Improved observation of the global water cycle with satellite remote sensing and neural network modeling

Une thèse présentée pour l'obtention du grade de Docteur

Sorbonne Université
École Doctorale des Sciences de l'Environnement d'Île de France (N° 129)

par

Matthew G. Heberger

Laboratoire d'Etudes du Rayonnement et de la Matière en Astrophysique et Atmosphères, UMR 8112
Observatoire de Paris

Présentée et soutenue publiquement le 12 janvier 2024
devant un jury composé de:

Hélène CHEPFER	Sorbonne University	Présidente du jury
Aaron BOONE	Météo France, Toulouse	Rapporteur
Frédéric FRAPPART	INRAE, Villenave d'Ornon	Rapporteur
Hélène BROGNIEZ	Paris Saclay University	Examinatrice
Ming PAN	Univ. of California at San Diego	Examinateur
Fabrice PAPA	IRD, Brasilia, Brazil	Examinateur
Filipe AIRES	LERMA/CNRS, Paris	Directeur de thèse

Acknowledgements

I had one of the the best advisors a doctoral candidate could hope for. Thank you, Filipe Aires, for having confidence in a “non-traditional student” and for your constant support and guidance. Thank you also to Victor Pellet for helpful conversations, coding advice, and careful reading of manuscripts.

I offer my sincere gratitude to the members of my thesis committee (*jury*) for your time, energy, and expertise : Hélène Chepfer, Aaron Boone, Frédéric Frappart, Hélène Brogniez, Ming Pan, and Fabrice Papa. To Drs. Pan and Papa, thank you for accompanying me since year one as part of my *Comité de Suivi*, and for your constructive input and encouragement.

I want to express gratitude to the Government of France and to the administration of President Emmanuel Macron for supporting climate research and for welcoming international researchers to France when they felt less than welcome in their home countries.

It has been an extraordinary honor and privilege to be a student at Sorbonne University and to do research at the Paris Observatory, both renowned institutions.

I humbly thank the many people that make these institutions run and who are too frequently unacknowledged – technicians, administrators, janitors, groundskeepers, cooks in the *cantine*, and many more. We could not do science without you.

Finally, my dearest thanks to my family, Michelle and Gabriel, for your unconditional love, support, and patience.

Declaration

This research was carried out under the direction of Dr. Filipe Aires from 2021 to 2023 at the Paris Observatory, within the *Laboratoire d'Etudes du Rayonnement et de la Matière en Astrophysique et Atmosphères*, or LERMA.

The supervision of the thesis was done by the Doctoral School of Environmental Sciences of Ile-de-France #129 at Sorbonne University.

Portions of the research described herein were funded by:

- European Space Agency contract number #4000136793/21/I-DT-lr
- ESTELLUS

Résumé en français

Titre: Amélioration de l'observation du cycle de l'eau à l'échelle globale grâce à la télédétection par satellite et à la modélisation par réseaux de neurones

Résumé : La télédétection par satellite est couramment utilisée pour observer le cycle hydrologique à des échelles spatiales allant des bassins fluviaux au globe terrestre. Pourtant, il reste difficile d'obtenir un bilan hydrique équilibré en utilisant des données de télédétection, ce qui met en évidence les erreurs et les incertitudes des données d'observation de la Terre. Cette recherche visait à améliorer les estimations des précipitations, de l'évapotranspiration, de l'écoulement, et du changement du stockage total de l'eau à l'échelle mondiale en utilisant une combinaison de méthodes analytiques (interpolation optimale, OI) et de méthodes de modélisation statistique, y compris les réseaux neuronaux (NN). Les modèles ont été entraînés sur un ensemble de 1 358 bassins fluviaux et validés sur un ensemble indépendant de 340 bassins et sur des observations in situ des précipitations, de l'évapotranspiration et du débit des cours d'eau. Les modèles sont étendus pour faire des prévisions à l'échelle du pixel dans des cellules de grille de 0,5° pour une couverture quasi mondiale. Les ensembles de données calibrées donnent des résidus de bilan hydrique plus faibles dans les bassins de validation : la moyenne et l'écart-type du déséquilibre sont de 11 ± 44 mm/mo lorsqu'ils sont calculés avec des données non corrigées et de 0,03 ± 24 mm/mo après calibrage par les modèles NN. Les résultats nous permettent de faire des estimations plus précises des composantes manquantes du cycle de l'eau, par exemple pour estimer l'évapotranspiration dans les zones non instrumentées, ou pour prédire le débit dans les bassins non jaugés. Les résultats peuvent également indiquer aux producteurs de données où leurs produits semblent incohérents par rapport à d'autres ensembles de données et où un étalonnage plus poussé pourrait apporter des améliorations. Enfin, cette recherche démontre l'utilisation des réseaux neuronaux et de l'apprentissage machine pour l'intégration des données satellitaires et pour l'étude du cycle de l'eau.

Mots clés : observation de la terre, télédétection, cycle de l'eau, hydrologie à grand échantillon, optimisation, calibration, apprentissage automatique, régression et classification, réseaux neuronaux, précipitations, évaporation, écoulement des eaux, débit des rivières.

Abstract

Satellite remote sensing is commonly used to observe the hydrologic cycle at spatial scales ranging from river basins to the globe. Yet, it remains difficult to obtain a balanced water budget using remote sensing data, which highlights the errors and uncertainties in earth observation (EO) data. This research aimed to improve estimates of precipitation, evapotranspiration, runoff, and total water storage change at the global scale using a combination of analytical methods (optimal interpolation, OI) and statistical modeling methods including neural networks (NN). Models were trained on a set of 1,358 river basins and validated them on an independent set of 340 basins and in-situ observations of precipitation, evapotranspiration, and river discharge. The models are extended to make pixel-scale predictions in 0.5° grid cells for near-global coverage. Calibrated datasets result in lower water budget residuals in validation basins: the mean and standard deviation of the imbalance is 11 ± 44 mm/mo when calculated with uncorrected EO data and 0.03 ± 24 mm/mo after calibration by the NN models. The results allow us to make more accurate estimates of missing water cycle components, for example to estimate evapotranspiration in un-instrumented areas, or to predict discharge in ungaged basins. The results can also indicate to data producers where their products seem incoherent with other datasets and where enhanced calibration could lead to improvements. Finally, this research demonstrates the use of neural networks and machine learning for the integration of satellite data and for the study of the water cycle.

Keywords: earth observation, remote sensing, water cycle, large-sample hydrology, optimization, calibration, machine learning, regression and classification, neural networks, precipitation, evaporation, runoff, river discharge.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Since the beginning of the space age in the 1960s, satellite remote sensing has been used to monitor water on the Earth and in the atmosphere. Today, there are dozens of earth-orbiting satellites collecting data that is translated into estimates of the major components of the water cycle (WC): precipitation, evapotranspiration, and changes in water storage.

The launch of the Gravity Recovery and Climate Experiment (GRACE) satellites in 2002 gave the scientific community the extraordinary capability of monitoring changes in the mass of water above and below the Earth's surface. The twin satellites do this by taking repeated, accurate observations of the Earth's gravity field. Information on water storage change had long been the "missing link" for calculating accurate water budgets. It is now possible to account for the water in river basins using data from satellite remote sensing plus river discharge from ground-based measurements. River discharge is the only one of the main water cycle components that is not routinely measured from orbit, although this is expected to change with the recent launch of the Surface Water and Ocean Topography (SWOT) mission on December 16, 2022.

The water cycle is an important field of study for earth scientists and for water resources planning and management. Analyses have relevance to drought, floods, agriculture, water supply, and more. And while enormous progress has been made in monitoring the water cycle via remote sensing, capturing a complete picture of the water cycle from space remains challenging.

Despite the many advances in sensor technology and calibration methods, earth observations datasets still have significant errors or biases. Two decades of research has shown that one cannot "close the water cycle," or create a balanced water budget using satellite data (Hegerl et al., 2015; Rodell et al., 2015). The usefulness of earth observation (EO) datasets has not been fully achieved because of this "incoherence" among various data products. McCabe et al. (2017) called the ability to monitor (and close) the water cycle "one of the outstanding challenges of hydrological remote sensing."

The research described in this thesis is an integrated, data-driven approach to balancing the water budget at the global scale using remote sensing data. I am seeking to optimize EO datasets describing the complete hydrologic cycle using statistical methods, *without the use of a simulation model*. I describe development and application of analytical and modeling methods to "recalibrate" remote sensing data so they result in a balanced water budget.

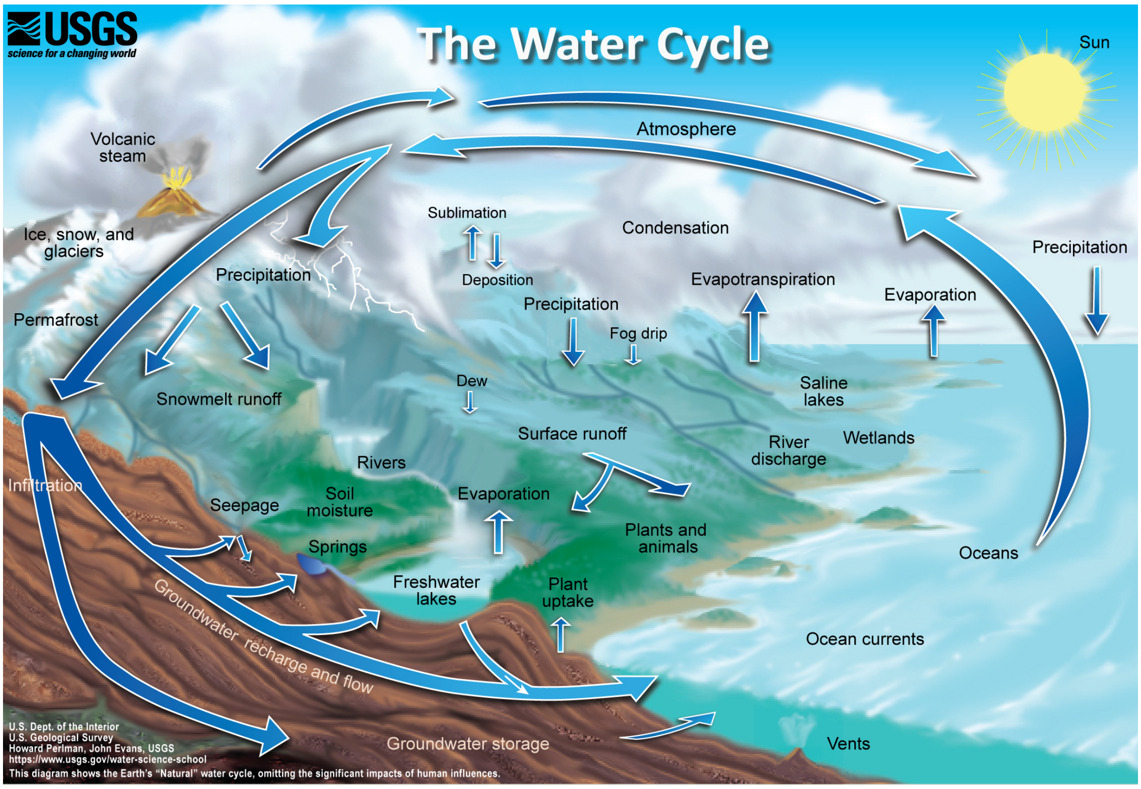
The methods are rooted in the "ensemble philosophy" – the idea that each dataset can contribute important information. The integrated approach optimizes WC components *simultaneously* rather than one at a time. In other words, to optimize observed precipitation data, we can make use of information about runoff and evapotranspiration. The premise is that these variables contain useful information for optimizing precipitation, because they are interrelated via the water cycle.

The output of my research is a new, more consistent dataset that quantitatively represents the water cycle over continental land surfaces. This is of scientific interest and practical relevance. The results show where EO datasets are less consistent, and where larger corrections are needed. This information can be used to help evaluate which datasets are best in a particular region, for example for water budgeting or modeling. The results should be of interest to data providers and algorithm developers, "to optimize existing water cycle products or identify deficiencies in current observations" (Dorigo et al., 2021).

For the remainder of this Introduction chapter, I aim to provide the context for this research, drawing from the fields of hydrology and remote sensing. I begin by briefly describing developments within the field of hydrology that have led, in the last few decades, to the era of large scale hydrology, or study of the water cycle at the global scale. I follow with a brief overview of the water cycle and describe the water budgets, a simple but powerful method used everywhere in water science and management. Next, I will give an overview of remote sensing of the water cycle. The treatment is of necessity superficial, as this is a large and diverse field. Finally, I describe various attempts that have been made to close the water cycle through a review of recent literature. These sections provide the background and motivation for my research.

1.1 The Water Cycle

Hydrology is the branch of science concerned with the movement and distribution of water through Earth's atmosphere, land surface, and subsurface. Some historians credit Leonardo da Vinci as the founder of modern hydrology (Rosbjerg & Rodda, 2019). Indeed, he performed pioneering measurements and put forth important ideas. But it was not until centuries later that the modern concept of the water cycle emerged. The French scientist Pierre Perrault performed a detailed accounting of rainfall and river flows of the Seine, described in his 1674 book *De l'Origine des Fontaines* (On the Origin of Springs). Perrault was the first to describe the water cycle more accurately, "stipulating that inflows to any area of any period of time shall always equal outflows in addition to the change in water storage" (Pfister, 2018).



U.S. Dept. of the Interior
U.S. Geological Survey
Howard Perlman, John Evans, USGS
https://www.usgs.gov/water-science-school
This diagram shows the Earth's "Natural" water cycle, omitting the significant impacts of human influences.

Figure 1.1: A typical representation of the natural water cycle, published by the US Geological Survey.

A typical representation of the natural water cycle is shown in Figure 1.1. This familiar version has been critiqued for its failure to show the human activities (Abbott et al., 2019), which can have a major influence on the water cycle. Three years after the publication of this critique, the U.S. Geological Survey published a new water cycle diagram (USGS, 2022b) that breaks with tradition by showing many human water uses. This new water cycle diagram, shown in Figure 1.2 is a fitting one for our current era, the Anthropocene – the recent geological period during which human activity has been the dominant influence on climate and the environment.

The water cycle diagram in Figure 1.2 includes another important innovation. It accurately reflects the conceptual model used by scientists and engineers by showing *pools* and *fluxes*. A pool is a volume of water stored in a particular zone, such as atmospheric water vapor, soil moisture or groundwater. Water moves

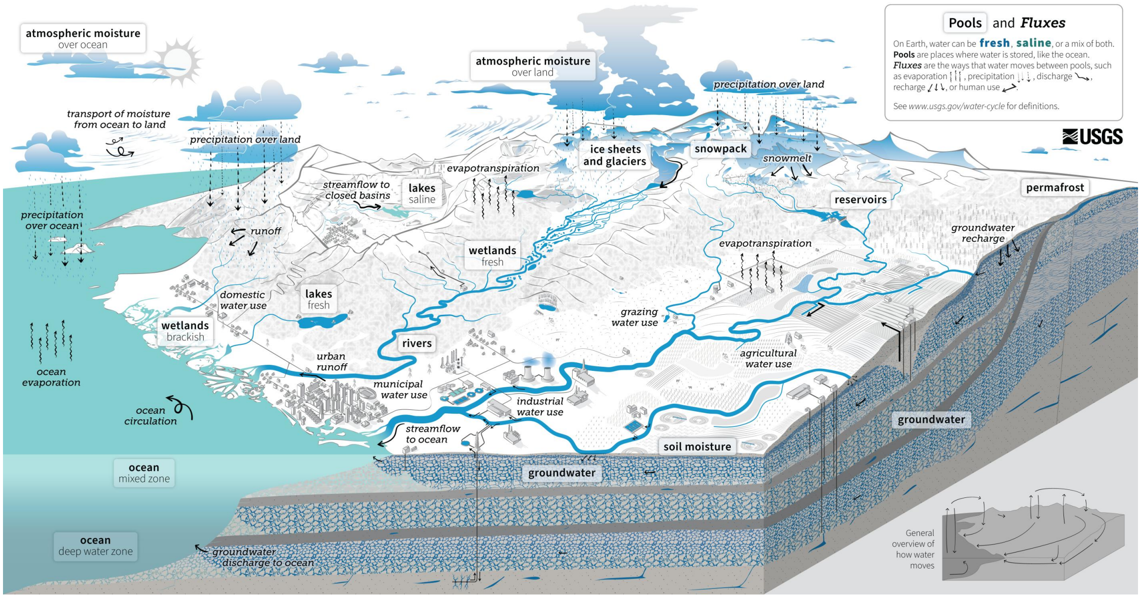


Figure 1.2: Detailed water cycle diagram published by the USGS in 2022, showing human influences.

from one pool to another via a flux (a flow across a boundary). For example, precipitation transports water from the atmosphere to the land surface, and *infiltration* describes movement from the surface into the soil. Conceptual models (and simulation models) of hydrologic systems vary in complexity, and may include dozens of pools and fluxes. As one example, some diagrams (and models) include *interception storage*, or water that is captured in tree canopies and stored (as droplets) on plant leaves. This phenomenon may be important for an accurate representation of the hydrology in certain regions, such as tropical rainforests, or at certain (short) time scales.

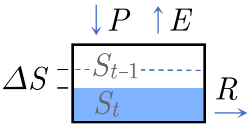


Figure 1.3: A simplified water budget showing the main water cycle components P, E, R, and ΔS.

For large-scale hydrologic investigations, it is necessary to zoom out and to simplify, ignoring many minor fluxes of water. The simplified conceptual model of the water cycle includes three fluxes plus the change in storage, as shown in Figure 1.3. It fully describes the fluxes into and out of any land area, such as a watershed or a grid cell. What is exciting is that three out of four of these components can now be measured by remote sensing. Further, the components of the water cycle can be described with simple equations or water budgets, described next.

1.1.1 Water Budgets

A water budget is the application of the law of conservation of mass in hydrology.[1] A simplified water budget for any land area (e.g., river basin, grid cell) includes the four main WC components: precipitation, P, evapotranspiration E, total water storage change (TWSC in the text and ΔS in equations), and runoff, R. By conservation of mass, the water budget can be stated:

$$P - E - \Delta S - R = 0 \tag{1.1}$$

According to scientists at the U.S. Geological Survey, "water budgets are tools that water users and managers use to quantify the hydrologic cycle. A water budget is an accounting of the rates of water movement and the change in water storage in all or parts of the atmosphere, land surface, and subsurface" (Healy et al., 2007, p. 6). An advantage of the water-budget equation is that it is simple, universal, and relies on few assumptions about the mechanisms of water movement and storage.

The water budget equation can be applied in principle over any area, or *accounting unit*, from 1 m² experimental plot, to a 7 million km² river basin.[2] Watersheds, or river basins, make convenient accounting units. In a watershed, we usually assume that there is no lateral inflow. (This is a strong assumption, and may not hold where there is groundwater flow across the basin boundary, or water diversions by canal or pipeline.)

Both P and E are regularly estimated (directly or indirectly) by remote sensing. Where surface flow is confined to a river channel, outflow R, is provided by observations of river discharge. While ΔS is not a flux (the flow of matter across a boundary), it is expressed in the same units of volume per time. Throughout this thesis, I refer to these three fluxes (P, E, and R), along with total water storage change (ΔS) as **water cycle components**, or WC components.

Overall, this research is a form of water budget analysis, conducted at a large scale and using large amounts of remote sensing data.

[1] Water is of course a molecule, not an element, so the law of conservation of mass does not strictly apply. Unlike for elements like carbon or nitrogen, water molecules are constantly being created or destroyed. Water molecules are split during photosynthesis, and new water molecules are formed in combustion and aerobic respiration, to cite just a few examples. However, it is universally assumed by hydrologists that these transformations are minimal compared to the scale of hydrologic fluxes, and are routinely ignored.

[2] Area of the Amazon River basin, the world's largest watershed.

1.1.2 Critiques of the Water Budget Method

Kampf et al. (2020) offer a critique of the water budget method. Practitioners ignore certain fluxes into and out of river where data are unavailable. But these are not always negligible. Examples include groundwater flow or man-made interbasin transfers. According to the authors, "such simplifying assumptions lead to missed opportunities for discovering where these unknowns in the water balance are important controls on streamflow." The authors advocate for expanding watershed monitoring networks to include previously unmonitored fluxes to better understand "how water moves through watersheds and between the surface and subsurface at multiple spatial and temporal scales."

Y. Liu et al. (2020) discuss the limitations of focusing solely on surface watershed boundaries. They identify two main factors that make "effective catchment areas" differ from those defined by surface topography. The first is *inter-catchment groundwater flow* – that is the movement of water into and out of the region. Subsurface flow, such as that shown in Figure 1.4, is hard to measure. In most areas, groundwater flow patterns are unknown. Yet, there is evidence that subsurface flow can be a significant contributor to the water budget in some locations (Healy et al., 2007). In practice, what we know about groundwater movement is extrapolated from modeling studies and observations at a limited number of observation wells, and estimates of groundwater fluxes have higher uncertainty than river discharge.

The second limitation of surface watershed boundaries identified by Y. Liu et al. (2020) is "limited connectivity within the catchment." This refers to portions of the surface watershed where water does not flow toward the outlet (i.e.: there are small endorheic basins or disconnected areas inside the watershed). (For a description of how I chose to handle this issue see Section 3.1.1 on page 91.)

A recent paper by Frame et al. (2023) questions the conventional wisdom of enforcing water budgets within the context of rainfall-runoff modeling. The authors state that "it might not be beneficial" for hydrologic models to enforce the conservation of water mass, arguing that it prevents hydrologic models from making accurate predictions due to errors in input (precipitation) and target (streamflow) data. They conjecture that this is a reason that machine learning models (which are not required to enforce closure) often out-perform catchment-scale models in terms of predicting river discharge. This strikes me as a radical proposal that is certain to generate controversy.

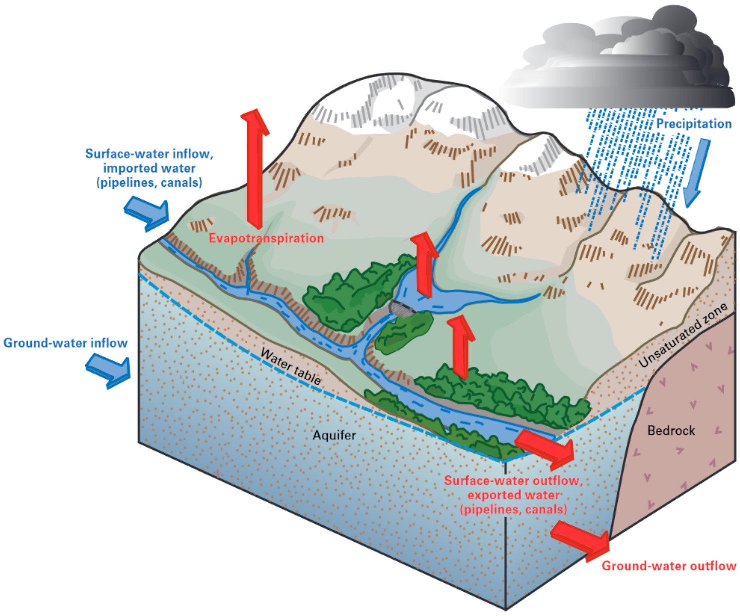


Figure 1.4: Water budget for part of a watershed. Reprinted from Healy et al. (2007). US government publication, public domain.

1.2 Remote Sensing of the Water Cycle

Remote sensing refers to any data collection from a distance, and includes medical imaging, radar, seismometers, etc. In this thesis, *remote sensing* refers to *earth observation* (EO) by satellites. The technology has interested hydrologists and water managers since the beginning of the satellite era. Lettenmaier et al. (2015) provides an excellent overview of developments. Some of the first satellite imagery was used to estimate snow cover in mountainous regions in 1968.

Satellites have become increasingly sophisticated in terms of the resolution of sensors and the number of bandwidths observed. Dozens of satellites are now dedicated to observation of the water cycle. Figure 1.5 begins to give an idea of the scale of the Earth Observation enterprise. A recent inventory stated that of 1,460 active satellites in orbit, 26% of these are dedicated to Earth Observation, and are operated by governments, the military, and commercial enterprises. The importance of these missions is highly recognized. Agencies begin planning for new satellite missions, and replacement satellites decades in advance, in order to provide continuous coverage and to maximize the quality of science and return on investment. The "concept-to-launch" timeline is now on the order of two decades (McCabe et al., 2017).

Satellite observations have many advantages over ground-based or in situ measurements. They offer broader spatial coverage, filling in the gaps between

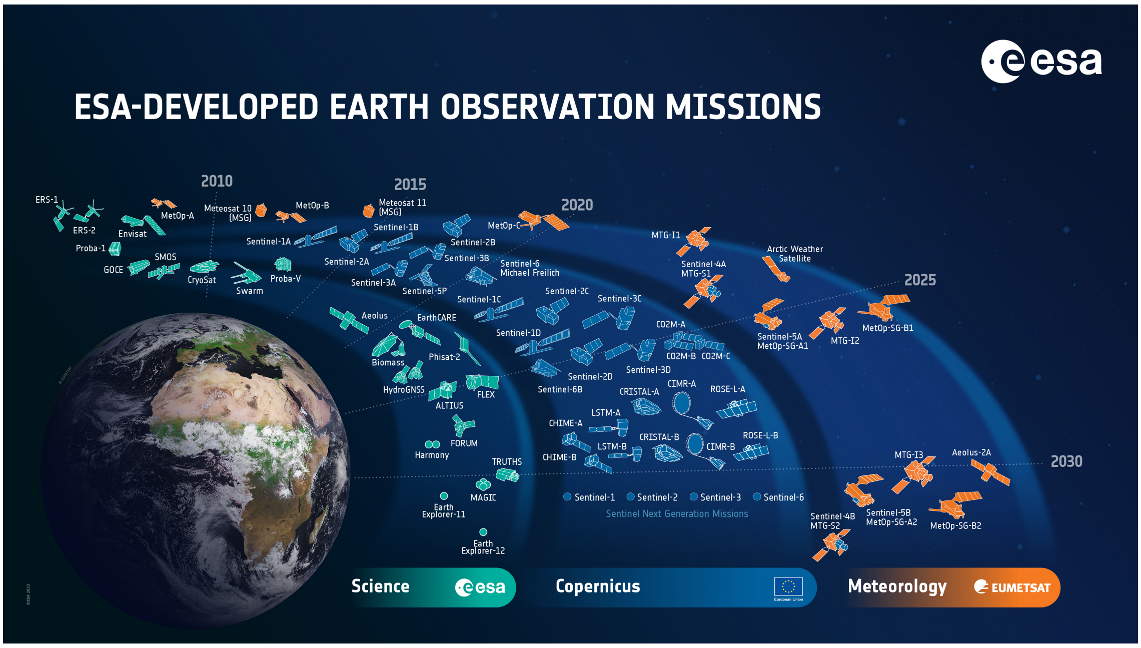


Figure 1.5: Earth observation missions developed by the European Space Agency

sparse ground stations. In remote locations or less-developed countries, remote sensing may be all that is available. Certain datasets have been published for decades, and their use is widespread, with important applications in flood forecasting, agriculture, water supply, climate modeling, and more (Lettenmaier et al., 2015). In the following sections, I give a brief overview of how satellites monitor the major components of the water cycle.

1.2.1 Precipitation

Precipitation refers to the downward flux of water from the atmosphere to the land surface. It includes rainfall and snow, and other forms of icy or frozen water like sleet and hail. Dew and fog drip are sometimes classified as precipitation (Healy et al., 2007, p. 36).The earliest precipitation measurements in the 1980s were based on infrared measurements of cloud-top temperatures, which are correlated with precipitation rate, oftentimes combined with measurements in the visible spectrum.

Precipitation has been described as *highly fractal*, as it can vary a great deal in space and time. As such, low-earth orbiting (LEO) satellites are at a disadvantage. Even with a constellation of satellites, there are usually gaps of several hours where no observations are available. Therefore, it has become common for data providers to supplement microwave data from LEO satellites with geostationary infrared satellites (Adler et al., 2018). While these observations are less accurate and have a lower resolution, they seamlessly cover much larger areas without interruption.

Each type of sensor has its advantages and disadvantages. Microwave sensors can detect emissions and lower-atmosphere scattering from rain, snow, and ice; infrared sensors measure precipitation indirectly observing cloud-top temperature and cloud height (J. Chen et al., 2020). Until 1997, retrievals relied on passive microwave observations (i.e. they measure naturally occurring microwave radiation emitted or reflected from the Earth's surface and atmosphere). The Tropical Rainfall Measuring Mission (TRMM) was the first to include active radar, which generates microwave signals that are transmitted toward Earth and are reflected or scattered. Active microwave sensors have proven so effective that they were included in subsequent missions like CloudSat in 2006 and the Global Precipitation Measurement (GPM) mission in 2014 (Kubota et al., 2020).

Certain EO datasets of precipitation incorporate station observations (Huffman et al., 1997, e.g., GPCP). Other datasets include model output. For example, one dataset I use in this analysis, the Multi-Source Weighted-Ensemble Precipitation (MSWEP), is not a pure remote sensing product but rather an "optimal merging" of gage observations, satellite observations, and reanalysis model output (Beck et al., 2019). A good overview of the current state of the precipitation observing system, challenges, and future directions is given by Levizzani and Cattani (2019).

EO precipitation datasets have been published since the 1980s, and are calibrated to an extensive network of rain gages across the globe. As a result, their errors and uncertainties are fairly well understood and well documented, at least over regions where station density is high (Lo Conti et al., 2014; Beck et al., 2020). Biemans et al. (2009) analyzed the uncertainty in precipitation datasets, comparing seven global gridded precipitation datasets. They found that the representation of seasonality is similar in all datasets, but estimates in mean annual precipitation vary widely, particularly in mountainous regions, the arctic, and over small basins. The average precipitation uncertainty (measured by distance from the ensemble mean) over river basins was estimated to be around 30%, with variations observed between basins. The authors further analyzed the effect of this uncertainty on basin runoff. They did this by applying the seven different datasets to force the uncalibrated dynamic global vegetation and hydrology model *Lund-Potsdam-Jena Managed Land* over 294 river basins worldwide. Unsurprisingly, there was considerable variance in model predictions of mean annual and seasonal discharge as a result of the different forcings. The authors conclude that it is important to consider precipitation uncertainty in water resources assessments, validation, and calibration of hydrological models. As I will explain further in Chapter 3, the statistical methods used in this research rely strongly on uncertainty estimates to optimally merge different datasets.

1.2.2 Evapotranspiration

Evaporation refers to the conversion of liquid water to water vapor. *Transpiration* is the loss of water vapor by plants via their stomata, the openings in leaves by which gases are exchanged. Transpiration is responsible for moving water from the soil into the atmosphere through plant growth and respiration. Because it is difficult to measure these two fluxes independently over land, they are often combined into the single term *evapotranspiration*. Thus, in the hydrological sciences, evapotranspiration refers to the upward flux of water vapor from land and water surfaces to the atmosphere.

On average, evapotranspiration is the second-largest water-budget component after precipitation. It is an important driver of the global climate, responsible for the exchange of water and energy from the land and sea surface to the atmosphere. It has been estimated that as a global average, evapotranspiration is about 60% to 75% of precipitation (Shiklomanov, 2009).

Evapotranspiration cannot be measured directly via remote sensing. Hydrologists have created a number of climatological methods for estimating evapotranspiration, using inputs such as daily temperature, relative humidity, or solar radiation. These methods vary from purely empirical to those with a more explicit grounding in theory. Examples of such methods include Thornthwaite, Jensen-Haise, Hamon, Penman-Monteith, and Priestley-Taylor (Shuttleworth, 1993).

Remote sensing data can provide the inputs to these relations, and satellite data on vegetation can help estimate the seasonal dynamics and relative magnitudes of evapotranspiration (Fisher et al., 2017). EO datasets of evapotranspiration have become more reliable and are widely used in science and water management, including water balance studies to crop performance monitoring (Mu et al., 2011). Yet, compared to precipitation stations, there are far fewer ground-based measurements of evapotranspiration (Fisher et al., 2008; Paca et al., 2019). It has also been shown that different algorithms yield substantially different outputs (M. Cao et al., 2021). Both of these factors (divergent algorithms and sparse ground stations) contribute to higher uncertainties in evapotranspiration than for other water cycle components.

1.2.3 Total Water Storage Change

The Gravity Recovery and Climate Experiment (GRACE) is an innovative joint mission of NASA and the German Aerospace Center (DLR). The first pair of GRACE satellites were in operation from 2002 to 2017, and a follow-on mission began in 2018. GRACE collects detailed observations of Earth's *gravity field anomalies*. Based on these anomalies, scientists are able to model how mass is distributed around the planet and how it varies over time (NASA Jet Propulsion Laboratory, 2018).

Most short-term changes in the Earth's gravity field are due to the movement of water on land and underground (Tapley et al., 2004). The gravimetric methods employed by GRACE have a solid basis in physics, yet researchers have not found an effective way to ground truth observations (Kusche et al., 2009; Reager et al., 2015). Furthermore, GRACE observations have a coarser spatial resolution than many EO data products. GRACE Level 3 data (estimates of liquid water equivalent) have been spatially filtered to remove random errors and systematic errors (Landerer & Swenson, 2012). The current mascon-based solutions improve upon the older spherical harmonic solutions, eliminating the north-south striping that plagued earlier releases (Scanlon et al., 2016). While the datasets have a relatively high 0.25° resolution, fine scale detail (i.e. values in a single pixel) are not likely to be meaningful, and data are more accurate when averaged over larger regions (Tapley et al., 2004).

GRACE data have been used in groundbreaking studies to analyze the terrestrial water budget, drought, climate change, and water management. GRACE data have been assimilated into land surface models (Zaitchik et al., 2008; Kumar et al., 2016) and have contributed to better prediction of groundwater availability and drought. GRACE allows researchers to document water stress and groundwater declines even in regions where data from monitoring wells are not available (Konikow, 2013; Richey et al., 2015; Zaki et al., 2019). Besides these examples, researchers have used GRACE data for many other applications in terrestrial hydrology – for a more complete overview see Jiang et al. (2014). Most importantly for this research, GRACE has made it possible to more completely monitor the water cycle, making it possible to perform water budget analyses. I describe over a dozen such studies in the literature review below.

1.3 Review of Literature on Closing the Water Cycle at the Global Scale

Historically, hydrologic investigations were conducted by engineers for practical purposes such as estimating reservoir yield or flood flows. In the last several decades, hydrologists have been increasingly interested in describing the movement of water at continental and global scales (Eagleson, 1994). Large-scale hydrology is concerned with exploring “spatial scales greater than a single river basin all the way up to the entire planet” (Cloke & Hannah, 2011). The sub-discipline of large-sample hydrology (LSH) often uses remote sensing data to collect data and inputs over large sets of river basins. LSH studies have made important contributions in extreme events, modeling, human impacts on hydrology, and climate change assessments (Addor et al., 2020).

Yet, the field of large-scale hydrology is not without its detractors. Kauffeldt et al. (2013) discuss the limitations of the large-scale approach to hydrology, describing some of the commonly used data and methods as “disinformative.” Among the problems identified by the authors are: (1) difficulty in accurately delineating river basin boundaries, (2) inconsistency in precipitation and evapotranspiration data, and inability to close the water balance, (3) modeled evapotranspiration that frequently exceeds physically realistic limits, and (4) basins where average runoff exceeds precipitation (not expected under natural conditions). Nevertheless, large-scales studies have made important contributions in our understanding of continental-scale water dynamics, regional water availability, and climate change impacts. Kauffeldt et al. (2013) stress the importance of screening datasets before modeling to eliminate biased datasets. This, they state, will increase confidence in the validity of model output and the chances of drawing robust conclusions from model simulations. By contrast, the philosophy behind this research is different. Rather than screening out bad datasets, or looking for the best combination, we use methods from statistics and optimization for data fusion. This is based on the idea that no dataset is perfect, and each can contribute valuable information.

1.3.1 Water Cycle Closure

The main goal in this thesis is to reconcile remote sensing data to “close the water cycle” or “balance the water budget”. I use both terms interchangeably, and they should be understood as referring to the same concept: reducing or eliminating the water cycle *imbalance*, $I = P - E - \Delta S - R$.

Many authors have affirmed the difficulty of balancing the water budget with observations of hydrologic fluxes. Kampf et al. (2020) note that even in a highly-instrumented experimental plot less than 10 m long, the water balance is uncertain. In their view, this is because "precipitation data have biases that are not correctable, evapotranspiration is difficult to measure, and subsurface components are rarely measured." The problem becomes even more difficult when scaled up to a large watershed with variable topography, vegetation, and aquifer properties.

The inability to close the water cycle at various spatial and temporal scales using remotely-sensed data has been widely discussed (Dorigo et al., 2021). A variety of approaches have been tested for either assessing or correcting the imbalance. I have sorted these efforts into a few broad categories, although some studies employ more than one of these methods.

Assimilation – Perhaps the most common approach focuses on *assimilation* of EO into hydrological models (see e.g., Yilmaz et al., 2011; Y. Zhang et al., 2016; Wong et al., 2021). In the field of weather forecasting, data assimilation refers to methods for reconciling numerical models with remote sensing observations. In brief, data assimilation continuously compares new data with an existing model, and the model is updated to reflect the new conditions. This is a large and important field with a rich literature, which however, I will not discuss further here, as my research focused on using methods other than simulation modeling.

The do-nothing approach – In one class of studies, scientists combine EO datasets without attempting to correct or reconcile them. The purpose may be to assess or document the bias or uncertainty in EO datasets. However, it is also common for scientists to assume that datasets are accurate (or at least unbiased) so that they may compute unknown water cycle components. For example, Rodell et al. (2011) estimated evapotranspiration, over seven large river basins, by assuming that mass is conserved and using the relation $E = P - \Delta S - R$. (The water budget method of predicting hydrologic fluxes is discussed further in Chapter 6.)

Best combination – Some analysts compile many different EO datasets, looking for those which are most representative for their region of interest. Studies may focus on a single variable like precipitation, and compare EO data to local observations (Huang et al., 2016). Or they may combine datasets looking for the combination that results in the lowest imbalance (Wong et al., 2021). In other cases, scientists seek the combination that best predicts a single variable. Lehmann et al. (2022) estimated ΔS as a function of observed and modeled P, E, and R over 189 large river basins, comparing predictions to GRACE observations. The authors looked for the best combination of input datasets that would maximize the fit between observed and predicted ΔS.

Bias-correction of individual water cycle components – Following this method, datasets are bias-corrected through comparison with in situ observations or modeled fluxes before being used in water cycle analyses. One example is provided by Schlosser and Houser (2007), who estimated global precipitation and evapotranspiration using bias-corrected reanalysis model data.

Ensemble-based methods – This approach involves averaging the water cycle components in each class. For example, multiple precipitation datasets are averaged to calculate an ensemble mean P. Examples in the literature include papers by Lorenz et al. (2014) and Lehmann et al. (2022). A weighted average can be used, with weights proportional to the analyst's confidence in a dataset. Where information on uncertainty is available, inverse-variance weighting is a popular option; however, detailed information about errors in EO datasets is rarely available (Tian & Peters-Lidard, 2010).

Include energy budget constraints – For the NASA Energy and Water Cycle Study (NEWS), a pair of studies demonstrated how to explicitly couple the energy and water cycles using satellite observations over both land and oceans (L'Ecuyer et al., 2015; Rodell et al., 2015). Thomas et al. (2020) refined NEWS water and energy balance estimates by analyzing error covariances for ocean turbulent heat flux products (Stephens et al., 2012).

Statistical optimization – This set of methods forces water budget closure without the use of a simulation model, instead using techniques drawn from statistics and optimization. Data-driven approaches to simultaneously optimizing multiple WC components can effectively close the water cycle, redistributing errors among the various components (e.g.: Pan & Wood, 2006; Aires, 2014, several more references given below). This will be the focus for the remainder of this literature review, and is the approach used in this thesis.

Table 1.1 summarizes recent studies that focused on closing the water cycle with remote sensing data using statistical optimization methods. The table lists the number of datasets used in each study, and the time period (temporal domain) of the analysis. Some of the studies in Table 1.1 use only remote sensing data as inputs, while others include data from models and in situ observations. The table includes a very brief description of the integration method used. Here, *integration* refers to the method for modifying EO datasets so that they result in closure.[3]

Many, but not all, of the studies listed in Table 1.1 provide estimates of the

[3] Authors use a variety of verbs to describe the process of modifying water cycle components to achieve closure. I have come across: *calibrate, harmonize, integrate, modify, recalibrate, reconcile, unify*.

uncertainty in the optimized water cycle components, often in terms of standard deviations or 95% confidence intervals. In such cases, I briefly describe the authors' approach to assessing uncertainty. In addition, many studies "ground truth" their results by comparing them to in situ observations. Where applicable, I include this information in the right-most column of Table 1.1.

Below, I provide some more details about some of these studies. My goal is to give a broad overview of recent work in the field, in order to put my research in context. This study is included in the final row of Table 1.1, and shows how my research expands upon previous work:

- larger number of river basins are analyzed,
- longer analysis time period,
- expanded set of evaluation data is used.

There are also methodological differences, to be described in Chapters 3 and 4. One unique aspect is my use of machine learning methods, which to the best of my knowledge, has not been used to date for closure of the water cycle using remote sensing data at the global scale.[4]

[4] Aires (2014) includes a brief discussion of the potential of neural network models for water cycle closure.

Table 1.1: Recent studies attempting to estimate a balanced water budget via remote sensing observations

	Number of datasets									
Study	**P**	**E**	**ΔS**	**R**	**Study domain**	**Spatial scale**	**Temporal domain**	**Integration method**	**Method for assessing uncertainties**	**Data used for validating results**
Pan and Wood (2006)	1	1	1	1	One heavily instrumented experimental area in Oklahoma, USA	0.5° grid cells	1997 - 1999	Constrained ensemble Kalman Filters and the VIC hydrologic model	"To evaluate the interpolation uncertainties, a "leave-one-basin-out" cross-validation procedure was performed."	in situ observed P, E, R, and soil moisture
Sheffield et al. (2009)	2	1	1	0	Mississippi River basin	basin	2003 - 2006	None. The authors estimated R = P - ET - ΔS. Bias correction applied to estimated R.	Relative error in discharge is the quadrature sum of errors in each component	in situ observed R (river discharge)
Azarderaksh et al. (2011)	4	2	1	1	Amazon + sub-basins	basin	2002 - 2006	No integration (searched for most consistent combination).	-	in situ observed R
Sahoo et al. (2011)	8	6	1	1	Global, 10 basins	basin	2003 - 2006	Constrained ensemble Kalman filter	Bias and RMSE are estimated by for different satellite precipitation (P) products with respect to the non-satellite merged product over ten river basins	-
Pan et al. (2012)	4	2	1	1	Global, 32 basins	basin	1984–2006	Constrained Kalman filter	-	-
Munier et al. (2014)	7	3	4	1	Mississippi Basin	basin	2002 - 2010	Simple weighting + post filtering (OI). The OI solution is approximated for each dataset with a single linear regression model.	-	in situ observed P and E, gridded datasets of spatially interpolated observed P.
Lorenz et al. (2014)	5	6	6	7	Global, 96 basins	basin	2003 - 2010	None, looked for best combination of datasets to predict R = P - ET - ΔS	-	-
Lorenz et al. (2015)	5	6	1	2	Global, 29 basins	basin	2005 - 2010	Ensemble Kalman filter and Constrained Ensemble Kalman Filter	Range of different estimates of a variable assumed to be a proxy for the uncertainty	in situ observed R
Rodell et al. (2015)	1	3	1	1	Global	continents	2002 - 2009	"maximum a posteriori solution," equivalent to OI (Aires 2014).	"The standard deviation across the original estimates is then taken to represent the uncertainty in the blended estimate."	Compared P, E, and R with those from 3 other published studies
Zhang et al. (2016)	5	6	1	2	Global	grid cells	2004 - 2007	Inverse variance weighting (equivalent to "simple weighting" in Aires, 2014) and Constrained Kalman Filter (CKF) method	none	R: observed discharge in 16 medium-sized basins
Munier and Aires (2018)	4	3	1	4	Global, 11 basins	basin, grid cell	2002 - 2010	Simple weighting + post filter (OI). Closure correction model, a 2- or 3-parameter regression to emulate the OI solution for each EO variable. Authors developed a 4 sets of equations for different climate zones they defined based on P, E, and vegetative cover (NDVI).	"the average of the corrected datasets (CCM with CIC) was considered as the reference to compute biases and uncertainties (standard deviation) of the corrected datasets, as well as correlation of errors."	in situ observed E
Zhang et al. (2018)	4	8	3	4	Global, 32 large basins	basin, grid cell	1984 - 2010	Constrained Kalman filter (CKF). Inputs from remote sensing, models, and observations.	"For the individual data products, their ensemble mean is taken as the best estimate for the variable, and the ensemble spread against the ensemble mean as a proxy for their uncertainty."	in situ observed R and E
Pellet et al. (2019a)	4	3	4	2	Mediterranean Basin, 6 sub-basins	basin	1980 - 2009	Simple weighting + post filter (OI). Scaling factor used to extrapolate results to sub-basin scale.	Assumed, based on standard deviation of annual predictions.	in situ observed P and E
Pellet et al. (2019b)	3	3	1	1	Southeast Asia, 5 basins	basin	1980 - 2015	Simple weighting + post filter (OI). Three-parameter regression used to extrapolate to time periods with missing data.	Comparison of the EO dataset to the OI result. Uncertainty assumed to equal the standard deviation of the residuals.	in situ observed R
Soltani et al. (2020)	1	1	1	0	1 basin - Central Basin of Iran	basin	2009 - 2016	No integration. Calculated R = P - ET - ΔS	-	None, although authors performed a sensitivity analysis on E.
Pellet et al. (2021)	4	3	5	1	Amazon River Basin	basin	2000 - 2015	Simple weighting + post filter (OI). Regression relationship used to extrapolate to sub-basin and pixel scale.	-	ΔS from Zhang et al. (2018); R from the models ERA-Land and CAMAFlood

Study	Number of datasets				Study domain	Spatial scale	Temporal domain	Integration method	Method for assessing uncertainties	Data used for validating results
	P	E	ΔS	R						
Abolafia-Rosenzweig et al. (2021)	5	3	4	3	Global, 24 basins	basin	2002 - 2014	Used 3 different methods: Proportional redistribution (PR), Constrained Kalman filter (CKF), and Multiple collocation (MCL). Applied all 3 methods to each possible combination of EO datasets.	Ensemble spread as a proxy of uncertainty for water budget estimates	Compared results to the assimilation model results of Zhang et al. (2018)
Luo et al. (2021)	5	4	3	0	1 basin - Tarim	basin	2003 - 2017	No integration (searched for best combination). Goal was to assess errors in EO datasets.	First order reliability method. Because their study basin is endorheic, there is no outflow, simplifying the water balance to 3 variables.	-

This study (2023)	3	3	3	1	Global, 1,698 basins	basin, grid cell	1980 - 2019	Inverse variance weighting (simple weighting) + Optimal Interpolation. (a) OI solution recreated with single linear regression model + parameter regionalization via spatial interpolation (kriging). (b) Neural network model to recreate the OI solution.	Uncertainty estimated by comparing differences between model results and optimized (OI) solution and to gridded interpolated observations of P.	In situ observed P (GHCN stations;
										R: river discharge at gages,
										E (FluxNet);
										Satellite observed ΔS (GRACE)

Several of the papers in Table 1.1 use Kalman filters to close the water cycle. In brief, a Kalman filter is an efficient recursive algorithm that estimates the state of a dynamic system based on a set of noisy measurements. The process was introduced by Rudolf Kalman in the 1960s, and is now widely used in many fields, including the Earth Sciences. The ensemble Kalman filter (EnKF) expands upon this approach by using multiple model state vectors to represent the system's state and its uncertainty. This is helpful when there are multiple sources of observations, for example from satellite remote sensing, ground radar, station observations, and ocean buoys. The EnKF method is also useful where systems have nonlinearity or high dimensionality, and is common in weather and climate models (Schillings & Stuart, 2017).

Pan and Wood (2006) performed a pioneering study using the constrained ensemble Kalman filter method to estimate balanced water budget components over a heavily instrumented experimental area in Oklahoma, in the United States. The authors used EnKF and assimilation with the VIC macroscale land surface model to "optimally redistribute" water budget imbalances. Sahoo et al. (2011) merged EO datasets for P, E, and ΔS over 10 large basins, using weighted values based on their errors, to mitigate errors in the individual satellite products. The authors use the constrained ensemble Kalman filter (CenKF) method to revise water cycle components in order to satisfy the water budget closure constraint. Pan et al. (2012) used similar methods, including CenKF to merge EO datasets over 32 global river basins. In addition to remote sensing data, the authors used in situ observations, land surface model simulations, and global reanalyses.

Following a somewhat different approach, Aires (2014) introduced an integration method called optimal interpolation (OI) that draws inspiration from inversion of satellite retrievals. It is a closed-form analytical solution that imposes a WC budget closure constraint. The OI modifies each of the WC components by an amount inversely proportional to its uncertainty. Aires showed that this constraint improves the estimation of the WC components in some places and times. In a following paper Munier et al. (2014) applied OI over the 3 million km^2 Mississippi River basin, revising satellite estimates for P, E, R, and ΔS. Optimal interpolation is at the basis of the methods used in this study, and is described more thoroughly in Section 3.8.

Multiple collocation (MCL) has also been used for solving the problem of balancing the water budget. MCL is an advancement of the triple collocation method, described well by Pan et al. (2015). Unlike methods relying on known errors in input products, MCL determines error levels based on the mutual distance (mean squared distance) among different observations, assuming their errors are uncorrelated. This a problem without a unique solution; the MCL algorithm seeks to find the best compromise in an over-constrained system. Another method for balancing water budgets is proportional redistribution (Abolafia-Rosenzweig et al., 2021). This method is perhaps the simplest of those described here, as it does not take into account uncertainties (information which is often unavailable), and simply redistributes water budget residuals to each component in proportion to its magnitude.

The techniques described above (best combination, Kalman filters, optimal interpolation, multiple collocation, and proportional redistribution) all have an important limitation. Because they require estimates of all four of the major water cycle components, they can only be applied over river basins, where estimates of river discharge can be used as a proxy for basin runoff. (For a discussion of the relationship between runoff and discharge see Section 2.5 on page 62.)

Munier et al. (2014) were among the first to deal with the problem of how to extend the closure analysis beyond the basin scale. What is noteworthy about their approach is that it is independent from any model. The authors created a simple linear model with auxiliary environmental variables to extend predictions to the global level at the pixel scale. In this paper, the auxiliary information was used in a fairly simple way. The authors did not us environmental data to divide basins into classes based on climate regime. (I hypothesized that predictions could be improved with a more complex model. In Section 4.4.2 I describe how environmental data are used as input variables to a neural network model.)

Later, Munier and Aires (2018) applied OI over 11 large river basins, from the 620,000 km² Colorado River basin to the 4.7 million km² Amazon. OI has also been shown to work well in optimizing satellite observations of the hydrologic cycle over river basins in the Mediterranean (Pellet et al., 2018), South Asia (Pellet, Aires, Papa, et al., 2019), and the Amazon (Pellet et al., 2021).

1.3.2 Machine Learning in the Hydrologic Sciences

Machine learning involves the use computer algorithms that learn from data. In contrast, in conventional computer programs, the programmer writes explicit instructions for how to treat data. One type of machine learning algorithm, the neural network, has been the basis for some of the most exciting recent advances in artificial intelligence and computer science. Examples include large language models like ChatGPT, capable of providing detailed and coherent responses to a wide variety of questions. Another example is image generators such as Dall-E and

MidJourney, which can produce extraordinary artwork based on a text “prompt” from the user. These models belong to a special class of “deep learning” models that ingest huge amounts of data during training. The large language model Chat-GPT 3.5 contains 175 billion parameters, and its successor v4 is rumored to contain 10 times more (Farseev, 2023).

Neural networks have been applied in the geosciences since the 1980s for a variety of tasks, from classification of vegetation to solving inverse radiative transfer function problems in remote sensing. Aires et al. (2001) used neural networks to estimate atmospheric water vapor, cloud liquid water, surface temperature, and emissivities over land from satellite microwave observations. PERSIANN (Ashouri et al., 2014) is a well-known precipitation dataset that uses a neural network model to estimate rainfall intensity based on data from geostationary infrared sensors and low-earth orbiting infrared sensors. Neural networks have also been used to estimate global evapotranspiration by merging satellite and ground-based observations (Shang et al., 2021).

Recently, dozens of papers have been published which use deep neural networks to predict river discharge. Rainfall-runoff models have many important applications, such as forecasting droughts and floods. Recent work has shown that general-purpose deep recurrent neural networks, such as long short-term memory (LSTM) models, can produce state-of-the-art hydrologic forecasts (Nearing et al., 2021), often outperforming conventional rainfall-runoff models.

A critique of such models is that they are black boxes, and while they may make accurate predictions, they do not offer any insight into the functioning of the hydrologic system (Xu & Liang, 2021). This lack of interpretability limits the use of ML models in critical decision-making processes where stakeholders require transparency and explanations for the predictions. It also limits their use for scenario-based planning, a major use for hydrologic models. (For example, how does basin runoff respond to a proposed land use change?). And since machine learning models are trained on historical data, they implicitly assume that future patterns and relationships will be similar to the past. Therefore, they are unlikely to make accurate predictions in a changing environment (Hong et al., 2021).

One promising approach involves hybrid machine learning models where some knowledge of the system is coded into the network model prior to training (Moshe et al., 2020). These so-called physics-informed neural networks (von Rueden et al., 2023) have shown promise in the Earth Sciences (although my experiments with them to close the water cycle were lackluster).

1.4 Research Questions and Objectives

This study has two main objectives. The first is to optimize hydrologic EO datasets and to calculate a balanced water budget at the river basin scale from these data. The second objective is to train models based on these results in order to make improved estimates of water cycle variables at the pixel scale. The output will be a harmonized gridded dataset for the variables P, E, ΔS, and R. These data will give a more complete global view of the water cycle and be useful for a variety of applications.

A third, stretch goal for the study was to test the model's ability to indirectly estimate missing water cycle components. For example, one can estimate GRACE-like TWSC by rearranging Equation 1.1 to give $\Delta S = P - E - R$. This would allow us to fill in missing data or to estimate water storage from before GRACE was launched in 2002. Similarly, one may estimate runoff in ungaged basins, a challenge that has preoccupied hydrologists for decades (Wagener et al., 2004). My hypothesis is that water cycle components that have undergone recalibration will allow more accurate prediction of missing hydrologic variables, and better understanding of the water cycle.

My analytical approach involves optimization of satellite observations for closure of the water budget. This involves two main steps. First, I use the optimal interpolation (OI) over a predefined set of river basins. The solution is an optimized set of WC components which satisfy the closure constraint.

Next, I train a set of models to emulate the OI solution, and which can be applied outside of the training basins, including at the pixel scale. I experiment with different types of models, including a simple model based on linear regression, and a more complex neural network model. The goal of the models is to *calibrate* EO datasets, with a goal of making them closer to the optimized version calculated by OI. This approach allows me to compare simple versus complex models for water cycle closure.

1.5 Organization of this Thesis

In this section, I give an overview of the organization of this document. This Introduction provided background and context for my research. The point to remember is that one cannot combine remote sensing datasets to create a balanced water budget. Solving this problem is an active field of research, but no consensus solution has emerged. Chapter 2 describes the datasets used in this research, covering all four major water cycle components: P, E, ΔS, and R. This includes data from remote sensing for data fusion and in situ observations that will be used later for evaluation. I also describe datasets describing environmental conditions: elevation, aridity, vegetative cover, and several others. These variables will be used as explanatory variables in the modeling to calibrate EO datasets.

Chapter 3 describes the methods used to create basin-scale water budgets, and details optimal interpolation (OI), a powerful analytical method for redistributing the water budget residual among water cycle components. I experiment with variants of the OI method, and conclude that using an affine error model yields the most consistent and realistic results. However, OI is needed, as it can only be applied over river basins where we have access to observations of all four water cycle components, including runoff. As the goal is to calibrate EO datasets at the pixel scale, another method is needed.

In Chapter 4, I describe two modeling methods that seek to recreate the OI solution, which are then used to make predictions at the pixel scale. The first method is based on fitting simple linear models, and using surface fitting methods to spatially interpolate the model parameters. The second method uses neural networks, a powerful machine learning method for estimating relationships among variables and fitting predictive models. Chapter 5 presents the results of these analyses. I analyze the modeling output in a number of ways in order to answer key questions. How much do the models change the original data? How well do we close the water cycle? Is the fit to in situ observations degraded or improved?

Chapter 6 extends and further evaluates the results presented in Chapter 5. The main emphasis is on making indirect estimates of a water cycle component via the other three components. For example, one may estimate discharge in an ungaged basin with $R = P - E - \Delta S$. There have been many studies attempting to do just this, usually with limited success. I show that using EO variables calibrated by the NN models developed here results in greatly increased predictive skill. Applications in record extension and filling missing data are discussed.

Finally, in the Conclusion, I summarize the main findings, and offer my thoughts on the significance and implications of this work. I describe which aspects of this work are novel, and help to advance research in the fields of hydrology and remote sensing. I discuss the strengths and limitations of the methods proposed here. Finally, I offer recommendations, including those for future study.

Chapter 2

Earth Observation Datasets

In this chapter, I describe the database of remote sensing-derived data, or Earth Observations (EO), assembled for analysis of the global water cycle. The first type of data are for the water cycle (WC) components. These data describe fluxes, the movement of water, over the earth's land surfaces, or changes in total water storage. Since the focus of this research is terrestrial hydrology, oceans are not considered. The second type of data describe environmental conditions, and include variables such as elevation, slope, and vegetative cover. Some of these datasets are derived entirely from satellite remote sensing data, while others are blended with in situ data or information from models.

This chapter begins with a discussion of the format and conventions for the datasets. To compare or merge EO datasets, they must have the same spatial and temporal resolution. Next, I describe the EO data sources for the four main WC components. Following this, I introduce a variety of environmental variables that serve to characterize local conditions and are used as input variables for neural network modeling described in Chapter 4.

The last section of this chapter presents a preliminary analysis of the EO datasets. I evaluated the quality and completeness of each dataset using the tools of exploratory data analysis – maps, time series plots, summary statistics. This helped to find and discard anomalous observations in some datasets and to verify that data had been imported and processed correctly.

2.1 Earth Observations of Water Cycle Components

Earth observation refers to the gathering of information about Earth's physical, chemical and biological systems via remote sensing technologies. Broadly, *remote sensing* involves monitoring and detecting the physical attributes of a region through the measurement of its reflected and emitted radiation, typically done from satellites or aircraft. In this thesis, I primarily use data collected from earth-orbiting satellites, although some of the datasets I drew upon also include information from in situ observations, models, and other sources.

First, we are concerned with what EO can tell us about the movement of water above, on, and below the earth's surface. The movement of water is expressed as a flow, or a *flux*. In the sciences, a flux is the flow of a substance into or out of a system. As described in Section 1.1.1, the water budget for any land area can be described with four main WC components: (1) Precipitation, P, (2) Evapotranspiration, E, (3) Total Water Storage Change (TWSC), or ΔS, and (4) Runoff, R. With measurements of these four variables, one can quantify the mass or volume of water flowing into and out of any arbitrary geographic region, such as a river basin or a grid cell.

2.1.1 Units

In this thesis, the four water cycle components are expressed as an area-normalized flux density, in units of depth of water per time. Units are in millimeters per month, or mm/mo. GRACE TWSC is not a flux per se, but can be expressed in the same units.

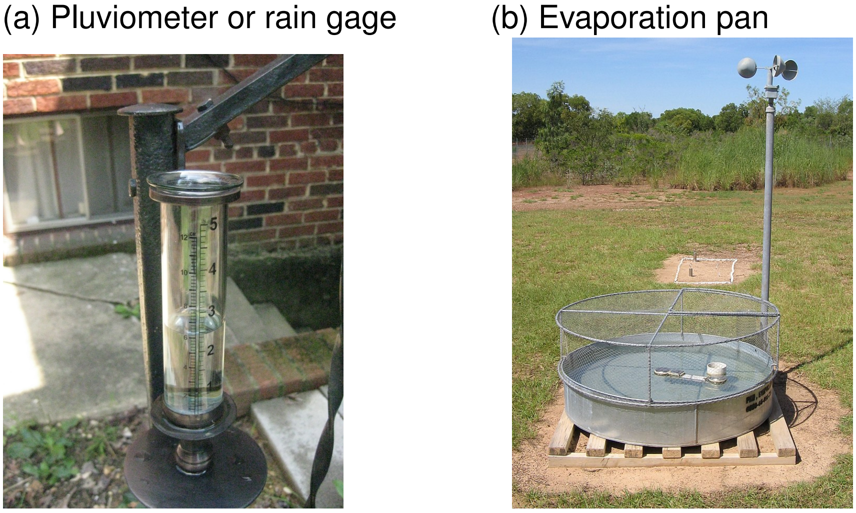
In principle, a variety of units can be used. According to the US Geological Survey, "water-budget equations can be written in terms of volumes (for a fixed time interval), fluxes (volume per time, such as cubic meters per day or acre-feet per year), or flux densities (volume per unit area of land surface per time, such as millimeters per day)" (Healy et al., 2007).

Using a flux density, in units of depth/time, is both a practical and mathematical convenience. It not only simplifies calculations, but it also allows us to directly compare fluxes over regions with different surface areas. The concept of depth per time is perhaps most intuitive with variables such as rainfall and evaporation, as we are accustomed to thinking of the vertical movement of water. Rain gages, or pluviometers (Figure 2.1(a)), are simple devices that measure the depth of water captured in a given time period. Similarly, evaporation is most often measured as the change in the depth of water in an evaporation pan (Figure 2.1(b)).

In the case of discharge or river flow, it is somewhat less intuitive to think of this as a flux density, in units of depth per time. In this case, the discharge, Q measured at a given location in m³/s is converted to runoff, R a flux density in mm/month by dividing by its basin area (and converting the units):

$$R\left(\frac{\text{length}}{\text{time}}\right) = Q\left(\frac{\text{length}^3}{\text{time}}\right) \times \frac{1}{A\left(\text{length}^2\right)} \tag{2.1}$$

The units for runoff, R are depth per time, just like P or E. We can think of the depth as a thin layer of water that is uniformly distributed over the basin. Unless care is taken to convert Q and A to compatible units, the results of Equation 2.1 will not be easy to interpret. Runoff in mm/month can be calculated from a



(a) Pluviometer or rain gage

(b) Evaporation pan

Figure 2.1: Important measurement devices for two fundamental hydrologic variables, which measure the change in depth of liquid water

conventional volumetric flow rate in m³/s by dividing by the land surface area in km², multiplying by the number of days in the month, and an appropriate conversion factor, as follows:

$$R\left(\frac{\text{mm}}{\text{month}}\right) = Q\frac{\text{m}^3}{\text{s}} \times \frac{1}{A,\,\text{km}^2} \times \frac{d,\,\text{days}}{\text{month}} \times \frac{\text{km}^2}{10^6\ \text{m}^2} \times \frac{1{,}000\ \text{mm}}{\text{m}} \ldots$$

$$\times \frac{60\ \text{s}}{\text{minutes}} \times \frac{60\ \text{minutes}}{\text{hour}} \times \frac{24\ \text{hours}}{\text{day}} = 86.4\frac{Q\cdot d}{A} \tag{2.2}$$

where d is the number of days in the month, a whole number between 28 and 31. Rearranging this equation allows us to calculate basin discharge, Q in m³/s as a function of the runoff, R, in mm/month:

$$Q\left(\frac{\text{m}^3}{\text{s}}\right) = \frac{R\cdot A}{86.4\ d} \tag{2.3}$$

where R is in units of mm/month and A is the basin area in km², and d is the number of days in the month. These are simple operations, but it helps to have them clearly documented, as all of the following calculations depend on having accurate input data.

2.1.2 Time Period and Temporal Resolution

All analyses described in this thesis were done with *monthly* data. Many remote sensing data products are available at higher temporal resolution (for example, hourly, daily, weekly). And in theory, water budget calculations can be performed at any time scale. However, there were practical and scientific reasons why I chose a monthly time scale. In practice, our ability to produce more frequent updates is limited by the GRACE data for water storage change, which are calculated monthly with a significant time lag (Rodell et al., 2018).

I chose to perform most analyses for the 20-year time period from January 2000 to December 2019, for a total of 240 months. The GRACE satellites were launched in 2002, and the first observations are available for April 2002. Nevertheless, I chose to begin the analysis in January 2000, as it is simple and memorable. Consequently, there are missing records at the beginning of time series in several of the datasets, but this has no effect on the analysis.

When I began my doctoral research in 2021, I would have liked to include more recent data, but I chose to end the analysis in 2019. There is a time lag associated with the publication of many of the datasets. In particular, there was little runoff data available for 2020 and thereafter, as we will see in Section 2.5.1.

For experiments in record extension (predictions of past conditions), I created a 40 year long database of observations from January 1980 to December 2019. Because there were fewer satellites in orbit in the 1980s and 1990s, the data for the first two decades is more sparse.

2.1.3 Spatial Resolution and Format

I used a common spatial projection and resolution for all gridded geospatial data. It is based a common equirectangular projection, or plain geographic latitude and longitude projection (sometimes called *plate carrée*). This projection is commonly used in the Earth sciences due to its simplicity, although it has certain tradeoffs, i.e., distortion of areas near the poles. One set of experiments was done with data at a spatial resolution of 0.25 degrees. The global grid consists of 720 rows and 1440 columns. Another set of experiments was done with data at a resolution of 0.5 degrees. This format is similar to the "climate modeling grid" (CMG) used by many climate researchers (Modis Land Team, 2021), although it is at a coarser resolution and thus less detailed. In this thesis, I refer to grid cells and pixels interchangeably, and these should be understood as referring to the same concept: a set of regular rectangular areas on the earth's surface.

The size of grid cells varies with latitude. Near the equator, each 0.25° grid cell is about 27 km on a side and has an area of about 770 km^2. As one moves north or south away from the equator and toward the poles, grid cells get smaller. At Rome's latitude of 42° N, grid cells are about 21 km wide (the height is unchanged) and have a surface area of about 576 km^2. The northernmost basins considered in our study (in Canada, Alaska, Norway, Finland, and Russia), extend to latitudes above 70° N, where a grid cell is less than 10 km wide and has an area of less than 270 km^2, only 35% of the size of a grid cell near the equator.

Some of the datasets I used were published at a higher spatial resolution, for example 0.05°. In such cases, I upscaled the data to decrease its spatial resolution and to make it consistent with the other datasets. Water storage data from the GRACE satellites are the limiting factor, with data products published at 0.25°. Methods are available to "downscale" these data to a higher spatial resolution, but scientifically, there is little to be gained from doing so, and any additional accuracy would be illusory.

2.1.4 Data Formats

Strictly speaking, the project database is not a single electronic file, but a collection of electronic data stored in an organized file structure with appropriate metadata, or descriptive details. Data were stored as floating-point numbers with double precision in Matlab `.mat` files. This is not necessarily the most efficient data format in terms of file size or storage space on disk, however, it eliminates the overhead of converting data or changing units. Matlab binary data files (.mat) can be read by Python, R, or other scientific software, often with the use of a library or plugin.

Note that I do not have the rights to republish all the data that I have collected. For example, with regards to river discharge data, the user agreement from the GRDC does not allow one to redistribute the data. However, my understanding is that one may share and distribute derivative works. This includes versions of the data that have been transformed or processed, such as monthly averages. I have created a public repository containing the database used in the modeling described herein. This should be useful to other researchers who wish to verify or replicate these results or conduct related research on the global water cycle. It is freely available at:

https://doi.org/10.5281/zenodo.8101659

Table 2.1: Datasets compiled for the four major components of the water cycle

Dataset	Begin	End	Temporal resol.	Spatial resol.	Citation	Download Link
Total Water Storage Anomaly						
GRACE-CSR	2002	present	month*	1.0°	Save (2020)	http://www2.csr.utexas.edu/grace/RL06_mascons.html
GRACE-JPL	2002	present	month*	1.0°	Landerer and Cooley (2021)	https://grace.jpl.nasa.gov/
GRACE-GSFC	2002	present	month*	1.0°	Loomis et al. (2019)	https://earth.gsfc.nasa.gov/geo/data/grace-mascons
Precipitation						
GPCP v2.3	1979	present	day	2.5°	Adler et al. (2018)	https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C00979
GPM-IMERG	2000	present	day	0.10°	Huffman et al. (2020)	https://disc.gsfc.nasa.gov/datasets/GPM_3IMERGDF_06/summary
MSWEP	1979	present	day	0.10°	Beck et al. (2019)	http://www.gloh2o.org/mswep/
Evapotranspiration						
GLEAM v3.5a	1980	present	day	0.25°	Martens et al. (2017)	https://www.gleam.eu/
GLEAM v3.5b	2003	present	day	0.25°	idem	https://www.gleam.eu/
ERA5	1950	present	3 hour	0.25°	Hersbach et al. (2018)	https://cds.climate.copernicus.eu
Observed evapotranspiration for evaluation						
FluxNet 2015	2002	2010	hour	point	Pastorello et al. (2020)	https://fluxnet.org/data/fluxnet2015-dataset/
Observed River Discharge – in situ						
GRDC	varies	varies	day	gage	BfG (2020)	https://www.bafg.de/GRDC
Australia	1970	2020	day	gage	Australia BOM (2020)	http://www.bom.gov.au/water/hrs/index.shtml
GSIM	varies	2016	month	gage	Gudmundsson et al. (2018)	https://doi.pangaea.de/10.1594/PANGAEA.887470
Runoff – synthetic						
G-RUN	1902	2019	monthly	0.5°	Ghiggi et al. (2021)	doi.org/10.6084/m9.figshare.12794075

2.1.5 Excluded Datasets

The Appendix contains a list and brief descriptions of dozens of EO datasets. There are many datasets in the literature that I chose *not* to use in my analysis. This includes well-known remote sensing datasets, including some that are considered state of the art, and which are widely used and highly-cited within the earth science community. In some cases, I excluded a dataset because it had inadequate temporal coverage, i.e., it was not available over the period from 2000 to 2019. For example, the SM2RAIN precipitation dataset (Massari et al., 2020) begins in 2004, while my goal was to maximize the overlap with GRACE observations, which began in April 2002. I excluded some datasets for insufficient geographic coverage. For example, CMORPH precipitation (P. Xie et al., 2019) covers latitudes 60°S to 60°N, yet many of our study's river basins are above 60°N and therefore outside of its coverage. In other cases, I excluded a dataset for insufficient temporal coverage.

It is worth briefly discussing what can be gained from adding more data to the analysis. Oftentimes, more data is better. This attitude is especially prevalent in the machine learning community, where large models are fed with massive amounts of data. However, the research described in this thesis focuses on method development. Indeed, it is possible to use the methods described here to create optimized water cycle components using the largest possible amount of available data, perhaps even customizing the inputs by region based on data availability.

2.2 Precipitation

Below, I describe the precipitation datasets I included in my analysis of the water cycle. A longer list of datasets that I considered, but did not include, is in Appendix A.

2.2.1 GPCP

The Global Precipitation Climatology Project (GPCP) is a global gridded dataset produced by an international consortium of researchers. It is based on multiple satellite observations that have been merged to estimate precipitation at the global scale (Adler et al., 2018). It has a long record, beginning in 1979. However, it also has a coarse spatial resolution, at 2.5°. GPCP is a well-known dataset that has been used for many studies and referenced in thousands of journal articles (Adler et al., 2018).

GPCP combines both passive microwave and infrared satellite data, as well as in situ data. Each class of data brings certain advantages. For instance, passive microwave data can penetrate clouds and provide precipitation estimates even in cloudy regions, while infrared data offers high spatial resolution and frequent updates due to geostationary positioning, albeit at a lower resolution. The authors also corrected for systematic wind-induced undercatch and wetting losses from rain gauges, as well as for orographic effects. I used version 2.3 of this dataset, which was updated in 2018.

2.2.2 GPM-IMERG

GPM-IMERG is the current multi-satellite precipitation product from NASA (Huffman et al., 2019). It stands for "Integrated Multi-satellitE Retrievals from the Global Precipitation Monitoring (GPM) satellite. GPM-IMERG offers several advancements and improvements over its predecessor TMPA (see Appendix). GPM-IMERG combines measurements in passive microwave and infrared from over a dozen different satellites, using morphing techniques and a Kalman filter, to provide accurate satellite-based precipitation estimates, supplemented by precipitation gauge analyses. It ingests data from the GPM core observatory satellite, which contains a dual-frequency precipitation radar and the GPM microwave imager. The dual-band precipitation radar provides better estimates of precipitation particle sizes and covers a wider range of precipitation rates compared to the single-band radar on the TRMM satellite (Y. Wang & Wu, 2022).

2.2.3 MSWEP

The Multi-Source Weighted-Ensemble Precipitation (MSWEP) is not a pure remote sensing product but an "optimal merging" of gage observations, satellite observations, and reanalysis model output (Beck et al., 2019). This dataset offers high spatial resolution of 0.1° and a maximum temporal resolution of 3 hours, which allows for a detailed analysis of precipitation patterns. MSWEP has been shown by the authors to exhibit more realistic spatial patterns in mean, magnitude, and frequency compared to other precipitation datasets (Beck et al., 2019). They also state that it provides more accurate precipitation estimates in mountainous regions, where other datasets tend to underestimate precipitation amounts. Because of its good spatial and temporal coverage, I chose MSWEP as an input in the analysis.

2.2.4 CPC Global Precipitation

The CPC Global Unified Gauge-Based Analysis of Daily Precipitation is a dataset of *rain gage measurements* (not remote sensing observations) published by NOAA's Climate Prediction Center (CPC). I include a brief description here, as it is a gridded dataset that is widely used in large-sample hydrology. This dataset covers 1979 to present at 0.5° resolution, with global coverage of land surfaces (not oceans). The goal of the project was to create a suite of unified precipitation products with consistent quantity and improved quality. This was achieved by combining all information sources available at CPC and using an optimal interpolation technique. The authors state that the dataset was more accurate than existing gage-based datasets available at the time, but that key uncertainties remain. In particular, there may be inconsistencies over regions where station networks changed, and less accuracy in places with high spatial variability in precipitation, such as in mountainous regions.

The methods behind this dataset are documented in a set of three articles:

- Interpolation algorithm: P. Xie et al. (2007).
- Gauge Algorithm Evaluation: (M. Chen et al., 2008)
- Construction of the Daily Gauge Analysis: M. Chen and Xie (2008)

I used the CPC dataset as an an additional source of data for evaluating the results of this reasearch, to be described in Section 5.4.5.

2.2.5 GHCN

This section describes a dataset of station observations which I used as an additional source of data to evaluate my results. Unlike the other datasets described here, it is not in grid format, nor is it based on remote sensing.

The Global Historical Climatology Network (GHCN) is a dataset reporting weather and climate variables at over 120,000 stations worldwide (Menne et al., 2012). Many stations only report precipitation, but many others report other variables such as temperature, pressure, humidity, cloud cover, etc. The dataset is maintained by the United States National Oceanic and Atmospheric Administration, National Centers for Environmental Information (NOAA NCEI).

The publisher provides a data inventory, containing the start and end dates for stations. I used this to select candidate stations with data from 2002 to 2019. This still left over 21,000 stations. I used Python scripts to download and process these data, calculate monthly averages, and to create a more detailed inventory of stations with precipitation data over the period of my analysis. The data provider has performed extensive and well-documented quality control on these data (Durre et al., 2008; Durre et al., 2010). Nevertheless, there are several subtleties and complications to dealing with precipitation observations. Examples include how to handle quality flags, what to do with precipitation logged as "trace,", and how to deal with missing daily data.

I used the following rules when compiling monthly precipitation for stations in the GHCN catalog. First, I only included stations where there are at least 60 months (5 years) of data. I only calculated the monthly average P where there are at least 25 days of data. I chose not to throw out months simply because there were a few missing daily observations. I handled missing daily data by assuming the rainfall on that day was the same as the average of other days in the same month. For example, if station data for January had 25 days of valid data, and 6 missing daily observations, the monthly precipitation is: $P_{\text{month}} = \sum P_{\text{daily}} \times \frac{31}{25}$.

There are more complex and sophisticated ways of filling in missing data (also referred to as imputation), but these were not considered here. The GHCN station data will be used for a secondary analysis, to verify that our modeling methods have not degraded the signal in EO datasets too much related to observations. The check against GHCN station data is just one of several evaluations of this kind. Therefore, it was not worth spending a great deal of time performing sophisticated analyses with these data.

The final set of GHCN station data for precipitation encompasses 21,880 stations. The distribution is highly uneven (Figure 2.2(a)). In Africa, South America, there are large blank spaces on the map. In Section 5.4.5 of this thesis, I compare calibrated P at the pixel scale to GHCN observed P. However, the geographic coverage of observations is extremely uneven. As shown in the Figure 2.2(b), there are many 0.5° grid cells where there are no stations at all, and others there are many pixels where there are a dozen or more stations, particularly in Germany, the Netherlands and other northern European countries and the United States (not shown).

2.3 Evapotranspiration

As shown in Table 2.1, I used 3 EO datasets of evapotranspiration for the data integration, and observations from flux towers for comparison and evaluation of my results. Each data source is listed in Table 2.1 and described in more detail below.

