Bridging GRACE/GRACE-FO Gap by Using Machine Learning to Increase Swarm Spatial Resolution

MSc thesis Space Exploration

Benjamin Harrison

Delft University of Technology

TU Delft

Bridging GRACE/GRACE-FO Gap by Using Machine Learning to Increase Swarm Spatial Resolution

by

Benjamin Harrison

Student Name	Student Number
Benjamin Harrison	4556828

Supervisor:	J.G. (João) De Teixeira da Encarnação
Thesis committee:	Dr. Ir. W. (Wouter) van der Wal
	Dr. R.P. (Richard) Dwight
Institution:	Delft University of Technology
Place:	Faculty of Aerospace Engineering, Delft

Cover Image: GRACE Mission (Source: NASA)

TU Delft

Preface

As I sit down to write this preface, I am filled with a mix of emotions. On the one hand, I am proud to have reached this point in my academic journey, having completed my Master's in Aerospace Engineering at the Delft University of Technology. On the other hand, I cannot help but feel a sense of nostalgia as I reflect on the incredible experiences and memories that I have made during my time here. Looking back, I realize that my time in Delft has not only been an academic journey but also a personal one, during which I have grown in ways I never imagined.

I owe a great deal of gratitude to my teachers, friends, and family, whose support has played a crucial role in the success of my education. In particular, I would like to thank my mom, who has been my unwavering support system throughout my life. She has always done everything in her power to provide me with every opportunity to grow and pursue my dreams, and I cannot thank her enough.

Finally, I would like to express my sincere gratitude to my thesis supervisor, João De Teixeira da Encarnação for the guidance and support with which he provided me with over the past year. His expertise and mentorship have been invaluable to me, and I could not have completed this thesis without his help.

Benjamin Harrison
Delft, April 2023

Executive Summary

The main scientific goal of the National Aeronautics and Space Administration (NASA) and German Aerospace Center (DLR) GRACE and GRACE Follow-On missions is to accurately monitor the Earth's gravitational field and its temporal variations. Their gravity field solutions and the resultant terrestrial water storage changes (TWSC) are helpful in improving our understanding of oceanography, calculating ice mass loss, tracking hydrology across terrestrial areas, and modelling climate change. After obtaining its final gravity field solution in May 2017, GRACE's successor, the GRACE Follow-On (GRACE-FO) mission, began its measurements in May 2018. This resulted in an 11-month gap between GRACE and GRACE-FO data. Previous studies have aimed at filling this data gap. One strategy for doing so is by deriving gravity field models from the Swarm satellite data. Mapping the geomagnetic and electric fields in the atmosphere was the main goal of the European Space Agency's (ESA) Swarm Earth Explorer mission, launched in 2013. However, according to a number of studies, GPS data from low-Earth orbit satellites like Swarm can be used to detect long-wavelength temporal changes in the Earth's gravitational field. This data has been used in research to bridge the GRACE/GRACE-FO data gap. While having a far lower spatial resolution than GRACE and GRACE-FO (only 1500 km as opposed to 350 km), Swarm is an independent source on TWSC. Another gap-filling strategy that has been explored in previous studies is the use of data-driven techniques such as machine learning. These algorithms use hydro-climatological maps such as precipitation and soil moisture as input, and aim at reproducing GRACE-like TWSC estimates during the gap. These maps result from land surface models such as the Global Land Data Assimilation System (GLDAS) NOAH model. The main benefit of using these hydro-climatological maps over Swarm data is that they offer data with a high level of temporal and geographical precision. The disadvantage of these observations is that they do not instantly reveal information about the Earth's gravitational field. Instead, they provide information on factors that could affect the local gravity signal, such as the movement of water over the surface. The goal of this thesis was to combine the notion of utilizing hydro-climatological data combined with Swarm to close the GRACE/GRACE-FO data gap. These two categories of data were merged as inputs to a neural network and assessed as predictors of GRACE-like TWSC. In order to enhance the effectiveness of each technique independently, data describing the changes in Earth's gravitational field was combined with high-resolution hydrological data. The Amazon Basin, which is a hydrologically active region, was selected as the study region.

The data used in this thesis consisted of GRACE gravity field models, Swarm-derived gravity field models, and hydro-climatological data. The GRACE data was obtained from the GRACE release 6 (RL06) produced at the Center for Space Research (CSR) of the University of Texas at Austin and consisted of 1°x 1°gridded maps of Equivalent Water Height (EWH). The Swarm-derived EWH maps were obtained from Swarm gravity field models that were created by combining four distinct solutions that were computed using various gravity field estimating techniques. They were synthesized to 1°x 1°gridded maps such that they were in agreement with GRACE. The hydro-climatological variables that were considered as input to the neural networks were precipitation, temperature, and soil moisture. This selection resulted from the literature study, in which they were proven to have a relationship with the gravity signals measured by GRACE. After analyzing the neural network performance with all possible combinations of the considered input variables, it was concluded that soil moisture was the best predictor. Therefore, soil moisture was used as the sole hydro-climatological input variable in the final simulations. It was also concluded that only using good-quality Swarm months for the training of the neural networks increased the performance of the model. This resulted in only 57 months of available data for training and testing.

Two different neural network (NN) architectures were designed for this thesis. The first one consisted of a fully connected Multilayer Perceptron (MLP) neural network with two hidden layers. The second neural network was a convolutional neural network (CNN) with four convolutional layers, followed by three fully connected layers. The two neural networks were each optimized in terms of architecture

and hyperparameters. For the first MLP model, it was found that the best performance was obtained when the model was trained using one month of input data at a time, as opposed to stacking multiple months. It was also concluded that normalizing the input data resulted in overfitting by the neural network. Lastly, the optimal hyperparameters of the model were obtained. The optimization of the CNN showed that a 3D CNN slightly outperformed a 2D CNN. As for the MLP model, stacking or normalizing the input data was not beneficial for the performance of the CNN. Once the two NNs were optimized, their performance in both the spatial and the temporal domain was compared. It was concluded that the MLP slightly outperformed the CNN while having a much lower training time. Therefore, it was decided to continue with the ANN for the rest of the thesis.

Analyzing the performance of the NN was done by performing three experiments, each consisting of 1000 simulations in which the NN was trained and the output was compared to the GRACE EWH time series. The spatial accuracy was defined as the average Nash-Sutcliffe Efficiency (NSE) over the validation data. The temporal agreement was expressed in terms of the correlation coefficient (CC). In the first experiment, the model was trained using only soil moisture data as input. The NN underfitted the GRACE time series and could not recreate its local minima and peaks. As a consequence, the validation and training NSE were 0.43 and 0.85, respectively. Only Swarm data was used to train the NN in the second experiment. The NN overfitted the GRACE data, resulting in a significant discrepancy in training and validation NSE of 0.99 and 0.6, respectively. The last experiment combined the SM data with the Swarm data as input to the NN. This caused the NN to neither underfit nor overfit the GRACE time series, yielding the highest validation NSE of 0.68 with a training NSE of 0.83. It also resulted in a high temporal agreement with GRACE, having a correlation coefficient (CC) of 0.95 between the NN output and the GRACE time series. It could thus be concluded that combining Swarm and hydro-climatological as input data outperforms the individual solutions from experiments 1 and 2. The results of this thesis were compared to the gap-filling results obtained in previous studies. This study outperforms the previous studies in terms of temporal correlation (CC) over the Amazon Basin and agrees with previous work in terms of spatial accuracy (NSE). This thesis has shown that combining both Swarm-derived gravity field data and hydro-climatological data with data-driven techniques is a promising alternative for filling the GRACE/GRACE-FO data gap, keeping in mind that the performance of this technique is currently limited by the data availability of Swarm.

Contents

Nomenclature

Abbreviations

Abbreviation	Definition
Adam	Adaptive Moment Estimation
ANN	Artifical Neural Network
ARMIA	AutoRegressive Integrated Moving Average
CC	Correlation Coefficient
DCNN	Deep Convolutional Neural Network
DLR	Deutsches Zentrum für Luft- und Raumfahrt (German Aerospace Center)
DNN	Deep Neural Network
ECMWF	European Centre for Medium-Range Weather Forecasts
GRACE	Gravity Recovery and Climate Experiment
GRACE-FO	GRACE Follow On - mission
LEO	Low-Earth Orbit
LS	Least Squares
LSM	Land Surface Model
LSTM	Long Short Term Memory
MLP	Multilayer Perceptron
MLR	Multiple Linear Regression
NASA	National Aeronautics and Space Administration
NSE	Nash-Sutcliffe Efficiency
RMSprop	Root Mean Squared propagation
RNN	Recurrent Neural Network
SARIMAX	Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors
SGD	Stochastic Gradient Descent
SMAP	Soil Moisture Active Passive
SSA	Singular Spectrum Analysis
STL	Seasonal and Trend Line
TWS	Terrestrial Water Storage
TWSC	Terrestrial Water Storage Changes

List of Figures

1

Introduction

Much research has relied on the data collected by the National Aeronautics and Space Administration (NASA) and German Aerospace Center (DLR) GRACE and GRACE Follow-On missions. Their gravity field solutions and derived terrestrial water storage changes (TWSC) have been demonstrated to be beneficial to the advancement of the understanding of oceanography, estimating ice mass loss, monitoring hydrology across terrestrial regions, and modelling climate change [31][30][33]. It has allowed scientists to measure the Earth's gravitational field with previously unheard-of spatial precision and to do so on a monthly basis. However, there are some limitations to the mission. There is considerable uncertainty in the lower GRACE spherical harmonics (SH) coefficients, especially in the $C_{2,0}$ term. Common practice is to replace this coefficient with a value obtained from Satellite Laser Ranging (SLR)[5]. Secondly, at the end of the GRACE mission, there was a degradation of the power system, leading to small data gaps of 1 month. Finally, the last GRACE model was obtained in May 2017, and its successor, the GRACE Follow-On (GRACE-FO) mission, started its measurements in May 2018. This means that there is an 11-month data gap between GRACE and GRACE-FO data.

The primary objective of the European Space Agency (ESA) Swarm Earth Explorer mission, which was launched in 2013, was to map the geomagnetic and electric fields in the atmosphere [14]. Three separate satellites make up the Swarm constellation, one of which has a higher orbital height than the other two and the last two fly in a pendulum formation. The dual-frequency Global Positioning System (GPS) receivers on each of these identical spacecraft enable exact orbit determination. Several studies have shown that long-wavelength temporal fluctuations in the Earth's gravitational field may be observed using GPS data from low-Earth orbit satellites like Swarm. This technique has been applied to the Swarm satellites in studies with the purpose of bridging the GRACE/GRACE-FO data gap. It allows for an alternative source of information on TWSC, albeit with a much lower spatial resolution, 1500km instead of 350km for GRACE and GRACE-FO.

GRACE TWSC data provides insights into how the Earth's surface mass distribution moves over time. These surface mass changes reflect or have an impact on a variety of meteorological and hydrological factors, including rainfall, temperature, and soil moisture. Land Surface Models (LSMs) and hydrological models incorporate such variables [26][10]. As a result, numerous studies have sought to create a relationship between these global LSMs and hydrological models and GRACE TWSC, with the goal of simulating GRACE-like TWSC data using the LSM and hydrological variables during the GRACE and GRACE-FO mission data gaps [22][21][29][28][12][1]. Most of these studies implemented data-driven techniques such as neural networks (NNs) and obtained some promising results. When compared to GRACE and Swarm data, the key advantage is that these LSMs provide data with a high temporal and spatial resolution. The fact that these measurements do not immediately provide information on the Earth's gravity field is a drawback, though. Instead, they offer details on variables that could impact the local gravity signal.

The purpose of this research was to merge the idea of using Swarm and hydro-climatological data to bridge the GRACE/GRACE-FO data gap. These two types of data were evaluated as predictors of

GRACE-like TWSC and combined as inputs to ML models. The term ”Models” is used to describe NN models unless specified otherwise. The area of study was chosen to be the hydrologically active Amazon Basin. In this way, direct information on the Earth’s gravity field was merged with high-resolution hydrological data to improve the performance of each strategy separately.

This all resulted in the following research question:

> ”Can the GRACE/GRACE-FO data gap be bridged by down-scaling Swarm gravity solutions to GRACE-like resolution by the use of ML with hydro-climatological input variables?”

This main research question could then be broken down into multiple sub-questions:

1. Which combination of hydro-climatological input variables is the best predictor for GRACE-like TWSC?
2. What ML model/architecture yields the best performance?
3. Does the addition of Swarm to the ML input data improve the performance of the model?

This report starts off by a literature review on the topic of GRACE/GRACE-FO and the previous GRACE/GRACE-FO gap-filling techniques, presented in Chapter 2. Chapter 3 introduces the different datasets that were used in this research. It also shows the different pre-processing steps that were taken in order to prepare the data before training the NNs. The various machine learning (ML) concepts and neural network architectures used in this project are presented in Chapter 4. After introducing the ML models and the relevant datasets, Chapter 5 describes the training and optimization of the NNs, as well as the evaluation metrics that were used to assess their performance. Once the ML models were optimised, the next step was to use them to fill the GRACE/GRACE-FO data gap and asses their gap-filling performances. This is presented in Chapter 6. The last chapter, Chapter 7, summarizes the research approach and answers the research questions. Chapter 7 also presents recommendations for future research.

2

Literature Review

Bridging the GRACE/GRACE-FO mission gap has been the topic of multiple studies in the past, as presented in this chapter. It brings an understanding of the mass transport over the 11-month period between the two missions and provides a backup technique for obtaining GRACE-like EWH in case GRACE-FO should ever become temporarily inoperative or even fail. After analyzing previous studies, four distinct strategies could be identified. Based on their methodologies and data, these are: extrapolating GRACE data, using Swarm-derived TWSC, using ground-based GPS data, and using data-driven methods with hydro-climatological input data. This chapter presents an overview of the literature on these topics. After presenting the previous studies, the drawn conclusions are discussed. From these conclusions, a research aim is presented.

2.1. Extrapolating GRACE

Li et al. (2019) [20] have used Singular Spectrum Analysis (SSA) time-series forecasting to bridge the GRACE/GRACE-FO data gap. They considered six regions of interest, all divided over China. They used GRACE spherical harmonic data, ranging from January 2003 to August 2016, provided by the University of Texas Center for Space Research (CSR) level-2 Release Level 06 (RL06). These spherical harmonic data sets were synthesized to surface mass changes expressed by equivalent water height (EWH). RL05 and RL06 Mascon data, obtained from JPL, were used for verification of the inverted EWH from the SH data. From the comparison between these TWSC and the Mascon results, they concluded that both the correlation coefficient (R) and the Nash-Sutcliffe efficiency (NSE) were the highest between the GRACE TWSC and the Mascon RL06. This holds for all six regions of interest. Therefore they used the Mascon RL06 data for the predictive analysis verification.

The principal components of the temporal hydrological signal variations were extracted by the use of SSA and were used to reconstruct the periodic term signals. They predicted GRACE-like TWSC from September 2016 to May 2019. The GRACE TWSC data from January 2003 to December 2012 were used as training data, whereas the data from January 2013 to August 2016 were used as validation data, assessing the accuracy of the SSA on different time scales. They found that in regions where GRACE TWSC experienced strong periodic signals, the accuracy of the SSA forecasting was high (SWC and TRHR in Table 2.1). For regions with a weak periodic signal and strong human factors, there was considerable uncertainty in the long-term results of the SSA (NCP in Table 2.1).

The results of the SSA were compared to an Autoregressive Moving Average (ARMA) model in Table 2.1. The Autoregressive Integrated Moving Average (ARIMA) model also performs well in the same regions with strong periodic signals, just like SSA. However, for all regions and time periods, the accuracy of the SSA is higher than for the ARIMA model.

Area	Method	Short-Term	Mid-Short Term	Medium-Term	Long-Term
NCP	SSA	0.51/0.26/2.37	0.45/0.20/2.38	0.36/0.08/2.34	0.34/2.27/0.03
	ARMA	0.36/0.12/2.58	0.30/0.07/2.56	0.27/0.03/2.40	0.26/0.02/2.29
LSWZ	SSA	0.87/0.61/4.27	0.85/0.59/3.89	0.70/0.44/4.41	0.58/0.31/5.55
	ARMA	0.73/0.54/4.63	0.76/0.57/3.98	0.63/0.39/4.61	0.65/0.42/5.11
HRB	SSA	0.67/0.35/2.16	0.65/0.32/2.26	0.61/0.24/2.38	0.56/0.09/2.54
	ARMA	0.59/0.32/1.39	0.53/0.23/1.51	0.43/0.05/1.66	0.28/−0.21/1.812
SWC	SSA	0.98/0.95/1.83	0.96/0.92/2.22	0.96/0.91/2.29	0.96/0.90/2.27
	ARMA	0.97/0.94/2.02	0.95/0.90/2.27	0.94/0.88/2.40	0.93/0.86/2.57
TRHR	SSA	0.97/0.95/1.73	0.96/0.92/2.10	0.95/0.91/2.30	0.95/0.90/2.37
	ARMA	0.94/0.89/2.55	0.91/0.84/3.08	0.90/0.81/3.33	0.91/0.82/3.14
TSMR	SSA	0.94/0.88/1.07	0.91/0.81/1.40	0.88/0.77/1.48	0.78/0.60/2.21
	ARMA	0.93/0.68/1.72	0.89/0.54/2.18	0.82/0.49/2.18	0.78/0.40/2.70

Table 2.1: Results of the SAA method compared to the ARMA results for different areas and different time scales. Structured as Correlation Coefficient/Nash-Sutcliffe Efficiency/Root Mean Squared Error [cm] [20]. The six regions are: North China Plain (NCP), Southwest China (SWC), Three-River Headwaters Region (TRHR), Tianshan Mountains Region (TSMR), Heihe River Basin (HRB), and Lishui and Wenzhou area (LSWZ)

The TWSC over the data gap was predicted using the SSA method from the 33 months prior to the gap, for all six regions. They compared the predicted results to the NASA Global Land Data Assimilation System (GLDAS) model (presented in Chapter 3) and found consistency between the two. Lastly, the predicted TWSC were compared to the first measurements of the GRACE-FO mission (from 06/2018 to 05/2019). An overall agreement was found, except for one of the six regions.

This method itself was of little importance to this research, as it only used data from GRACE itself. Therefore the predicted TWSC are not influenced by external events that might be captured by Swarm data or other data sets. There are however some elements in the article that were interesting for this project. The conclusion that regions with strong periodic signals and few human interventions perform better when it comes to forecasting TWSC played an important role in the selection of the Amazon basin as the study region.

2.2. Swarm Mission

The ESA Swarm Earth Explorer mission was launched in 2013 with as its main goal the surveying of Earth's geomagnetic field, together with the atmosphere's electric field[14]. The Swarm constellation consists of three individual satellites, with one at a higher orbital altitude than the other two flying in a pendulum formation. These identical satellites are each equipped with a dual-frequency GPS receiver that allows for precise orbit determination. Multiple studies have proven that long-wavelength temporal variations in Earth's gravity field can be observed by using GPS data from low-Earth orbit satellites such as Swarm[7]. Other studies that have applied this principle to the Swarm satellites, with a goal to bridge the GRACE/GRACE-FO data gap, are presented in this chapter.

2.2.1. Gravity Field Models from Swarm

Lück et al. (2018) [23] computed time-variable gravity fields from 37 months of kinematic Swarm orbits. They applied the integral equation approach with short arcs. They used these gravity fields to compute mass changes over 4 river basins, Greenland and the ocean. These mass changes were compared to solutions from GRACE.

They used Swarm Level 2 kinematic orbits provided by van den Ijssel et al. (2015)[32] to compute the gravity fields. They modelled air drag, solar radiation pressure, and Earth radiation pressure since the Swarm spacecraft's accelerometers are impacted by bias variations. These modelled accelerations were used in gravity estimation to replace the accelerometer measurements. Some post-processing was applied to the estimated gravity fields. They tested the impact of replacing the C_{20} coefficient by a value from Satellite Laser Ranging (SLR) but chose not to implement it for the assessment of the capability of Swarm alone to bridge the GRACE/GRACE-FO data gap. Next, they corrected for geo-

center motion by adding all degree 1 coefficients. They applied the same correction for glacial isostatic adjustment that was applied to the GRACE data with which they compared the results. Lastly, they employed an ocean or land mask, depending on what they wanted to analyse, to separate the ocean grid cells from the land grid cells. This avoids signal leakage and biases.

They estimated both monthly gravity field solutions and constant, trend, annual and semiannual (CTAS) signal terms directly. For both types of solutions, they estimated the spherical harmonics up to degree/order (d/o) 40 and evaluated them up to d/o 12.

They found that Swarm-derived time series behave similarly to GRACE-derived TWSC, but are overall noisier than the GRACE solutions. Important to note that in July 2014 there was an update to the Swarm GNSS receiver, after which the quality of the solutions improved. As for performance indicators, they computed the Root Mean Square Error (RMSE) between the Swarm-derived TWSC and the GRACE solutions, as well as the ratio of the variance of the GRACE time series to the RMSE. They call this the signal-to-noise (SNR) ratio in the given region. They found that the quality of the solutions was higher in larger regions and regions with higher signal strength, such as the ocean and the Amazon Basin. Lastly, they analysed the performance of Swarm solutions towards filling gaps of different lengths in the GRACE data. In Table 3.1, the values in the upper row represent the number of months in the gap and the data inside the table is the RMSE (mm of EWH). The values between brackets represent the percentages of solutions in which the CTAS solution outperforms GRACE extrapolation (which they call 'interpolated').

	1	3	6	12	18
GRACE (interpolated)	0.9	1.1	1.1	1.2	1.8
Swarm (CTAS)	1.4 (13.5 %)	1.5 (17.1 %)	1.6 (6.3 %)	1.6 (3.8 %)	1.6 (80.0 %)
Swarm (monthly)	3.3	3.7	3.8	3.9	3.8

Table 2.2: Gap filling performance of extrapolating GRACE, Swarm derived CTAS solutions and monthly Swarm solutions [23]

They found that for gaps smaller than one-year extrapolating GRACE data performs better than the Swarm solutions, in general. However, for longer gaps (18 months), the Swarm solutions perform better than extrapolating GRACE in 80% of all cases. Next to this, it is clear that estimating CTAS solutions outperform the monthly Swarm solutions for every gap length.

Richter et al. (2021) build further on the approach of Lück et al. (2018) [23] and focused on a reconstruction approach to improving the spatial resolution while implementing a priori information from when the GRACE mission was still in operation.

They decomposed the GRACE data to obtain three leading spatial modes. They then projected the Swarm solutions to these spatial modes. The reconstructed solutions are thus spatially constrained by the GRACE spatial modes. They compared these reconstruction solutions to GRACE solutions and Swarm-only solutions from Lück et al. (2018) over large basins, Antarctica and Greenland, and found that the reconstruction reduced the variance and improved the resolution of the Swarm gravity field solutions. The TWSC changes from the Swarm-reconstructed gravity fields had a much lower RMSE with respect to GRACE. To d/o 12, RMSE over the Mississippi basin decreased from 0.09 m for the Swarm-only solution to 0.04 m for the Swarm-reconstructed solution. If the gravity fields were evaluated up to higher d/o, the reconstructed fields outperformed the noisy Swarm-only solutions. For d/o 40, the global RMSE decreased from 0.29 m for the Swarm-only solutions to 0.08-0.02 m for the reconstructed gravity fields. This increase in d/o for which reasonable monthly solutions can be derived improved the spatial resolution of TWSC maps. Another important conclusion was that the reconstruction approach does not seem to be affected by the high ionospheric activity during the 2014-2015 period, which negatively impacted the Swarm performance.

The main takeaway from these studies is the fact that these Swarm-derived TWSC behave similarly to GRACE-derived TWSC, albeit at a lower spatial resolution. This motivates the use of Swarm as an

input to the ML model aimed at predicting GRACE-like TWSC over the mission gap. Another interesting aspect is the performance indicators they used, consisting of the RMSE and SNR. Lück et al. (2018) [23] also stated that the Swarm TWSC perform better over large basins such as the Amazon basin, which again acted as a motivation to choose this as the region of interest for this research.

2.2.2. SLR and Swarm Combination

Meyer et al. (2019) [24] have combined monthly solutions from satellite laser ranging (SLR) to geodetic satellites with Swarm-derived solutions in order to bridge the GRACE/GRACE-FO gap. They focused on the region over Greenland.

Only SH coefficients of degrees 2 to 5 and of degree 6 and order 1 can be determined from SLR alone, on a monthly scale. Therefore a combined solution with Swarm was set up to evaluate the gravity signals to higher degrees. In this combination, the SLR SH coefficients to d/o 10 are set up, to avoid omission and commission errors. They used Swarm derived gravity fields that are sensitive to temporal gravity variations up to d/o 13, based on Teixeira da Encarnação et al. (2016) [7]. The SLR, Swarm and GRACE data were all derived at the Astronomical Institute of the University of Bern (AIUB)

They performed a combination of Swarm and SLR on the normal equation level. By doing so, the correlations between gravity field coefficients and other parameters were taken into account, making it superior to a combination on the solution level. The weighting was performed based on variance component estimation (VCE). The resulting SHC were transformed to TWSC in terms of 1°-grids of EWH. For sake of comparison to the unconstrained SLR solutions, they truncated all the evaluated gravity fields at the same maximum degree 6. Lastly, they integrated the results over the area to obtain mass estimates over Greenland. The results from different SLR-Swarm weight combinations are shown in Figure 2.1. The lines in the plots represent the difference in ice mass between GRACE estimates and different combinations of SLR and SWARM estimates. The combinations consisted of SLR only, Swarm only, three different relatively weighted SLR and Swarm combinations (x*SLR+SWARM) and an equally weighted SLR and Swarm combination (SLR+SWARM).

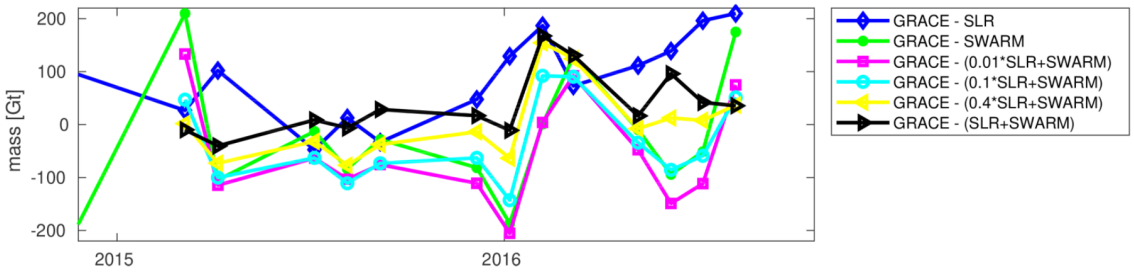


Figure 2.1: Mass over Greenland derived from different SLR-Swarm combinations compared to GRACE estimates Meyer et al. [24]

They found that an SLR-Swarm combination with close to equal weights (black line in Figure 2.1) has the highest performance, even better than weights based on VCE. This is reflected by the black line in Figure 2.1 being closest to 0 over time. Based on this, they concluded that the SLR-Swarm combined gravity field can be used to bridge the GRACE/GRACE-FO data gap for large-scale mass loss in the polar region.

2.2.3. Multi Approach Gravity Field Models

Teixeira da Encarnação et al. (2020) [8] have developed high quality gravity field models (GFMs) from Swarm by combining four different methods: celestial mechanics approach (CMA), decorrelated acceleration approach (DAA), short-arc approach (SAA), and the improved energy balance approach (IEBA). They continue on the work by Teixeira da Encarnação et al. (2016) [7], in which the first three of these methods were already combined.

They showed that the combined solution is superior to any of its individual solutions. Whereas the combination of the three solutions in Teixeira da Encarnação et al. (2016) [7] was performed by aver-

aging, this study applied a more advanced variance component estimation (VCE) method to determine the weights of the individual solutions. Next to this, they also tested the implementation of kinetic baselines (KBs) between the two Swarm spacecraft in pendulum formation. They concluded that the addition of KBs did not significantly improve the resulting GFMs.

The combined solutions were derived from 4 individual GFMs produced at 4 different institutions, each based on one of the four approaches described above. The GRACE data used for comparison consists of the RL06 GRACE and GRACE-FO GFMs from the Center for Space Research (CSR). In order to isolate the time time-variable gravity component, the Combined GRACE Gravity Model 05 static GFM was used. For months where GRACE and Swarm data are compared, the Swarm solutions were linearly interpolated to the epoch of the GRACE solutions. In terms of post-processing, the $C_{2,0}$ coefficient was replaced by an estimate value from SLR and an ocean mask was applied.

The VCE weights were derived on both the solution level and the normal equation level. It was determined that the weights derived at the solution level yielded a solution that was in better agreement with GRACE than the weights derived at the normal equation level. For the derivation of the VCE weights, the individual solutions were considered up to d/o 20. However, the maximum degree of the combined solution is d/o 40.

They found that over the ocean, Swarm is able to identify long term trends and global ocean mass. Swarm is not able to resolve monthly signals at GRACE-like spatial resolutions over the ocean. Over land areas, the GFM from Swarm agree much better with the solutions from GRACE. They found that it is possible to identify the majority of mass transport processes monitored by GRACE. They evaluated the performance of the Swarm GFMs over eight large basins and Greenland. They found an average correlation coefficient of 0.79 between Swarm and GRACE TWSC over the nine regions they considered. Over the Amazon basin, a correlation of 0.95 was found.

