2010-0000

Kalman Filter in Real Time URBIS

Date	Jan 2010
Author(s)	R. Kranenburg
Projectnumber	
Keywords	NOx, Real Time URBIS, Uncertainty, Rijnmond
Target	

Contents

1 Introduction

In this report the use of a Kalman filter in the Real Time URBIS model will be discussed. The Real Time URBIS model is a model which calculate the concentration NO_x in a city or an industrialized region. The concentration of NO_x is assumed to be equal to the sum of the concentrations for NO and NO_2. Nitrogen oxides are formed by the burning of fossil fuels in traffic and industry, they will arise if nitrogen from the air and the fuels reacts with oxygen. These nitrogen oxides reacts under influence of sunlight to air pollution, like smog and acid rain. Nitrogen oxides can also causes trouble for the eyes and lungs. Therefore the European commission has set out limit values for the concentrations of NO_2, thus it is important to have a good view on the concentrations NO_x and with that on the concentrations NO_2.

The Real Time URBIS model simulates the NO_x concentration by adding emissions from different sources like traffic, residents, shipping and industry. In this report a Kalman filter will be used to link the model simulations with a series of measurements made on 9 different stations. With this link a better simulation for the concentration NO_x can be given. Also a statistical uncertainty interval of the concentration can be given.

In Chapter 2, a more detailed explanation of the Real Time URBIS model is given. In Chapter 3, a statistical uncertainty of the model is made as well as ideas for the Kalman filter. In Chapter 4, the general use of a Kalman filter will be explained. In Chapters 5 and 6, the Kalman filter is applied on the Real Time URBIS model. First only on the background source, later on all the sources. In the last part of Chapter 6, the uncertainty intervals calculated with the Kalman filter will be connected with the population to give a functional application of this method.

2 Model and Measurements

2.1 Real Time URBIS model

Real Time URBIS is a model to determine the concentration of NO_x in a city or an industrialized region. The model calculates a concentration for the whole region, based on factors like wind, temperature and time on every hour. This study focuses on the Rijnmond area around Rotterdam; the domain of the study is shown in Figure 2.1.

The basis of the Real Time URBIS model is the URBIS model. The URBIS model calculates for the whole Rijnmond area the annual average concentration NO_x for the whole year. Detailed information about the URBIS model may be found in (Wesseling and Zandveld, 2003).

With the Real Time URBIS model, the annual average concentrations are used to get an average concentration for every hour. The state of the Real-Time-URBIS consists of the NO_x concentrations in a large number of grid points in the domain. Mathematically, the state is described by a vector:

$$\underline{c_k} \tag{2.1}$$

where k denotes the hour. In this study, the state vector is defined on about 94096 grid points covering the Rijnmond area, irregularly distributed over the grid. The state is computed as a linear combination of standard concentration fields, each representing the mean concentrations due to emissions from a particular source (traffic, ships, industry and residents) given a certain wind direction and wind speed. This is given in the state equation:

$$\underline{c_k} = M\underline{u_k} \tag{2.2}$$

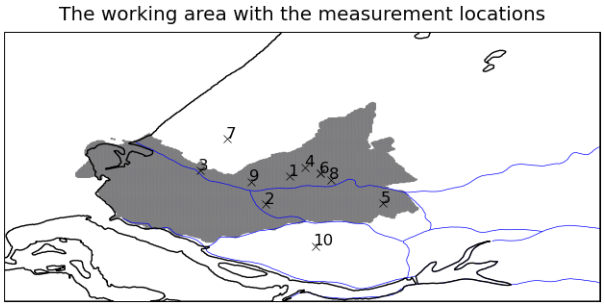
where each column of the matrix M is one of the standard concentration fields. In the model version used in this study, the total number of standard fields is 88, valid for 11 different source categories, 4 different wind directions, and 2 different wind speeds. Plots of all standard fields are included in Appendix A. The elements of vector u_k represent the weight of each standard field in the concentration at hour k. The weight depends on the meteorological conditions (wind direction, wind speed, temperature) and the moment (month, day of the week, hour).

In (Kranenburg, 2009) a more detailed description of the Real Time URBIS model could be found. Note that the Real Time URBIS model described in (Kranenburg, 2009) has an underlying URBIS model valid for the year 2000, while in this report the underlying URBIS model is valid for the year 2006. This URBIS model for 2006

has one large difference with the URBIS model for the year 2000. The model for 2000 consists of 10 different source categories instead of 11 in 2006, since an extra source in category traffic is added, namely source 'Zone Cards'. Further is the source category 'boundary' is renamed into 'rest'.

2.2 Measurements

In the Rijnmond area, there are also 11 locations where the concentrations NO and NO_2 are measured. The sum of these two concentrations is called NO_x. The locations of these 11 measurement stations are also shown in Figure 2.1. Locations 1-6 are sites operated by DCMR[1], the locations 7-11 are sites operated by RIVM[2]. Measurement stations 6 and 11 are located directly next to each other, in the Real Time URBIS model both locations are at the same grid point. Only 9 of these locations are in the domain covered in this study, locations 7 (Schipluiden) and 10 (Westmaas) are outside of the domain and will only be used as background stations. As will described in Chapter 3, the measurements on these 9 locations could be used to estimate the uncertainty of the model, by comparing the model results with the observations.



DCMR - Locations	
1	Schiedam
2	Hoogvliet
3	Maassluis
4	Overschie
5	Ridderkerk
6	Bentinckplein

RIVM - Locations	
7	Schipluiden
8	Schiedamsevest
9	Vlaardingen
10	Westmaas
11/6	Bentinckplein

Figure 2.1: *Domain of the working area for Real Time URBIS*

[1]DCMR: Dienst Centraal Milieubeheer Rijnmond. *www.dcmr.nl*
Environmental protection agency for the Rijnmond area around Rotterdam
[2]RIVM: RijksInstituut voor Volksgezondheid en Milieu. *www.rivm.nl*
Dutch institute for public health and environment

3 Statistical Uncertainty of the Real Time URBIS Model

3.1 Introduction

In (Kranenburg, 2009), a method is described to compute a bias correction for the simulations made by the Real Time URBIS model, by comparing the model simulations with the observations made on the measurement stations. After application of the Real Time URBIS model, the simulation is adjusted with a value dependent on the different meteorological conditions (wind direction, wind speed, temperature) and the moment (month, day of the week, hour). This correction is typically an example of post-processing; after applying the model, the model results are corrected with the aid of the measurements. In addition, from the dependencies on the meteorological conditions and the moment, the origin of the uncertainties in the Real Time URBIS model could be found. In this chapter, the same method is applied on the Real Time URBIS model with the underlying URBIS model for 2006. For the year 2006 all the differences between the observations and the model simulations are calculated. All these differences are used to make a correction on the results of the Real Time URBIS model.

3.2 Log-normal distributions

For all 9 measurement stations in the area, the observations are plotted in a histogram. This is shown in the left panel of Figure 3.1. In the right panel of Figure 3.1, the simulations for the grid points comparing with the measurement stations are plotted in a histogram. It is interesting that both the observations as well as the model simulations have a log-normal distribution. For this reason all corrections should be done in the log-domain. The main advantage of working in the log-domain is that when there is a correction added to the state, this correction is made on the logarithm of the concentration. After correction, the logarithm of the concentration could become negative but the corresponding concentration itself can not. In fact, an additive correction on the logarithm of the concentration is the same as a fractional correction on the absolute concentration:

$$\ln(c_k) \rightarrow \ln(c_k) + \mu \tag{3.1a}$$

$$c_k = e^{\ln(c_k)} \rightarrow e^{\ln(c_k)+\mu} = e^{\mu} c_k \tag{3.1b}$$

where c_k is the concentration at time k and μ is the correction term. Since the correction factor e^{μ} is always positive, the concentrations will remain positive too. Detailed information about corrections in the log-domain is given in (Kranenburg, 2009).

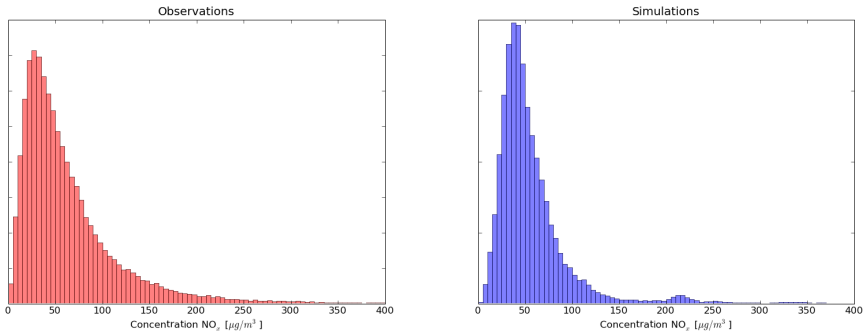


Figure 3.1: *Both the observations and the model simulations have a log-normal distribution.*

3.3 Uncertainty of the Real Time URBIS model

The method from (Kranenburg, 2009) to describe the uncertainty of the model is now applied on the Real Time URBIS model, with underlying URBIS model for 2006. For the year 2006 all logarithms of the observations made on the 9 monitoring locations are compared with logarithms of the model simulations. In total there are at most $9\ \text{stations} \times 8760\ \text{hours} = 78840$ of those differences. Due to some missing measurements or meteorological data, for 2006 there are only 67080 differences. With all these differences, the correction on the results of the Real Time URBIS model is made.

3.3.1 Structural bias

First the differences between the model results and the observations did not have mean zero, thus there is a systematical error in the model. This structural error causes a constant correction ($\mu = \mu_c$) on the logarithm of the model simulation, which corresponds with a constant fractional correction of the absolute concentration.

3.3.2 Wind direction dependency

The differences between observations and the logarithms of the simulations made with the constant corrected model are plotted with respect to the wind direction. For each of the 36 wind directions, all differences which appear during that wind direction are taken. In Figure 3.2 all the means per wind direction μ_i are plotted as blue dots.

$$\mu_i = \frac{1}{n_i}\sum_{k=1}^{n_i}\left(\ln\left(y_{i,k}\right) - \ln\left(c_{i,k}^m\right)\right) \tag{3.2a}$$

$$\sigma_i = \sqrt{\frac{1}{n_i}\sum_{k=1}^{n_i}\left(\left(\ln\left(y_{i,k}\right) - \ln\left(c_{i,k}^m\right)\right) - \mu_i\right)^2} \tag{3.2b}$$

where n_i represents the number of differences that appears during wind direction i, y_i represents the observations and c_i^m the model simulations, during wind direction i. The standard deviations σ_i of the differences per wind direction are represented by the length of the error bars in Figure 3.2. Per wind direction, Figure 3.2 shows a 1σ uncertainty interval for the difference between the logarithm of the observation and the logarithm of the model simulation. The green line forms the correction which is added to the logarithm of the model simulations. The correction on the model is now a function of the wind direction ϕ, thus $\mu = \mu_c + \mu_{wdir}(\phi)$.

This green line is a composed sinus function, that fits best on the differences between the model simulations and the observations. This best fitting is made with the blue dots and the wind rose in Figure 3.3, the weights for each blue dot are given in this wind rose. When a wind direction occurs a lot, the weight must be larger in the calculation of the best fitting sinus.

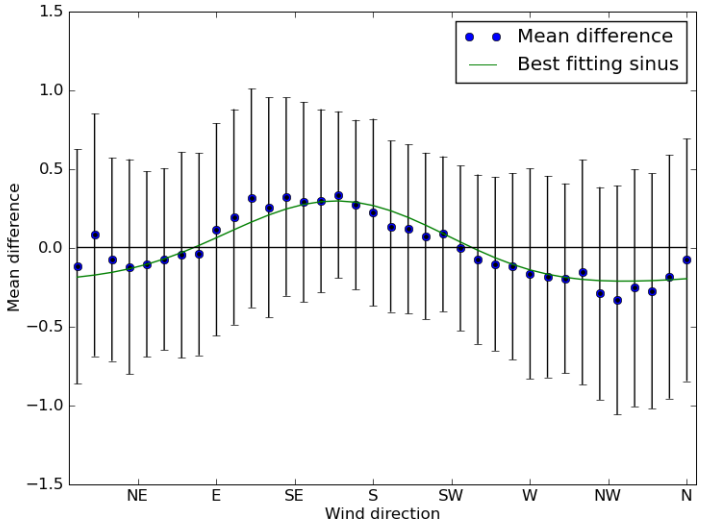


Figure 3.2: *Mean differences between the logarithms of the observations and the logarithms of the model simulations against the wind direction after constant correction.*

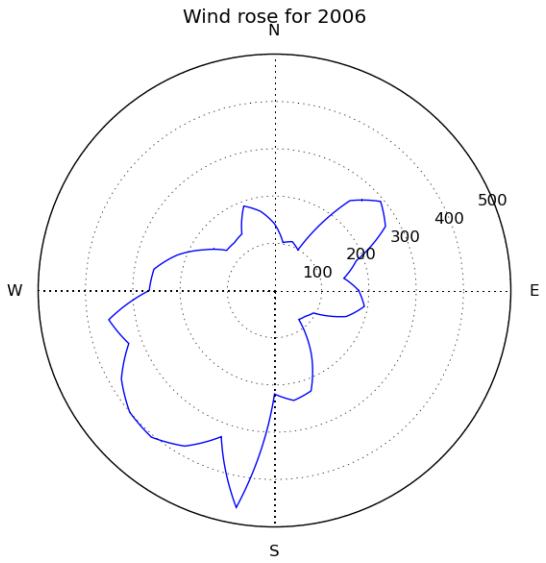


Figure 3.3: *Wind rose over 2006*

3.3.3 Hourly differences

After the correction on the wind direction, all the differences are plotted for every hour of the day in Figure 3.4. In this figure, the mean of all differences per hour of the day is plotted with a blue dot and the standard deviation is represented by the width of the error bars, computed similar to equations 3.2a and 3.2b. The green line is again a composed sinus function, which fits best on the blue dots. Because of missing measurements or meteorological data, not every hour has the same contribution in calculating the best fitting sinus. This best fitting sinus forms the correction added to the logarithms of the model simulations as a function of the hour of the day. The total correction on the model is now built from three parts: a constant part, a function dependent on the wind direction and a function dependent on the hour of the day (h).

$$\mu \quad = \quad \mu_c + \mu_{wdir}\left(\phi\right) + \mu_{hour}\left(h\right)$$

After this correction, the differences were plotted against the other input parameters wind speed, temperature, month and day of the week. It was found that the differences are no longer dependent on one of these input parameters.

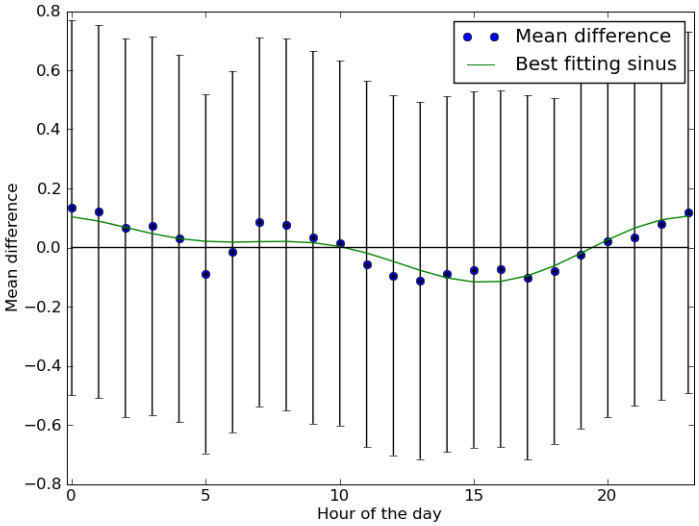


Figure 3.4: *Mean difference between the logarithms of the observations and the logarithms of the model simulations against the hour of the day, after correction on the wind-direction*

3.3.4 Standard deviation of the differences

The standard deviation of the differences was found to be a function of the wind speed. This is shown in Figure 3.5. The blue dots represent the standard deviation of all differences as a function of wind speed. This is done with an equation similar to equation 3.2b. The red line is the best fitting exponential function on the standard deviations per wind speed. In the calculation of this best fitting exponential the number of times that a wind speed occurs is also taken. When a wind speed occurs a lot, the weight must be larger in the calculation of the best fitting exponential.

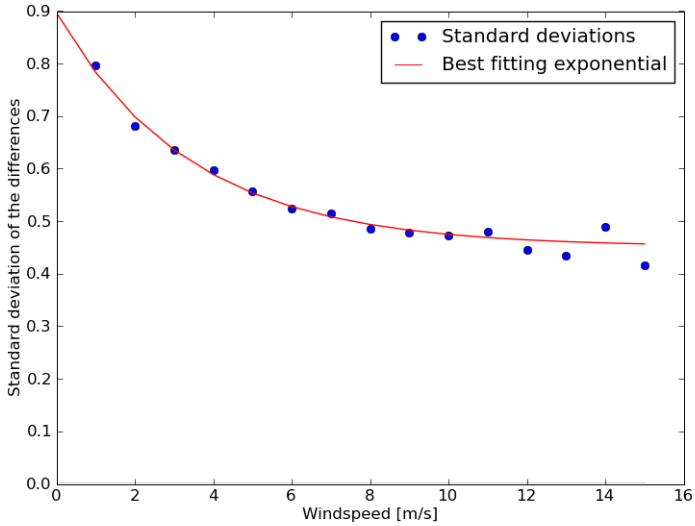


Figure 3.5: *The standard deviations of the differences between the logarithms of the observations and the model simulations against the wind-speed, after correction on the hour of the day*

3.4 Discussion

With the method described above, an uncertainty interval for the concentration NO_x can be given. In Figure 3.6 the 1σ uncertainty interval for the first week of 2006 is given for the grid cell with station Schiedam. The red dots represents the observations available in that period. The described method gives a useful approximation of the uncertainty of the model, however the uncertainty of the model is very large at some times.

