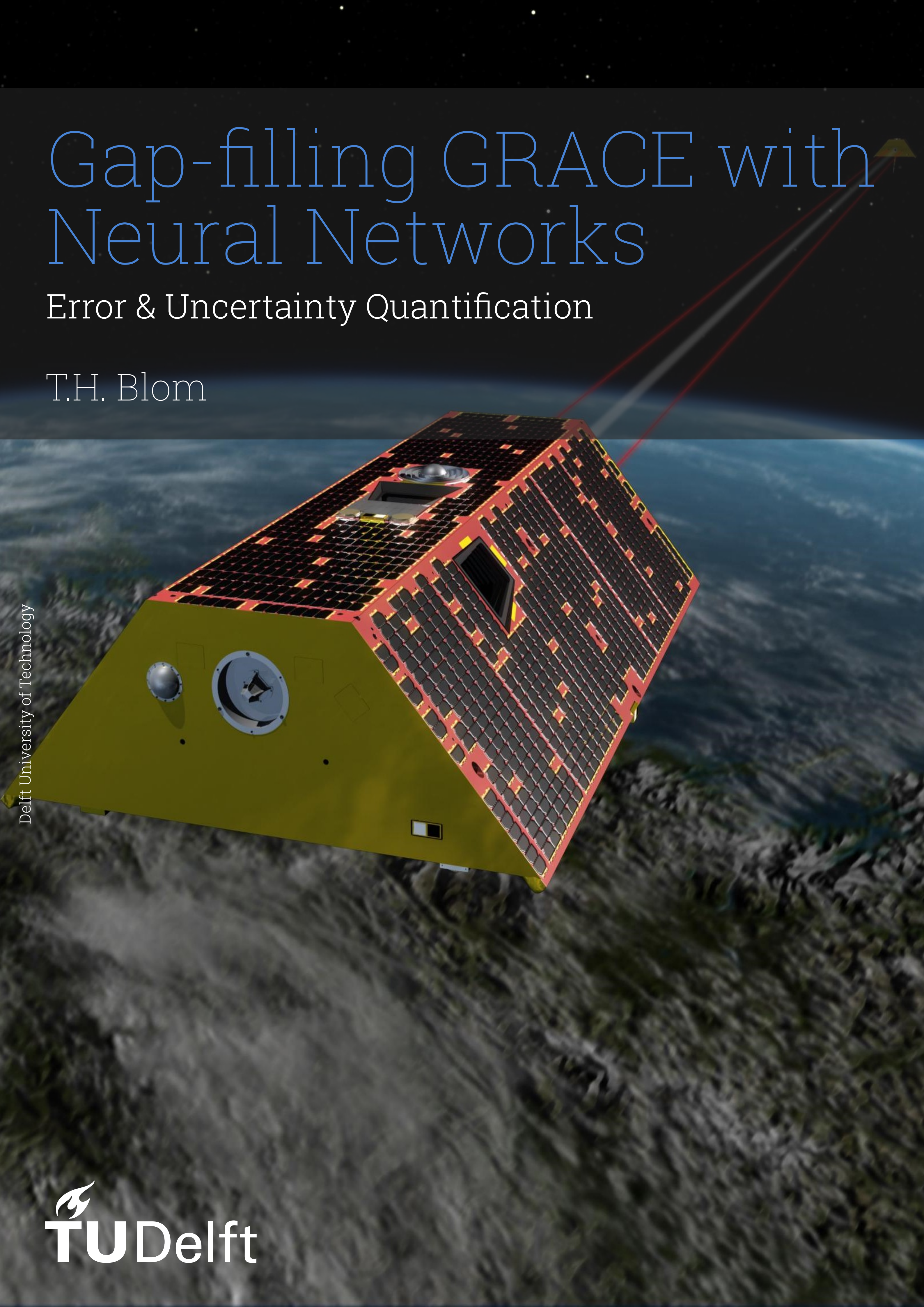
Gap-filling GRACE with Neural Networks

Error & Uncertainty Quantification

T.H. Blom



Delft University of Technology

Gap-filling GRACE with Neural Networks

Error & Uncertainty Quantification

by

T.H. Blom

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Monday January 24, 2025 at 09:00 AM.

Student number:	4793021	
Project duration:	February 12, 2024 – December 19, 2024	
Thesis committee:	Dr. ir. J. G. De Teixeira da Encarnação,	TU Delft, supervisor
	Dr. N. A. K. Doan,	TU Delft
	Dr. ir. M. J. Heiligers	TU Delft

Cover:	©NASA, GRACE-FO (PIA22445)
Style:	TU Delft Report Style, with modifications by Daan Zwaneveld

TU Delft

Preface

Writing this thesis has been quite the journey. I have learned more than I could have anticipated at the start of this project: both in technical areas, such as machine learning and spherical harmonic synthesis, and in understanding myself and my working methods. One particularly important lesson I have learnt, is to think before asking questions and to formulate my words more carefully. This is a skill I am still striving to master.

I would like to express my gratitude to my supervisor, João, for his time and invaluable guidance throughout this process. Furthermore, I would also like to thank Anh Khoa for his valuable perspective on neural networks and machine learning. I am also deeply thankful to my father and Martin Barragan for dedicating their valuable time to proofreading my thesis. Finally, I extend my sincere thanks to my girlfriend, mother, and brothers for their support during this journey.

T.H. Blom
Delft, January 2025

Contents

Abstract

Since the launch of the dual-satellite Gravity Recovery and Climate Experiment (GRACE) mission by NASA and DLR in 2002, the mission has become invaluable to providing information about the redistribution of mass on Earth's surface over time. Using an inter-satellite ranging system, the gravity field of Earth is estimated on a monthly basis. The temporal changes in these estimates allow for the redistribution of mass on Earth's surface to be quantified in terms of changes of equivalent water height (EWH) on a monthly basis. This information has been key to understanding fresh water river basin cycles, monitoring loss of ice mass in Greenland and Antarctica, and observing changes in groundwater storage. These insights help build an understanding of the effects of climate change on Earth. Unfortunately, data collection of the GRACE mission has not been complete followed by a long gap between GRACE and GRACE-FO. These gaps in the data are a hindrance to creating accurate climate models.

Estimating the data that should have been observed during these gaps is called gap-filling. The focus of this work is on using models that rely on alternative sources of data which are available during these gaps. Alternative sources are advantageous because these can contain information about irregular events that occurred during these gaps. Using gap-filled estimates, climate models can be made more accurate and thus lead to a better understanding of climate change. Attempts have been made to fill the gaps using neural networks (e.g. Harrison, 2023; Keleş, 2022). Whilst there is quantification of the errors in these models, there has been little analysis of the epistemic uncertainties inherent to neural networks and the aleatoric uncertainties in the input data. This discussion is important because the projection of climate into the future can be made more precise using error and uncertainty estimates of the spatial variables used.

This thesis addresses methods and calculations to quantify the errors and uncertainty of gap-filled data through use of neural networks. Neural networks based on the architecture of Harrison (2023) that use ESA's Swarm EWH data and NASA's Global Land Data Assimilation System (GLDAS) soil moisture data as inputs to predict GRACE EWH are analysed. Through quantification of the errors in these datasets, the effects of aleatoric uncertainty are simulated. Using different seed numbers, the epistemic uncertainties are quantified and investigated. It is found that generating additional training data through use of the quantified input errors can lead to reduced errors and uncertainty in the gap-filled data.

An experiment is designed which consists of training 20400 models to predict EWH over a selection of relevant river basins, being Amazon, Congo, Mississippi, and Nile river basins. The combination of all model outputs is used to quantify the errors and uncertainty in the gap-filled data. It is conclusively shown that neural networks outperform Swarm in predicting EWH on the basin level showing reductions in root-mean-squared (RMS) of model EWH w.r.t. GRACE EWH in the order of 2 to 3 [cm].

Training a neural network using additionally sampled data results in significant RMS error reductions relative to Swarm EWH. The sampling of additional data using input error quantifications gives the neural networks information about the errors in the input data. Furthermore, the uncertainty of models trained with additionally sampled data leads to reductions in uncertainty with respect to models trained without additionally sampled data. It is also shown that neural networks trained with additional data are less sensitive to epistemic uncertainties. The models trained with additionally sampled training datasets approach error levels within GRACE's own indication of errors. From this it is included that these neural networks successfully produce GRACE-like data.

Given the promising results of sampling additional training data, it is recommended that this avenue is explored further by also adding noise to the target GRACE data. Furthermore, recommendations are also made to improve the error quantification of Swarm and the GLDAS soil moisture data. The final recommendation is to also investigate other neural network types such as convolutional and bayesian convolutional neural networks and quantify their errors and uncertainties using the methods presented.

List of Abbreviations

Abbreviation	Value
GLDAS	Global Land Data Assimilation System
GPS	Global Positioning System
GRACE	Gravity Recovery and Climate Experiment
GRACE-FO	GRACE Follow-On
EWH	Equivalent water height
LSM	Land Surface Model
SH	Spherical Harmonics
CSR	Center for Space Research
DLR	German Aerospace Centre
ESA	European Space Agency
GFZ	German Research Center for Geosciences
JPL	Jet Propulsion Laboratory
NASA	National Aeronautics and Space Administration
BCNN	Bayesian convolutional neural network
BNN	Bayesian neural network
CNN	Convolutional neural network
NN	Neural network
ReLU	Rectified Linear Unit
CC	Correlation coefficient
MSE	Mean-squared error
NSE	Nash-Sutcliffe efficiency
RMS	Root mean-squared
RMSE	Root mean-squared error
SNR	Signal-to-noise ratio

List of Symbols

Symbol	Value
ξ	Equivalent water height
C	Spherical harmonic coefficient
S	Spherical harmonic coefficient
l	Degree
m	Order
f	Spherical harmonic functional
F	Spherical harmonic scaling factor (in Chapter 2) or fraction (in Chapter 6) or factor (in Chapter 7) or F-statistic (in Chapter 9)
Θ	Latitude
ϕ	Longitude
$\overline{P}$	Normalized Legendre polynomial
k	Load Love number
R	Radius
ρ	Density or Pearson CC
Θ	Soil moisture
N	Amount
σ	Standard deviation
ϵ	Error
μ	Mean
κ	Trainable parameter of softplus activation function
α	Seed number unique to training of NN
λ	Seed number unique to testing of NN
β	Basin
η	Number of additional training data sets
b	Coastal buffer size
r	Gaussian smoothing radius
T	Total time
m	Month
$\mathcal{M}$	Available GRACE months
$\mathcal{G}$	Gap months
γ	Training data fraction

List of Figures

List of Tables

1 Introduction

The National Aeronautics and Space Administration (NASA) and the German Aerospace Centre (DLR) launched the Gravity Recovery and Climate Experiment (GRACE) mission in 2002. The objective of this mission was to collect data about Earth's gravity field for the purpose of understanding the effect of climate change on Earth. The GRACE mission consists of two satellites that fly about 220 [km] apart in a near-polar orbit at an altitude of approximately 500 [km]. They have a micro-wave ranging system between them which reports the distance between the two satellites. Due to mass variations on Earth's surface, the accelerations both satellites experience are not the same. As a result, the distance between the satellites is not constant. Institutes such as Jet Propulsion Laboratory (JPL), German Research Center for Geosciences (GFZ), and Center for Space Research (CSR) use these distance measurements to estimate monthly gravity field solutions in the form of spherical harmonic (SH) coefficients (Dahle et al., 2019; Save, 2019; Yuan, 2019).

As a result of battery issues, power was conserved to prolong the mission. To do so, data collection had to be turned off for some periods. Subsequently, there are periods without gravity field solutions. These battery issues led to the decommissioning of the GRACE mission in 2017. Eleven months later, a new and improved satellite mission continued the role of GRACE, namely GRACE Follow-On (GRACE-FO). Henceforth, the name GRACE will be used to reference both GRACE and GRACE-FO, considering the two missions as a single one. Figure 1.1 shows when GRACE data is available. There are two types of data gaps in the GRACE data:

- Smaller gaps within the GRACE or GRACE-FO mission of a duration between 1 and 95 days.
- A single larger gap of eleven months between the GRACE and GRACE-FO mission (between the gray dashed lines in Figure 1.1).

The data availability in Table B.1 shows specifically which months are available and which months are not available.

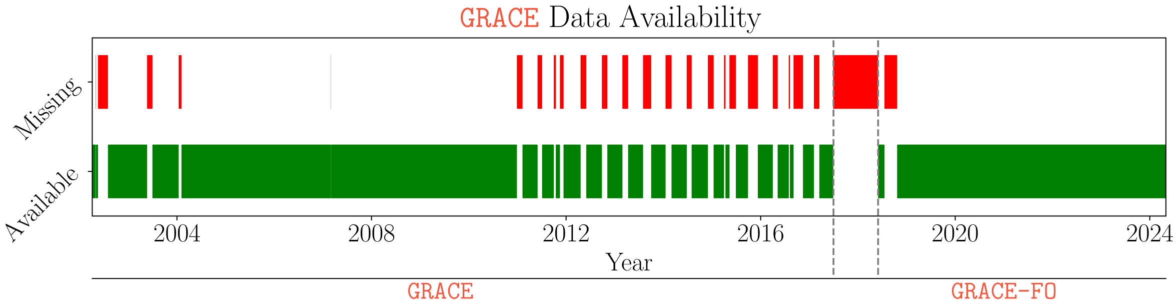


Figure 1.1: A visualization of when GRACE data is available and when it is not.

The handling of the GRACE data gaps is known as gap-filling and has become a popular research topic over the past few years. Gap-filling is defined as estimating the data that would have been observed during the missing months. In the context of GRACE gap-filling, GRACE-like data is data with a similar spatial resolution as GRACE. GRACE has been gap-filled using machine learning methods (e.g. Harrison, 2023; Keleş, 2022). However, the discussion on the errors and particularly the uncertainty of these GRACE-like results has been minimal. There is little to no discussion on the randomness introduced in the training process of neural network (NN)s (e.g. Harrison, 2023; Keleş, 2022).

This thesis covers gap-filling GRACE using a fully-connected NN, with a specific focus on the errors and uncertainty of the resulting GRACE-like data. Error and uncertainty estimates will allow researchers using the GRACE-like data to know the extent to which they can trust the gap-filled data. Climate models making projections into the future can be made more precise using error and uncertainty estimates of the spatial variables used (Wu et al., 2022). Based on this need for error and uncertainty estimates, the following two research questions are formulated:

- What are the errors and uncertainty of GRACE-like data produced by NNs?
- How does the inclusion of additional training data generated, using errors in the auxiliary datasets, affect the errors and uncertainty of GRACE-like data produced by NNs?

The general concept of a NN is to relate input datasets to target output datasets. When target data is not available, a trained NN with input data can be used to make a prediction of what the target data looks like. In this case, we want a NN to produce GRACE-like data for the months in which GRACE data is unavailable. The input data must be some dataset that is related to the GRACE data, such that the NN can learn the relationship between the two datasets and produce GRACE-like data when only the input data is available. For this reason, there are two requirements on the input data:

- To be available during the available GRACE months and the months which have to be estimated.
- To contain information about the mass distribution on Earth's surface.

If the input data is not available during GRACE's missing months, then a NN can not be used to produce GRACE-like data during these months. If the input datasets do not contain in part the same signals as the GRACE data, a NN will not be able to deduce a relationship between the input datasets and the GRACE data.

For this thesis, two input datasets are used as input to the NNs. One input dataset originates from the Swarm mission launched by European Space Agency (ESA) in 2013. The objective of the Swarm mission is to study Earth's magnetic field. However, Teixeira da Encarnação et al. (2016) use the GPS data collected by Swarm to create kinematic orbits. These orbits are used to determine monthly SH gravity field solutions of the Earth. Teixeira da Encarnação et al. (2020) conclude that over land the agreement of Swarm and GRACE gravity field solutions is within 4 [cm] equivalent water height (EWH) and that the SH coefficients of monthly Swarm solutions are reliable up to degree and order (d/o) 12. Several other gap-filling attempts with NNs have utilized Swarm data (Forootan et al., 2020; Harrison, 2023; Keleş, 2022).

As the changes in gravity field solutions from month to month are dominated by hydrological processes on Earth's surface, hydrological data products can also effectively be used as inputs for producing GRACE-like data (Harrison, 2023; Keleş, 2022; Mo et al., 2022). NASA's Global Land Data Assimilation System (GLDAS) provides monthly data products regarding Earth's surface which are known as land surface models (LSMs) (Rodell et al., 2004). These LSMs contain spatial variables such as soil moisture, snow cover, leaf area index, and more. For this thesis, the soil moisture variable is utilized. This data is available for all GRACE months.

In Chapter 2, the concept of SH synthesis is developed. Furthermore, NNs are explained in the context of this thesis. Additionally, the concept of errors and uncertainty are defined and distinguished. In Chapter 3, the research in the field of GRACE gap-filling is presented. In Chapter 4, a subsequent research gap is identified resulting in research questions and a research plan is formulated along with research requirements. Chapter 5 discusses how the research questions are answered. An important part in estimating the errors and uncertainty of the GRACE-like data requires that errors of GRACE EWH, Swarm EWH and GLDAS soil moisture data are quantified, as presented in Chapter 6. The NN hyperparameters and decisions are justified in a sensitivity analysis, elaborated upon in Chapter 7. The implementation of SH synthesis and NNs are verified in Chapter 8. Thereafter, the gap-filling results are presented and discussed in Chapter 9. Conclusions are presented in Chapter 10. Finally, Chapter 11 outlines a series of recommendations for future research.

2
Background Information

The synthesis of GRACE and Swarm SH gravity field solutions is discussed in Section 2.1. Section 2.2 describes a NN in the context of GRACE gap-filling with the aforementioned auxiliary datasets. Finally, errors and uncertainty are defined and distinguished in Section 2.3.

2.1. Spherical Harmonic Synthesis

This section describes how to produce gridded equivalent water height (EWH) maps using SH gravity field solutions. This gridded data results from applying SH synthesis to Level-2 data products such as those from GRACE and Swarm. The main steps to produce a gridded map from a SH gravity functional are:

1. To scale the SH coefficients such that the gravity functional represents EWH.
2. (Optional) To apply Gaussian smoothing to gravity functional by scaling the SH coefficients.
3. To perform SH synthesis to the (smoothed) EWH gravity functional which yields a gridded EWH map.

The inversion of the gravity field to a mass distribution is non-unique. This is because the observed gravitational potential represents the integrated distribution of mass below Earth's surface in the radial direction (Wahr et al., 1998). It is assumed that temporal changes in Earth's gravity field on a monthly scale are dominated by hydrological processes on Earth's surface because water is the only element on Earth that moves in large enough quantities (mass) on the monthly time-scale to be observable by GRACE (Wahr et al., 1998). This assumption is valid if the effects of solid body, ocean, and atmospheric tides are ignored. These effects are modelled and removed from the GRACE solutions by the institutes producing them (Dahle et al., 2019; Save, 2019; Yuan, 2019). Effects such as non-tidal atmospheric and ocean changes are more difficult to model and hence may still corrupt the gravity field solutions. It it assumed that these effects are negligible.

Dimensionless Gravity Functional

The Level-2 data products from both GRACE and Swarm are normalized dimensionless SH coefficients: $\overline{C}_{l,m}$ and $\overline{S}_{l,m}$. Each data product has some maximum degree l_{max} and maximum order m_{max} (for GRACE and Swarm $m_{max} = l_{max}$) for which coefficients are estimated. The dimensionless gravity functional $f(\theta, \phi)$ uses these coefficients and is shown in Equation 2.1. The parameters θ and ϕ are the latitude and longitude coordinates of the point to be evaluated, respectively.

$$f(\theta,\phi) = \sum_{l=0}^{l_{max}} \sum_{m=0}^{l} \overline{P}_{l,m}(\cos\theta)[\overline{C}_{l,m}\cos(m\phi) + \overline{S}_{l,m}\sin(m\phi)] \tag{2.1}$$

$C_{2,0}$ Replacement

The $C_{2,0}$ coefficient of Earth's gravity field represents Earth's oblateness. GRACE's estimates for this coefficient are unreliable due to its near-polar orbit which limits its ability to resolve low-degree (large-scale) North to South gravity field variations. It is suggested to replace GRACE's $C_{2,0}$ by more reliable values obtained from satellite laser ranging (SLR) (Cheng & Ries, 2017).

Equivalent Water Height Functional

The hydrological mass redistribution is approximated by the difference in Stokes coefficients between an observed monthly gravity field solution and a static gravity field solution. These residual Stokes coefficients are henceforth denoted as $\Delta\overline{C}_{l,m}$ and $\Delta\overline{S}_{l,m}$. The difference in surface mass per square meter (observed vs. static) is often expressed in terms of EWH. EWH is the height of a vertical block of water with an equivalent mass to the measured one. The scaling factor, $F_l^{\texttt{EWH}}$ for the EWH gravity functional is degree-dependent and is shown in Equation 2.2, where ρ_{e} and $\rho_{\texttt{water}}$ are the average density of the Earth and water respectively.

$$F_l^{\texttt{EWH}} = (2l+1)\frac{R_e\rho_{\text{e}}}{3\rho_{\texttt{water}}} \quad \forall l \in [0, 1, ..., l_{max}] \tag{2.2}$$

Each change in surface mass will result in some load-induced deformation of Earth's crust; even if this deformation is very small, it is not negligible and produces a secondary change in the gravitational field of the Earth. This violates the assumption that the changes in the gravitational field are only being caused by surface mass redistribution, which is what needs to be quantified. Load Love numbers (degree-dependent), k_l, are used to determine how much of $\Delta\overline{C}_{l,m}$ and $\Delta\overline{S}_{l,m}$ is attributed to load-induced deformation and actual differences in surface mass (Wahr et al., 1998). When using these Love numbers, Equation 2.2 gains an additional term as shown in Equation 2.3. Using the updated scaling factor, $F_l^{\texttt{EWH}}$, the non-dimensional gravity functional can be scaled into an EWH gravity functional, $E(\theta,\phi)$ shown in Equation 2.4.

$$F_l^{\texttt{EWH}} = (2l+1)\frac{R_e\rho_{\text{e}}}{3\rho_{\texttt{water}}}\frac{1}{1+k_l} \quad \forall l \in [0, 1, ..., l_{max}] \tag{2.3}$$

$$E(\theta,\phi) = \sum_{l=0}^{l_{max}} F_l^{\texttt{EWH}} \sum_{m=0}^{l} \overline{P}_{l,m}(\cos\theta)[\Delta\overline{C}_{l,m}\cos(m\phi) + \Delta\overline{S}_{l,m}\sin(m\phi)] \tag{2.4}$$

Smoothing

The ability for a satellite to observe a particular part of the gravity field for a specific degree, l scales with $\frac{1}{R_e+h}\left(\frac{R_e}{R_e+h}\right)^l$. Where h is the orbital altitude of the satellite. For increasing degree l, the contribution of high-degree components of the gravity field to the acceleration experienced by the satellites decreases exponentially (Wahr et al., 1998). However, the size of the estimated high-degree coefficients in the SH solutions are often disproportionately high in comparison to low-degree coefficients. This discrepancy arises because the signal strength of higher-degree components is much weaker relative to the noise in the estimated gravity field solutions, resulting in a lower signal-to-noise ratio (SNR). Additionally, as a result of GRACE not sampling the gravity field on Earth everywhere at once, signals with a shorter period than 30 days might remain unresolved during the estimation process of SH coefficients and leak into the high-degree SH coefficients inflating their values and apparent size.

The power at higher SH degrees (representing shorter wavelengths, typically above degree 30-40, depending on the monthly solution) is disproportionately high compared to the lower degrees. For increasing degree l, the contribution of a high-degree component of the gravity field to the acceleration of the satellites becomes exponentially smaller. Therefore, regardless of h the higher degrees are more difficult to determine accurately in comparison to the lower degrees (Wahr et al., 1998). This means that the higher degrees are more likely to have a lower SNR.

To reduce the contribution of less accurate high-degree coefficients and increase the SNR, spatial averaging using Gaussian smoothing is recommended (Wahr et al., 1998). Gaussian smoothing is introduced into the gravity functionals as degree-dependent weights, W_l. Equation 2.5 and Equation 2.6 are required to determine these weights recursively (r is the smoothing radius). The smoothing coefficients, W_l, are introduced into the EWH gravity functional, Equation 2.7, in a similar way as the scaling factors, Equation 2.3.

$$b = \frac{ln(2)}{1 - \cos(r/R_e)} \tag{2.5}$$

$$\begin{aligned} W_0 &= 1 \\ W_1 &= \frac{1 + e^{-2b}}{1 - e^{-2b}} - \frac{1}{b} \\ W_{l+1} &= -\frac{2l+1}{b} W_l + W_{l-1} \end{aligned} \tag{2.6}$$

$$E(\theta, \phi) = \sum_{l=0}^{l_{max}} F_l^{\texttt{EWH}} W_l \sum_{m=0}^{l} \overline{P}_{l,m}(\cos\theta)[\Delta\overline{C}_{l,m}\cos(m\phi) + \Delta\overline{S}_{l,m}\sin(m\phi)] \tag{2.7}$$

Figure 2.1 shows the effect of smoothing on a Swarm gravity field solution scaled to EWH. The middle column of plots corresponds to a GRACE gravity field solution with a smoothing radius of 450 [km] which will be considered as a solution with a high SNR. The left and right columns correspond to Swarm gravity field solutions with smoothing radii: 0 [km] and 750 [km] respectively. In the left-top triangle plot, the SH coefficients' magnitudes are of similar size (mostly orange and yellow) for degrees up to 30 and increase (from orange to dark-red) beyond this degree. This is indicative of a solution with a low SNR because the higher degrees typically contain more noise and are in this case of the largest magnitude. None of the signals over land visible in the lower middle map can be observed in the lower left map. The lower left map is dominated by noise and the EWH alternates between values of above 1 [m] (dark-red) and below -1 [m] (dark-blue). The smoothed Swarm solution (right column), shows more similarities with the high SNR GRACE solution. The top right triangle plot shows that power decreases gradually (from orange to dark-blue) for increasing degree just as the GRACE solution does (top middle triangle plot). The lower right map contains similar signals over land as the GRACE map (lower middle map) does, such as negative EWH (blue) over the Amazon and Greenland and the high EWH (red) over the middle of Africa and the North-East of North-America. The similarity in signals over land show that smoothing gravity field solutions with a low SNR leads to the solutions with a higher SNR.

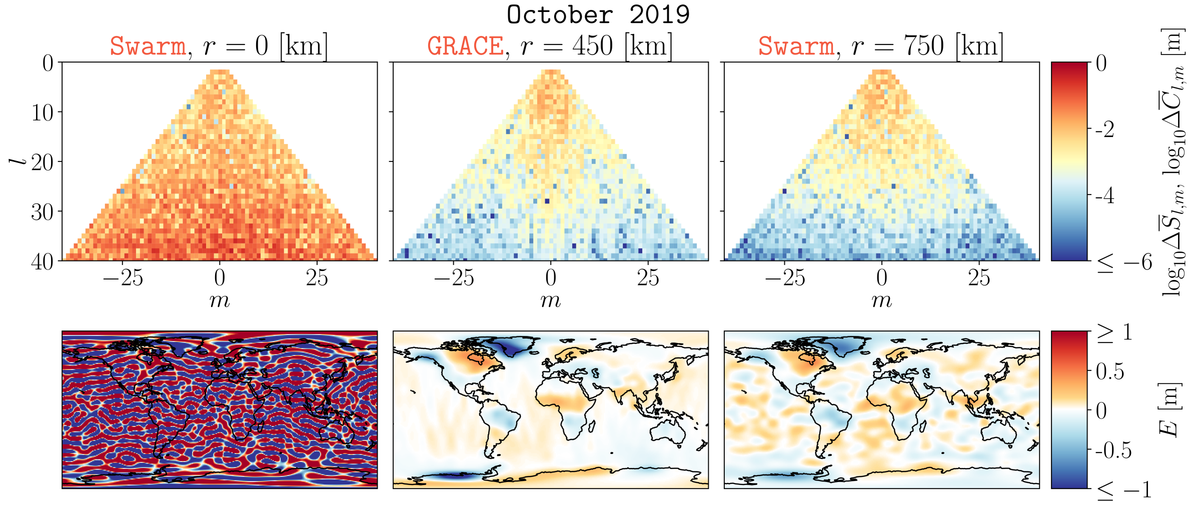


Figure 2.1: Comparison of three gravity field solutions relative to GGM05C for the month October 2019: Swarm without smoothing ($r = 0$ [km], left column), GRACE with smoothing ($r = 450$ [km], middle column), and Swarm with smoothing, $r = 750$ [km] (right column). The top row contains triangle plots of SH coefficient values for each degree and order in EWH. The bottom row contains gridded EWH maps for each gravity field solution.

Synthesis

The smoothed EWH gravity functional, $E(\theta, \phi)$, is converted to a grid composed of cells whose centres are described by latitude-longitude pairs. The highest spatial resolution (East-West and North-South surface distance between neighbouring grid points) is a function of the maximum degree and approximately $\frac{l_{max}}{20 \cdot 10^3}$ [km] (half-wavelength of SH on the surface of the Earth) (Wahr et al., 1998). Irrespective of this, the gravity functional is continuous and can be evaluated at arbitrarily dense grids.

Permanent-Tide System

The tidal potential generated by the Moon, Sun, and other planets on Earth has both a permanent component and a time-dependent (periodic) component on Earth's potential. These tidal potentials also lead to induced deformations. These tidal effects are removed such that the monthly gravity field solutions are unaffected. There are three ways in which the tidal potentials and deformations are dealt with, called permanent-tide systems. In a conventional tide-free system both the permanent and periodic time components are removed (McCarthy & Petit, 2004). This solution considers the Earth as if it were isolated in the universe. The second permanent tide-system is called the mean-tide system as it only removes the periodic component of the tidal potential and deformation. It represents Earth as if it only experiences a permanent tidal potential and deformation. The third permanent tide-system, zero-tide, removes both the periodic and permanent components of the tidal potential. It differs in comparison to the tide-free system through the way it deals with the permanent tidal deformation. In the zero-tide system, the deformation due to the permanent component is still present in contrast to the tide-free system where it is not. The zero-tide system is considered as a more natural representation of Earth (Mäkinen, 2021).