This research is very interesting because it presents another well-performing Swarm solution (when it comes to replacing GRACE) that could be used as an input to the ML models in this research. Next to this, it discussed some post-processing methods that might be of use such as applying ocean masks. Lastly, it introduced a new performance indicator under the form of the correlation coefficient.

2.3. Ground-Based GPS

Another alternative to estimate surface mass variations is inverted GPS measurements. The main advantage of this method is that ground-based GPS stations provide near real-time data on the vertical displacement of the Earth's surface. Next to this, provided that the stations lie in a dense network, the spatial resolution can sometimes be as high as 50 km[4]. This sort of data might be used in this research since it is accessible before, during, and after the GRACE/GRACE-FO gap. Two examples of studies that have compared inverted GPS measurements to GRACE are discussed in this chapter.

Rietbroek et al. (2014) [25] derived low-degree surface loading data from a global GNSS network. The surface deformation was inverted into spherical harmonics of surface mass. This inverted data was then tested as a potential gap filler for the GRACE/GRACE-FO mission gap.

They used data from 216 globally distributed GPS stations. The data was provided on a normal equation level. They filtered out the data from the GPS stations that had less than a year's worth of data. The inverted surface loading data was truncated at degree 10. The GRACE data used for the reference solution was the RL05 GRACE from GFZ, in the form of weekly normal equation systems. Simulated ocean bottom pressure data computed by the Finite Element Sea-Ice Model (FESOM) was also used in the reference solution.

For the sake of comparison, they set up a reference solution in which GRACE, GPS and FESOM are combined. A joint inversion scheme was used to create this combined solution. The weighting of the individual solutions was done by VCE.

They evaluated three methods to fill a gap in the GRACE/GRACE-FO mission. To asses the performance of the gap filling methods, two data gaps were simulated. The first gap was over the 2006 calendar year, thus having a length of one year. The second gap had a length of 1.5 year and was centered around March 2010. They did not specifically list the motivation behind this choice.

The first method (GPS-I) only used GPS data from within the GRACE gap. From this GPS data, surface loading coefficients were obtained up to degree 10. In the second method (GPS-C), correction parameters were applied to the GPS-I solution. These correction parameters were based on an overlapping period with the GRACE mission outside the gap, before and/or after the gap. For the 1-year gap, the overlapping periods were at both ends and had an individual length of 0.75 year each. Over these overlapping periods, a degree-1 polynomial was estimated, together with the annual cosine and sine amplitudes. The surface loading coefficients during the gap were then expanded by use of these correction parameters. For the second gap, a similar strategy was used, only the overlapping period lies before the gap and had a length of 1.5 year. The parameters consisted of an estimated bias and annual sinusoid. The last method (REF-S) consisted of a seasonal fit to the joint inversion solution. The motivation for the short overlapping periods was that they wanted to test whether a minimal amount of data would already be able to fill the gap sufficiently.

As a performance indicator they computed the signal-to-noise ratio (SNR). The signal term came from the reference solution, whereas the noise was defined as the error between the simulated solution over the gap and the reference solution. The spherical harmonics coefficient-wise SNRs are presented in Figure 2.2 for the two types of gaps and the three methods.

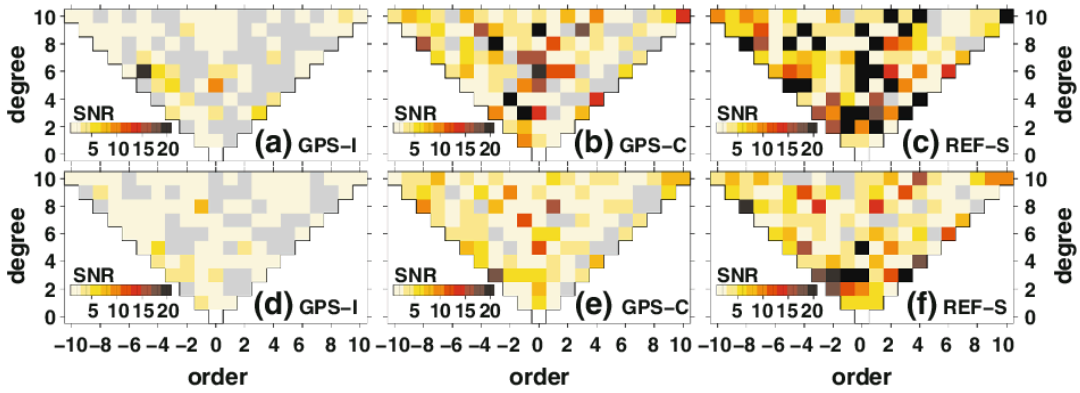


Figure 2.2: The coefficient wise SNR of the different gap filling methods for the two considered gaps. From left to right: GPS-I, GPS-C, REF-S. From top to bottom: 1-year gap, 1.5-year gap.

In Figure 2.2, the grey tiles represent the coefficients for which the SNR lies below 0.8. They concluded that including the correction parameters increases the gap-filling capabilities of the GPS solutions. However, the more simple REF-S method yielded even better results. The main disadvantage of this REF-S method is that it is restricted to a simple harmonic variation and a trend. They also concluded that the GPS-I and GPS-c methods might be useful to study time-variable signals, given that the region of interest consists of a dense GPS network.

A similar study was performed by **Zhong et al. (2020)** [34] over Southwest China. Their goal was to evaluate the performance of the GPS-inverted TWSC under different GPS stations' density distributions. They compared the inverted TWSC to GRACE and GRACE-FO data and demonstrated the feasibility of filling the GRACE/GRACE-FO gap.

They obtained the GPS vertical displacement data from the Crustal Movement Observation Network of China (CMONOC). The data set consisted of measurements from 85 GPS stations spread over Southwest China. The GRACE data used for comparison consisted of both TWSC inferred from SH as well as GRACE mascon surface mass variations. The monthly SH up to d/o 60 from the CSR RL06 were

used. The mascon solutions also came from RL06 provided by CSR.

Overall they found that the GPS-inverted measurements agreed with the GRACE solutions relatively well. They also found that the GPS solutions in the areas with a denser GPS station distribution outperformed the solutions from the areas with a more sparse distribution of stations. To minimise biases between the GPS-inverted data and GRACE/GRACE-FO data, they applied a scale factor method. This increased the accuracy of the GPS-inverted surface mass variations when compared GRACE. They concluded that the scaled GPS-inverted solutions can help bridge the GRACE/GRACE-FO data gap.

Their conclusion that the GPS-inverted models agree well with the GRACE solutions, allows it to be considered as a new type of input data for this research. Important to keep in mind is that the choice of the study area will be heavily influenced by the GPS measurements, as an area with a dense GPS network is required for better performance. Another interesting element of their research is the way in which they asses the gap-filling performance.

2.4. Data-Driven Techniques

With the rise of popularity of data-driven techniques such as machine learning, there have been multiple studies that have applied these types of algorithms to try and fill up the GRACE/GRACE-FO mission gap. In some of these studies, a single algorithm is selected and developed, whereas in other studies a comparison of different algorithms is presented. Most of these studies also focus on specific regions of interest. The main similarity between them is that they all try to find a relationship between GRACE TWSC data and climatological/hydrological parameters (such as temperature, water storage, precipitation) which are available during the GRACE/GRACE-FOR mission gap. In the following sections, an overview of these studies and their data-driven techniques is presented.

2.4.1. Multilayer Perceptron Artificial Neural Network

Long et al. (2014)[22] developed a Multilayer Perceptron (MLP) Artificial Neural Network (ANN) that was trained to generate GRACE-like TWSC from Land Surface Model (LSM) precipitation/temperature measurements. With this model, they simulated three decades of TWSC prior to 2012. They concluded that this model can also be used to bridge the gap between the GRACE/GRACE-FO missions. A simple overview of their ANN can be found in Figure 2.3

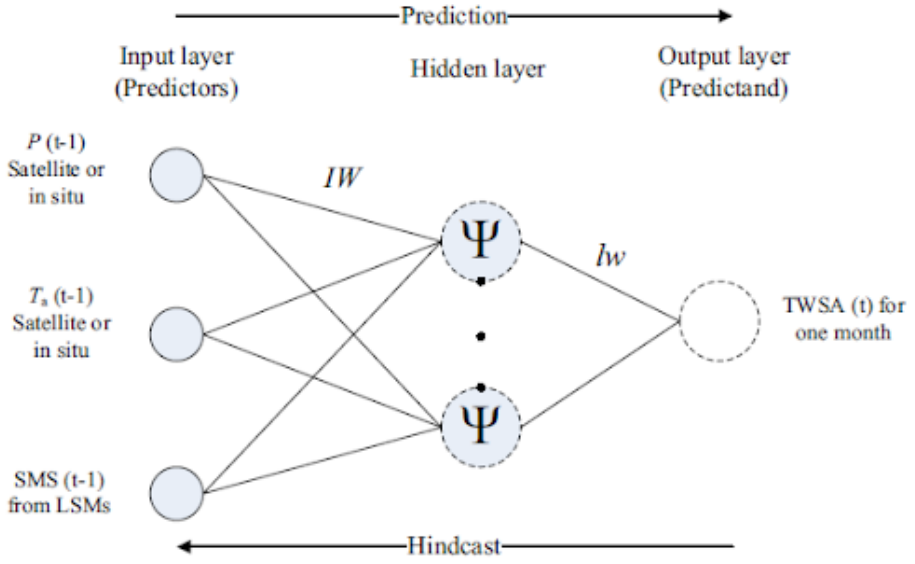


Figure 2.3: The architecture of the MLP model from [22]. P is the precipitation, T_a the ambient temperature, and SMS stands for Soil Moisture Storage from the NOAH LSM. IW and lw represent connection weights, and ψ represents the sigmoid transfer function.

This is a quite straightforward ANN that consists of just a single hidden layer. The division between training, validation and testing data was 60%, 20%, and 20% respectively.

They considered three predictor combinations, consisting of the NOAH LSM, in situ temperature measurements and in situ precipitation measurements. They found that the combination of NOAH soil moisture storage (SMS) and in situ precipitation as predictors yield the best performance of the ANN model. This model predicts TWSC with a coefficient of determination (R^2) of 0.91, a bias of -21 mm and an RMSD of 28 mm compared to GRACE TWSC. This is on a basin scale. On smaller scales, the performance decreases because of larger uncertainties in GRACE TWSC.

They presented an example of a rather straightforward ANN applied to the GRACE/GRACE-FO problem. This may serve as a starting point for the development of ML models. It also already presents two hydro-climatological parameters that influence the TWSC and might serve as input to the models used for this research.

2.4.2. ANN vs. ARX vs. MLR

Li et al. (2020)[21] developed a methodological framework to compare different data-driven methods that aim to predict GRACE-like TWSC, based on their relationship to climatological and hydrological variables. They considered an artificial neural network (ANN), a multiple linear regression (MLR) model and an autoregressive exogenous (ARX) model. They first decomposed the GRACE observed TWSC and climate data into temporal modes and spatial patterns by applying either principal component analysis (PCA) or independent component analysis (ICA). Next, they applied time series analysis such as Seasonal-Trend decomposition using LOESS (STL) or Least-Squares (LS) fitting in order to further decompose the individual modes into three component trends (linear, seasonal, residual). Lastly, the above-described data-driven methods were used to predict GRACE-like TWSC in the GRACE/GRACE-FO data gap, based on the relationship between the temporal modes of GRACE and the climate data. This was done for every possible combination of decomposition strategy, time series analysis, and ML model. They considered the world's 26 main river basins as regions of interest.

For the ANN, a Multi-layer perceptron (MLP) model was chosen. It consisted of the input and output layers and 1 hidden layer with 7 neurons. The output layer represents each decomposed component of the GRACE temporal modes. The input layer contains the 3 climate predictors that have the largest correlation with the predicted target component. The ARX model and the MLP model both used the

same inputs/outputs as the ANN model, for sake of comparison.

They found that both the ANN and the ARX yielded the most accurate results but tend to overfit. The MLR was identified to be more robust, at a cost of slightly lower correlation with the GRACE data (0.86 for the best basin). They also state that a region size of 10-15 million km^2 is optimal for their model.

2.4.3. DNN vs.MLR vs.SARIMAX

Sun et al. (2020)[29] developed 3 models (Deep Neural Network (DNN), MLR, Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors (SARIMAX)) based on 3 GRACE solutions (3 SH and 3 Mascons). They used a precipitation model, temperature model and NOAH TWS as training data for the model. They reconstructed TWSA data over 60 basins globally at a spatial resolution of 1 degree. The flowchart of their model development is shown in Figure 2.4

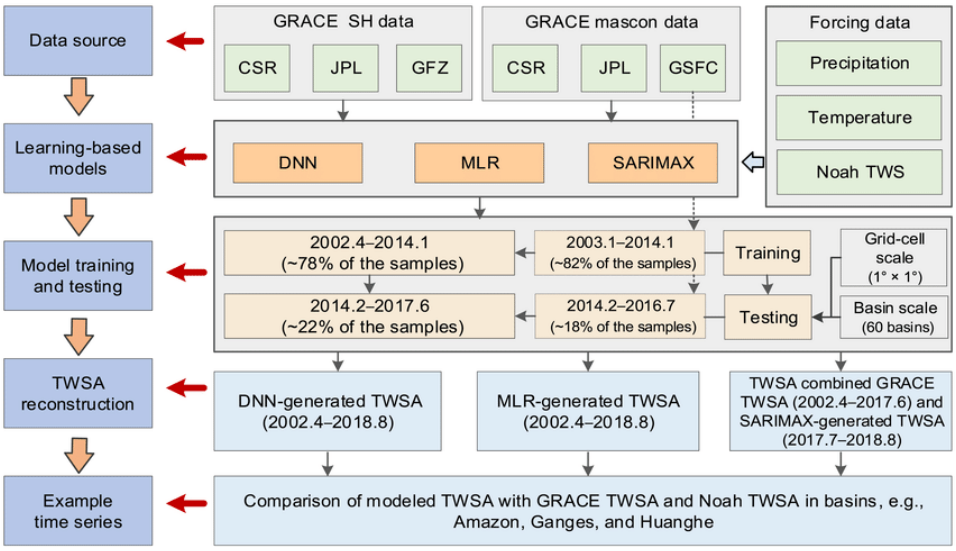


Figure 2.4: Flowchart of the three models developed by Sun et al. (2020) [29]

DNNs can be viewed as ANNs, but with deeper hidden layers. The deeper layers enable the DNN model to find more complex relationships between the input and output data. The SARIMAX model is an extension of the data-driven ARIMA model that includes seasonal components and outside variables (the input variables in this case). It is not a machine learning technique but a time series forecasting technique that identifies correlations in the data. It is important to note that there is a time lag of 1 month between the input data (temperature and precipitation) and the GRACE TWSC. The division of the data set in training, validation and testing data was kept constant throughout the study.

They found that the performance of models is affected by the climate and irrigation. In general, the performance is higher in humid and low-intensity irrigated basins. Next to this, also basin size and other hydro-climatological processes influence the performance of the models. The highest-performing models were identified to be the DNN and the SARIMAX models (correlation coefficient (CC): 0.88-0.95). The DNN models yielded slightly higher results than the SARIMAX. In terms of GRACE data, the mascons solutions performed better than SH solutions over the basins, but there were more differences between the different mascon solutions than between the different SH solutions.

Interesting for my research is the DNN model itself, as this is a more complex version of the previously mentioned ANN. Next to this, the conclusion that this model performs better over humid and low-intensity irrigated areas drove the selection of the Amazon basin as the study region.

2.4.4. Deep Convolutional Neural Networks

Sun et al. (2019)[28] developed 3 Deep Convolutional Neural Networks (DCNNs) (CGG16, Unet, and SegnetLite) to reconstruct the mismatch between the NOAH TWS and GRACE mascon observed TWS. This mismatch was then used to correct the NOAH TWS. Their area of interest was India. Their model development is presented in Figure 2.5.

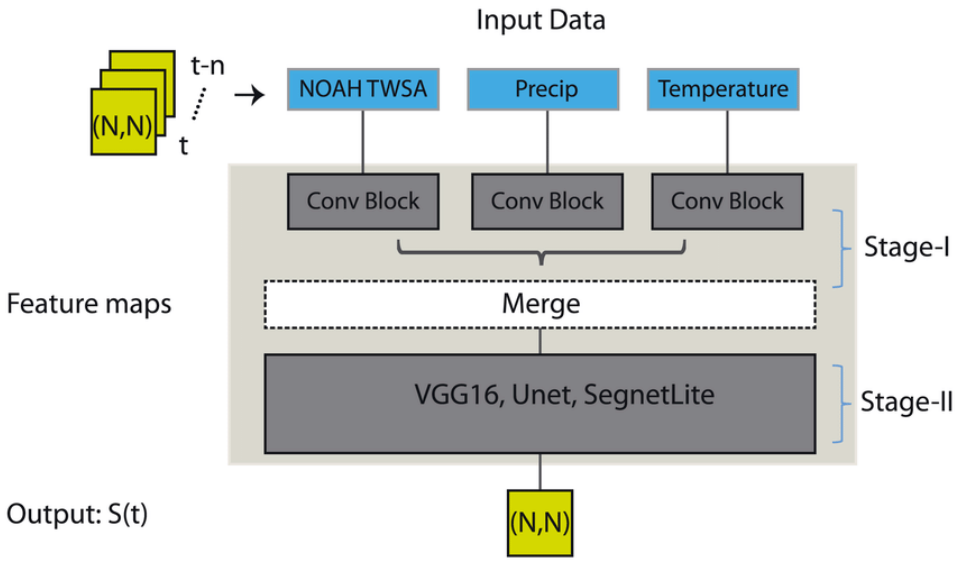


Figure 2.5: The architecture of the DCNN developed by Sun et al. (2019)[28]. Stage-I consists of a convolution state, Stage-II of a specific convolutional network.

Convolutional neural networks are a variation of neural networks that are very good at detecting patterns. As opposed to the straightforward MLP models, CNNs make use of hidden convolutional layers, which filter the data to identify these patterns. The input parameters were converted to a 2-D image and stacked on top of the images from the previous two-time steps to implement temporal correlations in the data resulting in a 3-D data set. Each of the CNNs described above has its own specific architecture. They differ from each other in terms of the number of layers and combinations of different layers such as convolutional layers, pooling layers etc. These concepts are introduced in Chapter 4. Figure 2.5 shows that the CNN consisted of two stages, where Stage-I consists of a convolution state, and Stage-II of a specific convolutional network. The first stage seeks to extract distinctive characteristics from each input, while the second stage aims to deep-learn the spatial and temporal patterns present in each input as well as the covariation patterns between the inputs.

All of the developed models were able to increase the quality of the NOAH TWS (both in magnitude and phase) by estimating its mismatch, both on the grid and basin scale. They found that additional predictors such as precipitation and temperature do not improve the results. The best-performing model was the Segnet model with a correlation of 0.946 and an NSE of 0.875.

What is most useful for this research is the way in which they stacked the 2-D maps to a 3-D data set which they use as an input. By doing so, the input layer will hold information on the previous time steps and thus hold temporal information next to spatial information.

2.4.5. Nonlinear Autoregressive Exogenous Model

Ferreira et al. (2019)[12] developed a Nonlinear Autoregressive Exogenous Model (NARX) neural network that learned the relationship between multiple hydro-climatic variables in order to reconstruct GRACE-like TWSC estimates back to 1979 over the West-Africa region. The hydro-climatic variables used to train the model are listed in the upper text box in **??**. Additionally, a similar model was built to TWS from the NOAH land surface model. The results of this model were used, together with the original data set, for validation of the GRACE-like TWSC results.

The NARX neural network looks at the previous values of both the TWSC from grace and the climate parameters and looks for regressions. These are then used to predict GRACE-like TWSC over the period GRACE was not in operation. The architecture of the neural network is shown in Figure 2.6.

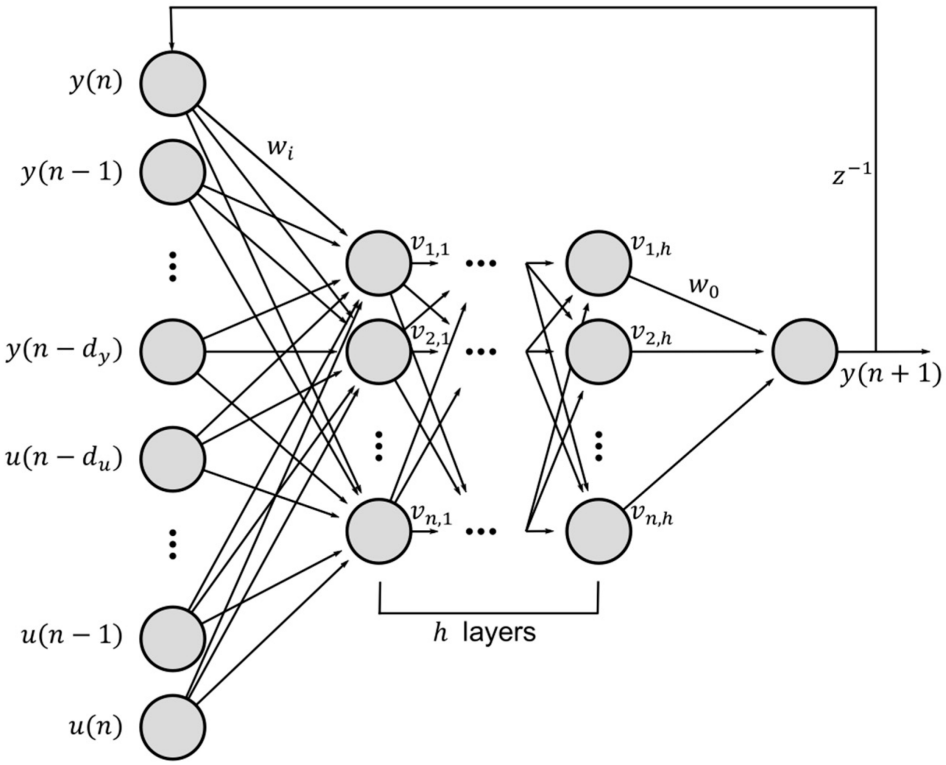


Figure 2.6: The architecture of the NARX neural network used by Ferreira et al. (2019)[12]. The number of hidden layers is equal to h=16. y(n) are the signals that are predicted and u(n) are the independent exogenous input signal at a discreet time step n. d_u and d_y are the are the input and output delays respectively. w_i is the weight of the input layer and w_o is the weight of the output layer. The *v*-values represent the parameters of the individual nodes in the h hidden layers.

The results show that the model was robust enough to be used for predicting GRACE-like TWSC estimates. The final TWSC were obtained with an RMSE of 11.83 mm/month, an NSE of 0.76 and a R^2 value of 0.89.

Ahmed et al. (2019)[1] also aimed to recreate and predict the GRACE TWSC time series by deriving a relationship between GRACE TWSC and climatological variables by using a NARX model. The climatological variables in this study consist of rainfall (R), evapotranspiration (ET), temperature (T) and vegetation index (NDVI). The architecture of the NARX model is shown in Figure 2.7.

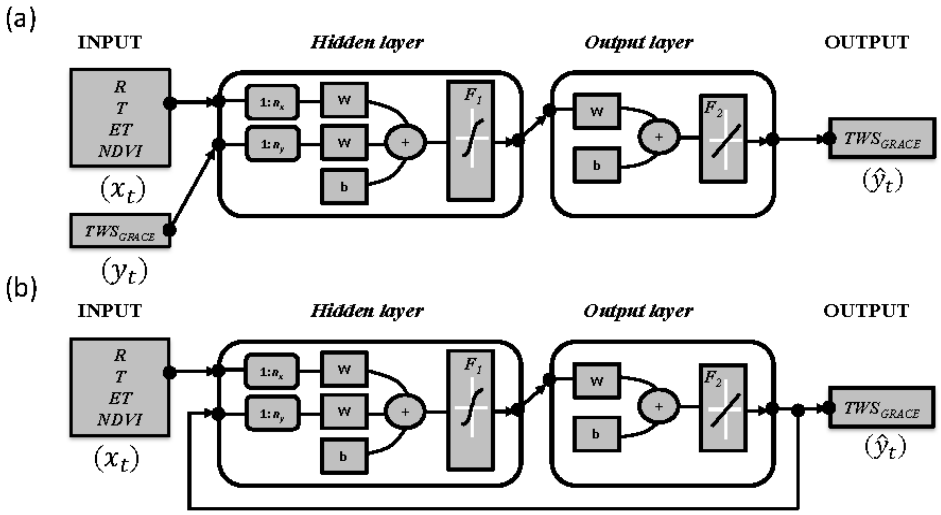


Figure 2.7: NARX neural network(s) used by Ahmed et al. (2019)[1] The w and b stand for the connection weights and bias, respectively. The F_1 denotes a hyperbolic tangent activation function and F_2 denotes a linear activation function.

The open-loop structure (a) was used in the training phase, while the closed-loop (b) structure depicts the trained model that was used to predict GRACE TWS time series. The model was used to forecast TWS over 10 African watersheds. In most of the watersheds, the performance was very good with NSE > 0.75 and R(scaled root mean squared error)<0.5. In some cases NSE>0.9 and R= 0.3. In some watersheds the performance was good (NSE>0.65 and Rstar<0.60) or satisfactory (NSE>0.5 and rstar<0.7).

These articles provided more options for hydro-climatological variables that might act as input parameters for the models in the Thesis. The performance indicators, R^2 and NSE, were also selected as performance indicators.

2.4.6. Recurrent Neural Network - Long Short-Term Memory

The last study presented in this chapter does not aim to represent GRACE data, but Soil Moisture Active Passive (SMAP) mission data. The NASA SMAP mission measures the top 5 cm of soil moisture globally. It is included here because the type of neural network they applied was considered for this research.

Fang et al. (2017)[11] developed a Recurrent Neural Network (RNN) - Long-Short Term Memory (LSTM) model that is able to forecast/hindcast SMAP-like top-surface moisture. The model uses atmospheric forcing, LSM and static physiographic attributes (such as sand, silt and clay percentages) as input. LSTM has a strong track record in solving time-domain tasks such as time series forecasting. A simplified overview of the model is presented in Figure 2.8.

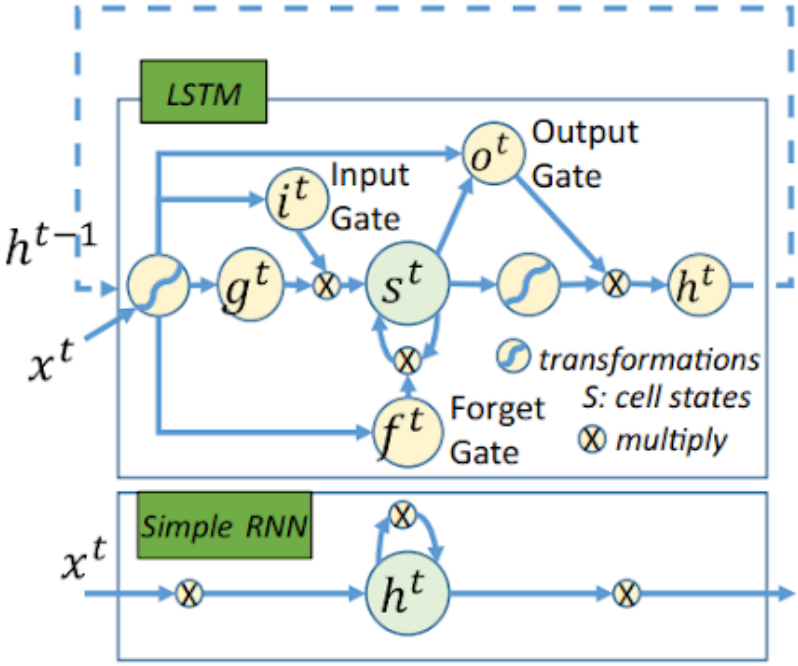


Figure 2.8: The architecture of the RNN-LSTM model used by Fang et al. (2017)[11]. The input vector for time step t is denoted by x^t and h^t denotes the hidden state. Unique to LSTM are the memory cells, denoted by s^t. They store and manipulate data via gates and control information flow between cells. The transformations from inputs to i, f , o are sigmoidal functions. From inputs to g and from s to h the transformation is a hyperbolic tangent function.

LSTM is a type of RNN that is especially powerful when it comes to sequenced data such as speech or video. It counters the fact that for long sequences of data, RNN does not include the states of multiple time steps in the past.

They found that the LSTM model performed better than MLR models, AR models, and one-layer NN models, in both space and time. This is given that the LSTM model is properly regularized.

2.5. Conclusion and Research Aim

After analyzing the studies that have tried to simulate data over the GRACE/GRACE-FO TWSC data gap, a research gap was found. The most promising studies either use Swarm as a replacement for GRACE-like data over the data gap, or train ML models to find a relationship between hydro-climatological data and GRACE. These two strategies both hold their respective advantages and disadvantages.

Using hydro-climatological data as predictors for GRACE has yielded some promising results. It has been proven that soil moisture, temperature and precipitation variations over hydrologically active regions correlate well with GRACE TWSC. Another advantage of these LSMs is that they often hold a high spatial and temporal resolution. They do however not capture gravity signals directly but rather provide information on the water on the Earth's surface which is the main force behind these short-term gravity signals captured by GRACE. It has been proven that Swarm-derived TSWC is able to represent these gravity signals captured by GRACE, albeit with much lower spatial resolution. As a result, it may be advantageous to integrate these two data sets as GRACE TWSC predictors.

