Analysis and Prediction of Long Term GNSS Height Time Series and Environmental Loading Effects

Master Thesis

Author(s):
Ruttner, Pia

Publication date:
2021-01-25

Permanent link:
https://doi.org/10.3929/ethz-b-000519390

Rights / license:
In Copyright - Non-Commercial Use Permitted

Master's Thesis

Analysis and Prediction of Long Term GNSS Height Time Series and Environmental Loading Effects

Pia Ruttner

January 25, 2021

Institute of Geodesy and Photogrammetry
Space Geodesy
ETH Zurich

Supervisors

Prof. Dr. Benedikt Soja
Dr. Roland Hohensinn

Acknowledgment

First, I dearly want to thank my supervisors, Dr. Roland Hohensinn and Prof. Dr. Benedikt Soja, who made this thesis possible guided me through it. Special thanks belong to Dr. Stefano D'Aronco, Dr. Jan Wegner and Laura Crocetti, who played a big part in helping me find my way around the machine learning part of this thesis. Furhtermore, I want to thank Jens, Jon and Nicholas, for their great job of proofreading my manuscript. Last, but most, I would like to thank Simon, who unconditionally hold my back free and never stoped encouraging and believing in me.

Abstract

GNSS height residuals often exhibit seasonal amplitudes, that can partly be explained by environmental surface loadings, such as hydrological, non-tidal atmospheric and non-tidal oceanic loading. In this thesis the state of the art procedure to reduce the height component of GNSS residuals in Europe by those environmental loadings is evaluated, with a focus on the residual RMS and amplitude reduction. On the one hand, the environmental loadings are subtracted directly from the residuals and on the other hand their annual component is reconstructed using Singular Spectrum Analysis (SSA), which is then used for the reduction. In the second part of this thesis, more complex relationships between GNSS height residuals and environmental influences are explored. Temporal Convolutional Networks (TCN) and Random Forests (RF) are used to model and predict the GNSS height residuals, using environmental loadings, raw meteorological data and tropospheric zenith delays as input features. The RMS and amplitude of the GNSS height residuals could be reduced to a median of up to 19% and 23% when using the original loading series, respectively. An RMS reduction of an average of 2% was obtained, by removing the loadings series after first reconstructing them with SSA. The majority of tested GNSS stations benefit from the inclusion of environmental parameters in their residual prediction. These yield improvements of up to 6% in prediction error, when compared to a prediction using GNSS residual data only.

Contents

CONTENTS

1 Introduction

More than two decades ago the first permanent geodetic Global Navigation Satellite System (GNSS) receivers were installed. Today, we benefit from long term position estimates, which allow the observation of GNSS signals down to a small millimeter range. The height component in station position residuals can provide detailed information into the Earth's crust vertical movements, which include long term trends, for example attributed to postglacial rebound, seasonal variations, as well as sudden, fast movements caused by earthquakes. The seasonal movements exhibit, among others, annual and semi-annual periodicities which can be partly attributed to vertical land motion due to environmental surface loadings, of which hydrological (HYDL), non-tidal atmospheric (NTAL), and non-tidal oceanic (NTOL) loading are included. The non-tidal term refers that these movements are not related to the mass attractions to the Sun and the Moon.

Already, Dam et al. (1994) found that around 24% of the total GNSS height variance can be explained through change in atmospheric pressure. In Dam et al. (1997), atmospheric and oceanic loadings were correlated with geoid deformation and therefore movement in GPS height residuals. Later, Dam et al. (2001) performed a more detailed analysis using global GNSS solutions, observing a correlation between vertical crustal movement due to continental water storage.

Gegout et al. (2010) undertook a study on how loading displacements can be identified and quantified in the GNSS solution. They also investigated if the a priori introduction of environmental loadings in the GNSS processing leads to an improvement in the estimation of the GNSS height component. In doing so, they were able to improve the GNSS height residuals in the northern, but not in the southern hemisphere, where most of the stations used in the study are located near the ocean. Gegout et al. (2010) thus claim that the remaining, not yet explainable, residuals could be related to mismodeled tropospheric propagation.

Further research has been performed with the goal to reduce displacement variability by including environmental loadings, a better understanding of error sources in general and an improvement of the geophysical interpretation of the GNSS signal. By doing so, it has been found that the quality of GNSS station positions propagate into the estimation of the International Terrestrial Reference Frame (ITRF) and its velocities, as well as in the estimation of earth orientation parameters (Collilieux et al., 2010; Santamaría-Gómez and Mémin, 2015).

More recent studies focus on the comparison of different loading data models (Bian, 2020; Jiang et al., 2013; Wu et al., 2020). It was also found that the assimilation of loading models with information gained from the Gravity Recovery and Climate Experiment (GRACE) lead to further improvement in root mean square error (RMS) reduction (Klos et al., 2020). Although, a high RMS reduction of the GNSS residual time series can also be misleading. Dong et al.

(2002), Klos et al. (2019), and Penna and Stewart (2003) showed that the variation at certain frequencies, at around 4-80 cycles per year, has to be taken with caution, as the signal rather contains systematic errors, than having real geophysical meaning.

While most studies concentrate on yearly fluctuations, there is also evidence that higher frequencies in the GNSS residuals can be explained through HYDL (Springer et al., 2019). Mémin et al. (2020) have also corrected GNSS data for non-tidal loadings by incorporating the ocean's response to air-pressure changes. In their global study, they found that the proportion of the variability in vertical land motion due to non-tidal loadings decreases, with decreasing latitudes. Up to today, the influence of environmental loadings on GNSS residual time series is still not fully understood. There is still no standardized method on how to include loading data into GNSS position solutions (Mémin et al., 2020), but on the other hand environmental loadings are not the sole explanation for seasonal patterns in GNSS residuals. Several studies suggest that the remaining residuals may correlate with errors in the tropospheric zenith delay estimation (Gegout et al., 2010; Mémin et al., 2020).

The goal of this thesis is to reduce the RMS error in GNSS height residuals through a posteriori introduction of environmental loadings into the station position estimate. As part of this, influences from the raw loading displacements, and with Singular Spectrum Analysis (SSA) reconstructed loading series, are evaluated. Moreover, using machine learning algorithms, the complex relationships of GNSS residuals and environmental influences, are explored. These algorithms utilize data from surface loading models, tropospheric zenith delays and raw meteorological data.

The thesis is broken up into five chapters. After this introduction, the second chapter presents the data used in this thesis, followed by a closer look into the most important employed methodologies. The fourth chapter presents the results and an in-depth analysis of these. The final chapter provides a summary and conclusion for the work performed in this thesis, and an outlook for where future research could go.

2 Data and Preprocessing

This Chapter provides an overview of the different datasets used and pre-processing applied, before the data is utilized within the main analysis.

2.1 GNSS Data

Two different GNSS height residual datasets are used at the core of this thesis. The first of which is the station height residuals from the EUREF Permentent GNSS Network (EPN) and the second being the 24 hour final solutions from the Nevada Geodetic Laboratory (NGL). Figure 2.1 shows an overview of the station distribution and information on the time series lengths.
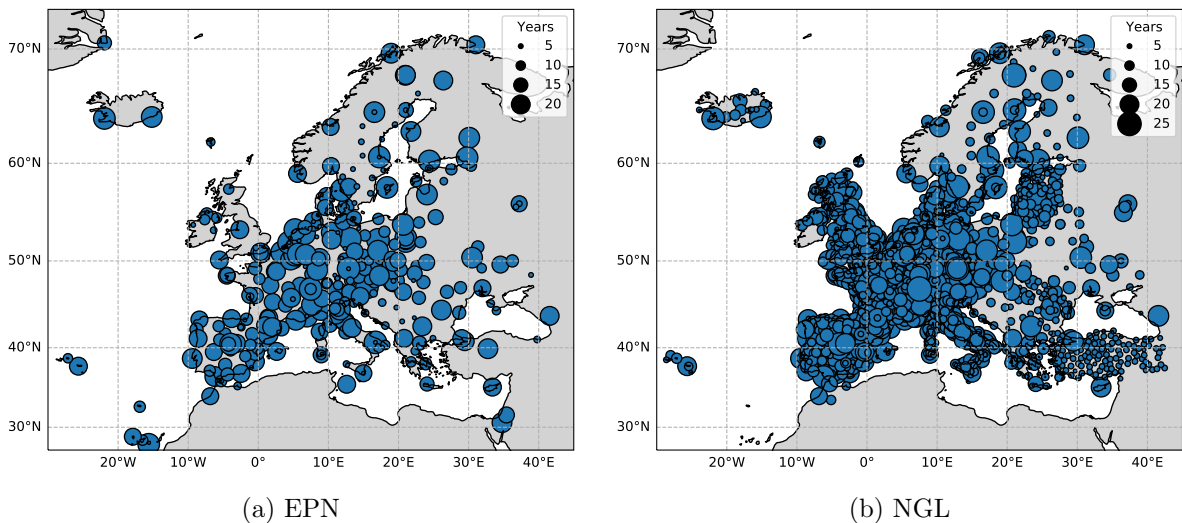

(a) EPN   (b) NGL

Figure 2.1: Station distribution of the (a) EPN and (b) NGL GNSS dataset. The different sizes of markers indicate the observation length at each station location

2.1.1 European Permanent Network

The multi-year European Permanent Network (EPN) solution (Bruyninx et al., 2019), provided by the Royal Observatory of Belgium (`http://epncb.oma.be/`), is composed of the epn-repro2 daily combinations until GPS week 1772, and of routine daily solutions afterwards.

Within their residuals, all discontinuities are removed and outliers are eliminated. No annual or semi-annual seasonal signals are removed. Nevertheless, it should be noted that the EPN solution is derived from a regional combination obtained from the Combination and Analysis of Terrestrial Reference Frames Software (CATREF) (Altamimi et al., 2007). This combination

procedure leads to the elimination of a common mode error, which causes a reduction of seasonal signals, compared to a global solution (Legrand et al., 2012).

From all available stations, those that are used that cover a time span of minimum 3.5 years and that have a maximum of 20% of missing observations. This results in a dataset containing 323 stations with an average data length of 14 years. All data within this dataset were captured between January 1996 to April 2020 and have daily sampling, with the residual values given in millimeters.

2.1.2 Nevada Geodetic Laboratory

The Nevada Geodetic Laboratory collects data from all available geodetic GNSS stations worldwide. They collect the raw data, process it and make it publicly available (Blewitt et al., 2018). As the focus of this study is on Europe, the EU-Plate fixed 24 hour final solutions are used. The dataset has a time span from January 1994 to October 2020, with a daily sampling rate and the observation values given in millimeters. The same filtering strategy that was used for the EPN dataset was also used here, which results in 2476 stations, and yields on average 10 years of data per station. Additionally, all outliers, jumps and linear trends are removed using the software package *Hector* (Bos et al., 2012).
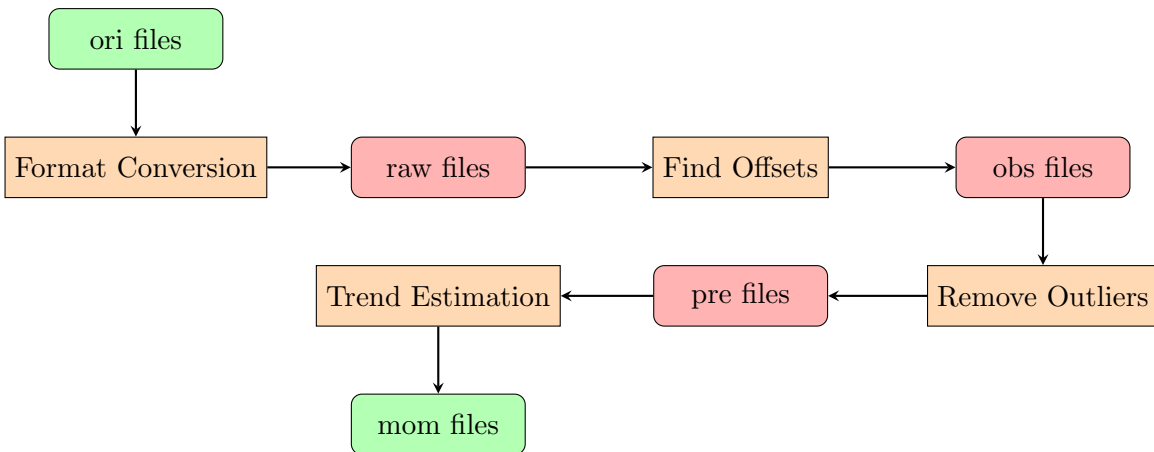

Figure 2.2: Flowchart of the NGL data preprocessing in Hector. The input and output files are marked in green. Orange blocks are the processing steps and intermediate files are colored in red.

Figure 2.2 illustrates the processing steps done with *Hector*, following the convention proposed in Bos and Fernandes (2019). The original *tenv3* format files (*ori files*), downloaded from NGL (http://geodesy.unr.edu/index.php), are converted into *mom* format (*raw files*) for acceptance into *Hector*. These are then used to find all offsets in the time series, which are stored in the header of the resulting *obs files*. Next, outliers are determined and the *pre files*

are created which have these outliers removed. Finally, the linear trend is estimated and stored within the *mom files*. These resulting files contain three columns. The first one represents the time in Modified Julian Date (MJD), the second contains the residuals and the third column represents the estimated trend. The underlying methodology used in the processing steps is described in Section 3.1.

2.2 Environmental Loading Data

The other major dataset used throughout this work is that describing the vertical displacements, caused by environmental surface loading effects. This data is taken from the Earth System Modelling Group Repository of Deutsche Geoforschungszentrum Potsdam (ESMGFZ).
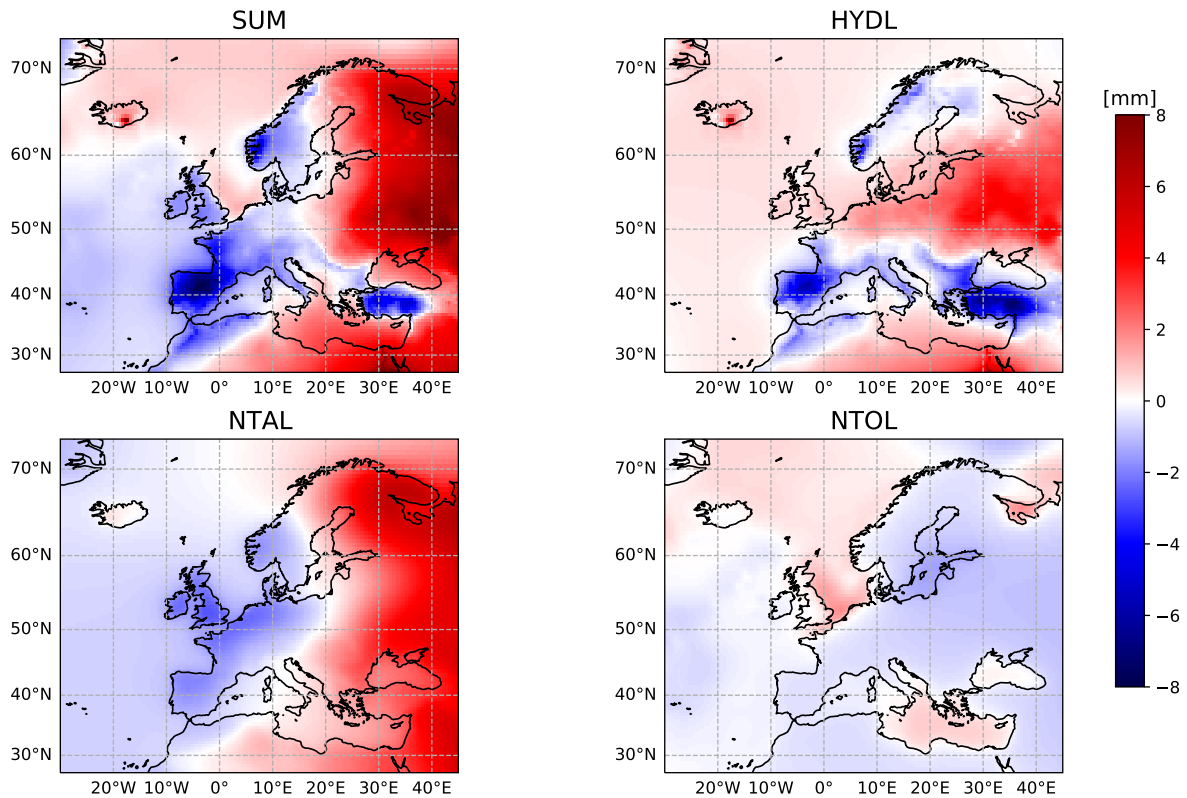

Figure 2.3: Vertical displacements of of environmental surface loadings, on the 17th of June 2015

Hydrological loadings (HYDL) are derived from the Land Surface Discharge Model (LSDM), taking into account surface water, snow, soil moisture, rivers, and lakes. The LSDM is in turn forced by atmospheric data from the European Center for Medium-Range Weather Forecasts (ECMWF) (Dill and Dobslaw, 2013; Dill, 2008). Non-tidal atmospheric (NTAL) and non-

tidel oceanic (NTOL) loadings are respectively derived from non-tidal parts of the atmospheric surface pressure of ECMWF and the ocean bottom pressure provided by the Max-Planck-Institute for Meteorology Ocean Model (MPIOM).

