Statistical error analysis of a nitrate deposition model

LEON ANDERSSON
Master's thesis

December 2014

Abstract

ECDS, Environment Climate Data Sweden, is a commitment that SMHI rendered the Science Council of Sweden. Their purpose is to assist with search, documentation and publication of data regarding environment and climate. ECDS offer services for data stored in their database. One such service that very useful is THREDDS, Thematic Realtilme Enviromental Distributed Data Service, which is a tool to make sections from large data sets and visualize data. The purpose of this Master's Thesis is to combine data from different sources in ECDS portal and create added value by co-evaluation. The essay is then used as a showcase for ECDS.

SMHI, the Swedish Meteorological and Hydrological Institute, monitors the deposition of many different pollutants using direct measurements and model calculations. Data from measurement stations are often very accurate but there are too few stations to provide knowledge about the geographic distribution of pollutants. Therefore an important compliment to measurements are model calculations, although these are less accurate.

SMHI uses a deterministic hierarchical model to predict deposition. The most important component of this model is the atmospheric transport of substances model, called MATCH, Multi-scale Atmospheric Transport and Chemistry.

To use data from a model it is important to know how well the predictions reproduce reality. The aim of this Master's Thesis is to perform an error analysis of the deposition model in wet-deposition of nitrate, NO_3, by comparing model calculations with real measurements. The error is constructed as the ratio between the model predictions and measurement data. To be able to explain the model error a linear regression model, using a subset of the input-parameters in MATCH as covariates, is constructed. At each observation site the same set of covariates is used but the regression parameters are re estimated. Building a second linear model in the spatial variation in the regression coefficient makes it possible to predict and analyze the error everywhere in Europe. Combining the two models results in a mixed-effect model.

Measurements data, model calculations and explanatory variables were mostly found in the ECDS's database. The error analysis is developed based on 149 different measurement stations and the 85×95 model grid of MATCH over Europe. The Mixed-effect model is tested at 9 randomly selected station sites and the results are promising.

Keywords

Nitrate, Mixed Effect Model.

Populärvetenskaplig sammanfattning

Efter Tjernobylkatastrofen 1987 utvecklade SMHI en modell, MATCH, för att spåra kemikalier i atmosfären. Syftet var att studera förorenat regn i Sverige. Idag används MATCH för att studera konsekvenser av luftföroreningar så som försurning och övergödning.

Modelldata används alltid som komplement till riktig mätdata. Mätdata från mätstationer är exaktare men mer kostsamma än modelldata. För att använda modelldata är det viktigt att veta hur trovärdig datan är. För att utvärdera detta jämförs modelldata med mätdata.

Genom att använda en statistisk regressions modell, bestående av klimatparametrar som kovariat, för att prediktera prediktionsfelet av MATCH kan man utvärdera vilka och hur mycket olika klimatparametrar inverkar på modellfelet.

För att använda felmodellen som en efterberarbetningsmodell måste man dock vid varje plats ha tillgång till depositions mätdata och klimatdata för att beräkna koefficienterna till kovariaten. Genom att konstruera en andra regressionsmodell som berskriver hur koefficienterna beror av spatiala variabler, kan prediktionsfelet av MATCH predikteras även där det inte finns deposition mätdata.

Den sammansatta felmodellen är en så kallad mixed effect model och parametrarna till denna modell har beräknats med hjälp av mätdata från 140 mätstationer spridda över Europa, climatdata och modelldata från MATCH. Felmodellen har sedan testats vid 9 mätstationer och resultatet är lovande.

Resultat från felanalysen visar att nederbörd är den största bidragande faktorn till prediktionsfelet av MATCH, sedan vind, luftfuktighet och temperatur. Genom att använda felmodellen för efterbearbetning förbättras prediktionerna avsevärt.

En vidareutveckling av examensarbetet vore att införa fler kovariat till koefficientmodellen. Det verkar nämligen som att koefficienterna beror av mer regionala parametrar. Landskapets utseende kan vara en avgörande faktor, är plattsen ett öppet fält, en skogsmiljö, en bergsmiljö eller en stadsmiljö.

Förord

Detta examensarbete skrevs under under första läsperioden av vårterminen, sommaren och första läsperioden av höstterminen år 2014 vid SMHI (Swedish Meteorological and hydrological Institute) i Norrköping och vid LTH i Lund vid institutionen Matematikcentrum.

Jag skulle vilja tacka min handledare på SMHI Cecilia Bennet, Civ.ing i Teknisk fysik och Britt Frankenberg, Civ.ing Teknisk fysik som alltid stöttat och gett mig värdefull feedback.

Jag är skyldig min examinator Johan Lindström, Civ.ing i statistik, ett stort tack till det akademiska bidraget och stödet i processen att skriva detta examensarbete.

Preface

This Master's Thesis was written during the second spring term, summer and first autumn term of 2014 at the SMHI (Swedish Meteorological and hydrological Institute) in Norrköping and at LTH in Lund at the institute Matematikcentrum.

I want to give my many thanks to my supervisors at SMHI Cecilia Bennet, PhD in physics and Britt Frankenberg, PhD in physics. who at all times has encouraged me and given me valuable feedback.

To my examiner Johan Lindström, PhD in statistics, I owe many thanks for academic input and support in the process of writing my thesis.

Contents

Chapter 1

Introduction

During the the late eighties SMHI begin development of a limited-area atmospheric transport model called MATCH, Multiple-Scale Atmospheric Transport and Chemistry Modeling System, to trace chemicals in the atmosphere. The development was motivated by the accident at Chernobyl in 1987 and by the need to study the deposition of acid rain in Sweden. In MATCH the atmosphere is divided in to a large number of gridboxes. In each gridbox an Eulerian framework is used to calculate the quantity of several different chemicals. Using discrete time step the quantities of these chemicals are updated by calculation of chemical transformations, inflow from surrounding gridboxes surface emissions from the outflow to surrounding gridboxes or deposition to the surface.[1]

Today MATCH is a tool to study the consequences of air pollution such as fertilization, acidification and greenhouse. The model traces components of chemicals in a three dimensional space and predicts where they are being deposited.

In order to evaluate a deterministic model's ability to describe reality the model predictions are compared to measurements. Building a statistical model using, e.g. regression, for the prediction errors allows us to evaluate how much of the error that can be explained by each covariate . In this Master's Thesis an error analysis is done on MATCH. The analysis is based on a model run of MATCH from 1980 to 2010 and measurements from 149 measurement stations in Europe. The substance that is studied is wet-deposition (deposition by precipitation such as rain or snow) of nitrogen.

The prediction error at each measurement station is modeled using a linear regression model with climate covariates that are input to MATCH. The spatial variation in regression coefficients at different stations of the covariates are then modeled as a stochastic field with a mean depending on spatial data and a covariance function depending on the distance between the sites. The purpose of the coefficients model is to be able to comment on the error where there are no measurement stations, making it possible to analyze the error everywhere in

Europe.

1.1 Outline of the report

Initially a general description of the measurement stations and the MATCH model is accounted together with data in Chapter 2. Then a theory Chapter follows, describing the MATCH prediction error model, variable selection and spatial dependence. In Chapter 3 the results are presented. Discussion, conclusions and further work are given in Chapter 4. Finally some appendices are attached; they contain additional results which are too detailed to be included in the results section.

Chapter 2

Data

In this Chapter the available data is presented. Data in this report consists of MATCH predictions, measurement data, climate data such as temperature, pressure, humidity, wind, cloudiness and spatial data regarding the measurement stations such as longitude, latitude, altitude and distance to coast. The climate data make up a subset of the input parameters used by MATCH and are here used as covariates to explain errors between MATCH predictions and observational data.

Measurements data, model calculations and explanatory variables were mostly found in the ECDS database. ECDS, Environment Climate Data Sweden, is a commitment that SMHI rendered the Science Council of Sweden. Their purpose is to assist with search, documentation and publication of data regarding enviroment and climate. ECDS offer services for data stored in their database. One such service that was very helpful is THREDDS, Thematic Realtilme Enviromental Distributed Data Service, which is a tool to make sections from large data sets and visualize data.

Lastly we explain how MATCH's prediction error is constructed.

2.1 MATCH predictions

MATCH predictions consists of daily data within the period 1980-01-01 to 2010-12-31 from a 85×95 grid covering Europe as seen in Figure 2.1. The predictions are mean nitrate deposition per square meter (g/m^2) of a grid box.

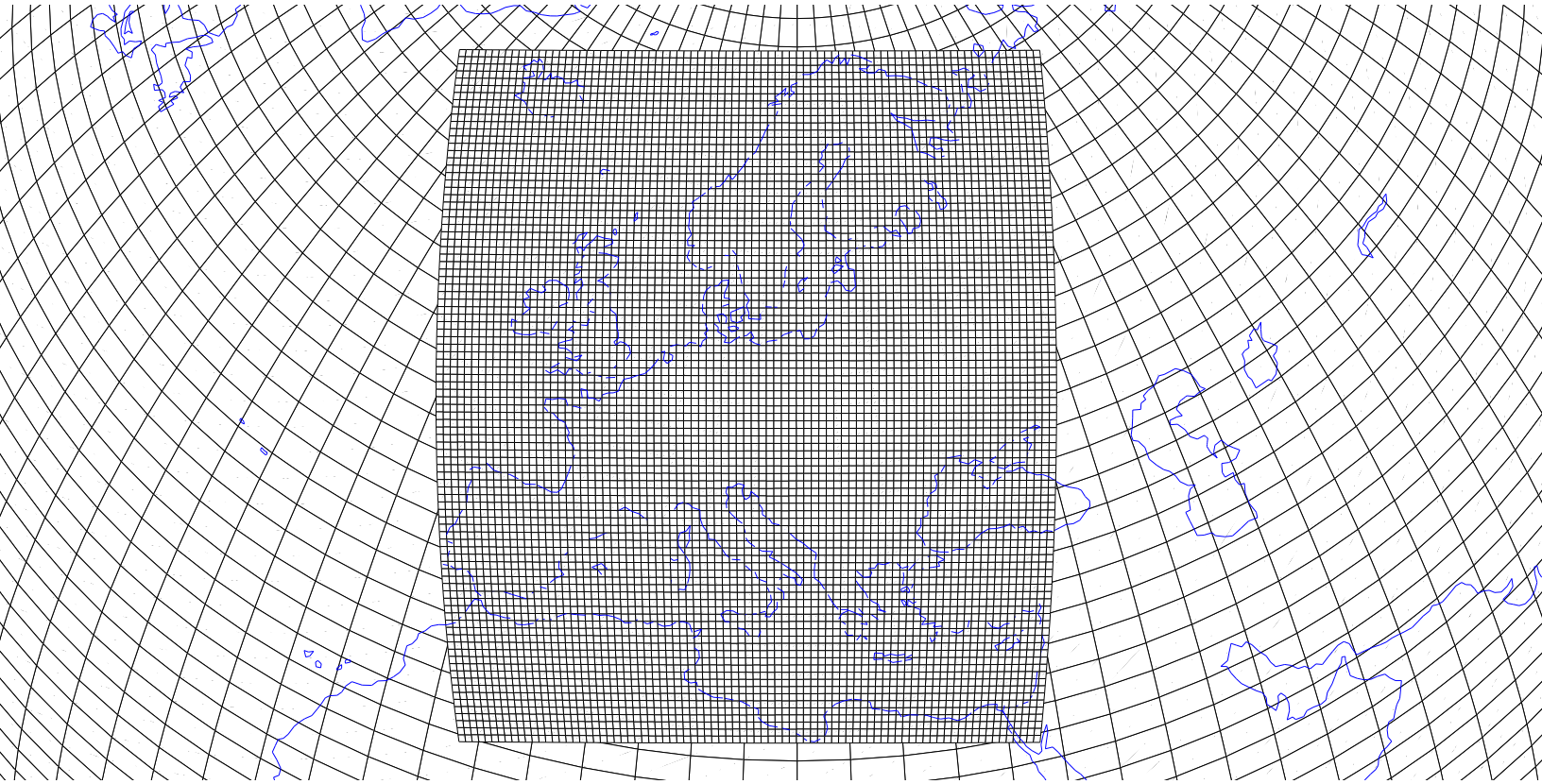

Figure 2.1: Grid of MATCH which has rotated $-39.3°$ latitude and $18°$ south pole

As seen in Figure 2.1 the longitudinal and latitudinal lines of the MATCH gird do not comply with the bigger grid which is the ordinary coordinate system. This is because the MATCH grid is given in a rotated coordinate system. The surface dimensions of the grid boxes are defined in a polar coordinate system. The atmosphere is divided uniformly according to these coordinates which means that the boxes will not have the same euclidean size. Using the ordinary polar coordinate system will for example result in mush smaller boxes in Sweden than in Greece. To avoid complications in the model the base coordinates are in a rotated coordinate system making the boxes more equally sized. All grid boxes have surfaces smaller than $49km \times 49km$ and do not differ much in size. The height of the boxes depends on the pressure where the box is placed. There are 61 levels of boxes, level one is at surface.

2.2 Measurement

The measurement data is from EMEP's measurement net, which consists of 315 MISU-samplers. A MISU-sampler is a lid-sampler designed to avoid dry-deposition [4]. The measurements are given as daily concentration of nitrate in precipitation, gram of nitrate per kilogram of water (g/kg), and the precipitation in millimeters (mm). Given this the concentration in gram per square meter is calculated.

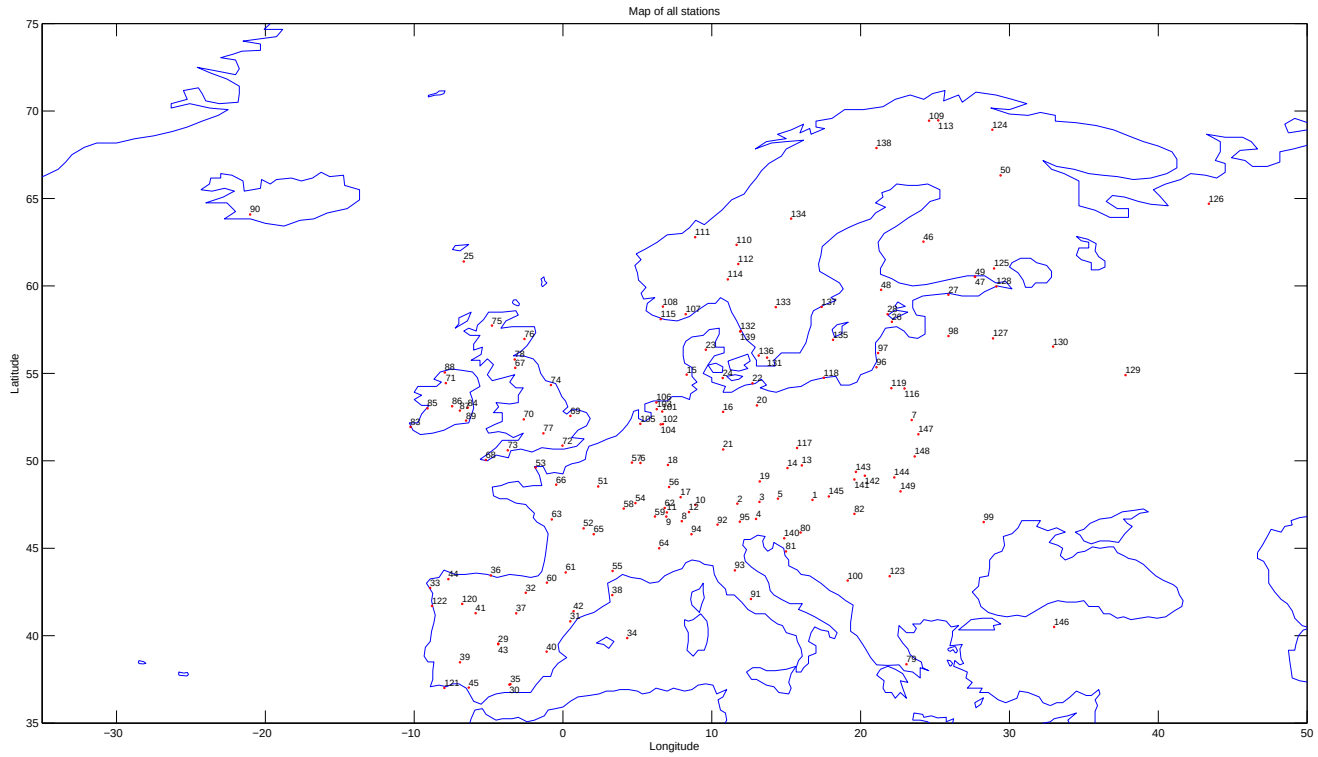

Figure 2.2: Map of the stations.

Out of 315 measurement stations from EMEP only 149 stations contained useful data of wet-deposited nitrate. The excluded stations contained either no data at all, too little data, too little data within the period of the MATCH simulation (1980-2010), just weekly or monthly data. There were also stations positioned too far from the MATCH grid. Figure 2.2 shows all the stations used for the study. A table of data regarding longitude, latitude and distance to sea is attached in the Appendix.

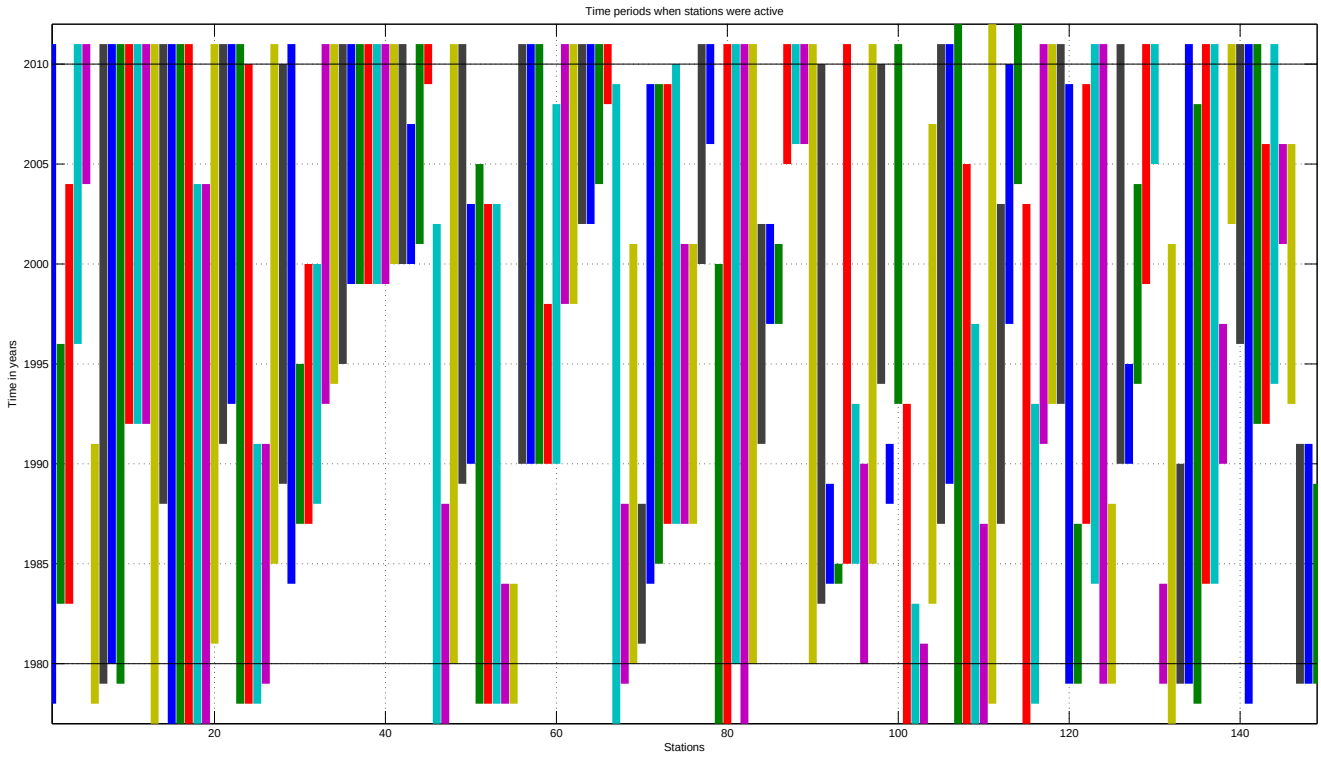

Figure 2.3: Time period when stations were active according to meta data. The black lines are the years 1980 and 2010.

As seen in Figure 2.3 the duration of the sampling at each station varies a lot. The periods in Figure 2.3 are according to the meta data.