The next objection is to decrease the uncertainty of the model by an improvement of the model. An indication for the largest inaccuracy of the model is given by the uncertainty analysis above. The differences between the observations and the model simulations are mostly dependent on the wind direction. For this reason it is assumed that the standard concentration fields for the source background are not accurate in the URBIS model. The source background corresponds with emission produced in the rest of the country which is blown into the Rijnmond area. Of course this source has a large dependency on the wind direction. In Chapter 5 a Kalman filter will be used to get better estimates of the background concentrations per wind direction. The advantage of using a Kalman filter is that also the uncertainty of the measurements is involved in the estimation. In Chapter 4 the working of a Kalman filter is explained.

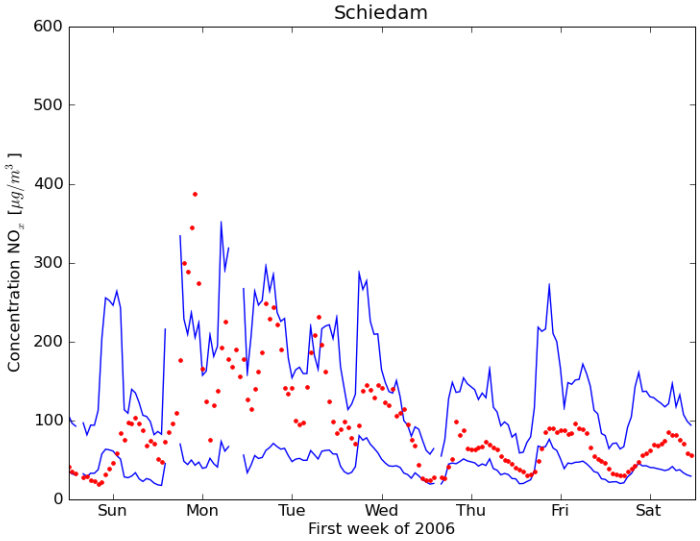


Figure 3.6: *Uncertainty interval for the first week on location Schiedam*

4 Kalman Filter

4.1 Introduction

A Kalman filter is mostly used to smooth random errors in the model of a dynamical system. In Figure 4.1 a schematic representation of the working of a Kalman filter is given. The simulations made by the model for time step k are corrected with the aid of a measurement on time step k. In this correction, also the uncertainties of the model and the measurements are taken into account. This application is very useful in a real time application such as the Real Time URBIS model.

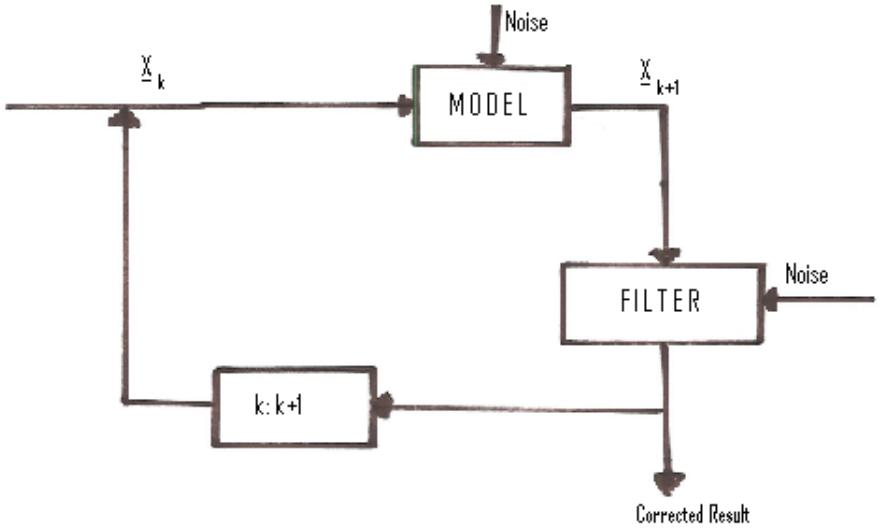


Figure 4.1: *Schematic representation of the Kalman filter*

4.2 Algorithm of Kalman Filtering

In this section the working of a Kalman filter is explained with a simple one dimensional example. The simulations made with the Real Time URBIS model, for location Schiedam are compared with a series of measurements on location Schiedam. This is done for the year 2006, in this year the Real Time URBIS model gives for 7906 of the 8760 hours a concentration NO_x. For the other hours, one of the meteo

input data is missing, thus the model cannot give a result. For 8603 of the 8760 hours there is an observation, on the other hours, the measurement was incorrect or missing. When the model does not give a result, the Kalman filter cannot give a result on that time step. When there is no measurement, the Kalman filter can give a result, which is computed from the previous time step. In the figures in this chapter there will be some 'holes', this is due to the missing model data.

4.2.1 Dynamical system

First of all it is important to have a well defined dynamical system. For the example in Schiedam, this dynamical system is given by:

$$\ln(c_k) = \ln(c_k^m) + \gamma_k \tag{4.1a}$$

$$\gamma_{k+1} = \alpha\gamma_k + \beta_k\omega_k \qquad \omega_k \sim N(0,1) \tag{4.1b}$$

In this equations the value c_k is the concentration of NO_x on time step k on location Schiedam. Every time step (hour) the Real Time URBIS model calculates a concentration NO_x on location Schiedam; these are called c_k^m. Because of the log-normal distribution of the concentration NO_x, the dynamical system deals with the logarithms of the concentration NO_x. More about this is discussed in Section 3.2 and in (Kranenburg, 2009). The parameter γ_k is an estimate for the difference between the logarithm of the real concentration and the logarithm of the model result, also called the perturbation on the model. With a Kalman filter these perturbations γ_k will be estimated.

This estimation does not lead to a computation of the optimal value for γ_k. Instead the result after application of the Kalman filter is that the value of γ_k can be found in a Gaussian distribution with mean $\hat{\gamma}_k$ and a variance p_k^2. With this Gaussian distribution, the value for $\ln(c_k)$ can be found in a Gaussian distribution with mean $(\ln(c_k^m) + \hat{\gamma}_k)$ and variance p_k^2. This all leads to an uncertainty interval for the logarithm of the concentration NO_x at every time step and with that, an uncertainty interval for the concentration NO_x .

For the perturbations, it is assumed that a perturbation at time k is correlated with the perturbation on the time step before, but that it also has a random component. A suitable mathematical description is an 'AR1' (auto-regressive 1) process, this is also called 'colored noise'. The temporal correlation is described by the parameter α, which also appears in the formula for the amplitude of the random contribution:

$$\beta_k = \sqrt{1-\alpha^2}\sigma_k \tag{4.2}$$

When $\alpha = 0$, the perturbation only has the random process, thus only white noise with standard deviation σ.

When α is close to one, the temporal correlation is strong and the fluctuations per time step are small. The value α is computed from:

$$\alpha = e^{-1/\tau} \tag{4.3}$$

where τ is a de-correlation scale. In this example the value of τ is chosen equal to 12, such that the perturbation is practically independent of the perturbation 12 time steps before.

4.2.2 Kalman filter form

For application of the Kalman filter, the dynamical system has to be written in the Kalman filter form:

$$\gamma_{k+1} = \alpha\gamma_k + \beta_k\omega_k \qquad \omega_k \sim N(0,1) \tag{4.4a}$$

$$\ln(y_k) = H(\ln(c_k^m) + \gamma_k) + \nu_k \qquad \nu_k \sim N\left(0, r_k^2\right) \tag{4.4b}$$

In here y_k is the observation on time step k, and H is the system operator, which projects the model state onto the observations. The observation error ν_k represents the error of the measurement, combined the instrumental error and the representation error, which is supposed to be Gaussian with zero mean and variance r_k^2.

For the example on location Schiedam, the system operator H is equal to 1, which means that the observation is just the model plus some perturbation. The value of r_k is assumed to be equal to 0.2. This means that the logarithms of the measurements has an uncertainty of 20%. Also the value σ_k is set to 0.2 too, which means that the perturbation on the model also has an uncertainty of 20 %.

The Kalman filter process could be started with initial values $\gamma_0 = 0$, and $p_0^2 = 0$; this is equivalent to the assumption that the expected concentration at time 0 equals the model result and the uncertainty is zero at this time.

4.2.3 Forecast step

In this first step of the Kalman filter, a forecasted mean $\hat{\gamma}_k^f$ of the perturbation is calculated with the mean from the previous time step. This forecasted mean is the expectation of γ_k.

$$\begin{aligned}\hat{\gamma}_{k+1}^f &= \mathrm{E}[\gamma_{k+1}] \\ &= \mathrm{E}[\alpha\gamma_k + \beta\omega_k] \\ &= \alpha\mathrm{E}[\gamma_k] + \beta\mathrm{E}[\omega_k] \\ &= \alpha\mathrm{E}[\gamma_k] \\ &= \alpha\hat{\gamma}_k\end{aligned} \tag{4.5}$$

where is used that $\mathrm{E}[\omega_k] = 0$. For the example of Schiedam the time correlation:

$$\alpha = e^{-1/12} \approx 0.92$$

Also a forecasted variance $\left(p_{k+1}^f\right)^2$ is calculated with the variance from the time step before.

$$\begin{aligned}
\left(p_{k+1}^f\right)^2 &= \mathrm{VAR}\left(\gamma_{k+1}\right) \\
&= \mathrm{E}\left[\left(\gamma_{k+1} - \mathrm{E}\left[\gamma_{k+1}\right]\right)^2\right] \\
&= \mathrm{E}\left[\left(\alpha\gamma_k + \beta_k\omega_k - \hat{\gamma}_{k+1}^f\right)^2\right] \\
&= \mathrm{E}\left[\left(\alpha\left(\gamma_k - \hat{\gamma}_k^f\right) + \beta_k\omega_k\right)^2\right] \\
&= \mathrm{E}\left[\alpha^2\left(\gamma_k^2 - \hat{\gamma}_k^f\right)^2 + 2\alpha\beta_k\left(\gamma_k - \hat{\gamma}_k^f\right)\omega_k + \beta_k^2\omega_k^2\right] \\
&= \alpha^2\mathrm{E}\left[\left(\gamma_k - \hat{\gamma}_k^f\right)^2\right] + 2\alpha\beta_k\mathrm{E}\left[\gamma_k - \hat{\gamma}_k^f\right]\mathrm{E}\left[\omega_k\right] + \beta_k^2\mathrm{E}\left[\omega_k^2\right] \\
&= \alpha^2\mathrm{E}\left[\left(\gamma_k - \mathrm{E}\left[\gamma_k\right]\right)^2\right] + \beta_k^2 \\
&= \alpha^2\mathrm{VAR}\left(\gamma_k\right) + \beta_k^2 \\
&= \alpha^2\left(p_k\right)^2 + \left(1 - \alpha^2\right)\sigma_k^2 \qquad (4.6)
\end{aligned}$$

where the independency of γ_k and ω_k is used, as well as $\mathrm{E}\left[\omega_k\right] = 0$ and $\mathrm{E}\left[\omega_k^2\right] = \mathrm{VAR}\left(\omega_k\right) = 1$

4.2.4 Analysis step

In the second step, the Kalman filter analyzes the results of the forecast step with an observation. A basic assumption in a Kalman filter is that the mean after the analyzing step $\hat{\gamma}_k^a$ is a linear combination of the forecasted mean and the difference between the logarithm of the observation and the logarithm of the model simulation. This results in an analyzed mean which is the forecasted mean plus a perturbation relative to the difference between the observation and its related simulation:

$$\hat{\gamma}_{k+1}^a = \hat{\gamma}_{k+1}^f + K_{k+1}\left(\ln\left(y_{k+1}\right) - H\left(\ln\left(c_{k+1}^m\right) + \hat{\gamma}_{k+1}^f\right)\right) \qquad (4.7)$$

The variance in this analyzing step $\left(p_{k+1}^a\right)^2$ is created by the variance from the forecast step and the variance from the representation error of the measurements.

$$
\begin{aligned}
\left(p_{k+1}^{a}\right)^{2} &= \mathrm{VAR}\left(\gamma_{k+1}\right) \\
&= \mathrm{E}\left[\left(\gamma_{k+1} - \mathrm{E}\left[\gamma_{k+1}\right]\right)^{2}\right] \\
&= \mathrm{E}\left[\left(\gamma_{k+1} - \hat{\gamma}_{k+1}^{a}\right)^{2}\right] \\
&= \mathrm{E}\left[\left(\gamma_{k+1} - \left(\hat{\gamma}_{k+1}^{f} + K_{k+1}\left(\ln\left(y_{k+1}\right) - H\left(\ln\left(c_{k+1}^{m}\right) + \hat{\gamma}_{k+1}^{f}\right)\right)\right)\right)^{2}\right] \\
&= \mathrm{E}\left[\left(\gamma_{k+1} - \left(\hat{\gamma}_{k+1}^{f} + K_{k+1}\left(H\left(\ln\left(c_{k+1}^{m}\right) + \gamma_{k+1}\right) + \nu_{k+1}\right)\right.\right.\right. \\
&\quad \left.\left.\left. - H\left(\ln\left(c_{k+1}^{m}\right) + \hat{\gamma}_{k+1}^{f}\right)\right)\right)^{2}\right] \qquad (4.8a) \\
&= \mathrm{E}\left[\left(\left(1 - K_{k+1}H\right)\left(\gamma_{k+1} - \hat{\gamma}_{k+1}^{f}\right) - K_{k+1}\nu_{k+1}\right)^{2}\right] \\
&= \mathrm{E}\left[\left(1 - K_{k+1}H\right)^{2}\left(\gamma_{k+1} - \hat{\gamma}_{k+1}^{f}\right)^{2}\right. \\
&\quad \left. - 2K_{k+1}\left(1 - K_{k+1}H\right)\left(\gamma_{k+1} - \hat{\gamma}_{k+1}^{f}\right)\nu_{k+1} + K_{k+1}^{2}\nu_{k+1}^{2}\right] \\
&= \left(1 - K_{k+1}H\right)^{2}\mathrm{E}\left[\left(\gamma_{k+1} - \mathrm{E}\left[\gamma_{k+1}\right]\right)^{2}\right] \\
&\quad + 2K_{k+1}\left(1 - K_{k+1}H\right)\mathrm{E}\left[\left(\gamma_{k+1} - \hat{\gamma}_{k+1}^{f}\right)\right]\mathrm{E}\left[\nu_{k+1}\right] + K_{k+1}^{2}\mathrm{E}\left[\nu_{k+1}^{2}\right] \\
&= \left(1 - K_{k+1}H\right)^{2}\left(p_{k+1}^{f}\right)^{2} + K_{k+1}^{2}r_{k+1}^{2} \qquad (4.8b)
\end{aligned}
$$

where the independency of γ_k and ν_k is used, as well as $\mathrm{E}\left[\nu_{k+1}\right] = 0$ and $\mathrm{E}\left[\nu_{k+1}^{2}\right] = 1$.

In line 4.8a, the Formula 4.4b is used. After this analyzing step, the values for $\hat{\gamma}_k^a$ and $\left(p_k^a\right)^2$ are the mean $\hat{\gamma}_k$ and the variance $\left(p_k\right)^2$ for the state of the system on time k.

A common choice for K_k is the minimum variance gain. For that gain, the value K_k is chosen such that the variance $\left(p_k^a\right)^2$ reaches a minimum. To obtain the minimum variance, the solution for K_{k+1} of

$$\frac{\partial\left(p_{k+1}^{a}\right)^{2}}{\partial K_{k+1}} = 0 \qquad (4.9)$$

has to be found. This is done in the next formula:

$$\begin{aligned}
\frac{\partial \left(p^a_{k+1}\right)^2}{\partial K_{k+1}} &= 0 \\
2K_{k+1}r^2_{k+1} - 2H\left(1 - K_{k+1}H\right)\left(p^f_{k+1}\right)^2 &= 0 \\
2K_{k+1}\left(r^2_{k+1} + H^2\left(p^f_{k+1}\right)^2\right) &= 2H\left(p^f_{k+1}\right)^2 \\
K_{k+1} &= \frac{H\left(p^f_{k+1}\right)^2}{H^2\left(p^f_{k+1}\right)^2 + r^2_{k+1}} \qquad (4.10)
\end{aligned}$$

Because of the second derivative

$$\frac{\partial^2 \left(p_{k+1}\right)^2}{\partial K^2_{k+1}} = 2H^2\left(p^f_{k+1}\right)^2 + 2r^2_{k+1} > 0 \qquad (4.11)$$

this extreme corresponds with a minimum.

4.2.5 Simple example of Kalman filtering

All steps of the Kalman filter are applied on location Schiedam for the first week of 2006. Figure 4.2 shows for the first week of 2006 all the model simulations and observations. At every hour the logarithm of the model result is shown together with the logarithm of the observation.

