Master's by Research in

Global Environmental Challenges

How can the water budget help us understand terrestrial water storage variations?

Fanny Lehmann

University of BRISTOL
Cabot Institute for the Environment

How can the water budget help us understand terrestrial water storage variations?

By

FANNY LEHMANN

supervised by Pr. J. L. Bamber and Dr. B. D. Vishwakarma

School of Geographical Sciences
UNIVERSITY OF BRISTOL

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of MASTER OF SCIENCE BY RESEARCH in the Faculty of Science.

SEPTEMBER 2021

Word count: 15991

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, first we attempt to close the global water budget using precipitation, evapotranspiration, and runoff data at the catchment scale. We use a large number of state-of-the-art datasets to estimate water storage changes. These estimates are then 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. Secondly, we developed an optimization framework using the water balance equation to reconstruct TWS time-series from 1981 to the start of the GRACE mission. The results were then compared with state-of-the-art TWS reconstructions.

We found that TWS changes obtained from the water balance equation were more accurate than the long-term mean of GRACE time-series in 99 % of the basins examined, and better than the monthly mean in 62% of the basins. By analyzing different combinations of the datasets that make up the water balance, we identified data products that performed well in certain regions. Although our TWS reconstruction was comparable with existing ones, we observed large uncertainties between reconstructions. This shows that a better understanding of underlying trends in datasets is needed, especially for the last decades, to improve our understanding of the spatiotemporal variations in TWS.

Dedication and acknowledgements

I would like to thank my supervisors Pr. J. Bamber and Dr. B. D. Vishwakarma for their dedicated support throughout the year, despite the context of distance learning. I am also really grateful to having been funded by the GlobalMass project for this work. I thank the GlobalMass team members for their questions and remarks on my research. I would like to extend my sincere thank to the examiners Pr. P. Bates and Pr. N. Sneeuw for their careful reading of my dissertation and the invaluable discussion that followed. Finally, I would like to thank G. Coxon for her supportive messages all along the year.

AUTHOR'S DECLARATION

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. The work is the candidate's own work with the assistance of the supervisors. As mentioned in the text, Chapter 2 has been submitted for publication and improved thanks to the referees's comments. Any views expressed in the dissertation are those of the author.

SIGNED: FANNY LEHMANN DATE: SEPTEMBER, 13, 2021

TABLE OF CONTENTS

List of Tables

List of Figures

Figure **Page**

CHAPTER 1

INTRODUCTION

"The most noticeable consequences of climate change will be impacts on the water cycle" says Rodell et al. (2015). Indeed, the water cycle describes the movement of water in space and time through the Earth system. Water reservoirs are of various size and availability. Therefore, quantifying usable amounts of water is essential to maintain human activities on our blue planet. Groundwater is a highly crucial fraction of water since it provides drinking and irrigation water in many regions of the world. However, large uncertainties are associated with our knowledge of the levels of groundwater and their evolution with time. This is why many studies have been conducted on the observed trends in available water and how it may be influenced by climate change.

1.1 The global water cycle

Water is present in three physical forms - ice, liquid, and vapour - on Earth and they are all crucial to ensure livable conditions. The oceans represent around 96.5 % of all water; they allow the circulation of heat and nutrients between continents. The remaining fraction of water is composed of saline groundwater (1 %) and fresh water (2.5 %). However, around two thirds of global fresh water are stored in ice sheets and glaciers that are not available for human consumption but highly contribute to weather and temperature regulation. Although the atmosphere holds a tiny fraction of fresh water, it is crucial in regulating surface temperature and precipitation. The remaining third of fresh water is distributed on Earth as surface water, soil moisture, and ground water; and is the main focus of this work.

As emphasized in Fig. 1.1, understanding the fluxes of water is perhaps even more important than quantifying the size of the water stores. Stephens et al. (2020) classifies those fluxes into three main categories. Firstly, precipitation and evapotranspiration represent the exchange

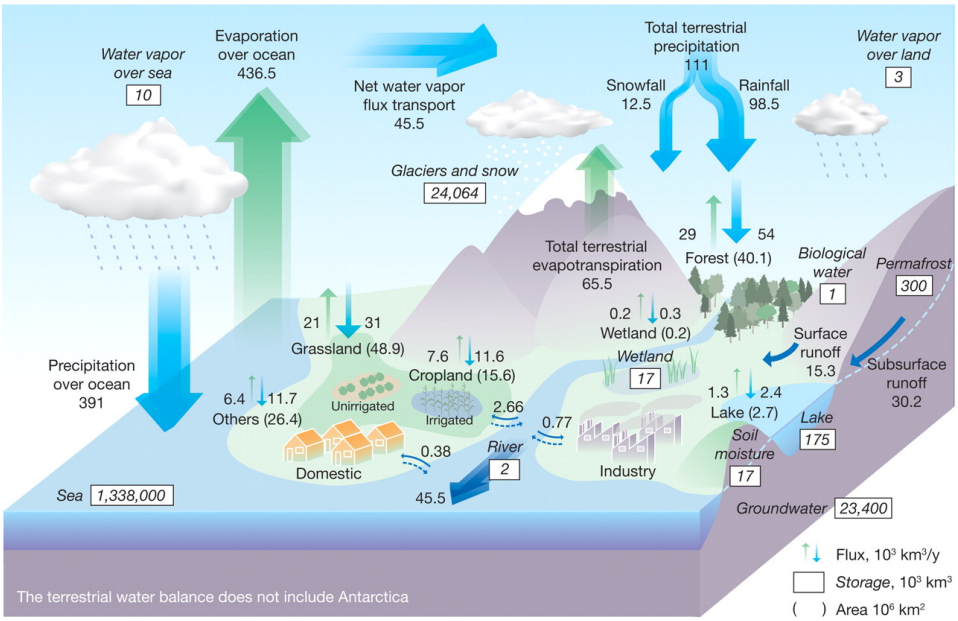


Figure 1.1: The main reservoirs of water on Earth with associated fluxes between them. Figure reproduced from Oki and Kanae (2006)

between the atmosphere and the Earth surface. Although precipitation is quite-well monitored in many regions, our observation systems are not yet able to provide consistent estimates of global precipitation and assess how climate change is affecting their intensity and frequency. Evapotranspiration is a much more complex phenomenon that brings water back to the atmosphere and is therefore difficult to quantify, as will be further explained in Chapter 2. The second category of fluxes is ice melt contributing to sea level rise and is not developed further in this work. Finally, large amounts of water are transferred from land to the ocean by means of runoff. Discharge measurements from gauge stations are available for hundreds of rivers across the globe but the availability and the quality of data is poor. Furthermore, global runoff may be underestimated due to the neglect of submarine groundwater flows (Stephens et al., 2020).

The water cycle can be described with the terrestrial water balance equation

$$\frac{dTWS}{dt} = P - ET - R. \tag{1.1}$$

This equation is the main focus of Chapter 2. It expresses changes in the terrestrial water storage (TWS) as the difference between water gained by precipitation (P), water returning back to the atmosphere through evapotranspiration (ET) and water flowing out of the catchment in the form of runoff (R).

1.2 Terrestrial Water Storage

TWS is defined as the sum of water stored in the form of snow, in canopy, as soil moisture, as groundwater, and as surface water (Scanlon et al., 2018). The relative contribution of each component varies depending on the climatic conditions of the region under study. For example, surface water (lakes, rivers, inundated foodplains, wetlands) dominates in the wet tropics like the Amazon basin, in arid basins with large rivers such as the Nile, as well as in subpolar regions, while soil moisture is especially critical in the tropics and mid-latitude regions. Globally, groundwater is a crucial component since it provides water to more than 2 billion of people and is largely used for irrigation (Famiglietti, 2014). Groundwater levels can be monitored via the depth of wells but this is notoriously difficult (Jasechko and Perrone, 2021).

The Gravity Recovery And Climate Experiment (GRACE) was launched jointly by NASA and the German Aerospace Center to overcome the lack of knowledge on invisible water. This satellite mission operated from March 2002 to October 2017 and was continued from May 2018 by the GRACE Follow-on mission. GRACE consists of two satellites orbiting at around 450 km above the Earth and measuring the distance between them. Since the velocity depends on the Earth gravitational field, it is a way to deduce gravity variations from the position of the two satellites. Those variations allow the detection of changes due to currents in the ocean, TWS, ice loss on ice sheets and glaciers, as well as variations within the Earth mantle. Therefore, GRACE led to unique assessments of the water cycle and especially evaluation of TWS. We nevertheless recall that inverting the 3D gravitational problem does not have a unique solution. Especially, the relatively high altitude of GRACE satellites prevents the detection of signals with a wavelength smaller than a few hundred kilometres. This is why the processing centres of GRACE measurements apply different techniques and advanced smoothing methods to compute estimates of TWS. Thus, there are limitations to the spatial resolution one should expect from GRACE measurements.

1.3 Investigating impacts of climate change

Groundwater evolves on inter-annual to decadal time scales and is therefore essential to observe in the context of climate change. Human activities cause groundwater depletion as is the case in the Ganges basin to sustain irrigation (Stephens et al., 2020). Dam management and the impoundment of water in artificial lakes also impact runoff of water towards the oceans. In addition, droughts are associated with TWS decrease in the Euphrates basin and in California and can be detected by GRACE (Rodell et al., 2018). Extracting linear trends from GRACE measurements allows to point out regions with severe declining trends that are likely caused by impacts of human activities. Out of the 34 large aquifers under study, Rodell et al. (2018) found 12 aquifers with decreasing TWS trends possibly or probably caused by climate change.

However, detecting such trends is not straight-forward as natural variability is also involved

in TWS variations. GRACE measurements are not yet long enough (less than 20 years with a 1-year gap between 2017 and 2018) to allow a proper distinction between natural variability and anthropogenic forcings. As shown by Vishwakarma et al. (2021), trends largely depend on the period used for the computation. Further, hydrological models do not help assessing linear trends since they lead to various and sometimes opposite results compared to GRACE (Scanlon et al., 2018). Therefore, several studies tried to extend TWS time-series prior the start of the GRACE mission. By enhancing our understanding of TWS variations over the last decades, we will be able to better predict the evolution of TWS under different climate change scenarios. Chapter 3 is dedicated to reconstructing TWS for this objective.

CHAPTER 2

ASSESSING THE WATER BUDGET CLOSURE AT THE GLOBAL SCALE

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. This 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. This chapter firstly review the use of the water balance equation to model hydrological components and assess the accuracy of datasets. This equation is then used to evaluate the imbalance error from a large number of datasets. This helps describing regions suffering from large uncertainties in the water budget components and others where the water balance equation could be used to achieve more ambitious goals.

A substantial part of Chapter 2 has been submitted as an article entitled "How well are we able to close the water budget on a global scale?" to Hydrology and Earth Systems Sciences and is currently in discussion for publication. Supporting code and data can be publicly accessed at `https://github.com/lehmannfa/water_budget_closure`.

2.1 Literature review on water budget closure

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 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.

2.1.1 Use of the water balance equation to estimate individual components

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 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 1.1.

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).

Apart from ET, our 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 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 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).

2.1.2 Challenges to improve the water budget closure

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 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) or highly impacted by human activities (e.g. the Yellow river basin, 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 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 ~ 200,000 km^2 established by 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 ~ 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 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 this chapter is thus to provide a revised overview of the water budget closure on a global scale. Section 2.2 presents the study area covering all parts of the globe (excluding Greenland and Antarctica) and the datasets. Then, section 2.3 details the metrics used to evaluate the water budget closure as well as the selection process for the best combinations. Moreover, section 2.4 explains the results and discusses previous studies.

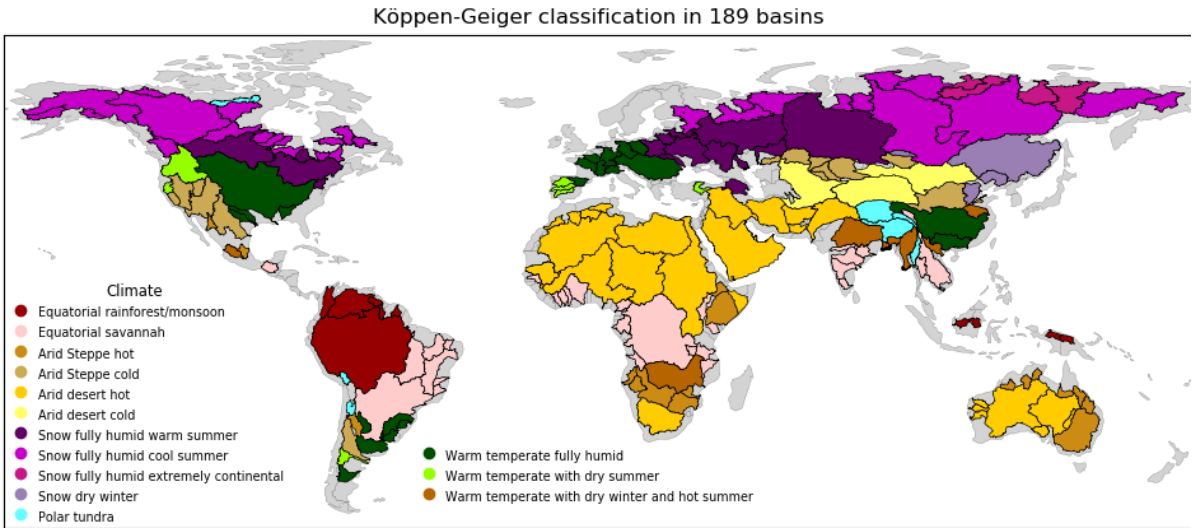


Figure 2.1: 189 basins larger than 63,000 km^2 with their corresponding climate zone

2.2 Data

2.2.1 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. 2.1 and their areas range from ~ 65,600 km^2 to ~ 5,965,900 km^2.

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

2.3 Methods

2.3.1 Water budget reconstruction

GRACE mascon fields were used to compute time-series of TWS anomalies relative to the mean between 2004 and 2009. Since equation 1.1 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 derivative was computed with centered finite difference (as in e.g., Long et al. (2014a) or

Pascolini-Campbell et al. (2020))

$$TWSC(t) = \frac{TWS(t+1) - TWS(t-1)}{2\Delta t}, \tag{2.1}$$

where Δt equals 1 month and $t-1$, t, $t+1$ are three consecutive months. Missing monthly values were filled with cubic 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 2.2 (Landerer et al., 2010)

$$\tilde{X}(t) = \frac{1}{4}X(t-1) + \frac{1}{2}X(t) + \frac{1}{4}X(t+1), \tag{2.2}$$

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.

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 1.1: $TWSC_{budget}(t) = P_{data_P}(t) - ET_{data_{ET}}(t) - R_{data_R}(t)$. This reconstruction was compared with the derivatives obtained from equation 2.1 and denoted $TWSC_{GRACE}(t)$. Since we used 11 precipitation, 14 evapotranspiration, and 11 runoff datasets, we finally evaluated 1694 combinations.

2.3.2 Metrics

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

$$RMSD = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(TWSC_{budget}(t) - TWSC_{GRACE}(t))^2}, \tag{2.3}$$

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

$$NSE = 1 - \frac{\frac{1}{T}\sum_{t=1}^{T}(TWSC_{budget}(t) - TWSC_{GRACE}(t))^2}{\frac{1}{T}\sum_{t=1}^{T}(TWSC_{GRACE}(t) - \overline{TWSC_{GRACE}})^2} = 1 - \frac{RMSD^2}{err_{cst}^2}, \tag{2.4}$$

where $\overline{TWSC_{GRACE}} = \frac{1}{T}\sum_{t=1}^{T} TWSC_{GRACE}(t)$ is the long-term mean of TWSC and err_{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 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 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) is then:

$$NSE_c = 1 - \frac{\frac{1}{T}\sum_{t=1}^{T}(TWSC_{budget}(t) - TWSC_{GRACE}(t))^2}{\frac{1}{T}\sum_{t=1}^{T}(TWSC_{GRACE}(t) - TWSC^m_{GRACE})^2} = 1 - \frac{RMSD^2}{err^2_{cyc}}, \tag{2.5}$$

where $TWSC^m_{GRACE}$ is the mean value for month m over all years and err_{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 2.4 and 2.5

$$NSE_c = \left(1 - \frac{err^2_{cst}}{err^2_{cyc}}\right) + \underbrace{\frac{err^2_{cst}}{err^2_{cyc}}}_{\gamma} NSE. \tag{2.6}$$

The γ factor describes the behaviour of the TWSC by comparison with the mean seasonal cycle. Basins with periodic seasonal cycles (i.e., low err_{cyc}) or large magnitudes (i.e., high err_{cst}) have larger γ. In those basins (e.g., the Amazon or Chad basins in Fig. **??**), extremely high NSE values are required to achieve a positive cyclostationary NSE, as can be seen in Fig. 2.2. Special attention must then be given when examining such basins to discriminate performances depending on the NSE or the cyclostationary NSE.

2.3.3 Selection of the most representative datasets

When estimating a water cycle component from the water balance equation 1.1, 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. 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

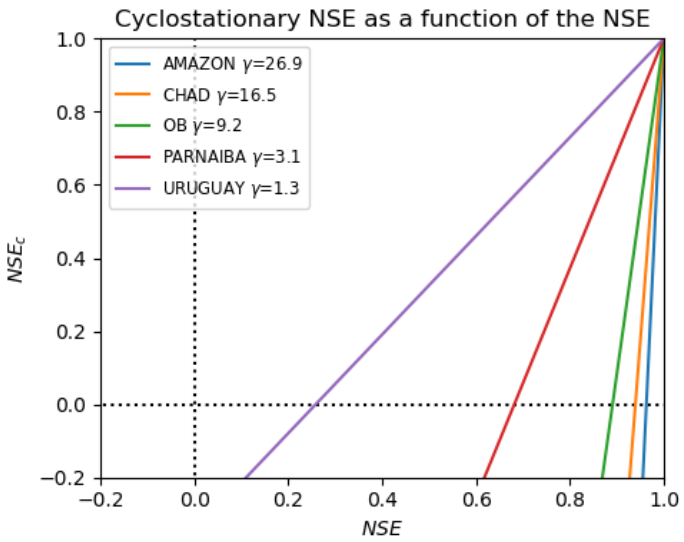


Figure 2.2: The cyclostationary NSE is related to the NSE through $NSE_c = 1 - \gamma + \gamma NSE$ where $\gamma = \frac{err^2_{cst}}{err^2_{cyc}}$

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 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 2.2 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

$$c_i^b = \max_{comb} NSE^b(comb) - NSE^b(combination_i), \tag{2.7}$$

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 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 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 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.

2.4 Results and discussion

2.4.1 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 2.3 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 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. 2.3). 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.

Figure 2.3 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 B.1 then shows the distribution of the maximum NSE. Although it has been explained in Section 2.3.2 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 large number of datasets, it is likely that

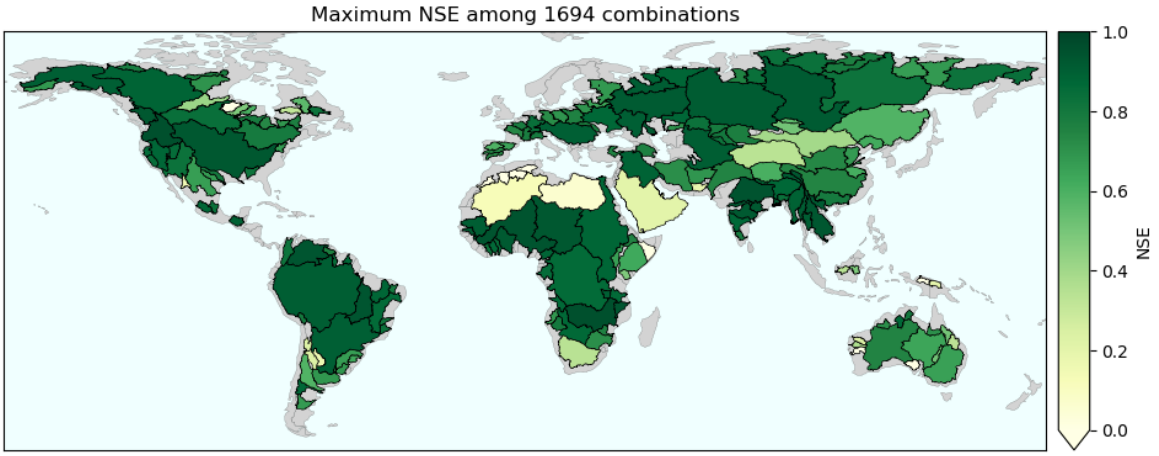


Figure 2.3: Maximum NSE per basin over all combinations. Green positive values mean that the budget reconstruction is a better approximation of GRACE TWSC than the long-term mean.

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.

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 was also used to assess the quality of reconstructed TWSC. Figure 2.4 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. 2.4).

When comparing Fig. 2.3 and Fig. 2.4, 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 reconstruction consistently underestimated the magnitude of TWSC (Fig. B.2). 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

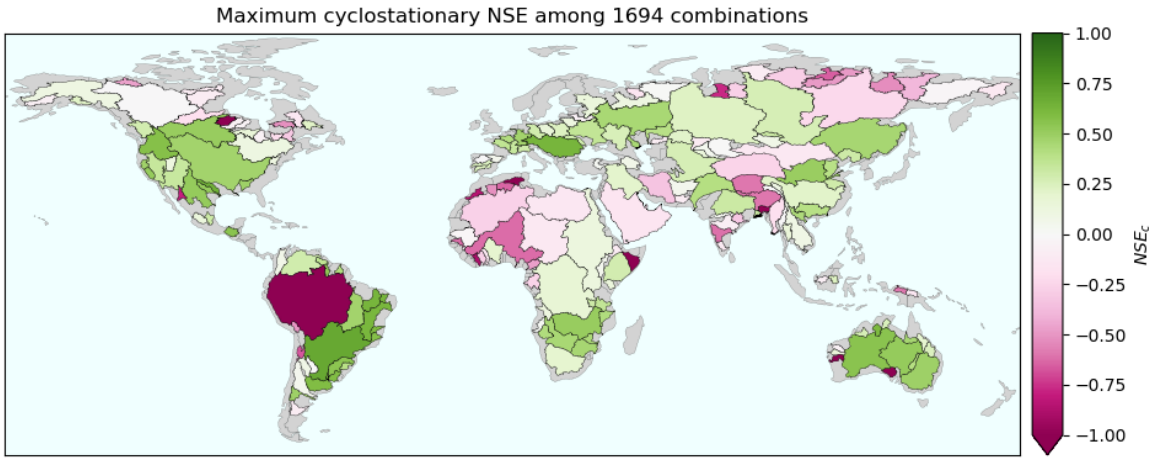


Figure 2.4: Maximum cyclostationary NSE per basin over all combinations. Green positive values mean that the budget reconstruction is a better approximation of GRACE TWSC than the mean monthly values.

excellent. However, when assessing the cyclostationary NSE ($maxNSE_c = -1.28$), it appeared that the mean monthly values were a better fit to GRACE values than the budget reconstruction (Fig. B.2).

The underestimation of annual variability in TWSC can be seen in the correlation plot between GRACE TWSC and our approximation (Fig. B.3). Due to the error in approximating the largest TWSC, the regression slope is 0.7, while 1 is the optimal value. Figure B.2 additionally shows that the water balance error is larger than GRACE uncertainty in 21% of months, meaning that the error is significant.

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. B.4). The regression slope was also almost perfect as shown in Fig. B.5. 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). 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.

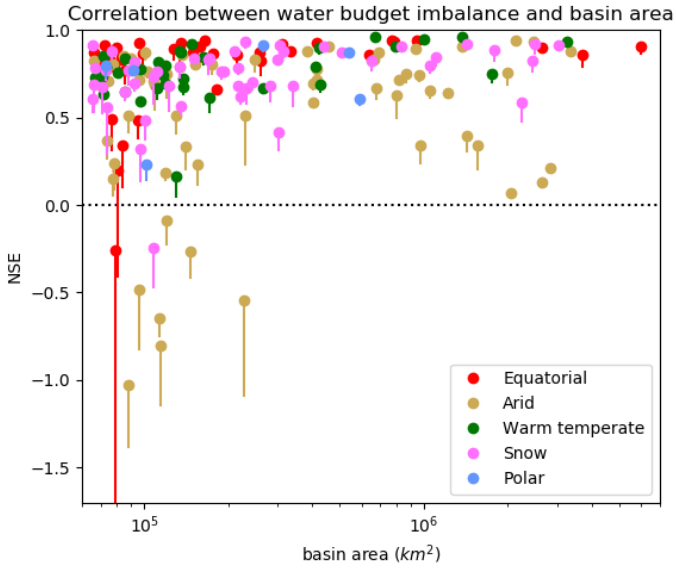


Figure 2.5: Each basin is represented by a bar between the maximum NSE (dot) and the 10th highest NSE

2.4.2 Variables influencing the water budget closure

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. 2.3). Furthermore, Fig. 2.5 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 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 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 2.5 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 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.