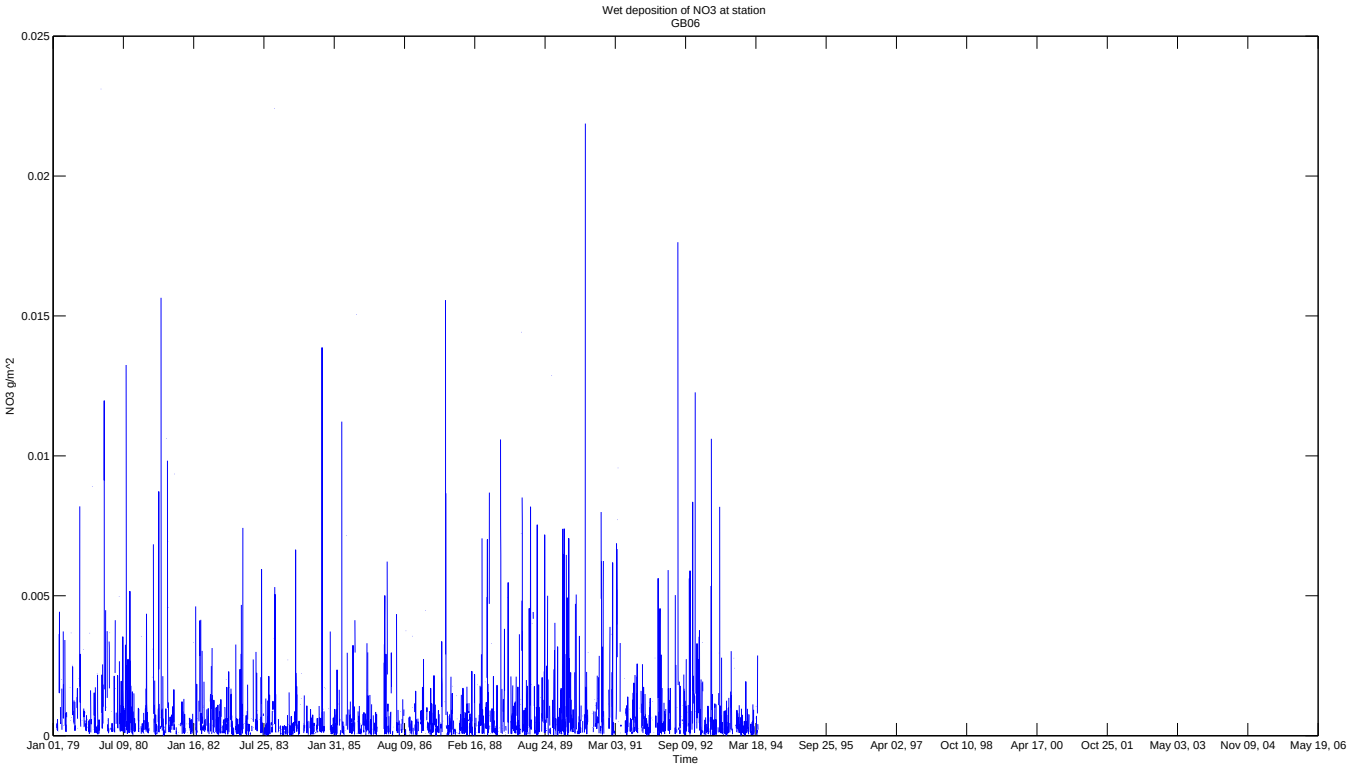

Figure 2.4: Wet deposition of NO^3 at station GB06.

Within these periods there are a lot times of gaps. An example of this is seen in Figure 2.4 where observations end in 1994, due to no rainfall or errors. The reason why the curve in Figure 2.4 is not continuous is because no data was captured at that time due to no rain or errors. The error analysis is only done on captured data which means blank spaces are not concerned.

2.2.1 Climate data

The climate data is originally from a re-analysis from ECMWF called ERA40. It contains daily data on a $96 \times 99 \times 61$ gird in the same rotated coordinate system as MATCH. The third dimension represents the height above surface defined through constant pressure levels, level 61 is the lowest level, i.e the surface. The climate data is from model calculations and represents the mean of each grid box.

The precipitation data is given at surface level in millimeters (mm). There

are three levels from where there is data of the temperature, surface level, level 60 and level 59. The temperature is given in Kelvin. The pressure (ps) is given as

$$p = p_0 + p_h \tag{2.1}$$

where p_0 is the pressure at surface level and p_h is the difference between the surface level pressure and the pressure at level h. The unit is Pascal (Pa). The mean humidity (qh) during a day is given at surface in kilogram water per kilogram air (kg/kg). Wind is given as mean daily value, in meters per second (m/s), and divided into two components, an easterly u-component and a northerly v-component. Cloud water contents (CWC) is given in kilogram water per cubic meter (kg/m^3) at level 60. Cloud cover (CC) is the fraction of the grid box that is covered by cloud at level 60. Total cloud cover (TCC) is the cloud cover for all the levels together.

2.2.2 Spatial data

Longitude, latitude and altitude for the stations are given as meta data in the EMEP and ERA40 data. Distance to coast is the smallest distance between the station and the coast line. Coordinates for the coast line were found on MathWorlds home page together with the altitude for sites where there are no stations.

2.3 Constructing the prediction error

Attempts of using a linear regression model to describe the additive error were made but variance of the residuals tended to be larger for larger errors. A better approach is to assume that the prediction error is multiplicative and constructing the prediction error as the ratio between the measurement data and the prediction.

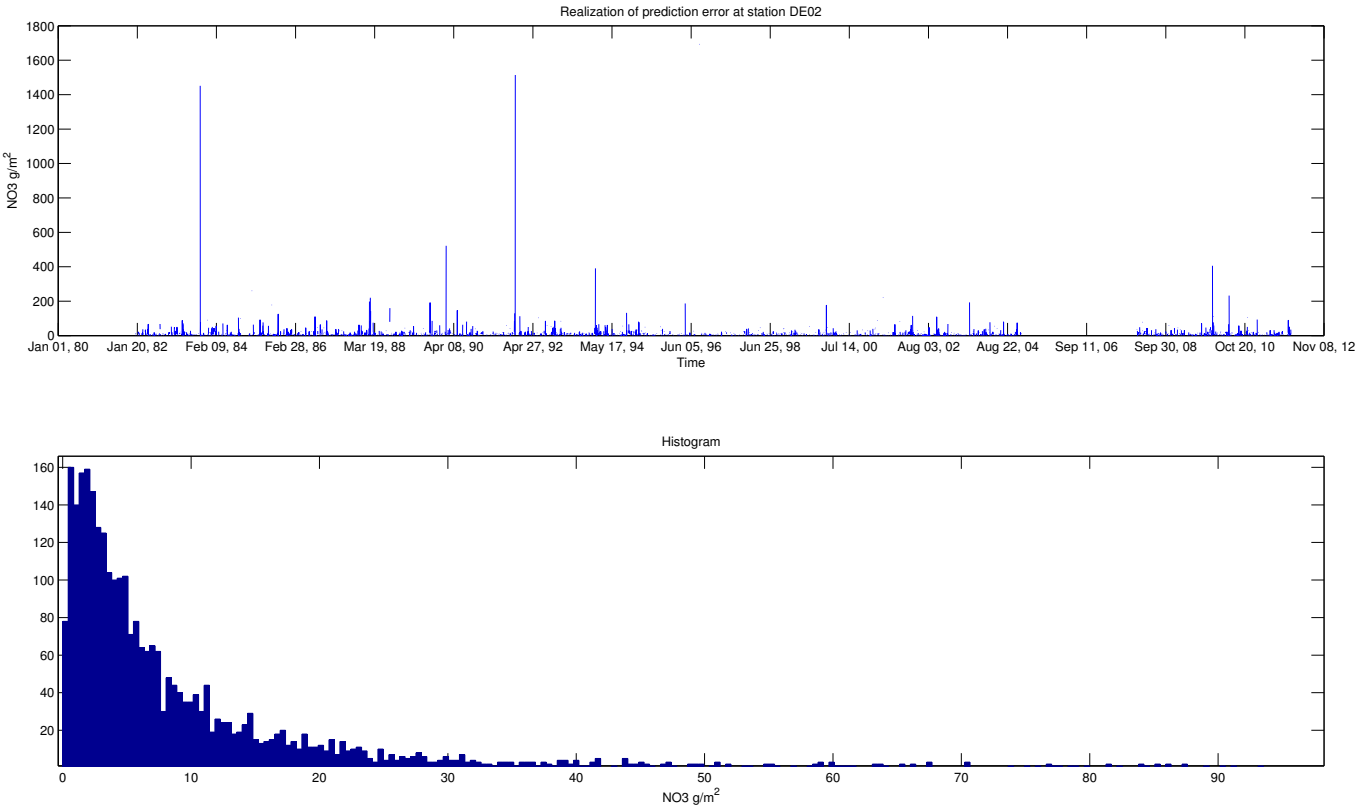

Figure 2.5: Example of the multiplicative error (measurement data divided with MATCH prediction) at station DE02 (prediction errors from other stations looks similar).

As seen in Figure 2.5 the distribution is skewed and has a heavy right tail which is hard to capture using only linear regression. To make the prediction error easier to model Box and Cox transformation is used [5]. To evaluate how the prediction error should be transformed the Box and Cox parameter is calculated for the prediction error at all stations.

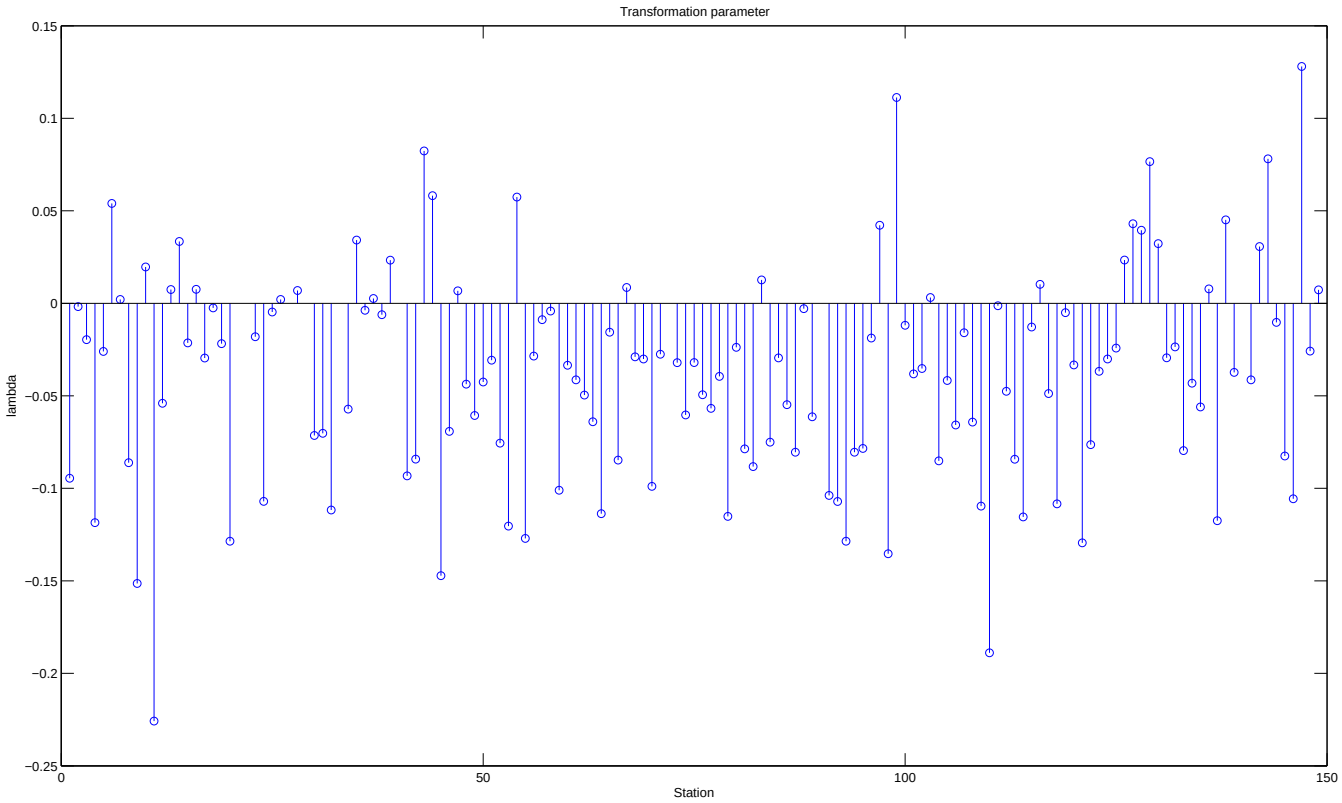

Figure 2.6: The value of the Box and Cox transformation parameter given at stations for valuation

In accordance with Figure 2.6 the transformation parameter is chosen to zero, i.e the log-transform is used, at all stations.

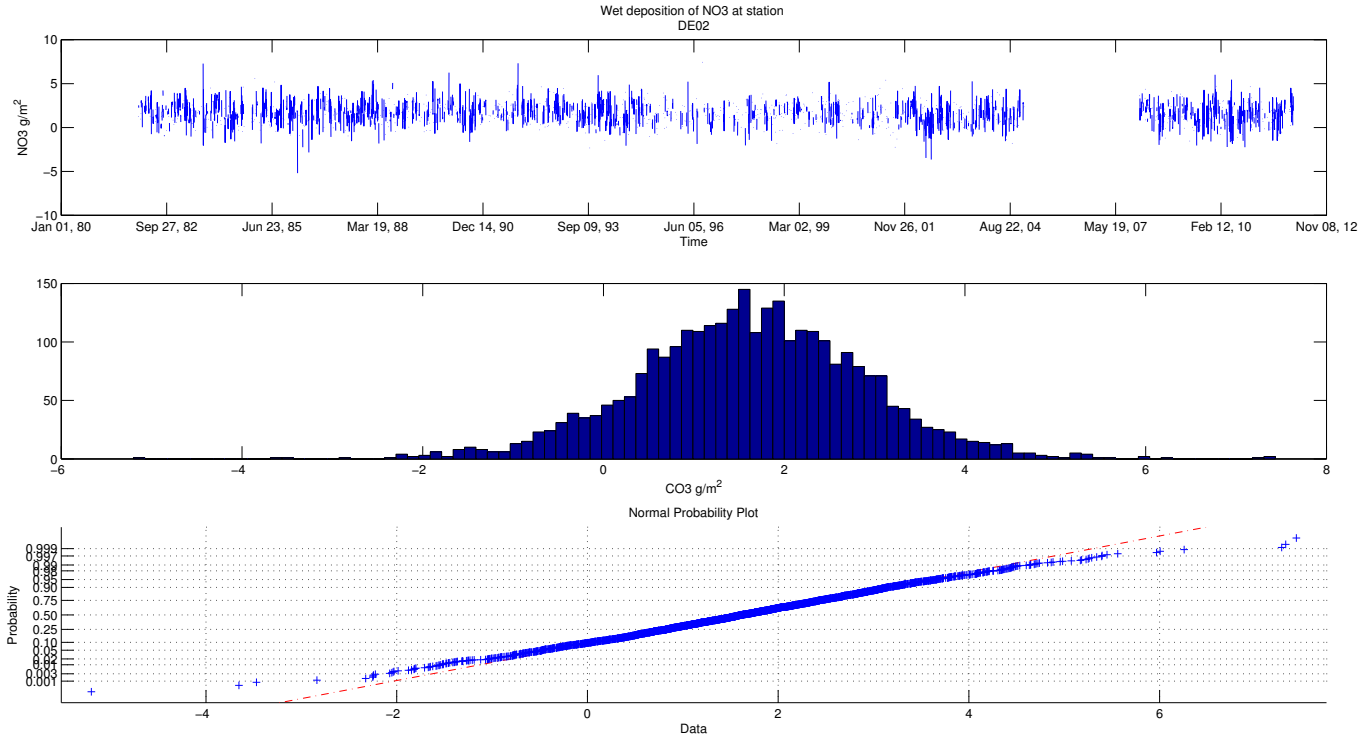

Figure 2.7: Example of the log-transformed multiplicative error at station DE02.

Chapter 3

Theory

This chapter contains theory necessary to construct the mixed-effect model used to predict the MATCH error as well as comments on how to chose covariates.

3.1 Prediction error model

In this section the hierarchic model is presented. It consists of a site model (or station model) and spatial models for the coefficients in the site model. The site model is a linear regression model that describing how the log-transformed prediction error depends on climate covariates at a each site. The spatial models describes how the coefficients of the covariates in the site model depends on spatial covariates. Both the models are linear for simplicity and only a selection of the climate covariates and the spatial covariates presented in the data chapter are used. The nonselected covariates are significantly proven to not depend on the prediction error or the coefficients. How this selection is done will be explained later in this chapter. Before we start we need some notations. Let t denote the time (in days) and let $\boldsymbol{s}$ denote the position on the earth in a three dimensional cartesian coordinate system (in mil) given by

$$\boldsymbol{s} = \boldsymbol{f}(\alpha, \phi, \lambda) = \begin{bmatrix} (R+\alpha)cos(\lambda)cos(\phi) \\ (R+\alpha)sin(\lambda)cos(\phi) \\ (R+\alpha)sin(\phi) \end{bmatrix} \tag{3.1}$$

where R is the radius of the earth, α the altitude, λ the longitude and ρ the latitude.

Site model

At position $\boldsymbol{s}$ the site model is given by

$$\begin{aligned} Y(t,\boldsymbol{s}) =& \boldsymbol{a}(t,\boldsymbol{s})\boldsymbol{\Theta}(\boldsymbol{s}) + \epsilon(t,\boldsymbol{s}) \\ =& \begin{bmatrix} a_1(t,\boldsymbol{s}) & a_2(t,\boldsymbol{s}) & \dots & a_p(t,\boldsymbol{s}) \end{bmatrix} \begin{bmatrix} \Theta_1(\boldsymbol{s}) \\ \Theta_2(\boldsymbol{s}) \\ \vdots \\ \Theta_p(\boldsymbol{s}) \end{bmatrix} + \epsilon(t,\boldsymbol{s}) \end{aligned} \tag{3.2}$$

where $Y(t,\mathbf{s})$ is the log-transformed prediction error, $a_1(t,\boldsymbol{s}), a_2(t,\boldsymbol{s})...a_p(t,\boldsymbol{s})$ the climate covariates, $\Theta_1(\boldsymbol{s}), \Theta_2(\boldsymbol{s})...\Theta_p(\boldsymbol{s})$ their coefficients and $\boldsymbol{\epsilon}(t,\boldsymbol{s})$ a Gaussian independent errors with variance $\sigma(\boldsymbol{s})$. This is just a linear regression model only depending on the temporally varying climate covariates and noise.

Spatial model

The spatial model is given by

$$\boldsymbol{\Theta}(\boldsymbol{s}) = \boldsymbol{b}(\boldsymbol{s})\boldsymbol{\beta} + \boldsymbol{u}(\boldsymbol{s}) = \begin{bmatrix} \boldsymbol{b}_1(\boldsymbol{s}) & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{b}_2(\boldsymbol{s}) & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{b}_p(\boldsymbol{s}) \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_p \end{bmatrix} + \begin{bmatrix} u_1(\boldsymbol{s}) \\ u_2(\boldsymbol{s}) \\ \vdots \\ u_p(\boldsymbol{s}) \end{bmatrix} \tag{3.3}$$

where

$$\boldsymbol{b}_i(\boldsymbol{s}) = \begin{bmatrix} b_1(\boldsymbol{s}) & \dots & b_{m_i}(\boldsymbol{s}) \end{bmatrix} \quad \text{and} \quad \boldsymbol{\beta}_i = \begin{bmatrix} \beta_{i,1} \\ \vdots \\ \beta_{i,m_i} \end{bmatrix} \tag{3.4}$$

contains the spatial covariates for the i:th coefficient and corresponding coefficients. The number of spatial covariates regarding the i:th coefficient of the site model is m_i. $\boldsymbol{u}(\boldsymbol{s})$ is a zero mean stochastic error vector. The errors at different sites have spatial dependence given by

$$\begin{aligned} \mathbb{C}(\boldsymbol{u}(\boldsymbol{s}), \boldsymbol{u}(\boldsymbol{s'})) =& \boldsymbol{M}(\|s - s'\|) \\ =& \begin{bmatrix} M_1(\|s-s'\|) & 0 & \dots & 0 \\ 0 & M_2(\|s-s'\|) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & M_p(\|s-s'\|) \end{bmatrix} \end{aligned} \tag{3.5}$$

where $\|\bullet\|$ is the L_2-norm. The covariance between the coefficients that regards different covariates is thus zero and the covariance between the coefficients that regards the same covariate depends only of the distance. All the functions $M_1(\|s-s'\|)\dots M_p(\|s-s'\|)$ belongs to the same family of functions and we will come back to them later in this chapter.