GRACE gravity solutions, the static gravity model GGM05, and the $C_{2,0}$ replacement coefficients are created using a zero-tide system. The Swarm gravity solution is created using a tide-free system. For a fair comparison between gravity field solutions, they should have the same permanent-tide system. Changing between the two systems solely affects the $C_{2,0}$ coefficient. The difference between the $C_{2,0}$ coefficients in both permanent-tide systems is shown in Equation 2.8 (McCarthy & Petit, 2004), where $k_{2,0}$ is Love number for degree 2 and order 0. For an an-elastic Earth, $k_{2,0}$ is 0.30190 [-].

$$\overline{C}_{2,0}^{\textit{zero tide}} - \overline{C}_{2,0}^{\textit{tide free}} = -0.31460 k_{2,0} \frac{1}{R_e \sqrt{4\pi}} \tag{2.8}$$

2.2. Neural Networks

In this thesis, neural network (NN)s are created that gap-fill GRACE using Swarm and soil moisture data. A NN can be seen as a function, $H(\mathbf{X}^{(m)})$, which transforms an auxiliary observation, $\mathbf{X}^{(m)}$, into a model prediction, $\mathbf{E}^{(m,\texttt{model})}$. $\mathbf{X}^{(m)}$ is some input data corresponding to particular month m. In this context, $\mathbf{X}^{(m)}$ is a matrix of size $2n \times o$ which contains two maps: an observation of Swarm EWH ($n \times o$) and an observation of soil moisture ($n \times o$). $\mathbf{E}^{(m,\texttt{GRACE})}$ is a matrix of size $n \times o$ which is estimated GRACE data. $\mathsf{E}^{(\texttt{GRACE})}$ is a tensor that contains the GRACE EWH maps for the set of months, $\mathcal{M}$ in which both GRACE and Swarm are available. X is the tensor containing the observations of Swarm EWH and soil moisture data for the set of months $\mathcal{M}$ and the set of months to be gap-filled, $\mathcal{G}$. The tensor of available auxiliary data X, tensor of available GRACE data, $\mathsf{E}^{(\texttt{GRACE})}$, and tensor of model predictions, $\mathsf{E}^{(\texttt{model})}$, are formally defined in Equation 2.9, Equation 2.10, and Equation 2.11 respectively.

$$\mathsf{X} = \left\{ \mathbf{X}^{(m)} \in \mathbb{R}^{2n \times o} \mid m \in \mathcal{M} \cup \mathcal{G} \right\} \tag{2.9}$$

$$\mathsf{E}^{(\texttt{GRACE})} = \left\{ \mathbf{E}^{(m,\texttt{GRACE})} \in \mathbb{R}^{n \times o} \mid m \in \mathcal{M} \right\} \tag{2.10}$$

$$\mathsf{E}^{(\texttt{model})} = \left\{ H(\mathbf{X}^{(m)}) \in \mathbb{R}^{n \times o} \mid m \in \mathcal{M} \cup \mathcal{G} \right\} \tag{2.11}$$

Training & Testing

The set of parameters required by the function H are known as the learnable parameters, $\mathcal{P}$. In the case of the aforementioned layers, these parameters are neuron biases and weights. And, if applicable, parameters part of the activation functions such as κ in the softplus function. Finding $\mathcal{P}$ such that the difference between the NN prediction, $\mathbf{E}^{(m,\texttt{model})}$ and the target data, $\mathbf{E}^{(m,\texttt{GRACE})}$ approaches zero is called **training** the NN. The set of observations of Swarm EWH and soil moisture data and the set of GRACE EWH data used to train the NN are called the **training data** and are denoted as $\mathsf{X}_{\texttt{train}}$ and $\mathsf{E}_{\texttt{train}}^{(\texttt{GRACE})}$ respectively. By computing a metric such as the root-mean-square (RMS) of $|\mathbf{E}^{(m,\texttt{GRACE})} - \mathbf{E}^{(m,\texttt{model})}|$, a derivative can be computed with respect to each parameter p in $\mathcal{P}$. This derivative, multiplied with a learning rate, l_r, is subtracted from the parameter which results in an updated parameter, $p_{\texttt{new}}$, as shown in Equation 2.12. This process of updating parameters to find an optimal set of parameters is called back-propagation.

$$p_{\texttt{new}} = p - l_r \frac{\partial \sqrt{\text{Mean}((|\mathbf{E}^{(m,\texttt{GRACE})} - \mathbf{E}^{(m,\texttt{model})})^2)}}{\partial p} \forall p \in \mathcal{P} \tag{2.12}$$

This process is repeated for each sample of $\mathbf{X}^{(m)}$ in $\mathsf{X}_{\texttt{train}}$. This cycle is again, repeated for a number of epochs, $n_{\texttt{epochs}}$ resulting in a final set of NN parameters. Depending on the selected NN architecture, there are various training strategies available. These strategies can include: dynamically updating the learning rate as the number of epochs progress or stopping the training process early if a particular performance level is reached. These strategies are implemented for the purposes of preventing over-fitting and ensure no computations are wasted. The parameters that characterize the training strategy are called **hyper-parameters**. These include: l_r and $n_{\texttt{epochs}}$.

If one were to use all observed X and $\mathsf{E}^{(\texttt{GRACE})}$ to train the NN, there is no way to assess how a NN performs when used to produce GRACE-like EWH in a month where there is no reference GRACE EWH available for comparison. To achieve an estimate of the performance, a portion of the data available input and target data are withheld from the training process and not used to optimize $\overrightarrow{P}$ (train the NN). The portions of input and target data that are withheld are called **testing data** and defined as $\mathsf{X}_{\texttt{test}}$ and $\mathsf{E}_{\texttt{test}}^{(\texttt{GRACE})}$ respectively. Estimating the performance of the NN over this testing data is called **testing** the NN.

The set of months used to train the NN, $\mathcal{M}_{\texttt{train}}$, and the set of months used to test the NN, $\mathcal{M}_{\texttt{test}}$, are allocated randomly to avoid the introduction of bias. The amount of data used for training is determined by a hyper-parameter, the training data fraction defined as γ. This is a hyper-parameter of the training process. Equations 2.13 through 2.17 govern the assignment of data to training and testing data. These equations ensure that data only occurs in the training set or in the testing set, not both and that all data is used. Finally, the data sets: $\mathsf{X}_{\texttt{train}}$, $\mathsf{E}_{\texttt{train}}^{(\texttt{GRACE})}$, $\mathsf{X}_{\texttt{test}}$, and $\mathsf{E}_{\texttt{test}}^{(\texttt{GRACE})}$ are formally defined in Equations 2.18 through 2.21.

$$\mathcal{M}_{\texttt{train}} \subset \mathcal{M} \tag{2.13}$$

$$\mathcal{M}_{\texttt{test}} \subset \mathcal{M} \tag{2.14}$$

$$\mathcal{M}_{\texttt{train}} \cap \mathcal{M}_{\texttt{test}} = \emptyset \tag{2.15}$$

$$|\mathcal{M}_{\texttt{train}}| = \left\lfloor \gamma|\mathcal{M}| \right\rfloor \tag{2.16}$$

$$|\mathcal{M}_{\texttt{train}}| + |\mathcal{M}_{\texttt{test}}| = |\mathcal{M}| \tag{2.17}$$

$$\mathsf{X}_{\texttt{train}} = \left\{\mathbf{X}^{(m)} \in \mathbb{R}^{2n \times o} \mid m \in \mathcal{M}_{\texttt{train}}\right\} \tag{2.18}$$

$$\mathsf{X}_{\texttt{test}} = \left\{\mathbf{X}^{(m)} \in \mathbb{R}^{2n \times o} \mid m \in \mathcal{M}_{\texttt{test}}\right\} \tag{2.19}$$

$$\mathsf{E}_{\texttt{train}}^{(\texttt{GRACE})} = \left\{\mathbf{E}^{(m,\texttt{GRACE})} \in \mathbb{R}^{n \times o} \mid m \in \mathcal{M}_{\texttt{train}}\right\} \tag{2.20}$$

$$\mathsf{E}_{\texttt{test}}^{(\texttt{GRACE})} = \left\{\mathbf{E}^{(m,\texttt{GRACE})} \in \mathbb{R}^{n \times o} \mid m \in \mathcal{M}_{\texttt{test}}\right\} \tag{2.21}$$

$\mathcal{P}$ and the hyper-parameters depend on the design of the NN. More details on these parameters are found in Section 3.2 and Chapter 7.

Dense Layer

The function, $H(\mathbf{X}^{(m)})$ performs the following of mathematical operations on the auxiliary observation $\mathbf{X}^{(m)}$. First, the input data is flattened into a vector $\overrightarrow{X}_m$. It is then passed through multiple layers, each of which functions as a transformation, mapping an input vector, $\overrightarrow{X}$, to an output vector, $\overrightarrow{Y}$. Figure 2.2 shows an example of a dense NN layer which accepts an input vector of length, i, and an output vector of length, j. The dense layer is composed of j neurons represented by each row in the figure. Each neuron has a parameter for its bias, b, and has a weight parameter, a, for each element in the input vector. The summation of each input element multiplied by its respective weight in combination with the bias is then passed into an activation function $n(z)$. The output of each activation function makes up the elements in the $\overrightarrow{Y}$. The layer in Figure 2.2 is called a dense layer because all input elements can contribute to each output element. The parameters in the dense layer are known as learnable parameters (defined a few paragraphs later).

By connecting a series of layers in a NN, complex relationships can be established between auxiliary data and target data. A requirement on consecutive layers is that the output size of the previous layer has to match the input size of the next layer.

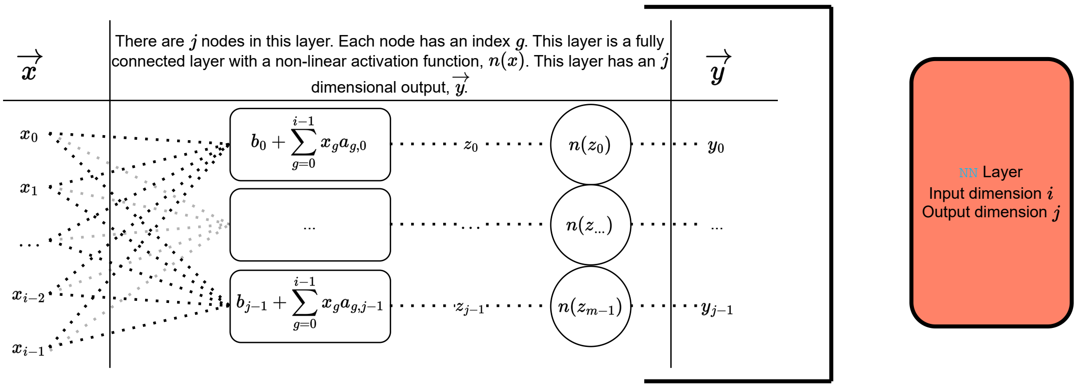

$\overrightarrow{x}$
There are j nodes in this layer. Each node has an index g. This layer is a fully connected layer with a non-linear activation function, $n(x)$. This layer has an j dimensional output, $\overrightarrow{y}$.
$\overrightarrow{y}$
x_0
x_1
...
x_{i-2}
x_{i-1}
$b_0 + \sum_{g=0}^{i-1} x_g a_{g,0}$
...
$b_{j-1} + \sum_{g=0}^{i-1} x_g a_{g,j-1}$
z_0
...
z_{j-1}
$n(z_0)$
$n(z_{...})$
$n(z_{m-1})$
y_0
...
y_{j-1}
NN Layer
Input dimension i
Output dimension j

Figure 2.2: A layer of a fully connected NN.

Activation Functions

Activation functions are introduced into the NN to provide non-linearity in the transformation of auxiliary data to some target prediction. This non-linearity is important because hydrological processes can show non-linear behaviour. Examples of these activation functions are: sigmoid (Equation 2.22), ReLU (Equation 2.23), and softplus (Equation 2.24). These functions are plotted in Figure 2.3. The sigmoid function (light blue line) is popular because it is analogous to a neuron firing. Its range is bound to [0,1]. For very large negative inputs its output approaches 0, and for very large positive numbers its output approaches 1. This function can be useful when trying to compress extreme values influenced by noise or errors. The ReLU function (green line) is a piece-wise function and can be interpreted as a high-pass filter. If z is below 0, the result is 0 and passed through directly to the output otherwise. A combination of high-pass filters may be useful in filtering out noise when creating a relationship between GRACE EWH and Swarm EWH or soil moisture data. The softplus function (orange dashed line) is a smooth approximation of the ReLU function and has a learnable parameter, κ. The softplus function is considered a viable alternative to the ReLU function in cases where the mapping of the input variables to the target variables has to be smooth in contrast to the ReLU function which can cause more abrupt changes in the output data.

$$n(z) = \sigma(z) = \frac{1}{1+e^{-z}} \tag{2.22}$$

$$n(z) = \text{ReLU}(z) = \max(0, z) = \begin{cases} z, & \text{if } z > 0. \\ 0, & \text{otherwise.} \end{cases} \tag{2.23}$$

$$n(z) = \text{softplus}(z) = \frac{1}{\kappa}\log(1 + e^{\kappa z}) \tag{2.24}$$

2.3. Defining Error & Uncertainty

Error and uncertainty are clearly distinguished: error is a measure of the difference between a prediction and some observation. Uncertainty is the degree to which predictions are similar if the same experiment is repeated. A low error means that a prediction is more accurate and if the predictions from independent experiments are close together, then there is a low uncertainty.

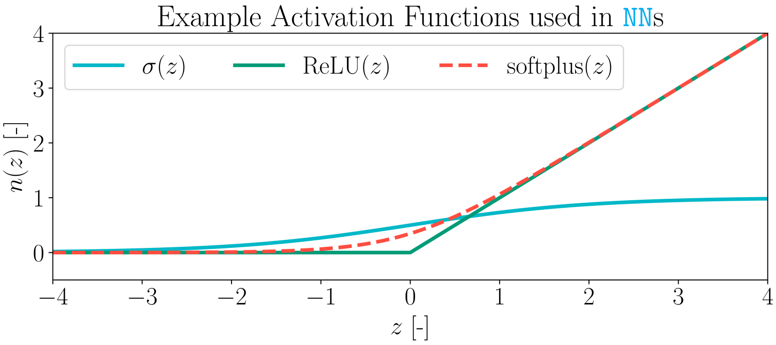


Figure 2.3: A visualization of the range of activation functions: sigmoid (Equation 2.22, light blue line), ReLU (Equation 2.23, dark blue line), and softplus (Equation 2.24, dashed red line, with $\kappa = 2$).

Two important characteristics of both error and uncertainty are accuracy and precision. In Figure 2.4, several shots are fired onto two targets. The red circle is the target and, in this analogy, the observation. The blue dots are the shots and, in this analogy, the predictions. The shots on the left target are an example of an accurate set of shots. On average these shots are close to the target. The shots on the right target are an example of a precise set of shots. These shots on average are further away from the bulls-eye. However, they are closer together, and hence, more precise.

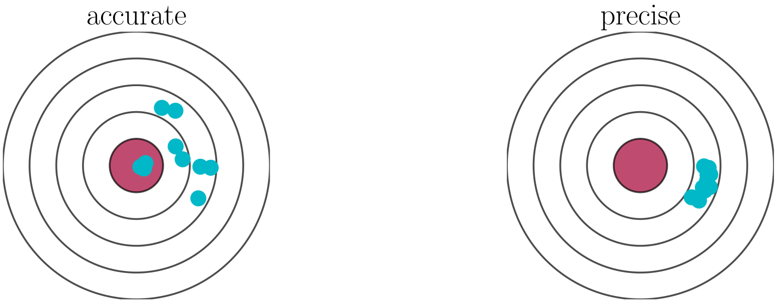


Figure 2.4: An example of an accurate set of shots (left) and an example of a precise set of shots (right).

Error

Error can only be measured when observations are available. This means that errors in GRACE-like data can only be measured for the months in which GRACE data is available, $\mathcal{M}$. For the months in which GRACE is not available, $\mathcal{G}$, the errors have to be estimated. In the context of GRACE gap-filling, the errors are quantified using three metrics: root-mean-square (RMS), Nash-Sutcliffe Efficiency (NSE), and the Pearson Correlation Coefficient (CC).

The first metric is the RMS of the difference between the predictions (GRACE-like data) and the observations (GRACE data). This is also known as the root-mean-square error (RMSE). Equation 2.25 defines the RMS by comparing T observations and predictions of EWH, E, for a single grid cell. Each observation for a time, t, is labelled as $E^{(t,\mathtt{GRACE})}$ and is compared to some model's prediction, $E^{(t,\mathtt{model})}$. The subscript and the superscript on the RMS indicate the observation and the prediction respectively.

$$\mathrm{RMS}_{\mathtt{GRACE}}^{\mathtt{model}} = \sqrt{\frac{\sum_{t=0}^{T-1}\left(E^{(t,\mathtt{model})} - E^{(t,\mathtt{GRACE})}\right)^2}{T}} \tag{2.25}$$

To account for the amplitude and variability of the data being modelled, the NSE metric was introduced in a paper by Nash and Sutcliffe (1970). The NSE value is intended for measuring how well a model predicts run-off as a function of time (Nash & Sutcliffe, 1970). Run-off is the fraction of precipitation that flows over land surface and eventually into streams, rivers, lakes, or oceans. The NSE value is calculated using Equation 2.26, where $\overline{E}^{(\texttt{GRACE})}$ is the mean of the observed EWH over time.

$$\texttt{NSE}^{\texttt{model}}_{\texttt{GRACE}} = 1 - \frac{\sum_{t=0}^{T-1}(E^{(t,\texttt{model})} - E^{(t,\texttt{GRACE})})^2}{\sum_{t=0}^{T-1}(E^{(t,\texttt{GRACE})} - \overline{E}^{(\texttt{GRACE})})^2} \tag{2.26}$$

Essentially, the squared error (model prediction vs observation) is being scaled with the inverse of the variability (an indication of amplitude) of the observed data series. An NSE value of 1 means that the modelled data is the same as the observed data. In the context of GRACE gap-filling, several authors have used this metric (e.g. Agarwal et al., 2023; Ali et al., 2024; Mo et al., 2022). Its advantage is that the NSE values for different river basins with varying seasonal amplitudes can be compared directly.

The third metric is the Pearson CC and is often used to determine the degree to which modelled and observed data match. Equation 2.27 shows how the Pearson CC is computed for a series of observations and predictions, where $\sigma_{(\texttt{GRACE})}$ and $\sigma_{(\texttt{model})}$ are the variances of each series respectively.

$$\texttt{CC}^{\texttt{model}}_{\texttt{GRACE}} = \frac{1}{\sigma_{(\texttt{GRACE})}\sigma_{(\texttt{model})}} \frac{1}{T} \sum_{t=0}^{T-1} \left((E^{(t,\texttt{GRACE})} - \overline{E}^{(\texttt{GRACE})})(E^{(t,\texttt{model})} - \overline{E}^{(\texttt{model})}) \right) \tag{2.27}$$

Two time-series score highly in the Pearson CC if the changes in their values over time are of similar sign and scale. This is in contrast to the NSE, which represents the individual errors per timestamp. This difference is highlighted in Figure 2.5 which shows how three models (orange dots) trying to mimic a particular time-series of observations (green line) score on the RMS (text value in figure), NSE (turquoise bar), and CC (red bar) metrics. The correlation is highest for the third model, 1.000 which has a bias in comparison to the observations (the orange dots are consistently below the green line for all t). The NSE is highest for the first model which has the noisiest results (the orange dots are scattered about the green line for all t) and the lowest RMS. This reinforces the advised usage of the RMS, NSE and CC metrics:

- RMS is a measure of absolute error and is useful when comparing the performance of models attempting to estimate EWH for the **same** location.
- NSE shows how close the modelled time-series is to the observed time-series scaled inversely with its amplitude (variance). This makes it suitable for comparing the performance of models estimating EWH for **different** locations. For example, the comparison of two models that gap-fill different basins is more fair using the NSE.
- CC shows how well the modelled time-series follows the changes in the observed time-series. It is useful for comparing models estimating EWH for the **same and different** locations.

Uncertainty

Uncertainty can only be measured by repeating an experiment multiple times. For this thesis, uncertainty is measured by computing the standard deviation, σ, of a set of predictions. The standard deviation is a measure of how much a set of predictions vary about their mean, μ. A larger standard deviation means a lower precision, and therefore, a higher uncertainty. This approach assumes that the mean, μ, is close to the target value and has no bias. If there is a bias, it is contained within the error criteria, such as the RMS. Therefore, it is important to discuss both error and uncertainty together when assessing a model making predictions of some phenomenon.

Two types of sources of uncertainty are defined for this thesis. The first is epistemic uncertainty which is caused by the method used, in this case a NN used to produce GRACE-like EWH data. The second, is aleatoric uncertainty which is caused by the data used to produce the results. In the case of this thesis, aleatoric uncertainty stems from the uncertainties in Swarm EWH data and GLDAS soil moisture data. The methods for quantifying these two types of uncertainty contributions are introduced in Section 5.3.

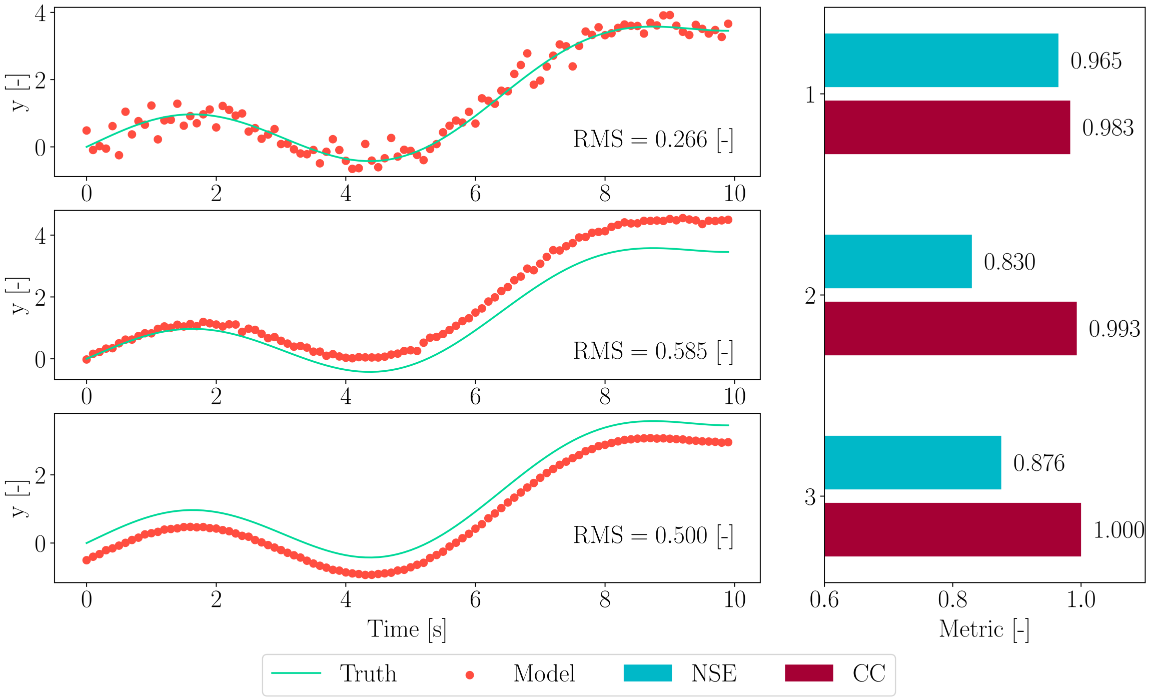


Figure 2.5: RMS, NSE, and CC metrics for three different models that attempt to mimic some observed time-series.

3

Literature Review

In Section 3.1, gap-filling methods are discussed, followed by a more in depth section on gap-filling with neural networks in Section 3.2. Finally, the different global regions covered by various sources of literature are investigated in Section 3.3.

3.1. Gap-filling Methods

Many studies have made attempts to fill the GRACE gaps (Forootan et al., 2020; Gu et al., 2023; Harrison, 2023; Keleş, 2022; Mo et al., 2022; F. Wang et al., 2021; Yi & Sneeuw, 2021). The methods used by these studies can be divided into two classes based on data used to fill the gaps. The first class only uses GRACE data and relies on recognizing trends and cycles. The second class uses GRACE data and one or more auxiliary datasets. In such a case, the auxiliary data sets must be available throughout the entire time series and they must not exhibit gaps. The gap-filling methods range from statistical methods, such as regression, to machine learning, in the form of NNs. Given that gap-filling with NNs is the focus of this thesis, a separate section has been dedicated to gap-filling with NNs in Section 3.2.

Singular Spectrum Analysis

One gap-filling method that does not rely on auxiliary data is called singular spectrum analysis (SSA). SSA uses singular value decomposition (SVD) to decompose a trajectory matrix, Y, into a diagonal, Σ, and orthonormal matrices, U and V, according to Equation 3.1 (Yi & Sneeuw, 2021).

$$Y = U\Sigma V^T \tag{3.1}$$

In the context of applying SSA to a time-series Y with N observations. The trajectory matrix can be formed as shown in Equation 3.2. In this case, the trajectory matrix has one hyper-parameter which is the window length, M. This parameter dictates how many consecutive observations are included in the trajectory matrix. K is the number of columns and is defined according to Equation 3.3.

$$Y = \begin{bmatrix} y_0 & y_1 & y_2 & \cdots & y_{K-1} \\ y_1 & y_2 & y_3 & \cdots & y_K \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_{M-1} & y_M & y_{M+1} & \cdots & y_{N-1} \end{bmatrix} \tag{3.2}$$

$$K = N - M + 1 \tag{3.3}$$

After decomposing Y, one can determine which eigentriples have the strongest signals. This is useful to filter out the weaker signals and focus only on the strongest signals. By keeping only the d strongest signals ($d \leq K$), one can create a reconstructed trajectory matrix (and in turn, a reconstructed time-series).

One requirement of SSA, is that the trajectory matrix (therefore also, the time-series) is complete (i.e. no gaps). To counter this, Yi and Sneeuw (2021) fill the GRACE gap by applying SSA in an iterative manner. They start with an initial guess of what the values of the time-series should be during the gap. Starting with only the strongest signal, they repeatedly reconstruct (update) the time-series until it no longer changes significantly. Then they increase the number of signals (d) they include by one. The aforementioned is repeated until $d = K$.

Figure 3.1 shows a simulated truth (green line) and observations (orange dots) made of this truth over time with an artificial gap (gray area) between 5 and 6.5 [s]. The observations have normal noise added to them. SSA is applied in the same aforementioned iterative manner to create a reconstructed signal (blue dots) that does span the artificial gap. It can be observed that SSA can filter out the observation noise because the reconstruction (blue dots) resembles the truth (green line). However, it is never the case that the truth is known. Only the difference between the observations and reconstruction, the residuals (shaded red area on lower plot), are available. The figure also shows an alternative truth (dashed green line) during the gap. This alternative truth experiences a non-regular event during the artificial gap. The reconstruction in the lower plot does not match this alternative truth demonstrating SSA's inability to capture such a non-regular event.

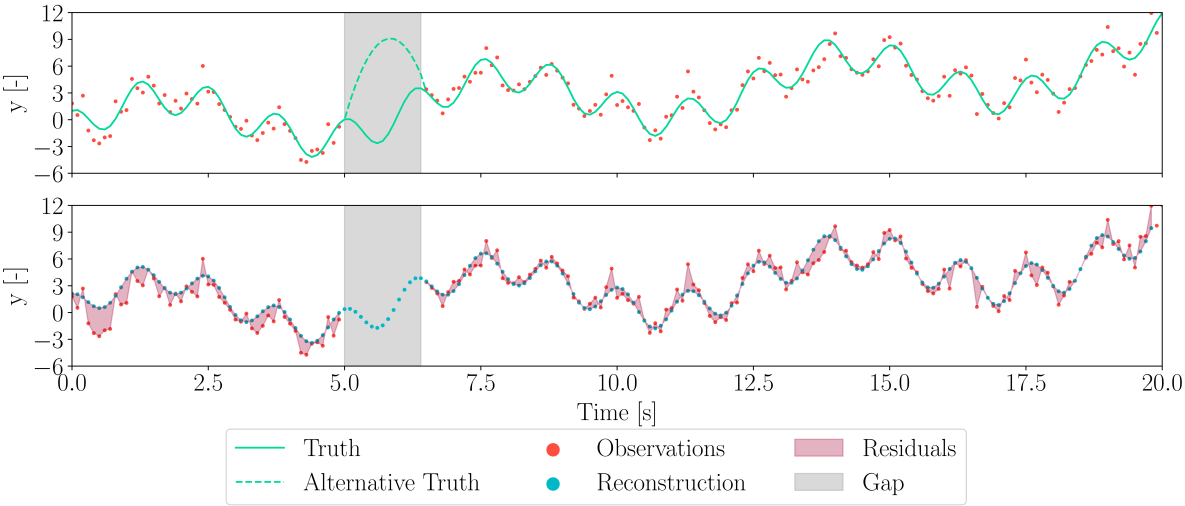


Figure 3.1: Simulated truth, simulated alternative truth and observations (upper plot). Observations, reconstruction using SSA, and residuals (lower plot).

Yi and Sneeuw (2021) apply different versions of an iterative SSA algorithm for smaller and larger gaps of a de-trended time-series of SH coefficients as opposed to a gridded EWH time-series. The window length, M, and minimum number of K that yields the minimum residuals are chosen for each SH coefficient. An analytical relationship between the d/o and K for a given M of 48 months is determined. The results show excellent agreement outside of the gap up to d/o of 30 degrees. The results are validated through a comparison to the Swarm gravity field solutions and find that their results consistently agree with Swarm when Swarm also agrees with GRACE. This is important because Swarm is available during the GRACE gap.