From the previous studies, it also became clear that data-driven ML models are very powerful when it comes to finding relationships between climate data and GRACE TWSC. Therefore it was decided to develop an ML model that would be trained to produce GRACE-like TWSC data by using a combination of hydro-climatological data and Swarm data as input data. "GRACE-like TWSC" means TWSC in terms of EWH with a spatial and temporal resolution similar to that of GRACE.

The Amazon basin was selected as the study region for this research. This is due to the fact that it is a big hydrologically active area with minimal human involvement. In addition, it has already been shown to be a very effective region for GRACE/GRACE-FO gap-filling experiments.

3

Data Selection & Pre-processing

This chapter introduces the different data sets that were used to train the neural networks, together with the GRACE & GRACE-FO products the models were trained to predict. The different pre-processing steps, that were taken to prepare the data for the training phase, are also presented.

3.1. Datasets

For this research, three different types of data were used. First of all, the GRACE & GRACE-FO data for which the data gap described must be bridged. As input to the models, both Swarm-derived TWS data and hydro-climatological data were used. The hydro-climatological variables that were considered consist of soil moisture, temperature, and precipitation.

3.1.1. GRACE & GRACE-FO

In March 2002, NASA and the German Aerospace Center (DLR) launched the Gravity Recovery and Climate Experiment (GRACE) mission. The goal of the twin-satellite mission was to provide monthly solutions of Earth's gravity field on a finer spatial resolution and with higher accuracy than previously possible. The satellites co-orbited each other at a distance of 200 km. They used low-low Satellite-to-Satellite tracking via K-Band Ranging System (KBR), measuring the inter-satellite ranges with very high accuracy (less than 10 μm)[17]. At lower frequencies, accelerometer errors and inaccuracies in the computed GRACE orbits are the major contributors to the GRACE error budget [13]. For higher frequency ranges (>14mHz), the contribution of ranging sensor errors becomes dominant [9]. The GRACE mission stopped its operations in October 2017, and the last gravity field model was on May 2017. Its successor, the GRACE-Follow On (FO) mission, started its operations in 2018. This results in a data gap of 11 months between June 2017 and April 2018, as illustrated by Figure 3.1 for the Amazon Basin, defined in Section 3.2.

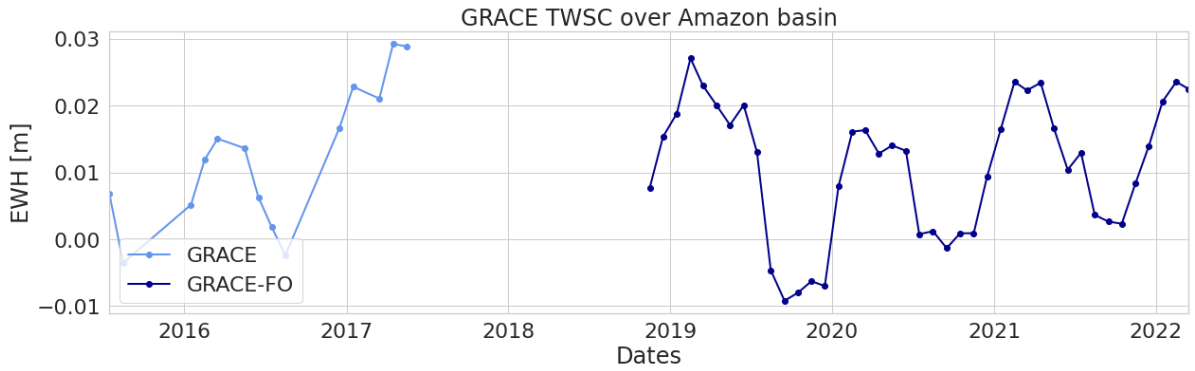


Figure 3.1: Time series plot of GRACE TSWC over the Amazon basin. Average grid cell value.

The GRACE-derived TWS data was obtained from the GRACE release 6 (RL06) produced at the Center for Space Research (CSR). A 750 km radius Gaussian smoothing was applied and the degree 1 coefficients were set to zero, ignoring the geo-centre motion. Lastly, the $C_{2,0}$ coefficient was replaced by a value from SLR[8]. An example of a monthly GRACE TWS map is shown in Figure 3.2.

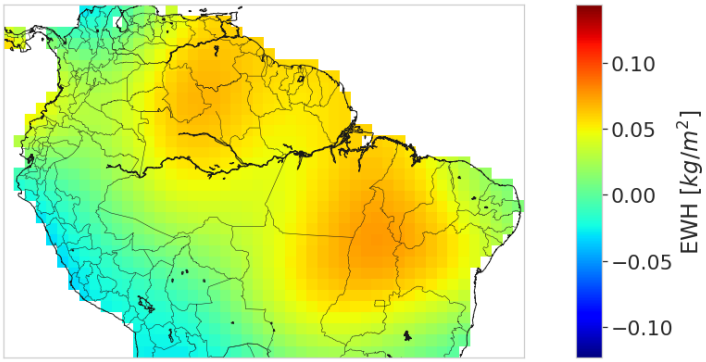


Figure 3.2: GRACE-derived TWS over the Amazon basin in terms of EWH, March 2022.

To ensure consistency with the Swarm and climate data, the monthly solutions for which GRACE data is available were interpolated to the same Epochs as the Swarm data, which is the middle of the calendar months.

Parametric GRACE MODEL

A second GRACE-derived EWH dataset was used in this research. This parametric GRACE model was obtained from Teixeira et al. (2020)[8], and is the result of a parametric regression of GRACE before smoothing. The model consists of 12 parameters: trend, bias, and five periods each represented by their respective sine and cosine components. For each SH coefficient, the 12 parameters were independently regressed linearly up to degree 40. Figure 3.3 shows how the parametric model compares to the real GRACE data.

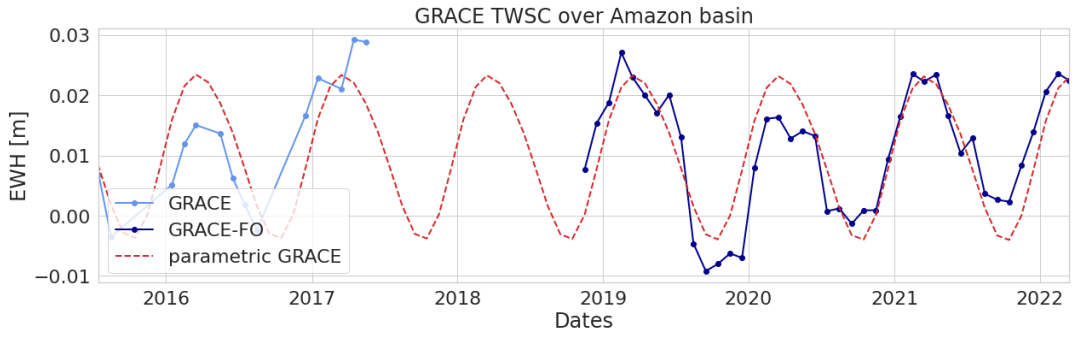


Figure 3.3: Time series of both GRACE TWSC and the parametric GRACE model. Average grid cell value.

The linear parametric GRACE model follows the GRACE seasonal variations relatively well. However, it fails to capture the local maxima/minima which are present in the GRACE TWSC time series. These are the sort of non-linearities that the models developed in this project hopefully catch. This parametric model is used for validation of the resulting time series over the gap, presented in Section 6.4.

3.1.2. Swarm

The Swarm-derived EWH maps were obtained from Teixera et al. [8]. They exhibit Swarm gravity field models that were created by combining four distinct solutions that were computed using various gravity field estimating techniques, as discussed in Section 2.2.3. It was shown that this combined solution outperforms any of the individual solutions. The maximum SH degree of the combined models

is 40. A Gaussian filter with a radius of 750 km was used over the land areas. The SH solutions where synthesized to 1°x 1°grids. Figure 3.4 shows an example map.

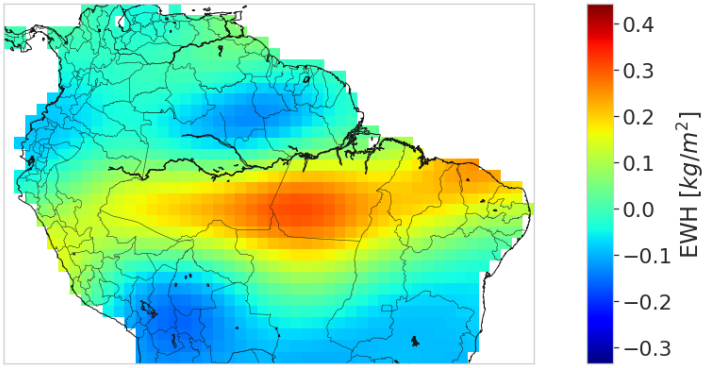


Figure 3.4: Swarm-derived TWS over the Amazon basin in terms of EWH, March 2022

Comparing the GRACE time series to the Swarm time series over the Amazon basin in Figure 3.5, it is clear that Swarm behaves much more erratic. This is a result of the larger uncertainties in the Swarm-derived EWH maps. However, the peaks and valleys in the Swarm time series agree with those from the GRACE time series, indicating a correlation. This is especially so for the GRACE-FO period.

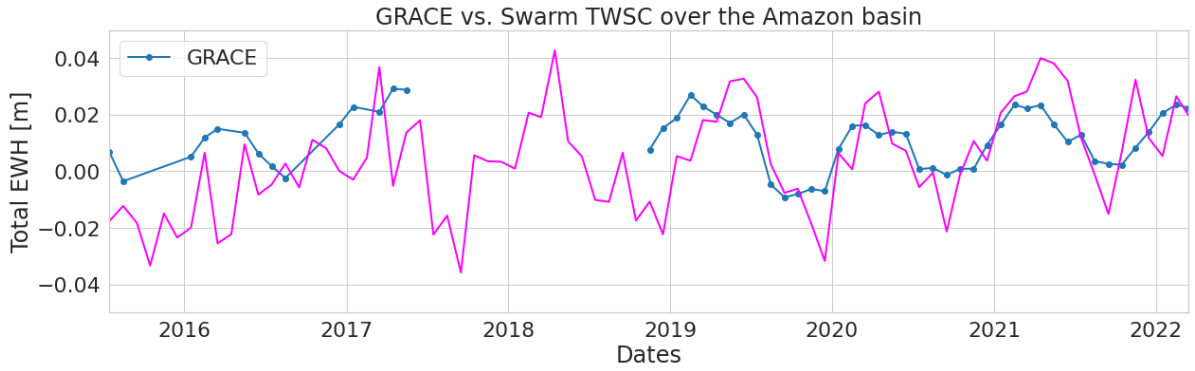


Figure 3.5: Average GRACE-derived TWSC compared to the Swarm-derived TWSC over the Amazon basin in terms of EWH.

3.1.3. Climate Data & Hydrological Models

With advanced land surface modelling and data assimilation techniques, the Global Land Data Assimilation System (GLDAS) ingests satellite and ground-based observational data products to produce the best possible fields of land surface states and fluxes [26]. The high-quality, worldwide land surface fields offered by GLDAS assist a number of applications for water resources, water cycle research, and present and proposed weather and climate prediction. The project produced an archive of modelled and observed global surface meteorological data, parameter maps, and output, including simulations of the Noah land surface model (LSM) with 1-degree and 0.25-degree resolutions from 1948 to the present. From this NOAH land surface model (Version 2), the climate & hydrological data sets were obtained from the Goddard Earth Sciences Data and Information Center[1]. For the purpose of this research, the 1°x 1°maps were used, which are consistent with the GRACE EWH maps.

Soil Moisture

The soil moisture data consists of monthly averaged daily soil moisture data between the depths of 0 cm and 200 cm, with units of kg/m^2. The data is averaged around the middle of the month. This data provides information on the amount of water that is stored under the surface of the Earth, which drives variations in the local gravity signal. How this soil moisture data influences the TWSC derived from

[1]Accessible via `https://disc.sci.gsfc.nasa.gov/`

GRACE, becomes clear from Equation 3.1.

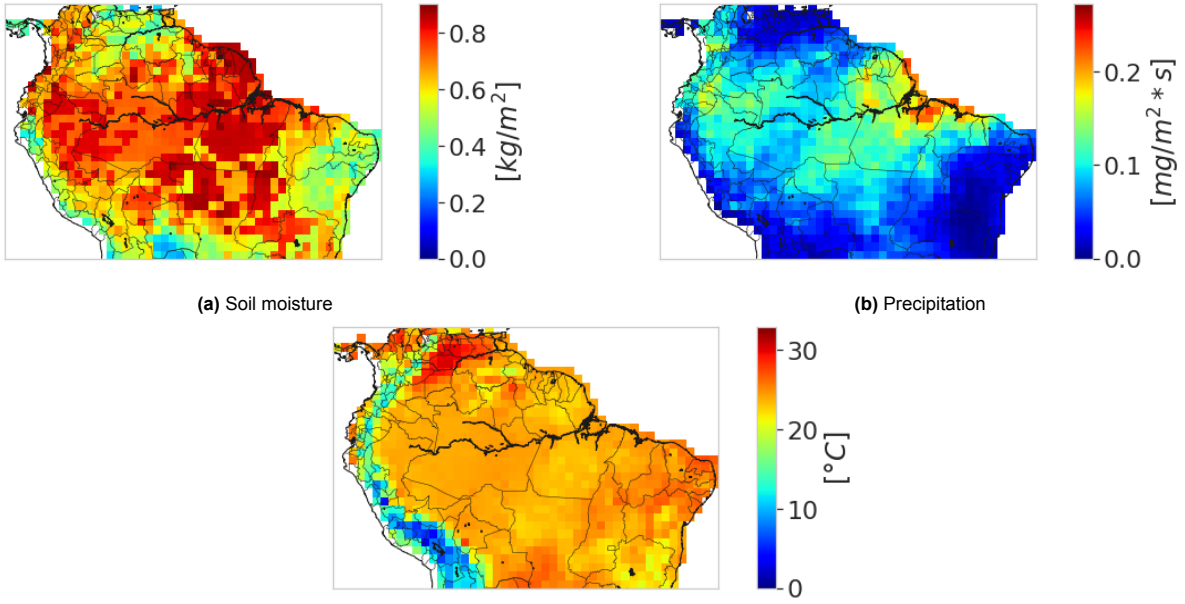
$$TWS = SM + SWS + CWS \tag{3.1}$$

SWS stands for snow water storage, CWS stands for canopy water storage. It is clear that in the absence of snow, soil moisture (SM) plays a large role in the total TWS in the Amazon basin

Precipitation
The rainfall data consists of monthly, 3-hourly averaged precipitation rate values, averaged around the middle of the month. This data is provided in kg/m^2s, but is converted to g/m^2s. Humphrey et al. (2017) suggested that at least 40% of the total variance of GRACE anomalies can be reconstructed from precipitation and temperature variability alone. Thus, in this study, precipitation and temperature are explored as additional predictors to help improve the model's performance.

Temperature
The NOAH LSM also provides monthly averaged temperature maps. These were also averaged around the middle of the month.

Figure 3.6 shows examples of these climate & hydrological maps over the Amazon Basin, for March 2022.



(a) Soil moisture

(b) Precipitation

(c) Temperature

Figure 3.6: Climate & hydrological data maps over the Amazon Basin, March 2022

3.1.4. Data Availability

The GLDAS-2.1 NOAH dataset provides hydro-climatological data starting from January 2000, the first GRACE gravity maps were produced in April 2002, and the first Swarm-derived GFMs are from December 2013. This means that the availability of Swarm data is a limiting factor for this research. The data availability is depicted in Figure 3.7.

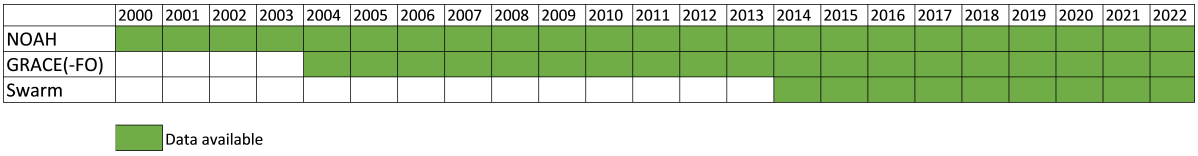


Figure 3.7: Availability of the various datasets used in the thesis.

Another challenge, in terms of data availability, is the quality of some of the GRACE months. Due to battery issues and other technical malfunctions, some of the months during the GRACE mission could not be used for training the model. Figure 3.8 zooms in on the operational period of Swarm and shows in which specific months the GRACE/GRACE-FO data is not available.

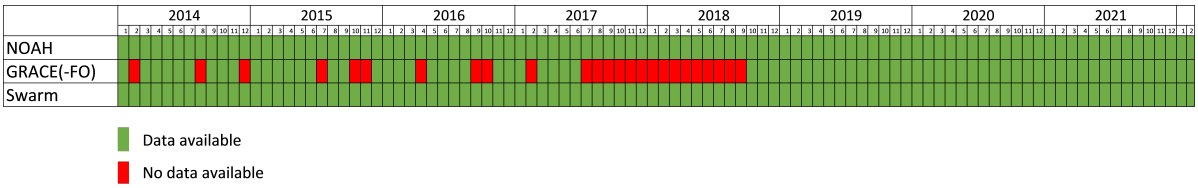


Figure 3.8: Availability of the data, limited by the Swarm data (starting in December 2013).

Combining the Swarm availability and the GRACE availability, 70 months for which hydro-climatological, Swarm and GRACE/GRACE-FO data are available were used for training of the NNs. This limited data availability was one of the main challenges for this project.

3.2. Pre-processing

Some steps were taken in order to prepare the data for the training of the NNs. These pre-processing steps, and their motivation, are discussed below.

Masking

The first step in preparing the input data before training the models is to filter out the region of interest. Therefore the boundaries of the Amazon Basin, as it will be used in this research, had to be defined. For this study, the Amazon basin was defined as the land region between 20°S-12N and 82°W-34W. This range was taken from the rectangular regions proposed by Giorgi and Francisco (2000)[15], and is visualized in Figure 3.9.

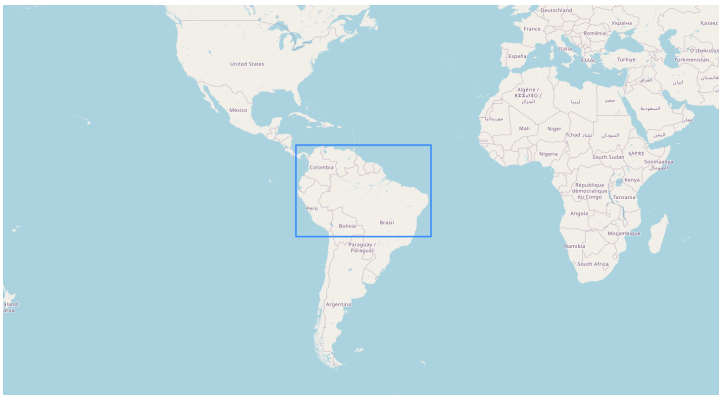


Figure 3.9: The Amazon Basin region, as defined by Giorgi and Francisco

All of the input data sets consist of 1°grids so this translates to maps of 33 x 49 grids consisting of 1617 grid cells.

Normalization

Different data-normalizing strategies exist. In order to reduce bias inside the neural network from one feature to another, it can be utilized to scale the data in the same range of values for each input feature. By starting the training process for each feature on the same scale, data normalization can help save training time. In modelling applications where the inputs are frequently on drastically different scales, it is very helpful. Two different normalization strategies were considered for the input features in this research [16].

The first strategy is called **min-max normalization**, where the samples are scaled between 0 and 1 Equation 3.2.

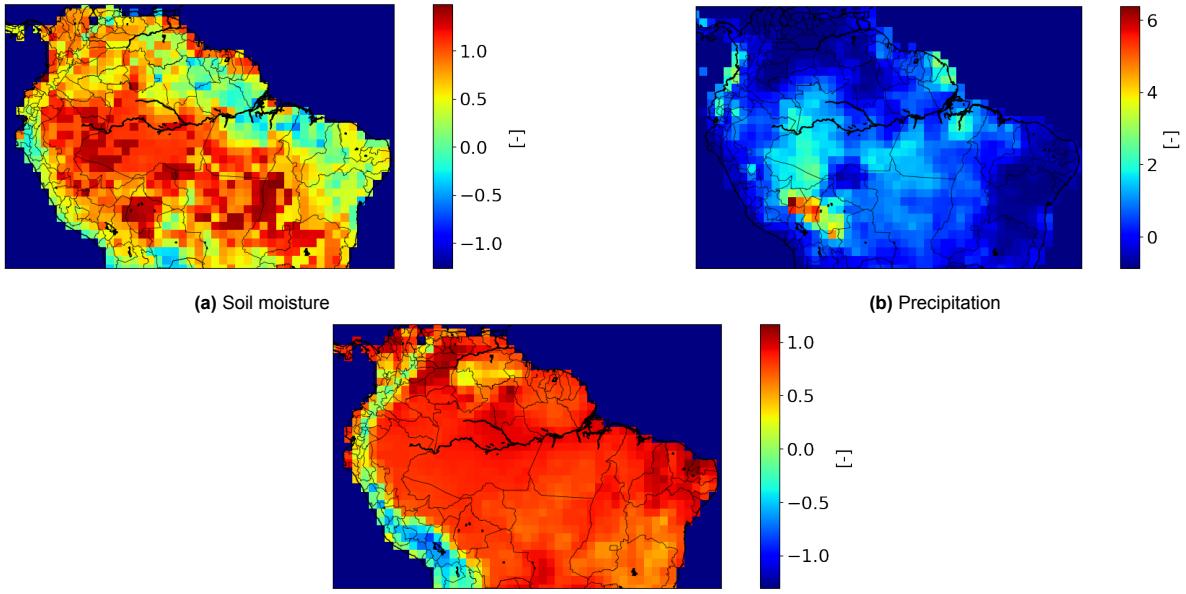
$$x' = \frac{(x - x_{\text{min}})}{(x_{\text{max}} - x_{\text{min}})} \tag{3.2}$$

Whenever $(x_{\text{max}} - x_{\text{min}}) = 0$, this indicates that the feature consists of a single value and it should be removed. The main advantage of the min-max normalization technique is that the relative relationships within the data set is preserved. However, because all values are being scaled, potential outliers may negatively influence the distribution of the normalized data set.

This effect of outliers is reduced in the second normalization strategy, called **statistical (or Z-score) normalization**. This strategy requires the mean and standard deviation for each separate input data set and then normalizes the data set by use of Equation 3.3.

$$x' = \frac{(x_i - \mu_i)}{\sigma_i} \tag{3.3}$$

The resulting data sets will each have a mean of 0, and a standard deviation of 1. This strategy was selected for the purpose of this research, as it is more robust to outliers in the data sets. Figure 3.10 shows the effect of this normalization strategy on the input data.



(a) Soil moisture

(b) Precipitation

(c) Temperature

Figure 3.10: Normalized climate & hydrological data maps over the Amazon Basin, January 2014

When compared to Figure 3.6, it is clear that the normalized values lie much closer together, while preserving most of the spatial features of the data.

Training and testing months

The next step is to divide the input data sets into training and testing months. The training months will be used to train the neural networks, while the testing months will be used to validate the performance of the model. In this way, unknown data is used to test the model, which provides an objective assessment of how well the model would perform when it is used to make predictions for new data. A rule of thumb for ANNs is that 70% of the data is used for training, while the remaining 30% is used for validation[3]. The more detailed validation process is discussed further in Chapter 5.

4

Machine Learning Models

Machine learning is a type of artificial intelligence with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it to learn complex relations between input data and target data for themselves. Artificial Neural Networks (ANN) are a type of machine learning that consists of computer systems that are designed to simulate the workings of the human brain. The main advantage of using such an ANN for this research is their ability to learn complex nonlinear relationships between datasets such as climate data and GRACE-like EWH. The following chapter starts by introducing the simplest form of ANNs, together with the basics of machine learning. Section 4.1 also presents the first model used in this research. Section 4.2 discusses convolutional neural networks (CNNs), and how they differ from the simpler ANNs. This section also shows the final model that was used for this research.

4.1. Multilayer Perceptron Neural Network

The first machine learning model consists of a fully connected Multilayer Perceptron Neural Network (MLP) neural network. This is the most basic form of an artificial neural network, consisting of an input layer, an output layer, and fully connected hidden layers in between. Each layer consists of a fixed number of nodes. A general architecture of an MLP NN is shown in Figure 4.1.

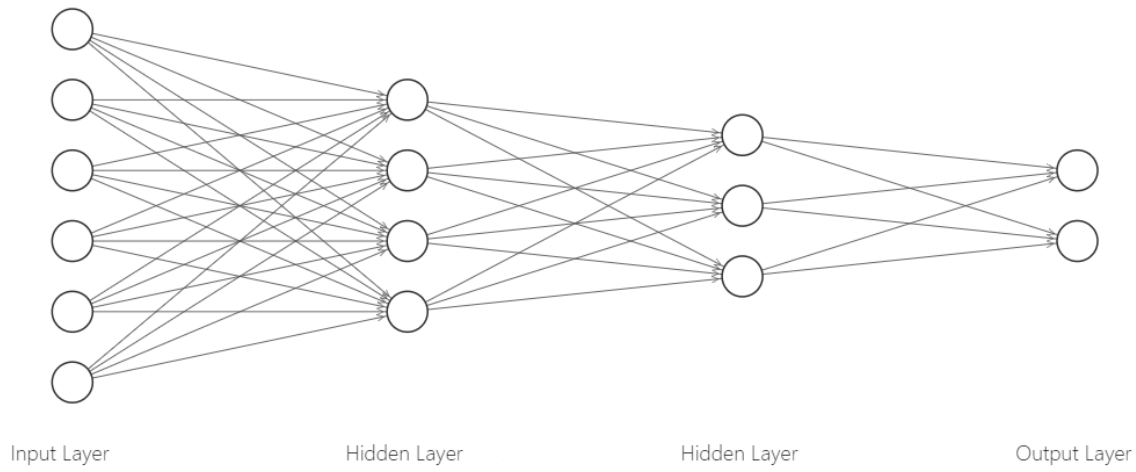


Figure 4.1: General architecture of an MLP network, consisting of an input layer, two hidden layers, and an output layer.

In general, NNs aim to mathematically represent the inner workings of the biological brain. The human brain is made up of a vast network of neurons, and these neurons (or nodes) serve as the fundamental building blocks for NNs. They are a combination of inputs, outputs, weights, biases, and nonlinear transfer functions or activation functions. The task of these neurons is to detect the nonlinearities in the

training data set. The network's ability to map these nonlinearities in the data is enabled by activation functions. Figure 4.2 depicts the activity of one neuron (left) and the link between neurons in the neural network (right).

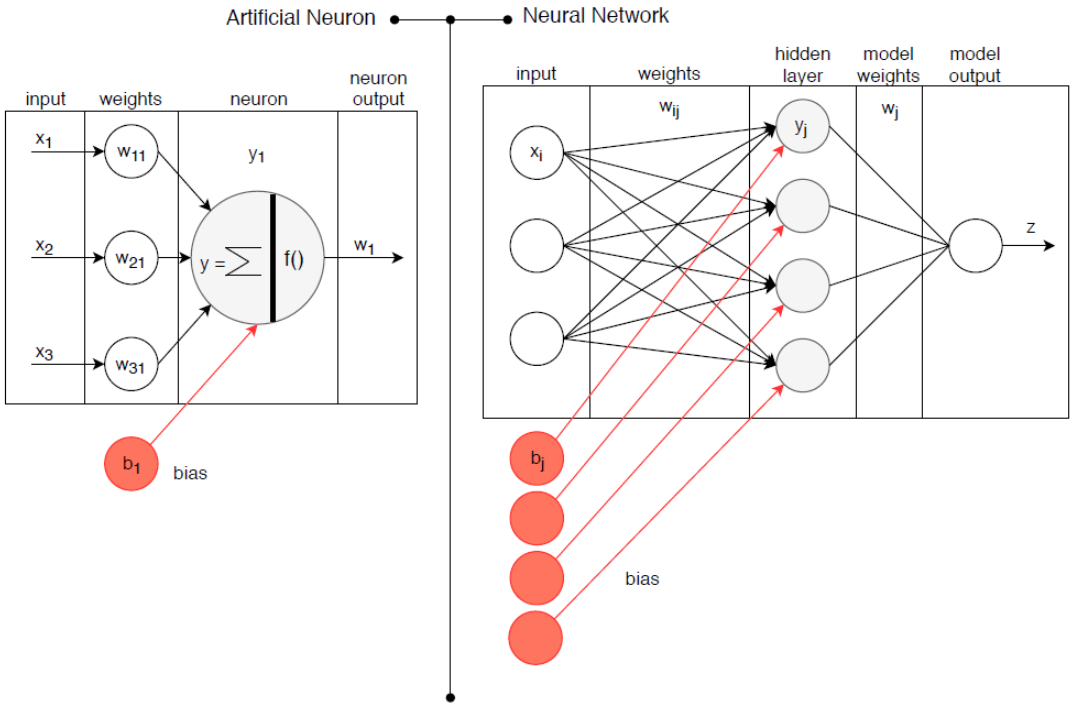


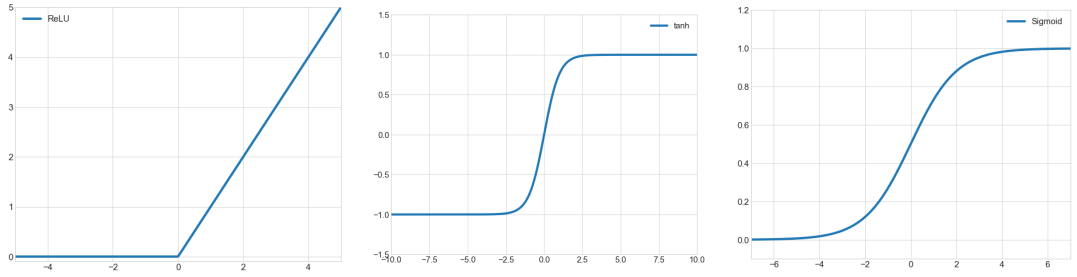
Figure 4.2: Diagram of a single neuron (left) and an ANN with a single hidden layer (right)[6].