Composite model

Inserting the equation for the spatial model (3.3) into the equation for the site model (3.1) gives

$$\begin{aligned} Y(t,\boldsymbol{s}) &= \boldsymbol{a}(t,\boldsymbol{s})(\boldsymbol{b}(\boldsymbol{s})\boldsymbol{\beta}+\boldsymbol{u}(\boldsymbol{s}))+\boldsymbol{\epsilon}(\boldsymbol{t},\boldsymbol{s}) \\ &= \boldsymbol{a}(t,\boldsymbol{s})\boldsymbol{b}(\boldsymbol{s})\boldsymbol{\beta}+\boldsymbol{a}(\boldsymbol{s})\boldsymbol{u}(\boldsymbol{s})+\boldsymbol{\epsilon}(\boldsymbol{t},\boldsymbol{s})=\boldsymbol{x}(\boldsymbol{s})\boldsymbol{\beta}+\boldsymbol{a}(\boldsymbol{s})\boldsymbol{u}(\boldsymbol{s})+\boldsymbol{\epsilon}(t,\boldsymbol{s}) \end{aligned} \quad (3.6)$$

where $\boldsymbol{x}(\boldsymbol{s})=\boldsymbol{a}(t,\boldsymbol{s})\boldsymbol{b}(\boldsymbol{s})$. The model has a random effect coming from $\boldsymbol{\epsilon}(t,\boldsymbol{s})$ which is random in time as well as space and a random effect coming from $\boldsymbol{u}(\boldsymbol{s})$ which is only random in space. The random effects can not be interpreted as one random effect without loss of generality. This is called a mixed (random) effect model. Both $\boldsymbol{\epsilon}(t,\boldsymbol{s})$ and $\boldsymbol{u}(\boldsymbol{s})$ are normally distributed and independent of each other.

3.1.1 Prediction error model on matrix form

Lets denote the number of measurement stations with r and the number of observations at the $i:th$ station with n_i. The position of the stations are $\boldsymbol{s}_1\dots\boldsymbol{s}_r$ and the prediction error is observed at times $t_{1,i},\dots,\quad t_{n_i,i}$ for the i:th station. The log-transformed prediction error at the $i:th$ station of the j:th observation time is $Y_{i,j}=Y(t_j,\boldsymbol{s}_i)$. Combining all observations at site i into a vector $\boldsymbol{Y}_i$ we have:

$$\boldsymbol{Y}_i = A_i\boldsymbol{\Theta}_i+\boldsymbol{\epsilon}_i \quad (3.7)$$

where

$$A_i=\begin{bmatrix}\boldsymbol{a}(t_1,\boldsymbol{s}_i)\\ \boldsymbol{a}(t_2,\boldsymbol{s}_i)\\ \vdots\\ \boldsymbol{a}(t_{n_i},\boldsymbol{s}_i)\end{bmatrix} \quad (3.8)$$

is a matrix whose columns' consist of each climate covariates for all observations times and $\boldsymbol{\Theta}_i=\boldsymbol{\Theta}(\boldsymbol{s_i})$ the coefficient vector at station i. The error vector $\boldsymbol{\epsilon}_i$ is stacked in the same way as $\boldsymbol{Y}_i$. Combining all the site models leads to a joint model, which can be expressed as

$$\boldsymbol{Y}=\begin{bmatrix}A_1 & 0 & \dots & 0\\ 0 & A_2 & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \dots & A_r\end{bmatrix}\begin{bmatrix}\boldsymbol{\Theta}_1\\ \boldsymbol{\Theta}_2\\ \vdots\\ \boldsymbol{\Theta}_r\end{bmatrix}+\begin{bmatrix}\boldsymbol{\epsilon}_1\\ \boldsymbol{\epsilon}_2\\ \vdots\\ \boldsymbol{\epsilon}_r\end{bmatrix}=A\boldsymbol{\Theta}+\boldsymbol{\epsilon} \quad (3.9)$$

where the error term $\boldsymbol{\epsilon}$ is stacked in the same way as $\boldsymbol{Y}$. The covariance matrix for $\boldsymbol{\epsilon}$ is

$$R=\begin{bmatrix}I\sigma_1 & 0 & \dots & 0\\ 0 & I\sigma_2 & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \dots & I\sigma_r\end{bmatrix} \quad (3.10)$$

where $\sigma_i=\sigma(\boldsymbol{s}_i)$ is the variance of the error at station i.

3.1.2 Spatial model

The $k:th$ coefficient at the $i:th$ station is given by

$$\Theta_{k,i} = \boldsymbol{b}_{k,i}\boldsymbol{\beta}_{k,i} + u_{k,i} \tag{3.11}$$

where $\boldsymbol{b}_{k,i} = \boldsymbol{b}_k(\boldsymbol{s}_i)$ is a column vector containing spatial covariates and $u_{k,i} = u_k(\boldsymbol{s}_i)$ is the spatial noise. Since the coefficients do not depend on the same covariates the length of $\boldsymbol{b}_{k,i}$ varies with k. The coefficient vector for the $i:th$ stations is then given by

$$\boldsymbol{\Theta}_i = \begin{bmatrix} \boldsymbol{b}_{1,i} & 0 & \dots & 0 \\ 0 & \boldsymbol{b}_{2,i} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \boldsymbol{b}_{l,i} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_l \end{bmatrix} + \boldsymbol{u}_i = \boldsymbol{B}_i\boldsymbol{\beta} + \boldsymbol{u}_i \tag{3.12}$$

where $\boldsymbol{u}_i = [\boldsymbol{u}_{1,i}^T, ..., \mathbf{u}_{\mathbf{l,i}}^{\mathbf{T}}]^{\mathbf{T}}$. For all stations the model can be expressed as

$$\boldsymbol{\Theta} = \boldsymbol{B}\boldsymbol{\beta} + \boldsymbol{u} = \begin{bmatrix} \boldsymbol{B}_1 \\ \boldsymbol{B}_2 \\ \vdots \\ \boldsymbol{B}_l \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \\ \vdots \\ \boldsymbol{u}_l \end{bmatrix} \tag{3.13}$$

where $\boldsymbol{\Theta} = [\boldsymbol{\Theta}_1^T, ..., \boldsymbol{\Theta}_r^T]^T$. The covariance matrix of $\boldsymbol{u}$ is denoted by

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{M}_{1,1} & \boldsymbol{M}_{1,2} & \dots & \boldsymbol{M}_{1,r} \\ \boldsymbol{M}_{2,1} & \boldsymbol{M}_{2,2} & \dots & \boldsymbol{M}_{2,r} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{M}_{r,1} & \boldsymbol{M}_{r,2} & \dots & \boldsymbol{M}_{r,r} \end{bmatrix} \tag{3.14}$$

where $\boldsymbol{M}_{i,j} = \boldsymbol{M}(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|)$.

3.1.3 Assembled model

The hierarchic model can now be summarized as

$$\begin{cases} \boldsymbol{Y} = \boldsymbol{A}\boldsymbol{\Theta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{R}) \\ \boldsymbol{\Theta} = \boldsymbol{B}\boldsymbol{\beta} + \boldsymbol{u}, \boldsymbol{u} \sim N(\mathbf{0}, \boldsymbol{D}). \end{cases} \tag{3.15}$$

The coefficient model inserted in the prediction error gives

$$\boldsymbol{Y} = \boldsymbol{A}(\boldsymbol{B}\boldsymbol{\beta} + \boldsymbol{u}) + \boldsymbol{\epsilon} = \boldsymbol{A}\boldsymbol{B}\boldsymbol{\beta} + \boldsymbol{B}\boldsymbol{u} + \boldsymbol{\epsilon} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{B}\boldsymbol{u} + \boldsymbol{\epsilon}. \tag{3.16}$$

where $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{B}$.

3.2 Mixed effects

Mixed random effects models or mixed models provide a powerful tool for analysis of grouped data which arise in many areas such as biology, economics, manufacturing and geophysics[8] and was developed by C. R. Henderson together with S. R. Searle in the fifties [5].

The prediction errors can be divided into groups consisting of prediction error from the same sites or stations. The data within a group is then exposed to the same random effect from $\boldsymbol{u}(\boldsymbol{s})$.

3.2.1 Mixed effect model

In general the mixed effect model is given by

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Bu} + \boldsymbol{\epsilon}. \tag{3.17}$$

where $\boldsymbol{\beta}$ is a deterministic vector which represents the fixed effects, $\boldsymbol{u}$ is a stochastic vector representing the random effect [5]. $\boldsymbol{u}$ and $\boldsymbol{\epsilon}$ are assumed independent, zero mean, multivariate normal random variables with covariance matrices $\boldsymbol{D}$ and $\boldsymbol{R}$.

3.2.2 Henderson's mixed model equations

The conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{u}$ is multivariate normal with expectation $\boldsymbol{X\beta} + \boldsymbol{Bu}$ and variance $\boldsymbol{R}$. The joint distribution of $\boldsymbol{Y}$ and $\boldsymbol{u}$ is then

$$f(\boldsymbol{Y}, \boldsymbol{u}) = f(\boldsymbol{Y}|\boldsymbol{u})f(\boldsymbol{u}) \propto e^{-\frac{1}{2}(\boldsymbol{Y}-\boldsymbol{X\beta}-\boldsymbol{Bu})^T\boldsymbol{R}^{-1}(\boldsymbol{Y}-\boldsymbol{X\beta}-\boldsymbol{Bu})}e^{-\frac{1}{2}\boldsymbol{u}^T\boldsymbol{D}^{-1}\boldsymbol{u}}. \tag{3.18}$$

The maximum likelihood estimates of $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are then given as the solution to the minimization problem

$$\arg\min_{\boldsymbol{\beta},\boldsymbol{u}}(\boldsymbol{Y} - \boldsymbol{X\beta} - \boldsymbol{Bu})^T\boldsymbol{R}^{-1}(\boldsymbol{Y} - \boldsymbol{X\beta} - \boldsymbol{Bu}) + \boldsymbol{u}^T\boldsymbol{D}^{-1}\boldsymbol{u}. \tag{3.19}$$

Setting the gradient of this expression to zero gives

$$\begin{cases} \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X\beta} + \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Bu} = \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Y} \\ \boldsymbol{B}^T\boldsymbol{R}^{-1}\boldsymbol{X\beta} + \boldsymbol{B}^T\boldsymbol{R}^{-1}\boldsymbol{Bu} + \boldsymbol{D}^{-1}\boldsymbol{u} = \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Y} \end{cases} \tag{3.20}$$

and the solution to these equations is

$$\begin{cases} \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{Y} \\ \hat{\boldsymbol{u}} = \boldsymbol{D}\boldsymbol{B}^T\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X\beta}) \end{cases} \tag{3.21}$$

where $\boldsymbol{V}$ is the variance of $\boldsymbol{Bu} + \boldsymbol{\epsilon}$, i.e $\boldsymbol{B^T D^{-1} B} + \boldsymbol{R}$ [5].

3.3 Spatial dependence

The Matérn covariance function is a family of functions describing the covariance between two points in a random field given the distance between them. It was suggested by Bertil Matérn in 1960 in his doctoral dissertation about forestry and has been used in spatial statistics, geostatistics, image analysis and other applications. The Matérn covariance between two points (or sites) $\boldsymbol{s}$ and $\boldsymbol{s'}$ is

$$M(\|\boldsymbol{s}-\boldsymbol{s'}\|) = \frac{\sigma^2}{\Gamma(\nu)}(\kappa\|\boldsymbol{s}-\boldsymbol{s'}\|)^{\nu}\mathbf{K}_{\nu}(\kappa\|\boldsymbol{s}-\boldsymbol{s'}\|), \sigma \geq \mathbf{0}, \kappa > \mathbf{0}, \nu > \mathbf{0} \quad (3.22)$$

where K_ν is a modified Bessel function [9]. The Matérn parameters are estimated for each of the residual fields $u_1(\boldsymbol{s}) \dots u_p(\boldsymbol{s})$ using a least square algorithm. Since the fields are unknown they are estimated as

$$\hat{\boldsymbol{u}} = \hat{\boldsymbol{\Theta}} - \hat{\boldsymbol{\beta}}\boldsymbol{X} \quad (3.23)$$

where $\hat{\boldsymbol{\Theta}}$ is the estimate of $\boldsymbol{\Theta}$ at each measurement station and $\hat{\boldsymbol{\beta}}$ the least square estimate of $\boldsymbol{\beta}$ given $\hat{\boldsymbol{\Theta}}$. Using the estimate of $\boldsymbol{D}$, $\boldsymbol{\beta}$ and $\boldsymbol{u}$, are estimated using (3.20).

The MATCH prediction error at some unobserved stations/sites is given by

$$\boldsymbol{Y'} = \boldsymbol{X'}\boldsymbol{\beta} + \boldsymbol{B'}\boldsymbol{u'} + \boldsymbol{\epsilon'}. \quad (3.24)$$

Let

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{u'} \\ \boldsymbol{u} \end{bmatrix} \quad (3.25)$$

and let

$$C = \mathbb{V}(\boldsymbol{U}) = \begin{bmatrix} \boldsymbol{C}_{1,1} & \boldsymbol{C}_{1,2} \\ \boldsymbol{C}_{2,1} & \boldsymbol{C}_{2,2} \end{bmatrix} \quad (3.26)$$

where $\boldsymbol{C}_{1,1} = \mathbb{V}(\boldsymbol{u'})$, $\boldsymbol{C}_{1,2} = \mathbb{C}(\boldsymbol{u'}, \boldsymbol{u})$, $\boldsymbol{C}_{2,1} = \mathbb{C}(\boldsymbol{u'}, \boldsymbol{u})$ and $\boldsymbol{C}_{2,2} = \mathbb{V}(\boldsymbol{u})$. The prediction of $\boldsymbol{Y'}$ is then given by

$$\hat{\boldsymbol{Y'}} = \boldsymbol{X'}\hat{\boldsymbol{\beta}} + \boldsymbol{B'}\hat{\boldsymbol{u'}} \quad (3.27)$$

where

$$\mathbb{E}(\boldsymbol{u'}|\boldsymbol{u}) = \boldsymbol{C}_{2,1}\boldsymbol{C}_{1,1}^{-1}\boldsymbol{u} \quad (3.28)$$

[6].

3.4 Covariate selection

A linear hypothesis test can be used to select covariates in regression models by testing if the total sum of squared error changes significantly when adding or removing a covariate. There are three common approaches to select the covariates, forward selection, backward selecting and stepwise fit.

3.4.1 Decomposition of the error

In order to construct a test to determine whether or not the error of a linear model has increased when one of the covariates is removed the error has to be decomposed. To explain, the linear model is written as:

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where A is the covariate matrix, $\boldsymbol{\beta}$ the coefficient vector $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{I}\sigma)$, the estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}$. The total sum of squares is given by

$$SS = (\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{\beta}),$$

the sum of squares for regression is given by

$$SSR = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \boldsymbol{A}^T \boldsymbol{A}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}),$$

and the sum of squares for error is given by

$$SSE = (\boldsymbol{Y} - \boldsymbol{A}\hat{\boldsymbol{\beta}})^T(\boldsymbol{Y} - \boldsymbol{A}\hat{\boldsymbol{\beta}}).$$

It holds that

$$SS = SSR + SSE$$

which is easiest seen by subtracting SSE from SS and simplifying the result. A linear hypothesis test of the regression coefficients can by formed as

$$B\beta = \mathbf{c}. \tag{3.29}$$

Say, for example, that one wanted to test the hypothesis $\beta_1 = 0$, then $\boldsymbol{B} = [1, ..., 0]$ and $c = 0$. $\boldsymbol{\beta}$ then has $p-1$ degrees of freedom, i.e the rest of the $p-1$ regression coefficients are not constrained. The best guess of the $bold symbol \beta$ under the hypothesis is then

$$\boldsymbol{\beta'} = \arg \min_{\substack{\boldsymbol{\beta} \\ \boldsymbol{B\beta}=\boldsymbol{c}}} SS = \arg \min_{\substack{\boldsymbol{\beta} \\ \boldsymbol{B\beta}=\boldsymbol{c}}} SSR \tag{3.30}$$

and the sum of squares under the constraint becomes

$$SS' = SSH + SSE \tag{3.31}$$

where the sum of squares for the linear hypothesis is $SSH = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta'})^T \boldsymbol{A}^T \boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta'})$ and it can be shown that

$$SS' \sim \chi(n - p + k), \quad SSE \sim \chi(n - p), \quad SSH \sim \chi(k).$$

The likelihood of the outcome of SS' can now be examined. In order to avoid uncertainty coming from the estimated variance σ^2 the test is configured as in Theorem 1.

Theorem 1. *The hypothesis can be tested using the F-statistic*

$$F := \frac{SSH/k}{SSE/(n-p)} \tag{3.32}$$

with large values of F evidence against the hypothesis. Thus at significance level α, we use critical region

$$F > F_\alpha(k, n-p), \tag{3.33}$$

where F_α is the survival function of the Fisher F-distribution. [5]

3.4.2 Stepwise selection

Forward selection

In the forward selection the initial model consists of only a constant term. For each possible covariate the p-value of expanding the model is computed. The covariate with lowest p-value is chosen as a candidate of being in the model. If the p-value of the candidate is higher than the significance level the covariate is rejected and no other covariate is selected since they all have higher p-values. If the p-value is lower than the chosen significance level the covariate is included in the model. The new model consists of the constant term and one covariate. The procedure is then repeated with the new model and the remaining unselected covariates. When no covareiates can be added to the model the selection is complete.

Backward selection

The initial model consists of a constant term and all the covariates. First the p-value of all covariates is calculated using the F-statistics. Then the covariate with highest p-value is chosen as a candidate for removal from the model. If the p-value of the candidate is lower than the significance level no covariate can be removed. If the p-value of the candidate is higher than the significance level the covariate is removed. The procedure is then repeated with the updated model and the remaining not removed covariates. When no more covariate can be removed the selection is done.

Stepwise fit

Once a variable is included in the forward selection procedure it can not be removed similarly in the backward selection procedure, once a variable is rejected it can not be selected again. The set of selected covariates will depend on the method. For example, if a combination of two covariates is better then a third covariate. In the Forward selection procedure only the third might be chosen and in the Backward selection procedure the combination is selected. In general, backward selection will take better account of the joint explanatory power of multiple covariates but lead to larger models.

A procedure to resolve these issues is to combine forward and backward selection. The algorithm starts with the forward selection stage, then follows it with a backward selection stage and alternates between the two until no further variables are introduced at the forward selection stage. The significant levels has to be chosen differently in the two stages which could make it easier for new covariates to enter the model in order to make it possible to prefer a combination of covariates over a single variable.

Chapter 4

Results

In this chapter the result of the implemented composite model is presented. 140 sites are used for estimation and the remaining 9 sites are used for validation. The validation stations, see Figure 4.1, are chosen randomly.

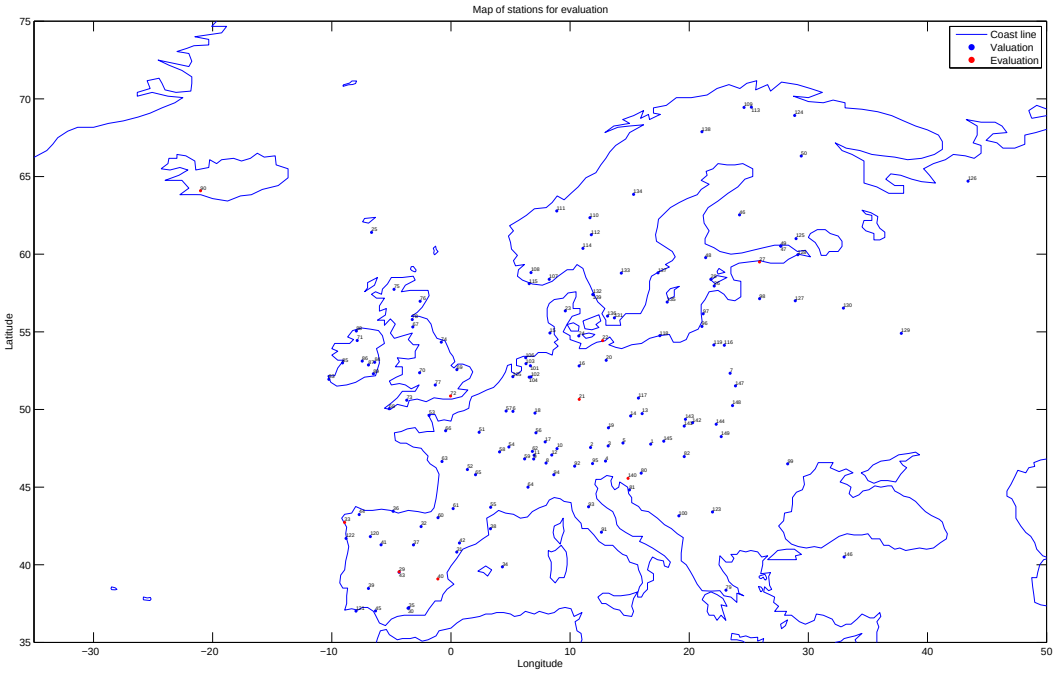

Figure 4.1: The indexed dots are the measurement stations, blue dots are stations used for estimation and red dots are stations for validation.

Seven of them are very close to the coast line and 5 of them are at high altitudes, see table 4.1, which makes it hard to predict the model error at these sites.

Name	Number	Long(deg)	Lat(deg)	Alt(m)	Distance to sea(km)
DE08	21	11	81	937	35
DE09	22	13	80	1	1
EE09	27	26	79	32	0
ES01	29	-4	79	917	31
ES05	33	-9	78	683	4
ES12	40	-1	78	885	7
GB07	72	0	74	8	2
IS02	90	-21	74	66	3
SI08	140	15	73	520	6

Table 4.1: Data regarding the validation stations containing the name of the station, station number, longitude, latitude, altitude and distance to the nearest coast line.