The different mass distribution models (LSDM, ECMWF, MPIOM) are then used with Green's function to calculate the elastic surface deformations. These computed loading displacements are available in meters on a 0.5°x0.5° grid. The sampling rate for HYDL is 24 hours whilst NTAL and NTOL use a sampling rate of 3 hours.

In order to obtain the loading data at the exact GNSS Station coordinates, they are downloaded using an interpolation-script, provided by ESMGFZ. This implements a bilinear interpolation based on the neighboring grid cells. Land/ocean tile information is also used to verify that no interpolation is done over the coastline to avoid adulterated results. In some areas, the original displacement time series of HYDL shows very large long-term drifts, which are reduced using a quadratic polynomial fit.
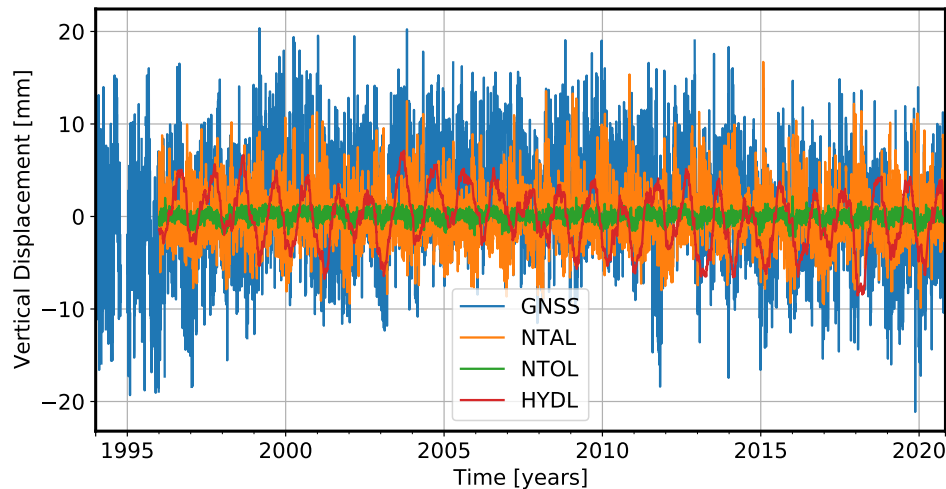

Figure 2.4: Vertical displacements of the station Zimmerwald, Switzerland (ZIMM), In blue the EPN GNSS residuals, in orange non-tidal atmospheric, in green non-tidal oceanic and in red hydrological loading.

Figure 2.3 shows the vertical displacements of the individual loadings for one particular day in June. The top-left figure shows the accumultaed loadings. The individual loadings are shown in the other three sub-plots. On the mainland the SUM is dominated by the NTAL and HYDL components. NTOL on the other hand, has lesser influence and is mostly observed around coasts. Figure 2.4 shows the temporal domain of HYDL, NTAL and NTOL, together with the EPN height residuals, at one example station in Zimmerwald, Switzerland (ZIMM). It can be seen that NTAL has a high variablity, almost at the level as the GNSS residuals. NTOL has the lowest amplitdues, compared to the other two components. This was expected as ZIMM is located in center of the european mainland. In contrast, HYDL shows a clear annual signal

with a rather high amplitude

2.3 European Climate Assessment and Dataset

Meteorological data is obtained from a pre-defined station subset of the European Climate Assessment and Dataset (ECAD).This includes daily mean temperature [Degree Celcius], precipitation amount [Millimeters], sea level pressure [Hectopascal], humidity [Percent] and radiation [Watt per Square Meter] (Tank et al., 2002). This dataset is blended, meaning that close stations, maximum 12.5km horizontal and 25m height differences, are merged to compensate for data gaps. Each GNSS is allocated to the closest meteo station, that has available time series for all parameters, and is not more than 50km horizontal distance away. The meteorological data is used in the second part of the thesis as an input feature for the GNSS residual predictions using machine learning, This usage also applies for the troposepheric zenith delay, described in the following paragraph.

2.4 Tropospheric Zenith Delay

Tropospheric total zenith delays are as well obtained from the Nevada Geodetic Laboratory. These are provided as millimetre values with a five minute sampling rate. Data is downloaded for all stations that are within both the EPN and NGL GNSS datasets, yielding a total of 290 stations.

3 Methodology

After the previous chapter introduced the datasets used, this chapter describes the basic methods which the presented analyses are based upon. The first part (Section 3.1) gives a recap of the basics in time series analysis, including the underlying trajectory model, Fourier transformation, the noise and stochastic model for Linear Trend Estimation, and Singular Spectrum Analysis. In Section 3.2, the general workflow, on how the proportion of environmental loadings in GNSS height residuals is evaluated, is explained. The last Section in this chapter (Section 3.3), gives a brief introduction into the time series prediction methods used for the second major analysis part of this thesis.

3.1 Time Series Analysis

3.1.1 Trajectory Model

By definition, coordinate time series are trajectories and, when describing a time series by a kinematic model, this model becomes a trajectory model (Montillet and Bos, 2020). The motion of GNSS stations can thus also be represented as a trajectory model, capturing the station displacements in a kinematic description (Bevis and Brown, 2014). The Standard Linear Trajectory Model (SLTM) is a way to describe the kinematic model as the sum of three different motions. Its basic form, is presented briefly in the following, after Montillet and Bos (2020). The station vector $\mathbf{x}(t)$, can be decomposed into:

$$\mathbf{x}(t) = \mathbf{x}_{trend} + \mathbf{x}_{jumps} + \mathbf{x}_{cycle}, \tag{3.1}$$

where $\mathbf{x}_{trend}$ is the continuous trend, $\mathbf{x}_{jumps}$ accounts for sudden positional jumps and $\mathbf{x}_{cycle}$ represents periodic displacements. The components can be described in more detail as:

$$\mathbf{x}_{trend} = \mathbf{x}(t_R) + \mathbf{v}(t - t_R) \tag{3.2}$$

$$\mathbf{x}_{jumps} = \sum_{j=1}^{n_J} \mathbf{b}_j H(t - t_j) \tag{3.3}$$

$$\mathbf{x}_{cycle} = \sum_{k=1}^{n_F} [\mathbf{s}_k \sin(\omega_k t) + \mathbf{c}_k(\omega_k t)] \tag{3.4}$$

The continuous trend $\mathbf{x}_{trend}$ is represented by a reference positon $\mathbf{x}(t_R)$ plus the velocity vector $\mathbf{v}(t - t_R)$, here assumed as constant, where t_R denotes a reference time. The jumps in position

$\mathbf{x}_{jumps}$ do not only contain real motions as coseismic jumps, but also artificial motions in position due to changes in GNSS hardware, for example the antenna. Each jump at time t_j is described by a vector $\mathbf{b}_j$, which includes the direction and magnitude of the jump, and the Heaviside step function H, all summing up to a number of jumps n_J. The periodic desplacement vector $\mathbf{x}_{cyclic}$ is modeled as Fourier series with the Fourier coefficients $\mathbf{s}_k$ and $\mathbf{c}_k$. The number of frequencies is denoted as n_F and $\omega_k = 2\pi/\tau_k$ is the angular frequency with period τ_k. The next paragraphs shall give more detailed information about Fourier transformation and the computation of the power spectrum.

3.1.2 Fourier Transform

The Fourier Analysis of a time series gives insight into its frequency domain, revealing information on contained periodicities with their corresponding amplitudes. The Fourier Transform is based on the idea that any signal with finite variance can be represented by a sum of periodic components. Nowadays, the Fast Fourier Transform (FFT) algorithm proposed by Cooley and Tukey (1965), is most commonly used, due to its fast computation, as suggested by its name.

Given a time series y_n with length N, its Fourier transform into Y_k coefficients, which represent the phase and amplitude of a period k/NT, with T as observation span, denotes as (Bracewell, 1978; Buttkus, 2012; Montillet and Bos, 2020):

$$Y_k = \sum_{n=0}^{N-1} y_n \cdot e^{-i2\pi \frac{kn}{N}} \qquad \text{for } k = [-N/2 + 1, \ldots, N/2] \tag{3.5}$$

From the complex numbers Y_k, the one-sided power spectral density S_k can be computed with:

$$\begin{aligned}
S_0 &= |Y_0|^2/f_s \\
S_{N/2} &= |Y_{N/2}|^2/f_s \\
S_k &= 2|Y_k|^2/f_s \qquad \text{for } k = [1, \ldots, N/2 - 1],
\end{aligned} \tag{3.6}$$

where f_s is the Nyquist frequency and each S_k has a corresponding frequency f_k:

$$f_k = \frac{k f_s}{N} \qquad \text{for } k = [0, \ldots, N/2] \tag{3.7}$$

When all S_k are summed up, the variance of the noise can be obtained.

3.1.3   Noise Model

The modeling of noise is important because no real world geodetic time series observation is perfect and therefore contains noise, which can be described as set of multivariate random variables. The noise is often modeled as a Gaussian probability density function. This has the advantage that, if the mean is assumed to be zero, the covariance matrix is sufficient to describe the stochastic properties of the noise.

The most common representation of noise in GNSS time series is a combination of white and power-law noise. The noise covariance matrix $\mathbf{C}$ is written as sum of the two noise models (Montillet and Bos, 2020):

$$\mathbf{C} = \sigma_{pl}^2 \mathbf{J}(\kappa) + \sigma_w^2 \mathbf{I}, \tag{3.8}$$

where σ_{pl} and σ_w are the amplitudes of the power-law and white noise, respectively. $\mathbf{J}$ is the general power-law covariance matrix with the spectral index κ. Due to $\kappa = 0$ in the case of white noise, the covariance matrix simplifies to the identity matrix $\mathbf{I}$.

### 3.1.4   Linear Trend Estimation with Log-Likelihood

Time series observations $\mathbf{y}$ can be described as (Montillet and Bos, 2020):

$$\mathbf{y} = \mathbf{g}(\mathbf{x}, t) + \mathbf{w} \tag{3.9}$$

with the trajectory model $\mathbf{g}(\mathbf{x}, t)$, $\mathbf{x}$ describing the model parameters and t the time, and $\mathbf{w}$ being the noise. When assuming a zero mean Gaussian probability density function, the noise $\mathbf{w}$ can be represented in the covariance matrix $\mathbf{C}$. The true values of neither $\mathbf{x}$ nor $\mathbf{C}$ are known. Only the observations $\mathbf{y}$ are available. Therefore, the goal is to maximize the likelihood function $L = f(\mathbf{x}, \mathbf{C}|\mathbf{y})$, by searching for the values of $\mathbf{x}$ and $\mathbf{C}$ that give the highest probability of observing $\mathbf{y}$. The most common representation of L is its logarithmic form:

$$ln(L) = -\frac{1}{2} \left[N \ln(2\pi) + \ln \det(\mathbf{C}) + (\mathbf{y} - \mathbf{g}(\mathbf{x}, t))^T \mathbf{C}^{-1} (\mathbf{y} - \mathbf{g}(\mathbf{x}, t)) \right] \tag{3.10}$$

This method is called Maximum Likelihood Estimation (MLE). The MLE is implemented in the software package *Hector* (Bos et al., 2012), where the trajectory model parameters $\mathbf{x}$, as well as the noise parameters κ, σ_{pl} and σ_w can be estimated.
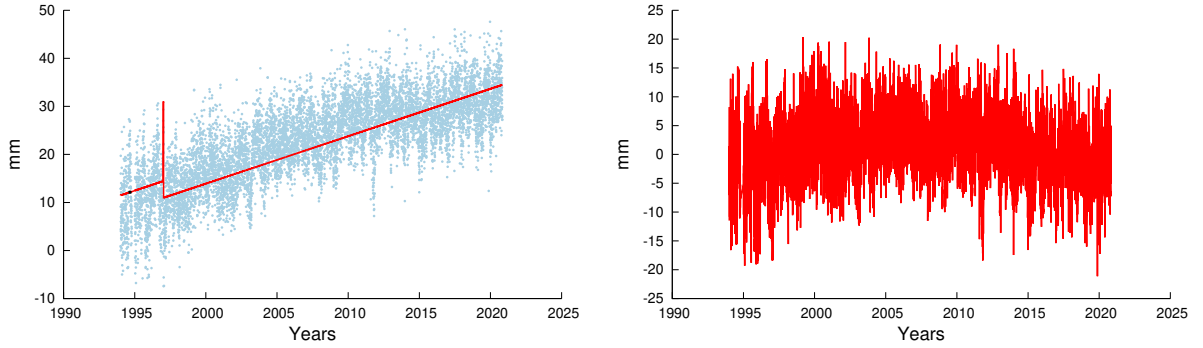
(a) Original observations and the estimated trajectory

(b) Residuals, obtained by subtracting the estimated trajectory from the original observations

Figure 3.1: NGL height observations and residual computation of station ZIMM (*figure adopted from Hector output (Bos et al., 2012)*)

In this work *Hector* is used for the preprocessing of the NGL dataset (Section 2.1.2). Figure 3.1 shows the NGL height time series of the example station ZIMM. On the left, the original observations (reduced to the first observation value) are shown in blue, together with the estimated trajectory model, using MLE, as red line. The configuration was set to estimate the trajectory model including jumps and a linear trend, while assuming a power-law plus white noise model. On the right, the resulting residuals are shown in red, which are computed by subtracting the estimated trajectory model from the original observations.

3.1.5 Singular Spectrum Analysis

Singular Spectrum Analysis (SSA) is a different spectral estimation method, independent of the previously described models. SSA is used to reconstruct the annual components of the environmental loading data as proposed by Klos et al. (2019). SSA is a data-driven method that does not require any prior assumptions about the analyzed time series (Broomhead and King, 1986; Vautard and Ghil, 1989; Vautard et al., 1992). The time domain information of the time series is used to model its Empirical Orthogonal Functions (EOF). Their first components represent the trend and seasonal signals, mostly annual and semi-annual, of the analyzed time series. SSA can be used for various applications in time series analysis, including smoothing, investigation of different periodicities a time series is composed of, removal of trends and seasonalities, or the reconstruction of a time series using only specific components.

The analysis of a time series with SSA can be divided into two main parts: decomposition and reconstruction. For the decomposition, the time series is embedded into a trajectory matrix. With the time series length N and a window size L, $K = N - L + 1$, lagged vectors are formed and arranged into a matrix $\mathbf{X}$:

$$\mathbf{X} = [X_1, \ldots, X_K] = (x_{ij})_{i,j=1}^{L,K} = \begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_K \\ x_2 & x_3 & x_4 & \cdots & x_{K+1} \\ x_3 & x_4 & x_5 & \cdots & x_{K+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & x_{L+2} & \cdots & x_N \end{pmatrix} \quad (3.11)$$

Each column is a lagged vector

$$X_i = (x_i, \ldots, x_{i+L-1})^T \quad (1 \le i \le K)$$

and is therefore a subseries of the original time series. The same applies to the rows. The resulting trajectory matrix is called *Hankel matrix*, which has the property that all anti-diagonal elements are equal.
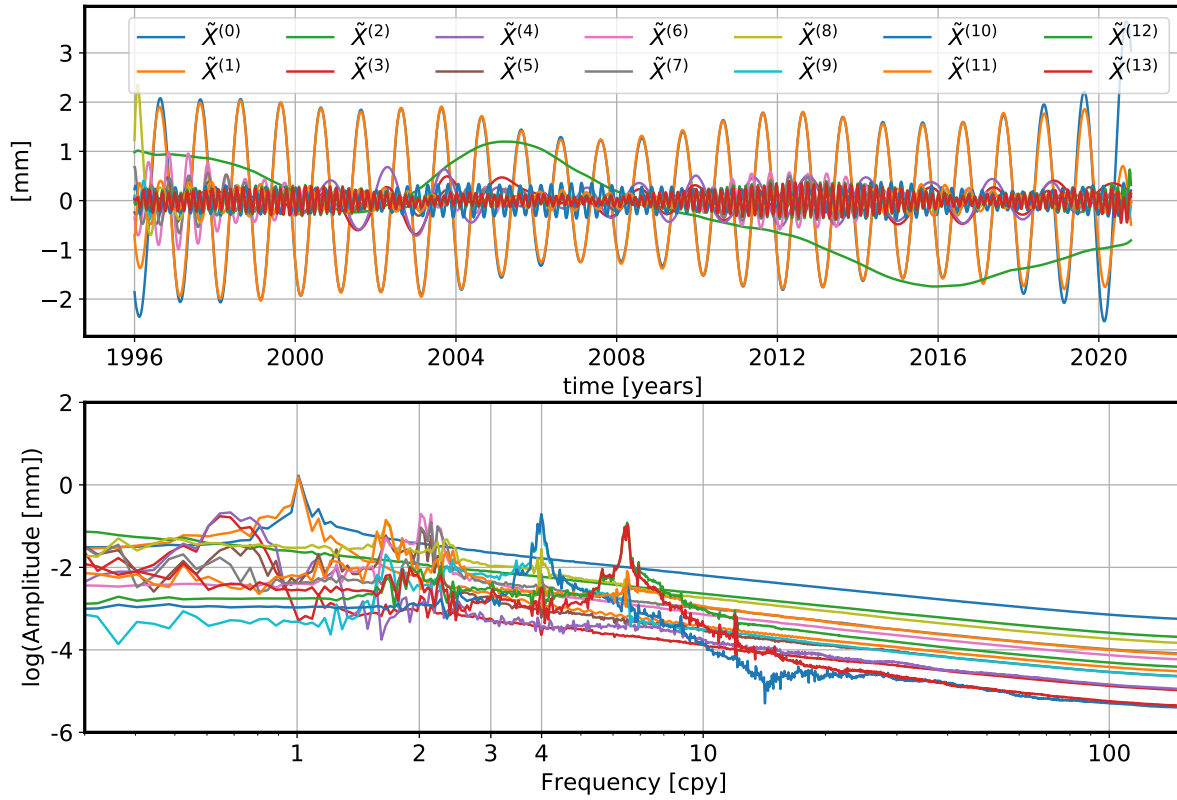
Figure 3.2: Time and frequency domain of the with SSA reconstructed elementary grouped components of the sum of all loadings at the station ZIMM

The next step is to apply Single Value Decomposition (SVD) to the trajectory matrix $\mathbf{X}$:

$$\mathbf{X} = \sum_{i=1}^{d} \sqrt{\lambda_i} U_i V_i^T \equiv \sum_{i=1}^{d} \mathbf{X}_i \tag{3.12}$$

where $d \leq L$ denotes the rank of the matrix, $\lambda_1, \lambda_2, \ldots, \lambda_L$ are the eigenvalues, arranged and sorted by magnitude, and $U_1, U_2, \ldots, U_L$ and $V_1, V_2, \ldots, V_L$ form an orthonormal system of left and right singular vectors of $\mathbf{X}$, respectively. Each $\mathbf{X}_i$ has rank 1 and is called *elementary matrix*, where the ensemble of $\sqrt{\lambda_i} U_i V_i^T$ is denoted as *eigentriple*.