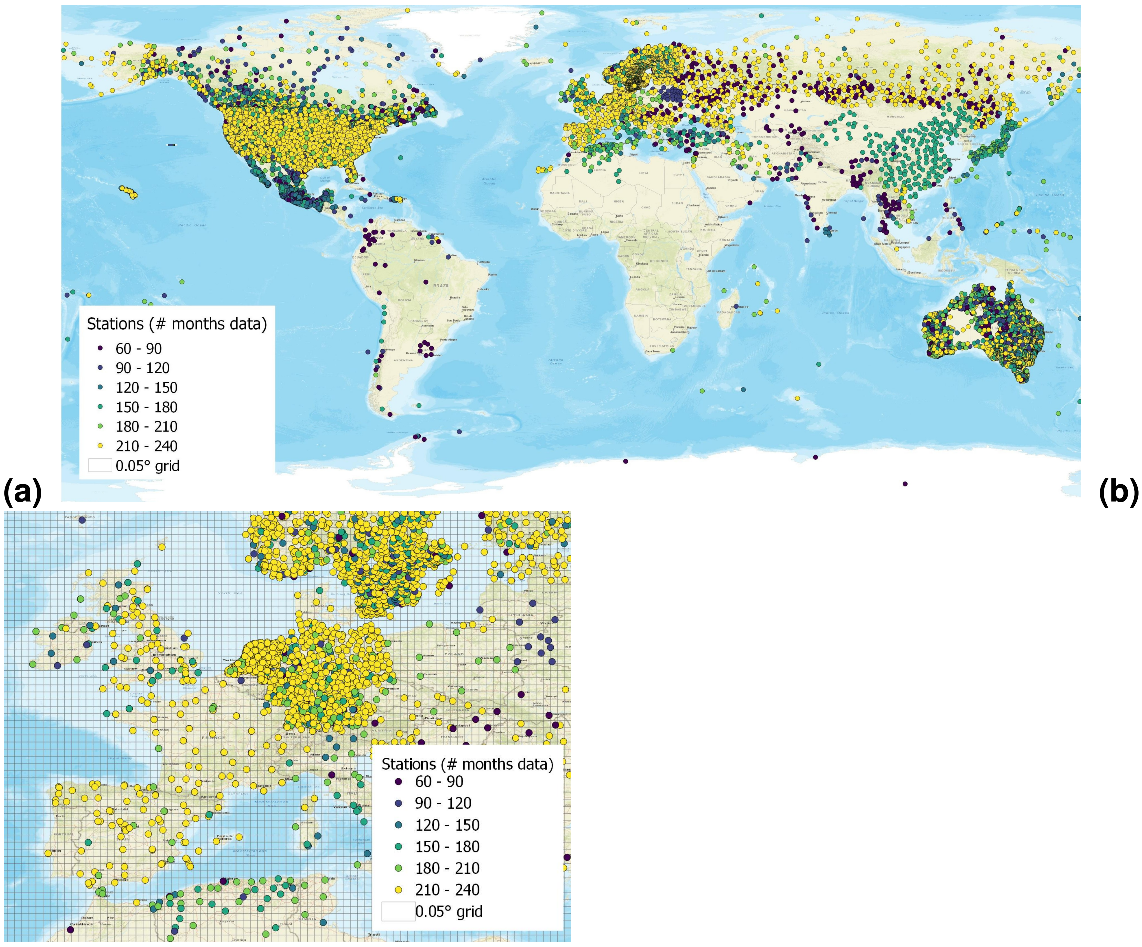


Figure 2.2: Selected stations providing precipitation data from the Global Historical Climatology Network (n = 21,880)

2.3.1 ERA5 Evapotranspiration

I also included a dataset that is not a purely remote-sensing based product, but based on the assimilation model ERA5, from the European Centre for Medium-Range Weather Forecasts (ECMWF). The model combines historical estimates (from both remote sensing and in situ observations) using an advanced modeling and assimilation system. ERA5 produces many variables describing the atmosphere, land, and ocean, at a resolution of up to a 30 km grid (Guillory, 2022). ERA5 estimates of E have been used in many recent hydroclimatic studies (see e.g., Tarek et al., 2020; Singer et al., 2021; Lu et al., 2021). ERA5 hydrological parameters have hydrological parameters that have effective units of "m of water per day" and so I multiplied by 1000 to convert to $\text{kg} \cdot \text{m}^{-2} \cdot \text{day}^{-1}$ or mm/day, and then multiplied by the number of days in each month to obtain units of mm/month.

2.3.2 GLEAM

The Global Land Evaporation Amsterdam Model (GLEAM) is a set of algorithms that estimates the various components that contribute to total evapotranspiration (Martens et al., 2017; Miralles et al., 2011; Hersbach et al., 2018). The authors used an empirical relationship, the Priestley-Taylor equation, to calculate potential ET based on satellite observations of surface net radiation and near-surface air temperature. GLEAM version 3.5a used reanalysis rather than satellite observations and covers 1980 to present. The updated version 3.5b relies more on remote sensing data and has a more limited temporal coverage of 2003 to present. GLEAM has global coverage of land surfaces at 0.25° resolution and a daily time step. Note: the latest version, GLEAM v3.6 was published in September 2022, after I completed much of the analysis described here.

GLEAM includes 10 components:

1. Actual Evaporation (E)
2. Soil Evaporation (Eb)
3. Interception Loss (Ei)
4. Potential Evaporation (Ep)
5. Snow Sublimation (Es)
6. Transpiration (Et)
7. Open-Water Evaporation (Ew)
8. Evaporative Stress (S)
9. Root-Zone Soil Moisture (SMroot)
10. Surface Soil Moisture (SMsurf)

2.3.3 Observed Evapotranspiration at Flux Towers

I used in situ data for validating the results of certain analyses. Several methods have been developed to measure evapotranspiration over land surfaces. The two main methods for measuring evapotranspiration rates at specific locations are with devices called lysimeters or via micrometeorological techniques (Healy et al., 2007). Micrometeorological techniques include eddy correlation, Bowen ratio/energy budget, and aerodynamic profile methods, which involves measuring the vertical flux of water vapor from the land surface to the atmosphere. While such methods are accurate, they are also expensive and require frequent maintenance.

I obtained in situ measurements of ET from flux towers, which belong to a class of micrometeorological measurement devices. Flux towers are outfitted with sensors and instruments mounted at various heights for measuring the vertical flux of water vapor from the land surface to the atmosphere. I selected data for 117 towers from the FluxNet2015 dataset, which compiles data from 212 global towers (Pastorello et al., 2020). Of the 212 stations, 34 did not contain data for latent heat flux, required for inferring evapotranspiration. After quality checking these data, I eliminated another 61 stations, either because the time record did not overlap our period of interest (2002 - 2019), was too short (less than 2 years), or there were other quality issues. This left us with 117 stations, with an average time series length of 5 years. The majority of selected towers are in Europe (51 towers) or North America (46), with fewer in Africa (2), Asia (6), Australia (9), and South America (3). Figure 2.3 shows the location of the selected flux towers.

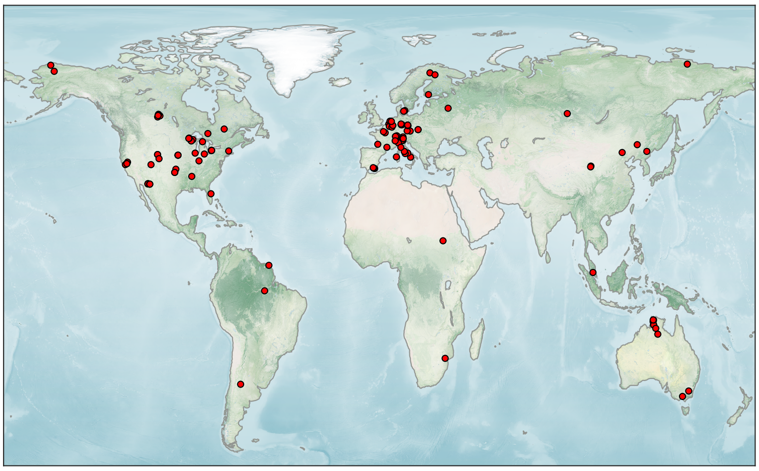


Figure 2.3: Map of the selected flux towers used in this study for their measurements of evapotranspiration.

The FluxNet database reports latent heat flux, in units of W/m². This can be converted to evapotranspiration for inclusion in the water balance model. The *latent heat of vaporization* of water is the energy required for water molecules to transition from liquid to gas, and varies according to water temperature as given by Shuttleworth (1993):

$$\lambda = 2501 - 2.361\, T_s \quad \left(\mathrm{J} \cdot \mathrm{kg}^{-1} \right) \tag{2.4}$$

As a simplification, a value of 2,450 J/g is commonly used, and has an error less than 2% over temperatures from 0° to 35°C. Therefore, the latent heat flux in W/m² can be converted to monthly average evapotranspiration as follows:

$$\begin{aligned}\frac{\mathrm{W}}{\mathrm{m}^2} = \frac{\mathrm{J}}{\mathrm{s}\cdot\mathrm{m}^2} \times \frac{\mathrm{g}}{2,450\,\mathrm{J}} \times \frac{\mathrm{cm}^3}{1\,\mathrm{g}} \times \left(\frac{1\,\mathrm{m}}{100\,\mathrm{cm}}\right)^3 \times \frac{1,000\,\mathrm{mm}}{\mathrm{m}} \times \frac{86,400\,\mathrm{s}}{\mathrm{day}} \\ = 0.0353\frac{\mathrm{mm}}{\mathrm{day}}\end{aligned} \tag{2.5}$$

Care must be taken in comparing E observed at flux towers to gridded hydroclimatic data. The challenge is in resolving the difference in scale differences between in situ data and grid cells. Flux tower observations are point estimates, taken at a single geographic location. One may easily find the pixel that intersects this point on a map to compare gridded data to the tower observations. Yet, the value in a pixel represents an average over an area, over which conditions may vary widely. At the scale of our model grid, a single 0.5° pixel has an area of about 3,000 km² near the equator. The land cover, vegetation, and topography over a grid cell may be substantially different from that of the flux tower site, making it challenging to generalize, and reducing the accuracy of comparisons between modeled fluxes and observations from flux towers. Nevertheless, such comparisons are still useful, and it is encouraging when a model recreates the observed temporal dynamics, even when the magnitudes of observed and modeled fluxes are not comparable.

2.4 Total Water Storage Change

Information on total water storage (TWS) for this research comes from the GRACE satellites. The first pair of satellites were in operation from 2002 to 2017, and a follow-on mission began in 2018.

It is convenient to refer to the *change* in water storage as equivalent to a flux, and it can be expressed in the same units as any other flow. GRACE provides the monthly TWS *anomaly*, expressed as land water equivalent (LWE) thickness surface mass anomaly, in units of cm or mm. GRACE does *not* estimate the total volume or mass of water in a region, but rather its *change* with respect to a historical baseline average. Nevertheless, the observations encompass water in all its forms and "represent the full magnitude of land hydrology and land ice" (Landerer, 2021).

Compared to precipitation and the other hydrologic variables, GRACE observations have a lower spatial and temporal resolution, limiting our analysis to a monthly time step.

Scientists at various institutions have developed different algorithms and sets of parameters for converting the measurements of the gravity field measured by

GRACE into estimates of TWSC. I obtained four GRACE data products, described in Table 2.2, and ultimately selected three of them for the analysis (from CSR, GSFC, and JPL). Each of these are Level 3 solutions[1] for LWE as measured by GRACE satellites. All the available datasets of TWS are nearly global (–89.5° to +89.5°), land surface only.

Among these three solutions shown in Table 2.2, there are differences in the algorithms and processing steps, which can produce slightly different results, particularly at regional scales. There are differences in the methods used to filter the data to remove noise, including atmospheric and oceanic variability. There are also differences in the methods and models used for post-processing. Such processing increases the accuracy of TWS anomalies by adjusting for changes in land surface due to seismic events or the isostatic rebound in regions formerly covered by glaciers that continue to uplift thousands of years after the glaciers have melted or receded. It is common for researchers to use multiple solutions to validate their findings and reduce the impact of any individual solution's uncertainties. Indeed, our approach relies on a neural network to extract the best information from each dataset.

I explored using the GRACE solution produced by GFZ, the German Research Centre for Geosciences in Potsdam (GeoForschungsZentrum), described by Kusche et al. (2009). However, I determined that this solution was less suitable than the three newer solutions. Original GRACE solutions were based on spherical harmonics are mathematical functions that represent the variations in the Earth's gravitational field. This method is useful for capturing large-scale features such as the overall distribution of mass and major geophysical phenomena. Conventional spherical harmonic solutions "typically suffered from poor observability of east-west gradients" (Watkins et al., 2015), resulting in pronounced north-south striping patterns in the data. These were usually removed by smoothing the data, which unfortunately causes some loss in signal. Further, the spherical harmonic method does not effectively account for localized variations in mass, especially in regions with strong mass anomalies.

The more recent GRACE solutions are based on calculations based on mass concentrations or *mascons*. Mascons are anomalies in the distribution of mass within a planet. Here, mascons refer to the regions on Earth where there are significant variations in the distribution of mass, which has a corresponding effect on the Earth's gravitational field. Such variations complicate the analysis of

[1] The levels refer to the amount of processing that has been done to the data. Level 1 is the raw satellite data. Level 2 has been processed to determine the gravity field. Level 3 is further processed to estimate changes in the amount of water in an area.

Table 2.2: GRACE datasets available for total water storage

Data Set	GRACE-CSR	GRACE-GFZ	GRACE-GSFC	GRACE-JPL
Units	m	m	m	m
Publisher	Univ. of Texas at Austin, Center for Space Research	Geoforschungs-Zentrum, Potsdam, Germany	NASA Goddard Space Flight Center	NASA Jet Propulsion Laboratory
Begin	2002-Apr-05	2002-Apr-04	2002-Apr-05	2002-Apr-05
End	near present	2017-Jun-29	near present	near present
Latitudes covered	–89.5º to +89.5º	–89.5º to +89.5º	–89.5º to +89.5º	–89.5º to +89.5º
Temporal resolution	quasi-monthly	quasi-monthly	quasi-monthly	quasi-monthly
Spatial resolution	0.25°	1.0°	0.5°	0.5°
Notes	"The data are represented on a 1/4 degree lon-lat grid, but they represent the equal-area geodesic grid of size 1x1 degree at the equator, which is the current native resolution of CSR RL06 mascon solutions."	RETIRED. No longer in production after mid 2017. This solution was based on the the conventional spherical harmonic method only, while the newer algorithms use mass concentration grids (mascons).	"This product is comparable to the JPL and CSR mascon products."	Data distributed at resolution of 0.5°, but the 3° mascons on which this solution is bas are evident when plotting.
Website	http://www2.csr.utexas.edu/grace/RL06_mascons.html	https://podaac.jpl.nasa.gov/dataset/TELLUS_GRAC_L3_GFZ_RL06_LND_v04	https://earth.gsfc.nasa.gov/geo/data/grace-mascons	https://grace.jpl.nasa.gov/
Citation	Save et al. (2016), Save (2020)	Landerer and Swenson (2012)	Loomis, et al. (2019)	Wiese et al. (2018)

GRACE data because they can distort the measurements of surface mass changes. The most recent techniques employ a gravity field basis function to separate the contributions in the signal from unequal distribution of the earth's mass from other factors such as water storage variations.

Guidance published by the GRACE science team states that the mascon-based solutions (CSR, GSFC, JPL) are better for use in hydrology and related fields like glaciology and oceanography, as they are more accurate, especially at smaller scales. The mascon gridded data products are recommended as they "suffer less from leakage errors than harmonic solutions, and do not necessitate empirical filters to remove north-south stripes, lowering the dependence on using scale factors to gain accurate mass estimates" (NASA Jet Propulsion Laboratory, 2020).

Figure 2.4 illustrates the increase in resolution that was obtained with the mascon-based solutions. Here, Save et al. (2016) used three different GRACE solutions to calculate the trend in total water storage. The newer mascon-based solution from CSR is shown in (a). In (b), the older GFZ solution (labeled TELLUS) is shown. Because of the post-filtering that was applied to eliminate striping, small-scale anomalies are smoothed and details are blurred. In (c), the JPL solution, one can see changes between neighboring 3° mascons.

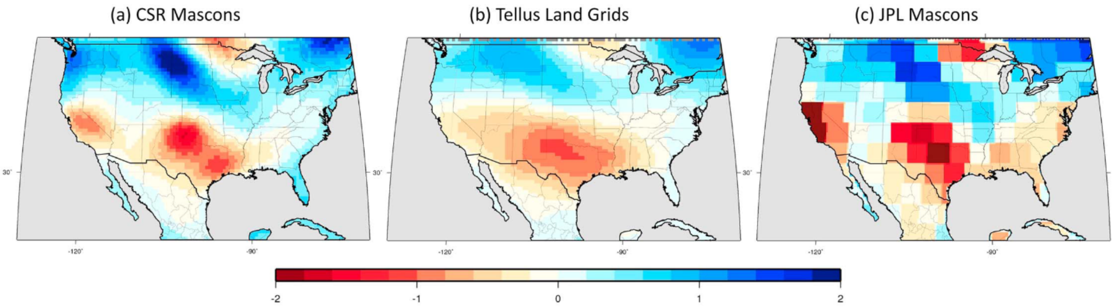


Figure 2.4: Trends in total water storage based on three different GRACE solutions. Reprinted from Save et al. (2016), Figure 7 (Creative Commons licensed).

While all three of the GRACE datasets I chose are based on the same satellite data, the solutions are based on different mascon grids. The details are complex, but interested readers are referred to detailed explanations in Loomis et al. (2019), Save et al. (2016), and Watkins et al. (2015). Official guidance from NASA has recommended that users average all three data center's solutions (Landerer & Cooley, 2021). This recommendation is backed by research that showed the ensemble mean (simple arithmetic mean of JPL, CSR, GFZ fields) led to a reduction in noise within the gravity field solutions, considering the scatter present in the available data (Sakumura et al., 2014).

And while the calculations for converting gravity field anomalies to changes in water storage have a solid basis in theory, researchers have not yet found an effective way to ground truth these observations (Kusche et al., 2009; Reager et al., 2015). GRACE observations have a relatively coarse spatial resolution compared to other EO datasets, such as precipitation. While the published datasets have a 0.25° resolution, the data have been spatially filtered to remove random errors and systematic errors (Landerer & Swenson, 2012). Thus, small scale details are not likely to be meaningful, and these data are more accurate over larger regions (Tapley et al., 2004).

In the following sections, I show some exploratory data analysis of the GRACE observations.

2.4.1 Analysis of Missing Data

The GRACE Level 3 data of TWS anomalies are geographically complete, with no gaps or missing pixels that are often seen in other EO datasets. However, there are gaps in the time series, where there are months with no data. In this section, I describe the causes for missing data. In Section 3.2.2, I describe a time series interpolation method to fill in missing data.

To understand when and why there are gaps in the GRACE dataset, it helps to know about the mission's history and the timeline of its operations. Figure 2.7 is an overview of available GRACE data. The GRACE satellites were launched in March 2002, and were originally designed for a 5-year lifetime. The first monthly data were published for April and May 2002. "Several months of instrument testing" resulted in no reported water storage data in June and July 2002, and two missing months of data in 2003 and 2004 (Herman et al., 2012). For the next 7 years, there is a continual record with no gaps.

In 2011, the batteries onboard GRACE began to fail. As a result, "the instruments have to be powered off during the maximum eclipse season, thus interrupting the nominal science data flow every 161 days for a period of approximately 3–4 weeks" (Flechtner et al., 2014). As a result, there are 1 to 2 month gaps every 6 months. The worsening condition of batteries resulted in longer data gaps beginning in September 2017 (Müller et al., 2019). GRACE was decommissioned in October 2017. With no power left to correct the orbits, the low-earth orbiting satellites were left to re-enter the atmosphere.

The follow-up mission, GRACE-FO, was launched in May 2018, and the first GRACE-FO data was reported for June 2018. From July 19 to October 16, 2018, an onboard instrument did not function properly, resulting in a gap in science data collection (Svarovsky, 2019). Other than this 3-month gap, the record has been continuous from mid 2019 to present. The time period chosen for this thesis is 2000 to 2019, with 20 years or 240 months. In total, there are monthly GRACE data for 181 months during this time period, or 75% of months. In Chapter 6, I examine the possibility of reconstructing GRACE-like TWS data for filling gaps and hindcasting for periods prior to 2002.

2.4.2 Temporal Resolution of GRACE

GRACE data are quasi-monthly. That is, published observations of TWS do not correspond precisely to calendar months, in terms of the effective beginning and end date of the observations. According to the data provider, NASA's Jet Propulsion Laboratory (JPL), "the generation of high-quality monthly gravity field solutions and mass change grids requires accumulation of satellite-to-satellite tracking data for about 30 days. However, within some months, a few days of data may not be used due to instrument issues, calibration campaigns etc." (NASA Jet Propulsion Laboratory, 2023). According to JPL, data users should "carefully assess whether the underlying sub-monthly sampling needs to be matched between the data sets." In other words, because the GRACE data are not truly monthly, this could lead to errors and uncertainties when trying to combine GRACE with other monthly datasets, as there is something of a time mismatch. Examples of the errors in GRACE observation dates is shown in Table 2.3.

Table 2.3: GRACE observations begin and end dates, showing the errors in the begin date and duration compared to calendar months.

GRACE observation number	Calendar month	GRACE observation begin date	GRACE observation end date	Observation duration (days)	Calendar month duration (days)	Error in duration (days)	Error in begin date (days)
1	2002-04	2002-04-04	2002-04-30	26	30	-4	3
2	2002-05	2002-05-02	2002-05-17	15	31	-16	1
3	2002-08	2002-07-31	2002-08-31	31	31	0	-1
4	2002-09	2002-08-31	2002-09-30	30	30	0	-1
5	2002-10	2002-09-30	2002-10-31	31	31	0	-1
6	2002-11	2002-10-31	2002-11-30	30	30	0	-1
7	2002-12	2002-11-30	2002-12-31	31	31	0	-1
8	2003-01	2002-12-31	2003-01-31	31	31	0	-1
9	2003-02	2003-01-31	2003-02-28	28	28	0	-1
10	2003-03	2003-02-28	2003-03-31	31	31	0	-1
11	2003-04	2003-03-31	2003-04-30	30	30	0	-1
12	2003-05	2003-04-30	2003-05-21	21	31	-10	-1
13	2003-07	2003-06-30	2003-07-31	31	31	0	-1
14	2003-08	2003-07-31	2003-08-31	31	31	0	-1
...							

Ideally, observations would be aligned with calendar months, but as we have seen this is not always the case. One suggested workaround for the issue is to average other hydrologic data at the same time scale, i.e., calculating the average based on the same days used by GRACE. There are advantages and disadvantages to this approach. The advantage is that it reduces one source of uncertainty, and makes the data more directly comparable. The disadvantage is that the results will be more difficult to interpret and to explain to others. For example, one would need to state that a result is for January 6 to February 3, 2017 rather than simply reporting January 2017. Another disadvantage is that certain datasets are not available at daily time step, and therefore could not be used in the analysis. Notably, two of the important discharge datasets used in my research, GSIM and GRUN, are published at the monthly time step only.

I analyzed the errors in the duration and start date of the quasi-monthly GRACE observations, with results shown in Figure 2.5. The histogram on the left shows the errors in the duration of the GRACE observation compared to the calendar month. For example, the first GRACE observation in April 2022 began on April 4, and ended April 30. This observation has a duration of 27 days, while the calendar month has 30 days, for a duration error of −3 days. The mode of the duration error is +1 day, which occurs in 142 of the 180 observations between 2002 and 2019. The duration errors range from −17 days to +13 days, and have an average of 0.0 days.

Because GRACE observations are roughly aligned with calendar months, I assigned each observation to its closest calendar month, rather than performing some type of interpolation to estimate monthly values from the raw data. This would only further compound the already large uncertainties associated with these data. This same assumption is made by other researchers in the field of global hydrology that work with GRACE data (see e.g., Zaitchik et al., 2008; Landerer et al., 2010).

On the right in Figure 2.5 is a histogram of the errors in the start date for GRACE observations. The vertical axis of the plot has been truncated to show greater detail among the low-frequency errors. For example, the first GRACE observation begins on the 4th of the month, for an error in the begin date of +3 days. The mode for the begin date error is −1 day, which occurs in 152 of the 180 observations included in our analysis. The start date error has an average of −0.1 days, and ranges from −13 days to +20 days. I concluded, in consultation with my thesis advisor, to assume that GRACE is approximately monthly and to combine it with other true monthly data. As discussed above, the majority of observations are aligned to calendar months to within a day, in terms of both the begin date and duration. Most observations have a relatively small error in the start date or duration of the observation. Errors in timing will contribute some additional uncertainty to GRACE observations.

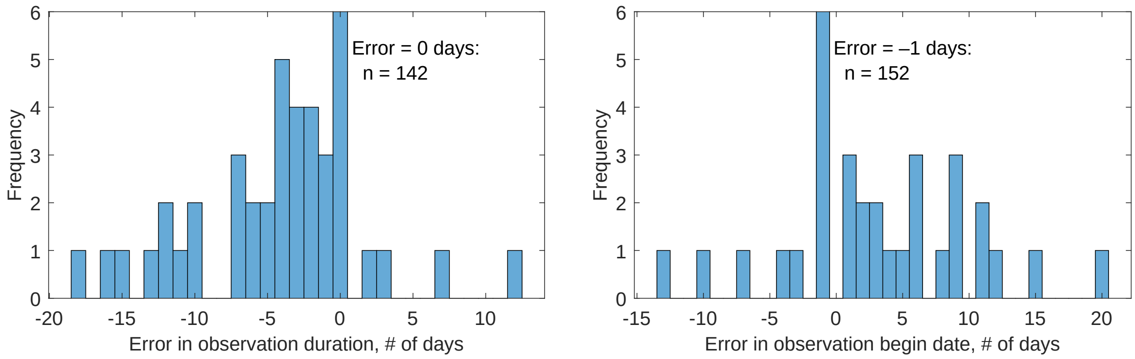


Figure 2.5: Errors in the begin date (left) and duration (right) of GRACE observations compared to calendar months

It was necessary to remove some GRACE observations, as they appeared to be largely redundant, overlapping other observations. Figure 2.6 illustrates the two observations I chose to delete. In these plots, the available GRACE observations are numbered beginning with 1 for the first quasi-monthly record in April 2022, using the same numbering as Table 2.3. The top plot in Figure 2.6 shows that observation number 110 covers the month of October 2011 quite well. The next observation, number 111, contributes little new information. It also does a poor job of representing November 2011, as the observation ends on November 15. Therefore, I decided to remove this observation from our input data. The bottom plot shows a similar situation for observation 145, which I also removed. The plots in Figure 2.6 also show that there are gaps in the GRACE dataset, with frequent missing months. The plots also show that the observations do not coincide perfectly with calendar months, as discussed above.

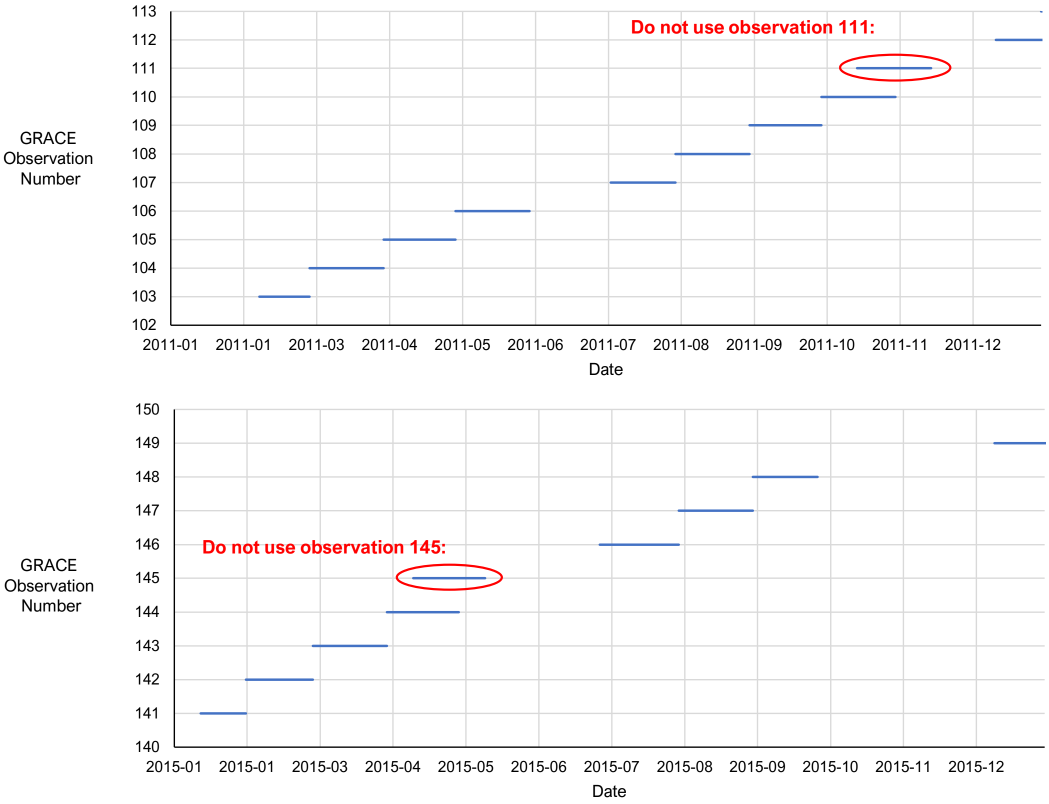
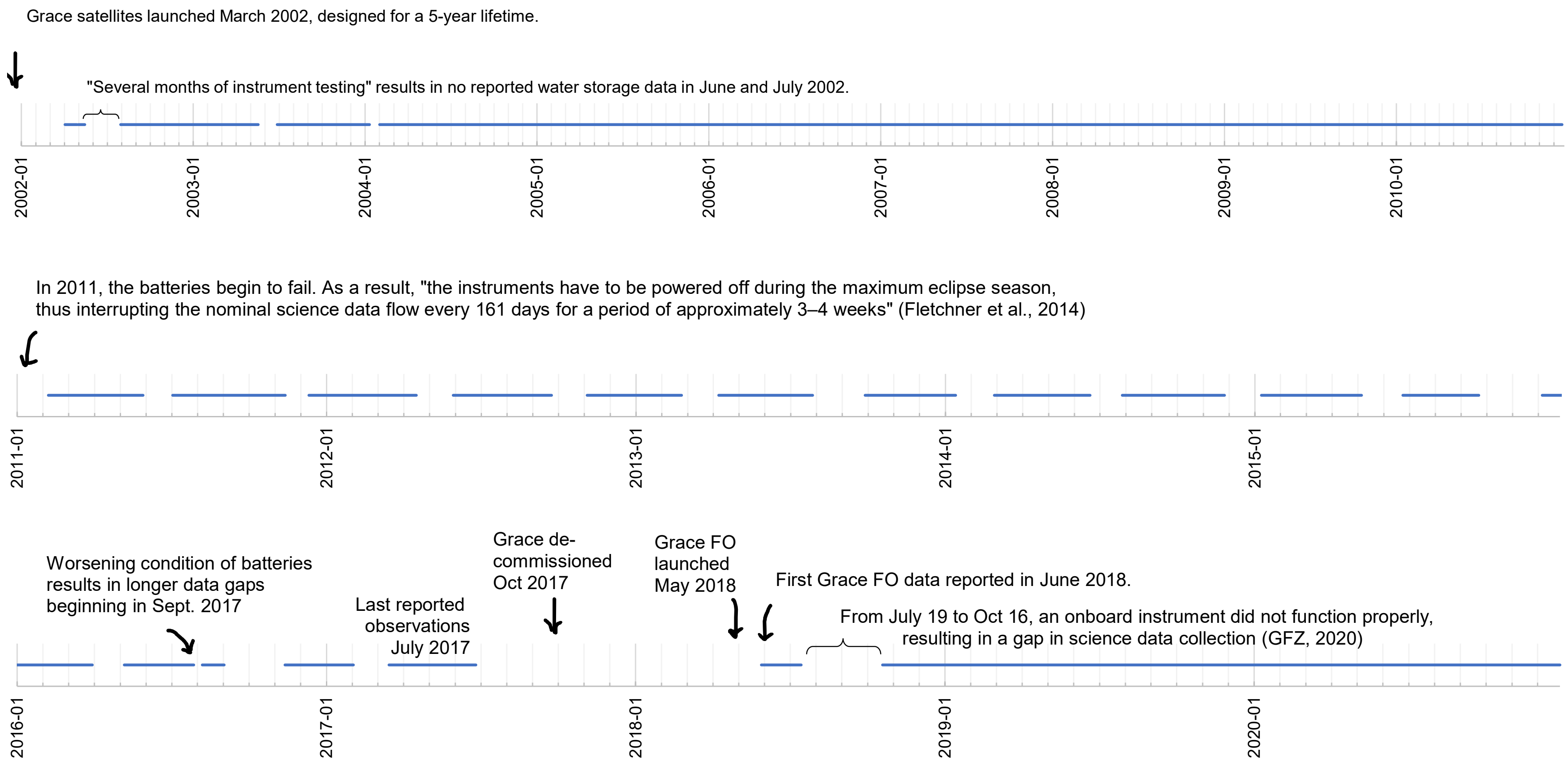


Figure 2.6: The time coverage of select GRACE observations, showing overlapping observations that were removed.



Grace satellites launched March 2002, designed for a 5-year lifetime.

"Several months of instrument testing" results in no reported water storage data in June and July 2002.

In 2011, the batteries begin to fail. As a result, "the instruments have to be powered off during the maximum eclipse season, thus interrupting the nominal science data flow every 161 days for a period of approximately 3–4 weeks" (Fletchner et al., 2014)

Worsening condition of batteries results in longer data gaps beginning in Sept. 2017

Last reported observations July 2017

Grace de-commissioned Oct 2017

Grace FO launched May 2018

First Grace FO data reported in June 2018.

From July 19 to Oct 16, an onboard instrument did not function properly, resulting in a gap in science data collection (GFZ, 2020)

Figure 2.7: Timeline of GRACE data availability and gaps.

2.4.3 Modeled TWSC

In order to compare and evalute the results of my research, I also gathered predictions of ΔS from recent modeling studies that reconstruct GRACE-like TWSC. Y. Zhang et al. (2018) used a land surface model and data assimilation techniques, first estimating the errors in each water budget component by comparison to in situ observations, then using a constrained Kalman filter to merge the datasets based on their error information, with a goal of minimizing the imbalance. This study produced global gridded datasets at 0.5° resolution, with monthly P, E, R, and ΔS for 1984–2010.

2.5 Runoff and River Discharge

Data on runoff provides the fourth and final flux in our simplified water balance (Equation 1.1. There is a long history of measuring the flow of water in rivers, also called discharge. Measurements of river levels were probably the first quantification of the water cycle. Sources describe the measurement of the water surface elevation by Egyptians in Pharaonic times as early as 1800 BC (National Research Council, 1991, p. 20). Other sources (e.g.: Pfister, 2018; Rosbjerg & Rodda, 2019) cite the work and writings of Leonardo da Vinci in the 1500s, who was among the first to measure river velocity and flow, and to speculate about the source of river flows.

In fact, modern methods for measuring streamflow have similarities to ancient methods. Measurements are made of flow velocity (m/s), and the cross-sectional area of the stream (m^2), and their product is a volumetric flow rate, in m^3/s. Because such measurements are difficult and time-consuming, hydrographers develop site-specific empirical relationships between water surface elevation, or "stage" and flow rate. In the near future, the scientific community expects to estimate river discharge based on high-accuracy measurements of water surface elevation and slope via the SWOT satellite launched in December 2022.(Durand et al., 2016; Prigent et al., 2016).

For this research, I used two types of runoff data: (1) estimated runoff from a statistical model, and (2) observations from the terrestrial monitoring of rivers. In the literature, the terms *runoff*, *river flow*, and *discharge* are sometimes used interchangeably. Usage also varies by discipline, often in subtle and confusing ways. Therefore, some clarification is in order.

River flow or **discharge** is an *in-situ* measurement, measured at *gages* (sometimes spelled *gauges*). River gages are typically installed and managed by national or regional governments or water management agencies. Gages use a variety of technologies to measure the instantaneous volumetric flow rate, expressed in units of volume per time such as m^3/s.

Runoff is all the water draining from a given land area. One component of runoff is overland flow – water from rainfall or snowmelt that does not infiltrate into the ground, and flows over the land surface. (Overland flow is also called *Hortonian flow*, after the pioneering 20th century American hydrologist Robert E. Horton.) Another component of runoff is subsurface flow, or the movement of groundwater out of a land area. Total runoff cannot be observed directly, but is sometimes monitored with tracers or estimated with mass balance approaches.

We follow Ghiggi et al. (2019) in assuming that, "at a monthly timescale the average catchment runoff can be assumed to equal the monthly streamflow [or river discharge] measured at the outlet divided by the catchment area." This approximation is valid when there is not a significant change in water storage during the month (such as in lakes or reservoirs), and there are no significant losses (such as withdrawals for irrigation or inter-basin transfers).

Under similar conditions (minimal change in storage and losses), one can calculate river discharge (for example for ungaged basins) from the runoff in the upstream area. Ghiggi et al. (2021) refer to "first-order river discharge estimates" obtained by calculating the spatial mean of the gridded runoff and over a basin and multiplying by the drainage area. This calculation is an inexact estimator of river discharge over a given time period, as it fails to consider the differing travel time of river flows in the basin. For example, a rainstorm that produces runoff in the headwaters of a large river basin may take days or weeks to reach the outlet. However, I follow Ghiggi et al. (2021) in assuming that, at a monthly timescale, the effect of water routing may be considered negligible except in very large basins.

2.5.1 Observed River Discharge

I sought to develop a large database of global gaged basins that would represent a range of geographic locations, environments, and basin sizes. The availability of river flow data is generally good across North America and western Europe, while observations are more scarce across much of Asia, Africa, and Latin America. Furthermore, in a troubling development for the field of large scale hydrology, river flow measurement has steadily decreased over the last few decades (Fekete et al., 2012).

I selected gages for our analysis based on data quality, geographic coverage, and location. I considered gages with a watershed area of 2,500 km^2 or greater. This corresponds to about 8 pixels in our 0.25° grid at mid-latitudes around 50°. This minimum threshold ensures that EO data are averaged across several grid cells, so that our estimates of basin-scale hydrologic fluxes are more robust. This is a particular concern for the water storage change datasets derived from the GRACE satellites which have a lower spatial resolution than many remote sensing datasets in the meteorological and hydrological sciences. I selected gages with a minimum temporal coverage of at least 6 observations during the period from 2002 to 2019. I also performed quality control of the runoff observations through a variety of plots and statistical summaries.

Figure 2.8 illustrates the location of the 2,056 river gages selected for the analysis.

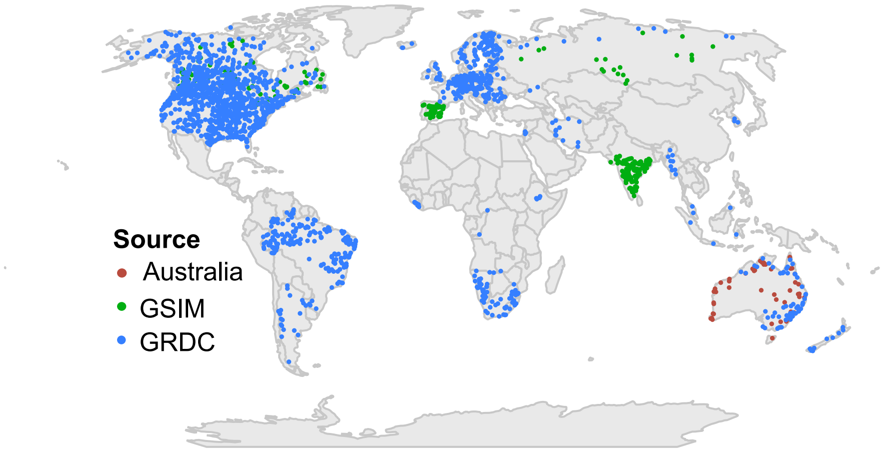


Figure 2.8: Map showing the location and data source of the 2,056 river flow gaging stations used in this analysis.

I obtained runoff data from 3 sources. First, I selected 1,737 gages from the Global Runoff Data Center (GRDC) and supplemented it with information from two other sources to fill in white spaces on the map (notably Asia and Australia). The GRDC, operating under the auspices of the World Meteorological Organization (WMO), is housed and operated by the German Federal Institute of Hydrology (BfG) WMO (1989). Their runoff database contains “historical mean daily and monthly discharge data and currently comprises river discharge data of well over 10,000 stations from 159 countries” (BfG, 2020). The GRDC database contains 10,361 stations, however the majority of these did not fit our criteria for spatial and temporal coverage. The GRDC acts as a clearinghouse for data but does *not* perform quality control. According to the GRDC website, “ownership of the data and responsibility for errors is with the data providers.” Among the problems I encountered were duplicate gages, records with few observations or implausible values, and abrupt shifts in the magnitude of flow.

Second, I obtained data for 272 gages from the Global Streamflow Indices and Metadata (GSIM) archive (Do et al., 2018; Gudmundsson et al., 2018). The creators of the GSIM database have expanded upon GRDC's database by adding information from 11 other publicly available databases. This includes research databases from Europe, Russia, China, and Thailand. It also included information from national databases in the USA, Canada, Brazil, Japan, Spain, Australia, and India. For the full list of sources, see Table 1 on page 768 in Do et al. (2018).

Finally, I obtained runoff data for 47 gages in Australia from the country's Bureau of Meteorology. I used gages that are a part of the bureau's set of Hydrologic Reference Stations (Australia BOM, 2020). This is a set of 467 "well-maintained river gauges of long, high quality streamflow records managed by Australian and State water agencies. The stations can be used to estimate trends in long-term and seasonal water availability from climate variability and change." I selected gages that matched our conditions for watershed size and dates, then downloaded data for these sites from the bureau's website and processed the data with a set of Python scripts.

The river flow data from the three sources are reported as daily averages. I calculated the monthly average by taking the ordinary arithmetic mean of the daily values. I used the threshold where a month must have at least 25 days of data to be included. I converted all data to common units, m^3/s, and then converted all flows to a normalized flux in mm/month by dividing by the watershed area (in km^2) and multiplying by a conversion factor. The spatial coverage of our final 2,056 river gages and basins is uneven across the globe. North America is over-represented with 1,111 gages (more than half the total of 2,056), as is Europe with 393 gages, while there are only 70 gages in Africa, 178 in Asia, and 195 in South America.

I selected gages with a minimum temporal coverage of at least 6 observations during the time period from 2002 to 2019. I performed quality control of observed runoff by making a variety of plots and statistical summaries. The GRDC acts as a clearinghouse for data but does not perform quality control.

I calculated monthly average runoff for months with at least 25 days of data. Volumetric flow rates in m^3/s were converted to area-normalized fluxes in mm/month by dividing by the land surface area in km^2 and multiplying by an appropriate conversion factor. The spatial coverage of our final 2,056 river gages (and their basins) is uneven across the globe (see Figure 2.8). North America is over-represented with 1,111 gages (more than half the total), as is Europe with

393 gages, while there are only 70 gages in Africa, 178 in Asia, and 195 in South America. China is among the most apparent data gaps. However, other countries are notable for their sparsity of data. For example, we have fewer observations in France (8 gages) than in smaller neighboring countries Switzerland (12 gages) and Belgium (12 gages).

The gage data in our database is also not consistent over time. There are 437 of our gages (out of 2,056) that have complete data coverage – each of these gages has 20 years of data over the 20-year period from 2000–2019. It is more common for our gages to be missing data from one or several years in this period. Figure 2.9 shows the distribution of gages by the number of years of data for the gage. The average length of data for our gages is 15.6 years, with a median of 17 years. There are 4 gages which have only one year's worth of data.

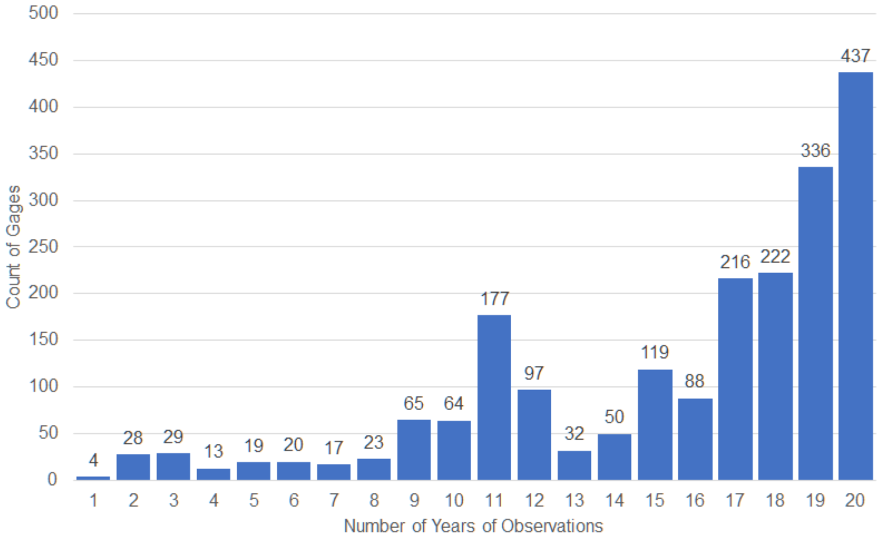


Figure 2.9: Distribution of data length in years for our 2,056 gages.

Our flow database is more complete during the first decade of our study's time period, from 2000–2009. This trend is illustrated in Figure 2.10. For example, in the year 2000, we have data at 1,892 gages, or 92% of our total gages. In the second decade, 2010–2019, the number of gages with flow data drops off. The last two years are especially sparse, with data at only 464 gages, or 23% coverage in 2019.

The reason for this trend is twofold. First, there is a lag in reporting of river flow data in many countries, and further there is a delay in sending these data to the GRDC. Second, there has been a decrease globally in the number of flow monitoring gages (Fekete et al., 2012). According to Fekete, "the number of monitoring stations peaked in the 1980s as a response to growing concerns about population growth and its impact on the environment, but as focus shifted toward climate

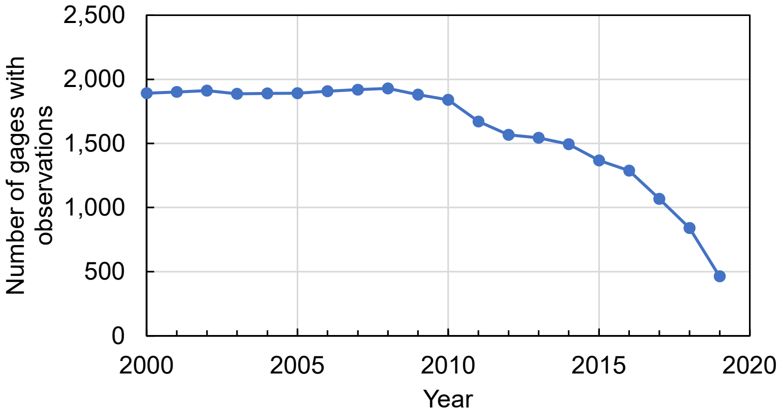


Figure 2.10: Number of gages with observations, by year.

change the commitment to continued operation of in situ monitoring networks diminished." The hydrologic science community is alarmed and dismayed by this trend. In situ measurements of river flow will never be completely replaced by remote sensing. Even with new platforms like SWOT, the scientific community will always need in situ data to calibrate and validate remote sensing observations.

I also examined the distribution of runoff data and its statistical properties. Figure 2.11 is a set of normal probability plots of the runoff dataset. In this figure, the runoff that has been normalized by converting the units to mm/month. Thus, we can compare runoff in basins of different sizes, and plot all of the data on a single figure. About 14% of the basins in our database have one or more records where the runoff is zero, at so-called *ephemeral* or *intermittent* rivers. This is not the same as missing data, which is stored in our database as NaN, or "not a number," a computer code for missing or corrupted data.

In terms of missing values, none of our time series is 100% complete. Every gage is missing data in one or more of the 240 months from 2000–2019. Nevertheless, more than 800 of the basins are at least 90% complete.

2.5.2 Synthetic Runoff

River discharge observations are limited, as they are only available at gaged locations. As an alternative, researchers have created gridded datasets that estimate runoff using statistical and machine learning methods. I used estimated runoff from GRUN Ensemble (Ghiggi et al., 2021). This dataset is a new version of the first GRUN (presumably for "Global RUNoff") dataset published in 2019 (Ghiggi et al., 2019). The authors created a global gridded dataset of monthly runoff using a machine-learning approach (random forest model), and based on precipitation

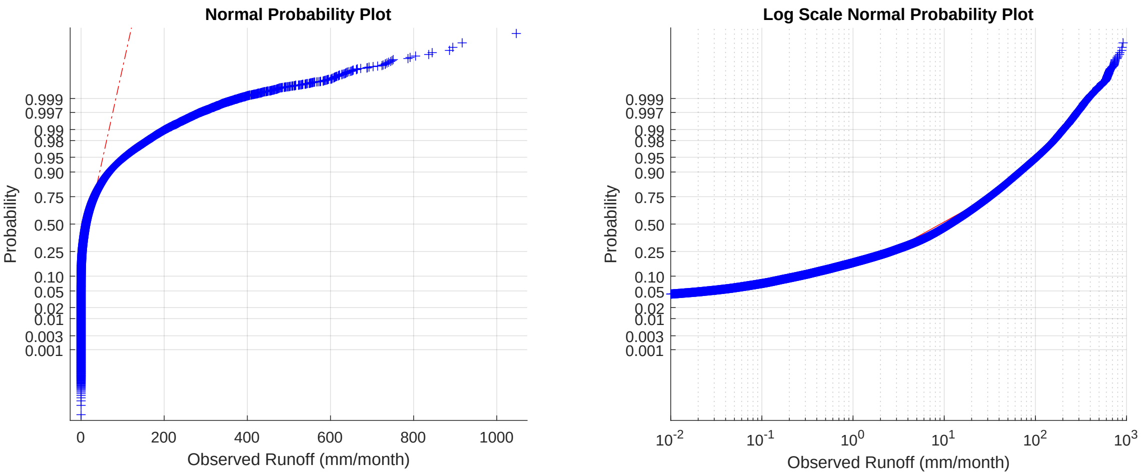


Figure 2.11: Normal probability plots of the normalized runoff dataset.

and evapotranspiration as predictor variables. For the 2021 GRUN Ensemble project, the authors used input data from 21 different sources, "including a set of atmospheric reanalysis, post-processed reanalysis and interpolated-stations data."

Ibarra et al. (2021) performed an independent evaluation of GRUN over a set of river basins in the Philippines. The GRUN model did not include data from the Philippines in its calibration or validation datasets, thus this served as a good independent quality check of GRUN. It also allowed testing under a variety of hydrologic conditions, due to the diversity of climates present.[2] Ibarra et al. (2021) compared GRUN predictions to observed discharge over 55 small tropical catchments with at least 10 years of data, extending back to 1946 in some cases. They found a significant but weak correlation ($R = 0.37$) and a "somewhat skillful prediction (volumetric efficiency = 0.36 and log(Nash–Sutcliffe efficiency) = 0.45)."[3]

I also performed an independent evaluation of GRUN against our 2,056 gages and found that GRUN is a relatively good fit to observed discharge. I first estimated the monthly discharge at the basin outlet by calculating the spatially averaged mean of gridded GRUN runoff. Then I calculated fit statistics comparing the observed and modeled flow time series. Figure 2.12 shows the distribution of two goodness-of-fit indicators comparing the time series of basin-averaged GRUN estimated runoff observed river discharge at 2,056 gaged basins. On the bottom of Figure 2.12, the maps show the geographic distribution of these same indicators.

[2] The Philippines is an island archipelago with a variable climate, stretching across 1,850 kilometers from north to south. Annual rainfall ranges from around 1,000 millimeters in mountain valleys to 5,000 millimeters along the east coasts of the major islands (US Library of Congress, 2005).

[3] The Nash–Sutcliffe efficiency and Kling-Gupta efficiency are goodness-of-fit indicators commonly used in the hydrologic sciences. I describe them in more detail in Sections 4.1.12 and 4.1.13.

I found that a median correlation $R = 0.84$ and median root mean square error, RMSE = 11.8 mm/mo, and 75% of gages had RMSE < 19 mm/mo. I also calculated a common fit indicator for modeled discharge, the Kling-Gupta Efficiency (KGE, explained in more detail in Section 4.1.13). Median KGE is 0.53, and 81% of gages have KGE > -0.41, the point at which a model's predictions are better than the mean of observations (Knoben et al., 2019). One word of caution is in order about the strength of this comparison. Many of the gages that the authors used to develop the GRUN dataset are the same as the gages I am using to judge its quality. Therefore, this is not a truly independent assessment. My conclusion is that GRUN has unprecedented skill in predicting runoff at the global scale, but that its accuracy is still limited. It appears that predictions are of only fair quality over certain zones, such as southeast Asian tropical island environments, as indicated by the results shown by Ibarra et al. (2021).

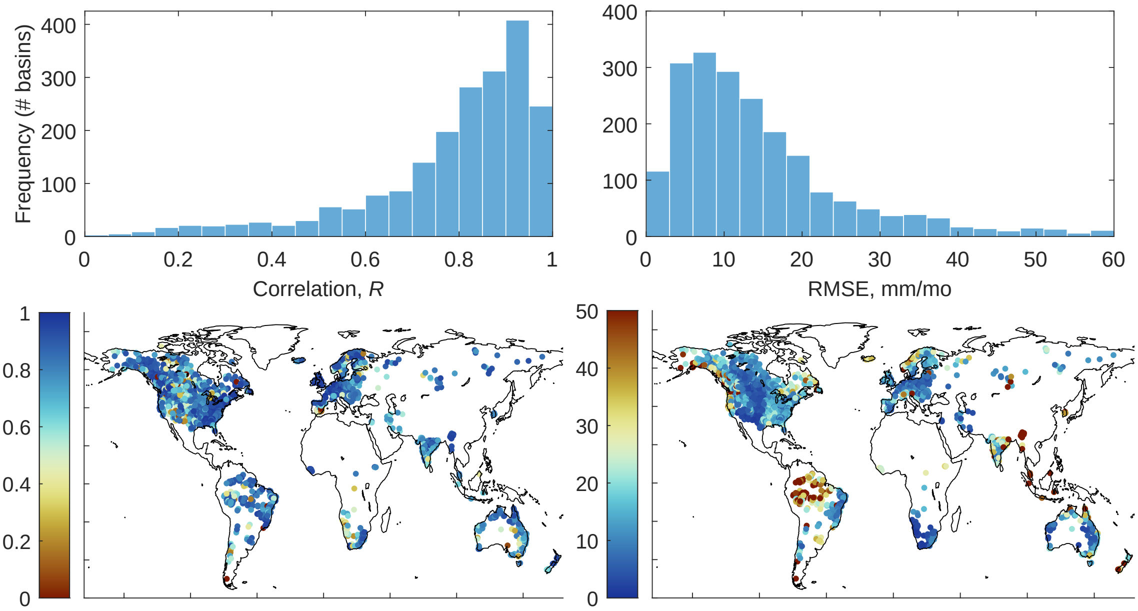


Figure 2.12: Comparison between GRUN estimated runoff and observed monthly-average river discharge at 2,056 gaged basins. Discharge data covering 2000–2019 was obtained from Australia's BOM, GRDC, and GSIM.

2.5.3 Modeled Runoff

I collected runoff data from two land-surface models in order to compare them to my results, to be described in Section 6.3. The first was the ERA5-Land model (Muñoz-Sabater et al., 2021). The model, from the ECMWF, consists of "global high resolution numerical integrations of the ECMWF land surface model driven by the downscaled meteorological forcing from the ERA5 climate reanalysis."

The authors state that, compared to previous versions of the model, it offers an improved description of the hydrological cycle, in particular better agreement with observed river discharges.

A second set of simulated runoff came from NASA's Global Land Data Assimilation System (Rodell et al., 2004). GLDAS drives the NOAH Land Surface Model (LSM). This hydrologic model has been operational since 1996 and has undergone continuous improvements to enhance its performance and accuracy. I added surface and subsurface runoff to calculate total runoff. For both datasets, I calculated the spatial mean over the project river basins, and converted units from kg/m² to mm/month.

2.5.4 Commentary on River Discharge Data in Large-Sample Hydrology

Collections of quality-controlled river discharge observations are essential in the field of global hydrology. Historically, researchers have had to spend a great deal of effort compiling and checking runoff data. It appears that many researchers are duplicating each others' work, which is both inefficient and a barrier to progress in hydrologic research. One cause for this is agencies that publish discharge data impose copyright or other restrictive data sharing agreements on data users. While this seems to be a legal and ethical gray area, most researchers appear to be conservative, and err on the side of not sharing data they have compiled. Furthermore, some national governments, such as India or China, have ceased publishing river discharge data, for reasons of security or national interest (Eyler, 2022). I believe that efforts to freely share global discharge data would spur a great deal of interesting work in global hydrology, climatology, remote sensing, and meteorology.

It is encouraging that there are a number of efforts underway in this area to compile discharge data, and associate it with meaningful metadata about the gage watersheds, often with the title CAMELS. In fact, a recent review has called this a new sub-discipline: large-sample hydrology, that "… relies on data from large sets (tens to thousands) of catchments to go beyond individual case studies and derive robust conclusions on hydrological processes and models" (Addor et al., 2017). In 2015, scientists at the United States Geological Survey (USGS), published a new dataset of flow and other meteorological data for 671 watersheds in the continental United States (Newman et al., 2015). A followup study (Addor et al., 2017) provided watershed attributes (related to topography, climate, land cover, soil, and geology) in a dataset called CAMELS, for "catchment attributes and meteorology for large-sample studies." Since then, similar datasets have been released by researchers in several different countries (Table 2.4). Finally, a group of researchers compiled these recent datasets into an omnibus dataset they called Caravan, for "a series of CAMELS". This dataset contains daily flow records for 6,830 gages in 12 countries between the years of 1980 and 2020.

Table 2.4: CAMELS runoff data sources that may be used in future studies

Dataset Name	Region	Year	# Gages	Reference
CAMELS	Continental USA	2017	671	Addor et al. (2017)
CAMELS-CL	Chile	2018	516	Alvarez-Garreton et al. (2018)
CAMELS-BR	Brazil	2020	3,679	Chagas et al. (2020)
CABRA	Brazil	2021	735	Almagro et al. (2021)
CAMELS-GB	Great Britain	2020	671	Coxon et al. (2020)
CAMELS-Aus	Australia	2021	222	Fowler et al. (2021)
LamaH-CE	Central Europe	2021	859	Fowler et al. (2021)
CARAVAN	multiple	2023	6830	Kratzert et al. (2023)

Easily-accessible datasets have led to a mini-boom in the use of machine learning in hydrology. Indeed, these data are ripe for exploration, and are ideally suited for practitioners of machine learning to explore. Artificial neural networks have been used to predict river discharge (as a form of black box rainfall-runoff model) since the early 1990s (Peel & McMahon, 2020). Recent advances in Long short-term memory (LSTM) networks have been able to make remarkably accurate predictions (Kratzert et al., 2019).

2.5.5 Runoff Data Limitations

In the water budget equation (Eq. 1.1), the runoff term R is meant to quantify the total flux of water leaving the watershed. Here, we use observed or modeled river discharge as a proxy for runoff. This is an imperfect measure on a number of counts. First, it does not account for subsurface flow (groundwater flow) which is not captured by gages (Fekete et al., 2002). Second, it does not account for man-made interbasin transfers or other human impacts. Yet, it is well-understood and well-documented that human alterations have large impacts on the natural water cycle (Vorosmarty et al., 2000; Hanasaki et al., 2006, see e.g.,). For example, interbasin transfers can have a confounding effect on our water budget calculations.

These are typically huge engineering projects, pipelines or canals that transfer water from one watershed to another. If the water is used primarily inside homes (and discharged to the ocean), it may not have any impact on our analysis at all. On the other hand, when water use is for irrigation, or to fill reservoirs, this could be a significant source of error. Studies conducted at the scale of a region or a watershed should make an effort to account for man-made fluxes into or out of the water. However, for this global study, this was not practical. Furthermore, such data are not always readily available, or easy to interpret. Some examples of large interbasin transfers include the Tajo-Segura interbasin water transfer in Spain, the California Water Project in the US, and China's South–North Water Transfer Project. (Note that this is just a few examples and is not at all comprehensive. Many more interbasin transfers exist.)

The national hydrologic agency in the United States, the USGS, provides special flags to alert data users when a river gage is affected by diversions or withdrawals. The agency has also published a dataset on a subset of 1,659 gages that are "relatively free of confounding anthropogenic influences," for the purposes of studying long-term variations in hydrology (Slack & Landwehr, 1992). Unfortunately, the GRDC does not provide a similar flag for its discharge database.

2.6 Environmental Indices

I also collected observations of ancillary environmental data as inputs to our NN model, listed in Table 2.5. The working hypothesis in my thesis research is that errors in EO data are the consequence of certain environmental conditions. For example, precipitation estimates are often biased in mountainous regions, or in relation to snow cover. If we provide the NN model with inputs that allow it to identify mountainous regions or snow covered regions, it should be able to make suitable corrections. In a previous study, Munier and Aires (2018) used ancillary environmental data to provide local correction factors for water budget components based on vegetation index (NDVI) and an aridity index (average $P - E$). Our hypothesis was a NN model fed with more environmental data could perform even better at correcting these errors and closing the water budget. The relationships are likely to be complex and non-linear, which a well-trained NN model should be capable of finding.

Each of the ancillary datasets in Table 2.5 consists of gridded geographic data. Certain of the environmental datasets are static; i.e., they do not vary over time (e.g., elevation, latitude). Other datasets are time-variable, such as solar radiation or vegetation indices. In all cases, I rescaled and reprojected environmental datasets as necessary to the standard project grid and calculated spatial means for river basins as described in Section 3.4. High-resolution data (e.g., 0.05°, 3600 x 7200 pixels), were upscaled to my standard 0.25° grid. While this was not strictly necessary, it lets me use the same data and workflows for computing basin means.

Table 2.5: Ancillary environmental data compiled at the river basin and pixel scale for use as inputs to an NN model.

#	Variable	Units	Source	Min	Median	Max
1	Aridity index	dimensionless	calculated	0.00032	0.57	2.9
2	Elevation, basin mean	meters	Amatulli et al. (2018)	15	520	5300
3	Latitude, basin centroid	decimal degrees	calculated	−50	27	75
4	Slope, basin median	dimensionless	Amatulli et al. (2018)	0	1.2	26
5	Vegetation Index, EVI	dimensionless	Didan (2015)	−0.17	0.21	0.7
6	Irrigated area (percent)	dimensionless	Siebert et al. (2015)	0	0.0006	0.76
7	Longitude, basin centroid	decimal degrees	calcualted	−160	27	180
8	Burned area (percent)	dimensionless	Giglio et al. (2020)	0	0	0.55
9	Snow cover (percent)	dimensionless	Hall and Riggs (2021)	0	0	100
10	Solar radiation	J/m^2	Hogan (2015)	0	18.1×10^6	346×10^6
11	Temperature	℃	Wan et al. (2021)	−45	26	57
12	Vegetation growth/senescence	dimensionless	calculated	−0.15	0.22	0.66

2.6.1 Aridity Index

The aridity index (or sometimes dryness index) is a widely used measure of the long-term hydro-climatic conditions of a region. It is variously described in the literature as a “degree of dryness” or a “degree of water deficiency.” Definitions vary, but it generally describes the ratio between available water and water demand for evaporation and water use by plants. The aridity index has been used in a wide range of studies and has been shown to be correlated with the runoff in basins (Arora, 2002). Methods of calculating the aridity index also vary. Among the most well-known and widely used definitions of the aridity index is the one proposed by Budyko (1974), $\phi = (E_0/P)$, where P is precipitation and E_0 is the reference evapotranspiration.

I used a contemporary global dataset from the Consultative Group on International Agricultural Research (CGIAR), which reports both annual average and monthly average aridity at the pixel scale over global land surfaces (Trabucco & Zomer, 2019). The annual average aridity index is shown in Figure 2.13. The CGIAR defines the aridity index as P/E. It is perhaps worth noting that some researchers call this the *humidity index*. The CGIAR’s definition is counterintuitive as it increases as climates get wetter. For dry environments $\phi \to 0$, while for wet, humid environments $\phi \to \infty$.

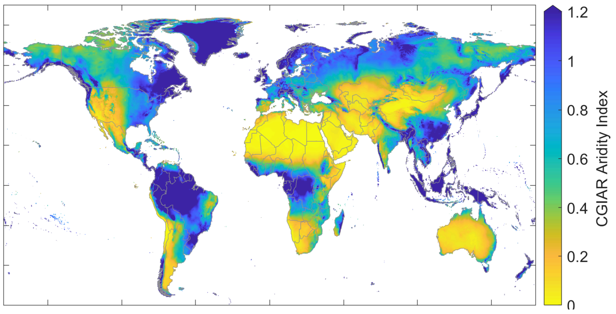


Figure 2.13: Global map of the CGIAR aridity index

2.6.2 Topographic Data: Elevation and Slope

I calculated the mean basin elevation and median slope using an open-source dataset of gridded global physiographic data published by Amatulli et al. (2018). This dataset combines several terrain-related variables, intended for environmental and biodiversity modeling.

I calculated the basin latitude and longitude based on the watershed polygon shapefiles in latitude and longitude coordinates using the software QGIS. The latitude and longitude are reported at the centroid of each river basin.

2.6.3 Vegetation

Satellite data is commonly used to identify the location and extent of vegetation by combining information from visible and near infrared wavelengths. These measures take advantage of the fact that chlorophyll in plant leaves absorbs and emits radiation in particular wavelengths. The first such measurement, which is still widely used, is the Normalized Difference Vegetation Index (NDVI). The NDVI is broadly an index of "greenness" that ranges from -1.0 to +1.0. While NDVI is not well correlated with photosynthesis rate or vegetation mass, it is used to "characterize the global range of vegetation states and processes" (NCAR, 2023). In other words, one cannot readily compare the NDVI in different areas to determine whether one location has more vegetation or biomass. However, there are rules of thumb for interpreting NDVI. According to the USGS Remote Sensing Phenology Program (Brown, 2018), here are typical NDVI values:

- $<$ **0.1**: Areas of barren rock, sand, or snow (usually)
- **0.2 to 0.5**: Sparse vegetation such as shrubs and grasslands or senescing crops
- **0.6 to 0.9**: Dense vegetation such as that found in temperate and tropical forests or crops at their peak growth stage

I obtained vegetation indices measured by the MODIS instrument onboard two different NASA satellites, Aqua and Terra. The Aqua satellite was launched in May 2002, and the first monthly data are available for July 2002. The Terra satellite was launched in December 1999, and the first monthly data are available for February 2000. Because of the better time coverage of the data from Terra, I chose to use this version for our analysis. Other remote sensing products related to vegetation are available. The most notable are the leaf area index (LAI) and fraction of absorbed photosynthetically active radiation (FAPAR).

I used the newer "enhanced" vegetation index EVI, rather than NDVI, as it "minimizes canopy-soil variations and improves sensitivity over dense vegetation conditions." Both indices provide a measure of the "composite property of leaf area, chlorophyll and canopy structure" (NCAR, 2023). The newer algorithm is thought by experts to be more accurate in areas of dense canopy (USGS, 2022a). Note that I used version 6 data; a newer version 6.1 became available in 2022, but was not yet complete. The EVI takes on theoretical values from –0.2 to +1.0.

I hypothesized that not only the amount of vegetation is important to the water cycle, but also the rate of vegetation growth or senescence. Therefore, I created a new ancillary variable by calculating the monthly rate of change of EVI shown. I refer to this rate herein as "vegetation growth/senescence" and it is shown in the final row of 2.5.

NDVI vs. EVI

Because EVI is a relatively new measurement, I did some analysis to better understand how it is different from NDVI, including creating summary statistics and maps. Figure 2.14 is a map of the correlation of NDVI vs EVI at the global scale. The map compares the correlations between the two time series in every pixel of the map. The green color means that in many places, the time series are highly correlated. However, large disagreement (as shown by orange and red colors) is seen in the Amazon and in the island region of Southeast Asia (Malaysia, Indonesia, Brunei, and Papua New Guinea).

Figure 2.15 is a histogram of the two datasets comparing the average values in all pixels over the period 2000 – 2019. The distributions are dissimilar, and EVI has a lower median value than NDVI, and fewer values in the highest part of the range between 0.6 and 1.

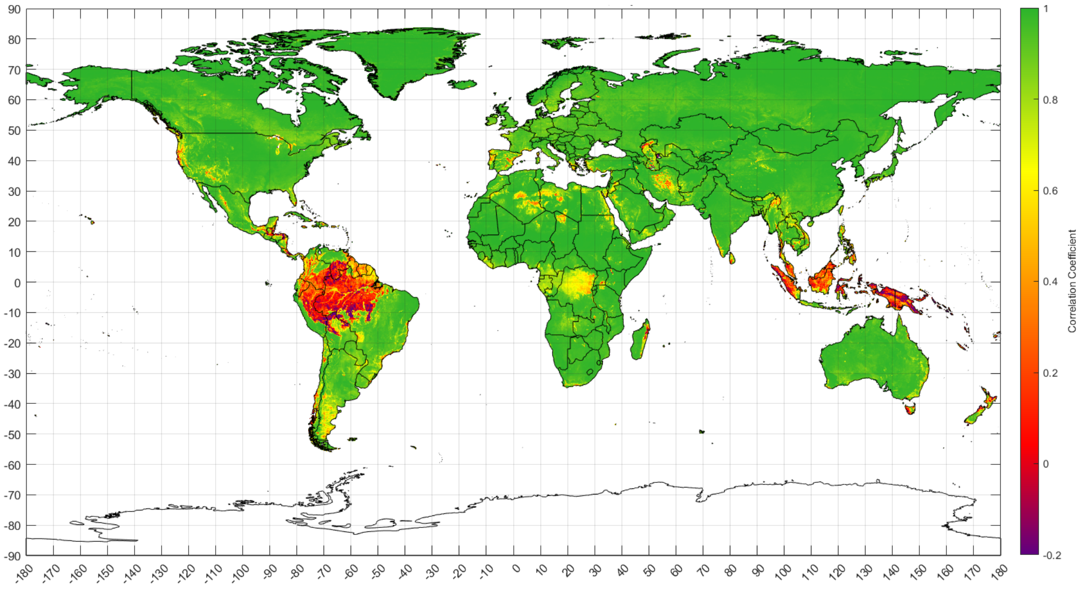


Figure 2.14: Correlation between NDVI and EVI time series at the pixel scale

NOAA AVHRR NDVI

For hindcasting applications, we wish to use an NN model to estimate TWSC prior to the launch of the first GRACE satellites in 2002 (this will be described in Section 4.4.2). I attempted to build a model with the most explanatory capability that

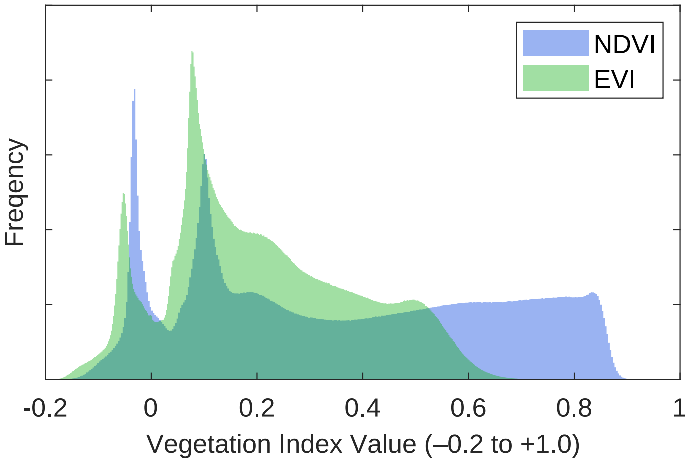


Figure 2.15: Distribution of average values of Terra/MODIS NDVI and EVI for 2000 to 2019

includes a variety of environmental variables, as described here. Unfortunately, the datasets NDVI and EVI described above are from the MODIS instruments onboard the Aqua and Terra satellites, and these particular datasets are not available before 2000. Therefore, I considered using an older vegetation dataset based on the Advanced Very High Resolution Radiometer (AVHRR).

The AVHRR is an instrument onboard a series of polar-orbiting NOAA weather satellites (NOAA 7, 9, 11, 14, 16, 17, and 18). NOAA publishes data on surface reflectance and NDVI based on AVHRR data. The dataset is daily, has global coverage, and a resolution of 0.05°. The data are available from June 24, 1981 to present (Vermote, 2018).

A related dataset is available from NOAA based on an instrument called the Visible Infrared Imaging Radiometer Suite (VIIRS). This instrument was meant to improve upon its predecessor AVHRR (Vermote, 2022). It is among the five instruments onboard the Suomi National Polar-orbiting Partnership (SNPP) satellite launched on October 28, 2011.

USGS Landsat Vegetation Indices

Another vegetation dataset with a long record is available from the Landsat satellites, operated by the United States Geological Survey (USGS). These data are very high resolution compared to the datasets described above, with a horizontal resolution of 1 arcsecond, which is equivalent to 1/3600°, or about 30 m near the equator. Vegetation from Landsat is interesting, largely because it spans many years, but it would require a great deal of specialized data processing.

Table 2.6 summarizes available datasets of global vegetation derived from satellite observations. Among the available data, there is a wide variety in the spatial and temporal resolutions. It is not immediately clear how to obtain a long record of EO vegetation data. Creating a high-quality dataset would involve comparing the spatio-temporal patterns among datasets and carefully intercalibrating them so that they can be combined.

Chu et al. (2022) published an article where they describe how their team created an extended time series of NDVI by combining data from different sensors, including AVHRR. I unsuccessfully attempted to acquire these data. The article states that data will be made available upon request. I requested the dataset from the authors and my messages went either undelivered or ignored. This phenomenon is unfortunately common in scientific publishing. One group of scholars that has examined data sharing practices opined: "statements of data availability upon (reasonable) request are inefficient and should not be allowed by journals." (Tedersoo et al., 2021).

Other sources of information on vegetation are available from models, but these were not considered here. Reanalysis models such as GLDAS or ERA5 include relatively simple vegetation dynamics based on assimilation of EO of leaf area index (LAI) and other observations. More detailed simulation is done in a class of models referred to as Dynamic Global Vegetation Models (DGVM). These models are designed to simulate the effects of changing climate on vegetation, and resultant changes to the hydrologic and biogeochemical cycles.

In conclusion, due to a lack of available vegetation data with a sufficiently long time record, I used EVI from MODIS/Terra as an input variable for the recent GRACE era (2002–2019), but dropped vegetation as an input variable in hindcasting experiments for 1980–1999.

2.6.4 Irrigated Area

Human activities can have a profound impact on the water cycle. Diversions and abstractions for irrigation may reduce runoff, take water from groundwater storage, and increase evapotranspiration through crop water use. For this reason, I hypothesized that irrigated area may be a useful explanatory variable. I estimated the percent of land area that is irrigated based on a global dataset by Siebert et al. (2015). The authors estimated irrigated area by "combining sub-national irrigation statistics with different data sets on the historical extent of cropland and pasture," and validated estimates with observations over the western United States. One limitation is that the data are through the year 2005. I chose to extract and use the values for this year, rather than attempting to do some kind of extrapolation of the

Table 2.6: Global remote sensing-based vegetation datasets

Name	Publisher	Begins	Ends	Time Res.	Spatial Res.	Source
NDVI						
AVHRR	NOAA	1981	present	day	0.05°	doi: 10.7289/V5ZG6QH9
AVHRR composite	USGS	1989	2019	10 day	1/12°	doi: 10.5066/F7707ZKN
GIMMS NDVI3g	NASA	1981	2015	half-month	1/12°	NASA (2019)
VIIRS	NOAA	2014	present	day	0.05°	doi: 10.25921/GAKH-ST76
Landsat	USGS	1982	present	16 days	1/3600°	USGS (2022a)
MODIS/Terra	USGS	Jan 2000	present	day, month	0.05°	doi: 10.5067/MODIS/MOD13C2.006
MODIS/Aqua	USGS	Jun 2002	present	day, month	0.05°	doi: 10.5067/MODIS/MYD13C2.006
EVI						
Landsat	USGS	1982	present	16 days	1/3600°	USGS (2022a)
MODIS/Terra	USGS	Jan 2000	present	day, month	0.05°	doi: 10.5067/MODIS/MOD13C2.006
MODIS/Aqua	USGS	Jun 2002	present	day, month	0.05°	doi: 10.5067/MODIS/MYD13C2.006

time series through 2019. Therefore, the data represents a snapshot of one point in time, limiting its accuracy. This dataset is shown in Figure 2.16.

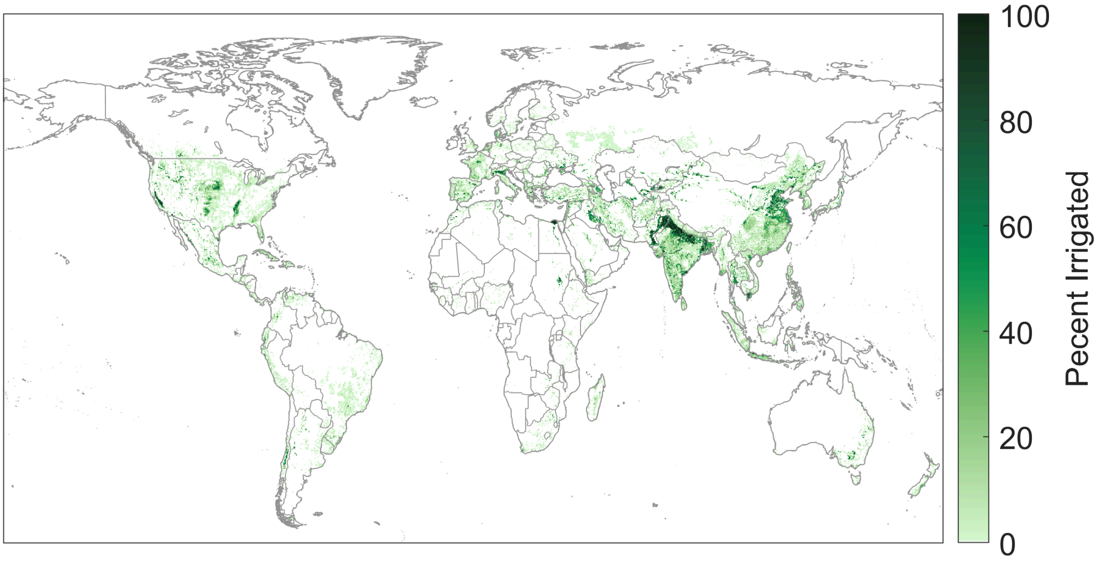


Figure 2.16: Global irrigation in 2005 from Siebert et al. (2015)

2.6.5 Land Surface Temperature

One could argue that land surface temperature is already indirectly included in my database, as it is among the input variables for evapotranspiration data products. Still, I chose to include this explanatory variable as an input to the neural network model as it has a strong, direct link on the hydrologic cycle. I obtained MODIS/Terra Land-Surface Temperature (Wan et al., 2021). As these data are available as monthly estimates on a 0.05° grid, including this variable was relatively straightforward.

2.6.6 Solar Radiation

Solar radiation has a strong effect on evaporation, vegetation growth, and the terrestrial energy balance. I obtained gridded monthly data for the variable “surface short-wave (solar) radiation downwards” from the ERA5 model (Muñoz-Sabater et al., 2021). Annual monthly solar radiation in Watts per square meter is shown in Figure 2.17.

