-term mean. The so-called cyclostationary NSE (Thor, 2013; Zhang, 2019) ~~then writes~~ is then:

$$NSE_c = 1 - \frac{\frac{1}{T}\sum_{t=1}^{T}(TWSC_{budget}(t) - TWSC_{GRACE}(t))^2}{\frac{1}{T}\sum_{t=1}^{T}(TWSC_{GRACE}(t) - TWSC_{GRACE}^m)^2} = 1 - \frac{RMSD^2}{\delta_{cyc}^2}, \quad (6)$$

where $TWSC_{GRACE}^m$ is the mean value for month m over all years and $\overline{err_{cyc}}$ δ_{cyc} is the deviation of GRACE TWSC from the periodic monthly signal. Similarly to the NSE, positive values of the cyclostationary NSE indicate a budget reconstruction better than the mean annual cycle, which measures the ability of the reconstruction to capture anomalous events (Lorenz et al., 2015; Tourian et al., 2017).

Moreover, one can express the cyclostationary NSE in terms of the NSE by combining equations 5 and 6

$$NSE_c = \left(1 - \underbrace{\frac{err_{cst}^2}{err_{cyc}^2}\frac{\delta_{cst}^2}{\delta_{cyc}^2}}\right) + \underbrace{\frac{\delta_{cst}^2}{\delta_{cyc}^2}}_{\gamma} NSE. \quad (7)$$

The γ factor describes the behaviour of the TWSC by comparison with the mean seasonal cycle. Basins with periodic seasonal cycles (i.e., low $\overline{err_{cyc}}$ δ_{cyc}) or large magnitudes (i.e., high $\overline{err_{cst}}$ δ_{cst}) have larger γ. In those basins (e.g., the Amazon or Chad basins in Fig. S1), extremely high NSE values are required to achieve a positive cyclostationary NSE, as can be seen in Fig. 2. Special attention must then be given when examining such basins to discriminate performances depending on the NSE or the cyclostationary NSE.

3.3 Selection of the most ~~represented~~ representative datasets

When estimating a water cycle component from the water balance equation 1, it is useful to know beforehand which datasets are more reliable to close the water budget in the region under study. This section aims to describe how such datasets can be selected. The NSE results were stored in a matrix where each row corresponded to a basin and each column to a combination.
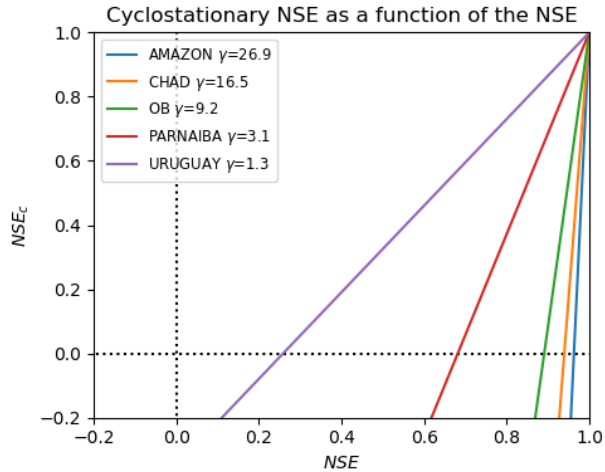
Figure 2. The cyclostationary NSE is related to the NSE through $NSE_c = 1 - \gamma + \gamma NSE$ where ~~$\gamma = \frac{err_{cst}^2}{err_{cyc}^2}$~~ $\gamma = \frac{\delta_{cst}^2}{\delta_{cyc}^2}$

Due to the matrix dimension (189×1694), an automated computation was needed to evaluate the combinations. This was achieved by introducing a cost function which represented the loss of accuracy when using any combination instead of the optimal one.

Our method can be summarised as follows:

1. compute the cost matrix to describe the performance of each combination

2. cluster basins into larger zones depending on the similarities between cost vectors

3. for each zone, select the combinations satisfying a maximum cost and extract the underlying datasets

In more details, the following steps were performed.

1. Using a cost function instead of the absolute metrics allowed us to overcome the lack of a NSE scale. On the one hand, there are significant differences between a combination leading to a budget reconstruction with a NSE close to 0 and another leading to an almost perfect reconstruction (NSE close to 1). These differences can be seen for example in terms of months where the budget reconstruction is within ~~GRACE confidence interval~~ the confidence interval from GRACE TWSCs. Therefore, we want to favour combinations leading to the highest NSE values. On the other hand, one cannot determine a NSE threshold assuring a satisfying reconstruction in all basins. Figure 2 shows that very high NSE values were needed in basins with large γ to outperform the monthly periodic signal. Consequently, a cost function evaluates the performances of a combination relatively to the largest NSE achievable in each basin. The cost function was then defined from the NSE by

$$c_i^b = \max_{comb} NSE^b(comb) - NSE^b(combination_i), \tag{8}$$

where the maximum was computed over all 1694 combinations. We emphasize that the cost was evaluated independently for each basin (denoted by the superscript b), allowing the maximum NSE to be different in each basin. For combinations leading

285 to a cost larger than 2 (*i.e.* a NSE below -1), the cost was restricted to 2. This limited the penalization of combinations with highly negative values but had no major influence on our results since we focused on the best performing combinations.

2. From the cost matrix, each basin could be represented by a vector of 1694 costs. The similarities between two basins b_1 and b_2 were evaluated based on the Euclidean distance between their respective cost vector, $d(b_1, b_2) = \sqrt{\sum_{i=1}^{1694}(c_i^{b_1} - c_i^{b_2})^2}$. For two basins to have a small Euclidean distance, each combination i should lead to a similar cost in all basins: either the

290 combination was satisfying in both cases ($c_i^{b_1} \simeq 0$ and $c_i^{b_2} \simeq 0$), or it did not perform well in both ($c_i^{b_1} \simeq 2$ and $c_i^{b_2} \simeq 2$). A hierarchical clustering algorithm was then applied to cluster basins so as to minimize the variance between cost vectors inside a cluster (Mueller et al., 2011).

3. Finally, the maximal cost for combinations to be considered as satisfying the water budget closure was chosen to be 0.1. This means that the difference between the RMSD of a suitable combination and the lowest RMSD over all combinations

295 is in average lower than $A/10$ where A is the mean seasonal amplitude of TWSC. This threshold guarantees that selected combinations have performances similar to the optimal combination. Then, in each cluster determined by the algorithm, we selected the combinations with a cost lower than 0.1 for all basins in the cluster. From the selected combinations, we extracted the underlying datasets of P, ET, and R. By reporting the number of combinations in which each dataset appeared, we could evaluate whether a dataset was clearly better than the others in a given region.

300 ## 4 Results and discussion

4.1 Water budget closure

In order to assess the global water budget closure, we first examined the best performances across all combinations. This means that for each basin, we reported the highest NSE among all 1694 combinations. Figure 3 shows the maximum NSE that can be achieved from a combination. Please note that a positive NSE was obtained over 99% of the total study area. Only 9 basins out

305 of 189 did not achieve a positive NSE for any combination. They were mainly hot arid deserts in Northern Sahara, Somalia, Australia, as well as two other basins in Papua New Guinea (Mamberamo basin) and Hayes basin (Canada) (Fig. 3). The poor performances in arid basins can be explained by limited precipitation and water storage variations that lead to a low signal-to-noise ratio. This is a major difference from previous studies where for example, Lorenz et al. (2014) found that only 29 basins out of 96 achieved a positive NSE.

310 Figure 3 can be interpreted as follows: all the basins with a positive NSE offer a budget reconstruction better than the long-term mean from GRACE TWSC. In addition, higher NSE values correspond to a better fit between reconstructed TWSC and GRACE TWSC. Figure S4 then shows the distribution of the maximum NSE. Although it has been explained in Section 3.2 that positive NSE should be interpreted cautiously, one can observe that 61% of the study area satisfied a NSE larger than 0.8 which is usually considered as very good performance (*e.g.* Henriksen et al. (2003), Samuelsen et al. (2015)). Being given the

315 large number of datasets, it is likely that cancellation of errors explains some of the good performances. The reader should remain cautious about this possibility when trying to reproduce our results and may use discrepancy measures such as the CV to examine datasets, as is explained in the following sections.
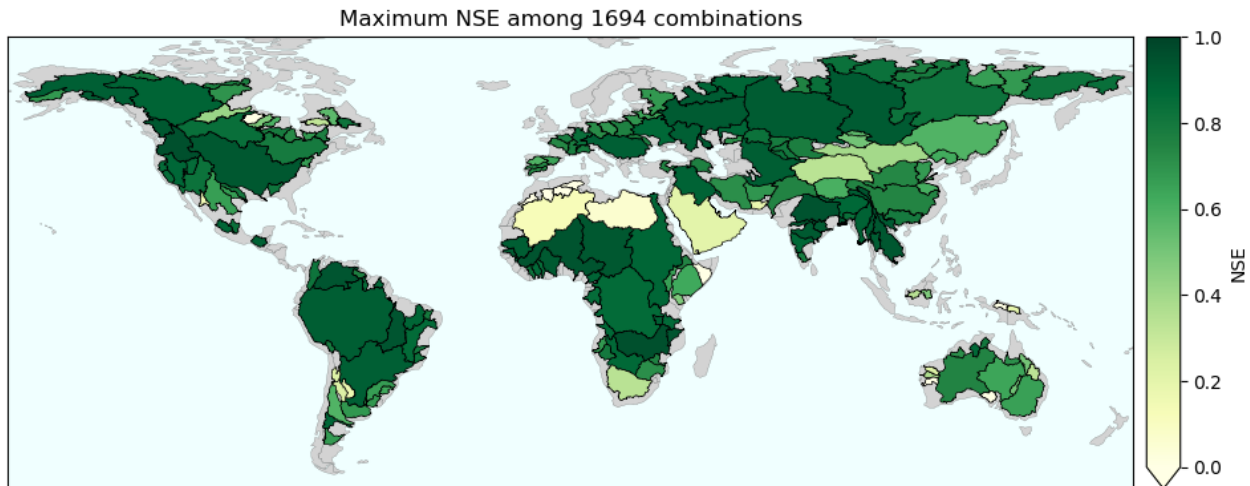
Figure 3. Maximum NSE per basin over all combinations. Green positive values mean that the budget reconstruction is a better approximation of GRACE TWSC than the long-term mean.

From its definition, the NSE can only be used to compare the budget reconstruction with the long-term mean. Since predicting intra-annual variations of TWSC would be more beneficial for hydro-meteorological studies, the cyclostationary NSE

320 was also used to assess the quality of reconstructed TWSC. Figure 4 shows that a positive maximum cyclostationary NSE was achieved over 62% of the study area. It means that in those basins, the reconstructed TWSC was better than the mean annual cycle obtained from GRACE TWSC. The budget reconstruction performed especially well in the continental United States and Central America, in most of Southern America except the Amazon and the Andes, in Southern Africa, Australia, Europe, West Russia, and East Asia (Fig. 4).