Increasing the number of hidden layers in a model, together with the number of nodes in these layers, increases the ability of the model to learn complex relationships in the training data. A trade-off must be made between simplicity and complexity in the model architecture. A simple model will need less computational effort to train, but might not be able to learn these complex relationships in the data. Conversely, a very complex model might not be able to find general relationships as it starts to overfit the training data.

The mathematical expression of what happens at a single node when data flows through the model is expressed by Equation 4.1.

$$y_j(t) = \psi\left(\sum_i w_{ij}x_i(t) + b_j\right) \tag{4.1}$$

The output of a node y_j is determined by the sum of the inputs to the node x_i multiplied by their according *weights* w_{ij} and the *bias* of the node. The bias b_j is a constant which is added to this feature-weight product and shifts the scalar resulting from Equation 4.1). This scalar is then fed to an activation function ψ, introducing nonlinearity to the model. Examples of the most popular activation functions used in ANNs are ReLU, tanh, Sigmoid. These are depicted in Figure 4.3.



(a) ReLU(x) **(b)** tanh(x) **(c)** sigmoid(x)

Figure 4.3: Popular activation functions

The input maps flow through the layers of the NN and reach the output layer where they are transformed into output maps. The mismatch between the model output and the values we want to predict is called the loss and is determined by the selected *loss function*. For this research, the Mean Squared Error (MSE) between the model output maps and the observed GRACE EWH maps is defined as the loss. The loss function $J\left(w^T, b\right)$ describes how the individual weights and biases of the network influence the overall loss and is presented by Equation 4.2.

$$J\left(w, b\right) = \frac{1}{m}\sum_{i=1}^{m} L\left(\hat{y}^{(i)}, y^{(i)}\right), \text{ where} \tag{4.2}$$

$$L\left(\hat{y}^{(i)}, y^{(i)}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - \hat{y}^{(i)}\right)^2 \tag{4.3}$$

In Equation 4.2, m represents the number of training data sets, whereas in Equation 4.3, n represents the number of grid cells in the output maps, and y and $\hat{y}$ represent the NN output and the target GRACE EWH maps respectively.

The training of the model is done by the use of *backpropagation*. Backpropagation is the procedure in which the weights and biases of the model are repeatedly adjusted with the aim of decreasing the loss. This could also be seen as minimising the loss function by way of gradient descent. This is the task of the *optimizer*, who updates the model's weights and biases in the opposite direction of the gradient of the cost function's relation to those parameters, minimizing the cost function J. The step size with which the optimizer updates the model is called the *learning rate* (LR) and it is a very important hyperparameter of the NN. A hyperparameter is a parameter in machine learning whose value is used to regulate the learning process, while other parameters, such as weights and biases, are often determined by training. A model can either use a fixed learning rate or a varying learning rate. With a fixed learning rate, the step size of the optimizer is predefined and will stay constant throughout the training of the NN. In some cases, a decreasing optimizer step size will be able to find better minima of the loss function and thus increase the performance of the NN. There exist various ways of decreasing the learning rate. In this project, three learning rate decreasing strategies were considered:

- **Plateau strategy**: decrease LR after model's accuracy stagnates
- **Multiplicative strategy**: decrease LR after every training epoch
- **Step strategy**: Decrease LR after n training epochs

How these strategies influence the performance of the NN, is presented in Chapter 6.

The optimizer most frequently used in ANNs is called the Adaptive Moment Estimation (Adam) optimizer. It is demonstrated via empirical data that Adam is a memory- and computationally-efficient stochastic optimization technique. The method has also been demonstrated to perform well in practice and perform favourably when compared to other gradient descent algorithms with adjustable learning rates [18]. The other two optimizers that were considered in this project were the Stochastic Gradient Descent (SGD) optimizer and the Root Mean Squared propagation (RMSprop) optimizer. These are two popular optimizers because they frequently update the NN parameters and require relatively little

memory.

This training process is repeated for each epoch of the training phase. In machine learning, an epoch refers to the full transit of training data through the algorithm. The number of epochs is a hyperparameter that controls the training process for the machine learning model. The NN training is stopped once the desired performance is obtained or the predefined max number of epochs is reached. The desired performance can be expressed in terms of loss or accuracy between the NN output and the targets. For this thesis, the number of epochs controlled the length of the training phase, as there is no prior desired performance.

4.1.1. Final Architecture

The optimal architecture of the MLP model used for this research was determined through trial and error. The number of nodes in the input and output layer is determined by the number of input grid cells and output grid cells respectively. This translates to 1617 nodes for the output layer (1 GRACE-like EWH map) and Nx1617 nodes for the input layer, with N being the number of input maps. The only way in which architectures could then differ from one another is by the number of hidden layers and their respective sizes.

The predictive accuracy of the MLP model increased with increasing complexity (e.g. more layers and larger layers) up to 3 hidden layers. After this, the model started to overfit the training data without increasing the performance any further. Higher complexity results in higher computational effort. Therefore the number of hidden layers was set to three. The number of nodes per hidden layer was chosen to be a multiple of 1617, such as the input and output layer, and was selected so that the number of nodes gradually decreases from Nx1617 to 1x1617. The number of nodes per considered architecture, dependent on the input data size, is presented in Table 4.1.

Layer	Number of nodes			
	N=1	N=2	N=3	N=4
input	1617	1617*2	1617*3	1617*4
hidden 1	1617	1617*2	1617*3	1617*4
hidden 2	1617	1617	1617*2	1617*3
hidden 3	1617	1617	1617	1617*2
output	1617	1617	1617	1617

Table 4.1: The number of nodes and layers per considered ANN architecture.

An example of such an ANN layout can be seen in Figure 4.4, were the size of the input layer is equal to 1617x4 due to the four input maps.

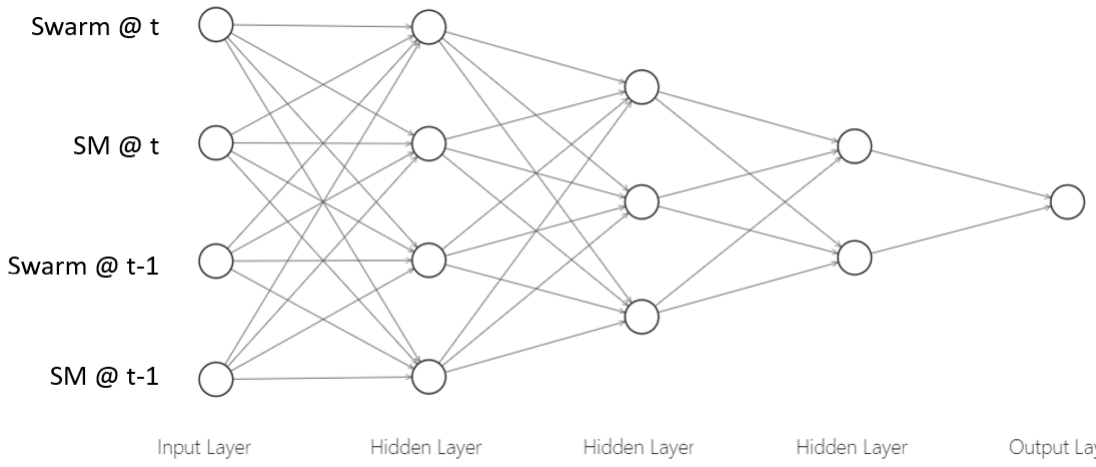


Figure 4.4: The MLP architecture used in this research. Each node in the figure represents 1617 nodes in the model.

Once this architecture is set, the hyperparameter tuning can be performed. Hyperparameter tuning is the process in which the optimal values of the hyperparameters, yielding the highest NN performance, are obtained by various optimization techniques. This is presented in Chapter 5.

4.1.2. Transfer Learning

One of the key challenges for this project was the scarcity of data. Despite the fact that GRACE and hydro-climatological data have been accessible since 2002, only data starting from 2014 was utilized to train the models. This is because Swarm data is only available after the end of 2013. Transfer learning was considered as a way of using this hydro-climatological and GRACE data in the period between 2002 and 2014. With transfer learning, a first NN is trained in producing GRACE-like EWH maps based on climate data from the pre-Swarm period. This pre-trained model is then used in a second training phase to produce similar output maps but this time trained with climate data and Swarm data from the Swarm period as input to the model. So instead of initializing a NN with random weights and biases (which is the case for the normal training), the initial weights and biases come from a model which is pre-trained using only climate data from before 2014 as input. This was done by simply copying the pre-trained model and appending the Swarm data to the input layer, doubling its size.

4.2. Convolutional Neural Network - CNN

The human visual system served as an inspiration for CNN's design, which aimed to extract fine details from inputs. In order to project an input image (or stack of images) onto a hierarchical set of feature maps, which can be viewed as nonlinear transformations of the input, CNN employs discrete convolution processes as its name implies. In order to extract spatial information (such as edges and corners) from each layer's input, the CNN deep learning model architecture uses convolutional hidden layers, as opposed to the fully connected hidden layers in the MLP model. As a result, CNN models are by nature well suited to learning multi-scale spatial patterns from several sources of gridded input, which is a difficult problem to address with MLP models that do not scale well on images. In this thesis, the input maps can be seen as images where the grid cells function as image pixels. Another advantage of using a CNN is that multiple maps at different time steps can be stacked into a 3D input layer, such that the NN does not only extract spatial but also temporal features. This is a concept that will be exploited in Chapter 5.

4.2.1. General Concepts

Before introducing the popular CNN architectures together with the model that was used for this research, it is important to realize how a CNN differentiates itself from other NNs such as an MLP.

Convolutional layers

The convolutional layer is the most significant part of the CNN architecture. It is made up of a number of convolutional filters (or kernels). The output feature map is produced by convolving the input image with these filters. Such a convolution step is depicted in Figure 4.5. The filter consists of a grid of discrete weights, shown in the 2x2 green matrices. As the filter scans over the image, it takes the dot product of the image values and the filter weights and creates an output feature map. The step size with which the filter scans the input maps is called the stride. To ensure that no information at the edges of the maps is lost, padding may be applied. In this research, zero padding was used by adding rows and columns of 0s to the input maps in such a way that the dimensions of the maps remain constant throughout the convolutional layers of the model. Other padding techniques might be used as well, but zero padding is the most common practice.

The filters in a convolutional layer each hold spatial features and by scanning through the input map, output maps are created that hold information on how the input maps respond to each of the convolutional filters. This is the idea behind these convolutional layers. Compared to the MLP model, where the NN did not respect the spatial structure of the input maps, this is a major advantage when it comes to image recognition tasks.

Pooling layers

Pooling aims at reducing large-scale feature maps to produce smaller feature maps in order to speed up the training process. At the same time, it keeps most of the dominating data (or characteristics)

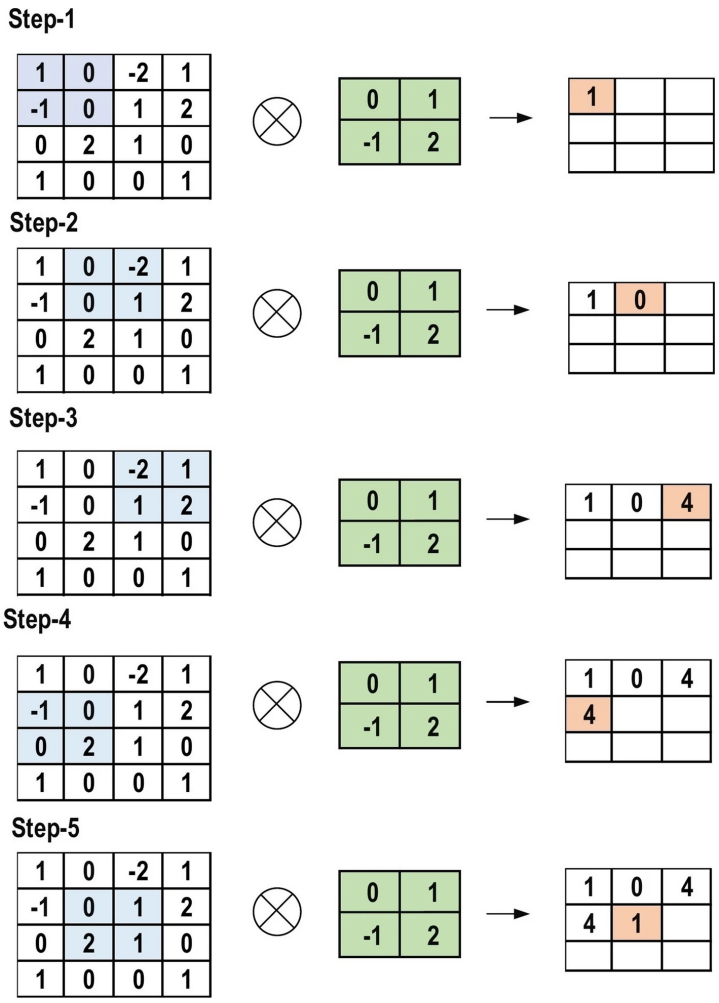


Figure 4.5: Different steps in a convolution, with the primary calculations of each step[3]. The input feature map is of dimension 4x4, the filter is of dimension 2x2 and the resulting feature map is of dimension 3x3.

during the entire pooling stage. Multiple pooling strategies are used in machine learning. The most used pooling techniques in CNNs are max pooling, average pooling and global average pooling. These three concepts are depicted in Figure 4.6.

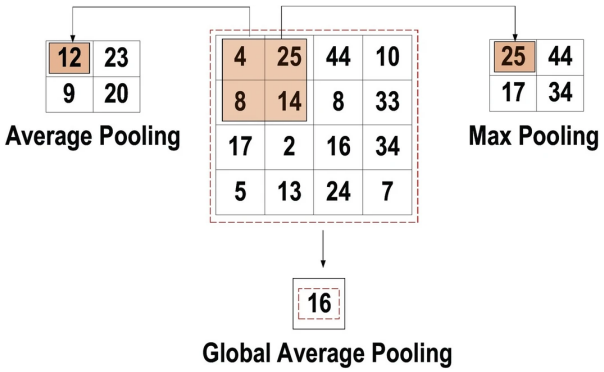


Figure 4.6: Max pooling, average pooling, global average pooling[3]

Dropout layers
A popular regularization technique in CNNs is the use of a dropout layer. This is a technique that comes back in almost all of the popular CNN architectures and aims at avoiding over-fitting by the often complex CNN networks. Adding a dropout layer to the CNN used in this research was therefore considered. With a dropout layer, a certain number of layer outputs are ”dropped out” or disregarded at random during training. This results in the layer appearing to have a different number of nodes and connections to the preceding layer for every training loop. Dropout causes the training process to become noisy, pushing the nodes within the model to find more general relationships between the input and output layers. The percentage of nodes that are being ”dropped” in a certain layer is manually set. For this project, the dropout percentages were set to 0, 25, 50, and 75 respectively.

4.2.2. Popular Architectures

The architecture of the CNN used in this research was based on previously developed CNN models. This subsection will provide an introduction to these models, AlexNet and VGG-16.

AlexNet
The first CNN architecture that was considered is the AlexNet model, published in 2012 [19]. It consists of 5 convolutional layers, with filter sizes of 11, 5, and 3 (x3) respectively. After each convolutional layer, a max pooling layer is added in order to decrease the size of the data flowing through the model. After the convolutional layers, 3 fully connected layers are added to the model. Figure 4.7 shows a simple representation of the AlexNet architecture.

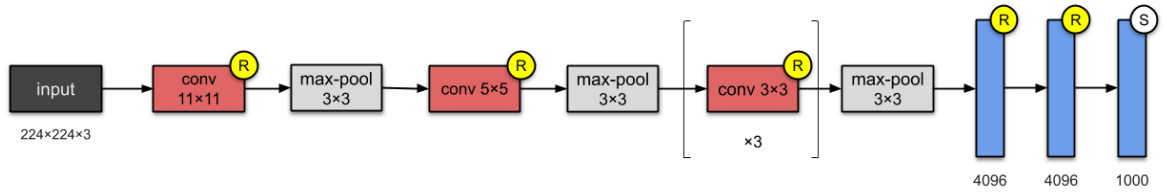


Figure 4.7: AlexNet Architecture[1]

Instead of the tanh function, which was the industry standard at the time, AlexNet employs Rectified Linear Units (ReLU), shown in **??** as activation function. The benefit of ReLU is that is decreases the training time by a factor of six when compared to models with tanh activation functions. Another

[1] https://towardsdatascience.com/illustrated-10-cnn-architectures-95d78ace614d

innovation at the time was the use of multiple GPUs for training, which decreased the training time even further. AlexNet took first place in the 2012 ImageNet competition[2].

VGG-16

The second CNN architecture that was considered, VGG-16, won the same ImageNet competition in 2015. The idea behind the VGG models was to implement a set of CNN design guidelines. Before VGG, the popular CNN architectures were all specifically found by trial and error, making it hard to play with and/or upscale. But with VGG some simple design principles were implemented [27]. Whereas AlexNet has 5 convolutional layers, VGG models have 5 convolutional stages. Each stage consists of multiple consecutive convolutional layers and pooling layers. It uses only 3x3 convolutional layers with a stride of 1, all max pool layers must be 2x2 with a stride of 2, and after each max pooling layer, the number of channels must be doubled. As for the AlexNet, the model was completed by a number of fully connected layers at the end. Figure 4.8 shows the simplified architecture of the VGG-16 model.

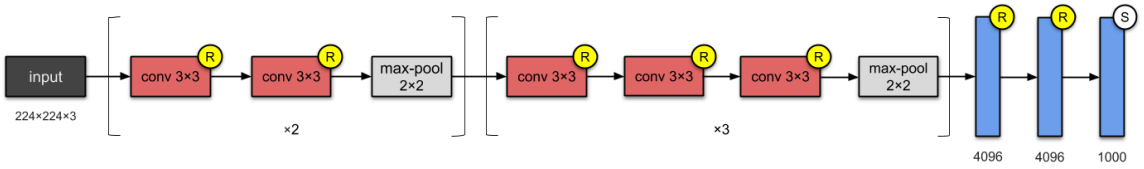
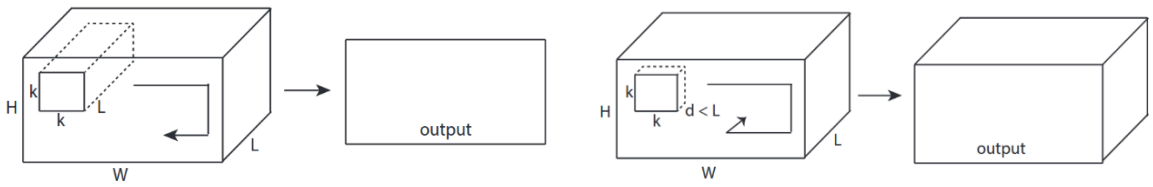


Figure 4.8: VGG16 Architecture[3]

The idea behind only using 3x3 convolutional layers arises from the fact that stacking 2 3x3 layers has the same effect as using 1 5x5 layer, while using computational effort. The second benefit of stacking 3x3 layers is the addition of extra ReLu activations in between these layers, allowing for more non-linearity.

4.2.3. 2D CNN vs. 3D CNN

Two different types of CNN architecture were considered for this project. A conventional CNN takes multiple channels of 2D maps as input, and a 3D CNN takes one single channel consisting of a 3D map as input. The 2D CNN typically takes as input multiple 2D maps, where each map represents an input feature. Each map is processed by a separate convolutional layer, and the results are combined into a single feature map. In contrast, the 3D CNN takes as input a single 3D map. The volume represents a sequence of 2D maps stacked onto each other. The convolutional layers in a 3D CNN operate on 3D kernels. The convolution that takes place after the input layer is visualised in Figure 4.9 for both cases.



(a) 2D CNN with 3D input layer.

(b) 3D CNN with 3D input layer

Figure 4.9: Simplified visualisation of a 2D and 3D convolutional operation, both with a 3D input layer. H and W represent the height and width of the input map respectively, k stands for the kernel size, and L stands for the number of input channels (a) or the depth of the input maps (b).

In Figure 4.9a, the input layer consists of multiple channels (L) of 2D maps. The 2D kernel scans these channels separately and combines the output into a 2D map. This map then flows through the CNN convolutional layers and the fully connected layers. For a 3D CNN, depicted in Figure 4.9b, the input maps are stacked into a 3D map, using one single channel. The 3D CNN uses 3D kernels that scan the 3D input map and reproduce 3D output maps. This is repeated at every convolutional layer.

[2] https://image-net.org/challenges/LSVRC/

[3] https://towardsdatascience.com/illustrated-10-cnn-architectures-95d78ace614d

4.2.4. Final Architecture

The conventional NNs used for image recognition are mostly models that take a 2D image as input and perform 2D convolutional operations in order to find features in this image. In this research, the input data does not consist of a single map for each month, but multiple maps consisting of their own type of data. In Section 4.1, it was shown that for the ANN this was solved by converting these maps to 1D arrays and concatenating them together to one single 1D input array. With CNNs it is possible to stack these 2D maps into a 3D array and perform 3D convolutional operations to find features across these maps. Another application of stacking is that the model can also search for temporal features by stacking maps from different time steps. Designing such a 3D CNN is a challenge because there are much fewer conventional 3D CNN architectures available when compared to 2D CNNs. Therefore it was chosen to design a 3D CNN from scratch while implementing principles from the AlexNet and VGG-16 models. The models also consist of an input layer followed by convolutional layers and fully connected layers at the end, as shown in Figure 4.10.

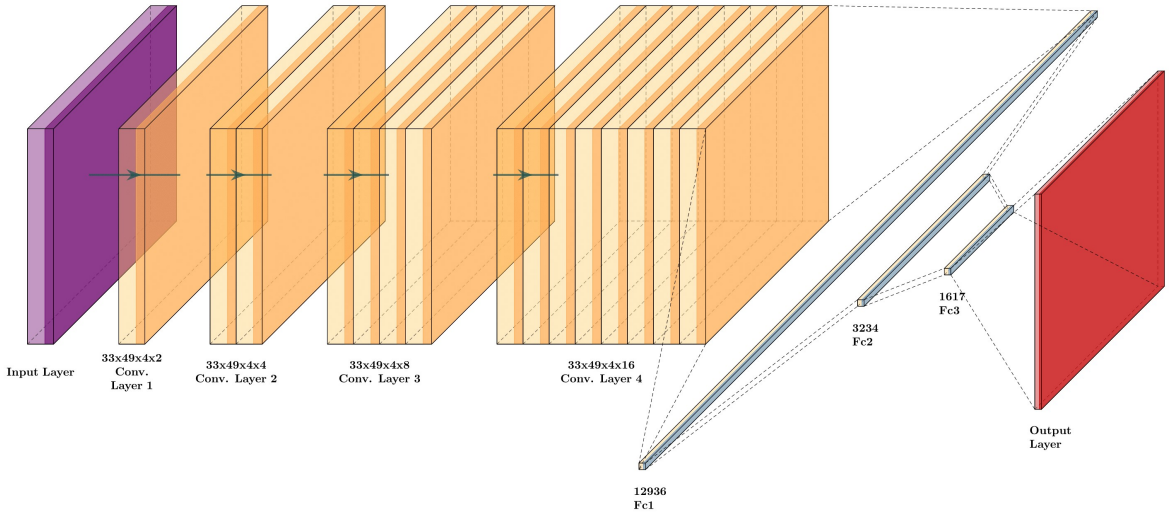


Figure 4.10: The 3D CNN model used for this research

The filter size of the convolutional layers was set to 3x3 as in the VGG-16 model. Next to this, the NN also uses ReLU activation functions on each layer of the model. The main difference from the previously discussed networks is the non-implementation of pooling layers. The main goal of pooling is to decrease the dimensions of your data to allow for faster training of the network. For this research, the training data set is quite limited and thus training time falls within the boundary of one hour. Therefore it was decided to not implement pooling as this might create the risk of losing small spatiotemporal features in the data.

4.2.5. Stacking the input maps

One of the main motivations behind the choice of a CNN for this project is the fact that the input maps can be stacked in a different way compared to the ANN. As described above, stacking the input maps for the ANN simply consisted of doubling the size of the 1D input array. The fact that we have 2D or 3D arrays as input to the CNN changes the stacking process in ways that might benefit the performance of the model. For the 2D CNN, stacking is performed by increasing the number of input channels. If we stack two months, each consisting of a Swarm and an SM map, on top of each other the number of channels increases from 2 to 4. The input maps thus still consist of multiple input channels with each a 2D array as input. The 3D input layer of the 3D CNN allows stacking to take place in a more straightforward way. Stacking for the 3D CNN was performed by stacking the extra maps onto the previous maps, increasing the size of the input layer. Table 4.2 provides an overview of the different input layer dimensions used for the 2D and 3D CNNs, for both stacking and no stacking.

	Number of input channels (N)	Size per input channel (WxLxD)	Total size of input layer (NxWxLxD)
2D CNN	2	33 x 49 x 0	2 x 33 x 49 x 0
3D CNN	1	33 x 49 x 2	1 x 33 x 49 x 2
2D CNN - stacked	4	33 x 49 x 0	4 x 33 x 49 x 0
3D CNN - stacked	1	33 x 49 x 4	1 x 33 x 49 x 4

Table 4.2: Overview of the input layer dimensions for the 2D and 2D CNNs, for both stacking and no stacking. Assuming 2 diferent input maps.

4.3. Neural Network Jacobian

Before the training of the ML models is discussed, one more concept needs to be introduced. One important application of neural networks is sensitivity analysis, which is the study of how changes in the input to a system affect its output. The neural network Jacobian is a tool used in a sensitivity analysis that measures the rate of change of the output of a neural network with respect to its inputs. To check which grid cells in the input maps have the largest influence on the output maps, the Jacobian matrix of the NN was computed. In short, all first-order partial derivatives of a multivariate function are collected in the Jacobian matrix. For a NN with input layer **x** and output layer **f(x)**, the Jacobian matrix is described by Equation 4.4.

$$\mathbf{J} = \frac{d\mathbf{f}(\mathbf{x})}{d\mathbf{x}} = \left[\frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_1} \cdots \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_u}\right] = \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\mathbf{x})}{\partial x_u} \\ \vdots & & \vdots \\ \frac{\partial f_v(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_v(\mathbf{x})}{\partial x_u} \end{bmatrix} \tag{4.4}$$

The Jacobian matrix holds the first-order partial derivatives of the output node values to the input node values. This matrix thus describes how local perturbations to the neural network input affect the output. By averaging these derivatives for each input node and plotting these values, the driving regions for both SM and Swarm over the study region can be visualized.

5

Optimization of Machine Learning Models

Once both of the ML architectures were defined, they still had to be optimized. This was done by finding the optimal combination of input data, tuning the hyperparameters of the models, and checking whether the performance is influenced by different strategies such as stacking input months or implementing transfer learning. Before the optimization could take place, the performance indicators were defined. This is described in Section 5.1.

5.1. Evaluation Metrics

The goal of the NN is to predict the GRACE-like EWH map of the Amazon basin for a given month. To check the quality of these predictions, evaluation metrics had to be defined. These metrics indicate the spatial and temporal accuracy of the NN's output maps.

NSE

The evaluation metric that represents the spatial accuracy of the predicted map is the Nash-Sutcliffe Efficiency (NSE) index. The NSE, represented by Equation 5.1, takes into account the observed values (o_i), the mean of the observed values ($\bar{o}$), and the predicted values (y_i) for all n grid cells of a single map.

$$\text{NSE} = 1 - \frac{\sum_{i=1}^{n} (y_i - o_i)^2}{\sum_{i=1}^{n} (o_i - \bar{o})^2} \tag{5.1}$$

The nominator in Equation 5.1 represents the sum-squared in the difference between the predicted and the observed, or GRACE, values. The denominator represents the variance in the observed time series. The possible values of NSE range between $-$ inf and 1. A predicted EWH map that completely matches the GRACE observed map has an NSE of 1. When the NSE is equal to 0, the predictions are as accurate as setting all grid values equal to the mean of the map.