In the first section the station model is constructed. Climate covariates are selected at each measurement station using the stepwise fit method. The number of times each covariate is selected for all the measurement stations are counted. The most frequently selected covariates are selected to be covariates in the station model. The coefficients are then calculated for each station by least square.

In the second section the spatial models are constructed. Spatial covariates are selected for each selected climate covariate in the site model. Here the stepwise fit method can not be used since the residuals are assumed to be dependent. Instead the spatial covariates for each of the climate covariate's coefficient model are selected by a so-called exploratory data analysis (EDA). When the spatial covariates for a climate covariate's model have been selected their coefficients are calculated by least square. The residuals are then used to estimate the parameters of the Matérn function describing the covariance between the climate covariate's coefficients.

In the third section, when the climate covariates in the station model and spatial covariates in the coefficient models are selected and parameters in Mat'ern functions are estimated, the mixed effect model is constructed by estimating the coefficient vector $\boldsymbol{\beta}$ and the spatial noise vector $\boldsymbol{u}$ using (3.20).

In the fourth section the composite model is validated. Estimating the spatial noise at the validation station given the spatial noise at the estimation station the mixed effect model is used to calculate the log-transformed model prediction error at the validation stations.

Finely the model is used to evaluate how the errors in the MATCH predictions depend on climate variables. How much of the log-transformed model error that is explained by a climate covariate is evaluated by comparing the predicted log-transformed model error to the predicted log-transformed model error when the climate covariate is excluding.

4.1 Station model

The variable selection varies a lot between the stations as seen in figure 4.2. Stepwise fit sometimes selects a covariate in favor of a set of covariates which could be the reason.

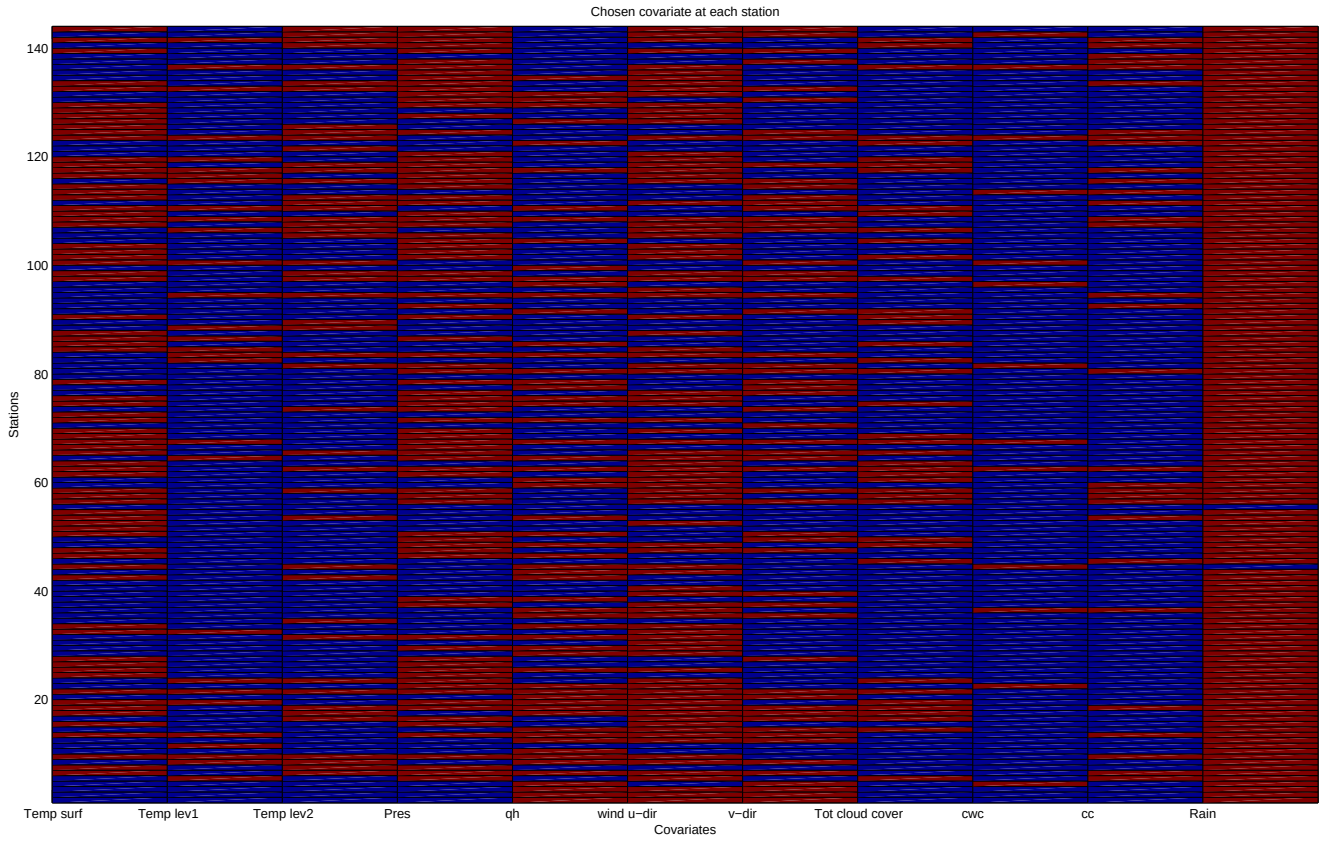

Figure 4.2: The selected climate covariates are red and the unselected are blue. qh stands air humidity, CWC stands for cloud water contents and cc stands for cloud cover.

The selected covariates are temperature at surface, pressure, air humidity, wind in u- and v- component, total cloud cover and rain. As seen in figure 4.3 this chose isn't obvious. Just one of the temperature covariates is chosen although the temperature at level 2 is selected at every third station. This is because the temperature at all levels is highly correlated.

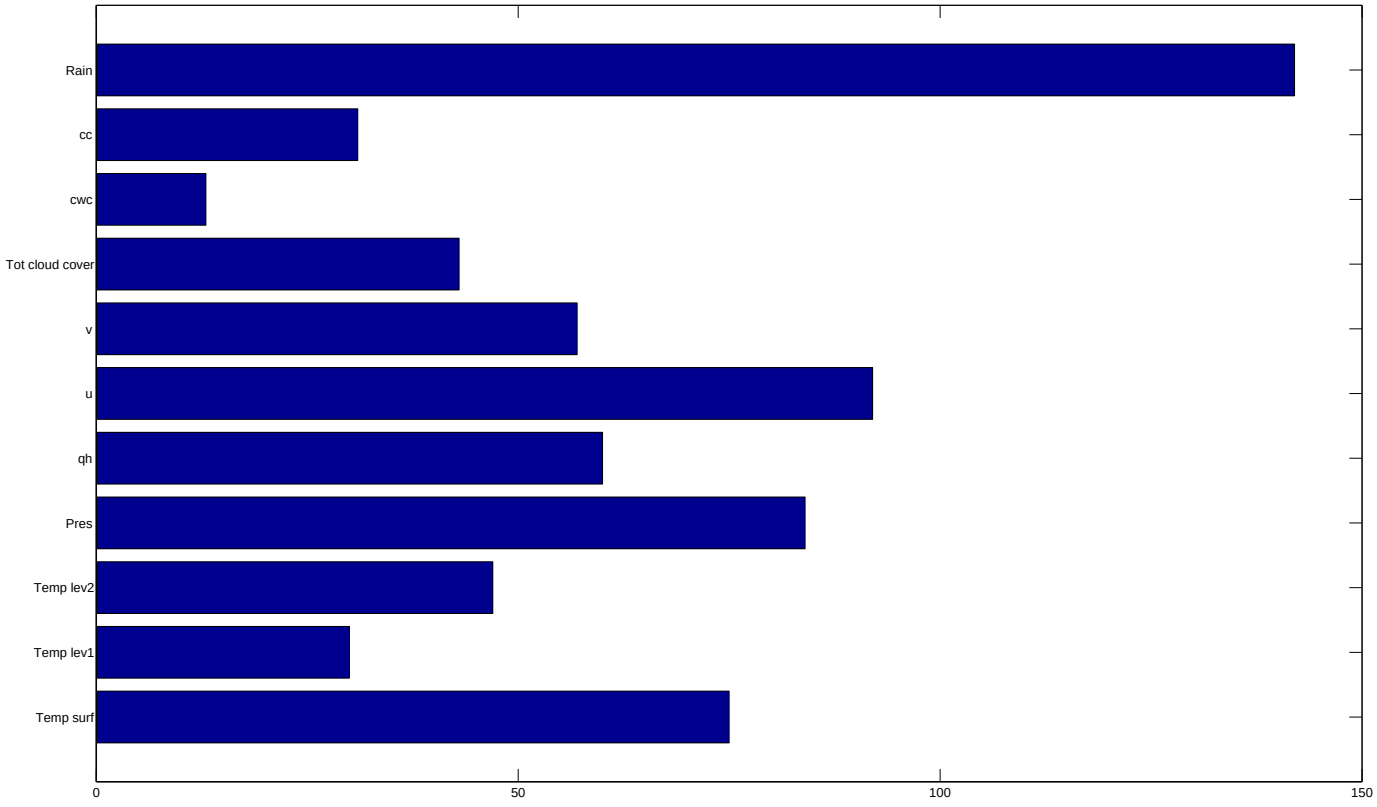

Figure 4.3: On the vertical axis are the covariates where cc is cloud cover, cwc is cloud water contents, qh is air humidity. The horizontal axis is the number of stations where the covariate was selected as an explanatory variable.

To evaluate the generalized model performance the prediction errors are examined. A good model should have residuals that are independent of the regressors and the variance of the residual should be constant. Since it is assumed that the noise is Gaussian and uncorrelated this should also be checked. The residual analyze is done visually by plotting the residual against the regressor, plotting the cross correlation between the residual and regressor, normal plot of the residual, the cross correlation of the residual and the realization.

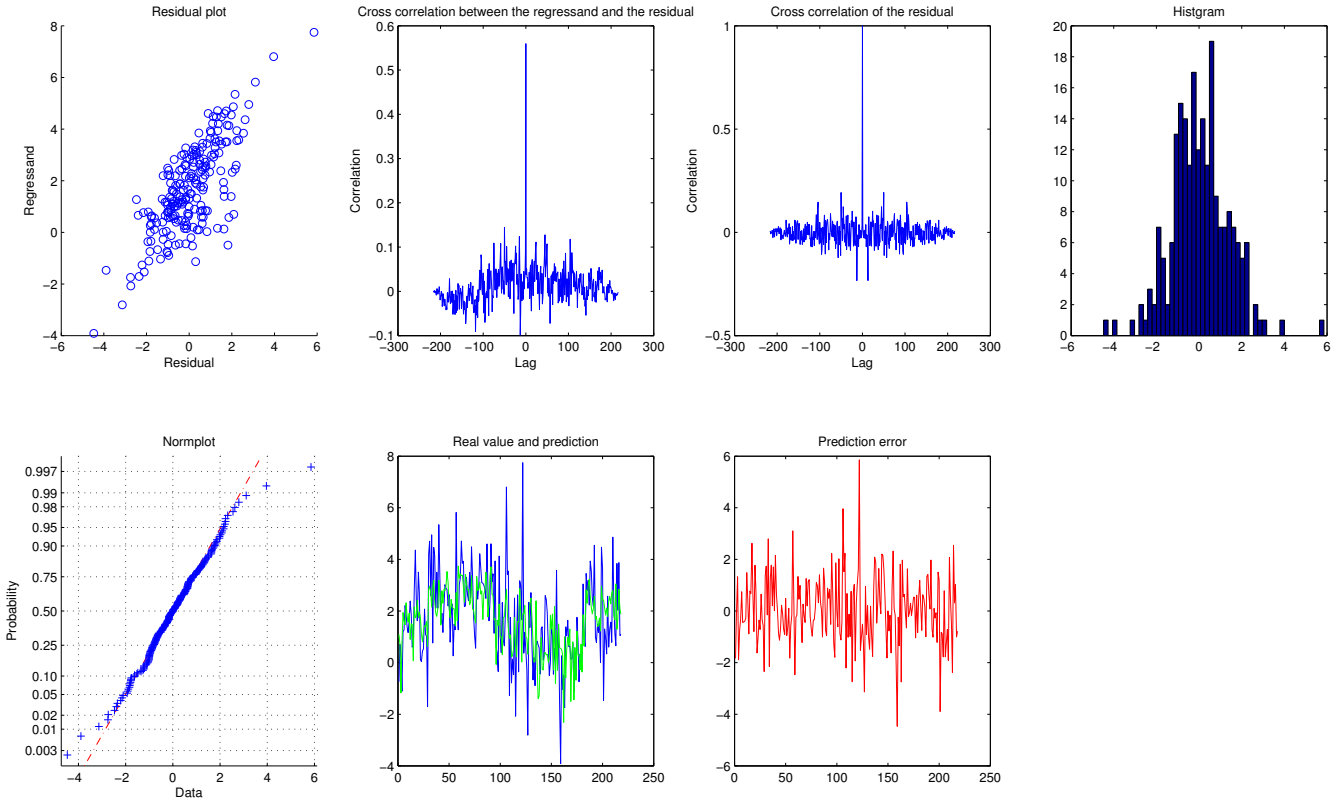

Figure 4.4: From left to right: residual plot, cross correlation with the regressor, cross correlation function, histogram and normplot of the residual. The sixth plot are the realization of the log-transformed model error (blue) and the prediction (green). The last subplot is the error of the predicted log-transformed model error. This is done at station AT48, the estimation of the covariate coefficients is done on 75 % of the data.

As seen in the first and sixth subplots of Figure 4.4 the residual is very dependent on the regressor. One can try a model consisting of a higher order polynomial but it want help much although making the model more complex.

4.2 Spatial model

Exploratory Data Analysis (EDA) is used to select spatial covariates for each coefficient model. In EDA the covariates are selected by plotting the variable of intreast together with the covariates. One wants to plot the variable of

interest together with as many covariates as possible at the same time in order to understand which effect each covariate gives. In this case each climate covariate's coefficients estimated at all stations are plotted on a map consisting of coast lines and height counter lines. One can the ideate if the coefficients depend on any of the spatial covariates.

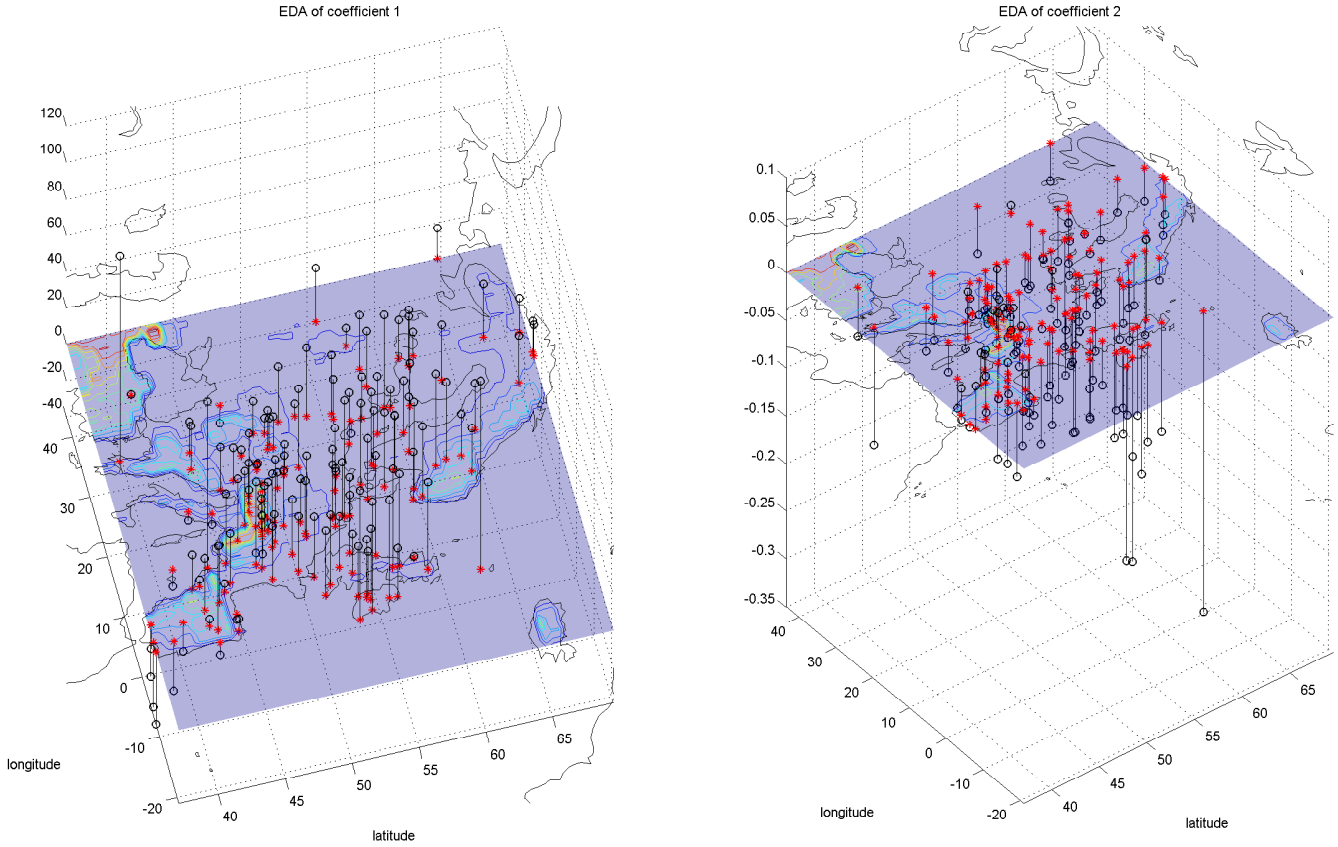

Figure 4.5: Coefficients for the intercept and the temperature at surface covariates

In the left subplot of Figure 4.5 the constant coefficients are plotted. It is hard to see by looking at one angle but it seems like there are dependences with the longitude and latitude. The value seems to be higher at stations more to the north-west than to the south-east apart from stations near the Mediterranean in Portugal and Spain. The chosen covariates are therefore longitude and latitude. In the right subplot of Figure 4.6 the coefficients of the temperature at surface

covariate are plotted. It seems like the values are lower near the coast line of the Atlantic sea but not at the Mediterranean sea. The values seems to be lowest in north-west and highest in south-east. The chosen covariates are therefore longitude and latitude.

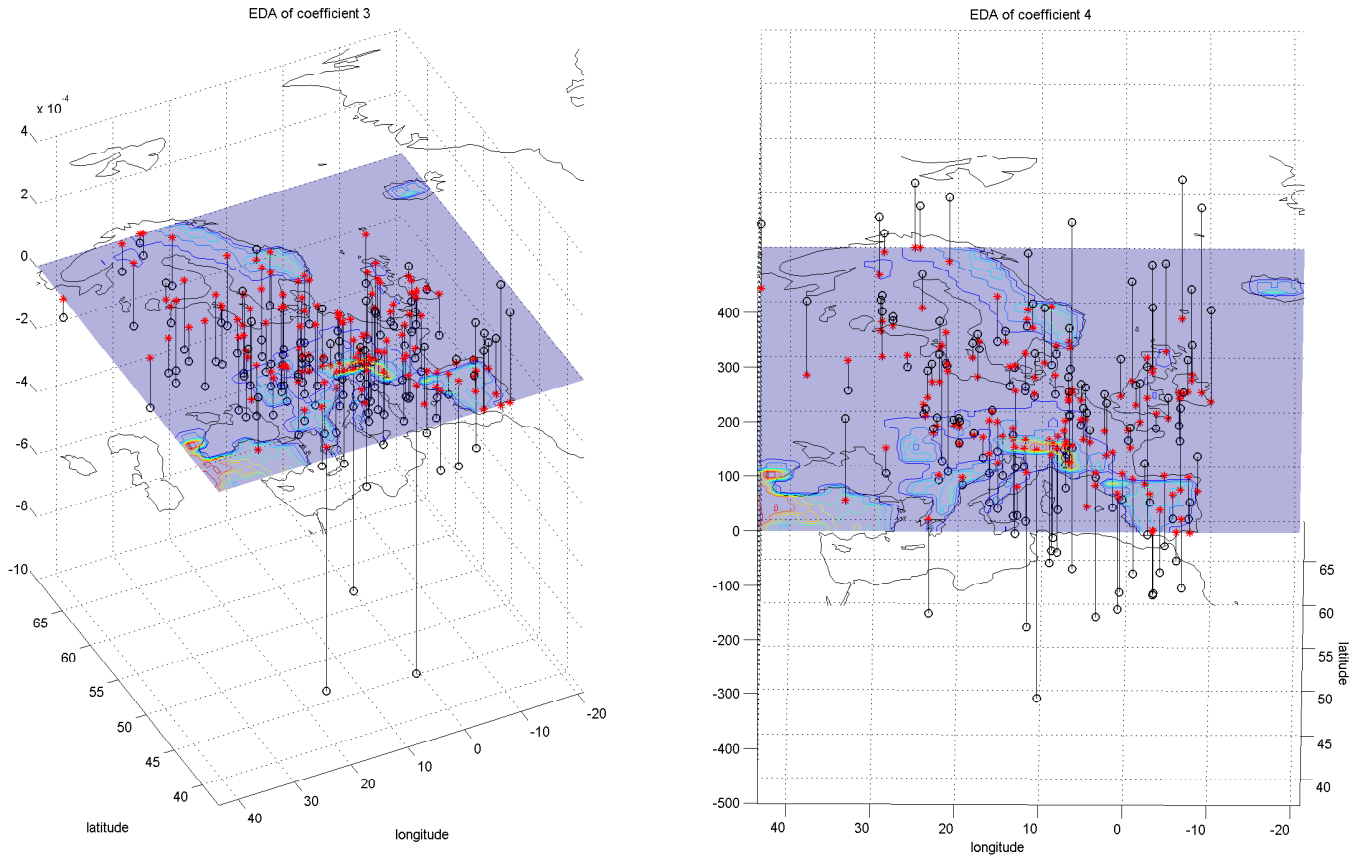

Figure 4.6: Coefficients for the pressure and the humidity both at surface

In the left subplot of Figure 4.6 the pressure coefficients are plotted. There seems to be a latitude dependence and perhaps longitude dependence. The values are likely to be higher for southerly stations. The chosen covariates are longitude and latitude.
In the right subplot of Figure 4.6 the air humidity coefficients are plotted. It appears that the coefficient depends on the longitude and latitude. The values seem to be higher at stations more to the north-west than to the south-east. The values differ dramatically from high to low when passing the north-west side

to south-east side of the mountain chains of Caucasus and the Pyrenees. The south-east side of these mountain chains are very mountainous so it is tempting to consider the altitude as a covariate but at many sites such as the station in Turkey and the stations in Norway the coefficients are high although they are not at higher altitude. The chosen covariates are longitude and latitude.