F. Wang et al. (2021) employ multichannel SSA (MSSA) to filter out the noise from GRACE data, indirectly filling the smaller GRACE gaps. In MSSA, each channel considered is a trajectory matrix time-shifted by a different amount. F. Wang et al. (2021) evaluate the filtering performance of three variations of MSSA: filling the gaps using interpolation, filling the gap iteratively with initial guesses of zero, and a third method which does not rely on interpolation nor iteration to fill the smaller gaps. Instead, it modifies the SSA method to directly handle the incomplete trajectory matrix by applying a weighted covariance matrix approach, where weights are assigned to the available data points to compute the singular value decomposition, ensuring that only the observed data contributes to the analysis. They find that the improved SSA method developed by Shen et al. (2015) can be adapted to MSSA and shows improved filtering and gap-filling capabilities (F. Wang et al., 2021).

Advantage of Auxiliary Data

Using auxiliary data offers new opportunities and advantages such as representing non-regular events that occur during the gaps. In the class of methods that do not use auxiliary data, there is no information about their occurrence. In the case of this thesis, auxiliary data sets are considered relevant when they contain information (signals) relevant to hydrological processes which may not be contained in the available GRACE EWH maps. The auxiliary datasets used by (Harrison, 2023) and this thesis are: Swarm data produced by ESA and soil moisture data sets produced by NASA's GLDAS (Harrison, 2023).

Downscaling
The maximum d/o for which Swarm geo-potential fields are accurate is considerably lower than the GRACE fields (Teixeira da Encarnação et al., 2020). Therefore, in studies where Swarm data is used to create GRACE-like data, the Swarm EWH maps are being down-scaled as their resolution is lower than that of GRACE. In the machine learning community this is known as super-resolution (Lepcha et al., 2023). For this thesis, the definition of downscaling is: a procedure in which a spatial input feature with a certain resolution is transformed resulting in a spatial output feature with a higher spatial resolution. An example is shown in Figure 3.2. After downscaling, the gridded map contains more small-scale information. The methods described here, transform and/or downscale the information contained in these auxiliary data sets to produce GRACE-like data.

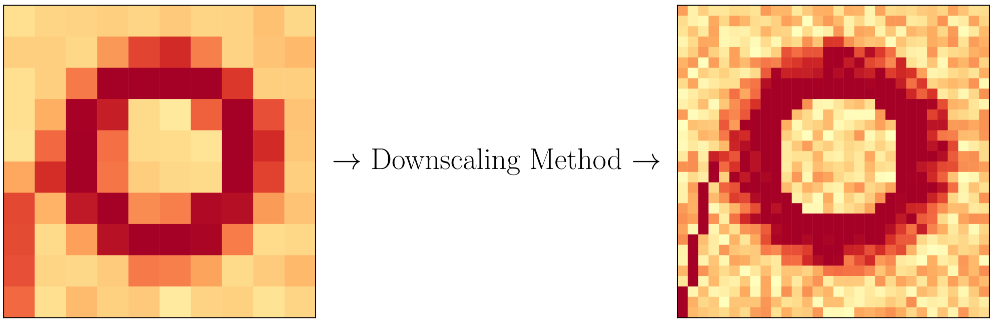


Figure 3.2: An example of a spatial map that is downscaled.

Component Analysis
Several authors employ a form of principal or independent component analysis as a gap-filling method while using Swarm or GLDAS as auxiliary data (Forootan et al., 2020; Gu et al., 2023). Principal component analysis (PCA) identifies the linear combinations of the auxiliary dataset that explain the variance of the data. Independent component analysis (ICA) identifies features of the auxiliary data that are statistically independent. Applied in the context of gap-filling, both methods require some form of iteration, starting with an initial guess of what the gap might be. This is not seen as a limitation because success is demonstrated by using an interpolation of GRACE data over the gaps as an initial guess (Forootan et al., 2020). Using ICA with Swarm data as the auxiliary source, they successfully filled GRACE data gaps and demonstrated ICA's effectiveness in reducing Swarm noise.

3.2. Gap-filling with Neural Networks

When considering gap-filling GRACE using NNs, a whole assortment of NN architectures have been investigated. This assortment includes: fully-connected NNs (Harrison, 2023), convolutional NNs (CNNs) (Harrison, 2023; Keleş, 2022), and Bayesian CNNs (BCNNs) (Keleş, 2022; Mo et al., 2022). The literature surrounding these architectures is covered in Subsection 3.2.1, Subsection 3.2.2, and Subsection 3.2.3, respectively.

3.2.1. Fully-connected Neural Network

A fully-connected NN is made up of multiple dense layers of neurons (introduced in Section 2.2). Harrison (2023) employs such a NN to gap-fill GRACE EWH over the Amazon river basin and finds that a combination of Swarm EWH data and GLDAS soil moisture data is sufficient for the task. The usage of other auxiliary data variables such as precipitation or temperature result in worse performances. The fully-connected NN is shown in Figure 3.3. The green maps represent input and output data. The blue blocks represent the input and output vectors. The orange blocks represent layers of the NN.

In the case of GRACE gap-filling, there are a few modifications that have to be made to the training and testing data when using a fully-connected NN. Both the GRACE data and the auxiliary data sets have a specific spatial representation, in this case, each variable is represented by a two-dimensional (longitude and latitude) map with a dimension of 33 by 49 degrees (a total of 1617 grid cells). The fully-connected NN is unable to directly use this data as input or for training, because each layer only takes a one-dimensional vector as input. Therefore, the GRACE EWH maps, Swarm EWH maps, and GLDAS soil moisture data are flattened into vectors. In the example shown, the input vector has a size of $2 \cdot 1617$. The output of the fully-connected NN is also a vector which means that the last step is to transform the vector back into a gridded map for interpretation.

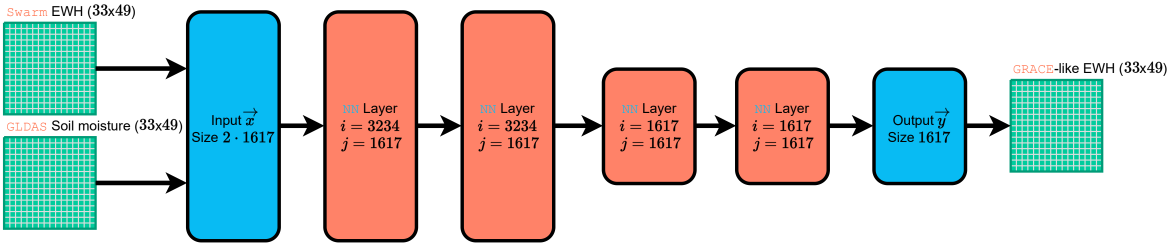


Figure 3.3: A fully-connected NN for gap-filling GRACE (Harrison, 2023).

3.2.2. Convolutional Neural Network

Where a fully-connected NN uses one-dimensional vectors as input, a CNN allows for multi-dimensional vector inputs. The advantage of a CNN over a NN is that a CNN typically requires less learnable parameters when it comes to large inputs (Montesinos-López et al., 2022). This means that the network is less prone to over-fitting as it has to create complex relationships with less parameters. A CNN is often comprised of convolutional layers along with pooling layers. The convolutional layer takes as input a tensor of dimension (C_ixW_ixH_i). C_i is the number of channels, in the case of GRACE gap-filling the number of channels is the number of auxiliary data sets used. The width and height of the data set, W_i and H_i, respectively, can be seen as the spatial dimensions of the input data. The subscript i is for input.

The convolutional layer uses k kernels of size (C_ixWxH). W and H are the width and height of the kernels. The values in the kernel, known as the weights, are the learnable parameters. This kernel is convoluted with each block of input data that is the same size as the kernel. The stride, s, determines whether blocks are skipped or not. One can also apply p_W and p_H padding to the input data so that the spatial dimension of the output tensor does not decrease in comparison to the input tensor. Padding increases the size of the input data by adding rows and columns to the edge of the input matrix. The padding and stride are useful tools in controlling the size of the output of a convolutional layer. On top of this, each convolution can be passed through a non-linear activation function, such as those described in Section 2.2. The size of the output of the convolutional layer is: (kxW_oxH_o). Where W_o and H_o are defined in Equation 3.4 and Equation 3.5 respectively. Each kernel can be seen as a feature detector, the number of output channels is defined as k. A larger value of k results in more trainable parameters and therefore, the training time of a CNN is increased. It also yields a more complex CNN that might be better at capturing relationships between auxiliary and target data. As there is a trade-off between training time and performance, the selection of k is best selected through a multi-objective sensitivity analysis.

$$W_o = (\frac{W_i - W + 2p_W}{s}) + 1 \tag{3.4}$$

$$H_o = (\frac{H_i - H + 2p_H}{s}) + 1 \tag{3.5}$$

An example of a simple convolutional layer is shown in Figure 3.4. A convolutional layer with a single ($k = 1$) kernel of size $1 \times 3 \times 3$ is shown. This kernel is convoluted with all equal-sized blocks in the padded input ($p_W = p_H = 1$) and then passed through a ReLU activation function (Equation 2.23). Through use of the padding, the output dimension of the convolutional layer is the same as the input dimensions. The convolutional layers are particularly important to reduce the amount of learnable parameters in comparison to fully-connected NNs.

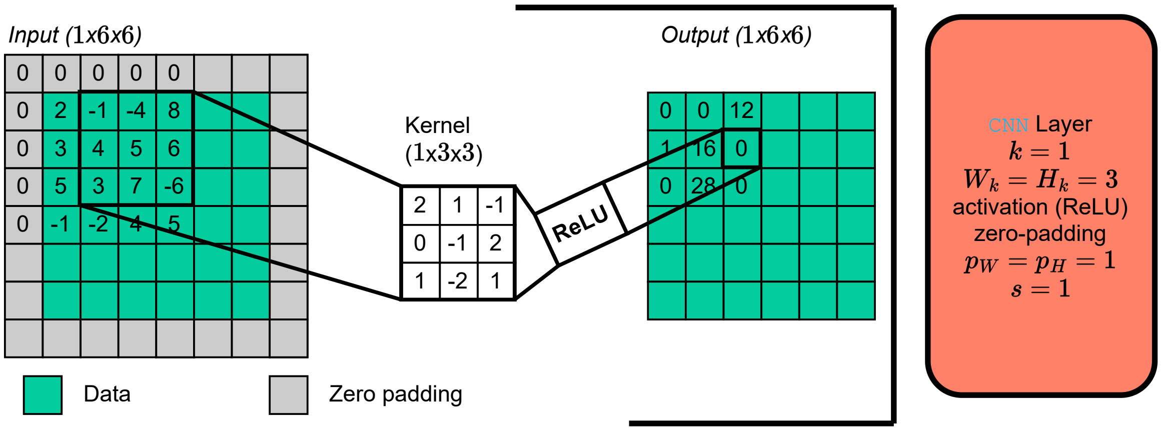


Figure 3.4: An example of convolutional layer in a CNN.

Pooling

It is also common to use pooling layers when there is a desire to reduce the dimension of the tensors throughout the CNN even further (Goodfellow et al., 2016). A pooling layer scans blocks of the input data just like the kernel, but instead of applying convolution to the block of input data it applies a simple operation such as taking the maximum value (max pooling) or averaging the values (average pooling). The pooling layer has no learnable parameters.

An example of a pooling layer that applies max pooling to gridded data is shown in Figure 3.5. In this figure a max pooling layer is applied reducing the dimensions of the input, $1 \times 8 \times 8$, to $1 \times 4 \times 4$. The max pooling layer is interesting as it introduces non-linearity into the CNN. However, the max pooling layer is more sensitive to changes in inputs than the average pooling layer is (Goodfellow et al., 2016).

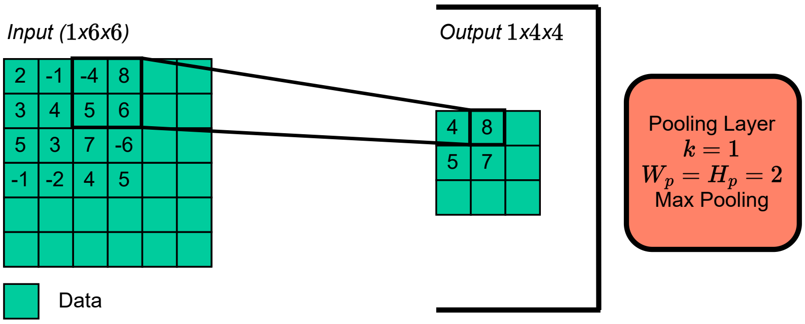


Figure 3.5: An example of pooling layer in a CNN.

Benefits from using pooling layers are the ability to reduce the dimensions of the features and introduce invariance to small translations in the input (Goodfellow et al., 2016). The invariance to small translations in the input leads to a lower sensitivity to noise. The dimension reduction is particularly useful when dealing with input variables which have different sizes. This is interesting when considering training a single CNN on multiple river basins of varying sizes.

CNN vs. NN

The study of Benjamin provides a good opportunity to discuss how using a CNN for gap-filling GRACE compares to fully-connected NNs. In this study, a CNN is designed based on two popular models called AlexNet (Krizhevsky et al., 2012) and VGG-16 (Simonyan & Zisserman, 2015). It is decided to not use pooling layers arguing that their use might destroy important information contained in the relatively small dataset of available input data. This reasoning may be erroneous because the CNN could be given layers with more kernels to account for the loss of information that pooling would introduce. Zero padding is applied to all convolution operations ensuring a constant map dimension throughout the CNN. Figure 3.6 shows the designed CNN which has three convolutional layers (top boxes) followed by two fully-connected dense layers (bottom boxes). Whilst optimizing the hyper-parameters of the CNN, the design-space (i.e. layer setup) is not explored/discussed. CNNs and fully-connected NNs are found to achieve similar accuracy levels. However, on average, the CNN has a training time that is 9 times as long in comparison to the training time of the NN.

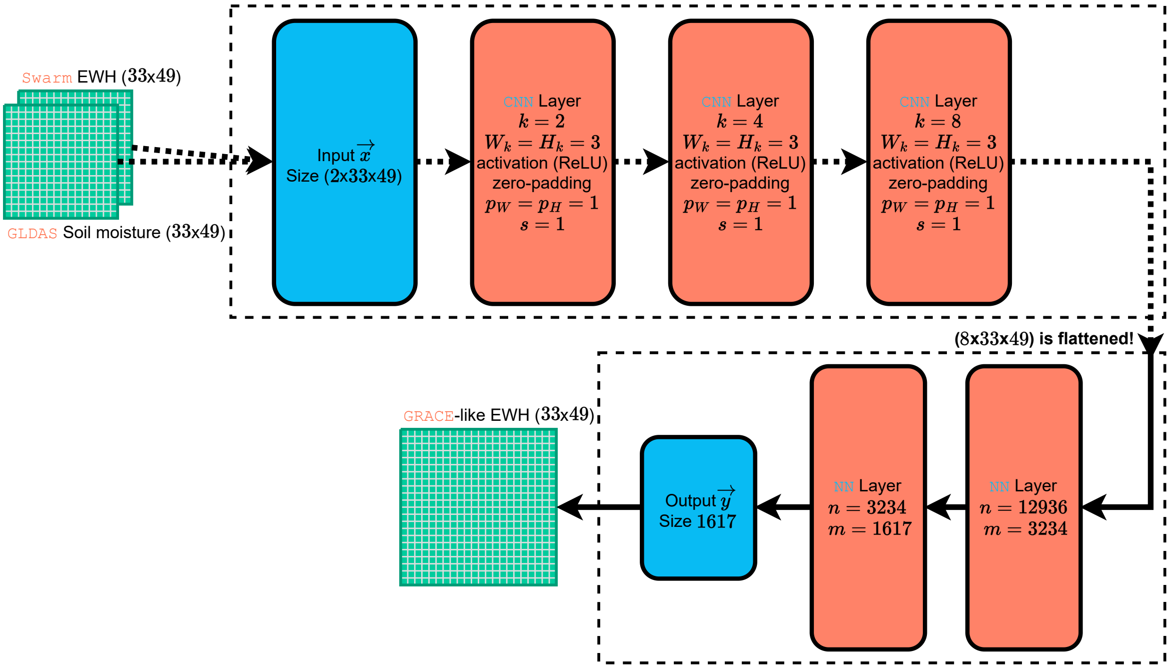


Figure 3.6: CNN inspired by AlexNet and VGG-16 for gap-filling GRACE (Harrison, 2023).

3.2.3. Bayesian Neural Network

BCNNs and Bayesian NNs (BNN) are types of NNs in which the weights and biases are replaced with random variables (Shridhar et al., 2018). This class of NNs is also able to provide some form of uncertainty estimation, namely, the variance associated with the output. The aleatoric and epistemic uncertainty components of this variance can be separated through decomposition (Kwon et al., 2020). Aleatoric uncertainty is the uncertainty in input data and epistemic uncertainty is the uncertainty generated by the use of the method itself (the BCNN).

Mo et al. (2022) seek to fill the GRACE gap in de-trended terrestrial water storage anomalies (TWSA) on a global scale. They find that using a BCNN, hydrological variables can be used as predictors to reproduce de-trended TWSA effectively and accurately. This method performs relatively well in arid regions in comparison to the other studies. They briefly cover the estimates of prediction uncertainty by the BCNN and find that these estimates do correlate with the errors. This is something that is extremely valuable as it means that the BCNN can predict where its errors might be largest. The BCNN outperforms global gap-filling attempts of three other studies (Humphrey & Gudmundsson, 2019; Li et al., 2021; Sun et al., 2020). Keleş (2022) also uses a BCNN to fill the GRACE gap on a global scale. They find that the BCNN outperforms their CNN. They attribute this higher performance to the tendency for BCNNs to prevent over-fitting.

3.3. Regions of Interest

Throughout the available GRACE gap-filling literature, there is a wide variety of areas on Earth that are covered. In some cases, the entire Earth is considered, this is often referred to as gap-filling GRACE on a global scale (e.g. Forootan et al., 2020; Gu et al., 2023; Keleş, 2022; Mo et al., 2022). When gap-filling a subset of Earth's area, the selected region usually does not cover seas or oceans (e.g. Agarwal et al., 2023; Harrison, 2023). This can be attributed to the fact that the changes of mass over land have larger amplitudes and therefore have a higher SNR. Wahr et al. (1998) reason that signals over the ocean have lower amplitudes as added water in the ocean tends to flow away or level out. Most researchers focus on particular regions for three reasons:

- To gain understanding of active river basins.
- To monitor ice mass loss.
- To monitor changes in groundwater storage.

River Basins

A river basin is often also referred to as a drainage basin and is defined as a region where all natural flowing water moves toward a single outlet. There are two groups of river basins:

- Those whose outlet is into the ocean.
- Those whose outlet is into a landlocked body of water.

The four largest river basins in the world are in the first category of river basins. These include, from largest to smallest: Amazon, Congo, Nile, and Mississippi (Lehner & Grill, 2013). Common examples of river basins in the second category are: Lake Chad and the Caspian Sea. Figure 3.7 shows where the mentioned river basins are on Earth.

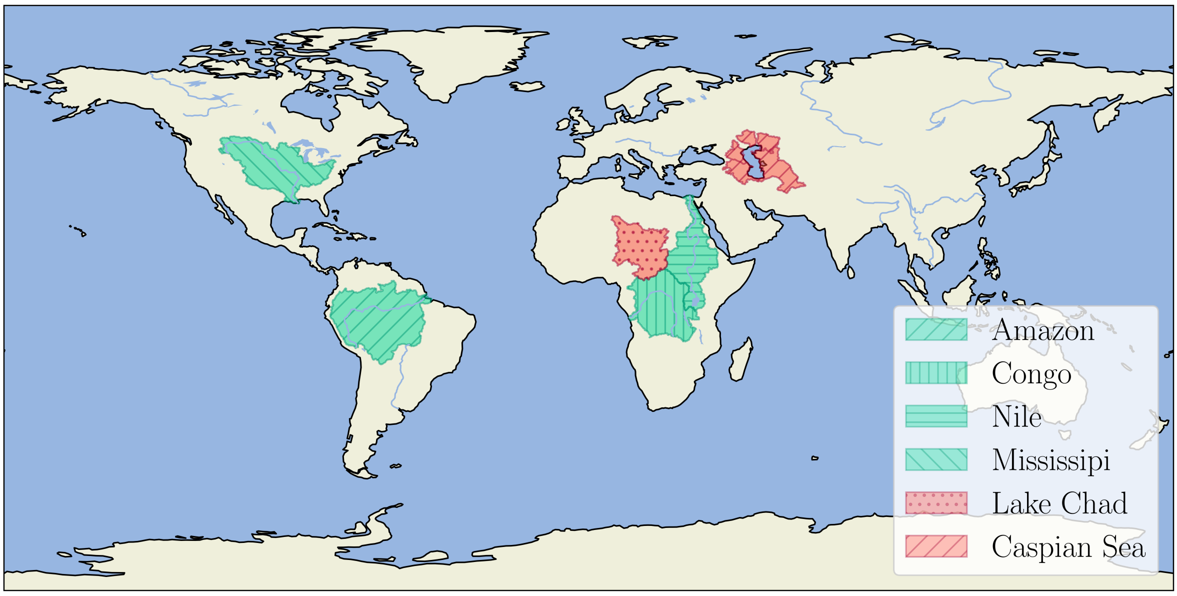


Figure 3.7: Six river basins on Earth: Amazon, Congo, Nile, Mississippi, Lake Chad, and the Caspian Sea. The basin outlines were obtained from Lehner and Grill (2013).

River basins are important to life on planet Earth and important to the transportation of freshwater. Freshwater flowing through river basins often has semi-annual and annual cycles. These cycles are caused by seasonal variations on Earth. Given their importance and activity, some GRACE gap-filling studies focus on gap-filling over these river basins (e.g. Harrison, 2023). Other studies that perform gap-filling on a global scale use river basins as regions over which to evaluate their model's performance (e.g. Forootan et al., 2020; Gu et al., 2023; Mo et al., 2022).

Ice Sheet Loss

The GRACE mission is widely used to measure ice sheet loss. The only two ice sheets in the world where the ice area is larger than $50{,}000$km^2 (located in Antarctica and Greenland) have been monitored using GRACE. From 2015 to 2017, the pace at which ice sheet mass was being lost in West Antarctica drastically decreased. Zhang et al. (2021) show that the Swarm data can be used to prove that the decreased pace did not persist during the GRACE gap. This insight suggests that the unusually strong transition from El Niño to La Niña were the main drivers for this temporary change in pace of ice sheet mass loss. Furthermore, it was noted that this indicates that the change in trends was not due to the change of satellites (GRACE to GRACE-FO) which might have introduced an intermission bias. These insights make gap-filling GRACE an important topic for understanding ice sheet loss.

Groundwater Storage

Several studies have been devoted to monitoring and modeling groundwater storage levels (e.g. Agarwal et al., 2023; Ali et al., 2024; Foroumandi et al., 2023). Collecting in-situ measurements of groundwater storage is often infeasible (Agarwal et al., 2023). In order to overcome the need for in-situ measurements, some studies suggest down-scaling GRACE data to groundwater storage levels or terrestrial water storage content (TWSC) (Agarwal et al., 2023; Y. Wang et al., 2024). Studies focusing on groundwater storage levels do so often for drought monitoring (e.g. Foroumandi et al., 2023). For this reason, studies related to groundwater storage levels frequently focus on regions with at-risk aquifers such as: California (Agarwal et al., 2023), Iran (Foroumandi et al., 2023), and the Indus Basin (Ali et al., 2024). Aquifers are layers under the surface of the Earth that are able to hold water and are often used as a source of water.

4

Research Proposal

Following the literature review (Chapter 3), a research gap is identified in Section 4.1. To fill this gap, research questions are formulated and motivated in Section 4.2. Research requirements are formulated in Section 4.3 to bound the research. Finally, in Appendix A the planning is outlined and reflected on post-completion of the research.

4.1. Research Gap

Whilst most articles do discuss the error associated with their NN-based gap-filling results, there is little to no discussion on the uncertainty (e.g. Harrison, 2023; Keleş, 2022; Mo et al., 2022). Articles employing BCNNs, which inherently provide some form of uncertainty estimation, do not discuss these estimates (e.g. Keleş, 2022; Mo et al., 2022). Furthermore, articles which use gap-filling methods that utilize auxiliary data do not consider errors in the auxiliary datasets used, such as Swarm, GLDAS or ERA5 (e.g. Harrison, 2023; Keleş, 2022; Mo et al., 2022). This lack of uncertainty estimation is a gap in the research field. Error and uncertainty estimates are important to climate models, especially when making projections into the future (Wu et al., 2022).

4.2. Research Questions

The first goal is to quantify the errors and uncertainty of GRACE-like data produced by NNs. The second goal is to reduce the uncertainty of GRACE-like data produced by NNs by generating additional training data using the errors in the auxiliary datasets. Changing the architecture of the NNs is also an option to reduce/affect the uncertainty of the NNs, but is deemed out of scope for this thesis. The two goals translate into the research questions shown below:

- What are the errors and uncertainty of GRACE-like data produced by NNs?
- How does the inclusion of additional training data generated, using errors in the auxiliary datasets, affect the errors and uncertainty of GRACE-like data produced by NNs?

To answer these research questions, the fully-connected NN architecture employed by Harrison (2023) is used because NN architecture design is not in the scope of this thesis. The uncertainty of this architecture's gap-filling capability is investigated for a variety of basins exhibiting different types of behaviour (selected in Section 5.2). This model is verified to ensure there are no technical implementation problems in Chapter 8. A sensitivity analysis is performed to determine an appropriate smoothing radius for both GRACE and Swarm EWH data (see Chapter 6) and to select the hyper-parameters of the fully-connected NN (see Chapter 7).

The latter sensitivity analysis shows that the performance of NNs is very sensitive to seed numbers used for random number generation during their training process. The sources of uncertainty are identified such that seed numbers can be used to quantify how they influence the uncertainty of the GRACE-like data produced by NNs. A total of 20400 models, each distinguished by a unique set of seed numbers, are trained to give insight into the varying errors across trained models.

To answer the first research question, the errors of the input datasets (Swarm EWH and GLDAS soil moisture) are used to sample additional testing data by adding normally distributed noise with the error as the standard deviation of this noise. To answer the second research question, the errors of the input data are used to sample additional training data in the same way additional testing data is sampled. The hypothesis being: if a NN has access to more training data, the uncertainty and errors of its output will be reduced. To clarify, the additional benefit of this method is that a NN now also receives information about the errors in the input data allowing it to put more or less weight on particular data points. By varying the amount of additionally sampled training datasets, the extent to which this hypothesis holds is investigated.

4.3. Research Requirements

Research requirements are formulated to constrain the research effort and output. One category of requirements constrains the method and are labelled with RR-M-x. Another category of requirements constrains the output of the research and are labelled with RR-O-x. The requirements are:

RR-M-1 Each NN gap-filling model is trained to predict GRACE-like data using GRACE EWH as target data.

RR-M-2 Each NN gap-filling model is trained using Swarm EWH and GLDAS soil moisture as input data.

RR-M-3 Each NN gap-filling model is a fully-connected NN.

RR-M-4 All NN gap-filling models have the same architecture and use the same hyper-parameters.

RR-M-5 A single NN gap-filling model is trained to produce GRACE-like data for a single river basin.

RR-O-1 GRACE-like EWH data are produced for all months in which Swarm is available and GRACE is not.

RR-O-2 Estimates of the uncertainty and error in the produced GRACE-like data are produced.

Requirements RR-M-1 and RR-M-2 constrain the data used to train and test the NNs. It is shown that a combination of Swarm and soil moisture data as input to a NN is effective for producing GRACE-like data (Harrison, 2023). Therefore, this research also limits itself to these datasets. RR-M-3 states that each NN is a fully-connected NN. The motivation for this requirement is two-fold: availability of other research for validation of results (Harrison, 2023; Keleş, 2022) and the lower training time (Harrison, 2023). To determine the uncertainty of the results, many models are trained and, for a fair comparison between these models, RR-M-4 is formulated to constrain all models to use the same architecture and hyper-parameters. Requirement RR-O-2 ensures that the errors and uncertainty in the GRACE-like data are produced in line with the research questions of this thesis.

5 Methodology

In order to answer the research questions, particular software is required to obtain the relevant data and to set up the NNs. Section 5.1 is dedicated to describing how the data for this research are acquired and how the NN architecture by Harrison (2023) is implemented.

Section 5.2 then motivates the selection of particular river basins. Section 5.3 describes the sources of uncertainty involved in gap-filling GRACE-like data using NNs and how this uncertainty is ”controlled” using seed numbers and random number generators. Section 5.4 describes how additional training data is sampled to answer the second research question. Finally, Section 5.5 provides an overview of the data collection strategy with some practical notes.

5.1. Data

The NNs rely on three data sets: monthly GRACE EWH maps, monthly Swarm EWH maps, and monthly GLDAS soil moisture maps. These data are discussed in the following subsections. They are defined for the same spatial resolution as all combinations of longitudes (spacing 1°) $\{-180^\circ, -179^\circ, ..., 180^\circ\}$ and latitudes (spacing 1°) $\{-90^\circ, -89^\circ, ..., 90^\circ\}$. To create these grids for all three data sets a code repository is developed. This repository contains Python scripts used to acquire and process the data sets. The remaining text in this section describes the required steps to acquire and process the data sets.