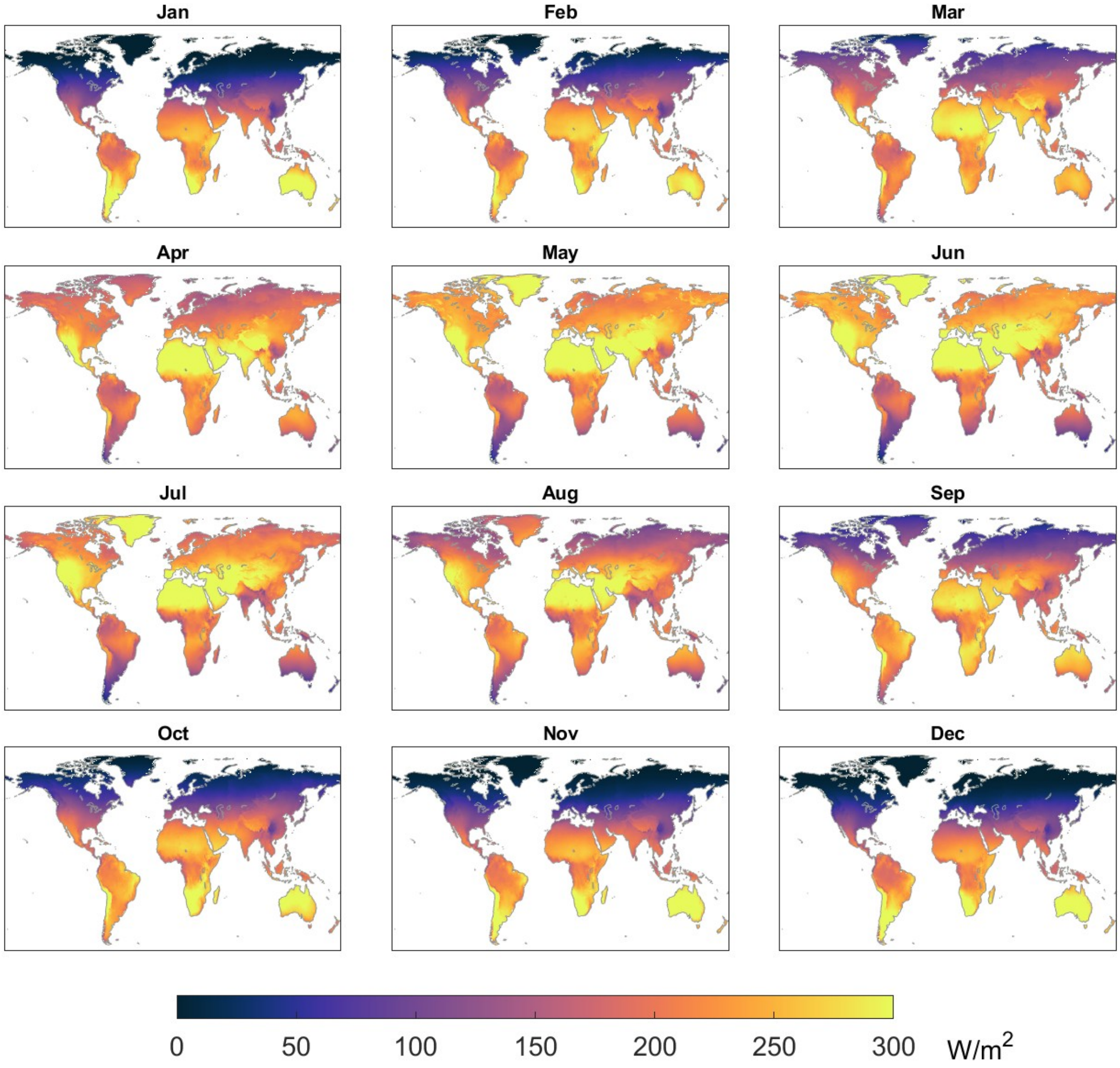


Figure 2.17: Average monthly solar radiation over the period 2000 to 2019, from the ERA5 model

2.6.7 Burned Area

Fire plays a role in altering the hydrologic cycle – burning of biomass releases water vapor to the atmosphere, and patterns of evapotranspiration are altered by changes to vegetation. Authors of a recent study concluded that "current and historical fires significantly affect terrestrial ecosystems, which can alter hydrologic fluxes" (F. Li & Lawrence, 2017). The authors attempted to quantify these changes, concluding that fire reduced global evapotranspiration by 0.6 × 10 km^3. This averages 0.003 mm/month, a seemingly insignificant amount (see Figure 2.19 for typical fluxes). Yet, fires can have a large local impact at certain places and at certain times. Therefore, I included an ancillary variable of *burned area* as a proxy for fire activity. I used data from the Terra and Aqua satellites' Moderate Resolution Imaging Spectroradiometer (MODIS) instrument. The presence of fire activity is determined through measurement of thermal anomalies measured from orbit. The creators of this dataset caution that burned area estimates have high uncertainty "due to nontrivial spatial and temporal sampling issues" (Giglio et al., 2020).

2.6.8 Snow Cover

The presence of snow and ice can negatively affect the accuracy of satellite observations of the water cycle (Kidd et al., 2012; Q. Cao et al., 2018). I hypothesized that including data on snow cover would allow the neural network model to make corrections in areas where fluxes are biased. I included a monthly variable on percent snow cover from the MODIS/Terra (Hall & Riggs, 2021), spatially averaged and upscaled to our 0.25° project grid.

2.7 Preliminary Analysis

In this section, I show the results of some preliminary analysis of the EO datasets. I performed a variety of exploratory data analysis and visualization of the EO datasets by creating maps, time series plots, and statistical summaries. This is a critically important step in the process of ingesting datasets from various sources, in order to verify that the various geographic transformations and unit conversions have been done correctly. This helps to see where datasets are in agreement, and when and where there is more divergence in their estimates. Above, I listed and described many EO datasets describing different components of the hydrologic cycle. I selected datasets for my analyses based on their spatial and temporal coverage and their data quality.

Figure 2.18 shows a snapshot of one month (January 2005) of the EO datasets used as input in our analysis. Some of the source datasets included data over both land and oceans. In these cases, I masked data over the oceans to facilitate visual comparison. Further, I excluded Antarctica from the datasets. The maps in Figure 2.18 show that the different datasets share many similarities in terms of the overall patterns, but many small differences are apparent upon close inspection. For example, precipitation according to GPCP appears smoother, while the other two datasets, which have a higher spatial resolution, show finer-grained patterns of high and low rainfall, particularly evident over the Amazon and southern Africa. Similarly, one can see differences in the spatial patterns of E and ΔS. River discharge, measured at gages, has a much sparser coverage, and the distribution of flows is highly skewed, with measured discharges covering several orders of magnitude from 0 to nearly 1,000 mm/month.

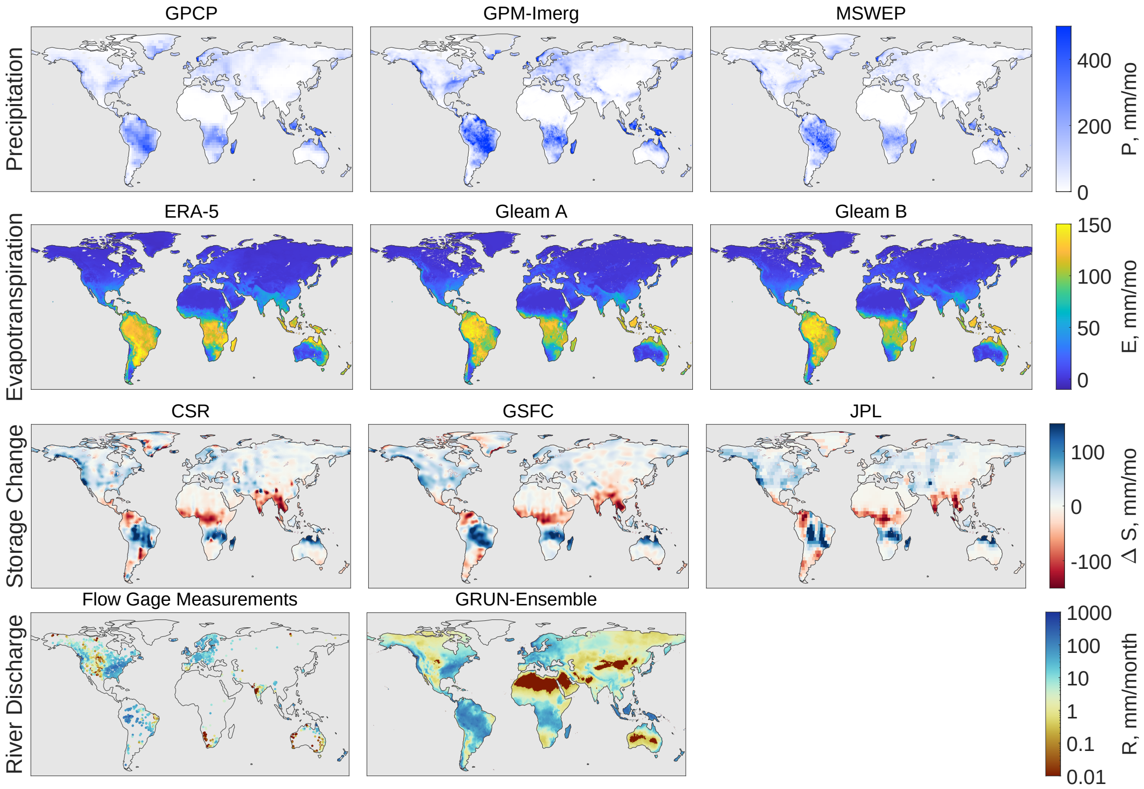


Figure 2.18: Maps of EO data for the month of January 2005. From top to bottom: observed precipitation, P, evapotranspiration, E total water storage change, ΔS and runoff, R.

2.7.1 Agreement and Disagreement among Datasets

Boxplots are useful for visualizing the distributions and central tendencies among the datasets. Figure 2.19 shows the distribution of values in the EO datasets used as input in our model. The boxes show the interquartile range, and the whiskers show the 10%-ile and the 90%-ile. Outliers are not shown, nor are the minimum or maximum values. The top boxplot in each set of observations is for all pixels over land within our analysis domain, which excludes Antarctica, Greenland, and areas above 73° North. The lower box in each set shows the distribution across our 2,056 basins. I calculated the basin mean fluxes from the gridded data using an area-weighted averaging method described in the Section 3.4. The statistics in Figure 2.19 were calculated over the 20-year period from 2000 to 2019. The marked differences among EO datasets is further evidence of the need for their calibration.

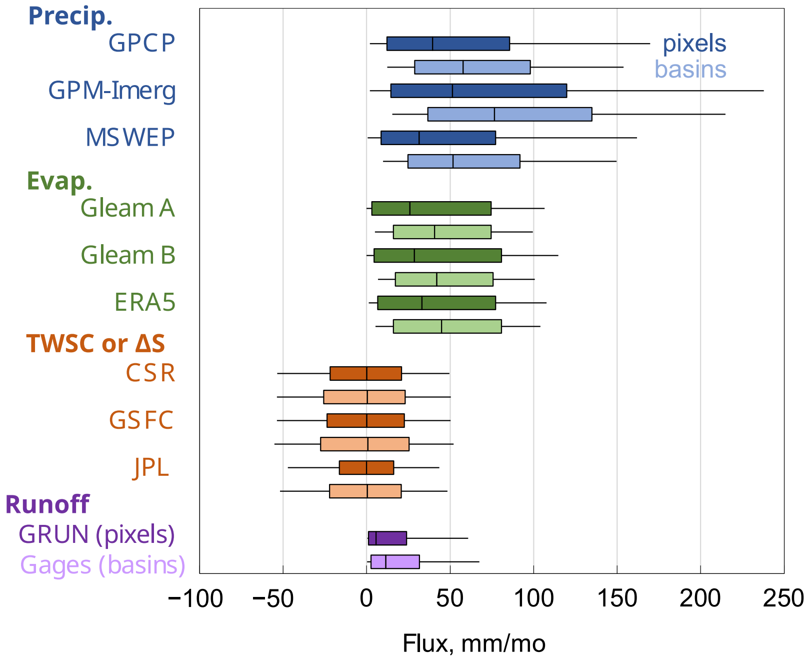


Figure 2.19: Boxplots showing the distribution of values in the EO datasets.

One can see in Figure 2.19 that, for most variables, the distribution of fluxes is greater over the pixels compared to the basins, with higher highs and lower lows. This is particularly the case for precipitation, but is also seen with evapotranspiration.

Calculating the mean flux over a basin tends to smooth out the extreme values and compress the distribution of observed values. (The gaged basins coverage, in terms of 0.25° pixels, is anywhere from 8 pixels to over 6,000 pixels, with a median of 19 and an average of 122 pixels in a basin.)

There are also differences in the distributions *within* each component of the water cycle. For example, GPM-IMERG contains higher observations of precipitation, with a higher 75- and 95-percentile than the other two datasets. The monthly water storage change, ΔS, is centered at about zero for each dataset. This is expected, as the storage in pixels and basins tends to fluctuate seasonally, and any long-term trend is small compared to the annual variations in storage. Runoff has the smallest magnitude of any of the hydrologic fluxes, with a low of 0 mm/month (no observed flow) to a 90%-ile of 68 mm/month, lower than the 90%-ile of P or E.

2.7.2 Trends in Total Water Storage

In this section, I analyze trends in total water storage, following methods used by Rodell et al. (2018). The purpose of this was two-fold: First, to verify that our handling of the data is correct by recreating the analysis in a well-known, peer-reviewed study based on GRACE data. The second purpose was to update the analysis by Rodell et al., which was based on observations through 2016. With five additional years of data, do the trends found by Rodell et al. continue or are they reversed? Furthermore, Rodell et al. (2018) used only one of the three available GRACE datasets, the one from JPL (Landerer & Cooley, 2021). How does the trend in water storage vary, based on the different datasets? What is the effect on storage trends of averaging the three datasets?

For this analysis, I followed Rodell et al. (2018) in calculating the trend using ordinary least squares regression. I calculated the slope and intercept for the TWS anomaly. I repeated the analysis first for the time period from 2002 to 2016 (matching the analysis by Rodell et al.) and then for an updated time period 2002 to 2021. Maps of the trend in TWS are presented in Figure 2.20. The maps appear to be quite similar, despite some minor differences in the colors. Some differences in the maps could come from the different versions of GRACE data. My calculations and mapping were done with updated GRACE data, release 06, while Rodell et al. used a previous version, release 05. NASA continually updates the algorithms and coefficients used for processing GRACE data and estimating TWS anomalies. According to NASA, "GRACE is a first-of-a-kind mission, so not surprisingly, revisions to the data processing are more frequent than for more mature satellite measurements."

Figure 2.21 illustrates the trend in GRACE total water storage, as calculated by the three different datasets: CSR, GSFC, and JPL. Maps at the right show the p-value, or the probability that the trend is different from zero. The overall patterns appear similar, but careful comparison of the maps reveals several differences. For example, the trends appear to be somewhat different over southern Africa. Zooming in on a single continent, such as South America in Figure 2.22, one sees distinct spatial patterns in each of the datasets. These patterns are artifacts of the data and processing used for each dataset. For example, with the CSR dataset, one can see the hexagonal grid of the mascons. The GSFC data has been smoothed with a Gaussian filter. The JPL dataset appears blocky, an artifact of the rectangular mascon 3° × 3° grid.

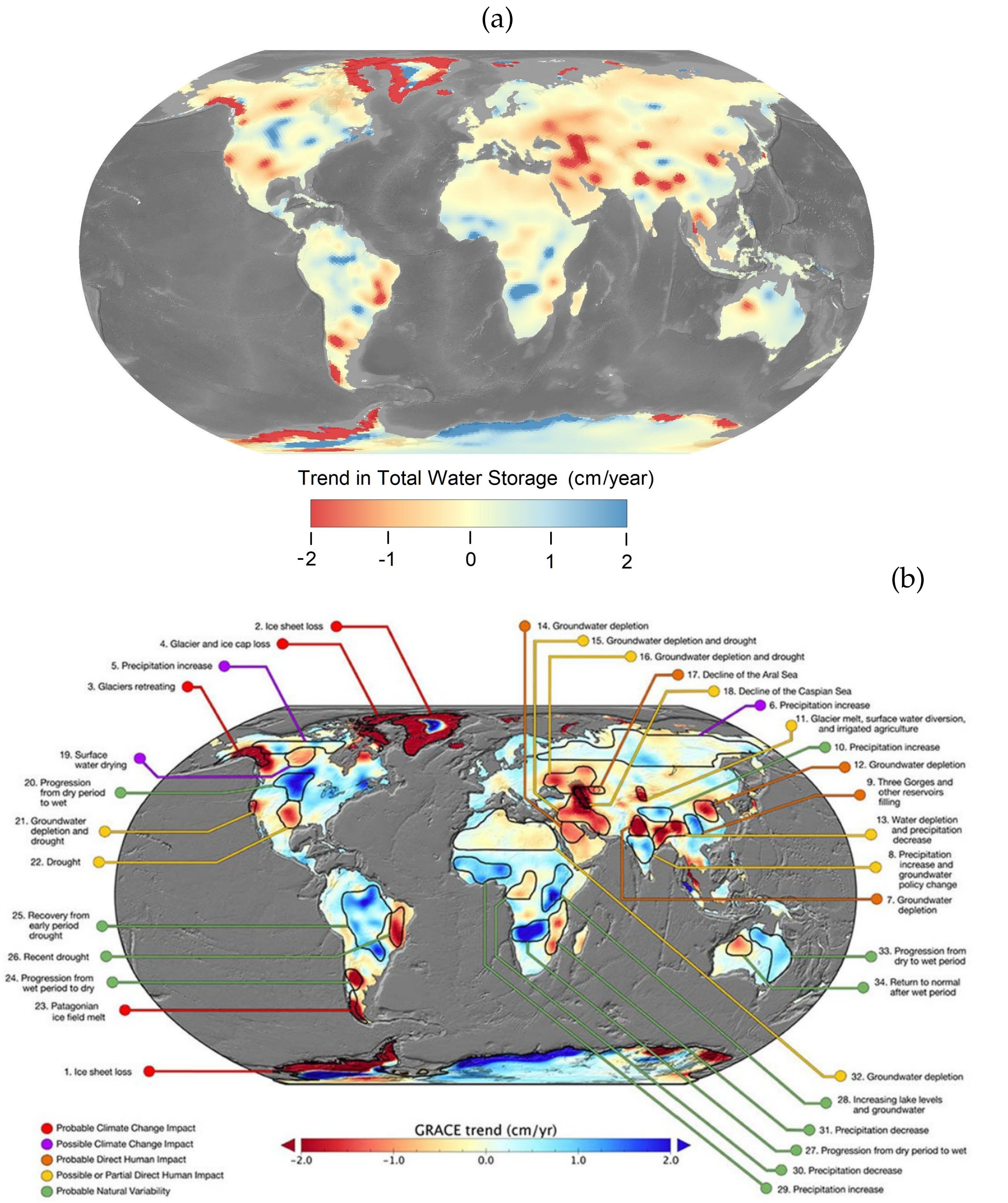

(a)
Trend in Total Water Storage (cm/year)
-2
-1
0
1
2
(b)
2. Ice sheet loss
4. Glacier and ice cap loss
5. Precipitation increase
3. Glaciers retreating
19. Surface water drying
20. Progression from dry period to wet
21. Groundwater depletion and drought
22. Drought
25. Recovery from early period drought
26. Recent drought
24. Progression from wet period to dry
23. Patagonian ice field melt
1. Ice sheet loss
14. Groundwater depletion
15. Groundwater depletion and drought
16. Groundwater depletion and drought
17. Decline of the Aral Sea
18. Decline of the Caspian Sea
6. Precipitation increase
11. Glacier melt, surface water diversion, and irrigated agriculture
10. Precipitation increase
12. Groundwater depletion
9. Three Gorges and other reservoirs filling
13. Water depletion and precipitation decrease
8. Precipitation increase and groundwater policy change
7. Groundwater depletion
33. Progression from dry to wet period
34. Return to normal after wet period
32. Groundwater depletion
28. Increasing lake levels and groundwater
31. Precipitation decrease
27. Progression from dry period to wet
30. Precipitation decrease
29. Precipitation increase
● Probable Climate Change Impact
● Possible Climate Change Impact
● Probable Direct Human Impact
● Possible or Partial Direct Human Impact
● Probable Natural Variability
GRACE trend (cm/yr)
-2.0
-1.0
0.0
1.0
2.0

Figure 2.20: Trends in total water storage based on GRACE-JPL, (a) calculated by the author; (b) from Rodell et al. (2018).

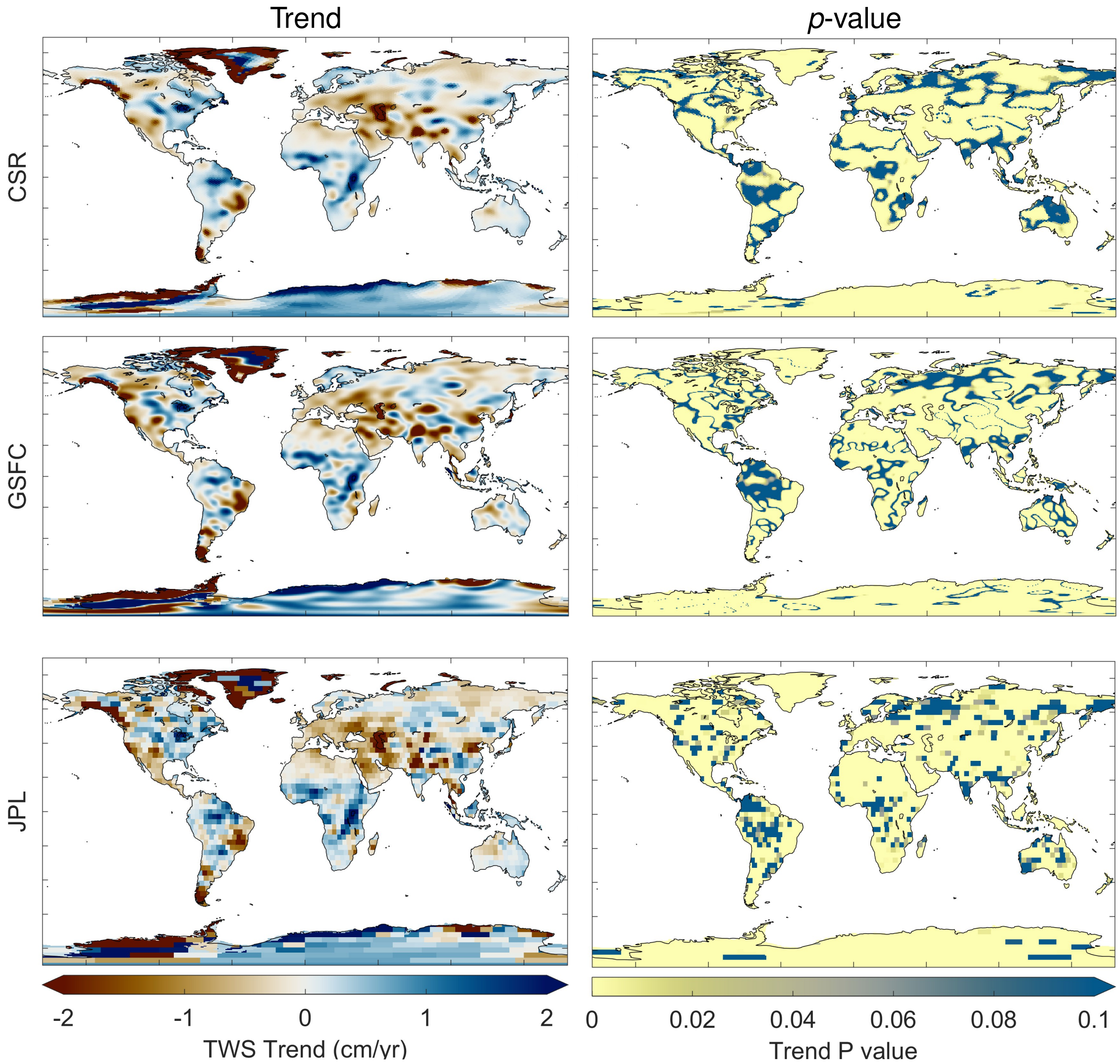


Figure 2.21: Pixelwise trends in total water storage and associated p-value calculated with the 3 GRACE solutions

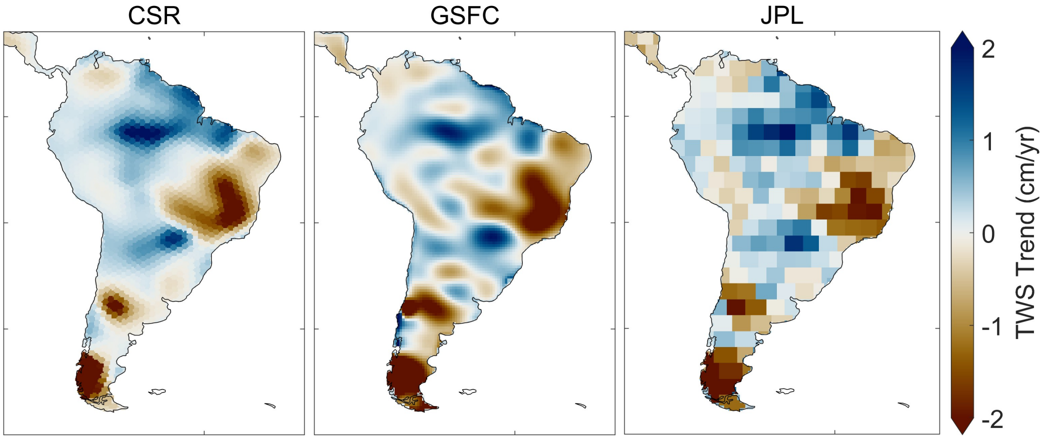


Figure 2.22: Trends in total water storage in South American based on the 3 GRACE solutions

Chapter 3

Balancing the Water Budget with Earth Observations

The overall goal of the research presented in this thesis is to improve remote-sensing datasets of the terrestrial water cycle. In Chapter 1, the Introduction, I showed that the water budget cannot be balanced using remote sensing data without incurring unacceptably large errors. This leads us to conclude that one or more of the remote sensing datasets for water cycle (WC) components are biased. In this chapter, I describe analytical methods that can be used to analyze earth observation (EO) datasets, and make modifications to them to derive new estimates for components of the global water cycle.

The first section in this chapter describes the derivation of appropriate geographic boundaries for the analysis, or the delineation of river basin boundaries. First, I describe a fast method for delineating watersheds upstream of a point of interest, in this case river gages. Next, I show a method for creating "synthetic" river basins of a given size. This method allows us to create a large set of basins with global coverage using gridded terrain datasets of flow direction. This method is useful when we are working with synthetic gridded runoff data. Here, we are not constrained to using gaged sites, and may define watershed outlets wherever it is convenient. Next, I describe the conversion of vector basin boundaries to gridded basin masks. Then I detail how these masks are used to calculate spatial means of earth observations and gridded environmental data. I developed an efficient algorithm for this calculation, as it was to be repeated several million times.

Some preprocessing of EO datasets was required before they could be used in the main analyses described below. This was particularly the case for the GRACE satellite data for total water storage (TWS). I describe methods for filling gaps or missing data, and describe methods for calculating the month-over-month total water storage change (TWSC).

Next, I describe methods for combining remote sensing datasets to develop a best ensemble estimate. Following this, I describe the optimal interpolation method (and its variants) for calibrating water cycle component such that the water budget constraint is satisfied. Optimal interpolation (OI) is a simple but powerful method, but it has a serious limitation – it can only be applied over basins where runoff is available. Thus it is not generalizable; it cannot be applied to ungaged basins or to grid cells. One of the key elements of my research was investigating various modeling approaches that could “recreate” the OI solution, and which could be applied more universally, at any location and at different geographic scales. I describe two classes of models. First, a set of statistical models based on regression and geographic extrapolation with surface fitting. Second, I describe a class of machine-learning methods, namely neural network models. With any model, we are concerned with how well its predictions match observations, i.e. its goodness of fit. Therefore, I describe methods for assessing model fit that can be applied to different classes of data such as time series and gridded EO data.

3.1 River Basin Delineation

River basins, or watersheds, are the most common geographic unit for analysis of the water cycle. In this thesis, I discuss analyses of the water cycle at the scale of two different geographic units: river basins and pixels. For the pixel scale analysis, I used the 0.25° and 0.5° grids described in the previous chapter. The disadvantage to studying the water balance at the pixel scale is the lack of observations of horizontal inflow or outflow. Hydrologists frequently make the simplifying assumption that there is no flow into a river basin. This means that there is no groundwater flow across basin divides, or in other words the patterns of groundwater flow in aquifers is similar to surface water flows. In other words, we assume that catchments are not “leaky” (Fan, 2019). Where there is no subsurface flow across the basin boundary, we may apply the conservation of mass principle and the water budget equation. This is the key advantage of performing analyses at the scale of river basins.

A watershed is defined as the area on the Earth’s surface where water drains to a common outlet, and is determined by the topography (or elevation) of the land surface. In principle, any point on land has a watershed, that is an upstream contributing area.

To define the river basins where our water budget analysis is be conducted, my goal was fully represent the diversity of environments around the world. Basins were chosen to cover a range of climatic conditions and ecosystem types. I used the following criteria: (1) Availability and duration of river discharge measures, (2) geographic coverage – the basins should be large enough that GRACE data is reliable, and (3) geographic location.

Geographic data for watershed footprints is essential for calculating the average for WC components over the watershed. I obtained basin geodata in shapefile format from the Global Runoff Data Center (GRDC), which covered many of our gaged basins (Lehner et al., 2008). However, some of these basin boundaries appeared to be inaccurate. I checked all basin geodata visually, overlaying the basin boundaries on topographic maps and aerial photographs. I found that some of the GRDC's basin delineations were inaccurate. Therefore, I created a new set of boundaries for every watershed using the methods described in the next section.

3.1.1 Watershed Delineation for Gaged Basins

Standard methods of automated (or computer-assisted) watershed delineation require large gridded (or raster) datasets that cover the entire basin to be mapped. For the best results, one should use the highest-resolution data that is available. The current state of the art for global terrain datasets is 90m resolution. However, delineating large watersheds with high resolution terrain data can be slow, requiring more than an hour for a single large watershed on a high-end desktop computer. The processing also requires more memory than is available on most desktop or laptop computers. I developed a hybrid method that is fast and accurate, which uses both vector- and raster-based data. I created a public repository to share the Python code (Heberger, 2022). My method uses a vector dataset called MERIT-Basins (Lin et al., 2019; Lin et al., 2021), where rivers are encoded as polylines and catchment boundaries as polygons. The MERIT-Basin dataset contains 5.9 million unit catchments. The average size of a unit catchment is 45 km^2.

Here is a brief description of the watershed delineation method. Finding the watershed upstream of a point is a 4-step process. First, I find the river segment that corresponds to the gage, or "snap the pour point." I created a simple algorithm that searched within a certain radius of the point to find the nearest river reach whose upstream area closely matched the reported drainage area of the gage. The process of relocating the watershed outlet point is known to be challenging (Lindsay et al., 2008; J. Xie et al., 2022), and dozens of gages required manual corrections. Second, I find the set of unit catchments that is upstream of our outlet, using a network traversal algorithm. Third, I perform a raster-based analysis to split the most downstream unit catchment, so that the watershed boundary coincides with the gage and does not include superfluous drainage area downstream of the gage.

For the detailed, raster-based calculations, I used the gridded elevation and flow-direction dataset MERIT-Hydro created by Yamazaki et al. Yamazaki et al. (2019). Finally, I merged the unit catchments in a geographic operation referred to as *dissolve* or *unary union*. The result is a single polygon that represents the drainage area upstream of the gage. I carefully reviewed the output of automated delineation routines, and made several manual corrections. I also compared the computed surface area with the reported drainage area for the gage. Where the areas differed by over 25%, I flagged these basins for more careful review.

One encounters many problems when performing watershed delineation. The initial results are often incorrect, and you have to go back and make adjustments and recalculate. This was fairly easy to do with my procedure because it is very fast and the results can be visualized immediately. Some of the delineated watersheds contain internal gaps or "donut holes." Many of these come from the source data. MERIT-Basins has many small gaps and slivers between unit catchments, often only a few pixels wide. These are obviously artifacts of the data processing, and do not appear to be meaningful. However, there are also larger donut holes inside watersheds, which represent internal sinks, out of which water cannot flow. As an example, Figure 3.1 a map of the Rio Grande watershed in the United States and Mexico, with outlet coordinates at $(26.05, -97.2)$. Between the two main branches, the Rio Grande in the west and the Pecos River in the east, there is an endorheic basin that runs north-south for 560 km from New Mexico to Texas. Within this basin, there are several alkaline lakes or *playas*, a tell-tale sign that water flows in and either seeps into the ground or evaporates.

How to handle these donut holes in delineated watersheds is somewhat of an outstanding question in hydrology. These gaps represent areas that are "disconnected" and do not contribute to surface water flow at the basin outlet. How you choose to handle them may depend on the intended analysis. If we are analyzing flood flows, we perhaps ought to exclude these areas. On the other hand, if you are studying groundwater recharge within the basin, precipitation that falls in these areas may be an important contributor to the

The set of gaged basins covers 47 million km², about 35% of the land surface below 73° North which is our study domain (Figure 3.2. I estimated this fraction based on a land mask from the GRACE project, clipped to the latitudes of our project area. Curiously, this layer includes Hudson Bay in Canada, and some areas of open water in the Arctic as land. This is not a concern for us, but does cause us to slightly overestimate the total land area. The total land area in our study area is 136 million km², and the area covered by our basins is approximately 47 million km². The footprint of all of our project basins is shown in Figure 3.2. On the map

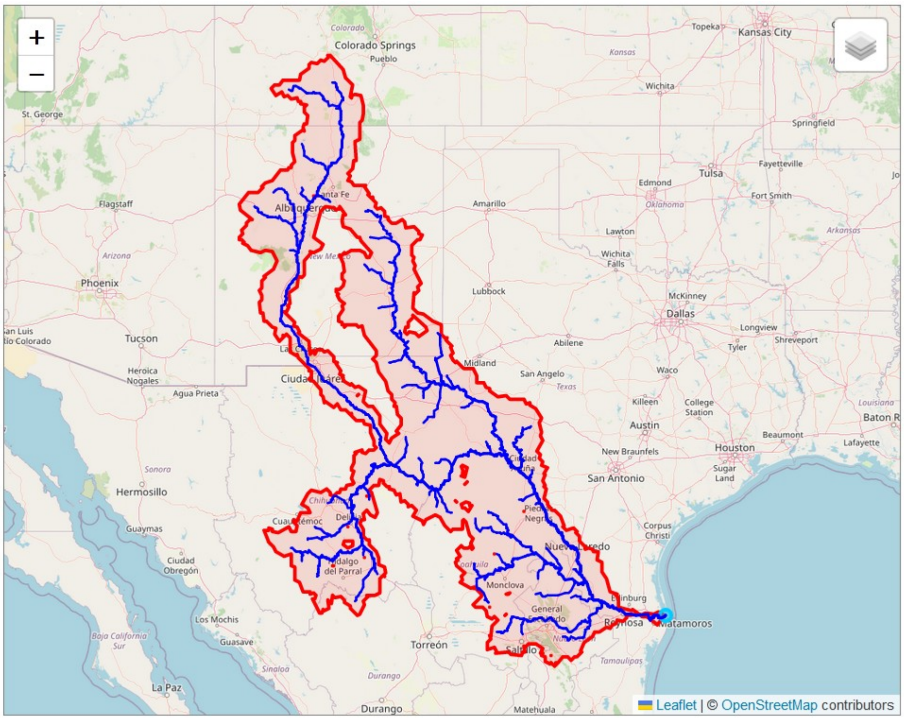


Figure 3.1: Example of internal gaps or “donut holes” in a delineated watershed.

in Figure 3.2, the land surface in the project area is shown as tan, while the ocean and large lakes are light blue.

I created two unique identifiers for each of the 2,056 project watersheds. First, I created a set of alphanumeric codes called the **basinCode**. These are 10-digit strings of text, where the first two digits identify the data source, followed by an underscore, and a number. The numbers are arbitrary, but unique. Furthermore, in Matlab code, it is easier to reference data with an integer, so we created a field called **basinID**, where basins are identified by an integer from 1,... 2,056.

basinCode	basinID
au_0000001	1
au_0000002	2
...	
gr_1259090	240

3.1.2 Watershed Characteristics

I performed some exploratory data analysis of the gaged watersheds to make sure that they looked valid and correct. First, I compared the delineated watershed area to the area reported by the data provider. The GRDC reports the watershed area for many, but not all, of its gages. However, there is little metadata describing the source of this information. I noticed that it often matches the area of the shapefiles

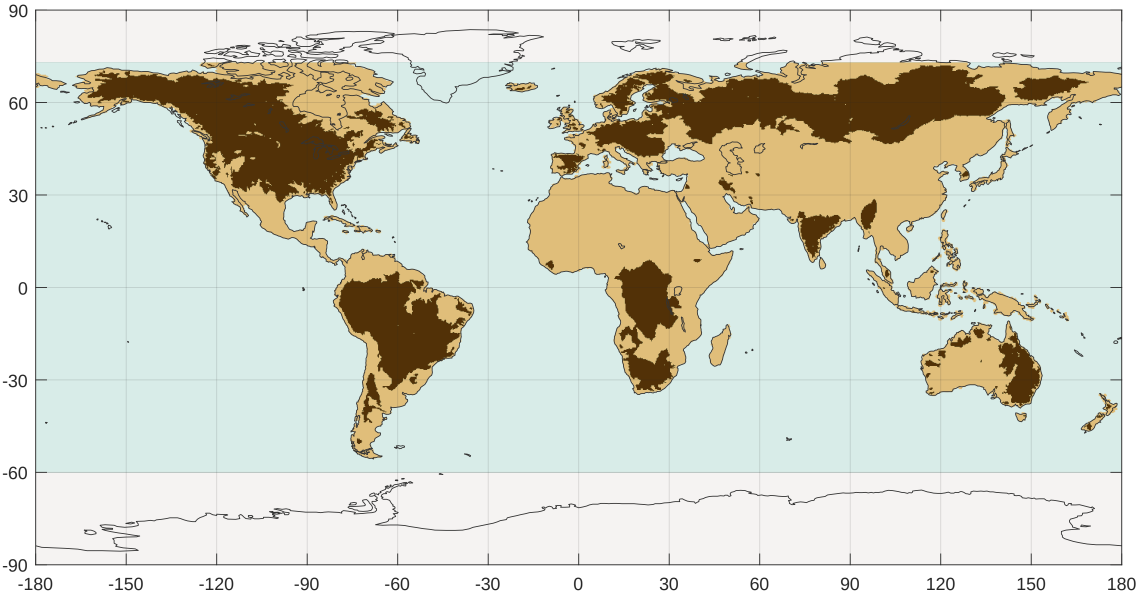


Figure 3.2: Distribution of areas of the 2,056 study river basins

produced by Lehner (2011) and which are provided by GRDC. In other cases, I presume that the watershed area was reported by the water agency or ministry in its country of origin. I suspect that for some older gages, it was calculated by an engineer or analyst using a topographic map and a planimeter.

Figure 3.3(a) shows the relationship between the delineated watershed area and the reported area, in km². Note that the points are clustered very tightly around the 1:1 line, indicating the relationship is very strong and errors are relatively low ($R^2 = 0.997$). Figure 3.3(b) is a histogram showing the percent differences between reported and calculated watershed areas. We could call the percent difference an error; however, it is possible that the reported area is inaccurate, especially if it was estimated with older data and more approximate methods. Overall, the delineation errors here are very low compared to other large-sample hydrology studies in the literature. I attribute this to the accurate data and methods used here in addition to the "human-in-the-loop" processing method that allowed me to quickly identify and fix errors.

Figure 3.4 is a histogram showing the distribution of the watershed area in square kilometers for the 2,056 basins selected for this study. Note that the horizontal axis is on a log scale. The distribution of basins is highly skewed, meaning that we have many small- and medium-sized basins, and there are fewer large basins. The largest basin in the study is for a gage on the Amazon River, at Obidos, Brazil, with an area of about 4.7 million km². Other large basins include the Congo River at Kinshasa (3.5 million km²), the Mississippi River at Vicksburg, Mississippi

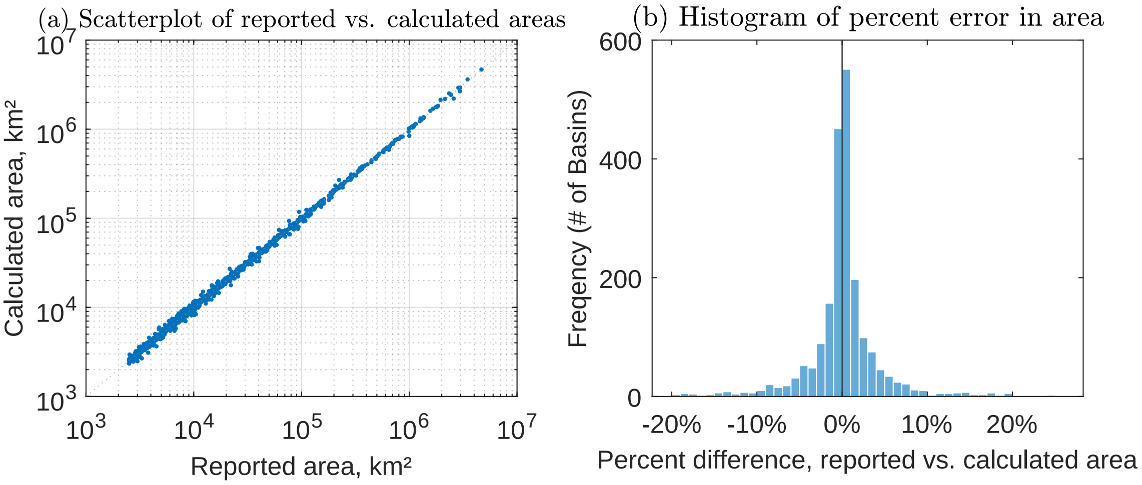


Figure 3.3: Comparison of reported and calculated watershed areas for the project's 2056 gaged basins

(3.0 million km²), and the Ob River at Salekhard, Russia (2.9 million km²).

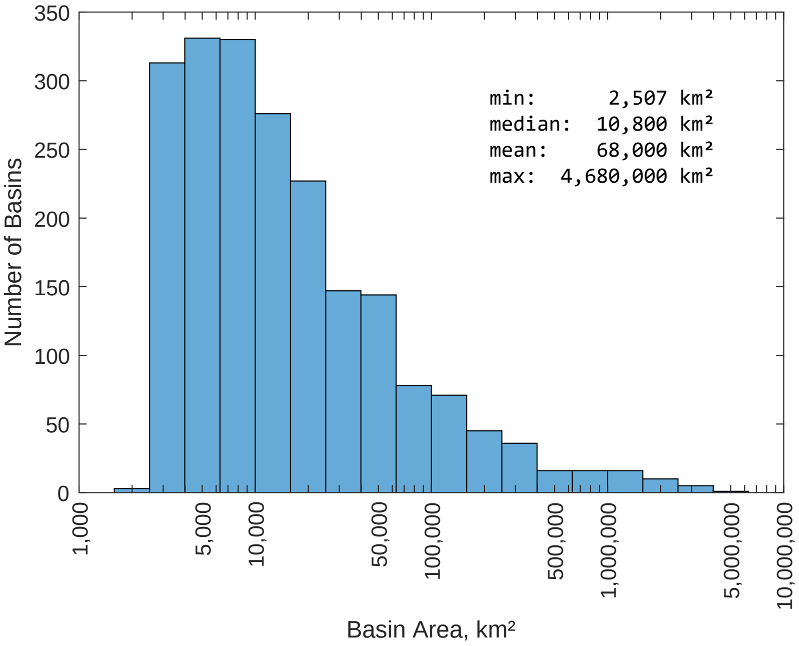


Figure 3.4: Distribution of areas of the 2,056 gaged river basins

3.1.3 Synthetic River Basins

In addition to the watersheds for the gaged basins, I created a second set of river basins for a set of experiments using gridded runoff data from GRUN. Here, we are not constrained by the location of river gages. Since we can calculate discharge by averaging the gridded runoff, we are free to define river basins anywhere. I refer to these *synthetic basins* to differentiate them from the gaged basins, which correspond to river flow gages. For the synthetic basins, their outlets do not correspond to a particular point of interest, such as a gage or an outlet. Rather, my goal was to create a collection of medium-size basins that cover much of the earth's land surface. Note however, that the basins correspond to real drainage patterns; it is merely that the outlet locations were chosen arbitrarily (by computer code) so that the basins are of a certain size that is roughly uniform.

To create these basins, I used a gridded flow direction dataset from the creators of the Global Flood Awareness System (Harrigan et al., 2020, GloFAS). GloFAS uses the LISFLOOD hydrologic model with inputs from ERA5 meteorological reanalysis data. I obtained raster data files that contain flow direction and upstream area on a 0.1° grid.[1]. These data are sufficiently detailed to represent the boundaries of mid-size river basins fairly accurately, and coarse enough that we can perform calculations quickly. The local drainage direction (LDD) file uses 8-direction (D8) encoding, using the conventions of PC-Raster environmental modeling software (Karssenberg et al., 2010). The direction of flow is coded as an integer from 1 to 9, as shown in Figure 3.5. In this scheme, a missing or NaN (not a number) value indicates outflow to the ocean, and a value of 5 means the pixel is a sink.

I used the open source Python library `pysheds` (Bartos et al., 2023) to create a flow accumulation grid. This is a standard data file that is used in terrain analysis and watershed delineation. There are two types of flow accumulation grids that are commonly used for watershed delineation. In the first type, each pixel contains a value representing the number of upstream cells. In the second type, each pixel contains its upstream drainage area, for example in km^2. I wrote a Matlab script to find the cells where the upstream area fell within a given size range. After some experimentation, and in consultation with my advisor, I chose to find basins ranging from 20,000 to 50,000 km^2. This routine required a data structure that inverts the downstream flow direction and instead reports the upstream cells. For this, I created a matrix listing the upstream neighbors of each cell. There are 6,480,000 cells (in the 3600 × 1800 grid), and each cell can have up to 8 upstream neighbors. Therefore, the upstream neighbor matrix has 6,480,000 and 8 columns. Using this information, I found the set of grid cells that defines the upstream drainage areas.

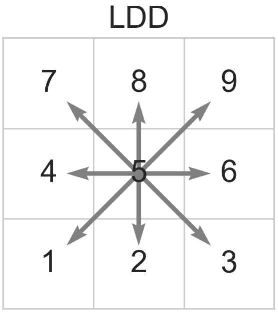


Figure 3.5: Flow direction encoding in the 0.1 ° resolution GloFAS local drainage direction raster.

[1] Data area available at https://confluence.ecmwf.int/display/CEMS/Auxiliary+Data

Figure 3.6 shows the 1,698 synthetic river basins. The color coding is for one of several sets of experimental partitions created for the training and validation of the neural network model, with the 80% of basins in blue for training, and 20% of basins in red for validation.

3.2 Pre-Processing of Total Water Storage Data

As described in Chapter 2, there are many missing records in the GRACE dataset of total water storage. In this section, I describe a simple method to fill in gaps in the record based on interpolation. I only use this method where I have reasonably high confidence in the interpolated value. In cases where the estimate is highly uncertain, more detailed methods based on modeling or water budget analysis may be more suitable for reconstructing GRACE-like TWS data, and filling in missing data. For one such example of methods to fill in missing data, see Yi and Sneeuw (2021). Other studies have focused on reconstructing the GRACE signal of TWS via modeling or regression-based methods, in order to hindcast, or create data from before the GRACE satellites were launched (see e.g., F. Li et al., 2021).

3.2.1 Calculating Total Water Storage Change

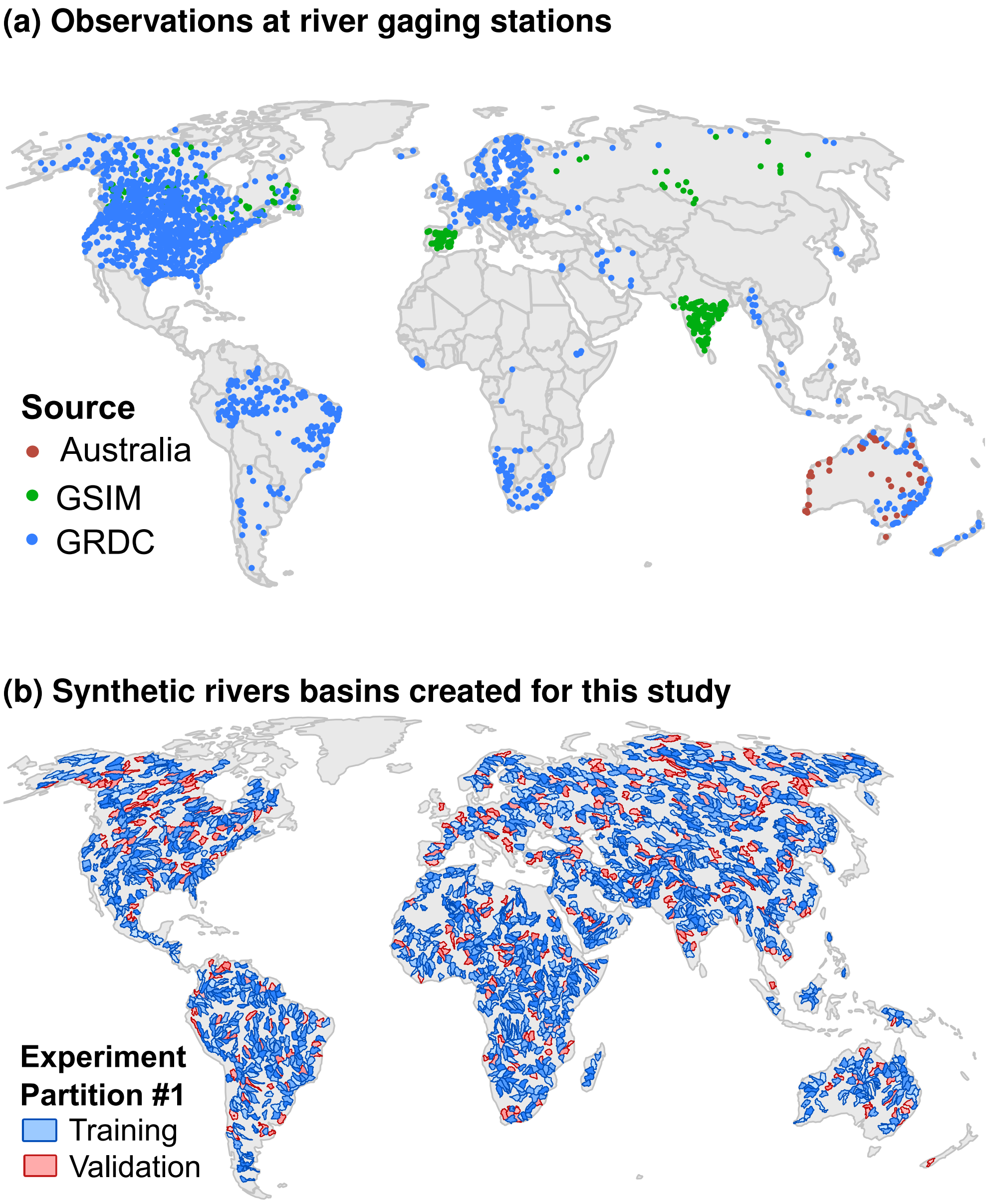
I calculated the month-over-month *rate of change* in water storage to provide the flux in mm/month. This converts the TWS anomaly to a flux equivalent, TWSC or ΔS. There are several methods for calculating the rate of change, but most researchers in this field use simple finite difference methods (see e.g., Landerer & Swenson, 2012; Biancamaria et al., 2019). I chose the simple backwards finite difference method.

$$\frac{\Delta S}{\Delta t} = \frac{S_t - S_{t-1}}{t - (t-1)} \tag{3.1}$$

Results from more complex methods such as fitting a cubic spline or using an "equivalent smoothing filter" (Landerer et al., 2010, see e.g.,) were comparable to those obtained with the simpler methods but often resulted in more missing observations, therefore I chose used the simple method in Equation 3.1.

3.2.2 Filling in Missing GRACE Data

The amount of missing data in the water storage time series threatened to limit our analyses. Therefore, we used methods from time series analysis regularly



(a) Observations at river gaging stations

(b) Synthetic rivers basins created for this study

Figure 3.6: Maps of this study's river basins: (a) 2,056 river flow gaging stations, corresponding to the basin outlets (basin boundaries not shown); (b) 1,698 synthetic river basins created for training and validating the neural network model.

used by earth scientists to fill in missing observations. The GRACE water storage observations begin in April 2002. Following a few missing months (gaps) in the beginning, and then there is a 6-year record with no gaps from 2004 to 2010. However, as the GRACE satellites aged, they were plagued by various issues such as failing batteries, and the satellites were powered down intermittently, resulting in intermittent 2-month-long gaps occurring regularly. The first satellite mission ended in 2016, and there is a 17-month gap until the follow-on mission was launched and began operating in 2018.

In the earth sciences, it is common for a dataset to be missing observations, for a variety of reasons, such as instrument failure, calibration problems. Instead of discarding these incomplete records, a commonly applied approach is to perform *imputation*, which involves filling in the missing values. I used standard time series analysis methods to fill in some missing observations via interpolation. In order to limit the uncertainty of interpolated observations, we did not allow the routine to fill gaps longer than 3 months. Further, we did not fill in missing observations where there is a change in slope in the time series before and after the gap. Because such a gap would represent either a peak or a trough, attempting to fill in this observation would be highly uncertain. As a result of the missing data, maps of ΔS for interpolated months appear patchy and incomplete. That is because I only filled in missing data with interpolation where the estimate has high confidence, as described above.

3.3 Upscaling of Gridded EO data

Certain datasets are published in a higher resolution than that of our analyses. In such cases, I used upscaling methods to create a lower-resolution versions of the high-resolution datasets. For example, I upscaled vegetation data (EVI and NDVI) from 0.05° to 0.25° resolution. There are several different methods available for doing this, with tools available in GDAL (`gdalwarp`), and using Matlab (`interp2` or `imresize` are two options).

For upscaling, I used a "correspondence matrix." This matrix contains a complete mixing of how pixels in the fine grid correspond to pixels in the coarse grid. This method is fast and gives"pixel-perfect" results. The correspondence matrix method only works when the high resolution is evenly divisible by the coarse resolution, for example going from 0.05° to 0.25° resolution. In such cases, the smaller pixels are fully contained within the larger pixels with no overlapping of edges. In this example, the detailed dataset is exactly 5 times the resolution of our target result. So each large pixel contains exactly 25 of the smaller pixels in the source dataset. We can simply take their arithmetic mean of 25 pixels to calculate the average in the large pixel. This approach (taking the average of overlapping grid cells) does not incorporate any smoothing or blurring. By contrast, these are features of some rescaling algorithms, which consider the values in neighboring cells when rescaling.

The algorithm I used takes missing values into account. Missing data are quite common, due to cloud cover, sunglint, or other factors. Because the vegetation datasets only cover land surfaces, there are no data over oceans and lakes. We should not allow a few missing values to prevent us from calculating the average over the larger pixel. In my experiments, doing so results in a large amount of missing data, especially near the coasts. Therefore, I used a compromise algorithm with the following rule. If a majority of pixels are available (do not contain the missing data flag), we calculate the average based on available data. For example, when we are upscaling from 0.05° to 0.25°, we must have valid data in 13 or more of the 25 small pixels.

What about the case where the lower resolution is not evenly divisible by the higher resolution? For example, suppose we are upscaling a dataset from 0.1° to 0.25° resolution. In this case, I used a two-step procedure. First, we rescale the high-resolution raster dataset using Matlab's `imresize` function. This is part of Matlab's Image Processing toolbox, and is originally intended for processing image data, but it works equally well on any sort of regular two-dimensional grid. We can instruct Matlab to use a "box-shaped kernel" over which to average the high-resolution data. We can also instruct it to omit NaN values ("not a number," a stand-in for missing data in a pixel). The risk here is we may end up with some pixels in the resulting lower-resolution output that are based on very few valid observations. At the upper limit, the result could be based on a single small pixel. Since such results are not very robust, I performed a separate calculation to count the number of pixels with missing data. If the percent of missing pixels is below a threshold, we declare that the calculation is low-quality, and discard it. I used the cutoff of 50%, meaning that if more than 50% of pixels are missing, the result is discarded. This allowed a good compromise between obtaining a complete coverage and a robust calculation of the mean.

3.4 Calculating Basin Means for EO variables

For each basin, I calculated the average P, E, ΔS, and R based on the gridded EO data products. I followed similar methods to those described by Kauffeldt et al. (2013). In brief, basin polygons are intersected with climate-data grid cells to calculate the fraction of precipitation (or other variable) that each cell contributes to the basin. This method relies on the assumption that there is no sub-grid variability, i.e. precipitation is assumed to be evenly distributed over each grid cell.

In geographic science, there are "zonal statistics" methods for calculating the statistics of gridded or raster datasets where they overlay or intersect a vector polygon. However, it was more efficient for me to use matrix math and perform the calculations in Matlab. To calculate the spatial weighted mean, I converted each basin polygon to a grid mask, where each pixel is a floating-point number representing the fraction of the pixel's area from 0 to 1 that is inside the basin. Because the surface area of pixels varies by latitude, I also use the pixel's area in our calculation via the Climate Data Toolbox for Matlab (Greene et al., 2019). The same routine was also used on all other gridded data products, such as the aridity index, vegetation indices, and elevation.

In this section, I describe how I calculated the average of gridded environmental data over watersheds. This is a common task in the environmental sciences, but it is worth describing for two reasons. First, this research required calculating basin means tens of thousands of times, so an efficient method is required. The first large-scale experiment described here used 2,056 basins, 10 variables, and 240 months; thus I repeated this calculation around 4.9 million times (actually somewhat less due to missing data). Second, this calculation is not always done correctly, even by experienced scientists. Witness a recent retraction in the prestigious journal *Nature*. The authors incorrectly calculated precipitation over river basins. The authors incorrectly used arithmetic averaging to calculate the mean, "instead of calculating a spatially weighted mean to account for the changing grid box size with latitude" (Marcus, 2022). As a result, the results were biased and the conclusions not supported, forcing the authors to retract the paper.

When working with data on a regular rectangular grid, it is important to understand that the area of the grid cells varies as a function of latitude. Figure 3.7 shows how the grid cells on the three-dimensional sphere of the Earth[2] are stretched and

[2] The Earth is not truly spherical. Rather, it is described as an *oblate spheroid*, with bumps and irregularities.

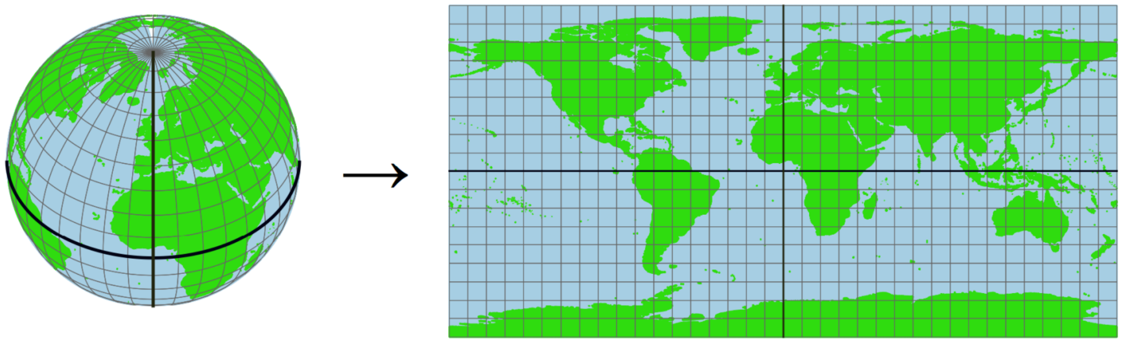


Figure 3.7: Grid cells on the earth vary in size due to projection distortions

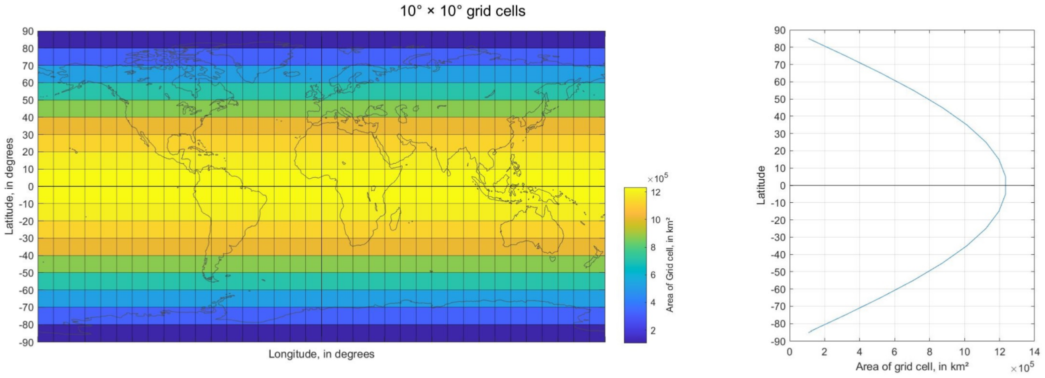


Figure 3.8: Area of grid cells varies by latitude

distorted when they are represented in two dimensions. Figure 3.7 shows 10° × 10° grid cells for clarity, rather than the smaller 0.25° or 0.5° grid cells of our EO data.

The surface area of grid cells is the maximum at the equator, and decreases as we move north or south, away from the equator and toward the poles. Figure 3.8 shows how the surface area of grid cells varies as a function of latitude. Again, this figure shows a 10° grid to demonstrate the concept.

To calculate the basin mean of an EO variable, we average the values of the pixels over the basin. However, because the pixels are of irregular size, we take a weighted average, where each pixel is weighted based on its surface area. Calculating the basin mean is a type of "zonal statistics," commonly calculated in geographic science. It is complicated somewhat by the fact that we are overlaying two distinct and incompatible data types. Basins are represented by vector polygons, and the earth observation datasets are grids or rasters. We can speed up the calculation of the spatial mean by converting the basin to a 'mask' in grid format. In this way, we can use fast and efficient matrix math to calculate spatially-weighted basin means.

I created two sets of masks for the basins:

- **Boolean**: Identifies pixels in the basin (1) or out of the basin (0). The output is a 720 × 1440 raster, where the value of each pixel is 0 (not in the basin) or 1 (in the basin). Pixels are in the basin if more than half of a pixel's surface falls intersects the basin.
- **Float**: The value of a pixel is the percentage of that pixel that intersects the basin. Pixel values vary from 0 (not in the basin) to 1 (completely in the basin). A pixel with a decimal value of say, 0.22, means that 22% of the pixel intersects the basin polygon, or 22% of the pixel is in the basin.

In Geographic Information System (GIS) software, converting a vector feature to a grid is called *rasterization*. GDAL is a widely used open-source library for performing many geoprocessing operations. I used the GDAL function `rasterize` to convert watershed polygons to a grid.

The initial results of the vector-to-grid conversion were unsatisfactory zoomed in to the pixel level. Since we are interested in some watersheds that cover only 8 pixels, we want to have the results as realistic as possible. The way that GDAL's `rasterize` algorithm determines whether a pixel is inside of a polygon is by looking at the centroid of the pixel. So you can encounter a situation where a polygon covers over half of pixel, but the pixel is not included in the raster output, because it was incorrectly classified as non-overlapping. Such considerations may be trivial when dealing with large features covering an area of thousands of pixels. But because some of our watersheds only cover about 8 pixels, I wished to avoid such rasterization errors. In other words, I was looking for "pixel-perfect" results. So I created a custom rasterization procedure using Python for QGIS. The following is an overview of the steps in the process.

There is a single input to my pixel-perfect rasterization algorithm: a vector polygon in shapefile format, representing the basin or watershed. Figure 3.9(a) shows an example of a watershed of GRDC gage 1159511 on South Africa's Vals River at Groodtraai. The polygon has an area of 7,801 km². The algorithm for calculating a higher-precision floating-point raster involves rasterizing at a 5x greater resolution, shown in Figure 3.9(b). This intermediate raster file has 25 times more pixels. In this example, to create a 0.25° raster mask, we first create a raster mask at 0.05°. In this intermediate file, cells have a value of 0 or 1. A single grid cell is covered by a 5 x 5 set of grid cells in the higher-resolution dataset. Then I calculate the sum of the small cells that are in each large cell. The result is the fraction of the large cell that overlaps the basin polygon.

The rasterization routine produces the outputs shown in Figure 3.9(c) and (d). The version on the left in (c) is a *floating point* raster, where the value in each cell represents the fraction of that pixel that is in the basin. The map on the bottom right (d) shows the *Boolean* (0/1 or true/false) version of the basin raster. The floating-point raster is able to preserve more information about the river basin's geographic coverage, which allows us to better estimate the WC components over the basin from gridded EO data.

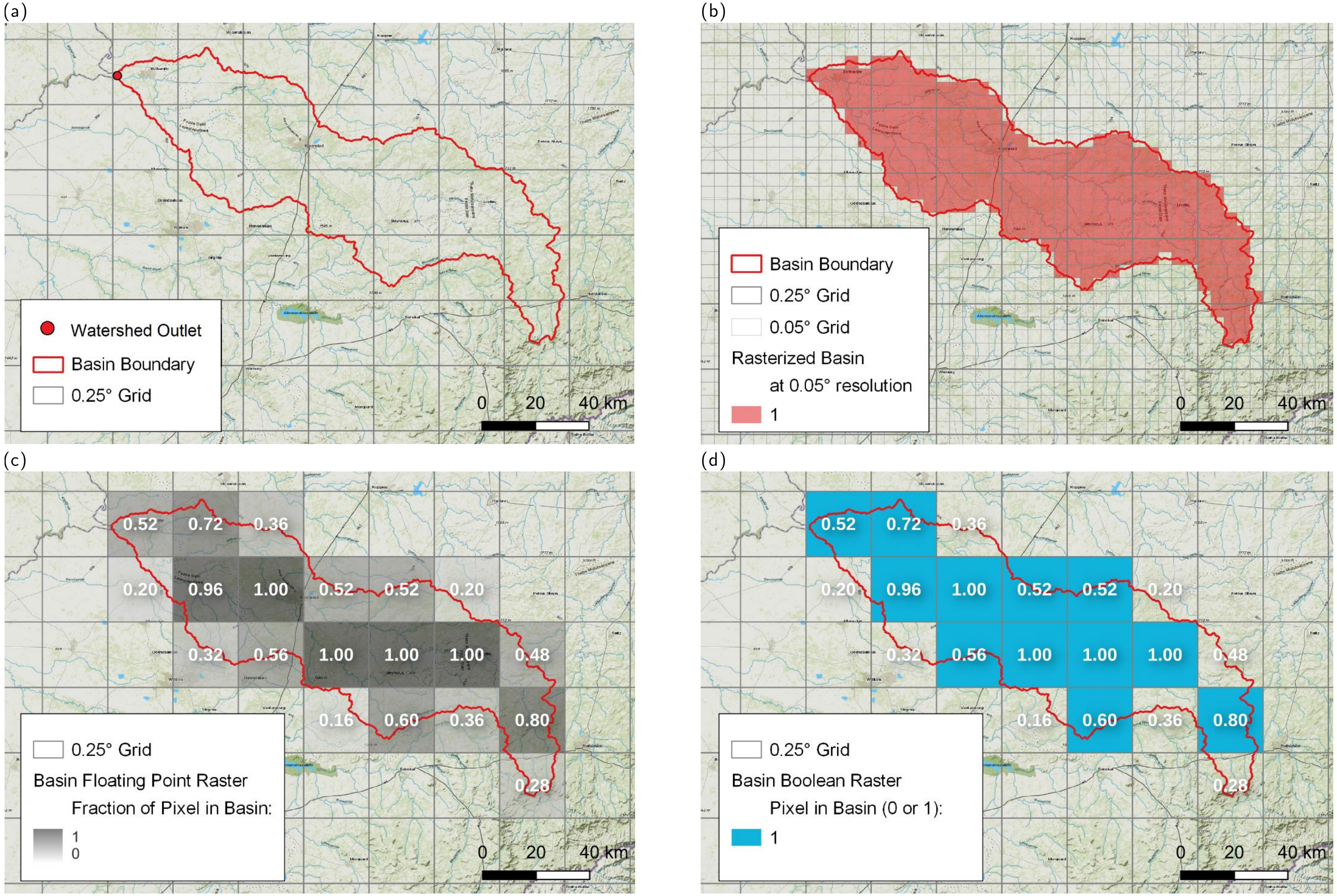


Figure 3.9: Rasterization of basin polygons to create basin masks, used for calculating the spatial averages over a basin of precipitation, evapotranspiration, or any gridded variable.

Table 3.1 summarizes the rasterized basin mask for our example basin, the Vals River watershed shown in Figure 3.9. Note that the Boolean mask contains fewer total pixels, but has a larger surface area. The Float mask contains more pixels in total, but most of the pixels have a value less than one, indicating that it is only partially contained in the basin. Overall, its weighted area (7,909 km^2 is lower, and much closer to the area of the vector polygon (7,901 km^2), a difference of only 0.1%.

Table 3.1: Statistics for basin masks

	# Pixels	Mask Area (km^2)	$\bar{P}$ (mm/month)
Vector		7,801	
Boolean	12	8,213	82.0
Float	20	7,909	78.6

As an example, I calculated the basin mean precipitation over the Vals River basin for February 2001, using CMORPH data, shown in Figure 3.10.

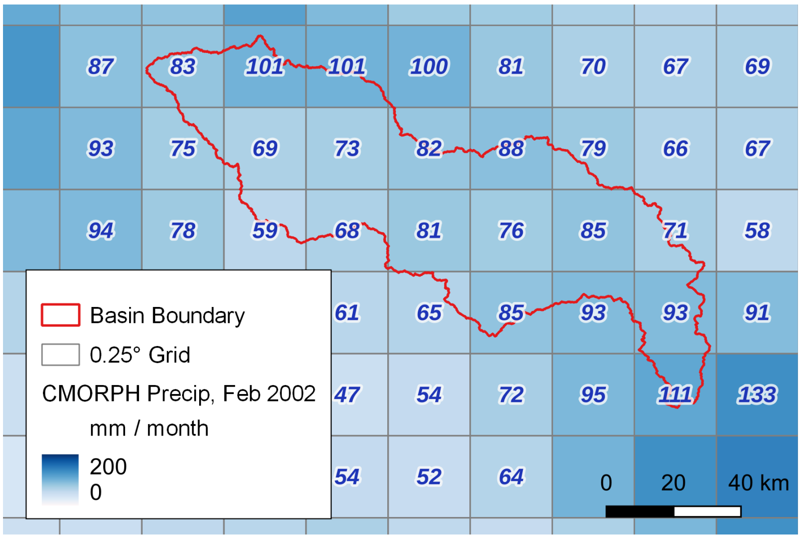


Figure 3.10: CMORPH precipitation over the Vals River basin in February 2001. Numbers in each grid cell represent the accumulated precipitation depth in mm in that month, or a downward vertical flux of water in mm/month.

The formula for calculating the weighted sum is as follows:

$$\bar{P}_b = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} B_{i,j}\, A_{i,j}\, P_{i,j}}{\sum_{i=1}^{m} \sum_{j=1}^{n} B_{i,j}\, A_{i,j}} \tag{3.2}$$

where we are summing over all the grid rows i from 1 to m, and over all columns j from 1 to n, B is the Boolean basin mask, A is the grid cell area in km², and P is the precipitation in mm/month. In principle, we can sum over all i and j, i.e. over every pixel on the planet, but the calculations are faster if we restrict i and j to those which are members of the basin, i.e. where $B = 1$. In practice, the Matlab code handles this automatically when we declare the grid mask as a *sparse matrix*. A sparse matrix is one where most of its elements are zero.

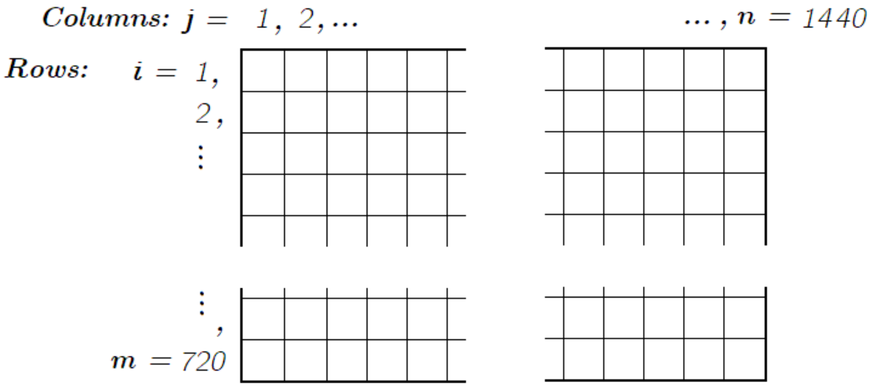


Figure 3.11: Indexing of grid cells.

Figure 3.11 shows the grid cell indexing, or numbering scheme used throughout the project. This is worth noting clearly here as different conventions are used by different agencies and research teams. In our grid, row numbering begins at the top of the globe (the north pole, or latitude +90° N) and increases as we move down or south. Rows are numbered from 1 to 720.[3] We use the variable i to refer to the row number, and the number of rows is m. Columns begin at longitude $-180°$, a line running north to south through the Pacific Ocean beginning in the north between Russia and Alaska. Columns are indexed from $j = 1, 2, \ldots, n$.

Finally, a note about the location of the pixels on the grid. The top edge of pixel (1, 1) is at 90° latitude, and its left edge is at $-180°$ longitude. The centroid of this pixel has latitude/longitude coordinates of (89.875, -179.875). Again, this is worth noting, as some data providers use a different convention of locating the centroid of the first column at -180. When this alternative convention is employed, an extra column of pixels is required to cover the globe. When this mapping method is used, a 0.25° global grid has 721 rows and 1441 columns. Compared to an ordinary edge-matched grid, there is a half-pixel width overlap where the eastern and western edge of the map join. These are small but important details that

[3] Throughout this thesis, I use 1 to begin numbering. This is the way it is done by mathematicians, rather than 0-based indexing that is favored by many computer programmers.

require careful attention of the programmer when dealing with data from multiple sources.

To calculate the area with the floating point basin mask, the formula is similar, but we multiply the terms in both the numerator and the denominator by the fraction in the basin:

$$\bar{P_f} = \frac{\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{n} F_{i,j}\, A_{i,j}\, P_{i,j}}{\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{n} F_{i,j}\, A_{i,j}} \tag{3.3}$$

where F is the floating point basin mask, and $0 \leq F \leq 1$.

I wrote efficient Matlab code using sparse matrices to speed up the calculation. The code also vectorizes two-dimensional matrices to avoid using loops in the code, which provides the biggest boost to calculation speed. Initial code implementations for calculating basin means of EO variables (obtained from other researchers in my laboratory) took 3 to 4 seconds on a laptop computer. Because I had to repeat this calculation many thousands of times, it was important to reduce the processing time. The new, efficient version reduced this to 15 to 20 milliseconds. This cut the total processing time from hours to minutes. This function, `calc_basin_mean()` is available in the code repository accompanying this thesis.

3.5 Preliminary Analysis of the Water Cycle Imbalance

In this section, we explore the earth observation (EO) datasets that we compiled to analyze the water cycle in river basins across the globe. I attempt to quantify the water budget with *uncorrected* EO data and show that it results in large budget residuals at both the pixel scale and river basin scale. The EO datasets referred to here as uncorrected are far from being raw data. Data on precipitation, for example, are typically calibrated to millions of observations collected at thousands of locations. Yet, our hypothesis is that further improvements are possible by cross-calibrating using data for other water cycle components.

This analysis reveals the extent to which there is a lack of consensus in predicting water cycle components. I also show the locations where the variance among different datasets is most pronounced.

The overarching goal of this research is to optimize earth observations of the water cycle. We saw previously that it is not possible to create a balanced water budget (or “close the water cycle”) using EO datasets. In principle, the fluxes plus change in storage over any given region should sum to zero, as expressed in Equation 1.1. However, this is rarely the case, leading us to the conclusion that one or more of the data layers contains errors. Recent research in hydrology and remote sensing has focused on how to combine or merge multiple datasets to reduce these errors. Our hypothesis is that there is a benefit of combining multiple *classes* of observations. Rather than focusing on calibrating individual water cycle components (for example precipitation), we seek to optimize all the components (P, E, ΔS, and R) simultaneously.

Figure 3.12 shows the water cycle imbalance for each of the 27 possible combinations of EO variables. The variables in this analysis are listed in Table 2.1 on page 43. There are 3 variables each for P, E, and ΔS, and 1 for observed R. The combinations are color-coded by the precipitation data source, which appears to have a large impact on the imbalance. Figure 3.12 also shows the simple weighted average solution as a black line, calculated as $I_{SW} = \bar{P} - \bar{E} - \overline{\Delta S}$. We can see that the simple step of averaging multiple datasets results in an imbalance that is less biased.

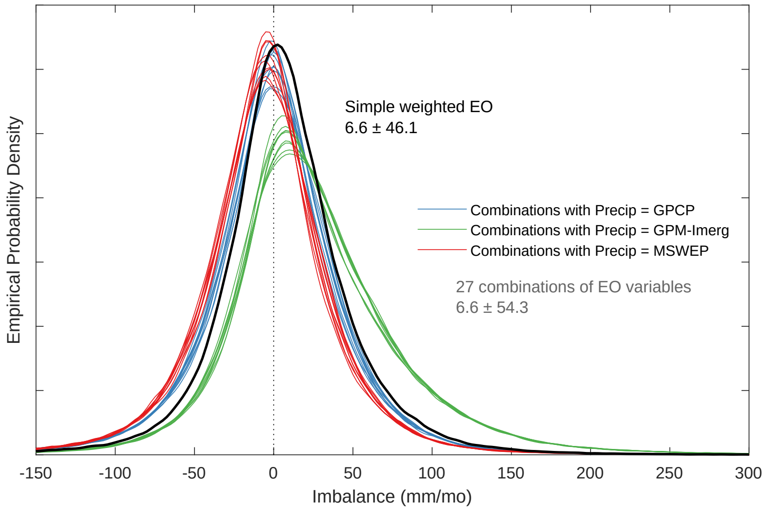


Figure 3.12: Distributions of the water cycle imbalance for each of the 27 possible combinations of EO variables, plus the simple weighted solution.

We can see from Figure 3.12 that the combinations that include precipitation from the GPM-IMERG dataset (light green lines) tend to have a higher imbalance than the other combinations, on average. The precipitation estimates are therefore more incoherent with the other water components. In addition, GPM-IMERG precipitation is higher on average than the other precipitation datasets. This can be seen in Figure 3.13, which shows the monthly mean precipitation over continental land surfaces. The GPM-IMERG dataset reports higher precipitation amounts, on average, across all months, compared to the GPCP and MSWEP datasets. This is illustrated in the set of maps in Figure 3.14. Each map shows the difference between average precipitation calculated in each grid cell, and the 3-member ensemble mean. Overall, GPM-IMERG reports higher precipitation across the globe as well, with the exception of the extreme west coast of South America, portions of Western North America, and around the Himalayan plateau in Asia. The differences can be quite large, up to 50 mm/month or more.