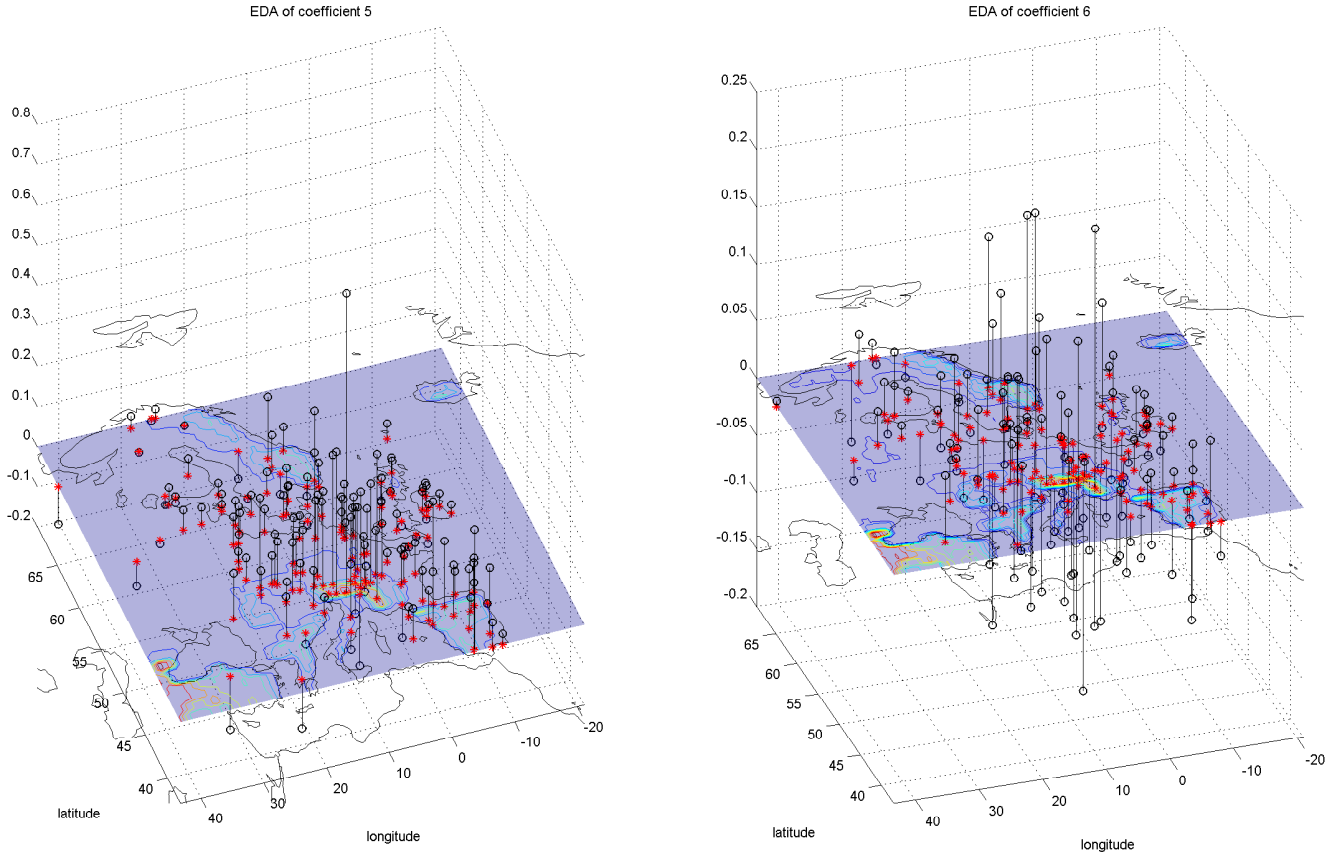

Figure 4.7: Coefficients for the wind in u- and v-direction at surface

In the left subplot of Figure 4.7 the wind in u-component coefficients are plotted. There is only dependency in altitude. The chosen covariate is therefore altitude.
In the right subplot of figure 4.7 the wind in v-direction coefficients are plotted. It is hard to sea any dependence at all.

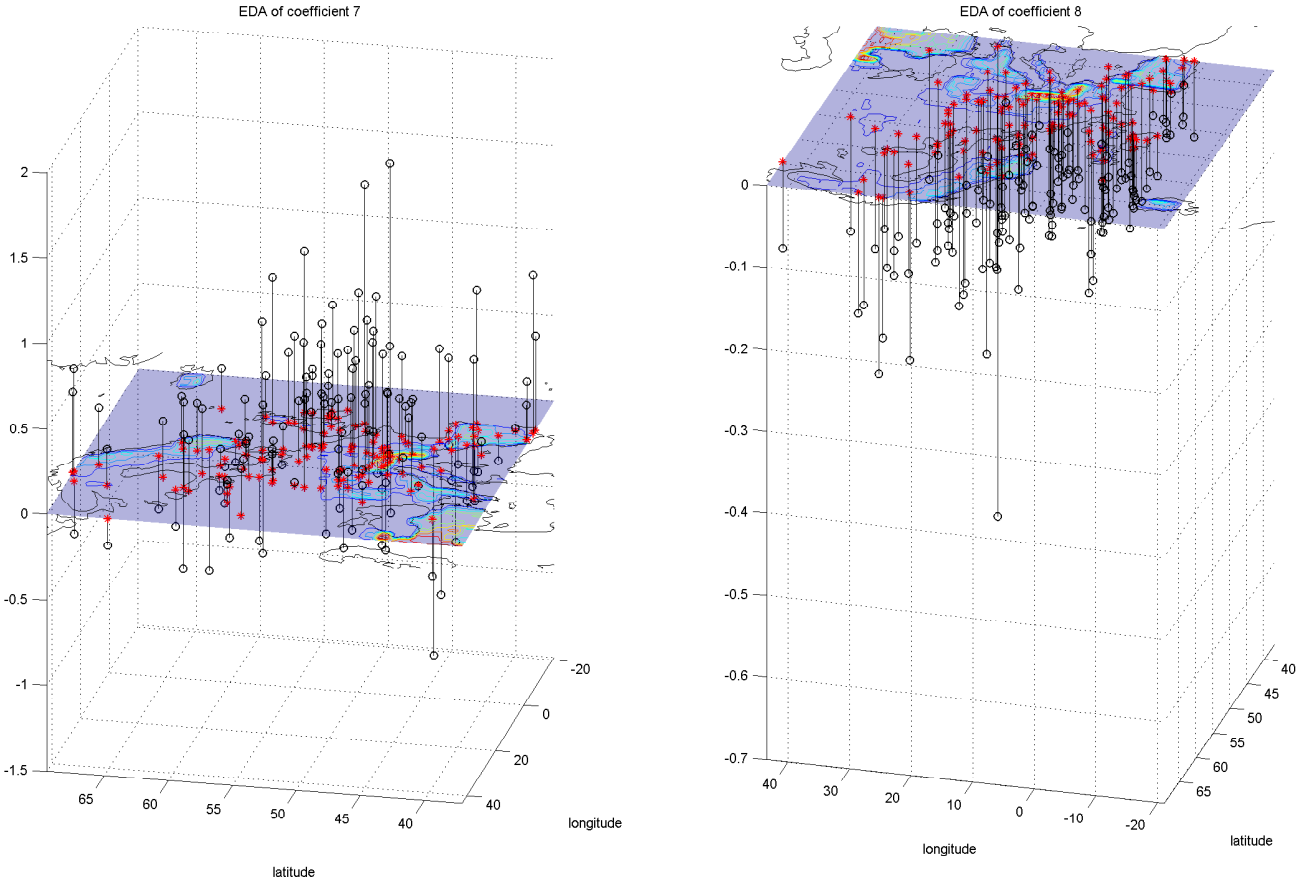

Figure 4.8: Coefficients for the total cloud cover and the rain

In left subplot of figure 4.8 the total cloud cover coefficients are plotted. There are no obvious dependencies other than a very vague trend of higher values to the south.
In right subplot of figure 4.8 the rain coefficients are plotted. Possible covariates are longitude and latitude. The values also look higher near the coast so perhaps distance to sea could be considerate as a covariate.

As said earlier, stepwise fit cannot be used but since the estimated correlations between the coefficients are very small (which will be seen later on in this chapter) the stepwise fit method is used to provide an indicator or hint of which covariates that could be chosen. The maximum p-value for a term to enter the model is set 0.05 and the maximum p-value for a term to be removed is set to 0.10.

The stepwise fit method suggests longitude and latitude for the interception coefficient and the temperature coefficients which also was the conclusion in the EDA. For the pressure at surface stepwise fit suggests only latitude, not longitude as selected in the EDA. For the humidity stepwise fit suggests latitude and altitude. The selected covariates from the EDA were longitude and latitude. Only altitude were selected as a covariate in the EDA of wind in u-component. Stepwise fit suggest both latitude and altitude. No covariates were chosen in the EDA of the wind in v-component coefficient but stepwise fit suggests altitude. In the EDA of the seventh coefficient, total cloud cover, no covariates were suggested. No covariate was suggested by stepwise fit. The rain coefficient was beveled to depend on longitude, latitude and distance to sea in the EDA. Stepwise fit only suggests longitude.

Taking the result from the stepwise fit method into account the conclusion, which is seen in table 4.2, is a bit different.

Coefficient	Lat	Long	Alt	Dist. sea
Intercept	•	•		
Temperature at sureface	•	•		
Pressure	•	•		
Humidity	•	•	•	
Wind in u-direction	•		•	
Wind in v-direction		•		•
TCC				•
Rain	•	•		•

Table 4.2: Table of chosen spatial covariates where the dots indicates that the chosen covariates.

4.3 Matérn covariance

The coefficient vectors, $\boldsymbol{\beta}_1 \dots \boldsymbol{\beta}_p$ for each selected climate covariates's coefficient model is calculated by least square and the residuals are used as estimates of $\boldsymbol{u}_1 \dots \boldsymbol{u}_p$.

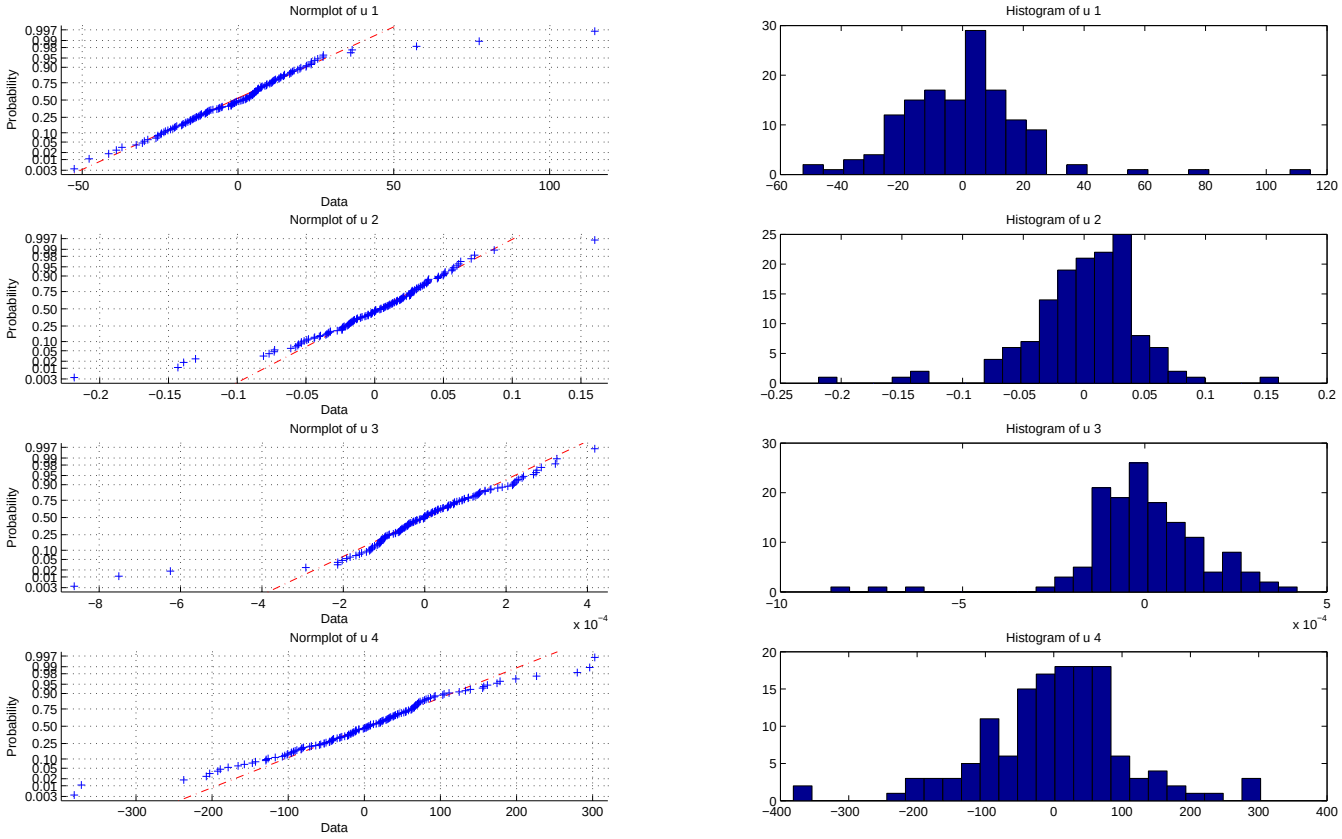

Figure 4.9: Residual 1-4 of the selecte covariate's coefficient models. These residuals regards the intercept, temperature at surface, pressure and air humidity covariate's coefficients models from top to bottom.

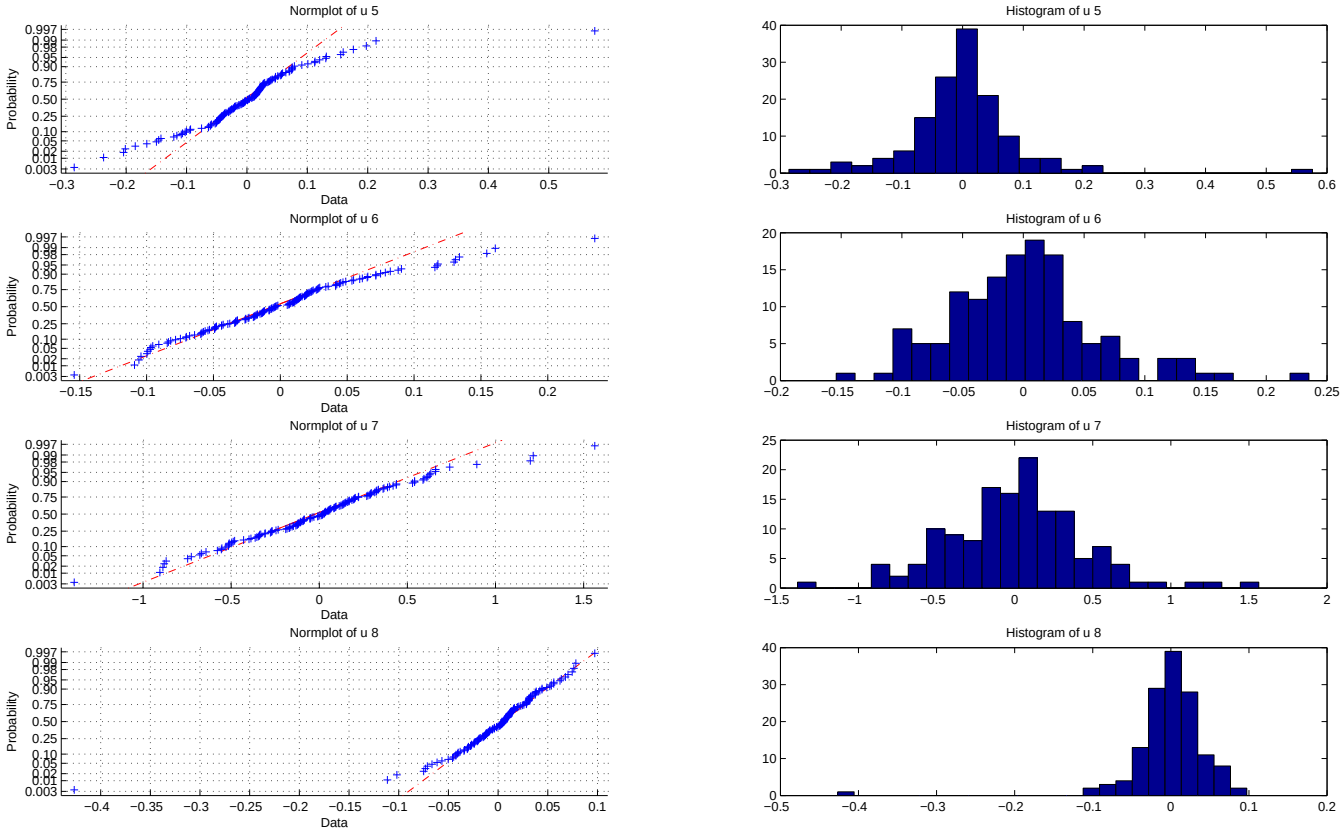

Figure 4.10: Residual 5-8 of the selecte covariate's coefficient models. These residuals regards the wind i u-component, wind in v-component, total cloud cover and rain covariate's coefficients models from top to bottom.

The estimate of each selected climate covariate's coefficient noise is normally distributed with zero mean as seen in the subplots of Figure 4.9 and Figure 4.9 (they are of course zero mean since least square is used).

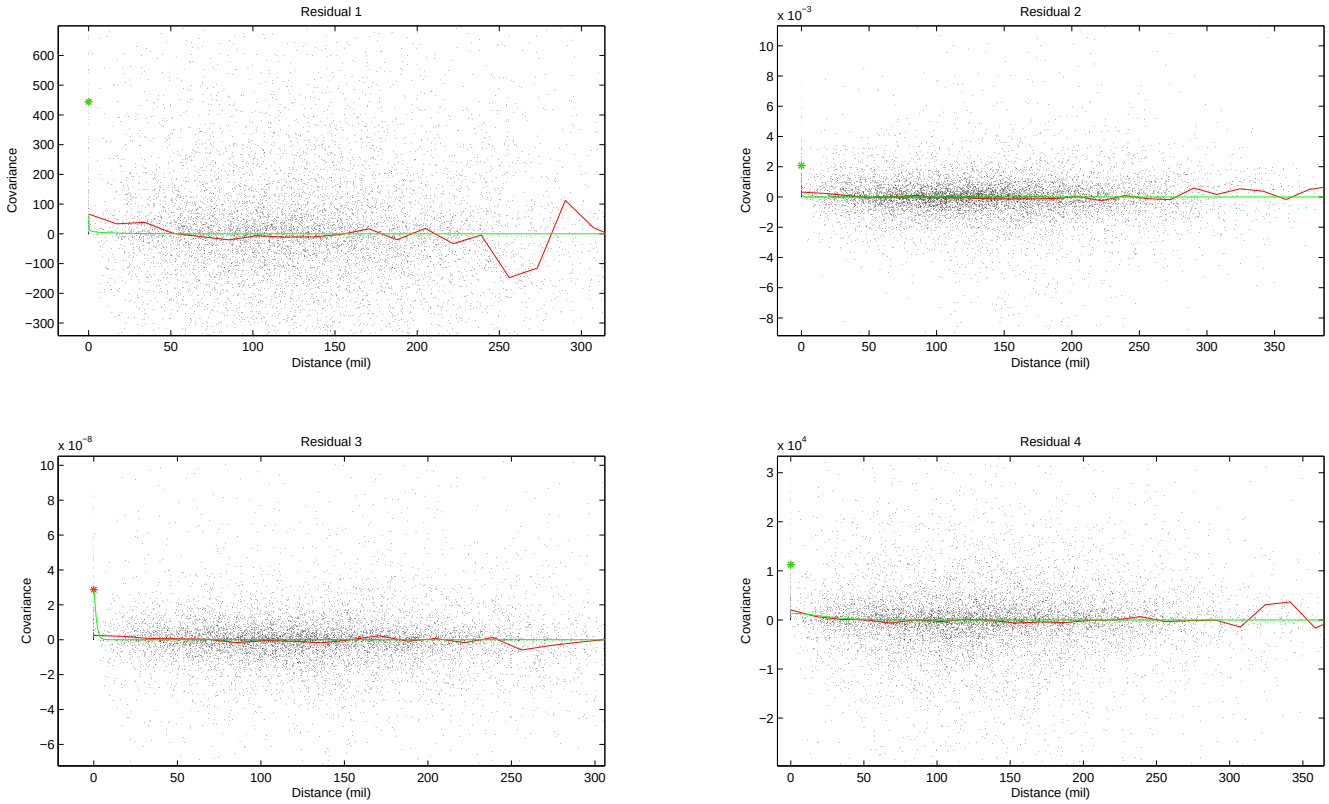

Figure 4.11: The green cuves are the estimated covariance function and the red curves the covariances calculated for a number of distances. These regards $\boldsymbol{u}_1$, $\boldsymbol{u}_2$, $\boldsymbol{u}_3$ and $\boldsymbol{u}_4$ from left to right.