Monthly EWH Maps

The monthly EWH maps are derived from monthly SH gravity field solutions. Those corresponding to GRACE (produced by CSR) and Swarm (produced by International Combination Service for Time-variable Gravity Fields (COST-G) (Jäggi et al., 2022)) are collected from ICGEM[1]. All collected solutions are synthesized to spatial grids and scaled to EWH according to the procedure described in Section 2.1. Within this procedure several choices are made:

- Load Love numbers are obtained from Wahr et al. (1998).
- A sensitivity analysis (Subsection 6.1.1) is performed to select the smoothing radius for the GRACE data (300 [km]) and the Swarm data (750 [km]).
- GGM05C is used as the static gravitational field to which the Swarm and GRACE solutions are compared to obtain EWH and also obtained from ICGEM[2] (Ries et al., 2016).
- For both Swarm and GRACE data the $C_{2,0}$ coefficients are replaced by those from Cheng and Ries (2017). This negates the need to match the permanent tide-system of Swarm SH solutions to those of GRACE SH solutions and GGM05C.
- Unlike all Swarm solutions, the GRACE solutions are not defined from the start of calendar month to the end of a calendar month. It is important that the GRACE and Swarm SH solutions describe Earth's gravity field for the same time domain. To achieve this, the GRACE SH solutions are interpolated

[1] https://icgem.gfz-potsdam.de/sl/temporal (Last accessed: June 2024)

[2] https://icgem.gfz-potsdam.de/tom_longtime (Last accessed: June 2024)

linearly to the Swarm solutions, with the interpolation epochs defined as the middle point in the respective solutions' time domains.

- To scale the SH solution from non-dimensional to EWH, several constants are required (see Equation 2.3). Table 5.1 shows what values are utilized for each constant.

Table 5.1: The assumed values of the constants required for scaling non-dimensional functional to EWH functional.

Constant	Symbol	Value	Units
Earth's equatorial radius	R_e	6378.137	[km]
Earth's average density	ρ_e	5513	[kgm^{-3}]
Average density of water on Earth's surface	ρ_water	1000	[kgm^{-3}]

Figure 5.1 shows the EWH maps of GRACE and Swarm for the month of April 2022 as an example. This figure verifies that Swarm and GRACE exhibit similar signals. Both GRACE and Swarm show a decrease in EWH in the West of Antarctica and in Greenland. Each map also show an increase in EWH over the Amazon, central Africa, and the North-East of North America and these variations are of similar amplitude and generally located in the same regions. Furthermore, the Swarm EWH map contains more variations over the ocean (see the strong red and blue bands from West to East over the ocean). The GRACE solution exhibits weaker variations over the ocean. This is an indication of GRACE being more accurate than Swarm, as further elaborated on in Subsection 6.1.1. The direction of striping for the GRACE solutions aligned North to South is explained by the configuration in which the multi-satellite mission orbits the Earth. The GRACE satellites are a in near polar orbit following each other, this leads to the GRACE mission not being able to distinguish acceleration changes in their cross-track direction (East-West).

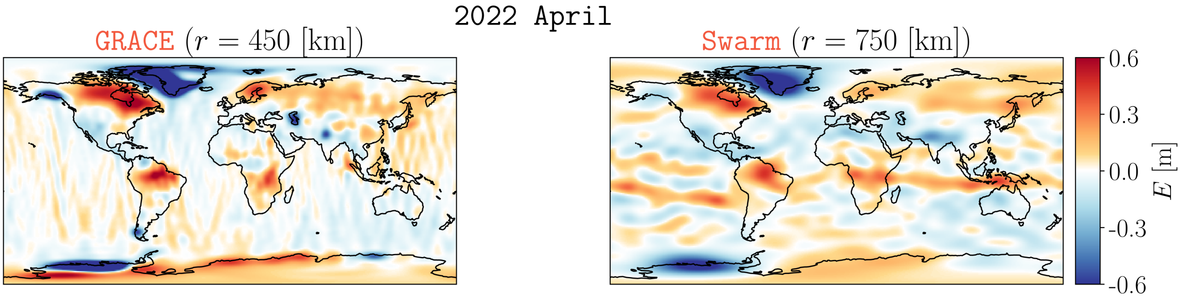

2022 April
GRACE $(r = 450\ [\mathrm{km}])$
Swarm $(r = 750\ [\mathrm{km}])$
0.6
0.3
0.0
-0.3
-0.6
E [m]

Figure 5.1: EWH maps for month April 2022: GRACE EWH (left) and Swarm EWH (right).

Monthly Soil Moisture Maps

Soil moisture, S, is a measure of the amount of water contained per square meter of soil in units [kg m^{-2}]. The required monthly soil moisture maps are acquired from NASA's EarthData data access portal. For the LSM, NOAH is chosen as it is the LSM for which the NN architecture used for this thesis is designed (Harrison, 2023). The soil moisture of the first 2 [m] of Earth's surface is reported in four separate sub-layers (0-0.1 [m], 0.1-0.4 [m], 0.4-1.0 [m], 1.0-2.0 [m]). Given that GRACE can not distinguish mass changes along the radial direction, the soil moisture content reported for the four separate sub-layers are summed.

Given that the GRACE EWH and Swarm EWH are represented by changes from month to month, changing the soil moisture maps to reflect monthly changes may prove easier for the NN to form relationships between the soil moisture and GRACE EWH data. As an alternative to the soil moisture maps, a second set of maps is created by subtracting the mean of all soil moisture maps from each individual map. This results in monthly maps of the change in soil moisture, ΔS. In Chapter 7, a sensitivity analysis is performed which leads to the decision to use the ΔS maps as opposed to the S maps.

Figure 5.2 shows the S and ΔS maps for April 2022. Where S only contains positive values, ΔS contains both positive and negative values. Comparing Figure 5.1 and Figure 5.2 leads to the observation that the ΔS maps contain some of the same signals as the EWH solutions. For instance, there is a decrease in both soil moisture and in EWH over Mexico and India. Additionally, there is increase in soil moisture and in EWH over the Amazon. There are also differences between the two data types. Over the North-East of North America there is no rise in soil moisture despite a significant increase in EWH. Where the soil moisture increases near the Caspian Sea, the EWH solutions show a negative change in that area. These differences do not have to mean that either of types of datasets are wrong because the soil moisture does not account for all moving water in a particular region. Amongst others, lakes, rivers, and aquifers can also lose and gain water leading to satellite EWH observations showing a different change in comparison to the ΔS maps.

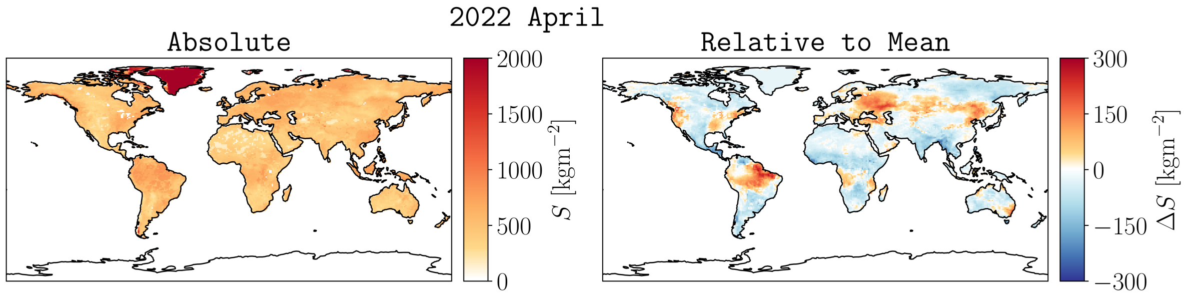


Figure 5.2: Soil maps for month April 2022: absolute soil moisture, S (left) and ΔS (right).

Neural Network Implementation

For this thesis, the NN architecture (see Figure 3.3) used by Harrison (2023) is implemented as a baseline NN. For the implementation of this NN, the Python library called TensorFlow[1] is used. For managing and implementing the NNs, a separate repository is created: `GapFillingNeuralNetworks`. The procedures for setting up the data for NNs, the custom NSE metrics for the NNs, and training the NNs are implemented in: `data.py`, `metrics.py`, and `neural_network.py`. The specific hyper-parameters used and other technical details regarding the NN are discussed in Chapter 7.

5.2. Regions of Study

Gap-filling on the global scale is out of the scope of this thesis. Instead, the focus is on smaller regions. River basins are an interesting subject for this study due to their variety in EWH signals. This variety is interesting as NNs are known for flexible pattern recognition. Basins with weaker EWH signals are expected to be challenging as their signal-to-noise ratio may be low. The four largest river basins are selected for this thesis: Amazon, Congo, Nile, and Mississippi. The motivation for selecting larger rather than small basins is that their are more grid points available for spatial analysis (see Chapter 9).

Figure 5.3 shows the mean EWH, $\mu(\mathbf{E}^{(m,\texttt{GRACE})}_{\texttt{basin}})$, as observed by GRACE over these basins for each month m. The Amazon river basin (turquoise line) has the largest mean EWH amplitude. There are different phases in the four basin mean EWH time series. For instance, the peak of EWH for the Amazon aligns with the trough of the Nile's mean EWH (red line). Some basins also show signs of secondary signals. For example, the Congo river basin (dark blue line) shows a second oscillation in its peaks.

Figure 5.4 shows the standard deviation of de-trended EWH spatially, $\sigma(\mathrm{E}^{(\texttt{GRACE},\texttt{detrend})})$, for each river basin, as measured by GRACE. The standard deviation of de-trended EWH is an indication of the amplitude of the measured EWH. The bulk of the Amazon's EWH amplitude with values up to 40 [cm] occurs in the North-East quadrant of the region (highlighted by the dashed green rectangle). In this region most sub-basins converge into the Amazon river which leads to a higher annual inflow and outflow with respect to upstream basins (Tourian et al., 2018). In the Congo river basin, the Southern (highlighted by solid green rectangle) and Northern part (highlighted by dashed green rectangle) experience the largest annual precipitation amplitudes driving the larger EWH amplitudes (Ndehedehe & Agutu, 2022). In the case of the Mississippi river basin, the largest standard deviation of EWH occurs where the Arkansas and Missouri river join the Mississippi river (in the dashed green rectangle). Another pattern of slightly higher standard deviation occurs along the Missouri river (highlighted by the solid green rectangle).

[1] https://www.tensorflow.org/

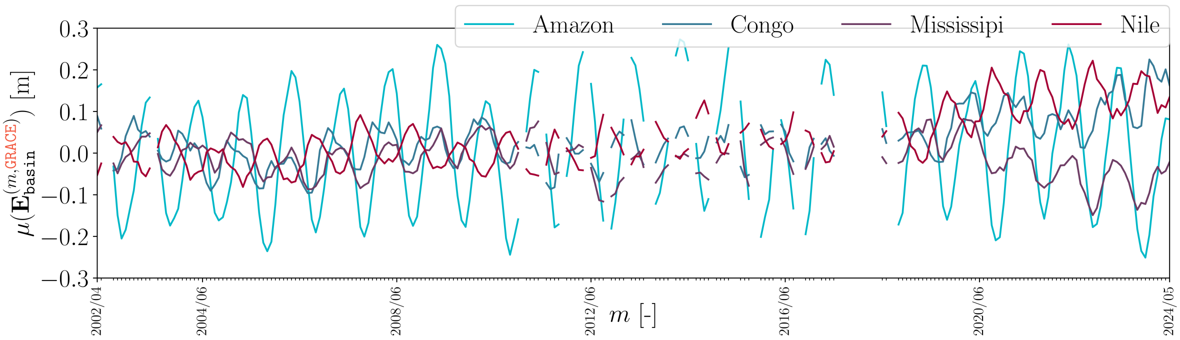


Figure 5.3: The mean EWH per basin for all available GRACE months.

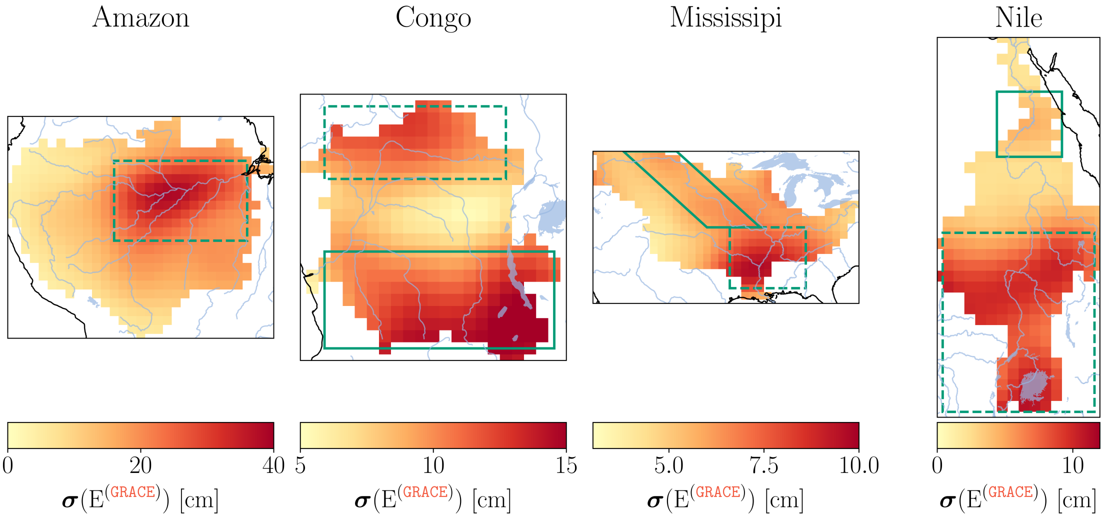


Figure 5.4: The standard deviation of de-trended EWH spatially per basin for all available GRACE months.

The Southern half of the Nile river basin has higher EWH amplitudes than the Northern half. The basin's main source is Lake Victoria (highlighted by the dashed green rectangle) which experiences consistent rainfall due to it being in the inter-tropical convergence zone. On the other hand, the Northern half of the Nile is relatively arid, experiencing little rainfall leading to reduced EWH amplitudes. These difference in rainfall and climate explain the contrast in EWH amplitude between the Nile's Southern and Northern half (Abd-Elbaky & Jin, 2019). There is an exception highlighted by a solid green rectangle. In this region, there are several dams amongst which the Aswan High Dam, it may be that the controlled release of water leads to larger variabilities in EWH in comparison to parts of the river further downstream and upstream.

Given the variety in amplitude (on average and spatially), phase shift, and frequency of the measured EWH, these four basins are considered diverse enough for this study. The basin outlines are obtained from Lehner and Grill (2013).

5.3. Quantifying Effects of Uncertainty

In Section 2.3 two types of uncertainty sources are defined: aleatoric and epistemic uncertainty. In the context of gap-filling GRACE EWH using NNs it is the architecture, parameters, and hyper-parameters of a NN that influence the extent of the epistemic uncertainty. The aleatoric uncertainty is driven by the uncertainties in the Swarm EWH and GLDAS soil moisture data used to generate the GRACE-like data. Using random number generators with recorded seed numbers, different instances of randomness can be generated and used to systematically investigate how specific sources of uncertainty contribute to the overall uncertainty of using NN's to gap-fill GRACE. Subsection 5.3.1 and Subsection 5.3.2 discuss the epistemic and aleatoric uncertainty sources, respectively.

5.3.1. Epistemic Sources

Within several parts of the training process of NNs, randomness is introduced. The parameters of a NN are initialized using a random number generator (parameter initialization). Furthermore, the training and testing data months are selected and ordered randomly prior to training each NN (data selection). The randomness involved in these two steps leads to different optimal parameters being found during the training process. This in turn can lead to different trained NN models producing different outputs. To quantify the epistemic uncertainty, two seed numbers will be used:

- $\alpha_{\texttt{weights}}$: seed number used to initialize parameters of NN.
- $\alpha_{\texttt{select}}$: seed number used to split and order the training and testing data.

5.3.2. Aleatoric Sources

Aleatoric uncertainty is the uncertainty related to the data used to train the NN. In this case, the uncertainty of auxiliary data: Swarm EWH and GLDAS soil moisture. When testing the NN's performance, it is important to recognize that the auxiliary data contain noise, $\boldsymbol{\xi}_x^{(m)}$ (with shape $2n \times o$) which is defined as the concatenation of the errors in Swarm EWH, $\boldsymbol{\xi}_E^{(m,\texttt{Swarm})}$ (with shape $n \times o$), and the errors in GLDAS soil moisture, $\boldsymbol{\xi}_{\Delta S}^{(m)}$ (with shape $n \times o$), as shown in Equation 5.1. $\boldsymbol{\xi}_E^{(m,\texttt{Swarm})}$ and $\boldsymbol{\xi}_{\Delta S}^{(m)}$ are quantified in Section 6.1 and Section 6.2 respectively.

$$\boldsymbol{\xi}_x^{(m)} = \begin{bmatrix} \boldsymbol{\xi}_E^{(m,\texttt{Swarm})} \\ \boldsymbol{\xi}_{\Delta S}^{(m)} \end{bmatrix} \tag{5.1}$$

It is expected that a NN with a low uncertainty will produce consistent results in spite of being tested with noisy data. As there is no possible way to gauge the true observational noise, it is decided to artificially generate noisy testing data to investigate the uncertainty in the process of using NNs to produce GRACE-like data. This results in a new definition of testing data sets, $\mathsf{X}_{\texttt{test}}$ and $\mathsf{E}_{\texttt{test}}^{(\texttt{GRACE})}$ as shown in Equation 5.2 and Equation 5.3 respectively. The added noise for each month, m, of the testing data is generated using a normal distribution with mean 0 and standard deviation $\boldsymbol{\xi}_x^{(m)}$. The noise is formally defined as $\mathcal{N}(0, \boldsymbol{\xi}_x^{(m)}, \lambda_x)$ and generated using a seed number, λ_x.

$$\mathsf{X}_{\texttt{test}} = \left\{ \mathbf{X}^{(m)} + \mathcal{N}(0, \boldsymbol{\xi}_x^{(m)}, \lambda_x) \in \mathbb{R}^{2n \times o} \mid m \in \mathcal{M}_{\texttt{test}} \right\} \tag{5.2}$$

$$\mathsf{E}_{\texttt{test}}^{(\texttt{GRACE})} = \left\{ \mathbf{E}^{(\texttt{GRACE},m)} \in \mathbb{R}^{n \times o} \mid m \in \mathcal{M}_{\texttt{test}} \right\} \tag{5.3}$$

Furthermore, the errors of GRACE EWH data are quantified in Subsection 6.1.2. These are used in Chapter 9 to determine if a model is performing as well as they should. If one were to fit exactly to the GRACE EWH data, then they are also fitting to the noise in GRACE which is not desirable. Therefore, if a model's errors are lower than GRACE EWH errors this means that the model has achieved the maximum agreement with GRACE.

5.4. Sampling Additional Training Data

The second research question aims to determine if the uncertainty of NN produced GRACE-like data can be reduced. One way to reach that goal is to test whether sampling additional training data for a NN leads to a reduced uncertainty. The observation of the input data can be defined as some combination of the truth and some observational errors or noise. Using estimates of the observational errors, normal noise is generated and used to created additionally sampled training data. In this additional data, the noise is only added to the input component, $\mathsf{X}_{\texttt{train}}$ whilst leaving the corresponding target data, $\mathsf{E}_{\texttt{train}}^{(\texttt{GRACE})}$, unchanged. The hypothesis is: if a NN is trained with access to more training data, then the NN will have a reduced uncertainty in its output GRACE-like EWH data in comparison to NNs trained only on the original data. To test the extent to which using additionally sampled training data supports this hypothesis, a new variable is introduced, η, the number of additionally sampled training sets.

The training data sets for auxiliary and target data are now redefined according to Equation 5.4 and Equation 5.5, respectively. These equations show that if $\eta = 0$, the training data is as originally defined in Equation 2.18 and Equation 2.20. If $\eta > 0$, the size of the training increases by a factor, $(1+\eta)$. The noise, $\mathcal{N}(0, \boldsymbol{\xi}_x^{(m)}, \alpha_x)$, will be generated using seed number, α_x.

$$\mathsf{X}_{\mathtt{train}} = \left\{\mathbf{X}^{(m)} \in \mathbb{R}^{2n\times o} \mid m \in \mathcal{M}_{\mathtt{train}}\right\} \cup \bigcup_{i=0}^{\eta-1} \left\{\mathbf{X}^{(m)} + \mathcal{N}(0, \boldsymbol{\xi}_x^{(m)}, \alpha_x) \in \mathbb{R}^{2n\times o} \mid m \in \mathcal{M}_{\mathtt{train}}\right\} \tag{5.4}$$

$$\mathsf{E}_{\mathtt{train}}^{(\mathtt{GRACE})} = \bigcup_{i=0}^{\eta} \left\{\mathbf{E}^{(\mathtt{GRACE},m)} \in \mathbb{R}^{n\times o} \mid m \in \mathcal{M}_{\mathtt{train}}\right\} \tag{5.5}$$

5.5. Data Collection Strategy

The uncertainty of the selected NN architecture is quantified by training $N_{\mathtt{models}}$ models. A model is characterized by the following seed numbers: $\alpha_{\mathtt{weights}}$, $\alpha_{\mathtt{select}}$, $\alpha_{\mathtt{x}}$. It is also defined by the basin, β, for which it is trained, and the number of additionally sampled training datasets, η. Table 5.2 provides an overview of the variables that define a model and their set of values. The sets for the variables β and η are pre-defined and their sizes, N_β and N_η are listed. Therefore, the number of models to train, $N_{\mathtt{models}}$ only has one remaining free variable, the number of training seeds, N_α. Equation 5.6 shows how $N_{\mathtt{models}}$ is defined and that it scales in a cubic manner with N_α for $N_\eta > 1$. This section covers how the number of training seed numbers, N_α, is selected accounting for a limited availability of computing power. It also describes the manner in which the models are evaluated. The number of testing seed numbers, N_λ, is not included in Equation 5.6 as the associated seed number is only used for testing, and not training the models.

Table 5.2: Each variable, the symbol it is represented by, the values the variable can be (set), and the size of each set each variable can take.

Variable	Symbol	Set	Set size, N
Basin	β	$\beta \in \left\{\text{Amazon}, \text{Congo}, \text{Nile}, \text{Mississippi}\right\}$	$N_\beta = 4$
Number of additionally sampled training data sets	η	$\eta \in \left\{0, 1, 2, 3, 4, 5\right\}$	$N_\eta = 6$
Training seed numbers (`weights`, `select`, `x`)	α	$\alpha \in \left\{0, 1, 2, ..., N_\alpha - 2, N_\alpha - 1\right\}$	N_α
Testing seed numbers (`x`)	λ	$\lambda \in \left\{0, 1, 2, ..., N_\lambda - 2, N_\lambda - 1\right\}$	N_λ

$$N_{\mathtt{models}} = N_\beta \cdot N_\alpha^2 (1 + (N_\eta - 1) \cdot N_\alpha) \tag{5.6}$$

Computational Effort

The entire experiment is performed on a computational server owned by the Technical University of Delft. In order to gauge the amount of time required to train all models on this server, a baseline experiment is set up to determine the training time for a model depending on the number of additionally sampled training datasets generated, η. The results of the baseline experiment performed on the server are shown in Figure 5.5. For each basin and $1+\eta$ (x-axis) a model is trained and its training time, t (y-axis), is plotted. The range of η values is from 0 to 5. The motivation for the selection of this value is that it allows for the determination of a relationship between η and NN error and uncertainty. A linear line is fitted through the baseline experiment results and its function, $t(\eta)$ in minutes, is shown in Figure 5.5. On average, for every additional training set that has to be processed, the training time of the model increases by approximately 9 minutes. The bias of the fitted line is approximately 1.5 minutes, which means that there is 1.5 minutes of overhead in each training run.

The total training time, $T_{\mathtt{training}}$, for all models is given by Equation 5.7. For $\eta = 0$, no additional training data is sampled and this value is not included in the summation term. The total training time is plotted for varying values of N_α in Figure 5.6. As $T_{\mathtt{training}}$ scales cubicly with N_α, the total training time quickly exceeds a 100 days when only using one CPU and $N_\alpha = 10$. To reduce the total training time,

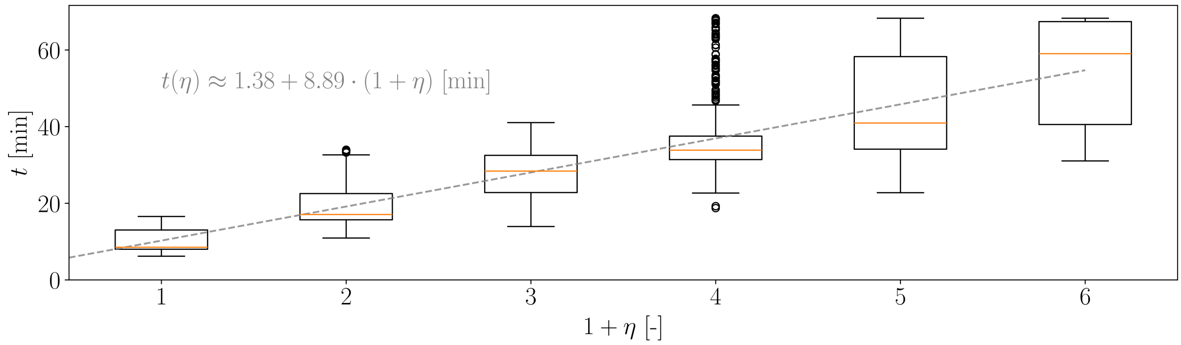


Figure 5.5: Training times per basin for each trained model for varying η. See also a linearly fitted trend line.

multiple CPU's are used to train the NN models. The server on which the experiment is run has multiple CPUs available. The amount of CPUs used is defined as $N_{\texttt{CPU}}$. The total training time is predicted for $N_{\texttt{CPU}}$ values of 1, 10, 25, and 50. Using only one CPU and $N_\alpha = 30$ results in a total training time of over 10^4 days (see red "X" in top right corner of Figure 5.6). Since this is impractical, it is decided to set $N_{\texttt{CPU}} = 50$ and $N_\alpha = 10$ for an approximate training time of 8 [days] (marked by green "X" in Figure 5.6) for a total of $20.4 \cdot 10^3$ trained models.

$$T_{\texttt{training}} = N_\beta \cdot N_\alpha^2 \left(t(\eta) + \sum_{\eta=1}^{5} N_\alpha t(\eta) \right) \tag{5.7}$$

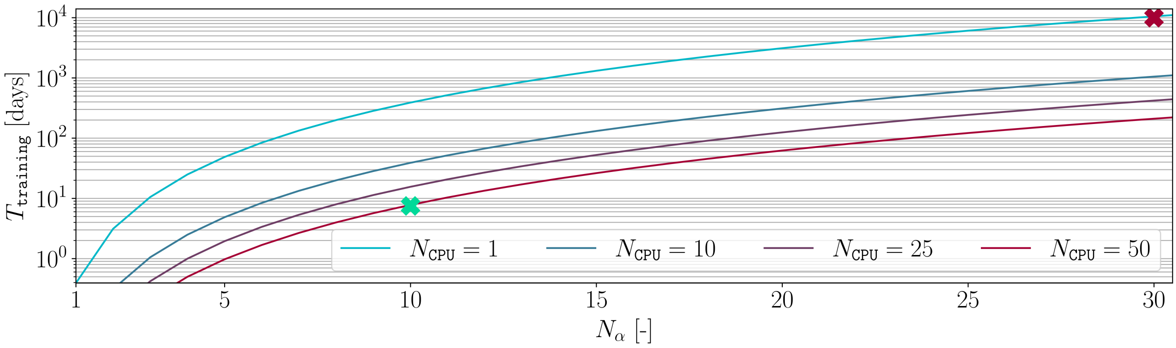


Figure 5.6: Total training time, $T_{\texttt{training}}$, for varying N_α and $N_{\texttt{CPU}}$.

Calculations for Error & Uncertainty Quantification

After training all models, they are all tested for N_λ different values of λ_x to provide insight into the aleatoric uncertainty (see Subsection 5.3.2). For this experiment, N_λ is set to 30. The number 30 is selected as it is the minimum recommended population size whose mean will approximate a normal distribution (Mascha & Vetter, 2018). This is important for the statistical analyses to hold. This means that the total number of test results, $N_{\texttt{results}}$ is $6.12 \cdot 10^5$. A test result is the tensor output, $\mathsf{E}_{\texttt{test}}^{(\texttt{model})}$, a model generates using tensor input data, $\mathsf{X}_{\texttt{test}}$.

For error calculation, $\mathsf{E}_{\texttt{test}}^{(\texttt{model})}$ is compared to $\mathsf{E}_{\texttt{test}}^{(\texttt{GRACE})}$ per model. There are several ways to represent these errors:

- NSE, RMS, and CC of the monthly mean of $\mathsf{E}_{\texttt{test}}^{(\texttt{model})}$ w.r.t. to monthly mean of $\mathsf{E}_{\texttt{test}}^{(\texttt{GRACE})}$. These are two numbers called, $\texttt{NSE}_{\texttt{test}}^{(\mu)}$, $\texttt{RMS}_{\texttt{test}}^{(\mu)}$, and $\texttt{CC}_{\texttt{test}}^{(\mu)}$ respectively.
- Spatial NSE, RMS, and CC of $\mathsf{E}_{\texttt{test}}^{(\texttt{model})}$ w.r.t. $\mathsf{E}_{\texttt{test}}^{(\texttt{GRACE})}$. These are three gridded maps called, $\mathbf{NSE}_{\texttt{test}}$, $\mathbf{RMS}_{\texttt{test}}$, and $\mathbf{CC}_{\texttt{test}}$, respectively. The bold notation indicates a spatial dimension.
- Temporal RMS of each month in $\mathsf{E}_{\texttt{test}}^{(\texttt{model})}$ w.r.t. each month in $\mathsf{E}_{\texttt{test}}^{(\texttt{GRACE})}$. This is a time-series of RMS values for all m in $\mathcal{M}$ denoted as, $\texttt{RMS}_{\texttt{test}}^{(m)}$.

When considering the errors of multiple models, statistics such as the mean are applied to multiple error metrics. For instance, if comparing the temporal RMS for multiple models, it will be denoted as $\mu(\mathtt{RMS}_{\mathtt{test}}^{(m)})$. Other statistics that are applied are the computation of standard deviation denoted by $\sigma()$ and the computation of a percentile, for example the 5^{th} percentile, $P_5()$.