325 When comparing Fig. 3 and Fig. 4, one can observe that despite a very high NSE, some basins could not reach a positive cyclostationary NSE. This happened especially in tropical basins like the Amazon, some catchments in Western Africa, India, and Myanmar. These basins illustrate i) the limits of the NSE and ii) the need for a complementary metric to evaluate the reconstruction. These two points corroborate the conclusions of Jain and Sudheer (2008). The Amazon basin exemplifies why the NSE should not be used alone to assess the water budget closure. In fact, even with the best combination, the budget

330 reconstruction consistently underestimated the magnitude of TWSC (Fig. S5). TWSC was too low in the wet season (January-March) and too high in the dry season (July-August). This indicates that the budget reconstruction was not good enough to capture the inter-annual as well as annual variability in TWS. Due to the large amplitude of TWSC in the Amazon basin ($[-100; 100$ mm/month]), the NSE was still very high ($maxNSE = 0.91$) and could mislead us into concluding that the budget reconstruction is excellent. However, when assessing the cyclostationary NSE ($maxNSE_c = -1.28$), it appeared that the mean

335 monthly values were a better fit to GRACE values than the budget reconstruction (Fig. S5).
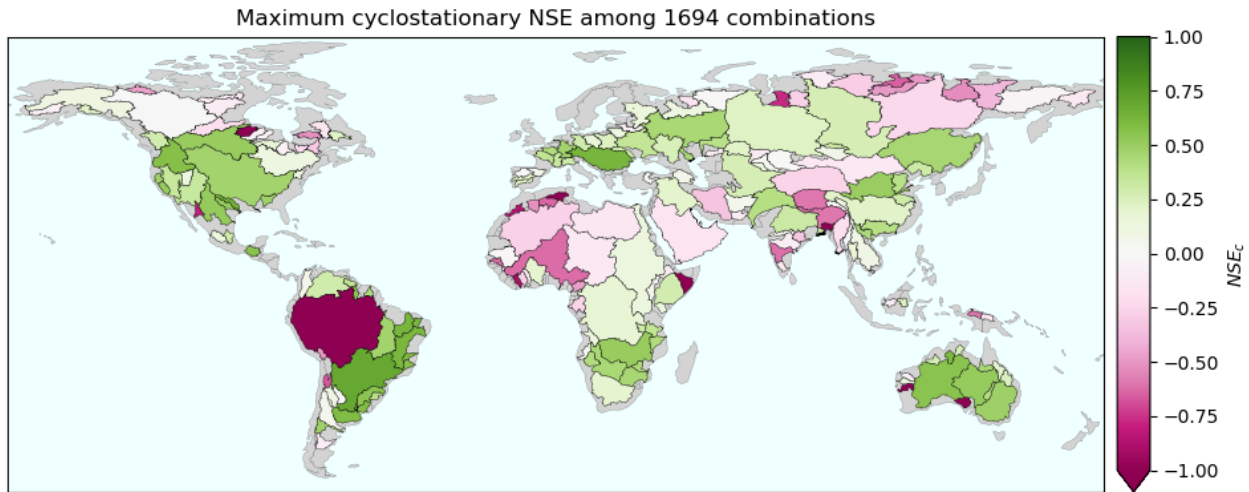
Figure 4. Maximum cyclostationary NSE per basin over all combinations. Green positive values mean that the budget reconstruction is a better approximation of GRACE TWSC than the mean monthly values.

The underestimation of annual variability in TWSC can be seen in the correlation plot between GRACE TWSC and our approximation (Fig. S6). Due to the error in approximating the largest TWSC, the regression slope is 0.7, while 1 is the optimal value. Figure S5 additionally shows that the water balance error is larger than GRACE uncertainty in 21% of months, meaning that the error is significant.

340 However, one should not conclude that all basins with a high NSE and negative cyclostationary NSE exhibit the same behaviour. The Niger basin is indeed another basin with a high NSE (0.94) and a negative cyclostationary NSE (-0.62). Contrary to the Amazon, there was no consistent pattern in the water closure error and the error was lower than GRACE uncertainty in 94% of months (Fig. S7). The regression slope was also almost perfect as shown in Fig. S8. In such a basin with low inter-annual variability, the error between GRACE TWSC and the mean monthly signal is very low (RMSD=6.6 mm/month).

345 Therefore, achieving a budget reconstruction more accurate than the monthly signal may be an unrealistic expectation.

In conclusion, while the cyclostationary NSE is useful to assess intra-annual variations in the budget reconstruction, it is not the best assessment tool for all the tropical basins with almost periodic TWSC. The regression slope between the reference and approximate TWSC can help in exhibiting consistent patterns in the water balance error.

4.2 Variables influencing the water budget closure

350 Several studies have limited their budget computation to large catchments only due to the general notion that the accuracy of budget closure increases with the size of the basin. We found that both small and large basins can achieve a high NSE (*cf.* Fig. 3). Furthermore, Fig. 5 proves that there is indeed no correlation between the maximum NSE and the basin area ($R^2 = 0.12$, $p = 0.12$). Although limiting their study to 10 large river basins worldwide, Sahoo et al. (2011) found no relationship between
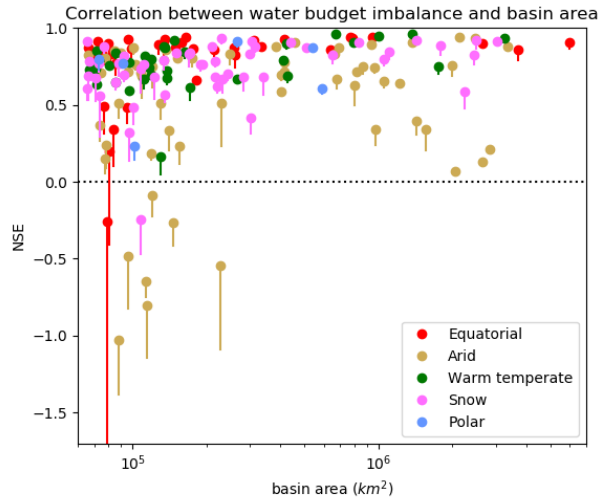
Figure 5. Each basin is represented by a bar between the maximum NSE (dot) and the 10th highest NSE

budget closure error and basin size. We extend this result and show that basins as small as 65,000 km^2 can close the water
355   budget. This result still holds if we evaluate the correlation between the basin area and the maximum cyclostationary NSE
($R^2 = 0.01$, $p = 0.90$).

     Figure 5 additionally indicates the consistency of our findings. Each basin was represented by a bar between the highest and
10th highest NSE values and the length of the bar was ~~lower~~ smaller than 0.15 in 90% of the basins. This means that several
combinations were able to close the water budget with similar imbalance errors.

360   Additionally, basins can be classified depending on their climate zone. Figure 6 shows the distribution of the maximum NSE
in each climate zone. Since the boxes (interquartile range) are of limited length (except for 'equatorial rain forest/monsoon'
and 'hot arid deserts'), this suggests that the imbalance error is rather consistent inside a given climate zone. In 'equatorial rain
forest/monsoon' climate zone, basins generally reached higher NSE values (map 3). However, this zone also contains small
Pacific islands (Papua New Guinea and Borneo) where runoff is much ~~more important~~ higher than evapotranspiration. Tables
365   S3 and S4 indicate that runoff was more uncertain (around 30% disagreements between datasets) than evapotranspiration
(around 18%) in those basins. Pacific islands with large runoff thus probably suffered from the poor runoff quality which led
to low NSE values.

     Hot arid deserts also have a large spread in the water budget imbalance (Fig. 6). Among those basins, some were entirely
desert (Arabian peninsula, Sahara, Somalia, South, and West Australia) with a low signal-to-noise ratio, as previously men-
370   tioned. Other basins were partially covered by steppe (Australia, Orange, around Indus) or equatorial savannah (Niger, Chad,
Nile). In those basins, precipitations occurred in the more humid subregions, thus increasing TWS variations. As a conse-
quence, the error in the datasets became less significant and allowed a proper budget reconstruction.
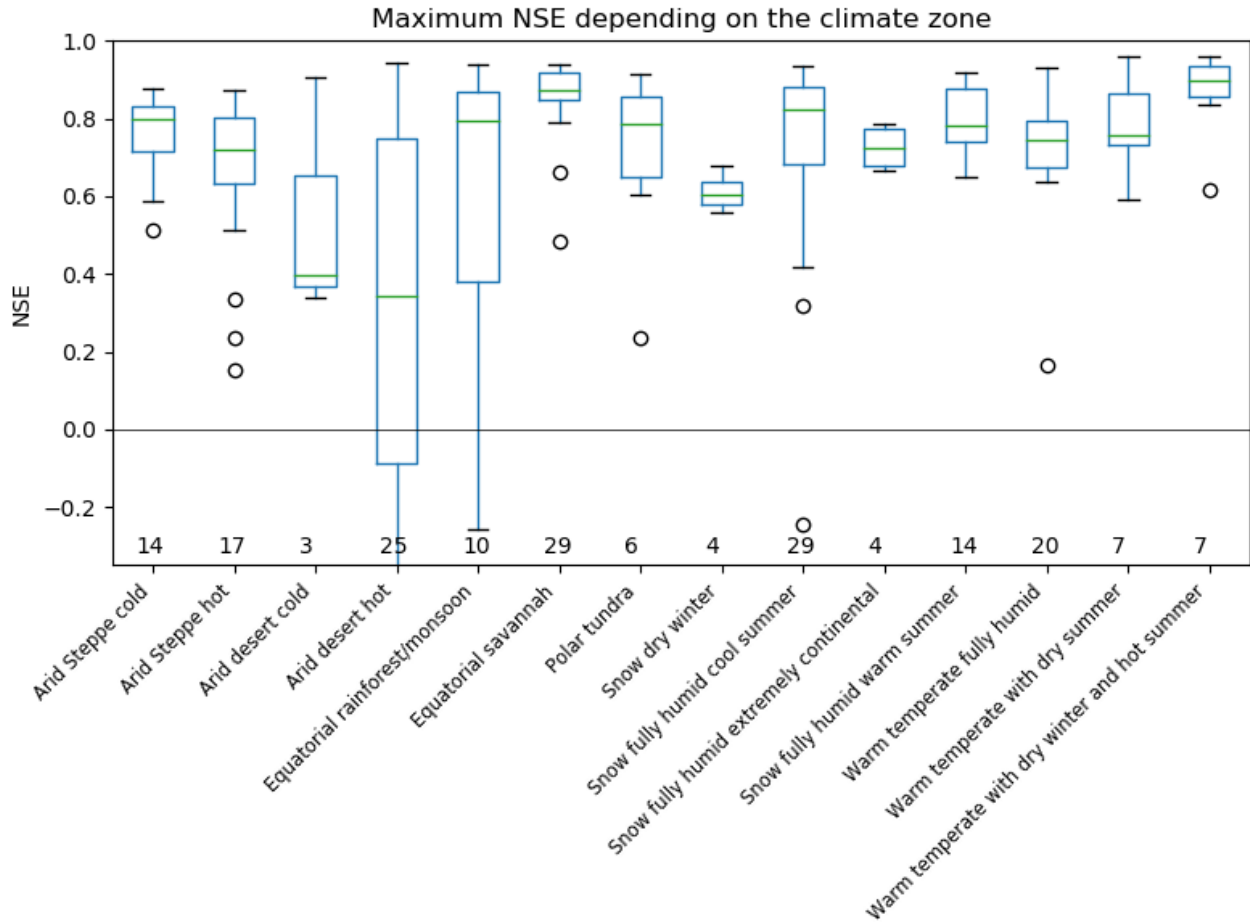
Figure 6. Boxplot of the maximum NSE per climate zone. The green line indicates the median, the box extends from the 1st quartile (Q_1) to the 3rd quartile (Q_3) while whiskers go from $Q_1 - 1.5(Q_3 - Q_1)$ (or the minimum value if higher) to $Q_3 + 1.5(Q_3 - Q_1)$ (or the maximum value if lower). Circles denote basins lying out of the whiskers. Figures represent the number of basins in each climate zone.