The first step of the reconstruction is the so called *eigentriple grouping*. The indices $1, \ldots, d$ are divided into m separate groups $I_1, \ldots, I_m$. One subset forms a matrix $\mathbf{X}_I = \mathbf{X}_{i_1} + \mathbf{X}_{i_2} + \ldots + \mathbf{X}_{i_p}$, with $i = 1, \ldots, p$. Eqn. (3.12) can therefore be redefined as

$$\mathbf{X} = \sum_{k=1}^{m} \mathbf{X}_{I_k} \tag{3.13}$$

The grouping is referred to as *elementary* if the number of groups m is equal to the rank d of the trajectory matrix $\mathbf{X}$.

In an idealized world the grouped matrices $\mathbf{X}_{I_k}$ in Eqn. (3.13) would be Hankel matrices and the reconstructed time series would be easy to obtain. Nevertheless, this is not the realistic case and therefore each grouped matrix first has to be transformed into a Hankel matrix using the *Hankelisation operator* $\hat{\mathcal{H}}$, which is also known as *diagonal averaging*. $\hat{\mathcal{H}}$ transforms the $L \times K$ matrix $\mathbf{X}_{I_k}$ to a Hankel matrix $\tilde{\mathbf{X}}^{(k)}$

$$\tilde{\mathbf{X}}^{(k)} = \hat{\mathcal{H}} \mathbf{X}_{I_k} \tag{3.14}$$

Let $\tilde{x}_{i,j}$ be an element of $\tilde{\mathbf{X}}^{(k)}$, with $s = i + j$. Each $\tilde{x}_{i,j}$ is computed by the average of all remaining elements of the anti-diagonal that includes $\tilde{x}_{i,j}$ with:

$$\tilde{x}_{i,j} = \begin{cases} \frac{1}{s+1} \sum_{l=0}^{s} x_{l,s-l} & 0 \leq s \leq L-1 \\ \frac{1}{L-1} \sum_{l=0}^{L-1} x_{l,s-l} & L \leq s \leq K-1 \\ \frac{1}{K+L-s-1} \sum_{l=s-K+1}^{L} x_{l,s-l} & K \leq s \leq K+L-2 \end{cases} \tag{3.15}$$

By applying Eqn. (3.15) to $\mathbf{X}_{I_k}$ the reconstructed series $\tilde{\mathbf{X}}^{(k)} = (\tilde{x}_1^{(k)}, \tilde{x}_2^{(k)}, \ldots, \tilde{x}_N^{(k)})$ is finally obtained. The original time series $x_1, x_2, \ldots, x_N$ can now be represented as sum of m

reconstructed time series:

$$x_n = \sum_{k=1}^{m} \tilde{x}_n^{(k)} \tag{3.16}$$

with $n = 1, 2, \ldots, N$.

One advantage of SSA is that amplitude and phase varying signals can be reconstructed, without any a priori assumptions. The only parameter that has to be chosen is the window size L, which determines the spectral resolution. If the lag window is too small, it overestimates seasonal, like annual and semi-annual, signals. If the time lag is chosen too large, it would underestimate these signals. Previous studies already explored the selection of optimal window sizes and showed that for time series, with annual and semi-annual signals, two or three years is a suitable window size (Chen et al., 2013; Klos et al., 2017). This allows for the obtainment of long-term changes as well as higher frequency seasonal variations. According to these suggestions in the literature, in this thesis the window size $L = 3$ years is chosen.
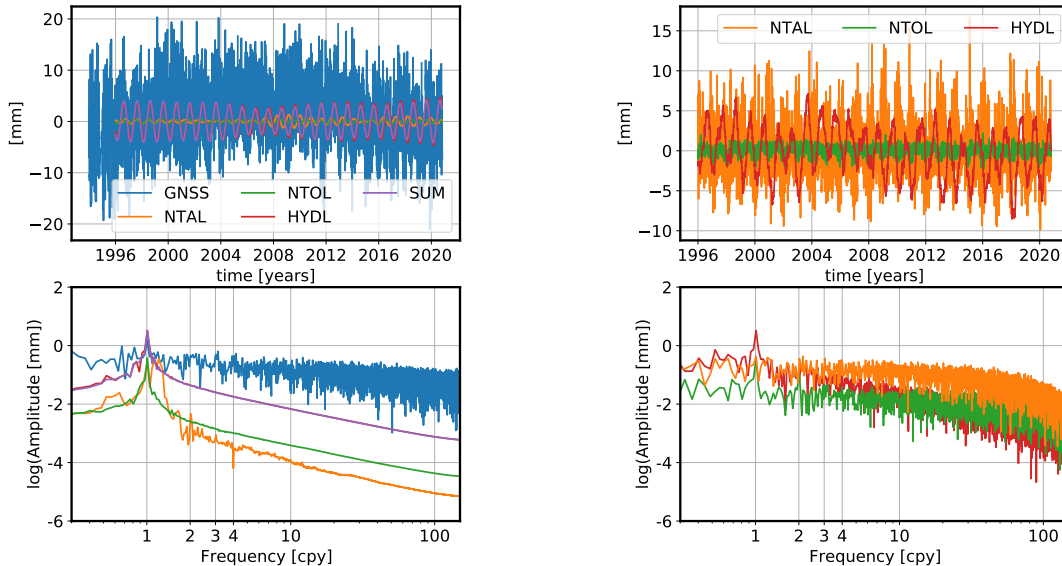
(a) SSA reconstructed annual components of environmental loadings and NGL residuals at the station ZIMM

(b) Fourier transform of original environmental loadings at station ZIMM

Figure 3.3: Frequency analysis of NGL residuals and environmental loadings

Figure 3.2 shows the first 10 elementary grouped components of the superposition of environmental loadings at the example station ZIMM. In the top sub-figure the different periodicities and trends of the different components can be seen in the time domain. The first two components represent the annual period, the third component the long term trend and the rest

accounts for higher periodicities, which can not easily be distinguished in the time domain. Therefore, in the bottom sub-figure, the Fourier transformed components are shown in cycles per year (cpy).

In Figure 3.3a the reconstructed annual components of the individual environmental loading displacements are shown together with the NGL residuals (in time and frequency domain) and Figure 3.3b shows the Fourier transform of the original environmental loadings for comparison.

3.2 Loading Reduction
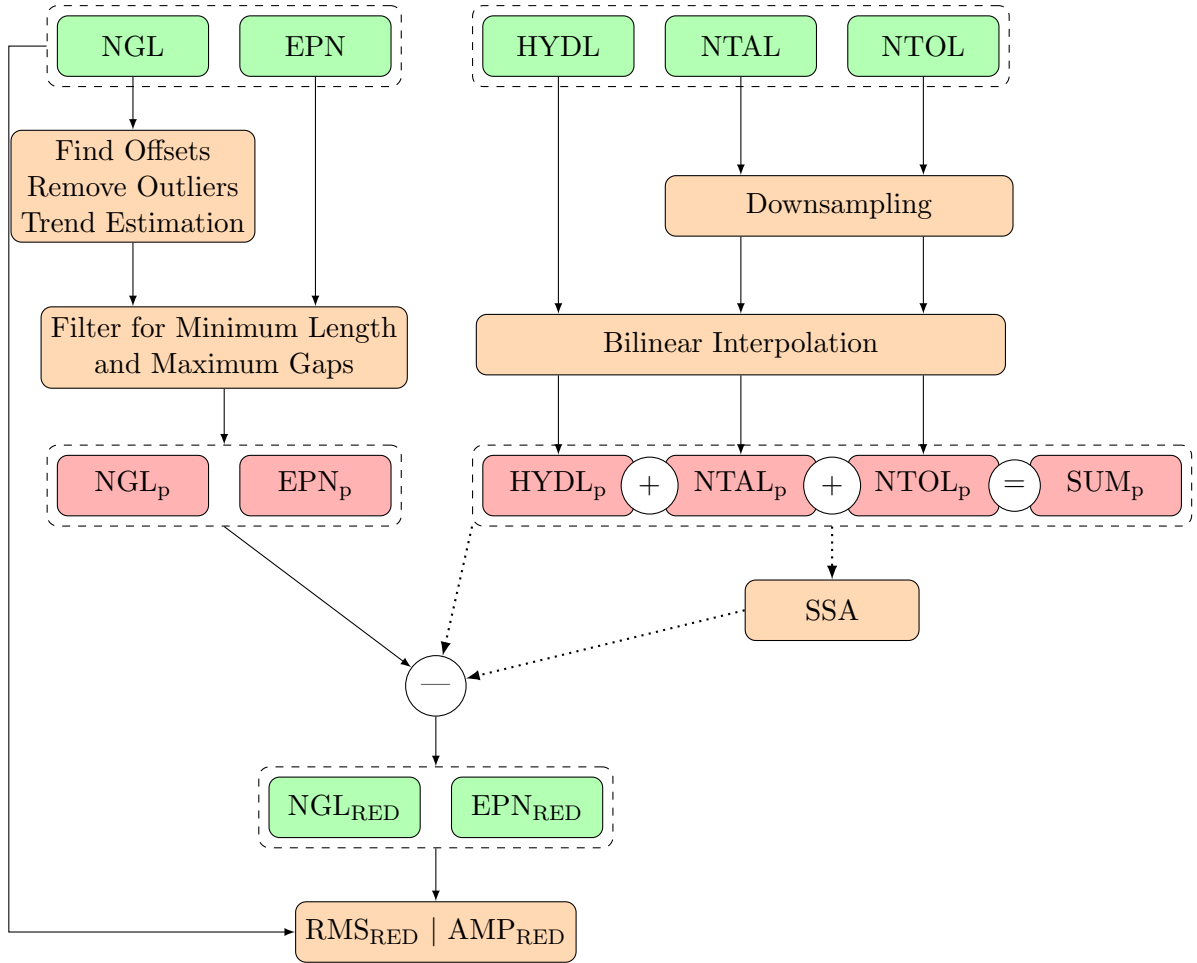
Figure 3.4: Overall workflow of RMS and AMP reduction. The main input and output values are marked in green, processing steps are colored in orange and intermediate products are depicted in red

In order to evaluate the proportion of environmental signals in the GNSS residuals the RMS and annual amplitude (AMP) reduction rates, RMS_{RED} and AMP_{RED}, are computed:

$$RMS_{RED} = \left(1 - \frac{RMS_{GNSS_{RED}}}{RMS_{GNSS}}\right) * 100 \tag{3.17}$$

$$AMP_{RED} = \left(1 - \frac{AMP_{GNSS_{RED}}}{AMP_{GNSS}}\right) * 100 \qquad \text{for } L = [HYDL, NTAL, NTOL, SUM], \tag{3.18}$$

where $GNSS$ is the original, accordingly preprocesed, GNSS time series and $GNSS_{RED}$ is the reduced GNSS series by the corresponding environmental loading effect, or the sum (SUM) of those. RMS stands for the computation of the root mean square, and AMP for the annual amplitude of the time series, estimated with Fourier transformation.

A summary of the whole procedure to obtain the RMS and AMP reductions, from the raw GNSS and environmental loading datasets, is illustrated in Figure 3.4. The raw input values are the GNSS height time series from NGL and EPN, as well as the environmental surface loadings HYDL, NTAL and NTOL. The input parameters are preprocessed to the corresponding datasets, colored in red. A more in depth explanation of the processing steps can be found in Chapter 2. After the environmental loading datasets are interpolated, the superposition (SUM) of all loadings is computed. Then, the GNSS residuals are reduced 1) by the original loading time series, and 2) only by the annual component of the loading time series, which is reconstructed, using SSA. The RMS reduction rates are calculated in both cases for all loading dates individually, and for the sum of all loadings together.

3.3 GNSS Residual Prediction

The prediction of the GNSS residuals from environmental influences is approached with two different supervised learning methods: 1) a deep learning algorithm using Temporal Convolutional Network (TCN), and 2) the ensemble machine learning method Random Forest (RF). To asses the performance of both algorithms, they are compared to a selected baseline algorithm, in this case Exponential Smoothing.

3.3.1 Temporal Convolutional Network

The very basics of a Temporal Convolution Network lie in the principles of sequence modeling, where an input sequence of length T, $x_0, \ldots, x_T$ is predicted to an output $y_0, \ldots, _T$. Important is, that for a prediction for y_t at time t, only previously observations $x_0, \ldots, x_t$ are used. A sequence modeling problem can be represented as a mapping function $f : \mathcal{X}^{T+1} \to \mathcal{Y}^{T+1}$, which maps the predicted observations $\hat{y}_0, \ldots, \hat{y}_T$ from an input $x_0, \ldots, x_T$. The sequence model network searches for the function f that minimizes the expected loss L between true

and predicted values:

$$L(y_0, \ldots, y_t, f(x_0, \ldots, y_T)) \tag{3.19}$$

TCNs use the advantages of low-level feature computation in a Convolutional Neural Network (CNN) and high-level feature processing in a Recurrent Neural Network (RNN) in one combined algorithm. Temporal information is captured at different levels with the following basic approaches: 1D convolutions gather information on the variability of low-level features over time, intermediate-level pooling enables a fast computation of long-term temporal patterns, and channel-wise normalization at a high level ensures improved robustness (Lea et al., 2016).

The realization of the previously introduced architecture results in two main characteristics. By using a 1D fully convolutional network architecture, the TCN is able to take an arbitrary sequence length as input and return the same sequence length as output. To avoid information leaking from the future to the past, all convolutions are *causal*, meaning that an element at time t is only convoluted with elements from the same time and earlier. Another specialty of the employed convolutions is the method of pooling in sequence modeling, called dilated convolutions. The dilation rate d allows for an exponentially large receptive field, while still using the same kernel size and number of parameters (Yan et al., 2020).
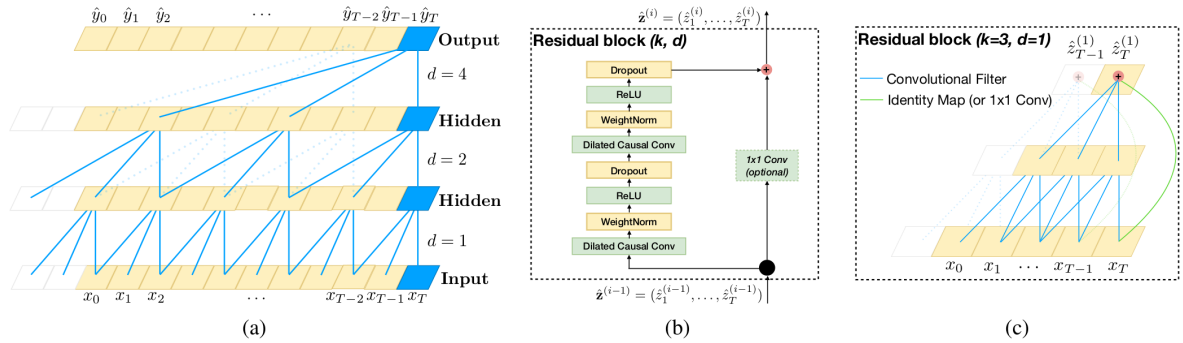
Figure 3.5: Components of a TCN algorithm. a) dilated causal convolution, b) residual block, c) example of residual connection (*source: Bai et al. (2018)*)

Figure 3.5 shows a graphical representation of the TCN architectural parts. In a), the concept of causal dilated convolutions is shown. An input sequence $x_0, \ldots, x_T$ is convolved with the dilation rates d and filter the size k. The dilated causal convolutions are embedded in so-called residual blocks. The schematic architecture of a residual block is depicted in b), and in c) an example residual connection is shown.