The mean NSE of the testing months was used as the primary indicator of the model's performance to find general relationships between input maps and GRACE maps. It was calculated after every training epoch and by visualizing it, it can be determined after what number of epochs the model starts to overfit. Overfitting takes place when the NN starts to remember the training targets and starts to reproduce them instead of finding general relations between the input and target data. The training accuracy becomes very high but the validation accuracy starts to drop.

CC

The Pearson correlation, also known simply as the correlation coefficient, calculates the linear connection between two continuously varying signals and displays it as a number between -1 (anti-correlated)

and 1 (perfectly correlated). The Pearson correlation coefficient (CC) was calculated by using Equation 5.2

$$\mathrm{CC} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(o_i - \bar{o})}{\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2 \sum_{i=1}^{n} (o_i - \bar{o})^2}} \tag{5.2}$$

In this research, the CC was used to quantify the temporal match between the predicted and the observed total EWH time series over the Amazon basin. Therefore y_i and o_i represent the predicted and observed total EWH, respectively, at a given month i. The mean values of the predicted time series are represented by $\bar{y}$, whereas $\bar{o}$ is the mean of the observed time series.

5.2. Optimal Input Data

The first step in the optimization of the NN, was to determine the set of input data that has the highest predictive power when it comes to GRACE-like EWH. Seven training phases of the NN were performed, one for each possible combination of the three types of hydro-climatological data:

1. Precipitation + temperature
2. Precipitation
3. SM + precipitation + temperature
4. SM + precipitation
5. SM + temperature
6. SM
7. Temperature

Each training phase consisted of 50 simulations. After the combination with the highest mean validation NSE was obtained, a more comprehensive hyperparameter tuning of the NN was performed. This is presented in Section 5.3.

In Figure 5.1, a box plot is used to depict the performance for each of the available input data combinations.

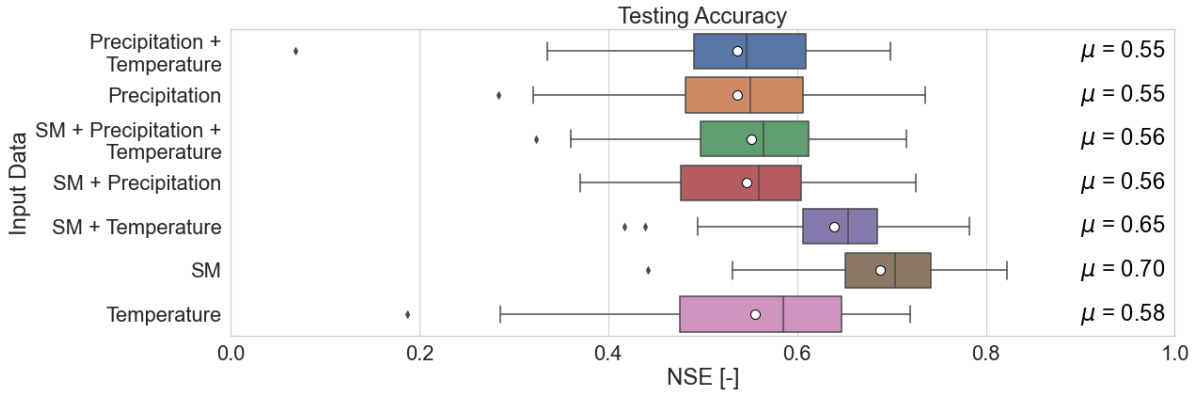


Figure 5.1: Boxplot of the final testing NSE for the different input data combinations

Precipitation ($\mu = 0.55$) is the worst-performing input variable, followed by temperature ($\mu = 0.58$) when it comes to predicting GRACE-like EWH. When soil moisture or a combination of soil moisture and temperature is used as input data, the highest accuracy is obtained ($\mu = 0.70$ and $\mu = 0.65$ respectively). Fewer types of input data translate to fewer input nodes and hence less computational effort. Because of the higher performance and the lower computational effort when compared to the "SM + temperature" combination, the soil moisture maps were chosen as the only hydro-climatological input data for this project. For the remainder of the optimization phase, the input maps consisted of a combination of SM and Swarm data from 2014-2022.

5.3. Hyperparameter Tuning

The effectiveness of a NN is influenced by a variety of parameters. The model's architecture is the most crucial component, which was optimized in Chapter 4. In addition to this, there are a number of parameters known as hyperparameters that are established before the model is trained.

5.3.1. Hyperparameters

The hyperparameters of a NN are the predefined parameters that influence its performance. For the models considered in this research, the hyperparameters are listed below:

Hyperparameter	Range
Learning rate [-]	0.00001-0.0001
Learning rate scheduler	None, Multiplicative, Step, Plateau
Batch size [-]	1-20
Optimizer	Adam, SGD, RMSprop
Epochs	50-250

Table 5.1: The hyperparameters of the models in this research, together with their range of values/types

The importance of these hyperparameters is described in Chapter 4. For the learning rate, the range of values resulted from trial and error with some preliminary experiments. There are only 57 months of data available for training, therefore a maximum batch size of 20 was chosen. The optimizers that were considered are the most popular ones in image recognition ML projects and are described in Section 4.1. The number of epochs also resulted from some preliminary experiments where it became clear that all experiments stagnated between 50 and 250 epochs.

5.3.2. Strategy

The hyperparameter tuning was performed by using the Optuna framework for Pytorch[2]. This framework only requires the user to define the hyperparameters and their respective range of possible values. The Optuna model will then perform a predefined number of studies, where each study consists of training the NN with a certain set of hyperparameters. This set of hyperparameters is altered after a study is completed, logging the performance of each study. The Tree-structured Parzen Estimator (TPE) is the default sampler in Optuna. By selecting locations nearer to previously successful results, Optuna utilizes TPE to search more effectively than a random search. The framework also provides numerical values for the importance of each hyperparameter, as well as contour plots illustrating these importances.

5.3.3. Final Hyperparameters

A first search was performed by the Optuna framework, consisting of 100 trials with the hyperparameter ranges for learning rate, batch size and optimizer listed in Table 5.1. The optimizer range consists of three distinctive optimizers, therefore the best one can be obtained by looking at the validation performances of each type of optimizer. This is presented in Figure 5.2, where the validation accuracy of the 100 trials is shown, sorted by optimizer.

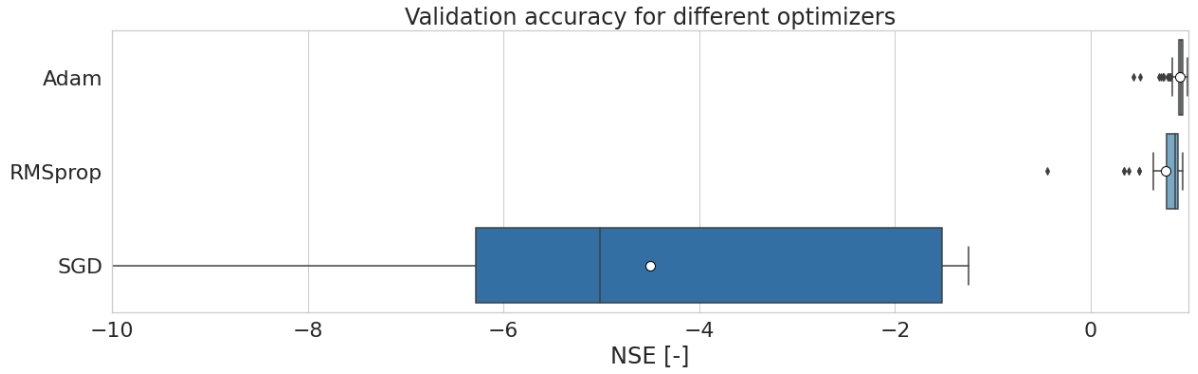


Figure 5.2: Validation accuracy over the first 100 optimization trials, sorted by optimizer.

From the three optimizers that were considered for this project, both the ”Adam” and the ”RMSprop” optimizers perform relatively well. Looking at the bottom line in Figure 5.2, it can be concluded that the ”SGD” optimizer does not work at all for this model. For the continuation of this project, it was decided to use the ”Adam” optimizer because of its slightly higher and less volatile performance compared to the ”RMSprop” optimizer.

Figure 5.2 also shows that the model optimization is finished when the out-of-sample performance stabilizes after 150 epochs. Decreasing the number of epochs from 250 to 150 resulted in lower training time while preserving a similar accuracy.

The next step was to determine the optimal learning rate and batch size for the model by creating a contour plot depicting the hyperparameters on the x- and y-axis, and the accuracy on the z-axis. For the first 100 trials, this contour plot is shown in Figure 5.3.

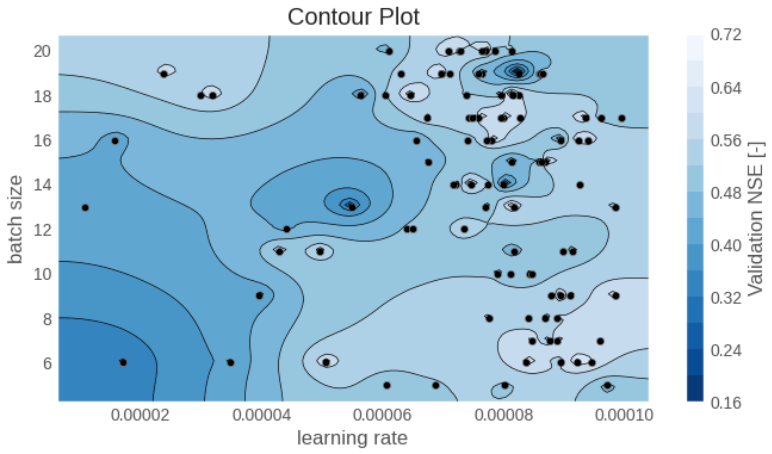


Figure 5.3: Contour plot of learning rate and batch size for the first 100 optimization trials.

This contour plot is the result of 100 trials in which the model is trained with a different learning rate-batch size combination for each trial. By plotting the final validation accuracy for each of these trials, a contour plot can help identify the optimal hyperparameter combinations. From the simulations shown in Figure 5.3, the optimal learning rate-batch size combination was found to be 0.000078 - 17.

To check whether a decreasing learning rate could improve the model’s performance, 200 additional trials were performed. In these trials, the batch size and initial learning rate were set based on the optimal values obtained from Figure 5.3, but the learning rate strategies described in Section 4.1 were implemented. The results are shown in Figure 5.4.

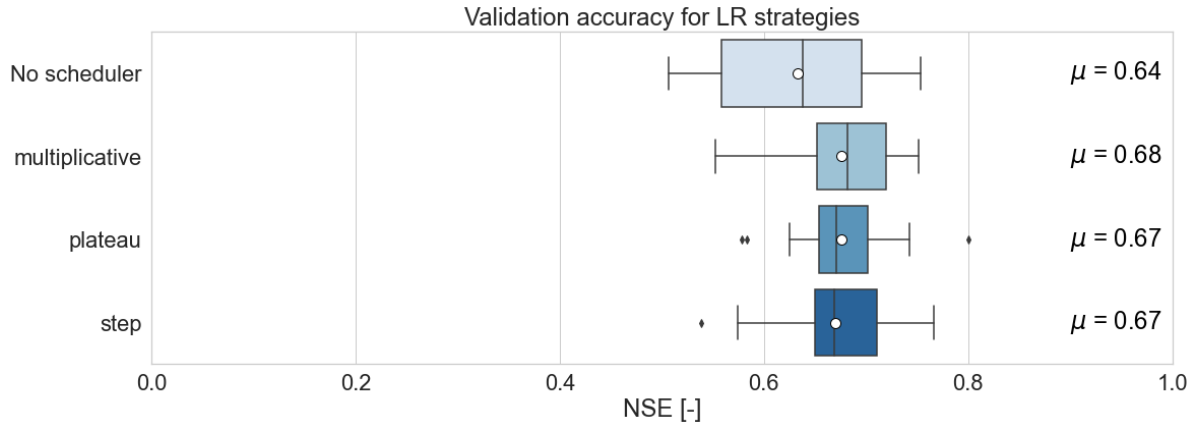


Figure 5.4: Boxplot of the final validation NSE for the different LR decreasing strategies.

Figure 5.4 shows that decreasing the learning rate over time generally increases the performance of the model and decreases the standard deviation of the results, compared to a fixed learning rate. There is no large difference between the three decreasing strategies in terms of average validation NSE (μ: 0.67-0.68). The multiplicative strategy was selected for this project, because of its slightly higher average accuracy. After completing the above steps, a NN optimized for this project was obtained. The final hyperparameters are listed in Table 5.2.

Hyperparameter	Optimal value
Learning rate [-]	0.00008
Learning rate decrease	Multiplicative
Batch size [-]	18
Optimizer	Adam
Epochs	80

Table 5.2: The hyperparameters of the optimized MLP model

5.4. Stacking Months

Up until this point, the models were trained by only using one month of Swarm and SM data. This means that the NNs were not provided with information on the temporal relationships, i.e. correlations, between two or more consecutive months. It was decided to stack prior months onto the current month and retrain the models in order to account for EWH variations in the temporal domain that may aid the NN to predict GRACE-like EWH maps. The MLP holds a 1-D input layer, therefore stacking the input maps was performed by adding *n x (1617 + 1617)* input nodes. With n the number of extra months being stacked, and 1617 (lat: 33° x lon: 49°) the number of nodes in one map (Swarm or SM). This is illustrated by Figure 4.4. The effect of stacking months on the validation accuracy and training accuracy is shown in Figure 5.5 and Figure 5.6 respectively.

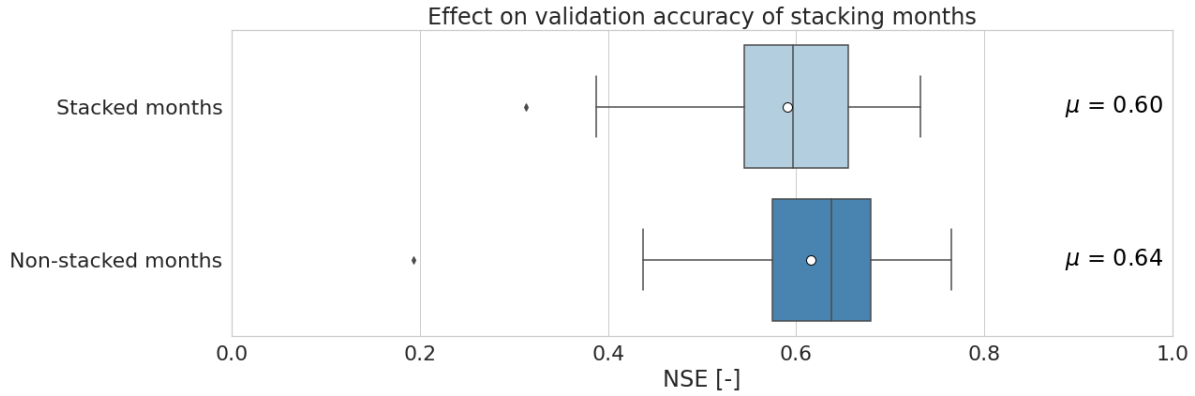


Figure 5.5: The effect of stacking on the validation accuracy. Result of 100 training simulations.

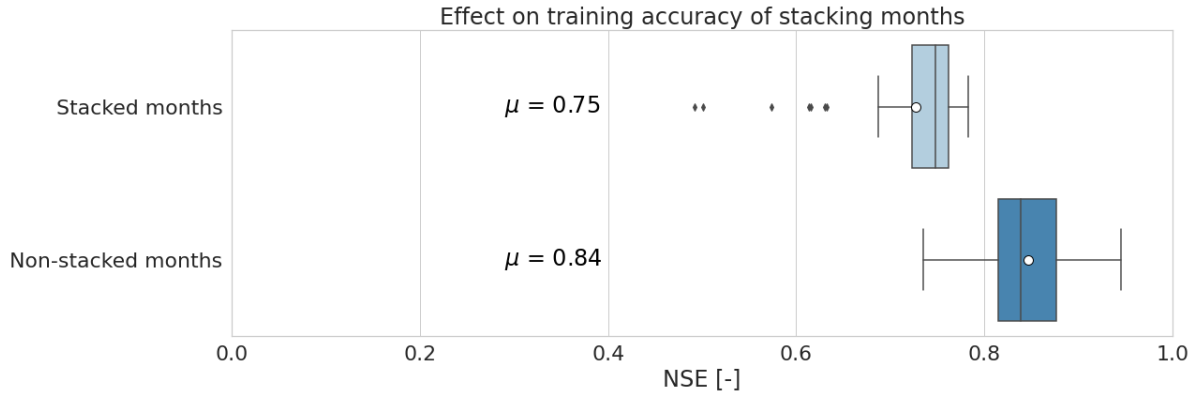


Figure 5.6: The effect of stacking on the training accuracy. Result of 100 training simulations.

Stacking the input months decreases both the validation accuracy as well as the training accuracy of the model. The decrease in validation accuracy is not caused by overfitting, because the training accuracy decreases as well. Therefore, the decrease in the model performance must be caused by the ANN not being able to handle the increase in information from the input layer. The problem becomes too complex for a straightforward MLP to handle. One way of solving this problem is to test stacking with a more complex NN, such as a CNN.

5.5. Transfer Learning

Figure 5.7 shows the effect of the transfer learning strategy, presented in Section 4.1, on the performance of the model. The first NN is trained for 30 epochs, after which the second training phase is initialized and the Swarm data is added to the input layer.

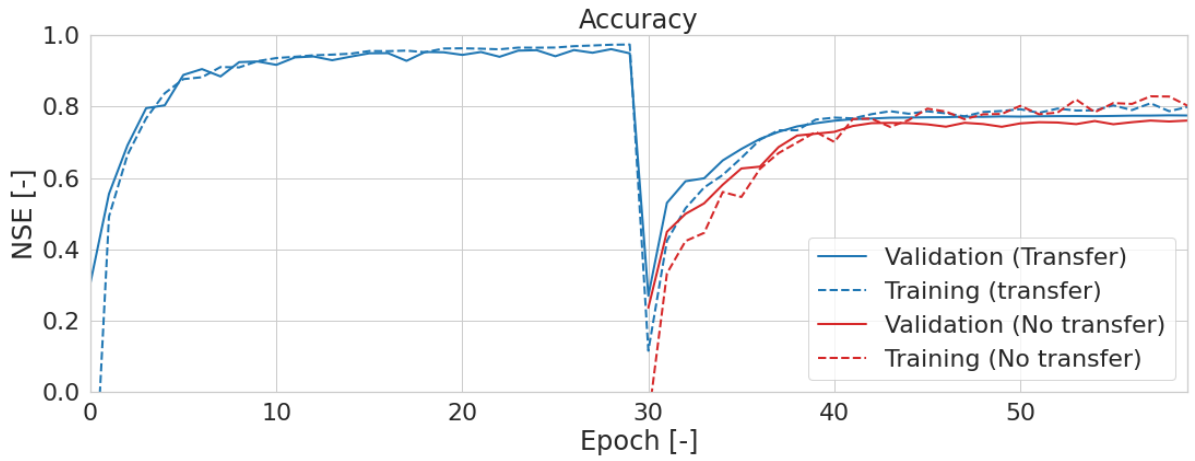


Figure 5.7: The effect of transfer learning on the training and the validation perfromance of the ANN.

From Figure 5.7, the first NN learns and stabilizes after around 20 training epochs. After 30 epochs, the transfer to the second NN is completed and the performance immediately drops for both the training and the testing data sets. The drop in performance follows from the fact that the input layer of the second NN is altered compared to the pre-trained NN, causing instability in the model. Although the hidden layers are identical, this instability in the first layer propagates through the layers of the model towards the final layer. This causes the second model to learn as if it was a completely new model, which it does. After around 20 new epochs (at epoch 50 in Figure 5.7), the model stabilizes and reaches accuracies comparable to the models described in the previous sections.

5.6. Influence of Swarm Quality

What causes the differences in training performance of the individual months may be explained by the data quality of the Swarm maps for these months. To verify this, the difference between each Swarm month and the interpolated GRACE solution is shown in Figure 5.8. The error is defined as the average absolute EWH difference over the Amazon basin between GRACE and Swarm for a single month.

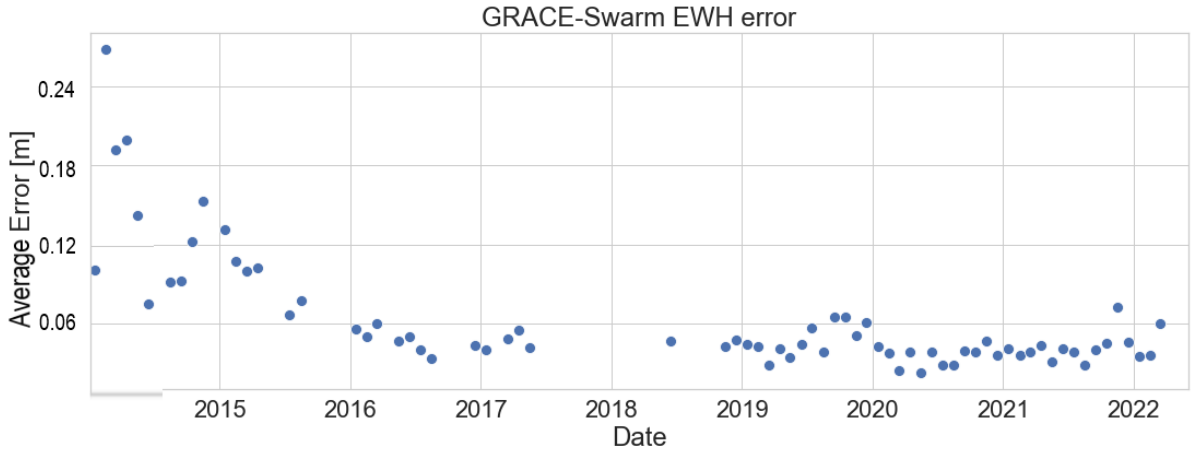


Figure 5.8: The difference between Swarm-derived EWH and GRACE-derived EWH over the Amazon basin.

The data quality of Swarm is relatively bad for the first year, after which it increases. After January 2016 the quality stabilizes and the errors become lower. Therefore, it was investigated whether excluding data related to the months from the beginning of the Swarm period could benefit the training of the models. The disadvantage of excluding the data from this period is that there are even fewer months available for training. This is a trade-off between data quality and availability. On the other hand, the data quality is so poor that it might not make sense to use it for the training of the NNs. The Swarm

months were divided into three categories, depending on their errors, as shown in Figure 5.9.

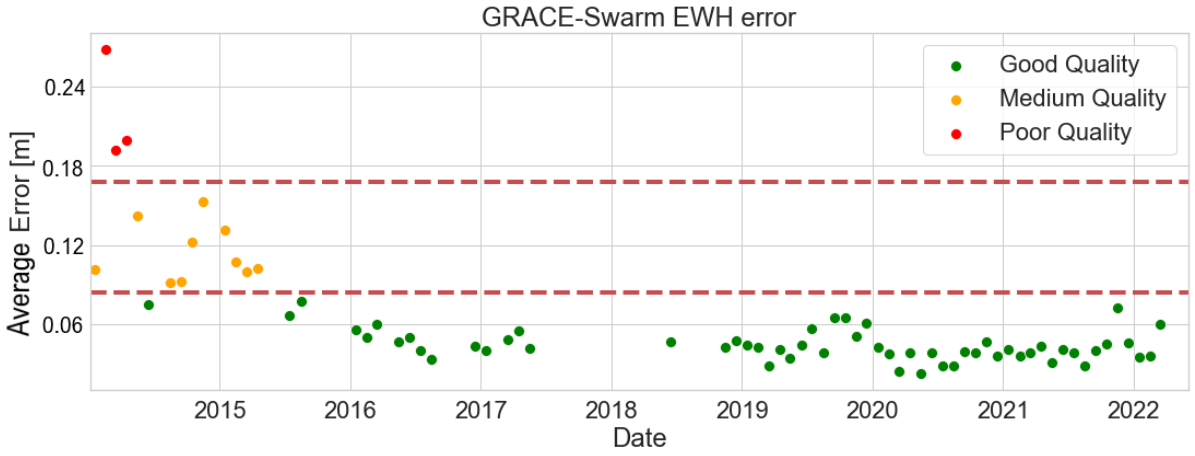


Figure 5.9: The difference between Swarm-derived EWH and GRACE-derived EWH over the Amazon basin, divided by data quality.

After dividing the months based on their data quality, the models were re-trained without excluding months and with the exclusion of poor and/or medium-quality months. The results are presented by Figure 5.10. Important to note that the number of testing months was kept constant at 12 for each simulation so that the effect of decreasing the number of training months was taken into account.

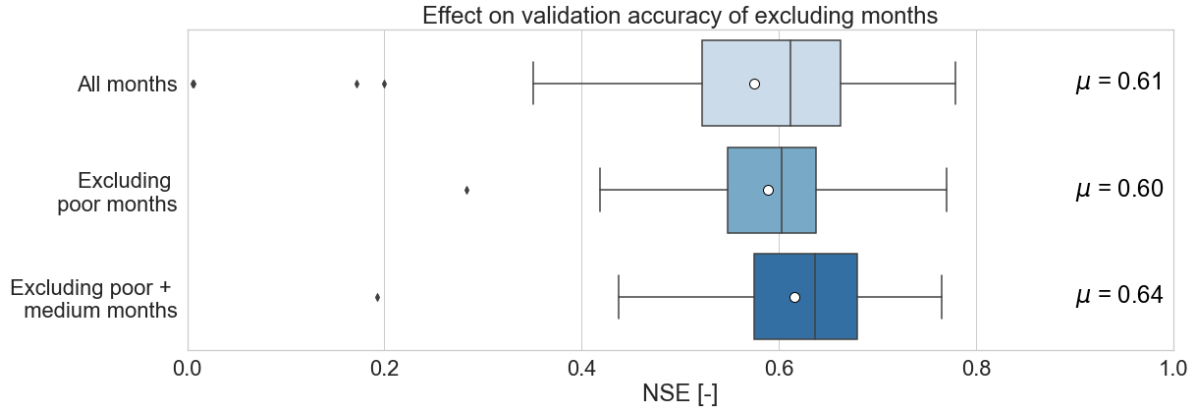


Figure 5.10: Validation accuracy for the different levels of Swarm data quality.

Excluding the poor-quality months does not improve performance; in fact, it slightly lowers the average NSE. Excluding both the low and medium-quality months raises the average NSE by 3%. This may not appear to be much at first glance, but recall that this performance was acquired with 13 months less training data. It was decided to keep using only high-quality months for NN training because doing so speeds up the training process.

5.7. Normalization of Input Data

The last step in optimizing the ANN is to check whether normalizing the input data prior to the training of the model is beneficial for the performance of the model. The statistical normalization was performed as described in Section 3.2. To assess its performance compared to non-normalized input data, 50 simulations with identical hyperparameters were performed. The validation performance and training performance are shown in Figure 5.11 and Figure 5.12 respectively.

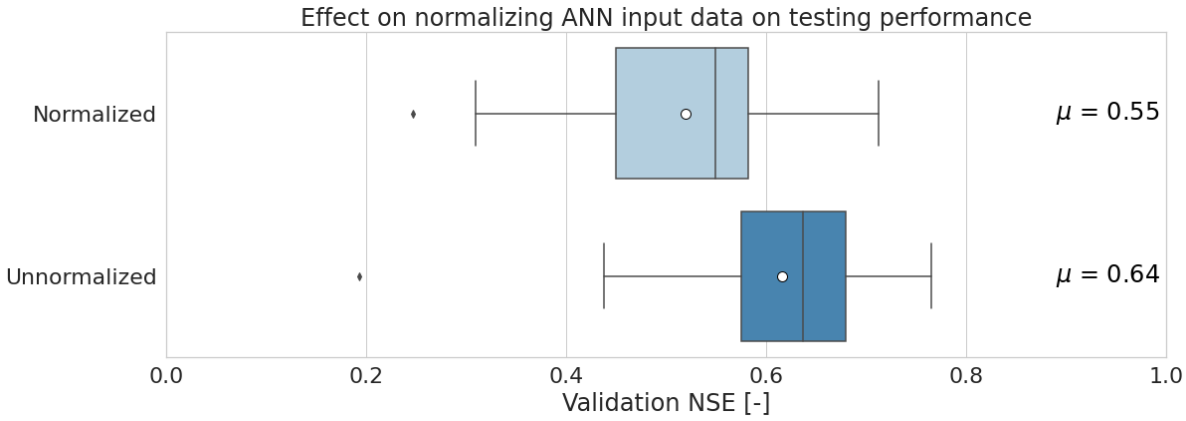


Figure 5.11: Boxplot of NN performance for both normalized and unnormalized input data. Both consist of 50 experiments each.