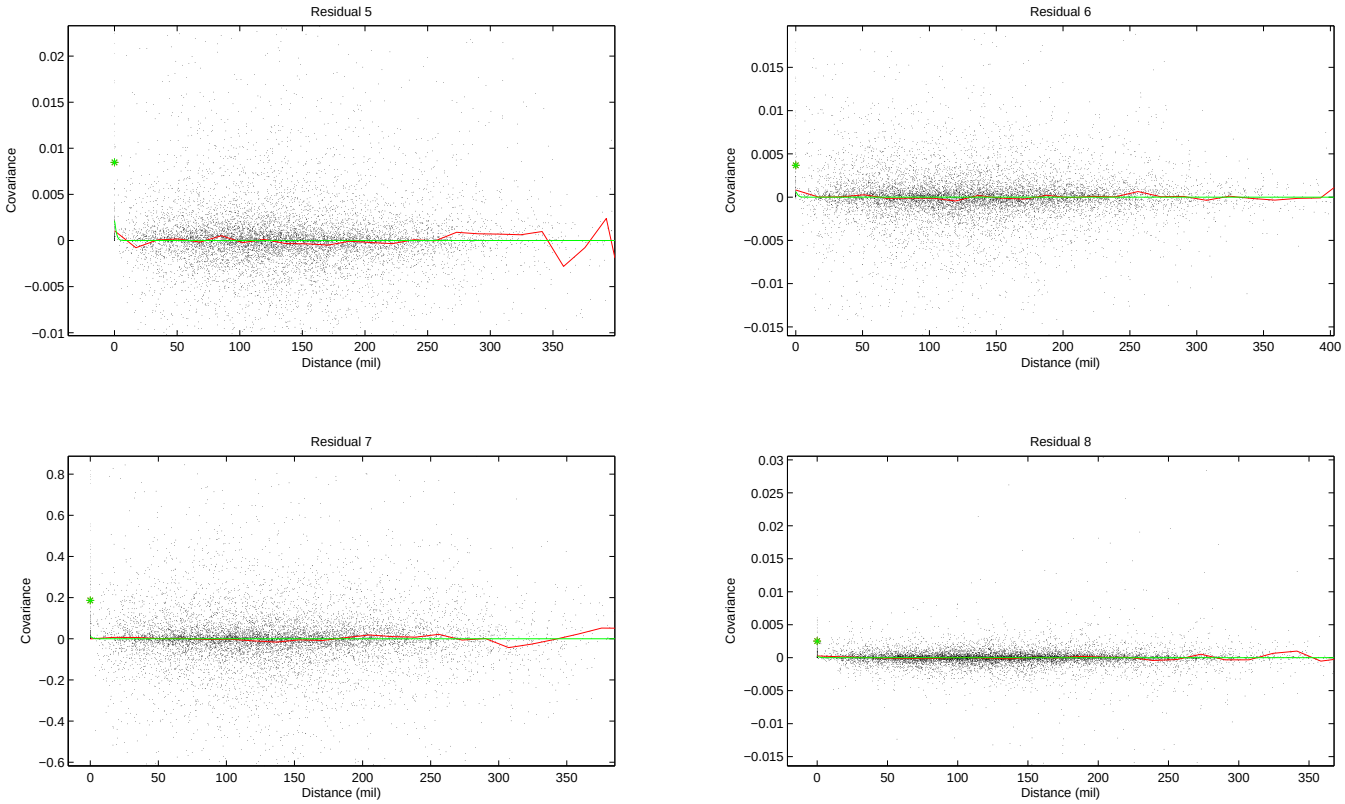

Figure 4.12: The green cuves are the estimated covariance function and the red curves the covariances calculated for a number of distances. These regards $\boldsymbol{u}_5$, $\boldsymbol{u}_6$, $\boldsymbol{u}_7$ and $\boldsymbol{u}_8$ from left to right.

In Figure 4.11 and Figure 4.14 the green curves are the estimated covariance function and the red curves the covariances calculated at 40 distance intervals of the selecte covariate's coefficient model residuals. The cloud consists of the residuals multiplied with each other other residual placed at the distance between them on the x-axis. The star is the variance. The green curve shows nearly any correlation between the coefficients at all while the red curves suggests more correlation. This comparison is made since the estimates of the covariance functions are rather doubtful with just 140 observations. It has been shown, by testing, that the best result are obtained by setting the covariance between the third covariate's coefficient residuals to zero as the red curve suggests.

4.4 Prediction of the coefficients

In this section the regression field for each selected covariate's coefficient is presented, first without using that the coefficients are correlated. The reason why this is done is to show how much the correlation effects the final result.

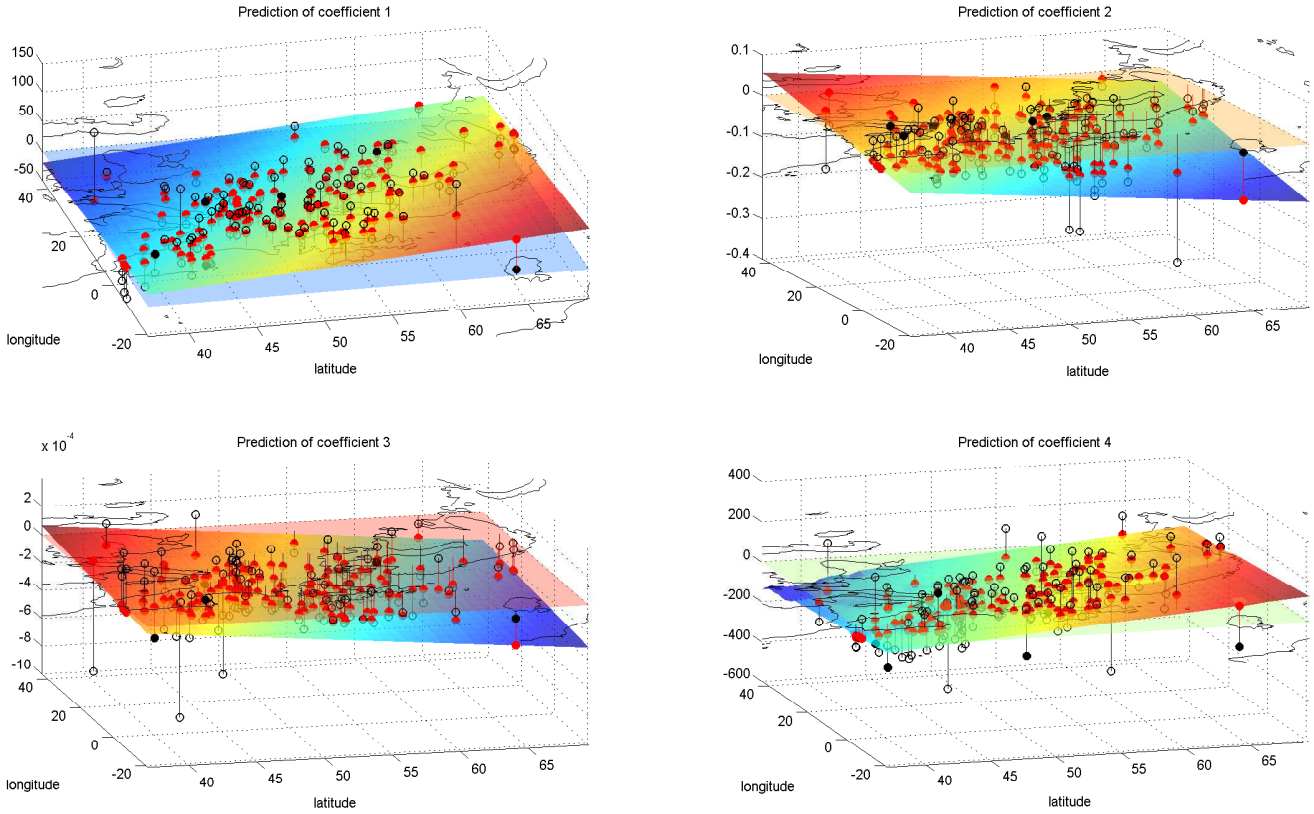

Figure 4.13: Coefficient predictions without using the correlations 1-4.

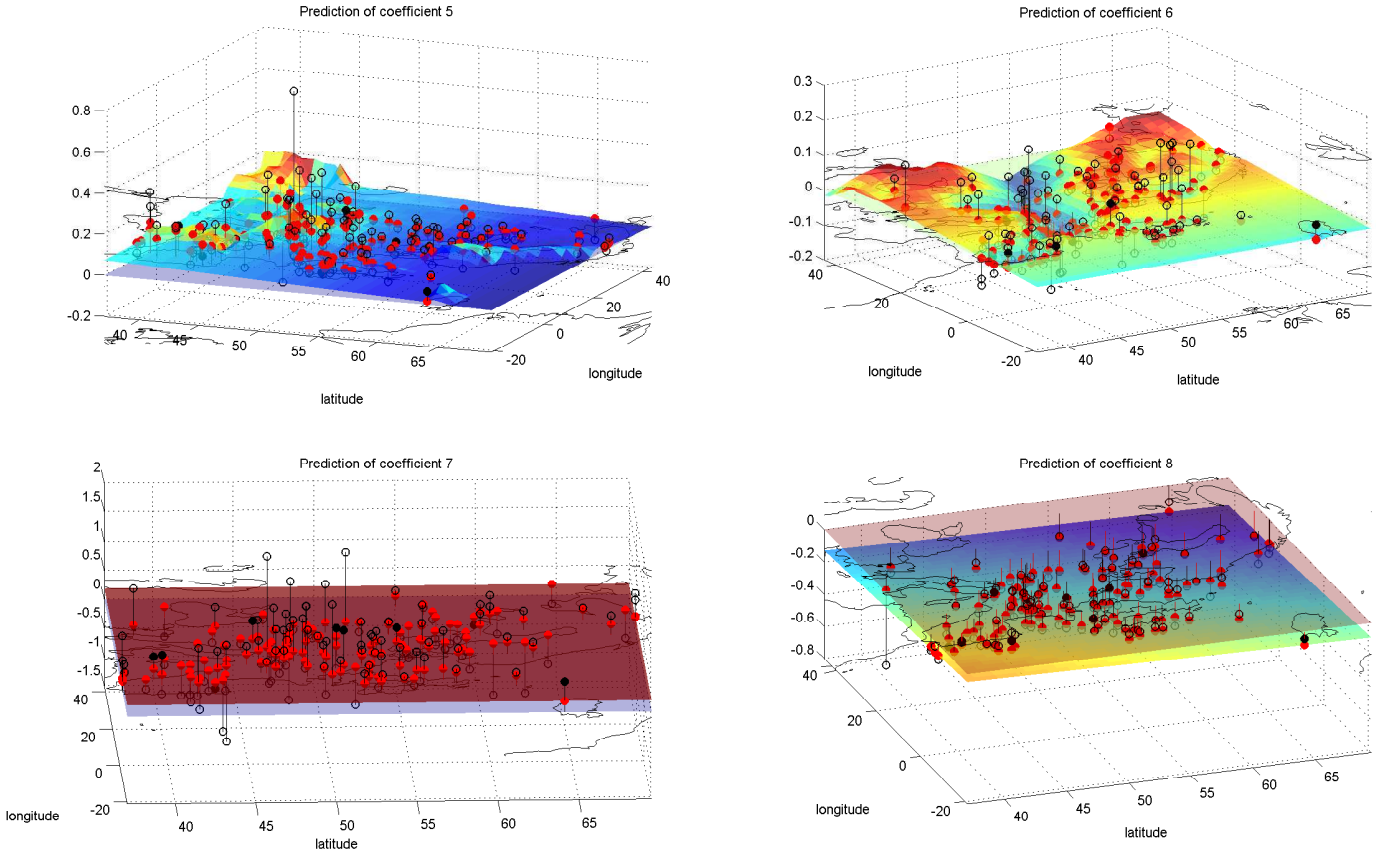

Figure 4.14: Coefficient predictions without using the correlations 5-8.

Figure **??** and Figure **??** are the predictions of the coefficients at all stations and the prediction coefficients at all points of the MATCH grid. The non-filled dots are the coefficients estimated at each station of the estimation set, the fillet black dots are the estimated coefficients at each station of the validation set, the red filled dots are predictions of the coefficient model and the surface are the coefficients predicted by the coefficient model for all points in the MATCH grid.

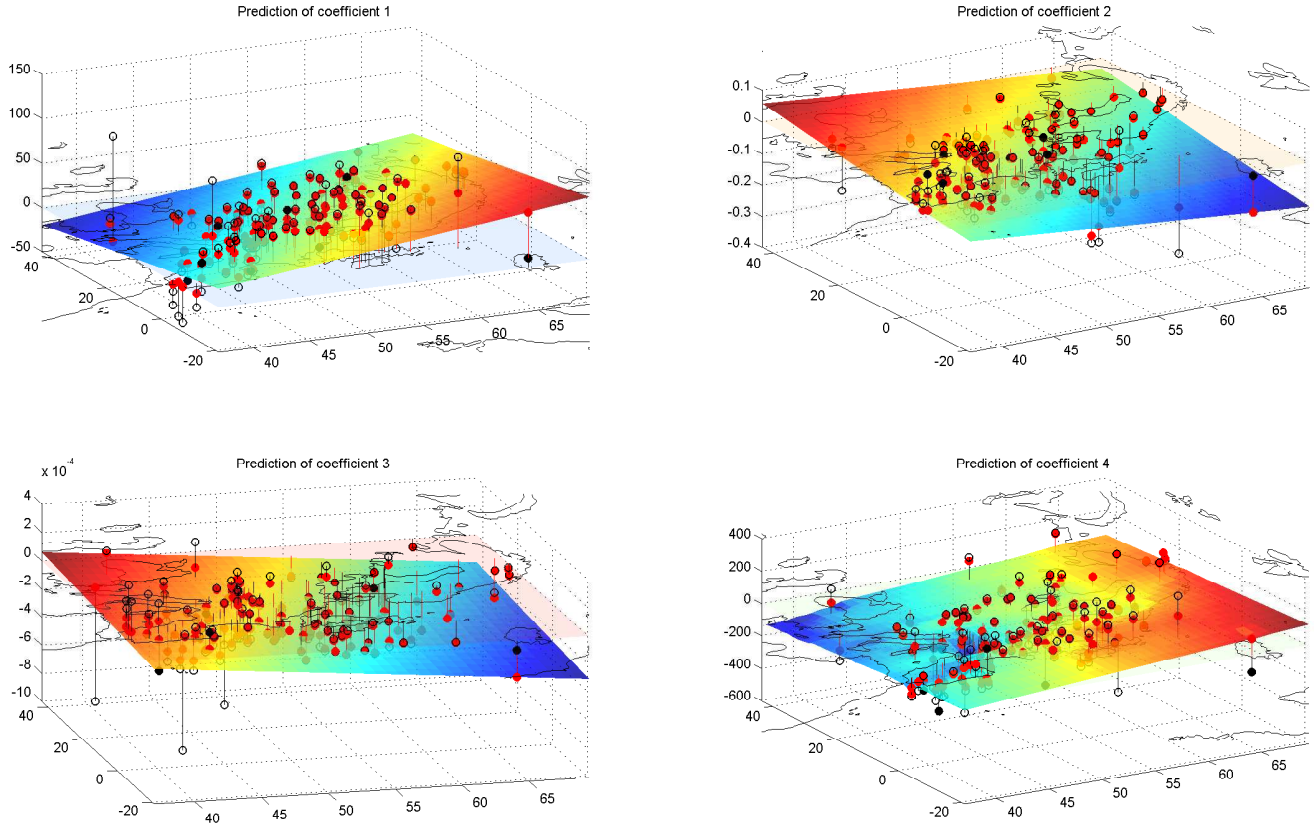

Figure 4.15: Coefficient predictions using the correlations 1-4.

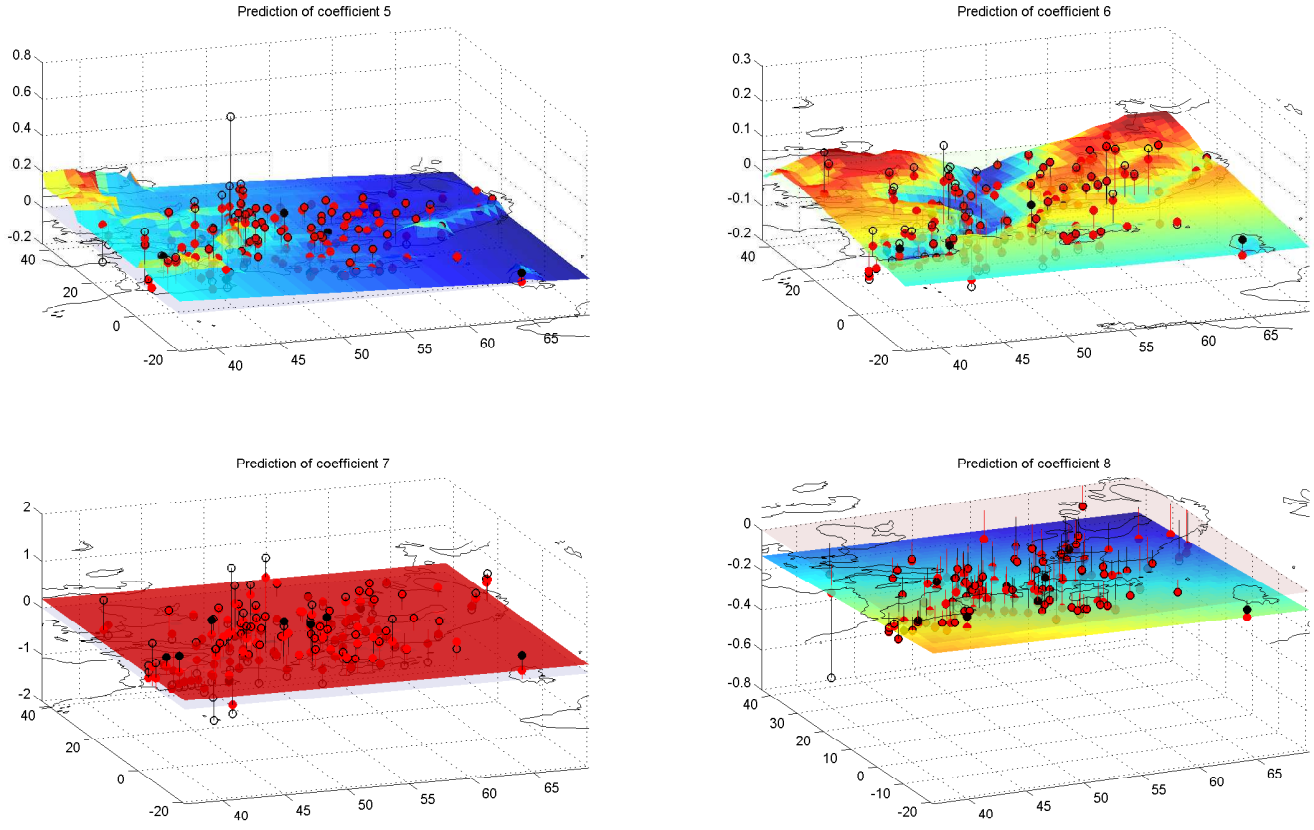

Figure 4.16: Coefficient predictions using the correlations 5-8.