3.3.2 Random Forest

A Random Forest (Breiman, 2001) is an ensemble of decision or regression trees and can be used for classification and for regression. A tree is a predictive model, based on a set of binary rules. The components of a tree are branches, nodes and leaves. At each node, starting with the root-node, the samples are divided into subsamples. Each splits a new branch of the tree with further nodes. If a node is not further split into sub-nodes it is called leave.
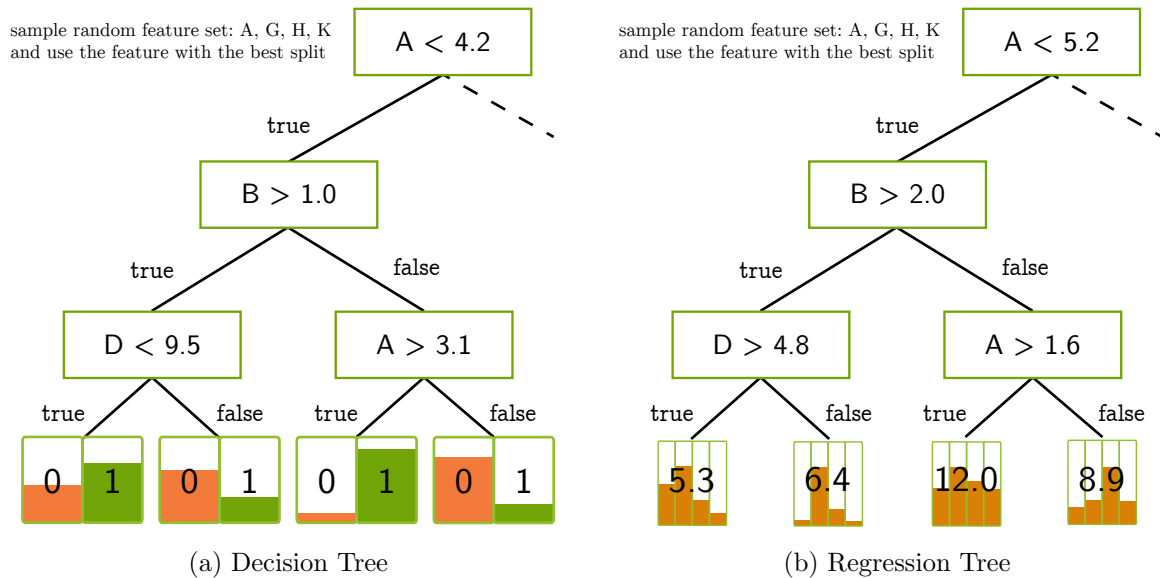

(a) Decision Tree   (b) Regression Tree

Figure 3.6: Structures of a decision and regression tree. The leaves of a decision vote for a class and the predicted value is the majority voted class. The leaves of a regression trees contain numerical values, which are averaged and returned for a prediction (*source: Buß (2020)*)

In the training phase the tree is build while learning the relationships between input and target data. In the prediction, the tree applies the same set of rules that was established in the previous step to map any given value to the best possible estimate.

A random forest is random in two ways. First, the individual trees are built using random sampling of training data. The samples are drawn with replacement, resulting in the possibility of using one sample multiple times within a tree, referred to as bootstrapping. Second, at each node, splits are computed on random samples of features, meaning that only a subset of all features is known when deciding on the split of a node.

The final prediction is a combination of the individual tree estimates, where each tree gives its own uncorrelated solution. In the case of RF regression the final estimate is computed as average over all separate estimates. The approach described above is known as bagging, which highly improves model generalization and avoids overfitting.

Figure 3.6 contains a graphic representation of a decision and a regression tree. The basic

structure and principles are the same for both trees. The difference is in the leaves: In the decision tree (Figure 3.6a), the leaves contain the predicted probabilities of the true class. The overall prediction is the class, which was most voted for in the individual leaves. The leaves of a regression tree (Figure 3.6b) contain numerical values. The returned predicted value is the average of all values in the leaves.

3.3.3 Baseline Algorithm - Exponential Smoothing

Exponential smoothing (Brown, 1959) is a univariate approach to time series forecasting. The basic idea is to smooth the data by taking a weighted average over the past observations, where the weight is decreasing exponentially over time. A one step ahead forecast is given by

$$\hat{y}_{T+1|T} = \alpha y_T + \alpha(1-\alpha)y_{T-1} + \alpha(1-\alpha^2)y_{T-2} + \dots \tag{3.20}$$

with the smoothing parameter $0 \leq \alpha \leq 1$.

Simple exponential smoothing is not able to deal with any trends or seasonalities. Holt (1957) improved the model by adding a parameter for linear trends and Winters (1960) again expanded it to include seasonality as well. The extended version is known as Triple Exponential Smoothing, or Holt Winters Smoothing. The basic forecast Eqn. (3.20) is now completed for three smoothing equations. One accounting for the level, one for the trend and one the seasonal influences, each having its own smoothing factor. The seasonal effect is specified with a certain frequency and can either be employed *additive* or *multiplicative*, depending on the nature of the signal.

3.3.4 Prediction Scores

The quality of predictions, returned by the machine learning model, can be evaluated in numerous ways. In this work the Root Mean Squared Error (RMSE) and the coefficient of Determination R-Squared (R2) are used. These metrics are computed as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2} \tag{3.21}$$

$$R2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \tag{3.22}$$

where N is the time series length, y are the true observed values of the time series, $\hat{y}$ the predicted value of y and $\bar{y}$ the mean value of y. The RMSE can fall in a range of $[0, \infty)$ and the

smaller the value the better. R2 can be understood as the correlation between the dependent and independent variables, and can range from $(-\infty, 1]$, where higher values are desired. If the model would always predict the mean of y, the RMSE would be the same as the variance of y and the R2-score would be 0. In case of perfect predictions the RMSE would become 0 and the R2-score 1. A model, performing worse than the mean of y, would achieve negative R2-scores.

4 Results and Discussion

This chapter presents the results, produced in the course of this work, using the methodologies that were introduced in the previous Chapter. According to the two major parts of the thesis, this Chapter first presents the obtained results of the RMS and amplitude reduction of the GNSS residuals, followed by the findings in the GNSS residual predictions.

4.1 Reduction of GNSS Residuals by Loadings

The first part of the presented study deals with the reduction of GNSS residuals, using environmental loading data. The aim is to explain the remaining seasonal signals, found in GNSS height residuals, by environmental surface loadings. The explainability of these residuals are measured with the RMS and annual-amplitude reduction rates (Eqn. 3.17 and 3.18).

4.1.1 RMS Reduction

In the following, two approaches for RMS reduction of GNSS residuals are discussed, that have been evaluated in this thesis. The approaches are 1) the reduction by environmental loadings directly and 2) the reduction by the SSA reconstructed annual signature (Section 3.1.5) of the environmental loadings. The workflow behind the obtained results can be found in Section 3.2.

Reduction by original loading series

At the beginning, the environmental loading data is subtracted directly from the GNSS residuals. The computed RMS reduction is illustrated in Figure 4.2, where on top the GNSS residuals are used from the EPN, and on the bottom from the NGL dataset. The EPN residuals are on average reduced by -10.9%, implying an *increase* of RMS. In contrast, the average of RMS reductions using NGL is at 19.1%, which is on the same level as found in previous literature. In the spatial distribution of NGL RMS reductions, a pattern can be recognized where the reduction rate increases in higher latitudes.

(a) HYDL   (b) NTAL   (c) NTOL
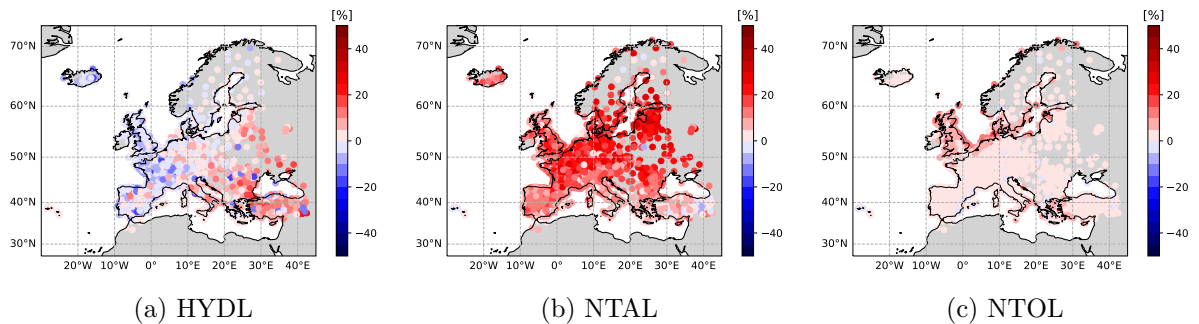
Figure 4.1: RMS reduction of the NGL time series, split into the influences of the individual environmental loadings
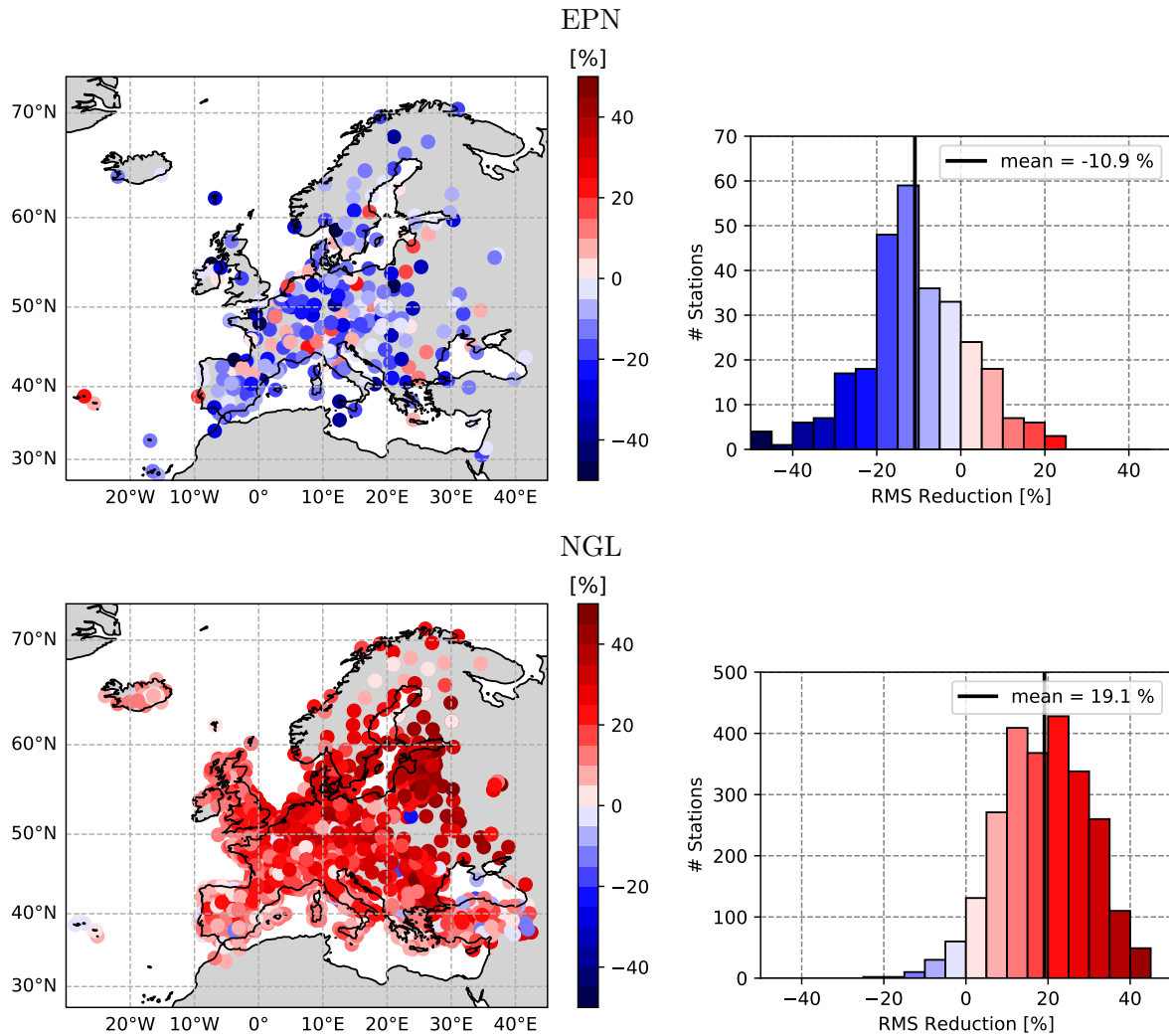
Figure 4.2: RMS Reduction of height residuals by the SUM of environmental loadings

A more detailed insight into the reduction of the NGL dataset is given in Figure 4.1, where the RMS reduction rate is divided into the individual environmental loading components. The RMS reduction by subtracting HYDL only, has no predominantly positive or negative influence. Unlike NTAL, which is the pulling factor to the overall positive RMS reduction. The magnitude of RMS reduction with NTAL, has also a spatial correlation, where the reduction rate increases with increasing latitudes. NTOL has the smallest influence, although mostly positive, with slightly higher rates in coastal regions. The breakdown of the EPN RMS reduction rates can be found in the Appendix (Figure 5.1).

The most plausible explanation for the deterioration of the EPN residual RMS after subtracting environmental loadings, can be traced back to the GNSS residual computation procedure. The EPN solution is obtained from a regional cumulative solution, which leads to a dampening

(a) EPN

(b) NGL
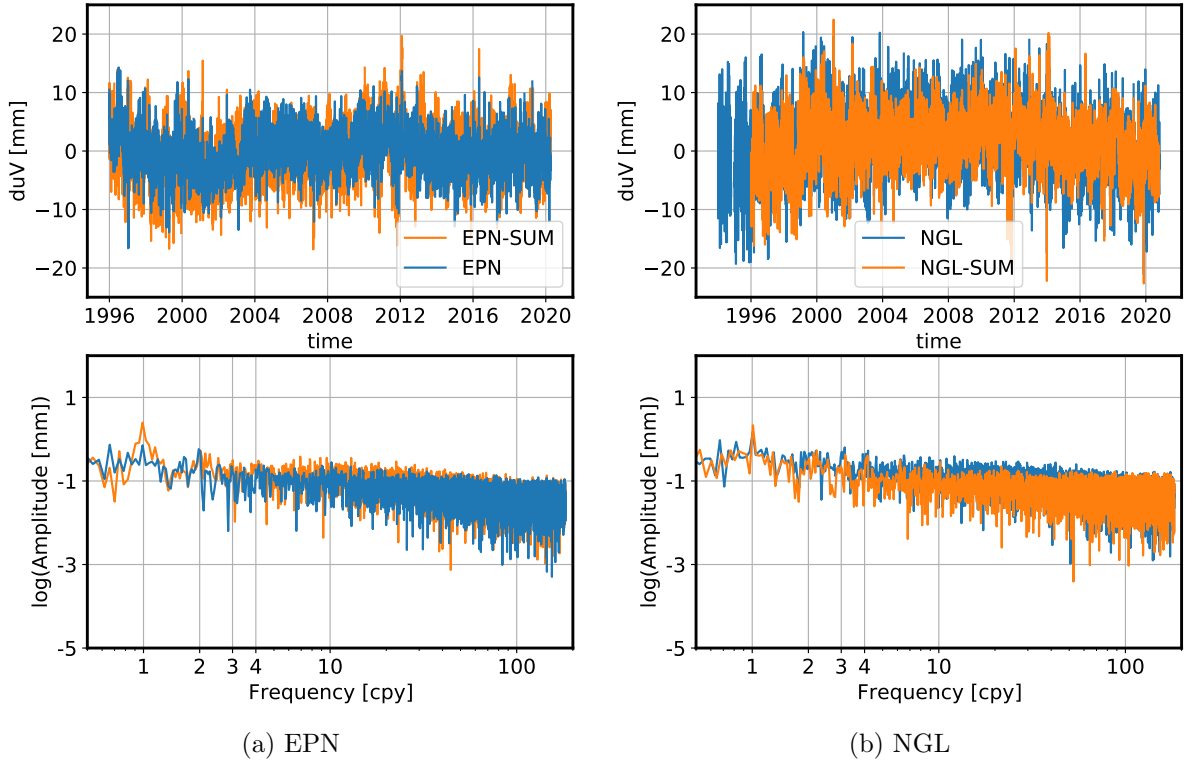
Figure 4.3: Residuals of station ZIMM before and after reduction of SUM loadings.

of annual and semi-annual amplitudes of the Up component due to the elimination of the common mode error (Legrand et al., 2012). The suspicion is confirmed, when looking at the frequency domain of a single time series in Figure 4.3a. The original annual amplitude is hardly prominent, whereas the supposedly reduced signal exhibits a boost at annual frequency. On the contrary, in Figure 4.3b there is a clear reduction in vertical motion of the reduced signal, but the annual amplitude is still not significantly reduced.