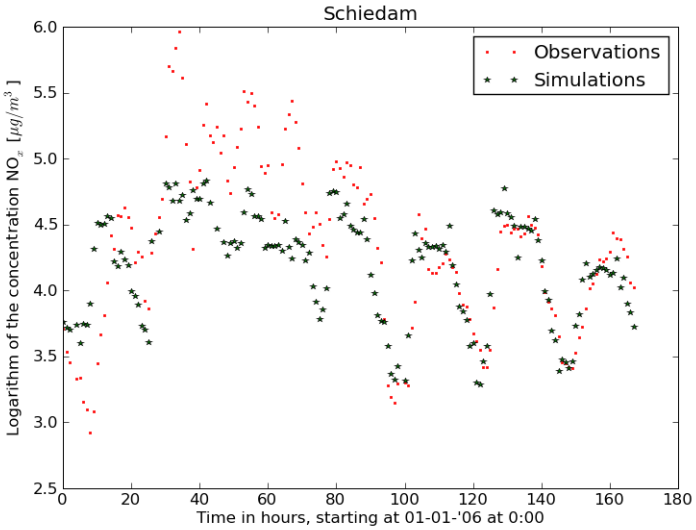


Figure 4.2: *Simulated and measured concentrations for the first week of 2006 on location monitoring location Schiedam*

In Figure 4.3 all steps of the Kalman filter are applied on the model results and the observations for the first week of 2006. The logarithm of the real concentration can

be found in a Gaussian distribution. The 1σ interval is given by the blue lines, this interval corresponds with

$$[\text{model result} + \hat{\gamma} - p, \text{model result} + \hat{\gamma} + p] \tag{4.12}$$

where $\hat{\gamma}$ is the mean after the Kalman filter and p corresponds with the square root of the variance after the Kalman filter. It is clear that in this case the uncertainty interval is mostly between the model result and the observation. In Subsection 4.3 is shown how this interval depends on the several input parameters r^2, α and σ^2.

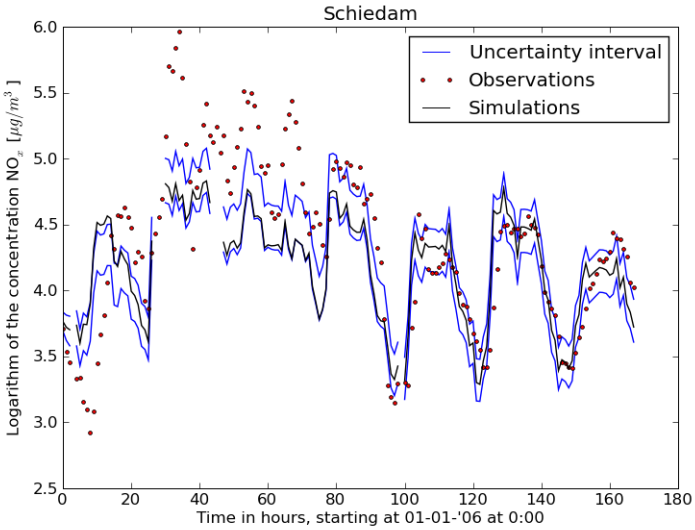


Figure 4.3: *Kalman filter applied on the first week of 2006 on location Schiedam*

In Figure 4.4 and 4.5 is shown what happens when there is a certain period without observations. When there is no observation, only the forecast step of the Kalman filter is applied. The mean value $\hat{\gamma}$ tends to the model results and the standard deviation p tends to the standard deviation of the model. In Figure 4.4 there are no observations analyzed, thus the uncertainty interval tends around the model and the width of the interval corresponds with the standard deviation of the model.

In Figure 4.5 there are no observations analyzed between time step 15 and time step 90. Between those time steps the uncertainty interval after the Kalman filter tends to the model. After time step 90, the Kalman filter analyzes the observations again and the intervals are again between the model results and the observations. In Figure 4.6 and 4.7 this phenomenon is better visible. In these figures, the differences between the model result and the measurement outcomes are shown, with black dots. The intervals in these figures are just the intervals $[\hat{\gamma} - p, \hat{\gamma} + p]$. When there are no measurements analyzed the mean of the interval has to tend to zero and the width of the interval corresponds with the variance of the model. If the observations are analyzed again the interval lies between zero and the black dots

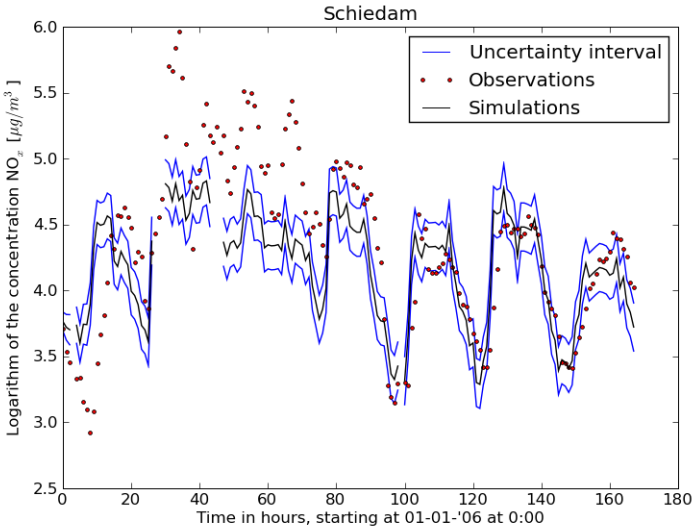


Figure 4.4: *Kalman filter applied on the first week of 2006 on location Schiedam, with no measurements analyzed.*

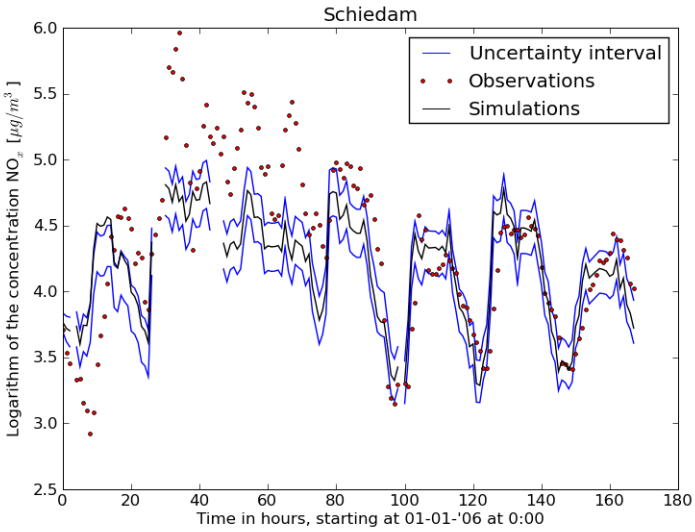


Figure 4.5: *Kalman filter applied on the first week of 2006 on location Schiedam, with no measurements analyzed from time step 15 till time step 90*

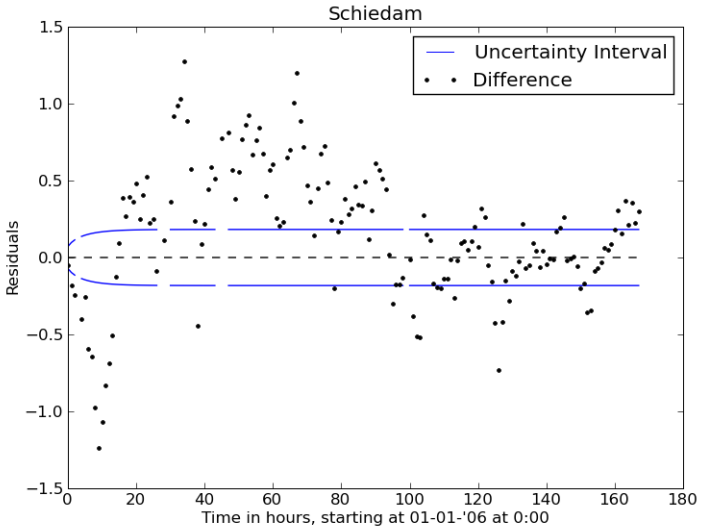


Figure 4.6: *Uncertainty interval of the perturbations with no measurements analyzed, the black dots are the differences between the outcomes of the measurements and the model.*

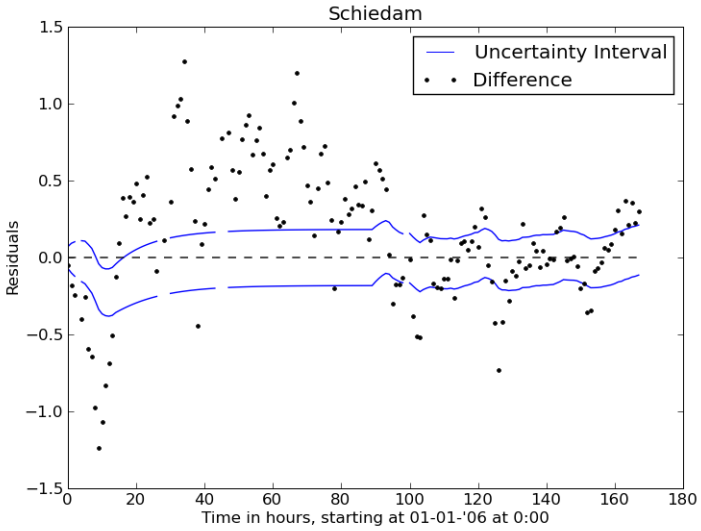


Figure 4.7: *Uncertainty interval of the perturbations with no measurements analyzed from time step 15 until time step 90*

4.3 Sensitivity Tests

In this example for Schiedam, some parameters can be changed to get a better view on their influence. When some parameters are changed the behavior of the Kalman filter is different.

4.3.1 Uncertainty of measurements (r)

The first parameter to change is the uncertainty of the measurements (r^2). In Figure 4.8 is shown what happens when there is respectively a small and a large uncertainty. The left panel of Figure 4.8 shows that when the uncertainty is small, the interval after filtering lies around the measurements. The width of this interval is also small, due to the small uncertainty of the measurements. The right panel in Figure 4.8 shows that, when the measurements have large uncertainty, the interval after filtering lies around the model results. The width of the interval is approximately as large as the uncertainty of the model.

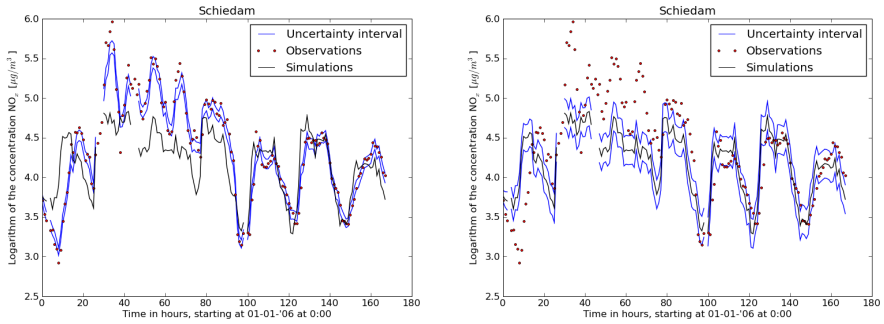


Figure 4.8: *Kalman filter applied on first week for location Schiedam with uncertainty of the measurements assumed small $r = 2\%$ (left panel) and large $r = 200\%$ (right panel)*

4.3.2 Time correlation parameter (α)

The second parameter to change is the time correlation parameter α, in Figure 4.9 the intervals of the perturbations are shown. In the left figure $\tau = 1$, thus $\alpha = e^{-1/\tau} \approx 0.37$, in the right figure $\tau = 250$, thus $\alpha = e^{-1/\tau} \approx 1.00$. The left panel of Figure 4.9 shows that, when α is small, there is hardly no time correlation. It is possible to get high fluctuations of the interval . If there are no observations analyzed from time step 15 till time step 90, the mean of the perturbation will tend rapidly to zero, the width of the interval will rapidly tend to the uncertainty of the model. The right panel of Figure 4.9 shows that when α is large the time correlation is large and the interval cannot make large fluctuations. For that reason the mean of the interval will tend slowly to 0 when there are no measurements analyzed from time step 15 till time step 90. Also the width of the interval will tend slowly to the width corresponding with the uncertainty of the model.

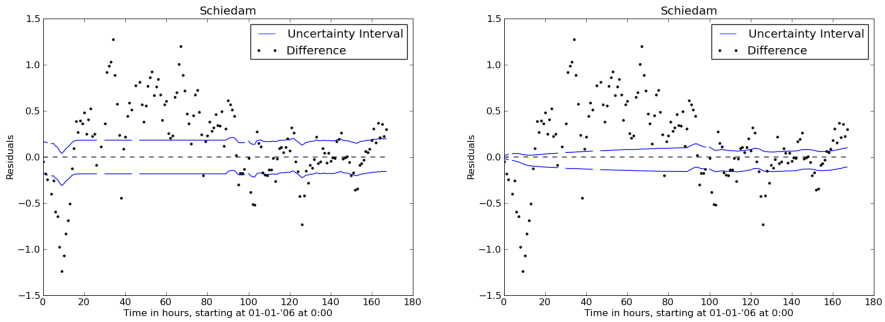


Figure 4.9: *Uncertainty interval for the perturbation for location Schiedam for the first week of 2006, with time correlation assumes small* $\alpha \approx 0.37$ *in the left panel and large* $\alpha \approx 1.00$ *in the right panel.*

4.3.3 Uncertainty of the model (σ)

The last parameter to change is the uncertainty of the model σ_k. In Figure 4.10 it is shown what happens when there is respectively a small and a large model uncertainty. In the left panel of Figure 4.10, the Kalman filter is applied with relatively small model uncertainty. The interval after filtering mostly follows the model and the width of the interval is also small because of the small uncertainty of the model. The right panel of Figure 4.10 shows the uncertainty interval when the model uncertainty is relatively high. The mean of the interval lies around the observations after filtering. The uncertainty interval has approximately the same width as the uncertainty interval corresponding with the uncertainty of the measurements.

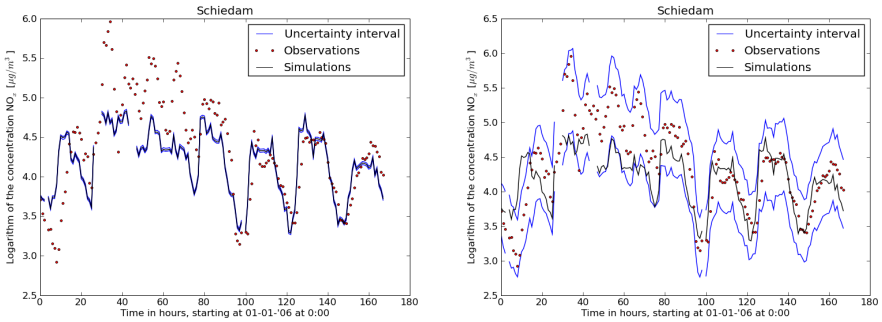


Figure 4.10: *Kalman filter applied on the first week of 2006 for location Schiedam, with model uncertainty assumed small* $\sigma_k = 2\%$ *in the left panel and large* $\sigma_k = 200\%$ *in the right panel*

4.4 Higher dimensional Kalman filtering

The algorithm described in Section 4.2 is an algorithm for a one dimensional problem. This algorithm can easily be extended to a higher dimensional problem. This

is also explained with an example of the Real Time URBIS model. In the Rijnmond area there are 9 locations where the concentration NO_x is measured. On each of these 9 locations there is a real concentration NO_x called $\underline{c}_k$(vector of length 9). Also the Real Time URBIS model gives for every hour a concentration NO_x on each location, called $\underline{c}_k^m$. Again it is necessary to work in the log-domain, thus $\underline{\gamma}_k$ is the perturbation on the model to estimate the real concentrations.

4.4.1 Dynamical system

The dynamical system of this problem

$$\underline{\ln(c)}_k = \underline{\ln(c^m)}_k + \underline{\gamma}_k \tag{4.13}$$

$$\underline{\gamma}_{k+1} = A\underline{\gamma}_k + \underline{\omega}_k \qquad \underline{\omega}_k \sim N(0, Q_k) \tag{4.14}$$

where $\underline{\ln(c)}$ stands for a vector with logarithms of the concentrations.

The dynamical system has become a matrix-vector equation, where matrix A replaces α as time correlation parameter. In the example over all locations A is a diagonal matrix with time correlations α_i on the diagonal representing the time correlations for each entry of $\underline{\Delta c}_k$. Q_k is a covariance matrix, built from the time correlation and the uncertainty of the model. The matrix Q is diagonal with elements $q_i^2 = \left(1 - \alpha_i^2\right)\sigma_i^2$.

4.4.2 Kalman filter form

The dynamical system has to be written in Kalman filter form

$$\underline{\gamma}_{k+1} = A\underline{\gamma}_k + \underline{\omega}_k \qquad \underline{\omega}_k \sim N(0, Q_k) \tag{4.15}$$

$$\underline{\ln(y)}_k = H\left(\underline{\ln(c^m)}_k + \underline{\gamma}_k\right) + \underline{\nu}_k \qquad \underline{\nu}_k \sim N(0, R_k) \tag{4.16}$$

where R_k is a covariance matrix with uncertainty of the measurements. The matrix R_k is also a diagonal matrix with elements r_i^2, the uncertainty of each entry of the vector with measurement outcomes $\underline{\ln(y)}_k$, H is now a higher dimensional system operator, which projects the model state onto the measurement outcomes. The result after Kalman filtering is again that the vector $\underline{\Delta c}_k$ can be found in a Gaussian distribution with mean $\underline{\hat{\gamma}}_k$ and covariance matrix P_k. With this Gaussian distribution, the logarithm of the concentration NO_x can be found in a Gaussian distribution with mean $\left(\underline{\ln(\hat{c}^m)}_k + \underline{\hat{\gamma}}_k\right)$ and covariance matrix P_k. P_k is a covariance matrix with covariances between the entries of state vector $\underline{\gamma}_k$. On the main diagonal of a covariance matrix are variances. From this diagonal, the uncertainty interval for each entry of $\underline{\gamma}_k$ can be computed by taking the square root of these variance. The value for the mean of $\underline{\gamma}_k$, called $\underline{\hat{\gamma}}_k$ is simply $\mathrm{E}\left[\underline{\gamma}_k\right]$.