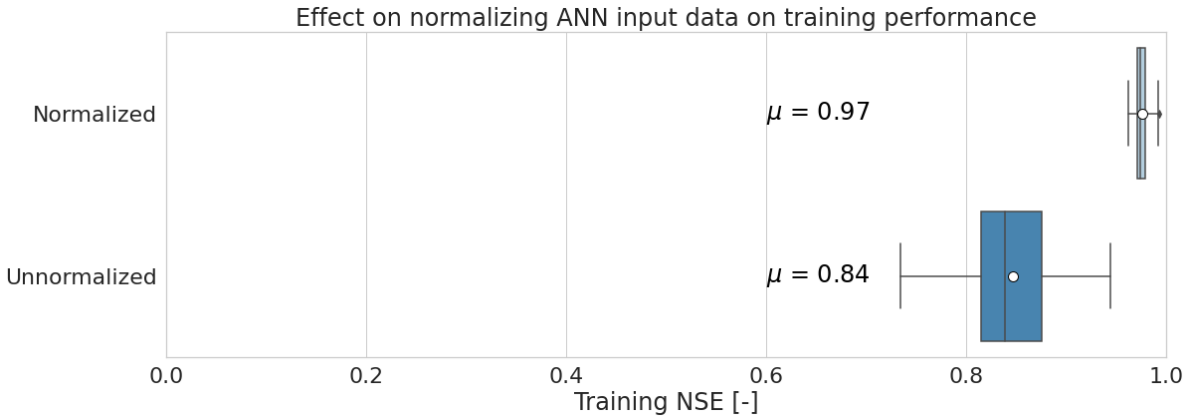


Figure 5.12: Boxplot of NN performance for both normalized and unnormalized input data. Both consist of 50 experiments each.

First of all, by analyzing Figure 5.11, it became clear that normalizing the input data decreases the validation performance of the model. The average validation NSE dropped 9% compared to the simulations with unnormalized input data. The cause behind this can be observed from Figure 5.12. The training accuracy of the model increased from 0.84 to 0.97 after the normalization of the input data. This is a clear sign that normalizing the input data causes the model to overfit the GRACE dataset. The reason for this overfitting is the fact that by standardizing the data, the effect of outliers is reduced, creating a much ”simpler” input dataset. This, together with the fact that the range of values in the dataset is decreased, makes it easier for the model to overfit the training months. The model starts to focus too much on reproducing these training months during the training of the NN, such that it fails to find general relationships between the input maps and the GRACE maps. This results in a decrease in the out-of-sample (validation) accuracy.

5.8. Final ANN Model

Once the optimization steps were completed, a final optimized ANN model was obtained. The final hyperparameters are listed in Table 5.3. This will be the ANN model that was used for the remainder of this project.

Hyperparameter	Optimal value
Learning rate [-]	0.00008
Learning rate decrease	Multiplicative (x0.98)
Batch size [-]	18
Optimizer	Adam
Swarm months	Only good quality
Stacking	No
Normalization	No

Table 5.3: The final (hyper)parameters of the optimized ANN model

Running 100 simulations with this model resulted in an average validation NSE of 0.68, an average training NSE of 0.83, and a CC of 0.95. The next step was to repeat these optimization steps, together with some new steps to find the optimal CNN for this project.

5.9. CNN

The optimised CNN model was obtained through a similar optimization process as described above. It was assumed that the best input data combinations (SM + Swarm) and optimizer (Adam) for the ANN would also be the best options for the CNN. The learning rate (strategy) and batch size were obtained in the exact same way as for the ANN. There were however some differences compared to the ANN, these are described in the following subsections.

5.9.1. 2D CNN vs. 3D CNN

As described in Section 4.2, there are two different architectural options for the CNN; a 2D CNN and a 3D CNN. In order to decide the optimal architecture, 50 simulations were performed for each of the two options. The validation and training performances are shown in Figure 5.13 and Figure 5.14 respectively.

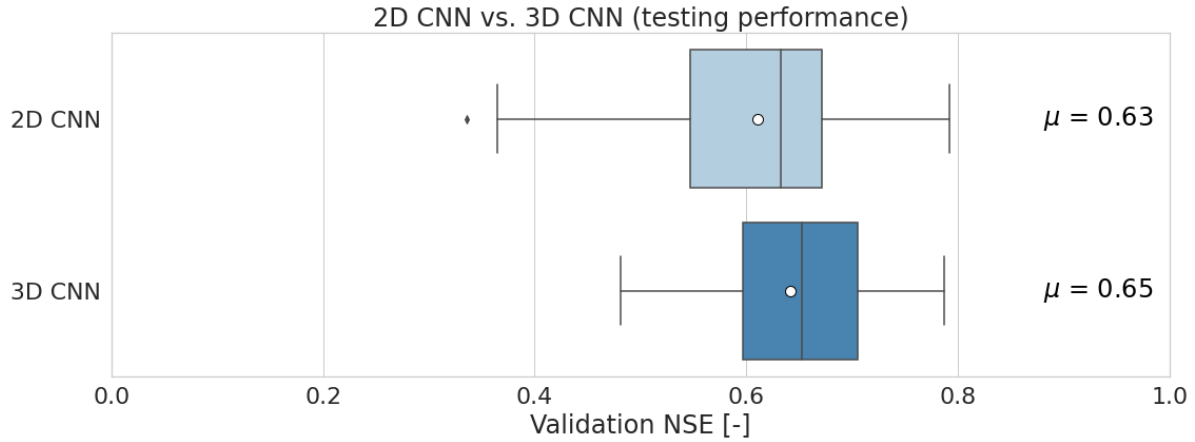


Figure 5.13: The validation accuracy of the 2D CNN and 3D CNN.

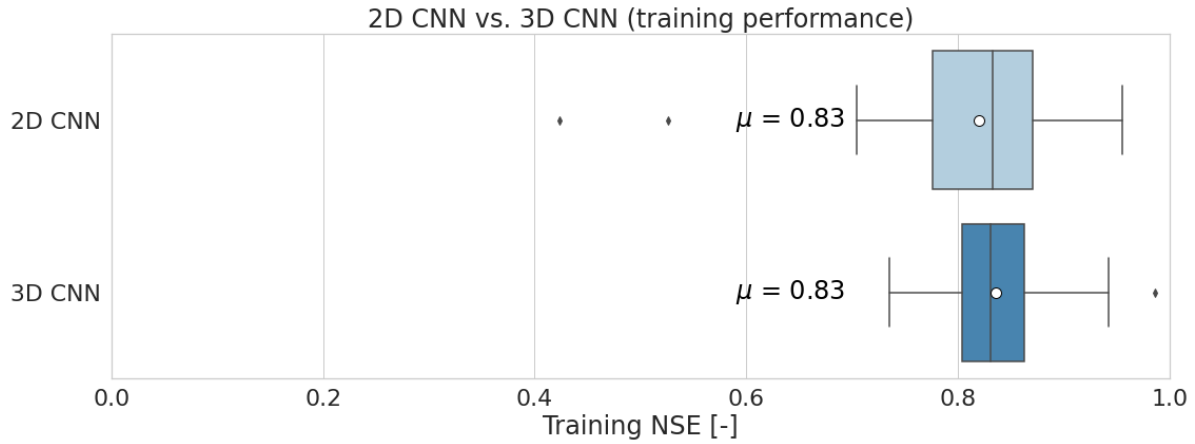


Figure 5.14: The training accuracy of the 2D CNN and 3D CNN

The 3D CNN slightly outperforms the 2D CNN with a validation accuracy of 2% higher, while obtaining a similar training accuracy. It can also be seen from the boxplots that for both the validation and training accuracies, the variance in results is lower for the 3D CNN.

5.9.2. Stacking of Input Maps

Also for the CNN, stacking the input months was considered to include temporal information in the input layer. As described in Section 4.2, stacking is done differently for the 2D CNN and the 3D CNN. The effect of stacking two input months on the validation accuracy for both architectures is shown in Figure 5.15.

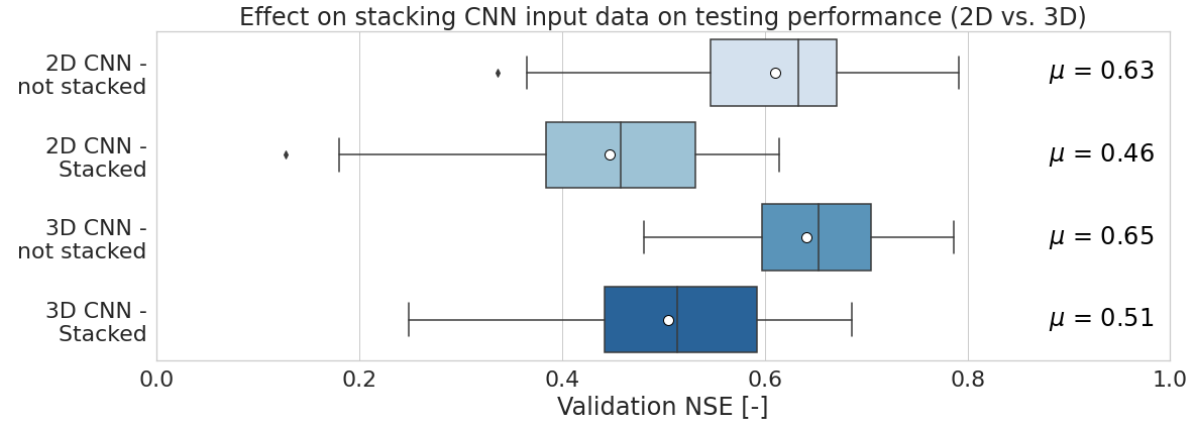


Figure 5.15: The validation accuracy of the 2D CNN and 3D CNN, with and without stacking. Each experiment consists of 50 simulations with optimised hyperparameters.

Stacking decreases the validation accuracy of the model, as shown in the 2nd and 4th rows in Figure 5.15. The 3D CNN slightly outperforms the 2D CNN (by 0.05 for the average NSE) when the input layers are stacked, but the performance still lies lower than without stacking ($\mu = 0.46$ vs. $\mu = 0.63$ for the 2D CNN and $\mu = 0.51$ vs. $\mu = 0.65$ for the 3D CNN). This could have two reasons: the model overfits the GRACE data when stacking takes place or the complexity of the model can not handle stacking.

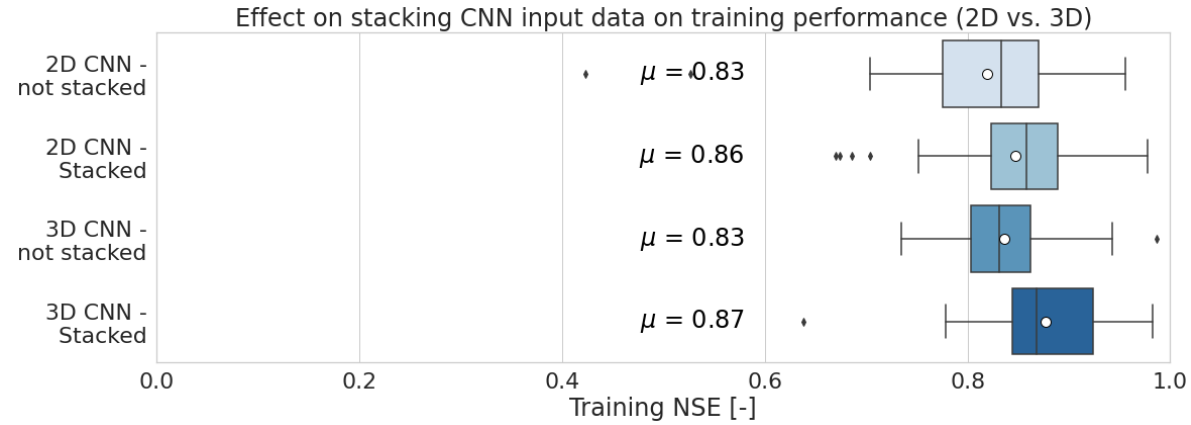


Figure 5.16: The training accuracy the 2D CNN and 3D CNN, with and without stacking. Each experiment consists of 50 simulations with optimised hyperparameters.

By analyzing the training performance, shown in Figure 5.16, it could be seen that the model indeed starts to overfit the data. The training accuracy increased with stacking (by 0.03 for 2D and 0.04 for 3D), while the validation accuracy drops significantly (by 0.17 for 2D and 0.14 for 3D). It was thus concluded that stacking the months in the input layer of the CNN was not beneficial for the results.

5.9.3. Dropout Level

As mentioned in Section 4.2, adding dropout layers to the CNN makes it harder for the model to overfit. To analyse the effect of adding dropout layers to the model, the dropout percentages for the CNN layers were set to 0, 25, 50, and 75 respectively. The effect on the CNN validation and training performances of these dropout layers is shown in Figure 5.17 and Figure 5.18 respectively.

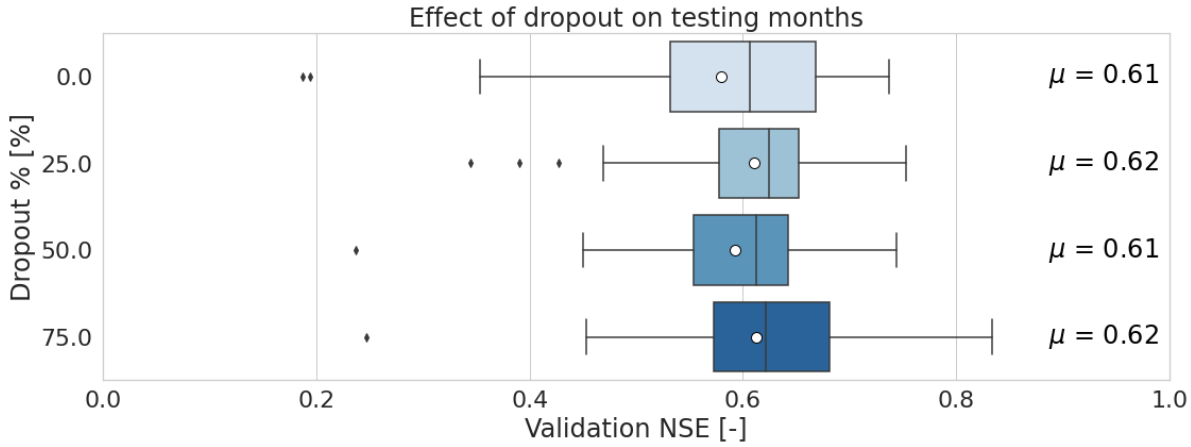


Figure 5.17: Effect of increasing the dropout percentage of the 3D CNN on the validation performance. Each experiment consisted of 50 simulations.

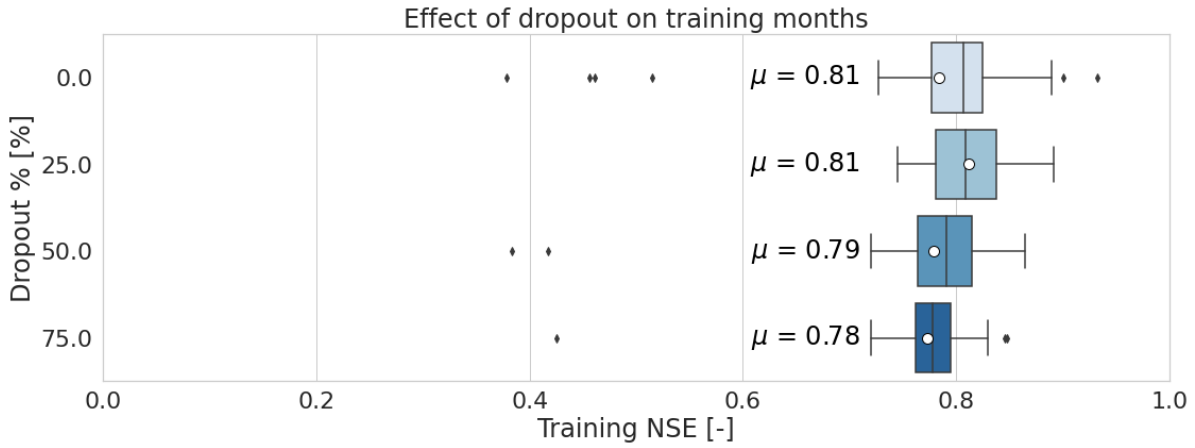


Figure 5.18: Effect of increasing the dropout percentage of the 3D CNN on the training performance. Each experiment consisted of 50 simulations.

Increasing the number of dropout cells did not influence the out-of-sample performance of the CNN. However, increasing the number of dropout cells slightly decreases the training accuracy, bringing it closer to the validation accuracy values. Reducing this difference between training and validation performance means slightly less overfitting is taking place. For the remainder of the project, a dropout percentage of 25% was used for the CNN.

5.9.4. Final CNN model

After going through the multiple optimization steps described above, the final CNN model was obtained. The final hyperparameters of the model are listed in Table 5.4.

Hyperparameter	Optimal value
Learning rate [-]	0.00012
Learning rate decrease	Multiplicative (x0.98)
Batch size [-]	12
Optimizer	Adam
Stacking	No
Normalization	No
Dropout [%]	25

Table 5.4: The hyperparameters of the optimized CNN model

After conducting 100 simulations, the CNN model resulted in an average validation NSE of 0.65 and an average CC of 0.96 compared to GRACE.

5.10. ANN vs CNN

Once both the ANN and the CNN were optimized, one of the two had to be selected for the continuation of the project. This decision was based on the performance of the neural network and the training time for one simulation. These values, resulting from 100 simulations, are listed in Table 5.5

		ANN	CNN
Validation NSE [-]	Average	0.68	0.65
	Best	0.79	0.78
Training NSE [-]	Average	0.83	0.87
	Best	0.94	0.98
Average Correlation Coefficient [-]		0.96	0.95
Training time [min/simulation]		2	18

Table 5.5: The average and best performances of the ANN and the CNN models, together with their average correlation coefficient and training time. CC is calculated with respect to GRACE EWH time series. Results of 100 simulations for each model.

Based on the values from Table 5.5, the performances of the two models lie very close to each other. This holds for both the spatial domain (NSE) and for the temporal domain (CC). The only significant difference between the models is in the training time. The more complex CNN takes around 9 times longer to train than the simpler ANN. These 16 minutes of extra training time are manageable for one or even a few simulations, but to obtain the final gap-filling time series, 1000 simulations had to be conducted. Therefore, it was decided to continue with the ANN for the rest of this research.

6

Performance of Neural Networks

The optimal ANN was used to fill the GRACE/GRACE-FO data gap. In doing so, three experiments were performed, each consisting of 1000 training simulations. The three experiments differed in terms of the input data that was used to train the model. The first experiment only used SM (2004-2022) as input to the model. In the second experiment, Swarm data (2014-2022) was the sole input to the model. The last experiment was the most crucial to answering the main research question of this project, as it combined both SM and Swarm data as input to the NN. The EWH time series resulting from these experiments are presented and discussed in this chapter, together with a sensitivity analysis of the NN output. Lastly, the results of this research will be compared to the results obtained in previous GRACE/GRACE-FO gap-filling studies.

An overview of the different input data combinations and the corresponding input layer sizes are presented in Table 6.1

Input data	Dimensions of input maps	Size of input layer [nodes]
Soil Moisture	33 x 49	1617
Swarm	33 x 49	1617
Soil Moisture + Swarm	2 x (33 x 49)	3234

Table 6.1: Overview of the size of the input maps and the corresponding size of the input layers for the considered combinations of input data.

6.1. Soil Moisture (2004-2022) as input data

From Section 5.2, it was clear that Soil Moisture (SM) data acts as the optimal hydro-climatological parameter when it comes to predicting GRACE EWH. This SM data is available for the entirety of the GRACE/GRACE-FO mission lifespan, unlike Swarm. Therefore it was decided to train the NNs using this SM data as the sole input to the model. This would, among others, provide more information on whether the performance of the model trained with both SM and Swarm data was hindered by the data availability of Swarm.

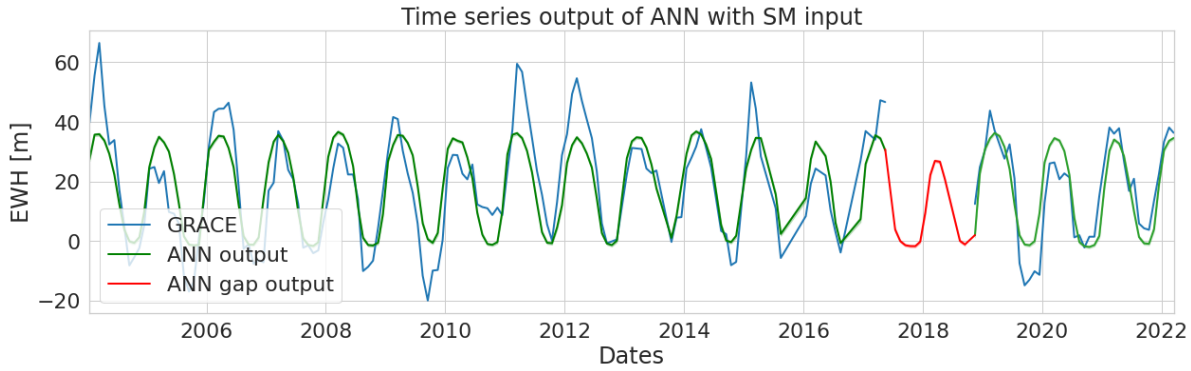


Figure 6.1: The Amazon Basin EWH time series of GRACE compared to the results from the ANN trained with only SM data. The ANN output line consists of the mean of 1000 simulations and its respective 95% confidence interval.

At first sight, the ANN performs relatively poorly in reproducing GRACE-like EWH. This is confirmed by the mean training and NSE values (0.58 and 0.41 respectively). It under-fits the GRACE time series data, reproducing a smooth line. It is able to capture seasonal variations in the EWH but is unable to catch small local peaks and valleys that can be observed in the GRACE time series. The reason behind this becomes clear by comparing the SM time series to GRACE, as can be seen in Figure 6.1. The SM time series follows the same periodicity as the GRACE time series but there are almost no local features present in the SM time series, unlike GRACE. If these local minima/maxima are not captured by the SM, it is impossible for the ANN to reproduce them from SM maps solely. For sake of comparison, a part of the parametric GRACE EWH was added to compare the smoothness of the SM time series to that of a linear model. It is clear that it leans more toward the parametric model than toward the more erratic GRACE time series.

Over-fitting is also partly avoided by the simplicity of the ANN. If an ANN would overfit the GRACE time series, it would essentially have to remember which SM map had which exact GRACE map as output. The amount of training months (140) is too large for the network to succeed in this, preventing, in part, over-fitting.

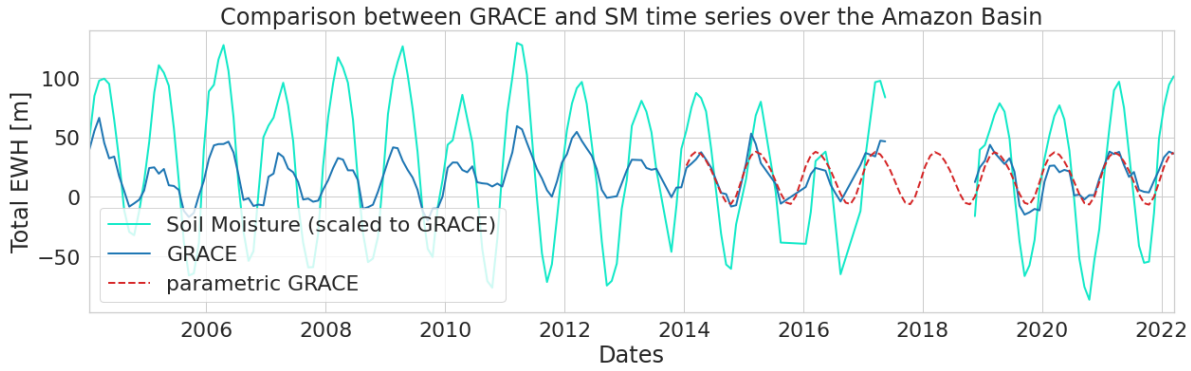


Figure 6.2: Caption

6.2. Swarm (2014-2022)

For the second experiment, the NNs were trained with Swarm maps as the only input to the model. As a result of the poor quality of the earliest Swarm EWH products, compared to GRACE (see Section 5.6), there were only 57 months of data available for training and validation of the NNs. Figure 6.3 shows the GRACE-like EWH resulting from the ANN trained with Swarm data as the only input.

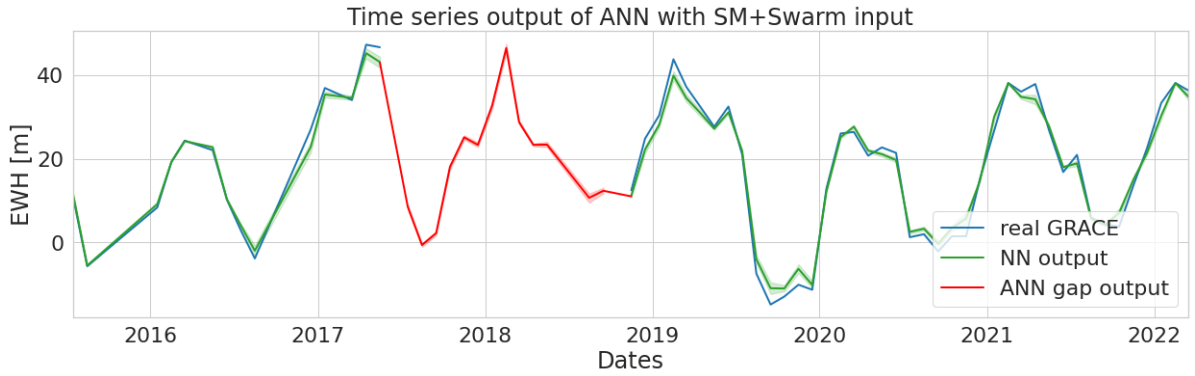


Figure 6.3: The Amazon Basin EWH time series of GRACE compared to the results from the ANN trained with only Swarm data. The ANN output line consists of the mean of 1000 simulations and its respective 95% confidence interval.

The NN time series agrees much better with GRACE compared to the time series from the previous experiment (Figure 6.1). There is much less under-fitting and the local minima/maxima in GRACE are better captured by the NN time series. This seems positive at first glance. However, keeping in mind the limited amount of training months, this could be a result of the model overfitting the GRACE time series. This can be verified by comparing the training performance to the validation performance, as is shown in Figure 6.4.

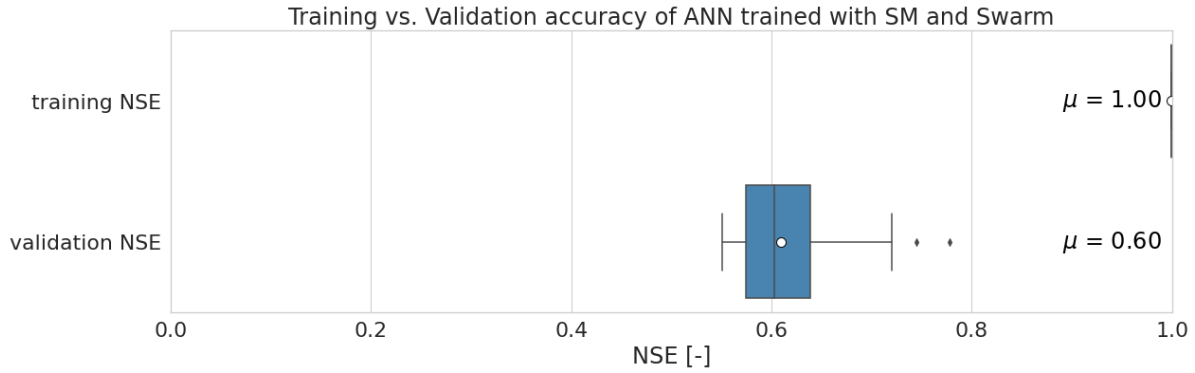


Figure 6.4: The training performance compared to the validation performance for the models trained with Swarm as the only input. Result of 100 simulations with optimised hyperparameters.

The average training accuracy of 1.00 indicates that for all the simulations, the NN manages to mimic the GRACE time series perfectly. The validation accuracy, on the other hand, has an average value of only 0.60. This proves that the NN trained with only Swarm data overfits the training dataset, and thus does not manage to find general relationships between the Swarm dataset and the GRACE EWH dataset. This is the opposite of the underfitting that happened in the first experiment.

To understand how the NN is able to overfit the GRACE training data by using Swarm input, the time series of Swarm and GRACE were considered. These are shown in Figure 6.5

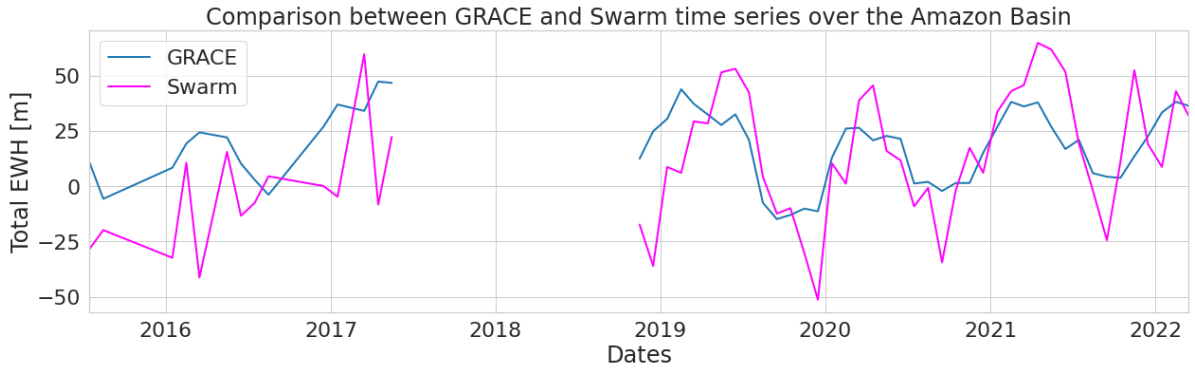


Figure 6.5: Amazon basin EWH time series of GRACE and Swarm. Only the months of "good" Swarm quality are shown.