However, both sides of Figure 4.3 contain another conspicuous effect. At frequencies around 4–80 cycles per year, the amplitudes are also amplified, respectively dampened, leading to a change in noise properties of the reduced signal. This finding is consistent with a study of Klos et al. (2019). They state that the spectral index can change to up to 0.5, which leads to an underestimation of velocity error. Additionally, they found that the magnitude of change in the noise spectral index depends on the proportion of contribution of the individual loading models. To overcome the issue of changed noise properties, they suggest to reconstruct the environmental loadings with SSA before subtracting them from the GNSS residuals. Accordingly, the annual components of the environmental loadings were reconstructed using SSA and in turn subtracted from the GNSS residuals. The resulting findings are presented in the next paragraphs.

Reduction by SSA reconstructed loading series

The resulting new RMS reductions are illustrated in Figure 4.4. In both cases, EPN and NGL, the RMS reduction rate become smaller. For NGL it is, with an average of 1.9%, close to zero, but slightly more stations are positively reduced than negatively. The majority of EPN residuals continue to experience an increase in RMS, with the rate somewhat improved. In the Appendix, the individual influences to the SSA RMS reduction are included (Figures 5.2-5.3).
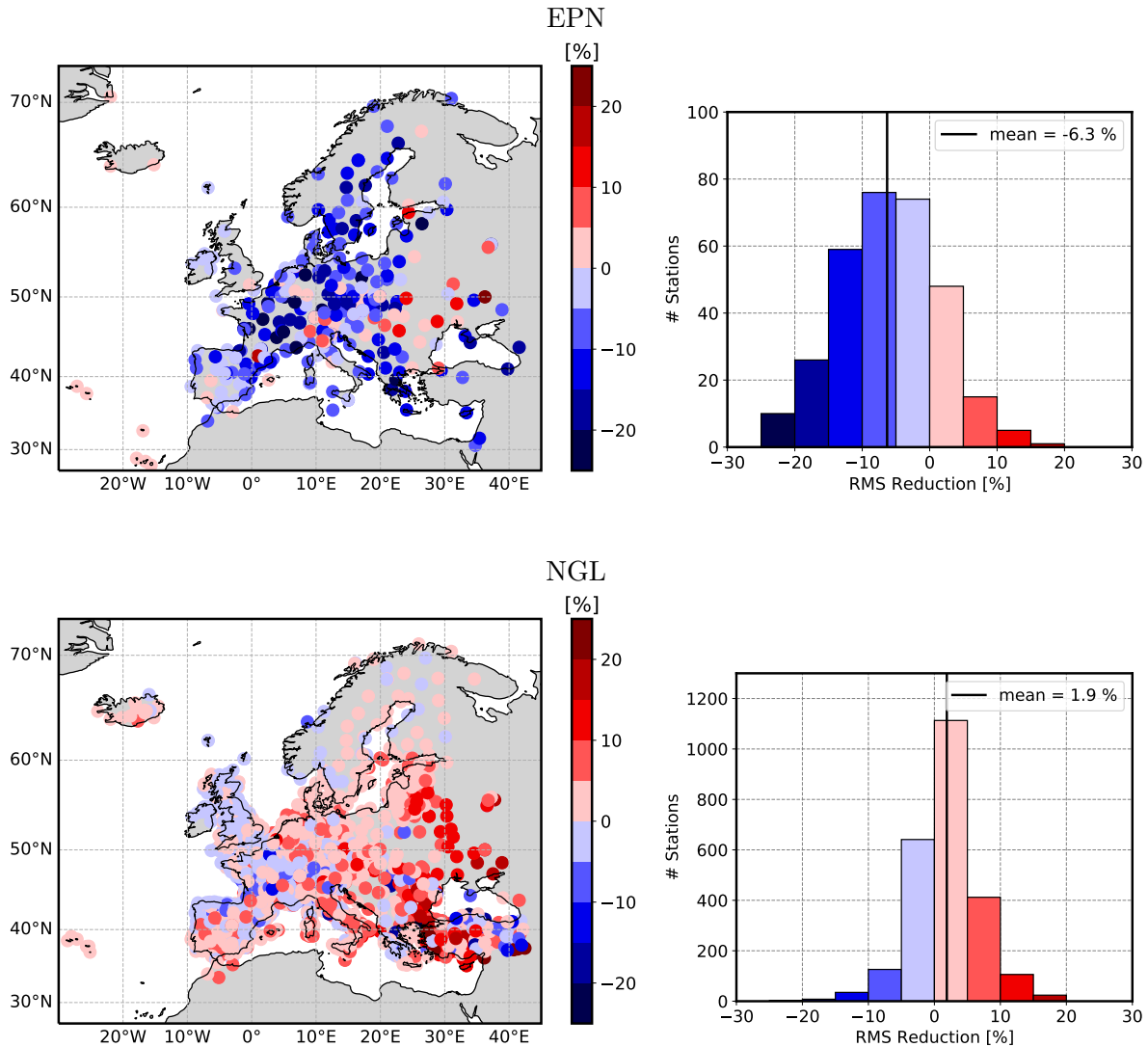
Figure 4.4: RMS Reduction of height residuals by the SUM of environmental loadings, after reconstructing their annual components with SSA

Figure 4.5 shows again the original GNSS residuals together with the reduced residuals, at the example of ZIMM. When comparing this Figure, to Figure 4.3, the biggest difference can be seen in the higher frequencies. While the GNSS residuals reduced by the original loading series,

show different amplitudes in the range of 4–80 cycles per year, the GNSS residuals reduced by the SSA reconstructed loadings, align with the frequency band of the original GNSS residuals.
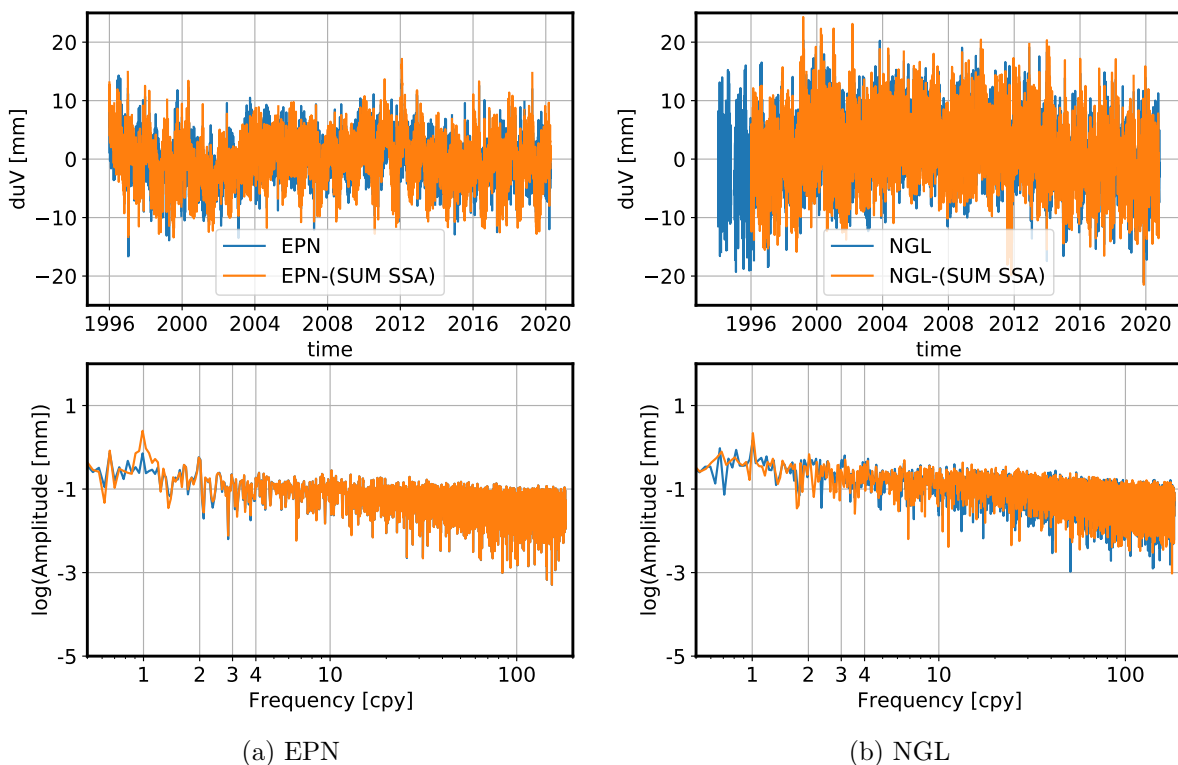
Figure 4.5: Residuals of station ZIMM before and after reduction of SSA reconstruced SUM loadings

4.1.2 Amplitude Reduction

In the last section, the overall RMS reduction rates were computed, giving reasonable results using NGL residuals. In the this section, the effect on the annual amplitude will be examined in more detail. Since the seasonal amplitudes in the EPN residuals are altered and the initial RMS reductions yield poor results, the following analysis is done on the NGL dataset only.
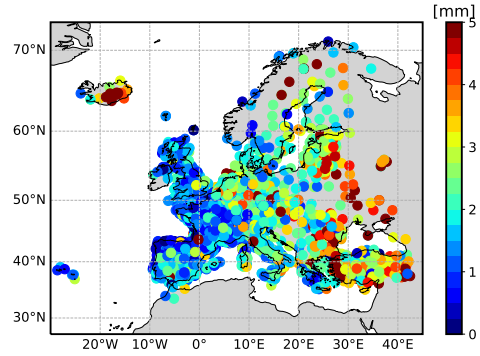
Figure 4.6: NGL annual amplitudes

(a) SUM

(b) HYDL

(c) NTAL

(d) NTOL
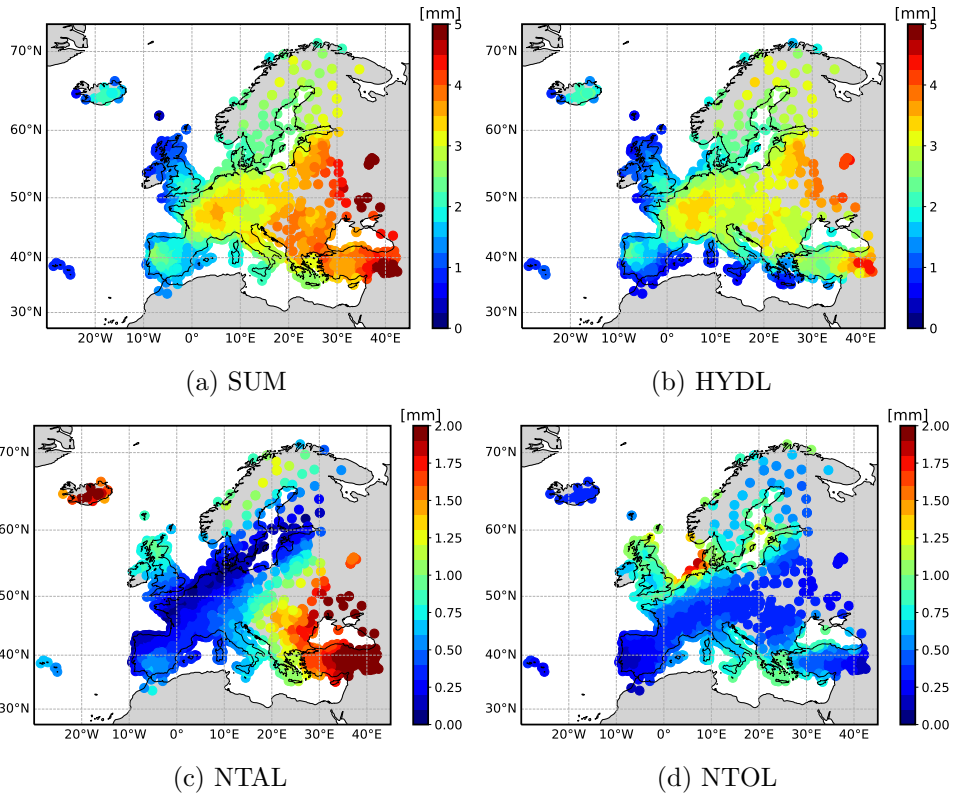
Figure 4.7: Annual amplitudes of the SUM and individual environmental loadings

The absolute yearly amplitudes of the NGL residuals, computed with a Fourier transformation, are shown in Figure 4.6. Furthermore, the superposition and individual environmental loadings are depicted in Figure 4.7. While a trend from west to east can be assumed for the GNSS amplitudes, the environmental loadings show a clear spatial pattern. The SUM amplitudes are mostly determined by HYDL, as it has the highest annual amplitudes. They present a strong trend from inland to the coast. NTAL and NTOL have both generally smaller amplitudes,

with a band of very low values from south-west, to north-east.

Figure 4.8 illustrates the overall annual reduction rates. At more than half of the stations the yearly amplitude is positively reduced. A few stations experience a very high amplitude amplification. Therefore, the mean over all amplitude reduction rates is slightly shifted. The median is thus computed as an additional metric, and displayed as a reference. Wu et al. (2020) found a high correlation of the amplitude reduction and phase differences between the GNSS residuals and the environmental loading series. Since the phase differences were not calculated in this work, it can only be assumed that the high amplification of the amplitudes in some cases, is also due to unfavourable phase differences. For further analysis, thus the corresponding phase differences would have to be calculated and taken into account.
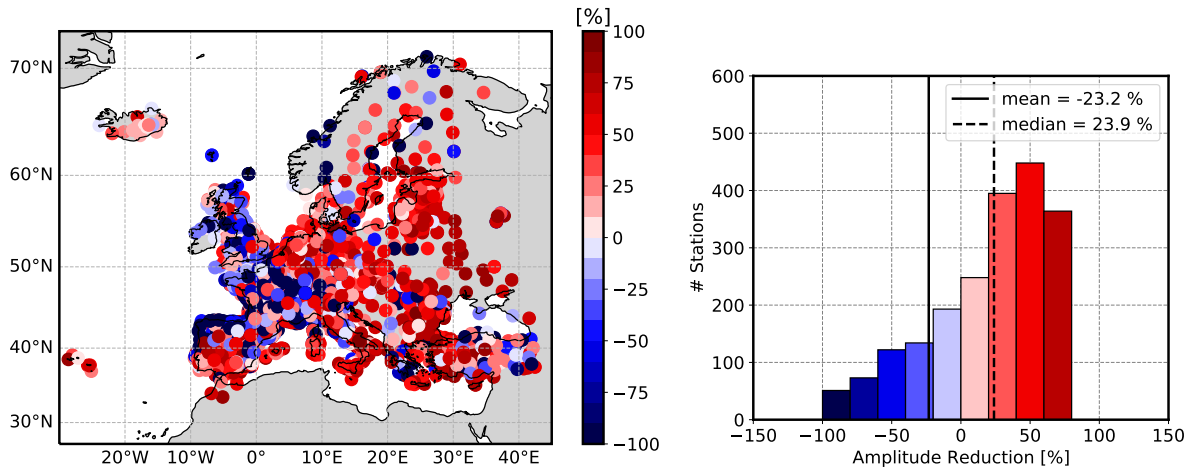
Figure 4.8: Amplitdue reduction rates of the NGL residuals after the subtraction of SUM of environmental loadings

(a) HYDL   (b) NTAL   (c) NTOL
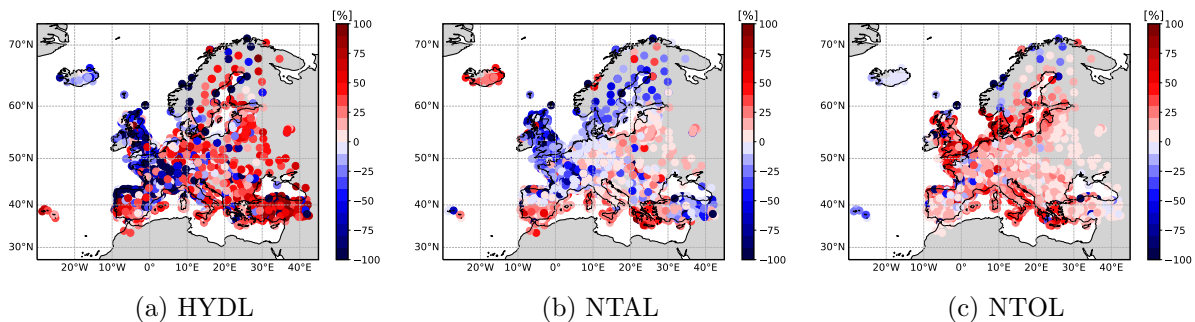
Figure 4.9: Amplitdue reduction rates of the NGL time series, split into the influences of the individual environmental loadings

In Figure 4.9 the amplitude reduction is divided into the individual contributions of HYDL, NTAL and NTOL. The part of HYDL (Figure 4.9a, looks very similar to the overall reduction,

27

indicating that it is the driving factor for all annual amplitudes. Striking is, that NTAL (Figure 4.9b) leads to a mostly negative amplitude reduction, in contrast to almost only positive contributions in the RMS reduction (Figure 4.1b). A spatial trend from south to north is again recognizable. As the GNSS amplitudes get rather amplified with increasing latitude, the assumption is made that the corresponding amplitudes are phase shifted with regard to the GNSS amplitudes. Furthermore, the comparably high positive contributions to the RMS reduction indicate that these are not primarily of annual nature. Although, the amplitudes of NTOL (Figure) are small, they have a big positive impact on the residual amplitude reduction, again with greater influence in coastal regions.