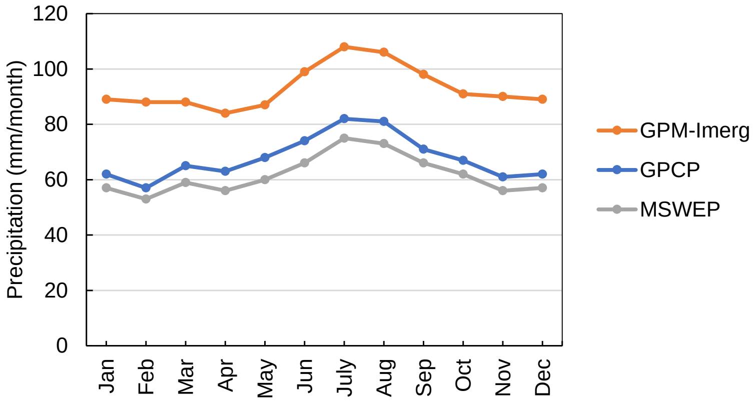


Figure 3.13: Monthly average precipitation across all terrestrial land surfaces (excluding Greenland and Antarctica) for the three precipitation datasets used in this study

Our goal is to make adjustments to remotely sensed observations of WC components such that the water cycle is balanced and hence improved. We seek to make these adjustments in a way that is methodical, mathematically rigorous, and has a solid basis in information theory. At the heart of our method is the idea that, for values with greater error or uncertainty, we have less confidence in their correctness, and therefore we will change them more. And for variables with a lower uncertainty, we are more confident that it is close to the correct value, and we will change them less.

Remote sensing of precipitation or another hydrologic variable involves a number of sources of uncertainty, which can be broadly categorized as *reducible* and *irreducible* uncertainties. Reducible uncertainties are those that can be reduced through improved measurement techniques, data processing algorithms, or calibration/validation procedures (Njoku & Li, 1999). Some sources of reducible uncertainty include errors in instrument calibration or in atmospheric correction algorithms, geolocation accuracy, and inversion algorithms. Irreducible uncertainties, on the other hand, are those that cannot be eliminated. Some sources of irreducible uncertainty are due to limitations of the instruments and sampling

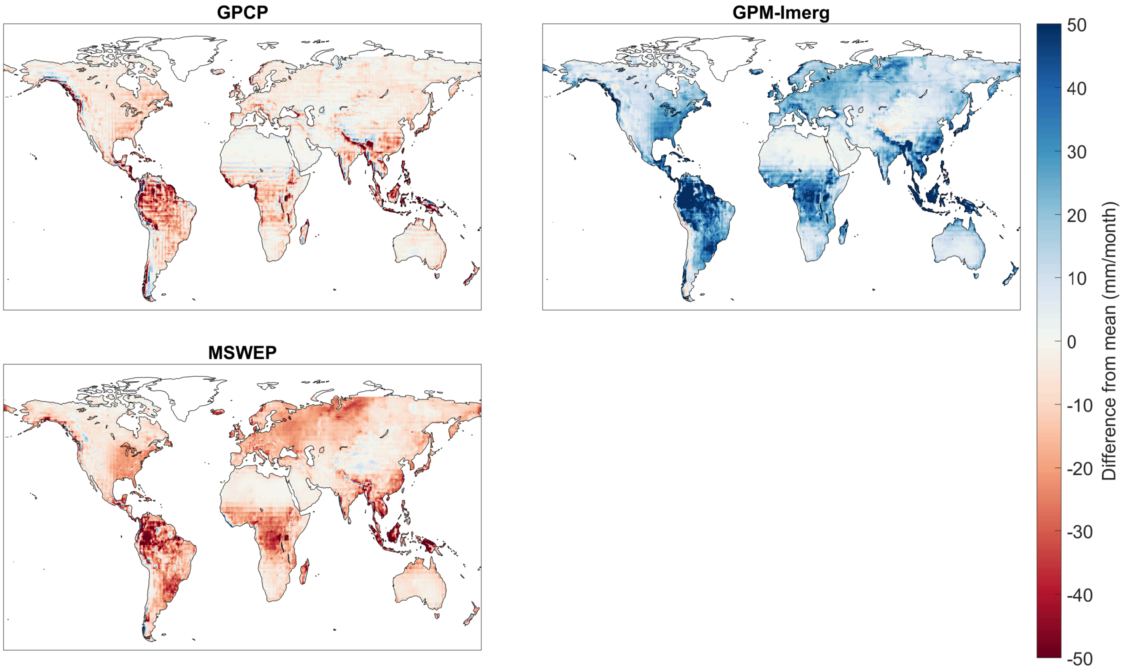


Figure 3.14: Maps of the difference between pixel mean precipitation and the ensemble mean for each dataset

rate. Variables such as precipitation are highly variable in space and time, and often cannot be fully captured by remote sensing.

I follow previous studies (Aires, 2014; Pellet, Aires, Munier, et al., 2019) in using Optimal Interpolation (OI) to integrate satellite-derived datasets and balance the water budget at the river basin scale. The OI method has been demonstrated and applied in several studies: over the Mississippi River Basin (Munier et al., 2014), over the Mediterranean Basin (Pellet, Aires, Munier, et al., 2019), and over five large river basins in South Asia (Pellet, Aires, Papa, et al., 2019).

The overall approach is based on forcing the imbalance, I, to equal zero. The OI calculation has two basic steps. First, we calculate a weighted average for each flux based on our input data, selected from among available EO datasets. This weighted average provides our best initial estimate for P, E, ΔS, and R. Next, the water budget residual is redistributed among each of the four water budget components using a post-filter matrix. OI makes adjustments to each variable in inverse proportion to the variable's uncertainty. This class of methods is well-known in the field of remote sensing, and their use in hydrology was introduced by Aires (2014). Such methods, well described by Rodgers (2000), are referred to as inverse methods and are widely used for remote sensing. In the following section, I introduce the mathematical notation and show how the calculations are done.

The equations below show how to redistribute the errors and balance the water budget for a *single observation*. In other words, we are dealing with the fluxes at one location and one time step. In this case, an observation is for a single river basin in a single month. This method does not contemplate distributing errors spatially among neighboring basins or pixels. Nor does it involve distributing errors over time, i.e. the month before or after the observation. Another important note: the equations and sample calculations shown here cover a single observation, with the inputs in the form of a vector. Multiple vectors are readily stacked into a matrix, and the calculations can be performed very efficiently, even with millions of observations.

3.6 Combining Multiple Estimates of Water Cycle Components

We begin with the problem of combining multiple satellite estimates of WC components. As we saw in Chapter 2, there are many earth observation datasets available for hydrologic variables. Indeed, I described over a dozen precipitation datasets, each of which are freely available and widely used. What does a practitioner do when faced with so much information? There are several common approaches. First, the analyst may choose a particular dataset out of custom or preference, or because a dataset is in a format that is compatible with existing software code. Other times, an analyst will do an *intercomparison* analysis to find the dataset which is most representative of their study area. A common approach is to compare the gridded data to in situ observations to determine which is the most representative of the target region (see e.g., Kidd et al., 2012; Duan et al., 2016; Huang et al., 2016). Another approach is to use the different datasets as inputs to a hydrologic model, and determine which ones result in the best predictions of observed runoff (Seyyedi et al., 2015; Guo et al., 2022).

Other analysts choose an ensemble approach, and simply average multiple datasets. The ensemble philosophy is that no-single “best-estimate” exists (Abolafia-Rosenzweig et al., 2021). This approach is commonly used in assessments of climate change. There are numerous climate models, and each gives somewhat different predictions of future climate. By considering the outputs of several models, the analyst can begin to see where there is consensus, and which outcomes are more uncertain. In the hydrologic sciences, rainfall-runoff modelers often combine data on precipitation from multiple sources (for example from radar, remote sensing, and gages). Using multiple sources allows the modeler to better capture the spatial and temporal variability of rainfall within the catchment (H. Liu et al., 2021). The ensemble approach is particularly valuable for flood forecasting, where the forcing is based on meteorological models that give divergent forecasts (Lee et al., 2019; Roux et al., 2020). Lorenz et al. (2014) raise the concern that biases in different datasets can cancel one another out, resulting in loss of information. One could argue that "cancelling out biases" is indeed the intent of the ensemble approach. When biases of individual datasets are smoothed out by averaging with other datasets, much has been gained in terms of accuracy.

There are also more systematic approaches to combining datasets, where the analyst seeks to extract the best information from each, often using some kind of weighting scheme. The simple weighting approach described below can be considered an example of this approach. Such approaches are valuable where certain datasets are known to be more or less accurate in certain regions, and can be given more or less weight. The accuracy of a given dataset may be determined with reference to in situ observations, by comparison to an ensemble mean, or via water balance calculations. An example of this approach is given by Lu et al. (2021). The authors created a global land evaporation dataset that incorporates information from multiple reanalysis models. Rather than taking a simple mean of the datasets, they used a method called reliability ensemble averaging (REA), which seeks to minimize errors by comparison to reference data (in situ and from remote sensing), and gives preference to consistencies among the products based on their coefficient of variation. Another example of this approach is the MSWEP precipitation data product (Beck et al., 2019), described in Section 2.2.3. The creators of this dataset have merged data from gauges, satellites, and reanalysis, seeking the highest-quality information at every location.

The inputs for our analysis are a series of climate and hydrologic observations, each of which is a hydrologic flux, in units of mass/time or water depth/time, which are equivalent. The inputs are:

- $[P_1, P_2, \ldots, P_p]$, p precipitation estimates;
- $[E_1, E_2, \ldots, E_e]$, e sources of information for evapotranspiration;
- $[\Delta S_1, \Delta S_2, \ldots, \Delta S_s]$, s sources of information for total water storage change;
- $[R_1, R_2, \ldots, R_r]$, r river discharge estimates.

The total number of observations is $n = p + r + e + s$, where n is the number of different datasets used as input. In practice, there is usually a single source of runoff data, thus $r = 1$, and this matrix contains a single column. River discharge is typically an in situ measurement, while all the other components are estimated satellite data products. The goal of OI is to combine these multiple estimates to obtain the best consensus of the water cycle state. For a detailed explanation of the mathematics behind OI, see Aires (2014) and Pellet, Aires, Munier, et al. (2019).

The "observing system" $\mathbf{Y}_\epsilon$ is an $n \times 1$ matrix (or column vector) that combines the n observations of WC components:

$$\mathbf{Y} = \left[P_1, P_2, \ldots, P_p, E_1, \ldots, E_e, \Delta S_1, \ldots, \Delta S_s, R_1 \ldots, R_r\right]^\mathsf{T} \tag{3.4}$$

The goal is to find the best estimator for the vector $\mathbf{X}$, a state vector representing the hydrologic cycle: This is the *truth* – the actual fluxes that occur in the environment and which we are trying to estimate from observations:

$$\mathbf{X} = \begin{bmatrix} P \\ E \\ \Delta S \\ R \end{bmatrix} \tag{3.5}$$

We seek to combine multiple observations in order to create an improved estimate of each flux. We can calculate any arbitrary linear combination from these data. A linear combination of variables refers to the expression formed by multiplying each variable by a constant coefficient and then adding the results. More formally, given variables $x_1, x_2, \ldots, x_n$ and coefficients $a_1, a_2, \ldots, a_n$, the linear combination of these variables is the expression: $a_1 x_1 + a_2 x_2 + \ldots + a_n x_n$. Here, $a_1, a_2, \ldots, a_n$ are constants.

In practice, many analysts choose a simple arithmetic mean. Or we can compute a weighted average if we believe that certain measurements are more accurate or reliable than others and thus should be given greater weight. A linear combination can be a useful way to combine measurements of the same variable obtained from different sources. If we have *a priori* information about how accurate a given dataset is, we can assign weights to the different estimates. In fact, it can be shown mathematically that this is the best, most robust way to estimate the mean, using the technique of inverse variance weighting, which is used widely in science and engineering.

The "observation operator" $\mathbf{A}$ is an $n \times 4$ matrix which serves to combine the observations into a formatted matrix for further analysis:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.6}$$

The matrix **A** lets us reformat our observations such that observations of each variable are aligned in the same column (in this order: $P, E, \Delta S, R$).

3.7 Simple Weighting

The simple weighted average is also used to combine multiple datasets of, for example, precipitation or evapotranspiration, into a single, unified estimate. Simple weighting, as introduced by Aires (2014) is an application of the method of inverse variance weighting (IVW) to the problem of calculating the best estimate that combines multiple remote sensing-derived water cycle components. IVW is a well-known technique used in many areas of science and engineering, commonly used in statistical meta-analysis or sensor fusion to combine the results from independent measurements (Hartung et al., 2008). It is a form of weighted average, where the weight on each observation is the inverse of the variance of that observation. Given multiple observations of precipitation, P_i, the IVW average, $\widehat{P}$ is calculated as follows. For the ith observation with variance σ_i^2, its weight is $\frac{1}{\sigma_i^2}$. The pooled weight is the sum of the individual weights:

$$\widehat{P} = \frac{\sum_i \frac{P_i}{\sigma_i^2}}{\sum_i \frac{1}{\sigma_i^2}}. \tag{3.7}$$

Here, the variance refers to the measurement of the precision of our measurements. We are not using variance in the sense of measuring how spread out a dataset is. (In this other sense of the term, the variance of a dataset is the average squared deviation of each data point from the mean of the entire dataset.) Instead, we are referring to variance as an evaluation of the precision of a measurement or experimental result. As such, the variance quantifies the degree of *uncertainty* in a measurement. So what we are calculating is how spread out are the measurement errors. In statistics textbooks, variance is usually described in the context of making repeated measurements of a single known variable. Each independent measurement will be slightly different, and the variance is how much noise there is among these different measurements. The formula for calculating variance in the context of measurement precision is:

$$Variance = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{3.8}$$

where:

- x_i is the ith measurement
- $\bar{x}$ is the mean of all the measurements
- n is the total number of measurements

In the context of satellite remote sensing, estimating the uncertainty of a data product is more complicated, and typically involves a combination of empirical validation, theoretical modeling, and expert judgment. Another important distinction between the simple textbook case and its application to remote sensing of WC components is that the variance is not likely to be constant, but to vary with the magnitude of the measurement and other factors. Because of this, the uncertainty is often expressed not as a constant but as a percentage, for example ±10%.

With inverse-variance weighting, we assign a heavier weight to observations with a lower variance, or with higher precision. We are saying that we are more confident in these observations and thus they should be given more importance when we calculate the average. However, there is still value in the other observations, even if they have a higher variance, and thus are more uncertain. It can be shown mathematically that this method for calculating the weighted average results in estimates with the lowest variance. However, we cannot guarantee that the result is unbiased, unless we can show that the measurements we are averaging are all unbiased.

For a simple case, assume that we have two observations of precipitation over an area, P_1 and P_2. We assume that the errors of each estimate have an average of zero (they are unbiased), and that they are normally distributed with standard deviations of σ_1 and σ_2 (or with variances of σ_1^2 and σ_2^2). The estimator of the mean precipitation based on these two observations with the least error (or the lowest variance) is given by:

$$\hat{P} = \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \right)^{-1} \left(\frac{P_1}{\sigma_1^2} + \frac{P_2}{\sigma_2^2} \right) \tag{3.9}$$

Doing some algebra to rearrange terms gives equation 6 in Aires (2014). Aires cites Rodgers (2000) as the source of this equation. Rodgers does not call the weighted averaging method IVW, but describes it as, "the familiar combination of scalar measurements x_l and x_2 of an unknown x, with variances σ_1^2 and σ_2^2 respectively," (Rodgers, 2000, Eq. 4.14, p. 67):

$$\hat{x} = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} x_1 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} x_2 \tag{3.10}$$

The equation looks slightly more complicated when $n = 3$:

$$\hat{P} = \frac{\dfrac{P_1}{\sigma_1^2} + \dfrac{P_2}{\sigma_2^2} + \dfrac{P_3}{\sigma_3^2}}{\dfrac{1}{\sigma_1^2} + \dfrac{1}{\sigma_2^2} + \dfrac{1}{\sigma_3^2}} \tag{3.11}$$

This can be rearranged and simplified as:

$$\hat{P} = \frac{\sigma_2^2 \sigma_3^2 P_1 + \sigma_1^2 \sigma_3^2 P_2 + \sigma_1^2 \sigma_2^2 P_3}{\sigma_1^2 + \sigma_2^2 + \sigma_3^2} \tag{3.12}$$

One of the key assumptions made when applying inverse-variance weighting is that the errors in the observations are independent, i.e. the errors in one variable are not correlated with the errors in any of the other variables. If the errors are correlated, then the multivariate form of inverse variance weighting, usually referred to as "precision-weighted" should be used (Hartung et al., 2008). In practice, information about covariance among the errors in remote sensing is not readily available.

In matrix form, simple weighting performs a weighted average of each class of hydrologic variable (P, E, ΔS, R), with the weights inversely proportional to their uncertainty. When no reliable estimates of the uncertainty are available, we may assign the same uncertainty to each variable within a class, and the calculation defaults to a simple arithmetic mean. For example, in the absence of detailed information on the uncertainty of EO precipitation datasets, we may assign the same uncertainty to $P_1, P_2, \ldots, P_n$.

$$\mathbf{X}_{sw} = \mathbf{K} \cdot \mathbf{Y}_{\varepsilon} \tag{3.13}$$

where **K** is a 4 x n matrix of weights that are created using the inverse variance weighting algorithm. This matrix can be constructed as follows. Consider the first row, $i = 1$, the entries are all zeros, except for columns $j = 1...p$: these entries represent the weights on the precipitation observations. The entries can be calculated for the first row as follows:

$$\mathbf{K}_{1,j} = \frac{1}{\sigma_j^2} \cdot \left(\sum_{k=1}^{p} \left(\frac{1}{\sigma_k^2} \right) \right)^{-1} \tag{3.14}$$

An alternative (equally correct) formulation is:

$$K_{1,j} = \left(\prod_{\substack{1 \leq k \leq p \\ k \neq j}} \sigma_k^2 \right) \left(\sum_{1 \leq k \leq p} \sigma_k^2 \right) \tag{3.15}$$

So the first entry (in row 1, column 1) will be given by:

$$\mathbf{K}_{1,j} = \frac{\frac{1}{\sigma_1^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} + \cdots + \frac{1}{\sigma_p^2}} \tag{3.16}$$

The variance of the simple-weighted variable is a function of the variances of the observations, as follows:

$$Var(\hat{P}) = \frac{1}{\sum_i \frac{1}{\sigma_i^2}} \tag{3.17}$$

The following section is an example demonstration of the simple weighting method. Suppose we have 4 observations of P, 2 observations of E, 3 observations of ΔS, and 1 observation of R, for $n = 10$. For a simple case, we assign the same variance to each of the classes of observation, i.e.: we assume that the precipitation variables each have the same uncertainty.

- Precipitation: $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 15.4$ mm/month
- Evapotranspiration: $\sigma_5 = \sigma_6 = 6.2$ mm/month
- Change in Water Storage: $\sigma_7 = \sigma_8 = \sigma_9 = 11.6$ mm/month
- Runoff: $\sigma_{10} = 1$ mm/month

In this case, the simple weighting matrix $\mathbf{K_{SW}}$ would be:

$$\mathbf{K_{SW}} = \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 1/3 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

So while **Equation 3.14** looks slightly complicated, it is just a convenient matrix form for averaging each class of hydrologic flux. In other words, we separately averaging the precipitation, evapotranspiration, change in storage, and runoff, using the IVW averaging method described above. It puts the weights into our $4 \times n$ matrix **K**. This lets us easily calculate **X**, our 4×1 matrix of average WC components computed from observations.

Now let us consider a slightly more complicated case. Suppose we had more information about the errors in our observations. Note we are still assuming that the errors are unbiased and that they are uncorrelated with one another.

$$\sigma_{1\ldots n} = \begin{bmatrix} 14.1 & 13.1 & 16.2 & 15.8 & 7.4 & 4.5 & 12.3 & 14.0 & 13.4 & 1.0 \end{bmatrix} \tag{3.18}$$

In this case, our matrix **K** would be as follows (with coefficients rounded to three decimal places):

$$\mathbf{K} = \begin{bmatrix} 0.269 & 0.312 & 0.204 & 0.215 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.270 & 0.730 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.382 & 0.295 & 0.322 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

In this case, the weights for each of the datasets are not equal. Rather, datasets with lower error will be weighted more highly in computing the average.

The matrix math shown above is simply one way to calculate the simple weighted average when we have multiple EO datasets estimating each of our WC components. The end result is a 4×1 vector **X** containing the 4 main fluxes in our simplified water balance: $P, E, \Delta S$, and R. While the estimated fluxes may be considered more reliable after having merged multiple inputs, there is no guarantee that they are coherent, or in other words, that they represent a balanced water budget by summing to zero. For this, additional calculations are needed.

3.8 Post Filtering

Our “first guess”, or best estimate for the water cycle state vector **X** is:

$$\mathbf{X_b} = \mathbf{X} + \varepsilon, \tag{3.19}$$

where ε is the error in the estimate, a 4 × 1 vector:

$$\varepsilon = \begin{bmatrix} err_P \\ err_E \\ err_{\Delta S} \\ err_R \end{bmatrix}. \tag{3.20}$$

The water balanced budget in Equation (1.1) can be expressed in matrix form as follows:

$$\mathbf{X}^T\mathbf{G} = 0, \tag{3.21}$$

where **G** is a 4 × 1 matrix that enforces the water balance as stated in Eq. (1.1):

$$\mathbf{G} = \begin{bmatrix} 1 & -1 & -1 & -1 \end{bmatrix}^{\mathsf{T}}. \tag{3.22}$$

After calculating the best first guess of the water budget $\mathbf{X_b}$, a post-filter is applied to enforce the water balance. In the section above, the best guess is from the simple weighted average method, so here, $\mathbf{X_b} = \mathbf{X_{SW}}$. Aires (2014) derived a solution for determining the linear combination of variables that satisfies the water budget constraint, weighting the contribution such that variables with lower error variance received greater weight. This method partitions the water balance residuals among the four water cycle components according to their uncertainty.

$$\mathbf{X}_{OI} = \mathbf{K}_{PF} \cdot \mathbf{X}_b, \tag{3.23}$$

where $\mathbf{X}_b$ is our first guess solution, which in our case is the simple weighted average, and $\mathbf{K}_{PF}$ is a 4 × 4 post-filter matrix:

$$\mathbf{K}_{PF} = \mathbf{I} - \mathbf{B}\mathbf{G}^T \left(\mathbf{G}\mathbf{B}\mathbf{G}^T\right)^{-1} \mathbf{G}, \tag{3.24}$$

where **I** is the 4 × 4 identity matrix, and **B** is the a priori error matrix of our simple weighted result. In this case, **B** only contains diagonal terms, as errors are assumed to be independent from one another for each of the 4 water cycle components. (In the case where the errors are correlated, the matrix **B** is the error covariance matrix.)

$$\mathbf{B} = \begin{bmatrix} \sigma_P^2 & 0 & 0 & 0 \\ 0 & \sigma_E^2 & 0 & 0 \\ 0 & 0 & \sigma_{\Delta S}^2 & 0 \\ 0 & 0 & 0 & \sigma_R^2 \end{bmatrix} \tag{3.25}$$

As $\mathbf{K}_{PF}$ is a $4\times$ 4 matrix, the post-filtering is simply performing a 4D linear transformation on our input matrix, the 4×4 simple weighted estimate of WC components P, E, ΔS, and R. The post-correction matrix K_{PF} is not invertible. It belongs to a class of matrices called singular or degenerate. This means that it is not possible to determine the left-side of Equation 3.23, or that there are infinite possible solutions.

The OI method is simple and effective, and has the advantage of not relying on any model. The post-filtered WC components in $\mathbf{X}_{OI}$ are always balanced, when the entries of $\mathbf{K}_b$ are any real numbers. While OI does an excellent job at enforcing the water budget the strict requirement to balance the water budget can occasionally produce unrealistic results. The OI method does not guard against returning negative values, which would be unrealistic for precipitation or runoff. Or it may produce values that are unrealistic because they are outside of the range of observations for a region, such as a P of 1,000 mm/month in a region that has never recorded this much rainfall in one month.

3.8.1 OI Relaxation Factor

Pellet, Aires, Munier, et al. (2019) introduced a relaxation factor into the OI equation, drawing upon previous work by Yilmaz et al. (2011) that relaxed the closure constraint during the assimilation. “This is an important feature because a tight closure constraint can result in high-frequency oscillations in the resulting combined dataset.” This modification makes the OI more flexible, but with a trade-off. With the relaxation factor, the OI will no longer force the imbalance to equal zero. Pellet et al. referred to the relaxation factor as Σ, the “tolerated WC budget residuals.” Here, I use the variable s to avoid confusion with the summation symbol. Pellet chose a value of $\sigma^2 = 2$ mm/month, or $s = 4$. We found that with our global dataset, this value is too low, and can give unrealistic results. I instead chose a value of $\sigma = 4$, or a $\Sigma = \sigma^2 = 16$. Allowing a slightly higher tolerated water budget residual gives a higher imbalance but more realistic values for the WC components. With the addition of the relaxation factor, modifications made by OI to the WC components are less aggressive. The imbalance has a mean near zero, and a standard deviation of $\sqrt{s}$. As

The new version of the post-filter matrix $\mathbf{K}_{PF}$ with the relaxation term is:

$$\mathbf{K}_R = \mathbf{I} - (\mathbf{B}^{-1} + \frac{1}{s}\mathbf{G}^{\mathrm{T}}\mathbf{G})^{-1} \cdot \frac{1}{s}\mathbf{G}^{\mathrm{T}}\mathbf{G} \tag{3.26}$$

As we increase s, modifications made to the input data are smaller and smaller, and the imbalance is larger. As $s \to \infty$, the solution approaches the null case, i.e. with the inputs unchanged and no reduction to the imbalance. Setting $s = 0$ makes $\mathbf{K}_R$ undefined.

3.8.2 Modifications to the OI post-filter

As we have seen in Equation 3.24, the variance (or estimated precision) of each flux in the water balance is an important determinant in the weights that will be assigned for redistributing the water balance residual. Therefore, it is worth considering how we shall estimate the variance in more detail.

The OI algorithm requires an a priori estimate of the error covariance matrix for our input variables, the remote sensing estimated fluxes. In practice, this information is rarely available, and therefore uncertainties are estimated by expert judgment or by computational experiments. Previous applications of OI assumed constant values for uncertainties, regardless of the season or the location. Such an assumption is defensible when analyzing a single river basin (the Mississippi in Munier et al., 2014), a single region, (Southeast Asia, in Pellet, Aires, Papa, et al., 2019), or the analysis is restricted to very large basins (11 basins, from the 620,000 km² Colorado River basin to the 4.7 million km² Amazon in Munier & Aires, 2018). However, this study aims to broaden the scope to have a truly global coverage, and our river basins cover a wide range of climates and hydrologic conditions, from highly arid to tropical rainforest.

With regards to errors in both discharge and EO estimates of WC components, there is considerable evidence that the error variance is not constant, but is proportional to the estimated flux. That is, it makes more sense to assign an uncertainty of say ±10% rather than to assign an uncertainty of 10 mm/month. Y. Liu et al. (2020) stated that discharge observations are assumed to contain "much smaller" uncertainty compared to the precipitation and actual evapotranspiration. This assertion is supported by the literature. Sauer and Meyer (1992) performed an error analysis of conventional in situ discharge measurements and concluded that errors are best expressed as a percentage of observed discharge. In a modeling study, Biemans et al. (2009) used an uncalibrated global rainfall-runoff model to assess the propagation of uncertainty in precipitation to predictions of discharge. Pre-modeling analysis showed that average uncertainty in precipitation over 294 global river basins is on the order of ±30%. Khan et al. (2018) systematically evaluated the uncertainties in datasets of evapotranspiration, integrating information from remote sensing (MOD16 and GLEAM), in situ measurements (flux towers) and reanalysis model output (GLDAS). They used the method of triple collocation to merge error information among the various datasets. The researchers converted absolute random errors in ET into relative uncertainties, calculated as the standard deviations of the error determined by their analysis, divided by the mean ET values.

Research by Tian and Peters-Lidard (2010) supports the idea that the uncertainty in satellite estimates of WC components are not fixed, but rather, they can be expressed as a percentage of the measurement. In this case, the authors analyzed an ensemble of precipitation data products in order to estimate the uncertainty over land and oceans. The authors created maps of the uncertainty, and rather than showing absolute values of the standard deviation from the ensemble mean, they chose to show the relative uncertainties as the ratios between the standard deviation and the ensemble mean precipitation. In other words, the uncertainty is represented as a percentage rather than as a value in mm. They also found that there are greater uncertainties over certain geographies: "complex terrains, coastlines and inland water bodies, cold surfaces, high latitudes and light precipitation emerge as areas with larger spreads and by implication larger measurement uncertainties."

This has also been referred to as an affine error model in the literature (Zwieback et al., 2012). This model accounts for both and additive and a multiplicative error, and is given as:

$$y_i^n = \beta_i x^n + \alpha_i + e_i^n \tag{3.27}$$

where y_i^n is the measurement number n by sensor i, x_n is the unknown variable and e_i^n is the corresponding error. The coefficient α_i is an additive bias term, and β_i is a multiplicative term. These coefficients can also be considered calibration terms. After they are determined, the difference between the sensor measurement and the true value is reduced to e_i^n, typically assumed to be Gaussian.

I adapted this error model to allow for larger errors in small measurements. Consider precipitation for example. In a given month, the EO dataset reports 0 rainfall. Nevertheless, a trace rainfall may have occurred and not been captured by satellite sensors. Therefore, I added a lower threshold at which a larger error is assumed. Here, we assign an uncertainty, σ, as a percentage of the flux. However, if the flux (in any given basin and in any given month) is below a threshold, then we will set it to a constant. I assigned the uncertainties for each water cycle component according to the following:

$$\begin{aligned} \sigma_P &= \begin{cases} 6 & \text{if } P < 30 \\ 0.2P & \text{if } P \geq 30 \end{cases} \\ \sigma_E &= \begin{cases} 6 & \text{if } E < 30 \\ 0.2E & \text{if } E \geq 30 \end{cases} \end{aligned} \tag{3.28}$$

I assigned a lower uncertainty to ΔS, as I do not want to the predicted values to depart too far from observations; as we will see later, one of the goals is to recreate the signal of total water storage with a statistical model, with the OI solution as the target. The errors in ΔS are in proportion to its absolute value, as it can take on both positive and negative values.

$$\sigma_{\Delta S} = \begin{cases} 3 & \text{if } \quad |\Delta S| < 30 \\ 0.1\Delta S & \text{if } \quad |\Delta S| \geq 30 \end{cases} \tag{3.29}$$

I assumed a slightly higher uncertainty for runoff, as we have lower confidence in the synthetic runoff provided by GRUN (dataset described in Section 2.5.2 on page 67. With runoff, we have more confidence in the estimates when R $\approx$ 0, and we do not want to make big changes to trace runoff, so we remove the threshold from the calculation of uncertainty.

$$\begin{aligned} \sigma_{R,OBS} &= \quad 0.2R_{OBS} \\ \sigma_{R,GRUN} &= \quad 0.4R_{GRUN} \end{aligned} \tag{3.30}$$

Finally, the OI can occasionally result in unrealistic values, such as negative precipitation or runoff. In such cases, we convert negative values to zero:

$$\begin{aligned} P &= \max(P, 0) \\ R &= \max(R, 0) \end{aligned} \tag{3.31}$$

3.9 Optimal Interpolation Results

This section presents the results of applying OI to two sets of water cycle observations, including 3 sources of precipitation, 3 sources of evapotranspiration, 3 sources of total water storage change, and:

1. Observed discharge over gaged basins
2. GRUN runoff over synthetic basins

First, let's look at how much we have changed the observations in each dataset with OI over the gaged basins. We can look at this information in a number of ways, each of which tells a different story about the data. Time series plots are perhaps the most intuitive. Figure 3.15 shows a set of plots for one of our 2,056 river basins. The data is for the White River at Petersburg, Indiana, United States, with a drainage area of 29,000 km^2. While no river basin is typical, this location does a good job demonstrating the output from our calculations as it has a long record of river discharge. The corrections made in this basin are relatively modest; over this region of the eastern United States, remote sensing datasets tend to be more reliable and well-calibrated due to the density and availability of in situ calibration data.

The analysis covers the 20-year period from 2000–2019, but we've zoomed in to a 6-year period to show more detail. The small plots on the right show the seasonality, or the monthly average of each variable. We can see that the simple weighted average (in black) tends to be in the middle of the EO variables (in gray). The OI solution tends to be close to SW, but slight perturbations have been applied. The changes are greater in months where there is a larger imbalance (bottom plots). With the standard OI algorithm, which I refer to here as "strict OI,", the Imbalance is always zero. For the OI solution with the relaxation factor, the Imbalance is not always zero, but it is much closer to zero than the imbalance calculated with uncorrected EO datasets. Thus, we see that OI has done exactly what we expect, which is to balance the water budget without departing too much from the values in the original EO datasets.

Next, we will look more globally at the changes that OI has made to the input data. Figure 3.16 shows a set of scatter plots, one for each EO input variable. The horizontal axes of each plot shows the uncorrected EO data. For example, in the first plot on the top left, the x-axis is for precipitation estimated by GPCP. The vertical axis is the OI solution for each water cycle component. In the first column, the data for the y-axis is the same for all three plots, because there is a single OI solution for P. The plots are for all months and all 2,056 basins (286,518 paired observations in total, with the color scale indicating greater density of points). We see that the majority of points are clustered around the 1:1 line, indicating that the adjustments made by OI are usually small ones. However, the cloud of points is also quite large, indicating that sometimes OI is making large adjustments to individual components in order to close the water cycle. Here, OI is imposing a

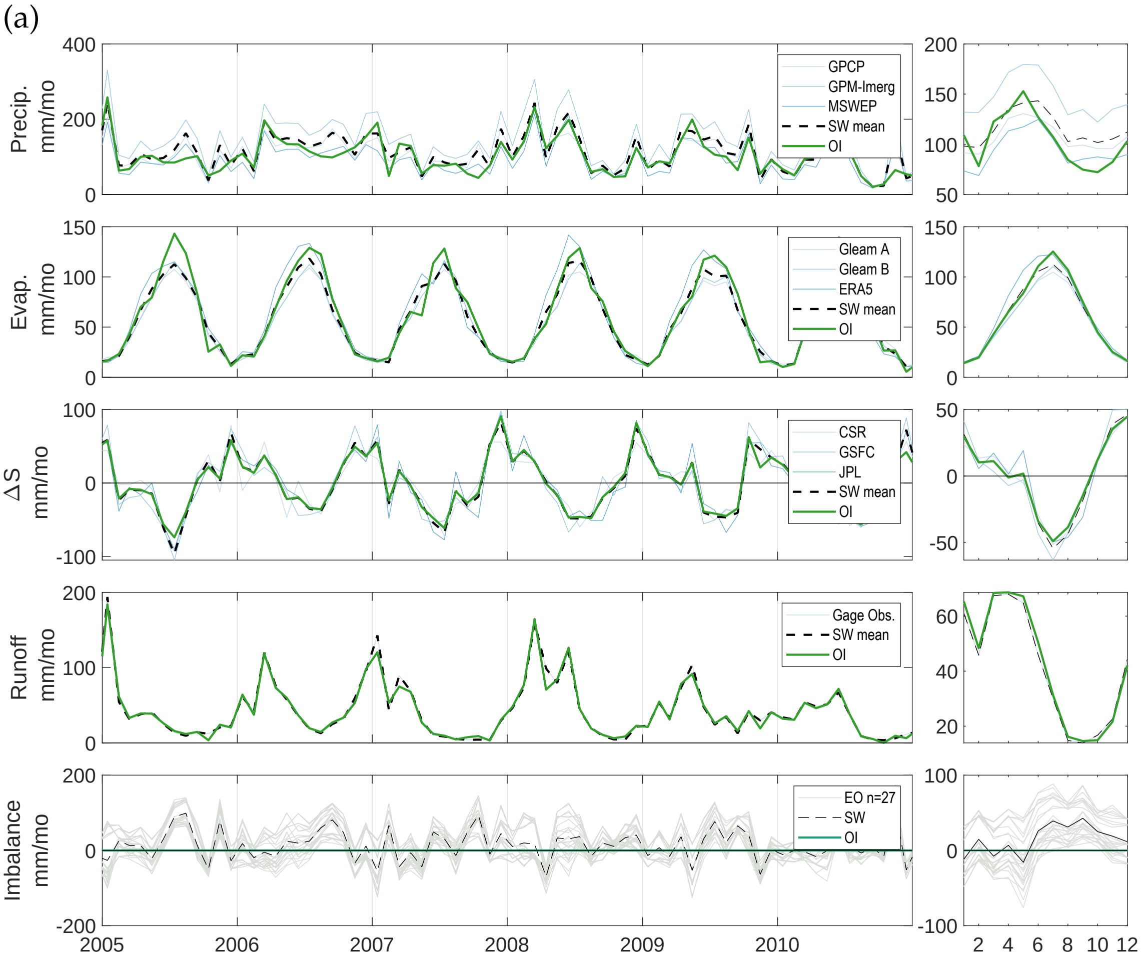


Figure 3.15: Time series plots of the four major water cycle components, showing remote sensing observations and the optimal interpolation solution at the White River at Petersburg, Indiana, United States (GRDC gage 4123202). The bottom plot shows the imbalance, or water cycle residual.

strong constraint, and occasionally large corrections are necessary. However, it is reassuring to know that such major corrections are relatively rare. Keep in mind that this is not an exercise in model fitting, where we are trying to maximize R^2. We set out to make modifications to the datasets, and these plots show the extent to which the observations have been modified.

The scatter plots in Figure 3.16 show the relationship between the inputs and the targets for the modeling methods to be described in the following section, Chapter 4. These methods will allow us to make corrections similar to OI, but can be made in ungaged basins, or where input data are incomplete.

Among the precipitation datasets, OI is making the largest changes to the GPM-IMERG dataset, with a root mean square difference (RMSD) of 49.8 mm/month. Figure 3.17 shows distributions of the changes made to each dataset by OI. The figures also report the mean and standard deviation for each dataset. The average

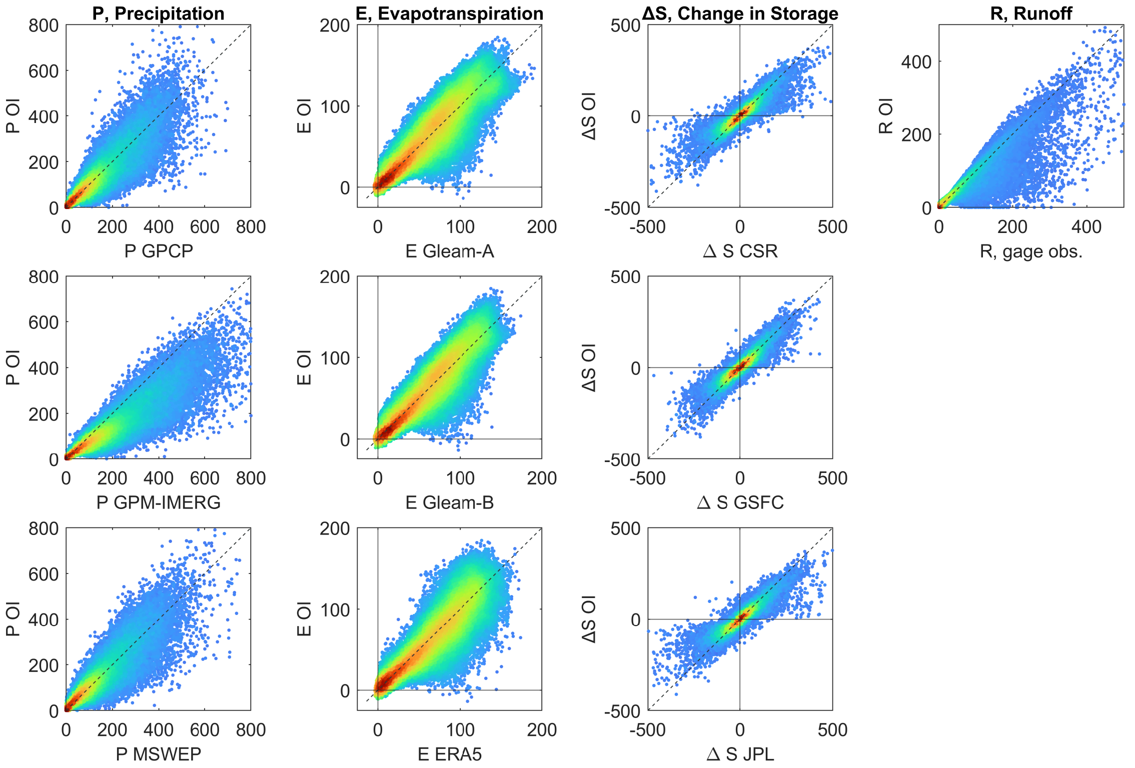


Figure 3.16: Scatter plots of uncorrected EO data vs. the OI solution, over the 2,056 gaged basins.

change is typically rather small, around 1 mm/month. The exception again is the dataset GPM-IMERG, where the OI solution is 27.6 mm/month higher, on average.

Finally, we may look at the geographic distribution of the changes OI makes to the datasets. Figure 3.18 shows the average difference between the simple weighted average of the EO datasets and the OI solution for each of the four water cycle components in each of the 2,056 basins. The average difference, equivalent to a bias when we are discussing model error, is calculated between the two time series as the mean of δ^i over all time steps i, where $\delta^i = P^i_{OI} - P^i_{SW}$ for all time steps i. Figure 3.18(a) shows the average corrections made by OI over the gaged basins, while part (b) shows the same over our synthetic river basins, where we are using runoff estimated by GRUN.

In Figure 3.18, each basin is mapped at its centroid, or approximate geographic center. Blue dots mean that $\delta > 0$, or the OI solution is higher on average than the SW mean, and red means that OI is lower than SW. We see that the OI solution for P is lower than SW over much of the globe. This means that observed precipitation is biased high, and that we have to revise P downward in order to close the water cycle. This is consistent with our previous observation that one of our

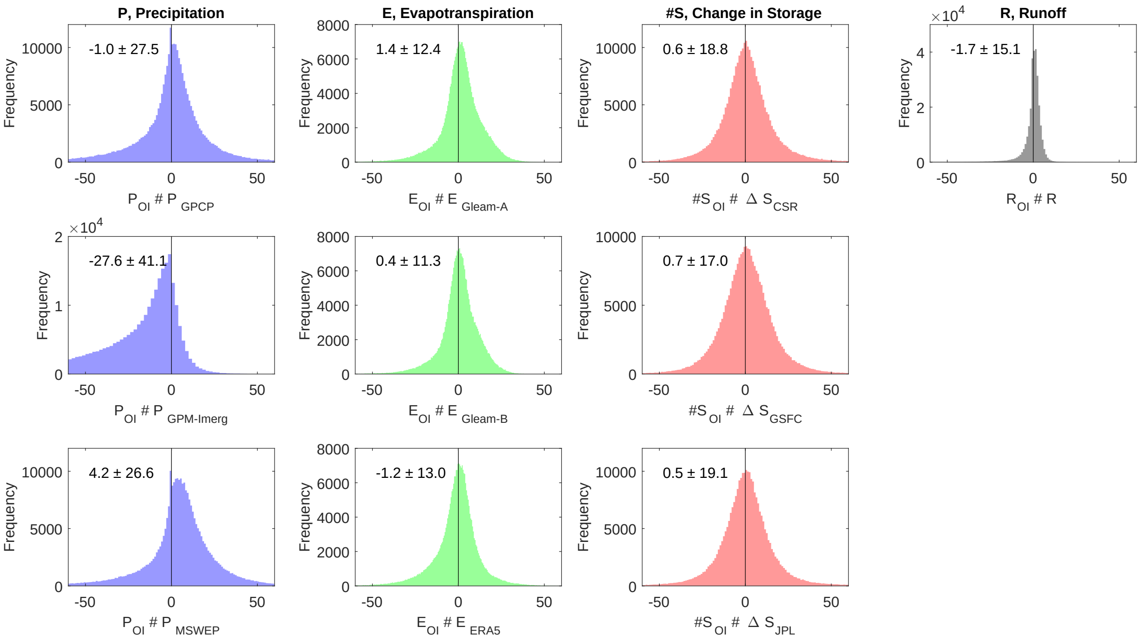


Figure 3.17: Distribution of the changes made to EO datasets by the OI algorithm. Represents all months over 2,056 gaged basins.

datasets, GPM-IMERG, consistently overestimates P over much of the globe. The corrections by OI are especially strong over parts of South America, the southern United States, India, and northern Australia. Corrections made to the other three water cycle components are lower in magnitude. Yet, distinct regional patterns emerge. There is a fairly consistent pattern over North America as we move from the northwest to the southeast. In the areas around Alaska and British Columbia, E is adjusted downward, and in the eastern and southern United States, E tends to be adjusted upward. Similar west to east patterns are apparent in South America and South Asia. Changes to runoff tend to be clustered in a few areas, most of which have colder climates. This is likely due to the influence of snow melt that contributes to runoff, and which may be inadequately quantified by GRACE.

We note largely the same patterns between (a) gaged basins and (b) synthetic basins. It is immediately apparent that the spatial coverage of the synthetic basins is much better. Indeed, this is one of the key motivations for working with this data. Nevertheless, in areas where both datasets have coverage, we see some differences. In particular, OI has adjusted runoff upwards in northeastern South America, while the changes made by OI to observed runoff is the opposite. We also see some differences in the changes made to evapotranspiration over South America. A thorough comparison is not possible because of the difference in spatial coverage between the two datasets.

3.10 Chapter 3 Conclusions and Discussion

This chapter described methods for combining remote sensing datasets to create a balanced water budget over river basins. I described methods for creating accurate river basin boundaries using topographic data, and for calculating spatial averages of gridded EO variables over the basin boundaries. Preliminary analysis of the EO data showed that there are serious inconsistencies among datasets, as there are large residuals to the water budget equation $P = E - \Delta S - R$.

I applied the optimal interpolation (OI) method over a collection of over 1,600 basins, a much larger number than previous studies using similar methods. The OI method leverages information from multiple satellite datasets, which may vary in accuracy by season or by location. This takes advantage of the idea that each dataset provides valuable information. Because the hydrologic conditions vary dramatically over the basins, I used a novel affine error model to make OI more flexible and return more realistic results. Maps of the difference between the EO datasets and the OI solution (Figure 3.18) can help illuminate areas where EO variables are less accurate and may need more detailed calibration.

pplying the OI method over thousands of watersheds is already a significant result. To my knowledge, this is the first study to apply these methods over such a large collection of river basins.

The OI method, despite its strengths, has certain limitations. OI imposes a very strong constraint, forcing closure of the water cycle by distributing the water balance residual. Any errors in one measurement will tend to propagate and infect the other components. Furthermore, it can only be applied over river basins, where we have access to river discharge data. One of the goals of this research was to find a method of optimizing EO variables at the pixel scale. In the following chapter, I describe modeling methods to extend the OI solution to the pixel scale.

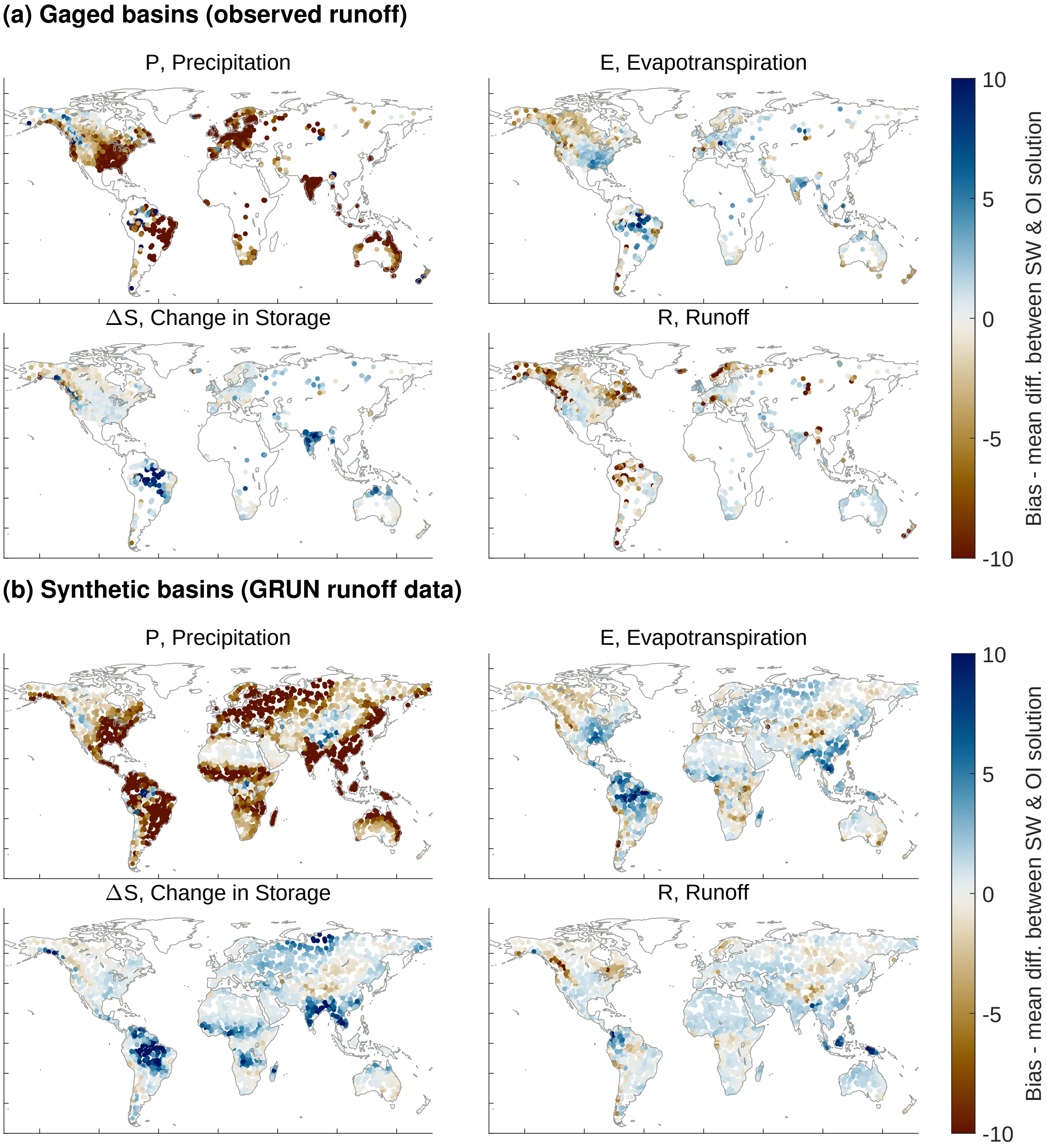

(a) Gaged basins (observed runoff)
P, Precipitation
E, Evapotranspiration
ΔS, Change in Storage
R, Runoff
10
5
0
-5
-10
Bias - mean diff. between SW & OI solution
(b) Synthetic basins (GRUN runoff data)
P, Precipitation
E, Evapotranspiration
ΔS, Change in Storage
R, Runoff
10
5
0
-5
-10
Bias - mean diff. between SW & OI solution

Figure 3.18: Map of the mean difference between the OI solution and SW average of observations for each of the four water cycle components.

Chapter 4

Modeling Approaches to Close the Water Budget

This chapter describes modeling approaches to balance the water budget with remote sensing datasets. In the previous chapter, we saw how optimal interpolation (OI) can be used to balance the water budget by making adjustments to each of the four main water cycle components: precipitation, P, evapotranspiration, E, total water storage change, ΔS, and runoff, R. A major disadvantage of this approach is that we must have all four of these variables. This usually means that OI can only be applied over river basins, where we have access to observed discharge at stream gages. I explored a workaround to this limitation by using a gridded dataset of synthetic runoff estimated by a statistical model (GRUN, Ghiggi et al., 2021). Even where we have access to runoff data, information on total water storage change (TWSC) is often a limiting factor. This water cycle component became available fairly recently, in 2002, with the launch of the GRACE satellites. So while there are a number of remote sensing datasets available from as early as 1980, we cannot use OI before 2002, because of missing data for TWSC. Further, there are a number of gaps in the GRACE record, as we saw in Section 2.4.1.

The main emphasis of the research for this thesis involved trying to find a model that can "recreate" the OI solution. Figure 4.1 shows a flowchart style overview of the steps in the modeling. In this chapter, I describe these two major modeling approaches. The first method uses simple linear regression models. The second approach uses more complex neural network models. Each class of models has distinct advantages and disadvantages, which will be discussed in this chapter. I applied these methods to both gaged basins and our synthetic basins where we used runoff from the GRUN dataset. The most useful model would be one that can be applied with one or more missing water cycle components, i.e., without R or ΔS. To this end, there is a major advantage to calibrating individual EO datasets, one at a time. The goal of the calibration is to make them closer to the OI solution and hence more likely to result in a balanced water budget. There are two main steps to the NN modeling method, as shown in Figure 4.1:

- Since it is preferable for the NN learning that these targets close the water budget, we will use the solution provided by optimal interpolation, following Pellet, Aires, Papa, et al. (2019).

- The output from OI creates the hydrologic time series that will be the target output used to train and validate the NN. These will be the four water main components of the water cycle, P, E, ΔS, and R.

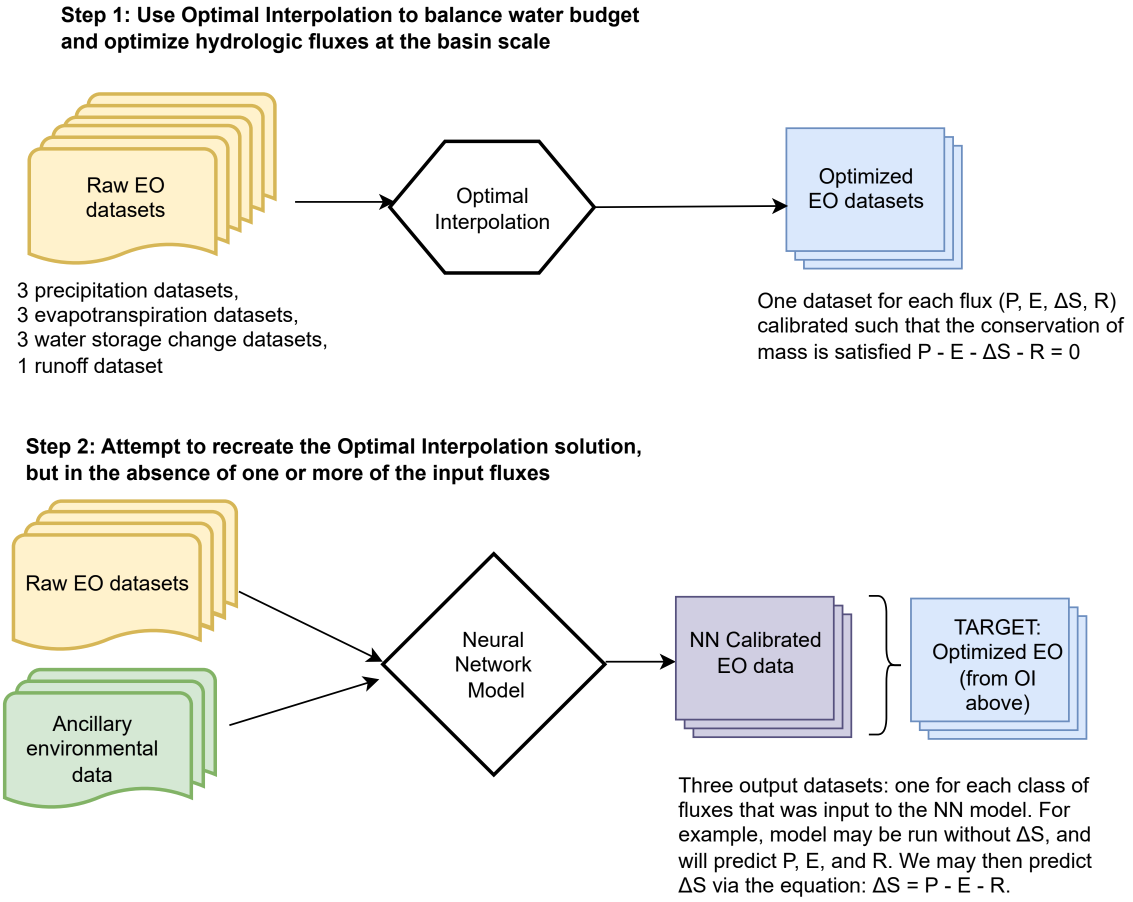

Step 1: Use Optimal Interpolation to balance water budget and optimize hydrologic fluxes at the basin scale
Raw EO datasets
Optimal Interpolation
Optimized EO datasets
3 precipitation datasets, 3 evapotranspiration datasets, 3 water storage change datasets, 1 runoff dataset
One dataset for each flux (P, E, ΔS, R) calibrated such that the conservation of mass is satisfied P - E - ΔS - R = 0
Step 2: Attempt to recreate the Optimal Interpolation solution, but in the absence of one or more of the input fluxes
Raw EO datasets
Ancillary environmental data
Neural Network Model
NN Calibrated EO data
TARGET: Optimized EO (from OI above)
Three output datasets: one for each class of fluxes that was input to the NN model. For example, model may be run without ΔS, and will predict P, E, and R. We may then predict ΔS via the equation: ΔS = P - E - R.

Figure 4.1: Overview of the two steps of the integration NN method.

Over the course of this research, I created many different models, with different configurations, sets of parameters, etc. How do we determine which of these is best? There is no single best indicator of goodness of fit of a model, or its predictive skill. I begin with a description of methods used for assessing model fit, and briefly describe the context in which they are useful. Recall that the overall purpose of this research is to more accurately estimate individual components of the hydrologic cycle. An additional goal is to use these predictions to better estimate long-term trends, particularly over large areas or areas with sparse ground-based measurements.

This chapter also includes a a discussion on the tradeoff between model complexity and its overall ability to make good predictions. A key consideration in the design and training of NNs is to avoid unnecessary complexity and overfitting, where a model fits the training data well, but performs poorly when asked to make predictions with new data.

In the following chapter, I present the results of the modeling analysis using techniques described in this chapter, including both regression-based models and neural networks.

4.1 Assessing Model Fit

In this section I describe indicators that used to assess the goodness-of fit between model predictions and observations. I also describe graphical methods of assessing model fit. No one indicator or plot is best, and often a combination of several should be used to choose the best model (Jackson et al., 2019). The methods apply to all of the modeling methods describe in the rest of the chapter.

During the course of my research, I obtained many solutions to the problem of creating a balanced water budget via remote sensing data. This involved many different model structures and parameterizations. This section is concerned with how to determine which of these is "the best." There is a rich literature on evaluating the goodness of fit of a model, or how well a simulation matches observations. Nevertheless, common practices vary in the fields of statistics, hydrology, and remote sensing. In each of these fields, prediction and estimation are important tasks, and we would like to know which model is the best at simulating or predicting "the truth," or is the closest fit to observations.

In the context of this research, there are several challenges to using conventional goodness-of-fit measures. First, the EO datasets that we are using in this study, have *already* been ground-truthed, or calibrated, to create a best fit to available in situ data. Therefore, it is unlikely that independently re-evaluating the fit to surface measurements will greatly improve these datasets. In other words, there would be little value to evaluating precipitation datasets against terrestrial observations from rain gages.

The second consideration when it comes to choosing a metric has to do with the geographic coverage of our meteorological and hydrological data. The output of our models is distributed across grid cells over the Earth's land surfaces. I searched for solutions that perform well across the globe – on different continents, and in different climate zones. The output of our model is also a time series (in river basins or pixels). Therefore I also looked for a solution that performs well in different seasons and times of the year. Overall, the following questions governed the search for a viable model for optimizing EO estimates of the water cycle:

- How well is our model predicting the solution obtained by Optimal Interpolation?
- How much does the model change the input EO data (P, E, ΔS, and R)
- Do the results balance the water budget? I.e.: What is the budget residual?
- What are the geographic and seasonal patterns of changes made to EO data by our model?

4.1.1 Errors and Residuals

In this thesis, I refer to two related concepts, measurement **errors** and modeled or estimated **residuals**. The error of an observation is the difference between the observation and the true value of the variable, which may not always be known or observable. Residuals are the differences between the observed values of the dependent variable and the values predicted by a model. Residuals are calculated as follows:

$$e_i = y_i - \hat{y}_i \tag{4.1}$$

where:

- $y_i = i^{\text{th}}$ observation
- $\hat{y}_i =$ model prediction corresponding to the i^{th} observation

4.1.2 Accuracy and Precision of Measurements and Model Predictions

There are two important concepts related to the quality of a measurement. The *accuracy* of a measurement refers to how close it is to the truth. A measurement's *precision* has to do with how repeatable a measurement is. In other words, if we repeat the same measurement multiple times, how close are the estimates to one another?

This is often illustrated with a target or bullseye, and clusters of points, and the analogy is that of a target shooter (presumably with a bow and arrow or a gun). When the cluster is centered on the bullseye at the center of the target, it is said that the shooter is accurate. When the points points are tightly clustered together, it is said that the shooter is precise. This has become a meme, as such figures can be found in many math, science, and engineering textbooks. However, I believe this is less than optimal as a visual aid for learners. First, the target is superfluously two-dimensional. We are not dealing with a scatterplot of x and y values. Rather,

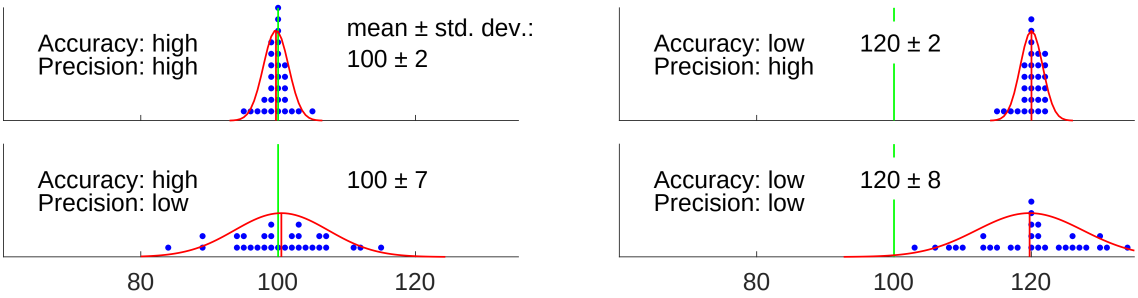


Figure 4.2: Illustration of the concepts of measurement accuracy and precision in one dimension.

the quantity that we are interested in is univariate – the distance from the target. It would may be more instructive to drop the analogy with shooting and instead describe a scientific measuremen.

As an example, suppose we are measuring the temperature at which water boils. The results may be displayed quite simply on a number line as a dot plot, as in Figure 4.2. On these plots, the green line represents the "truth" (100 °C, the boiling point of water at sea level). The blue dots represent 20 individual measurements.

Suppose the four plots represents measurements with four different thermometers, or done by four different teams, each of which is more or less skilled or careful. The series of measurements looks something like: 101, 99.5, 98.5, 100.5, etc. The red bell curve is the probability density function for a normal distribution fit to the observations, with a vertical red line at the observed mean. Where the mean is close to the truth, we say that the measurements are accurate. Where the measurements are spread out relative to one another, we say that they are less precise. On the top right, we can see that one of the four teams takes careful measurements. The observations are all close to one another. But they are all quite far from the truth, so while the precision is high, the accuracy is low. Conversely, in the bottom left plot, the observations are centered on the truth. On average, the measurements are good. However, the values are spread out, or we would say there is a lot of variability in the observations.

In the context of a model, the analog to accuracy is *bias*, and the analog to precision is the *standard error*, equivalent to the standard deviation of the residuals. It is common for time series models to either over-predict, under-predict the target variable (Jackson et al., 2019). Such a model is said to be *biased*.

This is also called root mean square difference RMSD, where difference refers to the residual between observations and model predictions. The plots in Figure 4.3 illustrate the concepts of the bias and standard error of model predictions.

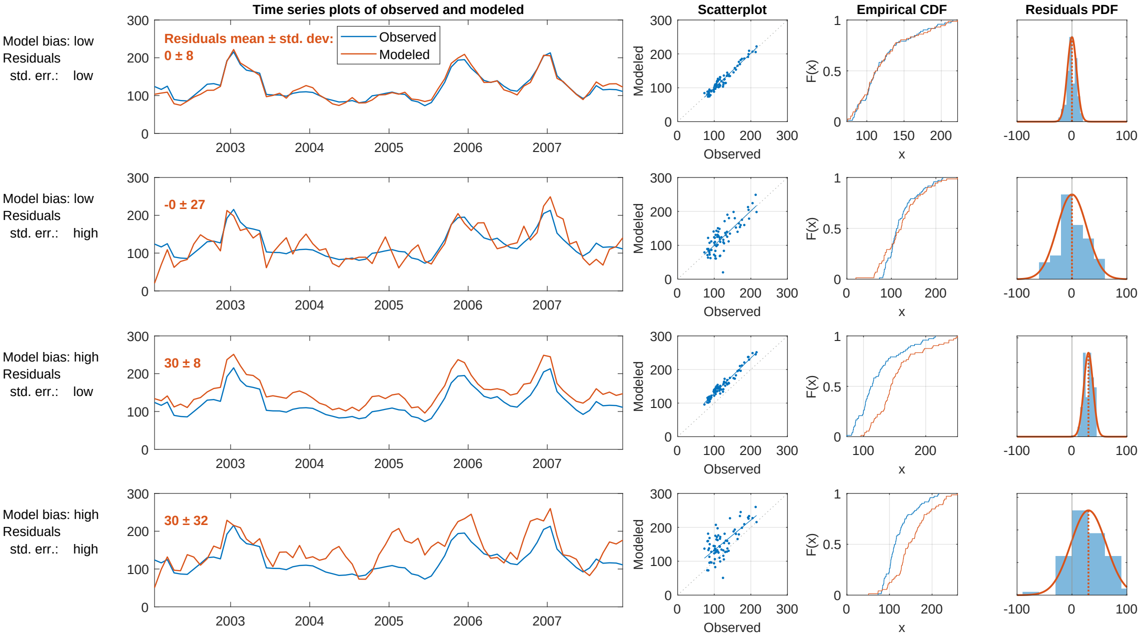


Figure 4.3: Illustration of the concepts of a predictive model bias and standard error.

4.1.3 Plots of Observed and Predicted Values

Time series plots are easily constructed and easy to understood. This is usually the first type of plot to inspect when working with sequential or time-variable data. Time series plots with separate lines for observed and predicted time series allow the viewer to quickly ascertain their similarity or difference. However, such plots can also be misleading or difficult to interpret. It is common for environmental data to have a skewed distribution, with frequent observations at lower values and fewer high values. In such cases, it is often helpful to display the vertical axis on a log scale. However, time series plots also have their limitations. Data often exhibit high levels of variability and noise, which can make it challenging to identify patterns or trends in the data. This can be particularly problematic when comparing observed and simulated data, as it may be difficult to determine whether any differences between the two datasets are meaningful or simply due to random variation. While visual inspection of the plots can provide some insight into the agreement between the two datasets, it does not provide a rigorous or objective assessment. For this reason, analysts rarely rely on such plots alone to judge the accuracy or reliability of models.

Scatter plots of observed versus predicted values are another useful graphical method for evaluating the model predictions. Thomann (1982) suggested that the most useful way to evaluate model fit is to plot the observed and predicted values, and calculate the regression equation and its accompanying statistics, slope,

a, intercept, b. While the discussion by Thomann was related to water quality models, the concepts are directly applicable to any type of predictive model. The best-fit model will be one where $a \approx 1$ and $b \approx 0$. We are also looking for a tight cluster of points about the 1:1 line ($x = y$), indicating a close agreement between observations and prediction (Helsel et al., 2020, Section 6.4). For this reason, in this manuscript scatter plots comparing observed and modeled data are square, have the same limits on the horizontal and vertical axes, and include a 1:1 line. When points fall mostly on one side of the 1:1 line, it is evidence of a biased model, which can be confirmed by calculating the mean or median residual as described below (see *Bias*, Section 4.1.7).

Empirical cumulative distribution function (ECDF) plots allow one to compare the distribution of datasets, and are useful for evaluating whether observations and predictions have comparable distributions. These plots are less intuitive than the plots described above, and require some orientation or training to read and interpret. By examining the vertical distance between the two ECDF curves at specific quantiles, researchers can assess the agreement or discrepancy between the two datasets at different points in the distribution. These plots plots also provide a visual representation of the tails of the distribution. By examining the behavior of the ECDF curves in the upper and lower tails, researchers can assess whether the model is able to predict extreme values or outliers that are present in the observed dataset.

The axes of an ECDF plot can be scaled according to the quantiles of a normal distribution, which allows the viewer to judge whether distributions are approximately normally distributed. When the y-axis of an emprical CDF plot can be transformed to a probability scale to create a Q-Q plot (so named as it plots theoretical quantiles vs. sample quantiles). When the normal distribution is used, it becomes a normal probability distribution. Similar plots can be constructed for any known probability distribution – common examples are the Weibull distribution. Hydrologists frequently use such plots, particularly for the study of high flows and flooding. On these plots, the plotted points (and the curve connecting them) becomes a non-exceedance probability, and are useful for estimating the expected frequency of high flows.

Histograms approximate the probability density function through sampling, and are a conventional way of visualizing the distribution of a dataset. By plotting the upper edge of histogram bins, or connecting their peaks, a histogram can be plotted as a line. This greatly facilitates comparing one distribution to another, as they can be overlaid on the same axes and maintain legibility.

We can also approximate the probability distribution function via a **kernel density plot**. Kernel density estimations (KDE), are a non-parametric method used to visualize the distribution of data. They provide a smoothed estimate of the probability density function (pdf) of a dataset, allowing for a more detailed understanding of the underlying distribution. The kernel density plot is created by placing a *kernel*, which is a smooth and symmetric function, at each data point and summing them to create a density estimate. The choice of kernel function and bandwidth, which determines the width of the kernel, is crucial in shaping the KDE and subsequent interpretations. A good balance must be struck between over-smoothing the data and hiding detail versus showing discontinuities.

4.1.4 Residual Plots

Plots of model residuals, e_i, versus time or the predicted variable are a valuable display of model fit. The goal is model residuals that are independent and randomly distributed. A good model will have a residuals pattern that looks like random noise, i.e. there should be no relationship between residuals and time (Helsel et al., 2020). If there is some structure in the pattern over time, it could be caused by seasonality, a long-term trend, auto-correlation among the residuals may be the cause. All of these are evidence that the model does not fully describe the behavior of the data. A plot of residuals versus *predicted* values is also useful for analyzing the structure of errors. Ideally, the variance of the residuals should be constant over the range of values of y. This is referred to as homoscedacity. When the variance is non-constant, or heteroscedastic, this is evidence that the model has not adequately captured the relationship among the variables.

4.1.5 Water Cycle Imbalance

The error in the water balance, or the residual, is also referred to as the *imbalance*. The imbalance I is calculated following Equation (4.2).

$$I = P - E - \Delta S - R. \tag{4.2}$$

One of the main motivators of this study is to discover a way to update or modify EO datasets to minimize the imbalance. So, we will judge a model in part by how close the imbalance is to zero. Therefore, the mean and standard deviation of the imbalance are two of our key indicators in judging the quality of our model. In summary, we seek to:

1. Minimize mean(I)
2. Minimize variance(I)

If we are only concerned with driving the imbalance to zero, the model may make wild and unrealistic changes to the input data. A "good" model should achieve a balance between modifying the input data the least amount necessary while also reducing the imbalance as much as possible. This is therefore a problem where we are seeking to simultaneously optimize more than one objective, and so other fit indicators are required in addition to the imbalance.

4.1.6 Mean Squared Error

It is customary to square the residuals before adding them to prevent positive and negative errors from canceling each other out. This also has the effect of penalizing model predictions more the further they are from the true value. The sum of squared errors, SSE, is:

$$\mathrm{SSE} = \sum_{i=1}^{n} e_i^2 \tag{4.3}$$

We are usually interested in the *average* error of a model's predictions. The mean square error (MSE) is calculated by averaging the squared residuals:

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} e_i^2 \tag{4.4}$$

One encounters MSE frequently as an objective function in the machine learning community. However, the units are usually not physically meaningful or intuitive as they are a squared quantity. Among earth scientists, it is customary to take the square root of the MSE, to express the error in the variable's original units.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}} \tag{4.5}$$

The RMSE is always 0 or positive. A value of zero indicates perfect model fit, and lower values indicate a better overall fit. A disadvantage of the RMSE is that comparisons cannot be made among different variables with incompatible units.

In recent hydrologic modeling literature, one frequently encounters a scaled form of the RSME. The RMSE standard deviation ratio (RSR) normalizes the RMSE by dividing by the standard deviation of observations. RSR combines error information with a scaling factor, thus the RSR value can be compared across populations, for example at sites with different flow rates and variances). The optimal value of RSR is 0, which indicates zero RMSE or residual variation and perfect model simulation. The lower the RSR, the lower the RMSE, and the better the model simulation performance.

$$\text{RSR} = \frac{\text{RMSE}}{\sigma} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}} = \frac{\sqrt{\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}}{\sqrt{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}} \tag{4.6}$$

4.1.7 Bias

Bias, in the context of model predictions, refers to a systematic error or deviation in the predictions that is not due to random chance but rather to a consistent and skewed pattern. A model that consistently overestimates or underestimate a variable is said to be *biased*. The best model will be one that is unbiased, that is, the average value of all the errors is zero. This is another way of saying that the expected value of the model prediction is the value of the response variable. Mathematically, bias of an estimator is defined as the difference between the expected value of the estimator and observed values, $\mathbb{E}(\hat{y} - y)$. Practically, the bias is readily estimated as the mean of the model residuals, as follows:

$$\text{Bias} = \text{Avg. prediction error} = \frac{1}{n}\sum_{i=1}^{n} e_i = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right) \tag{4.7}$$

Eqation 4.7 is equivalent to calculating the bias as the mean of predictions minus the mean of observations. Some analysts prefer to use the median rather than the mean, as it is a more robust estimator of central tendency with skewed data or in the presence of outliers.

It is helpful to scale the bias by dividing it by the mean of the observations. In this way, the bias can be compared across populations. For example, we may wish to compare flow predictions across basins with different average flows. The percent bias, or PBIAS is:

$$\text{PBIAS} = \frac{\text{bias}}{\bar{y}} \times 100 = \left[\frac{\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)}{\sum_{i=1}^{n} y_i}\right] \times 100 \tag{4.8}$$

All is not lost when models output biased results. If the bias is consistent, it can often be corrected. Simple methods of bias correction involve adding or subtracting a constant value to model output. More complex methods include quantile matching or cumulative distribution functions (CDF) transformation. The most complex methods of bias correction involve the use of a statistical or machine learning model to "post-process" model results (King et al., 2020). Bias correction is particularly common in climate science, especially when looking at the local or regional impacts of climate change. In general, global general circulation models (GCMs) do a poor job of predicting local conditions due to their coarse grid cells, their inability to resolve sub-grid scale features such as topography, clouds and land use. However, downscaling results often suffer from bias, and therefore bias correction methods are commonly used (Sachindra et al., 2014).

Remark

Somewhat confusingly, the term bias is used in different ways in the fields of statistics and machine learning. This will be discussed below in Section 4.2 on the bias-variance tradeoff. In this context, bias does not refer to a statistic of the residuals, as defined here. Rather, it refers to the overall fit of a model relative to the data used to train the model. Additionally, in the fields of artificial intelligence and machine learning, one frequently encounters the term bias to describe discrimination and unfairness in model predictions. This is an important aspect of AI that needs to be urgently addressed; fortunately, it has no impact on the research being described here.

4.1.8 Variance

The term *variance* is used in different ways across various fields in the natural sciences and machine learning. In univariate statistics, variances is a measure of how much spread there is in a variable's values. More specifically, variance is calculated as the average of the sum of squared deviations from its mean, as follows:

$$\text{Variance} = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{4.9}$$

where:

- x_i is the ith measurement
- $\bar{x}$ is the mean of all the measurements
- n is the total number of measurements

In this section, we are concerned with assessing the quality of models. The variance can be readily used to quantify the spread in prediction errors, e. To calculate the residual variance, we simply substitute the residuals, e for x in Equation 4.9. The custom is to refer to the sample estimate of the variance as s^2, and the variance of a population as σ^2. The sample standard deviation, s, is the most common measure of sample variance, and is simply the square root of the estimated variance. The standard deviation of residuals is also referred to as the standard error.

Overall, we are interested in finding a model that minimizes both bias and standard error (has low residual variance). The concepts of bias and standard error are analogous to accuracy and precision, which are commonly used when discussing observation or measurement methods.

Variance is also the name for a key concept in statistics and machine learning that will be discussed further in Section 4.3.7 on Cross-Validation. In this context, variance does not refer to a statistic of a random variable, as we have defined it here. Rather, it refers to how much a model fit to data changes each time we use a new sample of observations for the training data set. I believe this unfortunate naming has caused significant and unnecessary confusion among students and practitioners.

4.1.9 Correlation Coefficient

The Pearson correlation coefficient is a well-known estimator of the linear association of two variables:

$$\mathrm{R} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{4.10}$$

Values for R range between -1 and +1. When $R = 1$, it means that a model's predictions are perfectly correlated with observations. It does not, however, mean that the predictions are perfect, or even unbiased. This is discussed in more detail in the discussion of the related indicator R^2 below.

A value of $R \approx 0$ means that the variables are not correlated with one another. A hypothesis test can be conducted using R, where the null and alternate hypothesis, $H_0 : R = 0$ or $H_A : R \neq 0$. However, this test is not valid in the presence of outliers or with skewed data (Helsel et al., 2020, p. 212). The test statistic is computed by Equation 4.11 and compared to the critical values of t-distribution with $n - 2$ degrees of freedom.

$$t_R = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \tag{4.11}$$

4.1.10 Coefficient of Determination

The coefficient of determination, R^2 gives the proportion of the variation in a dependent variable that can be explained by an independent variable.

A general definition of the coefficient of determination is given by:

$$R^2 = 1 - \frac{SS_{residuals}}{SS_{total}} \tag{4.12}$$

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{4.13}$$

For a simple linear regression model, R^2 is directly related to the correlation coefficient, R, and can be calculated simply as the square of R. In this case, R^2 will range from 0 to 1. For other types of predictive models, R^2 can take on negative values. The definition of the coefficient of determination gives rise to useful rules of thumb for its interpretation. A baseline model that consistently predicts the mean value $\bar{y}$ will yield an R^2 value of 0. Models with predictions less accurate than this baseline will exhibit a negative R^2.