The Swarm EWH time series is clearly much more erratic compared to the SM time series shown in Figure 6.2. The peaks in the Swarm time series are more or less in sync with those in the GRACE time series. As a result of these distinctive local minima/maxima, the NN is able to remember which Swarm month corresponds to which GRACE month and therefore yields a high training accuracy. This, together with the fact that 45 months of training is sufficient for the NN to remember, results in the overfitting that can be observed from Figure 6.3 and Figure 6.4.

6.3. Soil Moisture + Swarm (2014-2022)

The last experiment was a combination of the first two experiments, described above. It tested the hypothesis of this thesis by combining the SM and Swarm data as input to the NNs. Similar to the second experiment, there were again only 57 months of SM and Swarm data used for training and validation of the model. The resulting EWH time series are shown in Figure 6.6.

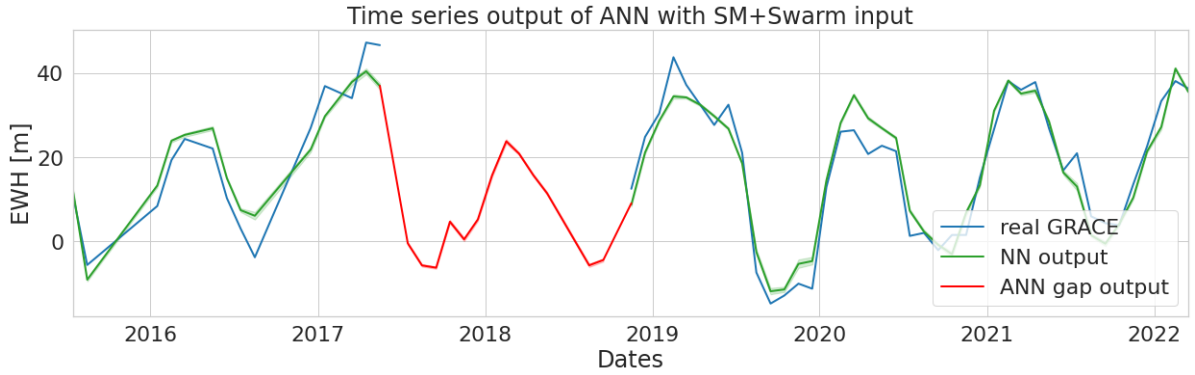


Figure 6.6: The Amazon Basin EWH time series of GRACE compared to the results from the ANN trained with SM and Swarm data. The ANN output line consists of the mean of 1000 simulations and its respective 95% confidence interval.

The NN output time series fits relatively well with the GRACE time series, with a CC of 0.95. From the CC, and looking at Figure 6.6, it can be concluded that performance between the NN trained with a combination of SM and Swarm lies between underfitting such as for experiment 1 and overfitting like experiment 2. This is confirmed by the training and validation performance shown in Figure 6.7.

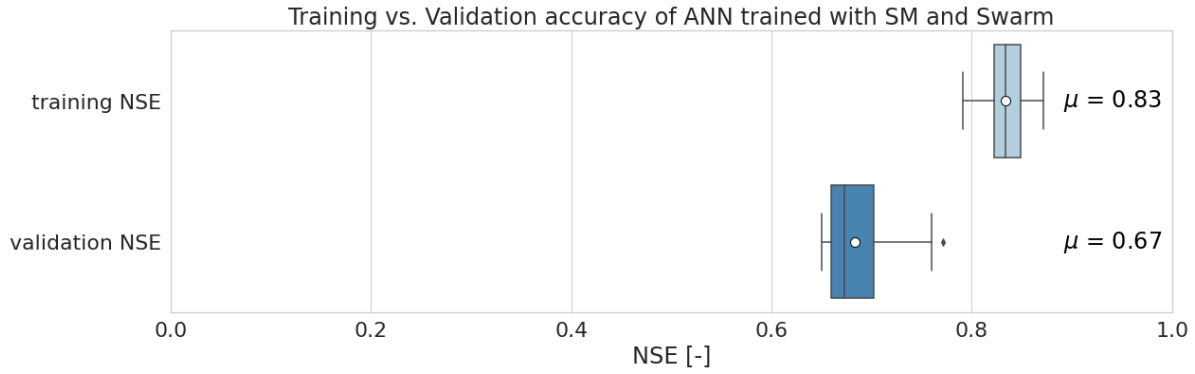


Figure 6.7: Training and validation accuracy of the ANN trained with both SM and Swarm as input data.

The training and validation performance lie much closer to each other than for the NN trained with Swarm only, indicating much less overfitting is taking place. The training and validation accuracy, together with the CC of the three different experiments are compared in Figure 6.8.

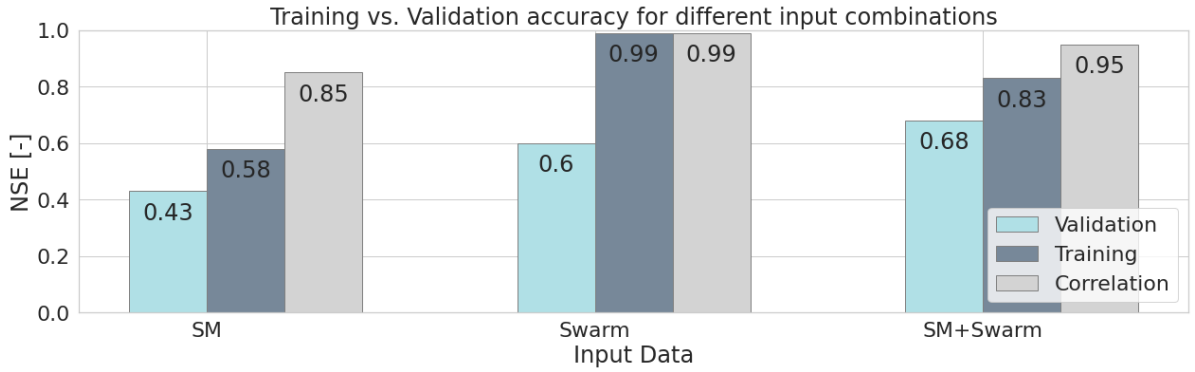


Figure 6.8: Comparison of the training accuracy, validation accuracy and temporal correlation with GRACE for the three different input data options.

The NN trained with both SM and Swarm yields the highest validation accuracy. This is because it is able to better capture the local GRACE minima and maxima than the NN model trained with SM only, and it is less prone to underfitting the GRACE data that the NN trained with Swarm only. It also yields a high temporal accuracy to the GRACE time series with a CC 0f 0.95.

Based on this, it was concluded that combining SM and Swarm data as input to the ANN is indeed beneficial when reproducing GRACE-like EWH. It yields the highest validation performance and a high CC, without overfitting GRACE.

6.4. Resulting Gap Time Series

Once the final gap-filling time series were obtained for the three experiments, they could be compared. Figure 6.9 shows the three time series together with GRACE around the gap.

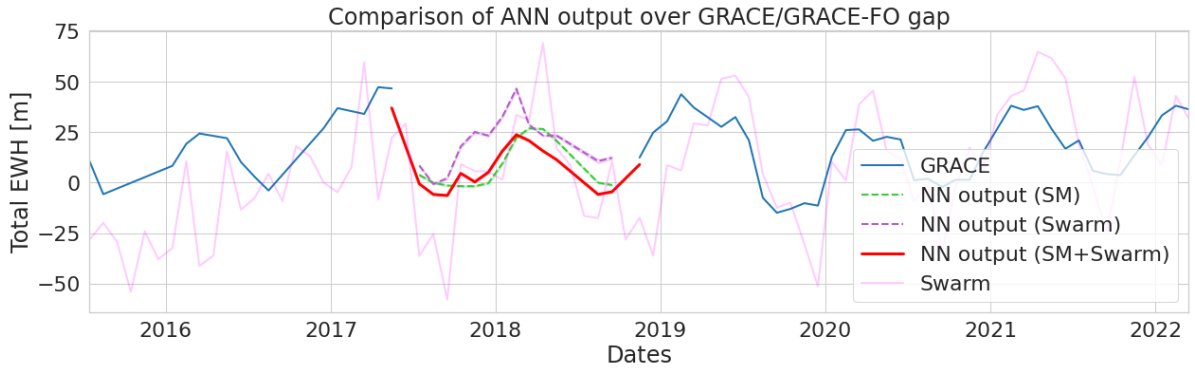


Figure 6.9: The gap-filling time series resulting from the three experiments, 100 simulations each.

All three of the gap time series show a peak around the beginning of 2018, which is in agreement with the seasonal variations in the Amazon Basin. The wet season starts around November, causing the EWH to rise. The end of the rain season is typically around March, just after the peak of the gap time series in February 2018. Another observation that can be made is the smoothness of the SM model output compared to the outputs that have Swarm as a data input. This is a result of the underfitting that played a role in experiment 1, creating a smooth time series without local minima or maxima. Swarm introduces a small decrease in EWH around October 2017, after which the EWH increases again when the wet season starts. Another interesting feature is that the ”Swarm only” time series holds higher EWH values than the other two gap-filling time series. This could be a result of the underestimation of GRACE EWH by SM data, as shown in Figure 6.1. This underestimation is however not present in the time series around the gap, illustrated by Figure 6.6. Therefore the reason for this difference between the ”Swarm only” EWH time series and the ”SM+Swarm” time series is most likely a result of the overfitting taking place when Swarm is the only input to the ANN.

Because of the GRACE/GRACE-FO data gap, there is no measured EWH data available over this period to compare the ANN results to. However, by analyzing the Amazon Basin SM over time, it is possible to gain insight into the intensity of the wet season over the data gap. This is shown in Figure 6.10.

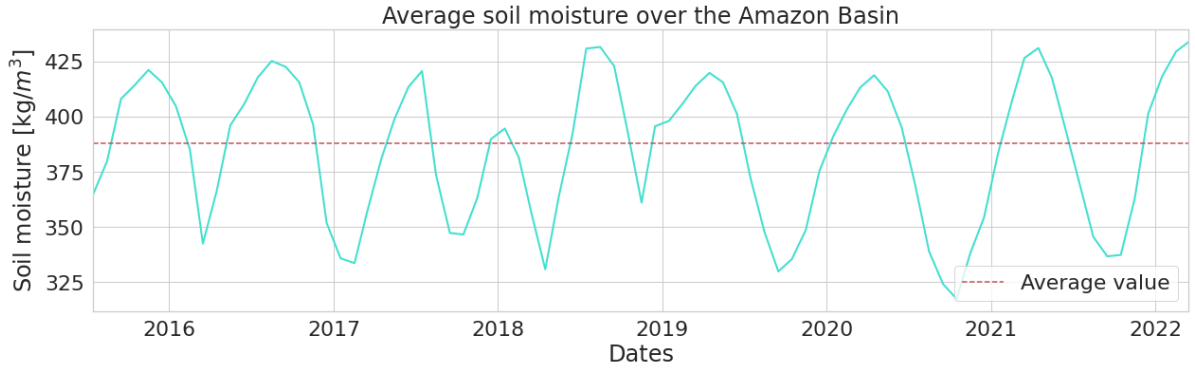


Figure 6.10: The average soil moisture content per grid cell over time, for the Amazon Basin.

For comparison sake, the average SM value is also plotted in Figure 6.10. When compared to the other peaks in time, the 2017/2018 wet season was a particularly weak one. This increases the confidence in the weak peak observed in the ANN output time series. It thus also validates the assumption that the difference between the ANN output (SM+Swarm) and the ANN output (Swarm), observed in Figure 6.9, is caused by overfitting for the Swarm-only input case.

Figure 6.11 compares the final ANN output to two other EWH time series over the data gap, namely the parametric GRACE model, the Swarm-derived EWH.

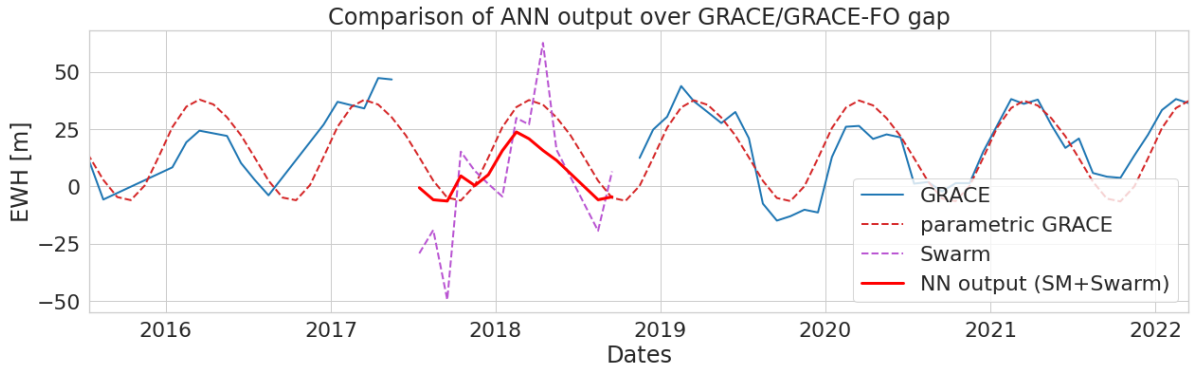


Figure 6.11: The gap-filling time series resulting from the experiment with SM ad Swarm as input maps, compared to Swarm EWH and the parametric GRACE model over the gap.

The ANN output agrees well with the general EWH time series shape over the data gap. All three time series show a peak in EWH in the first half of 2018. The ANN output EWH lies lower than the parametric GRACE model, confirming the weak 2018 wet season observed from Figure 6.10.

6.5. Spatial Distribution of Correlation

After analyzing the temporal performance of the ANN output EWH time series, the spatial performance was analyzed in more detail. This helps identify the regions over the Amazon Basin where the ANN is able to reproduce GRACE-like EWH more accurately. In these regions, the resulting ANN EWH time series agrees better with GRACE/GRACE-FO than in other areas. This can be visualized by plotting the CC for each grid cell in the study region, as shown in Figure 6.12.

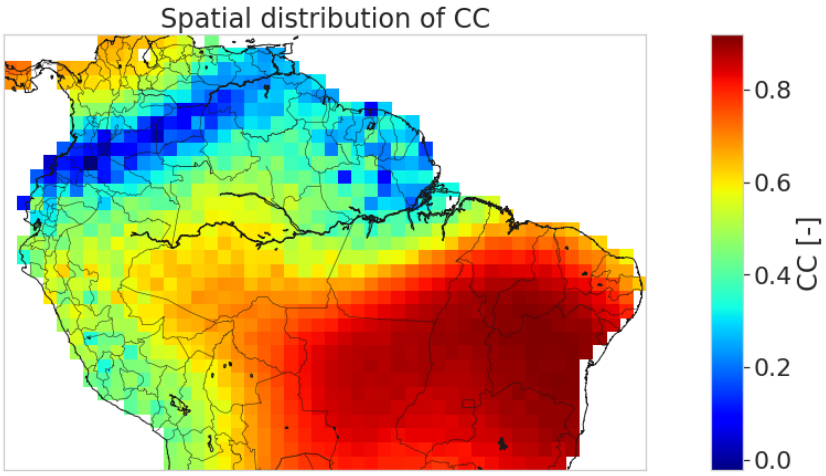
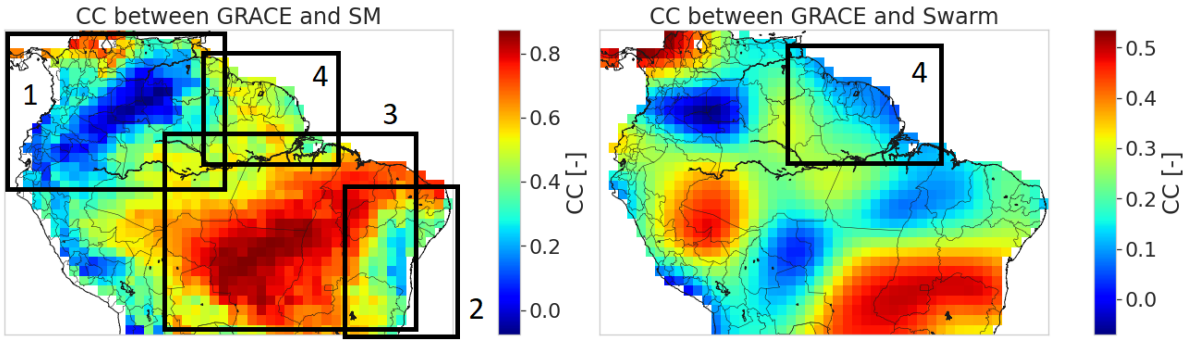


Figure 6.12: The temporal correlation between the ANN output and GRACE, over the Amazon Basin.

The NN is best able to predict GRACE-like EWH trends in the southeastern part of the Amazon Basin, with CC values in the range of 0.7-1.0. In the northwestern part of the study region, there is an area where the NN does not manage to produce accurate EWH trends compared to GRACE (CC<0.2). To investigate the cause of this imbalance in the temporal accuracy of the NN, the temporal agreements between the input data (SM + Swarm) and GRACE in these areas were considered. Figure 6.13 shows the spatial distribution of the CC between GRACE and the input maps over the Amazon Basin.



(a) Soil Moisture **(b)** Swarm EWH

Figure 6.13: The temporal correlation between the input data and GRACE, over the Amazon Basin.

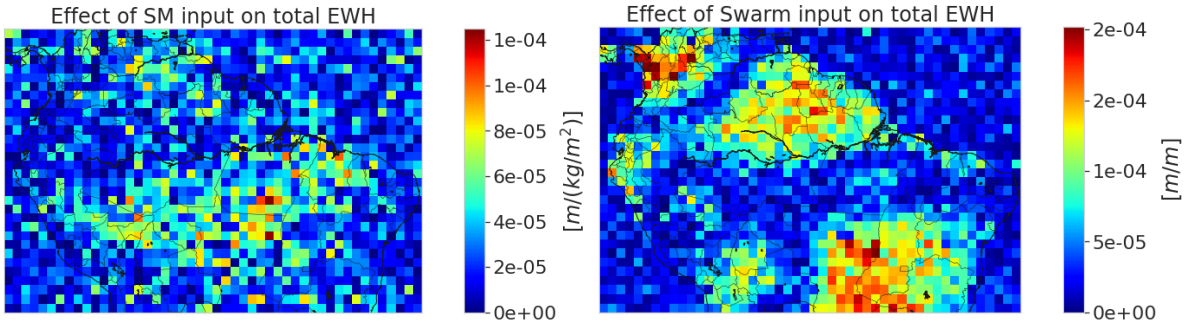
The extreme values in the temporal correlation between GRACE and the NN output in Figure 6.12 can also be observed from Figure 6.13a (box 1 and box 3). There are however smaller features that differ between Figure 6.12 and Figure 6.13a, such as those in box 2 and box 4. The NN results in higher CC values on the rightmost part of the study region (in Figure 6.12), whereas the correlation between SM and GRACE in this region is relatively low (box 2 in Figure 6.13a). Another interesting feature can be observed from box 4. This is a region where there is a higher correlation between GRACE and SM than between GRACE and the NN output. This might however be explained by the low correlation between Swarm and GRACE over the same region (Figure 6.13b box 4).

6.6. Sensitivity Analysis

This section takes a closer look at which parameters drive the NN results and how robust these results are. The first step was to analyze which areas in the input maps influence the resulting EWH trend the most. This was done by computing the Jacobian of the ANN. Secondly, the error over the ocean and the effect on the total NN accuracy are discussed.

6.6.1. Neural Network Jacobian

As introduced in Section 4.3, the Jacobian of a NN is used to analyze the sensitivity of the NN output to small changes in its input layer. By plotting the values from the NN Jacobian matrix, these sensitivities can be visualized and compared. In this way, the areas that influence the NN results the most can be identified. This is shown in Figure 6.14 for the ANN trained with both SM and Swarm.



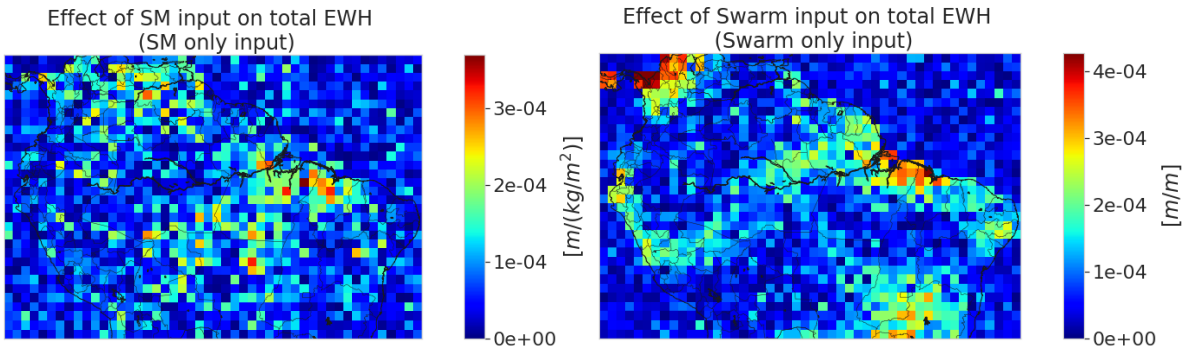
(a) Soil Moisture **(b)** Swarm

Figure 6.14: Absolute values of the averaged Jabobian values of output layer w.r.t to the grid cells of the SM (left) and Swarm (right) input maps. Each value on the map represents the sensitivity of the total EWH to that particular grid cell in the input layer.

Depending on the input data type, the regions in the input maps with the greatest average effect on the ANN output maps vary. For the SM maps (Figure 6.14a) the NN output is most sensitive to the central region of the study area. This is different for the Swarm maps (Figure 6.14b).

In Figure 6.15, the similar sensitivities are plotted for the ANNs trained with only SM and only Swarm

respectively. Even though the results from these ANNs were found to be inferior to the combined solution, comparing these sensitivity plots to those in Figure 6.14 provides useful information on how the NN sensitivity changes when combining multiple input data types.



(a) ANN trained with SM

(b) ANN trained with Swarm

Figure 6.15: The averaged Jabobian values of output layer w.r.t to the grid cells of the input maps, for the ANN models with a single input datatype. Each value on the map represents the sensitivity of the total EWH to that particular grid cell in the input layer.

Comparing Figure 6.15a to Figure 6.14a and Figure 6.15b to Figure 6.14b shows the difference in sensitivity of the NN output between the ANN trained with a single input data type and the ANN trained with both SM and Swarm. For both the SM (a) and Swarm (b) sensitivity maps, the driving regions are more spread over the study region for the SM + Swarm ANN compared to the climate-only ANN, where the driving regions are more concentrated. Secondly, when comparing the SM sensitivity maps to the Swarm sensitivity maps in Figure 6.15, it can be observed that the ANN output maps are more sensitive to ocean grid cells in the SM maps than ocean grid cells in the Swarm maps. This sensitivity to input ocean grid cells, whose values are all equal to 0, induces an error in the final result. This error is discussed in the next subsection.

6.6.2. Error Over Ocean

Before the input maps were fed into the NNs, an ocean mask was applied. This was done by setting all the values for grid cells over the ocean area to zero. Ideally, the NN should learn that the values for these grid cells in the output maps should always be equal to zero. This is not perfectly captured by the NNs, introducing a small error in the results. To quantify this error, the NSE was divided into two separate values; one for the land area (NSE_{land}) and one for the ocean area (NSE_{ocean}). The results are shown in Figure 6.16

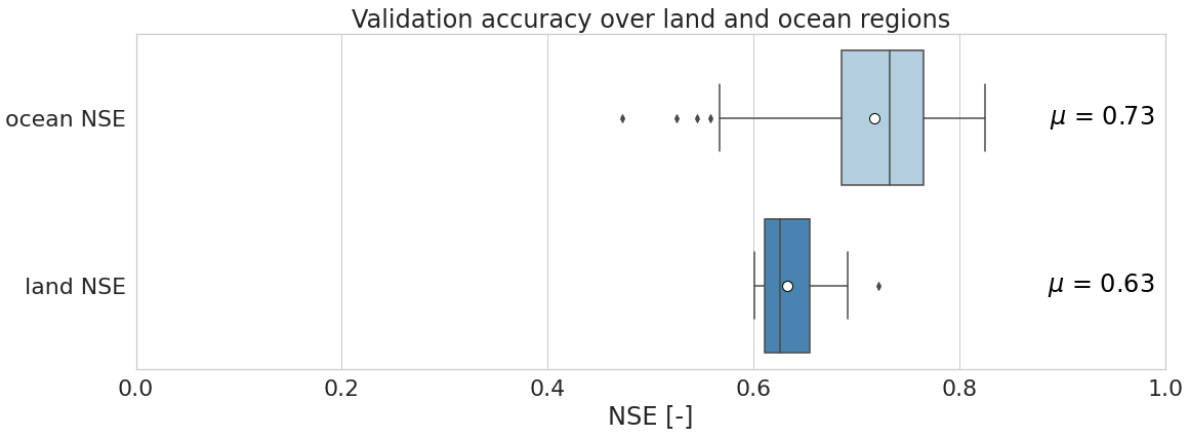


Figure 6.16: Boxplot of the validation accuracy over the land grid cells and the ocean grid cells.

Two observations can be made from Figure 6.16. Firstly, the average ocean NSE is smaller than 1.

This indicates that the NN does not fully account for the fact that the ocean grid cell values should be equal to 0 for each month. There exists a 27% error over the ocean in the NN output. When compared to the land NSE, the accuracy over the ocean is clearly higher, which was to be expected keeping in mind that the ocean grid cells should all be equal to zero and are therefore easy to predict. The second conclusion is that the average accuracy over land is actually 4%-5% lower than the total average accuracy shown in Figure 6.7.

6.7. Comparison to Previous Studies

After the results obtained by the NN were analyzed, it was time to compare these results to the previous studies discussed in Chapter 2. Table 6.2 summarizes the gap-filling results of these studies, together with their respective study regions, methods and input data combinations.

Study	Study region	Method	Input data	Results
Li et al. (2019)	Heihe River Basin (China)	SSA/Arima	GRACE	NSE = 0.35 CC = 0.67
Richter et al. (2021)	Mississippi Basin	PCA	Swarm	RMSE = 0.02 m
Teixeira da Encarnacao (2020)	Amzon Basin	Time series analysis	Swarm	CC = 0.95
Long et al. (2014)	Upper Mekong river	ANN	SM, precipitation	CC = 0.91 RMSE = 0.028 m
Li et al. (2020)	Amazon Basin	MLR	Precipitation, temperature, SM, climate indices, SST	CC: 0.93-0.97
		ANN		CC: 0.91-0.96
		ARX		CC: 0.91-0.95
Sun et al. (2019)	India	CNN	Precipitation, temperature, NOAH SM	NSE = 0.87 CC = 0.94
Sun et al. (2020)	Amazon Basin	SARIMAX	Precipitation, temperature, SM	NSE: 0.59-0.89
		DNN		
Fereira et al. (2019)	West-Africa	NARX	Precipitation, temperature, SM, climate indices, evaporation	NSE: 0.78-0.91 CC: 0.78-0.91
Ahmed et al. (2019)	African Watersheds	NARX	Precipitation, temperature, climate indices, evaporation	NSE: 0.54-0.94 CC: 0.79-0.97
Harrison (2023)	Amazon Basin	ANN (MLP)	SM, Swarm	NSE = 0.68 CC = 0.95
		CNN		NSE = 0.65 CC = 0.96

Table 6.2: The performance of the previous gap-filling studies considered for this research.

These results are visualized by use of a bar plot in Figure 6.17.