Overall, the RMS of the GNSS residuals can be reduced by 19% on average, when using the original environmental loadings. This leaves 81% still unexplained. Gegout et al. (2010) and Mémin et al. (2020), among other sources, identify a mismodeled tropospheric delay as a possible explanation for this remaining 81%.

4.1.3 Correlation of GNSS Residuals and Tropospheric Zenith Delay

In order to get a first impression of the relationship between the tropospheric total zenith delay and the GNSS residuals, the Pearson Correlation Coefficient is computed and depicted in Figure 4.10. As the tropospheric delay typically has a very strong algebraic correlation with the height component in the GNSS adjustment, initially, the linear and seasonal trends were removed form both input signals using *Hector* (Bos et al., 2012). In higher latitudes, where the atmospheric activity is higher, it can be assumed that the errors of the troposphere in the GNSS height residuals are higher as well.
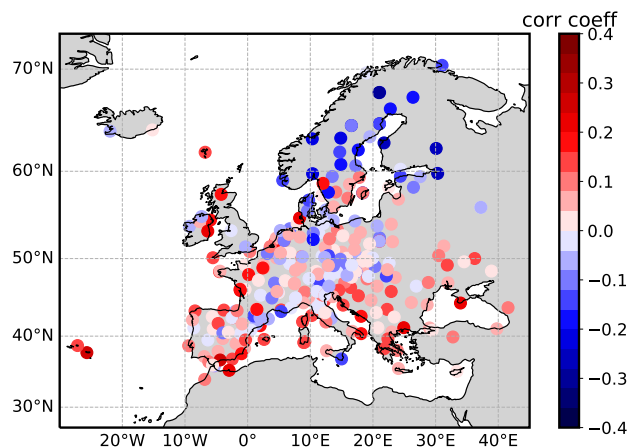
Figure 4.10: Pearson Correlation Coefficient of EPN residuals and tropospheric total zenith delay, both reduced by a linear, and a constant amplitude seasonal (annual and semi-annual), trend in *Hector*

4.2 Prediction of GNSS Residuals

This section describes the results obtained for the modeling and prediction of GNSS residuals using environmental data as input features. These features include the environmental surface loading models, the tropospheric zenith delay and raw meteorological data. For the GNSS height residuals, the NGL dataset is used. Figure 4.11 gives an overview of the observation and model values at the example station COMO. The locations of all tested stations are shown in Figure 4.12. For simplicity, the meteorological parameters, tropospheric zenith delay, and loading data are together denoted as *environmental* data, in the remainder of this section.

To evaluate the influence of the environmental data on the GNSS residual predictions, three different combinations are used as input features: 1) GNSS residuals together with environmental data, 2) GNSS residuals only and 3) environmental data only. The output is always the GNSS residuals.
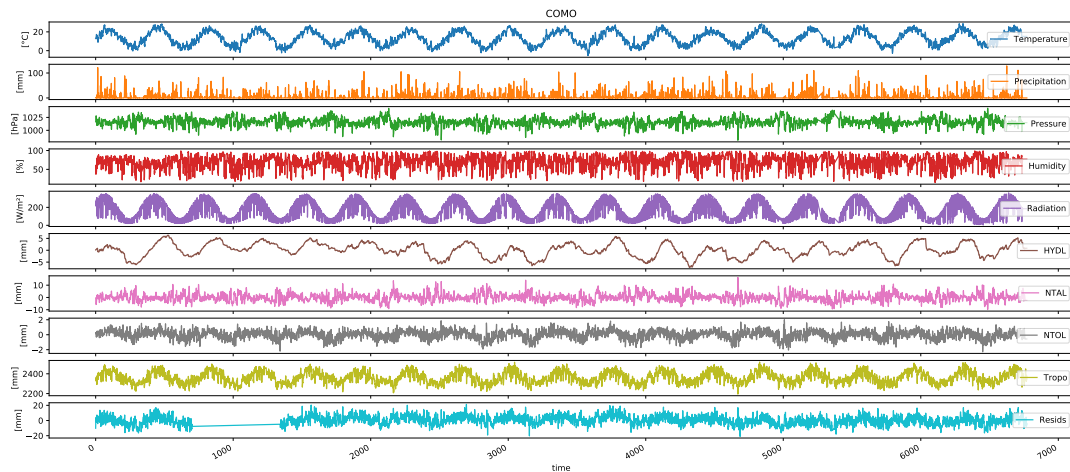
Figure 4.11: Overview of all possible input data for station COMO
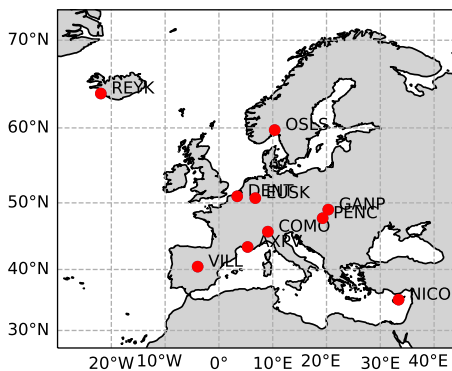
Figure 4.12: Test station locations

4.2.1 Temporal Convolutional Network
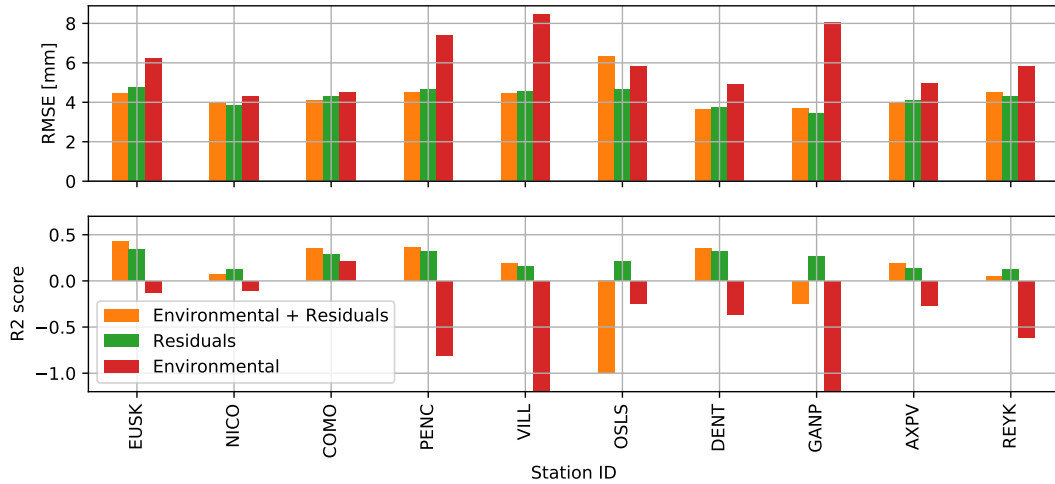
Figure 4.13: RMSE and R2 scores of TCN predictions

Figure 4.13 gives an illustrative overview of the achieved accuracies using TCN. For the majority of stations, the RMSE improves by including environmental data as input, in addition the GNSS residuals. The biggest improvement can be obtained in station EUSK, where the RMSE of the predictions, using environmental data and GNSS residuals together is 6% lower, than when using GNSS residuals as input data alone. The predictions from environmental data only, are mostly worse than the other two variants, although they come close in some cases. In the previous sections it was shown that a RMS reduction up to 19% on average, could be explained by environmental surface loadings. Therefore, it is acceptable if the forecast does not exceed the other variants, where residuals are included as input information as well. On the bottom part of Figure 4.13 the R2 scores are represented. The y-scale is cut-off on the negative side, to make a better comparison of the values around and above zero. The R2-score is closely related to RMSE, describing how well the model is explained with the predictions, a more detailed description of both measures is given in Section 3.3.1. A negative R2 score indicates that the variation around the predicted model is worse than the total variance, meaning that predicting the mean would be more accurate.

Correspondingly, two correlation matrices, representing the input data correlations for stations COMO and OSLS, are illustrated in Figure 4.14. Station COMO is the only station where the TCN achieves a positive R2 score for the prediction with environmental data only as input. Whereas station OSLS yields the lowest scores, when predicting the GNSS residuals with the residuals as input, together with environmental data. A difference can be seen in the correlation matrices, where station OSLS has very low correlations with the meteorological data. In contrast, station COMO has a high negative correlation with sea level pressure,
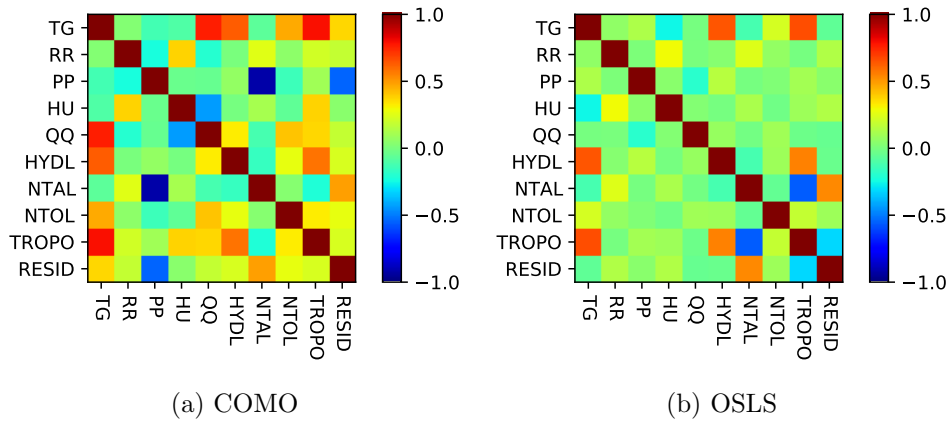
Figure 4.14: Input feature correlation matrix, with TG = Temperature, RR = Precipitation, PP = Sea level Pressure, HU = Humidity and QQ = Radiation.

as well as higher values with HYDL and NTOL. Another difference can be spotted in the correlation with tropospheric zenith delay, which is positively correlated in the case of COMO, but negatively correlated at station OSLS. In terms of the location OSLS (Oslo, Norway) is much closer to the ocean than COMO (Como, Italy). Stations located near the coast, are subject to greater uncertainties in the modeling of oceanic and specific tropospheric propagation effects (Gegout et al., 2010).
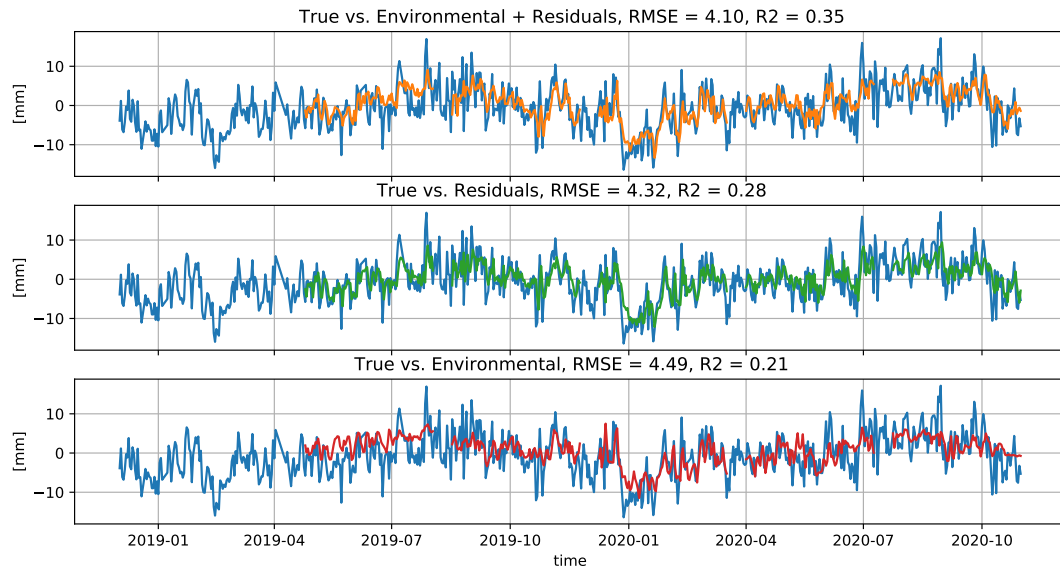
Figure 4.15: Comparison of TCN predictions at station COMO, in each case the true signal is in blue and the predicted values in the other color.
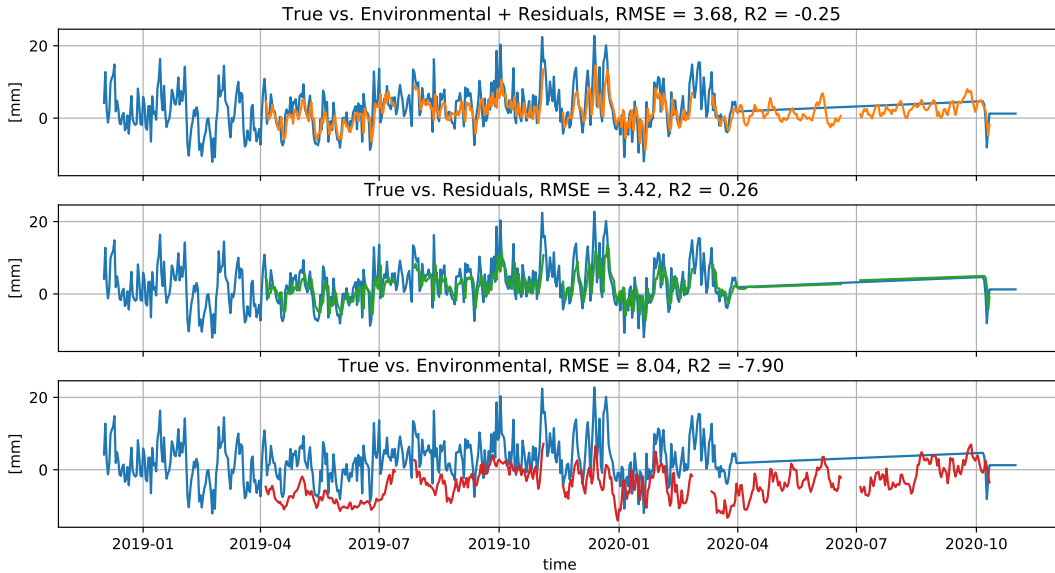
Figure 4.16: Comparison of TCN predictions at station GANP, in each case the true signal is in blue and the predicted values in the other color.

A closer look at the predicted values, compared to the true time series is taken in Figure 4.15 for the example of station COMO. The predictions with residuals only, and those together with environmental data look very similar, which is also reflected in close values of RMSE and R2. In the case of COMO, the RMSE of the predictions could be reduced by 5% compared to the univariate prediction with residuals only. The values, predicted with environmental data only, seem to approach the original time series quite well, although in some parts there seems to be a vertical offset, as seen around the middle of 2019.

A second station is examined at in detail, namely GANP, which achieves very low scores when including environmental data as input. When looking into its time series in Figure 4.16, it can be seen that there is a big data gap, that covers the majority of year the 2020. As the TCN implementation does not account for missing data, the time series were linearly interpolated, which works well for small data gaps. In the case of GANP, the results can not be considered valid and are excluded in the remainder of this chapter. This effect also shows the essence of clean preprocessing of all used data.

4.2.2 Random Forest

Random Forest, which is a widely used algorithm, is used in this thesis as a comparison. The results are similar to those of a TCN, although the RF predictions are almost identical in the cases of using residuals only, or a combination of residuals and environmental data as input features. This indicated that the RF cannot draw much from environmental data as additional

information and the predictions are mainly shaped by the residuals. Nevertheless, the RF seems to be more robust, as also the weak predictions do not spike as much. The RMSE of predictions using environmental data only, does not exceed the prediction RMSE, obtained when using the GNSS residuals in the input features, by 20%, on average. Figure 4.17 gives again the overview of achieved scores.
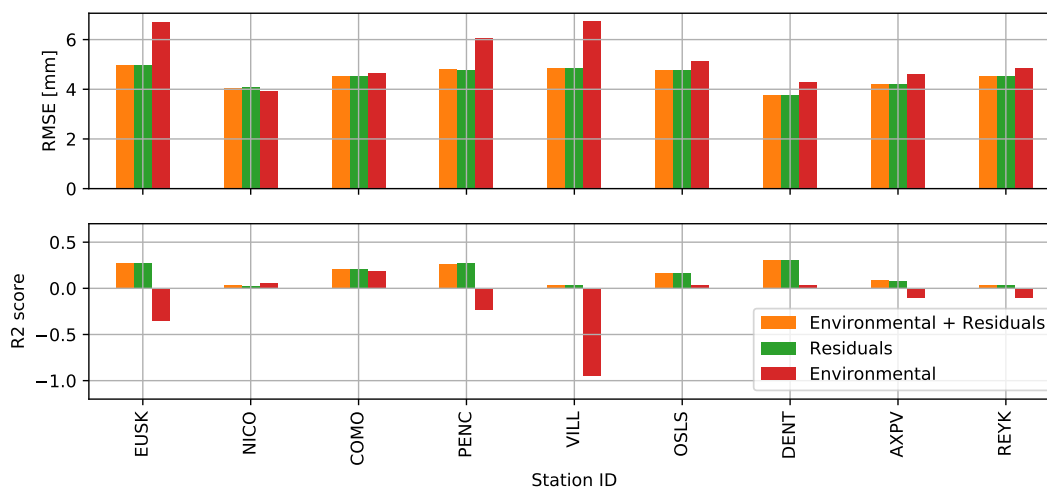

Figure 4.17: RMSE and R2 scores of RF predictions

4.2.3 Comparison of Algorithms with Baseline

A comparison of a novel approach with an established method, referred to as baseline, is crucial for any time series forecasting problem, to have a reference to measure how well the implemented models are actually performing. If the model to be tested is not better than the simple comparative algorithm, one must either optimize the used method and its settings, or one has to consider pursuing a different approach altogether. In this study, exponential smoothing (introduced in Section 3.3.3) is chosen as baseline comparison. Since exponential smoothing is a univariate method, the univariate approaches of the other algorithms, using residuals only as input feature, are used for comparison.