Additionally, basins can be classified depending on their climate zone. Figure 2.6 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 2.3). However, this zone also contains small Pacific islands (Papua New Guinea and Borneo) where runoff is much higher than evapotranspiration. Tables A.1 and A.2 indicate that runoff was more uncertain

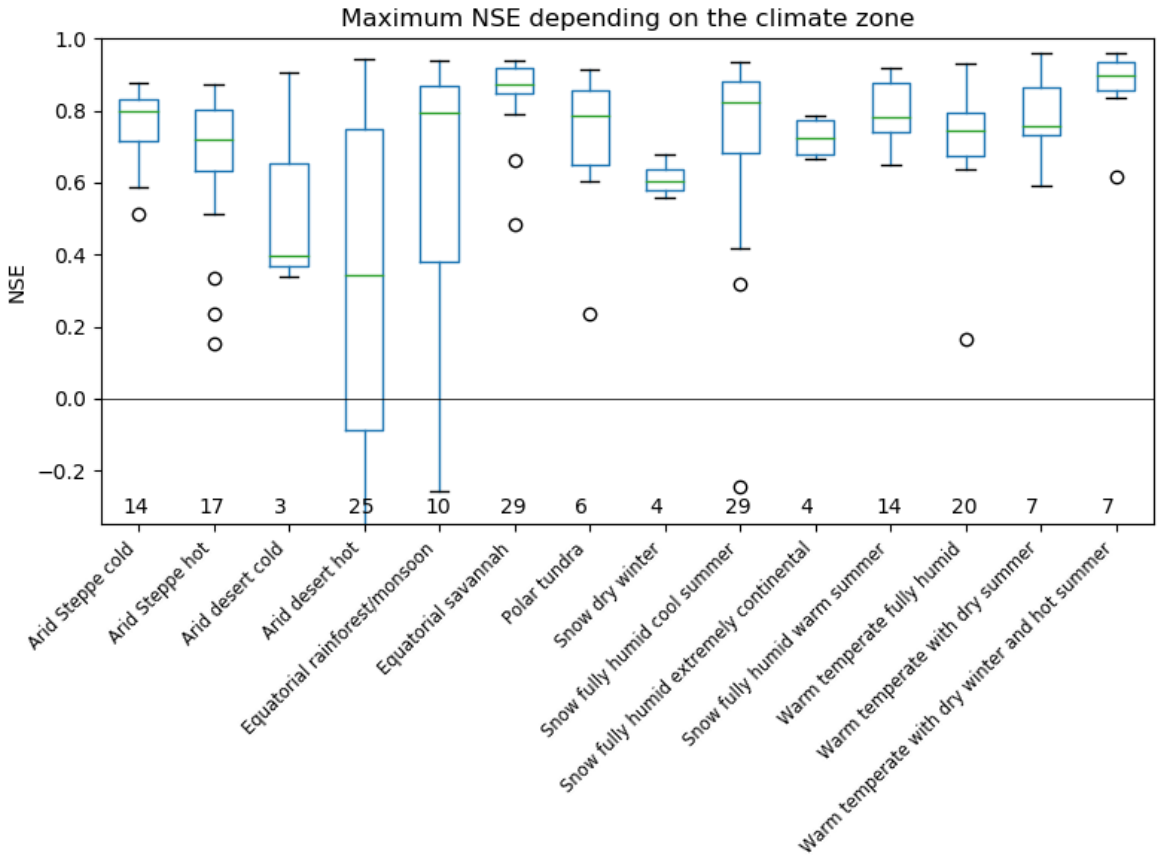


Figure 2.6: Boxplot of the maximum NSE per climate zone. The green line indicates the median, the box extends from the 1st quartile (Q_1) to the 3rd quartile (Q_3) while whiskers go from $Q_1 - 1.5(Q_3 - Q_1)$ (or the minimum value if higher) to $Q_3 + 1.5(Q_3 - Q_1)$ (or the maximum value if lower). Circles denote basins lying out of the whiskers. Figures represent the number of basins in each climate zone.

(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. 2.6). Among those basins, some were entirely desert (Arabian peninsula, Sahara, Somalia, South, and West Australia) with a low signal-to-noise ratio, as previously mentioned. Other basins were partially covered by steppe (Australia, Orange, around Indus) or equatorial savannah (Niger, Chad, Nile). In those basins, precipitation occurred in the more humid subregions, thus increasing TWS variations. As a consequence, the error in the datasets became less significant and allowed a proper budget reconstruction.

2.4.3 Overall combinations performances

Although a majority of basins achieved a positive cyclostationary NSE, they differed greatly in terms of the number of combinations 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. B.6). 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 2.1 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 combinations (35.5 million km^2 with a positive NSE_c from the total study area of 96.6 million km^2). Figure 2.7 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. 2.4 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 2.1 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 CLSM forced with version 2.0 of GLDAS (in particular PGF precipitation). Table 2.1 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 2.1 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. 2.7) 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 2.1 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. 2.7, major differences are observed in Europe, Western Russia, and Alaska. This can be explained by disagreement between precipitation from GPCP and PGF. For instance, CLSM2.1 yielded only low NSE values in most of Eastern Europe whereas version 2.0 of the

Table 2.1: 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.

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 2.1 was not significantly modified by discriminating basins on the area satisfying a NSE larger than 0.5 (usually considered as good performances)

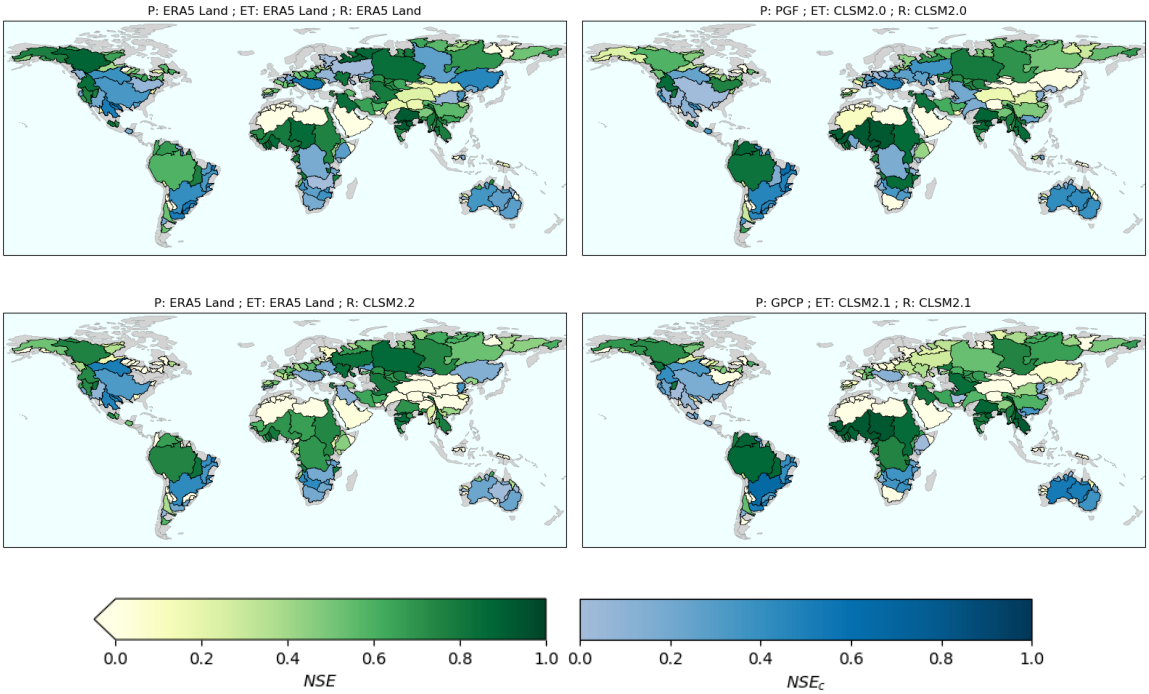


Figure 2.7: NSE and cyclostationary NSE with the first combinations in Table 2.1. Basins with a positive cyclostationary NSE are represented with blue shades corresponding to the NSE_c. Remaining basins are depicted in green, according to their NSE.

instead of a positive cyclostationary NSE. This ensures the reliability of the method used to highlight the most consistent combinations.

2.4.4 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. For this analysis, we focus on a subset of 132 basins out of the 189, where an excellent budget closure could be achieved (maximum NSE larger than 0.8 or maximum NSE_c larger than 0.1).

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. B.12, B.13, B.14, and B.15. 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.

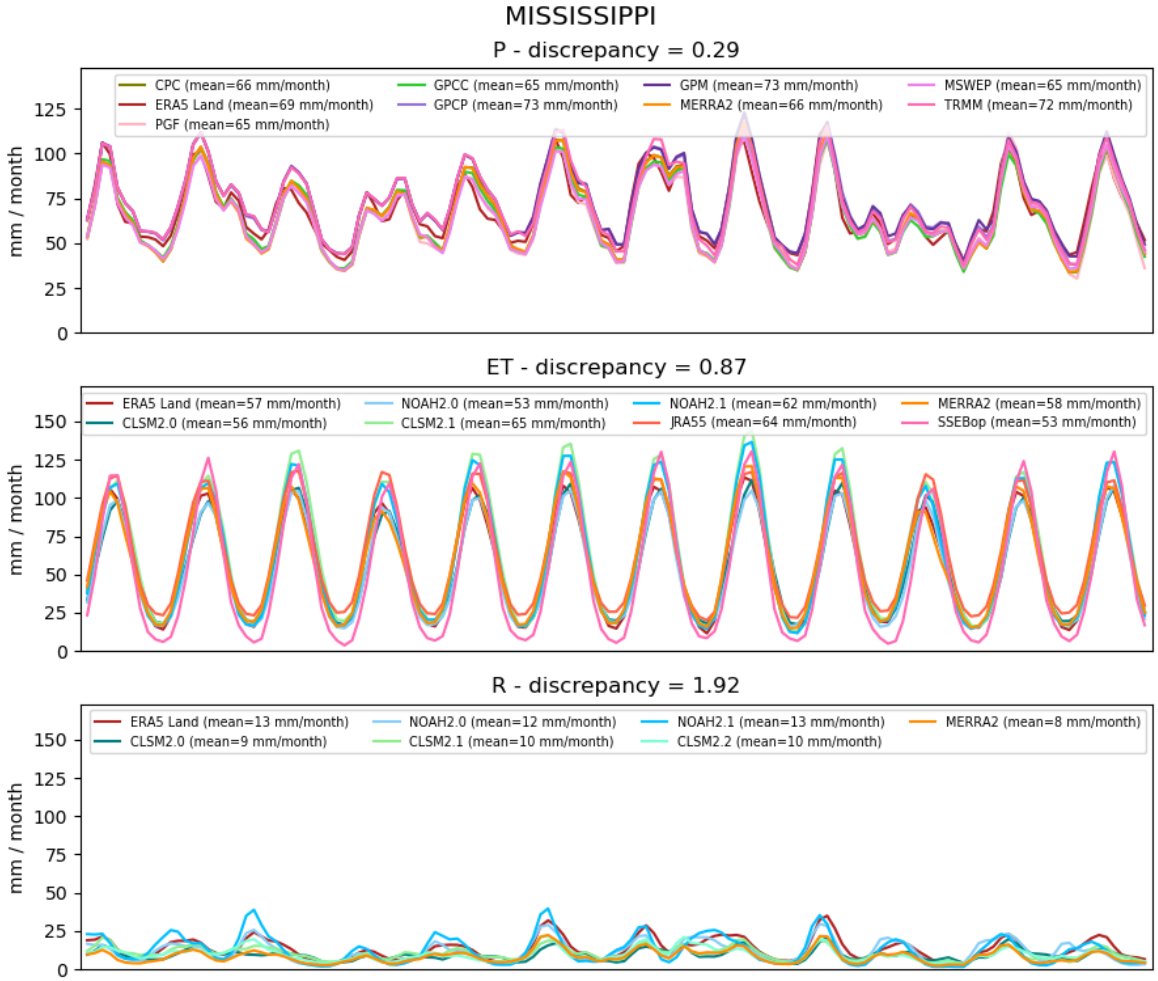


Figure 2.8: Datasets appearing in suitable combinations in the Mississippi basin (cost lower than 0.1). The discrepancy is similar to the coefficient of variation, except that the numerator is the difference between the maximum and minimum values instead of the standard deviation.

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. 2.8 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 performance, we chose to gather basins achieving the water budget closure for similar combinations. Those regions were determined

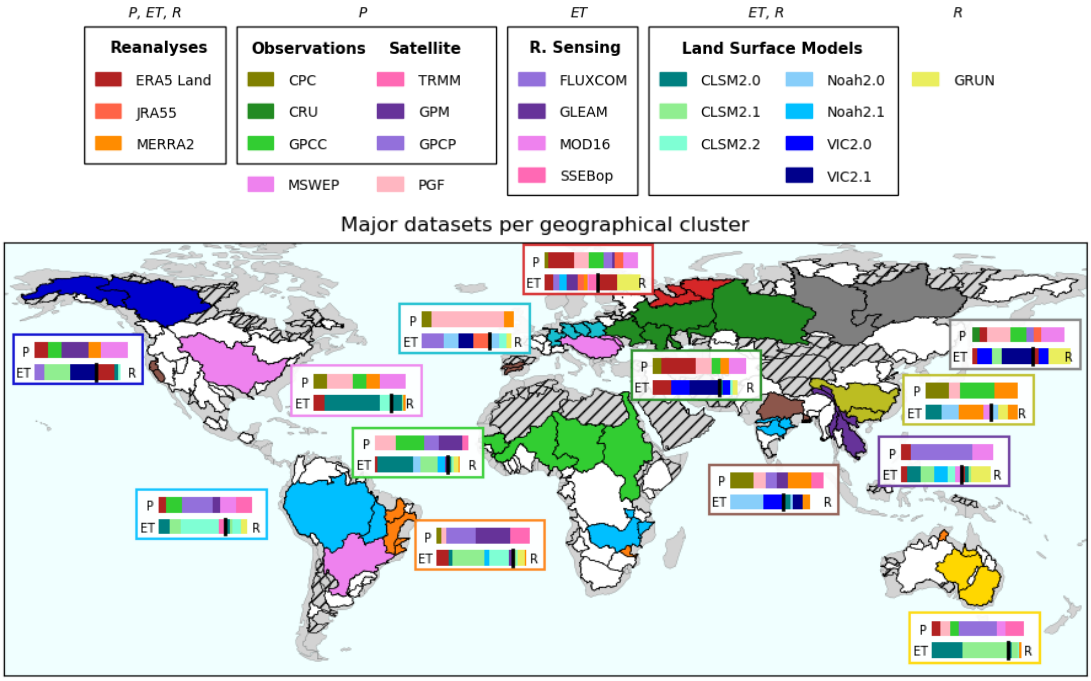


Figure 2.9: Datasets appearing in combinations that satisfy a cost lower than 0.1 for all basins inside the cluster. The 13 clusters highlighted in Fig. B.7 are shown with different colors. For each cluster, the top line of each box represents precipitation datasets. The left part of the bottom line 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).

with the hierarchical clustering described in section 2.3.3. The 132 selected basins with a good water budget closure are depicted in the dendrogram Fig. B.7 and clusters represent 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. B.7 and are shown with the same basin colors on the map in Fig. 2.9.

Basins clustered together in the dendrogram Fig. B.7 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 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).

For each of the 13 clusters, we selected combinations yielding a cost lower than 0.1 in every basin of the region. Figure 2.9 shows which datasets can be used in combination to satisfy the water balance. It first appears that among the precipitation datasets, the rain-gauge-based GPCC

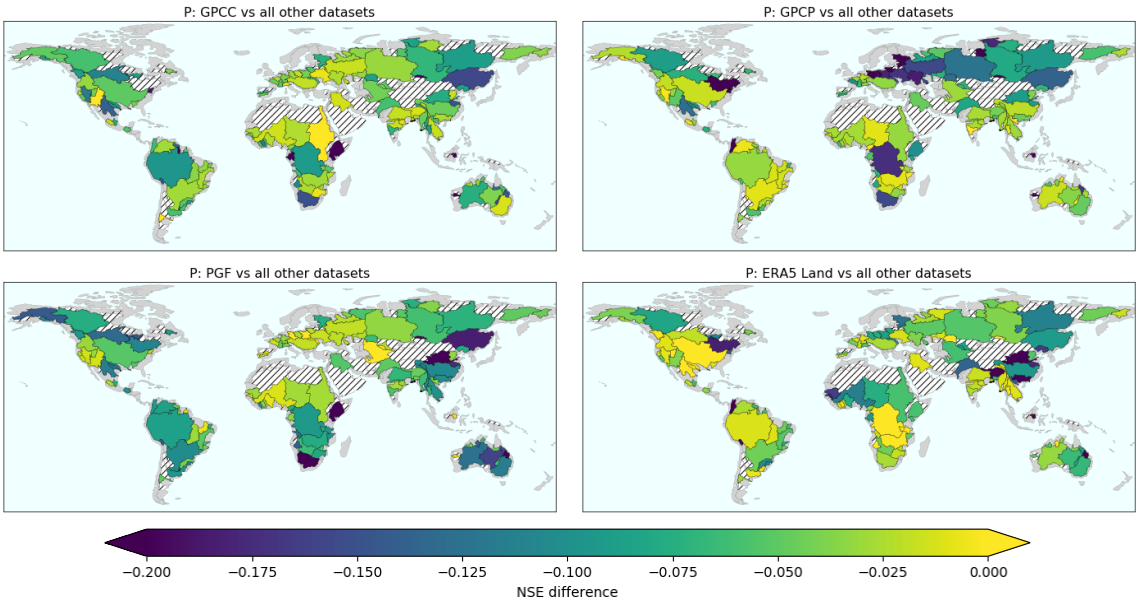


Figure 2.10: The mean of the 10th highest NSE with combinations comprising the reference dataset (*i.e.* GPCC, GPCP, PGF, or ERA5 Land) is compared to the mean of the 10th highest NSE excluding the reference dataset. Yellow indicates basins where the reference dataset is similar to or better than other 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).

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 possibilities.

Figure 2.10 (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 precipitation datasets. Surprisingly, Fig. 2.10 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 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. 2.9, PGF precipitation were able to close the water budget predominantly in Europe, as well as in Central Africa.

For comparison, Fig. B.8 indicates that CRU which never appears in the map 2.9 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. 2.9). Since MERRA2 is based on CPC

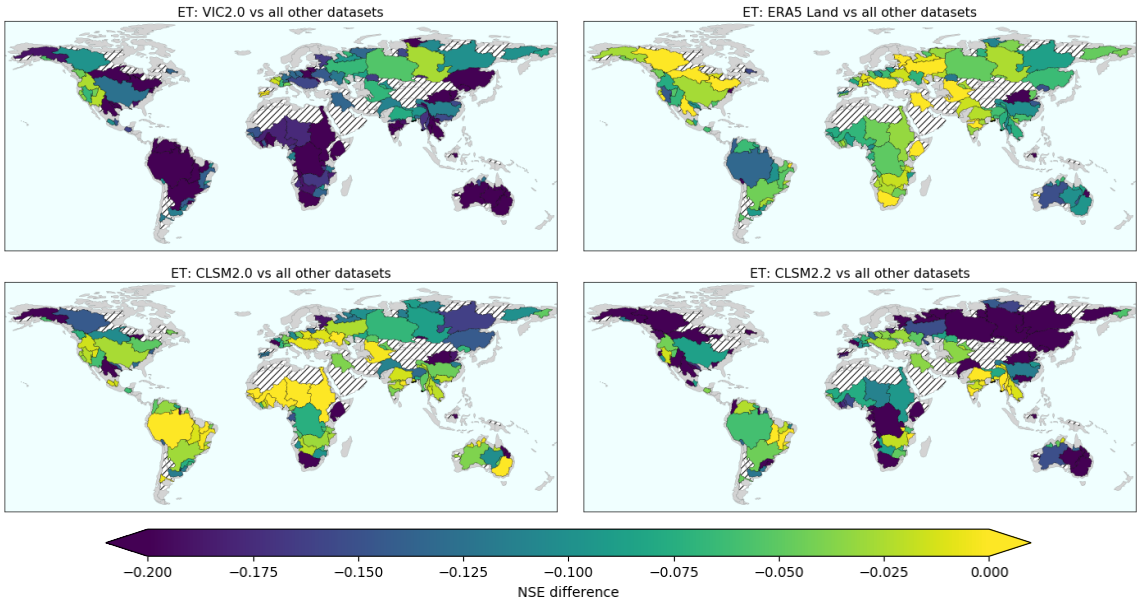


Figure 2.11: Same as 2.10 but for evapotranspiration datasets.

observations (except in Africa where slight variations can be seen in Fig. B.8), 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. B.8).

Fig. 2.9 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. 2.11). We found that VIC produces smaller evapotranspiration than other datasets, along with higher runoff. CLSM was also consistently found in Fig. 2.9. Version 2.0 and 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 precipitation, ERA5 Land evapotranspiration is an excellent dataset in most of the regions except the Amazon basin, China, and Australia (Fig. 2.11).

Evapotranspiration from version 2.2 of CLSM 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. 2.11). An example of this behaviour is given in Fig. B.10 where highly negative values appear in autumn. Since this dataset assimilates GRACE measurements and was validated against GRDC observations, these results are unexpected and should be further explored by the scientific community.

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. B.9). 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 the water budget closure. The poor performances of MOD16 have already been highlighted by e.g., Pascolini-Campbell et al. (2020) in the USA, Bhattarai et al. (2019) in India.

The evaluation of runoff datasets in Fig. 2.12 confirms the differences exhibited for evapotranspiration (Fig. 2.11). 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. 2.12 that this 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.

The low NSE decays on Fig. 2.12 indicate that the version 2.2 of CLSM provide accurate runoff estimations, which is the main objective of this dataset (Li et al., 2019). However, Fig. 2.12 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.

2.5 Conclusion of Chapter 2

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).

Based on these findings, we believe that the water balance equation can be used to reconstruct TWS prior GRACE era. As mentioned, there is a risk of error cancellation, care should be taken to not obtain right results for the wrong reasons. Chapter 3 will focus on methods to reach this goal and the limitations towards its achievement.

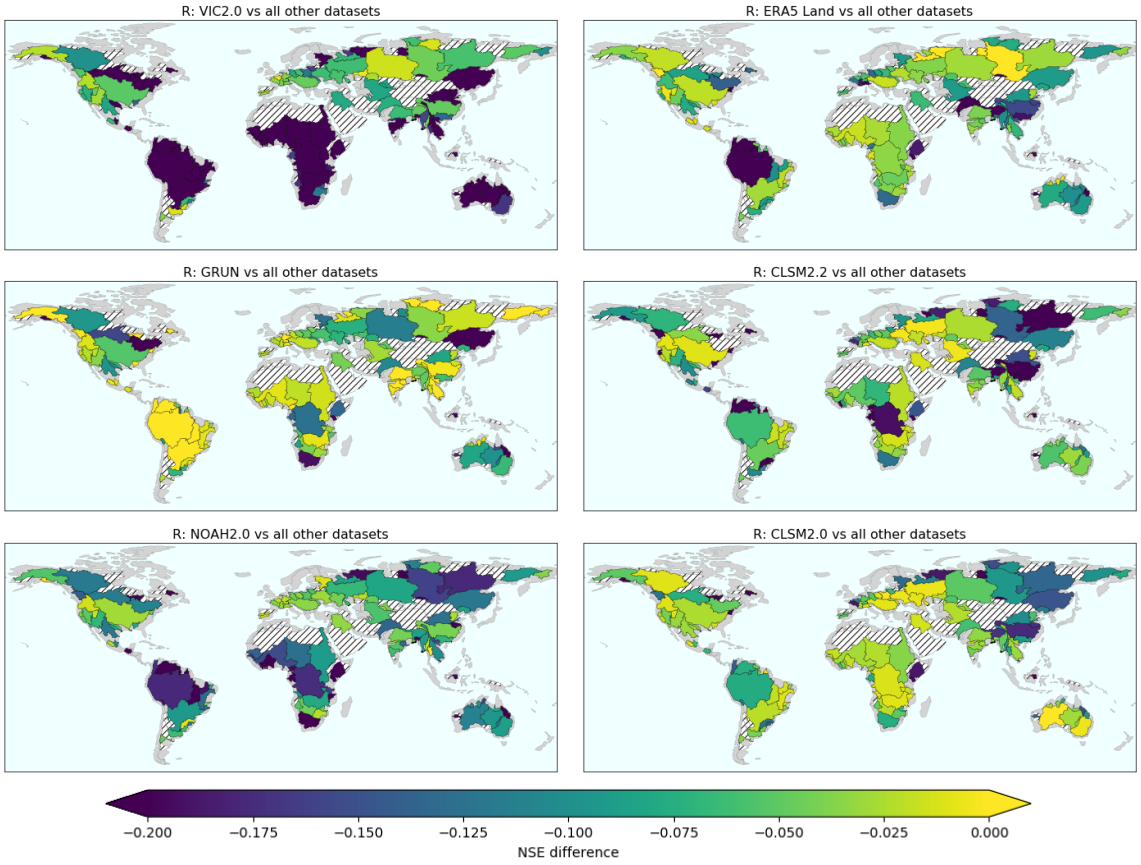


Figure 2.12: Same as 2.10 but for runoff datasets.