4.3 Overall combinations performances

Although a majority of basins achieved a positive cyclostationary NSE, they differed greatly in terms of the number of com-
375 binations yielding positive values. As an example, 839 combinations satisfied a positive NSE_c in the Sao Francisco basin
while only 94 did so in the neighbouring Tocantins basin (Fig. S9). Therefore, we wanted to evaluate the ability of a single
combination to close the water budget worldwide. To do so, we evaluated the total area of basins with a positive cyclostationary
NSE for each combination. Table 1 shows the 20 combinations leading to the largest area.

It appears that choosing all three variables (P, ET, and R) from ERA5 Land yields significantly better results than the other
380 combinations (35.5 million km^2 with a positive NSE_c from the total study area of 96.6 million km^2). Figure 7 indicates that

ERA5 Land performed well in the Central and Eastern United States of America (USA), but it failed to provide the positive NSE_c of Fig. 4 in the mountainous Western basins (Columbia, Great basin). Again comparing with the best possible results, ERA5 Land performed quite poorly in the equatorial region of South America (Amazon basin and above), in Central Eurasia (around the Ob, Aral sea, and Indus basins), as well as in several basins in Europe.

Knowing that there exists at least one combination giving a positive cyclostationary NSE in 62.3 million km^2, Table 1 shows that even the best combinations were far from approaching this number. This confirms that it is for now clearly impossible to achieve a good water budget closure with a single combination (Gao et al., 2010; Lorenz et al., 2014).

The second best combination in terms of area satisfying a positive cyclostationary NSE was the ~~Catchment Land Surface Model (CLSM)~~ CLSM forced with version 2.0 of GLDAS (in particular PGF precipitations). Table 1 shows that 30.8 million km^2 reached a positive NSE_c with this combination. Similar observations as ERA5 Land can be made generally, with good performances in Central and Eastern USA, South East America, and Australia. CLSM2.0 was more consistent than ERA5 Land in Europe but less so in Africa.

When looking at the following combinations, it appeared that their performances were more similar, compared to the differences observed between the two best combinations. Table 1 also shows that each variable has a determining impact on the water budget closure. Indeed, choosing for example CLSM2.2 for runoff instead of ERA5 Land (as shown in the left column of Fig. 7) led to poorer results in Alaska, Asia, and central Africa while it improved NSE values around the Amazon basin.

Concerning GLDAS LSMs, it is clear in Table 1 that CLSM was a globally better LSM than Noah and VIC. We also noted that when using all variables from the same LSM, GLDAS 2.0 was globally better than version 2.1 for all LSMs (CLSM, Noah, and VIC). As illustrated in the right column of Fig. 7, major differences are observed in Europe, Western Russia, and Alaska. This can be explained by disagreement between precipitations from GPCP and PGF. For instance, CLSM2.1 yielded only low NSE values in most of Eastern Europe whereas version 2.0 of the same model achieved a positive cyclostationary NSE. This last finding reflects the conclusion of e.g., Mueller et al. (2011) and Zaitchik et al. (2010) that forcing variables have a considerable influence on land surface models outputs.

We also point out that the ranking in Table 1 was not significantly modified by discriminating basins on the area satisfying a NSE larger than 0.5 (usually considered as good performances) instead of a positive cyclostationary NSE. This ensures the reliability of the method used to highlight the most consistent combinations.

4.4 Datasets suitable in given regions

In the previous section, numerous combinations of global datasets were evaluated. This section aims to describe regions where some datasets are more suitable than others to close the water budget. In a given basin, we defined as *suitable datasets* those appearing in combinations leading to a cost (difference between the maximum NSE and the NSE for a specific combination) lower than 0.1. This threshold was chosen to ensure that only the highest performing combinations were considered as suitable. ~~We study only~~ For this analysis, we focus on a subset of 132 basins out of ~~the~~ 189, ~~those with~~ where an excellent budget closure could be achieved (maximum NSE larger than 0.8 or maximum NSE_c larger than 0.1).
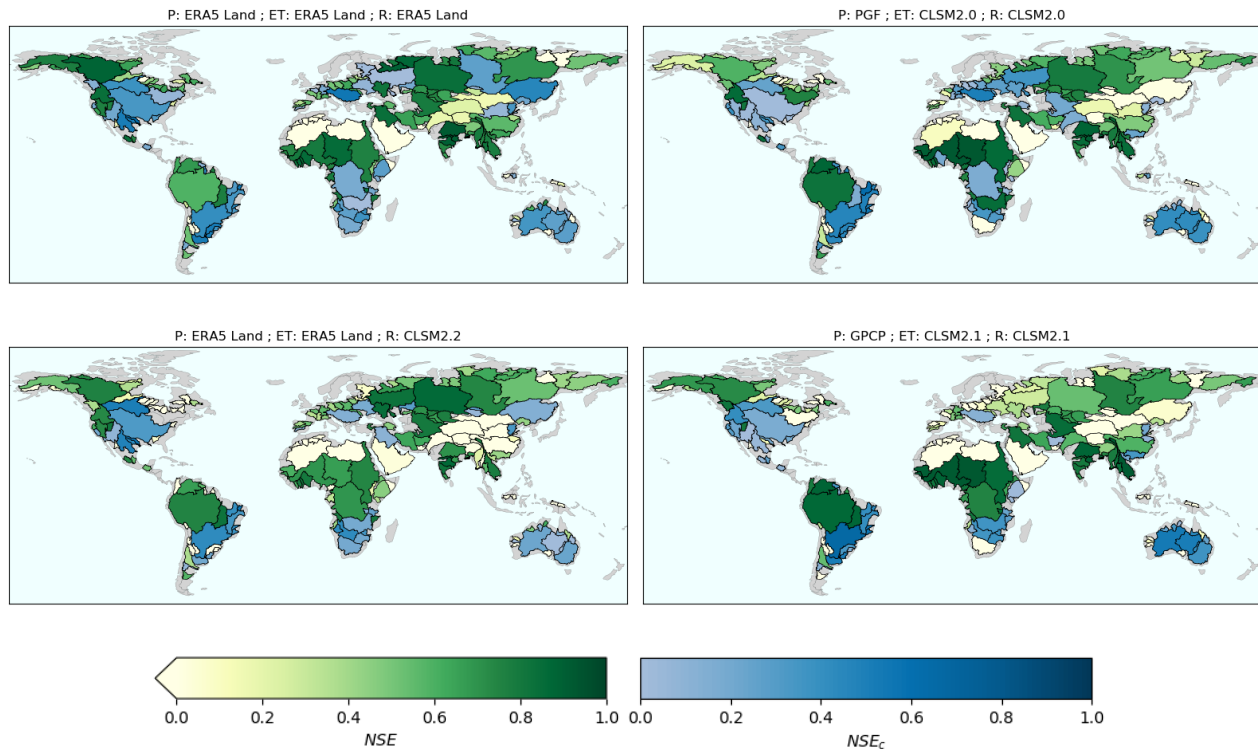
Figure 7. NSE and cyclostationary NSE with the first combinations in Table 1. Basins with a positive cyclostationary NSE are represented with blue shades corresponding to the NSE_c. Remaining basins are depicted in green, according to their NSE.

In general, many combinations were below the maximum cost: at least 112 combinations were suitable in 50% of the basins, at least 185 combinations in 25% of the basins. For a detailed review of suitable datasets in each basin, the reader is referred to Fig. S17, S18, S19, and S20. Although there was a large choice of combinations to close the water budget, two basins with similar characteristics only had a few suitable combinations in common. This makes a global and comprehensive evaluation of datasets more complex.

In addition, we observed that suitable datasets in a basin could generally not be mixed, suggesting that some cancellation bias occurred. As an example, Fig. 8 shows that suitable datasets in the Mississippi basin have considerably different seasonal cycles. Combining a precipitation dataset with high amplitude (GPCP) with low runoff (CLSM2.2) could close the water budget if associated with a high evapotranspiration (CLSM2.1, leading to $NSE_c = 0.32$) but not with a low evapotranspiration (Noah2.0, $NSE_c = -1.8$). Since there is no reason to consider a dataset as more reliable than others in the absence of unbiased observations, care must be taken when combining suitable datasets.