Figure 4.18 displays the achieved scores per station of the tested algorithms. In all test stations TCN performed best, followed by RF and both outperforming exponential smoothing. In Table 4.1 all results from the GNSS residual predictions are again numerically summarized, as mean over all test stations. Station GANP is excluded in the overall evaluation, as the poor prediction scores, are not related to the models, but to an error in data preprocessing. Interestingly, the mean RMSE of TCN in the case where environmental data was used additionally to GNSS residuals, is higher than the RMSE of the univariate prediction, although in the majority of stations the RMSE could be reduced when including environmental data as input feature
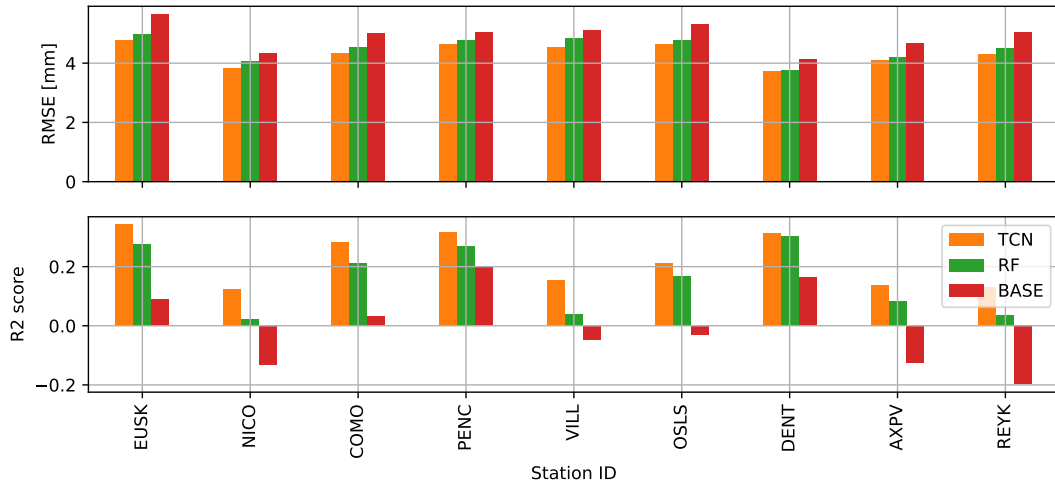
Figure 4.18: RMSE and R2 scores of all tested algorithms, using GNSS residuals only as input features

as well. This means that the individual predictions at most slightly improve, and the cases where the predictions are worse, have a significant higher RMSE. In the case of the RF the results presented in Figure 4.17/Table 4.1 indicate that a model trained with residuals only is sufficient to improve the performance of the prediction. Furthermore, these results show that a combination of residuals and environmental data does not lead to a significant improvement. However, the RF does perform better than the TCN, when using environmental data alone as input features.

In general, the results and conclusions must be viewed with the utmost caution, due to the small subset of test stations. There is still a lot of room for improvement, in terms of data preprocessing and selection, as well as the tuning of the machine learning models themselves.

			TCN	RF	BASE
Env. + Resid.		RMSE	4.35	4.40	-
		R2	0.07	0.16	-
	Resid.	RMSE	4.23	4.41	4.83
		R2	0.22	0.16	-0.01
	Env.	RMSE	6.05	5.23	-
		R2	-1.25	-0.38	-

Table 4.1: Prediction scores of all algorithms and input feature variations (excluding station GANP)

5 Conclusion and Outlook

This thesis concentrated on the influence of environmental data on GNSS height residuals. In the first part it was investigated how much the RMS of the GNSS residuals can be decreased when reducing the GNSS residuals by environmental surface loadings, including hydrological (HYDL), non-tidal atmospheric (NTAL) and non-tidal oceanic loadings (NTOL). The RMS reductions were computed using environmental loading datasets from the Earth System Modelling Group Repository of Deutsche Geoforschungszentrum Potsdam (ESMGFZ) on two different GNSS datasets, the residuals of the European Permanent Network (EPN) and the Nevada Geodetic Laboratory (NGL).

The most important findings of the RMS reductions were:

- The RMS of GNSS height residuals of the NGL dataset could be reduced to an average of up to 19%, when subtracting the superposition of environmental loadings.

- A RMS reduction of an average of 2% was achieved, for the same GNSS dataset, when subtracting the reconstructed annual component of the environmental loadings using Singular Spectrum Analysis (SSA).

- Both approaches were not fully sufficient to improve the RMS of the EPN height residuals. This can be attributed to the reduction of common mode signals in the computation of the EPN residual solution, where local combinations are used.

In the second part of this work the NGL GNSS height residuals where modeled and predicted using a Temporal Convolutional Network (TCN) and Random Forest (RF). The input features where different combinations of the GNSS residuals themselves, environmental surface loadings (HYDL, NTAL, NTOL), tropospheric zenith delays and raw meteorological paramters (Temperature, Precipitation, Humidity, Sea Level Pressure, Radiation). The results of the machine learning approaches showed that

- TCN and RF outperformed exponential smoothing, which was taken as comparative baseline algorithm.

- TCN was able to improve its prediction error in 6 out of 9 test stations when including environmental loadings, raw meteorological parameters and tropospheric zenith delays additionally to the GNSS residuals in the input features.

- The best prediction scores using the aforementioned environmental influences, without using GNSS residuals in the input parameters, were achieved with RF. On average, this prediction error is still up to 20% higher, compared to the prediction error when predicting the GNSS residuals, using RF, from themselves.

In further studies it is suggested to take a closer look into phase offsets between environmental surface loadings and GNSS height residuals. Another promising approach could be the further investigation on the assimilation of models, combining the environmental surface displacements with gravitational data from the Gravity Recovery and Climate Experiment (GRACE) (Klos et al., 2020).

The machine learning approach, proposed in this work, should be understood as first steps into this direction. A closer look can be taken into further improvements of the utilized model parameters, or the evaluation of using other models overall. Future considerations would also include a more in depth analysis and validation of the first available results by processing a larger set of test stations and run through more variations of input feature combinations. This approach would allow to gain a better understanding on the impact of different environmental influences on the GNSS height residuals.

Bibliography

Altamimi, Z. et al. (2007). "ITRF2005: A new release of the International Terrestrial Reference Frame based on time series of station positions and Earth Orientation Parameters". In: *Journal of Geophysical Research: Solid Earth* 112 (B9). DOI: https://doi.org/10.1029/2007JB004949.

Bai, Shaojie, J. Zico Kolter, and Vladlen Koltun (Apr. 19, 2018). "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling". In: *arXiv:1803.01271 [cs]*.

Bevis, Michael and Abel Brown (Mar. 2014). "Trajectory models and reference frames for crustal motion geodesy". In: *Journal of Geodesy* 88.3, pp. 283–311. DOI: 10.1007/s00190-013-0685-5.

Bian, Yankai (Aug. 20, 2020). "Comparisons of GRACE and GLDAS derived hydrological loading and the impacts on the GPS time series in Europe". In: *Acta Geodynamica et Geomaterialia*, pp. 297–310. DOI: 10.13168/AGG.2020.0022.

Blewitt, Geoffrey, William C. Hammond, and Corné Kreemer (2018). "Harnessing the GPS Data Explosion for Interdisciplinary Science". In: *Eos*.

Bos, M. S. and R. M. S. Fernandes (Jan. 2019). "Hector user manual version 1.7.2". In:

Bos, M S et al. (2012). "Fast error analysis of continuous GNSS observations with missing data". In: p. 10.

Bracewell, Ronald (1978). *The Fourier Trandform and its Applications*. Vol. 31999. McGraw-Hill New York.

Breiman, Leo (Jan. 2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32.

Broomhead, D.S. and Gregory P. King (June 1986). "Extracting qualitative dynamics from experimental data". In: *Physica D: Nonlinear Phenomena* 20.2, pp. 217–236. DOI: 10.1016/0167-2789(86)90031-X.

Brown, Robert Goodell (1959). *Statistical forecasting for inventtory control*. McGraw-Hill.

Bruyninx, Carine et al. (Oct. 2019). "GNSS metadata and data validation in the EUREF Permanent Network". In: *GPS Solutions* 23.4, p. 106. DOI: 10.1007/s10291-019-0880-9.

Buttkus, Burkhard (Dec. 6, 2012). *Spectral Analysis and Filter Theory in Applied Geophysics*. Google-Books-ID: 2knuCAAAQBAJ. Springer Science & Business Media. 665 pp.

Buß, Jens Björn (2020). "Bad Moon Rising? – Studies on the Performance of the First G-APD Cherenkov Telescope under Bright Light Conditions using SiPMs for Gamma-Ray Observations". PhD thesis. TU Dortmund University.

Chen, Q. et al. (Dec. 1, 2013). "Singular spectrum analysis for modeling seasonal signals from GPS time series". In: *Journal of Geodynamics*. SI: Geodetic Earth System 72, pp. 25–35. DOI: 10.1016/j.jog.2013.05.005.

Collilieux, Xavier et al. (Jan. 4, 2010). "Impact of loading effects on determination of the International Terrestrial Reference Frame". In: *Advances in Space Research* 45.1, pp. 144–154. DOI: 10.1016/j.asr.2009.08.024.

Cooley, James W. and John W. Tukey (1965). "An Algorithm for the Machine Calculation of Complex Fourier Series". In: *Mathematics of Computation* 19.90. Publisher: American Mathematical Society, pp. 297–301. DOI: 10.2307/2003354.

Dam, T. M. van et al. (June 1, 1997). "Predictions of crustal deformation and of geoid and sea-level variability caused by oceanic and atmospheric loading". In: *Geophysical Journal International* 129.3, pp. 507–517. DOI: 10.1111/j.1365-246X.1997.tb04490.x.

Dam, T. van et al. (2001). "Crustal displacements due to continental water loading". In: *Geophysical Research Letters* 28.4, pp. 651–654. DOI: https://doi.org/10.1029/2000GL012120.

Dam, Tonie M. van, Geoffrey Blewitt, and Michael B. Heflin (1994). "Atmospheric pressure loading effects on Global Positioning System coordinate determinations". In: *Journal of Geophysical Research: Solid Earth* 99 (B12), pp. 23939–23950. DOI: 10.1029/94JB02122.

Dill, R. and H. Dobslaw (2013). "Numerical simulations of global-scale high-resolution hydrological crustal deformations". In: *Journal of Geophysical Research: Solid Earth* 118.9. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/jgrb.50353, pp. 5008–5017. DOI: 10.1002/jgrb.50353.

Dill, Robert (2008). "Hydrological model LSDM for operational Earth rotation and gravity field variations". In: p. 37.

Dong, D. et al. (2002). "Anatomy of apparent seasonal variations from GPS-derived site position time series". In: *Journal of Geophysical Research: Solid Earth* 107 (B4). _eprint:

https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2001JB000573, ETG 9–1–ETG 9–16. DOI: https://doi.org/10.1029/2001JB000573.

Gegout, P. et al. (2010). "Modeling and Observation of Loading Contribution to Time-Variable GPS Sites Positions". In: *Gravity, Geoid and Earth Observation*. Ed. by Stelios P. Mertikas. International Association of Geodesy Symposia. Berlin, Heidelberg: Springer, pp. 651–659. DOI: 10.1007/978-3-642-10634-7_86.

Holt, Charles C. (1957). "Forecasting seasonals and trends by exponentially weighted moving averages". In: *International Journal of Forecasting* 20.1, pp. 5–10. DOI: 10.1016/j.ijforecast.2003.09.015.

Jiang, Weiping et al. (July 2013). "Comparative analysis of different environmental loading methods and their impacts on the GPS height time series". In: *Journal of Geodesy* 87.7, pp. 687–703. DOI: 10.1007/s00190-013-0642-3.

Klos, Anna, Machiel S. Bos, and Janusz Bogusz (Nov. 22, 2017). "Detecting time-varying seasonal signal in GPS position time series with different noise levels". In: *GPS Solutions* 22.1, p. 21. DOI: 10.1007/s10291-017-0686-6.

Klos, Anna et al. (2019). "Estimates of Vertical Velocity Errors for IGS ITRF2014 Stations by Applying the Improved Singular Spectrum Analysis Method and Environmental Loading Models". In: *Geodynamics and Earth Tides Observations from Global to Micro Scale*. Series Title: Pageoph Topical Volumes. Cham: Springer International Publishing, pp. 229–246. DOI: 10.1007/978-3-319-96277-1_18.

Klos, Anna et al. (2020). "Quantifying Noise in Daily GPS Height Time Series: Harmonic Function Versus GRACE-Assimilating Modeling Approaches". In: *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5. DOI: 10.1109/LGRS.2020.2983045.

Lea, Colin et al. (2016). "Temporal Convolutional Networks: A Unified Approach to Action Segmentation". In: *Computer Vision – ECCV 2016 Workshops*. Ed. by Gang Hua and Hervé Jégou. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 47–54. DOI: 10.1007/978-3-319-49409-8_7.

Legrand, J. et al. (2012). "Comparison of Regional and Global GNSS Positions, Velocities and Residual Time Series". In: *Geodesy for Planet Earth*. Ed. by Steve Kenyon, Maria Christina Pacino, and Urs Marti. International Association of Geodesy Symposia. Berlin, Heidelberg: Springer, pp. 95–103. DOI: 10.1007/978-3-642-20338-1_12.

Montillet, Jean-Philippe and Machiel S. Bos, eds. (2020). *Geodetic Time Series Analysis in Earth Sciences*. Springer Geophysics. Cham: Springer International Publishing. DOI: `10.1007/978-3-030-21718-1`.

Mémin, Anthony, Jean-Paul Boy, and Alvaro Santamaría-Gómez (Feb. 14, 2020). "Correcting GPS measurements for non-tidal loading". In: *GPS Solutions* 24.2, p. 45. DOI: `10.1007/s10291-020-0959-3`.

Penna, N. T. and M. P. Stewart (2003). "Aliased tidal signatures in continuous GPS height time series". In: *Geophysical Research Letters* 30.23. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029 DOI: `https://doi.org/10.1029/2003GL018828`.

Santamaría-Gómez, Alvaro and Anthony Mémin (Aug. 1, 2015). "Geodetic secular velocity errors due to interannual surface loading deformation". In: *Geophysical Journal International* 202.2. Publisher: Oxford Academic, pp. 763–767. DOI: `10.1093/gji/ggv190`.

Springer, Anne et al. (Oct. 2019). "Evidence of daily hydrological loading in GPS time series over Europe". In: *Journal of Geodesy* 93.10, pp. 2145–2153. DOI: `10.1007/s00190-019-01295-1`.

Tank, A. M. G. Klein et al. (2002). "Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment". In: *International Journal of Climatology* 22.12. _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.773, pp. 1441–1453. DOI: `https://doi.org/10.1002/joc.773`.

Vautard, R. and M. Ghil (May 1, 1989). "Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series". In: *Physica D: Nonlinear Phenomena* 35.3, pp. 395–424. DOI: `10.1016/0167-2789(89)90077-8`.

Vautard, Robert, Pascal Yiou, and Michael Ghil (Sept. 15, 1992). "Singular-spectrum analysis: A toolkit for short, noisy chaotic signals". In: *Physica D: Nonlinear Phenomena* 58.1, pp. 95–126. DOI: `10.1016/0167-2789(92)90103-T`.

Winters, Peter R. (Apr. 1, 1960). "Forecasting Sales by Exponentially Weighted Moving Averages". In: *Management Science* 6.3. Publisher: INFORMS, pp. 324–342. DOI: `10.1287/mnsc.6.3.324`.

Wu, Shuguang et al. (Aug. 31, 2020). "Comparative Analysis of the Effect of the Loading Series from GFZ and EOST on Long-Term GPS Height Time Series". In: *Remote Sensing* 12.17, p. 2822. DOI: `10.3390/rs12172822`.

Yan, Jining et al. (May 15, 2020). "Temporal Convolutional Networks for the Advance Prediction of ENSO". In: *Scientific Reports* 10.1. Number: 1 Publisher: Nature Publishing Group, p. 8055. DOI: `10.1038/s41598-020-65070-5`.