CHAPTER 3

RECONSTRUCTING TERRESTRIAL WATER STORAGE BEFORE 2002

As water scarcity becomes an increasing threat in many regions of the world, there is a growing need for a better understanding of the temporal variations of TWS. This knowledge requires, among others, longer time-series of TWS than what are currently available. Although some reconstructions of TWS based on GRACE findings exist, they have not been compared. Therefore, this chapter firstly provides an evaluation of existing TWS reconstructions. Then, it presents a novel optimization approach that extends TWS using the water balance equation. Even though our method has several limitations, we show that there are still large uncertainties surrounding TWS reconstructions and further work is required to estimate TWS variations before GRACE era.

3.1 Literature review on TWS reconstruction

Various methods have been developed to estimate TWS, ranging from hydrological models that simulate separate components of TWS to artificial neural networks and regression methods that design relationships between various predictors and TWS. The water budget equation has also been used to provide an alternative estimate of TWS prior GRACE era. Among the numerous studies elaborating on GRACE findings to extend, predict or downscale TWS measurements, we focused on the ones reconstructing TWS prior 2002 (summarised in Tab. 3.1). Except the following subsection that explains the concepts of land surface and water resource models, we do not present results of studies relying on model simulations of TWS.

3.1.1 Land surface models and hydrological models

Hydrological models can be divided in two main categories: Land Surface Models (LSMs) and Global Hydrological and Water Resource Models (GHWRMs). The former generally use a spatial grid to represent the Earth, assuming that all variables are constant within a grid cell. They can be coupled with atmospheric models and are mainly designed to simulate fluxes between the land, the ocean, and the atmosphere (Scanlon et al., 2018). The Noah and Variable Infiltration Capacity (VIC) models are two such LSMs that are used in this study (Chen et al., 1996; Liang et al., 1994). They do not represent all components from TWS since they only model plant canopy surface water, snow melt and snow precipitation, and soil moisture with layers of varying depth but only up to 10 meters (Li et al., 2020a).

In contrast, the Catchment Land Surface Model (CLSM) does not use a regular grid parameterisation of the Earth but instead sets cells from hydrological catchments defined by topography (Koster et al., 2000). In addition to the components included in Noah and VIC, CLSM also models subsurface water storage changes. However, it does not simulate rivers and streams that can have large contributions to TWS in wet climates (Li et al., 2019), nor does it simulate groundwater. Since they do not simulate all components of TWS, LSMs cannot be expected to accurately simulate TWS. In addition, the fact that they focus on fluxes rather than on storage means that even when being modelled, elements of TWS may be wrong (Milly et al., 2010).

To compensate the lack of TWS representation in LSMs, several methods have been developed to extrapolate GRACE measurements through LSMs. Sun et al. (2019) used deep convolutional neural networks to analyze the error between TWS simulated by Noah and GRACE measurements. They showed that their method was able to significantly improve Noah TWS in India, a region with strong water depletion due to human activities that cannot be captured by LSMs (Chen et al., 2016). Similarly, Jing et al. (2020) used two ensemble learning methods (random forest and extreme gradient boost) to learn the relationship between components of TWS simulated by the Catchment Land Model (the predecessor of CLSM) and GRACE TWS. They reconstructed TWS from 1979 in the Nile river basin and showed coherent results with other datasets and climate indices.

GHWRMs are conceptually different from LSMs since they try to model all components of water storage with an emphasis on human interaction with water resources. Two common GHWRMs are the PCR-GLOBWB (Sutanudjaja et al., 2018) and the WaterGAP models (Döll et al., 2003). They combine physically based simulations for e.g. evaporation or water infiltration, with models of human intervention (e.g. water withdrawal, irrigation water, return flows). Despite their complex framework, GHWRMs have been shown to be less accurate than GRACE. For example, Schellekens et al. (2017) suggested that GHWRMs underestimate TWS because of lacking water stores. In addition, Scanlon et al. (2018) revealed that models generally produce smaller linear trends than those estimated from GRACE measurements. As an example in the Ganges basin, trends from GRACE were comprised between -12 and $-17 \mathrm{km}^3/\mathrm{yr}$ while the

WaterGAP model produced a smaller trend of $-6.6\text{km}^3/\text{yr}$, PCR-GLOBWB led to an opposite trend of $4.8\text{km}^3/\text{yr}$, and LSMs yielded various results from -5.3 to $7.4\text{km}^3/\text{yr}$.

3.1.2 Artificial Neural Networks

Since the early nineties, Artificial Neural Networks (ANNs) have been used in hydrology for diverse purposes such as rainfall-runoff modeling or precipitation forecasting (ASCE, 2000). An ANN is an arrangement of nodes that tries to mimic synaptic connections in the human brain. Nodes are gathered in layers that characterize the ANN. Most ANNs are called feedforward multilayers since they contain one input layer, one or several hidden layers, and one output layer. The input layer collects variables used as predictors. Each node is connected to one or several nodes in the following layer by the means of weights. The output layer corresponds to the objective variable one wants to predict. The design of an ANN (the number of layers and number of nodes per layer) is generally done via trial and error and the training process consists in learning the weight associated with each node (ASCE, 2000).

Due to the complexity of TWS variations in space and time, ANNs have been widely used to predict and reconstruct TWS from GRACE measurements. Long et al. (2014b) designed an ANN predicting TWS at month t from precipitation, air temperature, and soil moisture at month $t-1$. They applied it in Southwest China to reconstruct TWS from 1979 to 2002 and found that TWS variations were mainly driven by soil moisture in this region. Later, Zhang et al. (2016) extended the same framework in the Yangtze river basin to analyse hydrological droughts.

Being given the variety of networks that can be used in conjunction with GRACE, Li et al. (2020b) proposed a unified framework to compare the performances of Multiple Linear Regression (MLR), ANN and autoregressive exogenous method (ARX, Ljung 1987). For each method, three predictors were chosen from a set of 17 climate indices, precipitation, air temperature, and sea surface temperature. While the MLR imposes a simple linear relationship between TWS and the three predictors, the ARX states that the output variable depends on its previous values in addition to current and previous values of predictors (exogenous variables). In a more complex setting, the ANN is able to exhibit non linear relationships between variables. Li et al. (2020b) found that the MLR was the most robust method to reconstruct TWS in a large number of basins.

3.1.2.1 Spatial decomposition

Since variables involve (potentially long) time-series of spatially gridded data, a dimension reduction process is necessary to compute the solution (Forootan et al., 2014). The spatial decomposition consists in separating time-varying maps in a spatial component (assumed to be constant in time) and temporal patterns. Li et al. (2020b) compares Principal Component Analysis (PCA, Wold et al. 1987) and Independent Component Analysis (ICA, Comon 1994) to perform the spatiotemporal decomposition. Contrary to the PCA that seeks orthogonal spatial components, the ICA imposes physically independent spatial patterns by introducing a rotation

matrix. Li et al. (2020b) showed that the PCA should be preferred to the ICA when reconstructing TWS.

Once the temporal components have been extracted, they cannot be directly related to the predictors as they involve processes with different time scales and each process is governed by distinct predictors. Therefore, several studies used time-series decomposition to adapt the methods to the temporal components (Humphrey et al., 2017; Humphrey and Gudmundsson, 2019; Li et al., 2020b, 2021).

3.1.2.2 Seasonal decomposition

Time-series of TWS are commonly decomposed into four components. A linear trend X_{linear} is firstly extracted using the Theil-Sen estimator (Theil, 1950; Sen, 1968). This method is more robust to outliers than usual linear regression (Humphrey et al., 2016). However, we point out that determining a linear trend from GRACE measurements is very challenging since the trend strongly depends on the period over which it is computed (Vishwakarma et al., 2021). Once the linear part has been removed from the signal, a Seasonal Trend decomposition using Loess (STL) is performed to extract the seasonal component $X_{seasonal}$ with LOESS smoothing filters (Cleveland et al., 1990). The remaining part is separated into high- and low-frequency signals (resp. $X_{residual}$ and $X_{interannual}$), with the low-frequency one containing only signals with a period larger than 12 months (Humphrey et al., 2016).

Li et al. (2020b) compared the STL with least squares regression to separate the temporal components and they found that least squares gave more robust reconstructions of TWS. Their analysis was performed in 26 large basins and Li et al. (2021) extended it to the global scale. They also took advantage of the GRACE-FO mission to test their prediction between 2018 and 2020, which was not possible in the previous studies. Those studies show that ANNs are reliable methods to reconstruct TWS prior 2002, if we assume that GRACE era is long enough to avoid overfitting in the training period. However, ANNs do not outline physical relationships between variables that could enhance our knowledge of TWS variations. Therefore, alternative methods based on prescribed associations between variables have been designed.

3.1.3 Statistical regression

Becker et al. (2011) built on existing methods reconstructing atmospheric or oceanic fields to design a method able to reconstruct TWS in the Amazon basin. They used a PCA decomposition to extract dominant spatial and temporal modes from GRACE TWS. Then, they established a correlation between in-situ discharge measurements and GRACE TWS. Since discharge records were available from 1980, they could extend the temporal components of TWS back to these dates. Finally assuming stationarity in spatial patterns, they reconstructed the spatial and the extended temporal time-series to obtain the reconstruction of TWS.

In order to provide a global reconstruction of TWS, Humphrey et al. (2017) designed a more general framework linking TWS to precipitation and temperature. They explicitly focused on the inter-annual and residual components of TWS that were shown to be correlated with inter-annual variability in precipitation (P) and temperature (T). At each grid cell (mascon), they assumed that TWS can be written as $TWS_{interannual+residual} = \beta_1 P^{\tau}_{interannual+residual} + \beta_2 T_{interannual} + \varepsilon$, with ε the error term and β_1, β_2 the coefficients to fit. A exponential decay τ was applied to precipitation to account for water store recharge induced by incoming water (Humphrey et al., 2016). Humphrey et al. (2017) showed that their reconstruction was better than four LSMs over GRACE period.

Elaborating on these findings, Humphrey and Gudmundsson (2019) proposed an alternative method based on the assumption of linear water storage. TWS on day t depends on its value on the previous day, deprived from some water that left the area (residence time depending on temperature) and augmented by precipitation, $TWS(t) = e^{-\frac{1}{\tau(t)}} TWS(t-1) + P(t)$. Although GRACE TWS and temperature were detrended to fit the parameters, long-term changes were not removed from the precipitation dataset and had major impacts on the results. In fact, using three precipitation datasets, Humphrey and Gudmundsson (2019) found three significantly different trends in the TWS reconstruction at the global scale over the common period from 1979 to 2016. Despite these known limitations, this dataset served as a reference in sea level (Frederikse et al., 2020) and TWS trend assessment studies (Vishwakarma et al., 2021).

Statistical regression studies have focused on inter-annual variability since this time scale is the most likely to be influenced by climatic variations (Humphrey et al., 2017). Authors warned that their reconstructions are not made to provide reliable long-term trends. This is also true for the ANN reconstruction of Li et al. (2021) that propagates GRACE linear trends back in time and predicts only interannual, seasonal, and residual components. The seasonal component is an additional limitation of such reconstructions since it is generally assumed to be constant in time (Becker et al., 2011; Humphrey et al., 2017; Humphrey and Gudmundsson, 2019).

3.1.4 Budget equations

Since budget equations involve variables that can be well-observed (at least in some areas), reconstructing TWS from budget equations appears as a promising approach. If we were able to obtain robust results, this method could especially be used to represent seasonal variations of TWS that are often ignored by other methods. As an example of budget equations applied to TWS estimation, the constrained Kalman filter has been used to enforce the closure of the water balance equation (Pan et al., 2012; Zhang et al., 2018). However, this method is limited to the reconstruction of TWSC that is involved in equation 1.1 while one would like to analyse TWS. Integrating TWSC to get TWS is an immense challenge since errors on TWSC are adding up during the integration process.

Zeng et al. (2008) proposed to correct ET before the integration to impose a null integral over the whole period to overcome this issue. This method resulted in high long-term variability in the two basins under study (the Amazon and Mississippi basins) with questionable physical meaning. More recently, Hirschi and Seneviratne (2017) used a 3-year running mean to remove long-term changes. By doing so, they provided a TWS reconstruction in 341 river basins that is quite well correlated with GRACE (mean anomalies correlation of 0.51). The impacts of the correction before the time integration are discussed in more details in section 3.3.

Contrary to the previously mentioned budget studies that rely on single datasets for P, ET, and R, our approach takes advantage of the large number of datasets examined in Chapter 2. To deal with integration errors creating artificial long-term trends, we propose an optimization framework that tries to match the GRACE trend while reducing the prediction error on TWS. Therefore, we provide a novel approach that helps to point out regions where existing reconstructions strongly disagree.

3.2 Data

The study area is the same in this chapter as the one described in section 2.2. 189 basins larger than 63.000km^2 were used to provide a global reconstruction of TWS. The reference value of TWS was provided by the JPL processing center under the form of mass concentration solutions RL06 (Watkins et al., 2015; Wiese et al., 2018). Due to the availability of datasets, TWS was reconstructed from 1981 to 2014.

3.2.1 Datasets used for hydrological variables

We used the datasets described in section 2.2 that cover the period 1981-2014. They include CPC (Chen and Xie, 2008), GPCC (Schneider et al., 2020), GPCP (Adler et al., 2018), MSWEP (Beck et al., 2019), and PGF (Sheffield et al., 2006) for precipitation. Due to the poor performances of CRU exhibited in chapter 2, CRU was not included in this chapter. To compensate from the lack of TRMM and GPM over the reconstruction period, we added the CHIRPS dataset (Funk et al., 2015). Similarly to TRMM and GPM, it combines satellite observations with rain-gauge measurements and spans the latitudes 50° S to 50° N. It is available at a 0.05° resolution from 1981 to near-present. For evapotranspiration data, we used GLEAM (Martens et al., 2017; Miralles et al., 2011) and FLUXCOM forced by two meteorological settings (CRUNCEP and GSWP3, (Jung et al., 2019)) while GRUN (Ghiggi et al., 2019) was used for runoff.

In addition, land surface models CLSM (Koster et al., 2000), NOAH (Chen et al., 1996), and VIC (Liang et al., 1994) were used from the version 2.0 of GLDAS (Li et al., 2020a) for evapotranspiration and runoff. Finally, reanalyses MERRA2 (Gelaro et al., 2017) and ERA5 Land (Muñoz-Sabater, 2019) were included for values of precipitation, evapotranspiration, and runoff over the reconstruction period.

3.2.2 Comparison data

One major challenge when assessing TWS reconstructions is the lack of comparison data prior to the GRACE launch in 2002. Therefore, alternative sources of data need to be found to assess the accuracy of our reconstruction. One first possibility is to explore the impacts of major climatic variations on TWS. On a global scale, El Niño/Southern oscillations (ENSO) are the most influential phenomena that can modify climate variables for several years. El Niño events correspond to warmer sea temperature in the Pacific ocean. By the means of ocean-atmosphere coupling, this influences precipitation and temperature worldwide. Major El Niño events occured in 1982-83, 1997-98, and 2014-16 (Wolter and Timlin, 1993, 1998). The strength of an ENSO event can be measured by the Multivariate ENSO index that involves sea level pressure, surface wind, sea surface temperature, surface air temperature, and total cloudiness fraction of the sky (Wolter and Timlin, 1993, 1998). Monthly values of the MEI from 1981 were acquired from `https://psl.noaa.gov/enso/mei.old/table.html`.

The mean sea level can also be used as an alternative way to assess TWS reconstructions globally. Indeed, since the atmosphere holds a negligible percentage of water, terrestrial water largely contributes to sea-level budget inter-annual variability (Stephens et al., 2020). Accurate measurements of sea level have become available with the generalized use of altimeters in 1993. We used monthly sea level anomaly grids from the European Space Agency as part of the Sea Level Climate Change Initiative project (ESA Sea Level CCI project team, 2017). Weighting each sea grid cell by its area, we computed a global time-series of sea level anomalies.

State-of-the-art TWS reconstructions were also used to compare our results. We included the global reconstructions based on statistical regression from Humphrey et al. (2017) and Humphrey and Gudmundsson (2019). For global results, the machine learning reconstruction of Li et al. (2021) was added to basin-scale results from Li et al. (2020b). In specific basins with sufficient records, we also used the budget reconstruction from Hirschi and Seneviratne (2017) (with the drift correction).

3.3 Methods

From the analysis of datasets conducted in chapter 2, we aim at reconstructing TWS by integrating TWSC. This section describes the method used to perform the integration (section 3.3.1) and to find the optimal time-series of TWS, firstly using integrals of TWSC from the water budget closure (section 3.3.2) and then developing an optimisation approach (section 3.3.3).

3.3.1 Time integration

In this subsection, we assume that we have an optimal time-series of TWSC (also called accumulation $a(t)$) in a given basin. We want to integrate it from 1981 to 2014 to obtain the TWS reconstruction. The reference period is GRACE era, starting in January 2003 to avoid the higher

uncertainties at the beginning of the mission and ending in November 2014. Therefore, we perform a backward integration using November 2014 as the initial value. Formally, we are solving a first-order ordinary differential equation

$$\begin{cases} \dfrac{ds}{dt} & = a(t) \\ s(t_0) & = TWS_{GRACE}(t_0) \end{cases} \tag{3.1}$$

where the accumulation $a(t)$ is assumed to be known over the period 1981-2014, t_0 is the initial time (November 2014), and $TWS_{GRACE}(t_0)$ is the initial value given by GRACE measurements. The integration is performed with the function `solve_ivp` from Scipy, 2020. We found that the Radau solver was well-adapted to the stiffness of the equation (Hairer et al., 1996). Indeed, solutions discretized with a 1-month time step can vary rapidly, therefore leading to numerical instability for some solvers. Equation 3.1 can then be considered as a stiff equation and care must be taken when choosing an adequate numerical solver.

To evaluate the accuracy of the integration, we define a reference TWS that is then differentiated to obtain $a(t)$. The solution $s(t)$ of equation 3.1 can then be compared to the reference TWS. We firstly observed that the time sampling of $a(t)$ was influencing the error on $s(t)$. In fact, the accumulation $a(t)$ is computed at a monthly time-scale and can show steep variations from month to month. Interpolating $a(t)$ with a cubic polynomial before performing the integration reduced the error for all basins. Figure B.16 shows for each basin the RMSD between the reference TWS_{GRACE} and the integral of its derivative $TWSC_{GRACE}$, with and without the cubic interpolation. Since all dots are below the diagonal, this shows that interpolating $a(t)$ always improved the results. The depicted results are obtained with 10 interpolation points per month; we found that the accuracy did not improve further when increasing the number of points.

In addition, we found that the choice of the initial time may significantly influence the error between $s(t)$ and the reference TWS. Due to the shape of the signal, it seems that starting the integration on a month with a large derivative (in absolute value) is more likely to lead to larger errors. Although this was not verified in every basin, it appeared preferable that the initial time did not correspond to an extremum of TWS. Figure B.17 exemplifies these findings in the Seine river basin. The error was higher when t_0 was maximising TWS (e.g. $RMSD = 19.7mm$ with t_0 being March 2014) compared to months with smoothest variations (e.g. $RMSD = 8.0mm$ with t_0 being November 2014). Note that the errors are small since the derivative is computed from the reference TWS. The initial time has even more impacts when the derivative is an approximation of the reference.

In the following, the accumulation $a(t)$ is cubic interpolated (with 10 points per month) before the time integration. Equation 3.1 is solved for initial times between June and November 2014, which gives 6 solutions s. Only the solution minimising the error is saved as the optimized TWS.

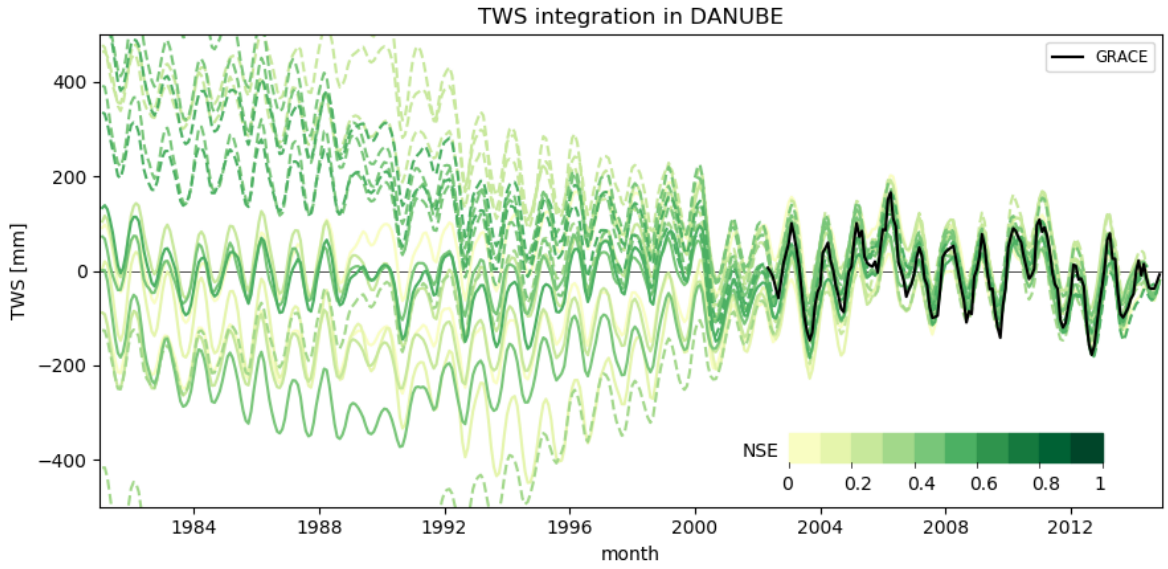


Figure 3.1: The integrated time-series $s(t)$ in the Danube basin for 18 combinations achieving a positive NSE over GRACE period. The darkness of the line indicates the NSE: darker colors correspond to lower errors. Dotted lines show combinations with a trend (1981-2002) larger than 2.5 times the trend observed over GRACE period (2003-2014).

3.3.2 Integration of individual TWSC time-series

In chapter 2, we showed that the water budget closure (on TWSC) was very satisfying since a positive NSE was achieved in 99% of the basins and 45% of them even reached a NSE larger than 0.8. These results did not change significantly when using datasets only available from 1981 (95% of basins achieved a positive NSE, 39% a NSE larger than 0.8). Therefore, it is logical to expect that integrating the $TWSC_{budget}$ can provide reasonable estimates of TWS from 1981 to 2003.

To exemplify our approach, we chose the Danube basin that led to a small imbalance error on TWSC: among the 384 combinations available from 1981 to 2014, 360 achieved a positive NSE and 167 a positive cyclostationary NSE. For each combination of P, ET, and R datasets, the accumulation $a(t)$ was computed from the water balance equation 1.1 and integrated with the method described in section 3.3.1. The time-series obtained (from 1981 to 2014), was then compared with GRACE measurements of TWS (2003-2014). We found that only 18 combinations satisfied a positive NSE over the period 2003-2014 (Fig. 3.1).

One of the main drawbacks of the time integration is the addition of errors between consecutive months. This is especially crucial when the accumulation $a(t)$ presents some constant bias to the reference $TWSC_{GRACE}$ since this bias becomes a linear trend after integration. To illustrate this point, Fig. 3.1 shows with dotted lines time-series with a trend larger than 2.5 times the trend over GRACE period (-0.25 mm/month, 95% CI [-0.51, 0] in the Danube basin). There is no physical reason why large trends would have occurred in the Danube basin between 1981 and

2014 and those can very likely be considered as artificial. This proves that even combinations with a low error over the GRACE period can deviate and exhibit unrealistic trends.

Zeng et al. (2008) and Hirschi and Seneviratne (2017) already faced this issue and proposed different techniques to overcome it. Zeng et al. (2008) corrected ET with a constant so that the integral of the corrected accumulation is 0 over the total integration period. However, figure B.18 shows that this method creates an artificial long-term variability. In the Danube basin, this correction suggests a decrease in TWS from 1981 to 1991, compensated by an increase from 1992 to 2000. Such variations do not correspond to any known modifications of TWS in the region. To preserve more inter-annual variability, Hirschi and Seneviratne (2017) proposed to remove a 3-year running mean from the accumulation $a(t)$ before integrating. Figure B.19 confirms that the different time-series have similar inter-annual variability. Nevertheless, it remains impossible to draw any conclusion on long-term variability from this approach.

3.3.3 Optimization problem

We propose an alternative approach that takes advantage of the large number of datasets available to estimate the accuracy of each dataset. In fact, one major limitation of the three methods presented in the previous Section 3.3.2 is that several time-series are obtained after integration but we are not able to exhibit the most plausible ones. We believe that if models relying on different sources of data provide similar values for P, ET, or R, then they are more likely to be close to the truth. We then design an optimization approach that dynamically tries to combine dataset to improve TWSC. The time integration is afterwards performed exactly as described in Section 3.3.1.