In order to provide a general overview of datasets ~~performances~~ performance, we choose to gather basins achieving the water budget closure for similar combinations. Those regions were determined with the hierarchical clustering described in section 3.3. The 132 selected basins with a good water budget closure are depicted in the dendrogram Fig. S10 and clusters represent
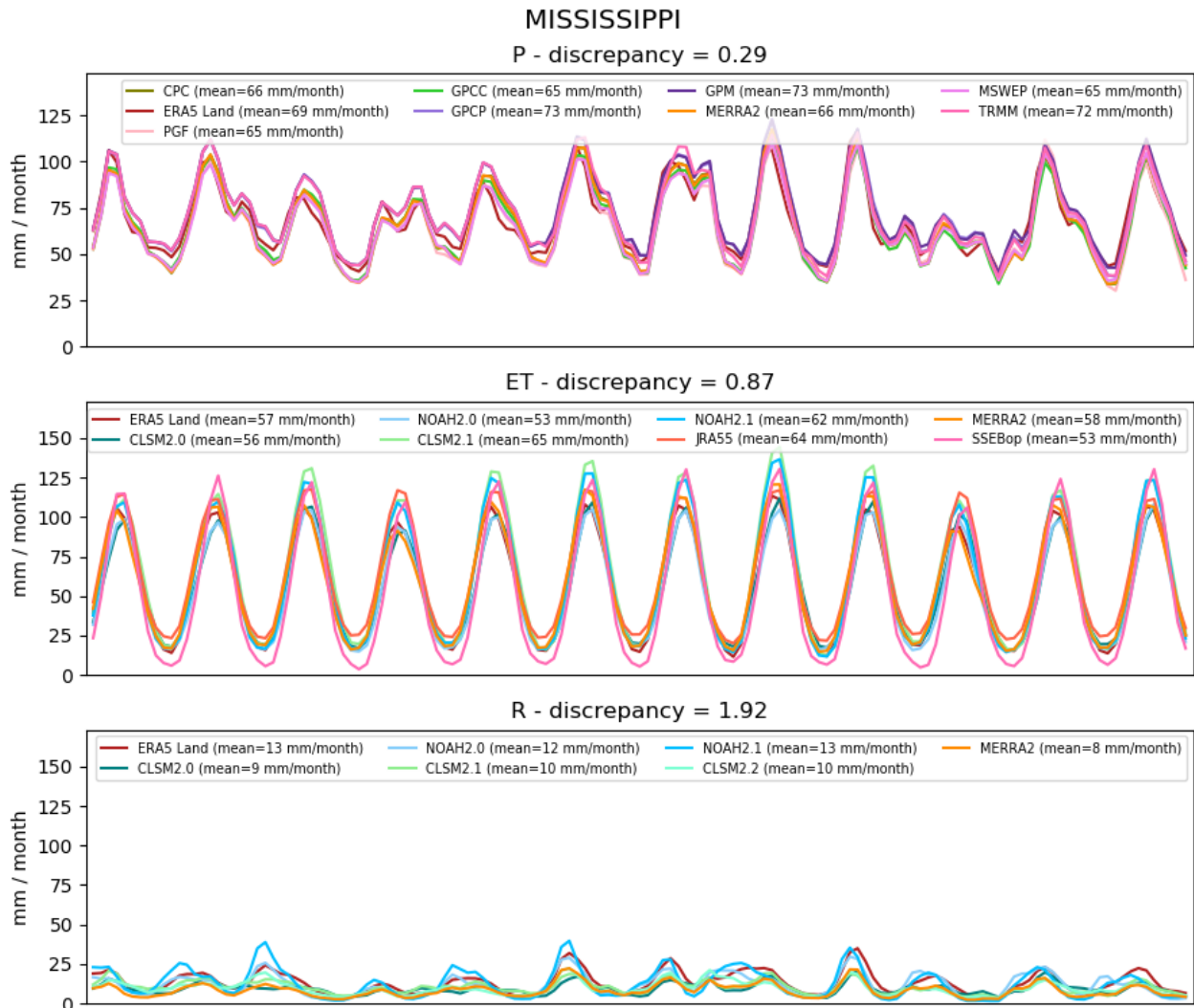
Figure 8. Datasets appearing in suitable combinations in the Mississippi basin (cost lower than 0.1). The discrepancy is similar to the coefficient of variation, except that the numerator is the difference between the maximum and minimum values instead of the standard deviation.

basins with similar costs for the same combinations. We chose 13 such clusters comprising major basins of the world to provide a precise but as succinct as possible overview of the datasets' performances. These clusters are denoted by the colored lines in Fig. S10 and are shown with the same basin colors on the map in Fig. 9.

Basins clustered together in the dendrogram Fig. S10 were either neighbouring basins (e.g., Eastern Europe or Eastern Australia) or basins with similar geographical conditions. It is therefore sensible that the same combinations performed well in
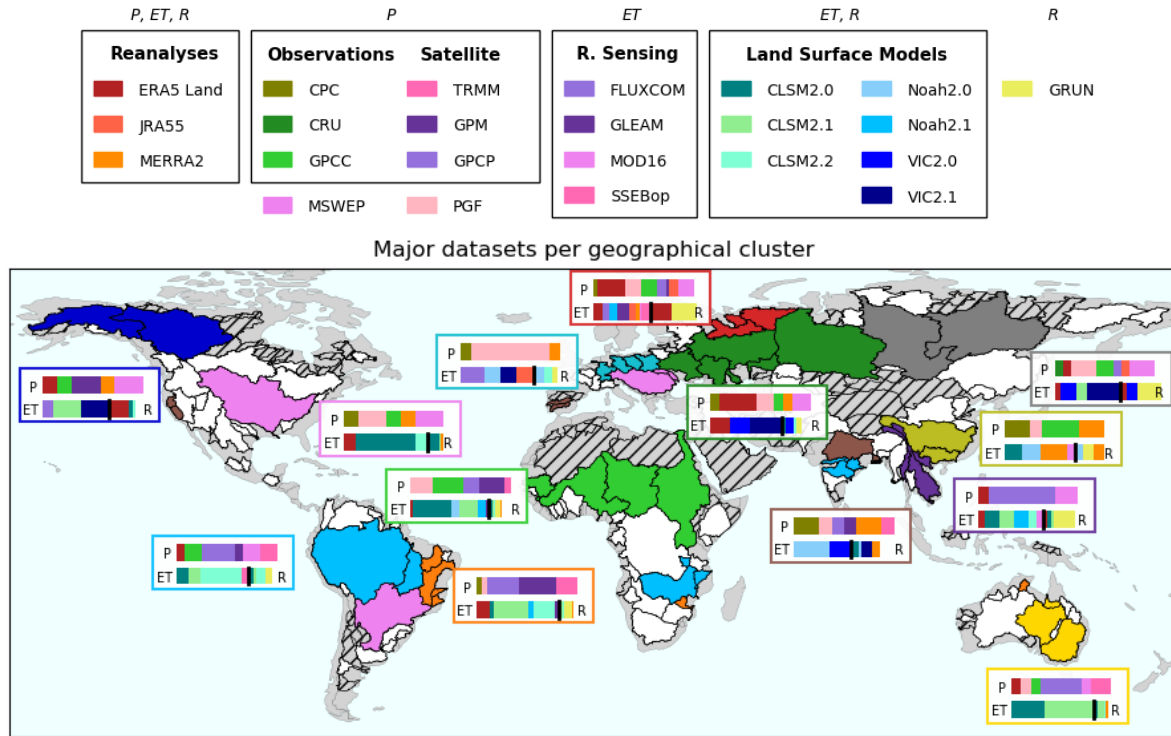
Figure 9. Datasets appearing in combinations that satisfy a cost lower than 0.1 for all basins inside the cluster. The 13 clusters highlighted in Fig. S10 are shown with different colors. For each cluster, the top line of each box represents precipitation datasets. The left part of the bottom line is evapotranspiration datasets while the right part is runoff. The limit between ET and R is symbolized by a black line located proportionally to the portion of ET in the mean annual water cycle of the corresponding region. Hatches show basins with a poor water budget closure (maximum NSE lower than 0.8 and maximum NSE_c lower than 0.1).

those basins. Among basins with similar characteristics, we pointed out large rivers in temperate regions (Mississippi, Parana, and Danube basins) or cold basins with different snow conditions (Yenisei, Lena, Mackenzie, Yukon, and Kolyma basins).

435   For each of the 13 clusters, we selected combinations yielding ~~to~~ a cost lower than 0.1 in every basin of the region. Figure 9 shows which datasets can be used in combination to satisfy the water balance. It first appears that among the precipitations datasets, the rain-gauge-based GPCC was often found in combinations satisfying the maximum cost, along with the satellite-augmented GPCP, reanalysis ERA5 Land, and the multi-source PGF. As a first approximation, those datasets are suitable for global water budget analyses. However, for regional analyses, a closer look at individual datasets is required to obtain all
440   possibilities.

Figure 10 (top left) shows the decay in NSE when using GPCC as the precipitation dataset. It confirms that GPCC was very close to the best-performing precipitations datasets. Surprisingly, Fig. 10 also indicates that although GPCP added satellite measurements to GPCC observations, it increased the water budget imbalance in Eastern Europe and western Russia, as well
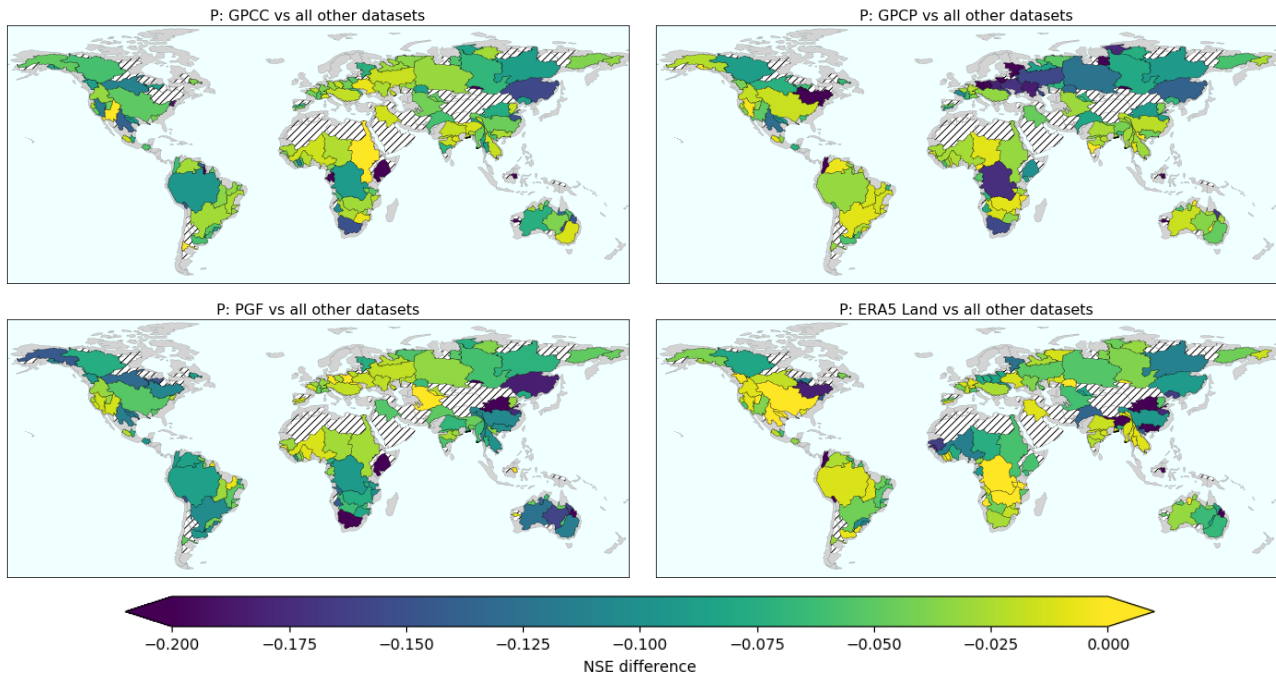
Figure 10. The mean of the 10th highest NSE with combinations comprising the reference dataset (*i.e.* GPCC, GPCP, PGF, or ERA5 Land) is compared to the mean of the 10th highest NSE excluding the reference dataset. Yellow indicates basins where the reference dataset is similar to or better than other precipitation datasets while blues show regions where it was significantly worse. Hatches show basins with a poor water budget closure (maximum NSE lower than 0.8 and maximum NSE_c lower than 0.1).

as in Congo and South Africa. GPCP performed notably well in South America, along with ERA5 Land that was one of the most consistent datasets for precipitation. The only region where ERA5 Land was not suitable was around China and Saint Lawrence basin. As shown in Fig. 9, PGF precipitation were able to close the water budget predominantly in Europe, as well as in Central Africa.