Uncertainty can only be quantified across multiple results (see Section 2.3), which are multiple sets of $\mathsf{E}_{\mathtt{test}}^{(\mathtt{model})}$. The uncertainty is quantified and represented in the following ways:

- The standard deviation of the mean EWH for each month, m, across all considered $\mathsf{E}_{\mathtt{test}}^{(\mathtt{model})}$ sets, $\sigma(E_{\mathtt{test}}^{(\mu,m,\mathtt{model})})$.

- The spatial standard deviation of the EWH for each month, m, across all considered $\mathsf{E}_{\mathtt{test}}^{(\mathtt{model})}$ sets, $\boldsymbol{\sigma}(\mathsf{E}_{\mathtt{test}}^{(\mathtt{model})})$.

The aforementioned uncertainties can also be calculated for $\mathsf{E}_{\mathtt{gap}}^{(\mathtt{model})}$, except for the errors as they require some GRACE data for quantification, which is unavailable for the months in $\mathcal{G}$.

To uncover which sources of uncertainty drive the model errors and uncertainties, models can be grouped. For example, if a group of models all have identical seed numbers for $\alpha_{\mathtt{x}}$ and $\lambda_{\mathtt{x}}$, then they can only vary in $\alpha_{\mathtt{weights}}$ and $\alpha_{\mathtt{select}}$. If the aforementioned uncertainty metrics are calculated for this group of models, then the resulting uncertainty can be attributed to the variation in $\alpha_{\mathtt{weights}}$ and $\alpha_{\mathtt{select}}$, defined as epistemic uncertainty. If grouping is applied in Chapter 9 it is indicated in-text.

6 Error Quantification

This chapter quantifies the errors in the datasets used to gap-fill GRACE EWH data. The errors in the auxiliary datasets (Swarm EWH and soil moisture data) are used to determine the uncertainty in the gap-filling results of the model (as described in Subsection 5.3.2). Furthermore, they are used in an attempt to reduce the uncertainty of these results as proposed in Section 5.4. The errors in the GRACE EWH data are used in Chapter 9 to evaluate the gap-filling performance of the model. Section 6.1 quantifies the error in the GRACE and Swarm EWH maps. Section 6.2 quantifies the errors for the GLDAS soil moisture maps.

6.1. GRACE & Swarm Data

First, the errors in GRACE and Swarm are reduced by smoothing the SH solutions prior to synthesis. The motivation for the selected smoothing radii is explained in Subsection 6.1.1. The spatial and temporal components of the error in the smoothed solutions are quantified in Subsection 6.1.2.

6.1.1. Smoothing

In the context of both GRACE and Swarm gravity field solutions, Section 2.1 describes the causes of noise, as contained in the higher degree SH terms. Through the application of Gaussian smoothing, the noise in the higher-degree terms in the SH solutions is reduced. A desired smoothing radius is achieved when the SNR of the smoothed gravity field solution is as high as possible. As the smoothing radius, r, approaches very large values, such as Earth's radius, R_e, the gravity field solution will approach its mean signal for all longitudes and latitudes (homogeneous sphere). All noise will be gone along with all information (contained in the signal) about spatial distribution of changes in Earth's gravity field. This is demonstrated in Figure 6.1 which shows the gridded map for the GRACE solution for the month January 2022 for smoothing radii: 0, 500, 3000, and 6378 [km]. For the lower two smoothing radii (left two plots), the EWH varies spatially as shown by the variation of colours red and blue, which represent different levels in EWH. On the other hand, the smoothing radii of 3000 and 6378 [km] (right two plots) show little to no variation in EWH spatially as indicated by the nearly homogeneous white maps. A balance has to be found such that as much noise as possible is removed, whilst maintaining the information with regard to the spatial distribution of true changes in Earth's gravity field.

A localized mass change over the ocean will disperse over the entire ocean because water tends to redistribute freely. Therefore, the standard deviation of EWH over the ocean, $\sigma(\mathbf{E}_{\texttt{ocean}})$, for a gravity field solution corresponding to a single month should be close to zero if GRACE or Swarm. If $\sigma(\mathbf{E}_{\texttt{ocean}})$ is close to 0, then there is no longer any signal leakage in the solution and only noise in the solutions. If $\sigma(\mathbf{E}_{\texttt{ocean}}) > 0$, this is considered an indication of the level of noise in the solution. Thus, one criterion by which the effectiveness of smoothing can be judged, is the extent to which it reduces $\sigma(\mathbf{E}_{\texttt{ocean}})$.

One result of smoothing a gravity field solution is that signals over land may spill over to signals across the ocean. This is called leakage and artificially increases $\sigma(\mathbf{E}_{\texttt{ocean}})$. This phenomenon is observed in Figure 6.1. The negative EWH (dark-blue) over Greenland corresponding to the solution of $r = 0$ [km] is blended with its surrounding region in the solution corresponding to $r = 500$ [km]. In the latter solution, the negative signals are now also more prevalent over the ocean along the coast

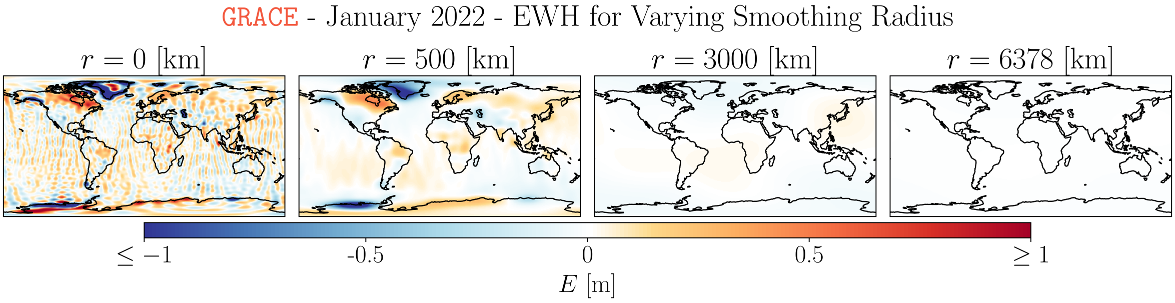


Figure 6.1: The GRACE EWH for the month January 2022 for varying smoothing radii, r.

of Greenland, whereas the noisy signals over the middle of the oceans are reduced. To remove the contribution of signal leakage to $\sigma(\mathbf{E}_{\mathtt{ocean}})$, a coastal buffer is implemented. A coastal buffer has a size b which is the distance from the true coastline to a buffered coastline. For calculations over the ocean, the area between the true coastline and the buffered coastline is not considered as part of the ocean. This means that as b increases, the area of what is considered ocean decreases. The standard deviation over the ocean with a coastal buffer of size b is denoted as $\sigma(\mathbf{E}_{\mathtt{ocean},b})$.

Figure 6.2 shows the EWH over North-America and Greenland corresponding to the GRACE solution for the month of January 2022 for smoothing radii of $r = 0$ [km] (upper-left plot) and $r = 500$ [km] (lower-left plot). In both plots, the true coastlines (black lines) are shown along with increasing coastal buffer sizes: $b = 250$ [km] (turquoise lines), $b = 500$ [km] (purple lines) and $b = 1000$ [km] (red lines). For the unsmoothed solution ($r = 0$ [km]), the land signals, such as the decrease in EWH over Greenland, are contained within the coastal buffer of $b = 250$ [km]. At the same time, this solution shows clear striping (alternating blue and orange) over the ocean. On the other hand, the solution corresponding to $r = 500$ [km] exhibits no visible striping (uniform white colour) over the ocean on the selected colour scale, while the negative EWH signal originating from Greenland is spread out further into the ocean, crossing the coastal buffer line corresponding to $b = 500$ [km].

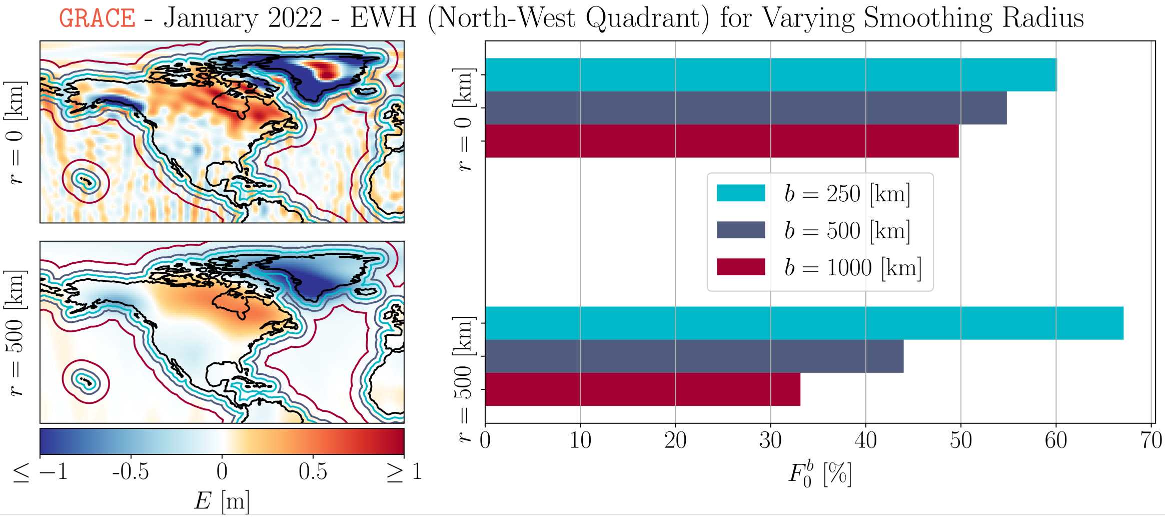


Figure 6.2: GRACE EWH solutions for month January 2022 over North-America and Greenland for smoothing radii $r = 0$ [km] (upper-right) and $r = 500$ [km] (lower-left) overlayed with buffered coastlines: $b = 250$ [km] (turquoise lines), $b = 500$ [km] (purple lines), and $b = 1000$ [km] (red lines). Additionally, the ratio of $\sigma(\mathbf{E}_{\mathtt{ocean},b}^{(202201,\mathtt{GRACE})})$ to $\sigma(\mathbf{E}_{\mathtt{ocean},0}^{(202201,\mathtt{GRACE})})$, F_0^b (right) for both smoothing radii.

On the right-hand side of Figure 6.2, the ratios of $\sigma(\mathbf{E}_{\mathtt{ocean},b}^{(202201,\mathtt{GRACE})})$ (with coastal buffer) to $\sigma(\mathbf{E}_{\mathtt{ocean},0}^{(202201,\mathtt{GRACE})})$ (no coastal buffer, $b = 0$ [km]) are plotted for the three coastal buffers and the two smoothing radii. These ratios are henceforth denoted as F_0^b. For $b = 250$ [km], the smoothed solution has a higher value for F_0^{250} than the unsmoothed solution. This is due to the signal leakage across the coastal buffer of 250 [km]. For the higher two coastal buffer sizes, the smoothed solution has lower values for both F_0^{500} and F_0^{1000} indicating that the signal leakage is removed. From this it is concluded that for selecting a smoothing radius for GRACE or Swarm, careful care has to be taken in selecting an appropriate coastal buffer as well when quantifying the variability over the ocean as a metric for the accuracy of a solution.

GRACE Smoothing Radius

The goal is to select a smoothing radius for the GRACE EWH solutions, such that $\sigma(\mathbf{E}_{\texttt{ocean}}^{(\texttt{GRACE})})$ (note that the subscript b is dropped) is minimized whilst minimizing r to preserve as much of the true signal over land. Figure 6.3 shows $\sigma(\mathbf{E}_{\texttt{ocean}}^{(\texttt{GRACE})})$ for GRACE EWH solutions as a function of r and b. Two combinations of r and b are highlighted (star markers) that are suitable for quantifying the error in GRACE. These star markers are considered suitable because they correspond to the combination of smallest smoothing radii to achieve a specific error bracket. If no coastal buffer is applied, a smoothing radius of 200 [km] (blue diamond marker) results in values of $\sigma(\mathbf{E}_{\texttt{ocean}}^{(\texttt{GRACE})})$ between 8 and 10 [cm] and is mainly influenced by signal leakage. Increase b to 300 [km] (blue triangle marker) and $\sigma(\mathbf{E}_{\texttt{ocean}}^{(\texttt{GRACE})})$ approaches the bracket of 4 to 6 [cm]. Increase the coastal buffer to 350 [km] (blue star marker) and the EWH gravity field solution reaches the ocean error bracket of 4 to 6 [cm]. Increasing b further leas to smaller changes in $\sigma(\mathbf{E}_{\texttt{ocean}}^{(\texttt{GRACE})})$, indicating that the signal leakage has been removed. For a smoothing radius of 300 [km], the lowest bracket of $\sigma(\mathbf{E}_{\texttt{ocean}}^{(\texttt{GRACE})})$ values (between 2 and 4 [cm]) is achieved for a coastal buffer of 400 [km] (red star marker).

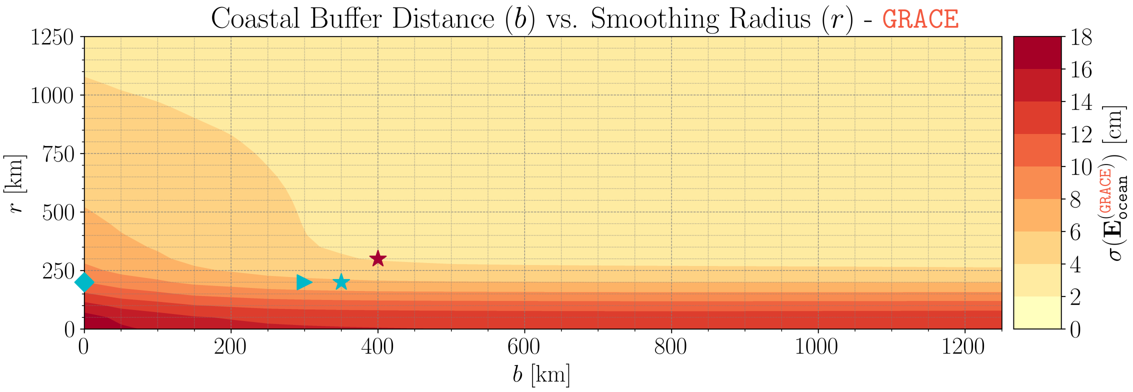


Figure 6.3: $\sigma(\mathbf{E}_{\texttt{ocean}}^{(\texttt{GRACE})})$ for GRACE for varying coastal buffer size b and smoothing radius, r.

For all GRACE solutions the smoothing radius of 300 [km] is selected along with corresponding coastal buffer, $b = 400$ [km] (corresponding to the green diamond). This means that a loss of spatial resolution over land (in comparison to $r = 200$ [km]) is accepted in favour of having a lower error, $\sigma(\mathbf{E}_{\texttt{ocean}}^{(\texttt{GRACE})})$ between 2 and 4 [cm].

Swarm Smoothing Radius

Figure 6.4 shows $\sigma(\mathbf{E}_{\texttt{ocean}}^{(\texttt{Swarm})})$ for Swarm EWH solutions as a function of r and b. There is no significant relation between b and $\sigma(\mathbf{E}_{\texttt{ocean}}^{(\texttt{Swarm})})$ as the colours on the plot vary mainly along the y-axis. The lowest levels of $\sigma(\mathbf{E}_{\texttt{ocean}}^{(\texttt{Swarm})})$ (between 0 and 40 [cm]) are achieved for Swarm smoothing radii of 450 [km] and upwards. The 40 [cm] ocean variability indication is larger than the worst ocean variability indication for GRACE of 20 [cm]. This reinforces the fact that Swarm EWH solutions are noisier than GRACE EWH solutions.

The smoothed Swarm EWH data and its errors will be used to train NNs to predict GRACE-like EWH data over basins which are on land. Therefore, it is decided to select a Swarm smoothing radius that minimizes the RMS of the difference between Swarm EWH and GRACE EWH over land, $\texttt{RMS}(\mathrm{E}_{\texttt{land}}^{(\texttt{GRACE})}, \mathrm{E}_{\texttt{land}}^{(\texttt{Swarm})})$. Figure 6.5 shows the mean and maximum of $\texttt{RMS}(\mathrm{E}_{\texttt{land}}^{(\texttt{GRACE})}, \mathrm{E}_{\texttt{land}}^{(\texttt{Swarm})})$ denoted as $\epsilon_{\texttt{mean}}$ (solid blue line) and $\epsilon_{\texttt{max}}$ (dashed blue line), respectively, for varying Swarm smoothing radii, r. The minimum of $\epsilon_{\texttt{mean}}$ is 4 [cm] and occurs at $r = 850$ [km] (red diamond marker). The minimum of $\epsilon_{\texttt{max}}$ is 20.7 [cm] and occurs at $r = 650$ [km] (red circle marker). Finally, a smoothing radius of 750 [km] is selected as this smoothing radius results in a mean error of 4.2 [cm] (green star on solid line) and a maximum error of 21.4 [cm]

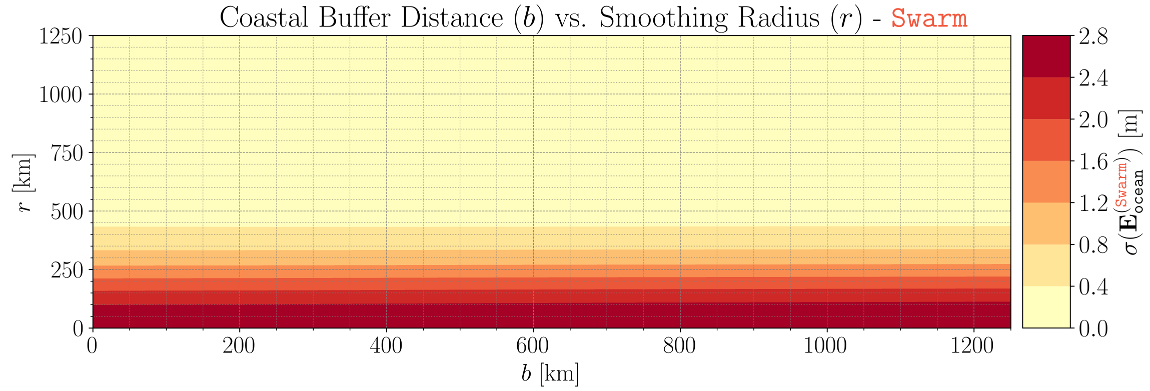


Figure 6.4: $\sigma(\mathbf{E}_{\texttt{ocean}}^{(\texttt{Swarm})})$ for Swarm for varying coastal buffer size b and smoothing radius, r.

(green star on dashed line). This solution is preferred over selecting $r = 850$ [km] because features of smaller spatial scale are maintained whilst still having a low mean error.

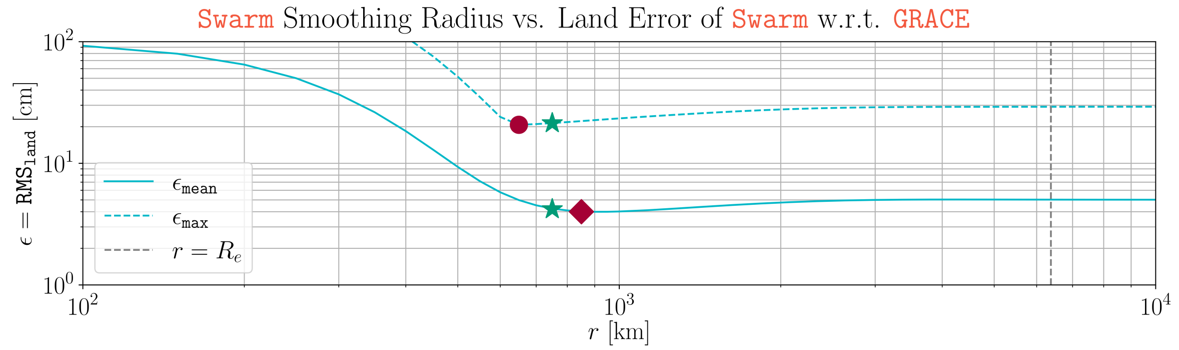


Figure 6.5: The mean and maximum RMS of the difference between Swarm EWH solutions and GRACE EWH solution (with smoothing radius of 300 [km]) for varying Swarm smoothing radii.

6.1.2. Error

To quantify the monthly errors in the GRACE EWH data, it is assumed that the standard deviation of linearly de-trended (removal of trend and bias) EWH over the ocean for a particular month, $\sigma(\mathbf{E}_{\texttt{ocean},400}^{(m,\texttt{GRACE},\texttt{detrend})})$ (coastal buffer is 400 [km] and smoothing radius is $r = 300$ [km]), is the best indication of the monthly errors in GRACE, $\xi^{(m,\texttt{GRACE})}$. These errors are used in Chapter 9 to evaluate whether a NN gap-filling model is performing as best as possible.

Figure 6.6 shows $\xi^{(m,\texttt{GRACE})}$ (green dots) for all available GRACE months. To show why the EWH is de-trended, $\xi^{(m,\texttt{GRACE})}$ based on EWH with a trend (faded red dots) is shown as well. After de-trending, $\xi^{(m,\texttt{GRACE})}$ no longer makes it appear as if GRACE's error are linearly increasing from 2008 onwards. The period of 2005-2009 is known as the nominal noise-bottom for GRACE as the errors are consistently low.

Examining GRACE's error over time, four types of outliers are distinguished. The first type, covers the first two months of GRACE data (highlighted by orange hexagons) which show high indications of ocean variability. In these months, the accelerometer data of GRACE-B was missing leading to higher errors in the estimate gravity field solution (Ries & Bettadpur, 2003). The second category of outliers (orange crosses) pertains to three months (December 2002 to February 2003) of GRACE data which suffered from data interruptions due to planned flight manoeuvrers (Ries & Bettadpur, 2003). The second category of outliers (solutions highlighted by orange squares) are caused by the GRACE satellites experiencing orbit resonance. Due to this resonance, the ground track coverage is low. This means that the satellites do not fly over as much of Earth's surface and the resulting EWH solutions are of a lower quality (McGirr et al., 2023). The final category of outliers (solutions highlighted by orange diamonds) are caused by the failure of an accelerometer in one of the GRACE satellites, GRACE-B. Due to this failure, the acceleration of GRACE-B is estimated by transplanting the acceleration of GRACE-A (Dahle, 2018). This transplantation is much less accurate resulting in high errors for these months. These four causes of error outliers

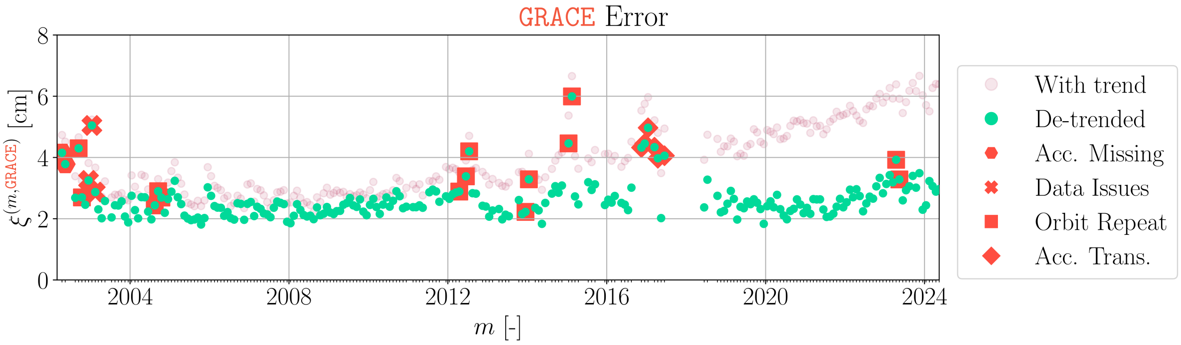


Figure 6.6: The temporal error of monthly GRACE EWH solutions ($r = 300$ [km]) based on the standard deviation of de-trended EWH over a buffered ocean ($b = 400$ [km]).

explain the outliers observed in Figure 6.6. Using this mission context, the results are analysed with a more complete picture in Chapter 9.

To generate noise for the training and testing data (as described in Section 5.3), the monthly error in the Swarm EWH data per basin, $\xi_{\mathtt{basin}}^{(m,\mathtt{Swarm})}$ are quantified. The goal of the NN's is to relate Swarm EWH data to GRACE EWH data, therefore, $\xi_{\mathtt{basin}}^{(m,\mathtt{Swarm})}$ is quantified by combining spatial RMS difference of GRACE EWH and Swarm EWH over each river-basin, $\mathtt{RMS}_{\mathtt{basin}}$ (Equation 6.1), and the monthly RMS difference between GRACE and Swarm over all land, $\mathtt{RMS}_{\mathtt{land}}^{(m)}$ (Equation 6.2), as shown in Equation 6.3. The former component is basin specific and has a spatial component. On the other hand, the latter component is a general indication of Swarm's quality over time.

$$\mathtt{RMS}_{\mathtt{basin}} = \mathtt{RMS}(\mathsf{E}_{\mathtt{basin}}^{(\mathtt{GRACE})} - \mathsf{E}_{\mathtt{basin}}^{(\mathtt{Swarm})}) \tag{6.1}$$

$$\mathtt{RMS}_{\mathtt{land}}^{(m)} = \mathtt{RMS}(\mathbf{E}_{\mathtt{land}}^{(m,\mathtt{GRACE})} - \mathbf{E}_{\mathtt{land}}^{(m,\mathtt{Swarm})}) \tag{6.2}$$

$$\xi_{\mathtt{basin}}^{(m,\mathtt{Swarm})} = [\mathtt{RMS}_{\mathtt{basin}}]^{\frac{1}{2}}[\mathtt{RMS}_{\mathtt{land}}^{(m)}]^{\frac{1}{2}} \tag{6.3}$$

To give context to the Swarm's land error w.r.t. GRACE, the mean of each month's error is plotted for each basin (left axis, each coloured line for a basin) in Figure 6.7. Additionally, the right-axis of this plot also shows the standard deviation of Swarm over the ocean per month (right axis, gray line). It can be observed, that the coloured lines, all exhibit higher errors at the start of the Swarm mission. In addition, the variability over the ocean of the Swarm solutions is also particularly high in the first year of the mission. These high errors are driven by the fact that during this period there was a high solar flux impacting the GPS signals received by the Swarm constellation (Encarnação & Visser, 2024). Upon discovery of this high solar flux, the GPS receivers of the Swarm satellites were tuned to be less susceptible to this radiation at the cost of a slightly lower accuracy. From 2014 to 2019 the solar flux decreases, only to start rising again to the even higher levels in 2024. As a result of the adjustment of the GPS receivers, the error indications of the Swarm solutions are much lower whilst experiencing much higher solar flux values from 2023 to 2024 than in the period of 2014 to 2015. The basin errors also reflect that the tuning of the GPS receiver did indeed decrease these errors. Swarm's basin errors shown in Figure 6.7 are used to evaluate the performance of the trained NN models in Chapter 9.

6.2. Soil Moisture Data

The error of the soil moisture data is defined as 10% of its own standard deviation and without a temporal component. Taking a fraction of the standard deviation as the error is justified by the assumption that the observational errors are proportional to the variability of the observed signal. A value of 10% is chosen because it is greater than 0% which assumes that some of the signal is noise. The value is also a lot less than 100% which assumes that most of the variability is due to true signal and not noise. Without literature arguing for a temporal error component to GLDAS soil moisture, it is deemed out of scope to deduce a method for quantifying said error component for this thesis. Without this component, the error in the soil moisture is modelled as constant for all months, $\xi_{\Delta S}^{(m)} = \xi_{\Delta S}$ (the superscript for month, m is dropped).

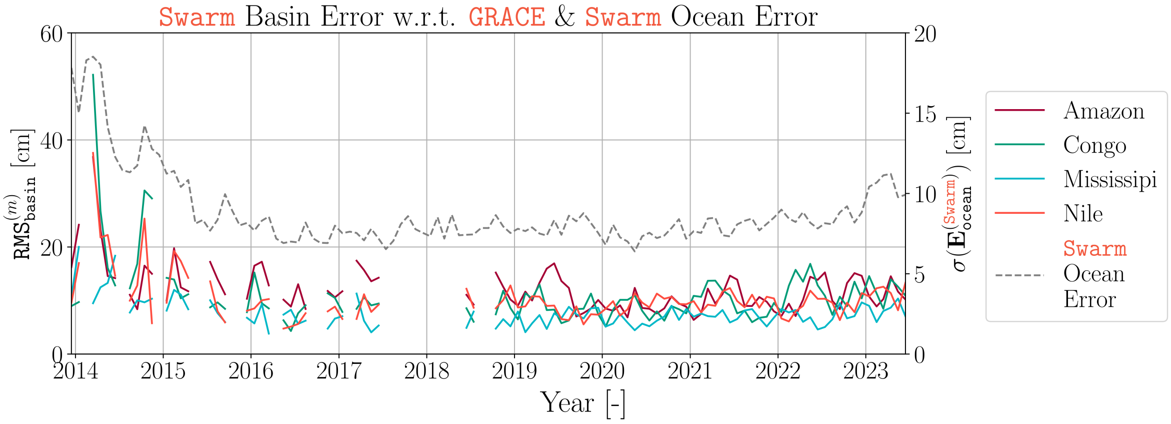


Figure 6.7: The temporal error of monthly Swarm EWH solutions ($r = 750$ [km]) based on the RMS difference with GRACE EWH solutions over each basin on the left axis. The temporal standard deviation over the ocean of Swarm EWH (gray line).

The error in soil moisture is shown in Figure 6.8. Over the East Amazon river basin the signals are particularly strong as indicated by the darker shades of red corresponding to approximately 15-20 [kgm^{-3}]. The Mississippi and Congo river basins indicate errors between 7.5 and 10 [kgm^{-3}]. Over the Nile river basins, the errors in the soil moisture are virtually zero as indicated by the yellow colour. This can be attributed to the dryness of the region.