It can be misleading to rely solely on R or R^2 as the sole determinants of model fit. In a number of cases a high R^2 belies a poor or biased model. The best model is one with a slope $m = 1$ and an intercept $a = 0$. Figure 4.4 illustrates three cases where $R^2 = 1$, but the model has a significant bias, indicated by a slope $a \neq 1$ or intercept $b \neq 0$.

4.1.11 Mean Absolute Error

Some authorities recommend the use of Mean Absolute Error (MAE) over the RMSE, as its interpretation is more straightforward. In this indicator, all residuals are given equal weight. This is unlike the MSE, where the residuals are squared, placing greater importance on larger errors. The MAE is simply the average of the absolute values of the residuals:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i| \tag{4.14}$$

4.1.12 Nash-Sutcliffe Model Efficiency

The Nash-Sutcliffe Model Efficiency (NSE herein, also sometimes E) is a fit indicator originally introduced for evaluating the skill of models that produce time series

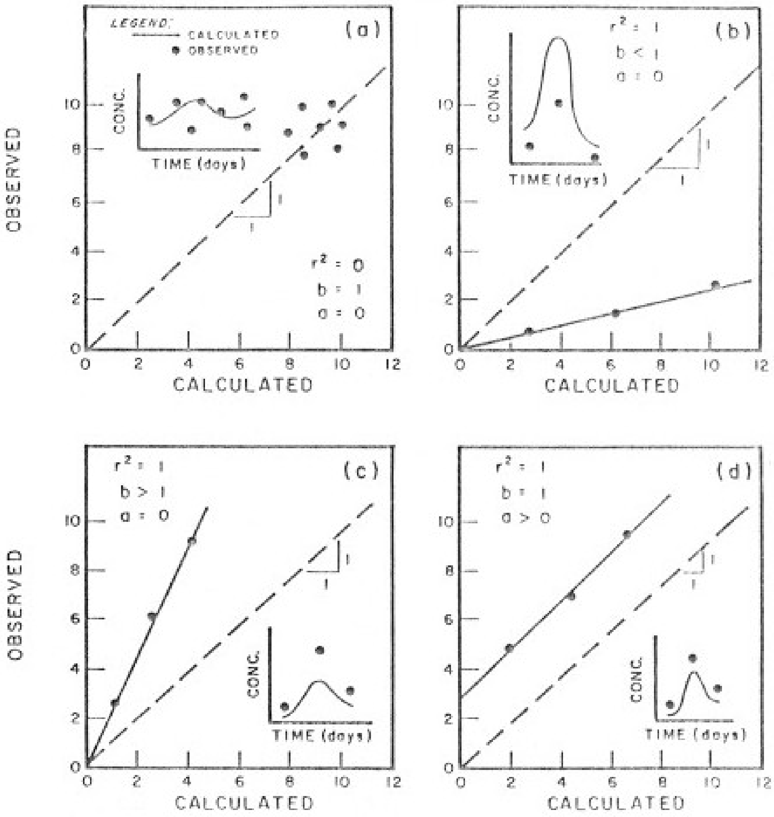


Figure 4.4: Possible cases in regression between calculated and observed values. Reprinted from Thomann (1982).

output. While the coefficient of determination (R^2) is useful in evaluating the fit of an ordinary least squares regression equation, the NSE is a better determinant of model fit when comparing observed and modeled time series. Nash and Sutcliffe (1970) originally proposed this statistic for evaluating rainfall-runoff models, and it is widely used in the hydraulic and hydrologic literature. Because of its advantages, it has become more widely used in the geosciences and climate research. NSE is a normalized form of the mean square error (MSE). Table 4.1 compares NSE to R^2, including notes on its interpretation. It is calculated as follows:

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n}(x_i - y_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{4.15}$$

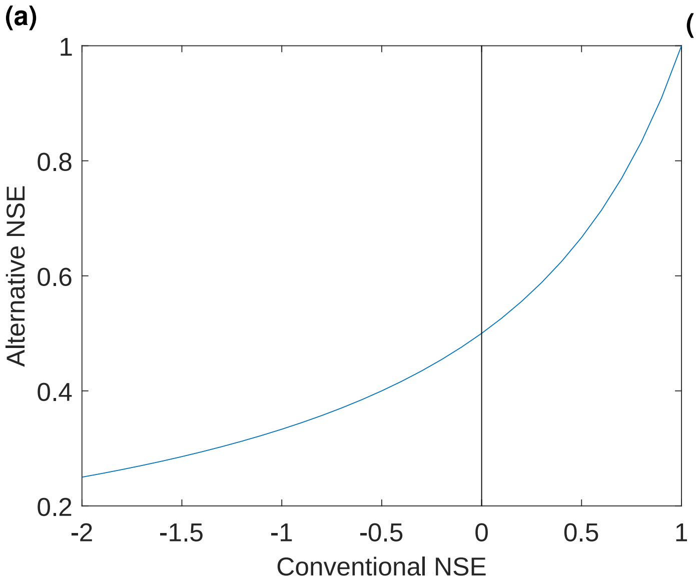
Table 4.1: Comparison of the Nash-Sutcliffe Efficiency, NSE, and the Coefficient of Determination, *R*²

Nash-Sutcliffe Model Efficiency, NSE	Coefficient of Determination, R^2
$\mathrm{NSE} = \frac{\sum(y-\bar{y})^2 - \sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}$	$R^2 = \frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\bar{y})^2}$
$-\infty < NSE \leq 1$	$0 \leq R^2 \leq 1$
NSE = 1 means model is perfect, i.e., $\hat{y}_i = y_i$ for all i	$R^2 = 1$ means model predictions are *perfectly correlated* with observations (but says nothing about bias)
NSE < 0 means model performs worse than the model $\hat{y} = \bar{y}$	$R < 0$ means a generic model performs worse than the model $\hat{y} = \bar{y}$ (but this is not the case for OLS regression)

An inconvenience of the NSE is that it occasionally takes on large negative values for poor fits. This makes it challenging to average values over many locations. Mathevet et al. (2006) introduced a useful variant of the NSE which has a lower bound of -1, and keeps the usual upper bound of +1.

$$\text{NNSE or NSE}' = \frac{\mathrm{NSE}}{2 - \mathrm{NSE}} \tag{4.16}$$

This variant is useful when calculating the average NSE, for example at multiple locations. Of course, another way around this limitation is to use an estimator of central tendency such as the median, which is more resistant to outliers. As a consequence of this transformation, positive values of NSE$'$ are somewhat lower than the conventional NSE. The difference is illustrated shown in Figure 4.5, with (a) a plot of the conventional vs. the alternative bounded NSE, and (b) a table comparing the values.



(b)

NSE	NSE′
1.0	1.0
0.9	0.81
0.7	0.54
0.5	0.33
0.0	0.0
-0.5	-0.2
-1.0	-0.33
-1×10^3	-0.998
-1×10^6	-0.999998

Figure 4.5: Comparison of the standard Nash-Sutcliffe model efficiency with its bounded version.

In a recent conference paper, Moshe et al. (2020) define a variant called NSE-persist. It follows the same general form, defining the model efficiency as:

$$E = 1 - \frac{\mathrm{RMSE}(\mathrm{residuals})}{\mathrm{RMSE}(\mathrm{BASELINE})} \tag{4.17}$$

where BASELINE is "a naive model that always predicts the future using the present reading of the target label measurement," or $\hat{y}_i^{t+1} = y_i^t$ (Moshe et al., 2020, p. 6).

This variant is useful for assessing assimilation models or certain types of machine learning models, such as long-short term memory (LSTM) neural networks. In such models, the prediction at time step t is a function of the observation at the previous time step $t - 1$. As such, the model predictions have high autocorrelation, and conventional metrics tend to inflate the predictive skill of the model.

Lamontagne et al. (2020) introduced a new variant formula for calculating NSE, which is meant to respond some of the critiques of the NSE in the literature. Their new formulations are more resistant to skewness and periodicity, common features of hydrologic data. In a series of Monte Carlo experiments with synthetic streamflow data, the authors show that their revised estimator is a significant improvement. As of this writing, the new estimator appears to have very limited use. Thus, I chose to use the conventional estimator of NSE in Equation 4.15.

4.1.13 Kling-Gupta Model Efficiency, KGE

Another variant on the NSE was introduced by Gupta et al. (2009) and further elaborated by Kling et al. (2012). The so-called Kling-Gupta Model Efficiency (KGE) is a composite indicator, combining three components – correlation, bias and variability. It is now widely used in the hydrologic sciences for comparing model output to observations.

$$KGE = 1 - \sqrt{(R-1)^2 + \left(\frac{s_x}{s_y} - 1\right)^2 + \left(\frac{\overline{y}}{\overline{x}} - 1\right)^2} \tag{4.18}$$

where:

- R is the correlation coefficient between the observations, x and the model predictions, y
- s_x is the standard deviation of the observations, s_y is the standard deviation of predictions
- $\bar{x}$ is the mean of observations, and $\bar{y}$ is the mean of y,

Like NSE, KGE ranges from $-\infty$ to +1. For the case where the model performance is equivalent to simply predicting the mean of observations $\hat{y} = \bar{x}$, we have NSE = 1 and KGE ≈ -0.41. Thus, these two related fit indicators are not equivalent. Further, if we choose the mean of a variable as our benchmark, we, performance over the range $-0.41 <$ KGE ≤ 1 would be considered reasonable as the model outperforms this benchmark.

Similar to the bounded version of NSE in Equation 4.16, some researchers prefer working with the bounded version of KGE (Mathevet et al., 2006). It has the advantage of being bounded by -1 and +1, and makes it simpler to average or compare values from multiple sites. The bounded version of KGE is:

$$\text{KGE}_B = \frac{\text{KGE}}{2 - \text{KGE}} \tag{4.19}$$

While KGE is designed as an improvement upon R^2 and NSE, it is not without its detractors. Lamontagne et al. (2020) state that its major flaw is that its components are based entirely on product moment estimators. This is acceptable when the data are well-behaved, i.e. normally distributed data. However, hydrologic data is often highly skewed (hence far from normally distributed), so the ratios of the product moment estimators “exhibit enormous bias, even for extremely large sample sizes in the tens of thousands and should generally be avoided" (Lamontagne et al., 2020).

4.1.14 Cyclostationary NSE

An alternative form of the NSE considers the departure of model predictions from the seasonal mean. This formulation was introduced by J. Zhang (2019) and used in a recent water-balance study by Lehmann et al. (2022). This is useful when the data we are modeling exhibits a strong seasonal cycle. This is the case for GRACE TWSC in many locations. Consider a naive model that can reproduce the seasonal cycle with fidelity but has no skill beyond this in predicting anomalies. Conventional fit indicators will inflate the skill of the model. Instead, it is more honest to assess the skill of the model in recreating the anomalies beyond the base seasonal signal. The Cyclostationary NSE (CNSE) is calculated as follows:

$$\mathrm{CNSE} = 1 - \frac{\sum_{i=1}^{n}\left(x_i - y_i\right)^2}{\sum_{i=1}^{n}\left(x_i - \tilde{x}\right)^2} \tag{4.20}$$

where:

- $\tilde{x}$ is the long-term monthly average of the variable x, i.e. the climatology
- $\bar{x}$ is the mean of x

This indicator is related to the Cyclostationarity Index (CI), an indicator of the strength of the seasonal signal introduced by J. Zhang (2019): the cyclostationarity index, or CI. The index is a ratio of the anomalies (non-seasonal variability) to the seasonal variability of an environmental variable:

$$\mathrm{CI} = 1 - \frac{\sum_{i=1}^{n}\left(x_i - \tilde{x}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \tag{4.21}$$

where $\tilde{x}$ is the long-term monthly average of the variable x (sometimes called the *climatology*) and $\bar{x}$ is the mean of x.

CI is dimensionless; a value of CI = 1 indicates perfect cyclo-stationarity (all variation is due to seasonal behavior). Conversely CI $\approx$ 0 indicates that non-seasonal behavior is dominant. J. Zhang (2019) used CI as an indicator of seasonality in the terrestrial water storage observations from GRACE.

In summary, a number of qualitative and quantitative measures can be used by researchers to evaluate and compare the performance of predictive models. Plots of observed and predicted (e.g.: time series plots and scatter plots) are the first and most important check on a model. Care must be used in interpreting model fit statistics; those which are dependent on units of observations and sample size (e.g., sum of squared errors) cannot be readily compared from one data set to the next. Even using a normalized quantity such as the coefficient of determination (R^2) may obscure bias in the model. Finally, the Nash-Sutcliffe model efficiency (NSE) was shown to be useful in comparing observed and modeled time series. Recent modifications to NSE make it an even more flexible, for example when averaging values across multiple cites, or where there is a strong seasonal signal among observations.

4.2 Model Selection and the Bias-Variance Tradeoff

In this section, I wish to give a brief introduction to an important subject in the field of statistical modeling and machine learning. For a more detailed introduction to the so-called bias-variance tradeoff, I refer the reader to the longer discussion in G. James et al. (2013, Section 2.2.2).

With any type of model, whether it is a simple linear regression or a complex machine learning model, we are interested in creating a tool or method for making predictions with new data. For example, suppose we are constructing a flood model to predict flooding given inputs such as temperature and rainfall. We would train or calibrate the model with historical data, seeking a good fit between model predictions and past observations of flooding. But our main interest is not in recreating the historical record, but in predicting future floods. Therefore, the usefulness of the model should be judged based on the accuracy of its predictions when fed with data it has not seen before, and which were not part of its training data. A model that provides a good fit to training data, but performs poorly with new data is said to be *overfit*. This is often the case with models that are overly complex or over-parameterized, or where the training dataset is small.

We can demonstrate this concept with an example. Figure 4.6 shows a simulation I created to demonstrate this concept. Here, we have a relationship $y = f(x)$ that we will use to generate samples. This demonstration is unrealistic of course. In nature, we don't usually know the true relationship between x and y. (The function, which I created arbitrarily to have an interesting curvy shape is $y = 6 - 0.5x \cdot \sin(x) \cdot e^{x/3}$.) I used this function to generate a training dataset. I sampled values of y at regular intervals on x, adding error to each point with normally distributed random noise. I then fit a polynomial to the training dataset via the method of least squares. The polynomials are of the following form:

$$\text{First order: } f(x) = a_1 x + b \tag{4.22}$$

$$\text{Third order: } f(x) = a_1 x + a_2 x^2 + a_3 x^3 + b \tag{4.23}$$

$$\vdots$$

$$\text{20th order: } f(x) = a_1 x + a_2 x^2 + \ldots + a_{20} x^{20} + b \tag{4.24}$$

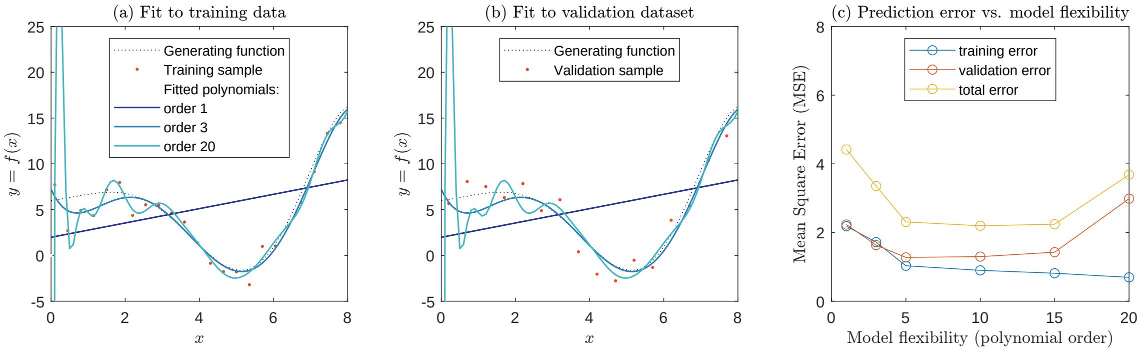


Figure 4.6: Illustration of the bias-variance tradeoff with increasing model flexibility

The polynomials fit to the training data are shown in Figure 4.6(a). In general, as we add more complexity to a model, it becomes more flexible, and can better fit the training data. However, this does not always translate to an improved fit to validation data. In this example, the relationship among the variables is poorly described by a first-order polynomial, which is a simple linear model (Equation 4.22). The mean square error (MSE) is high for both the training data set and the validation data set. A third-order polynomial (Equation 4.23) is able to capture more of the curvilinear relationship between x and y, and has a correspondingly lower MSE, for both the training and validation datasets. As we increase the polynomial order, the curve we fit to the data becomes more flexible. With a twentieth-order polynomial, the curve is very wiggly, and attempts to go through individual data points in the training set. As a result, the training MSE is the lowest of all the curves we tested. However, while the result fits this individual training set very well, it generalizes poorly. Here, the fit polynomial is chasing after the noise, or random errors around the parent relationship. The resulting curve is a poor fit to the validation dataset, and the validation MSE is high. For the higher-order polynomials, the model is overfit to the training data.

In this example, we see that as we increase the flexibility of the model, we better fit the training data. Yet at a certain point, the model overfits the training data, and lacks generality. The evidence for this is that it performs poorly at making predictions with data that were not part of its training. This phenomenon is referred to as the *bias-variance tradeoff*.

The name "bias-variance tradeoff" is somewhat unfortunate, as it reuses two common terms in statistics and modeling in a way that is inconsistent with their use in univariate statistics. Both terms have a specific mathematical definition, described above in Sections 4.1.7 and 4.1.8. In the context of the bias-variance tradeoff, it helps to forget for a moment these mathematical definitions for a moment and think of the plain-language meaning of the words bias and variance.

Here, the *bias* of a model is the error that is introduced by the model's simplified representation of the real-world problem. In this context, bias refers to model error more generally, not the difference between the means (or medians) of observations and predictions. This is perhaps best shown with a simple example. Consider the simple curve-fitting experiment in Figure 4.7. Here, we are trying to model the relationship between the independent variable, x, and a response variable y with a linear regression line. The simple linear model is not able to capture the true (curved) relationship between x and y, and the fit is poor. A machine learning practitioner would say that the model has high bias. Yet, we know from statistics that an ordinary least squares (OLS) regression fit by the method of moments is an unbiased linear estimator of y. Indeed, when the regression line is fit by OLS, the mean of the residuals (e_i values) is exactly zero. Visual evidence for this is shown in the histogram in Figure 4.7(b). Furthermore, the mean of the predictions ($\hat{y}_i$ values) equals the mean of the observed responses (y_i values). So a statistician would conclude that the model has no bias. (However, he or she would also conclude that the model form is incorrect by inspecting a plot of the residuals versus x, as described above in Section 4.1.4.)

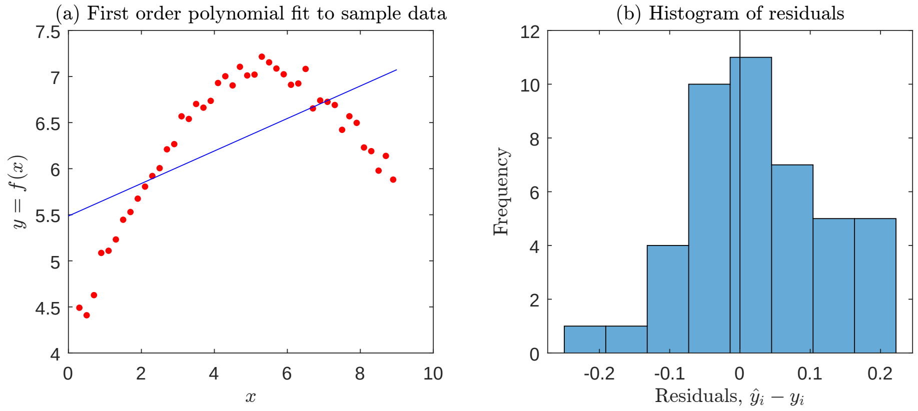


Figure 4.7: Illustration of the competing definitions of "bias" in the domains of statistics and machine learning.

In the context of the bias-variance tradeoff, *variance* refers to how much the model parameters would change if we used a different training data set. In this context, it again helps to think of the plain-language meaning of the word variance. We are not talking about the mathematical definition for the statistic of a random variable, σ^2, described above in Section 4.1.8. Rather, it refers to how much the model fit to data changes each time we use a new sample of observations for the training data set. To demonstrate this concept, I repeated a variant of the experiment above with synthetic data. Again I drew random samples from the generating function, adding random errors to each point. This time I generated five different training datasets. For each training dataset, I fit polynomials of various orders from 1 to 20. The results are shown in Figure 4.8.

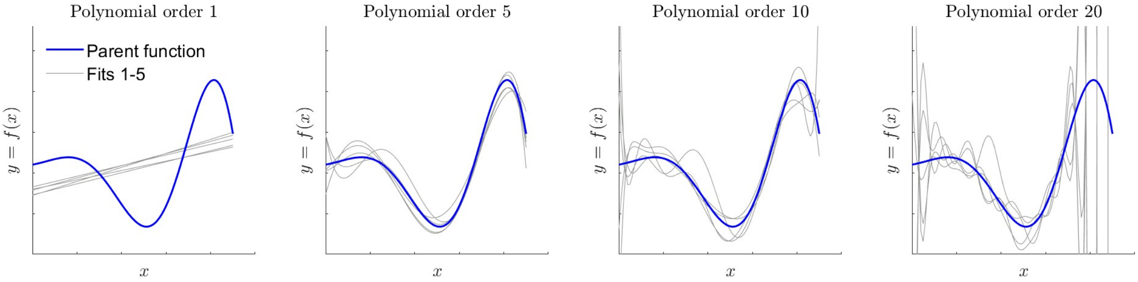


Figure 4.8: Illustration of the variance among models of varying flexibility

The dark line in Figure 4.8 is the true relationship between x and y, which was used to generate training data via random sampling. The light gray lines are the polynomials fit to the five different training sets. Each was fit by the method of least squares. We can see that the fit of the first-order polynomial is relatively stable. In other words, each of the five lines is relatively similar. In this case, the model fit is said to have low variance. As we increase the order of the polynomial, there is greater variety in the shapes of the curves. In essence, the fitting algorithm is trying to find the best fit to each training data set (the training data are not shown in Figure 4.8). At the most extreme, the 20th order polynomial oscillates wildly to try to fit individual data points.[1] Because this polynomial has more parameters, it is much more flexible. Here, a machine learning practitioner would say that there is high variance. This is distinct from saying that the residuals have high variance, or are widely spread about their mean.

The bias-variance tradeoff is a fundamental consideration whenever one is fitting a model to data. The goal of the modeler is to minimize the expected test error. To do so, one must select a modeling method that results in both

[1] When the dataset is small compared to the number of model parameters, the risk of overfitting is especially high. Indeed, an n^{th} order polynomial can pass directly through n points, providing a "perfect" fit.

low variance **and** low bias. This is the central challenge across various fields of statistics and machine learning (G. James et al., 2013). One can easily fit a model with extremely low bias but high variance by fitting a model that passes through every data point. Such a model lacks *generalizability*. On the other hand, one can fit a model with extremely low variance but high bias, for example a model of the form $y = \bar{x}$. However, such a model is not useful as a predictor, thus lacking *utility*.

Hastie et al. (2008) present the problem in a more formal and mathematical way by defining a relationship between an outcome variable Y and independent variable(s) x as:

$$y = f(X) + \epsilon \tag{4.25}$$

where ϵ is the irreducible error which is normally distributed and has a mean of 0. Our job as modelers is to fit a model $\hat{f}(x)$ that best approximates the true relationship $f(x)$, which is unknown to us. The error associated with a given prediction at a point x_0 is:

$$\mathrm{Err}(x_0) = E\left[(y_0 - \hat{f}(x_0))^2\right] \tag{4.26}$$

In general, the error of a model prediction stems from three sources:

$$\begin{aligned} \mathrm{Err}(x) &= \mathrm{Bias}^2 + \text{ Variance + Irreducible Error} \\ \mathrm{Err}(x) &= \left(E[\hat{f}(x)] - f(x)\right)^2 + E\left[\left(\hat{f}(x) - E[\hat{f}(x)]\right)^2\right] + \mathrm{Var}(\epsilon) \end{aligned} \tag{4.27}$$

Equation 4.27 defines the expected test MSE. This is the average test MSE one would get by repeatedly estimated f over many training sets, and testing each function at x_0. The overall expected test MSE is calculated by averaging $E\left[(y_0 - \hat{f}(x_0)^2\right]$ over all the values of x_0 in the test set. In this equation, the squared bias term is the difference between the modeled average and the true mean for x. The variance term is the expected squared deviation of $\hat{f}(x_0)$ from its mean.

To achieve the minimum expected test error, we must select a model that has both low variance and low bias. Variance is always non-negative, and the squared bias term is also non-negative. We can conclude from Equation 4.27 that the expected test MSE cannot ever be less the irreducible error, or $\mathrm{Var}(\epsilon)$.

I also wanted to show that the issues discussed here apply equally to the selection of neural network models. Above, I created a set of sample training

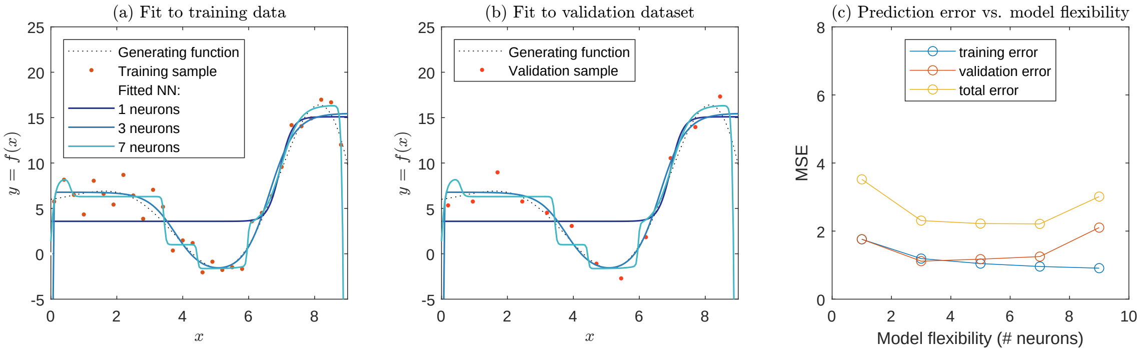


Figure 4.9: Illustration of the bias-variance tradeoff with a simple neural network model.

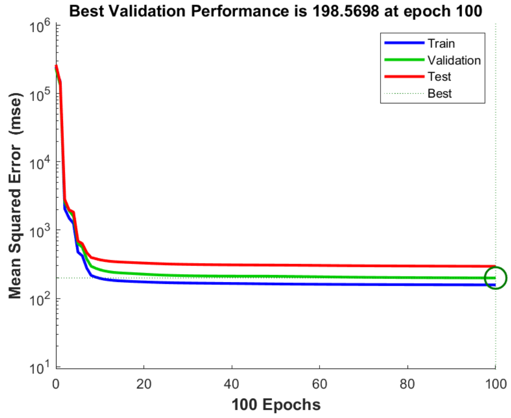


Figure 4.10: Example learning curve created with Matlab's Deep Learning Toolbox.

data and fit a set of polynomials. Here, I repeat the same experiment, using a set of neural network models with different numbers of neurons in a single hidden layer. The results are shown in Figure 4.9. The parent generating function which represents the true relationship between x and y is the same as the one shown in Figure 4.6. There is evidence of overfitting in the network with 9 neurons, as the curve appears to be chasing individual data points in the training set, causing it to oscillate.

It is worth noting that the default settings of contemporary machine learning algorithms provide guardrails against overfitting. In Matlab, for example, the default setting for fitting a feed-forward neural networks is to automatically divide the dataset into three sets, for training, testing, and validation. Users can also provide their own custom partition information. The training algorithm is programmed to stop when the test data set error rate stops going down or begins to increase (example in Figure 4.10. This makes it more difficult for even naïve users to overfit neural network models. Nevertheless, by customizing the settings, one can remove these guardrails for demonstration purposes, such as in the example above.

4.3 Regression Modeling Methods

The first set of analyses to balance the water budget involves a simple class of models to optimize water cycle components. Recall that the goal is to *calibrate* EO variables so they are closer to the OI solution, which results in a balanced water budget. Here, we attempt to create models to optimize each individual EO variable, such that they will be closer to the OI solution. For example, we are looking for an equation to transform GPCP precipitation, P_{GPCP} data so it more closely matches P_{OI}. Essentially, we are seeking a model $P' = f(P_{GPCP})$ where our objective is $P' \to P_{OI}$. Here, the function f is estimated by relatively simple linear models.

In this section, I describe the development of *single* linear regression models, so-called because there is one input variable. In later sections, I investigate the use of more complex models for f that are non-linear and include more input data. The regression models have between one and three coefficients that are fit with a variety of parametric, non-parametric and optimization methods. The next problem is how to generalize these results so that they can be applied anywhere, for example in basins that were not part of the training set, or at the pixel scale, which is discussed in the following section.

4.3.1 Ordinary Least Squares

The first models are linear regression models of the following form:

$$\begin{gathered} y_i = ax_i + b + \epsilon_i \\ \text{for } i = 1 \text{ to } n \end{gathered} \tag{4.28}$$

where:

- y is the ith observation of the response variable,
- x is the ith observation of the explanatory variable,
- a is the is the slope (the change in y with respect to x),
- b is the intercept
- ϵ_i is the random error or residual for the ith observation
- n is the sample size, or number of observations

The parameters (a and b) are fit by the standard ordinary least squares method, which minimizes the squared difference between predicted and observed values. Here we are plotting the best-fit line between one of our input EO variables and the OI solution for the variable. The example in Figure 5.1 shows P_{OI} versus P_{GPCP} over three example river basins. The equation of the regression line tells us the best estimate (according to the linear model) of P_{OI} for a given value of P_{GPCP}. The equation can therefore allow us to calibrate P_{GPCP}, or nudge it in the direction that will be closer to the OI solution.

In our case, we fit an equation for each of 10 variables, and in each of our experimental river basins.

$$P_{GPCP,cal} = a \cdot P_{GPCP} + b \tag{4.29}$$

where $P_{GPCP,cal}$ is the calibrated version of the GPCP precipitation time series, P_{GPCP} in a given basin. The slope and intercept parameters, a and b are estimated independently in each of the 1,698 river basins.

There are certain advantages to a linear regression model. It is simple, explainable, and only requires us to fit two parameters. However, there are also disadvantages. A linear model may not adequately describe the relationship between our dependent variable and response variable. Furthermore, our data may not follow all the assumptions of this method. For example, the prediction errors should be normally distributed, independent, and homoscedastic (have constant variance over x). Nevertheless, linear models may still be useful and appropriate, even when all of these assumptions are not strictly met. It depends on the intended use of the model. See Table 4.2 for a list of assumptions which should be met based on the intended purpose of an OLS model (reprinted from Helsel et al., 2020, p. 228).

A variant on OLS regression involves forcing the fitted line to pass through the origin at (0, 0). In such case, the regression equation is simplified to a one-parameter model:

$$y_i = ax_i + \epsilon_i \tag{4.30}$$

This follows Equation 4.28 but simply drops the intercept, b. Fitting a model with slope only, and no intercept, is called regression through the origin (RTO). It turns out that RTO is a surprisingly controversial subject among statisticians and scientists (Eisenhauer, 2003). Some authorities state that it is an appropriate model when the dependent variable is necessarily zero when the explanatory variable is zero. For example, suppose we are modeling the height of a tree as a function of its trunk's circumference. A tree with a circumference of zero cannot have a non-zero height. Regardless, it is nonsensical to talk about a tree with zero circumference. Other authorities insist that a linear model without an intercept is

Table 4.2: Assumptions necessary for the purposes to which ordinary least squares (OLS) regression is applied (reprinted from Helsel et al., 2020)

	Purpose			
Assumption	Predict *y* given *x*	Predict *y* and a variance for the prediction	Obtain best linear unbiased estimator of *y*	Test hypotheses, estimate confidence or prediction intervals
Model form is correct: *y* is linearly related to *x*.	X	X	X	X
Data used to fit the model are representative of data of interest.	X	X	X	X
Variance of the residuals is constant (homoscedastic). It does not depend on *x* or on anything else such as time.	-	X	X	X
The residuals are independent of *x*.	-	-	X	X
The residuals are normally distributed.	-	-	-	X

meaningless and inappropriate. Nevertheless, statisticians have shown with rigor that the coefficient of determination, R^2, cannot be calculated with RTO. While alternative formulations for R^2 have been proposed, it is not appropriate to use it to compare the OLS and RTO models. Instead, Eisenhauer (2003) suggests that it is appropriate to compare the standard errors of the OLS and RTO regressions (i.e. MSE or RMSE). I would add that one of the fit metrics commonly used by hydrologists (NSE or KGE, described above) would allow for a fair comparison of a RTO and OLS regression.

4.3.2 Outlier Detection

A well-known drawback with OLS regression is that individual data points can have a disproportionate impact on the regression line. To overcome this limitation, there are two main remedies. First, one can simply remove outliers, and repeat the regression analysis until satisfactory results are obtained. Or, one may use an alternative form of regression that is more resistant to outlier effects, such as the non-parametric regression methods described in the following section. This brief section introduces the concept of leverage, and discusses methods to detect outliers in bivariate datasets.

Outliers which have a large effect on the outcome of a regression are said to have high *leverage*. The leverage of the i_{th} observation in a simple regression is calculated by:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SS_x} \tag{4.31}$$

where SS_x is the sum of squared deviations for x, or $\sum_i (x_i - \bar{x})^2$. From this definition of h, we can deduce that the further an observation is from the mean, the greater its leverage will be. The leverage of a data point, h_i, will always have a value from $1/n$ to 1, and the average leverage for all observations is always equal to $(p+1)/n$.

Observations are often considered to have high leverage when $h_i > \frac{3p}{n}$, where p is the number of coefficients in the regression model. (Some statisticians prefer a lower value of $2p/n$.) In the case of single-variable regression, $p = 2$, as we are estimating two coefficients, the slope and intercept. An observation with high leverage will exert a strong influence on the regression slope. According to an influential text by USGS statisticians Helsel et al. (2020), “observations with high leverage should be examined for errors,” however, they note that an observation with high leverage is not a sufficient reason to remove the observation from the analysis.

Another widely used measure of influence is Cook's D (Helsel et al., 2020, p. 241):

$$D_i = \frac{e_i^2 h_i}{ps^2 \left(1 - h_i\right)^2} = \frac{e_{(\mathrm{i})}^2 h_i}{ps^2} \tag{4.32}$$

where:

- h_i is the leverage of observation i,
- e_i is the residual,
- p is the number of estimated parameters in the model (for SLR, $p = 2$), and
- s^2 is the variance of the residuals

Cook's D is used to determine the degree of influence of an observation by comparing its value to the critical value in an F-distribution for $p + 1$ and $n - p$ degrees of freedom. In our case, single linear regression with $n > 30$, for a two-tailed test and a 10% statistical significance ($\alpha = 0.1$), the critical value of $D \approx 2.4$.

Remark

I tried 4 different methods to detect outliers: the two described above, plus *bivariate leverage* and a method called DDFITS. I was not sure which was best, so this was an interesting and valuable experiment. The first method (leverage) produced good results, so I decided to use that. Two of the other methods flagged many more outliers, and I concluded that it would result in omitting too much of the data.

4.3.3 Non-parametric Regression

Non-parametric regression techniques are useful where the data do not follow the assumptions required for ordinary, parametric regression methods described above. In general, non-parametric statistics refers to a class of methods that do not make assumptions about the underlying distribution of the data. Among the benefits of these methods are the ability to deal with data with small datasets, outlier, and errors that are not normally distributed. Theil-Sen regression is a robust method for estimating the median of y given x. Compare this to ordinary least squares regression, which estimates the *mean* of y given x. The Theil-Sen line is widely used in water resources and more recently in other disciplines (Helsel et al., 2020).

The Theil-Sen line is estimated as the slope and intercept of the median of y as follows:

$$\hat{y} = \hat{a}x + \hat{b} \tag{4.33}$$

To compute the Theil-Sen slope, $\hat{a}$, one compares each point to all other points in pairwise fashion. For each set of (x, y) points, the slope $\Delta y / \Delta x$ is calculated. Then, the estimate of the slope of the Theil-Sen regression line is the median of all pairwise slopes:

$$\hat{a} = \text{median}\,\frac{(y_j - y_i)}{(x_j - x_i)} \tag{4.34}$$

for all $i < j, i = 1, 2, \ldots, (n-1), j = 2, 3, \ldots, n$.

Several methods for calculating the intercept have been put forth, but the most common is:

$$\hat{b} = y_{med} - \hat{b} \cdot x_{med} \tag{4.35}$$

where x_{med} and y_{med} are the medians of x and y. Further details, including how to compute P-values and confidence intervals for Theil-Sen regression coefficients are given by Helsel et al. (2020).

4.3.4 Variants to Linear Regression

In addition to the 1- and 2-parameter SLR models described above, I also tried an alternate 3-parameter model. This model specifically avoids the problem of predicting negative precipitation or runoff, which occasionally occurs with an SLR model with 2 parameters (slope and intercept). Such a model was used in the context of water cycle studies by Pellet, Aires, Munier, et al. (2019). This model has an exponential term on the intercept that prevents the model from changing values of $x = 0$:

$$y = a \cdot x + b\left(1 - e^{-\frac{x}{c}}\right) \tag{4.36}$$

where the variables a, b, and c are model parameters to be fit. Fitting the model is not done with the ordinary least squares method as in SLR, but can readily be fit with an optimization algorithm, such as the `fit` function in Matlab.

4.3.5 Transformation of Input Variables

Another common method to improve the performance of a regression model is to apply a transformation to the dependent variable, the response variable, or both. The same methods can and should be applied for neural network models. In their book on neural networks, Bishop (1996) state that it is "nearly always advantageous" to pre-process the input data, applying transformations, before it is input to a network.

Hydrologists commonly log-transform observations to make their distribution less highly skewed and to reduce the number of outliers. Taking $\log(x)$ or $\sqrt{x}$ can also solve the problem of predicting negative values. However, there are a more methodical ways of seeking the best transformation. It is a good statistical practice to transform the independent variable, x, such that the strength of the linear association is maximized. The conventional way to do this is via Tukey's ladder of powers, suggested by 20th century statistician John Tukey, where the linear model is applied to the transformed variable:

$$y = ax^{\theta} + b \tag{4.37}$$

where θ is a power transformation. One may choose any value for θ. Normally, when $\theta = 0$, it results in no transformation, as $x^0 = 1$. Tukey suggested that it is convenient to substitute $\log(x)$ when $\theta = 0$. A list of common transformations is shown in Table 4.3.

One is not limited to using the values for λ in Table 4.3. A related approach for transforming a variable seeks to make its distribution as close to normal as possible. The Box-Cox transformation for a variable x is given:

$$x'_{\lambda} = \frac{x^{\lambda} - 1}{\lambda} \tag{4.38}$$

where λ can take on any value, positive or negative. Equation 4.38 can be interpreted as a scaled version of the Tukey transformation above. Again, where $\lambda = 0$, the transformation is $x'_{\lambda} = \log(x)$. In some cases, the *response* variable y in a regression is also transformed, in addition to the independent variable x. In such cases, care must be exercised, because one cannot compare the variances as λ varies.

A practical complication arrives when the data to be transformed has values ≤ 0, as the result of the transformation is undefined for some values of λ. A common practice is to add a constant to x such that $x > 0$ for all values x. To facilitate these calculations, I wrote a simple Matlab function to apply a shift δ, to

Table 4.3: Ladder of powers, from Helsel & Hirsch (2020)

θ	**Transformation**	**Name**	**Comment**
Used for negatively skewed distributions			
i	x^i	i^{th} power	-
3	x^3	Cube	-
2	x^2	Square	-
Original units			
1	x	Original units	No transformation.
Used for positively skewed distributions			
$1/2$	$\sqrt{x}$	Square root	Commonly used.
$1/3$	$\sqrt[3]{x}$	Cube root	Commonly used. Approximates a gamma distribution.
0	$\log(x)$	Logarithm	Very commonly used. Holds the place of x^0
$-1/2$	$-1/\sqrt{x}$	Negative square root	The minus sign preserves the order of observations.
-1	$-1/x$	Negative reciprocal	-
-2	$-1/x^2$	Negative squared reciprocal	-
$-i$	$-1/x^i$	Negative i^{th} reciprocal	-

a vector, and to return the shifted vector (containing only positive values), and the value of the shift, δ applied. I found it convenient to make δ have a minimum of 1, even when all values of $x > 0$.

4.3.6 Spatial Interpolation

In the sections above, I describe several statistical methods to calibrate water cycle observations so that the more closely match the optimal interpolation solution and thus balance the water budget. These methods are all applied over river basins, where we have been able to apply OI. Now the challenge is how to extrapolate these results to other locations, such as ungaged basins or even individual grid cells.

A common approach in the sciences is to fit a surface to geographic point data in order to estimate or predict values at non-sampled locations. This is usually referred to as "spatial interpolation." When the locations are outside the range of the original data, it is referred to as frequently "extrapolation." Such methods are widely used in the geosciences to create continuous gridded data from point observations (for example rain gages, evaporation pans, or wind speed measurement devices). Spatial interpolation can be considered a quantitative application of Tobler's laws of geography, which states that "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970).

One potential complication is that the river basins are polygons. I calculated the centroid of the river basins, equivalent to their center of mass. In this way, the river basins can be represented as points by their x and y coordinates, or their longitude and latitude.

In the hydrologic sciences, spatial interpolation is used for "parameter regionalization." This approach allows the analyst to transfer parameters locations where models have been fit to new locations. It is used for estimating flood frequency (England et al., 2019), low-flow quantiles (Schreiber & Demuth, 1997), and hydrologic model parameters (L. D. James, 1972; Abdulla & Lettenmaier, 1997; Garambois et al., 2015). One approach for transferring information from gauged to ungauged basins involves identifying the relationship between model parameters and catchment characteristics (Bárdossy & Singh, 2011). The most common approach is to fit a multivariate linear regression regression model. Some recent research suggests that linear models may not be the best approach to finding complex and nonlinear relationships between model parameters and watershed properties. Song et al. (2022) demonstrated the effectiveness of a machine learning approach, using a gradient boosting machine (GBM) model been used to characterize the relationship between rainfall-runoff model parameters and soil and terrain attributes.

In the field of large-sample hydrology, a recent example of parameter regionalization is provided by Beck et al. (2020). In this study, the authors calibrated the HBV hydrologic model on a set of 4,000+ small headwater catchments. They then developed a set of transfer equations relating the parameters to a suite of 8 environmental and climate variables. This allowed the authors to create a map at 0.05° resolution of parameters for the HBV model, which improved predictions in catchments over which the model had not been calibrated.

There are a number of available surface fitting algorithms available directly in Matlab. These include:

- **Nearest Neighbor** - in this method, values are assigned to match that of the closest observation. The result resembles a Voronoi diagram or Thiessen polygons, which are widely used in the hydrologic sciences for spatial interpolation of point data. The resulting surface is discontinuous.
- **Natural Neighbor** - Similar to the nearest neighbor, but the resulting surface is C^1 continuous except at sample points.
- **Linear** - This method fits a different linear between sets of three points. Surface is C^0 continuous.
- **Cubic Spline** - Fits a cubic spline between sets of three points. Surface is C^1 continuous.
- **Biharmonic** - Belongs to a family of polyharmonic spline fitting algorithms. The surface is a linear combination of Green functions, and is C^2 continuous.

Inverse distance weighting is a commonly used method for spatial interpolation for observations collected at specific points. This method assigns values to points by taking a weighted average of its neighbors, where the weights are smaller the greater the distance. The formula for Inverse Distance Weighting (IDW) in two dimensions can be represented as follows:

$$Z(x,y) = \frac{\sum_{i=1}^{n} \frac{Z_i}{d_i^p}}{\sum_{i=1}^{n} \frac{1}{d_i^p}} \tag{4.39}$$

where:

- $Z(x,y)$ is the estimated value at the target location (x,y).
- Z_i is the known value at observed data point (x_i, y_i).

- d_i is the Euclidean distance between the target and observation points, given by $\sqrt{(x - x_i)^2 + (y - y_i)^2}$
- (x, y) are the coordinates of the target location.
- (x_i, y_i) are the coordinates of the known data point i.
- p is the positive exponent that determines the influence of distance on the weighting.

The IDW method is relatively simple and intuitive, but has certain disadvantages. Points too far away may have disproportionate influence. Some implementations limit the number of points contributing information to those within a certain radius. Also, depending on the choice of the value for p, the method may either smooth out or overemphasize small-scale variations more than desired. I used a Matlab function for inverse-distance weighting from a user contribution to the Matlab File Exchange (Fatichi, 2023). The code is from a reputable author and works as intended.

Kriging is a more complex method for spatial interpolation that takes into account not only distances between points but also the spatial correlation or covariance structure of the variable being interpolated. The method assumes that the values of a variable at nearby locations are more correlated than those at distant locations, and this correlation can be modeled using a variogram. A variogram is a plot of the semivariance versus point distance, where the semivariance is half of the average squared difference between the values of a variable at pairs of locations separated by a specific distance. The semivariance is given as:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} (z(x_i) - z(x_i + h))^2 \tag{4.40}$$

where:

- $\gamma(h)$ is the semivariance at lag distance h.
- $N(h)$ is the number of pairs of data points separated by a distance h.
- $z(x_i)$ is the value of the variable at location x_i.
- $z(xi + h)$ is the value of the variable at a location h units away from x_i.

There are several functions for kriging in the Matlab File Exchange, but none of them worked with my version of the software or my datasets. Therefore, I used a kriging function available in QGIS (Conrad, 2008). The function is actually a part of the SAGA GIS program, but its functions are available via the QGIS Processing Toolbox. Because I wanted to interpolate many surfaces (for several variables, and for k-fold cross-validation, discussed below), I used Python scripting to run the kriging operation in batch mode. It is worth noting that this toolbox has many parameters, and it appears that choosing a good set of parameters to get reasonable results is as much an art as a science. Nevertheless, there is little available documentation, and the online manual contains very few details or explanation. It seems that, based on some online research, many analysts prefer to use the GSTAT library (E. J. Pebesma & Wesseling, 1998) to perform kriging and related analyses. GSTAT was formerly standalone software for spatio-temporal analysis, which has since been ported to R and Python (Mälicke et al., 2021; E. Pebesma & Graeler, 2023). I would recommend to other analysts to give preference to using GSTAT for kriging, as there is a detailed user manual, and a variety of introductions and tutorials are available online.

4.3.7 Resampling and Cross Validation

Cross-validation is a technique used in machine learning and model evaluation. It belongs to a class of resampling methods, referred to by statisticians as “an indispensable tool in modern statistics” (G. James et al., 2013). The procedure involves repeatedly extracting samples from a training set. After each sample is drawn, a model of interest is retrained using that sample. The purpose is to learn about how well the model fits the training data. Such methods are especially valuable with small datasets, where the fit may be highly influenced by which observations are used to train the model. Resampling methods are computationally intensive, but they have become more common as computers have gotten faster. For a readable introduction to the concepts here, the reader is referred to G. James et al. (2013) or for a slightly more detailed treatment, Hastie et al. (2008).

In k-fold cross-validation, the dataset is divided into k subsets of approximately equal size. The training and testing process is then repeated ’k’ times, each time using a different subset as the testing data and the remaining subsets as the training data. It is a common practice to calculate statistics on the model residuals, such as the standard deviation, based on the set of results. This allows the modeler to partially quantify the variability in the skill of the model (although it does not take into account all sources of error).

I used a form of k-fold cross-validation is used to evaluate the accuracy of different spatial interpolation methods. I divided the set of synthetic basins into 20 different partitions. For each partition, 80% of the basins were randomly assigned to the training set, with the remaining 20% of basins assigned to the training dataset. For each of the 20 datasets, I fit a regression model to each of the 10 EO variables to estimate it’s OI-optimized value. For example, the regression model estimates P_{OI} as a function of P_{GPCP}. I created surfaces based on the regression parameters in order to estimate these parameters in non-sampled locations. (In this case, there are two regression parameters, slope and intercept, a and b. While I also experimented with 1- and 3-parameter regression models, I did not perform cross-validation for these models.)

Next, I used this surface to interpolate the parameter values for the training basins. The interpolation is done by looking up the value a point representing the latitude and longitude of the basin centroid. Then I estimate the accuracy of the interpolation, or the goodness of fit by comparing the interpolated parameter to its actual value. All in all, this method allows us to determine which spatial interpolation performs the best. It also allows us to estimate the bias and variance in estimates of the regression parameters.

4.4 Neural Network Modeling

In this section, I describe a flexible modeling framework based on neural networks. Similar to the regression-based models described above, the goal is to *calibrate* EO variables to produce a balanced water budget at the global scale. I give a brief introduction to neural networks (NNs), and describe the particular class of of NN model, feedforward networks, which was the focus of the modeling done here. In the final section go on to describe the particulars of the models I created for this task.

At a high level, a neural network takes in input data, such as images or text, and processes it through one or more layers of interconnected "neurons." A neuron is a mathematical function that performs a calculation on the input data and returns an output value. Each layer of neurons focuses on a specific aspect of the input data, and the output of each layer is passed on to the next layer for further processing. This hierarchical structure allows the neural network to learn and extract increasingly complex features from the input data.

A strength of NN models is that no knowledge of the physical processes is needed to create them – they are completely data-driven. This lets us apply NN models to many different problems. Another key advantage of neural network models is their extraordinary flexibility – they are able to simulate non-linear behavior and complex interactions among variables. A disadvantage to NNs is that their parameters do not have any clear physical interpretation, unlike in a conventional hydrologic or climate model, where parameters typically represent real-world phenomena (e.g. a model parameter may represent the rate at which water infiltrates into soil, in m/day).

Neural network models can be used for classification tasks (such as identifying the characters in handwriting, or determining whether an image is a cat or a dog). In such cases, the output is generally a number from 0 to 1 indicating the probability, or how confident the NN is in its predictions. For example, an output of 0.94 indicates a 94% probability that an image is of a cat. NN models can also be used to fit a function to data. In this sense, they are comparable to a regression model, albeit more powerful and flexible.

While NNs can be trained for different tasks, such as classification, or predicting the next word in a chat conversation, here we are primarily concerned with NNs as a *function*. An NN as a function makes predictions of a dependent scalar variable based on one or more independent or predictor variables. Neural network models can be set up to predict a single outcome variable. However, a NN model can also be configured to predict multiple variables, a key difference from the regression models discussed above. As such, NNs are an application of multivariate statistics, a branch of statistics that involves simultaneous observation and analysis of more than one outcome variable. By contrast, a multiple regression, or often, multivariate regression model, is *not* a type of multivariate regression, as it predicts a single response variable, $y = f(x_1, x_2, \ldots x_n)$. By contrast, a multivariate statistical model, can simultaneously estimating four output variables with a single model, e.g. $y_1, y_2, \ldots = f(x_1, x_2, \ldots)$. In our case example, we are interested in creating a prediction model for the four variables that are the main components of the water cycle: P, E, ΔS, and R.

I chose to use a particular type of neural network, feed-forward neural network, appropriate for modeling a quantitative response (Bishop, 1996). In older literature, this has been referred to as a Multi-Layered Perceptron (MLP) (Rumelhart et al., 1987).

Over the next page or two, I describe the fundamentals of how a neural network works. As has been noted by Hagan and Demuth (2002), the vocabulary and notation for describing neural networks varies a great deal in the literature. The authors suppose the reason is that papers and books come from many different disciplines – engineering, physics, psychology and mathematics – and authors use terminology particular to their domain. They note that "many books and papers in this field are difficult to read, and concepts are made to seem more complex than they actually are." The mathematics behind neural networks are actually not that complicated. Rather, it is the size and scope of modern NNs that make them complex.

An analyst can treat a neural network as a black box, a sort of Swiss Army knife for solving all sorts of problems in science and engineering. Indeed, with modern software and free tutorials available online, this is easier than ever. But I believe there is a benefit to really understanding how a network works, as this helps to better grasp its capabilities and limitations. For a short, readable history of the development of NNs, I refer the reader to the introduction to Chapter 10 in G. James et al. (2013). For those interested in a step-by-step introduction to NNs, including many coding exercises, the text by Kim (2017) is a valuable resource. Finally, a recent article in the journal *Environmental Modeling and Software* provides good context for the use of neural networks in the environmental sciences (Maier et al., 2023). The authors seek to dispel some of the myths around neural networks, and defend their use for prediction and forecasting.

The heart of a NN model is the *neuron*, sometimes called a *node*, *hidden unit* or a *perceptron*. Figure 4.11 shows a single neuron, and the mathematical functions it uses to transform the input variables. Overall, the neuron takes a set of *m* inputs, and produces an output, *a*, referred to as its *activation*.

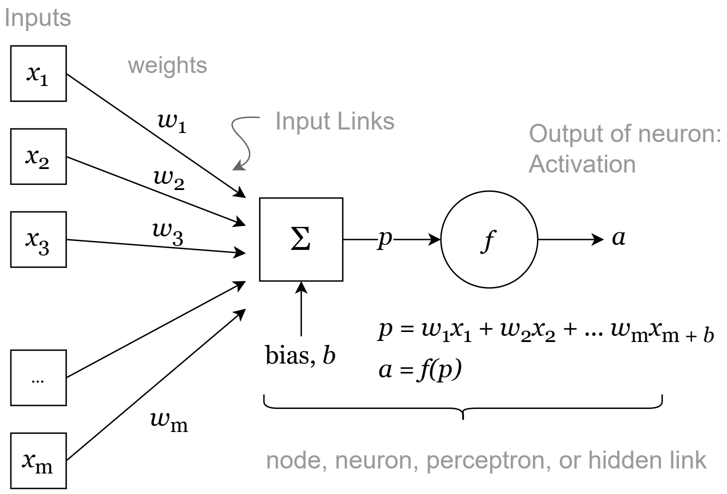

Inputs
x_1
x_2
x_3
...
x_m
weights
w_1
w_2
w_3
w_m
Input Links
Σ
p
f
a
Output of neuron: Activation
bias, b
$p = w_1x_1 + w_2x_2 + \dots w_mx_m + b$
$a = f(p)$
node, neuron, perceptron, or hidden link

Figure 4.11: Overview of a single neuron with multiple inputs

Each of the variables x from 1 to m are scalars real numbers ($x \in \mathbb{R}$). The neuron has a set of weights, w. The number of weights corresponds to the number of inputs. The values for these weights are assigned during training. The weights are used to calculate a linear combination of the inputs, essentially a weighted average of the inputs. The output of this average then has an offset added to it, referred to as a bias, b. The offset may be a positive or negative real number. The result of this calculation is p, a linear combination of the inputs plus a bias:

$$p = x_1w_1 + x_2w_2 + \ldots + x_mwx_m + b \tag{4.41}$$
$$= \sum_{i=1}^{m} x_iw_i + b \tag{4.42}$$
$$= \mathbf{xw} + b \tag{4.43}$$

In Equation 4.43, we suppose that the inputs x_1 to x_m have been combined into a column vector, and the weights w_1 to w_m are a row vector. Equation 4.43 is equivalent to:

$$p = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \cdot \begin{bmatrix} w_1 & w_2 & \cdots & w_m \end{bmatrix} + b \tag{4.44}$$

The value p, which is a linear combination of the inputs plus a constant, is then fed to the activation function, $g(p)$. The output of the neuron, sometimes called its "activation," is given as:

$$a = f(p) \tag{4.45}$$

A variety of different activation functions g can be used. Previously, sigmoid-shaped activation functions were favored (G. James et al., 2013), such as the logarithmic sigmoid:

$$g(z) = \frac{e^p}{1 + e^p} = \frac{1}{1 - e^{-p}} \tag{4.46}$$

The logarithmic sigmoid takes inputs from $-\inf$ to $+\inf$ and converts them to values between 0 and +1. Another common sigmoid function is the hyperbolic tangent function, *tanh*:

$$g(z) = \frac{e^p - e^{-p}}{e^p + e^{-p}} \tag{4.47}$$

Today, many deep learning practitioners favor the rectified linear unit (ReLU) activation function. Because it is fast to compute and easy to store, it is well-suited to training very large NN models. It is calculated as follows:

$$g(z) = (z)_+ = \begin{cases} 0 & \text{if } z < 0 \\ z & \text{otherwise} \end{cases} \tag{4.48}$$

Some other common activation functions available in Matlab are shown in

Figure 4.12.

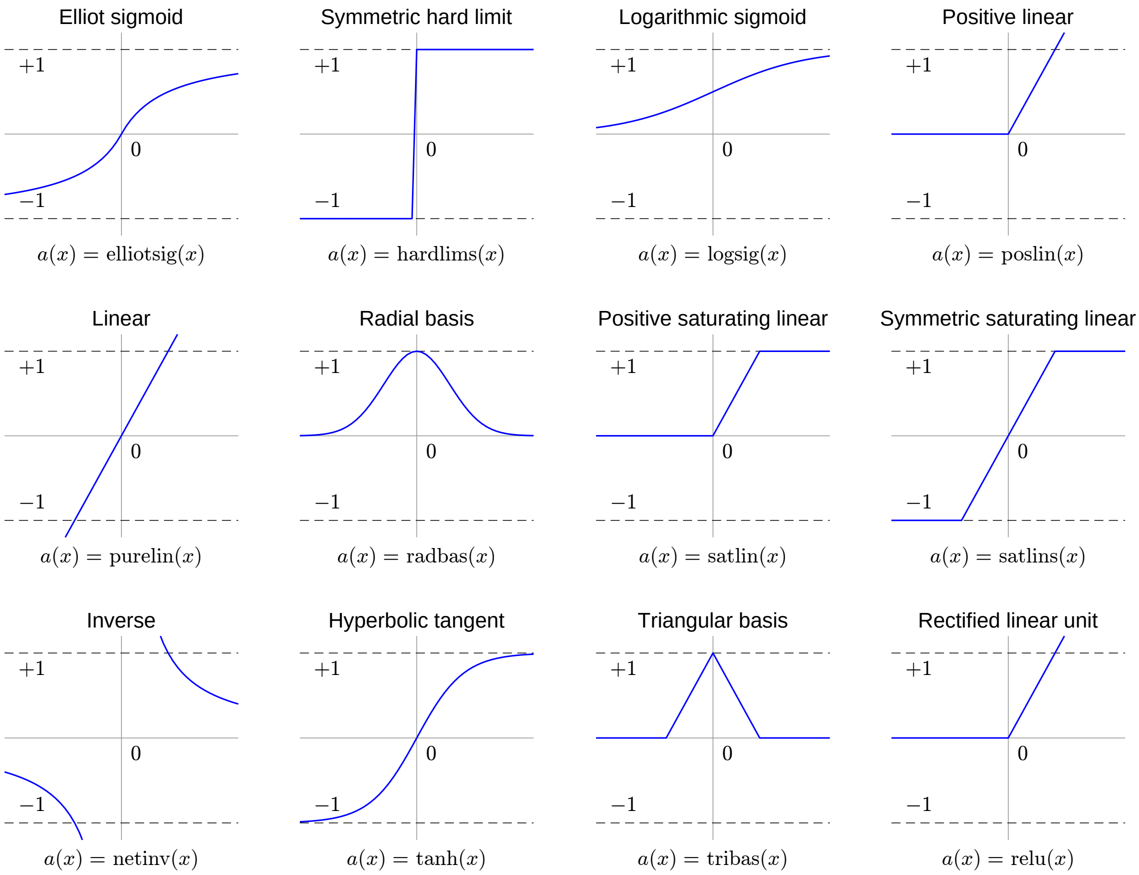


Figure 4.12: A sampling of the neural network activation functions available in Matlab

Typically, a single neuron is not enough to model most relationships. Thus, neurons are made to operate in parallel, in a "layer." In NN jargon, the inputs and outputs are also layers, while layers of neurons that are internal to the model are called *hidden layers*. A network with a single hidden layer of neurons is shown in Figure 4.13. Like the previous diagram, we have a set of inputs x. Here, we are using n neurons to transform the inputs, and we have n activation ouputs, $a_1 \dots a_n$. Each neuron has a weight for its connection to each input, thus we have a total of $m \times n$ weights. (This example would be called *fully connected*, as every input is connected to every neuron.) Each neuron also has its own bias, b. For convenience, we can assemble these into the rectangular $m \times n$ matrix W. The output of this hidden layer is n activations, each of which is a scalar value.

The final portion of the NN architecture takes the activations output by the hidden layer and converts it to the output or prediction. The output of the NN comes from the *output layer*. In a simple case, we have a single target, and the NN model is predicting a single output variable, y. Similar to the way we treated the input layer, the output is a linear combination of the activations, plus a bias term.

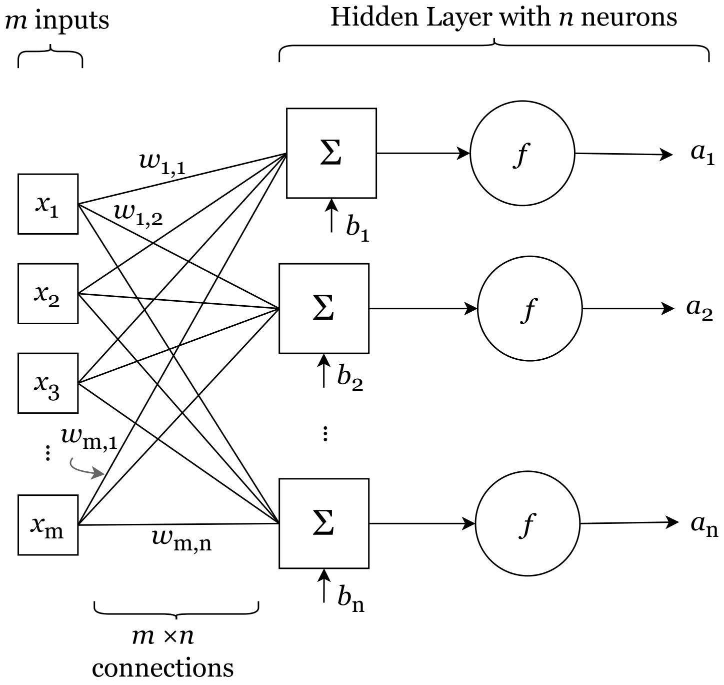


Figure 4.13: Example neural network layer with multiple neurons

So the output layer can be expressed as follows:

$$y = a_1h_1 + a_2h_2 + \ldots + a_nh_n + \beta \tag{4.49}$$

Where h is the output layer weight and β is the output layer bias. The output layer does not use a non-linear activation function, unlike the hidden layer. The output is not normally scaled between 0 and 1, but should be a real number in the range of target variable that we are trying to predict. So the final model is linear in the derived variables a that are output by the hidden layer. It is often the case where we wish for our NN model to output more than one variable. In such a case, the output layer should contain a node for each desired output. Figure 4.14 shows such a network.

Note that there it is not necessary for there to be any correspondence between the number of inputs, the number of neurons, and the number of outputs. For example, we may construct a model with many inputs and a single output. For example, suppose the input is a square image, with 256×256 pixels. If we use each pixel as an input to an NN model, we have $256 \times 256 = 65,536$ inputs. In this case, we have 1 input layer, which contains 65,536 input variables.

Suppose the purpose of the model is to predict whether the image is of a cat. In this case, we can create a model with out output layer, containing one output variable. Typically, the output is a value from 0 to 1, indicating the probability that the image is a cat (according to the model). Now suppose we create a new version of the model to guess between cat, dog, and horse. In this case, the model will still have a single output layer, with 3 outputs: $P(\text{cat})$, $P(\text{dog})$, and $P(\text{horse})$. So we see that the number of inputs and outputs is determined by the problem statement and the available training data.

How does one decide how many hidden layers to include, and the size (number of neurons) in each? There is no formula for determining the best network size and architecture. Indeed, this is an active area of research. The best number of hidden layers and neurons could be determined by trial and error, by reviewing the relevant literature, or by consulting with an experienced NN modeler. Indeed, certain patterns have emerged, where it seems that certain model configurations are well-suited to different problem domains, such as image and video classification, speech and text modeling, and so on

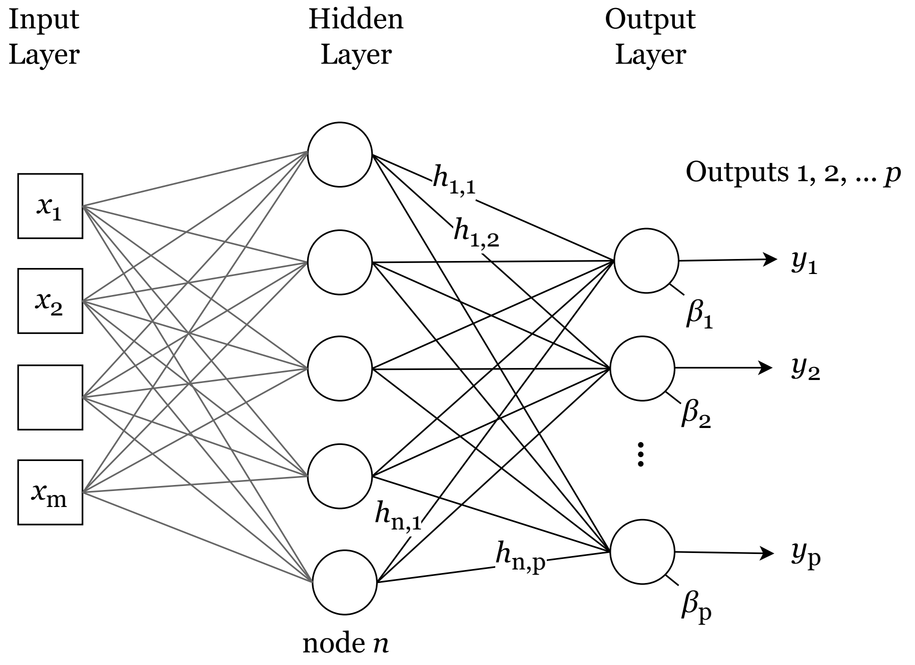


Figure 4.14: Example of a neural network with multiple output variables.

In the example in Figure 4.14, there are p outputs, so we have p output nodes. In general, NN architecture is fully connected, meaning that every neuron in a layer has a connection with every neuron in the preceding layer. Therefore, we have $n \times p$ weights connecting the hidden layer to the output layer. Further, each node in the output layer has a bias, β.

These weights and biases are adjusted during the learning process, so that the network can learn to recognize patterns in the input data and make accurate predictions. The original developers of neural networks were trying to simulate the way neurons in an organism function, with the nodes conceptually similar to nerve cells and the connections analogous to synapses, the connections between nerve cells. When the activation function $a = g(x)$ approaches one, they are *firing*, while activations near zero are *silent* or inactive.

As we add more neurons to the hidden layer, we can simulate more complex behavior, including non-linear responses and interactions. In theory, one can recreate any continuous function given enough neurons and sufficient training (G. James et al., 2013). A pair of landmark papers demonstrated and proved this to be the case. Cybenko (1989) showed that a linear combinations of sigmoidal (S-shaped) can approximate any continuous function of n variables. This paper, "Approximation by superpositions of a sigmoidal function," specifically used continuous and bounded sigmoid activation functions. In a related paper by Hornik et al. (1989), "Multilayer feedforward networks are universal approximators," the authors show proof that one can approximate any continuous function with arbitrary precision using a single-layer neural network. This makes NNs a class of "universal approximators." For functions that are discontinuous (not defined for all x), more than one layer is required to approximate. In practice, it has been found that, to simulate complex relationships, it is usually more efficient to add additional hidden layers to the NN. The resulting models tend to have fewer overall parameters and take less time to train, according to Hagan and Demuth, 2002

According to Hagan and Demuth (2002), it took around 30 years between the invention of neural networks and the addition of an extra hidden layer to the single-layer neural network. Practically speaking, this had to do with the difficulty in training such networks – "a proper learning rule for the multi-layer neural network took quite some time to develop." This problem was solved with the popularization of the back-propagation algorithm. A paper published in *Nature* in 1986 experimentally demonstrated the usefulness of the back-propagation method for training neural networks (Rumelhart et al., 1986). The back-propagation algorithm provided a systematic way to adjust the weights of hidden nodes to reduce the model error. However, the ideas behind back-propagation had been developed earlier. Schmidhuber (2015) provides a readable account of the history of neural networks, tracing the invention of back-propagation to a Finnish Master's student Seppo Linnainmaa in 1970, whose FORTRAN code implements backpropagation, although not for an NN. Important contributions related to backpropagation in NNs were made in the early 1980s by Paul J. Werbos, D.B. Parker, and Yann LeCun.

Another important consideration is the optimization method that seeks to minimize the errors by modifying model parameters. These methods are based on the concepts of gradient descent, which have an esteemed history in mathematics, being first described by Cauchy (1847), and applied to systems of Euler-LaGrange equations in the Calculus of Variations (Fraser, 2005). For this research, I used Matlab's implementation of the back-propagation with the Levenberg-Marquardt optimization algorithm. The algorithm dynamically adjusts a damping factor to balance between gradient descent (steepest descent) and the Gauss-Newton method (approximating the Hessian matrix), allowing it to efficiently converge to a local minimum of the objective function while avoiding convergence issues such as divergence or slow convergence (Hagan & Demuth, 2002, p. 12-19).

Hagan and Demuth (2002) state that a two-layer network with a sigmoid first layer and a linear second layer can be trained to approximate most functions. Contemporary transformer-based architectures, used for language processing can have many more hidden layers. It is not uncommon to see networks with dozens, hundreds, or even thousands of layers in the most advanced and specialized architectures.

Fitting a neural network requires estimating the weights and biases in the hidden layers and output layer in order to minimize prediction errors, which we calculate with a given *loss function*. When the model is estimating a quantitative response (rather than categorical, as in classification problems), analysts typically choose a squared-error loss function, i.e. the sum of squared errors (Equation 4.4). The parameters are chosen to minimize the total overall error through the process of *training*.

What about model architecture? How does one decide the number of hidden layers, how many neurons to place in each, and what type of activation function to use? These are sometimes called a model's *hyperparameters*. Hyperparameters describe the overall structure of the model, and are chosen by the analyst. Deciding on a model's structure, and how it will be trained, are decisions that must be made in the planning stages. By contrast, the model's parameters are the weights and biases associated with the hidden layers, which are determined automatically during the training. Indeed, these connections are made by the computer using automated methods, making neural networks an example of *machine learning*. Choosing the best type and size of model for a particular problem is the subject of great deal of ongoing research.

As G. James et al. (2013) note, careful tuning of networks can often result in performance improvements, "but the tinkering process can be tedious, and can result in overfitting if done carelessly.” Indeed, we must be constantly aware of the risk of overfitting a model. In Section 4.2, I discussed the bias-variance tradeoff. It is a common that a model can be made to perform well with training data – such a model is said to have low bias. However, the model may perform poorly when asked to make predictions based on data it has not seen during training – this is referred to as having a high variance.

Certain aspects of the NN architecture will be determined by the specification of the problem. We set the number of inputs based on the number of input variables we wish to include in our model. The number of outputs, and hence the number of nodes in the input layer, is based on how many variables we are trying to predict. For example, the NN models in this thesis usually have four outputs representing the four major fluxes of the hydrologic cycle: precipitation, evapotranspiration, runoff, and total water storage change (P, E, R, ΔS).

There are many variations on the fairly simple NN models I have described above. These include recurrent neural network (RNN) models, useful for modeling time series and other datasets with auto-correlation. One type of RNN, the Long short-term memory (LSTM) has recently shown good performance in modeling runoff, outperforming conventional rainfall-runoff models (Kratzert et al., 2019). Convolutional neural networks (CNNs), which have recently enjoyed great success in image classification tasks.

A variety of methods have been developed to prevent overfitting of NN models. One method involves selectively dropping certain network connections during training or “pruning” the network. Another strategy involves early stopping of the training process. Typically, the training algorithm continues iterating over an unlimited number of “epochs” until a given criteria is met, i.e. the gradient of the error function is less than a threshold value. Another approach involves limiting the training to a fixed number of epochs. Constraining the number of training epochs can be roughly equivalent to choosing a less complex model (Heberger, 2012).

Training algorithms use random seeds to initialize a set of weights and biases for the network. In some circumstances, training may not produce an optimal outcome. This is due to the possibility of reaching a local minimum of the performance surface. Hagan and Demuth (2002) recommend restarting the training at several different initial conditions and selecting the network that produces the best performance. the authors point to research showing that five to ten restarts will “almost always produce a global optimum.”

Above, I described how machine learning practitioners divide or “partition” training data into sets for training, testing, and validation. Resampling methods like k-fold validation allow us to create multiple models based on different subsets of our data. The purpose of this is to better understand how the model performs with new, unseen data. Typically, after such experiments, the results are saved and the models themselves are discarded. Another approach involves performing multiple training runs and making use of all of the trained networks, for example by averaing the result. This has been referred to as a "committee of networks." (Hagan & Demuth, 2002) write that "the performance of the committee will usually be better than even the best of the individual networks." In general, ensemble methods use multiple model predictions, combined in some way (e.g., averaging) to make final predictions. Such methods are widely used in machine learning and are considered among the most effective techniques for improving the performance and robustness of predictive models (Alber et al., 2019).

Common methods for combining information from multiple model runs include *bagging* and *boosting*. Bagging, short for "bootstrap aggregating" involves training multiple instances of the same model on different subsets of the training data (bootstrapped samples) and averaging their predictions (Hastie et al., 2008). Boosting is an iterative technique that was originally designed for classification problems, but can be applied to regression problems as well. With this method, models are trained in sequence, and each subsequent model focuses on examples that models before it had difficulty classifying.

4.4.1 Transformations on Input Data

I analyzed the input data for the NN modeling, and found that most variables were not normally distributed. For one set of experiments, I first normalized the all input variables, using Box-Cox transformations (described in Section 4.3.5) to transform and rescale the input variables. Figure 4.15 shows the empirical probability distribution of the EO variables before and after transformation. In the top set of plots in Figure 4.15(a), we can see that most of the variables have a skewed distribution, with many more low values compared to higher values. This is typical of many environmental observations. For runoff, the most frequent observation, or the mode, is zero. This is because we have many ephemeral rivers in our database, where there is no measurable runoff at certain times of the year. The exception is with ΔS, or total water storage change, which is roughly symmetrical about zero.

The distribution after applying a Box-Cox transformation to the data, is shown in Figure 4.15(b). I have superimposed a theoretical normal PDF over the histogram of values, with the mean and standard deviation of the transformed data. Here, the goal was to make the data more approximately normally distributed, which has been achieved. The runoff dataset appears to be the furthest from normal after transformation. This is because of the large number of zero values that affects the transformation calculations. Nevertheless, the distribution looks similar to a truncated normal distribution, albeit with an anomalously large spike at zero. It was not necessary to rigorously test for normality (however, this would be desirable in a regression analysis). In such case, one can check for normality with a probability plot correlation coefficient significance test (Vogel, 1986), or with the well-known Kolmogorov–Smirnov or Shapiro-Wilkes hypothesis tests.

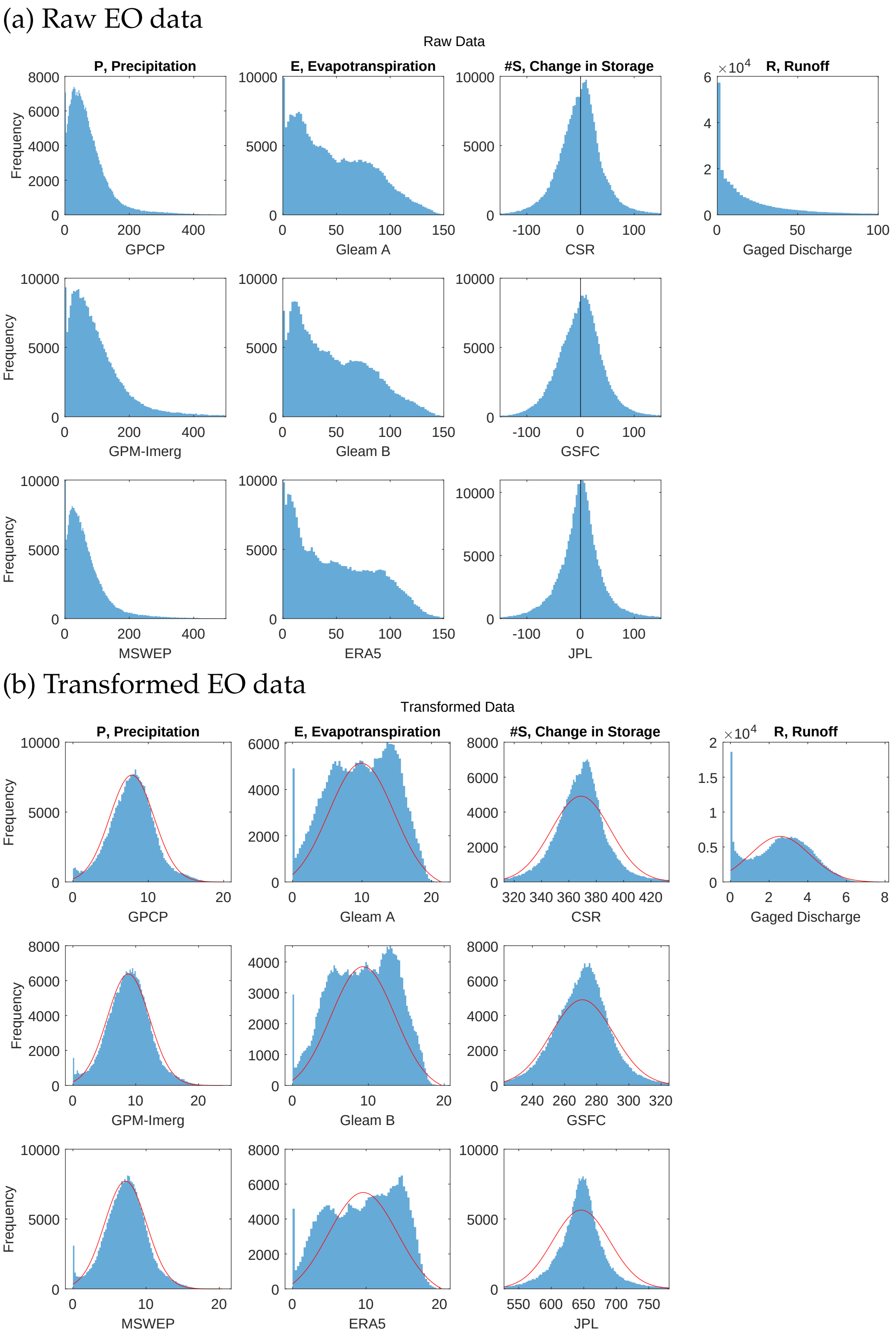
4.4.2 Neural Network Model Architecture

Here I describe the two main sets of NN models that I created and trained to close the water cycle. Both sets of models are at the basin scale. The first set uses runoff data from the 2,056 gaged basins, which are defined based on the drainage areas upstream of a flow measurement gage, as described in Section 2.5.1 on page 63. The second set of NNs use runoff data from 1,398 synthetic river basins. For this analysis, I used a gridded runoff data product, GRUN, and thus I was free to create river basins with outlets at arbitrary locations, as described in Section 3.1.3 on page 95.