In a given basin, each precipitation dataset provides a different time-series $P_i(t)$ for $i \in \{1, \cdots, n_P\}$ ($n_P = 8$ precipitation datasets, detailed in Section 3.2.1). Similarly, there are $n_{ET} = 8$ evapotranspiration and $n_R = 6$ runoff datasets. From the water balance equation 1.1, we assume that TWSC at month t can be written as

$$TWSC_{opt}(t) = \sum_{i=1}^{n_P} \alpha_i P_i(t) - \sum_{j=1}^{n_{ET}} \beta_j ET_j(t) - \sum_{k=1}^{n_R} \gamma_k R_k(t) + \sum_{m=1}^{12} c_m \mathbb{1}_{month(t)=m} + \varepsilon(t). \tag{3.2}$$

We showed in Chapter 2 that some basins present systematic monthly errors, such as an overestimation of TWSC in the driest months (June-September) in the Amazon basin (Fig. B.2). We try to correct these effects by allowing a monthly correction c_m that is kept constant between years. Coefficients c_m are constrained to be lower than 20% of the mean value of the corresponding month. The random variable ε accounts for measurements errors as well as model approximations.

The weights α_i, β_j, and γ_k represent the accuracy of each dataset. Since we showed that some datasets constantly under- (or over-) estimate hydrological variables in some parts of the world, we should allow the algorithm to compensate these errors. Therefore, instead of constraining the sum of weights to equal 1, some deviation is possible. The constraint on weights becomes $\left|\sum_{i=1}^{n_P} \alpha_i - 1\right| \leq \delta$, with $\delta = 0.05$. A similar constraint is applied to β_j and γ_k.

The variable we are trying to optimize is $\boldsymbol{x} = (\alpha_1 \cdots \alpha_{n_P}\ \beta_1 \cdots \beta_{n_{ET}}\ \gamma_1 \cdots \gamma_{n_R}\ c_1 \cdots c_{12}) \in \mathbb{R}^n$. We denote $\boldsymbol{y} = (TWSC_{opt}(t_m) \cdots TWSC_{opt}(t_M))$ the objective vector containing estimates of TWSC from January 2003 (t_m) to November 2014 (t_M). Then, equation 3.2 becomes in notation matrix

$$\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{\varepsilon}, \tag{3.3}$$

where each column of A contains values of one dataset for all months. However, the aim of this method is not only to optimize TWSC but mainly to reconstruct TWS. Therefore, our interest lies in the integral $\boldsymbol{Y}$ of $\boldsymbol{y}$, satisfying for all t, $Y'(t) = y(t)$. By definition, $Y(t)$ can be written

$$\begin{aligned} Y(t) &= Y(t_m) + \int_{t_m}^{t} y(\tau) d\tau \\ &= Y(t_m) + \int_{t_m}^{t} (A\boldsymbol{x})(\tau) d\tau + \eta(t), \end{aligned}$$

using equation 3.3 and denoting η the integral of ε. Involving the optimized water balance equation 3.2 yields

$$Y(t) = Y(t_m) + \int_{t_m}^{t} \left(\sum_{i=1}^{n_P} \alpha_i P_i(\tau) - \sum_{j=1}^{n_{ET}} \beta_j ET_j(\tau) - \sum_{k=1}^{n_R} \gamma_k R_k(\tau) + \sum_{m=1}^{12} c_m \mathbb{1}_{month(\tau)=m} \right) d\tau + \eta(t).$$

Then, inverting the sum and the integral gives

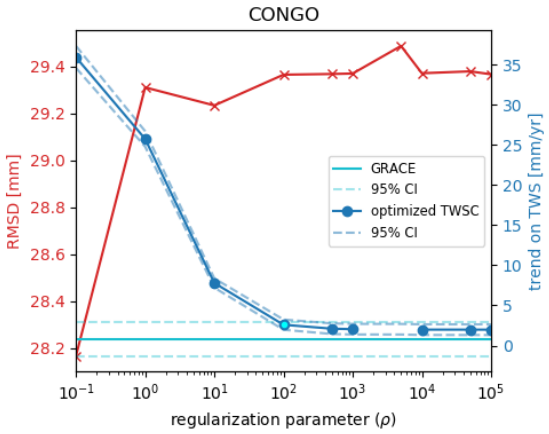
$$Y(t) = Y(t_m) + \sum_{i=1}^{n_P} \alpha_i \underbrace{\int_{t_m}^{t} P_i(\tau) d\tau}_{\overline{P}_i(t)} - \sum_{j=1}^{n_{ET}} \beta_j \underbrace{\int_{t_m}^{t} ET_j(\tau) d\tau}_{\overline{ET}_j(t)} - \sum_{k=1}^{n_R} \gamma_k \underbrace{\int_{t_m}^{t} R_k(\tau) d\tau}_{\overline{R}_k(t)} + \sum_{m=1}^{12} c_m \int_{t_m}^{t} \mathbb{1}_{month(\tau)=m} d\tau + \eta(t),$$

where $\overline{P}_i(t)$ is the accumulation of precipitation provided by the dataset i between the months t_m and t, similarly for $\overline{ET}_j$ and $\overline{R}_k$. In the algorithm, $\overline{P}_i(t)$ is computed with a trapezoidal rule for each month. This formulation leads to a new matrix $\overline{A}$ that can be deduced from A and that allows to rewrite equation 3.3 in its integrated form

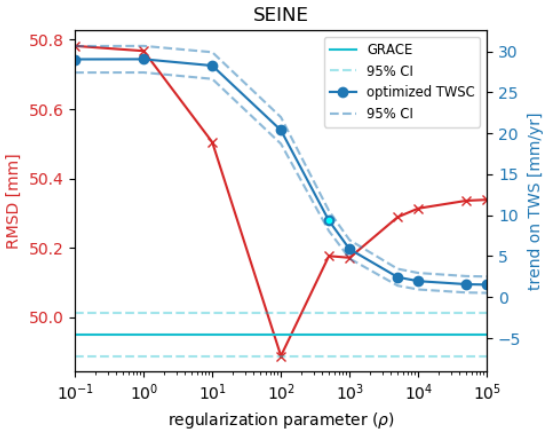
$$Y(t) = Y(t_m) + \overline{A}\boldsymbol{x}(t) + \eta(t) \tag{3.4}$$

The reference vector for $\boldsymbol{Y}$ is $\boldsymbol{Y_{GRACE}} = (TWS_{GRACE}(t_m) \cdots TWC_{GRACE}(t_M))$, the time-series of GRACE measurements. Since we are aiming at minimizing the error between $\boldsymbol{Y}$ and $\boldsymbol{Y_{GRACE}}$, the loss function should contain the mean squared error between those vectors $\|\boldsymbol{Y} - \boldsymbol{Y_{GRACE}}\|_2^2 = \sum_{t=t_m}^{t_M} (Y(t) - Y_{GRACE}(t))^2$. It was mentioned in the previous Section 3.3.2 that artificial long-term trends created by the integration of biases in TWSC were a major issue. Since there is no perfect approach to allow long-term variability while limiting unrealistic trends, we chose to constrain the trend of our TWS reconstruction to approach GRACE trend λ_{GRACE}. This is done by adding a regularization term to the loss function:

$$\mathscr{L}(\boldsymbol{Y}) = \|\boldsymbol{Y} - \boldsymbol{Y_{GRACE}}\|_2^2 + \rho \cdot (\lambda(\boldsymbol{Y}) - \lambda_{GRACE})^2 \tag{3.5}$$



(a) The linear trend $\lambda(\boldsymbol{Y})$ converges towards GRACE trend λ_{GRACE}



(b) $\lambda(\boldsymbol{Y})$ converges outside of GRACE trend confidence interval.

Figure 3.2: RMSD between GRACE TWS and the optimized TWS (red curve) and linear trend (dark blue curve) as a function of the regularization parameter ρ. GRACE trend λ_{GRACE} is shown with the light blue line. 95% confidence intervals (CI) are represented with dotted lines and derived from the Theil-Sen estimator for the linear trend.

Note that the mean squared error is summed over GRACE period $[t_m, t_M]$ while the trend of the TWS reconstruction $\lambda(\boldsymbol{Y})$ is computed over the whole integration period as the mean of $\boldsymbol{y}$, $\lambda(\boldsymbol{Y}) = \frac{1}{M+1}\sum_{t=t_0}^{t_M} y(t)$. From equations 3.3 and 3.4, the loss function can be rewritten in terms of coefficients $\boldsymbol{x}$

$$\mathscr{L}(\boldsymbol{x}) = \sum_{t=t_m}^{t_M} \left(\overline{A}x(t) - (Y_{GRACE}(t) - Y_{GRACE}(t_m))\right)^2 + \rho \cdot \left(\frac{1}{M+1}\sum_{t=t_0}^{t_M} Ax(t) - \lambda_{GRACE}\right)^2 \tag{3.6}$$

The minimization problem was solved using the `cvxpy` library (Diamond and Boyd, 2016; Agrawal et al., 2018). The optimal value for the regularization parameter was found by exploring different values of ρ in $\{0.1, 1, 10, 100, 500, 1000, 5000, 10\,000, 50\,000, 100\,000\}$. In the majority of basins, $\lambda(\boldsymbol{Y})$ converged towards λ_{GRACE} in the sense that the 95% confidence intervals of $\lambda(\boldsymbol{Y})$ and λ_{GRACE} intersected for ρ large enough (Fig. 3.2a). We then chose the smallest value of ρ such that they intersected. In other basins, $\lambda(\boldsymbol{Y})$ did converge but the limit was not inside λ_{GRACE} confidence interval. The selected ρ was the smallest value to reach the convergence criterion: the consecutive difference of $\lambda(\boldsymbol{Y})$ is smaller than 0.5mm/yr.

To summarise, our approach consists in gathering time-series of all available datasets for P, ET, and R in a given basin. For each value of the regularization parameter ρ, we run the optimization algorithm to find the vector of weights $\boldsymbol{x}$ that minimizes the loss function. Using the method described above, we select the best regularization parameter and save the corresponding $\boldsymbol{x}$. From this vector of weights, we finally compute the time-series of monthly TWSC values using equation 3.2 and integrate it following the method described in section 3.3.1. This yields to a time-series of monthly TWS for each basin under study.

3.4 Results and discussion

We firstly present some optimization results about our algorithm. In terms of performance comparison, our reconstruction of TWS is compared with GRACE results over the training period (2003-2014). Then, longer assessments of our reconstruction on a global scale are conducted against the MEI index and the mean sea level. Finally, comparing our results with state-of-the-art reconstructions provides a complementary approach to evaluate our method.

3.4.1 Optimization details

During the optimization process, the algorithm tries to minimize the loss function by weighting the datasets with appropriate coefficients. It is interesting to note that this yields to a proper selection of datasets since the majority of datasets are given a zero coefficient. Figure B.20 shows that only two precipitation datasets were selected in the Amazon basin (coefficient α equal to 0.9 for PGF and 0.12 for GPCP). Similarly, only CLSM2.0 had a significant coefficient for evapotranspiration and runoff. One can note that those datasets appeared as suitable in the Amazon basin in the previous chapter (see maps 2.10, 2.11, and 2.12). Therefore, our method is able to exhibit datasets bringing the most useful contribution to TWS.

However, we note that the datasets coefficients depend on several parameters including the regularization coefficient ρ. Figure B.21 shows this parameter in each basin. The median value of ρ is 100, showing that the constraint of approaching GRACE trend is moderate. We also point out that there is no correlation between the regularization parameter selected with our approach and GRACE trend.

3.4.2 Comparison with GRACE measurements

GRACE measurements of TWS are the most accurate values to assess our reconstruction. However, they are also the objective values of the optimization algorithm and therefore cannot serve as an independent evaluation. Nevertheless, they can be used as a first assessment of our reconstruction's accuracy in the training period. Figure 3.3 shows excellent performances of our reconstruction in a majority of basins since the median NSE is 0.85. One can also notice poor performances in West-European basins, mainly caused by strong errors at the beginning of GRACE ERA (2002-2003). This may be related to the small size of European basins that increases leakage errors.

On a global scale, we computed the global TWS as the mean of basin-scale TWS weighted by the basin area. To better represent intra-annual variations, we obtained deseasonalised time-series of global TWS by removing the mean annual cycle from the full time-series. Figure 3.4 shows a good correlation between GRACE and our reconstruction (correlation coefficient R^2 = 0.85 for the deseasonalised, NSE_c = 0.29 for the full time-series). On a basin scale, the median value of correlation between deseasonalised time-series is 0.87 (0.79 and 0.93 for the

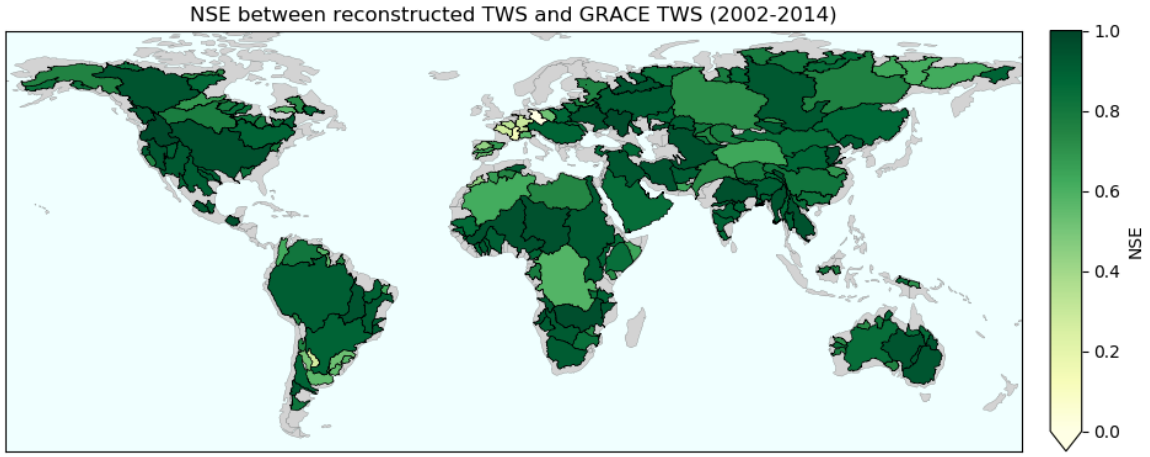


Figure 3.3: NSE between GRACE TWS and our optimized TWS

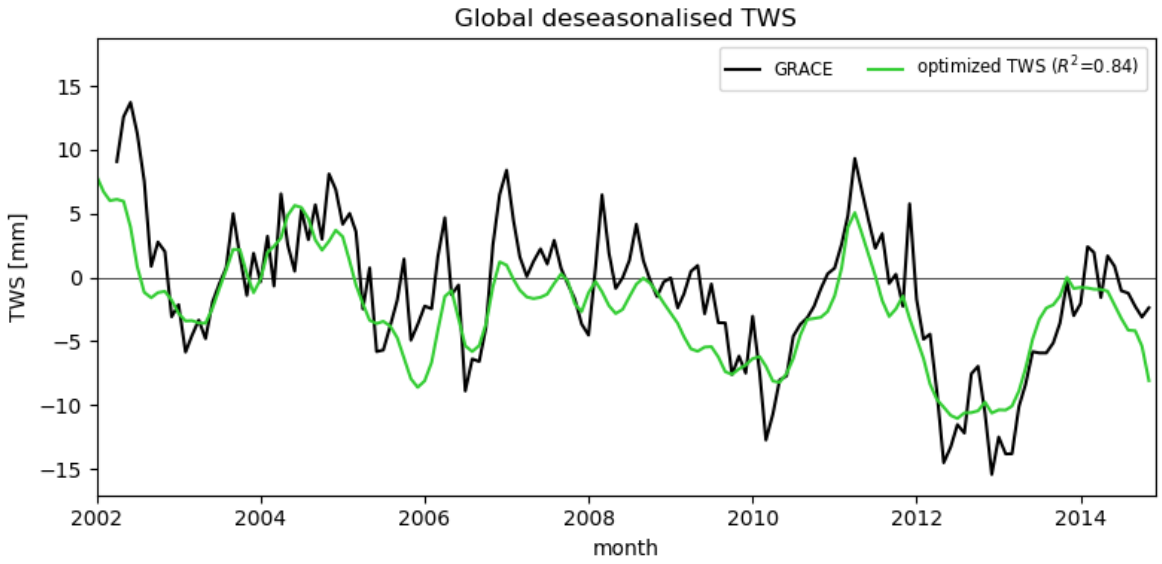


Figure 3.4: Deseasonalised time-series of global GRACE TWS and our optimized TWS.

25th and 75th percentiles). This is higher than the results from Hirschi and Seneviratne (2017) who obtained a mean correlation between deseasonalised time-series of 0.51.

3.4.3 Inter-annual variability

The inter-annual component of TWS corresponds to the non-linear long-term trend (see Section 3.1.2.2). It is the more likely component to be influenced by meteorological variables while the anthropogenic impacts are assumed to be negligible at this time scale (Humphrey et al., 2016). Since ENSO events largely drive climate variations, the MEI is an appropriate index to analyse TWS variations. Figure 3.5 shows that TWS and the MEI generally follow opposite variations,

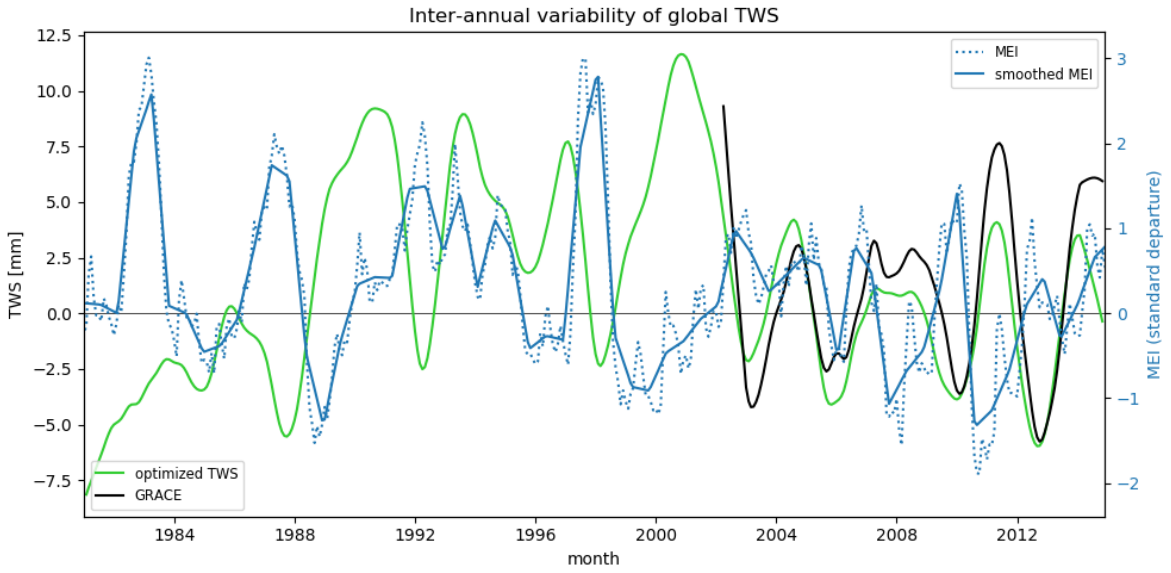


Figure 3.5: Inter-annual components of global TWS from GRACE (black) and from our optimized reconstruction (blue). Original time-series of the MEI (dotted line) and the time-series smoothed by removing the highest Fourier coefficients (blue line).

which is the expected behaviour (Phillips et al., 2012). Indeed, the correlation coefficient between the optimized TWS and the smoothed MEI is -0.22. One can notice on Fig. 3.5 a strong decrease of TWS around 1998 corresponding to a major El Niño event.

On a global scale, comparing our TWS reconstruction against mean sea level also indicates that our reconstruction is able to catch the major inter-annual variations. While Fig. 3.6 shows seemingly opposite variations between global TWS and sea level anomalies, the correlation coefficient is too low to conclusively argue that our reconstruction is perfect. This poor result may come from artificial inter-annual variations induced by the water budget integration (as can be seen in Fig. 3.8). Since anomalies only remove linear trends, long term variations may still be present in the reconstruction that do not have a physical interpretation with respect to mean sea level. In addition, a more sophisticated smoothing may be needed to account for high-frequency variations of TWS that may not transmit from the land to the ocean.

3.4.4 Global comparison with state-of-the-art datasets

Deseasonalised time-series of global TWS are shown in Fig. 3.7 from three state-of-the-art datasets. One can see that our reconstruction lies in the middle of the spread of various time-series, suggesting the plausibility of our findings. All reconstructions agree on a decrease in global TWS around 1998 followed by an increase until 2000. Since Li et al. (2021) extends the linear trend of GRACE period prior 2002, their reconstruction shows the strongest linear trend. On the contrary, Humphrey et al. (2017) rely on detrended data and therefore cannot exhibit

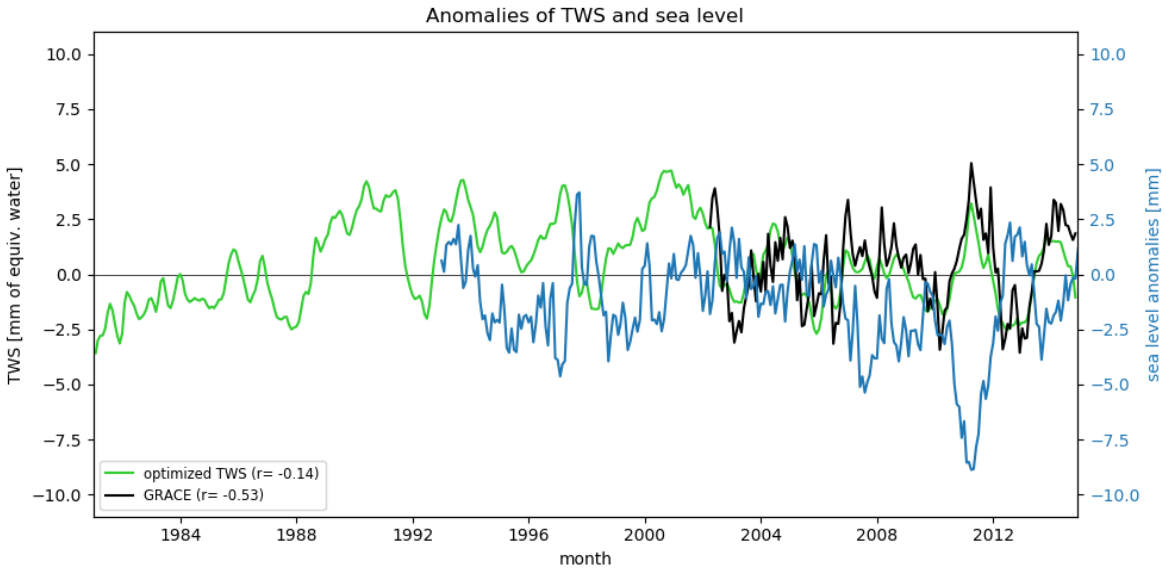


Figure 3.6: Global anomalies time-series (detrended and deseasonalised) of our TWS reconstruction (green) and GRACE (black) normalized by the total land area against global mean sea level anomalies (blue). The correlation coefficients r of TWS with sea level are shown in the legend.

any long-term trend. Similarly, trends that may be seen in Humphrey and Gudmundsson (2019) reconstructions come from underlying trends in precipitation datasets and do not necessarily reflect true TWS trends.

Our reconstruction does not show any linear trend but suggests a slight increase at the beginning of the reconstruction period (1981-1991) followed by a trough in 1993, visible in all reconstructions. The global mean of TWS variations hides the problems at the basin scale. Fig. 3.8 shows a significant difference between the reconstructions. The TWS reconstruction of Hirschi and Seneviratne (2017) has more intra-annual variations than the others but lower amplitudes.

In terms of inter-annual variations, our reconstruction does not strongly agree with the linear trend found by Li et al. (2021). Despite including GRACE trend in the loss function, the Mackenzie and Parana basins show that the long-term trend of our reconstruction is not linear. A decrease in TWS has been observed during GRACE period in the Mackenzie basin, largely due to ice melt and surface water drying (Rodell et al., 2018). It is not very likely that the opposite trend occurred between 1981 and 1996 as our reconstruction suggests.