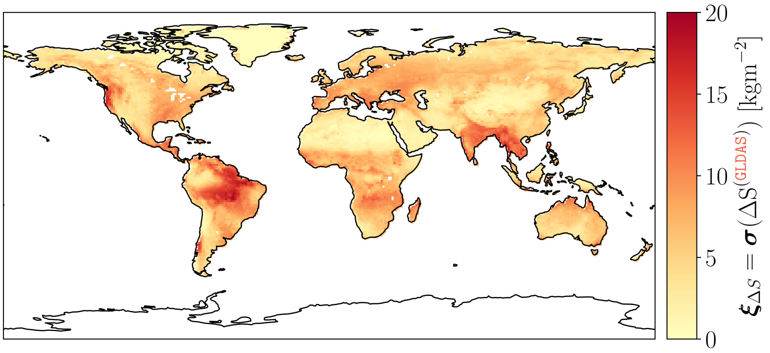

20
15
10
5
0
$\xi_{\Delta S} = \boldsymbol{\sigma}(\Delta \mathrm{S}^{(\mathtt{GLDAS})})$ [kgm^{-2}]

Figure 6.8: 10% of the standard deviation of the GLDAS NOAH soil moisture data.

7

Selecting Neural Network Hyper-parameters

A fully-connected NN is implemented for the investigation of the uncertainty of GRACE-like data produced by NNs. Hyper-parameters characterize the training process of a NN because they influence: its duration and the final weights and biases of the trained NN. The latter is what determines the performance of the NN. A sensitivity analysis is performed in order to select and set particular hyper-parameters required to train a NN. The architecture shown in Figure 3.3 is used as a blueprint for the NN. First, the required hyper-parameters are discussed in Section 7.1. To improve the performance of the architecture, several augmentations are considered in Section 7.2. In this context, a particular choice is considered an augmentation if it is not strictly required to create and train a NN. Section 7.3 explains the method of optimization used to select the hyper-parameters and augmentations. Finally, Section 7.4 explains and motivates the selected hyper-parameters and augmentations.

7.1. Required Hyper-parameters

All hyper-parameters considered in the sensitivity analysis fall under one of three categories: `Integer`, `Continuous` (assume floating point values) or `Categorical` (assume discrete categories). Table 7.1 contains the list of hyper-parameters that are considered in this study. For the `Integer` and `Continuous` hyper-parameters, the range of values is shown. In case of an `Integer` hyper-parameter, all integers between the range bounds are considered. In the case of a `Continuous` hyper-parameter, all real numbers between the range bounds are considered as viable selections. For the `Categorical` hyper-parameters, the categories are shown. The selected ranges are chosen by evaluating selected hyper-parameters in literature and then creating a range about these hyper-parameters (Harrison, 2023).

Table 7.1: Required hyper-parameters, their types, and the explored ranges/categories.

Type	Hyper-parameter	Range / Categories
`Integer`	Number of epochs	$[50,...,300]$
`Integer`	Batch size	$[1,...,30]$
`Continuous`	Fraction of data used for training, γ	$[0.7,...,0.9]$
`Continuous`	Initial learning rate	$[0.00001,...,0.01]$
`Continuous`	Multiplicative learning rate	$[0.9,...,1.0]$
`Categorical`	Activation function for internal layers	[sigmoid (Equation 2.22), softplus (Equation 2.24), ReLU (Equation 2.23)]
`Categorical`	Activation function for output layer	[sigmoid, softplus, ReLU]

7.2. Augmentations

Table 7.2 shows the considered augmentations. One of the augmentations is a Continuous parameter, called F_l, which is a factor by which the size of inner layers (i.e., those that are not the input and output ones) are increased or decreased: [0.5,...,1.5]. Increasing them leads to a greater complexity and therefore greater ability to learn, but doing so in excess leads to over-fitting. It also leads to increased training times as more computations have to be performed during the optimization process.

Table 7.2: The investigated augmentations, their types, and the explored ranges/categories.

Type	Augmentation	Range / Categories
Continuous	Factor by which the size of inner layers is increased, F_l (decreased for values lower than 1).	[0.5,...,1.5]
Categorical	Add dense linear layer after output.	[yes, no]
Categorical	Use $\Delta\Theta$ instead of Θ for soil moisture.	[yes, no]
Categorical	Mask back-propagation gradient with basin land mask.	[yes, no]

The other three considered augmentations are Categorical in the sense that they are either implemented (”yes”) or not implemented (”no”). It is hypothesized that the addition of a dense linear layer (activation function, $n(z) = z$) can lead to improved performance. If the NN has a sigmoid function for the activation of the output layer, each element in a GRACE-like map will be in the range of [0,...,1]. EWH typically has both negative and positive values which may in magnitude be larger than 1. Therefore, adding a linear layer may lead to improved results because the range of the NN output data (GRACE-like data) can be scaled to a more favourable range than [0,...,1].

Another augmentation that is added to the list of hyper-parameters is whether to use monthly soil moisture w.r.t. to a mean, $\Delta\Theta$ instead of absolute monthly soil moisture, Θ for the soil moisture part of the input data. The GRACE EWH maps represent a change in gravity field relative to a mean gravity field. The goal of a NN is to relate auxiliary input data to output data. The closer the auxiliary input data matches the output data in information content, the less complex a NN needs to be to find a relationship between the two. Therefore, using soil moisture maps that also represent changes with respect to a mean, could require a less complex NN to establish a relationship between the soil moisture part of the input and the desired GRACE-like data.

The goal of the NN is to gap-fill over a particular basin. The fourth augmentation is a question of whether to apply a mask to the back-propagation process (used for training the NN) such that the difference between auxiliary and GRACE EWH data is only minimized over the output elements of the grids that are actually inside the basin. This might be beneficial for two reasons. First, the NN has to build less relationships between input and target data as the size of the latter will have decreased. This might result in a less complex NN which reduces training time and prevents over-fitting. The second reason is more basin specific. The shape of the datasets is determined by the rectangular grid of latitudes and longitudes that fit around the basin in question. The basins used for this thesis are again shown in Figure 7.1. The green cells represent the basin and the red cells represent the area outside the basin which are masked if the augmentation is implemented. The rectangular outline of each basin (entire image) contains some cells which are over water (hashed black lines). Attempting to predict GRACE EWH over the ocean is not useful, as the signals seen over the oceans are noise-dominated (see Section 6.1). This means, that if the mask is applied, i.e. only the green cells of value 1 are used for back-propagation, the NN does not have to establish a relationship between auxiliary data and GRACE’s noisy signal over the ocean. This could result in a less complex NN being required for the same level of performance.

7.3. Optimization

The Optuna[1] framework is utilized to determine the optimal selection of hyper-parameters and augmentations. Within the range of possible values for all the hyper-parameters and augmentations (search-space), the framework draws different combinations of values for these hyper-parameters, samples. Each sample, is evaluated. In this case, a sample evaluation is the training of a NN and subsequent testing of NN performance. The sampling strategy determines what the next drawn sample will be

[1] https://optuna.org/

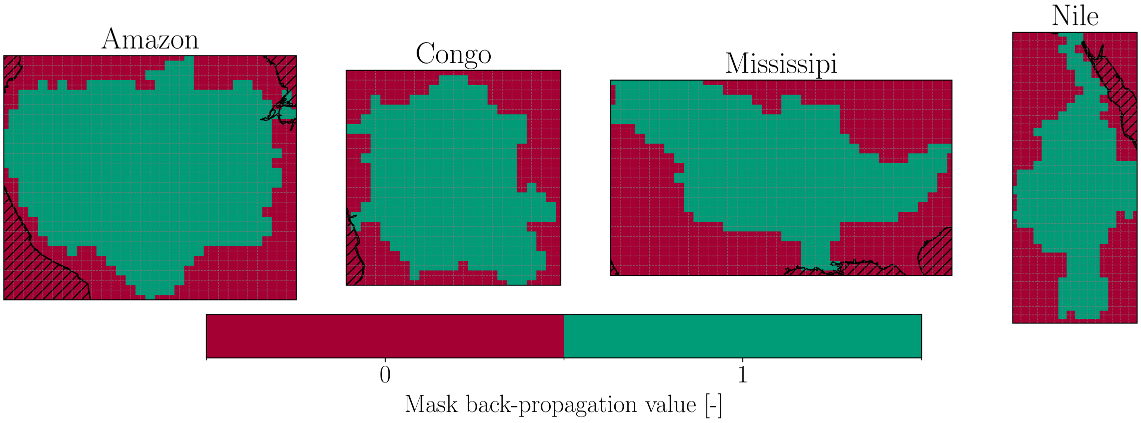


Figure 7.1: The masks applied to the back-propagation process if the considered augmentation is applied per basin. The hashed black lines indicate seas or oceans.

based on the performance of the previous samples. One such strategy is to draw samples at random. However, some sampling strategies can find optimal solutions more quickly. Depending on the search-space, the Optuna framework recommends different sampling strategies. As the NN requires, `Integer`, `Continuous`, and `Categorical` variables, the Tree-Structured Parzen Estimator (TPE) sampling algorithm is recommended.

The TPE sampler creates probability distributions for hyper-parameters that are promising and unpromising and draws new samples based on these probability distributions. It balances exploration and exploitation through use of a threshold parameter which dictates when a hyper-parameter is considered promising or not. If there is a higher threshold for promising values, then the algorithm is considered ”greedy” because it exploits promising solutions without considering that less promising solutions may lead to the best solution. A lower threshold for promising values leads to an algorithm which explores more as it does not lay too much focus on the most promising solutions. This may lead finding more optimal solutions at the cost of a longer time to find the most optimal solution.

In the context of NN's there is one problem that the TPE sampling algorithm faces. There is a significant amount of epistemic uncertainty in the results of a NN when not fixing any seed numbers for the random number generators. Figure 7.2 shows that for a fixed set of required hyper-parameters, models trained with varying seed numbers exhibit different test NSE values. When only varying $\alpha_{\texttt{weights}}$, the spread is between 0.555 and 0.570 [-]. When only varying $\alpha_{\texttt{select}}$ the spread is between 0.53 and 0.57 [-]. The larger spread when varying $\alpha_{\texttt{select}}$ suggests that the NN is more sensitive to the uncertainty caused by which data is selected for training. When both seeds are varied at once, the range of NSE values is between 0.5 and 0.57 [-]. This shows that the effects of both seeds interact and compound.

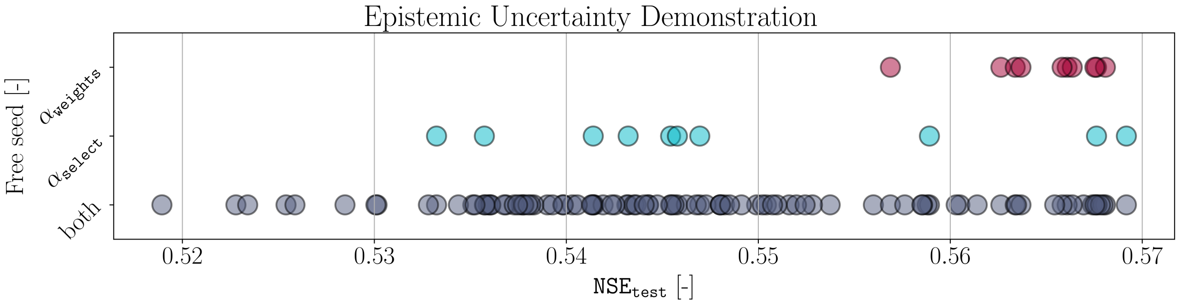


Figure 7.2: Test NSE values for trained models with varying seed numbers and constant model hyper-parameters and choices.

Finding a relationship between the hyper-parameters and NN performance is a computationally expensive task for the TPE sampler. Introduce epistemic uncertainty into the process and the distinction between promising and unpromising samples becomes less clear. This leads to the TPE sampler requiring more samples to find an optimal set of hyper-parameters. Fixing the seed numbers is, while logical, not a good solution as this could result in cherry-picking the best results and ignoring the uncertainties involved in NNs. This would inevitably result in a sub-optimal set of hyper-parameters being chosen.

If time and computational effort are not limiting, each hyper-parameter should be sampled multiple times to account for uncertainty. The Optuna framework would run the experiment detailed in Chapter 5 for all possible combinations of hyper-parameters and choices detailed in Table 7.1 and Table 7.2. Assuming that, for the continuous variables, the domain is made finite with 10 equidistant samples, then the total amount of possible combinations is 5.4×10^9 [-]. The total training time per combination, $T_{\texttt{training}}$, required to evaluate a model's uncertainty is estimated at approximately 10 [days] utilizing 50 CPU's. This would mean that finding the optimal set of hyper-parameters and augmentations that results in the lowest uncertainty, would take approximately 150 million years. Even if the TPE sampler could reduce the amount of combinations to test to the order of thousands, finding the optimal NN design would take 150 years, which is still not considered feasible. Therefore, it is decided to accept the set of hyper-parameters and augmentations determined by Optuna knowing that they are likely sub-optimal due to epistemic and aleatoric uncertainty. This is acceptable because estimating the uncertainty of the selected NN is the goal of this thesis, while finding a true optimal set of hyper-parameters is not.

7.4. Final Hyper-parameters & Augmentations

Table 7.3 shows the hyper-parameters and augmentation decisions selected by the TPE sampler. Providing an explanation for the framework's choice of hyper-parameters is difficult. For example, explaining why a batch size of 20 was chosen as opposed to a batch size of 19 is not realistic because there is no direct insight into the logic of a choice made by the optimizer.

All Categorical augmentations were selected by the Optuna framework. As predicted, the addition of a dense linear layer is in combination with the selection of the softplus function for the hyper-parameter of the activation function of the output layer. Furthermore, the NN is shrunk by 20% in size. This means that the Optuna framework found its best solutions using a NN with less learnable parameters.

Table 7.3: The hyper-parameters and augmentations selected by the Optuna framework for the NN.

Type	Hyper-parameter	Selection
Integer	Number of epochs	200 [-]
Integer	Batch size	20 [-]
Continuous	Fraction of data used for training, γ	0.8 [-]
Continuous	Initial learning rate	0.000576 [-]
Continuous	Multiplicative learning rate	0.9985 [-]
Categorical	Activation function for internal layers	Softplus
Categorical	Activation function for output layer	Softplus
Type	**Augmentation**	**Selection**
Categorical	Additional dense linear layer	Yes
Categorical	$\Delta\Theta$ in stead of Θ for soil moisture	Yes
Categorical	Back-propagation mask	Yes
Continuous	Factor for size of inner layers, F_l	0.8 [-]

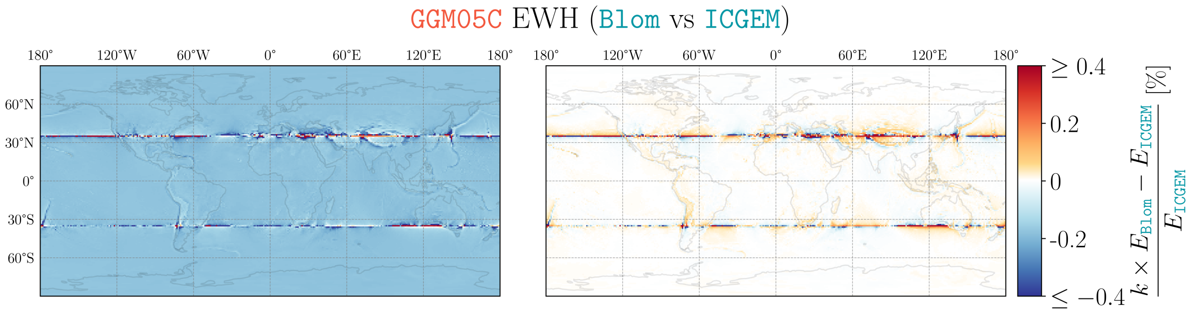
8 Model Verification

This chapter focuses on the verification of the implemented algorithms. In Section 8.1, the implementation of spherical harmonic synthesis as outlined in Section 2.1 is verified. The functioning of the trained NNs is verified in Section 8.2.

8.1. Spherical Harmonic Synthesis Implementation

To verify the implementation of spherical harmonic synthesis, the calculation service[1] of ICGEM is used. As a reference case, the static gravity field model, GGM05C is scaled to EWH and synthesized. This results in a grid, hereafter named $\mathbf{E}_{\texttt{Blom}}$. To validate $\mathbf{E}_{\texttt{Blom}}$, ICGEM's calculation service is used to produce a gridded EWH map corresponding to GGM05C, hereafter named $\mathbf{E}_{\texttt{ICGEM}}$. The subscripts Blom and ICGEM are used to distinguish between the authors of the grids.

The left plot in Figure 8.1 shows the fractional difference between $\mathbf{E}_{\texttt{Blom}}$ and $\mathbf{E}_{\texttt{ICGEM}}$ relative to $\mathbf{E}_{\texttt{ICGEM}}$ expressed as a percentage. Three phenomena can be observed. First, there is a constant difference of approximately -0.2 [%] for all longitudes and latitudes indicated by the blue color across the map. Secondly, the contours of several topographic features on Earth can be observed, such as the Himalayas and the Andes. Finally, at approximately ± 35 [deg] latitude there are strong differences between $\mathbf{E}_{\texttt{Blom}}$ and $\mathbf{E}_{\texttt{ICGEM}}$ indicated by the dark-blue and dark-red pixels. Barring the strong differences at approximately ± 35 [deg] latitude, all differences between $\mathbf{E}_{\texttt{Blom}}$ and $\mathbf{E}_{\texttt{ICGEM}}$ are within ± 0.3 [%].


GGM05C EWH (Blom vs ICGEM)
180°
120°W
60°W
0°
60°E
120°E
180°
60°N
30°N
0°
30°S
60°S
≥ 0.4
0.2
0
-0.2
≤ −0.4
$k \times E_{\texttt{Blom}} - E_{\texttt{ICGEM}}$ / $E_{\texttt{ICGEM}}$ [%]

$k = 1$ $\qquad$ $k = 1.0018$

Figure 8.1: The fractional difference between $\mathbf{E}_{\texttt{Blom}}$ and $\mathbf{E}_{\texttt{ICGEM}}$ relative to $\mathbf{E}_{\texttt{ICGEM}}$ expressed as a percentage. The left plot shows the original difference, the right plot shows the difference after scaling $\mathbf{E}_{\texttt{Blom}}$ by a factor 1.0018 [-].

The constant percentage difference between $\mathbf{E}_{\texttt{Blom}}$ and $\mathbf{E}_{\texttt{ICGEM}}$ can be removed by applying a scaling factor, k to $\mathbf{E}_{\texttt{Blom}}$. The factor, k, is deduced by computing the mean difference between $\mathbf{E}_{\texttt{Blom}}$ and $\mathbf{E}_{\texttt{ICGEM}}$ over the 10 by 10 pixels in the South-West of the globe: in this region there appear to be no irregular differences but only a constant difference. The constant factor, k, is calculated to be 1.0018 [-]. The

[1] https://icgem.gfz-potsdam.de/calcgrid

effectiveness of this factor is shown in the right plot in Figure 8.1, where most of the map now colours white. This corresponds to a negligible difference. The only constant scaling factors in Equation 2.3 are the equatorial radius of the Earth, R_e, and the ratio of densities of Earth and water, $\frac{\rho_e}{\rho_{\texttt{water}}}$. Of these two constants, ICGEM's calculation service only allows the specification of R_e. This means that the constant factor difference may be attributed to a different value for $\frac{\rho_e}{\rho_{\texttt{water}}}$ being used by ICGEM. The second plot shows that only the second and third differences observed remain after the constant scaling factor is applied.

Figure 8.2 shows the RMS of the percentage difference between $\mathbf{E}_{\texttt{Blom}}$ and $\mathbf{E}_{\texttt{ICGEM}}$ per latitude for both scaling factors ($k = 1$ and $k = 1.0018$ are represented by red and green lines respectively). After applying the constant scaling factor, the percentage difference is reduced to values below 0.01 [%]. The strongest difference that remains with percentage differences up to 100 [%] is located near the ± 35 [deg] latitudes. These differences remain unexplained.

The strong differences near the ± 35 [deg] latitudes can only effect the results of gap-filling the Mississippi river basin as it is the only basin within these latitudes. Given that the method of synthesis applied for this thesis is applied consistently for both Swarm and GRACE data, these strong differences are neglected. The constant scaling factor difference is considered negligible given that the resulting percentage difference is only -0.2 [%]. Therefore, the implementation of the EWH functional and the synthesis of spherical harmonic coefficients is considered verified.

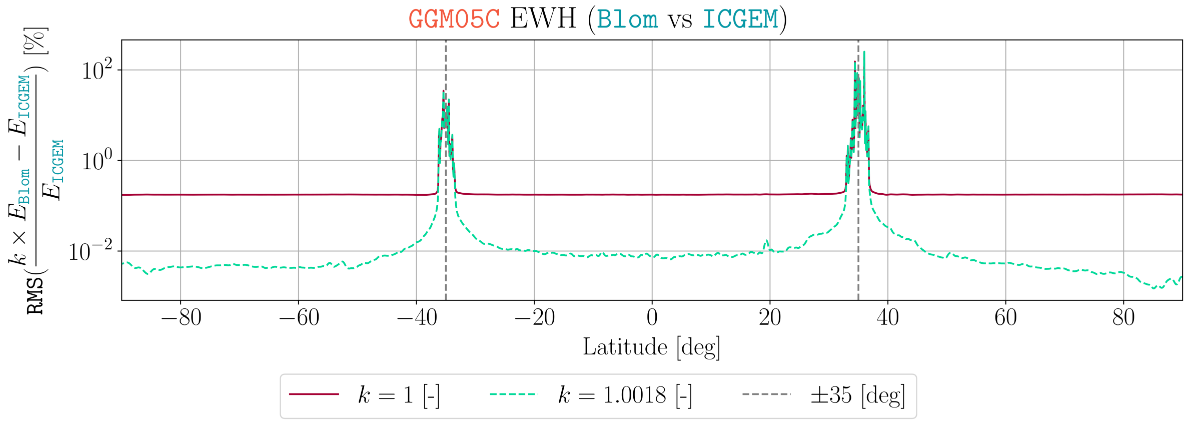


Figure 8.2: Latitude dependent RMS difference of both absolute and percentage different between $\mathbf{E}_{\texttt{Blom}}$ and $\mathbf{E}_{\texttt{ICGEM}}$ for different scaling factors: $k = 1$ [-] and $k = 1.0018$ [-].

8.2. Neural Network Implementation

The process of verification of the selected NN to work as intended is defined as the ability of a NN to relate some lower resolution EWH data to some higher resolution EWH data, such that it can be used for gap-filling GRACE. For each basin, a model trained with a unique set of seed numbers and η value is selected randomly. These models are shown in Table 8.1.

Table 8.1: Randomly selected models, the basin they are trained for, the eta they are trained with and their seed characteristics.

Model	Basin	η	α_x	$\alpha_{\texttt{weights}}$	$\alpha_{\texttt{select}}$	λ_x
Amazon_000_000_000_006_002	Amazon	0	0	0	6	2
Congo_004_007_001_003_012	Congo	4	7	1	3	12
Mississippi_005_002_001_009_019	Mississippi	5	2	1	9	19
Nile_001_004_001_006_019	Nile	1	4	1	6	19

To confirm that these models work as intended, the RMS difference between the model mean EWH predictions over a basin and GRACE mean EWH over a basin is computed for all the model's training and testing months. Swarm is also compared to GRACE in the same way. Figure 8.3 shows the calculated errors for Swarm (red dots) and each model (each row is a model and identified by basin, β, errors are represented by blue dots). For the Congo, Mississippi, and Nile river basins the models' errors are a magnitude lower than Swarm for each month. This means that the model has a lower error than Swarm and that the models successfully capture GRACE's signal. For the Amazon basin, the model outperforms Swarm in all but one training month in 2015. All other errors are also a magnitude lower. This magnitude difference in error over the testing months for all basins, shows that the NNs have low errors over testing months which they have not been trained on. Therefore, the implementation training NNs is considered verified.

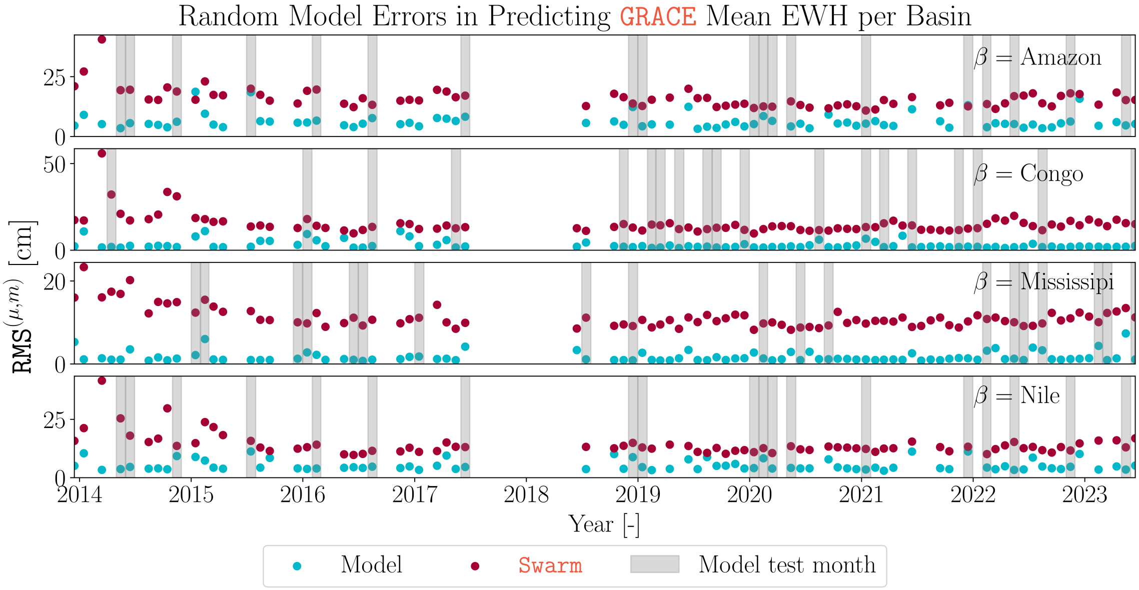


Figure 8.3: Model and Swarm RMS errors w.r.t. GRACE for train (white background) and test (gray background) months.

9 Results

This chapter presents the analysis of the $20.4 \cdot 10^3$ trained NN GRACE gap-filling models. First, the performance of the models is evaluated by their ability to predict mean EWH (one number per month) over each basin as opposed to the exact EWH per grid cell (one number per month per spatial grid cell) in Section 9.1. Then, the models are analyzed for their ability to predict GRACE-like EWH temporally and spatially in Section 9.2 and Section 9.3, respectively.

9.1. Errors & Uncertainty of Predicting GRACE Mean EWH

To get a general idea of the performance of the trained NNs, their ability to predict the mean GRACE-like EWH per basin per month is evaluated. For evaluation, each basin is considered as if it had no spatial dimensions. Predicting mean EWH is easier than per spatial grid cell because averaging over a basin reduces the impact of spatial variability and local errors, effectively smoothing out noise and biases present in individual grid cells. Without the spatial dimension, Figure 9.1 shows how well the models predict mean GRACE-like EWH for each basin by showing the distribution of the NSE (top row), RMS (middle row), and CC (bottom row) values for both training and testing. The blue dots represent the mean model error. The whiskers represent the bounds within which the top 95 [%] of the model errors lie. The orange and green dashed lines represents Swarm's and GLDAS soil moistures' mean error with respect to the GRACE mean EWH. The error for soil moisture is not included for the NSE and RMS errors as these metrics can only compare data with the same units.

The average of all models outperforms Swarm and soil moisture in predicting the mean EWH per basin. This is because all the blue dots are above the dashed lines in the NSE and CC plots and they are below the dashed lines in the RMS plots. For the Congo, Mississippi and Nile river basins the whiskers do not cross the dashed lines. This means that for these river basins, more than 95 [%] of the models outperform Swarm and soil moisture. For the Amazon river basin, the whiskers do overlap with the dashed orange lines. This means that less than 95 [%] of the models outperform Swarm.

The larger the whiskers are, the higher the uncertainty in the models. For all river basins, the mean error and the uncertainty in the training data are lower than in the testing data. This is expected because the models are trained to fit the training data. Testing data is unseen and therefore, more difficult to predict.

The mean model test errors shown in the aforementioned plots are shown in Table 9.1. This table also shows the mean basin amplitude in centimetres. The average model's RMS values is less than 22 [%] of the corresponding mean basin amplitude. To answer the second research question, the effect of increasing η on a model's performance is evaluated in Subsection 9.1.1. To explain phenomena observed in Subsection 9.1.1, Subsection 9.1.2 is dedicated to exploring the role seed numbers play in the errors and uncertainty of the produced models. In Subsection 9.1.3, gap-filling results when predicting mean EWH over a basin are presented.