4.4.3 Forecast step

In the forecast step the mean $\underline{\hat{\gamma}}^f_{k+1}$ is computed with the mean $\underline{\hat{\gamma}}_k$ from the time step before.

$$\begin{aligned}
\underline{\hat{\gamma}}^f_{k+1} &= \mathrm{E}\left[\underline{\gamma}_{k+1}\right] \\
&= \mathrm{E}\left[A\underline{\gamma}_k + \underline{\omega}_k\right] \\
&= A\mathrm{E}\left[\underline{\gamma}_k\right] + \mathrm{E}\left[\underline{\omega}_k\right] \\
&= A\mathrm{E}\left[\underline{\gamma}_k\right] \\
&= A\underline{\hat{\gamma}}_k
\end{aligned} \tag{4.17}$$

where is used that $\mathrm{E}\left[\omega_k\right] = 0$.

The covariance matrix P^f_{k+1} of $\underline{\gamma}_k$ is as in one dimension a function of the covariance matrix from the time step before.

$$\begin{aligned}
P^f_{k+1} &= \mathrm{COV}\left(\underline{\gamma}_{k+1}\right) \\
&= \mathrm{E}\left[\left(\underline{\gamma}_{k+1} - \mathrm{E}\left[\underline{\gamma}_{k+1}\right]\right)\left(\underline{\gamma}_{k+1} - \mathrm{E}\left[\underline{\gamma}_{k+1}\right]\right)^{\mathrm{T}}\right] \\
&= \mathrm{E}\left[\left(\left(A\underline{\gamma}_k + \underline{\omega}_k\right) - A\underline{\hat{\gamma}}_k\right)\left(\left(A\underline{\gamma}_k + \underline{\omega}_k\right) - A\underline{\hat{\gamma}}_k\right)^{\mathrm{T}}\right] \\
&= \mathrm{E}\left[A\left(\underline{\gamma}_k - \underline{\hat{\gamma}}_k\right)\left(\underline{\gamma}_k - \underline{\hat{\gamma}}_k\right)^{\mathrm{T}}A^{\mathrm{T}} + \underline{\omega}_k\left(\underline{\gamma}_k - \underline{\hat{\gamma}}_k\right)A^{\mathrm{T}}\right. \\
&\quad \left. + A\left(\underline{\gamma}_k - \underline{\hat{\gamma}}_k\right)\underline{\omega}^{\mathrm{T}}_k + \underline{\omega}\underline{\omega}^{\mathrm{T}}_k\right] \\
&= A\mathrm{COV}\left(\underline{\gamma}_k\right)A^{\mathrm{T}} + \mathrm{E}\left[\underline{\omega}_k\right]\mathrm{E}\left[\left(\underline{\gamma}_k - \underline{\hat{\gamma}}_k\right)\right]A^{\mathrm{T}} \\
&\quad + A\mathrm{E}\left[\left(\underline{\gamma}_k - \underline{\hat{\gamma}}_k\right)\right]\mathrm{E}\left[\underline{\omega}^{\mathrm{T}}_k\right] + \mathrm{COV}\left(\underline{\omega}_k\right) \\
&= AP_kA^{\mathrm{T}} + Q_k
\end{aligned} \tag{4.18}$$

where the independency of $\underline{\omega}_k$ and $\underline{\gamma}_k$ is used, as well as $\mathrm{E}\left[\underline{\omega}_k\right] = 0$ and $\mathrm{COV}\left(\underline{\omega}_k\right) = Q_k$.

4.4.4 Analysis step

In the analyzing step the results of the forecast step are analyzed with outcomes of a series of measurements. The mean from the forecast step is analyzed with a linear Kalman gain K, such that the mean after the analyzing step is similar with the one dimensional case.

$$\underline{\hat{\gamma}}^a_{k+1} = \underline{\hat{\gamma}}^f_{k+1} + K_{k+1}\left(\underline{\ln\left(y\right)}_{k+1} - H\left(\underline{\ln\left(c^m\right)}_{k+1} + \underline{\hat{\gamma}}^f_{k+1}\right)\right) \tag{4.19}$$

The covariance matrix after the analyzing step is as in the one dimensional case a function of the covariance matrix from the forecast step.

$$
\begin{aligned}
P^a_{k+1} &= \mathrm{COV}\left(\underline{\gamma}_{k+1}\right) \\
&= \mathrm{E}\left[\left(\underline{\gamma}_{k+1} - \mathrm{E}\left[\underline{\gamma}_{k+1}\right]\right)\left(\underline{\gamma}_{k+1} - \mathrm{E}\left[\underline{\gamma}_{k+1}\right]\right)^{\mathrm{T}}\right] \\
&= \mathrm{E}[\left(\underline{\gamma}_{k+1} - \underline{\hat{\gamma}^a}_{k+1}\right)\left(\underline{\gamma}_{k+1} - \underline{\hat{\gamma}^a}_{k+1}\right)^{\mathrm{T}}] \\
&= \mathrm{E}\left[\left(\underline{\gamma}_{k+1} - \left(\underline{\hat{\gamma}^f}_{k+1} + K_{k+1}\left(\underline{\ln(y)}_{k+1} - H\left(\underline{\ln(c^m)}_{k+1} + \underline{\hat{\gamma}^f}_{k+1}\right)\right)\right)\right)\right. \\
&\quad \left.\left(\underline{\gamma}_{k+1} - \left(\underline{\hat{\gamma}^f}_{k+1} + K_{k+1}\left(\underline{\ln(y)}_{k+1} - H\left(\underline{\ln(c^m)}_{k+1} + \underline{\hat{\gamma}^f}_{k+1}\right)\right)\right)\right)^{\mathrm{T}}\right] \\
&= \mathrm{E}\left[\left(\underline{\gamma}_{k+1} - \underline{\hat{\gamma}^f}_{k+1} - K_{k+1}\left(H\left(\underline{\ln(c^m)}_{k+1} + \underline{\gamma}_{k+1}\right) + \underline{\nu}_{k+1}\right.\right.\right. \\
&\quad \left.\left.\left. - H\left(\underline{\ln(c^m)}_{k+1} + \underline{\hat{\gamma}^f}_{k+1}\right)\right)\right) \\
&\quad \left(\underline{\gamma}_{k+1} - \underline{\hat{\gamma}^f}_{k+1} - K_{k+1}\left(H\left(\underline{\ln(c^m)}_{k+1} + \underline{\gamma}_{k+1}\right) + \underline{\nu}_{k+1}\right.\right. \\
&\quad \left.\left.\left. - H\left(\underline{\ln(c^m)}_{k+1} + \underline{\hat{\gamma}^f}_{k+1}\right)\right)\right)^{\mathrm{T}}\Bigg] \\
&= \mathrm{E}\left[\left((I - K_{k+1}H)\left(\underline{\gamma}_{k+1} - \underline{\hat{\gamma}^f}_{k+1}\right) - K_{k+1}\underline{\nu}_{k+1}\right)\right. \\
&\quad \left.\left((I - K_{k+1}H)\left(\underline{\gamma}_{k+1} - \underline{\hat{\gamma}^f}_{k+1}\right) - K_{k+1}\underline{\nu}_{k+1}\right)^{\mathrm{T}}\right] \\
&= \mathrm{E}\left[(I - K_{k+1}H)\left(\underline{\gamma}_{k+1} - \underline{\hat{\gamma}^f}_{k+1}\right)\left(\underline{\gamma}_{k+1} - \underline{\hat{\gamma}^f}_{k+1}\right)^{\mathrm{T}}(I - K_{k+1}H)^{\mathrm{T}}\right] \\
&\quad -\mathrm{E}\left[(I - K_{k+1}H)\left(\underline{\gamma}_{k+1} - \underline{\hat{\gamma}^f}_{k+1}\right)\underline{\nu}^{\mathrm{T}}_{k+1}K^{\mathrm{T}}_{k+1}\right] \\
&\quad -\mathrm{E}\left[K_{k+1}\underline{\nu}_{k+1}\left(\underline{\gamma}_{k+1} - \underline{\hat{\gamma}^f}_{k+1}\right)^{\mathrm{T}}(I - K_{k+1}H)\right] \\
&\quad +\mathrm{E}\left[K_{k+1}\underline{\nu}_{k+1}\underline{\nu}^{\mathrm{T}}_{k+1}K^{\mathrm{T}}_{k+1}\right] \\
&= (I - K_{k+1}H)\,\mathrm{E}\left[\left(\underline{\gamma}_{k+1} - \mathrm{E}\left[\underline{\gamma}_{k+1}\right]\right)\left(\underline{\gamma}_{k+1} - \mathrm{E}\left[\underline{\gamma}_{k+1}\right]\right)^{\mathrm{T}}\right](I - K_{k+1}H)^{\mathrm{T}} \\
&\quad +K_{k+1}\mathrm{E}\left[\underline{\nu}_{\mathrm{k}+1}\underline{\nu}^{\mathrm{T}}_{\mathrm{k}+1}\right]K^{\mathrm{T}}_{k+1} \\
&= (I - K_{k+1}H)\,\mathrm{COV}\left(\underline{\gamma}_{k+1}\right)(I - K_{k+1}H)^{\mathrm{T}} + K_{k+1}\mathrm{COV}\left(\underline{\nu}_{k+1}\right)K^{\mathrm{T}}_{k+1} \\
&= (I - K_{k+1}H)\,P^f_{k+1}(I - K_{k+1}H)^{\mathrm{T}} + K_{k+1}R_{k+1}K^{\mathrm{T}}_{k+1} \qquad (4.20)
\end{aligned}
$$

where the independency of $\underline{\gamma}_k$ and $\underline{\nu}_k$, as well as $\mathrm{E}[\nu_k] = 0$ and $\mathrm{COV}(\underline{\nu}_k) = R_k$

Also in this higher dimensional problem it is a common use to take for K_k the gain that minimizes the variance P^a_k in l_2 norm. This gain is expressed similar tot the one dimensional case

$$K_k = P^f_k H^{\mathrm{T}}\left(HP^f_k H^{\mathrm{T}} + R_k\right)^{-1} \qquad (4.21)$$

More information about higher dimensional Kalman filtering can be found in (Segers, 2002)

5 Kalman Filter on Background Concentrations

5.1 Introduction

In the Real Time URBIS model, a simulation is made for the mean concentration NO_x per hour. This results in a model simulation for the concentration NO_x, called $\underline{c}_k^m$. These simulation is made with the aid of the underlying URBIS model by the following formula:

$$\underline{c}_k^m = M\underline{u}_k \tag{5.1}$$

where each column of M corresponds with a standard concentration field, computed by the URBIS model, shown in Appendix A. The vector $\underline{u}_k$ represents the weight for every standard concentration field.

The figures in Appendix A shows that the standard concentration fields for the emission source 'background from the rest of the Netherlands' are the same for every wind direction and wind speed. This is contradicting with the ideas of Chapter 3. In that chapter, it was shown that the differences between the model simulations and the observations depend on the wind direction and the wind speed. This analysis is done for all stations, thus it is assumed that the dependency on the wind direction is the same for the whole area. It is possible that the dependency on the wind direction is caused by inaccurate local emission source, but it is not likely that an inaccurate local emission source influences all stations.

Therefore the emission from source 'Rest of the Netherlands' is marked as the inaccurate emission source. This source is typically a source that has to be dependent of the wind direction and the wind speed. It is likely that the wind dependency found in Chapter 3, is caused by the lack of wind dependency in the source 'background from the rest of the Netherlands' in the URBIS model. In this chapter the standard concentration fields for this emission source will be recalculated with a Kalman filter. The idea is that the new concentration fields have different patterns for each of the four wind-directions and two wind-speeds.

Figure 3.2 gives an idea of how the standard concentration fields have to be changed. When the wind is from direction north-west the model simulation is too high, thus the concentration fields from directions west and north have to be lower. When the wind is from direction south-east, the model simulation is too low, thus the standard concentration field from directions east and south has to be higher. Figure 3.5 gives the idea is that when the wind speed is high, the concentration is lower because of a larger dilution of the emission.

This application of the Kalman filter will also lead to an uncertainty interval of the total concentration NO_x for every time and location in the Rijnmond area. The total

concentration is only changed by a correction on the background. Because of the nearly constant background on the whole area, the change on the total concentration is nearly the same on every location. Also the width of the uncertainty interval is nearly the same for every location.

5.2 Kalman filter

To make a correction on the standard concentration fields, every field gets a correction factor e^{γ_i}. These factors are larger then zero, thus there is no problem with negative concentrations. Adding these corrections to the model from Equation 5.1 leads to the following equation for the corrected model:

$$\underline{c_k} = \sum_{i=1}^{88} \mu_{i,k} \underline{m_i} e^{\gamma_i} \tag{5.2}$$

In this equation vectors $\underline{m_i}$ are the columns of M, representing the standard concentration fields. In the log-normal distribution the expected concentration $\mathrm{E}[\underline{c_k}]$ is given by:

$$\mathrm{E}[\underline{c_k}] = \sum_{i=1}^{88} \mu_{i,k} \underline{m_i} e^{\underline{\hat{\gamma}} + 1/2 \underline{sd}^2} \tag{5.3}$$

where $\underline{\hat{\gamma}}$ is the median of $\underline{\gamma}$ and $\underline{sd}$ is the standard deviation of each entry of $\underline{\gamma}$.

5.2.1 Dynamical system

In Chapter 3, the idea is that the background concentrations are not accurate in the model. The other fields were supposed to be good enough, thus the correction factor on those fields are stated equal to one ($\gamma = 0$). This leads to the following equation:

$$\sum_{i=1}^{8} \mu_{i,k} \underline{m_i} e^{\gamma_i} + \sum_{i=9}^{88} \mu_{i,k} \underline{m_i} e^{0} \tag{5.4}$$

The vectors m_i for $i = 1..8$ corresponds with the standard concentration fields for the source: 'background from the rest of the Netherlands', these fields have a correction e^{γ_i}. The second term of Equation 5.4 is not dependent of any γ_i, thus a constant called $\underline{c}_k^{m,d}$ is introduced to describe the concentration calculated by the model for all sources, different from the source 'background from the rest of the Netherlands':

$$\underline{c}_k^{m,d} = \sum_{i=9}^{88} \mu_{i,k} \underline{m_i} \tag{5.5}$$

Because of the log-normal distribution of the model simulations, a transformation to the logarithms of the simulations is required:

$$\ln\left(\underline{c}_k^m\right) = \ln\left(\sum_{i=1}^{8}\left(\mu_{i,k}\underline{m}_i e^{\gamma_i}\right) + \underline{c}_k^{m,d}\right) \tag{5.6}$$

This is a non-linear equation for $\underline{\gamma}$. The Kalman filter requires a linear model, therefore a linearization of this equation is made around $\underline{\gamma} = \underline{0}$:

$$\begin{aligned}\ln\left(\sum_{i=1}^{8}\left(\mu_{i,k}\underline{m}_i e^{\gamma_i}\right) + \underline{c}_k^{m,d}\right) &= \ln\left(\underline{c}_k^{m,b} + \underline{c}_k^{m,d}\right) \\ &+ \left[\frac{\mu_{j,k}\underline{m}_j}{\underline{c}_k^{m,b} + \underline{c}_k^{m,d}}\right]_{j=1}^{j=8}\underline{\gamma} + \mathcal{O}\left(\underline{\gamma}\cdot\underline{\gamma}\right)\end{aligned} \tag{5.7}$$

where $\underline{c}_k^{m,b}$ is the concentration calculated by the model for the source: 'background from the rest of the Netherlands':

$$\underline{c}_k^{m,b} = \sum_{i=1}^{8}\mu_{i,k}\underline{m}_i \tag{5.8}$$

In Equation 5.7, the quotient of two vectors is defined as the element wise quotient.

The dynamical system for the background concentration will then become the following.

$$\ln\left(\underline{c}_k\right) = \ln\left(\underline{c}_k^{m,b} + \underline{c}_k^{m,d}\right) + \left[\frac{\mu_{j,k}\underline{m}_j}{\underline{c}_k^{m,b} + \underline{c}_k^{m,d}}\right]_{j=1}^{j=8}\underline{\gamma}_k \tag{5.9}$$

$$\underline{\gamma}_{k+1} = A\underline{\gamma}_k + \underline{\omega}_k \qquad \underline{\omega}_k \sim N\left(0, Q_k\right) \tag{5.10}$$

The first equation is the linearization of the equation for the logarithm of the concentration, the second equation is the auto correlation process for the series of perturbations $\left\{\underline{\gamma}_k\right\}_{k=1}^{k=n}$, with $n = 8760$, the number of hours in a year. In the Kalman filter, an estimate of the uncertainty interval of the vector $\underline{\gamma}_k$ will be found, this uncertainty interval of $\underline{\gamma}_k$ will then be use to get a better uncertainty interval for the total concentration at time k.