Additional basins are shown in Fig. B.22, B.23, and B.24. Our reconstruction is generally comparable with state-of-the-art datasets. Surprising peaks or troughs appear in some basins like Congo, Indus, or Ob. As an example, the Congo basin suffers from a lack of precipitation monitoring that creates large uncertainties in the datasets. This may suggest that our algorithm selected poor-quality datasets to perform the reconstruction. For basins with a well-established TWS decrease over the GRACE era (Euphrates, Ganges, and Yellow river from human extraction, Yukon from ice melt), constraining our algorithm to approach GRACE trend leads to reasonable

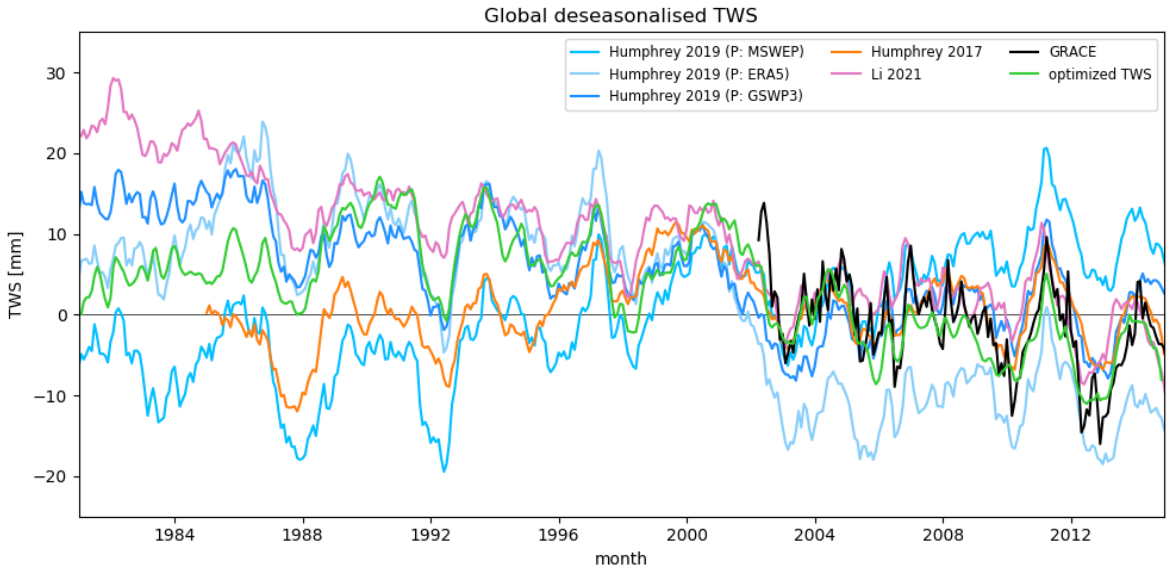


Figure 3.7: Deseasonalised global TWS from Humphrey and Gudmundsson (2019) with three precipitation forcings (MSWEP, ERA5, and GSWP3), Humphrey et al. (2017), Li et al. (2021), GRACE, and our reconstruction

results.

The main conclusion from this comparison with other available datasets is that our approach provides coherent results with a strong physical constraint in the background. Since the reconstruction depends on the quality of other products of water budget components, there are still large uncertainties in TWS predictions prior to the GRACE era. Therefore, we can conclude that it is hard to select one reconstruction over the others.

3.4.5 Limitations

Although the available data to evaluate our reconstruction suggest that it performs reasonably well, several limitations remain. We firstly point out that the training period has a strong influence on the TWS reconstruction. Figure B.25 indeed shows significantly different interannual variations when the training period is modified. Although starting in 2003 or 2004 leads to similar reconstructions, shorter training periods present larger errors. Figure B.26 indeed confirms that basin-scale reconstructions have larger errors compared to GRACE observations when the length of the training period is reduced.

A second limitation of our approach is that selecting the datasets that seemed more promising in Chapter 2 (in the sense that they have the smallest NSE difference, as shown in maps 2.10, 2.11, and 2.12) does not improve results. Although we recognize that the quality of data products varies as we go back in time, we would assume that datasets leading to a better budget closure from 2003 to 2014 would be more accurate to reconstruct TWS before 2003. This may suggest that our

method to identify suitable datasets in Chapter 2 is not suited to select datasets. More in depth analysis of the quality of input data may help discriminating datasets for TWS reconstruction. Additionally, the dependency on the training period may also mean that closing the water budget and reconstructing TWS are two distinct goals that cannot be achieved with the same methods.

Finally, we acknowledge that the availability of datasets is a limiting factor in our analysis. Indeed, many datasets starting before the GRACE era (at least in 1981 as we selected them) do not provide results after 2014, as is the case for the version 2.0 of GLDAS or GRUN. Therefore, we could not use longer records of GRACE measurements or even outputs from the GRACE-Follow on mission to evaluate our reconstruction.

3.5 Conclusion of Chapter 3

We have presented a reconstruction of TWS based on the water balance equation. Our method combines weighted versions of precipitation, evapotranspiration, and runoff datasets to fit GRACE measurements during the training period (2003-2014). It also allows seasonal correction of water variations. To avoid the accumulation of errors due to the time integration, our method adds a regularization term that penalizes reconstructions with an unrealistic linear trend.

We demonstrated that our reconstruction is coherent with existing reconstructions that are based on statistical regression or machine learning. It is also coherent with independent validation data such as the Multivariate ENSO Index and mean sea level anomalies, although correlation coefficients were low. At the basin scale, the analysis of inter-annual variations was made difficult by the presence of non-linear trends in our reconstruction. This analysis nevertheless shows that further work is needed to claim significant improvement in our understanding of the spatiotemporal variations in TWS before the start of the GRACE mission.

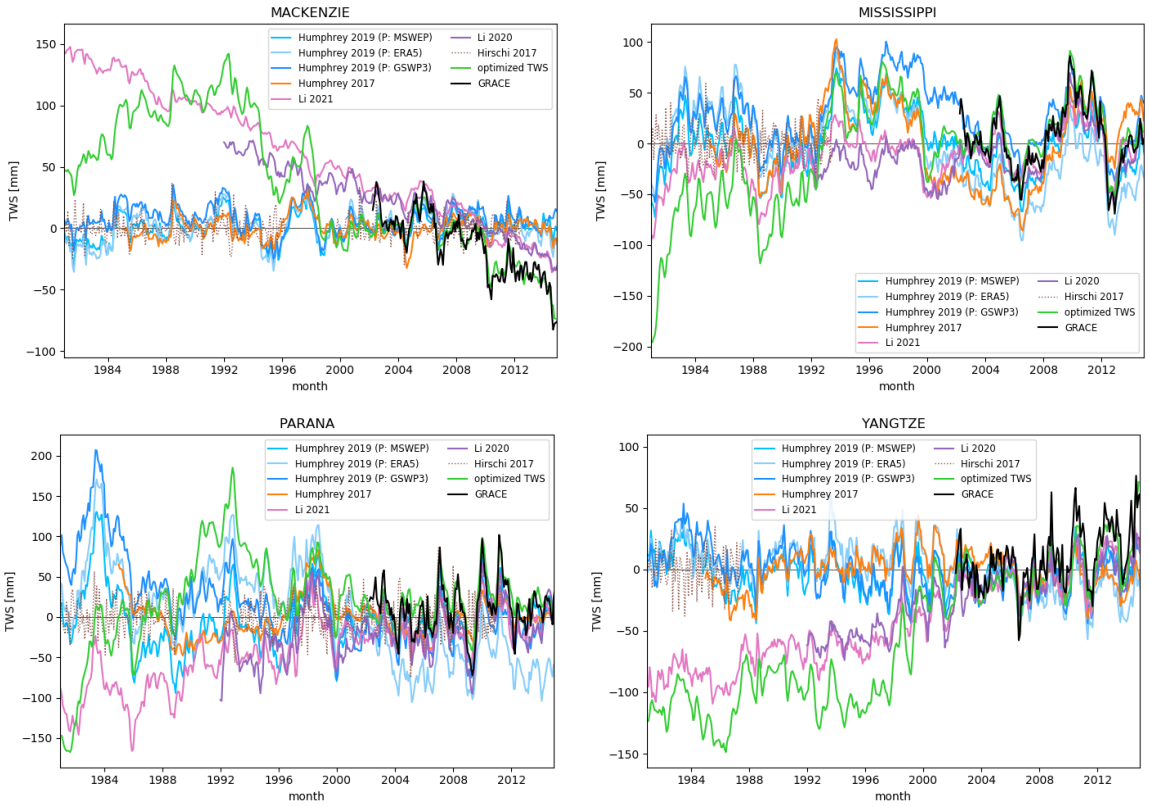


Figure 3.8: Deseasonalised TWS in the Mackenzie, Mississippi, Parana, and Yangtze basins.

Study	Area	GRACE	Extension	Method	Results
Zeng et al. (2008)	Amazon and Mississippi	2002-2006	1970-2006 (Amazon), 1928-2006 (Mississippi)	integrate P-ET-R over the reconstruction period while correcting ET so that TWS has a null integral over the whole period	• good agreement with GRACE in terms of mean seasonal cycle • Amazon: large recharge (+800mm) between 1979 and 1985 and large decline in TWS between 1985 and 1995 • Mississippi: large decline (-500mm) between 1928 and 1940, general rise from 1957
Becker et al. (2011)	Amazon	2002-2008	1980-2008	remove seasonal cycles from river discharge and TWS time series PCA, assume stationarity of spatial models compute time-varying amplitude from the relationship between R and TWS	• correlation with ISBA-TRIP hydrological model (smoothed time-series) is 0.9 • no trend in reconstructed TWS and lower levels during ENSO events • high correlation between first spatial mode and SOI
Long et al. (2014b)	Southwest China	2003-2010 (train) 2011-2012 (test)	1979 - 2002	ANN between TWS at time t and precipitation, air temperature, soil moisture storage (GLDAS Noah) at time $t-1$	• test period: R^2 = 0.91, bias~-21mm, RMSD~28mm • TWS seems to be mainly driven by soil moisture variations (in sub-basins, more differences over GRACE period than for the reconstruction period)
Zhang et al. (2016)	Yangtze river basin	2003-2012 (60% training, 20% validation, 20% testing)	1979-2012	ANN same as Long et al. (2014b)	• testing period: R^2 = 0.89, NSE=0.83. • poorer performances in lower subbasins • analysis of hydrological droughts

Hirschi and Seneviratne (2017)	341 basins larger than 35 000km^2	not applicable	1979-2015	integration of the water balance equation (atmospheric and terrestrial) long-term changes removed with a 3-year running mean ERA-Interim data and discharge gauge measurements	• mean anomalies correlation with GRACE is 0.51
Humphrey et al. (2017)	global (mascon scale)	2002-2015	1985-2015	daily precipitation and temperature from ERA-Interim temporal decomposition and exponential decay filter on P model parameters β_1, β_2, τ are calibrated at each GRACE mascon using Markov chain Monte-Carlo simulations	• non-seasonal TWS at Obidos (Amazon): R^2 = 0.61 (sees the 1998 drought) • "at least 40% of the total variance of GRACE anomalies can be reconstructed from P and T variability alone" • R^2 and NSE better with the reconstruction than 4 LSMs • R^2 with detrended GMSL is -0.88
Humphrey and Gudmundsson (2019)	global (mascon scale)	2003-2014	1901-2014 with GSWP3 1979-2016 with MSWEP and ERA5	precipitation from MSWEP, GSWP3, and ERA5; temperature from ERA5 and GSWP3 TWS on day t depends on TWS on day $t-1$, the residence time of the water store (dependent on temperature) and precipitation over day t model parameters are involved in the residence time model and in the calibration with GRACE (monthly scale, linear trends and seasonal cycles removed, as Humphrey et al. (2017))	• long-term trends highly dependent on (non-detrended) precipitation datasets • higher R^2 and NSE at mascon scale in densely gauged regions, preference for MSWEP or ERA5 • R^2 with GMSL around -0.65 (detrended deseasonalised)

Li et al. (2020b)	26 large basins	2002-2011 (training) 2011-2017 (testing)	1992-2002	spatiotemporal decomposition: ICA or PCA seasonal decomposition of temporal modes: Least squares or STL prediction of each mode from 3 variables among precipitation, temperature, sea surface temperature, and 17 climate indices: ANN, ARX, MLR	• multiple training errors are significantly lower with MLR • least squares should be preferred to STL, and PCA preferred to ICA
Li et al. (2021)	global	2002-2017	1979-2020 (2018-2020 used as a test period)	extension of Li et al. (2020b) use GRACE linear trend in the reconstruction	• good correlation with satellite laser ranging TWS (1992-2002) • correlation with global sea level (deseasoned) is -0.56, -0.63, -0.59 (1992-2001, 2002-2017, 2017-2019)

Table 3.1: abbreviations used: SOI=Southern Oscillation Index, ENSO=El Niño Southern Oscillation, ICA=Independent Component Analysis, PCA=Principal Component Analysis, STL=Seasonal Trend decomposition based on LOESS, ARX=Autoregressive Exogenous model, MLR=Multiple Linear Regression. Refer to section 3.1 for explanations of these terms.

CHAPTER 4

CONCLUSION

Terrestrial Water Storage is a crucial component of the water cycle that provides water for human needs. Severe declining trends have been observed by the GRACE satellite mission in several regions of the world (Rodell et al., 2018). However in the absence of longer records, it remains difficult to detect variations caused by natural variability from those induced by human impacts on climate change. The water balance equation is a tool that describes the fluxes of water in the forms of precipitation, evapotranspiration, runoff, and TWS variations. Knowing the four variables of the equation, it can be used to evaluate the imbalance error and deduce the accuracy of the datasets. Next, assuming that we selected well-purposed datasets, the equation can serve to reconstruct the unknown variable, such as TWS prior to GRACE mission.

In chapter 2, 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. We demarcated river catchments where the usual metrics (NSE, cyclostationary NSE) were of limited interest to evaluate the imbalance error. 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.

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 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.

In chapter 3, we developed an optimization approach to reconstruct TWS prior GRACE era based on the water balance equation. We used a weighting method to try to avoid the aforementioned cancellation of errors. Since the water budget equation involves derivatives of TWS, the reconstruction approach needs to integrate estimates of the derivatives. Due to the frequent biases in the estimates, the integration process leads to accumulation of biases that create artificial linear trends in the TWS reconstruction. Despite these challenges, we found that our reconstruction was coherent when compared to existing reconstructions. However, this comparison showed large uncertainties among TWS reconstructions.

In conclusion, 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 while being aware of the risk of error cancellation. From our comparison of existing TWS reconstructions, one cannot claim to be able to understand the sources of TWS variations. Therefore, further work is needed to obtain a reconstruction of TWS long enough to predict a probable evolution of TWS under climate change scenarios. Alternative validation data are especially crucial to perform this task.

APPENDIX A

ADDITIONAL TABLES

A.1 Additional tables for Chapter 2

Table A.1: 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

The first value is the mean annual cycle averaged over all datasets while the second one is the standard deviation of mean annual values over all datasets

Table A.2: 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

The first value is the mean annual cycle averaged over all datasets while the second one is the standard deviation of mean annual values over all datasets

APPENDIX B

ADDITIONAL FIGURES

B.1 Additional figures for Chapter 2

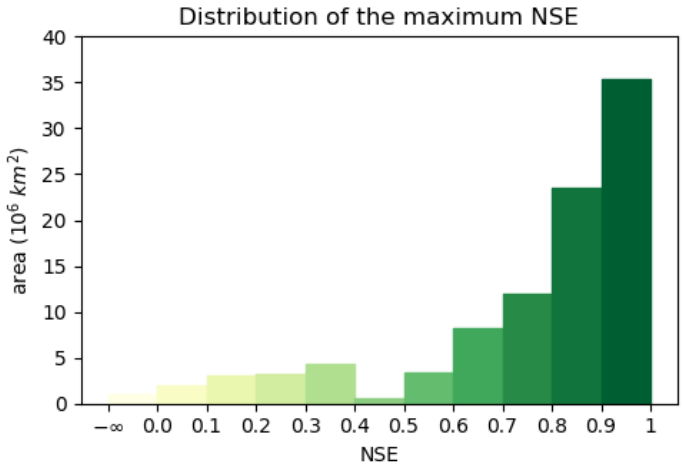


Figure B.1: Distribution of the maximum NSE over all combinations in terms of basin area

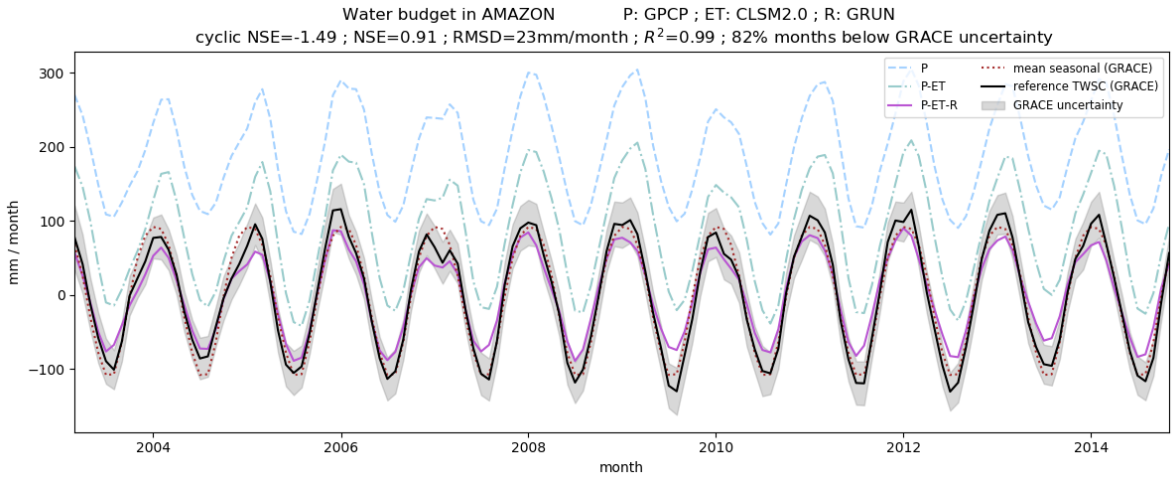


Figure B.2: Components of the water budget in the Amazon basin for the combination leading to the highest NSE

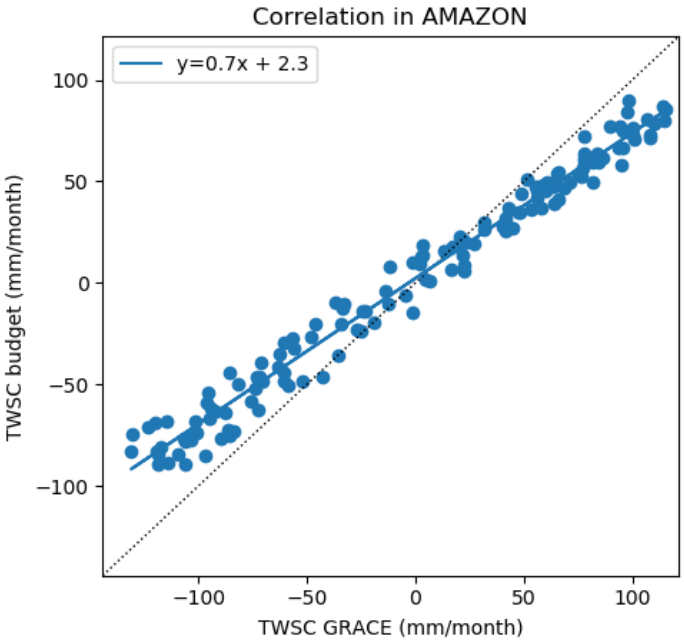


Figure B.3: Correlation between monthly values of GRACE TWSC and the budget reconstruction in the Amazon basin, with the combination leading to the highest NSE (NSE=0.92 and cyclostationary NSE=-1.28)

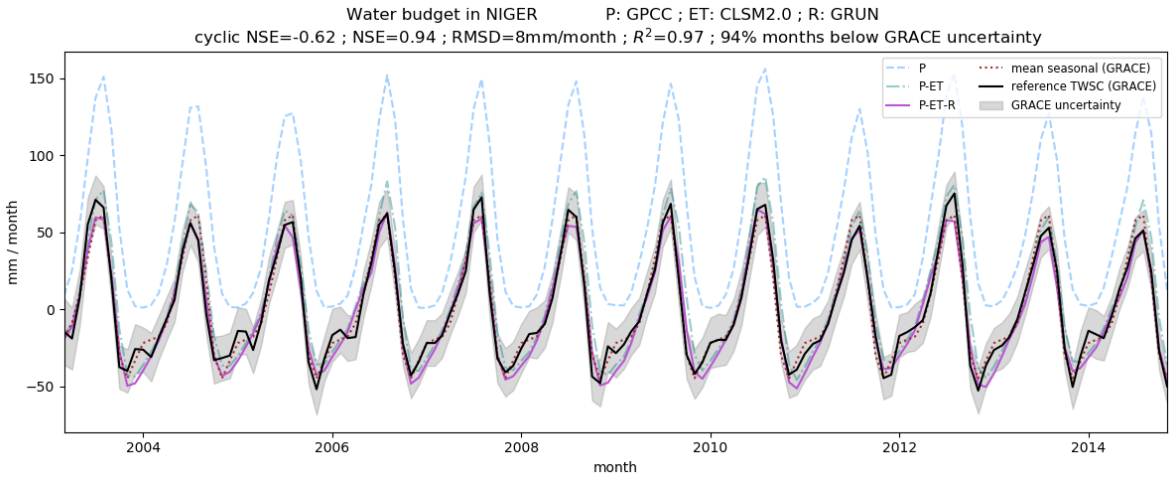


Figure B.4: Components of the water budget in the Niger basin for the combination leading to the highest NSE

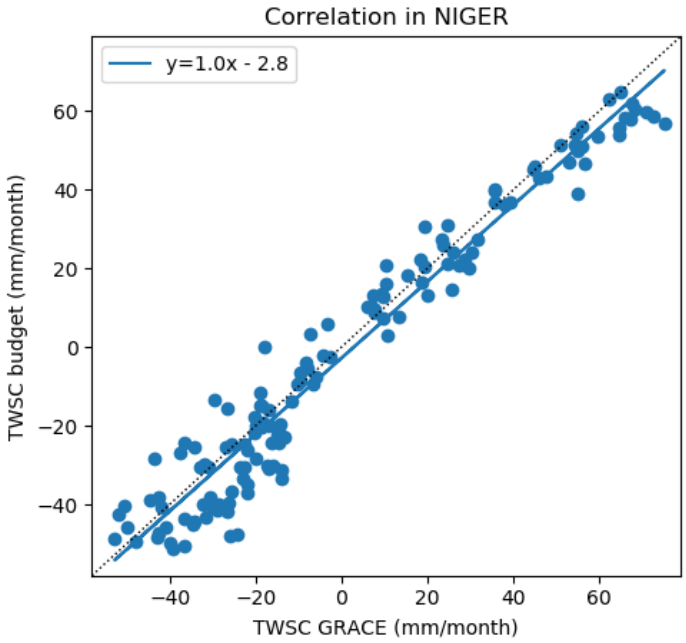


Figure B.5: Correlation between monthly values of GRACE TWSC and the budget reconstruction in the Niger basin, with the combination leading to the highest NSE (NSE=0.94 and cyclostationary NSE=-0.62)

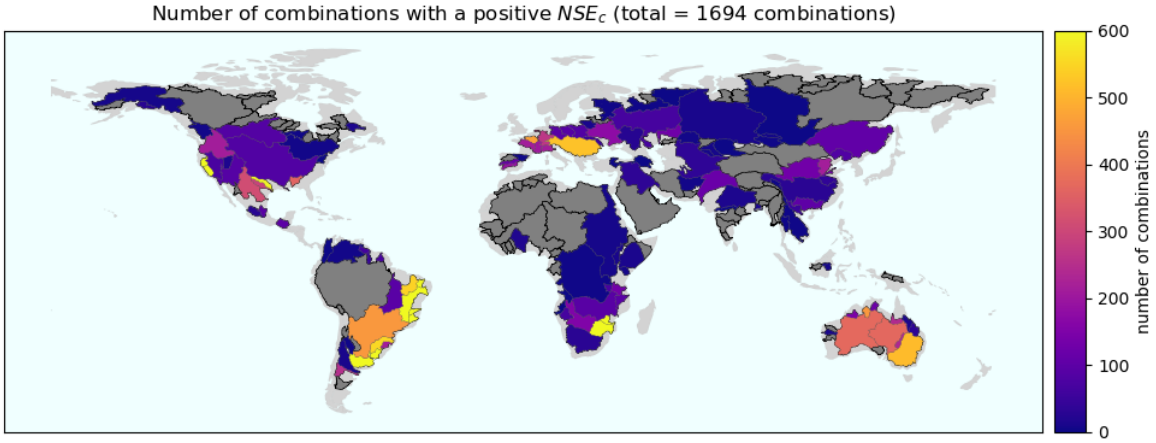


Figure B.6: Number of combinations yielding a positive cyclostationary NSE in each basin. Grey means that no combination achieved a positive value.