For comparison, Fig. S12 indicates that CRU which never appears in the map 9 performed very poorly compared to other datasets. Harris et al. (2020) mentioned that no homogenization of data was performed in CRU data. It also uses climatology values when measurements are missing, making it more appropriate for global analyses. The other rain-gauge-based dataset CPC was mainly suitable in Europe and China (see Fig. 9). Since MERRA2 is based on CPC observations (except in Africa where slight variations can be seen in Fig. S12), similar conclusions can be drawn for MERRA2. In addition, using GPM instead of TRMM (where we recall that GPM includes and extends TRMM results) improved the water budget closure. Finally, there was no overwhelming advantage in choosing the multi-source MSWEP dataset. It is consistent in Europe and South America but should be avoided in snow-dominated regions of Eastern Russia and Alaska (Fig. S12).

Fig. 9 clearly shows that evapotranspiration from the land surface model VIC should be chosen in Russian snow-dominated basins, with a preference for version 2.0 compared to 2.1. However, this dataset should not be used in hotter regions such as

South America, Africa, or Australia (Fig. S11). We found that VIC produces smaller evapotranspiration than other datasets, along with higher runoff. ~~The Catchment Land Surface Model~~ CLSM was also consistently found in Fig. 9. Version 2.0 and

460 2.1 performed similarly (except in Europe where version 2.0 was better as already mentioned) and were especially suitable in equatorial (South America, Subsaharan Africa, Australia) and some temperate regions (South-Eastern Europe and the USA). Similar to precipitations, ERA5 Land evapotranspiration is an excellent dataset in most of the regions except the Amazon basin, China, and Australia (Fig. S11).

~~Same as 10 but for evapotranspiration datasets.~~

465 ~~Evapotranspiration from~~ Evapotranspiration from CLSM version 2.2 ~~of the Catchment Land Surface Model~~ provided a good water budget closure in most of South America, Europe, and especially South Asia. However, it led to unrealistic low values in snow-dominated basins (see Fig. S11). An example of this behaviour is given in Fig. S15 where highly negative values appear in autumn. Since this dataset assimilates GRACE measurements and was validated against GRDC observations, this may reflect overfitting of runoff that is better constrained than evapotranspiration, therefore leading to unrealistic ET values.

470 When examining specific evapotranspiration datasets (FLUXCOM, GLEAM, MOD16, and SSEBop), it appeared that GLEAM led to almost optimal NSE values in Africa and Europe (Fig. S14). We also compared the newly released version 3.5 of GLEAM with the older v3.3 used in this study and found that the new version slightly improved the budget closure in every basin (not shown). FLUXCOM was also consistent in North and South America, Europe, western Russia, and South Asia, though it was outperformed by CLSM and ERA5 Land. Finally, SSEBop and MOD16 brought little improvement to

475 the water budget closure. The poor performances of MOD16 have already been highlighted by e.g., Pascolini-Campbell et al. (2020) in the CONUS, Bhattarai et al. (2019) in India.

The evaluation of runoff datasets in Fig. S13 confirms the differences exhibited for evapotranspiration (Fig. S11). VIC was mainly suitable in temperate and snow regions even if it performed quite poorly in some snow-dominated basins (Nelson, Saint Lawrence, Pechora, among others) due to overestimation of runoff during summer. It is also clear from Fig. S13 that this

480 LSM is not well-suited for equatorial and arid basins in South America (except some temperate basins in the extreme South), Africa, Australia, and part of Asia. In those basins, the machine-learning model GRUN was exceptionally good, especially outperforming others in South America. In addition, except in the Amazon basin and China where it has already been said that ERA5 Land was not appropriate, this reanalysis yielded a good runoff estimation.

~~Same as 10 but for runoff datasets.~~

485 The low NSE decays on Fig. S13 indicate that the CLSM version 2.2 ~~of the Catchment Land Surface Model~~ provide accurate runoff estimations, which is the main objective of this dataset (Li et al., 2019). However, Fig. S13 shows that it did not improve the water budget closure achieved by version 2.0 of this same model. In some basins like Congo, the water budget imbalance increased.

In a selection of 10 large basins with sufficient temporal coverage of GRDC gauge measurements (Amazon, Congo,

490 Mackenzie, Mississippi, Ob, Orange, Parana, Volga, Yenisei, and Yukon), we additionally evaluated the maximum NSE (and cyclostationary NSE) that could be obtained using GRDC records as the only source of runoff data. We found that the water budget closure slightly improved in 6 basins and significantly improved in 3 basins. The only basin where a slight decrease

could be observed was the Orange basin. This suggests that users interested in using discharge measurements should not see the water budget closure worsening compared to the datasets we used, but care needs to be taken in ensuring the discharge data are of sufficient quality and completeness for the basin of interest.

5 Conclusions

We assessed the ability of various precipitation, evapotranspiration, and runoff datasets to close the water balance equation against satellite observed terrestrial water storage anomalies on a global scale. Our analysis was comprehensive as a large number of global datasets were used to prepare 1694 combinations for closing the water balance in 189 catchments investigated. We found that the TWSC prediction was better than the long-term mean for 99% of the study area and better than the monthly mean in 62% of the study area. This illustrates that we can close the water balance equation in most of the regions if we choose certain datasets for budget components, which is a novel finding in terms of our previous understanding (Lorenz et al., 2014; Sahoo et al., 2011). We demarcated river catchments where the usual metrics (NSE, cyclostationary NSE) were of limited interest to evaluate the imbalance error.

Although the lowest imbalance error possible was generally small, we found that none of the 1694 combinations assessed succeeded in closing the water budget worldwide. Some combinations performed better in some regions but underperformed in others. The combination with all the budget components from reanalysis ERA5 Land was the best in terms of achieving a positive cyclostationary NSE over the largest fraction of the area under investigation. Individual components (P, ET, and R) of ERA5 Land were also close to the best performing datasets, except for around the Amazon basin and Eastern China.

The Catchment Land Surface Model additionally appeared as a suitable dataset in many regions excluding snow-dominated basins. However, version 2.2 of this LSM, which assimilates GRACE data, performed poorly compared to its previous versions. In some snow-dominated basins, it even led to highly unrealistic ET values during the cold season. Despite being designed for better runoff estimates, this latest version did not bring much improvement to other runoff datasets in terms of the water imbalance error. In contrast, GRUN, a machine learning runoff dataset, considerably reduced the imbalance error in several basins, with the best performances being detected in South America, South Asia, and some Arctic basins in Russia and Alaska.

We have presented a comprehensive overview of our ability to close the global water balance with the help of a wide range of water budget components disseminated for scientific studies. For each water budget component, we also assessed the performance of individual datasets with respect to the other datasets available, which helped us to infer the quality of the dataset when closing the water budget. We also found that the water balance can close due to a cancellation of errors in budget components, therefore, caution should be practiced when closing the water budget over a catchment or region and a large number of datasets should be explored to avoid obtaining the right results for wrong reasons. We hope that our analysis will help fellow researchers in finding the most appropriate datasets for water budget analysis in different parts of the world.

Code availability. Our code is made available at https://github.com/lehmannfa/water_budget_closure.

Table 1. Combinations with the largest area covered with a positive cyclostationary NSE

	total area with $NSE_c > 0$ (in million km^2)	total area with $NSE > 0$ (in million km^2)
P: ERA5 Land ; ET: ERA5 Land ; R: ERA5 Land	35.5	89.7
P: PGF ; ET: CLSM2.0 ; R: CLSM2.0	30.8	90.2
P: ERA5 Land ; ET: ERA5 Land ; R: CLSM2.2	24.5	79.7
P: PGF ; ET: NOAH2.0 ; R: CLSM2.0	23.9	90.9
P: GPCP ; ET: CLSM2.1 ; R: CLSM2.1	23.4	79.2
P: ERA5 Land ; ET: ERA5 Land ; R: GRUN	22.7	81.3
P: MSWEP ; ET: CLSM2.0 ; R: CLSM2.0	21.8	78.5
P: ERA5 Land ; ET: ERA5 Land ; R: CLSM2.0	21.7	78.6
P: ERA5 Land ; ET: ERA5 Land ; R: MERRA2	21.7	76.6
P: GPM ; ET: CLSM2.1 ; R: CLSM2.1	21.1	80.1
P: GPCP ; ET: CLSM2.1 ; R: CLSM2.0	20.8	78.4
P: GPCC ; ET: CLSM2.0 ; R: CLSM2.0	20.4	79.4
P: ERA5 Land ; ET: ERA5 Land ; R: NOAH2.0	19.8	84.4
P: GPM ; ET: CLSM2.1 ; R: CLSM2.0	19.0	79.4
P: MERRA2 ; ET: MERRA2 ; R: MERRA2	18.8	92.1
P: GPM ; ET: NOAH2.1 ; R: NOAH2.0	18.8	81.0
P: GPM ; ET: CLSM2.1 ; R: CLSM2.2	18.7	71.2
P: GPCP ; ET: CLSM2.1 ; R: CLSM2.2	18.5	74.6
P: TRMM ; ET: CLSM2.1 ; R: CLSM2.1	18.5	56.7
P: PGF ; ET: NOAH2.0 ; R: CLSM2.2	18.4	86.3
...	...	...
P: PGF ; ET: VIC2.0 ; R: VIC2.0	16.1	87.6
...	...	...
P: PGF ; ET: NOAH2.0 ; R: NOAH2.0	16.0	92.4
...	...	...
P: GPCP ; ET: NOAH2.1 ; R: NOAH2.1	13.3	82.6
...	...	...
P: ERA5 Land ; ET: CLSM2.2 ; R: CLSM2.2	10.8	57.8
...	...	...
P: JRA55 ; ET: JRA55 ; R: JRA55	8.7	72.2
...	...	...
P: GPCP ; ET: VIC2.1 ; R: VIC2.1	7.1	75.6

Combinations are ranked by decreasing area of basins with a positive cyclostationary NSE. Italics indicate combinations where P, ET, and R are from the same model.