Unlike the linear models described above, the NN model is more easily able to integrate ancillary information to constrain the spatial dimension of our overall water cycle database. Previous work has shown the imbalance, or error in the water budget, varies with location and is correlated with environmental indices. Munier and Aires, 2018 showed that the errors are correlated with a vegetation index and an aridity index. This is evidence that feeding ancillary data to our neural network could help improve the accuracy of its predictions.

I experimented with a number of NN architectures. While the one shown in Figure 4.16 is among the simpler models that I tried, it performed the best. On the left are the model inputs, the uncorrected EO datasets, and on the right are the targets, the solution from OI that results in a balanced water budget. I chose a modular architecture with separate calibration and mixture steps that allows us to investigate the outputs of individual layers as we may gain useful information from each:

- First, a set of NNs serves to *calibrate* the individual inputs, or to transform them such that they more closely match the OI solution that satisfies the



(a) Raw EO data

(b) Transformed EO data

Figure 4.15: Empirical probability distribution of EO variables before and after normalization. The first set of plots are in native units of mm/month. The units on the second set of plots are not meaningful. A normal distribution superimposed in red with the mean and standard deviation of the transformed data.

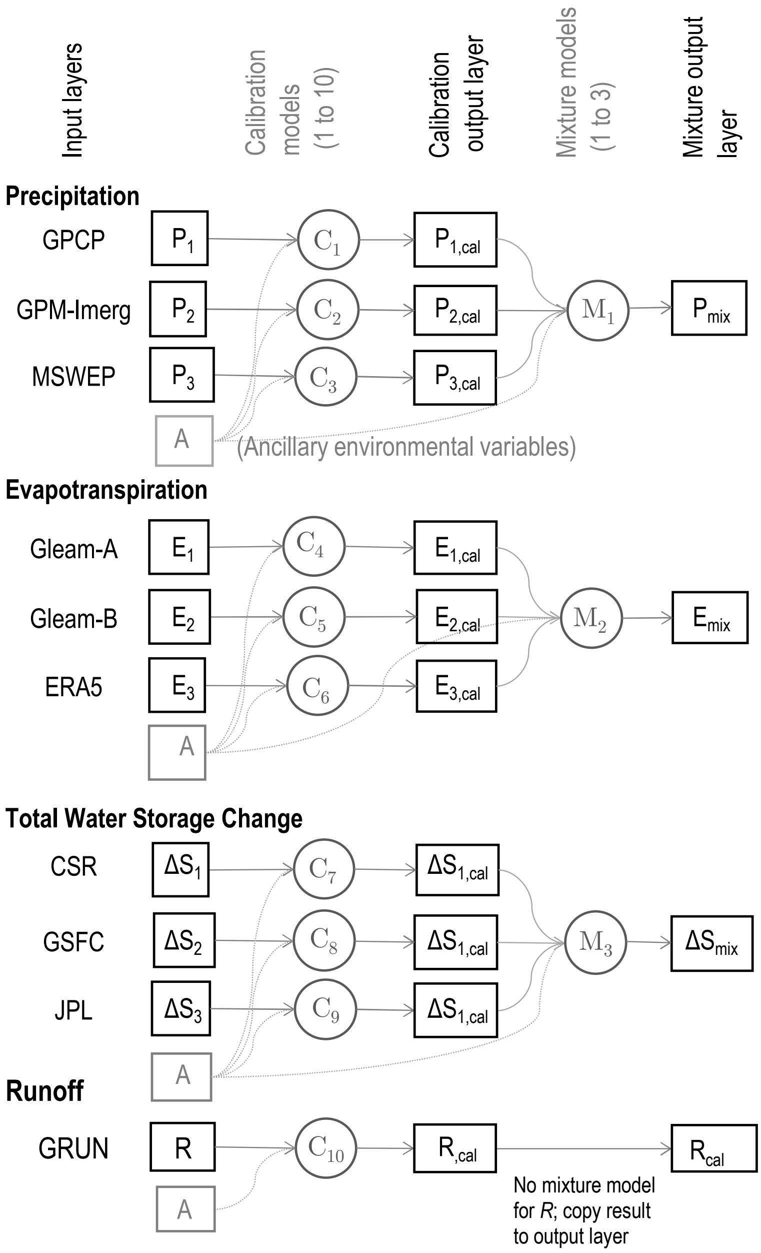


Figure 4.16: Neural network model architecture for calibration then mixture of EO datasets.

water balance constraint. For example, the output of the first calibration sub-model in Figure 4.16, $P_{1,cal}$, is a function of P_1 and the environmental indices (ancillary variables). In this way, each EO product can be optimized independently to each other. This allows running the NN in various configurations with different numbers of input variables (e.g., when one input variable is missing).

- Next, the *mixture* NNs combine information output by the calibration layer to estimate P, E, ΔS, and R. The NN seeks the best compromise among the calibrated EO datasets to fit the target, the OI solution.

A database with paired input and target data is required to train and test the NN model, as well as to select the best model architecture and find the best set of model parameters. For the set of NNs shown in Figure 4.16, each of the 10 calibration networks has 13 inputs (1 EO variable and 12 ancillary environmental variables), 10 neurons in the hidden layer, and 1 neuron in the output layer. The outputs of the calibration layer are calibrated EO datasets, which are useful in their own right, as they should better balance the water budget. Further, they are inputs to the mixture model layers. These layers also have 10 neurons in the hidden layer and 1 neuron in the output layer. For example, the inputs to the precipitation mixture model are calibrated P from each of the three calibration models plus the ancillary variables. Again the target is the OI solution for P calculated previously. In the following section, we evaluate the results of the 10 calibration NNs (1 calibration per EO dataset), and the output of 4 mixture NNs (1 mixture per WC component).

The number of neurons in the hidden layers and the number of hidden layers controls the complexity of the model. I experimented with a range of network sizes and configurations, and found that the fit does not improve with more neurons. Estimation of the optimal parameters of the NN was performed during the training stage using back-propagation and the Levenberg-Marquardt algorithm (Hagan & Demuth, 2002). I trained the model on a set of 1,358 basins and validated the model over a set of 340 independent basins (for an 80/20 split between training and validation). I corrected any physically implausible negative values for P or R by setting them zero. Finally, outputs for ΔS and R were smoothed with a 3-month moving mean filter to remove high-frequency noise from the predictions. I also performed the equivalent smoothing on validation datasets in order to ensure a fair comparison.

The goal of the trained NN model is to optimally combine EO datasets, extracting the "best" information from each one in different environments. The NN model can then be used to make predictions, not only over the basins for which it was trained, but in other river basins for which input data is available. I validated the model by examining errors over basins that were not part of the training data, and where river discharge measurements are available. (In this case, there were 1,358 training basins, and 340 validation basins.)

I trained the NN model with two main configurations, representing the two "eras" described above. For the "contemporary" era, the 20 years from 2000–2019, the inputs are all of the EO datasets shown in Table 2.1. For the "hindcasting" era, 40 years from 1980 to 2019, fewer EO datasets are available, for the simple reason that there were not as many satellites in orbit making observations. This model, with a smaller number of inputs, is shown in Figure 4.16. While the contemporary NN uses three precipitation datasets, the hindcasting model only uses two. In addition, the hindcasting model drops one ancillary variable: EVI. (The vegetation index EVI is not available from before the Aqua and Terra satellites. The Terra satellite was launched on December 18, 1999, and the first vegetation data are available for February 2000.)

4.4.3 Pixel-Scale Predictions with the NN model

Once the NN model has been calibrated (and validated) over a set of global river basins, it can be used to make predictions at the pixel scale, over all global land surfaces (excluding Greenland and area above 70° North). I was concerned that input variables assembled over the pixels would contain extreme values outside the range of inputs used to train the NN model. Recall that the training data had been averaged over basins containing tens to thousands of pixels. This averaging would tend to smooth out extreme values. Feeding the model input data outside the range of training data is an extrapolation problem, and could result in unrealistic and unreliable values in the output.

I compared the distribution and statistics of the input data compiled for basins and for pixels as a check on how large the extrapolation problem is likely to be. Figure 4.17 shows empirical probability distributions (kernel density plots) for the EO variables that are input to the NN model. In the legend of each figure, the pair of values in brackets is the minimum and maximum value observed over the domain (pixels or river basins). For the most part, the distributions are overlapping. That is, the values for the EO variables in the pixels are mostly present in the basing training data. It is worth noting that the high outlier values in the pixels are well outside of the range of the basin-averaged data. Because of this, the predictions in those pixels will be asking the model to make extrapolations outside of the range of the training data and will be more uncertain. However, the number of observations outside the range of training data is small. I did a small analysis to count the number of pixel-based observations that are outliers in this sense, and report the percentages for each variable in Table 4.4.

Table 4.4: Percentage of observations among pixel-scale EO variables that are outside the range of basin-averaged training dataset

Dataset	Perc. $<$ Min. Obs.	Perc. $>$ Max Obs.
Precipitation		
GPCP	–	0.007%
GPM-IMERG	–	0.014%
MSWEP	–	0.02%
Evapotranspiration		
GLEAM A	6.1%	0.02%
GLEAM B	4.4%	0.008%
ERA5	4.5%	0.02%
Total Water Storage Change		
CSR	0.006%	0.0009%
GSFC	0.004%	0.0002%
JPL	0.001%	0.0005%

Figure 4.18 shows similar information for the ancillary environmental data. Here, we can see that the distributions for these variables are largely overlapping. The notable exception is surface area, in square kilometers. The area of the pixels is much smaller than that of the basins. In fact, the range of areas is non-overlapping. For this reason, I dropped area from the NN model inputs.

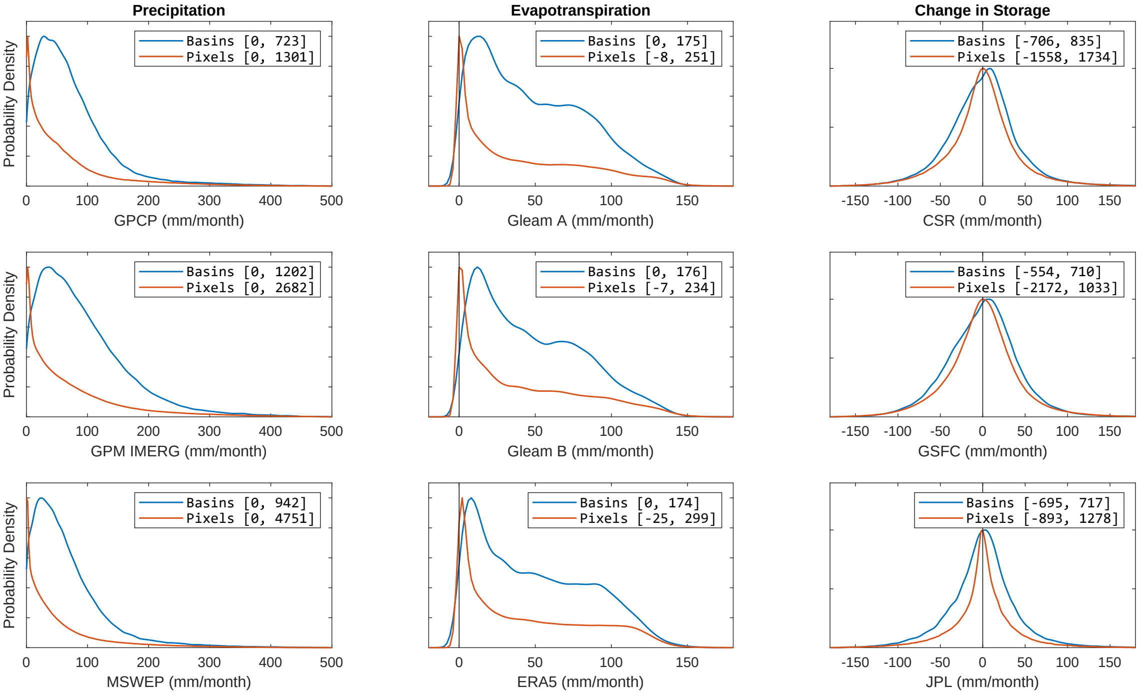


Figure 4.17: Distribution of EO variable values over training basins and over global land pixels.

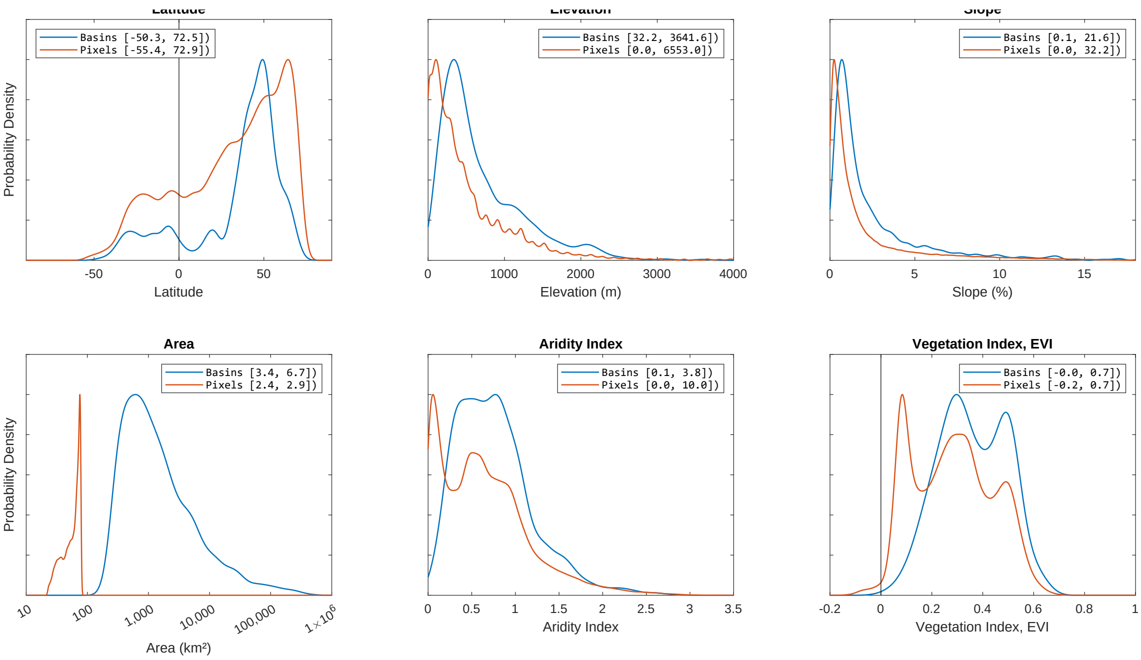


Figure 4.18: Distribution of ancillary variable values over training basins and over global land pixels.

Chapter 5

Results of Modeling to Balance the Water Budget

This chapter presents the results of the modeling analyses, whose goal is to calibrate earth observation (EO) datasets so that they combine to create a balanced water budget. The research described here focused on two distinct modeling methods which seek to emulate or recreate the results of optimal interpolation (OI). The OI analytical method is powerful and effective at balancing water budgets, but it can only be applied over river basins where data are available for all four of the major water cycle components: P, E, ΔS, and R. Previously, Chapter 3 described the OI method in detail. Chapter 4 described the methods for developing models to recreate the OI solution. A key advantage of the models described here is their ability to calibrate the variables individually. In other words, we do not need all 4 of the water cycle components, as we do with OI.

The two classes of models described here use different methods but share a common objective: to take a set EO data as inputs and output a new, calibrated version. The first modeling method involved fitting linear regression models for each variable over a set of global river basins. I describe the methods for this analysis in Section 4.3. Results of the regression analyses over individual rivers are generalized by the creating surfaces of the fitted regression parameters that can be used for spatial interpolation. Spatial interpolation methods are described in Section 4.3.6. I refer to this suite of analyses (regression + parameter regionalization) as "**the regression method**." I also use the abbreviation **Regr.** in tables and figures.

The second method is referred to in this chapter as **neural network modeling**, and with the abbreviation **NN**. Strictly speaking, the NN model is also a kind of regression, as it involves modeling the relationship between a dependent variable and one or more independent variables. Nevertheless, the form and structure of the NN model differs from the conventional regression models described previously. Furthermore, the NN model also includes more input variables, as I have included several environmental variables (elevation, slope, vegetation, etc.) as inputs.

My goal was to first train models at the basin scale, then use the trained models to make predictions in (1) ungaged basins and (2) at the pixel scale over global land surfaces. In this chapter, I begin with a discussion of how the models were trained. As is the case for both types of models described here, the models are not calibrated to fit environmental observations. Rather, they are trained to fit the OI solution for the water cycle components that was obtained earlier and that results in a balanced water budget. Section 5.1 focuses on the results of the regression modeling method. This section includes a discussion of outlier detection and removal, the results of non-parametric Theil-Sen regression and an alternative 3-parameter model. I present the results for each regression model and explain how the final regression model was chosen. The last subsection, 5.1.4, describes the process of parameter regionalization, or using surface-fitting algorithms to spatially interpolate model parameters to new locations outside of the location of original training basins. I methodically tested several surface-fitting techniques and found that kriging, among the more detailed and complex methods, produced the best results with my data.

The following section, 5.2 discusses calibration and training of NN models to fit the OI data. Over the course of my research, I created hundreds of models over varying configurations and sizes. I limit the discussion here to the final phase of model selection. At this point, I had found an NN model architecture that worked well, and focused on the final selection of input variables and setting the network hyperparameters (e.g.: number of layers and neurons). The remainder of this chapter focuses on the model results and compares the results of the two modeling methods. I explore the models' performance in a series of plots and maps. Further, I examine how the models make changes to the input data, and how well the results fit the target OI solution.

Both models can be extended to the pixel scale quite successfully. I explore the geography of the imbalance with a set of maps, to see whether the models perform better in some locations than others. Finally, I analyze how well the EO data fits in situ observations, comparing the fit before and after calibration. This analysis helps to verify that the calibration of EO variables has not degraded the signal too much.

Overall, both modeling methods result in substantial improvements to EO datasets, making them more coherent and helping to close the water cycle. Nevertheless, the NN model outperforms the regression method in terms of overall reduction in the water budget residual. Nevertheless, the regression method enjoys the advantage of being relatively simple, requiring less input data, and is perhaps more readily explainable.

5.1 Regression Model Development

I explored the relationship between the EO variables and their respective OI solutions with various forms of linear regression described in Section 4.3. I found that ordinary-least squares (OLS) regression with two parameters, slope and intercept, performed well. Removing a small number of outliers prior to fitting the regression lines further enhanced the fit. The results of OLS regression were similar to results obtained with non-parametric Theil-Sen regression.

Figure 5.1 is an example of the regression for a single parameter over 3 randomly selected river basins. In the first plot, we can see the influence of high outliers on the OLS regression. The non-parametric Theil-Sen regression is more resistant to outliers and has a more realistic slope. Next we will explore the effect of removing outliers before fitting the regression line.

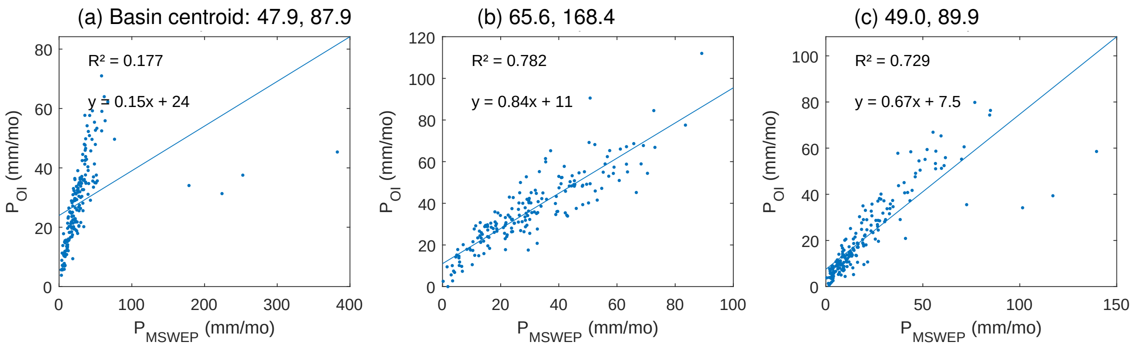


Figure 5.1: Regressions of OI precipitation against observed precipitation from MSWEP over three river basins. The regression line on the plot was fit by the method of ordinary least squares, without removing any outliers. The basins' centroid coordinates are shown above each plot.

5.1.1 Outlier Detection and Removal

As can be seen in the left-most plot in Figure 5.1, there are occasional outliers in the EO datasets that interfere with fitting an OLS regression line. When one or a few points have a large influence on the fit of the line, it can result in a poor fit to the majority of the data, as discussed in Section 4.3.2. In order to detect and remove outliers, I calculated two indicators, univariate leverage and Cook's *D*, for each paired set of predictor and explanatory variables. That is, for each EO dataset (10) and each training basin (1,358). Figure 5.2 shows an example regression where we have removed the outliers which exert strong leverage on the fit of the regression line. The plots show the same data from the left plot in

Figure 5.1 above. In (a), each point has been color-coded according to its leverage h, calculated by Equation 4.31. In this case, we have $n = 184$ points, and so we use an outlier threshold of $h = 16/n = 0.087$. In Figure 5.2(b), we have the same data, but with the four high outliers removed. The OLS regression line is a much better fit to the remaining data, with $R^2 = 0.77$. The slope of the regression line $a = 0.84$ much closer to 1. Because the fitted relationship will be used to calibrate the EO variable, an extreme high or low value is undesirable as it would make large changes to the input.

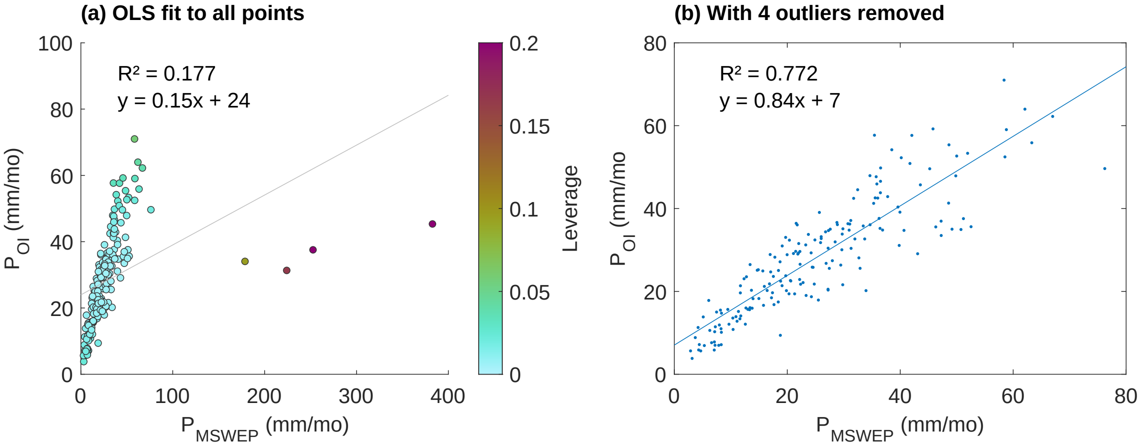


Figure 5.2: Regression analysis before and after removing outliers

The number of outliers varied by EO variable. The number of outliers also varied based on the method used for outlier detection. Figure 5.3 shows the distribution in the number of outliers over 1,358 training basins for each of the 10 EO variables. The mode, or most common value, for the number of outliers is 0, for all 10 EO variables. Runoff has the most outliers, with an average of 1.24 outliers in each basin. There tend to be more outliers among the variables for P, and fewer outliers among the E and ΔS variables. This is expected as the datasets for P and R tend to be more highly skewed to the right, with many observations close to zero, and occasional observations that are much higher. Within each class of variable, the shape of the distribution is similar. Again, this is expected, as the variables are correlated with one another, and tend to exhibit similar behavior.

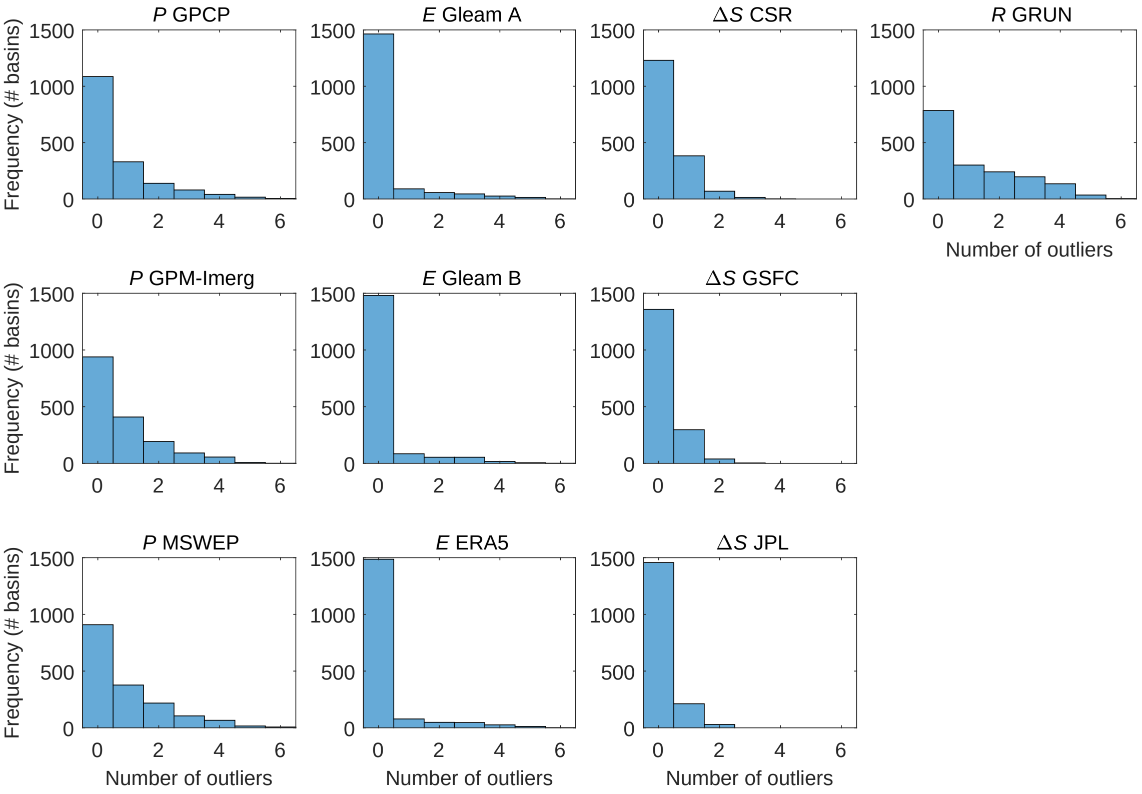


Figure 5.3: Distributions for the number of outliers in the 1,358 training basins for each of the 10 EO variables.

5.1.2 Regression Results

I calculated the regression parameters relating EO data to the OI solution for each of the 1,358 training basins and for each of the 10 EO variables using the various methods described in Section 4.3. These methods included:

1. Ordinary least squares (OLS, 2 parameters)
2. OLS Regression through the origin (RTO, one parameter)
3. Theil-Sen regression, a non-parametric method that is more robust when faced with outliers and skewed data (2 parameters)
4. Alternative regression, $y = a \cdot x + b\left(1 - e^{-\frac{x}{c}}\right)$, which prevents predicting negative values (3 parameters)

The results for the alternative 3-parameter regression appeared to be acceptable when we examine the fits in individual basins. An example is shown for one basin in Figure 5.4. In this basin, the fitted regression lines for each of the 3 methods look similar. However, when we zoom into the lower values (at right), we see that both the OLS and Theil-Sen regression lines have an intercept below zero. Consequently, the fitted relationship predicts negative values for precipitation values less than about 15 mm/month. The alternative 3-parameter regression has the desirable quality of passing through the origin and never predicting negative values for precipitation. The one-parameter RTO regression also passes through the origin, and never predicts negative P, however, it appears to consistently overestimate low values of P.

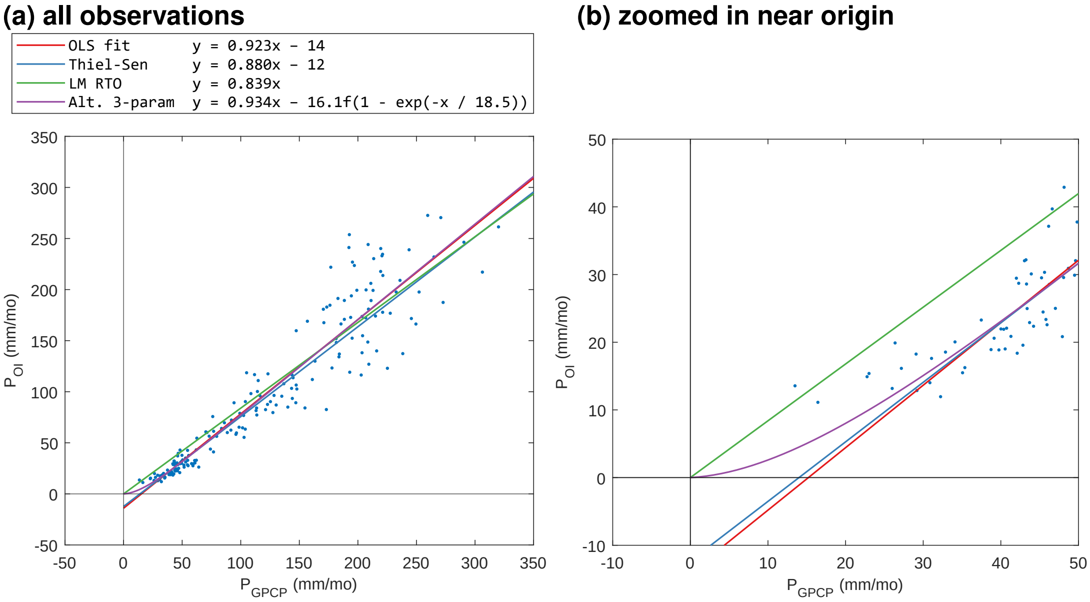


Figure 5.4: Example fits between the EO variable and the OI solution for 3 regression-type models. Example for GPCP precipitation over Pinquen River basin in Peru with centroid at coordinates (-12.4, -71.3).

Despite the desirable qualities of the 3-parameter regression, maps of the parameters show there is a lot of variability. The fitted values vary over multiple orders of magnitude. They also lack a consistent geographic pattern needed to fit a surface and extrapolate parameter values to new locations. Figure 5.5 shows a map of the parameter c and the distribution of its values over the training basins. Note that the vertical axis of the histogram is on a log scale. Most of the values of the parameter c are relatively low, but the distribution is highly skewed to the right, with several high outliers where $c > 14,000$. Because of this large variability, it is not possible to create a smooth interpolated surface. Therefore, I did not consider the results of the alternative 3-parameter regression method in further analyses. It is also worth noting that fitting the 3-parameter equation is much slower and less efficient than either OLS or Theil-Sen regression. Fitting this equation over 1,358 training basins and 10 variables using Matlab's **fit** function took about an hour on a laptop computer, versus a few seconds for the other forms of regression. This makes it slightly less practical for testing and running experiments like resampling and cross-validation. However, this is a relatively minor concern. The main problem with this method is the variability in fitted parameter values and the lack of a consistent geographic pattern.

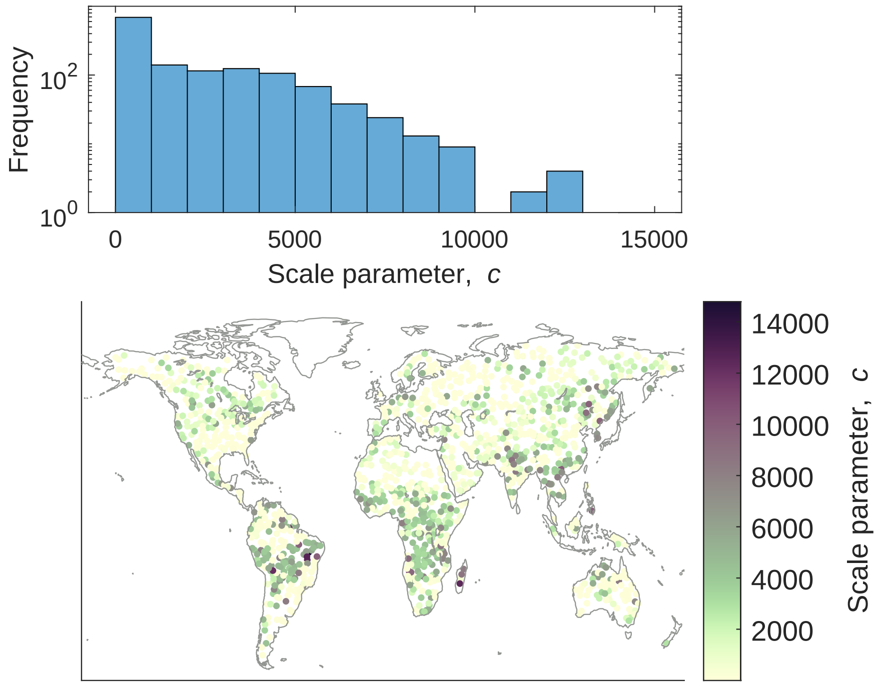


Figure 5.5: Distribution of values of the scale parameter, c, in the alternative 3-parameter regression over the training basins for the precipitation variable GPCP.

5.1.3 Nonparametric Regression Results

We saw in Section 4.3.3 that the non-parametric Theil-Sen regression method is more robust in the face of outliers and skewed data. For our dataset, the results of Theil-Sen regression are similar to those of OLS regression performed after outliers have been removed. Figure 5.6 shows the distribution of the regression parameters (slope and intercept) for the different methods under consideration. In Figure 5.6(a), we can see that the distribution of slopes among the various methods is similar. The first type, OLS regression, tends to have slightly lower slopes on average. The distribution of slopes for OLS regression after outlier removal and Theil-Sen regression are similar. For the one-parameter RTO (with no intercept term), the slopes tend to be somewhat higher on average. Figure 5.6(b) shows the distribution of the intercept terms. It is also worth noting that the Theil-Sen method also has the greatest density of intercepts close to zero. One of the effects of outliers on this dataset is to make the intercept move further from the intercept.

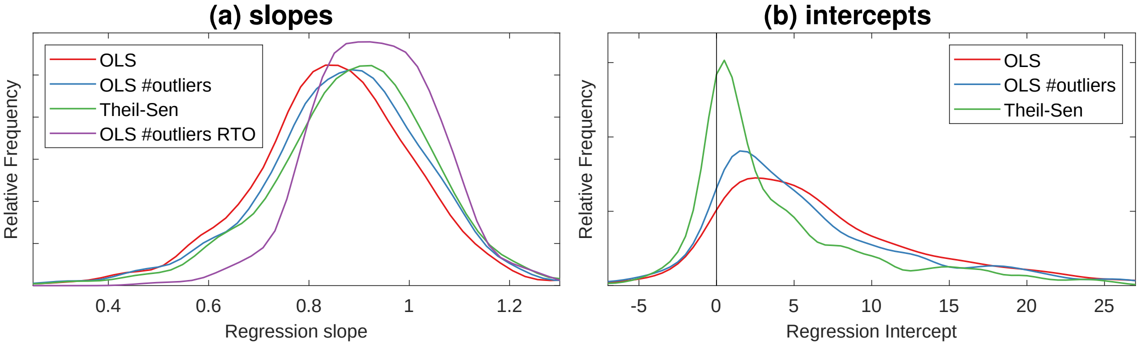


Figure 5.6: Distribution of fitted regression parameters for the precipitation variable GPCP over the test basins for four regression methods.

5.1.4 Spatial Interpolation of Regression Parameters

The regression analysis gives us a simple, straightforward method for adjusting EO datasets such that they are closer to the OI solution for water cycle components which satisfy the water cycle closure constraint. For example, estimates of precipitation from the datasets GPCP, GPM-IMERG, and MSWEP are made to more closely approximate OI precipitation. Next, we turn to the problem of extrapolating the results of the regressions so that we may perform the calibrations outside of our training basins. To do so, we use spatial interpolation methods described above in Section 4.3.6. This method of spatial interpolation can be considered an example of *parameter regionalization*, a method often used in the hydrologic sciences.

Several techniques exist for conducting spatial interpolation. These methods vary in terms of complexity and flexibility. I experimented with seven different methods. Figure 5.7 demonstrates the effect of different spatial interpolation methods. The maps show the spatial interpolation of the OLS regression slope parameter for one of our three precipitation variables, GPCP. The interpolated surfaces cover the global land surface, but Figure 5.7 is zoomed in on Asia to show more detail.

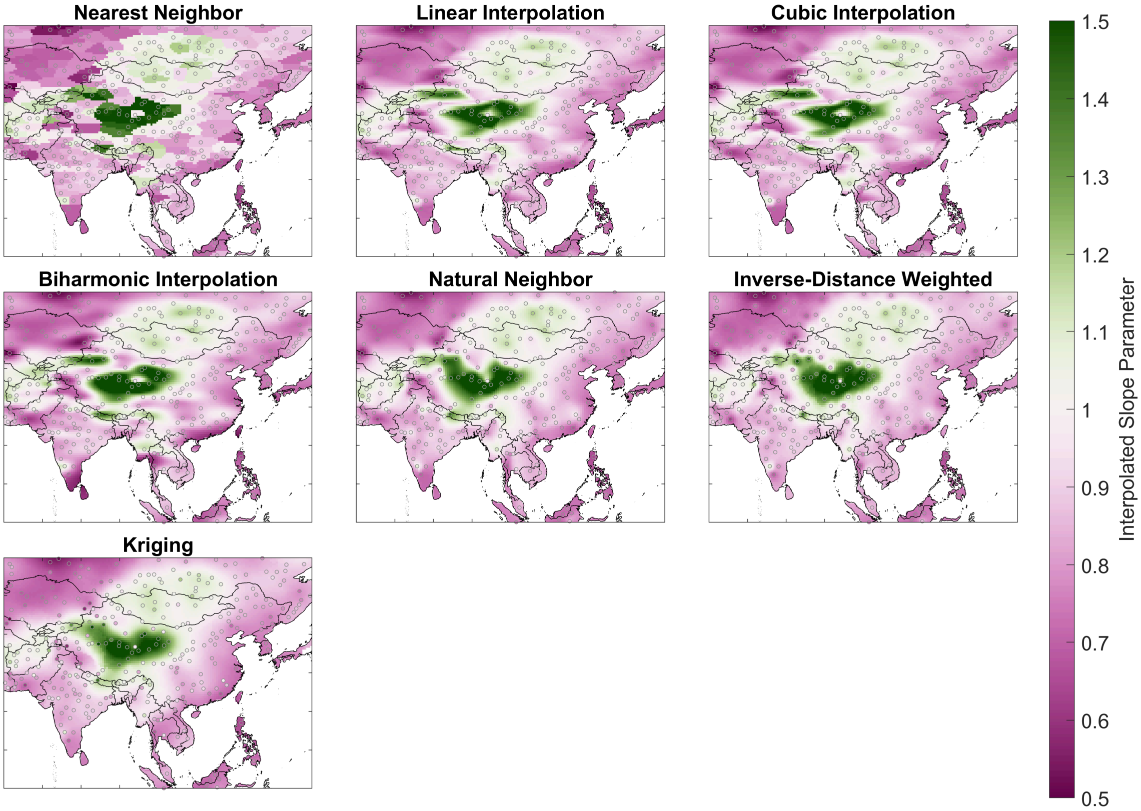


Figure 5.7: Examples of different surface fitting algorithms. The 7 surfaces shown here were fit to the OLS regression slope parameter for calibrating GPCP precipitation over training basins, shown as light gray circles on the maps.

The fitted surfaces vary in terms of their smoothness, shape, and level of fine-grained detail. It is not readily apparent which of these surfaces is the best fit to our data. I believe that many scientists and engineers simply choose a surface-fitting algorithm that is customary within their discipline or within their organization. For example, nearest-neighbor methods are commonly used by hydrologic modelers for spatial interpolation of rainfall data, while kriging is common in geology and mining. I used a more thorough and analytical approach to method selection based on resampling and cross-validation. My method searched for the method that is best at predicting the correct value at locations that were not used to fit the surface.

For the cross-validation experiment, I created 20 different partitions, separating the 1,698 synthetic river basins into training and validation sets. I used an 80/20 split, with 1,358 training basins and 340 validation basins. Basins were assigned at random in each set. As such, my resampling method is not a strict k-fold cross validation, which divides the samples into k contiguous blocks.

I fit the regression model for all 10 variables in each of the 1,698 basins in my training dataset. Then, I fit surfaces for the regression slope parameter. For each of the 20 surface-fitting experiments, the surface was fit using a training set consisting of 80% of the basins, $n_{train} = 1,358$. I evaluated the quality of the fit on the other 20% of basins ($n_{validation} = 340$) by calculating the root mean square error (RMSE). I repeated this experiment 20 times, for each experimental partition. The result is a set of fit statistics for each method and each variable.

Table 5.1 summarizes the results of the cross-validation, showing the average bias error for each method and for each. The table reports the average RMSE (out of 20 trials) for each variable and each surface-fitting method. In Table 5.1, the best RMSE (i.e., the lowest average value) for each variable is in bold. I also calculated the average RMSE across all 10 EO variables and the rank for each method from 1 to 7.

Table 5.1: Summary of the cross-validation experiment to determine which surface fitting method best predicts regression model parameters for calibrating EO variables. Table entries are the root mean square error in mm/month for 340 validation basins

	P GPCP	P GPM- IMERG	P MSWEP	E Gleam A	E Gleam B	E ERA5	ΔS CSR	ΔS GSFC	ΔS JPL	R GRUN	Avg. RMSE	Avg. Rank
Nearest Neighbor	0.118	0.082	0.140	0.210	0.113	0.140	0.086	0.088	0.141	0.316	0.143	7
Natural Neighbor	0.105	0.075	0.126	0.137	**0.088**	0.126	0.078	0.072	0.115	**0.257**	0.118	5
Linear	**0.096**	**0.066**	0.122	0.155	0.095	0.113	0.065	0.065	0.115	0.270	0.116	2
Cubic Spline	0.098	0.068	0.123	0.161	0.096	0.113	0.064	0.064	0.117	0.277	0.118	6
Biharmonic	0.101	0.070	0.123	0.156	0.095	0.112	**0.061**	**0.062**	0.118	0.272	0.117	3
IDW	0.101	0.069	0.129	0.140	0.093	0.118	0.074	0.074	0.115	0.264	0.118	4
Kriging	0.097	0.067	**0.122**	**0.137**	0.088	**0.112**	0.065	0.065	**0.110**	0.266	**0.113**	1

The performance of the various spatial interpolation methods is relatively similar, with values for the RMSE that are similar to within a few percent. The exception is the nearest neighbor method, whose performance is poor compared to the six other methods. Interestingly, the *natural neighbor* method, which adds some smoothing between stations, significantly outperforms nearest neighbor. The relatively simple linear interpolation method is the second best, slightly outperforming more complex methods such as the cubic spline and biharmonic interpolation. Kriging is the most complex method and can fit the most flexible surface. It is the clear winner, outperforming the other six methods on average. Nevertheless, it is not always the best. For some variables, another interpolation method performs better. For example, for ΔS CSR, the biharmonic interpolation method returns the lowest validation RMSE. Yet, across the 10 EO variables, kriging gives the least error overall. I would note that the kriging method has many parameters, and could perhaps be improved even further through extensive trial and error.

We now have a relatively straightforward model for calibrating EO variables in ungaged basins or at the pixel scale. Spatial interpolation gives us the regression parameters at the pixel scale. We can use the fitted surface to look up the best value for the regression equation at any location over any continental land surfaces. Figure 5.8 shows an example of performing the calibration, for a single month and a single variable. Here, we are using the 2-parameter OLS model to calibrate E from Gleam-A, where the parameters for each pixel over land are determined by the surface fit described above. In Figure 5.8(a), we have the original, uncorrected estimates of E from Gleam-A. This is the input data, prior to calibration by the models. The second map (b) shows the calibrated E. A third map, Figure 5.8(c) shows the magnitude of the adjustment. Finally, the histogram in (d) shows the distribution of the changes made in all pixels for this month. Positive values indicate that calibrated E is higher than the input data. The distribution of adjustments made to E for this variable in this month is asymmetrical; pixels where the method is increasing E outnumber those where E is being decreased.

The example shown in Figure 5.8 shows just 1 of 10 variables, and 1 out of 240 months in our temporal domain from 2000 to 2019. Nevertheless, it is interesting to look into the particularities of the changes made to the EO dataset, as this is the main purpose of this research project. However, this pattern of corrections depends on the variable, as each has a different set of calibration parameters. In this case, for Gleam-A, calibration is making the largest increases in E over the central United States and eastern Asia, while the largest decreases are over the northern arctic region in Canada, Alaska, and the European Nordic countries. The changes are typically relatively small compared to the magnitude of observed E, usually less than 10% of the uncorrected observation. Below, we will assess the results of this method in terms of how much it reduces the water cycle imbalance, and compare it to a different modeling method based on neural networks.

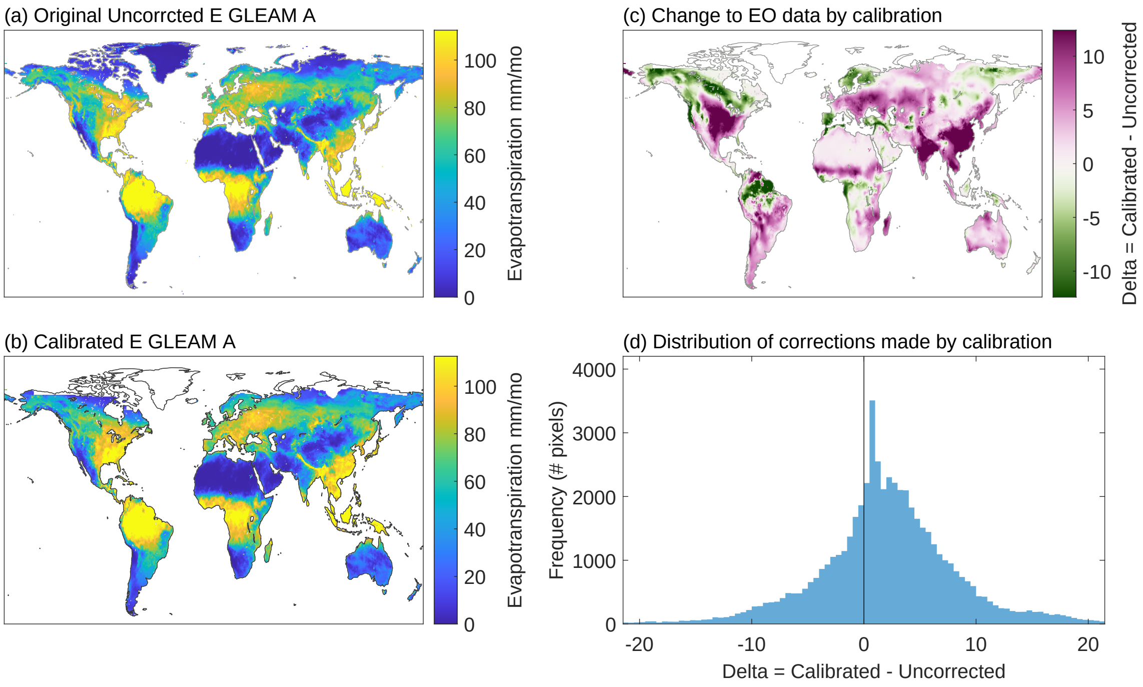

(a) Original Uncorrcted E GLEAM A
100
80
60
40
20
0
Evapotranspiration mm/mo
(c) Change to EO data by calibration
10
5
0
-5
-10
Delta = Calibrated - Uncorrected
(b) Calibrated E GLEAM A
100
80
60
40
20
0
Evapotranspiration mm/mo
(d) Distribution of corrections made by calibration
4000
3000
2000
1000
0
Frequency (# pixels)
-20
-10
0
10
20
Delta = Calibrated - Uncorrected

Figure 5.8: Demonstration of the EO calibration with the regression-based method for Gleam-A evapotranspiration for a single month, May 2007. All units are in mm/month.

5.2 Neural Network Model Development

The NN model whose architecture was shown in Figure 4.16 has two sets of outputs. First, the calibration layer calibrates each of the 10 EO variables individually. Next, the mixture models combine information from individual members of each class of the calibrated EO datasets (P, E, ΔS, and R). The output of each of the 4 mixture models is a calibrated water cycle component in units of mm/month. Recall that the target for the NN models is the OI solution, a set of water cycle components (P, E, ΔS, and R), which satisfy the water cycle closure constraint, or which result in a closed water budget over the training river basins.

In the course of my research, I tested many different configurations and architectures. After converging on an NN model architecture that yielded good results, I tested a number of minor variants, summarized in Table 5.2. This table reports the mean and standard deviation of the water cycle imbalance across the 340 validation basins, and all available monthly observations from 2000 to 2019. I looked at other fit indicators, such as mean square error, in assessing model fit. However, the imbalance gives a good high-level view of our main objective, which is to close the water cycle. (The numbering of the models is arbitrary and refers to my naming system for Matlab scripts and data files.)

Among the variants I tested, some models included larger or smaller networks.

For example, NN 12-4 decreased the number of neurons from 40 to 10. As a result, the model fit and imbalance were degraded. I also tried larger networks, such as 12-6a with three hidden layers. This did significantly improve the fit or lower the imbalance, but the larger model takes longer to train and to run. I also experimented with different activation functions in the hidden layer (see Figure 4.12). Changing from the *tansig* function to the rectified linear (ReLu) function or a pure linear function made the fit and the imbalance slightly worse.

I chose NN 12-5 as the best model from among those described in Table 5.2. Overall, the performance of the networks in Table 5.2 varies somewhat, but not dramatically. The models all shared a similar architecture (Figure 4.16, and only vary in terms of the details. It is worth noting that the values in the table represent a single training run, and these values vary somewhat when the training is repeated. With each training run, Matlab's training algorithm chooses a different set of initial parameters using a random number seed. Therefore, small differences in the outputs shown in Table 5.2 are not likely to be meaningful.

Table 5.2: Experimental neural network configuration trials. Shows experiments with different model configurations, and the resulting water cycle imbalance over the 340 validation basins. The best results have a mean imbalance near zero and a lower standard deviation.

Model	**Description**	**Imbalance (mm/mo)**	
		mean	std. dev.
NN 12-1	Calibration + Mixture, with 5 ancillary variables. Small networks: 1 hidden layer, 20 neurons	−0.9	31
NN 12-2	Ancillary variables are normalized via Box-Cox transformation	−0.8	31
NN 12-3	Adds longitude to ancillaries	−0.5	31
NN 12-4	As NN3, but with smaller network (10 neurons)	2.4	33
NN 12-5	Added an extra 6 ancillary variables: irrigated area, burned area, snow cover, solar radiation, temperature, vegetation growth/senescence (dV/dt)	-0.4	27
NN 12-5b	changes activation function in the hidden layer from tansig to ReLu	−1.2	27
NN 12-5c	changes activation function in the hidden layer to pure linear	−0.3	36
NN 12-6	4 mixture networks only; no intermediate calibration step	−0.6	30
NN 12-6a	Same as NN 6, but bigger network: 3 hidden layers with 30, 10, and 3 neurons in each layer, respectively	-0.7	27
NN 12-7	same as 6, but the ancillaries are normalized	−1.0	27

The final NN model architecture (Figure 4.16) contains individual networks each with 1 hidden layer and 20 neurons. I experimented with different numbers of layers and neurons, and determined that this configuration minimizes the cost function (mean squared error) in the training and testing sets. Figure 5.9 shows the results of an experiment in training a network with a single layer and a varying number of neurons between 5 and 40. The lines show the root mean square error (RMSE) for each of the water cycle components over the validation set. There is a slight decline in the error as we increase the number of neurons from 5 to 20, although the trend is not strong. Adding more layers and neurons beyond 20 does not result in lower error. However, larger models with more parameters were not necessarily worse; I did not see any evidence of over-training. This is largely due to the smart training algorithms in Matlab, which stops training when validation errors begin to increase. However, we also do not see evidence that the larger models with more neurons perform any better. Therefore, it was more parsimonious to choose the smaller model with fewer parameters. Furthermore, smaller models are faster to train and to run.

The plots in Figure 5.9 also show the effect of including the ancillary environmental variables on the quality of the output of the NN model. Models that include environmental information consistently had lower overall error and greater ability to reduce the water cycle imbalance.

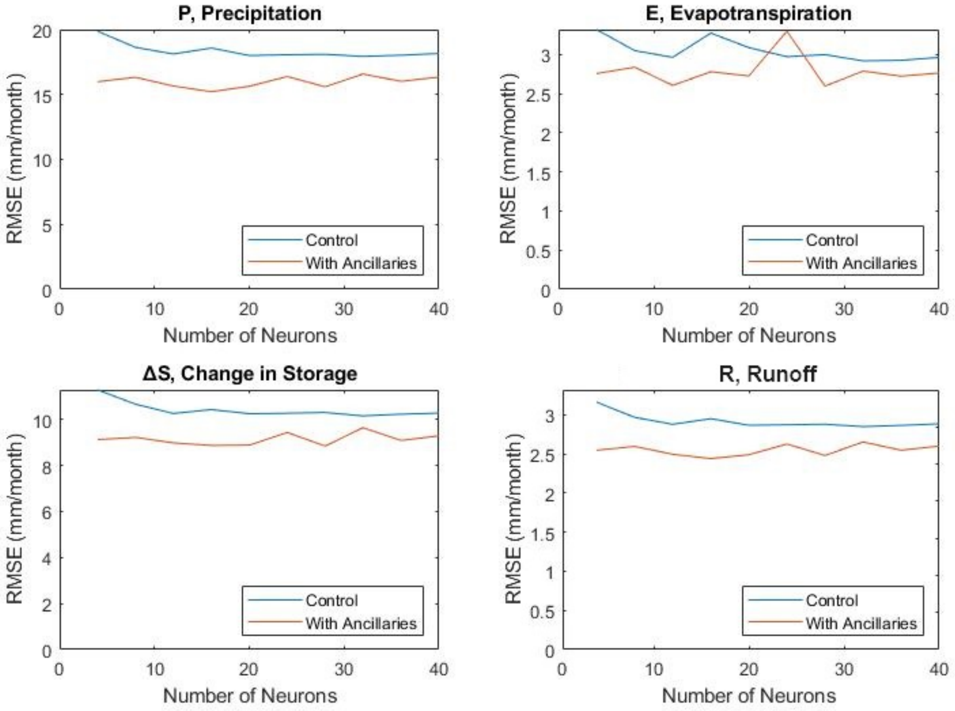
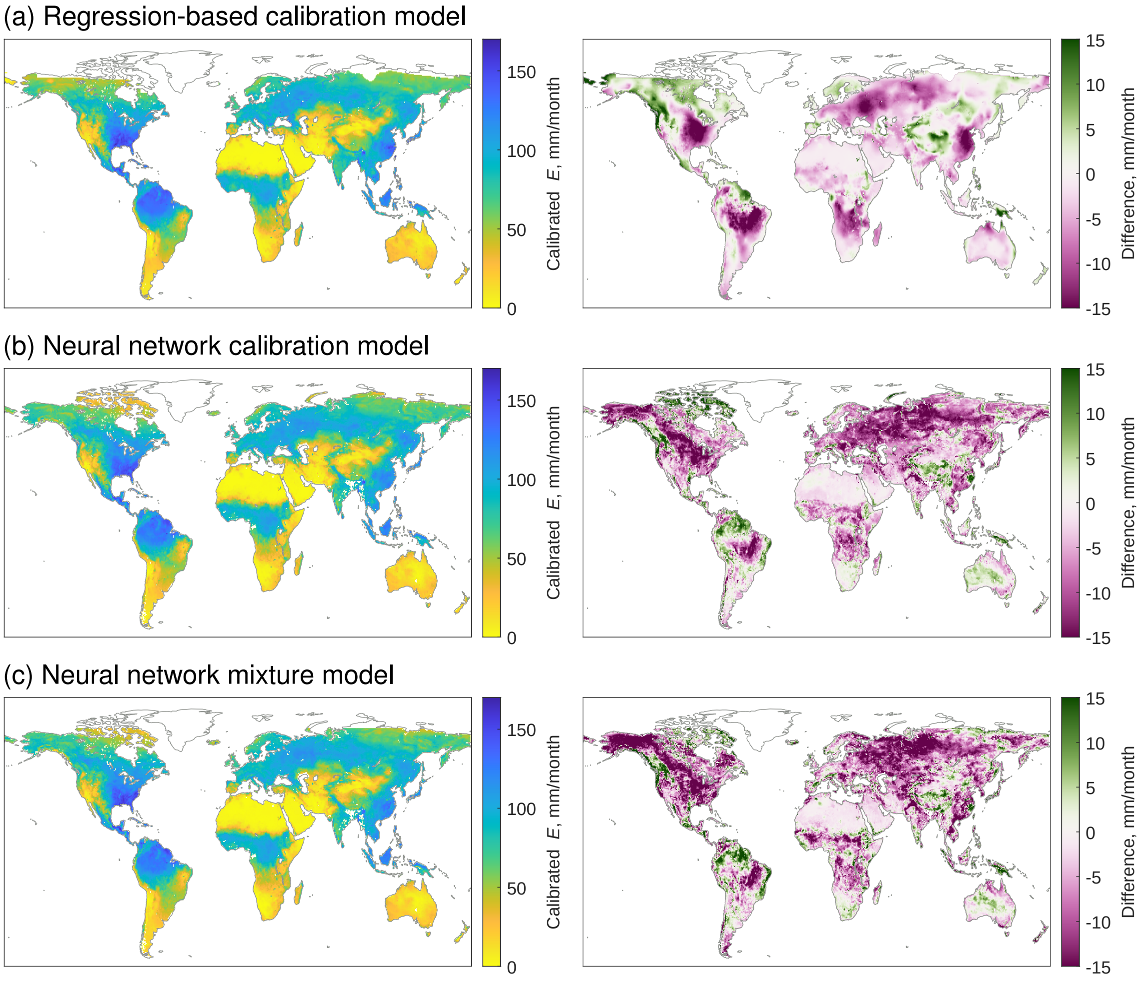


Figure 5.9: Validation set error vs. the number of neurons

5.3 Calibration of EO Variables at the Pixel Scale

A stated goal of this research has been to create a model capable of making good predictions at the pixel scale. In other words, the model should be transferable from the basin-scale data to similar data compiled at the pixel scale, and still be capable of calibrating EO datasets in a way that is realistic and helps close the water cycle. For the regression-based model, making pixel-scale predictions is straightforward. For each pixel over continental land surfaces, we have a simple linear relationship with two parameters, a slope and an intercept. The parameters are stored in 360 × 720 matrices for the 0.5° global grid. Thus, we have a total of 20 such matrices: 10 EO variables × 2 parameters for each. The matrices contain values for grid cells over land surfaces only. Grid cells contain NaN (not a number) for oceans and high northern latitudes outside of our model's domain.) The calculation involves simple matrix multiplication and addition, and is fast and efficient. Applying the NN model is more computationally intensive, but not burdensome, and can be calculated in a few seconds on a laptop computer. While the models were calibrated over the time period from 2002–2019, they can be readily applied to time periods before or after this. In the following chapter, Section 6.2, I test the ability of our model to "hindcast" GRACE-like total water storage prior to the satellites' launch in 2002.

The full set of pixel-scale results covers 10 EO variables, monthly over a 20 year period. While this is too much information to display on the page, I show an example for a single variable in a single month in Figure 5.10. The maps show the calibrated E from GLEAM-B, via the regression method in Figure 5.10(a), and the NN method in (b). The maps in Figure 5.10(c) show E calibrated by the NN mixture model, which combines information from three different calibrated EO datasets. To the right of each figure is the percent difference between the uncorrected data and the calibration. All units displayed in the maps are in mm/month.


(a) Regression-based calibration model

(b) Neural network calibration model

(c) Neural network mixture model

Figure 5.10: Pixel scale calibrated evapotranspiration for July 2004 calculated by (a) regression model, (b) NN calibration model, and (c) NN mixture model. The maps in (a) and (b) are calibrated versions of GLEAM-A, while (c) combines information from 3 EO datasets.

Based on the maps in Figure 5.10, the global pattern of evapotranspiration looks somewhat similar for all three calibrations. However, there are small differences that are more readily visible in the maps on the right side of Figure 5.10, which show the difference between the uncorrected GLEAM-B dataset and the calibrated version. The geographic pattern of differences based on the regression-based model appear to be smoother, a result of the spatial interpolation of model parameters used to make adjustments to the EO dataset. There is more fine-grained spatial detail in the NN-based model results. This is also not surprising, as the NN model includes a wider range of inputs that includes environmental variables that vary spatially over short distances, such as slope, elevation, and vegetative cover.

We can also see some differences among the modeling methods when we focus on a particular region. In North America, the regression-based calibration appears somewhat different from the NN models, particularly in the northwest (Canada and Alaska). There also appear to be some differences in the sign of the change across northern Asia and Australia.

The one-month snapshot of a single variable in Figure 5.10 is not enough to describe the effect of the calibration as a whole. In the following section, I explore how much the calibration changes the original EO datasets and how close the calibrated EO variables are to the OI solution. I also look at the overall impact of the calibration closing the water cycle, at both the basin scale and pixel scale.

5.4 Comparison of the Two Modeling Methods

Thus far, we have applied two different modeling approaches for calibrating earth observation (EO) data of the four main water cycle components. The first method was based on linear regression and spatial interpolation of regression model coefficients, and the second method uses neural network NN modeling. Both of the methods seek to recreate the optimal interpolation solution, which has the key disadvantage of only being applicable over river basins where runoff data is available. Both of the modeling methods were trained on basin-scale data. After the models have been trained, the trained models can be applied to calibrate EO data at the pixel scale. The goal of both methods can make the datasets more coherent, resulting in a lower overall water cycle imbalance. However, which method performs best? In this section, I compare the performance of the regression and neural network models to determine which is more successful, according to several different criteria.

To visualize the results, let us begin with a set of time series plots over selected river basins. This is a simple and intuitive way to review the results, letting us quickly visualize the inputs and outputs of the model. However, because the models were trained and validated over thousands of basins, it is not practical to view all the results in this way. The remainder of the figures present the results more globally, integrating information from all the modeled basins. The results shown here are focused on the set of 340 validation basins, as we are interested in how well the model performs with data that were not a part of its training.

Figure 5.11 is an example showing the inputs and outputs of the analyses over one river basin, the White River in the United States. Here, the river basin coincides closely with the drainage area for the gage at St. Petersburg, Indiana (GRDC gage 4123202, or USGS gage 03374000), with an area of 29,000 km^2. While there is no "typical" river basin, this location has a long record of river discharge, and so it does a good job demonstrating the output from our calculations. Further, over this region of the United States, remote sensing datasets tend to be more reliable and well-calibrated, due to the density and availability of in situ calibration data. One generally expects EO data to be more accurate compared to areas of sparse in situ data, for example parts of Africa, Asia, and Latin America.

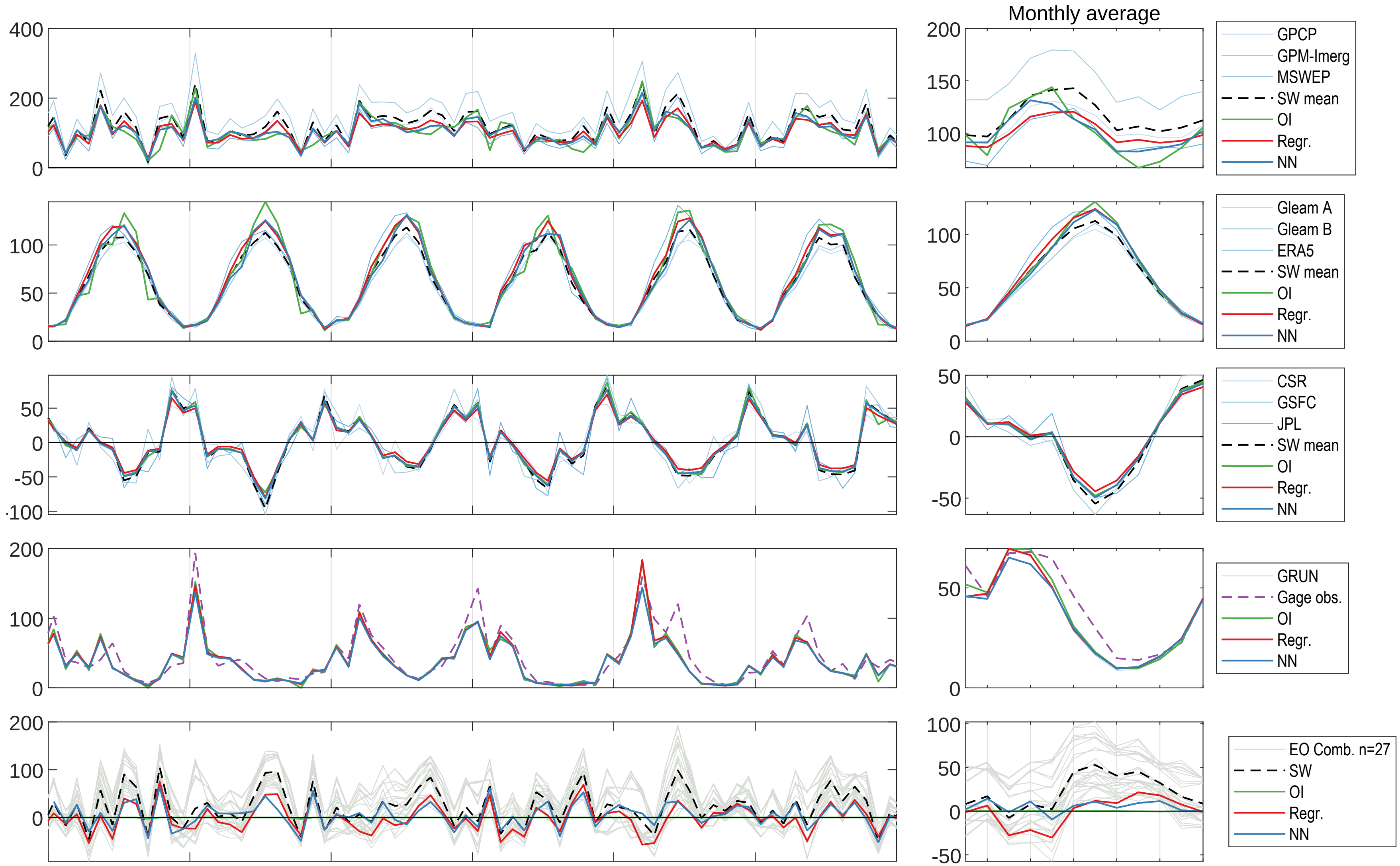


Figure 5.11: Time series plots of hydrologic fluxes over a single basin, the White River in Indiana, USA. Left: observed (pastel colors), combined via optimal interpolation (OI, red), and estimated by the neural network model (NN, blue). Right: corresponding seasonality (monthly averages).

Figure 5.11 shows time series plots of the inputs (observed hydrologic fluxes, in gray) and the outputs of optimal interpolation (green), the regression-based model (red) the NN model (blue). Times series are presented, from top to bottom, for: P, E, ΔS, R, and the water budget Imbalance, $I = P - E - R - \Delta S$. The seasonality of each dataset is shown in monthly average plots on the right. These full time series cover 2000 to 2019, but here I have zoomed in on the 6-year period from January 2004 to December 2009 in order to show more detail. Despite this, there is a lot of data, and it is not possible to clearly distinguish the datasets, as they are frequently overlapping. However, with careful inspection, certain patterns begin to emerge. For example, there is significant disagreement among the 3 precipitation datasets. In particular, GPM-IMERG tends to show higher values than the other two datasets. In contrast, the evaporation datasets are more consensual, at least at this location. The three GRACE datasets for TWSC are also highly correlated with one another. This is expected as each of the datasets is calculated from the same satellite data using different methods.

In Figure 5.11, the bottom time series for the water cycle residual or *imbalance*, I: The gray lines show each of the 27 possible combinations of the datasets ($3P \times 3E \times 3\Delta S \times 1R$). The imbalance of the various combinations of EO datasets is large: the seasonal I can reach ±50 mm/month depending on the combination of datasets. It is the objective of the integration technique to reduce this imbalance as much as possible. The imbalance from the OI solution (in green) is equal to zero by definition. This is why this solution is chosen as a target for the NN integration. The regression and NN optimizations both result in a significant improvement of the imbalance.

In this one example, we can see that the original EO datasets are not modified too much by the models, which is an important feature. In the following sections, we will look into this in more detail, examining (a) how much the NN model has changed the input data, and (b) how closely the output matches the OI solution. Next, we will look at how well the results close the water cycle by examining the remaining imbalance. Finally, we will examine the results of applying the models for making predictions at the pixel scale.

5.4.1 How Much Does Calibration Change EO Data?

We are interested in seeing how much the NN model has changed the EO data. Recall that the goal was to reduce the water cycle imbalance by making changes to the EO data, thus making them more coherent with one another, i.e., resulting in a balanced water budget. However, we would like these changes to be as small as possible, while still achieving the objective of a balanced water budget. If there are certain locations where large changes are necessary, it indicates that one or more of the components has a large bias. Our approach uses data fusion and statistical modeling to reconcile errors. However, the old adage "garbage in, garbage out" applies here. These methods do not have the discernment to treat very large errors. Figure 5.12 shows the relationship between the inputs to our models and the output for each of the four major fluxes predicted by the NN mixture model. On the horizontal or x axis we have the uncorrected EO dataset, and on the vertical or y axis, we have the calibrated version, after being processed by (a) the regression model or (b) the neural network model.

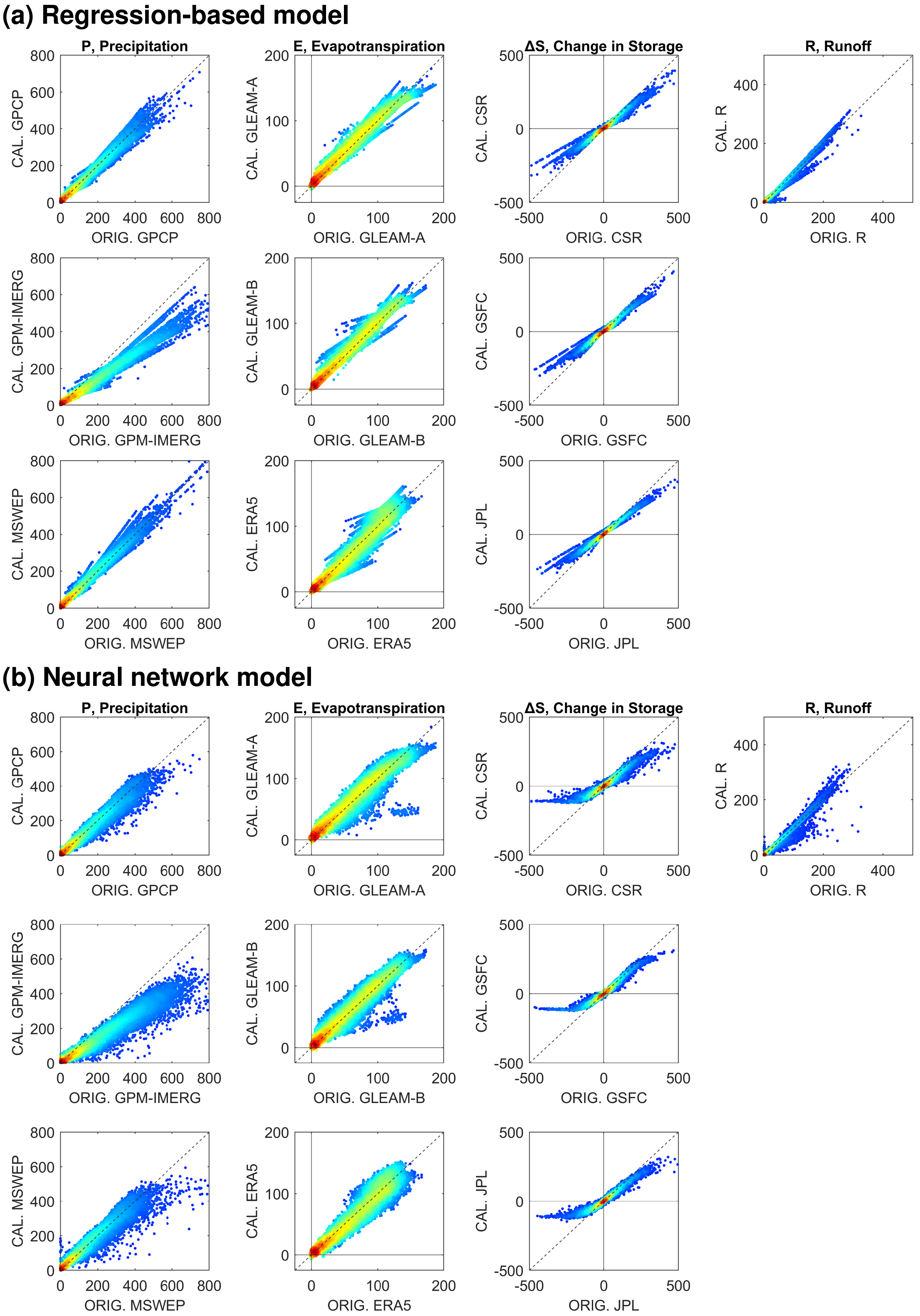


Figure 5.12: Scatter plots showing changes made to EO data by the NN mixture model. All units are in mm/month.

The scatter plots in Figure 5.12 show *all* monthly observations and predictions over the 340 validation basins between 2000 and 2019. Each of the 75,000 data points is an observation for one month and one river basin, with darker colors indicating greater density of points. (Note that there is some missing data where one or more input variables was missing.) Reviewing these graphs, several noticeable patterns emerge, offering some insights to the changes our model makes to the input data.

For the precipitation datasets, the input/output relationship appears to be mostly linear, while the cloud of points has somewhat of a cone shape, meaning that the model is making bigger changes to higher values of P. For GPCP and MSWEP, the points are mostly clustered around the 1:1 line. In contrast, the slope for GPM-IMERG is less than 1; here the NN model is consistently reducing the values. This is evidence that the GPM-IMERG dataset may have a positive bias, i.e., it is over-predicting P in some regions.

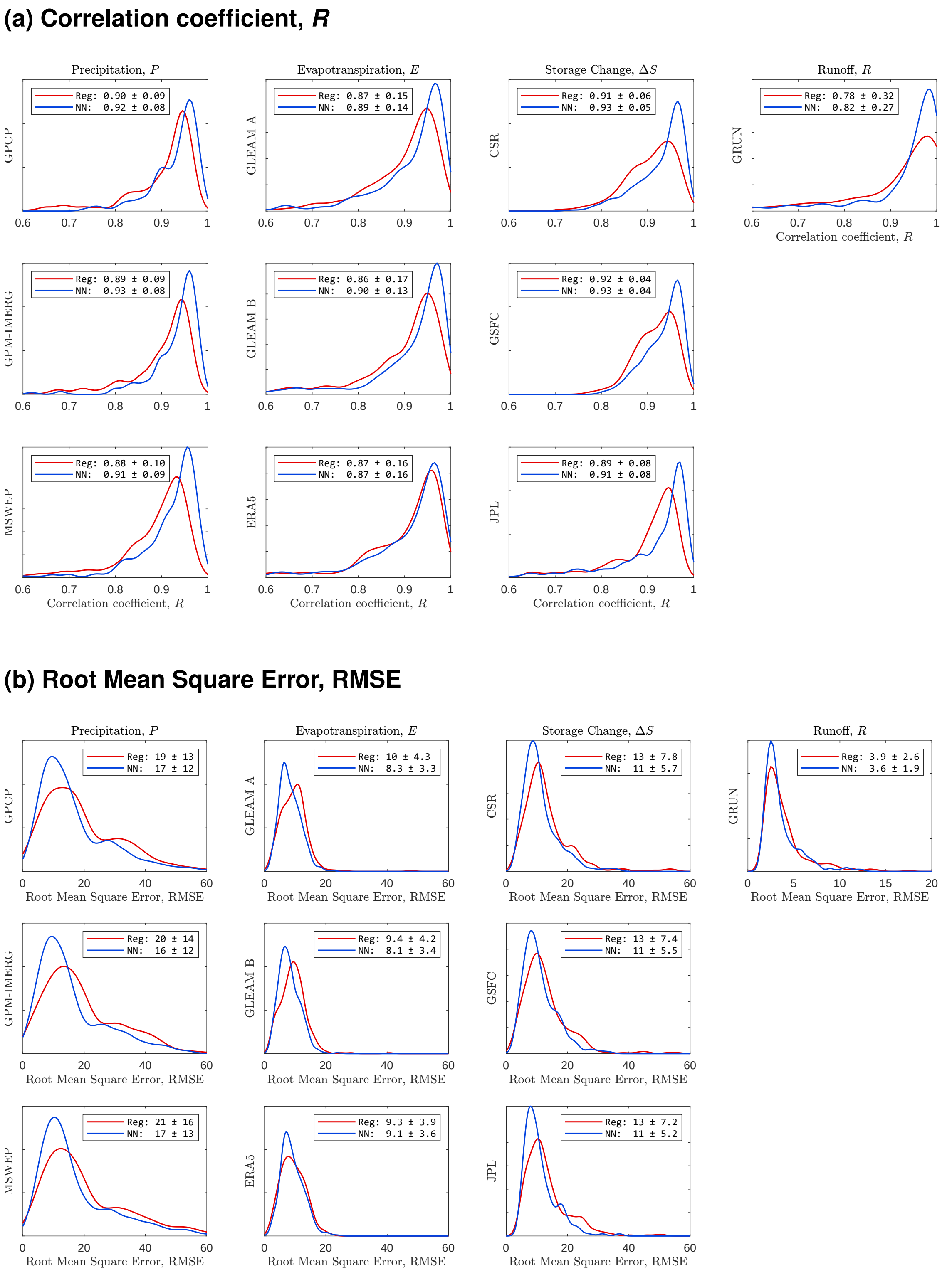
With respect to the changes made to ΔS, the changes made by the NN model have a distinct sigmoid shape, which commonly occurs with NN models like ours which use a non-linear sigmoid function *tansig*. Both the regression and NN models appear to be dampening the signal: extreme high and low values are being squeezed into a smaller range. However, the relationship appears to be more linear in the lower range of values. Very high and low values of (ΔS ¿ 100 mm/month) are unusual, representing less than 2% of observations, and tend to occur in cold climates where snow and ice accumulate and melt.