The interpretation of the dynamical system is now that the logarithm of the real concentration is the logarithm of the model simulation plus a correction on the background. The correction on the background is a time correlated process, the temporal correlation is calculated in Section 5.4. The covariance matrix Q_k is built from the temporal correlation and the model uncertainty. Matrix Q_k is a diagonal matrix with

on the main diagonal elements q_i^2. This is a colored noised process driven by a white noise process, assuming that both the temporal correlation and the uncertainty of the model is independent of time.

$$q_i = \sqrt{1-\alpha_i^2}\sigma_i \tag{5.11}$$

where σ_i corresponds with the overall uncertainty of the perturbations.

5.2.2 *Kalman filter form*

The dynamical system in Equations 5.9 and 5.10 has to be written in a Kalman filter form. There are 9 series of measurements $\underline{y}$, which are made on the 9 measurement stations in the domain. This series of measurements have to be compared with the model results.

This leads to the following system of equations in Kalman filter form.

$$\underline{\gamma}_{k+1} = A\underline{\gamma}_k + \underline{\omega}_k \tag{5.12}$$

$$\ln\left(\underline{y}_k\right) = H\left(\ln\left(\underline{c}_k^{m,b} + \underline{c}_k^{m,d}\right) + \left[\frac{\mu_{j,k}\underline{m}_j}{\underline{c}_k^{m,b} + \underline{c}_k^{m,d}}\right]_{j=1}^{j=8} \underline{\gamma}_k\right) + \underline{\nu}_k \qquad \underline{\nu}_k \sim N\left(0, R_k\right) \tag{5.13}$$

Matrix H is the system operator which projects the model state onto the observations. The covariance matrix R represents the uncertainty of the logarithms of the measurements, combined the instrumental error and the representation error. This matrix R is a diagonal matrix with diagonal elements r_i^2, the values for r_i will be estimated in Section 5.3. To simplify notations, the system is rewritten to:

$$\underline{\gamma}_{k+1} = A\underline{\gamma}_k + \underline{\omega}_k \tag{5.14}$$

$$\underline{\tilde{y}}_k = \tilde{H}_k\underline{\gamma}_k + \underline{\nu}_k \qquad \underline{\nu}_k \sim N\left(0, R_k\right) \tag{5.15}$$

where vector $\underline{\tilde{y}}_k$ and matrix $\tilde{H}_k$ are defined by:

$$\underline{\tilde{y}}_k = \ln\left(\underline{y}_k\right) - H\ln\left(\underline{c}_k^{m,b} + \underline{c}_k^{m,d}\right) \tag{5.16}$$

$$\tilde{H}_k = H\left[\frac{\mu_{j,k}\underline{m}_j}{\underline{c}_k^{m,b} + \underline{c}_k^{m,d}}\right]_{j=1}^{j=8} \tag{5.17}$$

5.2.3 *Forecast of background correction*

On this Kalman filter form the algorithm for the Kalman filter can be applied. The forecast step gives then the following formulas for the expected median $\underline{\hat{\gamma}}_k^f$ and the variance P_k^f of $\underline{\gamma}_k$.

$$\underline{\hat{\gamma}}^f_{k+1} = A\underline{\hat{\gamma}}_k \tag{5.18}$$

$$P^f_{k+1} = APA^{\mathrm{T}} + Q_k \tag{5.19}$$

5.2.4 Analysis of background correction

In the analyzing step, the filter makes a comparison with a series of measurements, in this case 9 measurements per time step for the 9 measurement stations in the domain. This leads to the following formulas for the expected median $\underline{\hat{\gamma}}^a_k$and variance P^a_k of $\underline{\gamma}_k$:

$$\underline{\hat{\gamma}}^a_{k+1} = \underline{\hat{\gamma}}^f_{k+1} + K_{k+1}\left(\underline{\tilde{y}}_{k+1} - \tilde{H}_{k+1}\underline{\hat{\gamma}}^f_{k+1}\right) \tag{5.20}$$

$$\begin{aligned} P^a_{k+1} &= \left(I - K_{k+1}\tilde{H}_{k+1}\right)\left(P^f_{k+1}\right)\left(I - K_{k+1}\tilde{H}_{k+1}\right)^{\mathrm{T}} \\ &+ K_{k+1}R_{k+1}K^{\mathrm{T}}_{k+1} \end{aligned} \tag{5.21}$$

where K_{k+1} is the Kalman gain that minimizes the variance P^a_{k+1}.

$$K_k = P^f_k\tilde{H}^{\mathrm{T}}_k\left(\tilde{H}_kP^f_k\tilde{H}^{\mathrm{T}}_k + R_k\right)^{-1} \tag{5.22}$$

The values $\underline{\hat{\gamma}}^a$ and P^a_k, are stated as the median and the covariance matrix for $\underline{\gamma}$ on time step k, and will be used as input for the next time step.

5.3 Uncertainty of the observations

The observation error (R in Equation 5.13) is an important parameter in the Kalman filter. Section 4.3 shows the influence on the solution when parameter r^2 is changed. Because of R is built from all r^2_i, the observation errors of each entry of the observation, the influence of covariance matrix R is also large.

The uncertainty of the measurements is assumed the square of a percentage (r_{frac}) of the outcome of the measurement:

$$R_{ii,k} = r^2_{frac}y^2_{i,k}$$

where r_{frac} will contain both the instrumental error and the representation error.

At location Bentinckplein in Rotterdam, there are two measurement stations located directly next to each other, one measurement station from DCMR and one from RIVM. With these two series of measurements an indication of the instrumental error can be found. In Figure 5.1, the logarithms of the observations on the two measurement stations at Bentinckplein are shown in a scatter plot. An assumption for the

logarithm of the real concentration at this location is the mean of the logarithms of the two observations.

$$\ln(y_k^r) = \frac{\ln(y_k) + \ln(z_k)}{2} \tag{5.23}$$

where y_k^r is the real concentration at time k and y_k, z_k are respectively the measured concentrations on the DCMR and the RIVM station.

In Figure 5.2, a histogram with differences between the logarithms of the observations at the DCMR station and the assumed logarithms of the real concentrations is shown. The red line is the probability density function of the normal distribution with mean 0 and standard deviation 0.08, this standard deviation is the same as the standard deviation of the differences plotted in the histogram. The peak of the histogram is not located on zero, which means that the annual mean concentration is not the same on both stations. The annual mean on the RIVM station is larger than the annual mean on the DCMR station. Although the normal distribution did not fit very well with the histogram, the assumption that the differences are normal distributed with standard deviation 0.08 is at least a good approximation. This will be used as an estimate for the instrumental error in r_{frac}, a random noise on the observation. The histogram for the RIVM station is the same as for the DCMR station, but then the negative version so that the histogram is mirrored in the y-axis.

This contribution of the representations error is not easy to calculate, this will be done by a method of trial and error. The Kalman filter will be run applicated with different values for $r_{frac} > 0.08$ to obtain the optimal value for r_{frac}.

The last assumption is that the observation error is the same for all stations, and not correlated between the stations. The matrix R is then a diagonal matrix with on the main diagonal elements $0.08y_j^2$. This corresponds to an uncertainty for the logarithm of the observation of 8%.

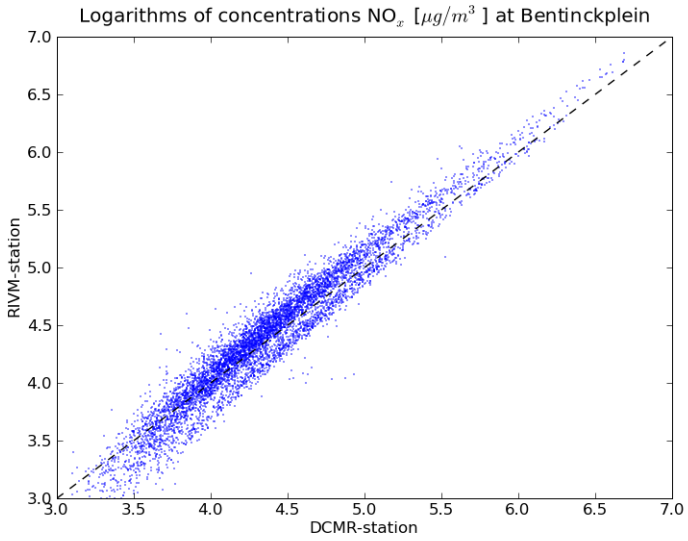


Figure 5.1: *The logarithms of the observations of the two monitoring stations at location Bentinckplein*

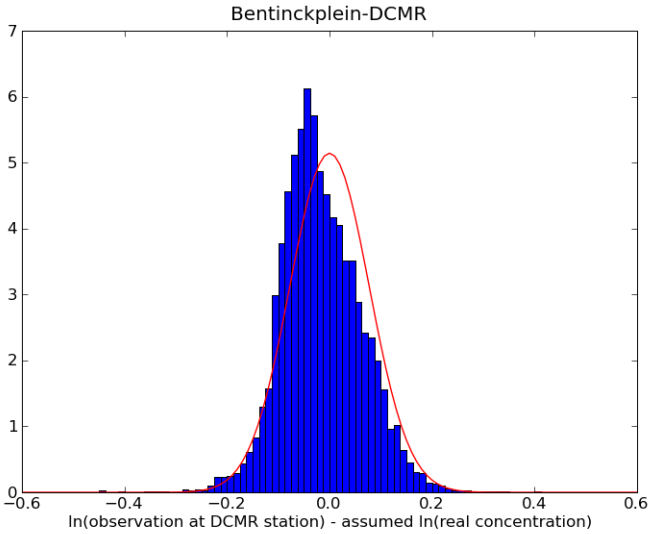


Figure 5.2: *Histogram of the differences between the logarithms of the observations and the assumed logarithm of the real concentration at location Bentinckplein. The red line is the probability density function of the normal distribution with mean zero and standard deviation 0.08.*

5.4 Time correlation parameter

Another important parameter is the temporal correlation. In the dynamical system given by equations 5.9 and 5.10, the matrix A contains the temporal correlation parameters $\alpha_{i,j}$ for the perturbation on the logarithm of the several background concentrations. The measurement locations Schipluiden and Westmaas, numbers 7 and 10 in Figure 2.1, are two locations, in a region far away from industry sources or main roads. These measurement sites are chosen to obtain estimates of the background concentrations. With the observations on this locations it is possible to get an estimate for the time correlation parameters.

In general, the correlation between two series of measurements $\{y_i\}_{i=1}^n$and $\{z_i\}_{i=1}^n$ could be computed with the following formula:

$$corr \quad = \quad \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i-\bar{y})}{\sigma_y}\frac{(z_i-\bar{z})}{\sigma_z} \tag{5.24}$$

where $\bar{y}$, $\bar{z}$ are the mean of the series $\{y_i\}_{i=1}^n$and $\{z_i\}_{i=1}^n$, and σ_y, σ_z are the standard deviations of the series $\{y_i\}_{i=1}^n$and $\{z_i\}_{i=1}^n$.

Assumed is that there is no correlation between the perturbations from different wind directions and wind speeds. The matrix A will then be a diagonal matrix with on the main diagonal elements α_i. An estimation for α_i is made with Equation 5.24 from two series of measurements $\{z_k\}_{k=1}^{n-m}$ and $\{z_k\}_{k=m}^n$, where z_k is the difference between the logarithm of the observation and the logarithm of the model simulation at time step k on location Schipluiden.

Another estimate for α_i is made with the differences on location Westmaas. For both locations this is done for $m \in [0, 60]$.

$$z_k = \ln(y_k) - \ln(c_k^m) \tag{5.25}$$

Both locations Westmaas and Schipluiden are outside the model domain, thus there is no model simulation. Because of both stations are assumed to be background stations, the concentration is caused mainly by the background. This background is assumed constant, therefore the calculation of the correlation could be done with $z_k = \ln(y_k)$.

For every period, the correlation is calculated for the perturbation on time k compared with the perturbation on time $k+m$. In Figure 5.3 these correlations are plotted with respect to the period m. This figure shows some peaks at period 24 hours, and period 48 hours. This means that the correlation has a daily pattern. This is a reasonable idea, because it is expected that the emission in Schipluiden and Westmaas is mostly produced by people living in Schipluiden and Westmaas.

A reasonable assumption is that the concentration on time step k does not depend on the concentration on time $k - m$ when m is large. Therefore the correlation between the perturbation on time k and the perturbation on time $k+m$ should go to zero when $m \to \infty$. Mathematically there is a correlation between the concentration on time step k and time step $k+m$, this can be understood by the time patterns in the emission and diurnal cycles in meteorological parameters. For example the concentration on Monday at 08:00 in the morning is roughly the same as the concentration at Tuesday 08:00 in the morning. Mathematically this gives a high temporal correlation for $\Delta t = 24$, but physically this concentrations are not realted. For that reason it is only important to look at the temporal correlation for a few hours.

In Figure 5.3 is a fitting exponential function drawn for the first few periods. In this case the formula for this function is $\alpha(\Delta t) = e^{-\Delta t/12}$. The de-correlation parameter $\tau = 12$ gives the idea that the concentration on time $k + 12$ is not dependent on the concentration at time k.

At last this de-correlation parameter $\tau = 12$ must be seen as an estimate, this is the temporal correlation for the concentration with varying wind speeds and wind directions. When the wind is with constant speed from the same direction, the correlation is perhaps different. In the application an optimal value for each α_i will be found by testing the Kalman filter with different values for each α_i.

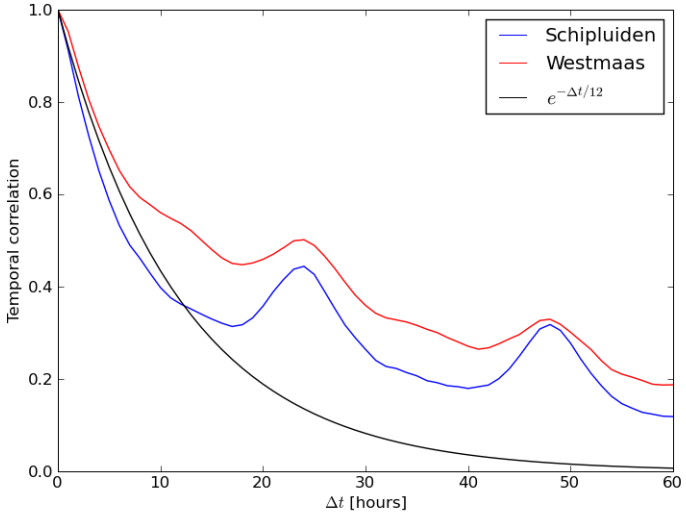


Figure 5.3: *The temporal correlation for background stations Schipluiden and Westmaas. The black line corresponds with correlation* $\alpha(\Delta t) = e^{-\Delta t/12}$

5.5 Kalman filter runs

The application of the Kalman filter, assuming the model uncertainty σ equal to 0.4, leads to a calculation of $\underline{\hat{\gamma}}_k$, the expected mediaan of vector $\underline{\gamma}_k$ and P_k the covariance matrix of vector $\underline{\gamma}_k$ at every time step k. In this application the vector $\underline{\gamma}$ represents the perturbations on the logarithms of the background concentration for four different wind directions and two different wind speeds. For the first week of 2006, the 1σ intervals of these eight perturbations are given in Figure 5.4. On every time step the weight $\mu_{i,k}$ for a standard concentration field is different, the values for $\mu_{i,k}$ are also given in Figure 5.4. When the contribution of a standard concentration field is high, the change in the correction factor γ_i is also high. If for a longer period a standard concentration field has no contribution, the mean of the correction factor γ_i tends to zero.

An interesting aspect of this result is that the values for γ_i are relatively high for some time steps, this means that the background concentration receives a relatively high correction factor for that time step. This is due to the fact that in this application it is assumed that the difference between the observation and the model simulation is completely depending on the background concentration. A better assumption is that when the difference between the observation and the model simulation is large, that there are some other errors in the model.

Another aspect is that the linearization of the dynamical system around $\underline{\gamma} = \underline{0}$ has accuracy $\mathcal{O}\left(\underline{\gamma} \cdot \underline{\gamma}\right)$, when $\underline{\gamma}$ become large the accuracy of the linearization decreases quadratically. For those reasons a screening process is implemented in the Kalman filter. When the difference between the observation and the model simulation is too high the analysis step will not be executed. The result is that the values for γ_i are

limited. This screening process is explained in Section 5.6.

In Figure 5.5 the problems with large values for γ_i are shown. In this figure, at every time step the concentrations are calculated with the values for $\underline{\gamma}_i$ and with Equations 5.2 and 5.4. In each figure the yellow line represents the largest contribution on the correction of the background $\max\left(\mu_{i,k}e^{\gamma_{i,k}}\right)$ for every times step k. In this figures it is clear that the concentrations after applying the Kalman filter are not accurate in the regions where the values for γ_i became large.