5.1 ~~Additional figures~~

525 ~~Distribution of the maximum NSE over all combinations in terms of basin area~~

~~Components of the water budget in the Amazon basin for the combination leading to the highest NSE~~

~~Correlation between monthly values of GRACE TWSC and the budget reconstruction in the Amazon basin, with the combination leading to the highest NSE (NSE=0.92 and cyclostationary NSE=-1.28)~~

~~Components of the water budget in the Niger basin for the combination leading to the highest NSE~~

530 ~~Correlation between monthly values of GRACE TWSC and the budget reconstruction in the Niger basin, with the combination leading to the highest NSE (NSE=0.94 and cyclostationary NSE=-0.62)~~

~~Number of combinations yielding a positive cyclostationary NSE in each basin. Grey means that no combination achieved a positive value.~~

~~132 basins with a maximum NSE larger than 0.8 or a maximum NSE_c larger than 0.1. The distance between basins is the~~
535 ~~Euclidean distance between the vector of costs for each combination. The height of the U-shaped link is proportional to this distance. Basins are clustered to minimize the intra-cluster variance and colored basins are those selected to plot Fig. 9~~

~~The mean of the 10th highest NSE with combinations comprising the reference dataset is compared to the mean of the 10th highest NSE excluding the reference dataset. Yellow indicates basins where the reference dataset is similar to or better than others while blues show regions where it was significantly worse. Hatches show basins with a poor water budget closure~~
540 ~~(maximum NSE lower than 0.8 and maximum NSE_c lower than 0.1).~~

~~The mean of the 10th highest NSE with combinations comprising the reference dataset is compared to the mean of the 10th highest NSE excluding the reference dataset. Yellow indicates basins where the reference dataset is similar to or better than others while blues show regions where it was significantly worse. Hatches show basins with a poor water budget closure (maximum NSE lower than 0.8 and maximum NSE_c lower than 0.1).~~

545 ~~Components of the water budget in the Mackenzie basin with all components from GLDAS2.2 CLSM (assimilating GRACE TWS)~~

~~Legend of Fig. S17, S18, S19, S20~~

~~Datasets appearing in combinations that satisfy a cost lower than 0.1 for each basin separately. The top line of represents precipitations datasets. The left part of the bottom line is evapotranspiration datasets while the right part is runoff. The limit~~
550 ~~between ET and R is symbolized by a black line located proportionally to the portion of ET in the mean annual water cycle of the corresponding region, explaining while the bottom line may have a length different than 100%. Basins are ordered according to hierarchical clustering (dendrogram in Fig. S10). The color legend for datasets can be found in Fig. S16~~

~~Following of Fig. S17~~

~~Following of Fig. S18~~
555 ~~Following of Fig. S19~~

5.1 ~~Additional tables~~

Components of the mean annual water cycle in Pacific islands P (mm/year) ET (mm/year) R (mm/year) SEPIK 3390 ± 653 1404 ± 223 2116 ± 597 MAMBERAMO 3578 ± 851 1340 ± 227 2406 ± 756 MAHAKAM 3163 ± 356 1359 ± 272 1911 ± 529 KAPUAS 3666 ± 204 1366 ± 266 2339 ± 480

560   Components of the mean annual water cycle in equatorial rain forest/monsoon basins in South America P (mm/year) ET (mm/year) R (mm/year) MAGDALENA 2339 ± 650 1157 ± 216 1373 ± 498 CUYUNI 2051 ± 269 1395 ± 223 766 ± 327 ESSEQUIBO 2121 ± 251 1314 ± 217 946 ± 405 MARONI 2312 ± 247 1406 ± 269 885 ± 364 AMAZON 2177 ± 172 1251 ± 196 958 ± 251 ORINOCO 2269 ± 289 1237 ± 200 1090 ± 315

Author contributions. J. B. and B. D. V. designed the experiment. F.L. implemented the code and wrote the paper with support from all
565 coauthors. All the authors contributed to the synthesis of results and key conclusions.

Competing interests. JLB and FL were supported by the European Research Council (ERC) under the European Union's Horizon 2020
research and innovation programme under grant agreement No 694188 (GlobalMass) and BDV is supported by the Marie Skłodowska-Curie
Individual Fellowship (MSCA-IF) under grant agreement no 841407 (CLOSeR).

Acknowledgements. We are grateful to Miss Megan Rounsley, who carefully proofread the manuscript. Concerning the datasets used, they
570 are all publicly available and the links to download them can be found in our Github repository. CPC Global Unified Precipitation data
provided by the NOAA/OAR/ESRL PSL, Boulder, Colorado, USA, from their Web site at https://www.psl.noaa.gov/data/gridded/data.cpc.
globalprecip.html. GPCP data provided by the NOAA/OAR/ESRL PSL, Boulder, Colorado, USA, from their Web site at https://psl.noaa.gov/
data/gridded/data.gpcp.html. CSR mascons were downloaded from http://www2.csr.utexas.edu/grace and GRACE/GRACE-FO JPL Mascon
data are available at http://grace.jpl.nasa.gov.

References

Adler, R. F., Sapiano, M. R. P., Huffman, G. J., Wang, J.-J., Gu, G., Bolvin, D., Chiu, L., Schneider, U., Becker, A., Nelkin, E., Xie, P., Ferraro, R., and Shin, D.-B.: The Global Precipitation Climatology Project (GPCP) Monthly Analysis (New Version 2.3) and a Review of 2017 Global Precipitation, Atmosphere, 9, https://doi.org/10.3390/atmos9040138, 2018.

Armanios, D. E. and Fisher, J. B.: Measuring water availability with limited ground data: assessing the feasibility of an entirely remote-sensing-based hydrologic budget of the Rufiji Basin, Tanzania, using TRMM, GRACE, MODIS, SRB, and AIRS, Hydrological Processes, 28, 853–867, https://doi.org/10.1002/hyp.9611, publisher: John Wiley & Sons, Ltd, 2014.

Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., van Dijk, A. I. J. M., McVicar, T. R., and Adler, R. F.: MSWEP V2 Global 3-Hourly 0.1° Precipitation: Methodology and Quantitative Assessment, Bulletin of the American Meteorological Society, 100, 473–500, https://doi.org/10.1175/BAMS-D-17-0138.1, 2019.

Bhattarai, N., Mallick, K., Stuart, J., Vishwakarma, B. D., Niraula, R., Sen, S., and Jain, M.: An automated multi-model evapotranspiration mapping framework using remotely sensed and reanalysis data, Remote Sensing of Environment, 229, 69–92, https://doi.org/10.1016/j.rse.2019.04.026, 2019.

Blöschl, G., Bierkens, M. F., Chambel, A., Cudennec, C., Destouni, G., Fiori, A., Kirchner, J. W., McDonnell, J. J., Savenije, H. H., Sivapalan, M., Stumpp, C., Toth, E., Volpi, E., Carr, G., Lupton, C., Salinas, J., Széles, B., Viglione, A., Aksoy, H., Allen, S. T., Amin, A., Andréassian, V., Arheimer, B., Aryal, S. K., Baker, V., Bardsley, E., Barendrecht, M. H., Bartosova, A., Batelaan, O., Berghuijs, W. R., Beven, K., Blume, T., Bogaard, T., Borges de Amorim, P., Böttcher, M. E., Boulet, G., Breinl, K., Brilly, M., Brocca, L., Buytaert, W., Castellarin, A., Castelletti, A., Chen, X., Chen, Y., Chen, Y., Chifflard, P., Claps, P., Clark, M. P., Collins, A. L., Croke, B., Dathe, A., David, P. C., de Barros, F. P. J., de Rooij, G., Di Baldassarre, G., Driscoll, J. M., Duethmann, D., Dwivedi, R., Eris, E., Farmer, W. H., Feiccabrino, J., Ferguson, G., Ferrari, E., Ferraris, S., Fersch, B., Finger, D., Foglia, L., Fowler, K., Gartsman, B., Gascoin, S., Gaume, E., Gelfan, A., Geris, J., Gharari, S., Gleeson, T., Glendell, M., Gonzalez Bevacqua, A., González-Dugo, M. P., Grimaldi, S., Gupta, A. B., Guse, B., Han, D., Hannah, D., Harpold, A., Haun, S., Heal, K., Helfricht, K., Herrnegger, M., Hipsey, M., Hlaváčiková, H., Hohmann, C., Holko, L., Hopkinson, C., Hrachowitz, M., Illangasekare, T. H., Inam, A., Innocente, C., Istanbulluoglu, E., Jarihani, B., Kalantari, Z., Kalvans, A., Khanal, S., Khatami, S., Kiesel, J., Kirkby, M., Knoben, W., Kochanek, K., Kohnová, S., Kolechkina, A., Krause, S., Kreamer, D., Kreibich, H., Kunstmann, H., Lange, H., Liberato, M. L. R., Lindquist, E., Link, T., Liu, J., Loucks, D. P., Luce, C., Mahé, G., Makarieva, O., Malard, J., Mashtayeva, S., Maskey, S., Mas-Pla, J., Mavrova-Guirguinova, M., Mazzoleni, M., Mernild, S., Misstear, B. D., Montanari, A., Müller-Thomy, H., Nabizadeh, A., Nardi, F., Neale, C., Nesterova, N., Nurtaev, B., Odongo, V. O., Panda, S., Pande, S., Pang, Z., Papacharalampous, G., Perrin, C., Pfister, L., Pimentel, R., Polo, M. J., Post, D., Prieto Sierra, C., Ramos, M.-H., Renner, M., Reynolds, J. E., Ridolfi, E., Rigon, R., Riva, M., Robertson, D. E., Rosso, R., Roy, T., Sá, J. H., Salvadori, G., Sandells, M., Schaefli, B., Schumann, A., Scolobig, A., Seibert, J., Servat, E., Shafiei, M., Sharma, A., Sidibe, M., Sidle, R. C., Skaugen, T., Smith, H., Spiessl, S. M., Stein, L., Steinsland, I., Strasser, U., Su, B., Szolgay, J., Tarboton, D., Tauro, F., Thirel, G., Tian, F., Tong, R., Tussupova, K., Tyralis, H., Uijlenhoet, R., van Beek, R., van der Ent, R. J., van der Ploeg, M., Van Loon, A. F., van Meerveld, I., van Nooijen, R., van Oel, P. R., Vidal, J.-P., von Freyberg, J., Vorogushyn, S., Wachniew, P., Wade, A. J., Ward, P., Westerberg, I. K., White, C., Wood, E. F., Woods, R., Xu, Z., Yilmaz, K. K., and Zhang, Y.: Twenty-three unsolved problems in hydrology (UPH) – a community perspective, Hydrological Sciences Journal, 64, 1141–1158, https://doi.org/10.1080/02626667.2019.1620507, 2019.

610 Chen, F., Mitchell, K., Schaake, J., Xue, Y., Pan, H.-L., Koren, V., Duan, Q. Y., Ek, M., and Betts, A.: Modeling of land surface evap-
      oration by four schemes and comparison with FIFE observations, Journal of Geophysical Research: Atmospheres, 101, 7251–7268,
      https://doi.org/10.1029/95JD02165, 1996.