List of Figures

List of Abbreviations

CATREF   Combination and Analysis of Terrestrial Reference Frames

CNN   Convolutional Neural Network

CPY   Cycles Per Year

ECAD   European Climate Assessment and Dataset

ECMWF   European Center for Medium-Range Weather Forecasts

EOF   Eompirical Orthogonal Functions

EPN   EUREF Permentent GNSS Network

ESMGFZ   Earth System Modelling Group Repository of Deutsche Geoforschungszen-
                trum Potsdam

FFT   Fast Fourier Transform

GNSS   Global Navigation Satellite System

GRACE   Gravity Recovery and Climate Experiment

HYDL   Hydrological Loading

ITRF   International Terrestrial Reference Frame

LSDM   Land Surface Discharge Model

MJD   Modified Julian Date

MLE   Maximum Likelihood Estimation

MPIOM   Max-Planck-Institute for Meteorology Ocean Model

NGL   Nevad geodetic Laboratory

NTAL   Nontidal Atmospheric Loading

NTOL   Nontidal Ocean Loading

R2   R-Squared, Coefficient of Determination

RF   Random Forest

RMS   Root Mean Square

RMSE   Root Mean Squared Error

RNN   Recurrent Neural Network

SLTM   Standard Linear Trajectory Model

SSA   Singluar Spectrum Analysis

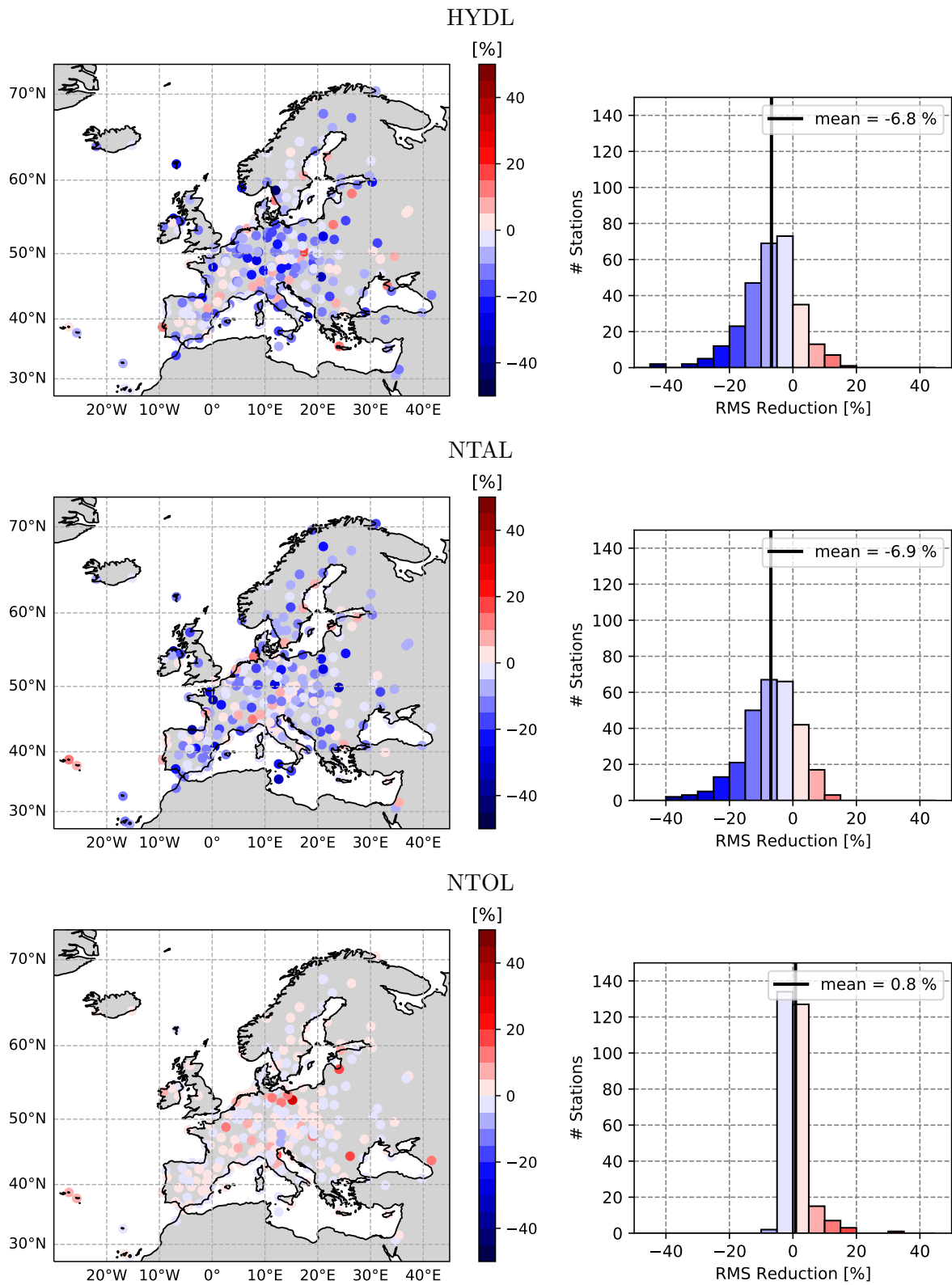TCN   Temporal Convolutional Network

Appendix

Figure 5.1: RMS reduction of the EPN time series, split into the individual environmental loadings
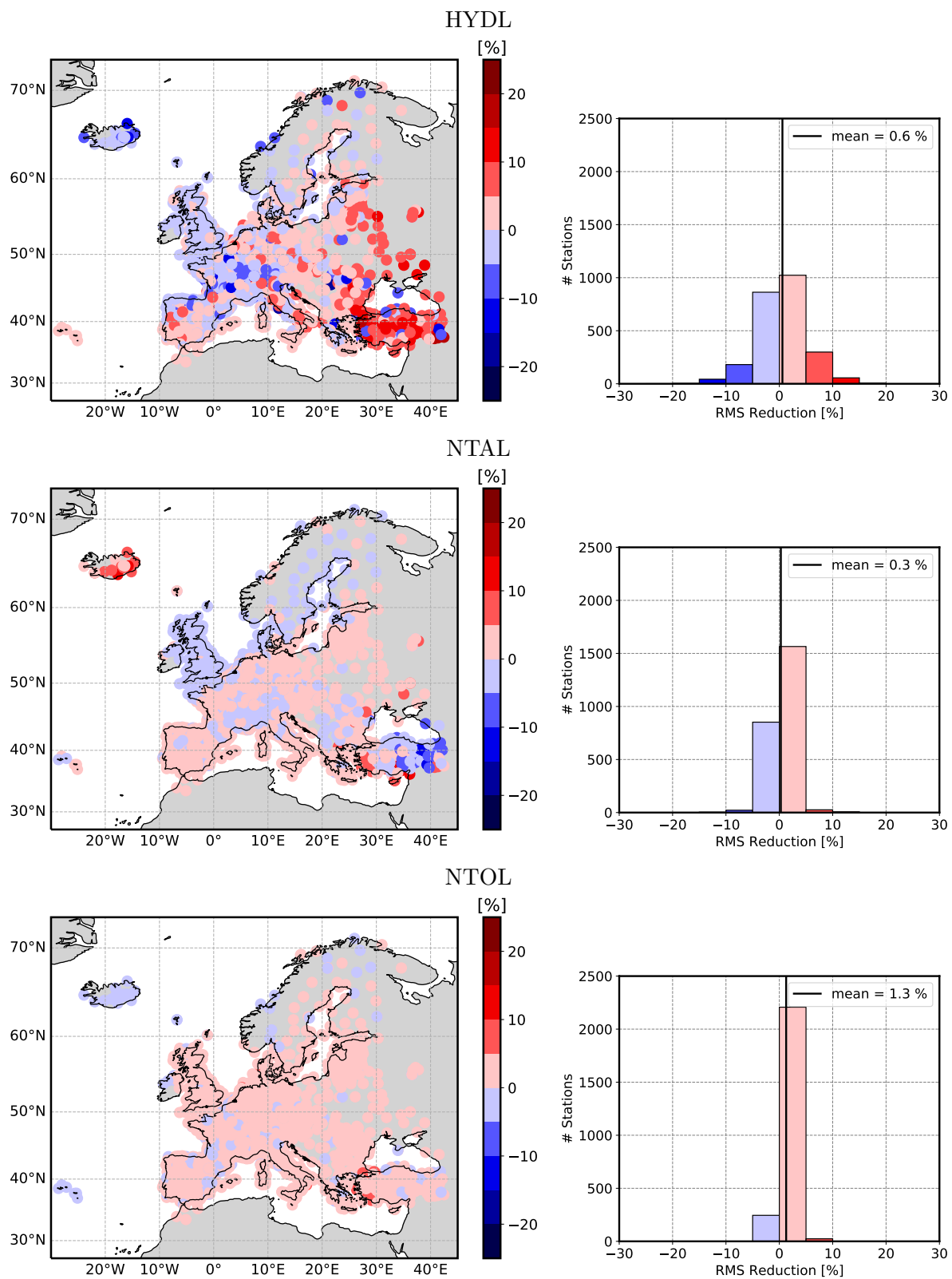
Figure 5.2: RMS reduction of the NGL time series, when first reconstructing the loadings with SSA
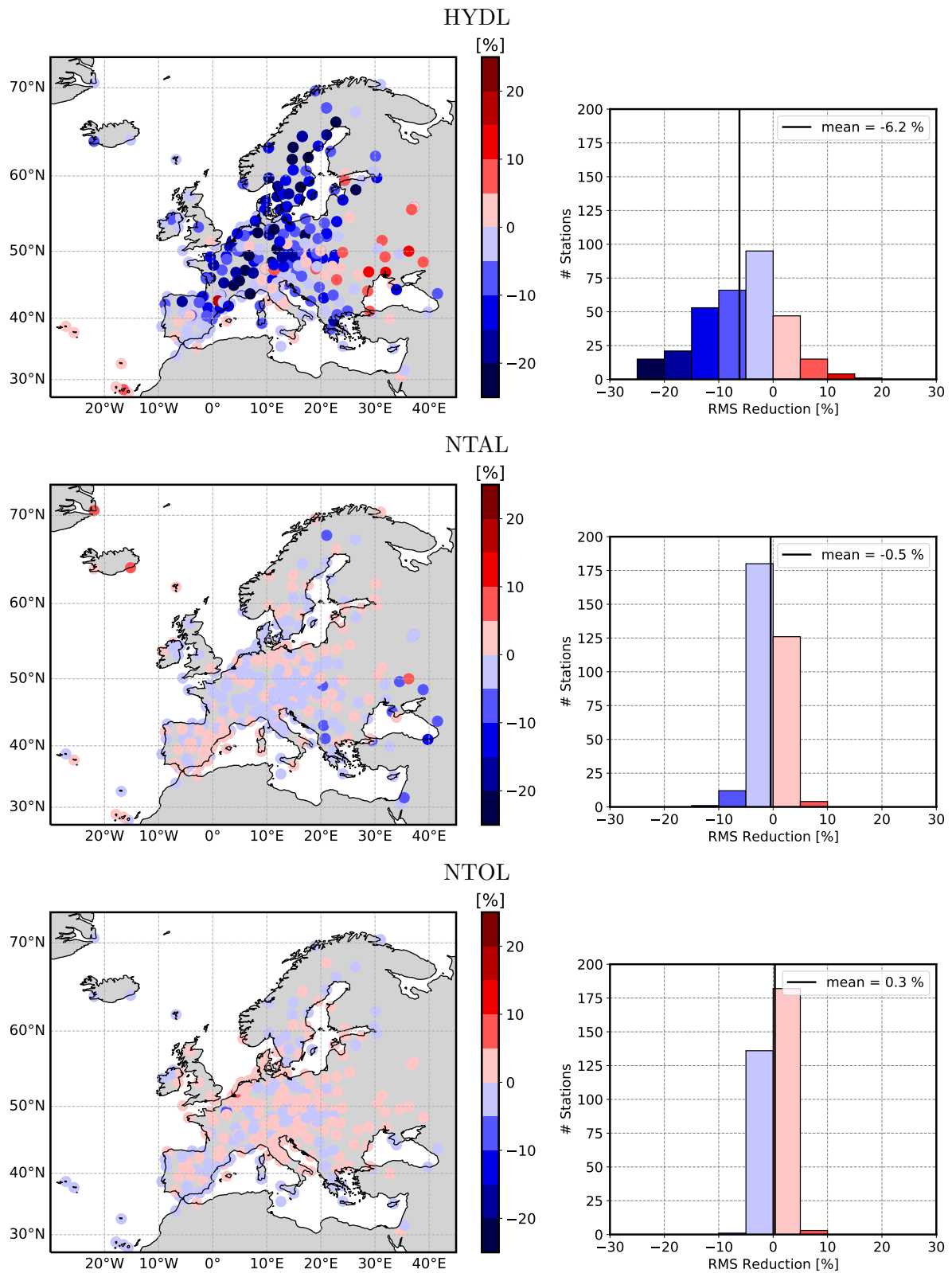
Figure 5.3: RMS reduction of the EPN time series, when first reconstructing the loadings with SSA
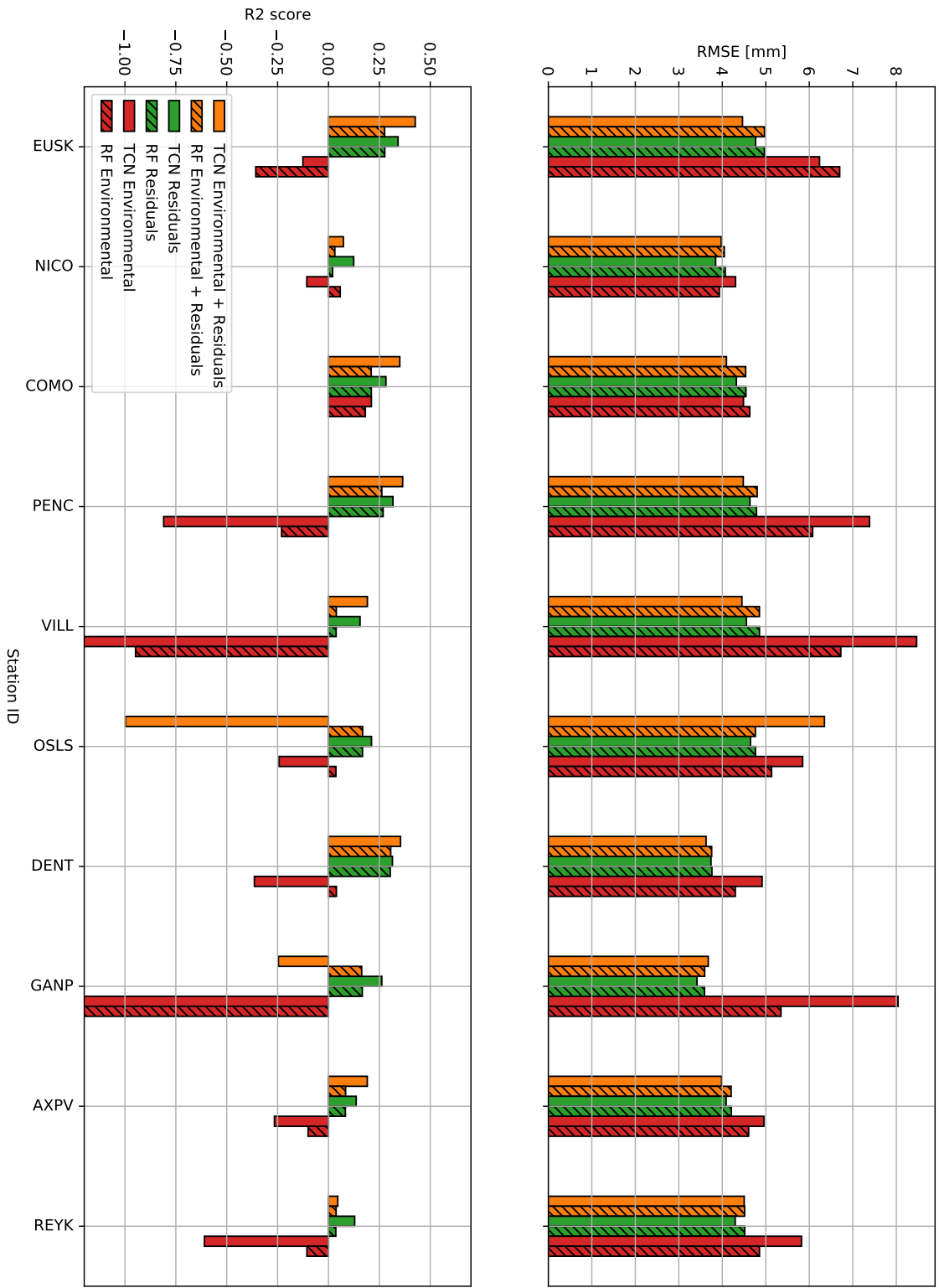
Figure 5.4: Comparison of TCN and RF prediction scores for all input variations

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

_____

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

ANALYSIS AND PREDICTION OF LONG TERM GNSS HEIGHT TIME SERIES AND ENVIRONMENTAL LOADING EFFECTS

Authored by (in block letters):
For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
Ruttner	Pia

With my signature I confirm that
  − I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
  − I have documented all methods, data and processes truthfully.
  − I have not manipulated any data.
  − I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**
Zürich, 25.01.2021	*Pia Ruttner*

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.