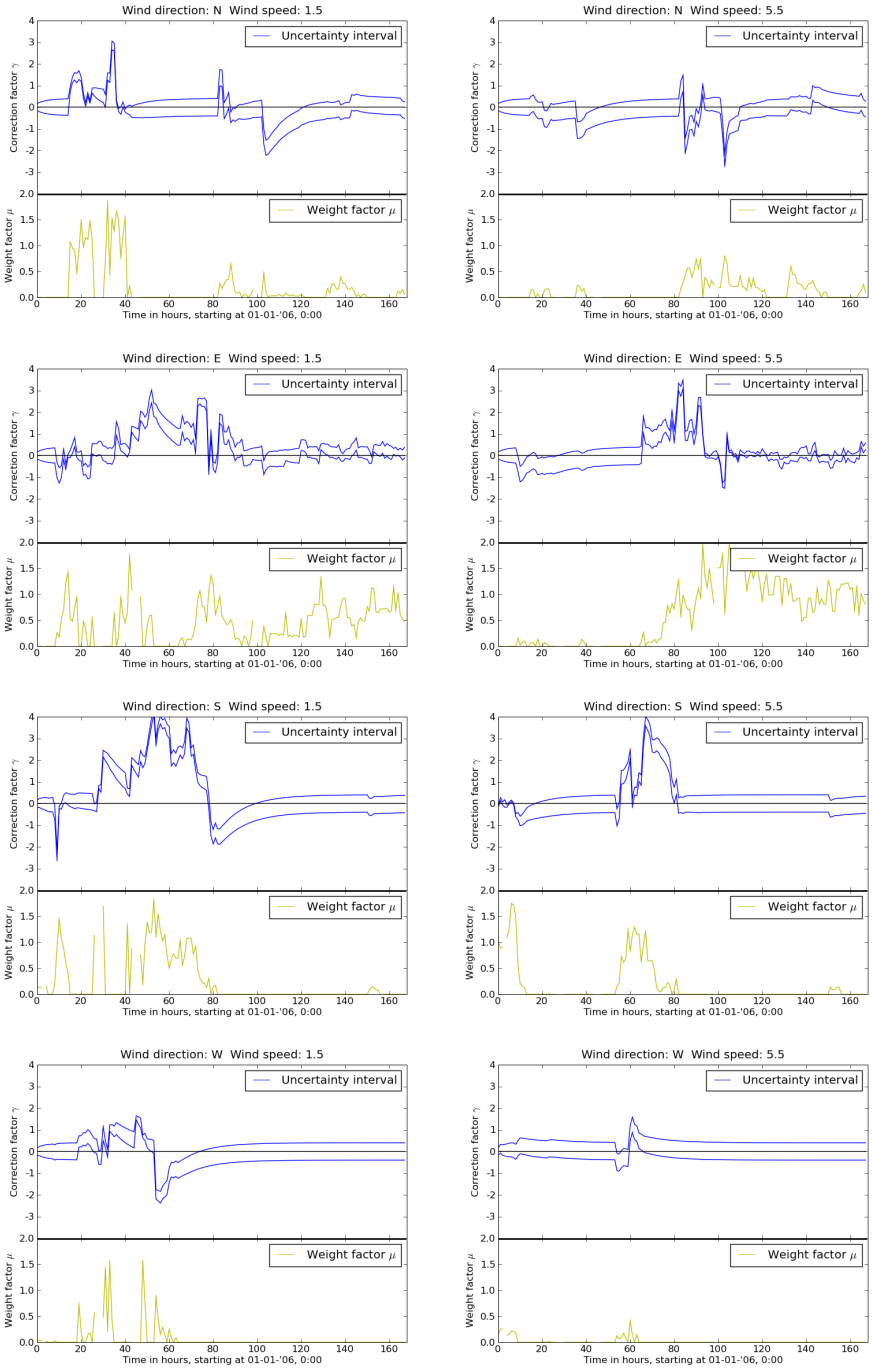


Figure 5.4: *Uncertainty intervals for the correction factors γ_i, together with the weights for each standard concentration field, for the first week of 2006*

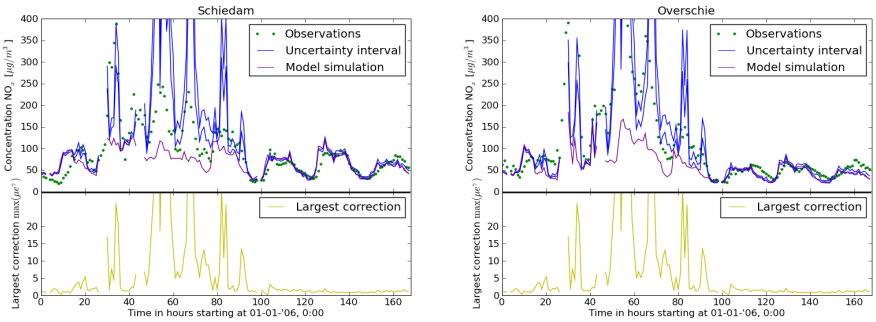


Figure 5.5: *Concentrations for the first week of 2006 after application of the Kalman filter on the background concentrations, for locations Schiedam and Overschie.*

5.6 Screening process

As mentioned in Section 5.5, for some time steps the difference between the observation and the model simulation could not be only explained by inaccuracies of the background concentrations. For that reason, a screening process is implemented in the Kalman filter. When the difference between the observation and the forecasted concentration is too high, the difference between the observation and the model simulation is not only caused by the inaccurate background, but also by some other sources or incidental occasions. If this situation occurs, the analysis step will only be executed on the observations which are close to the forecasted mean, such that the values for γ_i will stay small. The result is that the background concentrations will not get large correction factors and the linearization still have good accuracy. It is also important to have a view on which observations are screened, this could give an idea of other inaccuracies in the model. For example, if many observations are screened in the weekend, the model have large uncertainty in the weekend.

To implement a screening process, a criterion has to be made, whether a difference between an observation and a model simulation is too high. For the Kalman filter on the background concentration the assumption is the following:

$$\underline{y}_k - \left(\sum_{i=1}^{8} \underline{m}_i \mu_{i,k} e^{\gamma^f_{i,k} + 1/2 (sd^f_{i,k})^2} + c^{m,d}_k \right) \quad \sim \quad N\left(0, P_{abs,k} + R_{abs,k}\right) \tag{5.26}$$

In here, $P_{abs,k}$ and $R_{abs,k}$ represents the variance for respectively the model simulations and the observations. The variance for the model is not known explicitely, due to the log-normal distribution. Therefore this variance is assumed to be equal to the square of the difference between the upper band and the median of the 1σ interval of the concentration. The variance of the measurements corresponds with an uncertainty of the measurements coupled with the uncertainty of the logarithm of the

observation, $r_{frac} > 0.08$.

$$P_{abs,k} = \left(\sum_{i=1}^{8} \underline{m}_i \mu_i e^{\gamma_{i,k}^f} + c_k^{m,d} - \left(\sum_{i=1}^{8} \underline{m}_i \mu_{i,k} e^{\gamma_{i,k}^f + \lambda_{i.k}^f} + c_k^{m,d} \right) \right)^2 \tag{5.27}$$

$$R_{abs,k} = (r_{frac} y_k)^2 \tag{5.28}$$

where $\lambda_{i,k}^f$ is the standard deviation of $\gamma_{i,k}^f$. With the assumption from Equation 5.26 a criterion is chosen whether an observation is 'good' enough:

$$\left(\underline{y}_k - \left(\sum_{i=1}^{8} \underline{m}_i \mu_{i,k} e^{\gamma_{i,k}^f + 1/2 (sd_{i,k}^f)^2} + c_k^{m,d} \right) \right)^2 < \beta^2 \left(P_{abs,k} + R_{abs,k} \right) \tag{5.29}$$

The parameter β defines the screening criterion. If the square of a difference is more than β^2 times the variance of model simulation plus the variance of the observations, the observation does not fit on the assumption that the difference is only caused by the inaccuracy of the background. In that case the observation is not involved in the analyzing process. This is a vector inequality, which means that when for an entry of both vectors this inequality holds, the observation corresponding with that entry is not involved in the analyzing step.

This screening process is implemented in the Kalman filter with parameter $\beta = 2$, this means that the square of a difference may not be larger then 4 times the sum of variations. The value $\beta = 2$ is chosen because in the normal distribution approximately 95% of the data lies in the 2σ interval.

The application of the Kalman filter with this screening process results in concentrations for Schiedam and Overschie for the first week of 2006, as shown in Figure 5.6. The largest correction $\max(\mu_{i,k} e^{\gamma_{i,k}})$ became much smaller, and the concentrations have less extremes. In these figures, it is also shown that a lot of observations are not taken into the analysis step, about 68% of the observations are not taken into the analysis step. A possibility is that the temporal correlation is too large, a large temporal correlation in the Kalman filter leads to a result without large fluctuations in the concentration. When the observations have large fluctuations, it is possible that many observations will be screened. If the temperal correlation is set with decorrelation parameter $\tau = 1$, there are still 66% of the observations are not taken into the analysis step.

Another idea is that the differences are not completely caused by the inaccuracy of the background concentrations. In Chapter 6 the Kalman filter is applied on all the different emission sources, to get a better estimate of all the different concentration fields of the URBIS model. The idea is that the large differences between the observations and the simulations are caused by inaccuracies of one of the standard concentration fields.

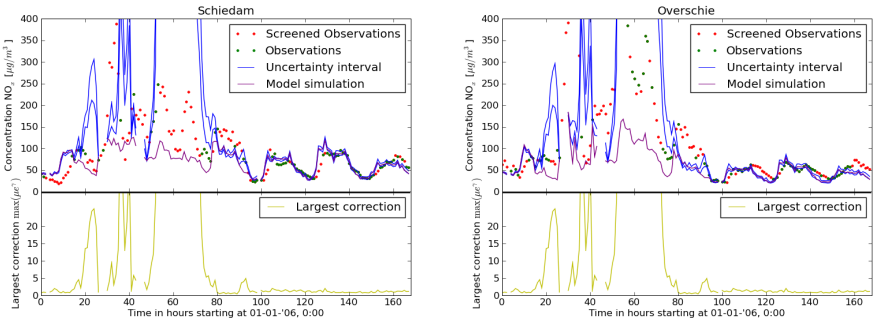


Figure 5.6: *Concentrations for the first week of 2006 after application of the Kalman filter on the background concentrations, with a screening process, at locations Schiedam and Overschie*

5.7 Discussion

The ideas from Chapter 3 were that the differences between the model simulations and the observations are caused by inaccuracies of the background. In this chapter, it is shown that this assumption does not hold for most of the differences. A correction on the background is not sufficient to eliminate the difference between the model simulation and the observation from a measurement location.

With the corrections made on the background, it is possible to create better standard concentration fields. Because of the large number of measurements which are not involved in the Kalman filter process, it is not expected that the new standard concentration fields for the background will be very accurate. For that reason, the Kalman filter will be applied to all different emission sources to get a 'better' standard concentration field for every source. This application will be explained in Chapter 6, together with the different runs to obtain the optimal values for each α_i, σ and R.

6 Kalman filter on all Concentration sources

6.1 Introduction

In Chapter 5, it was shown that a correction on the background concentrations leads to a better estimate of the real concentrations for only a small number of time steps. For the other time steps the Kalman filter did not give a correction on the background because the difference between the observation and the model simulation was not only caused by the inaccurate background.

Therefore an additional analysis on the differences between the observations and the model simulations is made. In Figure 6.1 is shown in which cases the differences between the observations and the model simulations on location Schiedam are relatively large. The red bar plots gives the percentage of the situations where the observation is more than 2 times the model simulation. The blue bar plots gives the total number of differences that occurs for every input parameter (wind direction, wind speed, temperature, hour of the day, day of the week and month of the year).

For the wind direction, a high percentage of the differences is relatively large when the wind is from the south-east, but the total number of wind directions from the south-east is not very high. Thus it is assumed that the contribution to the total inaccuracy is not very large. For the wind speed, a high percentage of the differences is relatively large when the wind speed is below 2 $^m/_s$. Also the total number of times that the wind speed is above 2 $^m/_s$, is relatively large. This suggests that the inaccuracies in the model when the wind speed is low, have a large contribution to the total inaccuracy.

Another notable parameter is the hour of the day, in the morning and the end of the evening there are relatively many large differences. This is an indication that there are some inaccuracies in the sources which are time dependent (traffic and emission produced by the residents of the Rijnmond area). The last interesting parameters are the day of the week and the month of the year. Only on Sunday there are not very much large differences, in the autumn and the winter are relatively many large differences. This is also an indication the time dependent sources have inaccuracies. In section 2.2 of Report (Kranenburg, 2009) was already mentioned that the sources industry and shipping could have a time dependency. Thus also the sources shipping and industry may have inaccuracies

The idea in this chapter is that the uncertainty of the model is caused by several different emission sources, therefore the Kalman filter will be applied on all the different sources. With this application all the standard concentration fields for all emission sources will be estimated. These estimates are again calculated by multiplying each field with a correction factor, which leads to a corrected state equation:

$$\underline{c}_k = \sum_{i=1}^{88} \mu_{i,k} \underline{m}_i e^{\gamma_i} \tag{6.1}$$

In this chapter it is no longer assumed that some of the entries of $\underline{\gamma}$ are equal to zero. The Kalman filter process will estimate all values of $\underline{\gamma}$ by a comparison of the model with the measurements.

An advantage of this application is that the uncertainty intervals for the total concentration is a combination of uncertainty intervals for the different emission sources. This leads to an uncertainty interval which is different for every location. For example, on locations where the concentration is mostly caused by emission from traffic, the uncertainty interval is approximately equal to the uncertainty interval of the concentration from traffic sources. If the uncertainty for the source traffic can be reduced, the uncertainty on all locations with high traffic emission will be reduced.

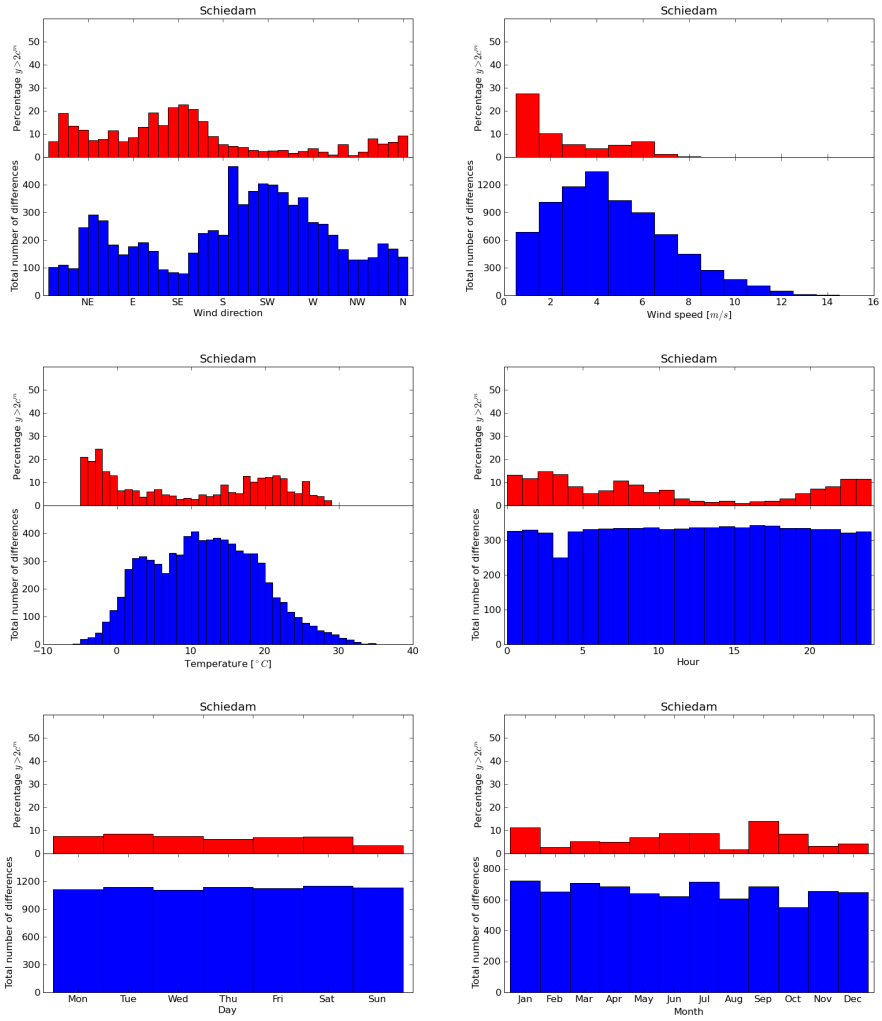


Figure 6.1: *Bar plots of relatively large differences between observations and model simulations at location Schiedam. In the lower graphs is the total number of differences plotted for each input parameter, in the upper graphs is the percentage given when the observation is more than two times the model simulation.*

6.2 Kalman filter

The application of the Kalman filter is nearly the same as in the application for the background concentrations. Every standard concentration fields gets a correction factor. So for each of the entries of $\underline{\gamma}$ a temporal correlation parameter has to be found. For some sources this will be difficult because there is no good series of measurements to calculate the correlation. In Section 6.4 the different values for α will be calculated. Not all the values for α declared exactly, also the uncertainty of the model σ is not known exactly. The uncertainty of the logarithms of the observations (R), estimated in Section 5.3 is also not known exactly. Therefore some sensitivity runs are done to find the optimal values for α, σ and R, this is explained in Section

6.5

6.2.1 Dynamical system

To create a dynamical system for $\underline{\gamma}$, it is again necessary to make a switch to the logarithms of the concentrations. This is done in the next formula:

$$\ln\left(\underline{c}_k\right) = \ln\left(\sum_{i=1}^{88} \mu_{i,k}\underline{m}_i e^{\gamma_i}\right) \tag{6.2}$$

This equation is non-linear for $\underline{\gamma}$, therefore a linearization is made around $\underline{\gamma} = \underline{0}$:

$$\ln\left(\sum_{i=1}^{88} \mu_{i,k}\underline{m}_i e^{\gamma_i}\right) = \ln\left(c_k^m\right) + \left[\frac{\mu_{j,k}\underline{m}_j}{\underline{c}_k^m}\right]_{j=1}^{j=88} \underline{\gamma} + \mathcal{O}\left(\underline{\gamma}\cdot\underline{\gamma}\right) \tag{6.3}$$

where $\underline{c}_k^m$ is the total concentration, calculated by the model. The dynamical system will then become:

$$\ln\left(\underline{c}_k\right) = \ln\left(\underline{c}_k^m\right) + \left[\frac{\mu_{j,k}\underline{m}_j}{\underline{c}_k^m}\right]_{j=1}^{j=88} \underline{\gamma} \tag{6.4}$$

$$\underline{\gamma}_{k+1} = A\underline{\gamma}_k + \underline{\omega}_k \qquad \underline{\omega}_k \sim N\left(0, Q_k\right) \tag{6.5}$$

Matrix A contains the temporal correlation parameters, these are calculated in Section 6.4. Covariance matrix Q_k is again a diagonal matrix, with diagonal elements q_i^2. This is a colored noise process, driven by a white noise process, like in the application for the background. Furthermore it is assumed that both the temporal correlation and the model uncertainty are independent of time.

$$q_i = \sqrt{1-\alpha_i^2}\sigma_i^2 \tag{6.6}$$

where σ_i corresponds with the model uncertainty for each entry of $\underline{\gamma}$.