Chen, J., Tapley, B., Rodell, M., Seo, K., Wilson, C., Scanlon, B. R., and Pokhrel, Y.: Basin-Scale River Runoff Estimation From GRACE
      Gravity Satellites, Climate Models, and In Situ Observations: A Case Study in the Amazon Basin, Water Resources Research, p. 21,
615 https://doi.org/10.1029/2020WR028032, 2020.

Chen, M. and Xie, P.: CPC Unified Gauge-based Analysis of Global Daily Precipitation, Cairns, Australia, 2008.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer,
      P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haim-
      berger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz,
620 B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., and Vitart, F.: The ERA-Interim reanalysis:
      configuration and performance of the data assimilation system, Quarterly Journal of the Royal Meteorological Society, 137, 553–597,
      https://doi.org/10.1002/qj.828, 2011.

Ek, M. B., Mitchell, K. E., Lin, Y., Rogers, E., Grunmann, P., Koren, V., Gayno, G., and Tarpley, J. D.: Implementation of Noah land surface
      model advances in the National Centers for Environmental Prediction operational mesoscale Eta model, Journal of Geophysical Research:
625 Atmospheres, 108, https://doi.org/10.1029/2002JD003296, 2003.

Fisher, J. B., Melton, F., Middleton, E., Hain, C., Anderson, M., Allen, R., McCabe, M. F., Hook, S., Baldocchi, D., Townsend, P. A., Kilic, A.,
      Tu, K., Miralles, D. D., Perret, J., Lagouarde, J.-P., Waliser, D., Purdy, A. J., French, A., Schimel, D., Famiglietti, J. S., Stephens, G., and
      Wood, E. F.: The future of evapotranspiration: Global requirements for ecosystem functioning, carbon and climate feedbacks, agricultural
      management, and water resources, Water Resources Research, 53, 2618–2626, https://doi.org/10.1002/2016WR020175, 2017.

630 Gao, H., Tang, Q., Ferguson, C. R., Wood, E. F., and Lettenmaier, D. P.: Estimating the water budget of major US river basins via remote
      sensing, International Journal of Remote Sensing, 31, 3955–3978, https://doi.org/10.1080/01431161.2010.483488, 2010.

Ghiggi, G., Humphrey, V., Seneviratne, S. I., and Gudmundsson, L.: GRUN: an observation-based global gridded runoff dataset from 1902
      to 2014, Earth System Science Data, 11, 1655–1674, https://doi.org/10.5194/essd-11-1655-2019, 2019.

GRDC: Major River Basins of the World - Global Runoff Data Centre, https://www.bafg.de/GRDC/EN/02_srvcs/22_gslrs/221_MRB/
635 riverbasins_node.html, 2020.

Harris, I., Osborn, T. J., Jones, P., and Lister, D.: Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset,
      Scientific Data, 7, 109, https://doi.org/10.1038/s41597-020-0453-3, 2020.

Henriksen, H. J., Troldborg, L., Nyegaard, P., Sonnenborg, T. O., Refsgaard, J. C., and Madsen, B.: Methodology for construction, cali-
      bration and validation of a national hydrological model for Denmark, Journal of Hydrology, 280, 52–71, https://doi.org/10.1016/S0022-
640 1694(03)00186-0, 2003.

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

Huffman, G. J., Adler, R. F., Bolvin, D. T., and Nelkin, E. J.: The TRMM Multi-Satellite Precipitation Analysis (TMPA), in: Gebremichael
645 M., Hossain F. (eds) Satellite rainfall Applications for Surface Hydrology, pp. 3–22, Springer, Dordrecht, https://doi.org/10.1007/
      978-90-481-2915-7_1, 2010.

Huffman, G. J., Bolvin, D. T., Braithwaite, D., Hsu, K., Joyce, R., Kidd, C., Nelkin, E. J., Sorooshian, S., Tan, J., and Xie, P.: NASA Global Precipitation Measurement (GPM) Integrated Multi-Satellite Retrievals for GPM (IMERG), p. 38, https://doi.org/10.5067/GPM/IMERG/3B-MONTH/06, 2019.

650 Jain, S. K. and Sudheer, K. P.: Fitting of Hydrologic Models: A Close Look at the Nash–Sutcliffe Index, Journal of Hydrologic Engineering, 13, 981–986, https://doi.org/10.1061/(ASCE)1084-0699(2008)13:10(981), 2008.

Jung, M., Reichstein, M., and Bondeau, A.: Towards global empirical upscaling of FLUXNET eddy covariance observations: validation of a model tree ensemble approach using a biosphere model, Biogeosciences, 6, 2001–2013, https://doi.org/10.5194/bg-6-2001-2009, 2009.

Jung, M., Koirala, S., Weber, U., Ichii, K., Gans, F., Camps-Valls, G., Papale, D., Schwalm, C., Tramontana, G., and Reichstein, M.: The 655 FLUXCOM ensemble of global land-atmosphere energy fluxes, Scientific Data, 6, 74, https://doi.org/10.1038/s41597-019-0076-8, 2019.

Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., Onogi, K., Kamahori, H., Kobayashi, C., Endo, H., Miyaoka, K., and Takahashi, K.: The JRA-55 Reanalysis: General Specifications and Basic Characteristics, Journal of the Meteorological Society of Japan. Ser. II, 93, 5–48, https://doi.org/10.2151/jmsj.2015-001, 2015.

Koren, V., Schaake, J., Mitchell, K., Duan, Q.-Y., Chen, F., and Baker, J. M.: A parameterization of snowpack and frozen 660 ground intended for NCEP weather and climate models, Journal of Geophysical Research: Atmospheres, 104, 19 569–19 585, https://doi.org/10.1029/1999JD900232, 1999.

Koster, R. D., Suarez, M. J., Ducharne, A., Stieglitz, M., and Kumar, P.: A catchment-based approach to modeling land surface processes in a general circulation model: 1. Model structure, Journal of Geophysical Research: Atmospheres, 105, 24 809–24 822, https://doi.org/10.1029/2000JD900327, 2000.

665 Kottek, M., Grieser, J., Beck, C., Rudolf, B., and Rubel, F.: World Map of the Köppen-Geiger climate classification updated, Meteorologische Zeitschrift, 15, 259–263, https://doi.org/10.1127/0941-2948/2006/0130, 2006.

Landerer, F. W., Dickey, J. O., and Güntner, A.: Terrestrial water budget of the Eurasian pan-Arctic from GRACE satellite measurements during 2003–2009, Journal of Geophysical Research, 115, D23 115, https://doi.org/10.1029/2010JD014584, 2010.

Li, B., Rodell, M., Kumar, S., Beaudoing, H. K., Getirana, A., Zaitchik, B. F., Goncalves, L. G., Cossetin, C., Bhanja, S., Mukherjee, A., 670 Tian, S., Tangdamrongsub, N., Long, D., Nanteza, J., Lee, J., Policelli, F., Goni, I. B., Daira, D., Bila, M., Lannoy, G., Mocko, D., Steele-Dunne, S. C., Save, H., and Bettadpur, S.: Global GRACE Data Assimilation for Groundwater and Drought Monitoring: Advances and Challenges, Water Resources Research, 55, 7564–7586, https://doi.org/10.1029/2018WR024618, 2019.

Liang, X., Lettenmaier, D. P., Wood, E. F., and Burges, S. J.: A simple hydrologically based model of land surface water and energy fluxes for general circulation models, Journal of Geophysical Research, 99, 14 415, https://doi.org/10.1029/94JD00483, 1994.

675 Liu, W., Wang, L., Zhou, J., Li, Y., Sun, F., Fu, G., Li, X., and Sang, Y.-F.: A worldwide evaluation of basin-scale evapotranspiration estimates against the water balance method, Journal of Hydrology, 538, 82–95, https://doi.org/10.1016/j.jhydrol.2016.04.006, 2016.

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

Long, D., Shen, Y., Sun, A., Hong, Y., Longuevergne, L., Yang, Y., Li, B., and Chen, L.: Drought and flood monitoring 680 for a large karst plateau in Southwest China using extended GRACE data, Remote Sensing of Environment, 155, 145–160, https://doi.org/10.1016/j.rse.2014.08.006, 2014b.

Long, D., Yang, Y., Wada, Y., Hong, Y., Liang, W., Chen, Y., Yong, B., Hou, A., Wei, J., and Chen, L.: Deriving scaling factors using a global hydrological model to restore GRACE total water storage changes for China's Yangtze River Basin, Remote Sensing of Environment, 168, 177–193, https://doi.org/10.1016/j.rse.2015.07.003, 2015.

685   Longuevergne, L., Scanlon, B. R., and Wilson, C. R.: GRACE Hydrological estimates for small basins: Evaluating processing approaches on the High Plains Aquifer, USA, Water Resources Research, 46, https://doi.org/https://doi.org/10.1029/2009WR008564, 2010.

Lorenz, C., Kunstmann, H., Devaraju, B., Tourian, M. J., Sneeuw, N., and Riegger, J.: Large-Scale Runoff from Landmasses: A Global Assessment of the Closure of the Hydrological and Atmospheric Water Balances, Journal of Hydrometeorology, 15, 2111–2139, https://doi.org/10.1175/JHM-D-13-0157.1, 2014.

690   Lorenz, C., Tourian, M. J., Devaraju, B., Sneeuw, N., and Kunstmann, H.: Basin-scale runoff prediction: An Ensemble Kalman filter framework based on global hydrometeorological data sets, Water Resources Research, 51, 8450–8475, https://doi.org/10.1002/2014WR016794, 2015.

Lv, M., Ma, Z., Yuan, X., Lv, M., Li, M., and Zheng, Z.: Water budget closure based on GRACE measurements and reconstructed evapotranspiration using GLDAS and water use data for two large densely-populated mid-latitude basins, Journal of Hydrology, 547, 585–599,
695   https://doi.org/10.1016/j.jhydrol.2017.02.027, 2017.

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

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

Monteith, J. L.: Evaporation and Environment, Symposia of the Society for Experimental Biology, pp. 205–234, https://repository.rothamsted.ac.uk/item/8v5v7/evaporation-and-environment, 1965.

Mu, Q., Zhao, M., and Running, S. W.: Improvements to a MODIS global terrestrial evapotranspiration algorithm, Remote Sensing of
705   Environment, 115, 1781–1800, https://doi.org/10.1016/j.rse.2011.02.019, 2011.

Mueller, B., Seneviratne, S. I., Jimenez, C., Corti, T., Hirschi, M., Balsamo, G., Ciais, P., Dirmeyer, P., Fisher, J. B., Guo, Z., Jung, M., Maignan, F., McCabe, M. F., Reichle, R., Reichstein, M., Rodell, M., Sheffield, J., Teuling, A. J., Wang, K., Wood, E. F., and Zhang, Y.: Evaluation of global observations-based evapotranspiration datasets and IPCC AR4 simulations: global land evapotranspiration datasets, Geophysical Research Letters, 38, https://doi.org/10.1029/2010GL046230, 2011.