9.1.1. Effect of Additional Training Data

The second research question reads: how does the inclusion of additional training data, generated using errors in the auxiliary datasets, affect the errors and uncertainty of GRACE-like data produced

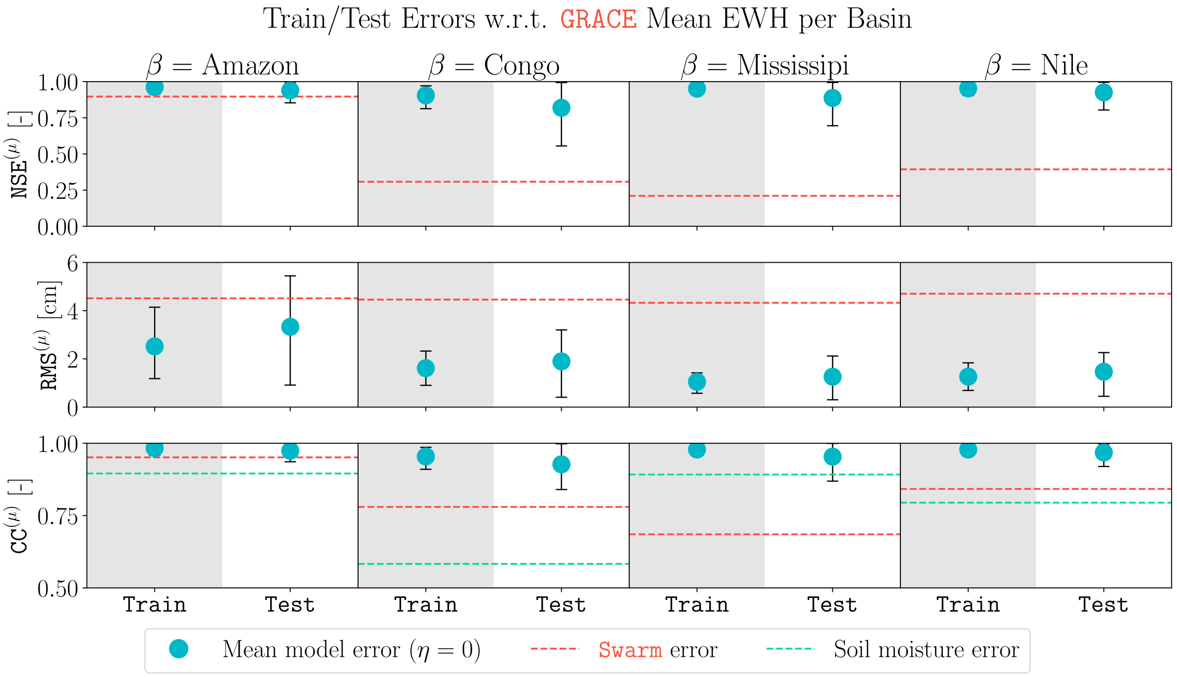


Figure 9.1: Mean training and testing NSE, RMS, and CC of all models ($\eta = 0$) w.r.t. GRACE basin mean EWH for each basin, β.

Basin	Basin amplitude	Mean Model Test Error ($\eta = 0$)		
β	$\mu(\sigma_{\xi}^{\mathtt{GRACE}})$ [cm]	$\mu(\mathtt{NSE}^{(\mu)})$ [-]	$\mu(\mathtt{CC}^{(\mu)})$ [-]	$\mu(\mathtt{RMS}^{(\mu)})$ [cm]
Amazon	18	0.940	0.974	3.3
Congo	11	0.819	0.927	1.9
Mississippi	6	0.887	0.954	1.3
Nile	8	0.926	0.968	1.5

Table 9.1: Mean model ($\eta = 0$) test errors (w.r.t. GRACE mean EWH data) for each basin, β.

by NNs? The hypothesis that is formulated states that increasing the amount of additional training data samples will result in a reduced error and uncertainty in the NN generated GRACE-like EWH data. Figure 9.2 shows the mean model test errors when predicting GRACE mean EWH data as a function of η (x-axis) per basin (each line): NSE, $\mu(\mathtt{NSE}^{(\mu)}_{\mathtt{test}})$ (left most plot), and RMS, $\mu(\mathtt{RMS}^{(\mu)}_{\mathtt{test}})$ (middle plot). The right-most plot shows the standard deviation of the model test RMS errors, $\sigma(\mathtt{RMS}^{(\mu)}_{\mathtt{test}})$, for each basin as a function of η. The standard deviation of the RMS represents the uncertainty in the error of the models for each combination of basin and η.

For all basins, the errors and uncertainties are lowest ($\mu(\mathtt{NSE}^{(\mu)}_{\mathtt{test}})$ highest, $\mu(\mathtt{RMS}^{(\mu)}_{\mathtt{test}})$ lowest, and $\sigma(\mathtt{RMS}^{(\mu)}_{\mathtt{test}})$ lowest) for models that have been trained with $\eta = 5$. The changes in mean error are significant from $\eta = 0$ to $\eta = 5$. The Amazon sees a drop in $\mu(\mathtt{RMS}^{(\mu)}_{\mathtt{test}})$ of about 1.5 [cm] and the other three basins of approximately 0.5 [cm]. The decrease in uncertainty is most significant for the Amazon river basin which sees a drop of $\sigma(\mathtt{RMS}^{(\mu)}_{\mathtt{test}})$ from 1.2 [cm] to 0.8 [cm]. The other basins see less significant drops of about 0.1 [cm] in RMS uncertainty.

A noteworthy outcome is that for the Amazon, increasing from $\eta = 0$ [-] to $\eta = 1$ [-], first leads to an increase in both error ($\mu(\mathtt{NSE}^{(\mu)}_{\mathtt{test}})$ decreases, $\mu(\mathtt{RMS}^{(\mu)}_{\mathtt{test}})$ increases) and uncertainty ($\sigma(\mathtt{RMS}^{(\mu)}_{\mathtt{test}})$ increases). The models trained for the Mississippi experience an increase in uncertainty going from $\eta = 0$ [-] to $\eta = 1$ [-]. On the other hand, the models trained for the Nile river basin experience an increase in error going from $\eta = 2$ [-] to $\eta = 3$ [-]. This means that the hypothesis that increasing η always reduces error and uncertainty does not always hold for increasing. To investigate these unexpected events, special attention is paid to the case $\eta = 1$ [-] in Subsection 9.1.2 and in Section 9.2 with a focus on the Amazon river basin.

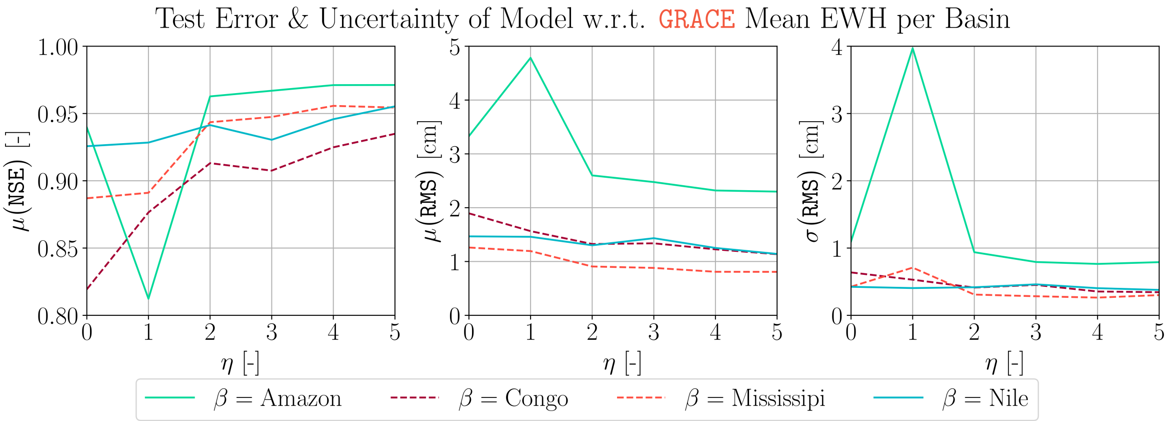


Figure 9.2: Mean testing errors NSE (left plot) and RMS (middle) w.r.t. GRACE for NN GRACE-like data per basin as a function of η. Standard deviation of testing RMS is shown in right plot.

9.1.2. Effect of Seed Numbers

To investigate the effect of epistemic and aleatoric uncertainties on the performance of the NN models, the relationship between the seed number and model uncertainty is calculated. Figure 9.3 contains the standard deviation of NSE test values for models trained with $\eta = 0$. These standard deviations represent uncertainty and are plotted for each basin (each row). Each figure on a given row has, on its axes, a different combination of these seed numbers: $\alpha_{\texttt{select}}$, $\alpha_{\texttt{weights}}$, and λ_x. Each pixel in each map represents the standard deviation of the models which are trained and tested with that specific pair of seed numbers. At the top of each column of figures, the amount of samples per pair, ρ_p, is indicated. As the values of the seed numbers themselves hold no meaning, they are withheld from the plot axes.

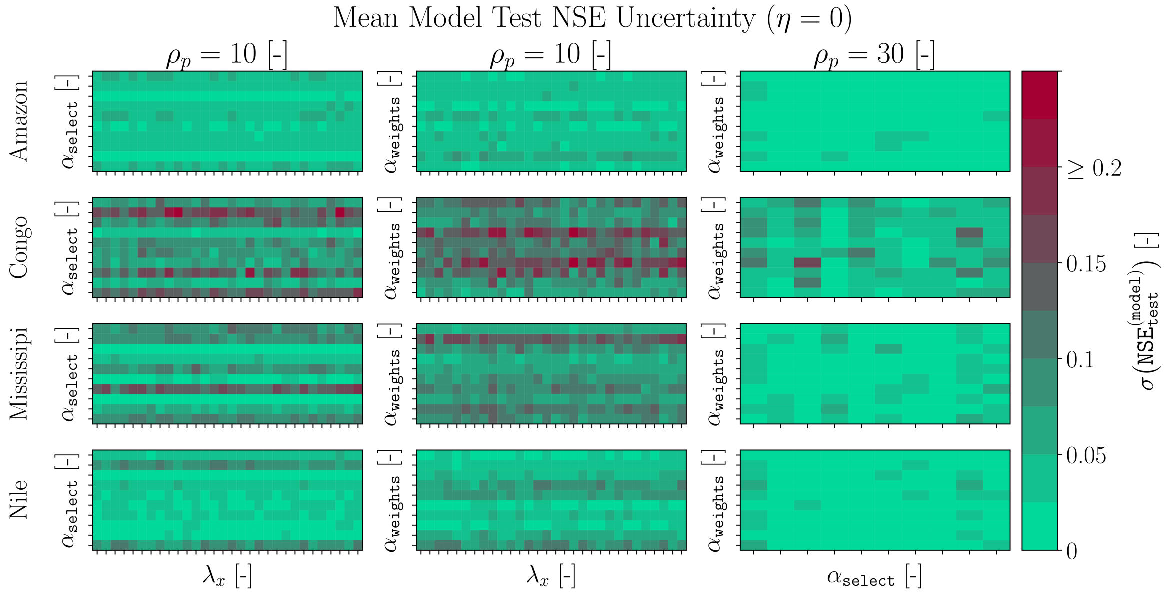


Figure 9.3: Standard deviation of model test NSE for models trained with $\eta = 0$ as a function of seed numbers, $\alpha_{\texttt{select}}$, $\alpha_{\texttt{weights}}$, and $\lambda_{x,\texttt{test}}$.

The first two columns in Figure 9.3 show significantly stronger variations in colour along the y-axes than along the x-axis for the Amazon, Mississippi, and Nile river basins. This shows that the seed numbers $\alpha_{\texttt{select}}$ and $\alpha_{\texttt{weights}}$ are stronger contributors to model uncertainty for these basins. The third column for these river basins, show most variation along the x-axis. This means that of the two epistemic uncertainty related seed numbers ($\alpha_{\texttt{select}}$ and $\alpha_{\texttt{weights}}$), $\alpha_{\texttt{select}}$ is a stronger predictor for model uncertainty for these river basins. This means that the final performance of these models is sensitive to which months are used for training and which months are not. It also means, that these models are less sensitive to the parameter initialization at the start of their training process.

For the Congo river basin, the plots in the first two columns still show more variation along the y-axes, but less distinctly. This is indicative of the aleatoric uncertainty (λ_x) playing a larger role for the Congo river basin than for the other three basins. In the third column and second row of Figure 9.3 there are a few sporadic pixels which are clearly darker than the other pixels. This means that there are Congo NN models trained with $\eta = 0$ and a specific combinations of $\alpha_{\texttt{select}}$ and $\alpha_{\texttt{weights}}$ that are more sensitive to input data noise than others.

In addition to using changes in colours in plots to quantify the significance of particular seed numbers, a statistical approach is taken in which F-statistics are used. These statistics can be used to determine whether particular variables have a stronger or weaker effect on a performance metric and if that effect is significant or not. In this process, the seed numbers are considered categorical because their order or value holds no significant meaning. To use categorical variables in F-statistics, target encoding is applied. Target encoding is applied wherein each category value (seed number and value) is replaced with the standard deviation of the performance metric observed for that seed number. In this case, the performance metric is the testing NSE. This approach was implemented using the Python library `scipy`'s `stats` module.

Figure 9.4 contains the computed F-statistic values, F (y-axes), for each seed number (represented by solid coloured lines) per basin (each subplot) as a function of η (x-axes). Only statistically significant values of F are included in the figure. It is also noted that for $\eta = 0$ it is impossible to have a value of F for the α_{x} seed number because there are no additionally sampled data sets used in training. A larger value of F indicates that a particular seed number explains more of the variance (uncertainty) in NSE test values. The F values for λ_{x} are the least strong explainers of the variability in the test NSE of the models indicated by the absence of red lines for the Mississippi and Amazon river basins. The red lines are only partially visible for the Congo and Nile river basins. This shows that the λ_{x} seed numbers are not a primary source of uncertainty for the NN's and means that the NN's are generally well equipped to deal with noise. The second lowest F values are attributed to the seed number α_{x} (green lines) which is responsible for generating noise on the additionally sampled training data. This indicates, again, that the selected NN architecture and hyper-parameters struggle less with aleatoric uncertainty sources.

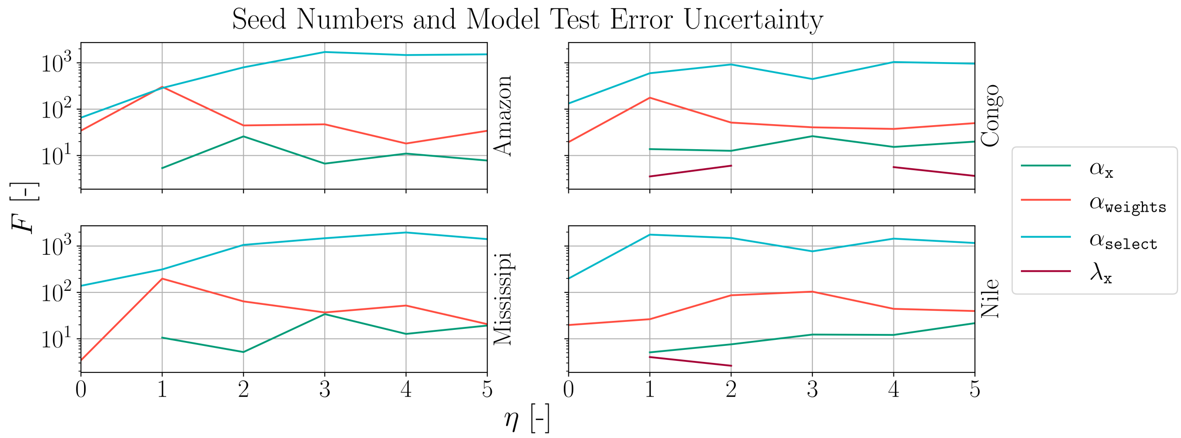


Figure 9.4: F-statistic F values (y-axes) for varying η (x-axes) per basin (each subplot) for target encoded seed numbers to explain variance in model test NSE.

In Figure 9.4, all combinations of basins and η (barring one combination) the blue line representing the F-statistic for the seed number $\alpha_{\texttt{select}}$ is largest which confirms that the selection of training and testing months is the most influential in determining the test NSE of a model. The combination of η and basin for which this does not hold is for $\eta = 1$ and the Amazon river basin. For this combination, the F value for $\alpha_{\texttt{weights}}$ (orange line) is just as large as the F value for $\alpha_{\texttt{select}}$. In Subsection 9.1.1 it is found that for this case, the hypothesis that increasing η always reduces uncertainty does not hold. Furthermore, in Subsection 9.1.1 an increase in uncertainty going from $\eta = 0$ to $\eta = 1$ for the Mississippi river basin is observed. In the lower left plot of Figure 9.4 corresponding to the Mississippi river basin, an increase in the F value of $\alpha_{\texttt{weights}}$ is observed as well. It is postulated that for several specific combinations of $\alpha_{\texttt{select}}$ and $\alpha_{\texttt{weights}}$ sub-optimal NN's are trained resulting in the observed increased uncertainty and increase in mean model errors. This means that the selection of training and testing data should perhaps not be randomized, or at least, be inspected before training a NN. Similarly, $\alpha_{\texttt{weights}}$ should be tested in combination with $\alpha_{\texttt{select}}$ prior to using a NN's results when gap-filling GRACE EWH.

9.1.3. Gap-filling GRACE Mean EWH

The models trained with $\eta = 5$ [-] appear to perform best in the previous section. In this section, the mean EWH predictions of the models trained with $\eta = 5$ [-] are analysed on a monthly basis in terms of their testing error and uncertainty. Figure 9.5 shows the mean EWH, $E^{(\mu,m)}_{\texttt{basin}}$ (on y-axis) per basin (each plot row) as observed by GRACE (dashed red line). It also shows the mean of all models' (with $\eta = 5$) test predictions (blue line) for each month (x-axis). The blue and red line are close together (or overlap) for all basins. This confirms the low observed error for all basins in Figure 9.2. Note that there are gaps in the models' predictions for particular months in which GRACE is available. These gaps occur because the 10 seed numbers for $\alpha_{\texttt{select}}$ lead to these months being selected as training months for all 10 seed numbers. Therefore, this figure which only uses test results, contains gaps. In future research it is recommended to carefully test the $\alpha_{\texttt{select}}$ seed numbers to ensure each month appears a certain number of times as a testing month.

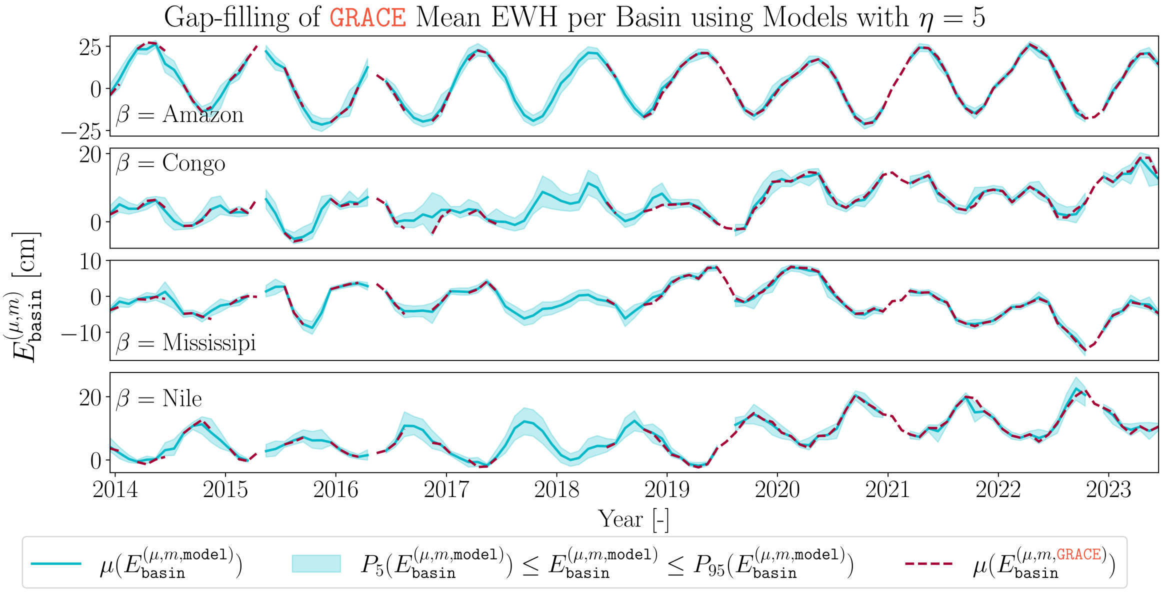


Figure 9.5: Mean GRACE EWH per basin per month up to 2019 along with mean of all models with $\eta = 5$ to fill gaps.

This figure also shows the range (shaded blue area) in which 90 [%] of the models' EWH predictions lie. This means that if a NN is trained to gap-fill GRACE mean EWH, then there is a significant probability that the model will predict GRACE EWH somewhere within the shaded blue area. The size of the shaded area indicates the uncertainty of the results for that month. If a particular month has a higher uncertainty it indicates that, for this month, the Swarm and soil moisture data are harder to relate to GRACE EWH data. There is no clear correlation between the uncertainty of months for which GRACE is available and the months for which GRACE is not available. Section 9.2 investigates what temporal events drive these uncertainties and errors.

9.2. Temporal Error

In this section, the performance of models on a monthly basis is investigated and is motivated by the opportunity to draw relationships between the errors and uncertainty of the models and the errors in the data used. In this section, the error metrics are not computed using mean EWH, but they are computed on a pixel by pixel basis. The focus lies on models trained with $\eta = 0$ [-], $\eta = 1$ [-] (in case of Amazon river basin), and $\eta = 5$ [-]. For each set of models, the mean of the models' test RMS values w.r.t. GRACE on a monthly basis is lower than Swarm's RMS w.r.t. GRACE. For this reason it is not included in Figure 9.6.

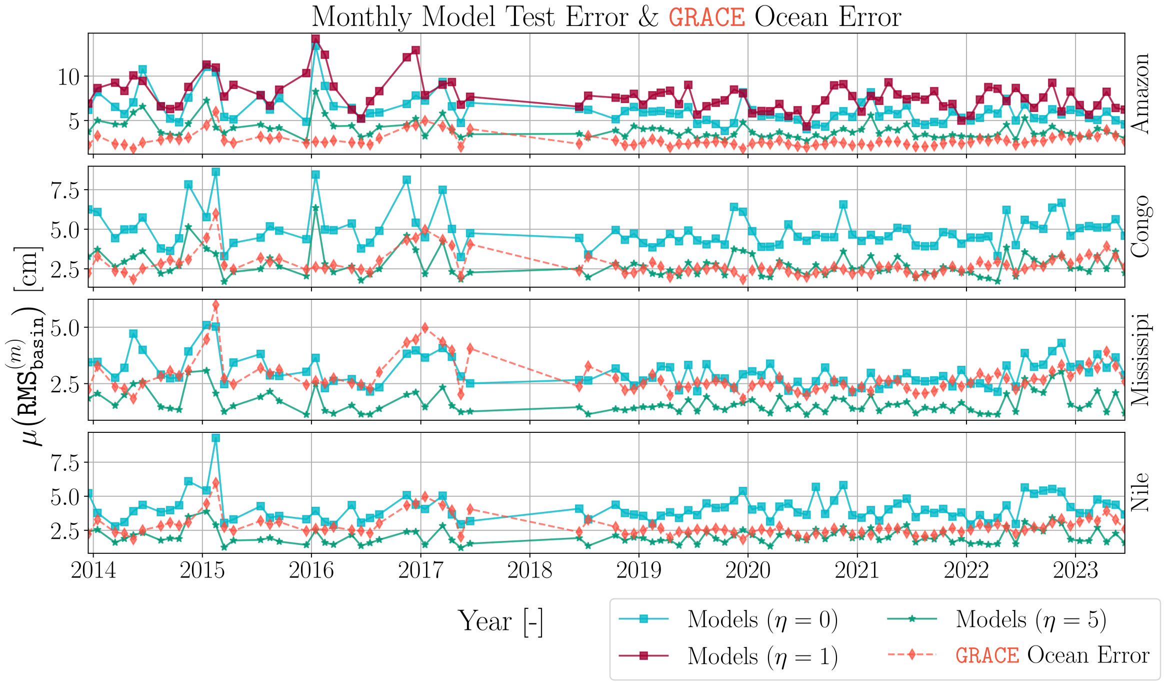


Figure 9.6: For each basin (each row), the mean of the models' RMS w.r.t. GRACE (y-axis) on a monthly basis (x-axis) for models trained with $\eta = 0$ (blue lines) and $\eta = 5$ (green lines). For the Amazon river basin the models with $\eta = 1$ (red lines) are included. GRACE's monthly ocean error is plotted on each row as well.

Figure 9.6 shows the mean of the models' test RMS values w.r.t. GRACE (y-axis) on a monthly basis (x-axis) for each basin (each row). It also contains GRACE's ocean error indication (green lines, quantified in Chapter 6). The models trained with $\eta = 5$ have a lower error than the models trained with $\eta = 0$ for all months (green line is always below blue line). A slight correlation between the errors in models with $\eta = 0$ and models with $\eta = 5$ is observed. For instance, the peaks at the 2016 and 2020 markers for the Congo river basin. This indicates, that if a model's error is high originally w.r.t. to other months, it is likely to still be relatively high after training with additionally sampled datasets.

Examining the figure, there is no consistent correlation between GRACE ocean errors and model errors for all months. In particular for most of the months where GRACE's error is relatively low. However, at the end of 2014 and beginning of 2015, and the period in 2017 leading up to the decommissioning of the original GRACE mission, the high GRACE errors do coincide with higher model errors. Models trained with $\eta = 0$ show more of an overlap than the models trained with $\eta = 5$. During the former period, the GRACE mission suffered from low ground track coverage due to being in a repeating orbit. The latter period has high errors due to one of the GRACE accelerometers not being functional. Therefore, there is an indication that during months where GRACE has high errors, it is more difficult for NN's to correlate Swarm and soil moisture data to noisy GRACE data. This case is not as strong for models trained with $\eta = 5$ which indicates that training with additionally sampled datasets can reduce sensitivity to these peaks in GRACE errors. Finally, it is also found that the models trained with $\eta = 5$ mostly show errors of the value or a lower value than the GRACE ocean errors. This indicates that for these months, the models are approaching GRACE-like resolution.

9.3. Spatial Error

A key feature of GRACE EWH data, is its spatial resolution. Therefore, it is important to investigate the spatial resolution of the trained models. Figure 9.7 shows the density distribution of NSE values w.r.t. GRACE per grid cell per basin for three cases: Swarm (orange line), models trained with $\eta = 0$ (green line), and models trained with $\eta = 5$ (blue line). The mean NSE value for each distribution is plotted using a vertical dashed line. If the dashed line is not visible, then the mean of that distribution is below 0. The mean of the spatial NSE values for the models with $\eta = 0$ are 0.413, 0.237, 0.310, and -0.803

for the Amazon, Congo, Mississippi, and Nile, respectively. These values are considerably lower than the values reported in Table 9.1. The key difference is that the GRACE EWH data is not averaged and hence more complex to predict. Furthermore, the models trained with $\eta = 5$ improve on both Swarm and the models trained with $\eta = 0$ with mean values of 0.68, 0.586, 0.587, and 0.370 for the Amazon, Congo, Mississippi, and Nile, respectively.

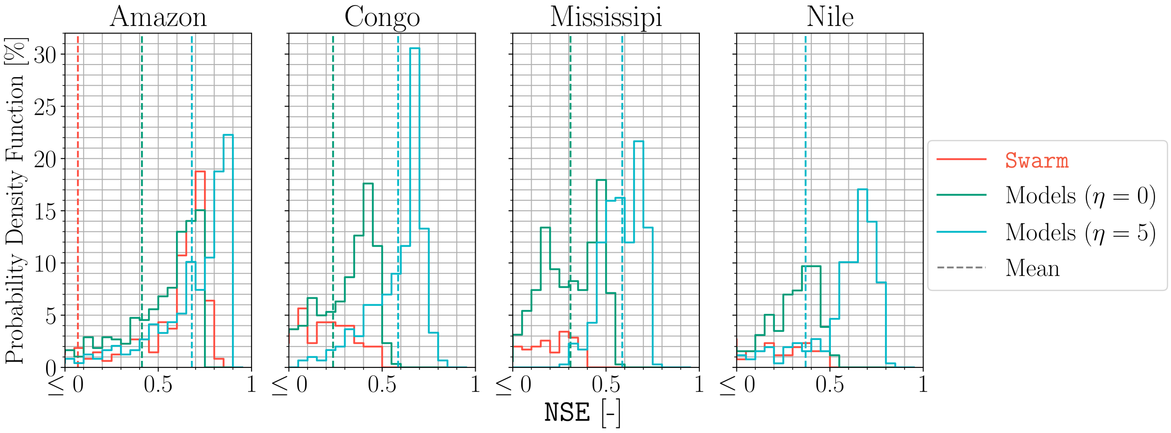


Figure 9.7: The probability density distribution for each river basin (each plot) of NSE values for Swarm (orange line), for models trained with $\eta = 0$ (green line) and for models trained with $\eta = 5$ (blue line).

Furthermore, when gap-filling the Amazon river basin using models trained with $\eta = 0$ it can be noted that the Swarm NSE distribution has a larger number of high NSE pixels than the models trained with $\eta = 0$. The mean of the models with $\eta = 0$ is still higher than Swarm's mean because the models trained with $\eta = 0$ improve primarily on Swarm's lower NSE values which is why the green line if above the orange line for lower NSE values. This indicates that using the models trained with $\eta = 0$ instead of Swarm to gap-fill the Amazon is a choice between having a larger or smaller discrepancy in spatial resolution across the basin.

Figure 9.8 plots the distribution of NSE values spatially. In its top row, the spatial NSE of Swarm w.r.t. GRACE is shown for each basin (left to right). The middle and bottom row contain the mean spatial NSE of models trained with $\eta = 0$ and $\eta = 5$ respectively. These maps reflect the observations made in Figure 9.7. For instance, for the Congo, Mississippi, and the Nile river basin, the models trained with $\eta = 0$ contain less red than each of the Swarm maps indicating a lower error just as observed in Figure 9.7. Over the Amazon river basin, the models trained with $\eta = 0$ only improve the NSE in the West as indicated by the reduced presence of red and there is a slight reduction in green over the North-East. This explains why the cumulative distributions functions of NSE for Swarm and models trained with $\eta = 0$ cross lines in Figure 9.7.

Furthermore, the models trained with $\eta = 5$ show more green and less red for all river basins in comparison to the models trained with $\eta = 0$. This confirms that on average the spatial errors for models trained with $\eta = 5$ are lower than those of models trained with $\eta = 0$. The maps in Figure 9.8 are a useful tool in examining where NN GRACE-like data have low errors. For instance, a researcher may want to use the gap-filled data spatially. If they have a requirement that the NSE values must be above 0.8, they may only want to rely on the gap-filling models for the Amazon river-basin and specifically focus on the North-East regions of the gap-filled data.