6.2.2 Kalman filter form

The dynamical system has to be written in Kalman filter form, for the implementation of the Kalman filter. There are still 9 series of measurements available, these series will be compared with the model simulations to get a better estimate of the NO_x concentration. The dynamical system in Kalman filter form is defined as follows:

$$\underline{\gamma}_{k+1} = A\underline{\gamma}_k + \underline{\omega}_k \qquad \underline{\omega}_k \sim N\left(0, Q_k\right) \tag{6.7}$$

$$\ln\left(\underline{y}_k\right) = H\left(\ln\left(\underline{c}_k^m\right) + \left[\frac{\mu_{j,k}\underline{m}_j}{\underline{c}_k^m}\right]_{j=1}^{j=88} \underline{\gamma}_k\right) + \underline{\nu}_k \qquad \underline{\nu}_k \sim N\left(0, R_k\right) \tag{6.8}$$

In this equations matrix H is the system operator which projects the model state onto the measurements. Covariance matrix R corresponds with the uncertainty of the logarithm of the measurements, the instrumental error combined with the representation error. Matrix R will be a diagonal matrix, the elements on the diagonal are estimated in Section 5.3.

To simplify the notations from Equations 6.7 and 6.8, the Kalman filter equations are written as follows:

$$\underline{\gamma}_{k+1} = A\underline{\gamma}_k + \underline{\omega}_k \qquad \underline{\omega}_k \sim N(0, Q_k) \tag{6.9}$$

$$\underline{\tilde{y}}_k = \tilde{H}\underline{\gamma}_k + \underline{\nu}_k \qquad \underline{\nu}_k \sim N(0, R_k) \tag{6.10}$$

where vector $\underline{\tilde{y}}$ and matrix $\tilde{H}$ are defined as follows:

$$\underline{\tilde{y}}_k = \ln\left(\underline{y}_k\right) - H\ln\left(\underline{c}_k^m\right) \tag{6.11}$$

$$\tilde{H} = H\left[\frac{\mu_{j,k}\underline{m}_j}{\underline{c}_k^m}\right]_{j=1}^{j=88} \tag{6.12}$$

6.2.3 *Forecast step*

In the forecast step, a prediction is made for the values of $\underline{\gamma}_{k+1}$ with information from the time step before. The expected median and variance of $\underline{\gamma}_{k+1}$ are given by the next formulas:

$$\underline{\hat{\gamma}}_{k+1}^f = A\underline{\hat{\gamma}}_k \tag{6.13}$$

$$P_{k+1}^f = AP_kA^{\mathrm{T}} + Q_k \tag{6.14}$$

6.2.4 *Analysis step*

In the analysis step the forecasted concentrations are compared with the outcomes of a series of measurements. Like in the application for the background concentrations, it is not expected that the values for $\underline{\gamma}$ will become very large. Also the linearization is around $\underline{\gamma} = \underline{0}$ of order $\mathcal{O}\left(\underline{\gamma}\cdot\underline{\gamma}\right)$, thus large values for γ_i will cause stability problems. For those reasons a screening process as in Section 5.6 is implemented. In Section 6.3 the screening process for this application is explained.

$$\underline{\gamma}_{k+1}^a = \underline{\gamma}_{k+1}^f + K_{k+1}\left(\underline{\tilde{y}}_{k+1} - \tilde{H}_{k+1}\underline{\hat{\gamma}}_{k+1}^f\right) \tag{6.15}$$

$$P_{k+1}^a = \left(I - K_{k+1}\tilde{H}_{k+1}\right)P_{k+1}^f\left(I - K_{k+1}\tilde{H}_{k+1}\right)^{\mathrm{T}} + K_{k+1}R_{k+1}K_{k+1}^{\mathrm{T}} \tag{6.16}$$

where K_{k+1} is again the minimal variance gain, the gain that minimizes P_{k+1}^adefined as follows:

$$K_{k+1} = P^f_{k+1}\tilde{H}^{\mathrm{T}}_{k+1}\left(\tilde{H}_{k+1}P^f_{k+1}\tilde{H}^{\mathrm{T}}_{k+1} + R_{k+1}\right)^{-1} \tag{6.17}$$

6.3 Screening process

When the difference between the observation y and the model simulation $\underline{c}^m_k$ is large, the Kalman filter will produce a large correction factor for one or more standard concentration fields. This is not wanted, because it is assumable that a large difference is caused by another inaccuracy in the model or by an incidental occasion. For example when a road is blocked, the traffic pattern is different and thus the emissions are different from the expectations calculated by the model. Like in the application for the background in Chapter 5, a criterion has to be made whether a measurement is good enough.

After the forecast step it is possible to make an uncertainty interval of the forecasted concentration, there is also an uncertainty interval for the observation. When both of these intervals has a empty intersection, the difference between the simulation and the observation is too large. If for both intervals the 2σ uncertainty interval is taken, the screening criterion corresponds with the screening criterion in Section 5.6.

$$\left[\sum_{i=0}^{88} m_i\mu_{i,k}e^{\gamma^f_{i,k}-\beta sd^f_{i,k}}, \sum_{i=0}^{88} m_i\mu_{i,k}e^{\gamma^f_{i,k}+\beta sd^f_{i,k}}\right] \cap [y_k - \beta r_{frac}y_k, y_k + \beta r_{frac}y_k] \neq \emptyset \tag{6.18}$$

The value for $rfrac$ is optimized in Section 6.5 and equal to 0.34. The application of the Kalman filter with this screening process with $\beta = 2$, results in a concentration for the first week of 2006 for locations Schiedam and Overschie as shown in Figure 6.2. Contradicting to Figure 5.6, only 12% of the observations are not executed in the analysis step of the Kalman filter. This means that the inaccuracies in the model could be well described by inaccuracies of the several standard concentration fields.

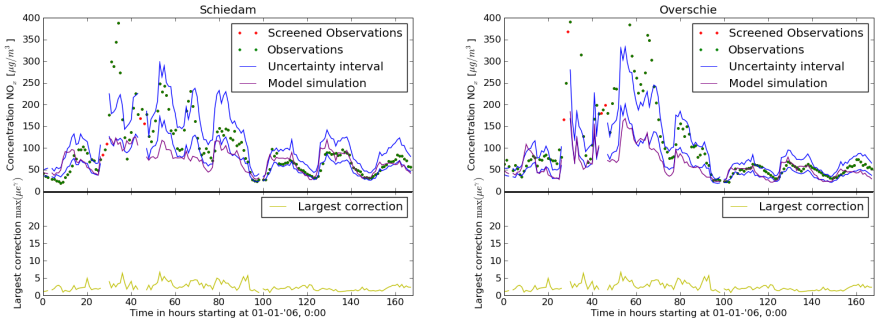


Figure 6.2: *Concentrations for the first week of 2006 after application of the Kalman filter on all the sources, with a screening process, at locations Schiedam and Overschie*

6.4 Correlation parameters α

In Section 5.4, an estimate value for the parameters α_i corresponding with the source background is calculated. In this section the same procedure will be done to obtain estimated values for the other parameters α_i. The temporal correlation for the source 'rest' is not possible to calculate with a series of measurements, therefore some runs of the Kalman filter has to be applied to get the optimal values for α_i corresponding with the source 'rest', this will be done in Section 6.5.

6.4.1 Traffic sources

The temporal correlation for the traffic sources will be obtained by looking at the measurements on locations Overschie and Ridderkerk. Location Overschie is close to main road A20 and Ridderkerk is close to main roads A15 and A16, both stations will give a good approximation of the emission from traffic sources.

In Figure 6.3, the temporal correlation is given for both stations, this is done with the same method as for the background sources, describes in Section 5.4. The best fitting exponential function has de-correlation parameter $\tau = 5$, thus an estimated value for each α_i corresponding with a traffic source is equal to $e^{-1/5}$.

The high peaks at 24 and 48 can be explained by the fixed traffic pattern. Each day the amount of traffic is roughly the same, thus there is a high mathematical temporal correlation for periods of one day.

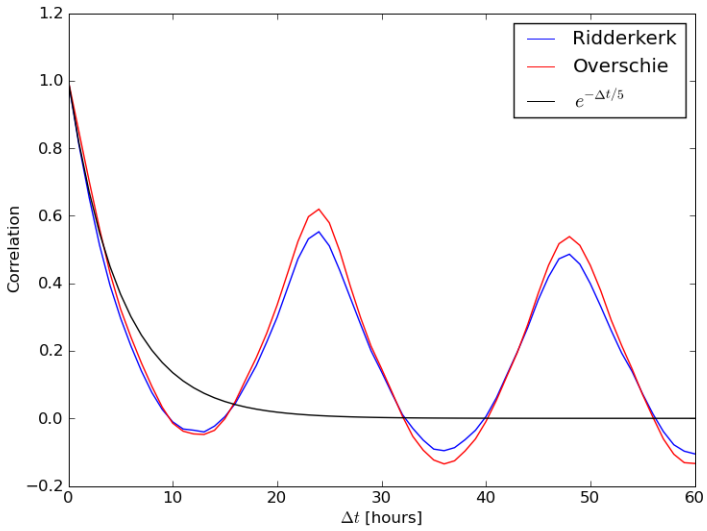


Figure 6.3: *The temporal correlation for traffic stations Overschie and Ridderkerk. The black line corresponds with the de-correlation parameter* $\tau_{tr} = 5$

6.4.2 *Industry source*

The temporal correlation for the source industry is not very easy to determine. There is no location with a dominating industry emission. The location Vlaardingen is the best location to calculate the correlation. This only results in an estimated correlation which is not very accurate. In Figure 6.4, the temporal correlation on location Vlaardingen is given. The best fitting exponential has de-correlation parameter $\tau = 10$, thus an estimated value for α_i corresponding with the source 'industry' is equal to $e^{-1/10}$.

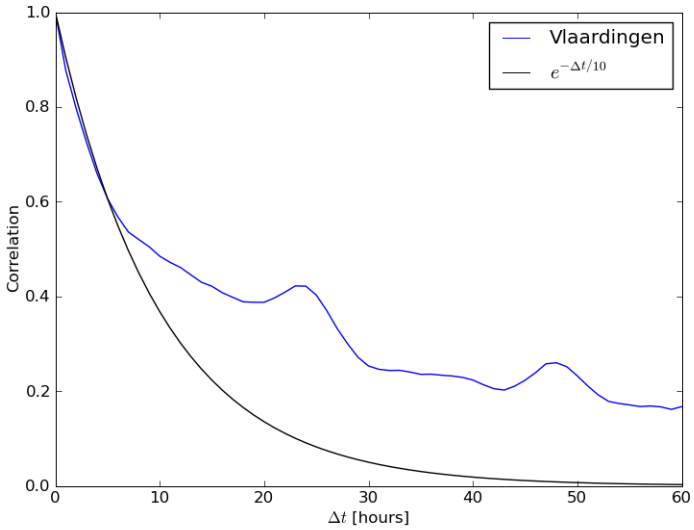


Figure 6.4: *The temporal correlation for station Vlaardingen, the station which matches best with the source industry. The black line has de-correlation parameter $\tau_{in} = 10$.*

6.4.3 *Shipping sources*

Also for the temporal correlation of the shipping sources, it is not easy to determine an estimate value for α_i. Location Maassluis is the best location to calculate the correlation, the temporal correlation at Maassluis is given in Figure 6.5. In here the same holds as for the industry, the de-correlation parameter $\tau = 8$ is only an inaccurate estimate. This leads to an estimated value for α_i corresponding with the shipping sources which is equal to $e^{-1/8}$.

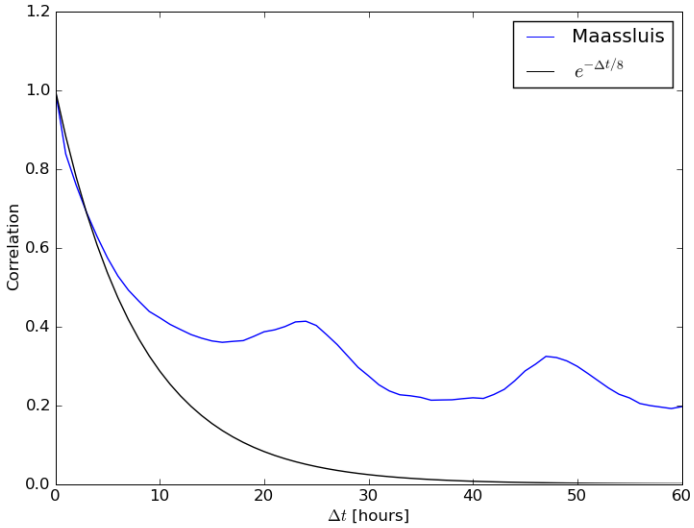


Figure 6.5: *The temporal correlation for station Maassluis, the station which corresponds best with the emission from shipping. The black line corresponds with de-correlation parameter* $\tau_{sh} = 8$.

6.4.4 Rest source

It is clear that the temporal correlation for the source 'rest' is not easy to declare. The only idea of this temporal correlation is that the de-correlation parameter τ_{re} will be around the values for the de-correlation parameters $\tau_{bg}, \tau_{tr}, \tau_{in}$ and τ_{sh}. The de-correlation parameter τ_{re} is estimated equal to 9, thus α_i corresponding with source 'rest' is estimated equal to $e^{-1/9}$

6.5 Sensitivity runs

In Sections 5.4 and 6.4, all the parameters $\alpha_{i,j}$ for the matrix A are not declared exactly. Also the parameter σ for the uncertainty of the model and the uncertainty of the model r_{frac} are not known exactly. Therefore the Kalman filter is applied for different values of $\tau_{bg}, \tau_{tr}, \tau_{sh}, \tau_{in}$ and different values of σ and r_{frac}. In Section 5.4, it was shown that an estimated value for τ_{bg} is equal to 12. In Section 6.4 the estimated values for $\tau_{tr}, \tau_{sh}, \tau_{in}$ and τ_{re} were found. The uncertainty of the model is assumed between 25 and 35%. The instrumental error in the observations calculated in Section 5.3 is equal to 8%, but the representation error may be larger, due to a grid with a low resolution. A trial and error process leads to the conclusion that the Kalman filter gives an optimal result with total uncertainty of the measurements between 25% and 35%. This will lead to applications of the Kalman filter with the following values:

$$\begin{aligned} \tau_{bg} &\approx 12 \\ \tau_{tr} &\approx 5 \\ \tau_{sh} &\approx 8 \\ \tau_{in} &\approx 10 \\ \tau_{re} &\approx 9 \\ \sigma &\in [0.25, 0.35] \\ r_{frac} &\in [0.25, 0.35] \end{aligned} \tag{6.19}$$

To obtain which combination of values for $\tau_{bg}, \tau_{tr}, \tau_{sh}, \tau_{in}, \sigma$ and r_{frac} is the best, there are three criteria which have to be optimized: The Root Mean Squared Error (RMSE), the mean of the differences between the Kalman filter results and the observations and finally the standard deviation of these differences.

6.5.1 Root Mean Squared Error (RMSE)

The first criterion is to minimize the value for RMSE

$$RMSE = \frac{1}{\sqrt{9n}} \sqrt{\sum_{k=1}^{n} \sum_{i=1}^{9} \left(y_{i,k} - c_{i,k}^{Kf} \right)^2}$$

where $c_{i,k}^{Kf}$ is the concentration after applying the Kalman filter. The number of time steps corresponds with the value of n, which is equal to 8760 for a whole year. The summation over i is up to 9, the number of measurement locations. The value for the RMSE is a measure for the absolute difference between the results after application of the Kalman filter and the observations. When this is minimized the Kalman filter results have the best comparison with the observations

6.5.2 Mean

Another criterion is the mean of the differences between the Kalman filter results and the observations:

$$Mean = \frac{1}{9n} \sum_{k=1}^{n} \sum_{i=1}^{9} \left(y_{i,k} - c_{i,k}^{Kf} \right)$$

The value for this mean is in the optimal situation equal to 0, in that case the Kalman filter results have the same mean as the observations.