710   Muñoz-Sabater, J.: ERA5-Land monthly averaged data from 2001 to present, ECMWF, https://doi.org/10.24381/CDS.68D2BB30, type: dataset, 2019.

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

Oki, T. and Kanae, S.: Global Hydrological Cycles and World Water Resources, Science, 313, 1068–1072,
715   https://doi.org/10.1126/science.1128845, 2006.

Oliveira, P. T. S., Nearing, M. A., Moran, M. S., Goodrich, D. C., Wendland, E., and Gupta, H. V.: Trends in water balance components across the Brazilian Cerrado, Water Resources Research, 50, 7100–7114, https://doi.org/10.1002/2013WR015202, 2014.

Pan, M., Sahoo, A. K., Troy, T. J., Vinukollu, R. K., Sheffield, J., and Wood, E. F.: Multisource Estimation of Long-Term Terrestrial Water Budget for Major Global River Basins, Journal of Climate, 25, 3191–3206, https://doi.org/10.1175/JCLI-D-11-00300.1, 2012.

720   Pascolini-Campbell, M. A., Reager, J. T., and Fisher, J. B.: GRACE-based Mass Conservation as a Validation Target for Basin-Scale Evapotranspiration in the Contiguous United States, Water Resources Research, 56, e2019WR026 594, https://doi.org/10.1029/2019WR026594, publisher: John Wiley & Sons, Ltd, 2020.

Penatti, N. C., Almeida, T. I. R. d., Ferreira, L. G., Arantes, A. E., and Coe, M. T.: Satellite-based hydrological dynamics of the world's largest continuous wetland, Remote Sensing of Environment, 170, 1–13, https://doi.org/10.1016/j.rse.2015.08.031, 2015.

725 Penman, H. L.: Natural evaporation from open water, bare soil and grass, Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences, 193, 120–145, https://doi.org/10.1098/rspa.1948.0037, 1948.

Reichle, R. H., Koster, R. D., De Lannoy, G. J. M., Forman, B. A., Liu, Q., Mahanama, S. P. P., and Touré, A.: Assessment and Enhancement of MERRA Land Surface Hydrology Estimates, Journal of Climate, 24, 6322–6338, https://doi.org/10.1175/JCLI-D-10-05033.1, 2011.

Reichle, R. H., Liu, Q., Koster, R. D., Draper, C. S., Mahanama, S. P. P., and Partyka, G. S.: Land Surface Precipitation in MERRA-2, Journal

730 of Climate, 30, 1643–1664, https://doi.org/10.1175/JCLI-D-16-0570.1, 2017.

Rodell, M. and Famiglietti, J. S.: Detectability of variations in continental water storage from satellite observations of the time dependent gravity field, Water Resources Research, 35, 2705–2723, https://doi.org/10.1029/1999WR900141, 1999.

Rodell, M., Houser, P. R., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C.-J., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., Entin, J. K., Walker, J. P., Lohmann, D., and Toll, D.: The Global Land Data Assimilation System, Bulletin of the American Meteorological

735 Society, 85, 381–394, https://doi.org/10.1175/BAMS-85-3-381, 2004.

Rodell, M., Beaudoing, H. K., L'Ecuyer, T. S., Olson, W. S., Famiglietti, J. S., Houser, P. R., Adler, R., Bosilovich, M. G., Clayson, C. A., Chambers, D., Clark, E., Fetzer, E. J., Gao, X., Gu, G., Hilburn, K., Huffman, G. J., Lettenmaier, D. P., Liu, W. T., Robertson, F. R., Schlosser, C. A., Sheffield, J., and Wood, E. F.: The Observed State of the Water Cycle in the Early Twenty-First Century, Journal of Climate, 28, 8289–8318, https://doi.org/10.1175/JCLI-D-14-00555.1, 2015.

740 Saemian, P., Elmi, O., Vishwakarma, B., Tourian, M., and Sneeuw, N.: Analyzing the Lake Urmia restoration progress using ground-based and spaceborne observations, Science of The Total Environment, 739, 139 857, https://doi.org/10.1016/j.scitotenv.2020.139857, 2020.

Sahoo, A. K., Pan, M., Troy, T. J., Vinukollu, R. K., Sheffield, J., and Wood, E. F.: Reconciling the global terrestrial water budget using satellite remote sensing, Remote Sensing of Environment, 115, 1850–1865, https://doi.org/10.1016/j.rse.2011.03.009, 2011.

Samuelsen, A., Hansen, C., and Wehde, H.: Tuning and assessment of the HYCOM-NORWECOM V2.1 biogeochemical modeling system

745 for the North Atlantic and Arctic oceans, Geoscientific Model Development, 8, 2187–2202, https://doi.org/10.5194/gmd-8-2187-2015, 2015.

Scanlon, B. R., Zhang, Z., Save, H., Sun, A. Y., Müller Schmied, H., van Beek, L. P. H., Wiese, D. N., Wada, Y., Long, D., Reedy, R. C., Longuevergne, L., Döll, P., and Bierkens, M. F. P.: Global models underestimate large decadal declining and rising water storage trends relative to GRACE satellite data, Proceedings of the National Academy of Sciences, 115, E1080–E1089,

750 https://doi.org/10.1073/pnas.1704665115, 2018.

Schneider, U., Becker, A., Finger, P., Rustemeier, E., and Ziese, M.: GPCC Full Data Monthly Version 2020 at 0.5°, https://doi.org/10.5676/DWD_GPCC/FD_M_V2020_050, 2020.

Senay, G. B., Bohms, S., Singh, R. K., Gowda, P. H., Velpuri, N. M., Alemu, H., and Verdin, J. P.: Operational Evapotranspiration Mapping Using Remote Sensing and Weather Datasets: A New Parameterization for the SSEB Approach, JAWRA Journal of the American Water

755 Resources Association, 49, 577–591, https://doi.org/10.1111/jawr.12057, 2013.

Sheffield, J., Goteti, G., and Wood, E. F.: Development of a 50-Year High-Resolution Global Dataset of Meteorological Forcings for Land Surface Modeling, Journal of Climate, 19, 3088–3111, https://doi.org/10.1175/JCLI3790.1, place: Boston MA, USA Publisher: American Meteorological Society, 2006.

Sheffield, J., Ferguson, C. R., Troy, T. J., Wood, E. F., and McCabe, M. F.: Closing the terrestrial water budget from satellite remote sensing,

760 Geophysical Research Letters, 36, https://doi.org/10.1029/2009GL037338, publisher: John Wiley & Sons, Ltd, 2009.

Sneeuw, N., Lorenz, C., Devaraju, B., Tourian, M. J., Riegger, J., Kunstmann, H., and Bárdossy, A.: Estimating Runoff Using Hydro-Geodetic Approaches, Surveys in Geophysics, 35, 1333–1359, https://doi.org/10.1007/s10712-014-9300-4, 2014.

Swann, A. L. S. and Koven, C. D.: A Direct Estimate of the Seasonal Cycle of Evapotranspiration over the Amazon Basin, Journal of Hydrometeorology, 18, 2173–2185, https://doi.org/10.1175/JHM-D-17-0004.1, place: Boston MA, USA Publisher: American Meteorological Society, 2017.

Tapley, B. D.: GRACE Measurements of Mass Variability in the Earth System, Science, 305, 503–505, https://doi.org/10.1126/science.1099192, 2004.

Thor, R.: Least-Squares prediction of runoff, 2013.

Tourian, M., Schwatke, C., and Sneeuw, N.: River discharge estimation at daily resolution from satellite altimetry over an entire river basin, Journal of Hydrology, 546, 230–247, https://doi.org/10.1016/j.jhydrol.2017.01.009, 2017.

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

Wahr, J., Molenaar, M., and Bryan, F.: Time variability of the Earth's gravity field: Hydrological and oceanic effects and their possible detection using GRACE, Journal of Geophysical Research: Solid Earth, 103, 30 205–30 229, https://doi.org/10.1029/98JB02844, 1998.

Wahr, J., Swenson, S., and Velicogna, I.: Accuracy of GRACE mass estimates, Geophysical Research Letters, 33, L06 401, https://doi.org/10.1029/2005GL025305, 2006.

Wan, Z., Zhang, K., Xue, X., Hong, Z., Hong, Y., and Gourley, J. J.: Water balance-based actual evapotranspiration reconstruction from ground and satellite observations over the conterminous United States: water balance-based observational ET reconstruction, Water Resources Research, 51, 6485–6499, https://doi.org/10.1002/2015WR017311, 2015.

Wang, H., Guan, H., Gutiérrez-Jurado, H. A., and Simmons, C. T.: Examination of water budget using satellite products over Australia, Journal of Hydrology, 511, 546–554, https://doi.org/10.1016/j.jhydrol.2014.01.076, 2014.

Watkins, M. M., Wiese, D. N., Yuan, D.-N., Boening, C., and Landerer, F. W.: Improved methods for observing Earth's time variable mass distribution with GRACE using spherical cap mascons: Improved Gravity Observations from GRACE, Journal of Geophysical Research: Solid Earth, 120, 2648–2671, https://doi.org/10.1002/2014JB011547, 2015.

Wiese, D. N., Yuan, D.-N., Boening, C., Landerer, F. W., and Watkins, M. M.: JPL GRACE Mascon Ocean, Ice, and Hydrology Equivalent Water Height Release 06 Coastal Resolution Improvement (CRI) Filtered Version 1.0, https://doi.org/10.5067/TEMSC-3MJC6, type: dataset, 2018.

Xie, J., Xu, Y., Gao, C., Xuan, W., and Bai, Z.: Total Basin Discharge From GRACE and Water Balance Method for the Yarlung Tsangpo River Basin, Southwestern China, Journal of Geophysical Research: Atmospheres, 124, 7617–7632, https://doi.org/10.1029/2018JD030025, 2019.

Zaitchik, B. F., Rodell, M., and Olivera, F.: Evaluation of the Global Land Data Assimilation System using global river discharge data and a source-to-sink routing scheme: source to sink routing for global models, Water Resources Research, 46, https://doi.org/10.1029/2009WR007811, 2010.

Zhang, J.: Assessing the statistical relations of terrestrial water mass changewith hydrological variables and climate variability, Ph.D. thesis, Universität Stuttgart, München, https://publikationen.badw.de/de/046188119/pdf/CC%20BY, 2019.

Zhang, Y., Pan, M., Sheffield, J., Siemann, A. L., Fisher, C. K., Liang, M., Beck, H. E., Wanders, N., MacCracken, R. F., Houser, P. R., Zhou, T., Lettenmaier, D. P., Pinker, R. T., Byteway, J., Kummerow, C. D., and Wood, E. F.: A Climate Data Record (CDR) for the

global terrestrial water budget: 1984–2010, Hydrology and Earth System Sciences, 22, 241–263, https://doi.org/10.5194/hess-22-241-2018, 2018.