Figure 4.15 and **??** are the predictions using the correlation functions to predict the the noise at the validation stations and at the points of the MATCH grid given the estimated noise at the estimated stations. This effect can only be seen in subplot 1 and 4 of **??**.

4.5 Prediction of the MATCH model error

In this section the model is validated by comparing the predictions with the MATCH predictions error at the validation stations. A simpler version of the MATCH prediction error model where the correlation between the coefficients are assumed to be zero also compared with the MATCH prediction error model and the MATCH prediction error to evaluate if we actually gain anything by assuming that the coefficients are correlated. We will refer to these models as model 1 and model 2 where model the last is the simpler model. At five of the nine randomly chosen validation stations the more advanced model is better. Only the result from station 33 and 140 are shown here, the others can be be seen in Appendix A.

4.5.1 Station 33

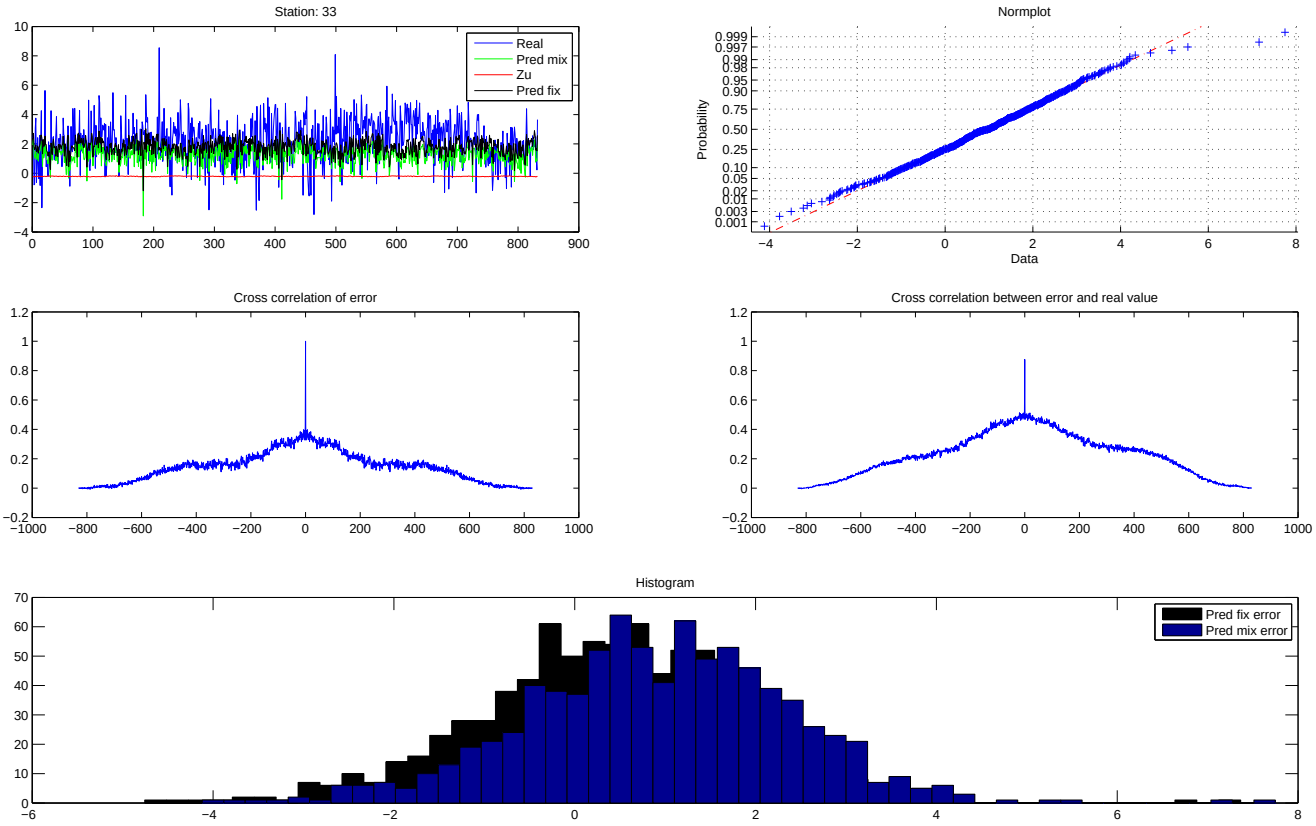

Figure 4.17: Subplot 1 consists of the real log-transformed MATCH error (blue), the prediction of model 1 (green), the prediction of model 2 (black) and the mixed effect (red). subplot 2-4 is a normal, cross correlation function, cross correlation with the real error of model 1 (from left to right). The last subplot consists of two histogram of the residuals of model 1 (blue) and model 2 (black).

As seen in subplot 1 and 5 of figure 4.17, regarding station 33, the predictions are very biased. The mixed effect makes the result worse, it pushes the estimates down. Station 33 is close to station 44 and 122 which makes it possible to calculate the mixed effect. The error of model 2 has 0.6% less mean quadratic error than model 1.

4.5.2 Station 140

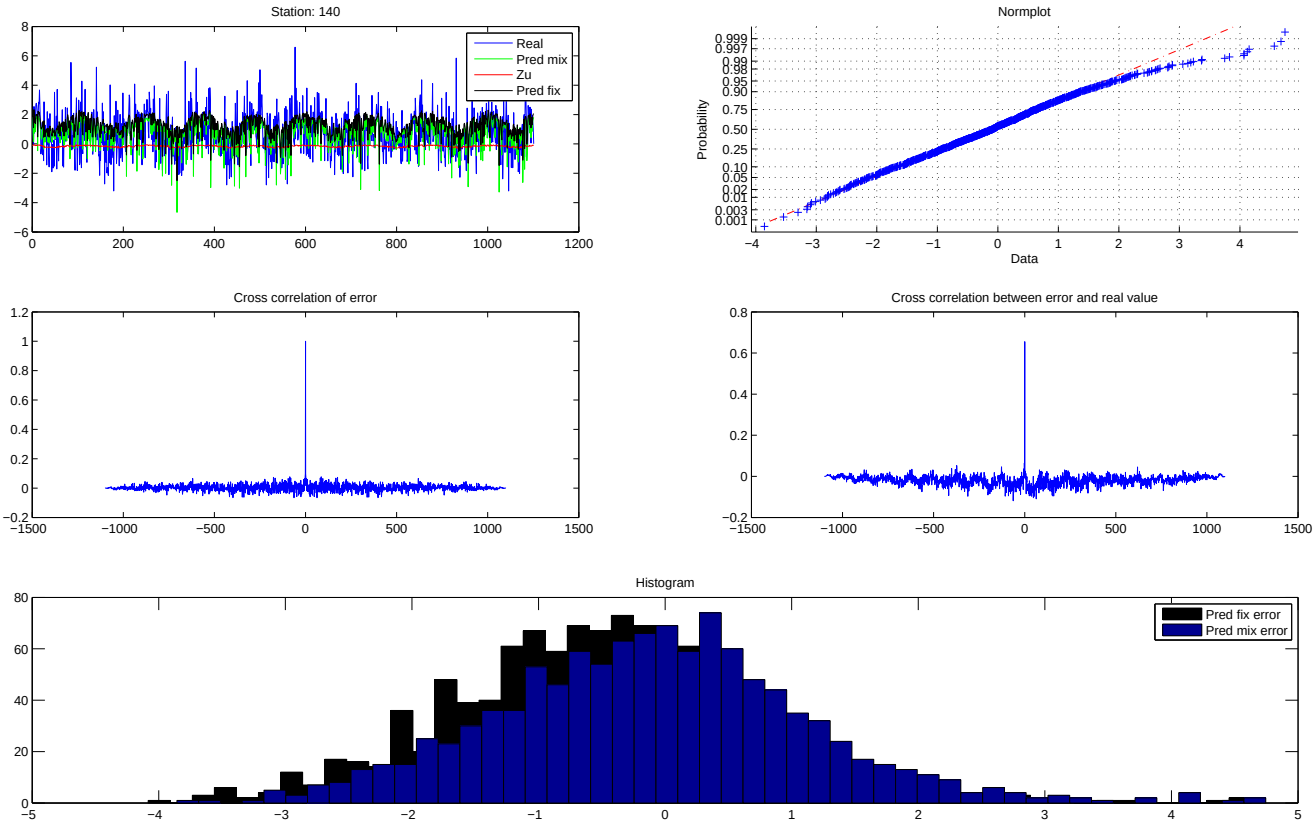

Figure 4.18: Subplot 1 consists of the real log-transformed MATCH error (blue), the prediction of model 1 (green), the prediction of model 2 (black) and the mixed effect (red). subplot 2-4 is a normal, cross correlation function, cross correlation with the real error of model 1 (from left to right). The last subplot consists of two histogram of the residuals of model 1 (blue) and model 2 (black).

Both the estimates are almost unbiased. The mean of the mixed effect is zero and varies with time. The mean quadratic error of model 2 is 6.1 % less than then mean quadratic error of model 1.

4.6 Error analysis

The error analysis is done by comparing the predicted MATCH prediction error with the predicted MATCH prediction error when one of the climate covariate is absent, i.e the model constructed without one of the covariates. The difference is then how much of the predicted MATCH prediction error that can be explained by the covariate.

Each covariate is removed in turn and the error predicted by the reduced model is compared to the complete model.

In this analysis the mean quadratic error is used. Let MQE_j be the predicted mean quadratic error at station j and $MQE_{i,j}$ be the predicted mean quadratic error at station j when climate covariate i is absent. The relative change of predicted mean quadratic error, at station j when climate covariate i is absent, is

$$RMQE_{i,j} = \frac{|MQE_{i,j} - MQE_j|}{MQE_j}. \tag{4.1}$$

Station nr	Temp(%)	Pressure(%)	Humidity(%)	Wind u(%)	Wind v(%)	TCC(%)	Rain(%)
21	1.85	5.67	7.71	16.15	2.10	1.17	11.55
22	1.25	0.22	0.20	4.14	2.38	1.20	10.89
27	4.87	4.87	6.86	0.15	2.56	0.63	21.15
29	7.34	9.57	7.64	9.81	6.89	7.22	11.51
33	5.01	7.65	3.87	7.89	0.72	1.54	1.97
40	1.00	6.23	11.76	6.56	0.45	0.33	24.72
72	1.10	0.98	0.82	0.80	0.91	0.59	8.96
90	1.82	3.24	0.45	6.11	0.94	0.60	8.36
140	1.40	0.93	4.68	0.61	0.89	0.79	35.44

Table 4.3: Relative change in mean quadratic error for each selected climate covariate at each station.

Rain gives the highest relative change in mean quadratic error with a mean of 14.95% for all validation stations, wind in u-component the second (5.80%), humidity the third (4.89%), pressure the fourth (4.37%), temperature the fifth (2.85%), wind in v-component the sixth (1.98%) and total cloud cover the seventh (1.56%).

Chapter 5

Conclusion

5.1 Discussion

According to the error analysis precipitation effects the MATCH prediction error most, then the wind in u-component, humidity, pressure, temperature, wind in v-component and last total cloud cover.

A mixed effect model for the log-transformed MATCH prediction error has been successfully implemented. Assumption of spatial dependence between the coefficients in the site model did not give any significant improvement of the model, maybe due to parameters of the covariance functions were not estimated well enough. Model 1 is not always stable, it can give very wrong predictions. Model 2 is therefore better.

5.2 Further work

The point was to do the error analysis at all the cell of the climate data grid but due lack of time the error analysis was just done at the validation stations. To be able to say how the error is effected by the climate covariates, and how these effects depend on the position in Europe this must be done.

The best way of estimating all parameters in the mixed effect model would be to maximize the assumed density function (3.18), given the data, with respect to $\boldsymbol{\beta}$, $\boldsymbol{u}$ and the parameters for the covariance functions. But this is a thesis itself.

In order to improve the log-transformed MATCH prediction error model more climate and spatial covariates must be found.

The model could be used as a post processing model to MATCH. The log-transformed MATCH prediction error most then be transformed back.

List of Figures

List of Tables

Appendix A

Test stations

A.0.1 Station 21

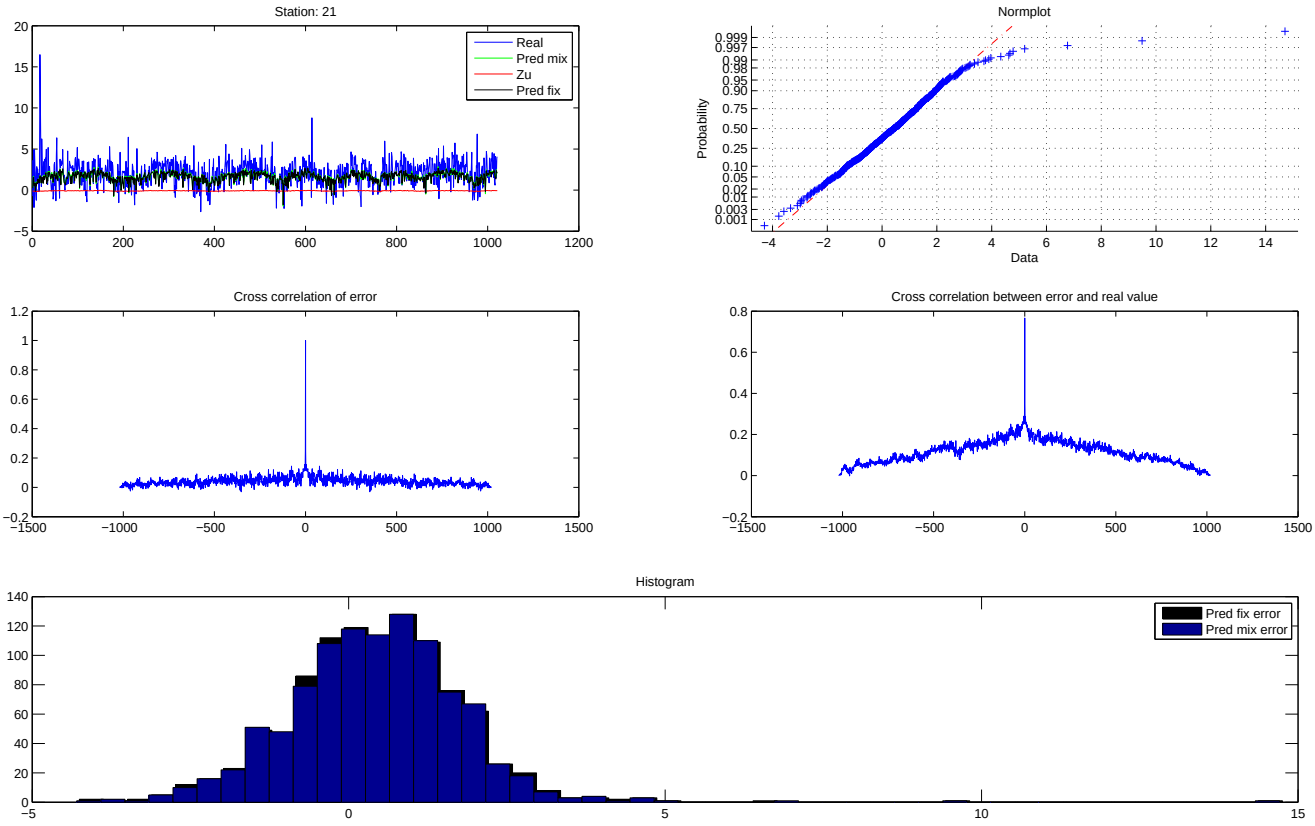

Figure A.1: Subplot 1 consists of the real log-transformed MATCH error (blue), the prediction of model 1 (green), the prediction of model 2 (black) and the mixed effect (red). subplot 2-4 is a normal, cross correlation function, cross correlation with the real error of model 1 (from left to right). The last subplot consists of two histogram of the residuals of model 1 (blue) and model 2 (black).

As seen in the first subplot of A.1 the mixed effect (the effect of the estimated noise in the coefficient model) is not even visible. As seen in figure **??** station 21 is placed in the middle of Europe with no stations near by which explains the absent mixed effect. The mixed effect model gives 2 % less mean quadratic error than the fixed effect model. None of the errors are unbiased. The residuals are uncorrelated and the cross-covariance with the MATCH error is very large (as expected).

A.0.2 Station 22

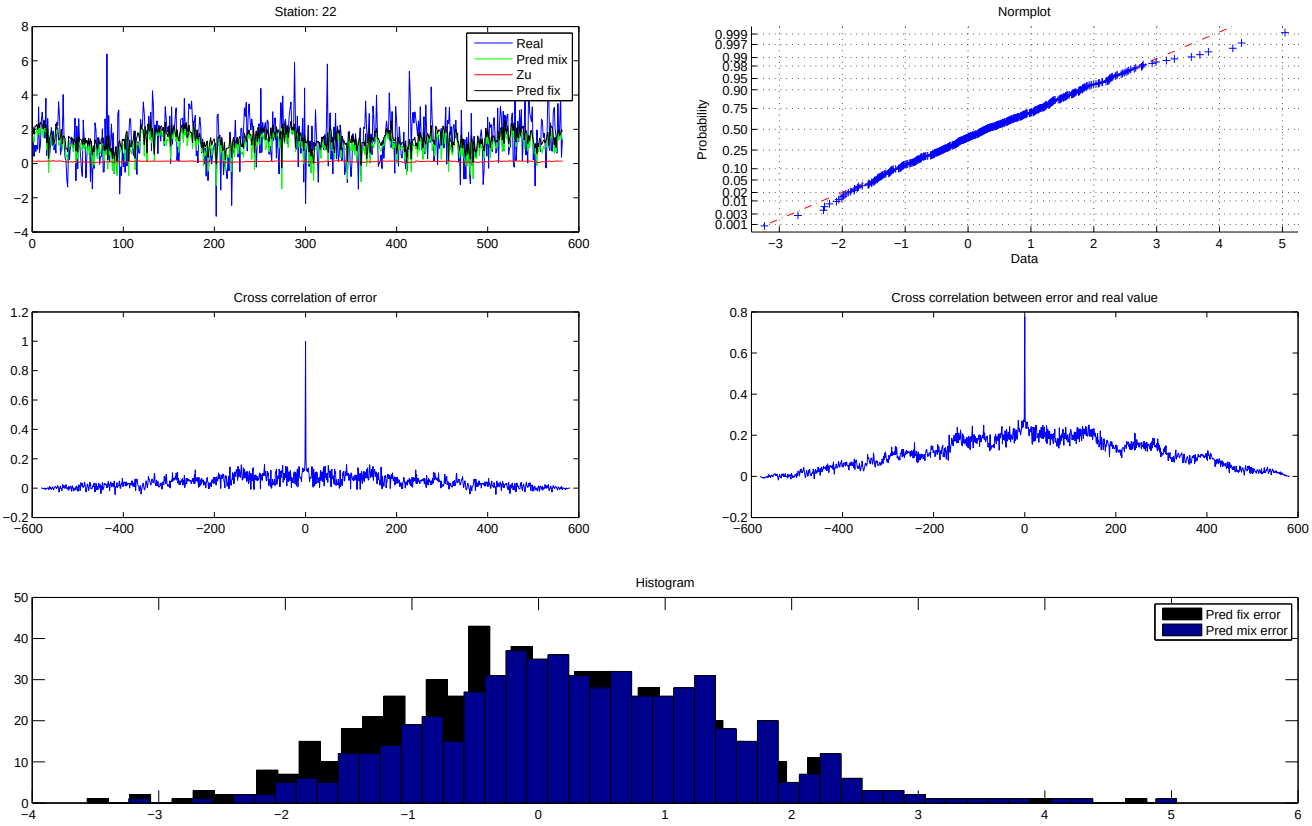

Figure A.2: Subplot 1 consists of the real log-transformed MATCH error (blue), the prediction of model 1 (green), the prediction of model 2 (black) and the mixed effect (red). subplot 2-4 is a normal, cross correlation function, cross correlation with the real error of model 1 (from left to right). The last subplot consists of two histogram of the residuals of model 1 (blue) and model 2 (black).

Station 22 is placed near station 20, 24, 131 and 136. As seen in the first subplot of Figure A.2 the mixed effect makes a difference. Model 1 gives approximately 3.6 % less mean quadratic error than model 2. The residuals are almost unbiased as seen in the last subplot.