For runoff, there is only one input dataset. On the scatterplots for R in Figure 5.12, there are several faint lines. It seems that there are multiple relationships, represented by the different lines on the plot of uncorrected versus calibrated R. This is a result of distinct changes being made to the inputs in different locations and environments, and is evidence that the NN is performing as intended.

5.4.2 Goodness of Fit to the Target OI solution

The role of both the classes of models is to calibrate EO variables so that they are closer to the OI solution. Training the models involves determining the set of model parameters or coefficients such that the model output is the best fit to this target. Therefore, it is important to evaluate how well the models are able to recreate the OI solution. I evaluated the goodness of fit between the model output and OI over the 340 validation basins, which were not used in the training of the model. Estimating the basin mean calibrated EO variables over the validation basins is a two-step process. First, the models are run to create output at the pixel scale, then the gridded data is spatially averaged over the validation basins. This step was repeated for all 340 validation basins and the 240 months from 2000 to 2019. I used the fast algorithm for calculating basin means described in Section 3.4. I then compared the computed time series for each of the basins to the OI solution, computing several goodness-of-fit indicators, and repeated this for all 10 variables.

Figures 5.13(a) shows the distribution of correlation coefficients, R, for each of the 10 EO variables. Figure 5.13(b) shows similar information, but here the fit indicator is the root mean square error, RMSE. The legends in each individual plot show the mean and standard deviation for the fit indicator across the 340 validation basins. For example, for the precipitation dataset GPCP, the correlation coefficient, R, has an average of 0.90 and a standard deviation of 0.09. Careful inspection of each of the plots shows that the NN model outperforms the regression model most of the time. For the ERA5 evapotranspiration, R is equivalent for both models. However, the NN model has a lower RMSE than the regression model. The NN model also has a smaller error variance, as indicated by the standard deviation of the RMSE over the 340 validation basins.



(a) Correlation coefficient, *R*

(b) Root Mean Square Error, RMSE

Figure 5.13: Goodness of fit between model output and the OI solution over the set of 340 validation basins, comparing the performance of the regression and neural network models. Top plot (a) shows the distribution of correlation coefficients, bottom (b) shows the root mean square error. The horizontal axes represent the indicator value, and the vertical axis quantifies the frequency or probability density.

Information on the fit is also summarized in Table 5.3, showing the *median* correlation coefficient, R and the mean root mean square error for the fits across the 340 validation basins. The values in the table make it clear that the NN model outperforms the regression model in terms of fit to the OI solution.

Table 5.3: Summary of the fit of calibrated EO data to the OI solution. Table entries are the medians for the fit statistic across 340 validation basins.

	Corr., R		RMSE	
	(0−1)		(mm/month)	
	Regr.	NN	Regr.	NN
P GPCP	0.93	0.95	16	12.8
P GPM-IMERG	0.92	0.95	16.1	12.6
P MSWEP	0.91	0.94	16.2	12.9
E GLEAM A	0.92	0.94	9.5	7.7
E GLEAM B	0.93	0.95	9.2	7.6
E ERA5	0.93	0.94	8.9	8.5
ΔS CSR	0.92	0.95	11.3	9.8
ΔS GSFC	0.92	0.95	11.0	9.4
ΔS JPL	0.92	0.95	11.2	9.6
R GRUN	0.93	0.96	3.2	2.9

5.4.3 Reduction of the Water Cycle Imbalance

The purpose of our modeling has been to calibrate EO variables so that they are more coherent and result in a balanced water budget. Therefore, the water cycle imbalance ($I = P - E - \Delta S - R$) is one of the most important indicators. Imbalance that is near zero is the best. A good model is one that makes the imbalance significantly closer to zero. The more the imbalance is reduced, the better the model is performing. Recall that the target for our model is the OI solution for the four water cycle components, which sum to a balanced water budget. Therefore, a model that is able to perfectly recreate the OI solution, will also perfectly close the water cycle. Figure 5.14 shows the distribution of the water cycle imbalance calculated by uncorrected EO data and by the two modeling methods. The basin mean imbalance is plotted as a kernel density. Here, the mean was calculated over each basin with all available monthly observations over the period from 2000 to 2019. The distributions show the spread of the basin mean imbalances over the 340 validation basins. The light gray lines represent the 27 possible combinations using the uncorrected EO datasets ($3P \times 3E \times 3\Delta S \times 1R$). The three gray lines that are to the right of the others are all calculated with the

GPCP precipitation dataset. As we have seen previously, GPCP tends to report higher precipitation than the two other precipitation datasets used in this study.

The regression-based method produces calibrated versions of each of the 10 variables ($P_1, P_2, \ldots R$). The light pink lines in Figure 5.14 (a) are the imbalances from the 27 possible combinations of the calibrated variables. Unlike the NN models, there is no standalone "mixture model." However, it is logical to combine the information output by the regression model by taking a simple average. I calculated the mean across each variable class, for example $\bar{P} = \text{mean}(P_1, P_2, P_3) =$ mean the average of P(GPCP), P(GPM-IMERG), and P(MSWEP). In Figure 5.14(a), the darker red line is the imbalance calculated with such averages, i.e.: $I = \bar{P} - \bar{E} - \bar{\Delta S} - \bar{R}$. The regression-calibrated mean significantly outperforms any of the individual combinations. Compared to the uncalibrated EO datasets, the calibrated datasets result in a significantly lower imbalance, on average. On the right side, Figure 5.14(b) shows the variability of the imbalance across, again across the 340 validation basins. Specifically, the plot shows the distribution of values of the standard deviation of the imbalance in the basins. The regression modeling has moved the central tendency of the imbalance closer to zero. In addition, the model has also reduced the variance of the imbalance in most basins. In other words, after calibration, the imbalance is closer to zero more frequently.

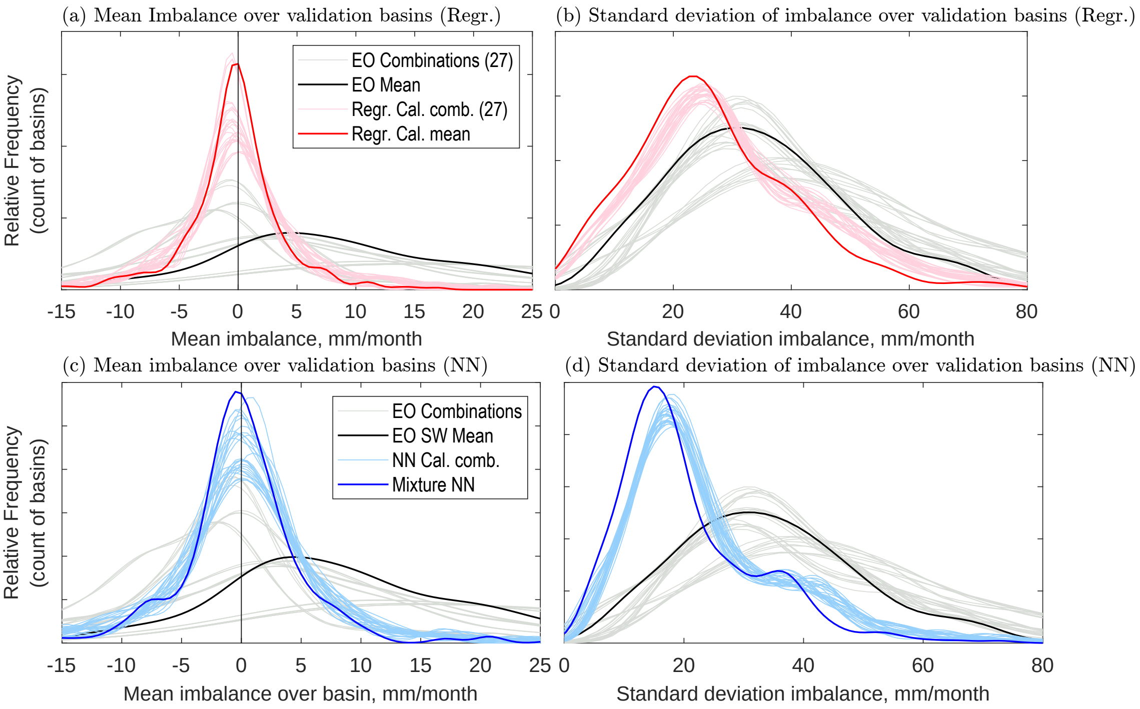


Figure 5.14: Water cycle imbalances over the 340 validation basins, shown as empirical probability distributions (kernel density plots) for (top) Regression-based model, (bottom) neural network model.

Figure 5.14(c) and (d) show the reduction of the water cycle imbalance achieved by the NN models over the 340 validation basins. Because the NN calibration model operates individual variables, we have 10 outputs ($3P \times 3E \times 3\Delta S \times 1R$). The light blue lines are the imbalance calculated via all 27 possible combinations of the NN-calibrated data. The dark blue line is for the NN mixture model. We can see that the NN models result in a substantially lower mean imbalance and a lower imbalance variance. The calibration models are a big improvement over uncorrected EO datasets. The mixture model is an improvement over the calibration alone. Comparing Figure 5.14(b) and (d), we see that the NN model is more effective than the regression model at reducing the variance of the imbalance. However, it is not clear from comparing Figure 5.14(a) and (c) which model has a greater effect on the mean imbalance. A different set of plots can help to elucidate this question, shown next.

The kernel density plots in Figure 5.15 are calculated with all available monthly observations from 2000 to 2019 over the 340 validation basins. As above, we see that both methods significantly reduce the imbalance. The imbalance calculated with the simple-weighted average of uncorrected EO datasets has an average of 11.1 mm/month. The distribution is wide and has "heavy tails," with around 3% of monthly observations having an imbalance of over 100 mm/month. Compared to the imbalance calculated with uncorrected EO datasets, both the regression and NN methods result in an average imbalance much closer to zero, and with a lower variance. From the plot in Figure 5.15, it is hard to say which method is better, as the mean imbalance is similar, with $\bar{I} = 0.8$ for regression and $\bar{I} = -0.6$ for the NN model. They are both quite close to zero, but on opposite sides of the origin.

Both methods reduce the variance in the imbalance, as indicated on the plot by the standard deviation. The variance of the imbalance is slightly lower with the NN model. While the difference is relatively small, it has strong statistical significance. Because of the large number of samples ($n > 56,000$), the results of a statistical hypothesis test are strongly conclusive. A two sample F-test for difference in sample variance rejects the null hypothesis of no difference between sample means with $p < 1 \times 10^{-33}$. This allows us to conclude that this is a small but real difference between the two methods. Therefore, we may conclude that the NN method results in water cycle imbalance closer to zero more frequently than the regression-based method.

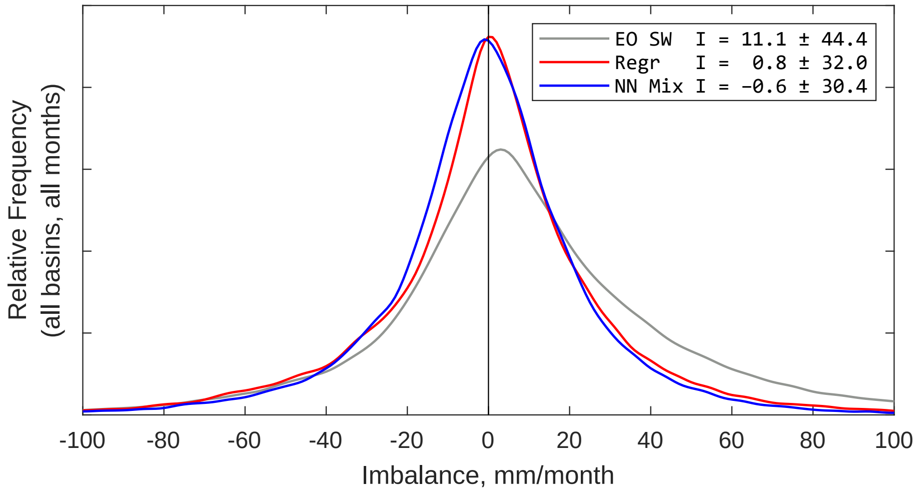


Figure 5.15: Comparison of the water cycle imbalance over validation basins for the two modeling methods

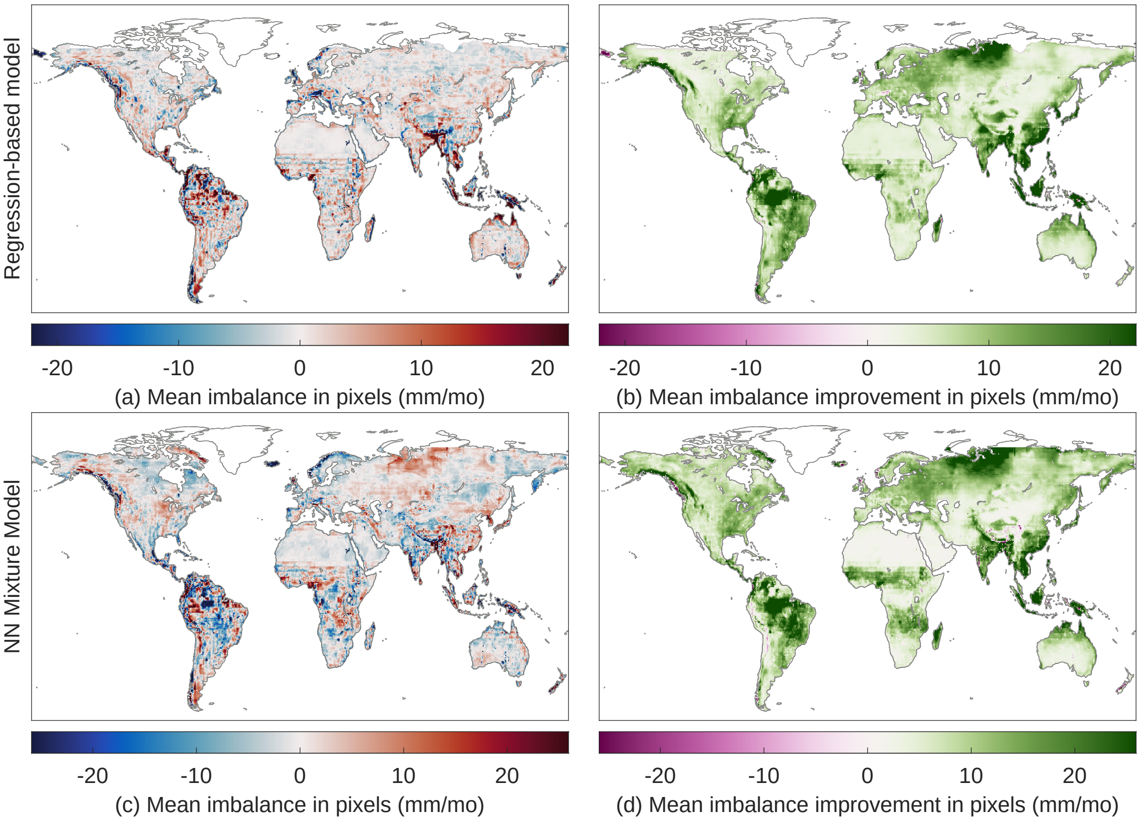
5.4.4 Geography of the Imbalance

Because we have access to all four water cycle components at the pixel scale, it allows us to calculate the imbalance over all land pixels. Figure 5.16 shows the change in the water cycle imbalance calculated from fluxes calibrated by the regression-based model (a) and the NN mixture model (c). The maps in Figure 5.16(b) and (d) show the difference in the imbalance compared to the imbalance with the simple weighted EO datasets. The NN model results in a lower water budget residual in many locations, with particularly large improvements over parts of the Amazon and southeast Asia. The imbalance is made worse in a few small locations, notably near the western coasts of Canada, Chile, England, and Norway.

Which of the modeling methods reduces the imbalance the most at the pixel scale? To answer this question, I compared the imbalance *before* and *after* calibration. I defined a "closure improvement factor," comparing I_{cal}, the imbalance with calibrated datasets, to I_{obs}, the imbalance based on the SW mean of EO datasets. The closure improvement factor and is calculated as:

$$F = |I_{obs}| - |I_{cal}| \tag{5.1}$$

The closure improvements factor, F, is a "convergence metric" that measures how much closer I is to zero after calibration. It uses the absolute value because it is the distance from zero that we care about here, not the sign of the difference from zero. The plot in Figure 5.17 summarizes the improvement in closure over all land pixels in our study domain (excluding Greenland and latitudes above



(a) Mean imbalance in pixels (mm/mo)

(b) Mean imbalance improvement in pixels (mm/mo)

(c) Mean imbalance in pixels (mm/mo)

(d) Mean imbalance improvement in pixels (mm/mo)

Figure 5.16: Map of the average water cycle imbalance in 0.5° pixels over the years 2000–2019. Top plots are for the regression-based model, bottom shows the NN mixture model. At left is the mean imbalance, and at right is the improvement (reduction) in the imbalance compared to water budgets with uncorrected EO data. (Green = better; purple = worse)

73° North). Positive values indicate that the imbalance is closer to zero (lesser in magnitude, the desired result), while a negative value means that the imbalance is further from zero (greater in magnitude, an undesirable result).

The NN model has a higher improvement factor, on average, which means that it results in a reduction in the imbalance over a larger number of pixels. We can conclude, therefore, that the imbalance reduction is greater on average with the neural network. The regression model reduces the water cycle imbalance by an average of 8.7 mm/month over global land pixels, while the NN model reduces the imbalance by 10.3 mm/month.

In the maps in Figure 5.16(b) and (d), we noted that the calibration very occasionally makes the imbalance worse on average. These are grid cells where $F < 0$. Fortunately, this is a rare occurrence. The imbalance is made better ($F > 0$) by the regression method over 99.1% of pixels, and in 98.3% of pixels for the NN model. One of the largest problematic areas, where the closure is degraded, is the Chukotka Peninsula in Russia's far east. This is largely due to the peculiarity

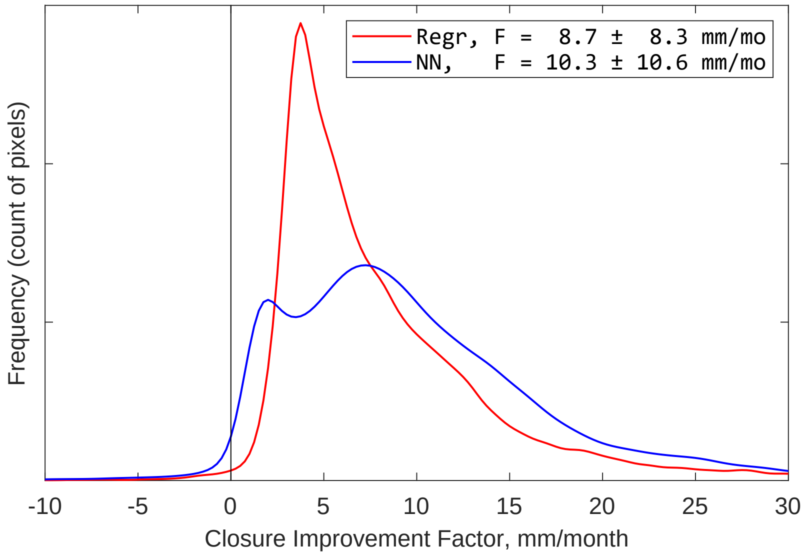


Figure 5.17: Improvement in the water cycle closure at the pixel scale for the two modeling methods. Higher values mean more improvement to the water cycle closure.

of where this region is located on our modeling grid. The standard global grid used widely in climate studies is centered on 0° longitude – the Prime Meridian passing through Greenwich, England. As a result, the left side of our map, at $-180°$, cuts off the eastern part of the Asian mainland. As a result, the far eastern portion of Russia appears as an island at the extreme northwest of our map. It is separated from the remainder of its physiographic province, which is on the far east or right side of our map. The regression + surface fitting model struggles to make accurate predictions for this region, as the model parameters are spatially extrapolated from across the Bering Strait from the US state of Alaska to the east. As a result, the calibration results are poor over the portion of the Russian Far East that is between the longitudes of -180 and -170.

Other areas where the NN model's calibration of EO data results in a degradation of the imbalance, rather than an improvement, include the southwest coast of South America in Chile, portions of China's Tarim Basin. These regions have unique climate conditions due to their geography, which the model appears to have limited capability to accurately represent. The model also fails to improve the imbalance in portions of Scotland, Iceland, and New Zealand. With the smaller island regions, the model has encountered a problem of extrapolation. I did not identify any river basins in our target size range (20,000 to 50,000 km^2) to use as training data. Thus, the model has not been trained to represent the conditions on these islands, and is extrapolating based on continental climates, sometimes thousands of kilometers away. Nevertheless, the areas where the imbalance is made worse is less than 2% of continental land surfaces.

5.4.5 Comparison to In Situ Observations

As an additional assessment of the calibration of EO variables, I compared the model output to in situ observations at ground-based stations. I performed this comparison to available data for 3 of the 4 water cycle components – precipitation, evapotranspiration, and runoff. In situ observations of the fourth component, ΔS, or total water storage change, are not directly measurable. In lieu of direct measurements, investigators have validated GRACE total water storage (TWS) by comparison with groundwater well levels (see e.g. Rodell et al., 2007) and land surface model simulations of soil moisture storage (see e.g.: Munier et al., 2012).

It is important to remember that the main purpose of the modeling was *not* to calibrate EO variables to improve the fit to observations. Data providers have already extensively calibrated data products to in situ observations. Indeed, some of these products have been continuously improved over decades. Thus, I do not expect that my "recalibration" will necessarily improve upon the original calibration, at least in terms of the fit to available ground-based observations. In fact, the goal is to make the datasets more coherent with one another, so that the water budget is balanced. Thus, comparing the datasets to in situ observations is simply an additional check. I performed this comparison both *before* and *after*. Ideally, the calibration will not introduce unacceptably large errors. I allow for the possibility that the modeling may degrade individual water cycle components somewhat. Nevertheless, the model-based calibration should make the EO datasets more coherent with one another.

The results of the comparison to in situ observations are summarized in Table 5.4. For assessing the fit to observation, correlation (R or R^2) would not be informative. The regression model applies a linear transformation to EO data, and the correlation with observations is the same before and after calibration. Therefore, it is more informative to look at alternative indicators such as the Nash-Sutcliffe Efficiency, NSE, described in Section 4.1.12. Percent bias (PBIAS, see Section 4.8 on page 140 is also an informative indicator, as it tells us how far the data are from the observed mean.

For the analysis of precipitation, I compared uncorrected EO estimates of P to observed precipitation at 21,880 stations in the Global Historical Climatology Network (GHCN, see Section 2.2.5). Then I compared the outputs of the calibration and mixture NN models to these same observations. I repeated the same procedure for E at 117 global flux towers, and for R, comparing NN predictions to discharge measured at gages (Section 2.5.1). I calculated fit statistics comparing the observed

Table 5.4: Evaluation of the NN model predictions for *P*, *E*, and *R*, showing the impact of the regression and neural network model calibration on the goodness of fit to observations. Table values are the median over *n* samples.

Dataset	NSE	RMSE, mm/mo	Percent Bias
Precipitation, at *n* = 21,880 stations			
GPCP (EO)	0.70	31.9	8.3%
GPCP (Regr. cal)	0.70	**31.7**	7.1%
GPCP (NN cal)	0.69	32.8	**3.9%**
GPM-IMERG (EO)	0.56	45.6	33.6%
GPM-IMERG (Reg. cal)	**0.76**	**27.2**	8.2%
GPM-IMERG (NN cal)	0.74	29.3	**6.4%**
MSWEP (EO)	**0.85**	**20.1**	**0.0%**
MSWEP (Regr. cal)	0.82	22.7	8.0%
MSWEP (NN cal)	0.77	25.9	4.3%
EO SW Mean	0.63	37.1	13.7%
Regr. cal. avg.	**0.78**	**25.9**	7.9%
NN mix.	0.75	28.0	4.5%
Evapotranspiration, at *n* = 117 flux towers			
GLEAM-A	0.65	21.4	3.6%
GLEAM-A (Regr. cal)	**0.70**	19.2	7.8%
GLEAM-A (NN cal)	0.69	**19.0**	**3.3%**
GLEAM-B	0.69	20.1	5.4%
GLEAM-B (Regr. cal)	0.67	19.9	**4.9%**
GLEAM-B (NN cal)	0.69	**18.5**	6.1%
ERA5	0.70	19.9	7.6%
ERA5 (Regr. cal)	0.68	20.9	7.8%
ERA5 (NN cal)	0.70	**19.4**	**6.0%**
EO SW mean	0.70	19.5	**3.9%**
Regr. cal. avg.	0.70	20.1	4.9%
NN Mixture Model	0.69	**19.4**	4.1%
Runoff, at *n* = 1,781 gages			
GRUN	0.52	12.7	−1.7%
GRUN (Regr. cal)	**0.53**	**12.4**	−**0.1%**
GRUN (NN cal)	0.50	12.7	−0.7%

and predicted time series at each measurement location. Table 5.4 reports the **median** of the fit statistic. For example, for E, I calculated 117 values of the correlation coefficient, R. For Gleam-A, the first row in the table, these values ranged from -0.11 to 0.98, with a median of 0.91. The models denoted by *Regr. cal.* have undergone calibration using the regression model, and rows labeled *NN cal.* have been calibrated by the NN model. Entries in bold text highlight the best value of each indicator for the variable.

Figure 5.18 shows the distribution of one fit statistic, for the three datasets and for the two modeling methods. We can see from this plot that both calibration methods have a slight positive impact on GPCP precipitation, decreasing the median RMSE. There is a stronger positive impact on GPM-IMERG precipitation, where both modeling methods reduce both the median and interquartile range of the RMSE. The only variable which is not positively impacted is MSWEP. This dataset, which combines information from remote sensing and reanalysis models, has already undergone extensive debiasing (Beck et al., 2019). Notably, the creators of MSWEP used the GHCN, among other sources, to debias the estimates of precipitation. This is the same dataset I am using here for validation. Therefore, it is likely that *any* modification to this dataset will make it depart further from the station data. With regards to the results that come from mixing multiple EO datasets, the NN mixture model and the mean of the 3 regression-calibrated results, both offer a significant improvement over the simple-weighted mean of the uncorrected EO datasets.

For E, the NN models appear to have generally *improved* the fit to observations collected at flux towers. The improvements are not large, and may not be important considering the caveats related to comparing point estimates to grid cell values. Nevertheless, it is a positive sign that our model does not degrade the signal, and in fact may be improving it.

The situation with discharge is largely reversed, and it appears that NN calibration is degrading the signal somewhat, albeit only slightly. Here, I calculated fit statistics against a set of gages with a strong runoff signal. From my original dataset of 2,506 gages, I excluded gages in arid regions where runoff is often at or near zero, leaving 1,781 gages. Geographically, the changes made to runoff data by the calibration, and fits to observations are not evenly distributed. Based on the change in RMSE, there is an improved fit to observations in 47% of basins, and a slight degradation of the fit in 53% of basins.

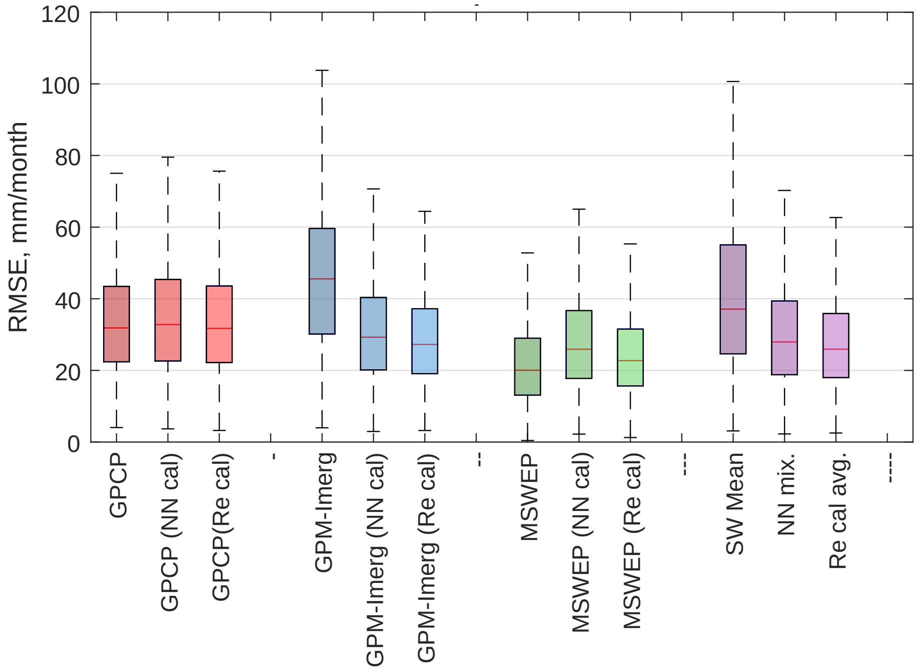


Figure 5.18: Boxplots summarizing the fits root mean square error between EO datasets and precipitation observations at 21,880 GHCN stations

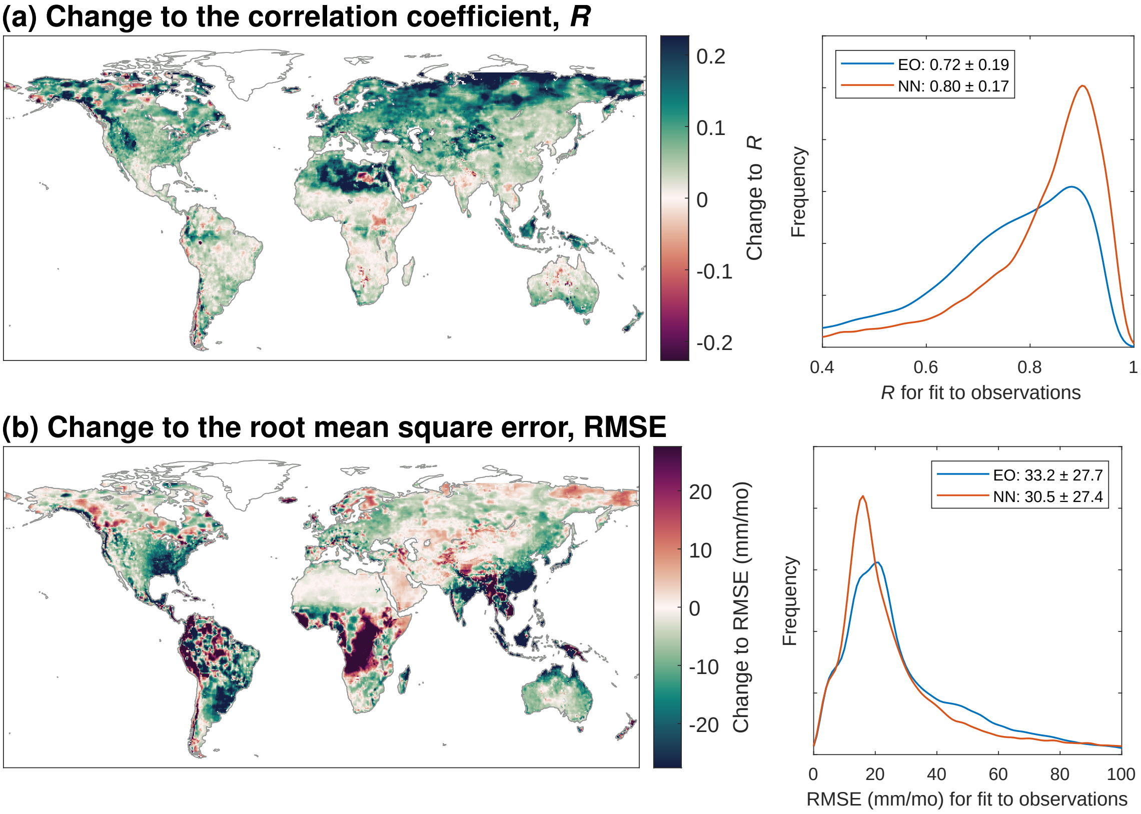
5.4.6 Comparison to a Gridded Precipitation Product

In the preceding section, I compared the EO datasets to in situ observations of P, E, and R. As an alternative to the comparison to station data (point locations), we may compare the results to gridded datasets, where a continuous coverage has been created based on interpolating station data. For precipitation, well-known datasets are WorldClim (see Appendix A) and CPC, described in Section 2.2.4. The data producers have relied on various algorithms and assumptions to interpolate station data across space and time. Using gridded data overcomes two problems with using the station data. First, it allows us to fill in empty places on the map. Second, it overcomes the sampling problem, i.e., where we have thousands of stations in Finland and Germany, and almost none in Africa. These datasets have been carefully compiled and are highly cited. Nevertheless, they are one step away from in situ observations and therefore have higher uncertainty. However, comparing the calibration model output to gridded precipitation data is an additional validation step for assessing the performance of the EO calibration model over global land surfaces.

Here, I focus on the CPC data, produced by the US NOAA Climate Prediction Center. The CPC dataset is usable right away, as it is in the same projection and spatial resolution as our calibrated EO datasets – geographic coordinates at

0.5° resolution. The results of the comparison to CPC precipitation are shown in Figure 5.19. This figure shows two fit indicators, *R* and RMSE. The kernel density plots at right show the distribution of fits across 54,509 pixels over land. Overall, the NN calibration results in a higher average correlation and lower RMSE at the pixel scale. The pixel-wise "delta" or change in the fit indicator is shown in the maps at the left in Figure 5.19, with green indicating an improvement, and red for a worsening of the fit. The NN calibration results in an increase in *R* over 90% of the pixels, and a decrease in RMSE in 66% of pixels.

We saw previously that the calibration of EO variables results in a lower water cycle imbalance. However, there was concern that in revising these variables to increase their coherency, it would have the undesirable side effect of degrading the signal. In other words, the model could force a trade-off, exchanging coherency for lower accuracy. However, as the above analyses have shown, the model-based calibration does not strongly degrade the fit to observations. In fact, in many circumstances, the fit to observations is actually improved.



(a) Change to the correlation coefficient, *R*

(b) Change to the root mean square error, RMSE

Figure 5.19: Effect of NN model calibration on the goodness of fit between EO precipitation and the CPC gridded precipitation data product. Maps and kernel density plots show the change in the fit of the simple-weighted (SW) ensemble mean of the 3 precipitation datasets before and after calibration by the NN model.

5.5 Chapter 5 Conclusions and Discussion

This chapter presented the results of two methods for the calibration of EO variables of the terrestrial water cycle. Both methods seek to recreate the solution from optimal interpolation over a set of over global river basins, and can be applied to make predictions at the pixel scale. In general, the NN model outperformed the regression-based model in terms of ability to recreate the OI solution and to close the water cycle. Table 5.5 summarizes the main points comparing the two main methods for water cycle closure explored in this thesis.

Table 5.5: Advantages and disadvantages of the two main methods for water cycle closure explored in this thesis.

Model Type	Advantages	Disadvantages
Regression + Parameter Regionalization	• Simple, understandable transformations of inputs. • Uses widely-used methods in earth science and hydrology	• Single-variable linear regression model may be too simple to describe complex relationships. • Requires little input data (only needs lat, lng coordinates of the training basin for extrapolation to new locations).
NN modeling	• Able to describe nonlinearities and complex relationships. • Environmental indices in the NN allows it to better localize, make predictions specific to different environments.	• “Black box” model lacks interpretability. • Requires a larger suite of input data.

The approach for optimizing water cycle variables used here contains certain innovative aspects. The regression method is paired with spatial interpolation for parameter regionalization. This is a useful method that is commonly used in hydrologic modeling, and I show here that it is effective for water cycle analyses with remote sensing data. Another advantage to the models used here is their modular structure – they can be used to calibrate one variable at a time. This is an advantage when confronted with with missing or incomplete data.

Optimized EO data that satisfies the closure constraint can also be used to develop more accurate water budgets. This has relevance "to the fields of agriculture, atmospheric studies, meteorology, climatology, ecology, limnology, mining, water supply, flood control, reservoir management, wetland studies, pollution control, and other areas of science, society, and industry" (Healy et al., 2007). Water-budget based methods can be used to estimate missing water cycle components. For example, we may estimate river discharge indirectly by rearranging Equation 1.1 as $R = P - E - \Delta S$. This indirect way of estimating water cycle components is also called the *water budget method*; I explore this in more detail in Chapter 6.

Chapter 6

Evaluation and Exploitation of the Calibrated EO Database

Previous chapters in this thesis described methods to optimize Earth Observations (EO) of the hydrologic cycle. These methods make adjustments to EO datasets so that they can be combined to create a balanced water budget. In the preceding chapter, I described how the models were applied at the pixel scale to create calibrated EO datasets with higher hydrological coherency than the uncorrected EO datasets.

In this chapter, I show how these calibrated EO data can be used to make improved predictions over ungaged or un-instrumented basins. For example, we may estimate runoff in ungaged basins or calculate total water storage change (TWSC, or ΔS) to fill in missing GRACE satellite observations. This opens up a number of possibilities for using the model-calibrated EO data. Water budget-based approaches can be used to estimate any one of the four components based on the other three, by rearranging the terms of the water balance equation, $P - E - \Delta S - R = 0$. This is an important application of the calibrated database. It can be seen too as an additional way of evaluating the calibrated data. As we will see below, many researchers have applied this method with remote sensing data, with varying degrees of success, as I will describe below.

The water budget approach is applied to estimate evapotranspiration, E, runoff, R, and total water storage change, TWSC or ΔS. (I did not attempt to calculate P by inference, as this variable is already calibrated to a large set of rain gages, and the model is unlikely to be useful for prediction.) My hypothesis was that estimating missing water cycle components via this method could be improved by using EO data that has been calibrated by the neural network (NN) model described in Chapters 4 and 5. I compared inferred predictions to observations and to the results from other modeling studies. This analysis can help to verify that the models are actually improving EO datasets. Are the predictions a better fit to observations? Are estimates an improvement over using uncorrected EO datasets?

The goodness of fit to observations is often modest, but I found that the fit is greatly improved when using NN-calibrated datasets rather than the original, uncorrected data. The significant improvement in predicting water cycle components with remote sensing data illustrates the usefulness and practical application of the methods described in this thesis.

6.1 Indirect Estimation of Evapotranspiration

Prior to GRACE, hydrologists used water budgets as one method of estimating long-term evapotranspiration over river basins. To do so, they assume that there is no trend in total water storage. When this is the case, then $\bar{E} = \bar{P} - \bar{R}$ over sufficiently long time scales (Lopez et al., 2015). A common practice in the northern hemisphere is to perform such calculations over a *water year*, from October 1 - September 30. However, observations from GRACE have shown that water storage in many regions is dynamic, and can vary significantly over annual and decadal time scales (Rodell et al., 2018). Therefore, the assumption that basin water storage is constant over long time scales appears to be invalid more often than previously thought (see Section 2.7.2 on page 85). Several recent studies have made use of GRACE data and the water-budget method to estimate evapotranspiration, using $E = P - \Delta S - R$.

Rodell et al. (2011) estimated E over seven large river basins and compared predictions to the output of several land surface and atmospheric models. They concluded that the uncertainty in GRACE ΔS is too high to produce useful monthly estimates, but that the method produces viable annual estimates of E.

Long et al. (2014) estimated E over river basins in Texas using GRACE data, observed runoff, and a variety of data sources for precipitation. The fit of predicted E was modest, with R^2 ranging from 0.21 to 0.64. Pascolini-Campbell et al. (2020) estimated monthly basin-scale E over 11 major river basins in the contiguous United States. The authors compared the results of what they called the “mass conservation ET estimate” to E from remote sensing data products and land surface models. They found that using this method and GRACE data consistently reproduced the seasonal pattern of E, but also resulted in higher estimates of E compared to other data sources. Yet, because this study did not include comparison to in situ observations, it is difficult to say which method is the best fit to observations, and hence the most accurate.

I calculated E by the water budget method with uncorrected EO datasets, then repeated the analysis with NN-calibrated EO datasets. I compared the results to E observed at 117 flux towers across the globe. The results of this analysis are shown in Figure 6.1.

I computed fit statistics comparing the time series of E observed at the tower

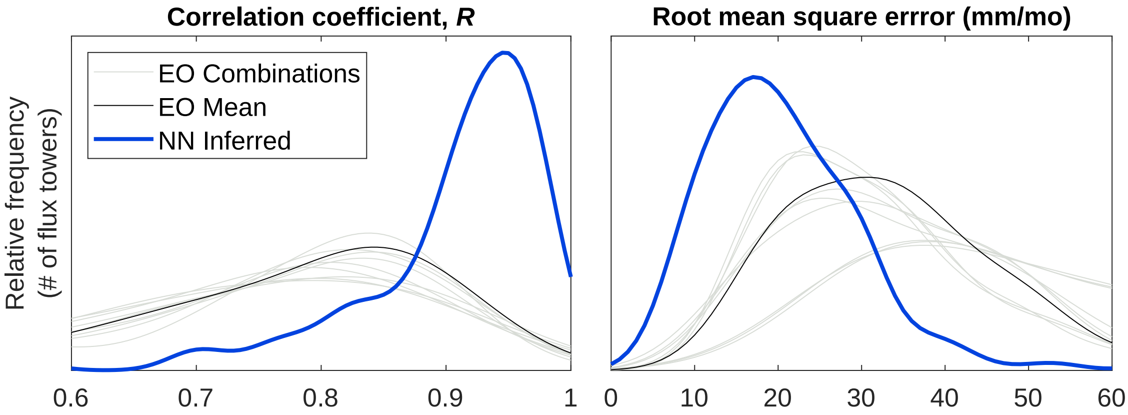


Figure 6.1: Empirical probability distribution plots of the correlation (left) and RMS error (right) between EO-based estimates of basin evapotranspiration and in situ observations at 117 flux towers.

to the time series from the corresponding grid cell in the calibrated EO data layer of E, and report the results in Table 6.1. Calculating E from *uncorrected* EO data results in a relatively poor fit. Among the 27 possible combinations, shown in gray on Figure 6.1, the correlation, R, has a mean and standard deviation of 0.64 ± 0.31. The quality of predictions varies depending on which EO database is used as input. Simply averaging multiple datasets has a slight positive effect. The NN calibration further helps improve the fit. Using the NN-corrected EO data to compute E improves the fit, with average $R = 0.87$.

For the sake of comparison I have also included direct estimates of E in Table 6.1. These appear at the top, under the heading **EO datasets**. This allows us to see how the water-budget method compares to direct estimates by remote sensing. When we use uncorrected EO data as inputs, it is much less accurate to estimate evapotranspiration by $E = P - \Delta S - R$. However, after NN calibration, the quality of water budget-based estimates rivals GLEAM or ERA5.

So, the calibration the EO datasets with the NN model allows us to make much more accurate predictions of E with the water budget method, compared to using uncorrected EO data. This shows that the NN-optimization of the water components P, ΔS and R makes them closer to the E in situ measurements. The results appear to be just as good as those obtained with current state-of-the-art remote sensing datasets. They also appear to be better than results reported in several recent studies cited above.

Table 6.1: Goodness of fit to evapotranspiration estimated by various methods, compared to observations at at 117 flux towers. Table entries are the median for the fit statistic over the sample. For EO combinations, table reports the median of the medians.

	Corr. *R*	RMSE mm/mo	Bias mm/mo
EO datasets			
GLEAM-A	0.91	21	2.0
GLEAM-B	0.93	20	2.5
ERA5	0.91	20	3.9
NN mixture model (this study)	0.92	19	2.1
***E* estimated indirectly using**			
EO Combinations (n=27)	0.75	34	8.1
EO Mean	0.78	32	11
NN calibrated EO (this study)	**0.92**	**19**	**0.3**

6.2 Indirect Estimation of Total Water Storage Change

We have seen how observations of TWS from GRACE contributed to more holistic study of water cycle. Previously, water storage could only be inferred or estimated indirectly. We have also seen that GRACE data has significant gaps (Section 2.4.1). There is also interest in reconstructing GRACE-like total water storage for periods prior to 2002, using a variety of methods.

Landerer and Swenson (2012) discuss the difficulty in comparing GRACE observations to the results of simulation models. GRACE TWS does not map directly to a state variables in land surface models, which many not fully simulate groundwater, glaciers, etc. or their simultation may be unrealistic “due to missing model physics” (Landerer & Swenson, 2012). Nevertheless, studies have found that GRACE estimates of TWSC are correlated with observed groundwater surface elevation changes (Rodell et al., 2018), soil moisture estimated by a land surface model (Scanlon et al., 2019), surface water extent (Papa et al., 2008), and reservoir volume (X. Wang et al., 2011).

Scanlon et al. (2019) assessed the correlation between GRACE observations and modeled water storage over 183 global river basins, using data from 7 global hydrologic and land surface models. For most of the models they considered, the researchers used the model estimates of soil moisture as a proxy for TWS. The authors concluded that discrepancies between observations and simulations are partly due to missing storage compartments in models (e.g., surface water and/or groundwater). In one of the more thorough analyses conducted to date, Biancamaria et al. (2019) compared observed TWS anomaly to modeled water storage from two hydrologic models in the Garonne river basin in France, and found correlation coefficients of around 0.9 and Nash-Sutcliffe Efficiency of around 0.7.

In another example, Lehmann et al. (2022) estimated ΔS by the water budget method over 189 large river basins, and compared predictions to GRACE observations. Rather than seeking to optimize the datasets, the authors looked for the best combination of inputs. The authors deemed their method successful because the Nash-Sutcliffe Efficiency, NSE > 0 in the majority of basins, which means that the model performed better than a constant at the mean of observations. This modest performance underscores the difficulty of estimating TWS based on other, unrelated, remote sensing observations.

Due to the lack of in situ data, I also evaluated the results of my NN model against results from other studies that predicted GRACE-like total water storage change using different methods, including those that were more sophisticated than the simple water balance method used here. Pan et al. (2012) estimated the water budget components using satellite observations in 32 globally distributed major basins for 1984–2006. Their approach used data assimilation techniques, first estimating the errors in each water budget component by comparison to in situ observations, then using a constrained Kalman filter to merge the datasets based on their error information, with a goal of minimizing the imbalance. Y. Zhang et al. (2018) employed a similar method at the pixel scale, rather than at the scale of the river basin. The authors concluded that the imbalance error is mainly due to disagreement among evapotranspiration estimates.

I obtained the results from Pan et al. (2012) by request to the author, and downloaded the data from Y. Zhang et al. (2018). I used geodata for Pan's 32 large basins (basin masks on a 1° grid) to calculate the spatial-averaged means for changes in storage over these basins. Because Zhang et al. produced global gridded estimates of TWSC, I could compare the results to GRACE observed TWSC at the pixel scale. For my NN model and Zhang's model, I averaged the estimated TWSC over the 32 large river basins used in Pan's study. Pan et al. (2012) includes data for the years 2000 - 2006, while Y. Zhang et al. (2018) covered 1984 - 2010.

Overall, ΔS predicted by the water-budget method using NN-calibrated EO data was a better fit to GRACE observed ΔS, according to two common goodness-of-fit measures (Figure 6.2). On these plots, the blue points represent the fit indicator in one basin, and the red line is the median. I compared the model fit to the simple-weighted average of the three GRACE solutions for TWS. These results are also reported in Table 6.2

Over these 32 large basins, Pan's model had a median correlation coefficient $R = 0.86$, compared to Zhang's $R = 0.90$, and $R = 0.94$ for my model. Pan's model had a median root mean square error, $RMSE = 12.0$, compared to $RMSE = 10.2$ for Zhang's model, and $RMSE = 8.0$ for my model. Thus, in these large river basins, my NN model is a better fit to GRACE observed TWSC. The comparison may not be entirely fair, as I have calibrated my NN model using recently published versions of GRACE, while Pan's model was calibrated to an older version of GRACE available in 2012.

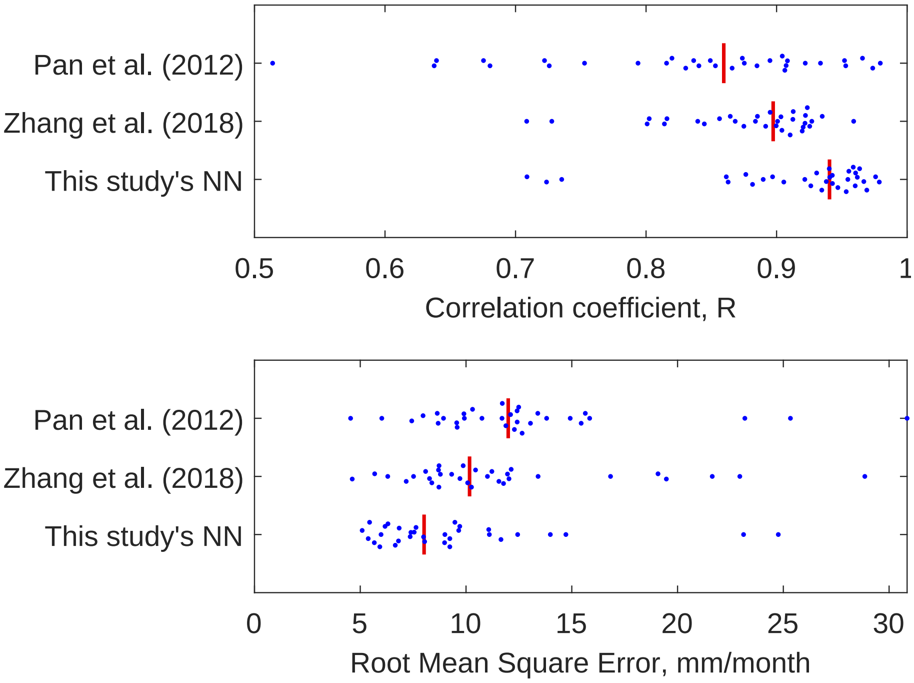


Figure 6.2: Goodness of fit between GRACE observed and modeled monthly TWSC inferred from my NN model predictions, and two recent assimilation model-based studies, over 32 large river basins. The vertical red line is the median of the 32 data points.

Figure 6.3 shows the fit to observed TWSC by my neural network model and the predictions by Y. Zhang et al. (2018). The map shows that the geographic patterns are similar, in terms of where the models produce better or worse fits to observations. Overall, my model has a slightly higher median correlation with observations, and a slightly lower root mean square error. However, my model performs poorly in more geographic areas.

The results in Table 6.2 show that my model's indirect estimates of ΔS are equivalent to the predictions by Y. Zhang et al. (2018), based on the fit to GRACE observations at the pixel scale over the overlapping time period 2002 to 2009. It

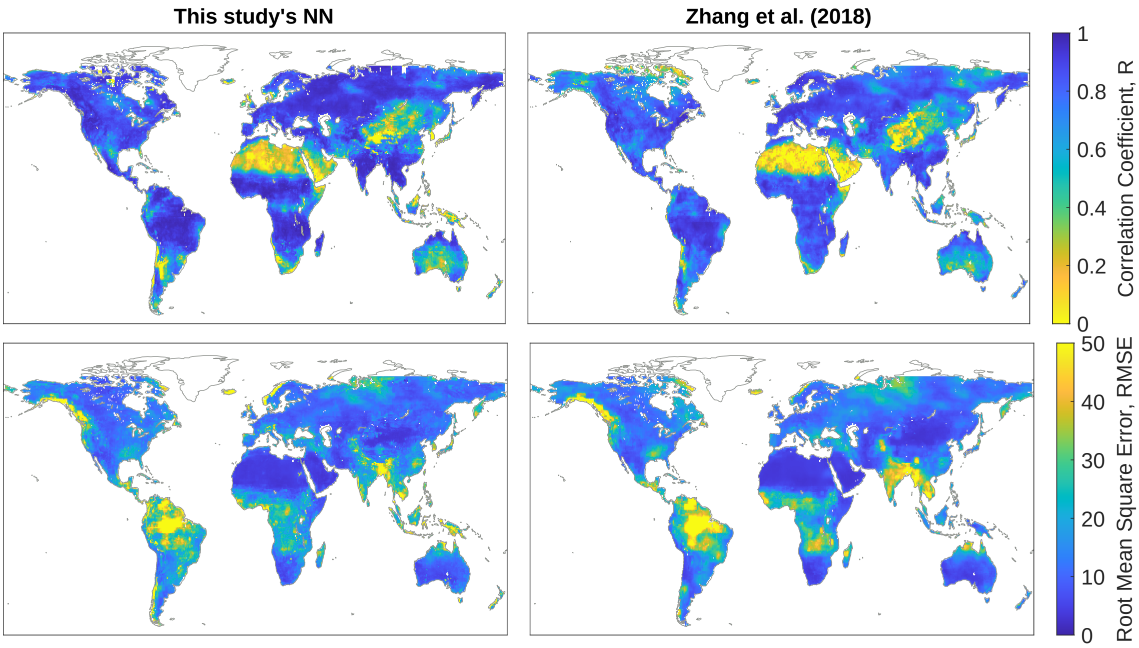


Figure 6.3: Maps of the correlation and root mean square error for predictions of TWSC from two sources: inferred by my NN predictions, and Zhang et al. (2018).

was not among the main goals of this research to predict TWSC. Nevertheless, my NN model is able to do so nearly as well as a state-of-the-art model.

Table 6.2: Goodness of fit to GRACE observations for total water storage change estimated indirectly by the water-budget method, in 57,286 land pixels. Table entries are the median for the fit statistic over the sample. For EO combinations, table reports the median of the medians.

TWSC, ΔS, estimated indirectly by	Corr. R	RMSE mm/mo	Bias mm/mo
EO combinations (n=9)	0.71	24	3.3
EO Mean	0.75	22	6.4
Zhang et al. (2018)	0.79	13	**0.09**
NN calibrated EO (this study)	**0.84**	13	0.18

At the pixel scale, the results of my NN predicted ΔS compare favorably to those predicted by Zhang. Figure 6.4 shows the empirical probability distribution for two fit indicators over all land pixels. The average correlation for Zhang is $R = 0.70$, while for my model, $R = 0.74$. My NN model's median correlation is slightly higher, with median $R = 0.84$ vs. Zhang's median $R = 0.79$.

A limitation of the water-budget method for estimating TWSC is that its inputs are hydroclimatic variables only. It does not include information on human influence on the water cycle, such as groundwater pumping, irrigation, withdrawals, or

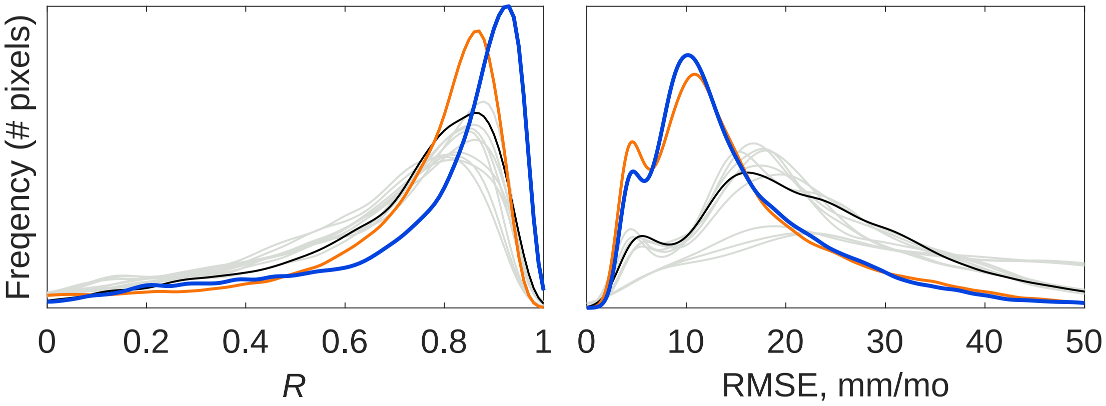


Figure 6.4: Empirical probability distribution plots of the correlation (left) and RMS error (right) between in situ observations and indirect EO-based estimates of basin-scale total water storage change.

interbasin transfers. Because of this, these methods will be less accurate in zones with extensive human impacts. In zones without anthropogenic influences, the results may help show how water storage responds to climate and meteorological forcing.

6.2.1 Correlation between TWSC and ENSO

Using the methods above, I reconstructed a signal of Total Water Storage Change, ΔS for the period 1982 - 2019, which includes 20 years prior to the launch of the GRACE satellites. In the section above, I showed that the reconstruction is a reasonably good fit to observations. It is interesting to examine the relationship between water storage and other climate variables. Since the late 19th century, scientists have "teleconnections" in weather and climate – the relationships or links between phenomena at widely separated locations of the globe.

Studies have used various methods to identify and analyze teleconnections in hydrology. For example, Martens et al. (2018) highlighted the need to consider teleconnections to accurately predict the fate of the terrestrial branch of the hydrological cycle. They used observational evidence to improve the representation of surface fluxes in Earth system models. Similarly, Rasouli et al. (2020) conducted variance, correlation, and singular spectrum analyses to identify hydroclimatic phases related to teleconnection patterns in a small headwater basin in Idaho, USA. Their study linked hydrological variations at local scales to regional climate teleconnection patterns.

With a nearly 40-year reconstruction of ΔS, it is interesting to analyze the relationship between water storage and well-known climate patterns. The El Niño Southern Oscillation (ENSO) is a quasi-periodic climate pattern that characterizes the warming and cooling of surface waters in the eastern tropical Pacific Ocean (El Niño) and its effect on air pressure across the equatorial Pacific Ocean (Southern Oscillation). During El Niño events, there is a warming of the ocean surface temperatures in the central and eastern Pacific, causing changes in atmospheric circulation and rainfall patterns experienced across the globe (Guimarães Nobre et al., 2019).

The ENSO cycle consists of two phases: El Niño and La Niña. For example, in South America, El Niño, there is often increased rainfall along the west coast of South America, leading to increased flooding and landslides. In North America, El Niño can bring above average precipitation in the southern United States, and drier than average conditions to parts of the Pacific Northwest. In Australia and Southeast Asia, El Niño is linked to reduced rainfall and drought. It is also linked to increased rainfall in parts of eastern Africa, while in India, El Niño is associated with reduced precipitation.

I downloaded ENSO indices from the NOAA (2023) and analyzed the correlation with my reconstructed 38-year dataset of TWSC. There are several different ENSO indices available, based on different variables, and calculated over different regions of the Pacific Ocean. I based the analysis here on the Multivariate El Niño/Southern Oscillation (ENSO) index (MEI.v2) index, which combines "five different variables (sea level pressure (SLP), sea surface temperature (SST), zonal and meridional components of the surface wind, and outgoing longwave radiation (OLR)) over the tropical Pacific basin (30°S to 30°N and 100°E to 70°W)."

The relationship we see here is consistent with what we know about how El Niño events can affect weather patterns and the water cycle in different parts of the world. Figure 6.5 shows the correlation between TWSC with the ENSO index MEIv2 over South America. The temporal behavior is also coherent with correlations positive or negative up to 0.5. The interpretation is that ENSO can explain up to 25% of the variability in the monthly TWSC.

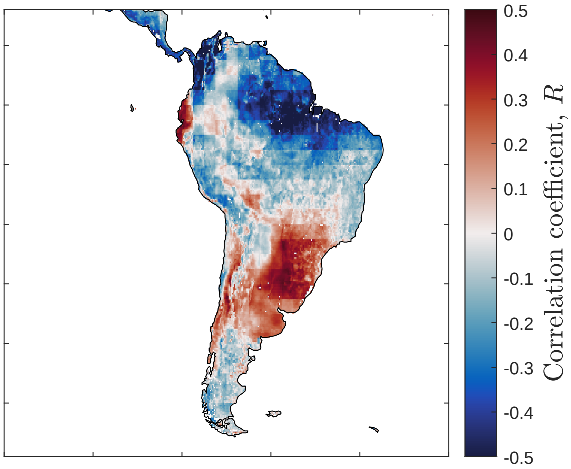


Figure 6.5: Correlation between NN calibrated ΔS and the ENSO index MEIv2, for 1980 - 2019, at the pixel scale over South America.

6.2.2 Estimating trends in Total Water Storage

One interesting application of the extended reconstruction of TWSC is to analyze trends. Reconstructing the signal of TWSC was not my main goal. However, combining the water budget method with NN-calibrated EO data allows us to reconstruct the signal of ΔS about as well as a state-of-the-art assimilation model. However, challenges remain in terms of reconstructing TWS. The variable predicted by my model, ΔS, is a change in the volume of water stored over time, in units of mm/month. Water managers are usually more interested in the time integral of this rate, or the total water storage, TWS. When I calculated this quantity by integrating ΔS, I found that even relatively small errors or biases in the signal of TWSC are compounded when calculating the integral, causing unacceptably large uncertainties.

It can be shown that any slight inaccuracies or biases are magnified when integrating to determine S. Thus, the estimates of the trend are highly uncertain. I do not believe that any of the current published reconstructions of GRACE-like water storage are sufficiently accurate to reliably estimate trends in water storage in the pre-GRACE era. This makes it difficult to reliably calculate trends in TWS given a reconstructed signal of TWSC. Thus, it is possible to use my NN model, or the other models described above, for *hindcasting*, or to make predictions of TWSC for before the GRACE satellites were launched in 2002. We can adequately reproduce the seasonal pattern, but it is impossible to accurately predict trends. Therefore, the usefulness of these predictions is severely limited.

For future research, there are certain strategies which may help to produce more robust reconstructions of TWSC, which could help improve the estimation of trends. One strategy is to combine both observed and modeled time series of ΔS to estimate the trend, using Bayesian estimation or Kalman filters. One could also try debiasing modeling results to fit the observations before using them to estimate the trend, e.g., with CDF matching, or quantile-quantile bias correction. However, I believe that trends in water storage estimated using climate data will always be suspect if they do not include human influences such as diversions and withdrawals, which have a major impact on the water cycle in many locations.

6.3 Indirect Estimation of Runoff

Considering the decline in river discharge monitoring in recent decades, alternative methods of estimating runoff are becoming more important. As we have done above with E and ΔS, we may use the water budget approach to predict runoff, R, from the other three water cycle components. Lorenz et al. (2014) refers to this method of estimating basin-scale runoff as the "hydrologic approach." One of the main advantages of this approach is that it "does not require runoff routing as it is taken care of by the water storage changes."

Researchers have largely been unsuccessful in trying to estimate river discharge using such water-budget based methods. Frequently, the magnitude of runoff is small compared to the other components, making the signal-to-noise ratio low. And, as noted by Lorenz et al. (2014), "the accuracy of the runoff estimates will be only as good as the least accurate dataset." Despite these difficulties, predicting basin runoff by indirect methods is a compelling topic of research. Developing a new, accurate method for prediction in ungaged basins would be considered a major breakthrough. Such predictions would be highly valuable in regions with limited measurement infrastructure, with potential applications in agricultural water management, drought and famine prediction, or predicting the impacts of climate change on future runoff (Chiew, 2010).

Some authors have made the simplifying assumption that, over sufficiently long time periods, $\Delta S = 0$ (i.e. no trend in storage), allowing one to estimate long-term average runoff as $R = P - E$ (Y. Liu et al., 2020). There are several recent studies where the authors use data from GRACE to provide this information. Syed et al. (2005) used GRACE data and P and E from a reanalysis model to estimate discharge over the Mississippi and Amazon basins. Overall, the fit was poor, and prediction errors were high. Nevertheless, the authors expressed confidence in the method, and hypothesized that observed discharge at gages may not adequately capture the flux out of the basin, which may be exiting via subsurface flow and other "unmonitored surface fluxes."

Sheffield et al. (2009) used GRACE and other remote sensing data to impute the discharge from the Mississippi River basin. Their results were fairly poor – the 95% confidence interval for estimated discharge in certain months ranged from -3 to +3 mm/day, equivalent to a range of $-100{,}000$ to +100,000 m^3/s. (The mean discharge of the Mississippi at Vicksburg is around 16,000 m^3/s.)

Gao et al. (2010) used the water budget method to infer runoff over 9 large river basins in the continental United States. The predictions also appear to be rather poor; while the authors do not report any fit statistics, simply noting that errors are "generally quite large, especially during the warm season."

Lorenz et al. (2014) used EO data to predict runoff over 96 global river basins ranging in sizes from 20,000 km² to 4 million km² (the Amazon). The authors concluded concluded that "the budget-based approaches do not provide realistic runoff estimates because of significant biases in the input datasets." The water budget-based model performed worse in river basins where the flows are low or nearly constant. Sneeuw et al. (2014) attempted to estimate river discharge using the same method, which they called "the hydrogeodetic approach," over 5 large river basins. The results were again relatively poor, with NSE > 0 (meaning the model outperforms the mean, $f(x) = \bar{x}$) in only 1 of the 5 basins. The authors expressed the hope that better runoff predictions would be possible in the future after "improvements in the quality of global hydrological and hydro-meteorological datasets."

J. Chen et al. (2020) estimated river runoff in the Amazon basin using data from GRACE, ERA5 reanalysis data, and satellite precipitation observations for 2003 through 2015. Their water-budget based estimates of runoff exceeded observations by about 30%. The authors speculate that there is a significant subsurface runoff flux that contributes in part to this difference. Indeed, there is evidence of significant groundwater flow in the aquifer beneath the Amazon, equivalent to around 3% of river flow (Pimentel & Hamza, 2011). However, this subsurface flow only accounts for about $1/3$ of the difference between observed flow and estimates by J. Chen et al. (2020).

In a more recent paper, Abolafia-Rosenzweig et al. (2021) predicted discharge over 24 global basins combining remote sensing and in situ observations. The authors concluded that they were not able to accurately predict discharge (R^2 ranged from 0.42 to 0.47), concluding that the uncertainties in other water budget components are " generally larger than the magnitude of Q [discharge] itself." However, the authors also found that adding the water cycle closure constraint contributed to improved predictions of discharge.

For this analysis, I calculated R indirectly using the three NN-calibrated water cycle components. The output is a gridded data layer of R at the pixel scale. We may then compute the spatial average to estimate river discharge in small- to mid-size river basins. Figure 6.6 and Table 6.3 compares the fit to observations of runoff calculated by inference from uncorrected remote sensing datasets, and by the calibrated EO data output by my NN model. The NN-based result is a significant improvement over using uncorrected EO data.

The uncertainty in runoff estimated by the water budget method is too high to consider this a reliable estimator of discharge in un-gaged basins. This is a signal-to-noise ratio issue. Runoff tends to be much smaller in magnitude than the other three water cycle components. However, the coherency between the WC components has been improved by the NN framework.

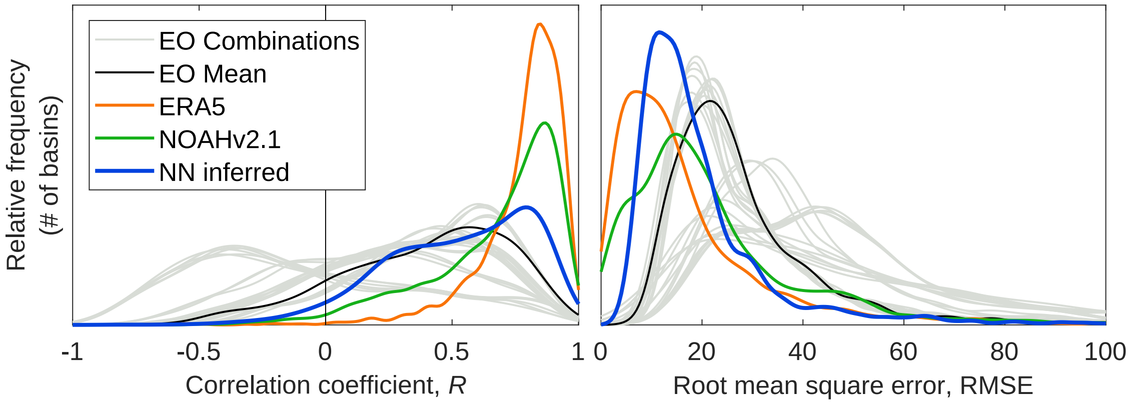


Figure 6.6: Empirical probability distribution plots of the correlation (left) and RMS error (right) between in situ observations and EO-based estimates of basin runoff.

In Figure 6.6, showing the fit to observations for runoff predicted by inference, there is a sub-ensemble in gray with first mode around -0.5. These lines have all been calculated with GLEAM-A. evapotranspiration dataset. Because use of this particular dataset tends to result in poorer predictions of observed runoff, one may choose to discard it in future water cycle analyses. Indeed, the data provider publishes different versions of GLEAM, as described in Section 2.3.2. GLEAM-A uses meteorologic inputs from reanalysis modeling, while GLEAM-B relies more on remote sensing data. In this context, predicting runoff via the water-balance method, GLEAM-A yields larger errors, so preference should be given to GLEAM-B.

Based on the results in Table 6.3, we can see estimated runoff using NN-calibrated EO data has a lower bias error than estimates made with uncorrected EO data. A summary of the percent bias errors over 1,781 river basins is shown in Figure 6.7. Recall that the bias measures the distance between the mean of observations and the mean of predictions. The percent bias is the percentage difference between the means of observations and predictions.

We saw in Figure 6.6 that inferences of *R* via the water balance that use the NN-calibrated data are significantly improved compared to using uncorrected EO datasets. Nevertheless, the accuracy of these predictions is still modest, and errors may be too high for many applications. Further, predictions of runoff

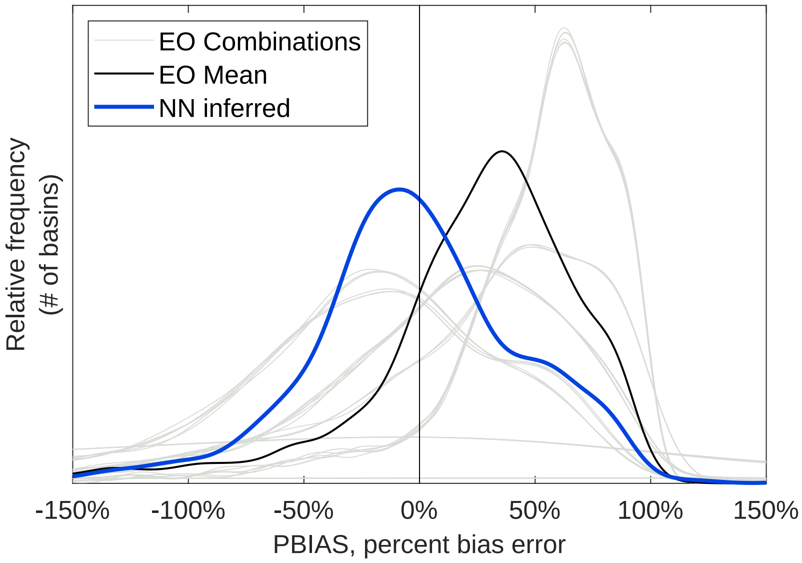


Figure 6.7: Distribution of the percent bias error in predicted runoff over 1,781 river basins. Predicted runoff was estimated indirectly by the water-budget method using EO datasets, before and after calibration by the neural network model.

via a monthly water-balance model would not be suitable for all applications. For example, flood warning would typically require hourly or at least sub-daily temporal resolution. However, such estimates could be useful in agricultural water management or famine early warning systems. I also investigated whether the NN calibration improves estimates of the long term mean of R. Even when the RMSE is too high, information on the mean runoff is still valuable information for prediction over ungaged basins. At least we may say that it provides a first-order estimate of runoff and river discharge.

Based on the simple-weighted average EO data, runoff estimates had a median bias error of -20 mm/month, compared to -3 mm/month using NN-calibrated data. For the sake of this analysis, let us suppose that estimates of discharge are adequate when the absolute value bias error is less than 50% (i.e.: the prediction is within 50% of the truth, regardless of whether the estimate is too high or too low). With uncorrected EO data, the estimated runoff had a bias error less than 50% in 1,022 out of 1,781 basins, or 57% of the time. After NN calibration of EO data, the number of basins where $|\text{PBIAS}| < 50\%$ increases to 1,261, or 71% of the total. Based on these statistics and Figure 6.7, we can see that NN calibration leads to a significant improvement of estimates of runoff made via the water-budget method.

Table 6.3: Goodness of fit between runoff estimated indirectly by the water-budget method and observed river discharge at 1,781 river gages. Table entries are the median for the fit statistic over the sample. For EO combinations, table reports the median of the medians.

Water cycle component	Corr. *R*	RMSE mm/mo	Bias mm/mo
Modeled runoff			
GRUN	0.45	21	8.2
ERA5	0.83	13	0.5
NOAHv2.1	0.76	18	1.6
Runoff estimated indirectly with			
EO Combinations (n=27)	0.29	31	2.4
EO Mean	0.45	24	8.2
NN calibrated EO (this study)	**0.57**	**16**	**1.0**

6.3.1 Estimating Discharge in Large River Basins via the Water Budget Method

As described above, several studies have used water-budget based methods to estimate discharge in large river basins, such as the Amazon and the Mississippi. We saw above that using NN calibrated EO datasets resulted in significant improvements in runoff prediction, compared to using uncorrected EO datasets. I tested the water budget method's ability to predict discharge in the Mississippi River basin, comparing the results to observations (USGS gage 07289000 at Vicksburg). Figure 6.8 shows the time series (left) and monthly average ± standard deviation (right) for predictions and observations. As can be seen with the light gray lines, *R* estimated with various combinations of EO datasets varies widely, and is often wildly inaccurate. Discharge estimated with the simple-weighted mean of EO datasets appears to be unbiased during the months of December through May, but exhibits a significant high bias from June to November. Calibrating the EO datasets with the NN model results in improved predictions of basin runoff over the Mississippi.

I believe that this result is better than the results in several of the papers cited above that predicted flows in the Mississippi using water-budget based methods. It is hard to say this definitively, as some of these papers describe their results qualitatively, without reporting fit statistics. It is also worth noting that my results are for a longer time period, from 2002 to 2019, with gaps where GRACE data are missing. Table 6.4 reports several fit statistics comparing predicted discharge with observations, for pre- and post-calibrated EO datasets. Overall, this method is able to predict annual mean discharge quite well, as evidenced by a low bias. The seasonal pattern is also mimicked with good accuracy. However, in terms of a predictive model, this method is not very strong. A Nash-Sutcliffe Efficiency (NSE) near zero means that a constant model equal to the long term mean performs equally well. The Cyclostationary NSE removes the seasonal cycle prior to estimating the goodness of fit. A CNSE < 0 indicates that this method does not do a good job predicting the anomalies.

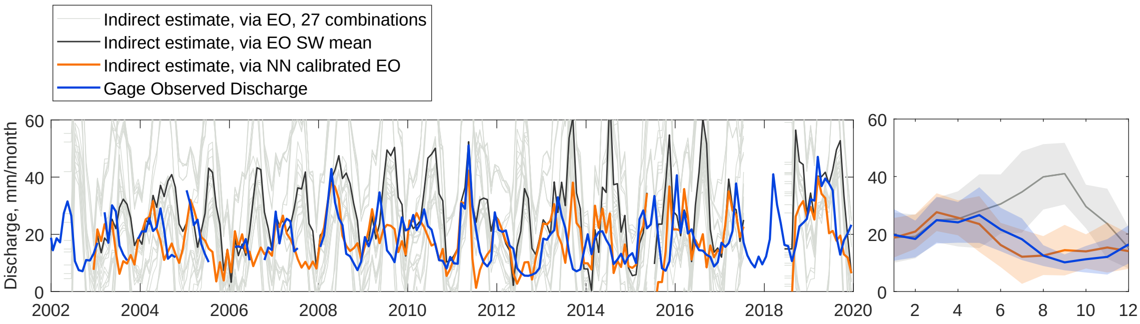


Figure 6.8: Time series plot and seasonality for monthly runoff for the Mississippi River at Vicksburg calculated from EO datasets, pre- and post-calibration by the NN model.

Table 6.4: Fit statistics for monthly runoff for the Mississippi River at Vicksburg calculated from EO datasets, pre- and post-calibration by the NN model.

	Pre-calibration	Post-calibration
Bias, mm/month	8.6	-0.6
RMSE, mm/month	16.8	8.6
Correlation, R	0.12	0.53
KGE	-0.1	0.53
NSE	-2.8	0.017
CNSE	-4.9	-0.55

6.4 Chapter 6 Conclusions and Discussion

In this chapter, I applied water-budget based methods for estimating missing water cycle components. With this method, we are solving for one unknown when we have three known variables in the equation $P - E - \Delta S - R = 0$. This method has been widely used in research and by practitioners.

For evapotranspiration, water-budget based methods predict observed E at flux towers as well as state-of-the-art methods based on remote sensing and more complex models. Indirect water-budget based methods can be used to reconstruct historic TWSC from 1982 to present. However, these results appear to be of relatively low quality. Time series are correlated with climate factors like El Niño, but should not be relied on to estimate trends. Predictions of runoff, while improved, cannot compete with land surface models in terms of predicting river discharge. Overall, this is an important extension of this research, and also an additional way of evaluating the results.

The quality of water-budget based estimates for a missing component is dependent on several factors, including the uncertainty of the 3 inputs variables and the signal to noise ratio. It is, for example, difficult to estimate discharge in basins where it is much less than the precipitation. Yet, the estimation of water cycle components is significantly improved compared to using uncorrected EO data. This is further evidence that the calibrated EO database has greater coherency and better describes the overall water cycle.

Chapter 7

Conclusions

In this final chapter, I first summarize the findings of this research and highlight its implications and significance. In the Perspectives section, I offer my recommendations based on these findings, including directions for future research.

7.1 Summary and Significance of Findings

This research explored methods of analyzing the global water cycle with remote sensing datasets. My goal was to reconcile these data to *close the water cycle*, or to reduce the overall error in estimating the water budget. This work built upon previous research and also contained several innovative aspects.

I applied a closed-form analytical solution, optimal interpolation (OI), that forces the water budget residual to equal zero. This approach has several advantages – it is straightforward to implement and has a basis in information theory, as it allocates errors in observations inversely proportional to their uncertainty. Compared to previous applications of the OI method, I applied it on a much larger scale, using over 1,600 river basins that I delineated based on topography, on every continent other than Greenland and Antarctica. Unlike with prior uses of OI, I used an affine error model that produces more realistic results under a range of hydrologic regimes. Yet, despite its advantages, OI can only be applied at the river basin scale where discharge observations are available.

I explored two methods for extrapolating the results of optimal interpolation to make predictions in ungaged basins and at the pixel scale. The first method involved fitting simple linear regression models over training basins, and then using spatial interpolation methods to evaluate model parameters over all global land surfaces. Second, I trained a nested set of neural network (NN) models to reproduce the results of OI. The NN models are able to ingest a large amount of information, and to find complex and non-linear relationships among variables (i.e.: remote sensing observations and carefully chosen environmental data).

The NN models outperformed simpler methods in terms of both fit to the OI solution and in terms of water budget closure. The model goodness of fit varies by location; it tends to be better over humid regions, and less accurate over the Arctic or over parts of Asia and South America. I also applied the NN model at the pixel scale and showed that the solution results in lower water cycle imbalance errors over most of the earth's land surfaces.

There are several potential applications for the resulting calibrated earth observation (EO) datasets. I explored the use of water budget-based methods for predicting missing water cycle components, useful for uninstrumented regions. This method predicts evapotranspiration observed at flux towers on every continent as well as a state-of-the art remote sensing dataset. The NN model's inferred predictions of runoff are a significant improvement over R calculated via raw EO datasets, but slightly less accurate than a statistical model calibrated to gage observations over specific river basins.