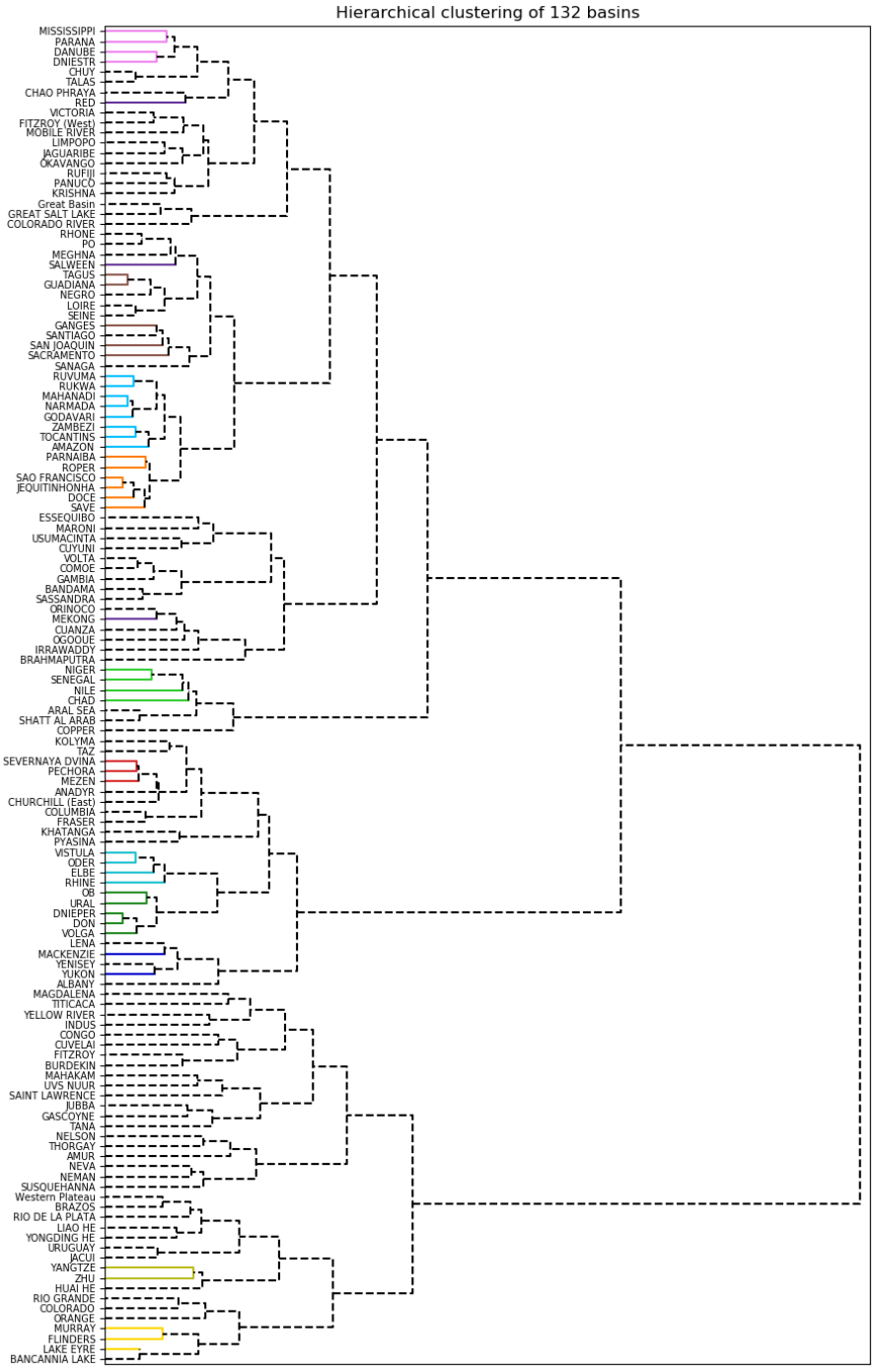


Figure B.7: 132 basins with a maximum NSE larger than 0.8 or a maximum NSE_c larger than 0.1. The distance between basins is the Euclidean distance between the vector of costs for each combination. The height of the U-shaped link is proportional to this distance. Basins are clustered to minimize the intra-cluster variance and colored basins are those selected to plot Fig. 2.9

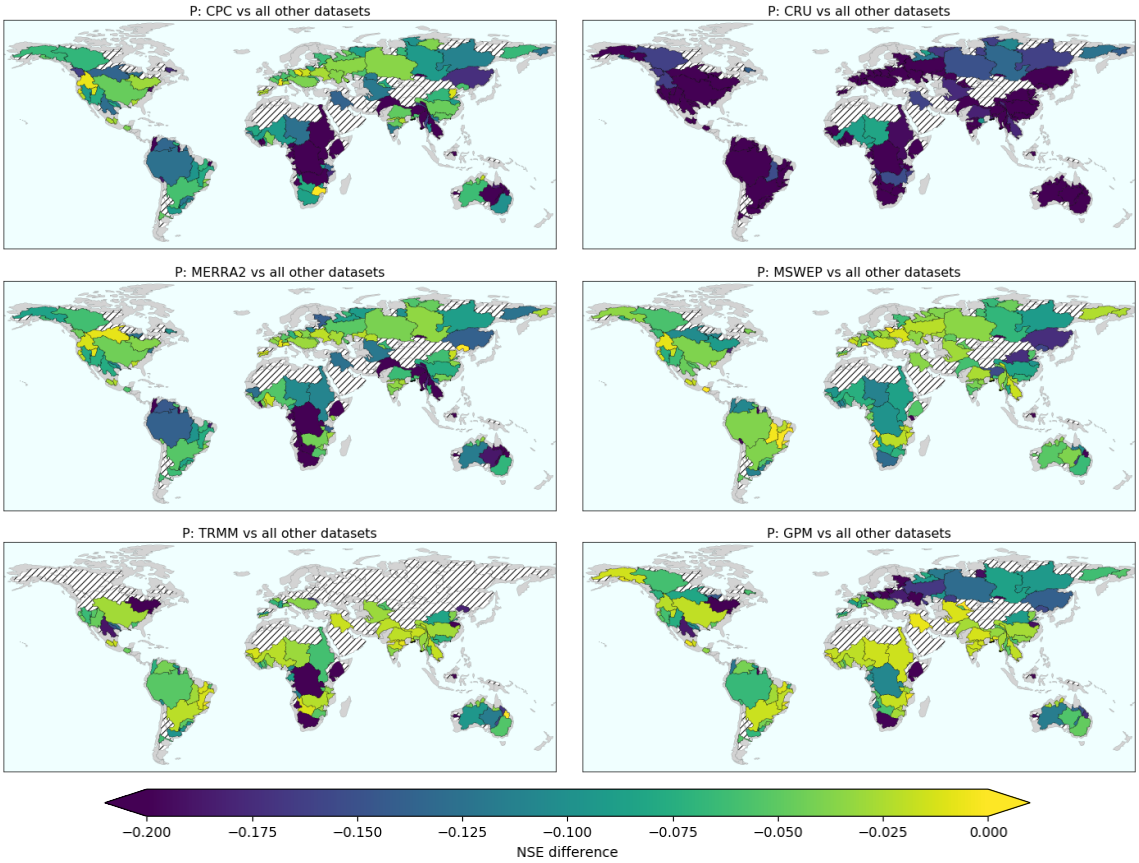


Figure B.8: The mean of the 10th highest NSE with combinations comprising the reference dataset is compared to the mean of the 10th highest NSE excluding the reference dataset. Yellow indicates basins where the reference dataset is similar to or better than others while blues show regions where it was 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).

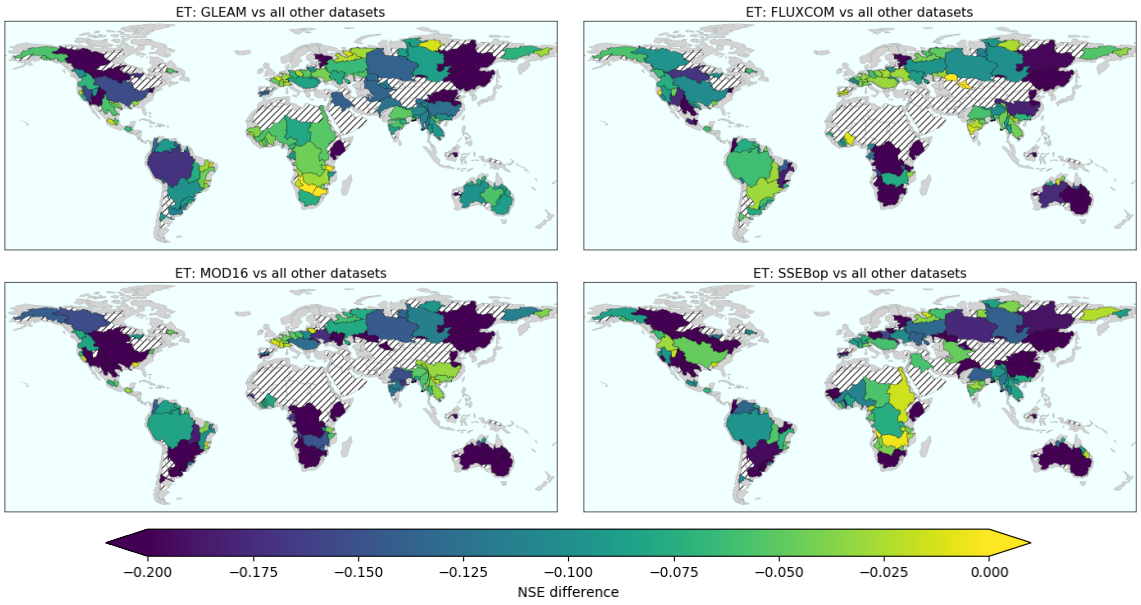


Figure B.9: The mean of the 10th highest NSE with combinations comprising the reference dataset is compared to the mean of the 10th highest NSE excluding the reference dataset. Yellow indicates basins where the reference dataset is similar to or better than others while blues show regions where it was significantly worse. Hatches show basins with a poor water budget closure (maximum NSE lower than 0.8 and maximum NSE_c lower than 0.1).

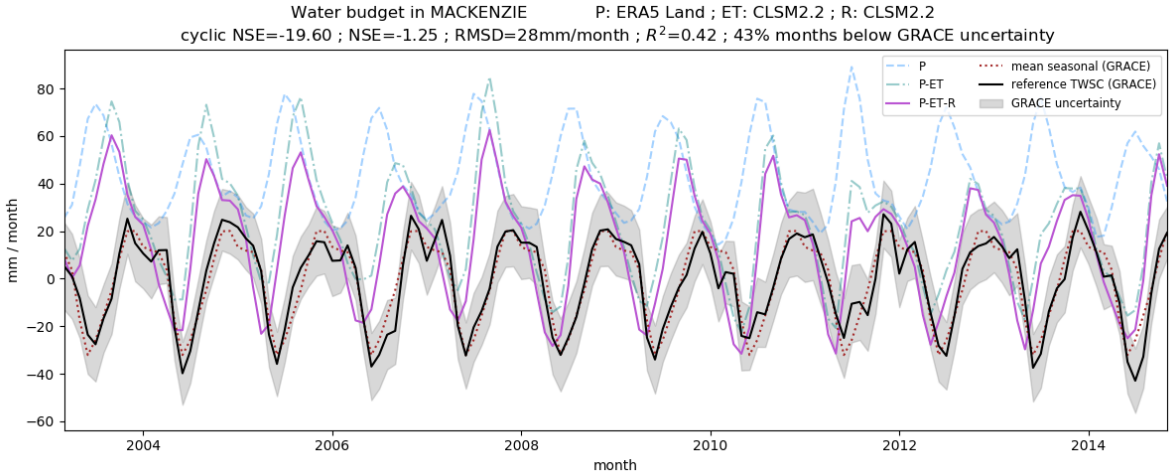


Figure B.10: Components of the water budget in the Mackenzie basin with all components from GLDAS2.2 CLSM (assimilating GRACE TWS)

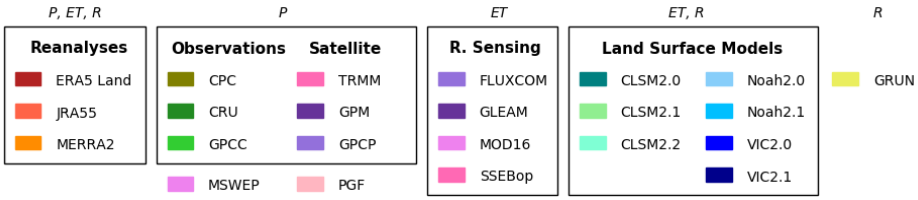


Figure B.11: Legend of datasets represented in Fig. B.12, B.13, B.14, B.15

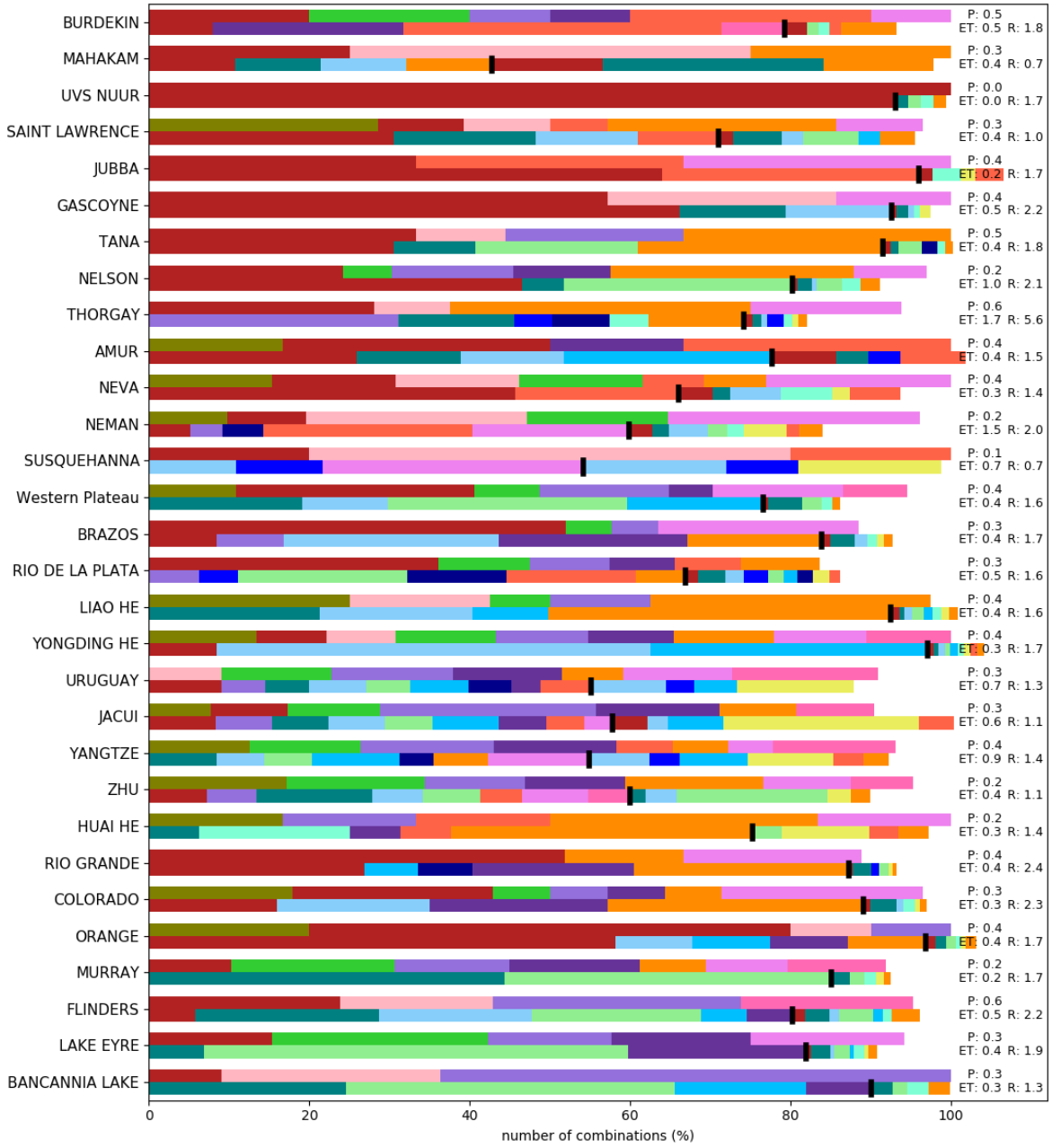


Figure B.12: Datasets appearing in combinations that satisfy a cost lower than 0.1 for each basin separately. The top line of each basin bar represents precipitations datasets. The left part of the bottom line is evapotranspiration datasets while the right part is runoff. The limit between ET and R is symbolized by a black line located proportionally to the portion of ET in the mean annual water cycle of the corresponding region, explaining while the bottom line may have a length different than 100%. Basins are ordered according to hierarchical clustering (dendrogram in Fig. B.7). The color legend for datasets can be found in Fig. B.11

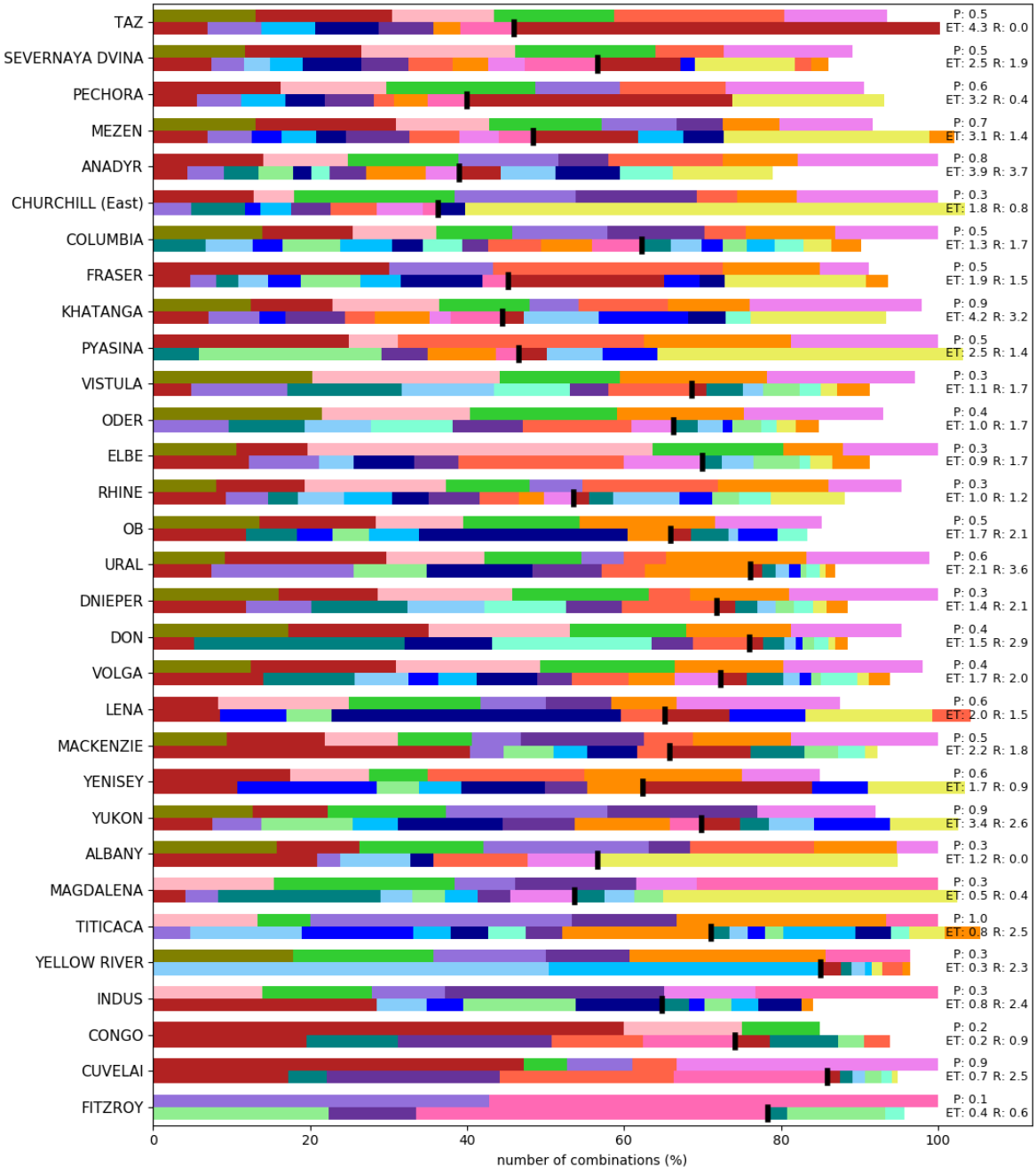


Figure B.13: Following of Fig. B.12

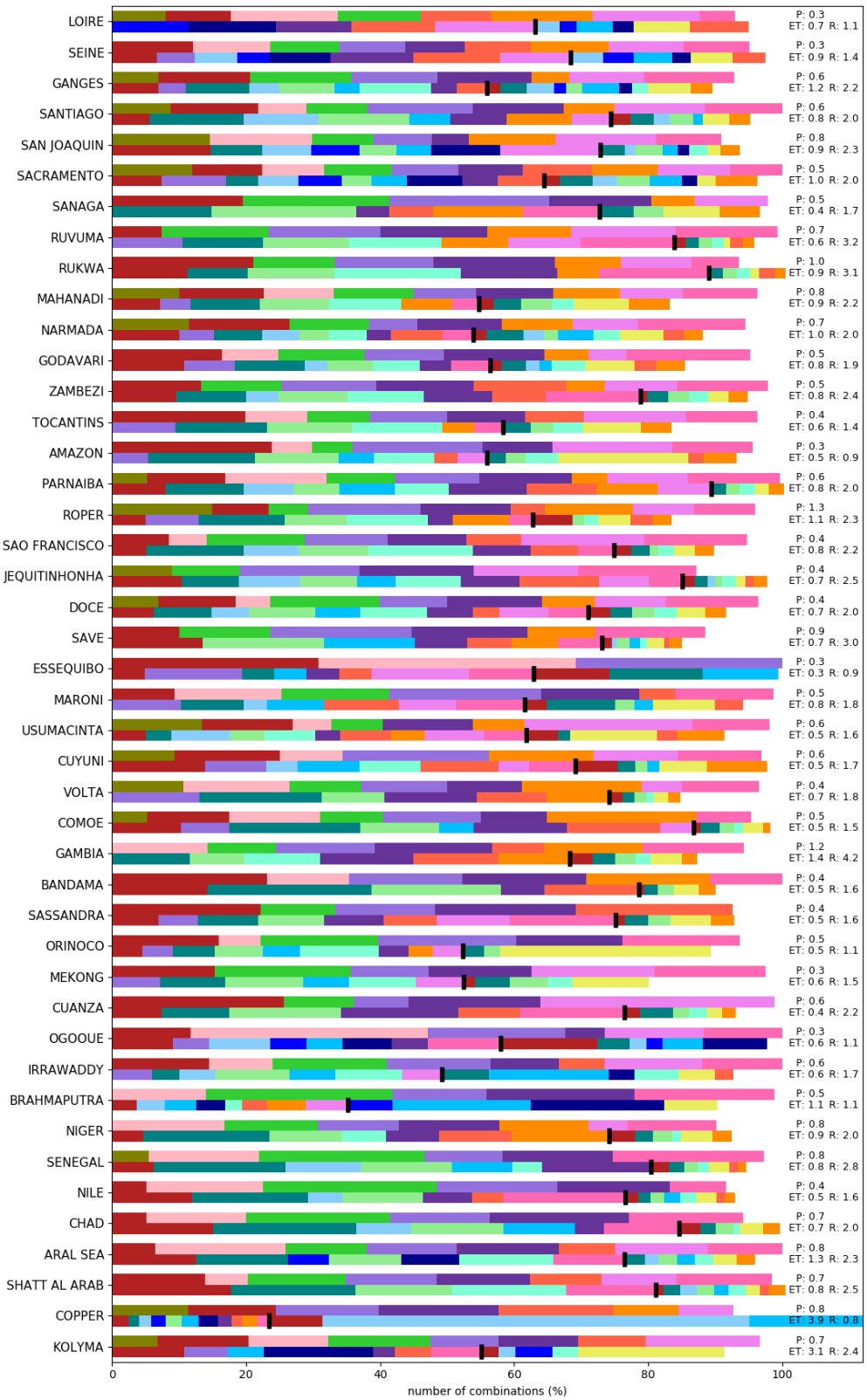


Figure B.14: Following of Fig. B.13

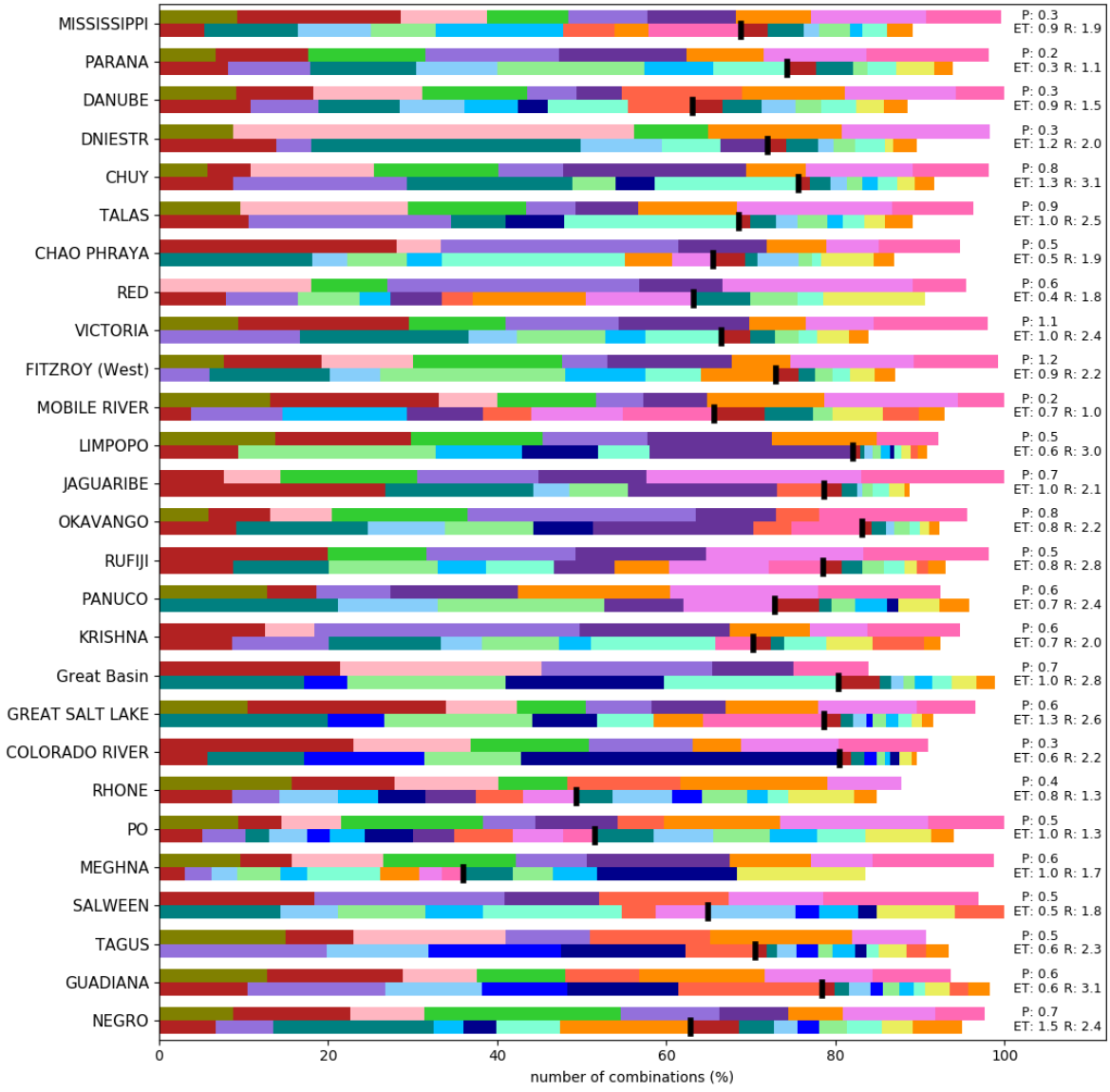


Figure B.15: Following of Fig. B.14

B.2 Additional figures for Chapter 3

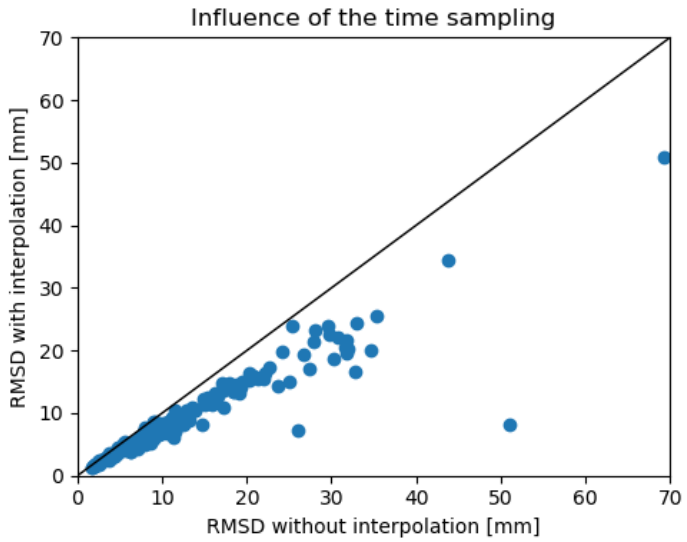


Figure B.16: In each basin, a dot shows the error between TWS_{GRACE} and the solution $s(t)$ of equation 3.1. All other parameters are kept constant.