Comparing the mean spatial NSE of the models trained with $\eta = 0$ and $\eta = 5$ to the river basin variability computed in Section 5.2 (see Figure 5.4) shows similarities in the observed spatial patterns. The regions where higher EWH variability is observed are highlighted by blue rectangles. The higher test NSE values in the North (dashed rectangle) and South (solid rectangle) of the Congo river basin are also observed in the spatial variability patterns caused by the higher levels of precipitation in the North and South of this basin. Similarly, models perform distinctly better over the Southern half of the Nile (dashed rectangle) which also experience much more precipitation than its Northern counterpart. Furthermore, models trained with $\eta = 5$ clearly perform better over a small part of the North of the Nile (solid rectangle). In Section 5.2, it is postulated that the operations of the Aswan High Dam and other dams may be causing an increase in EWH variability in this region. Computing the Pearson CC between spatial standard deviation of de-trended GRACE EWH and the mean model test NSE values

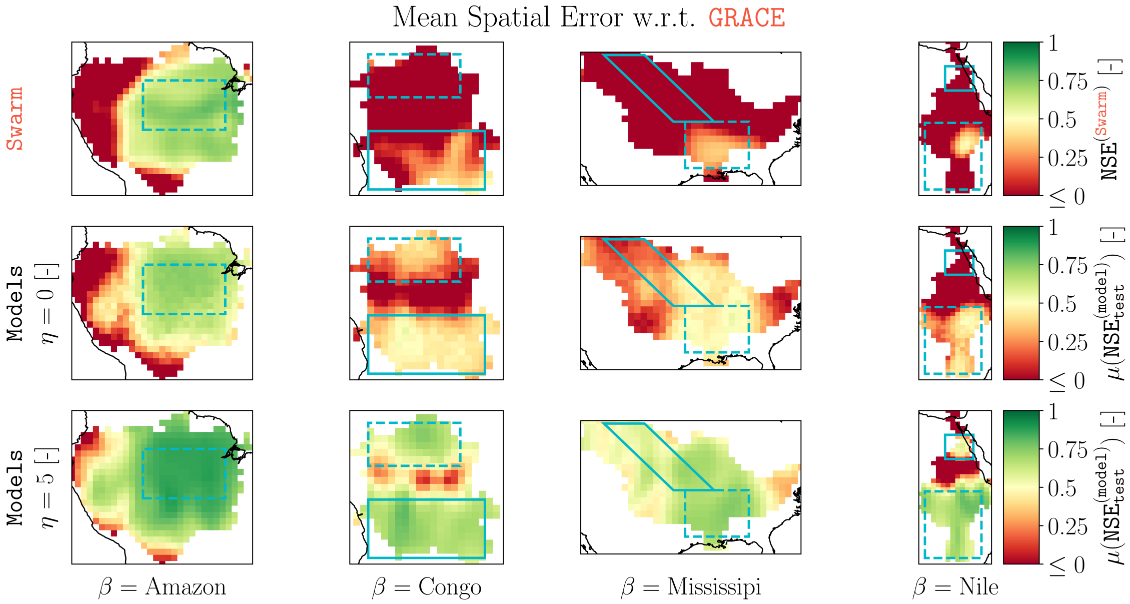

Mean Spatial Error w.r.t. GRACE
Swarm
Models
$\eta = 0$ [-]
Models
$\eta = 5$ [-]
β = Amazon
β = Congo
β = Mississipi
β = Nile
1
0.75
0.5
0.25
≤ 0
$\mathrm{NSE}^{(\mathtt{Swarm})}$ [-]
$\mu(\mathrm{NSE}_{\mathrm{test}}^{(\mathtt{model})})$ [-]

Figure 9.8: The NSE of Swarm (top row), the mean of model ($\eta = 0$) test NSE (middle row) and the mean of model ($\eta = 5$) test NSE (bottom row) spatially for each river basin. Dashed and solid rectangles are areas with high EWH variability discussed in Section 5.2.

for each basin results in values between 0.70 and 0.82 [-]. This correlation indicates that the spatial distribution of model (models trained with $\eta = 0$ and with $\eta = 5$) error is driven by EWH signal variability.

10 Conclusion

The first research question reads: what are the errors and uncertainty of GRACE-like data produced by NNs? When predicting GRACE mean EWH over a basin, the 400 generated models that do not utilize additional training data have mean test NSE values of 0.940, 0.819, 0.887, and 0.926 for the Amazon, Congo, Mississippi, and Nile river basins respectively. All these models are consistently and significantly more accurate than Swarm's NSE values for these basins. Therefore, when predicting the mean EWH over a basin during a gap period, it is better to use a trained NN than to use Swarm EWH data directly to fill the gaps.

The second stated research question is: how does the inclusion of additional training data generated, using errors in the auxiliary datasets, affect the errors and uncertainty of GRACE-like data produced by NNs? Models trained with $\eta = 5$ on average have NSE values of 0.970, 0.935, 0.954, and 0.955 for the Amazon, Congo, Mississippi, and Nile river basins respectively. This shows that sampling with additional training data leads to reductions in error for all basins w.r.t. models trained with $\eta = 0$. The increase in quality of the models is particularly significant for the Congo river basin.

Additionally, aleatoric uncertainty is found to play the least significant role in determining model errors and uncertainty. This means that NN's are well-equipped to deal with noisy input data. It is also shown that the strongest predictor for model uncertainty is of an epistemic nature and is the seed number related to selecting testing and training data. This is attributed to some months being more suitable for training rather than testing. Training with additional data is also shown to reduce this uncertainty and the uncertainty introduced by the other seed numbers. However, for some combinations of the aforementioned seed number and the seed number for NN parameter initialization, there is an increase in uncertainty. When using five additionally sampled training datasets this irregularity does not appear. This means that when sampling five additional training data, there is less concern for epistemic and aleatoric uncertainty. For instance, the optimization of hyper-parameters in Chapter 7 is plagued by epistemic uncertainty. This epistemic uncertainty can be reduced by training models with additionally sampled training data and therefore, result in a more optimal hyper-parameter selection.

In the temporal and spatial sense it is also found that models trained with $\eta = 5$ outperform the models trained with $\eta = 0$ both in terms of error and uncertainty. For some combinations of lower η values and basin, the spatial and temporal analyses reflect the negative effect on the error and uncertainty of particular combinations of seed numbers used for selecting training and testing data and seed numbers used for parameter initialization. This leads to the conclusion that for lower values of η emphasis has to be placed on proper selection of seed numbers. When training with five additionally sampled training datasets, less caution has to be taken in the selection of seed numbers. The reason why lower values of η (such as $\eta = 1$ for the Amazon river basin) have a higher chance of being sensitive to the $\alpha_{\texttt{weights}}$ seed number in contrast to models with $\eta = 0$ and $\eta = 5$ is not discovered. In the temporal analysis it is also found that the models trained with $\eta = 5$ exhibit error indications of the same size or lower for all months w.r.t. GRACE's ocean errors. From this it can be concluded that in the temporal sense, the models trained with $\eta = 5$ have achieved GRACE-like resolution.

Furthermore, using the NN models trained with $\eta = 5$ to predict GRACE-like EWH spatially is recommended only over local subsets of the river basins as indicated in the bottom row of Figure 9.8 where strong spatial variations in mean error are visible. Depending on the requirement on the local NSE value, a researcher may or may not decide to use GRACE-like EWH spatially. It is also found that spatial signal variability (indication of amplitude) over a basin strongly correlates with a NN's spatial error. Regions with active rainfall or points where rivers join have a higher NN gap-filling performance. This means that predicting smaller signals in, for example, arid regions such as the Northern half of the Nile river basin, is more difficult for a NN even when trained with additionally sampled data. It is strongly advised to examine the mean NSE maps in Figure 9.8 to determine whether NN spatially gap-filled data should be used or not.

11 Recommendations

Recommendations for future research in the field of error and uncertainty quantification of GRACE gap-filling using Swarm EWH and GLDAS soil moisture using machine learning are split into three categories.

The first category of recommendations is with regard to the type of NN used for gap-filling. The motivation for not selecting a CNN in Chapter 4 is that CNN's are a lot more computationally expensive than fully-connected NNs (Harrison, 2023). However, there is no mention of this in the other study that gap-fills GRACE using CNNs (Keleş, 2022). CNNs can be more efficient if they are made small enough (Goodfellow et al., 2016). It is recommended that this experiment is performed for gap-filling with CNNs with a focus on finding a more optimal CNN architecture. It is also recommended that BCNNs are investigated as they can provide uncertainty estimates on their output data (Kwon et al., 2020; Shridhar et al., 2018).

The second category covers the error quantification of the datasets used. For the quantification of error in Swarm land agreement with GRACE is used. However, the agreement of Swarm with GRACE over different basins may have different optimal smoothing radii for Swarm. This may result in better Swarm agreement with GRACE on the basin level and hence better results for the NN gap-filling. It is recommended to investigate whether a temporal component of error in the soil moisture data can be quantified. This might result in more accurate noise being sampled for this dataset.

The final category covers the use of noise to generate additional training data. Sampling additional training data is done only by generating noise on the auxiliary data. This shows promising results in reducing both the errors and uncertainty in the NN gap-filling models. The NN models trained with five additionally sampled datasets show reduced errors and uncertainty. It is further postulated that adding noise to the target data in the additionally sampled data sets will further reduce the errors and uncertainty in the models. This is because the NN not only receives information about the errors in the auxiliary data, but also in the target data (GRACE EWH data). Therefore, it is recommended to investigate if adding noise to the GRACE EWH target data, when sampling additional training data reduces the errors and uncertainty further. Furthermore, to fully verify the extent to which sampling additional training data with noise reduces uncertainty and errors, it is recommended that a scaling factor is added to the noise such that the level of noise and its effect on uncertainty and errors can be investigated. Finally, it is recommended to investigate why lower values of η such as $\eta = 1$ in combination with the Amazon river basin is more susceptible to a larger uncertainty contribution due to the $\alpha_{\texttt{weights}}$ seed number.

A

Reflection on Planning

After the literature review a list of tasks with an associated planning was created to answer the research questions. The tasks tabled below outline an ambitious set of tasks to meet the originally set goals. These goals involved comparing the performance of fully-connected NNs and BCNNs. The tasks were divided into six categories:

- O , *Planning.* This category covers tasks that were intended for planning and preparing work ahead of time.
- A , *Auxiliary Data.* This category covers all tasks related to external sources of data and the effort required to obtain or process that external data. This category also includes input data error quantification tasks.
- B , *NN Models.* Tasks in this category cover the creation, verification, validation, and analysis of all NNs.
- D , *Uncertainty Quantification.* This category contains tasks related to quantifying the uncertainty in the created NNs.
- R , *Reporting.* This category covers tasks associated with progress reporting on the generation of the thesis and all its components.
- P , *Presenting.* This category covers all tasks associated with generating and practising a presentation around the thesis.

Table A.1 outlines the tasks devised for each category and their assigned task identities.

A few weeks after the literature review the random number seed related uncertainties were discovered when performing task B.2 . This lead to a change of scope for the thesis from attempting to analyse the uncertainty of a single NN model to investigating the uncertainty across many models. The amount of time this would take, lead to the scrapping of the creation of a BCNN entirely. Additionally, this ensemble based method of generating many models to estimate the gap-filling uncertainty led to the tasks in task set D being removed from the planning as well.

The tasks in task sets: O , P , and R have all been completed. This can be attributed to their generic nature. These tasks are common tasks to complete for any thesis. Additionally, the tasks in task set A have been completed. The word 'uncertainty' is used originally where later this was to clearly distinguish uncertainty and error. This shows that even after the literature review leading up to the final green-light submission, substantial conceptual re-framing had to be done.

Looking back at all the tasks, the original goals of this thesis were rather ambitious. New discoveries in the areas of methodology, NN training and analysis lead to the planning becoming more short-term and frequently changing. Realizing the need to include additional tasks to ensure the thesis to remain truly meaningful, was a cause for some frustration, as it always felt like planning could not be performed far ahead because the plans were always changing. One of the main lessons learned to address this problem was to take a step back every once in a while from the in-depth work back to high-level conceptual thinking. This significantly helped to frame the tasks in relation to the research goals. The main takeaway and message to future self: take even *more* steps back and look to the topic from different angles as it will allow to yield a more accurate planning and possibly an even more robust result.

Table A.1: Overview of tasks and their descriptions.

Category	Task Identity	Task Description
0	0.1	Create list of model result plots that will be created.
0	0.2	Devise model verification method (synthetic experiment).
0	0.3	Devise sensitivity analysis method.
0	0.4	Devise model (pseudo-)validation method.
0	0.5	Classify basins based on cycles and observed trends.
A	A.1	Create Swarm EWH maps.
A	A.2	Create Swarm EWH uncertainty maps.
A	A.3	Acquire GLDAS soil moisture maps.
A	A.4	Acquire GLDAS soil moisture uncertainty maps.
A	A.5	Acquire other gap-filling study results.
A	B.1	Create baseline fully-connected NN model (M0).
B	B.2	Verify baseline M0.
B	B.3	Plot M0 results.
B	B.4	Perform sensitivity analysis on M0 and create M1.
B	B.5	Create baseline BCNN model (M2).
B	B.6	Verify M2.
B	B.7	Plot M2 results.
B	B.8	Perform sensitivity analysis on BCNN model and create M3.
B	B.9	Validate all models.
D	D.1	Propagate distributions analytically.
D	D.2	Propagate statistical moments analytically.
D	D.3	Propagate n samples.
D	D.4	Compare uncertainties to inherent UQ of M1.
R	R.0	Handle feedback.
R	R.1	Work on writing mid-term version of thesis.
R	R.2	Submit mid-term version of thesis.
R	R.3	Work on writing greenlight version of thesis.
R	R.4	Submit greenlight version of thesis.
P	P.1	Create greenlight review presentation.
P	P.2	Practice greenlight review presentation.
P	P.3	Create defence presentation.
P	P.4	Practice defence presentation.

B

Available & Missing GRACE Months

2002/04	2002/05	🟧 2002/06	🟧 2002/07	2002/08	2002/09	2002/10	2002/11
2002/12	2003/01	2003/02	2003/03	2003/04	2003/05	🟧 2003/06	2003/07
2003/08	2003/09	2003/10	2003/11	2003/12	2004/01	2004/02	2004/03
2004/04	2004/05	2004/06	2004/07	2004/08	2004/09	2004/10	2004/11
2004/12	2005/01	2005/02	2005/03	2005/04	2005/05	2005/06	2005/07
2005/08	2005/09	2005/10	2005/11	2005/12	2006/01	2006/02	2006/03
2006/04	2006/05	2006/06	2006/07	2006/08	2006/09	2006/10	2006/11
2006/12	2007/01	2007/02	2007/03	2007/04	2007/05	2007/06	2007/07
2007/08	2007/09	2007/10	2007/11	2007/12	2008/01	2008/02	2008/03
2008/04	2008/05	2008/06	2008/07	2008/08	2008/09	2008/10	2008/11
2008/12	2009/01	2009/02	2009/03	2009/04	2009/05	2009/06	2009/07
2009/08	2009/09	2009/10	2009/11	2009/12	2010/01	2010/02	2010/03
2010/04	2010/05	2010/06	2010/07	2010/08	2010/09	2010/10	2010/11
2010/12	🟧 2011/01	2011/02	2011/03	2011/04	2011/05	🟧 2011/06	2011/07
2011/08	2011/09	2011/10	🟧 2011/11	2011/12	2012/01	2012/02	2012/03
2012/04	🟧 2012/05	2012/06	2012/07	2012/08	2012/09	🟧 2012/10	2012/11
2012/12	2013/01	2013/02	🟧 2013/03	2013/04	2013/05	2013/06	2013/07
🟧 2013/08	🟧 2013/09	2013/10	2013/11	2013/12	2014/01	🟧 2014/02	2014/03
2014/04	2014/05	2014/06	🟧 2014/07	2014/08	2014/09	2014/10	2014/11
🟧 2014/12	2015/01	2015/02	2015/03	2015/04	🟧 2015/05	🟧 2015/06	2015/07
2015/08	2015/09	🟧 2015/10	🟧 2015/11	2015/12	2016/01	2016/02	2016/03
🟧 2016/04	2016/05	2016/06	2016/07	2016/08	🟧 2016/09	🟧 2016/10	2016/11
2016/12	2017/01	🟧 2017/02	2017/03	2017/04	2017/05	2017/06	🟧 2017/07
🟧 2017/08	🟧 2017/09	🟧 2017/10	🟧 2017/11	🟧 2017/12	🟧 2018/01	🟧 2018/02	🟧 2018/03
🟧 2018/04	🟧 2018/05	2018/06	2018/07	🟧 2018/08	🟧 2018/09	2018/10	2018/11
2018/12	2019/01	2019/02	2019/03	2019/04	2019/05	2019/06	2019/07
2019/08	2019/09	2019/10	2019/11	2019/12	2020/01	2020/02	2020/03
2020/04	2020/05	2020/06	2020/07	2020/08	2020/09	2020/10	2020/11
2020/12	2021/01	2021/02	2021/03	2021/04	2021/05	2021/06	2021/07
2021/08	2021/09	2021/10	2021/11	2021/12	2022/01	2022/02	2022/03
2022/04	2022/05	2022/06	2022/07	2022/08	2022/09	2022/10	2022/11
2022/12	2023/01	2023/02	2023/03	2023/04	2023/05	2023/06	2023/07
2023/08	2023/09	2023/10	2023/11	2023/12	2024/01	2024/02	2024/03
2024/04	2024/05						

Table B.1: This table contains the list of available (shaded green) and missing (shaded orange) GRACE months denoted in format: YYYY/MM.

Bibliography

Abd-Elbaky, M., & Jin, S. (2019). Hydrological mass variations in the nile river basin from grace and hydrological models. *Geodesy and Geodynamics*, *10*(6), 430–438. https://doi.org/https://doi.org/10.1016/j.geog.2019.07.004

Agarwal, V., Akyilmaz, O., Shum, C., Feng, W., Yang, T.-Y., Forootan, E., Syed, T. H., Haritashya, U. K., & Uz, M. (2023). Machine learning based downscaling of grace-estimated groundwater in central valley, california. *Science of The Total Environment*, *865*, 161138. https://doi.org/https://doi.org/10.1016/j.scitotenv.2022.161138

Ali, S., Ran, J., Khorrami, B., Wu, H., Tariq, A., Jehanzaib, M., Khan, M. M., & Faisal, M. (2024). Downscaled grace/grace-fo observations for spatial and temporal monitoring of groundwater storage variations at the local scale using machine learning. *Groundwater for Sustainable Development*, *25*, 101100. https://doi.org/https://doi.org/10.1016/j.gsd.2024.101100

Cheng, M., & Ries, J. C. (2017). The unexpected signal in grace estimates of $C_{2,0}$. *Journal of Geodesy*, *91*, 897–914. https://doi.org/10.1007/s00190-016-0995-5

Dahle, C. (2018). *Release notes for gfz grace level-2 products - version rl06* (tech. rep.). GFZ German Research Centre for Geosciences.

Dahle, C., Flechtner, F., Murböck, M., Michalak, G., Neumayer, K.-H., Abrykosov, O., Reinhold, A., & König, R. (2019). *GRACE-FO D-103919 (Gravity Recovery and Climate Experiment Follow-On): GFZ Level-2 Processing Standards Document for Level-2 Product Release 06* (Scientific Technical Report STR - Data No. 19/09) (Rev. 1.0, June 3, 2019). GFZ German Research Centre for Geosciences. Potsdam.

Encarnação, J., & Visser, P. (2024, December). *Multi-approach gravity field models from swarm gps data: Signal and error in the swarm models up to 2024-09-30* (tech. rep.) (Prepared and checked by João Encarnação; approved by Pieter Visser). Delft University of Technology (TU Delft), Astronomical Institute of the University of Bern (AIUB), Astronomical Institute Ondrejov (ASU), Institute of Geodesy Graz (IfG), and Ohio State University (OSU).

Forootan, E., Schumacher, M., Mehrnegar, N., Bezděk, A., Talpe, M. J., Farzaneh, S., Zhang, C., Zhang, Y., & Shum, C. K. (2020). An iterative ica-based reconstruction method to produce consistent time-variable total water storage fields using grace and swarm satellite data. *Remote Sensing*, *12*(10). https://doi.org/10.3390/rs12101639

Foroumandi, E., Nourani, V., Jeanne Huang, J., & Moradkhani, H. (2023). Drought monitoring by downscaling grace-derived terrestrial water storage anomalies: A deep learning approach. *Journal of Hydrology*, *616*, 128838. https://doi.org/https://doi.org/10.1016/j.jhydrol.2022.128838

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning* [http://www.deeplearningbook.org]. MIT Press.

Gu, Y., Huang, F., Huang, J., Yuan, H., Yu, B., & Gao, C. (2023). Filling the gap between GRACE and GRACE follow-on observations based on principal component analysis. *Geophysical Journal International*, *236*(3), 1216–1233. https://doi.org/10.1093/gji/ggad484

Harrison, B. (2023). *Bridging grace/grace-fo gap by using machine learning to increase swarm spatial resolution* [Master's thesis, Delft University of Technology].

Humphrey, V., & Gudmundsson, L. (2019). Grace-rec: A reconstruction of climate-driven water storage changes over the last century. *Earth System Science Data*, *11*(3), 1153–1170. https://doi.org/10.5194/essd-11-1153-2019

Jäggi, A., Meyer, U., Lasser, M., Jenny, B., Lopez, T., Flechtner, F., Dahle, C., Förste, C., Mayer-Gürr, T., Kvas, A., Lemoine, J.-M., Bourgogne, S., Weigelt, M., & Groh, A. (2022). International combination service for time-variable gravity fields (cost-g). In J. T. Freymueller & L. Sánchez (Eds.), *Beyond 100: The next century in geodesy* (pp. 57–65). Springer International Publishing.

Keleş, M. (2022). *Filling the data gap between grace and grace follow-on missions using deep learning algorithms* [Master's thesis, Istanbul Technical University].

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, 1097–1105.

Kwon, Y., Won, J.-H., Kim, B. J., & Paik, M. C. (2020). Uncertainty quantification using bayesian neural networks in classification: Application to biomedical image segmentation. *Computational Statistics & Data Analysis*, *142*, 106816. https://doi.org/https://doi.org/10.1016/j.csda.2019.106816

Lehner, B., & Grill, G. (2013). Global river hydrography and network routing: Baseline data and new approaches to study the world's large river systems. *Hydrological Processes*, *27*(15), 2171–2186. https://doi.org/https://doi.org/10.1002/hyp.9740

Lepcha, D. C., Goyal, B., Dogra, A., & Goyal, V. (2023). Image super-resolution: A comprehensive review, recent trends, challenges and applications. *Information Fusion*, *91*, 230–260. https://doi.org/https://doi.org/10.1016/j.inffus.2022.10.007

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 [e2021GL093492 2021GL093492]. *Geophysical Research Letters*, *48*(8), e2021GL093492. https://doi.org/https://doi.org/10.1029/2021GL093492

Mäkinen, J. (2021). The permanent tide and the international height reference frame ihrf. *Journal of Geodesy*, *95*(9). https://doi.org/10.1007/s00190-021-01541-5

Mascha, E., & Vetter, T. (2018). Significance, errors, power, and sample size: The blocking and tackling of statistics. *Anesthesia and analgesia*, *126*, 691–698. https://doi.org/10.1213/ANE.0000000000002741

McCarthy, D. D., & Petit, G. (Eds.). (2004). *Iers conventions (2003)* [Paperback]. Verlag des Bundesamts für Kartographie und Geodäsie.

McGirr, R., Tregoning, P., Allgeyer, S., McQueen, H., & Purcell, A. P. (2023). Interplay of altitude, ground track coverage, noise, and regularization in the spatial resolution of grace gravity field models. *Journal of Geophysical Research: Solid Earth*, *128*(1), e2022JB024330. https://doi.org/https://doi.org/10.1029/2022JB024330

Mo, S., Zhong, Y., Forootan, E., Mehrnegar, N., Yin, X., Wu, J., Feng, W., & Shi, X. (2022). Bayesian convolutional neural networks for predicting the terrestrial water storage anomalies during grace and grace-fo gap. *Journal of Hydrology*, *604*, 127244. https://doi.org/https://doi.org/10.1016/j.jhydrol.2021.127244

Montesinos-López, O., Montesinos, A., & Crossa, J. (2022, January). Convolutional neural networks. https://doi.org/10.1007/978-3-030-89010-0_13

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part i - a discussion of principles. *Journal of Hydrology*, *10*. https://doi.org/10.1016/0022-1694(70)90255-6

Ndehedehe, C. E., & Agutu, N. O. (2022). Historical changes in rainfall patterns over the congo basin and impacts on runoff (1903–2010). In *Congo basin hydrology, climate, and biogeochemistry* (pp. 145–163). American Geophysical Union (AGU). https://doi.org/https://doi.org/10.1002/9781119657002.ch9

Ries, J., & Bettadpur, S. (Eds.). (2003, October). *Grace science team meeting proceedings 2003* [Held October 8–10, 2003].

Ries, J., Bettadpur, S., Eanes, R., Kang, Z., Ko, U., McCullough, C., Nagel, P., Pie, N., Poole, S., Richter, T., Save, H., & Tapley, B. (2016). *Development and evaluation of the global gravity model ggm05* (CSR Report No. CSR-16-02). Center for Space Research, The University of Texas at Austin.

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

Save, H. (2019). *GRACE Follow-On: Gravity Recovery and Climate Experiment Follow-On, CSR Level-2 Processing Standards Document for Level-2 Product Release 06* (CSR GRFO-19-01 (GRACE-FO D-103920)). Center for Space Research, The University of Texas at Austin. Austin, TX.

Shen, Y., Peng, F., & Li, B. (2015). Improved singular spectrum analysis for time series with missing data. *Nonlinear Processes in Geophysics*, *22*(4), 371–376. https://doi.org/10.5194/npg-22-371-2015

Shridhar, K., Laumann, F., & Liwicki, M. (2018, December). *A comprehensive guide to bayesian convolutional neural network with variational inference* [Doctoral dissertation]. https://doi.org/10.13140/RG.2.2.21142.09287

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition.

Sun, Z., Long, D., Yang, W., Li, X., & Pan, Y. (2020). Reconstruction of grace data on changes in total water storage over the global land surface and 60 basins [e2019WR026250 2019WR026250]. *Water Resources Research*, *56*(4), e2019WR026250. https://doi.org/https://doi.org/10.1029/2019WR026250

Teixeira da Encarnação, J., Visser, P., Arnold, D., Bezdek, A., Doornbos, E., Ellmer, M., Guo, J., van den IJssel, J., Iorfida, E., Jäggi, A., Klokocník, J., Krauss, S., Mao, X., Mayer-Gürr, T., Meyer, U., Sebera, J., Shum, C. K., Zhang, C., Zhang, Y., & Dahle, C. (2020). Description of the multi-approach gravity field models from swarm gps data. *Earth System Science Data*, *12*(2), 1385–1417. https://doi.org/10.5194/essd-12-1385-2020

Teixeira da Encarnação, J., Arnold, D., Bezdek, A., Dahle, C., Doornbos, E., van den IJssel, J., Jäggi, A., Mayer-Gürr, T., Sebera, J., Visser, P., & Zehentner, N. (2016). Gravity field models derived from swarm gps data. *Earth, Planets and Space*, *68*. https://doi.org/10.1186/s40623-016-0499-9

Tourian, M. J., Reager, J. T., & Sneeuw, N. (2018). The total drainable water storage of the amazon river basin: A first estimate using grace. *Water Resources Research*, *54*(5), 3290–3312. https://doi.org/https://doi.org/10.1029/2017WR021674

Wahr, J., Molenaar, M., & Bryan, F. (1998). Time variability of the earth's gravity field: Hydrological and oceanic effects and their possible detection using grace. *Journal of Geophysical Research: Solid Earth*, *103*(B12), 30205–30229. https://doi.org/https://doi.org/10.1029/98JB02844

Wang, F., Shen, Y., Chen, Q., & Wang, W. (2021). Bridging the gap between grace and grace follow-on monthly gravity field solutions using improved multichannel singular spectrum analysis. *Journal of Hydrology*, *594*, 125972. https://doi.org/https://doi.org/10.1016/j.jhydrol.2021.125972

Wang, Y., Li, C., Cui, Y., Cui, Y., Xu, Y., Hora, T., Zaveri, E., Rodella, A.-S., Bai, L., & Long, D. (2024). Spatial downscaling of grace-derived groundwater storage changes across diverse climates and human interventions with random forests. *Journal of Hydrology*, *640*, 131708. https://doi.org/https://doi.org/10.1016/j.jhydrol.2024.131708

Wu, Y., Miao, C., Fan, X., Gou, J., Zhang, Q., & Zheng, H. (2022). Quantifying the uncertainty sources of future climate projections and narrowing uncertainties with bias correction techniques [e2022EF002963 2022EF002963]. *Earth's Future*, *10*(11), e2022EF002963. https://doi.org/https://doi.org/10.1029/2022EF002963

Yi, S., & Sneeuw, N. (2021). Filling the data gaps within grace missions using singular spectrum analysis [e2020JB021227 2020JB021227]. *Journal of Geophysical Research: Solid Earth*, *126*(5), e2020JB021227. https://doi.org/https://doi.org/10.1029/2020JB021227

Yuan, D.-N. (2019). *GRACE Follow-On: Gravity Recovery and Climate Experiment Follow-On, JPL Level-2 Processing Standards Document for Level-2 Product Release 06* (JPL D-103921). Jet Propulsion Laboratory, California Institute of Technology. Pasadena, CA.

Zhang, C., Shum, C. K., Bezděk, A., Bevis, M., de Teixeira da Encarnação, J., Tapley, B. D., Zhang, Y., Su, X., & Shen, Q. (2021). Rapid mass loss in west antarctica revealed by swarm gravimetry in the absence of grace [e2021GL095141 2021GL095141]. *Geophysical Research Letters*, *48*(23), e2021GL095141. https://doi.org/https://doi.org/10.1029/2021GL095141