A.0.3 Station 27

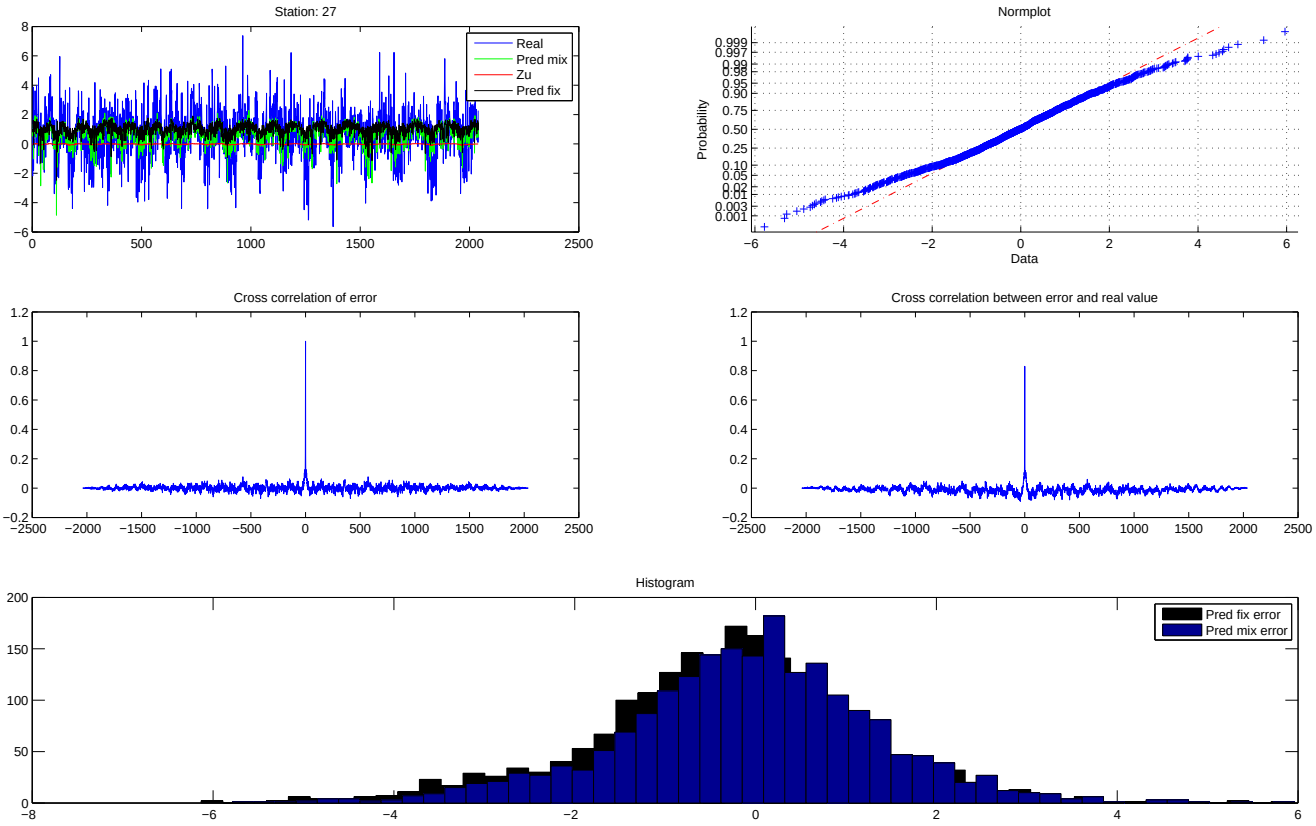

Figure A.3: Sub plot 1 consists of the real error (blue), the prediction made with out the mixed effect (black), the prediction med with the mixed effect and the mixed effect $B\mathbf{u}$ (red). subplot 2-5 is a normal, cross correlation function, cross correlation with the real error and a histogram of the mixed (blue) and fixed (black) effect prediction errors

Station 27 is placed near Estonia not far from the coast with no stations near by. The mixed effect model gives more than 7.6 % less mean quadratic error than the fixed effect model but not because of the mixed effect which is barely apparent as seen in Figure A.2. The prediction of the mixed effect model is almost unbiased.

A.0.4 Station 29

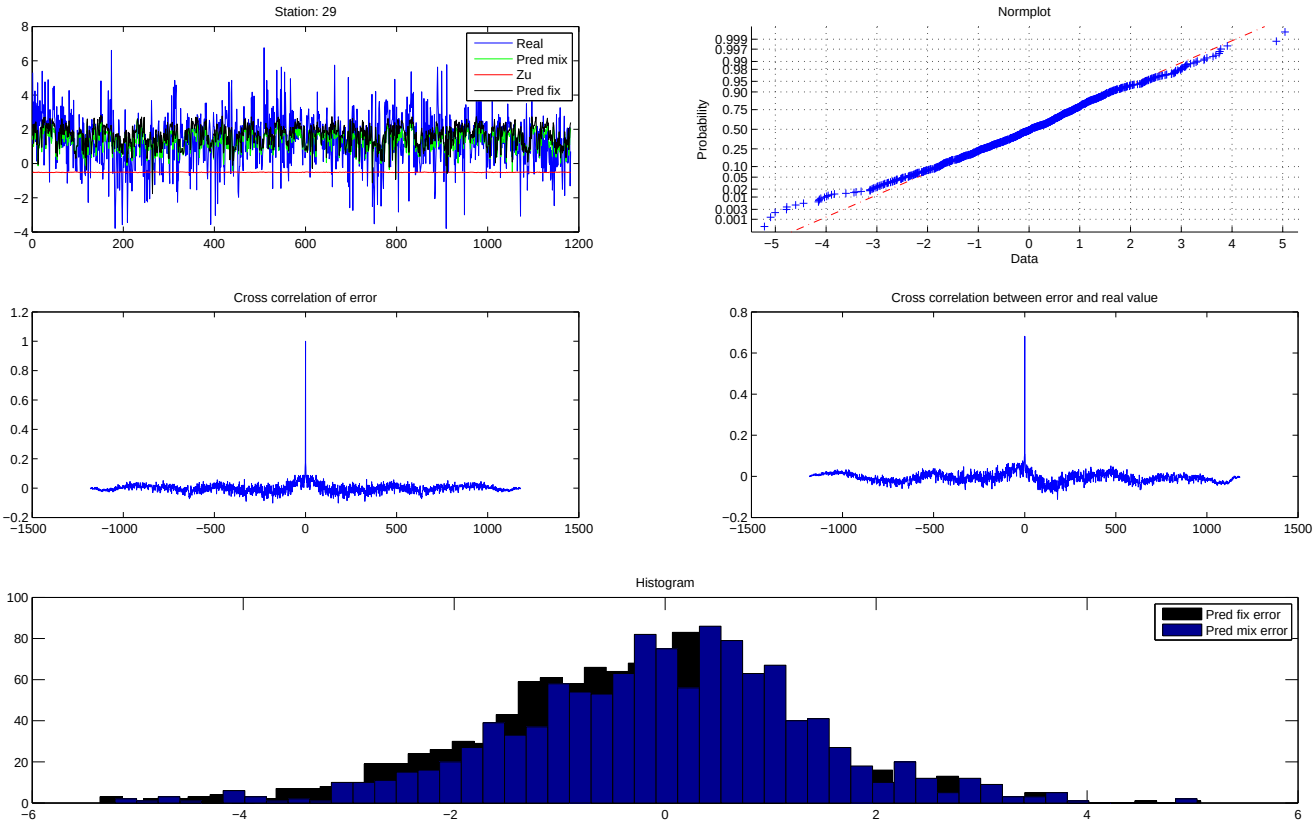

Figure A.4: Subplot 1 consists of the real log-transformed MATCH error (blue), the prediction of model 1 (green), the prediction of model 2 (black) and the mixed effect (red). subplot 2-4 is a normal, cross correlation function, cross correlation with the real error of model 1 (from left to right). The last subplot consists of two histogram of the residuals of model 1 (blue) and model 2 (black).

Station 29 and 43 are very close to each other. The correlation between the coefficients are large enough to estimate the mixed effect. Model 1 gives 3% less mean quadratic error than in model 2. As seen in Figure A.4 the mixed effect is almost constant with a value of approximately -0.3. The prediction of model 1 is almost unbiased.

A.0.5 Station 40

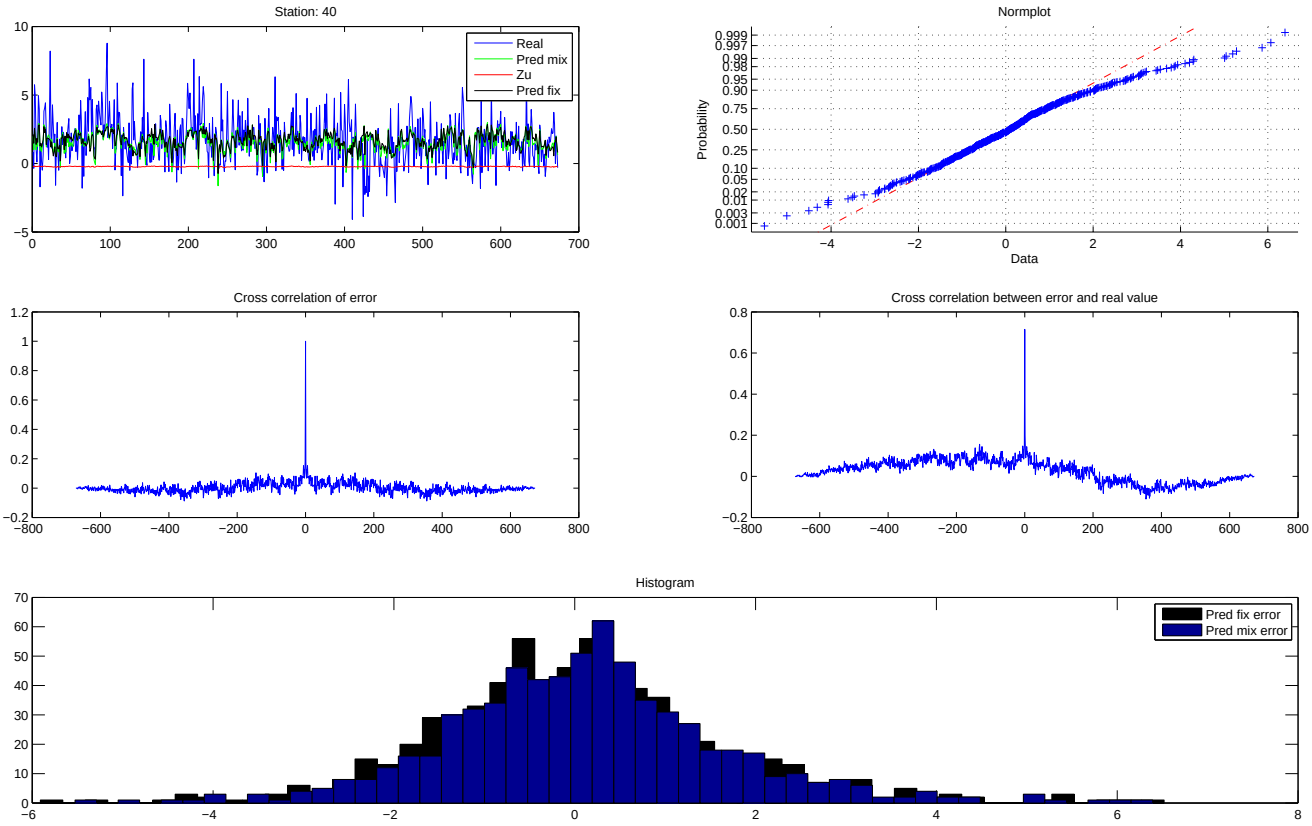

Figure A.5: Subplot 1 consists of the real log-transformed MATCH error (blue), the prediction of model 1 (green), the prediction of model 2 (black) and the mixed effect (red). subplot 2-4 is a normal, cross correlation function, cross correlation with the real error of model 1 (from left to right). The last subplot consists of two histogram of the residuals of model 1 (blue) and model 2 (black).

The prediction error of model 1 is 8.6% less than the error of model 2. The prediction error of model 1 is almost unbiased as seen in the last sub plot of Figure A.5. The mixed effect makes not much difference, it is almost constant around -0.25.

A.0.6 Station 72

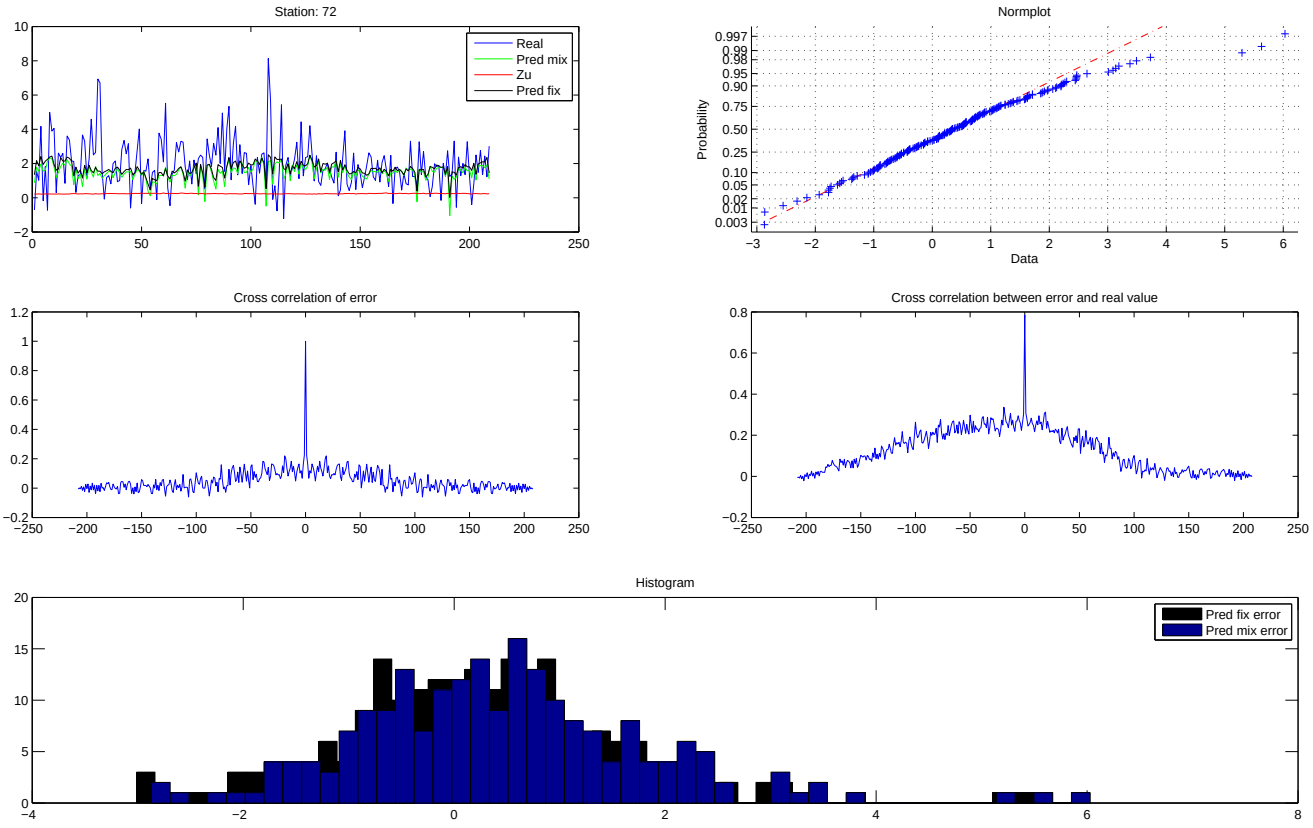

Figure A.6: Subplot 1 consists of the real log-transformed MATCH error (blue), the prediction of model 1 (green), the prediction of model 2 (black) and the mixed effect (red). subplot 2-4 is a normal, cross correlation function, cross correlation with the real error of model 1 (from left to right). The last subplot consists of two histogram of the residuals of model 1 (blue) and model 2 (black).

The predictions at station 72 are both biased, model 1 less so than model 2. Nonetheless model 2 has 0.2% less mean quadratic error.

A.0.7 Station 90

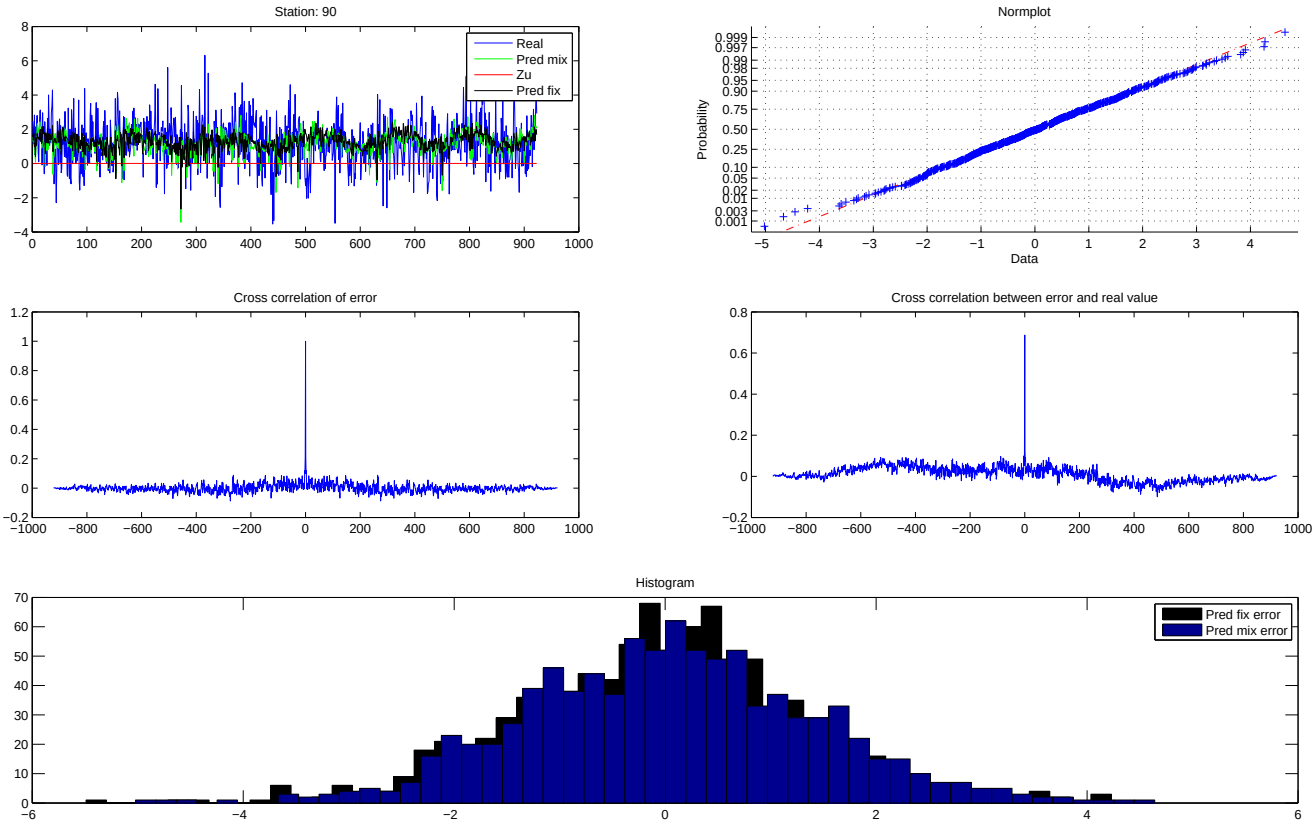

Figure A.7: Subplot 1 consists of the real log-transformed MATCH error (blue), the prediction of model 1 (green), the prediction of model 2 (black) and the mixed effect (red). subplot 2-4 is a normal, cross correlation function, cross correlation with the real error of model 1 (from left to right). The last subplot consists of two histogram of the residuals of model 1 (blue) and model 2 (black).

As seen in Figure reffig:station8 there is hardly any mixed effect helping model 1. Both prediction errors are unbiased. The prediction error of model 2 has 7% less mean quadratic error than model 2.

Bibliography

[1] Lennart Robertson and Joakim Langner(1998) *An Eulerian Limited-Area Atmospheric Model*
(Online)Available:
http://journals.ametsoc.org/doi/pdf/10.1175/1520-0450

[2] ECDS *ECDS*
(Online)Available:
http://www.smhi.se/ecds/about-ecds

[3] unidata *unidata*
(Online)Available: http://www.unidata.ucar.edu/software/thredds/current/tds/TDS.html

[4] Gunilla Pihl Karlsson, Sofie Hellsten, Per Erik Karlsson, Cecilia Akselsson and Martin Ferm *Kvävedeposition till Sverige, Jämförelse av depositionsdata från Krondroppsnätet Luft- och nderbördskemiska nätet samt EMEP*
(Online)Available: http://www.ivl.se/download/
18.3175b46c133e617730d800015394/1350484316686/B2030.pdf

[5] N.H. Bringham, Johan M. Fry (2010)*Regression, Linear Models in Statistics*
Imperial College, London and the University of East London

[6] N.H. Bringham, Johan M. Fry (2010)*Regression, Lear Models in Statistics*
Imperial College, London and the University of East London. Page 119.
regression1s119

[7] N.H. Bringham, Johan M. Fry (2010)*Regression, Lear Models in Statistics*
Imperial College, London and the University of East London, page 119

[8] Jose' C. Pinheiro, Douglas M. Bates (2000)*Mixed-Effects Models in S and S-PLUS* Springer

[9] Alan E. Gelfand, Peter j. Diggle, Montserrat Fuentes Peter Guttorp (2010)*Handbook of Spatial Statistics* CRC Press

[10] Andreas Jakobsson (2013)*An introduction to Time Series Modeling*
Studentlitteratur AB, Lund

[11] MathWorks (2014)*Stepwisefit*
(Online)Available: http://www.mathworks.se/help/stats/stepwisefit.html