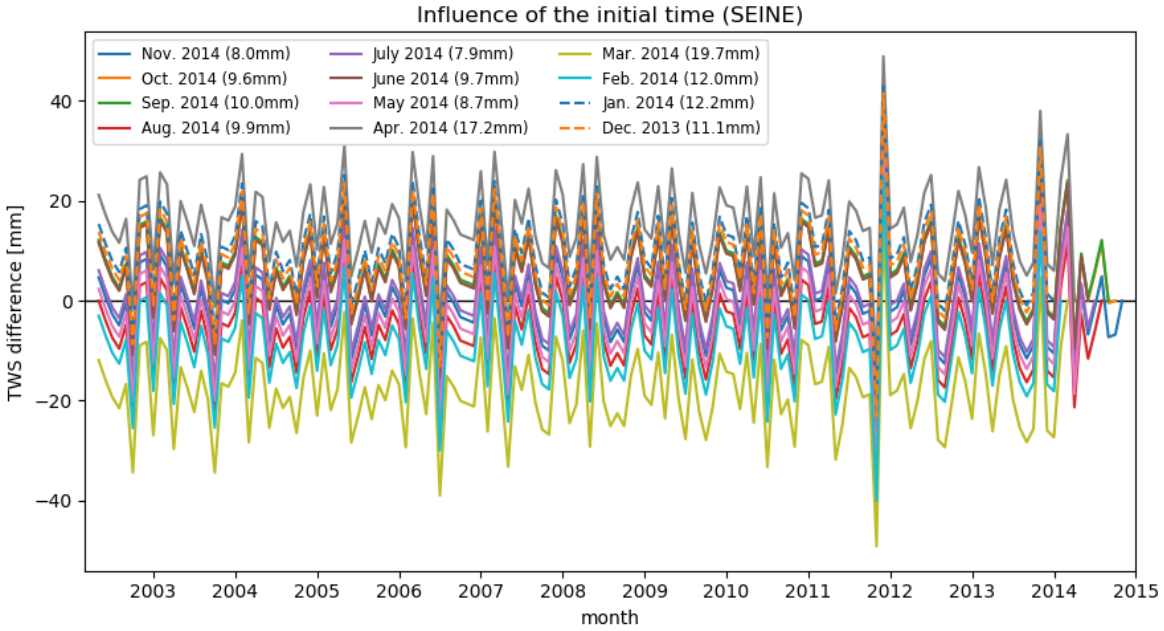


Figure B.17: For 12 different initial times t_0, the difference between TWS_{GRACE} and the solution $s(t)$ of equation 3.1 is shown (with $a(t)$ being the cubic interpolated derivative of TWS_{GRACE}). The RMSD between s and TWS_{GRACE} is given for each initial time.

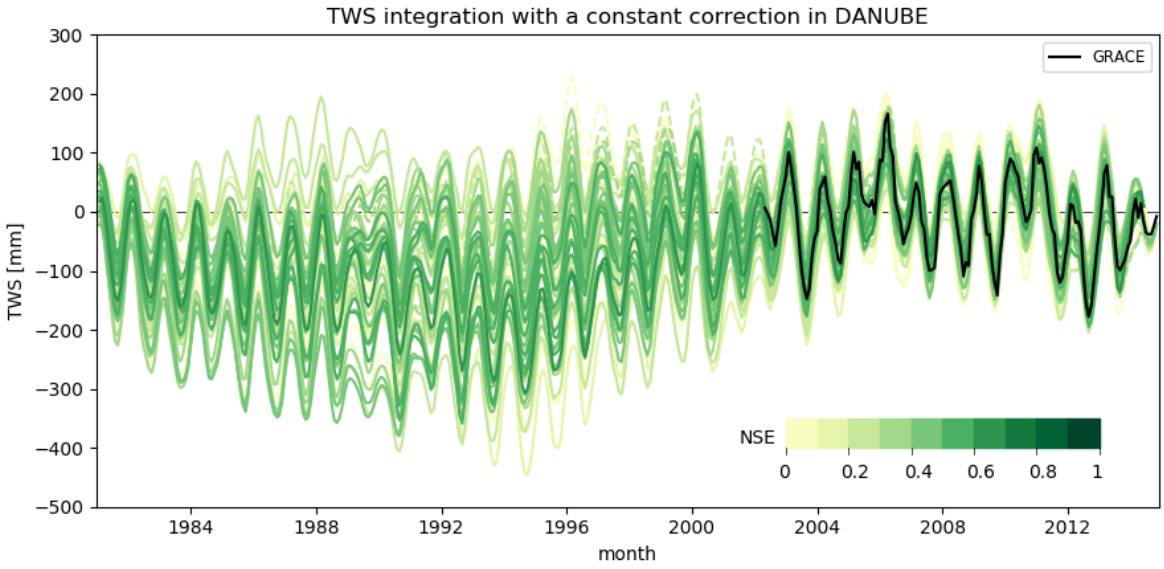


Figure B.18: Integration of TWSC obtained from the water balance equation and corrected by a constant to obtain a null integral over the whole integration period (method from Zeng et al. (2008)). Only combinations with a positive NSE on TWS are shown. The darkness of the line corresponds to the misfit to GRACE.

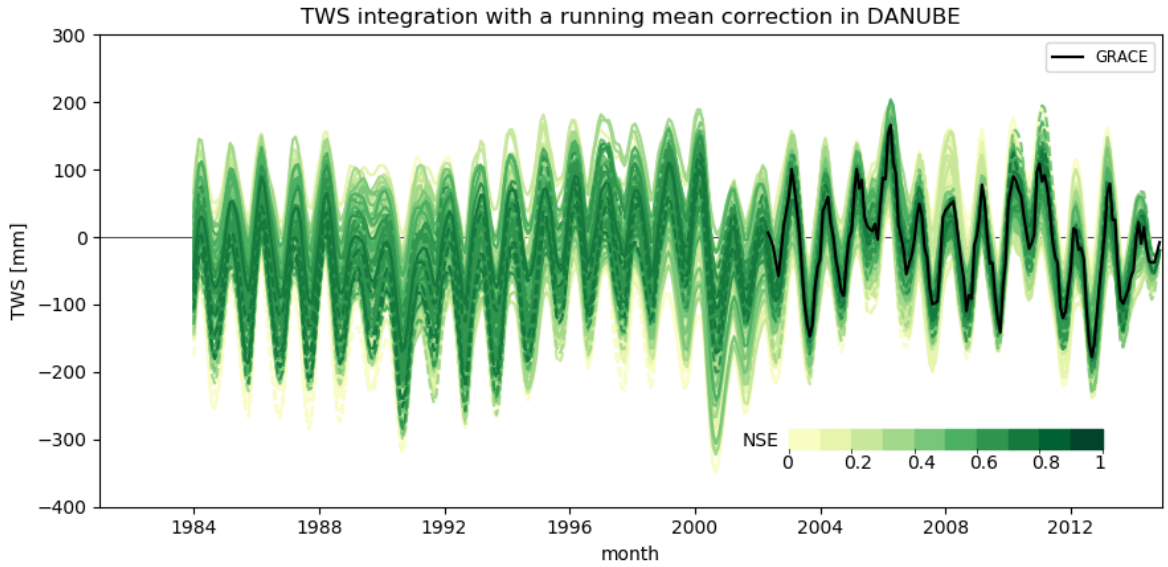


Figure B.19: Integration of TWSC obtained from the water balance equation and corrected by subtracting a 3-year running mean (method from Hirschi and Seneviratne (2017)). Only combinations with a positive NSE on TWS are shown. The darkness of the line corresponds to the misfit to GRACE.

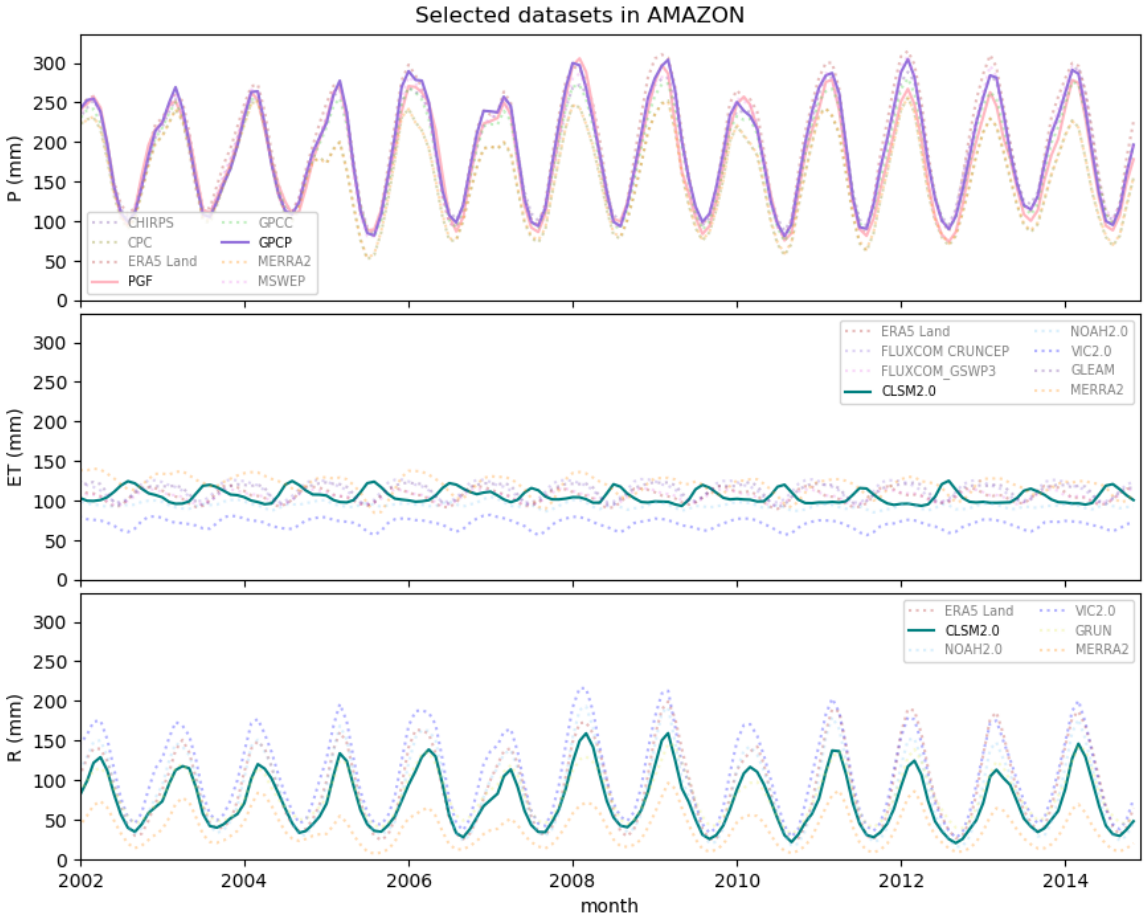


Figure B.20: Timeseries of all datasets in the Amazon basin. Colored lines show datasets with a regression coefficient larger than 0.05 after the optimization algorithm. Datasets with dotted lines have a corresponding coefficient smaller than 0.05 and thus almost did not influence the optimized TWSC.

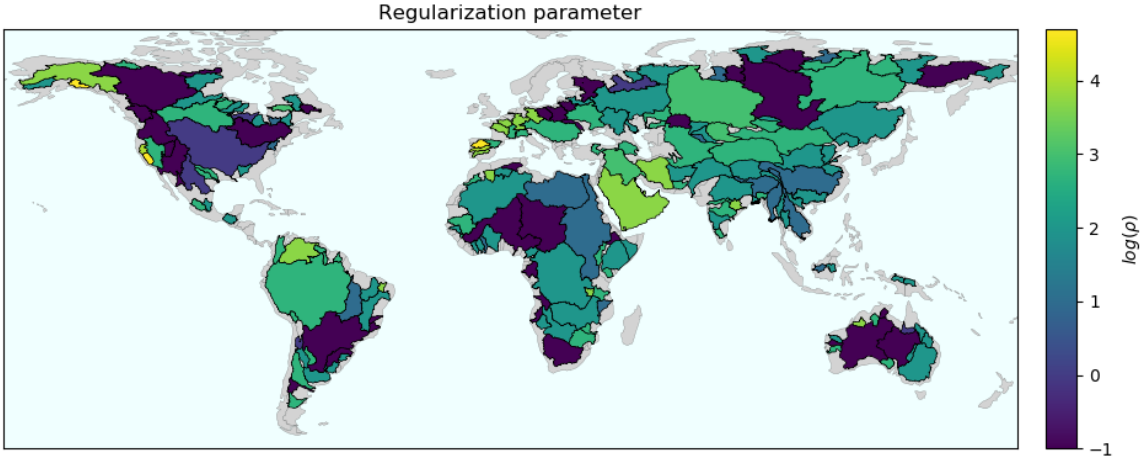


Figure B.21: Regularization parameter ρ chosen with the method described in Section 3.3.3

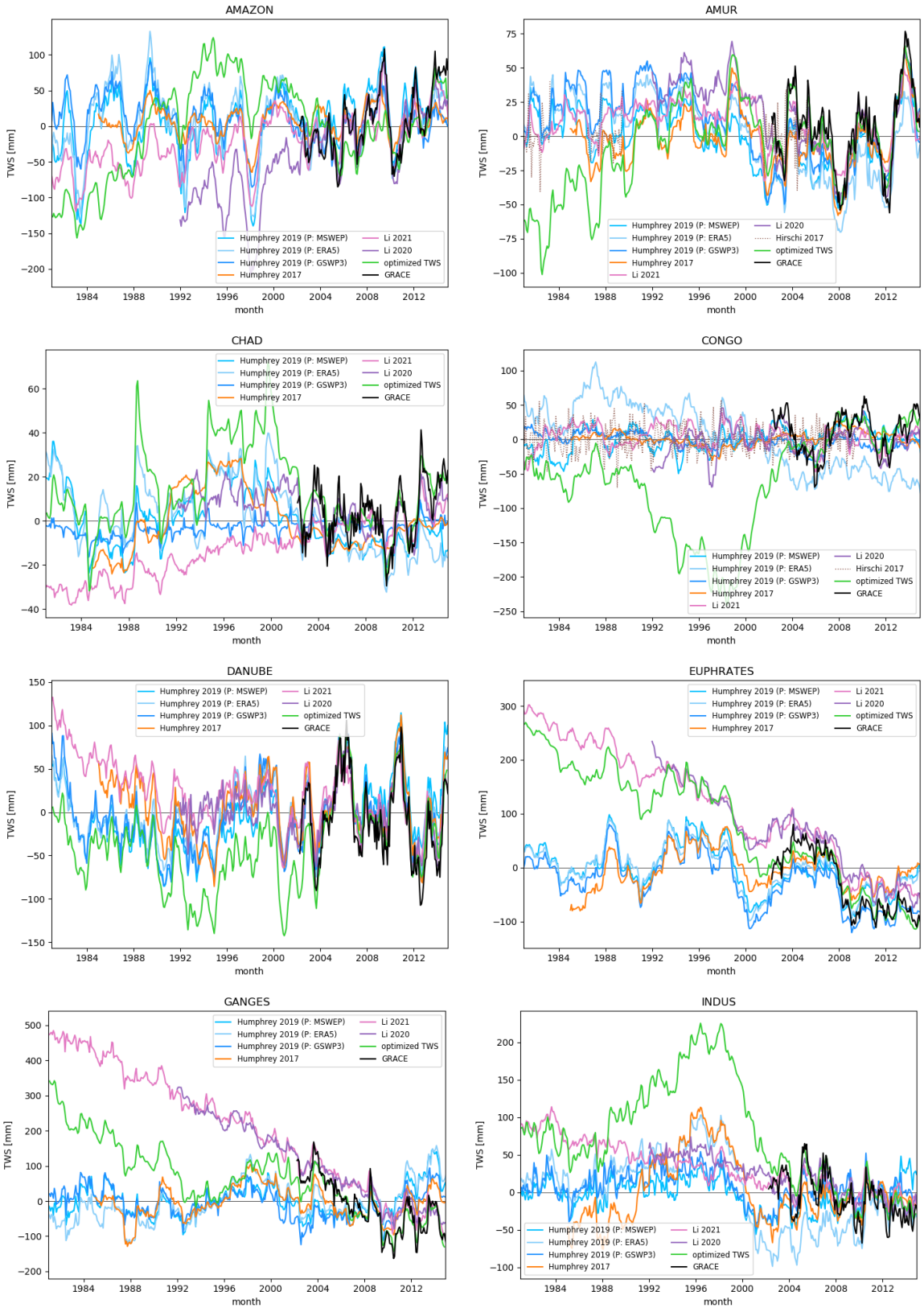


Figure B.22: Comparison of our TWS reconstruction with state-of-the-art datasets ((Humphrey et al., 2017), (Humphrey and Gudmundsson, 2019), (Li et al., 2020b), (Li et al., 2021), (Hirschi and Seneviratne, 2017)) in several basins. All time-series are deseasonalised