I also explored the capability of using the water budget method with calibrated EO data to fill in missing observations of total water storage change (TWSC) from 1980 to 2002, before GRACE observations are available. I showed that this method can predict TWSC as well as a state-of-the-art global assimilation model. These results are informative, as they reasonably recreate the seasonal runoff signal and interannual variability. However, greater accuracy is needed to predict trends in total water storage. I do not believe that any of the methods described in the literature can reliably recreate TWSC from climate data alone. One cause is that small bias errors in total water storage change are compounded when integrating to calculate trends in TWS. Another reason is that water storage is profoundly impacted by human activities, which are not well monitored.

Overall, the methods discussed in this thesis are effective at reconciling remote sensing data and improving water cycle closure. Yet, the methods and outputs do have certain limitations. The results have fairly coarse spatial and temporal resolutions (monthly, 0.5°). They provide a global view, or one that can be applied over large basins, but may not be suitable for highly local applications. The analysis is not quite global as I did not analyze Antarctica, Greenland, or Arctic regions above 73° North.

Regions with permanent snow and ice defy conventional hydrologic analysis – alternative techniques and methods are needed for studying cold regions. My analysis covers a longer time period than other recently-published global water cycle studies, but still only covers from the launch of GRACE in 2002 to 2019. I showed how the methods described herein can extend the calibration of EO datasets to previous time periods, but I also showed that extrapolating this information to estimate trends is risky and unreliable.

One of the difficulties in the record extension of total water storage anomalies is that it is not solely a function of climate variables – it is highly affected by human activities, such as water diversions and groundwater withdrawals. A better understanding of how these activities affect the water cycle would require detailed modeling that incorporates information on human activities, e.g., population densities, dam construction, reservoir levels and operations, and irrigation intensity.

7.2 Perspectives

In this section, I offer some recommendations based on my research, including potential directions for future investigation.

The availability of river discharge observations is an important limiting factor for large-scale hydrologic analysis. In the near future, data from the SWOT satellites will create exciting opportunities for similar lines of research, providing runoff estimates at many more locations than are currently monitored. After a few years of SWOT data have been acquired, the methods described here could be combined with new data for record extension of SWOT-like discharge back to 2002, the beginning of the GRACE era.

To date, most studies analyzing the water cycle with remote sensing data have been done on a small number or river basins. My early efforts at integration of water cycle data focused on using observed river discharges. The results lacked generalizability, however, due to very little training data over Africa, Asia, and parts of South America. It would be valuable to extend the research performed here with new sources of river discharge data. Continental scale studies could readily be performed over North America or Europe. Regional studies could be performed over countries where there is sufficient data, such as Brazil or Chile. For researchers with privileged access to data, China or India would make interesting case studies.

The relatively simple statistical model I described in Chapters 4 and 5 performed almost as well as a more complex neural network model. More could be done to create more detailed parameter regionalization models. For example, we could explore fitting a seasonal regression model, where we fit a different equation for each month or each quarter.

Subsurface flow is neglected in large-sample and global studies, yet it is known that this is a significant flux in some regions. The methods developed in this thesis could help to better characterize those fluxes. Studies in endorheic basins also offer unique opportunities, as there is no outflow, thus simplifying the water budget to three components. There is an opportunity to perform more detailed analyses over such basins using data from GRACE and other satellites.

In Chapter 6, we saw that the calibrated database leads to better predictions of water cycle variables using simple water budget based methods. It would be interesting to see whether calibrated datasets could improve the predictions of more complex simulation models. This could be tested by using EO the datasets before and after calibration as forcing for a hydrologic model. One could choose either an uncalibrated model or a model that can be calibrated automatically. My hypothesis is that the calibrated data will result in lower residuals and better fits, but this would need to be tested.

Additional information could be added in terms of sub-components of the water cycle. For example, a more detailed water cycle model could include soil moisture from the SMOS or SMAP missions (Kolassa et al., 2016), or surface water extent from GIEMS (Prigent et al., 2016) or SWOT.

Ground-based observations are critical to our understanding of the water cycle and will be essential for calibrating and interpreting the data returned by SWOT. Greater cooperation and funding is needed to expand and maintain the in situ discharge observation network. Discharge is an "integrating" variable that is key to understanding the water cycle, the effects of climate change, and for calibrating and validating the next generation of satellite measurements. To maximize the value of these data, better data sharing and quality assurance is essential.

GRACE is unique among remote sensing products used in hydrology, as it is the only mission that directly measures the variable of interest: the mass of water, and how it changes over time. GRACE integrates information about all of the water in a region – groundwater, surface water, soil moisture, etc. As one of the mission scientists has noted, "GRACE cannot feasibly be replicated by ground-based observations" (Rodell et al., 2015). The mission has fostered innovative science, spawned dozens of applications, and resulted in hundreds of publications. It is my hope that governments and space agencies provide continued support for this and similar missions.

This research demonstrated the use of neural networks and machine learning for the integration of satellite data and for the study of the water cycle. Artificial intelligence is rapidly evolving, and new developments could be tested for better modeling of the water cycle. Study of the water cycle at large scales is constrained by the uncertainty in EO datasets. However, alternative approaches in deep learning such as long short-term memory (LSTM) neural network models may be able to exploit the temporal information in EO time series for better predictions of water cycle variables.

Acronyms

AI	Artificial intelligence
AI	Aridity Index
AGU	American Geophysical Union
Aqua	Satellite launched by NASA in 2002 to study the water cycle
AVHRR	Advanced Very High Resolution Radiometer
CAMELS	Catchment Attributes and Meteorology for Large-sample Studies
CARAVAN	A dataset of catchment attributes and meteorology, combines several CAMELS datasets
CI	Confidence Interval
CI	Cyclostationarity Index
CCI	Climate Change Initiative, short name for the ESA Programme on Global Monitoring of Essential Climate Variables
CDF	Cumulative distribution function
CDR	Climate Data Record
CGIAR	(formerly) Consultative Group for International Agricultural Research
CMORPH	CPC Morphing Technique, a global precipitation dataset
CPC	Climate Prediction Center, an office of the US National Weather Service
CMG	Climate Modeling Grid
CSIRO	Australia's Commonwealth Scientific and Industrial Research Organisation
CSR	Center for Space Research at the University of Texas at Austin
ECMWF	European Centre for Medium-Range Weather Forecasts
ED 129	L'École Doctorale des Sciences de l'Environnement d'Île de France
EGU	European Geophysical Union
ENSO	El Niño Southern Oscillation
EO	Earth Observation
ERA5	Fifth generation ECMWF atmospheric reanalysis of the global climate
ESA	European Space Agency
ET	Evapotranspiration
EVI	Enhanced vegetation index
FAPAR	Fraction of absorbed photosynthetically active radiation
GBM	Gradient boosting machine

GDAL	Geospatial Data Abstraction Library
GHCN	Global Historical Climatology Network
GIEMS	Global Inundation Extent from Multi-satellites
GIS	Geographic Information System (software)
GLDAS	Global Land Data Assimilation System
GLEAM	Global Land Evaporation Amsterdam Model (Miralles et al., 2011)
GPCP	Global Precipitation Climatology Project (Adler et al., 2018)
GPM-Imerg	Global Precipitation Monitoring, Integrated Multi-satellitE Retrievals (Huffman et al., 2020)
GRACE	Gravity Recovery and Climate Experiment
GRDC	Global Runoff Data Center
GRUN	Global gridded runoff dataset (Ghiggi et al., 2019)
GSFC	Goddard Space Flight Center
GSIM	Global Streamflow Indices and Metadata Archive (Do et al., 2018)
HBV	Hydrologiska Byråns Vattenbalansavdelning (a hydrologic simulation model from Sweden)
HDF	Hierarchical Data Format
HydroSHEDS	Hydrological Data and Maps Based on Shuttle Elevation Derivatives at Multiple Scales (Lehner et al., 2008)
IAHS	International Association of Hydrological Sciences
IDW	Inverse distance weighted
IQR	Interquartile range
IVW	Inverse-variance weighting
JPL	Jet Propulsion Laboratory
KGE	Kling-Gupta Efficiency
LAI	Leaf area index
LERMA	*Laboratoire d'Etudes du Rayonnement et de la Matière en Astrophysique et Atmosphères*. (Laboratory for the Study of Radiation and Matter in Astrophysics and Atmospheres)
LISFLOOD	Hydrologic model created by the European Union Joint Research Center
LSH	Large sample hydrology
LSTM	Long short-term memory
LWE	Liquid water equivalent
MERIT	Multi-Error Removed Improved-Terrain (Yamazaki et al., 2017)
ML	Machine learning
MLP	Multi-Layer Perceptron

MSE	Mean squared error
MODIS	Moderate Resolution Imaging Spectroradiometer
NaN	Not a number
NASA	(US) National Aeronautics and Space Agency
NetCDF	Network Common Data Form
NDVI	Normalized Difference Vegetation Index
NetCDF	network Common Data Form (format for environmental data)
NN	Neural network
NOAA	US National Oceanic and Atmospheric Agency
NSE	Nash-Sutcliffe Model Efficiency
OI	Optimal interpolation
OLS	Ordinary least squares
PDF	Probability distribution function
PERSIANN	Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks
RMSD	Root mean square difference
RMSE	Root mean square error
RTO	Regression through the origin
SLR	Single linear regression
SMAP	Soil Moisture Active Passive (a NASA satellite mission)
SM2RAIN	an algorithm for estimating rainfall from soil moisture data (Massari, 2020)
SMOS	Soil Moisture and Ocean Salinity (an ESA satellite mission)
SW	Simple weighted mean
SSE	Sum of squared errors
SWOT	Surface Water Ocean Topography Mission
Terra	A multi-national earth observation satellite launched in December 1999
TMPA	Tropical Rainfall Measuring Mission Multi-satellite Precipitation Analysis
TWCS	Total Water Storage Change
TWS	Total Water Storage
UMR	*Unité Mixte de Recherche* (Joint research unit)
USGS	United States Geological Survey
WC	Water cycle
WMO	World Meteorological Organization

Bibliography

Abbott, B. W., Bishop, K., Zarnetske, J. P., Minaudo, C., Chapin, F. S., Krause, S., Hannah, D. M., Conner, L., Ellison, D., Godsey, S. E., Plont, S., Marçais, J., Kolbe, T., Huebner, A., Frei, R. J., Hampton, T., Gu, S., Buhman, M., Sara Sayedi, S., . . . Pinay, G. (2019). Human domination of the global water cycle absent from depictions and perceptions. *Nature Geoscience*, *12*(7), 533–540. https://doi.org/10.1038/s41561-019-0374-y

Abdulla, F. A., & Lettenmaier, D. P. (1997). Development of regional parameter estimation equations for a macroscale hydrologic model. *Journal of Hydrology*, *197*(1-4), 230–257. https://doi.org/10.1016/S0022-1694(96)03262-3

Abolafia-Rosenzweig, R., Pan, M., Zeng, J. L., & Livneh, B. (2021). Remotely sensed ensembles of the terrestrial water budget over major global river basins: An assessment of three closure techniques. *Remote Sensing of Environment*, *252*, 112191. https://doi.org/10.1016/j.rse.2020.112191

Addor, N., Do, H. X., Alvarez-Garreton, C., Coxon, G., Fowler, K., & Mendoza, P. A. (2020). Large-sample hydrology: Recent progress, guidelines for new datasets and grand challenges. *Hydrological Sciences Journal*, *65*(5), 712–725. https://doi.org/10.1080/02626667.2019.1683182

Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017). The CAMELS data set: Catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences*, *21*(10), 5293–5313. https://doi.org/10.5194/hess-21-5293-2017

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., & Shin, D.-B. (2018). The Global Precipitation Climatology Project (GPCP) Monthly Analysis (New Version 2.3) and a Review of 2017 Global Precipitation. *Atmosphere*, *9*(4), 138. https://doi.org/10.3390/atmos9040138

Aires, F. (2014). Combining datasets of satellite-retrieved products. Part I: Methodology and water budget closure. *Journal of Hydrometeorology*, *15*(4), 1677–1691. https://doi.org/10.1175/JHM-D-13-0148.1

Aires, F., Prigent, C., Rossow, W. B., & Rothstein, M. (2001). A new neural network approach including first guess for retrieval of atmospheric water vapor, cloud liquid water path, surface temperature, and emissivities over land from satellite microwave observations. *Journal of Geophysical Research: Atmospheres*, *106*(D14), 14887–14907. https://doi.org/10.1029/2001JD900085

Alber, M., Buganza Tepole, A., Cannon, W. R., De, S., Dura-Bernal, S., Garikipati, K., Karniadakis, G., Lytton, W. W., Perdikaris, P., Petzold, L., & Kuhl, E. (2019). Integrating machine learning and multiscale modeling—perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences. *npj Digital Medicine*, *2*(1), 115. https://doi.org/10.1038/s41746-019-0193-y

Almagro, A., Oliveira, P. T. S., Meira Neto, A. A., Roy, T., & Troch, P. (2021). CABra: A novel large-sample dataset for Brazilian catchments. *Hydrology and Earth System Sciences*, *25*(6), 3105–3135. https://doi.org/10.5194/hess-25-3105-2021

Alvarez-Garreton, C., Mendoza, P. A., Boisier, J. P., Addor, N., Galleguillos, M., Zambrano-Bigiarini, M., Lara, A., Puelma, C., Cortes, G., Garreaud, R., McPhee, J., & Ayala, A. (2018). *The CAMELS-CL dataset: Catchment attributes and meteorology for large sample studies – Chile dataset* (preprint). Hydrometeorology/Instruments and observation techniques. https://doi.org/10.5194/hess-2018-23

Amatulli, G., Domisch, S., Tuanmu, M.-N., Parmentier, B., Ranipeta, A., Malczyk, J., & Jetz, W. (2018). A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. *Scientific Data*, *5*(1), 180040. https://doi.org/10.1038/sdata.2018.40

Arora, V. K. (2002). The use of the aridity index to assess climate change effect on annual runoff. *Journal of Hydrology*, *265*(1-4), 164–177. https://doi.org/10.1016/S0022-1694(02)00101-4

Ashouri, H., Hsu, K., Sorooshian, S., Braithwaite, D., Knapp, K., Cecil, L., Nelson, B., & Prat, O. (2014). PERSIANN-CDR: Daily Precipitation Climate Data Record from Multisatellite Observations for Hydrological and Climate Studies. *Bulletin of the American Meteorological Society*, *96*. https://doi.org/10.1175/BAMS-D-13-00068.1

Australia BOM. (2020). Hydrologic Reference Stations update 2020. Retrieved December 22, 2022, from http://www.bom.gov.au/water/hrs/update_2020.shtml

Bárdossy, A., & Singh, S. K. (2011). Regionalization of hydrological model parameters using data depth. *Hydrology Research*, *42*(5), 356–371. https://doi.org/10.2166/nh.2011.031

Bartos, M., Smith, T. J., Itati01, Debbout, R., Kraft, P., & Huard, D. (2023). Pysheds: 0.3.5. https://doi.org/10.5281/ZENODO.3822494

Beck, H. E., Pan, M., Lin, P., Seibert, J., van Dijk, A. I. J. M., & Wood, E. F. (2020). Global Fully Distributed Parameter Regionalization Based on Observed Streamflow From 4,229 Headwater Catchments. *Journal of Geophysical Research: Atmospheres*, *125*(17). https://doi.org/10.1029/2019JD031485

Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., Van Dijk, A. I., McVicar, T. R., & Adler, R. F. (2019). MSWEP V2 global 3-hourly 0.1 precipitation: Methodology and quantitative assessment. *Bulletin of the American Meteorological Society*, *100*(3), 473–500. https://doi.org/10.1175/BAMS-D-17-0138.1

BfG. (2020). BfG - The GRDC. Retrieved December 22, 2022, from https://www.bafg.de/GRDC/EN/01_GRDC/grdc_node.html

Biancamaria, S., Mballo, M., Le Moigne, P., Sánchez Pérez, J. M., Espitalier-Noël, G., Grusson, Y., Cakir, R., Häfliger, V., Barathieu, F., Trasmonte, M., Boone, A., Martin, E., & Sauvage, S. (2019). Total water storage variability from GRACE mission and hydrological models for a 50,000 km2 temperate watershed: The Garonne River basin (France). *Journal of Hydrology: Regional Studies*, *24*, 100609. https://doi.org/10.1016/j.ejrh.2019.100609

Biemans, H., Hutjes, R. W. A., Kabat, P., Strengers, B. J., Gerten, D., & Rost, S. (2009). Effects of precipitation uncertainty on discharge calculations for main river basins. *Journal of Hydrometeorology*, *10*(4), 1011–1025. https://doi.org/10.1175/2008JHM1067.1

Bishop, C. M. (1996). *Neural Networks for Pattern Recognition*. Oxford University Press.

Brown, J. (2018). NDVI, the Foundation for Remote Sensing Phenology. Retrieved September 22, 2023, from https://www.usgs.gov/special-topics/remote-sensing-phenology/science/ndvi-foundation-remote-sensing-phenology

Budyko, M. I. (1974). *Climate and Life*. Academic Press.

Cao, M., Wang, W., Xing, W., Wei, J., Chen, X., Li, J., & Shao, Q. (2021). Multiple sources of uncertainties in satellite retrieval of terrestrial actual evapotranspiration. *Journal of Hydrology*, *601*, 126642. https://doi.org/10.1016/j.jhydrol.2021.126642

Cao, Q., Painter, T. H., Currier, W. R., Lundquist, J. D., & Lettenmaier, D. P. (2018). Estimation of Precipitation over the OLYMPEX Domain during Winter 2015/16. *Journal of Hydrometeorology*, *19*(1), 143–160. https://doi.org/10.1175/JHM-D-17-0076.1

Cauchy, A. (1847). Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris*, *25*(1847), 536–538. Retrieved October 6, 2023, from https://cs.uwaterloo.ca/~y328yu/classics/cauchy-en.pdf

Chagas, V. B. P., Chaffe, P. L. B., Addor, N., Fan, F. M., Fleischmann, A. S., Paiva, R. C. D., & Siqueira, V. A. (2020). CAMELS-BR: Hydrometeorological time series and landscape attributes for 897 catchments in Brazil. *Earth System Science Data*, *12*(3), 2075–2096. https://doi.org/10.5194/essd-12-2075-2020

Chen, J., Tapley, B., Rodell, M., Seo, K.-W., Wilson, C., Scanlon, B. R., & Pokhrel, Y. (2020). 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*, *56*(10). https://doi.org/10.1029/2020WR028032

Chen, M., & Xie, P. (2008). CPC unified gauge-based analysis of global daily precipitation. https://ftp.cpc.ncep.noaa.gov/precip/CPC_UNI_PRCP/GAUGE_CONUS/DOCU/

Chen, M., Shi, W., Xie, P., Silva, V. B. S., Kousky, V. E., Wayne Higgins, R., & Janowiak, J. E. (2008). Assessing objective techniques for gauge-based analyses of global daily precipitation. *Journal of Geophysical Research*, *113*(D4), D04110. https://doi.org/10.1029/2007JD009132

Chiew, F. H. S. (2010). Lumped Conceptual Rainfall-Runoff Models and Simple Water Balance Methods: Overview and Applications in Ungauged and Data Limited Regions. *Geography Compass*, *4*(3), 206–225. https://doi.org/10.1111/j.1749-8198.2009.00318.x

Chu, D., Shen, H., Guan, X., & Li, X. (2022). An L1-regularized variational approach for NDVI time-series reconstruction considering inter-annual seasonal similarity. *International Journal of Applied Earth Observation and Geoinformation*, *114*, 103021. https://doi.org/10.1016/j.jag.2022.103021

Cloke, H. L., & Hannah, D. M. (2011). Large-scale hydrology-advances in understanding processes, dynamics and models from beyond river basin to global scale. *Hydrological Processes*, *25*(7), 991–1200. https://doi.org/10.1002/hyp.8059

Conrad, O. (2008). SAGA-GIS Module Ordinary Kriging. Retrieved August 27, 2023, from https://saga-gis.sourceforge.io/saga_tool_doc/2.2.3/statistics_kriging_0.html

Coxon, G., Addor, N., Bloomfield, J. P., Freer, J., Fry, M., Hannaford, J., Howden, N. J. K., Lane, R., Lewis, M., Robinson, E. L., Wagener, T., & Woods, R. (2020). CAMELS-GB: Hydrometeorological time series and landscape attributes for 671 catchments in Great Britain. *Earth System Science Data*, *12*(4), 2459–2483. https://doi.org/10.5194/essd-12-2459-2020

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, *2*(4), 303–314. https://doi.org/10.1007/BF02551274

Didan, K. (2015). MODIS Vegetation Index Products (NDVI and EVI). Retrieved April 15, 2021, from https://modis.gsfc.nasa.gov/data/dataprod/mod13.php

Do, H. X., Gudmundsson, L., Leonard, M., & Westra, S. (2018). The Global Streamflow Indices and Metadata Archive (GSIM) – Part 1: The production of a daily streamflow archive and metadata. *Earth System Science Data*, *10*(2), 765–785. https://doi.org/10.5194/essd-10-765-2018

Dorigo, W., Dietrich, S., Aires, F., Brocca, L., Carter, S., Cretaux, J.-F., Dunkerley, D., Enomoto, H., Forsberg, R., & Güntner, A. (2021). Closing the water cycle from observations across scales: Where do we stand? *Bulletin of the American Meteorological Society*, *102*(10), E1897–E1935. https://doi.org/10.1175/BAMS-D-19-0316.1

Duan, Z., Liu, J., Tuo, Y., Chiogna, G., & Disse, M. (2016). Evaluation of eight high spatial resolution gridded precipitation products in Adige Basin (Italy) at multiple temporal and spatial scales. *Science of The Total Environment*, *573*, 1536–1553. https://doi.org/10.1016/j.scitotenv.2016.08.213

Durand, M., Gleason, C. J., Garambois, P. A., Bjerklie, D., Smith, L. C., Roux, H., Rodriguez, E., Bates, P. D., Pavelsky, T. M., Monnier, J., Chen, X., Di Baldassarre, G., Fiset, J.-M., Flipo, N., Frasson, R. P. d. M., Fulton, J., Goutal, N., Hossain, F., Humphries, E., . . . Vilmin, L. (2016). An intercomparison of remote sensing river discharge estimation algorithms from measurements of river height, width, and slope. *Water Resources Research*, *52*(6), 4527–4549. https://doi.org/10.1002/2015WR018434

Durre, I., Menne, M. J., Gleason, B. E., Houston, T. G., & Vose, R. S. (2010). Comprehensive Automated Quality Assurance of Daily Surface Observations. *Journal of Applied Meteorology and Climatology*, *49*(8), 1615–1633. https://doi.org/10.1175/2010JAMC2375.1

Durre, I., Menne, M. J., & Vose, R. S. (2008). Strategies for Evaluating Quality Assurance Procedures. *Journal of Applied Meteorology and Climatology*, *47*(6), 1785–1791. https://doi.org/10.1175/2007JAMC1706.1

Eagleson, P. S. (1994). The evolution of modern hydrology (from watershed to continent in 30 years). *Advances in Water Resources*, *3*(18), 16. https://doi.org/10.1016/0309-1708(94)90019-1

Eisenhauer, J. G. (2003). Regression through the origin. *Teaching Statistics*, *25*(3), 76–80. https://doi.org/10.1111/1467-9639.00136

England, J. F., Jr., Cohn, T. A., Faber, B. A., Stedinger, J. R., Jr, W. O. T., Veilleux, A. G., Kiang, J. E., & Robert R. Mason, J. (2019). *Guidelines for determining flood flow frequency — Bulletin 17C*. U.S. Geological Survey. https://doi.org/10.3133/tm4B5

Eyler, B. (2022). Science Shows Chinese Dams Are Devastating the Mekong. *Foreign Policy*. Retrieved May 22, 2022, from https://foreignpolicy.com/2020/04/22/science-shows-chinese-dams-devastating-mekong-river/

Fan, Y. (2019). Are catchments leaky? *WIREs Water*, *6*(6), e1386. https://doi.org/10.1002/wat2.1386

Farseev, A. (2023). Is Bigger Better? Why The ChatGPT Vs. GPT-3 Vs. GPT-4 'Battle' Is Just A Family Chat. Retrieved October 6, 2023, from https://www.forbes.com/sites/forbestechcouncil/2023/02/17/is-bigger-better-why-the-chatgpt-vs-gpt-3-vs-gpt-4-battle-is-just-a-family-chat/

Fatichi, S. (2023). Inverse Distance Weight. Retrieved August 25, 2023, from https://fr.mathworks.com/matlabcentral/fileexchange/24477-inverse-distance-weight

Fekete, B. M., Looser, U., Pietroniro, A., & Robarts, R. D. (2012). Rationale for Monitoring Discharge on the Ground. *Journal of Hydrometeorology*, *13*(6), 1977–1986. https://doi.org/10.1175/JHM-D-11-0126.1

Fekete, B. M., Vörösmarty, C. J., & Grabs, W. (2002). High-resolution fields of global runoff combining observed river discharge and simulated water balances. *Global Biogeochemical Cycles*, *16*(3), 15–1–15–10. https://doi.org/10.1029/1999gb001254

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., . . . Wood, E. F. (2017). 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. https://doi.org/10.1002/2016WR020175

Fisher, J. B., Tu, K. P., & Baldocchi, D. D. (2008). Global estimates of the land–atmosphere water flux based on monthly AVHRR and ISLSCP-II data, validated at 16 FLUXNET sites. *Remote Sensing of Environment*, *112*(3), 901–919. https://doi.org/10.1016/j.rse.2007.06.025

Flechtner, F., Morton, P., Watkins, M., & Webb, F. (2014). Status of the GRACE Follow-On Mission. In U. Marti (Ed.), *Gravity, Geoid and Height Systems* (pp. 117–121). Springer International Publishing. https://doi.org/10.1007/978-3-319-10837-7_15

Fowler, K. J. A., Acharya, S. C., Addor, N., Chou, C., & Peel, M. C. (2021). *CAMELS-AUS: Hydrometeorological time series and landscape attributes for 222 catchments in Australia* (preprint). Hydrology and Soil Science – Hydrology. https://doi.org/10.5194/essd-2020-228

Frame, J. M., Kratzert, F., Gupta, H. V., Ullrich, P., & Nearing, G. S. (2023). On strictly enforced mass conservation constraints for modelling the Rainfall-Runoff process. *Hydrological Processes*, *37*(3), e14847. https://doi.org/10.1002/hyp.14847

Fraser, C. G. (2005). Leonhard Euler, book on the calculus of variations (1744). In *Landmark Writings in Western Mathematics 1640-1940* (pp. 168–180). Elsevier. Retrieved October 6, 2023, from https://www.sciencedirect.com/science/article/pii/B9780444508713500930

Gao, H., Tang, Q., Ferguson, C. R., Wood, E. F., & Lettenmaier, D. P. (2010). Estimating the water budget of major US river basins via remote sensing. *International Journal of Remote Sensing*, *31*(14), 3955–3978. https://doi.org/10.1080/01431161.2010.483488

Garambois, P., Roux, H., Larnier, K., Labat, D., & Dartus, D. (2015). Parameter regionalization for a process-oriented distributed model dedicated to flash floods. *Journal of Hydrology*, *525*, 383–399. https://doi.org/10.1016/j.jhydrol.2015.03.052

Ghiggi, G., Humphrey, V., Seneviratne, S. I., & Gudmundsson, L. (2021). G-RUN ENSEMBLE: A Multi-Forcing Observation-Based Global Runoff Reanalysis. *Water Resources Research*, *57*(5). https://doi.org/10.1029/2020WR028787

Ghiggi, G., Humphrey, V., Seneviratne, S. I., & Gudmundsson, L. (2019). GRUN: An observation-based global gridded runoff dataset from 1902 to 2014. *Earth System Science Data*, *11*(4), 1655–1674. https://doi.org/10.5194/essd-11-1655-2019

Giglio, L., Schroeder, W., & Hall, J. V. (2020). MODIS Collection 6 Active Fire Product User's Guide Revision C. https://modis-fire.umd.edu/files/MODIS_C6_Fire_User_Guide_C.pdf

Greene, C. A., Thirumalai, K., Kearney, K. A., Delgado, J. M., Schwanghart, W., Wolfenbarger, N. S., Thyng, K. M., Gwyther, D. E., Gardner, A. S., & Blankenship, D. D. (2019). The Climate Data Toolbox for MATLAB. *Geochemistry, Geophysics, Geosystems*, *20*(7), 3774–3781. https://doi.org/10.1029/2019GC008392

Gudmundsson, L., Do, H. X., Leonard, M., & Westra, S. (2018). The Global Streamflow Indices and Metadata Archive (GSIM) – Part 2: Quality control, time-series indices and homogeneity assessment. *Earth System Science Data*, *10*(2), 787–804. https://doi.org/10.5194/essd-10-787-2018

Guillory, A. (2022). ERA5. Retrieved March 10, 2023, from https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5

Guimarães Nobre, G., Muis, S., Veldkamp, T. I. E., & Ward, P. J. (2019). Achieving the reduction of disaster risk by better predicting impacts of El Niño and La Niña. *Progress in Disaster Science*, *2*, 100022. https://doi.org/10.1016/j.pdisas.2019.100022

Guo, X., Wu, Z., He, H., & Xu, Z. (2022). Evaluating the Potential of Different Evapotranspiration Datasets for Distributed Hydrological Model Calibration. *Remote Sensing*, *14*(3), 629. https://doi.org/10.3390/rs14030629

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of hydrology*, *377*(1-2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Hagan, M. T., & Demuth, H. B. (2002). *Neural Network Design* (2nd). https://hagan.okstate.edu/nnd.html

Hall, D., & Riggs, G. (2021). MODIS/Terra Snow Cover Monthly L3 Global 0.05Deg CMG, version 6.1. https://doi.org/10.5067/MODIS/MOD10CM.061

Hanasaki, N., Kanae, S., & Oki, T. (2006). A reservoir operation scheme for global river routing models. *Journal of Hydrology*, *327*(1), 22–41. https://doi.org/10.1016/j.jhydrol.2005.11.011

Harrigan, S., Zsoter, E., Alfieri, L., Prudhomme, C., Salamon, P., Wetterhall, F., Barnard, C., Cloke, H., & Pappenberger, F. (2020). GloFAS-ERA5 operational global river discharge reanalysis 1979–present. *Earth System Science Data*, *12*(3), 2043–2060. https://doi.org/10.5194/essd-12-2043-2020

Hartung, J., Knapp, G., & Sinha, B. K. (2008). *Statistical meta-analysis with applications*. Wiley.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Second Edition). Springer. https://hastie.su.domains/Papers/ESLII.pdf

Healy, R., Winter, T., LaBaugh, J., & Franke, O. (2007). *Water Budgets: Foundations for Effective Water-Resources and Environmental Management* (Circular No. 1308). U.S. Geological Survey. https://pubs.usgs.gov/circ/2007/1308/

Heberger, M. (2012). Australia's Millennium Drought: Impacts and Responses. In P. H. Gleick (Ed.), *The World's Water Volume 7: The Biennial Report on Freshwater Resources* (pp. 97–126). Island Press.

Heberger, M. (2022). Mheberger/delineator: 1.0. https://doi.org/10.5281/ZENODO.7314287

Hegerl, G. C., Black, E., Allan, R. P., Ingram, W. J., Polson, D., Trenberth, K. E., Chadwick, R. S., Arkin, P. A., Sarojini, B. B., Becker, A., Dai, A., Durack, P. J., Easterling, D., Fowler, H. J., Kendon, E. J., Huffman, G. J., Liu, C., Marsh, R., New, M., . . . Zhang, X. (2015). Challenges in Quantifying Changes in the Global Water Cycle. *Bulletin of the American Meteorological Society*, *96*(7), 1097–1115. https://doi.org/10.1175/BAMS-D-13-00212.1

Helsel, D. R., Hirsch, R. M., Ryberg, K. R., Archfield, S. A., & Gilroy, E. J. (2020). *Statistical Methods in Water Resources* (tech. rep. 4-A3). U.S. Geological Survey. Reston, Virginia. Retrieved May 22, 2020, from http://pubs.er.usgs.gov/publication/tm4A3

Herman, J., Davis, A., Chin, K., Kinzler, M., Scholz, S., & Steinhoff, M. (2012). Life with a weak Heart; Prolonging the GRACE Mission despite degraded Batteries. *AIAA Meeting Papers, SpaceOps 2012*, 12. https://elib.dlr.de/76288/1/id1275101-Paper-002.pdf

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., & Thépaut, J-N.\. (2018). Copernicus Climate Data Store —. https://doi.org/10.24381/cds.adbb2d47

Hogan, R. (2015). Radiation Quantities in the ECMWF model and MARS. https://www.ecmwf.int/sites/default/files/elibrary/2015/18490-radiation-quantities-ecmwf-model-and-mars.pdf

Hong, J., Lee, S., Lee, G., Yang, D.-S., Bae, J. H., Kim, J., Kim, K.-S., & Lim, K. J. (2021). Comparison of Machine Learning Algorithms for Discharge Prediction of Multipurpose Dam. *Water*, *13*(23). https://doi.org/10.3390/w13233369

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Huang, D.-Q., Zhu, J., Zhang, Y.-C., Huang, Y., & Kuang, X.-Y. (2016). Assessment of summer monsoon precipitation derived from five reanalysis datasets over East Asia. *Quarterly Journal of the Royal Meteorological Society*, *142*(694), 108–119. https://doi.org/10.1002/qj.2634

Huffman, G. J., Adler, R. F., Arkin, P., Chang, A., Ferraro, R., Gruber, A., Janowiak, J., McNab, A., Rudolf, B., & Schneider, U. (1997). The Global Precipitation Climatology Project (GPCP) Combined Precipitation Dataset. *Bulletin of the American Meteorological Society*, *78*(1), 5–20. https://doi.org/10.1175/1520-0477(1997)078⟨0005:TGPCPG⟩2.0.CO;2

Huffman, G. J., Bolvin, D. T., Braithwaite, D., Hsu, K., Joyce, R., Kidd, C., Soroosh Sorooshian, Jackson Tan, & Xie, P. (2020). *Algorithm Theoretical Basis Document (ATBD) Version 06: NASA global precipitation measurement (GPM) integrated multi-satellite retrievals for GPM (IMERG)* (tech. rep.). National Aeronautics and Space Administration. Greenbelt, MD. https://gpm.nasa.gov/sites/default/files/2020-05/IMERG_ATBD_V06.3.pdf

Huffman, G. J., E.F. Stocker, D.T. Bolvin, E.J. Nelkin, & Jackson Tan. (2019). GPM IMERG Final Precipitation L3 1 day 0.1 degree x 0.1 degree V06. https://doi.org/10.5067/GPM/IMERGDF/DAY/06

Ibarra, D. E., David, C. P. C., & Tolentino, P. L. M. (2021). Technical note: Evaluation and bias correction of an observation-based global runoff dataset using streamflow observations from small tropical catchments in the Philippines. *Hydrology and Earth System Sciences*, *25*(5), 2805–2820. https://doi.org/10.5194/hess-25-2805-2021

Jackson, E. K., Roberts, W., Nelsen, B., Williams, G. P., Nelson, E. J., & Ames, D. P. (2019). Introductory overview: Error metrics for hydrologic modelling – A review of common practices and an open source library to facilitate use and adoption. *Environmental Modelling & Software*, *119*, 32–48. https://doi.org/10.1016/j.envsoft.2019.05.001

James, G., Witten, D., Hastie, T., & Tibshirani, R. (Eds.). (2013). *An introduction to statistical learning: With applications in R*. Springer. https://www.statlearning.com/

James, L. D. (1972). Hydrologic modeling, parameter estimation, and watershed characteristics. *Journal of Hydrology*, *17*(4), 283–307. Retrieved October 19, 2023, from https://www.sciencedirect.com/science/article/pii/0022169472900893

Jiang, D., Wang, J., Huang, Y., Zhou, K., Ding, X., & Fu, J. (2014). The Review of GRACE Data Applications in Terrestrial Hydrology Monitoring. *Advances in Meteorology*, *2014*. https://doi.org/10.1155/2014/725131

Kampf, S. K., Burges, S. J., Hammond, J. C., Bhaskar, A., Covino, T. P., Eurich, A., Harrison, H., Lefsky, M., Martin, C., McGrath, D., Puntenney-Desmond, K., & Willi, K. (2020). The Case for an Open Water Balance: Re-envisioning Network Design and Data Analysis for a Complex, Uncertain World. *Water Resources Research*, *56*(6). https://doi.org/10.1029/2019WR026699

Karssenberg, D., Schmitz, O., Salamon, P., De Jong, K., & Bierkens, M. F. (2010). A software framework for construction of process-based stochastic spatio-temporal models and data assimilation. *Environmental Modelling & Software*, *25*(4), 489–502. https://doi.org/10.1016/j.envsoft.2009.10.004

Kauffeldt, A., Halldin, S., Rodhe, A., Xu, C.-Y., & Westerberg, I. K. (2013). Disinformative data in large-scale hydrological modelling. *Hydrology and Earth System Sciences*, *17*(7), 2845–2857. https://doi.org/10.5194/hess-17-2845-2013

Khan, M. S., Liaqat, U. W., Baik, J., & Choi, M. (2018). Stand-alone uncertainty characterization of GLEAM, GLDAS and MOD16 evapotranspiration products using an extended triple collocation approach. *Agricultural and Forest Meteorology*, *252*, 256–268. https://doi.org/10.1016/j.agrformet.2018.01.022

Kidd, C., Bauer, P., Turk, J., Huffman, G. J., Joyce, R., Hsu, K.-L., & Braithwaite, D. (2012). Intercomparison of High-Resolution Precipitation Products over Northwest Europe. *Journal of Hydrometeorology*, *13*(1), 67–83. https://doi.org/10.1175/JHM-D-11-042.1

Kim, P. (2017). *MATLAB Deep Learning*. Apress. https://doi.org/10.1007/978-1-4842-2845-6

King, F., Erler, A. R., Frey, S. K., & Fletcher, C. G. (2020). Application of machine learning techniques for regional bias correction of snow water equivalent estimates in Ontario, Canada. *Hydrology and Earth System Sciences*, *24*(10), 4887–4902. https://doi.org/10.5194/hess-24-4887-2020

Kling, H., Fuchs, M., & Paulin, M. (2012). Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *Journal of Hydrology*, *424*, 264–277. https://doi.org/10.1016/j.jhydrol.2012.01.011

Knoben, W. J. M., Freer, J. E., & Woods, R. A. (2019). Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrology and Earth System Sciences*, *23*(10), 4323–4331. https://doi.org/10.5194/hess-23-4323-2019

Kolassa, J., Gentine, P., Prigent, C., & Aires, F. (2016). Soil moisture retrieval from AMSR-E and ASCAT microwave observation synergy. Part 1: Satellite data analysis. *Remote Sensing of Environment*, *173*, 1–14. https://doi.org/10.1016/j.rse.2015.11.011

Konikow, L. F. (2013). Groundwater Depletion in the United States (1900−2008). *Scientific Investigations Report*. https://doi.org/10.3133/sir20135079

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019). Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning. *Water Resources Research*, *55*(12), 11344–11354. https://doi.org/10.1029/2019WR026065

Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L., Hassidim, A., Klotz, D., & Nevo, S. (2023). Caravan-A global community dataset for large-sample hydrology. *Scientific Data*, *10*(1), 61.

Kubota, T., Aonashi, K., Ushio, T., Shige, S., Takayabu, Y. N., Kachi, M., Arai, Y., Tashima, T., Masaki, T., Kawamoto, N., Mega, T., Yamamoto, M. K., Hamada, A., Yamaji, M., Liu, G., & Oki, R. (2020). Global Satellite Mapping of Precipitation (GSMaP) Products in the GPM Era. In V. Levizzani, C. Kidd, D. B. Kirschbaum, C. D. Kummerow, K. Nakamura, & F. J. Turk (Eds.), *Satellite Precipitation Measurement* (pp. 355–373). Springer International Publishing. Retrieved December 21, 2021, from http://link.springer.com/10.1007/978-3-030-24568-9_20

Kumar, S. V., Zaitchik, B. F., Peters-Lidard, C. D., Rodell, M., Reichle, R., Li, B., Jasinski, M., Mocko, D., Getirana, A., De Lannoy, G., Cosh, M. H., Hain, C. R., Anderson, M., Arsenault, K. R., Xia, Y., & Ek, M. (2016). Assimilation of Gridded GRACE Terrestrial Water Storage Estimates in the North American Land Data Assimilation System. *Journal of Hydrometeorology*, *17*(7), 1951–1972. https://doi.org/10.1175/JHM-D-15-0157.1

Kusche, J., Schmidt, R., Petrovic, S., & Rietbroek, R. (2009). Decorrelated GRACE time-variable gravity solutions by GFZ, and their validation using a hydrological model. *Journal of Geodesy*, *83*(10), 903–913. https://doi.org/10.1007/s00190-009-0308-3

Lamontagne, J. R., Barber, C. A., & Vogel, R. M. (2020). Improved Estimators of Model Performance Efficiency for Skewed Hydrologic Data. *Water Resources Research*, *56*(9). https://doi.org/10.1029/2020WR027101

Landerer, F. W., & Swenson, S. C. (2012). Accuracy of scaled GRACE terrestrial water storage estimates. *Water Resources Research*, *48*(4). https://doi.org/10.1029/2011wr011453

Landerer, F. W. (2021). CSR TELLUS GRACE Level-3 Monthly Land Water-Equivalent-Thickness Surface Mass Anomaly Release 6.0 version 04 in netCDF/ASCII/GeoTIFF Formats. https://doi.org/10.5067/TELND-3AC64

Landerer, F. W., & Cooley, S. S. (2021). *Gravity Recovery and Climate Experiment Follow-on (GRACE-FO) Level-3 Data Product User Handbook*. Jet Propulsion Laboratory, California Institute of Technology.

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

L'Ecuyer, T. S., Beaudoing, H. K., Rodell, M., Olson, W., Lin, B., Kato, S., Clayson, C. A., Wood, E., Sheffield, J., & Adler, R. (2015). The observed state of the energy budget in the early twenty-first century. *Journal of Climate*, *28*(21), 8319–8346. https://doi.org/10.1175/JCLI-D-14-00556.1

Lee, K. T., Ho, J.-Y., Kao, H.-M., Lin, G.-F., & Yang, T.-H. (2019). Using ensemble precipitation forecasts and a rainfall-runoff model for hourly reservoir inflow forecasting during typhoon periods. *Journal of Hydro-Environment Research*, *22*, 29–37.

Lehmann, F., Vishwakarma, B. D., & Bamber, J. (2022). How well are we able to close the water budget at the global scale? *Hydrology and Earth System Sciences*, *26*(1). https://doi.org/doi.org/10.5194/hess-26-35-2022

Lehner, B. (2011). *Derivation of watershed boundaries for GRDC gauging stations based on the HydroSHEDS drainage network* (Report No. 41). Global Runoff Data Centre. Koblenz, Germany. Retrieved September 2, 2021, from https://www.bafg.de/GRDC/EN/02_srvcs/24_rprtsrs/report_41.html?nn=201764

Lehner, B., Verdin, K., & Jarvis, A. (2008). New Global Hydrography Derived from Spaceborne Elevation Data. *Eos, Transactions American Geophysical Union*, *89*(10), 93–94. https://doi.org/10.1029/2008EO100001

Lettenmaier, D. P., Alsdorf, D., Dozier, J., Huffman, G. J., Pan, M., & Wood, E. F. (2015). Inroads of remote sensing into hydrologic science during the WRR era. *Water Resources Research*, *51*(9), 7309–7342. https://doi.org/10.1002/2015WR017616

Levizzani, V., & Cattani, E. (2019). Satellite Remote Sensing of Precipitation and the Terrestrial Water Cycle in a Changing Climate. *Remote Sensing*. https://doi.org/10.3390/rs11192301

Li, F., & Lawrence, D. M. (2017). Role of Fire in the Global Land Water Budget during the Twentieth Century due to Changing Ecosystems. *Journal of Climate*, *30*(6), 1893–1908. https://doi.org/10.1175/JCLI-D-16-0460.1

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

Lin, P., Pan, M., Beck, H. E., Yang, Y., Yamazaki, D., Frasson, R., David, C. H., Durand, M., Pavelsky, T. M., Allen, G. H., Gleason, C. J., & Wood, E. F. (2019). Global

Reconstruction of Naturalized River Flows at 2.94 Million Reaches. *Water Resources Research*, *55*(8), 6499–6516. https://doi.org/10.1029/2019WR025287

Lin, P., Pan, M., Wood, E. F., Yamazaki, D., & Allen, G. H. (2021). A new vector-based global river network dataset accounting for variable drainage density. *Scientific Data*, *8*(1), 28. https://doi.org/10.1038/s41597-021-00819-9

Lindsay, J. B., Rothwell, J. J., & Davies, H. (2008). Mapping outlet points used for watershed delineation onto DEM-derived stream networks. *Water Resources Research*, *44*(8). https://doi.org/10.1029/2007WR006507

Liu, H., Tolson, B. A., Newman, A. J., & Wood, A. W. (2021). Leveraging ensemble meteorological forcing data to improve parameter estimation of hydrologic models. *Hydrological Processes*, *35*(11), e14410. https://doi.org/10.1002/hyp.14410

Liu, Y., Wagener, T., Beck, H. E., & Hartmann, A. (2020). What is the hydrologically effective area of a catchment? *Environmental Research Letters*, *15*(10), 104024. https://doi.org/10.1088/1748-9326/aba7e5

Lo Conti, F., Hsu, K.-L., Noto, L. V., & Sorooshian, S. (2014). Evaluation and comparison of satellite precipitation estimates with reference to a local area in the Mediterranean Sea. *Atmospheric Research*, *138*, 189–204. https://doi.org/10.1016/j.atmosres.2013.11.011

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

Loomis, B. D., Luthcke, S. B., & Sabaka, T. J. (2019). Regularization and error characterization of GRACE mascons. *Journal of Geodesy*, *93*(9), 1381–1398. https://doi.org/10.1007/s00190-019-01252-y

Lopez, O., McCabe, M., & Houborg, R. (2015). Evaluation of multiple satellite evaporation products in two dryland regions using GRACE. *MODSIM2015, 21st International Congress on Modelling and Simulation*, 1379–1385. https://repository.kaust.edu.sa/handle/10754/621070

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

Lu, J., Wang, G., Chen, T., Li, S., Hagan, D. F. T., Kattel, G., Peng, J., Jiang, T., & Su, B. (2021). A harmonized global land evaporation dataset from model-based products covering 1980–2017. *Earth System Science Data*, *13*(12), 5879–5898. https://doi.org/10.5194/essd-13-5879-2021

Maier, H. R., Galelli, S., Razavi, S., Castelletti, A., Rizzoli, A., Athanasiadis, I. N., Sànchez-Marrè, M., Acutis, M., Wu, W., & Humphrey, G. B. (2023). Exploding the myths: An introduction to artificial neural networks for prediction and forecasting. *Environmental Modelling & Software*, *167*, 105776. https://doi.org/10.1016/j.envsoft.2023.105776

Mälicke, M., Möller, E., Schneider, H. D., & Müller, S. (2021). Scikit-gstat: A scipy flavoured geostatistical variogram analysis toolbox. https://doi.org/10.5281/ZENODO.4835779

Marcus, A. (2022). NASA researchers retract Nature paper on climate change and evapotranspiration. Retrieved August 3, 2023, from https://retractionwatch.com/2022/03/03/nasa-researchers-retract-nature-paper-on-climate-change-and-evapotranspiration/

Martens, B., Miralles, D. G., Lievens, H., Van Der Schalie, R., De Jeu, R. A., Fernández-Prieto, D., Beck, H. E., Dorigo, W. A., & Verhoest, N. E. (2017). GLEAM v3: Satellite-based land evaporation and root-zone soil moisture. *Geoscientific Model Development*, *10*(5), 1903–1925. https://doi.org/10.5194/gmd-10-1903-2017

Martens, B., Waegeman, W., Dorigo, W. A., Verhoest, N. E. C., & Miralles, D. G. (2018). Terrestrial evaporation response to modes of climate variability. *npj Climate and Atmospheric Science*, *1*(1), 1–7. https://doi.org/10.1038/s41612-018-0053-5

Massari, C. (2020). GPM+SM2RAIN (2007-2018): Quasi-global 25km/daily rainfall product from the integration of GPM and SM2RAIN-based rainfall products. https://doi.org/10.5281/zenodo.3854817

Massari, C., Brocca, L., Pellarin, T., Abramowitz, G., Filippucci, P., Ciabatta, L., Maggioni, V., Kerr, Y., & Fernandez Prieto, D. (2020). A daily 25 km short-latency rainfall product for data-scarce regions based on the integration of the Global Precipitation Measurement mission rainfall and multiple-satellite soil moisture products. *Hydrology and Earth System Sciences*, *24*(5), 2687–2710. https://doi.org/10.5194/hess-24-2687-2020

Mathevet, T., Michel, C., Andréassian, V., & Perrin, C. (2006). A bounded version of the Nash-Sutcliffe criterion for better model assessment on large sets of basins. *IHS Publication 307*, 211–220. https://iahs.info/uploads/dms/13614.21--211-219-41-MATHEVET.pdf

McCabe, M. F., Rodell, M., Alsdorf, D. E., Miralles, D. G., Uijlenhoet, R., Wagner, W., Lucieer, A., Houborg, R., Verhoest, N. E. C., Franz, T. E., Shi, J., Gao, H., & Wood, E. F. (2017). The future of Earth observation in hydrology. *Hydrology and Earth System Sciences*, *21*(7), 3879–3914. https://doi.org/10.5194/hess-21-3879-2017

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., & Houston, T. G. (2012). An Overview of the Global Historical Climatology Network-Daily Database. *Journal of Atmospheric and Oceanic Technology*, *29*(7), 897–910. https://doi.org/10.1175/JTECH-D-11-00103.1

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

Modis Land Team. (2021). MODIS Grids. Retrieved February 27, 2023, from https://modis-land.gsfc.nasa.gov/MODLAND_grid.html

Moshe, Z., Metzger, A., Kratzert, F., Elidan, G., Morin, E., Nevo, S., & E-Yyaniv, R. (2020). HydroNets: Leveraging River Network Structure and Deep Neural Networks for Hydrologic Modeling. *Eighth International Conference on Learning Representations*, 5. https://doi.org/10.5194/egusphere-egu2020-4135

Mu, Q., Zhao, M., & Running, S. W. (2011). Improvements to a MODIS global terrestrial evapotranspiration algorithm. *Remote Sensing of Environment*, *115*(8), 1781–1800. https://doi.org/10.1016/j.rse.2011.02.019

Müller, K., Löw, S., Herman, J., Gaston, R., & Davis, a. (2019). End–of–Life Power Management on the GRACE Satellites with Several Failed Battery Cells. *70th International Astronautical Congress (IAC)70th International Astronautical Congress (IAC)*, 1–12. https://www.researchgate.net/profile/Sebastian-Loew-2/publication/336721493_End-of-Life_Power_Management_on_the_Grace_Satellites_with_several_failed_Battery_Cells/links/5dc57ac44585151435f794f2/End-of-Life-Power-Management-on-the-Grace-Satellites-with-several-failed-Battery-Cells.pdf

Munier, S., Becker, M., Maisongrande, P., & Cazenave, A. (2012). Using GRACE to detect Groundwater Storage variations: The cases of Canning Basin and Guarani Aquifer System. *International Water Technology Journal*, *2*(1). https://hal.science/hal-01162472/

Munier, S., & Aires, F. (2018). A new global method of satellite dataset merging and quality characterization constrained by the terrestrial water budget. *Remote Sensing of Environment*, *205*, 119–130. https://doi.org/10.1016/j.rse.2017.11.008

Munier, S., Aires, F., Schlaffer, S., Prigent, C., Papa, F., Maisongrande, P., & Pan, M. (2014). Combining datasets of satellite-retrieved products for basin-scale water balance study: 2. Evaluation on the Mississippi Basin and closure correction model. *Journal of Geophysical Research: Atmospheres*, *119*(21), 12, 100–12, 116. https://doi.org/10.1002/2014JD021953

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., & Thépaut, J.-N. (2021). *ERA5-Land: A state-of-the-art global reanalysis dataset for land applications* (preprint). Data, Algorithms, and Models. https://doi.org/10.5194/essd-2021-82

NASA. (2019). NDVI from NASA ARC ECOCAST GIMMS NDVI3g v1p0: Version 1.0. https://iridl.ldeo.columbia.edu/SOURCES/.NASA/.ARC/.ECOCAST/.GIMMS/.NDVI3g/.v1p0/.ndvi/

NASA Jet Propulsion Laboratory. (2018). Overview - Monthly Mass Grids. Retrieved February 28, 2023, from https://grace.jpl.nasa.gov/data/monthly-mass-grids

NASA Jet Propulsion Laboratory. (2020). Which GRACE(-FO) data set should I choose? Retrieved July 25, 2023, from https://grace.jpl.nasa.gov/data/choosing-a-solution

NASA Jet Propulsion Laboratory. (2023). GRACE & GRACE-FO - Data Months / Days. Retrieved July 13, 2023, from https://grace.jpl.nasa.gov/data/grace-months

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

National Research Council. (1991). *Opportunities in the Hydrologic Sciences*. The National Academies Press. https://doi.org/10.17226/1543

NCAR. (2023). NDVI and EVI: Vegetation Indices (MODIS). https://climatedataguide.ucar.edu/climate-data/ndvi-and-evi-vegetation-indices-modis

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., & Gupta, H. V. (2021). What Role Does Hydrological Science Play in the Age of Machine Learning? *Water Resources Research*, *57*(e2020WR028091). https://doi.org/10.1029/2020WR028091

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R. J., Blodgett, D., Brekke, L., & Arnold, J. R. (2015). Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: Data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, *19*(1), 209–223. https://doi.org/10.5194/hess-19-209-2015

Njoku, E., & Li, L. (1999). Retrieval of land surface parameters using passive microwave measurements at 6-18 GHz. *IEEE Transactions on Geoscience and Remote Sensing*, *37*(1), 79–93. https://doi.org/10.1109/36.739125

NOAA. (2023). Multivariate ENSO Index Version 2 (MEI.v2). https://psl.noaa.gov/enso/mei/

Paca, V. H. d. M., Espinoza-Dávalos, G. E., Hessels, T. M., Moreira, D. M., Comair, G. F., & Bastiaanssen, W. G. M. (2019). The spatial variability of actual evapotranspiration across the Amazon River Basin based on remote sensing products validated with flux towers. *Ecological Processes*, *8*(1), 6. https://doi.org/10.1186/s13717-019-0158-8

Pan, M., Fisher, C. K., Chaney, N. W., Zhan, W., Crow, W. T., Aires, F., Entekhabi, D., & Wood, E. F. (2015). Triple collocation: Beyond three estimates and separation of structural/non-structural errors. *Remote Sensing of Environment*, *171*, 299–310. https://doi.org/10.1016/j.rse.2015.10.028

Pan, M., Sahoo, A. K., Troy, T. J., Vinukollu, R. K., Sheffield, J., & Wood, E. F. (2012). Multisource estimation of long-term terrestrial water budget for major global river basins. *Journal of Climate*, *25*(9), 3191–3206. https://doi.org/10.1175/JCLI-D-11-00300.1

Pan, M., & Wood, E. F. (2006). Data assimilation for estimating the terrestrial water budget using a constrained ensemble Kalman filter. *Journal of Hydrometeorology*, *7*(3), 534–547. https://doi.org/10.1175/JHM495.1

Papa, F., Güntner, A., Frappart, F., Prigent, C., & Rossow, W. B. (2008). Variations of surface water extent and water storage in large river basins: A comparison of different global data sources. *Geophysical Research Letters*, *35*(11), 2008GL033857. https://doi.org/10.1029/2008GL033857

Pascolini-Campbell, M. A., Reager, J. T., & Fisher, J. B. (2020). GRACE-based mass conservation as a validation target for basin-scale evapotranspiration in the contiguous United States. *Water Resources Research*, *56*(2), e2019WR026594. https://doi.org/10.1029/2019WR026594

Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.-W., Poindexter, C., Chen, J., Elbashandy, A., Humphrey, M., Isaac, P., Polidori, D., Reichstein, M., Ribeca, A., van Ingen, C., Vuichard, N., & Zhang, L. (2020). The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. *Scientific Data*, *7*(1), 225. https://doi.org/10.1038/s41597-020-0534-3

Pebesma, E., & Graeler, B. (2023). Gstat: Spatial and Spatio-Temporal Geostatistical Modelling, Prediction and Simulation. Retrieved August 27, 2023, from https://cran.r-project.org/web/packages/gstat/index.html

Pebesma, E. J., & Wesseling, C. G. (1998). Gstat: A program for geostatistical modelling, prediction and simulation [Publisher: Elsevier]. *Computers & Geosciences*, *24*(1), 17–31. https://doi.org/10.1016/S0098-3004(97)00082-4

Peel, M. C., & McMahon, T. A. (2020). Historical development of rainfall-runoff modeling. *WIREs Water*, *7*(5). https://doi.org/10.1002/wat2.1471

Pellet, V., Aires, F., Mariotti, A., & Fernández-Prieto, D. (2018). Analyzing the Mediterranean Water Cycle Via Satellite Data Integration. *Pure and Applied Geophysics*, *175*(11), 3909–3937. https://doi.org/10.1007/s00024-018-1912-z

Pellet, V., Aires, F., Munier, S., Fernández Prieto, D., Jordá, G., Dorigo, W. A., Polcher, J., & Brocca, L. (2019). Integrating multiple satellite observations into a coherent dataset to monitor the full water cycle–application to the Mediterranean region. *Hydrology and Earth System Sciences*, *23*(1), 465–491. https://doi.org/10.5194/hess-23-465-2019

Pellet, V., Aires, F., Papa, F., Munier, S., & Decharme, B. (2019). Long-term Total Water Storage Change from a Satellite Water Cycle Reconstruction over large south Asian basins. *Hydrology and Earth System Sciences Discussions*, 1–30. https://doi.org/10.5194/hess-2019-262

Pellet, V., Aires, F., Yamazaki, D., & Papa, F. (2021). Coherent Satellite Monitoring of the Water Cycle Over the Amazon. Part 1: Methodology and Initial Evaluation. *Water Resources Research*, *57*(5), 21. https://doi.org/10.1029/2020WR028647

Pfister, L. (2018). Leonardo da Vinci's conceptualisations of the water cycle, 13913. Retrieved October 30, 2023, from https://ui.adsabs.harvard.edu/abs/2018EGUGA..2013913P

Pimentel, E. T., & Hamza, V. M. (2011). Indications of an underground river beneath the Amazon River: Inferences from results of geothermal studies. *12th International Congress of the Brazilian Geophysical Society*. Retrieved November 8, 2023, from https://www.earthdoc.org/content/papers/10.3997/2214-4609-pdb.264.SBGF_3139

Prigent, C., Lettenmaier, D. P., Aires, F., & Papa, F. (2016). Toward a High-Resolution Monitoring of Continental Surface Water Extent and Dynamics, at Global Scale: From GIEMS (Global Inundation Extent from Multi-Satellites) to SWOT (Surface Water Ocean Topography). *Surveys in Geophysics*, *37*(2), 339–355. https://doi.org/10.1007/s10712-015-9339-x

Rasouli, K., Scharold, K., Mahmood, T. H., Glenn, N. F., & Marks, D. (2020). Linking hydrological variations at local scales to regional climate teleconnection patterns. *Hydrological Processes*, *34*(26). https://doi.org/10.1002/hyp.13982

Reager, J., Thomas, A., Sproles, E., Rodell, M., Beaudoing, H., Li, B., & Famiglietti, J. S. (2015). Assimilation of GRACE Terrestrial Water Storage Observations into a Land Surface Model for the Assessment of Regional Flood Potential. *Remote Sensing*, *7*(11), 14663–14679. https://doi.org/10.3390/rs71114663

Richey, A. S., Thomas, B. F., Lo, M.-H., Reager, J. T., Famiglietti, J. S., Voss, K., Swenson, S., & Rodell, M. (2015). Quantifying renewable groundwater stress with GRACE. *Water Resources Research*, *51*(7), 5217–5238. https://doi.org/10.1002/2015WR017349

Rodell, M., Houser, P. R., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C.-J., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., Entin, J. K., Walker, J. P., Lohmann, D., & Toll, D. (2004). The Global Land Data Assimilation System. *Bulletin of the American Meteorological Society*, *85*(3), 381–394. https://doi.org/10.1175/BAMS-85-3-381

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. (2015). The observed state of the water cycle in the early twenty-first century. *Journal of Climate*, *28*(21), 8289–8318. https://doi.org/10.1175/JCLI-D-14-00555.1

Rodell, M., Chen, J., Kato, H., Famiglietti, J. S., Nigro, J., & Wilson, C. R. (2007). Estimating groundwater storage changes in the Mississippi River basin (USA) using GRACE. *Hydrogeology Journal*, *15*(1), 159–166. https://doi.org/10.1007/s10040-006-0103-7

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

Rodell, M., McWilliams, E. B., Famiglietti, J. S., Beaudoing, H. K., & Nigro, J. (2011). Estimating evapotranspiration using an observation based terrestrial water budget. *Hydrological Processes*, *25*(26), 4082–4092. https://doi.org/10.1002/hyp.8369

Rodgers, C. D. (2000). *Inverse Methods for Atmospheric Sounding: Theory and Practice* (Vol. 2). World Scientific. https://doi.org/10.1142/3171

Rosbjerg, D., & Rodda, J. (2019). IAHS: A brief history of hydrology. *History of Geo- and Space Sciences*, *10*, 109–118. https://doi.org/10.5194/hgss-10-109-2019

Roux, H., Amengual, A., Romero, R., Bladé, E., & Sanz-Ramos, M. (2020). Evaluation of two hydrometeorological ensemble strategies for flash-flood forecasting over a catchment of the eastern Pyrenees. *Natural Hazards and Earth System Sciences*, *20*(2), 425–450. https://doi.org/10.5194/nhess-20-425-2020

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1987). Learning Internal Representations by Error Propagation. In D. E. Rumelhart & J. L. McLelland (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations* (pp. 318–362). MIT Press. https://ieeexplore.ieee.org/servlet/opac?bknumber=6276825

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Sachindra, D. A., Huang, F., Barton, A., & Perera, B. J. C. (2014). Statistical downscaling of general circulation model outputs to precipitation—part 2: Bias-correction and future projections. *International Journal of Climatology*, *34*(11), 3282–3303. https://doi.org/10.1002/joc.3915

Sahoo, A. K., Pan, M., Troy, T. J., Vinukollu, R. K., Sheffield, J., & Wood, E. F. (2011). Reconciling the global terrestrial water budget using satellite remote sensing. *Remote Sensing of Environment*, *115*(8), 1850–1865. https://doi.org/10.1016/j.rse.2011.03.009

Sakumura, C., Bettadpur, S., & Bruinsma, S. (2014). Ensemble prediction and intercomparison analysis of GRACE time-variable gravity field models. *Geophysical Research Letters*, *41*(5), 1389–1397. https://doi.org/10.1002/2013GL058632

Sauer, V. B., & Meyer, R. W. (1992). *Determination of error in individual discharge measurements* (Open-File Report No. 92-144). US Geological Survey; Books and Open-File Reports Section [distributor], https://pubs.usgs.gov/of/1992/ofr92-144/

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

Save, H., Bettadpur, S., & Tapley, B. D. (2016). High-resolution CSR GRACE RL05 mascons. *Journal of Geophysical Research: Solid Earth*, *121*(10), 7547–7569. https://doi.org/10.1002/2016JB013007

Scanlon, B. R., Zhang, Z., Rateb, A., Sun, A., Wiese, D., Save, H., Beaudoing, H., Lo, M. H., Müller-Schmied, H., Döll, P., van Beek, R., Swenson, S., Lawrence, D., Croteau, M., & Reedy, R. C. (2019). Tracking Seasonal Fluctuations in Land Water Storage Using Global Models and GRACE Satellites. *Geophysical Research Letters*, *46*(10), 5254–5264. https://doi.org/10.1029/2018GL081836

Scanlon, B. R., Zhang, Z., Save, H., Wiese, D. N., Landerer, F. W., Long, D., Longuevergne, L., & Chen, J. (2016). Global evaluation of new GRACE mascon products for hydrologic applications. *Water Resources Research*, *52*(12), 9412–9429. https://doi.org/10.1002/2016WR019494

Schillings, C., & Stuart, A. M. (2017). Analysis of the Ensemble Kalman Filter for Inverse Problems. *SIAM Journal on Numerical Analysis*, *55*(3). https://doi.org/10.1137/16m105959x

Schlosser, C. A., & Houser, P. R. (2007). Assessing a Satellite-Era Perspective of the Global Water Cycle. *Journal of Climate*, *20*(7), 1316–1338. https://doi.org/10.1175/JCLI4057.1

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Schreiber, P., & Demuth, S. (1997). Regionalization of low flows in southwest Germany. *Hydrological Sciences Journal*, *42*(6), 845–858. https://doi.org/10.1080/02626669709492083

Seyyedi, H., Anagnostou, E. N., Beighley, E., & McCollum, J. (2015). Hydrologic evaluation of satellite and reanalysis precipitation datasets over a mid-latitude basin. *Atmospheric Research*, *164-165*, 37–48. https://doi.org/10.1016/j.atmosres.2015.03.019

Shang, K., Yao, Y., Liang, S., Zhang, Y., Fisher, J. B., Chen, J., Liu, S., Xu, Z., Zhang, Y., Jia, K., Xiaotong Zhang, Junming Yang, Xiangyi Bei, Xiaozheng Guo, Ruiyang Yu, & Zijing Xie. (2021). DNN-MET: A deep neural networks method to integrate satellite-derived evapotranspiration products, eddy covariance observations and ancillary information. *Agricultural and Forest Meteorology*, *308*, 108582. https://doi.org/10.1016/j.agrformet.2021.108582

Sheffield, J., Ferguson, C. R., Troy, T. J., Wood, E. F., & McCabe, M. F. (2009). Closing the terrestrial water budget from satellite remote sensing. *Geophysical Research Letters*, *36*(7). https://doi.org/10.1029/2009gl037338

Shiklomanov, I. A. (2009). World Water Balance. In *Hydrological Cycle, Volume 2* (p. 6). UNESCO - Encyclopedia Life Support Systems (UNESCO-EOLSS). https://books.google.fr/books?id=KAiCCwAAQBAJ

Shuttleworth, A. (1993). Evaporation. In D. Maidment (Ed.), *Handbook of Hydrology* (1st edition, Chapter 4). McGraw Hill.

Siebert, S., Kummu, M., Porkka, M., Döll, P., Ramankutty, N., & Scanlon, B. R. (2015). A global data set of the extent of irrigated land from 1900 to 2005. *Hydrology and Earth System Sciences*, *19*(3), 1521–1545. https://doi.org/10.5194/hess-19-1521-2015

Singer, M. B., Asfaw, D. T., Rosolem, R., Cuthbert, M. O., Miralles, D. G., MacLeod, D., Quichimbo, E. A., & Michaelides, K. (2021). Hourly potential evapotranspiration at 0.1° resolution for the global land surface from 1981-present. *Scientific Data*, *8*(1), 224. https://doi.org/10.1038/s41597-021-01003-9

Slack, J. R., & Landwehr, J. M. (1992). *Hydro-climatic data network (HCDN); a U.S. Geological Survey streamflow data set for the United States for the study of climate variations, 1874-1988* (USGS Numbered Series No. 92-129). U.S. Geological Survey ; Copies of this report can be purchased from USGS Books and Open-File Reports Section, https://doi.org/10.3133/ofr92129

Sneeuw, N., Lorenz, C., Devaraju, B., Tourian, M. J., Riegger, J., Kunstmann, H., & Bárdossy, A. (2014). Estimating runoff using hydro-geodetic approaches. *Surveys in Geophysics*, *35*(6), 1333–1359. https://doi.org/10.1007/s10712-014-9300-4

Song, Z., Xia, J., Wang, G., She, D., Hu, C., & Hong, S. (2022). Regionalization of hydrological model parameters using gradient boosting machine. *Hydrology and Earth System Sciences*, *26*(2), 505–524. https://doi.org/10.5194/hess-26-505-2022

Stephens, G. L., Li, J., Wild, M., Clayson, C. A., Loeb, N., Kato, S., L'Ecuyer, T., Stackhouse, P. W., Lebsock, M., & Andrews, T. (2012). An update on Earth's energy balance in light of the latest global observations. *Nature Geoscience*, *5*(10), 691–696. https://doi.org/10.1038/ngeo1580

Svarovsky, A. (2019). Gravity Recovery and Climate Experiment-Follow-On (GRACE-FO) Mission. Retrieved July 13, 2023, from https://www.gfz-potsdam.de/en/section/global-geomonitoring-and-gravity-field/projects/gravity-recovery-and-climate-experiment-follow-on-grace-fo-mission

Syed, T. H., Famiglietti, J. S., Chen, J., Rodell, M., Seneviratne, S. I., Viterbo, P., & Wilson, C. R. (2005). Total basin discharge for the Amazon and Mississippi River basins from GRACE and a land-atmosphere water balance. *Geophysical Research Letters*, *32*(24), L24404. https://doi.org/10.1029/2005GL024851

Tapley, B. D., Bettadpur, S., Ries, J. C., Thompson, P. F., & Watkins, M. M. (2004). GRACE measurements of mass variability in the Earth system. *Science*, *305*(5683), 503–505. https://doi.org/10.1126/science.1099192

Tarek, M., Brissette, F., & Arsenault, R. (2020). Large-Scale Analysis of Global Gridded Precipitation and Temperature Datasets for Climate Change Impact Studies. *Journal of Hydrometeorology*, *21*, 1–54. https://doi.org/10.1175/JHM-D-20-0100.1

Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä., Pedaste, M., Raju, M., Astapova, A., Lukner, H., Kogermann, K., & Sepp, T. (2021). Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific Data*, *8*(1), 192. https://doi.org/10.1038/s41597-021-00981-0

Thomann, R. V. (1982). Verification of water quality models. *Journal of the Environmental Engineering Division, American Society of Civil Engineers*, *108*(5), 923–940. https://doi.org/10.1061/JEEGAV.0001352

Thomas, C. M., Dong, B., & Haines, K. (2020). Inverse Modeling of Global and Regional Energy and Water Cycle Fluxes using Earth Observation Data. *Journal of Climate*, *33*(5), 1707–1723. https://doi.org/10.1175/JCLI-D-19-0343.1

Tian, Y., & Peters-Lidard, C. D. (2010). A global map of uncertainties in satellite-based precipitation measurements. *Geophysical Research Letters*, *37*(24). https://doi.org/10.1029/2010GL046008

Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, *46*, 234. https://doi.org/10.2307/143141

Trabucco, A., & Zomer, R. (2019). Global Aridity Index and Potential Evapotranspiration (ET0) Climate Database v2. https://doi.org/10.6084/m9.figshare.7504448.v3

US Library of Congress. (2005). Country Profile: Philippines. Retrieved July 11, 2023, from https://web.archive.org/web/20050717172656/http://lcweb2.loc.gov/frd/cs/profiles/Philippines.pdf

USGS. (2022a). Landsat Enhanced Vegetation Index. Retrieved February 28, 2023, from https://www.usgs.gov/landsat-missions/landsat-enhanced-vegetation-index

USGS. (2022b). New USGS diagram re-envisions how Earth's most precious commodity cycles the planet. Retrieved July 11, 2023, from https://www.usgs.gov/news/national-news-release/new-usgs-diagram-re-envisions-how-earths-most-precious-commodity-cycles

Vermote, E. (2018). NOAA Climate Data Record (CDR) of AVHRR Normalized Difference Vegetation Index (NDVI), Version 5. https://doi.org/10.7289/V5ZG6QH9

Vermote, E. (2022). NOAA Climate Data Record (CDR) of VIIRS Normalized Difference Vegetation Index (NDVI), Version 1. https://doi.org/10.25921/GAKH-ST76

Vogel, R. M. (1986). The Probability Plot Correlation Coefficient Test for the Normal, Lognormal, and Gumbel Distributional Hypotheses. *Water Resources Research*, *22*(4), 587–590. https://doi.org/10.1029/WR022i004p00587

von Rueden, L., Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., Kirsch, B., Pfrommer, J., Pick, A., Ramamurthy, R., Walczak, M., Garcke, J., Bauckhage, C., & Schuecker, J. (2023). Informed Machine Learning – A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems. *IEEE Transactions on Knowledge and Data Engineering*, *35*(1), 614–633. https://doi.org/10.1109/TKDE.2021.3079836

Vorosmarty, C. J., Green, P., Salisbury, J., & Lammers, R. B. (2000). Global water resources: Vulnerability from climate change and population growth. *Science*, *289*(5477), 284–288. https://doi.org/10.1126/science.289.5477.28

Wagener, T., Sivapalan, M., McDonnell, J., Hooper, R., Lakshmi, V., Liang, X., & Kumar, P. (2004). Predictions in ungauged basins as a catalyst for multidisciplinary hydrology.

Eos, Transactions American Geophysical Union, *85*(44), 451–457. https://doi.org/10.1029/2004EO440003

Wan, Z., Hook, S., & Hulley, G. (2021). MODIS/Terra Land Surface Temperature/Emissivity Monthly L3 Global 0.05Deg CMG V061. https://doi.org/10.5067/MODIS/MOD11C3.061

Wang, X., de Linage, C., Famiglietti, J. S., & Zender, C. S. (2011). Gravity Recovery and Climate Experiment (GRACE) detection of water storage changes in the Three Gorges Reservoir of China and comparison with in situ measurements. *Water Resources Research*, *47*(12). https://doi.org/10.1029/2011WR010534

Wang, Y., & Wu, Q. (2022). Comparison of Multi-Satellite Precipitation Data from the Global Precipitation Measurement Mission and Tropical Rainfall Measurement Mission Datasets: Seasonal and Diurnal Cycles (S. Bonafoni, Ed.). *Advances in Meteorology*, *2022*, 1–20. https://doi.org/10.1155/2022/6404243

Watkins, M. M., Wiese, D. N., Yuan, D.-N., Boening, C., & Landerer, F. W. (2015). Improved methods for observing Earth's time variable mass distribution with GRACE using spherical cap mascons. *Journal of Geophysical Research: Solid Earth*, *120*(4), 2648–2671. https://doi.org/10.1002/2014JB011547

WMO. (1989). *The Global Water Runoff Data Project: Workshop on the Global Runoff Data Set and Grid Estimation*. World Meteorological Organization.

Wong, J. S., Zhang, X., Gharari, S., Shrestha, R. R., Wheater, H. S., & Famiglietti, J. S. (2021). Assessing Water Balance Closure Using Multiple Data Assimilation–and Remote Sensing–Based Datasets for Canada. *Journal of Hydrometeorology*, *22*(6), 1569–1589. https://doi.org/10.1175/JHM-D-20-0131.1

Xie, J., Liu, X., Bai, P., & Liu, C. (2022). Rapid Watershed Delineation Using an Automatic Outlet Relocation Algorithm. *Water Resources Research*, *58*(e2021WR031129). https://doi.org/10.1029/2021WR031129

Xie, P., Chen, M., Yang, S., Yatagai, A., Hayasaka, T., Fukushima, Y., & Liu, C. (2007). A Gauge-Based Analysis of Daily Precipitation over East Asia. *Journal of Hydrometeorology*, *8*(3), 607–626. https://doi.org/10.1175/JHM583.1

Xie, P., Robert Joyce, Shaorong Wu, Yoo, S.-H., Yelena Yarosh, Fengying Sun, & Roger Lin. (2019). NOAA Climate Data Record (CDR) of CPC Morphing Technique (CMORPH) High Resolution Global Precipitation Estimates, Version 1. https://doi.org/10.25921/w9va-q159

Xu, T., & Liang, F. (2021). Machine learning for hydrologic sciences: An introductory overview. *WIREs Water*, *8*(5), e1533. https://doi.org/10.1002/wat2.1533

Yamazaki, D., Ikeshima, D., Sosa, J., Bates, P. D., Allen, G. H., & Pavelsky, T. M. (2019). MERIT Hydro: A High-Resolution Global Hydrography Map Based on Latest Topography Dataset. *Water Resources Research*, *55*(6), 5053–5073. https://doi.org/10.1029/2019WR024873

Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J. C., Sampson, C. C., Kanae, S., & Bates, P. D. (2017). A high-accuracy map of global terrain elevations: Accurate Global Terrain Elevation map. *Geophysical Research Letters*, *44*(11), 5844–5853. https://doi.org/10.1002/2017GL072874

Yi, S., & Sneeuw, N. (2021). Filling the Data Gaps Within GRACE Missions Using Singular Spectrum Analysis. *Journal of Geophysical Research: Solid Earth*, *126*(5). https://doi.org/10.1029/2020JB021227

Yilmaz, M. T., DelSole, T., & Houser, P. R. (2011). Improving land data assimilation performance with a water budget constraint. *Journal of Hydrometeorology*, *12*(5), 1040–1055. https://doi.org/10.1175/2011JHM1346.1

Zaitchik, B. F., Rodell, M., & Reichle, R. H. (2008). Assimilation of GRACE Terrestrial Water Storage Data into a Land Surface Model: Results for the Mississippi River Basin. *Journal of Hydrometeorology*, *9*(3), 535–548. https://doi.org/10.1175/2007JHM951.1

Zaki, N. A., Haghighi, A. T., Rossi, P. M., Tourian, M. J., & Kløve, B. (2019). Monitoring Groundwater Storage Depletion Using Gravity Recovery and Climate Experiment (GRACE) Data in Bakhtegan Catchment, Iran. *Water*, *11*(7). https://doi.org/10.3390/w11071456

Zhang, J. (2019). *Assessing the statistical relations of terrestrial water mass change with hydrological variables and climate variability* (PhD Thesis). Universität Stuttgart. Germany. https://doi.org/0.18419/opus-10474

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., & Wood, E. F. (2018). A Climate Data Record (CDR) for the global terrestrial water budget: 1984–2010. *Hydrology and Earth System Sciences*, *22*(1), 241–263. https://doi.org/10.5194/hess-22-241-2018

Zhang, Y., Pan, M., & Wood, E. F. (2016). On creating global gridded terrestrial water budget estimates from satellite remote sensing. In *Remote Sensing and Water Resources* (pp. 59–78). Springer. https://link.springer.com/chapter/10.1007/978-3-319-32449-4_4

Zwieback, S., Scipal, K., Dorigo, W., & Wagner, W. (2012). Structural and statistical properties of the collocation technique for error characterization. *Nonlinear Processes in Geophysics*, *19*(1), 69–80. https://doi.org/10.5194/npg-19-69-2012