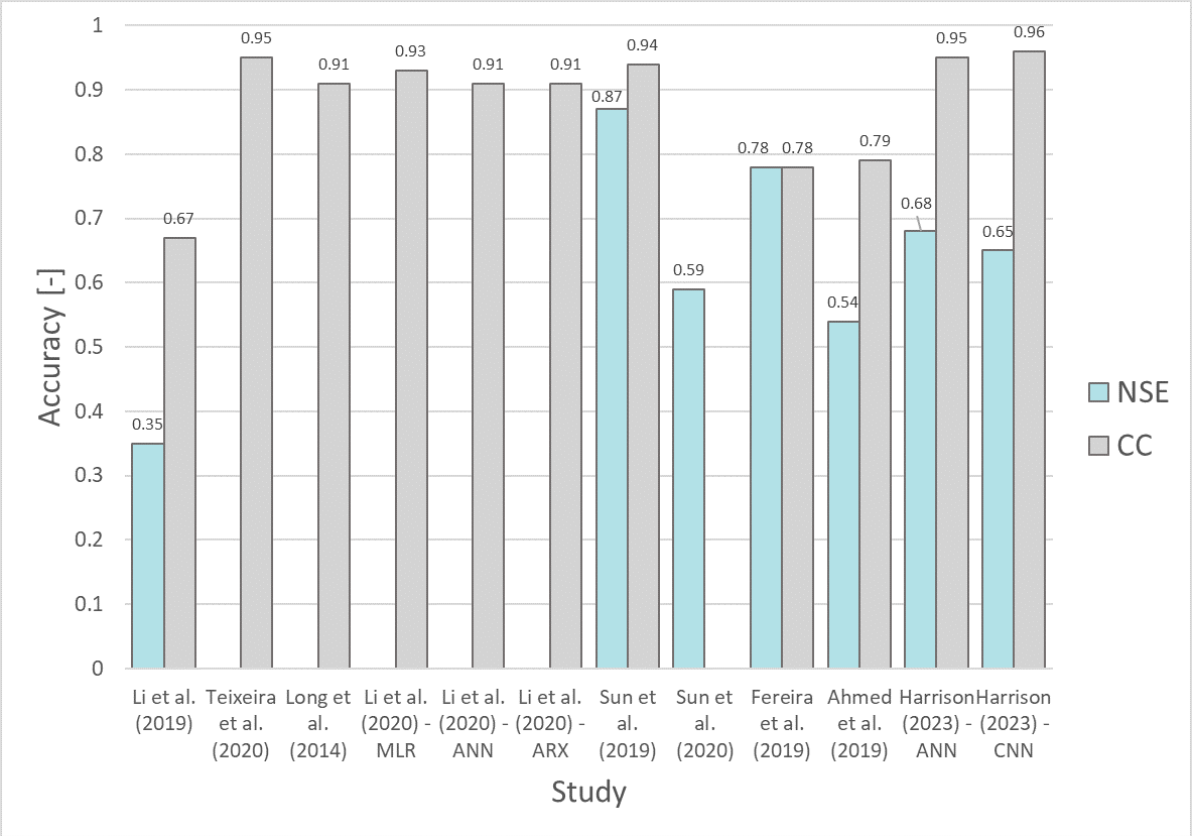


Figure 6.17: Bar plot of accuracies obtained in the previous gap-filling studies, compared to the accuracies obtained in this study.

It becomes clear that in terms of gap-filling performance, data-driven studies with hydro-climatological input data are the main competitors for this research. These studies outperform the studies that do not use data-driven techniques in both spatial and temporal accuracy (0.35 vs. 0.70 for average NSE, and 0.81 vs. 0.88 for average CC). By looking at the last column in Table 6.2 and Figure 6.17, the CC obtained in this research (0.96-0.97) competes with the highest values previously obtained (0.94-0.95). However, in terms of NSE, it is hard to state which previous studies outperform this research and vice versa. Most of these studies do not use NSE as a performance indicator. The studies that did use NSE in their analysis, often listed a range of NSE values, making it hard to compare. The most relevant studies in this list are the three studies that also use the Amazon Basin as a study region. If only these studies are considered, the results from this research agree with the previous research, both in terms of temporal accuracy (CC: 0.95 vs. 0.59-0.97) and spatial accuracy (NSE:0.68 vs. 0.59-0.89).

6.8. Comparison of Spatial Resolution

One of the important aspects of the research question and objectives is the down-scaling of Swarm data. This means that the ANN output maps have a higher spatial resolution than the Swarm input maps. This becomes clear when we compare the ANN output maps to the Swarm maps over the gap period. The ANN output maps over the GRACE/GRACE-FO gap can all be found in Appendix A. Figure 6.18 compares the ANN output map with the Swarm input map for the first month of the data gap.

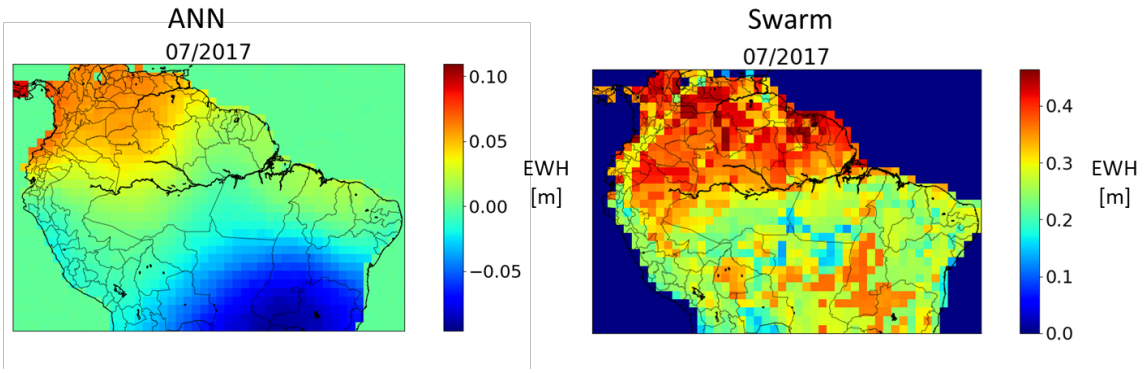


Figure 6.18: Comparison of the ANN output map and the Swarm map for the first month of the GRACE/GRACE-FO data gap.

The resolution of the ANN map is clearly higher than that of the Swarm map. This follows from the fact that the ANN is trained to mimic the GRACE data as closely as possible, therefore also producing maps with GRACE-like resolution. This is validated by comparing Figure 6.18 to Figure 3.2. This confirms the statement that Swarm data and hydro-climatological data were combined as input data to an ML model to fill the GRACE-/GRACE-FO data gap by **increasing the spatial resolution of Swarm-derived EWH maps**.

7

Conclusion & Recommendations

The goal of this research was to check whether the combination of hydro-climatological data and Swarm-derived EWH data could be used to bridge the GRACE/GRACE-FO EWH gap, using ML models. Section 7.1 summarizes the steps that were taken in this research and what resulted from them. From the work performed in this thesis, new questions arose. These, together with suggestions for future research are presented in Section 7.2

7.1. Conclusions

The first step was to decide which area of interest the research would center on. As described in Section 2.5, the Amazon basin was chosen as the study area because it is well-known for hydro-climatological research and has been demonstrated to have strong gravimetric signals that are well observed by both Swarm and GRACE. It has also been a region of interest in previous GRACE/GRACE-FO gap-filling studies, where it was shown that there was a strong correlation between GRACE and Swarm-derived EWH over the Amazon Basin

The best GRACE predictor of the hydro-climatological variables taken into consideration was found to be SM (Section 6.3). This was established by training a NN to replicate GRACE-like EWH with multiple input variable combinations. The simulations revealed that SM produced the best validation performance with the least amount of training time.

The first model that was developed was a Multi-Layer Perceptron ANN consisting of an input layer, three hidden layers, and an output layer. After optimization of the model's hyperparameters, the average validation accuracy after 1000 simulations was found to be 0.63. The simulation with the best performance yielded a validation NSE of 0.78. A more complicated CNN was built in addition to the ANN to see whether this complex model might improve validation any further. Two versions of the CNN, a 2D CNN and a 3D CNN were considered. The training accuracy of the 3D CNN was comparable to the 2D CNN, while the validation accuracy was somewhat greater. The complexity of the ANN is not a limiting factor to the performance of the NN, as shown by the fact that the 3D CNN did not considerably outperform the ANN. It was determined that the ANN is the best architecture for this project because of its short training time compared to the CNN (Section 5.10).

Stacking was considered to improve the performance, but it was concluded that this was not beneficial to the validation accuracy because it caused the models to overfit the training data (Section 5.9.2). The same holds for standardizing the input maps (Section 5.7).

Transfer learning was also considered as a way to include SM data prior to the Swarm period, but it did not increase the performance of the model as it caused a sudden drop in the model's performance after including the Swarm data in the input layer(Section 5.5).

It was also concluded that by omitting months of poor and medium Swarm EWH quality, the perfor-

mance of the model could be improved, even though 13 fewer months of training data were available (Section 5.6).

Finally, three gap-filling experiments of 1000 simulations each were performed with the optimised ANN. In the first experiment, the model was trained by only using SM from the 2004-2022 period as input data. The NN under fitted the GRACE time series and was not able to reproduce the local minima and maxima of GRACE. This resulted in a low validation and training accuracy of 0.43 and 0.85 respectively.

In the second experiment, only Swarm data from 2014-2022 were used to train the NN. The NN overfitted the GRACE data causing a large difference between the training and validation accuracy of 0.99 and 0.6 respectively. As a result of the NN not being able to find a general relation between the Swarm input set and the desired GRACE data, it should not be used to fill the GRACE/GRACE-FO data gap.

The last experiment combined the SM data with the Swarm data from 2014-2022 as input to the NN. This caused the NN to neither underfit nor overfit the GRACE time series, yielding the highest validation accuracy of 0.68 with a training accuracy of 0.83. It also resulted in high temporal accuracy with a CC of 0.95 between the NN output and the GRACE time series (Section 6.3).

The temporal accuracy of the NN EWH time series was found to be highest over the southeastern part of the study region and lowest over the northwestern part. This is most likely driven by the temporal correlation between the input data and GRACE over these regions.

With these results, it is possible to answer the research questions that were drawn up at the beginning of this research:

"Can the GRACE/GRACE-FO data gap be bridged by down-scaling Swarm gravity solutions to GRACE-like resolution by the use of ML with hydro-climatological input variables?"

Combining SM and Swarm as input data to a NN in order to create a GRACE-like EWH time series is feasible. A temporal Pearson correlation of 0.95 can be obtained, albeit with an average spatial accuracy (NSE) of 0.68.

1. *Which combination of hydro-climatological input variables is the best predictor for GRACE-like TWSC?*
   SM is the best hydro-climatological variable when it comes to predicting GRACE TWSC. This was proven by training the ANN with multiple combinations of hydro-climatological variables. The model with SM as only input data outperformed the other models and also resulted in the lowest training time.
2. *What ML model/architecture yields the best performance?* The ANN with three hidden layers of size 1617x4 - 1617x3 - 1617x2 was determined to be the best for this research. This model was optimised and the hyperparameters are listed in Table 5.3. The CNN architecture's performance was similar, but it took on average 9 times longer to train.
3. *Does the addition of Swarm to the ML input data improve the performance of the model?* The addition of Swarm is crucial to the performance of the model. Without Swarm, the model under-fits the GRACE time series and is not able to find local maxima and minima. By adding Swarm, the model captures these features more easily, increasing both the validation accuracy and training accuracy.

Lastly, the results of this research were compared to the gap-filling results obtained in previous studies. This study agrees with the previous studies in terms of both temporal correlation and spatial accuracy over the Amazon Basin.

7.2. Recommendations for Future Work

During the course of this research, new possibilities arose. They were deemed to be out of the scope of this project but might be interesting to future research on ML methods to fill the GRACE/GRACE-FO data gap.

1) Re-do simulations in future
The longer Swarm and GRACE-FO operate at the same time, the more training data will become available. Running the simulations with more available Swarm months will point out whether the accuracy obtained in this research is limited by the data availability or by the ability of SM and Swarm to reproduce accurate GRACE-like EWH. If the data set becomes too large or complex for the ANN to handle, there might be a clear difference between the ANN and CNN model in terms of performance, in which case the CNN becomes favourable over the ANN.

2) Perform similar research for other regions
Neural networks are very powerful at learning specific tasks. The NNs created in this project are optimized to learn relationships between SM and swarm EWH on the one hand, and GRACE EWH on the other, but only for the Amazon basin. Other regions might experience different EWH variations that could be better captured by Swarm and/or SM. Therefore performing similar research in other hydrologically active regions will be interesting.

3) In-depth sensitivity analysis on Swarm uncertainty
The Swarm-derived EWH maps used as input to the models do not consist of exact values. They consist of estimates, together with measures of error. For this research, the estimates were assumed to be true values, not taking into account their uncertainty in these Swarm values. It might be interesting to perform a sensitivity analysis to assess how robust the NN is to uncertainty in the Swarm input maps.

Bibliography

[1] M. Ahmed, M. Sultan, T. Elbayoumi, and P. Tissot. Forecasting grace data over the african watersheds using artificial neural networks. *Remote Sensing*, 11, 7 2019. ISSN 2072-4292. doi: 10.3390/rs11151769.

[2] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[3] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, L. Farhan, and et al. Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions. *Journal of Big Data*, 8(1), 2021. doi: 10.1186/s40537-021-00444-8.

[4] D. F. Argus, Y. Fu, and F. W. Landerer. Seasonal variation in total water storage in california inferred from gps observations of vertical land motion. *Geophysical Research Letters*, 41:1971–1980, 3 2014. ISSN 00948276. doi: 10.1002/2014GL059570.

[5] M. Cheng and J. Ries. The unexpected signal in grace estimates of c_{20}. *Journal of Geodesy*, 91: 897–914, 8 2017. ISSN 0949-7714. doi: 10.1007/s00190-016-0995-5.

[6] J. Claes. *Thermosphere Modelling Using Machine Learning*. PhD thesis, 2019.

[7] J. T. da Encarnação, D. Arnold, A. Bezděk, C. Dahle, E. Doornbos, J. van den IJssel, A. Jäggi, T. Mayer-Gürr, J. Sebera, P. Visser, and N. Zehentner. Gravity field models derived from swarm gps data. *Earth, Planets and Space*, 68, 12 2016. ISSN 1880-5981. doi: 10.1186/s40623-016-0499-9.

[8] J. T. da Encarnação, P. Visser, D. Arnold, A. Bezdek, E. Doornbos, M. Ellmer, J. Guo, J. van den IJssel, E. Iorfida, A. Jäggi, J. Klokocník, S. Krauss, X. Mao, T. Mayer-Gürr, U. Meyer, J. Sebera, C. K. Shum, C. Zhang, Y. Zhang, and C. Dahle. Description of the multi-approach gravity field models from swarm gps data. *Earth System Science Data*, 12, 6 2020. ISSN 1866-3516. doi: 10.5194/essd-12-1385-2020.

[9] P. Ditmar, J. T. da Encarnação, and H. H. Farahani. Understanding data noise in gravity field recovery on the basis of inter-satellite ranging measurements acquired by the satellite gravimetry mission grace. *Journal of Geodesy*, 86:441–465, 6 2012. ISSN 0949-7714. doi: 10.1007/s00190-011-0531-6.

[10] Y. Fan. Climate prediction center global monthly soil moisture data set at 0.5° resolution for 1948 to present. *Journal of Geophysical Research*, 109(D10), 2004. doi: 10.1029/2003jd004345.

[11] K. Fang, C. Shen, D. Kifer, and X. Yang. Prolongation of smap to spatiotemporally seamless coverage of continental u.s. using a deep learning neural network. *Geophysical Research Letters*, 44, 11 2017. ISSN 0094-8276. doi: 10.1002/2017GL075619.

[12] V. Ferreira, S. Andam-Akorful, R. Dannouf, and E. Adu-Afari. A multi-sourced data retrodiction of remotely sensed terrestrial water storage changes for west africa. *Water*, 11, 2 2019. ISSN 2073-4441. doi: 10.3390/w11020401.

[13] F. Flechtner, K.-H. Neumayer, C. Dahle, H. Dobslaw, E. Fagiolini, J.-C. Raimondo, and A. Güntner. What can be expected from the grace-fo laser ranging interferometer for earth science applications? *Surveys in Geophysics*, 37:453–470, 3 2016. ISSN 0169-3298. doi: 10.1007/s10712-015-9338-y.

[14] E. Friis-Christensen, H. Lühr, D. Knudsen, and R. Haagmans. Swarm – an earth observation mission investigating geospace. *Advances in Space Research*, 41:210–216, 1 2008. ISSN 02731177. doi: 10.1016/j.asr.2006.10.008.

[15] F. Giorgi and R. Francisco. Uncertainties in regional climate change prediction: A regional analysis of ensemble simulations with the hadcm2 coupled aogcm. *Climate Dynamics*, 16(2-3):169–182, 2000. doi: 10.1007/pl00013733.

[16] T. Jayalakshmi and A. Santhakumaran. Study about statistical normalization and back propagation. *Novel Research Aspects in Mathematical and Computer Science Vol. 2*, page 33–42, 2022. doi: 10.9734/bpi/nramcs/v2/2208b.

[17] J. Kim and S. W. Lee. Flight performance analysis of grace k-band ranging instrument with simulation data. *Acta Astronautica*, 65:1571–1581, 12 2009. ISSN 00945765. doi: 10.1016/j.actaastro.2009.04.010.

[18] D. P. K. Kingma and J. L. Ba. *Internal Conference on Learning Representations*, page 1–13. URL `https://arxiv.org/abs/1412.6980`.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. doi: 10.1145/3065386.

[20] Li, Wang, Zhang, Wen, Zhong, Zhu, and Li. Bridging terrestrial water storage anomaly during grace/grace-fo gap using ssa method: A case study in china. *Sensors*, 19, 9 2019. ISSN 1424-8220. doi: 10.3390/s19194144.

[21] F. Li, J. Kusche, R. Rietbroek, Z. Wang, E. Forootan, K. Schulze, and C. Lück. Comparison of data□driven techniques to reconstruct (1992–2002) and predict (2017–2018) grace□like gridded total water storage changes using climate inputs. *Water Resources Research*, 56, 5 2020. ISSN 0043-1397. doi: 10.1029/2019WR026551.

[22] D. Long, Y. Shen, A. Sun, Y. Hong, L. Longuevergne, Y. Yang, B. Li, and L. Chen. Drought and flood monitoring for a large karst plateau in southwest china using extended grace data. *Remote Sensing of Environment*, 155, 12 2014. ISSN 00344257. doi: 10.1016/j.rse.2014.08.006.

[23] C. Lück, J. Kusche, R. Rietbroek, and A. Löcher. Time-variable gravity fields and ocean mass change from 37 months of kinematic swarm orbits. *Solid Earth*, 9, 3 2018. ISSN 1869-9529. doi: 10.5194/se-9-323-2018.

[24] U. Meyer, K. Sosnica, D. Arnold, C. Dahle, D. Thaller, R. Dach, and A. Jäggi. Slr, grace and swarm gravity field determination and combination. *Remote Sensing*, 11, 4 2019. ISSN 2072-4292. doi: 10.3390/rs11080956.

[25] R. Rietbroek, M. Fritsche, C. Dahle, S. E. Brunnabend, M. Behnisch, J. Kusche, F. Flechtner, J. Schröter, and R. Dietrich. Can gps-derived surface loading bridge a grace mission gap? *Surveys in Geophysics*, 35:1267–1283, 11 2014. ISSN 15730956. doi: 10.1007/s10712-013-9276-5.

[26] M. Rodell, P. R. Houser, U. Jambor, J. Gottschalck, K. Mitchell, C.-J. Meng, K. Arsenault, B. Cosgrove, J. Radakovich, M. Bosilovich, and et al. The global land data assimilation system. *Bulletin of the American Meteorological Society*, 85(3):381–394, 2004. doi: 10.1175/bams-85-3-381.

[27] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale-image recognition. *ICLR 2015 Conference Paper*, 2015. doi: 10.48550/arXiv.1409.1556.

[28] A. Y. Sun, B. R. Scanlon, Z. Zhang, D. Walling, S. N. Bhanja, A. Mukherjee, and Z. Zhong. Combining physically based modeling and deep learning for fusing grace satellite data: Can we learn from mismatch? *Water Resources Research*, 55, 2 2019. ISSN 0043-1397. doi: 10.1029/2018WR023333.

[29] Z. Sun, D. Long, W. Yang, X. Li, and Y. Pan. Reconstruction of grace data on changes in total water storage over the global land surface and 60 basins. *Water Resources Research*, 56, 4 2020. ISSN 0043-1397. doi: 10.1029/2019WR026250.

[30] B. D. Tapley, S. Bettadpur, J. C. Ries, P. F. Thompson, and M. M. Watkins. Grace measurements of mass variability in the earth system. *Science*, 305:503–505, 7 2004. ISSN 0036-8075. doi: 10.1126/science.1099192.

[31] B. D. Tapley, M. M. Watkins, F. Flechtner, C. Reigber, S. Bettadpur, M. Rodell, I. Sasgen, J. S. Famiglietti, F. W. Landerer, D. P. Chambers, J. T. Reager, A. S. Gardner, H. Save, E. R. Ivins, S. C. Swenson, C. Boening, C. Dahle, D. N. Wiese, H. Dobslaw, M. E. Tamisiea, and I. Velicogna. Contributions of grace to understanding climate change. *Nature Climate Change*, 9:358–369, 5 2019. ISSN 1758-678X. doi: 10.1038/s41558-019-0456-2.

[32] J. van den IJssel, J. Encarnação, E. Doornbos, and P. Visser. Precise science orbits for the swarm satellite constellation. *Advances in Space Research*, 56:1042–1055, 9 2015. ISSN 02731177. doi: 10.1016/j.asr.2015.06.002.

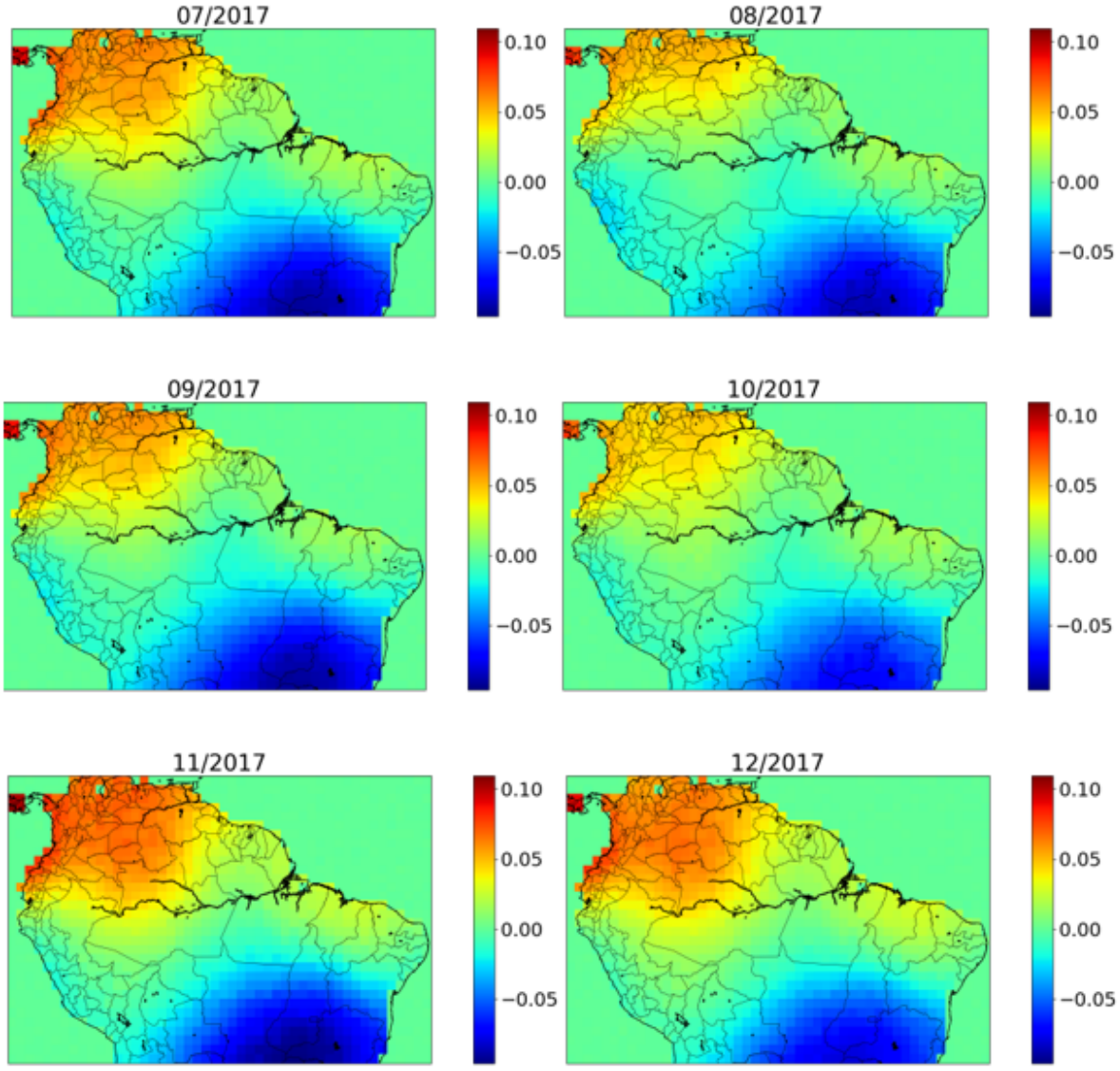
[33] J. Wahr, M. Molenaar, and F. Bryan. Time variability of the earth's gravity field: Hydrological and oceanic effects and their possible detection using grace. *Journal of Geophysical Research: Solid Earth*, 103:30205–30229, 12 1998. ISSN 01480227. doi: 10.1029/98JB02844.

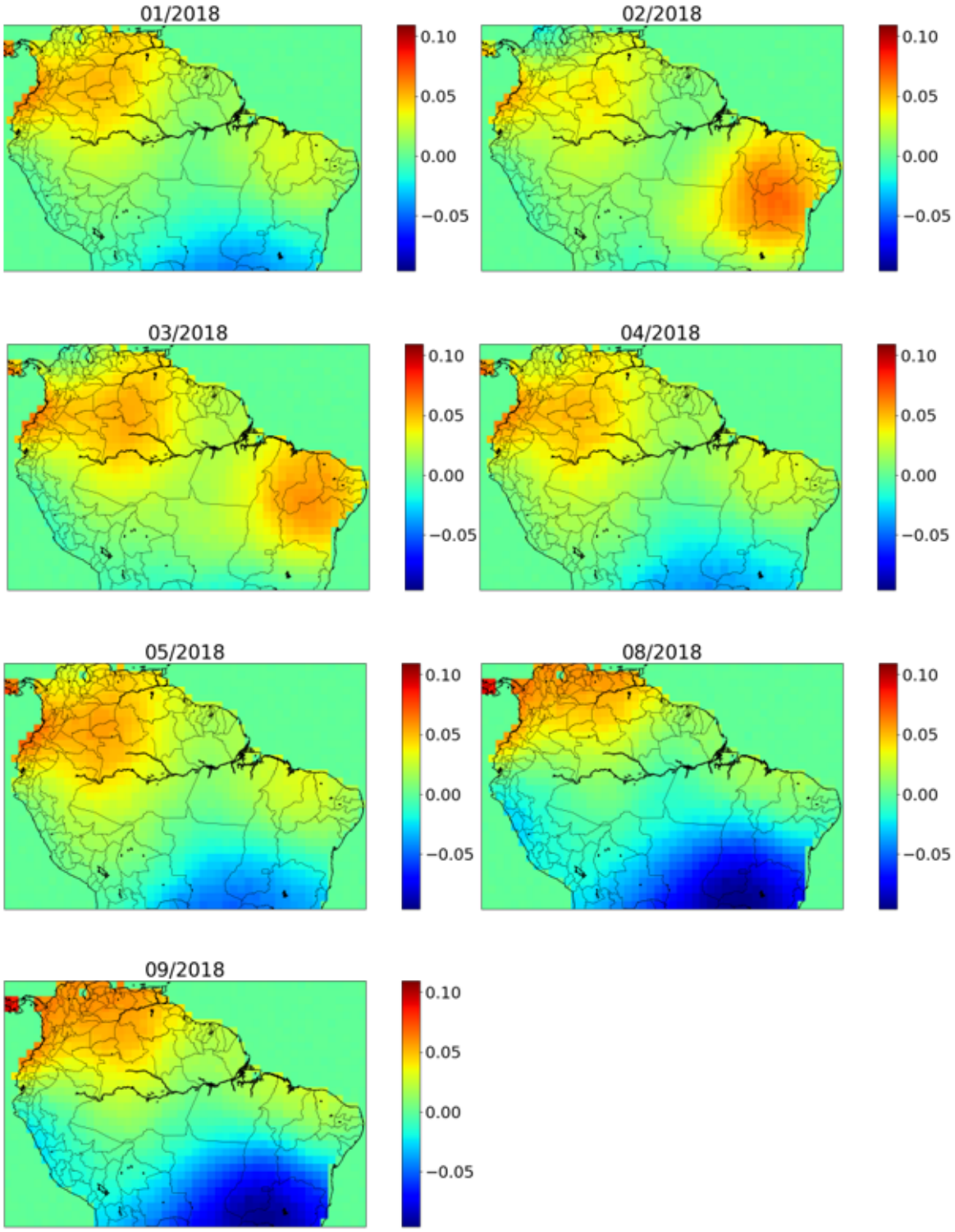
[34] B. Zhong, X. Li, J. Chen, Q. Li, and T. Liu. Surface mass variations from gps and grace/gfo: A case study in southwest china. *Remote Sensing*, 12, 6 2020. ISSN 2072-4292. doi: 10.3390/rs12111835.

A

ANN output EWH maps over the GRACE/GRACE-FO data gap

The figures listed below show the output of the trained ANN over the Amazon basin during the GRACE/GRACE-FO mission gap. The values in the graphs represent the average EWH values for a grid cel **in m**.


07/2017
0.10
0.05
0.00
−0.05
08/2017
0.10
0.05
0.00
−0.05
09/2017
0.10
0.05
0.00
−0.05
10/2017
0.10
0.05
0.00
−0.05
11/2017
0.10
0.05
0.00
−0.05
12/2017
0.10
0.05
0.00
−0.05


01/2018
0.10
0.05
0.00
−0.05
02/2018
0.10
0.05
0.00
−0.05
03/2018
0.10
0.05
0.00
−0.05
04/2018
0.10
0.05
0.00
−0.05
05/2018
0.10
0.05
0.00
−0.05
08/2018
0.10
0.05
0.00
−0.05
09/2018
0.10
0.05
0.00
−0.05