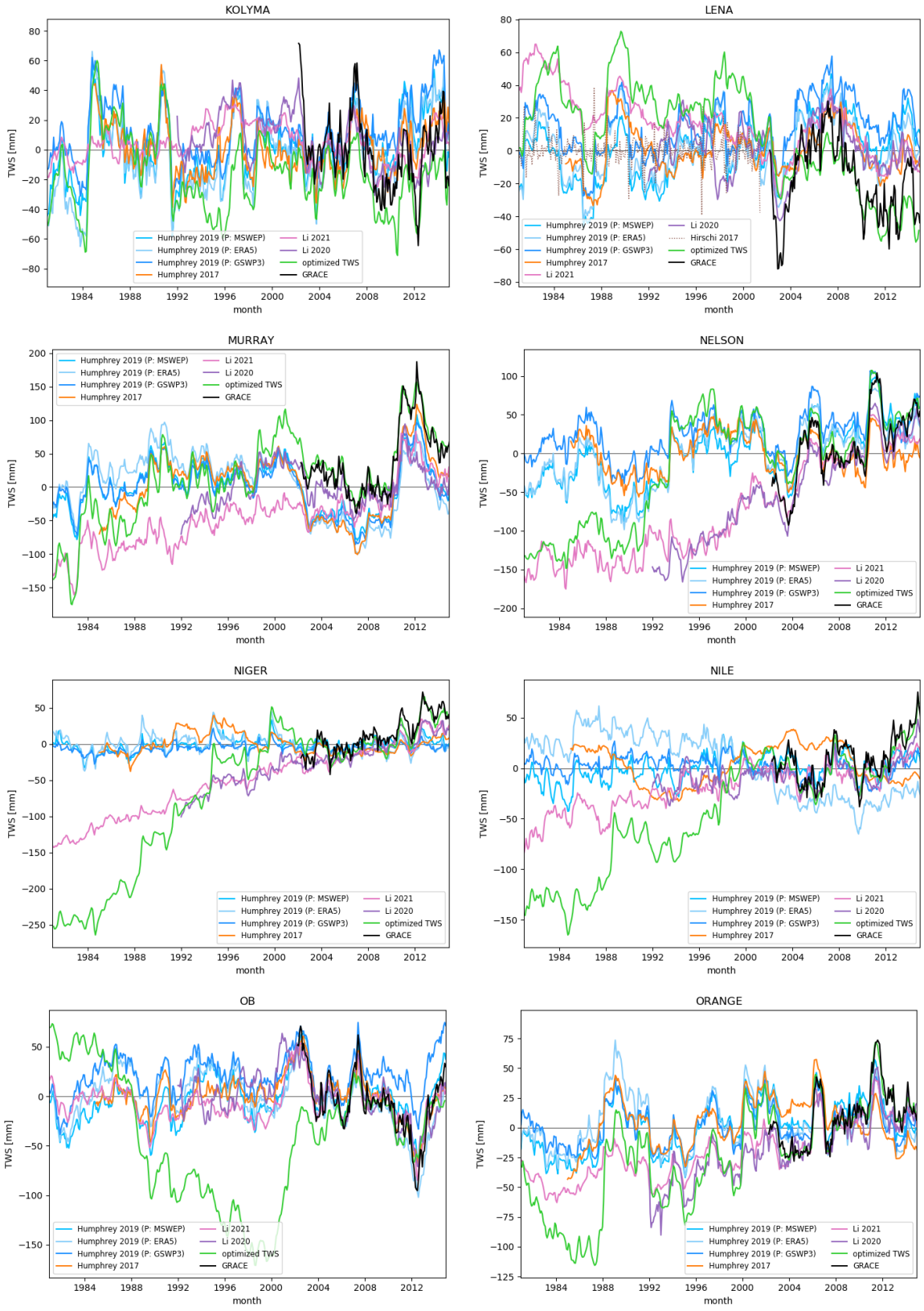


Figure B.23: Following of Fig. B.22

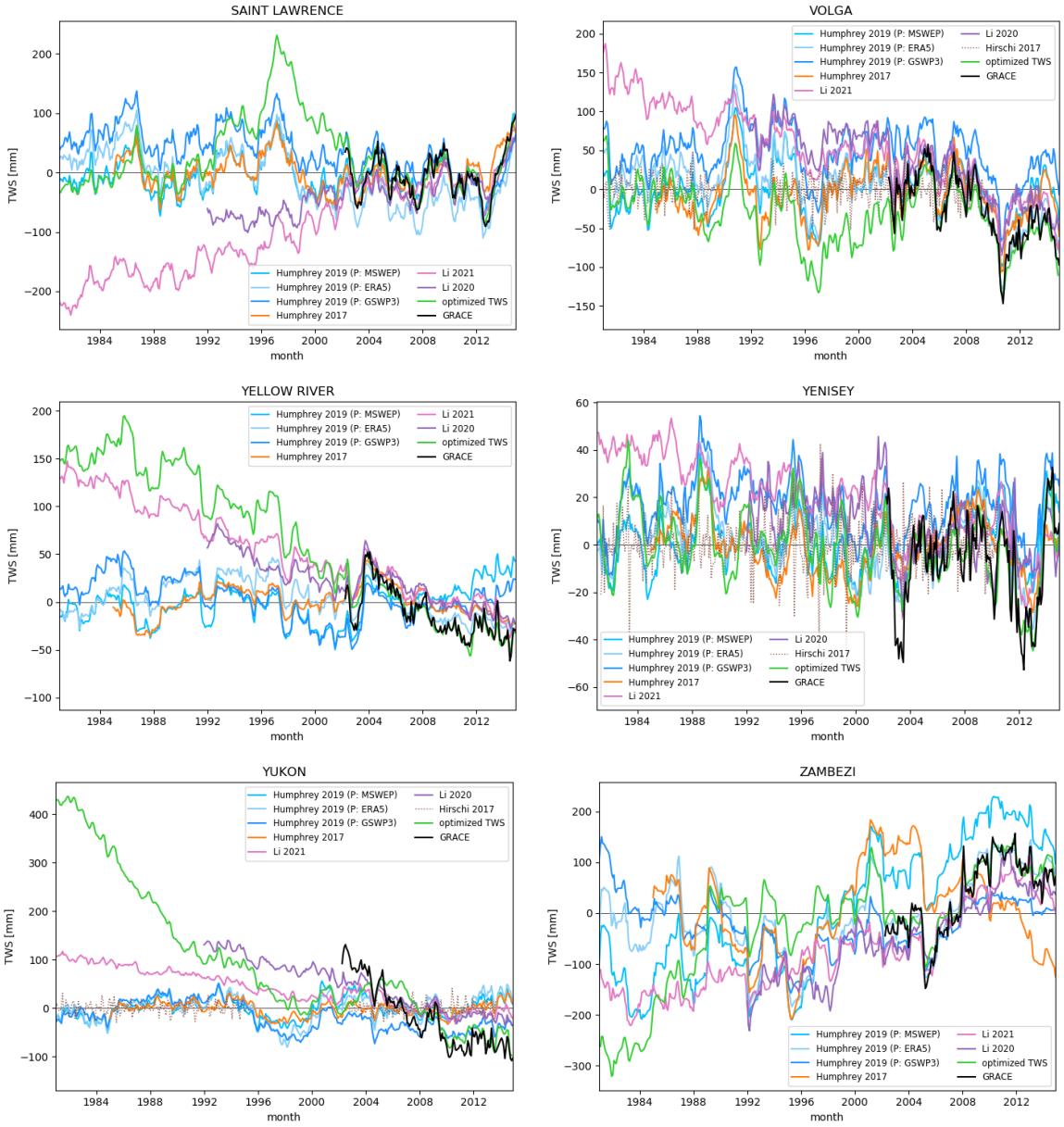


Figure B.24: Following of Fig. B.23

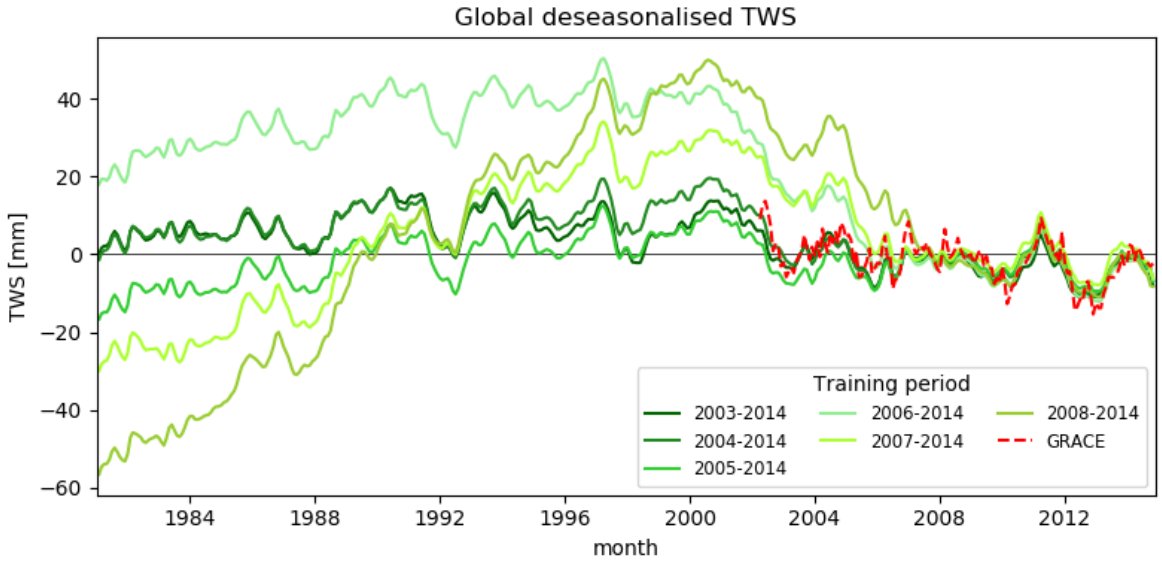


Figure B.25: Global deseasonalised time-series of our TWS reconstruction when varying the training period.

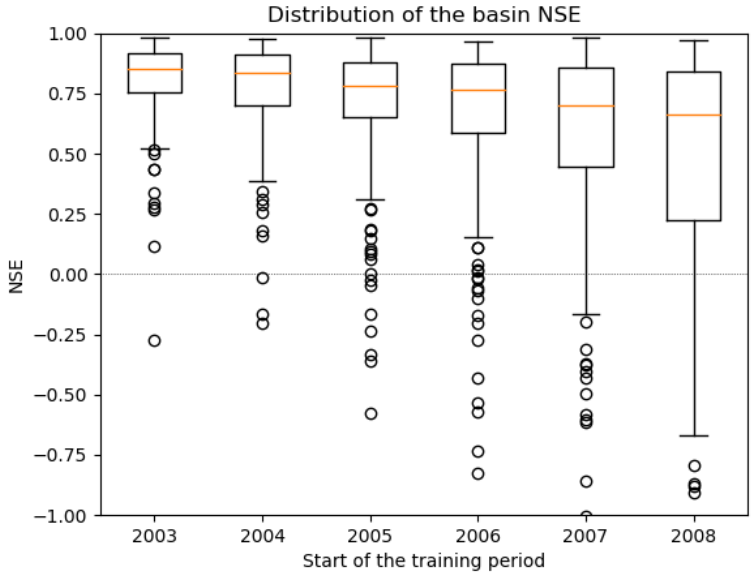


Figure B.26: Each column represents a different training period, starting on the indicated year and ending in 2014. The boxplot shows the NSE against GRACE observations computed from 2003 to 2014 in each basin (median in orange, the box extends from the 25th to the 75th percentile).

BIBLIOGRAPHY

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(4), 2018. doi: 10.3390/atmos9040138.

Agrawal, A., Verschueren, R., Diamond, S., and Boyd, S. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60, 2018. doi: 10.1080/23307706.2017.1397554.

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 (3):853–867, 2014. doi: 10.1002/hyp.9611. Publisher: John Wiley & Sons, Ltd.

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(3):473–500, 2019. doi: 10.1175/BAMS-D-17-0138.1.

Becker, M., Meyssignac, B., Xavier, L., Cazenave, A., Alkama, R., and Decharme, B. Past terrestrial water storage (1980–2008) in the Amazon Basin reconstructed from GRACE and in situ river gauging data. *Hydrology and Earth System Sciences*, 15(2):533–546, 2011. doi: 10.5194/hess-15-533-2011.

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, 2019. doi: 10.1016/j.rse.2019.04.026.

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(10):1141–1158, 2019. doi: 10.1080/02626667.2019.1620507.

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 evaporation by four schemes and comparison with FIFE observations. *Journal of Geophysical Research: Atmospheres*, 101(D3):7251–7268, 1996. doi: 10.1029/95JD02165.

Chen, J., Famigliett, J. S., Scanlon, B. R., and Rodell, M. Groundwater Storage Changes: Present Status from GRACE Observations. *Surveys in Geophysics*, 37(2):397–417, 2016. doi: 10.1007/s10712-015-9332-4.

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*, page 21, 2020. doi: 10.1029/2020WR028032.

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

Cleveland, R. B., Cleveland, W. S., McRae, J. E., and Terpenning, I. STL: A seasonal-trend decomposition procedure based on Loess. *Journal of Official Statistics*, 6(1):3–73, 1990.

Comon, P. Independent component analysis, A new concept? *Signal Processing*, 36(3):287–314, 1994. doi: 10.1016/0165-1684(94)90029-9.

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., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, 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(656): 553–597, 2011. doi: 10.1002/qj.828.

Diamond, S. and Boyd, S. CVXPY: A Python-Embedded Modeling Language for Convex Optimization. *Journal of machine learning research : JMLR*, 17, 2016.

Döll, P., Kaspar, F., and Lehner, B. A global hydrological model for deriving water availability indicators: model tuning and validation. *Journal of Hydrology*, 270(1-2):105–134, 2003. doi: 10.1016/S0022-1694(02)00283-4.

ESA Sea Level CCI project team. Sea Level CCI ECV dataset - 1993-2015 (v2.0), 2017. URL `https://catalogue.ceda.ac.uk/uuid/142052b9dc754f6da47a631e35ec4609`. type: dataset.

Famiglietti, J. S. The global groundwater crisis. *Nature Climate Change*, 4(11):945–948, 2014. doi: 10.1038/nclimate2425.

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(4):2618–2626, 2017. doi: 10.1002/2016WR020175.

Forootan, E., Kusche, J., Loth, I., Schuh, W.-D., Eicker, A., Awange, J., Longuevergne, L., Diekkrüger, B., Schmidt, M., and Shum, C. K. Multivariate Prediction of Total Water Storage Changes Over West Africa from Multi-Satellite Data. *Surveys in Geophysics*, 35(4):913–940, 2014. doi: 10.1007/s10712-014-9292-0.

Frederikse, T., Landerer, F., Caron, L., Adhikari, S., Parkes, D., Humphrey, V. W., Dangendorf, S., Hogarth, P., Zanna, L., Cheng, L., and Wu, Y.-H. The causes of sea-level rise since 1900. *Nature*, 584(7821):393–397, 2020. doi: 10.1038/s41586-020-2591-3.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., and Michaelsen, J. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific Data*, 2(1):150066, 2015. doi: 10.1038/sdata.2015.66.

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(14):3955–3978, 2010. doi: 10.1080/01431161.2010.483488.

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M. G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., da Silva, A. M., Gu, W., Kim, G.-K., Koster, R., Lucchesi, R., Merkova, D., Nielsen, J. E., Partyka, G., Pawson, S., Putman, W., Rienecker, M., Schubert, S. D., Sienkiewicz, M., and Zhao, B. The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *Journal of Climate*, 30(14):5419–5454, 2017. doi: 10.1175/JCLI-D-16-0758.1.

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(4):1655–1674, 2019. doi: 10.5194/essd-11-1655-2019.

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

Hairer, E. E., Wanner, G., and Wanner, G. *Solving ordinary differential equations II*. Springer series in computational mathematics. Springer, Berlin, 2nd rev. ed. edition, 1996. ISBN 3-540-60452-9 978-3-540-60452-5.

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(1):109, 2020. doi: 10.1038/s41597-020-0453-3.

Henriksen, H. J., Troldborg, L., Nyegaard, P., Sonnenborg, T. O., Refsgaard, J. C., and Madsen, B. Methodology for construction, calibration and validation of a national hydrological model for Denmark. *Journal of Hydrology*, 280(1-4):52–71, 2003. doi: 10.1016/S0022-1694(03)00186-0.

Hirschi, M. and Seneviratne, S. I. Basin-scale water-balance dataset (BSWB): an update. *Earth System Science Data*, 9(1):251–258, 2017. doi: 10.5194/essd-9-251-2017.

Humphrey, V., Gudmundsson, L., and Seneviratne, S. I. A global reconstruction of climate-driven subdecadal water storage variability. *Geophysical Research Letters*, 44(5):2300–2309, 2017. doi: 10.1002/2017GL072564.

Humphrey, V. and Gudmundsson, L. GRACE-REC: a reconstruction of climate-driven water storage changes over the last century. *Earth System Science Data*, 11(3):1153–1170, 2019. doi: 10.5194/essd-11-1153-2019.

Humphrey, V., Gudmundsson, L., and Seneviratne, S. I. Assessing Global Water Storage Variability from GRACE: Trends, Seasonal Cycle, Subseasonal Anomalies and Extremes. *Surveys in Geophysics*, 37(2):357–395, 2016. doi: 10.1007/s10712-016-9367-1.

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

Jasechko, S. and Perrone, D. Global groundwater wells at risk of running dry. *Science*, 372(6540):418–421, 2021. doi: 10.1126/science.abc2755.

Jing, W., Di, L., Zhao, X., Yao, L., Xia, X., Liu, Y., Yang, J., Li, Y., and Zhou, C. A data-driven approach to generate past GRACE-like terrestrial water storage solution by calibrating the land surface model simulations. *Advances in Water Resources*, 143:103683, 2020. doi: 10.1016/j.advwatres.2020.103683.

Jung, M., Koirala, S., Weber, U., Ichii, K., Gans, F., Camps-Valls, G., Papale, D., Schwalm, C., Tramontana, G., and Reichstein, M. The FLUXCOM ensemble of global land-atmosphere energy fluxes. *Scientific Data*, 6(1):74, 2019. doi: 10.1038/s41597-019-0076-8.

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(D20):24809–24822, 2000. doi: 10.1029/2000JD900327.

Kottek, M., Grieser, J., Beck, C., Rudolf, B., and Rubel, F. World Map of the Köppen-Geiger climate classification updated. *Meteorologische Zeitschrift*, 15(3):259–263, 2006. doi: 10.1127/0941-2948/2006/0130.

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):D23115, 2010. doi: 10.1029/2010JD014584.

Li, B., Rodell, M., Kumar, S., Beaudoing, H. K., Getirana, A., Zaitchik, B. F., Goncalves, L. G., Cossetin, C., Bhanja, S., Mukherjee, A., 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(9):7564–7586, 2019. doi: 10.1029/2018WR024618.

Li, B., Beaudoing, H., Rodell, M., and NASA/GSFC/HSL. GLDAS Catchment Land Surface Model L4 monthly 1.0 x 1.0 degree V2.0, 2020a. URL `https://disc.gsfc.nasa.gov/datacollection/GLDAS_CLSM10_M_2.0.html`. type: dataset.

Li, F., Kusche, J., Rietbroek, R., Wang, Z., Forootan, E., Schulze, K., and Lück, C. Comparison of Data-Driven Techniques to Reconstruct (1992–2002) and Predict (2017–2018) GRACE-Like Gridded Total Water Storage Changes Using Climate Inputs. *Water Resources Research*, 56(5), 2020b. doi: 10.1029/2019WR026551.

Li, F., Kusche, J., Chao, N., Wang, Z., and Löcher, A. Long-Term (1979-Present) Total Water Storage Anomalies Over the Global Land Derived by Reconstructing GRACE Data. *Geophysical Research Letters*, 48(8), 2021. doi: 10.1029/2021GL093492.

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(D7):14415, 1994. doi: 10.1029/94JD00483.

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, 2016. doi: 10.1016/j.jhydrol.2016.04.006.

Ljung, L. System Identification: Theory for the User. 1987.

Long, D., Longuevergne, L., and Scanlon, B. R. Uncertainty in evapotranspiration from land surface modeling, remote sensing, and GRACE satellites. *Water Resources Research*, 50(2): 1131–1151, 2014a. doi: 10.1002/2013WR014581.

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

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, 2015. doi: 10.1016/j.rse.2015.07.003.

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(11), 2010. doi: https://doi.org/10.1029/2009WR008564.

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(6):2111–2139, 2014. doi: 10.1175/JHM-D-13-0157.1.

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(10):8450–8475, 2015. doi: 10.1002/2014WR016794.

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, 2017. doi: 10.1016/j.jhydrol.2017.02.027.

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(5):1903–1925, 2017. doi: 10.5194/gmd-10-1903-2017.

Milly, P. C. D. C., Cazenave, A., Famiglietti, J. S., Gornitz, V., Laval, K., Lettenmaier, D. P., Sahagian, D. L., Wahr, J. M., and Wilson, C. R. Terrestrial Water-Storage Contributions to Sea-Level Rise and Variability. In Church, J. A., Woodworth, P. L., Aarup, T., and Wilson, W. S., editors, *Understanding Sea-Level Rise and Variability*, pages 226–255. Wiley-Blackwell, Oxford, UK, 2010. ISBN 978-1-4443-2327-6 978-1-4443-3451-7. doi: 10.1002/9781444323276.ch8. URL `http://doi.wiley.com/10.1002/9781444323276.ch8`.

Miralles, D. G., Holmes, T. R. H., De Jeu, R. A. M., Gash, J. H., Meesters, A. G. C. A., and Dolman, A. J. Global land-surface evaporation estimated from satellite-based observations. *Hydrology and Earth System Sciences*, 15(2):453–469, 2011. doi: 10.5194/hess-15-453-2011.

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(6), 2011. doi: 10.1029/2010GL046230.

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

Nash, J. and Sutcliffe, J. River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, 10(3):282–290, 1970. doi: 10.1016/0022-1694(70)90255-6.

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

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(9):7100–7114, 2014. doi: 10.1002/2013WR015202.

on Application of Artificial Neural Networks in Hydrology, A. T. C. Artificial Neural Networks in Hydrology. I: Preliminary Concepts. *Journal of Hydrologic Engineering*, 5(2):115–123, 2000. doi: 10.1061/(ASCE)1084-0699(2000)5:2(115).

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(9):3191–3206, 2012. doi: 10.1175/JCLI-D-11-00300.1.

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(2):e2019WR026594, 2020. doi: 10.1029/2019WR026594. Publisher: John Wiley & Sons, Ltd.

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, 2015. doi: 10.1016/j.rse.2015.08.031.

Phillips, T., Nerem, R. S., Fox-Kemper, B., Famiglietti, J. S., and Rajagopalan, B. The influence of ENSO on global terrestrial water storage using GRACE: USING GRACE TO DETECT ENSO SIGNAL. *Geophysical Research Letters*, 39(16):n/a–n/a, 2012. doi: 10.1029/2012GL052495.

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(24):6322–6338, 2011. doi: 10.1175/JCLI-D-10-05033.1.

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(9): 2705–2723, 1999. doi: 10.1029/1999WR900141.

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(21):8289–8318, 2015. doi: 10.1175/JCLI-D-14-00555.1.

Rodell, M., Famiglietti, J. S., Wiese, D. N., Reager, J. T., Beaudoing, H. K., Landerer, F. W., and Lo, M.-H. Emerging trends in global freshwater availability. *Nature*, 557(7707):651–659, 2018. doi: 10.1038/s41586-018-0123-1.

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*, 739:139857, 2020. doi: 10.1016/j.scitotenv.2020.139857.

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(8):1850–1865, 2011. doi: 10.1016/j.rse.2011.03.009.

Samuelsen, A., Hansen, C., and Wehde, H. Tuning and assessment of the HYCOM-NORWECOM V2.1 biogeochemical modeling system for the North Atlantic and Arctic oceans. *Geoscientific Model Development*, 8(7):2187–2202, 2015. doi: 10.5194/gmd-8-2187-2015.

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(6):E1080–E1089, 2018. doi: 10.1073/pnas.1704665115.

Schellekens, J., Dutra, E., Martínez-de la Torre, A., Balsamo, G., van Dijk, A., Sperna Weiland, F., Minvielle, M., Calvet, J.-C., Decharme, B., Eisner, S., Fink, G., Flörke, M., Peßenteiner, S., van Beek, R., Polcher, J., Beck, H., Orth, R., Calton, B., Burke, S., Dorigo, W., and Weedon, G. P. A global water resources ensemble of hydrological models: the eartH2Observe Tier-1 dataset. *Earth System Science Data*, 9(2):389–413, 2017. doi: 10.5194/essd-9-389-2017.

Schneider, U., Becker, A., Finger, P., Rustemeier, E., and Ziese, M. GPCC Full Data Monthly Version 2020 at 0.5°, 2020. URL `https://opendata.dwd.de/climate_environment/GPCC/html/fulldata-monthly_v2020_doi_download.html`.

SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Sen, P. K. Estimates of the Regression Coefficient Based on Kendall's Tau. *Journal of the American Statistical Association*, 63(324):1379–1389, 1968. doi: 10.1080/01621459.1968.10480934.

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(13):3088–3111, 2006. doi: 10.1175/JCLI3790.1. Place: Boston MA, USA Publisher: American Meteorological Society.

Sheffield, J., Ferguson, C. R., Troy, T. J., Wood, E. F., and McCabe, M. F. Closing the terrestrial water budget from satellite remote sensing. *Geophysical Research Letters*, 36(7), 2009. doi: 10.1029/2009GL037338. Publisher: John Wiley & Sons, Ltd.

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(6):1333–1359, 2014. doi: 10.1007/s10712-014-9300-4.

Stephens, G. L., Slingo, J. M., Rignot, E., Reager, J. T., Hakuba, M. Z., Durack, P. J., Worden, J., and Rocca, R. Earth's water reservoirs in a changing climate. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 476(2236):20190458, 2020. doi: 10.1098/rspa.2019.0458.

Sun, A. Y., Scanlon, B. R., Zhang, Z., Walling, D., Bhanja, S. N., Mukherjee, A., and Zhong, Z. Combining Physically Based Modeling and Deep Learning for Fusing GRACE Satellite Data: Can We Learn From Mismatch? *Water Resources Research*, 55(2):1179–1195, 2019. doi: 10.1029/2018WR023333.

Sutanudjaja, E. H., van Beek, R., Wanders, N., Wada, Y., Bosmans, J. H. C., Drost, N., van der Ent, R. J., de Graaf, I. E. M., Hoch, J. M., de Jong, K., Karssenberg, D., López López, P., Peßenteiner, S., Schmitz, O., Straatsma, M. W., Vannametee, E., Wisser, D., and Bierkens, M. F. P. PCR-GLOBWB 2: a 5 arcmin global hydrological and water resources model. *Geoscientific Model Development*, 11(6):2429–2453, 2018. doi: 10.5194/gmd-11-2429-2018.

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(8):2173–2185, 2017. doi: 10.1175/JHM-D-17-0004.1. Place: Boston MA, USA Publisher: American Meteorological Society.

Tapley, B. D. GRACE Measurements of Mass Variability in the Earth System. *Science*, 305(5683): 503–505, 2004. doi: 10.1126/science.1099192.

Theil, H. A Rank-Invariant Method of Linear and Polynomial Regression Analysis. *Proceedings of the Royal Netherlands Academy of Sciences*, 1950. doi: 10.1007/978-94-011-2546-8_20. Series Title: Advanced Studies in Theoretical and Applied Econometrics.

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, 2017. doi: 10.1016/j.jhydrol.2017.01.009.

Vishwakarma, B. D., Bates, P., Sneeuw, N., Westaway, R. M., and Bamber, J. L. Re-assessing global water storage trends from GRACE time series. *Environmental Research Letters*, 16(3): 034005, 2021. doi: 10.1088/1748-9326/abd4a9.

Vishwakarma, B., Devaraju, B., and Sneeuw, N. What Is the Spatial Resolution of GRACE Satellite Products for Hydrology? *Remote Sensing*, 10(6):852, 2018. doi: 10.3390/rs10060852.

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(B12):30205–30229, 1998. doi: 10.1029/98JB02844.

Wahr, J., Swenson, S., and Velicogna, I. Accuracy of GRACE mass estimates. *Geophysical Research Letters*, 33(6):L06401, 2006. doi: 10.1029/2005GL025305.

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(8):6485–6499, 2015. doi: 10.1002/2015WR017311.

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, 2014. doi: 10.1016/j.jhydrol.2014.01.076.

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(4):2648–2671, 2015. doi: 10.1002/2014JB011547.

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, 2018. URL `https://podaac.jpl.nasa.gov/dataset/TELLUS_GRACE_MASCON_CRI_GRID_RL06_V1`. type: dataset.

Wold, S., Esbensen, K., and Geladi, P. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3):37–52, 1987. doi: 10.1016/0169-7439(87)80084-9.

Wolter, K. and Timlin, M. S. Monitoring ENSO in COADS with a seasonally adjusted principal component index. pages 52–57, 1993.

Wolter, K. and Timlin, M. S. Measuring the strength of ENSO events: How does 1997/98 rank? *Weather*, 53(9):315–324, 1998. doi: 10.1002/j.1477-8696.1998.tb06408.x. Publisher: John Wiley & Sons, Ltd.

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(14):7617–7632, 2019. doi: 10.1029/2018JD030025.

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(6), 2010. doi: 10.1029/2009WR007811.

Zeng, N., Yoon, J.-H., Mariotti, A., and Swenson, S. Variability of Basin-Scale Terrestrial Water Storage from a PER Water Budget Method: The Amazon and the Mississippi. *Journal of Climate*, 21(2):248–265, 2008. doi: 10.1175/2007JCLI1639.1.

Zhang, D., Zhang, Q., Werner, A. D., and Liu, X. GRACE-Based Hydrological Drought Evaluation of the Yangtze River Basin, China. *Journal of Hydrometeorology*, 17(3):811–828, 2016. doi: 10.1175/JHM-D-15-0084.1.

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

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., Bytheway, 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(1):241–263, 2018. doi: 10.5194/hess-22-241-2018.