6.5.3 Standard deviation

The final criterion is the standard deviation of the differences between the Kalman filter results and the observations:

$$Std \quad = \quad \frac{1}{\sqrt{9n}} \sqrt{\sum_{k=1}^{n} \sum_{i=1}^{9} \left(\left(y_{i,k} - c_{i,k}^{Kf} \right) - Mean \right)^2}$$

In the optimal situation the value for this standard deviation is equal to 1.

6.5.4 Optimal results for RMSE, mean and standard deviation

To obtain the optimal combination of parameters, the values for $RMSE$, $Mean$ and Std are plotted with respect to the 7 parameters $\tau_{bg}, \tau_{tr}, \tau_{sh}, \tau_{in}, \tau_{re}, \sigma$ and r_{frac}. This is shown in Figure 6.6, for each of the parameters the values for the $RMSE$, $Mean$ and Std are plotted with respect to the possible values for the parameters. The plots with respect to parameters τ_{sh}, τ_{in} and τ_{re} are not shown in Figure because the values for $RMSE$, $Mean$ and Std are nearly constant for every value of $\tau_{sh,}\tau_{in}$ and τ_{re}. In the optimal application of the Kalman filter these values will be taken equal to the estimated values from Section 6.4, thus $\tau_{sh} = 8, \tau_{in} = 10$ and $\tau_{re} = 9$.

For the optimal values of the other parameters the plots in Figure 6.6 have to be analyzed. For the parameter τ_{bg}, the standard deviation decreases if τ_{bg} increases. The mean is equal to zero when $\tau_{bg} \approx 7$, the $RMSE$ will decrease slowly for large τ_{bg}. Therefore a good option will be $\tau_{bg} = 10$. The same analysis could be done for τ_{tr} and leads to the option $\tau_{tr} = 3$.

The parameter σ has a larger influence on all three criteria, to optimize the standard deviation the value for σ must be around 0.36. For an optimal value of the mean, the value of σ must be around 0.32, the $RMSE$ is decreasing when σ is increasing. This results in a value for σ stated equal to 0.35. The analysis of the parameter r_{frac} states that the standard deviation will be optimal for $r_{frac} \approx 0.38$, the mean is close to zero when $r_{frac} \approx 0.34$. The value for $RMSE$ is minimal when $r_{frac} \approx 0.28$, with this analysis, the value for r_{frac} is stated equal to 0.34.

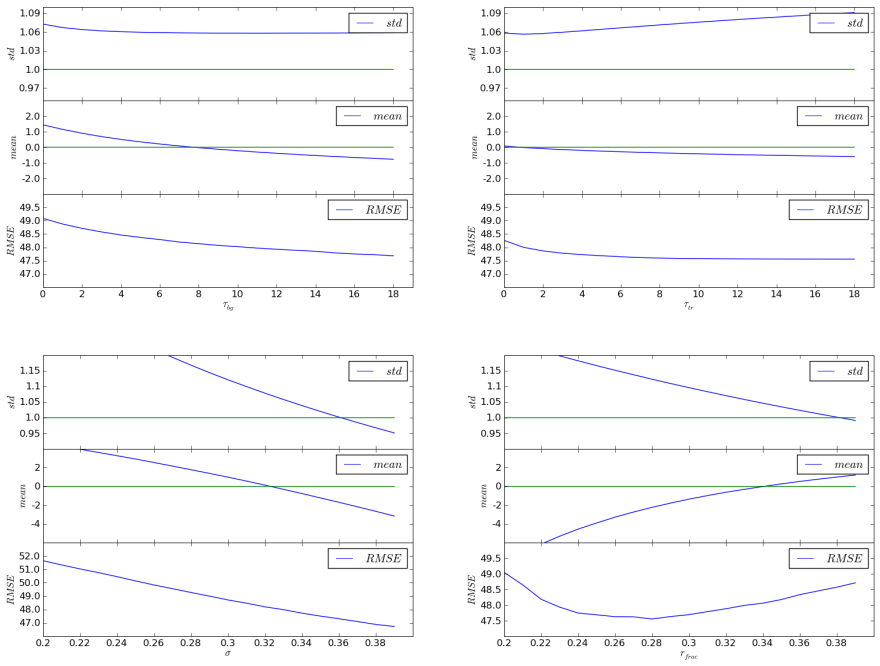


Figure 6.6

6.6 Connection with population

After the application of the Kalman filter, it is possible to calculate an uncertainty interval for the concentration NO_x for each grid cell in the area of interest. The next objective is to connect the interval on a certain grid cell with the number of people living in that grid cell.

6.6.1 Population density

A map of the population of the area is given in Figure 6.7. This figure represents the density of postal zip codes per grid cell instead of the number of people per grid cell. The total number of zip codes in this area is equal to 595.396. According data from CBS [1], the total number of residents in this region is equal to 1.186.306 on the first of January of 2006. Thus the average number of people per zip code is equal to 1.99. Further in this report it is assumed that the number of people per zip code is equal, thus the number of residents in a grid cell is 1.99×the number of zip codes.

$$pop_j = 1.99 \times \#\, of\ zip\, codes_j \tag{6.20}$$

[1] CBS: Centraal Bureau voor de Statistiek. $www.cbs.nl$
Dutch organization for statistics

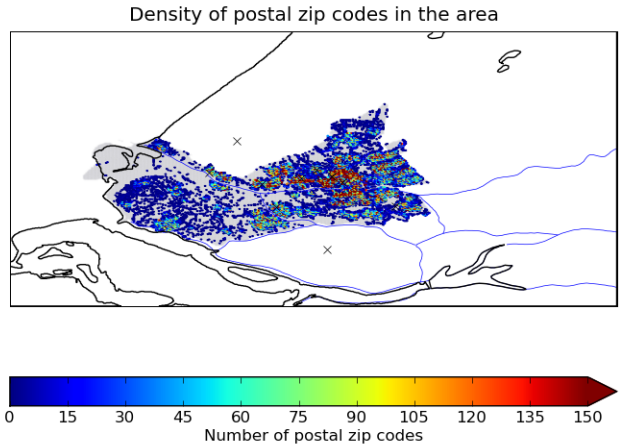


Figure 6.7: *Density of postal zip codes in the DCMR area*

6.6.2 Mean width of the uncertainty intervals connected with population density

For every grid cell, on every hour an uncertainty interval is calculated by the Kalman filter application. The width of these intervals is a measure for the uncertainty of the concentration NO_x, when the width of the interval is small, the estimate of the concentration NO_x is 'good'. The idea is now to have small intervals on locations where the population density is high, in that case there is a good estimate of the exposure of the population on the concentration NO_x. The width of an uncertainty interval in grid cell j on time k is the upperband of the 1σ interval minus the lower band of the 1σ interval:

$$w_{j,k} = \sum_{i=1}^{88} \mu_{i,k} m_{i,j} e^{\gamma_{i,k} + sd_{i,k}} - \sum_{i=1}^{88} \mu_{i,k} m_{i,j} e^{\gamma_{i,k} - sd_{i,k}} \tag{6.21}$$

where $m_{i,j}$ is de standard concentration of emission source i in grid cell j.

In Figure 6.8, the annual mean $\bar{w}_j$ of $\{w_{j,k}\}_{k=1}^{8760}$is plotted for each grid cell:

$$\bar{w}_j = \frac{1}{8760} \sum_{k=1}^{8760} w_{j,k} \tag{6.22}$$

These annual means for each grid cell can be compared with the population density on each grid cell, this comparison is shown in Figure 6.9. On the x-axis are the values of $\bar{w}_j$, on the y-axis are the number of people living in a grid cell with that annual mean. For $\bar{w}_j \in [w_i, w_{i+1}]$, the number of people for that width range of $\bar{w}_j$ is equal to:

$$\sum_{j=1}^{n_{gc}} pop_j \mathcal{I}_{\{\bar{w}_j \in [w_i, w_{i+1}]\}} \tag{6.23}$$

where n_{gc} is the number of grid cells and $\mathcal{I}$ is the indicator function.

In Figure 6.8, it is shown that relatively many grid cells have an annual mean above 40. This large uncertainty mostly occurs on main roads and industrial regions, therefore there are not that many people living in grid cells with a large annual mean of uncertainty.

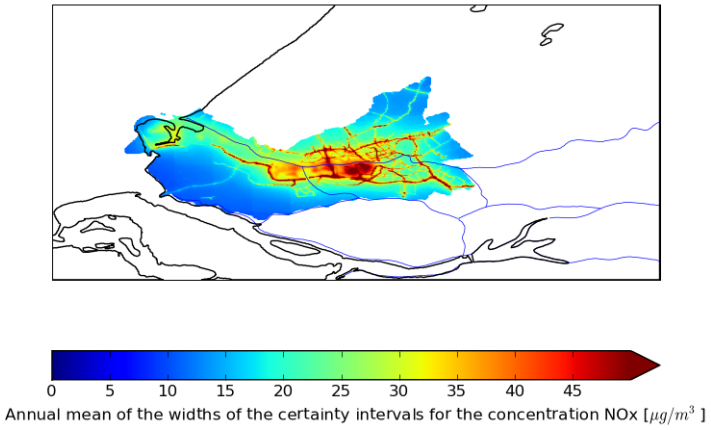


Figure 6.8: *The values for $\bar{w}_j$ over the whole area of interest*

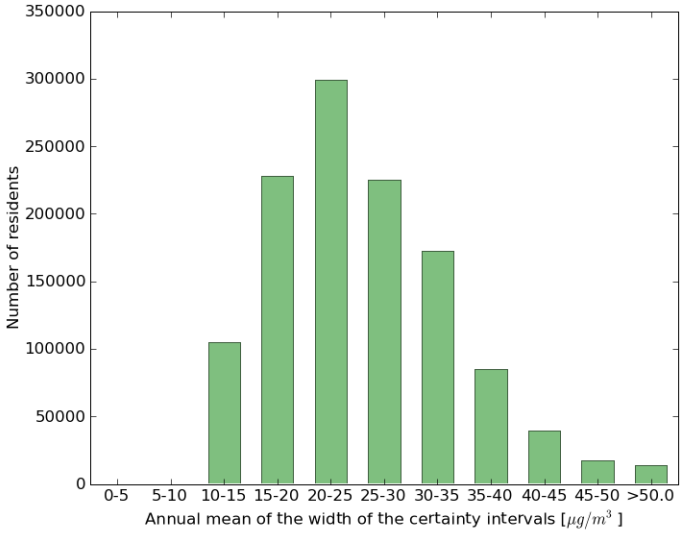


Figure 6.9: *Histogram of the number of people per grid cell against the annual mean $\bar{w}_j$. On the x-axis are ranges of $\bar{w}_j$, on the y-axis are the number of people living in a grid cell with $\bar{w}_j$ in that range.*

6.6.3 Large widths of the uncertainty intervals connected with population density

When the width of an uncertainty interval is large, it is difficult to give a good estimation of the real concentration NO_x, also the estimate for the exposure will not be accurate. For that reason it is interesting to look for each grid cell at the number of times that the uncertainty interval has a large width.

Therefore a definition of 'large' has to be made. The mean of all the widths of all uncertainty intervals over a whole year is equal to 26.0, the standard deviation of all those means is 14.2. Therefore an uncertainty interval has a 'large' width if the width is above the mean plus two times the standard deviation, thus 26.0 + 2×14.2 = 54.4. Now it is possible to count for every grid cell the number of times that the width of the uncertainty interval is above 54.4.

$$\#\, large\ uncertainties_j = \sum_{k=1}^{8760} \mathcal{I}_{\{w_{j,k}>54.4\}} \tag{6.24}$$

In Figure 6.10, the number of large widths of the uncertainty intervals are shown for each grid cell in the whole region. This number of large uncertainties can also be compared with the population density, Figure 6.11 shows this comparison. On the x-axis are the amount of large uncertainties, on the y-axis are the number of people living is a grid cell with that amount of large uncertainties.

For $\#\, large\ uncertainties \in [t_i, t_{i+1}]$ with t_i the number of times that a large uncertainty exists, the number of people is equal to:

$$\sum_{j=1}^{n_{gc}} pop_j \mathcal{I}_{\{\#\ large\ uncertainties_j \in [t_i, t_{i+1}]\}} \tag{6.25}$$

In Figure 6.10, it is shown that for relatively many grid cells there are many number of times with large uncertainty. Most of the grid cells with a large number of times with high uncertainty are on main roads or industrial regions, thus again not many people have relatively many large uncertainties. This is shown in Figure 6.11. The last two peaks in the histogram shows the people with a large amount of large uncertainties, this corresponds with the people living near main roads and nearby industrial regions.

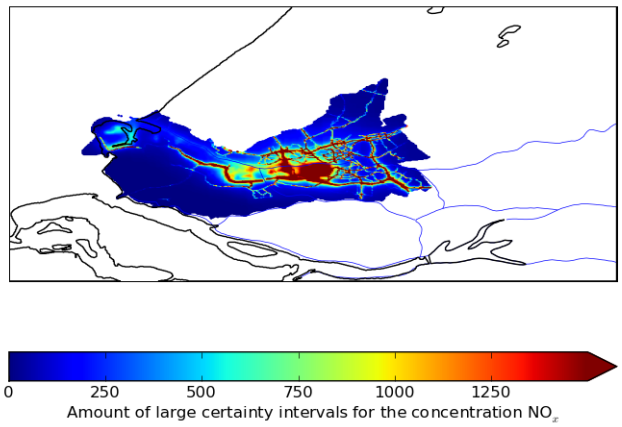


Figure 6.10: *Amount of large values for $w_{j,k}$ over the year 2006 for the whole domain*

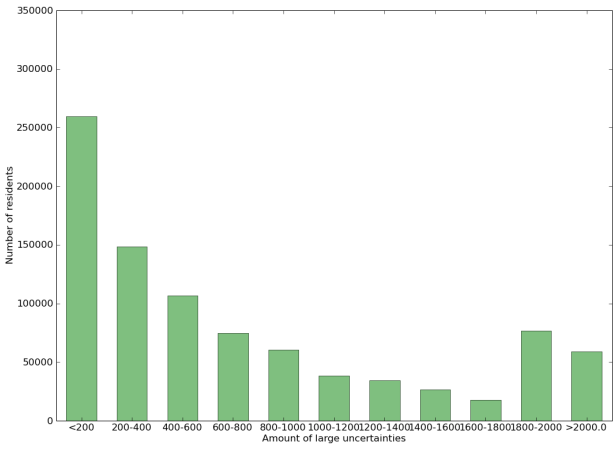


Figure 6.11: *Histogram of number of people per grid cell against the amount of large uncertainties for each grid cell. On the x-axis are the ranges of $\#\ of\ large\ uncertainties$, on the y-axis are the number of people living in a grid cell with $\#\ of\ large\ uncertainties$ in that range.*

7 Conclusion and discussion

In Chapter 3, it is shown that the inaccuracies of the Real Time URBIS model are dependent of the wind direction, the wind speed and the hour of the day. For that reason a Kalman filter is applied on the standard concentration field from the URBIS model to eliminate these inaccuracies and to connect the model with a series of measurements.

In Chapter 5, the Kalman filter was only applied on the background concentrations, this was not sufficient to eliminate all the inaccuracies. Therefore in Chapter 6, the Kalman filter is applied on all emission sources. This application is good enough to connect the model to almost all measurements, nearly 90% of the measurements are taken into account. The application of the Kalman filter results in an uncertainty interval for the concentration NO_x for the whole domain covered by DCMR. The uncertainty interval has a large width on the main roads and on the industry region around Pernis. On this locations the concentration is relatively large, thus also the absolute uncertainty will be large.

With this Kalman filter it is now possible to correct the Real Time URBIS model on every hour with the series of measurements. This correction will be different for each location, because of the differences in the sources compositions in each grid cell.

Another application of this method is that, it is possible to connect the uncertainty intervals with the population in the area. The idea is that on crowded locations the uncertaintyinterval should be as small as possible. Therefore a further investigation could be, adding some monitoring locations such that the width of the uncertainty interval will change. With this method it is possible to create an optimal setting of measurement locations.

Bibliography

R. Kranenburg. Statistische onzekerheidsanalyse van Real Time URBIS voor het Rijnmondgebied. TNO-Report 2009-01347, TNO, May 2009.

A.J. Segers. Data assimilation in atmospheric chemistry model using Kalman filtering. Technical report, TU Delft, 2002.

P. Wesseling and P.Y.J. Zandveld. URBIS Rotterdam Rijnmond; A pilot study. TNO-rapport 2003/245, TNO, 2003.

A Standard concentration fields

Standard concentration fields for the 11 different sources in the URBIS model, each source has 8 standard concentration field valid for 4 different wind directions (N, E, S, W) and 2 different wind speeds (1.5 m/s and 5.5 m/s).

Figure A.1: Emission source: Abroad

Figure A.2: Emission source: Car

Figure A.3: Emission source: Residents Rijnmond

Figure A.4: Emission source: Industry

Figure A.5: Emission source: Rest

Figure A.6: Emission source: Background from the rest of the Netherlands

Figure A.7: Emission source: Roads far

Figure A.8: Emission source: Roads nearby

Figure A.9: Emission source: Ships inland

Figure A.10: Emission source: Ships sea

Figure A.11: Emission source: Zonecards

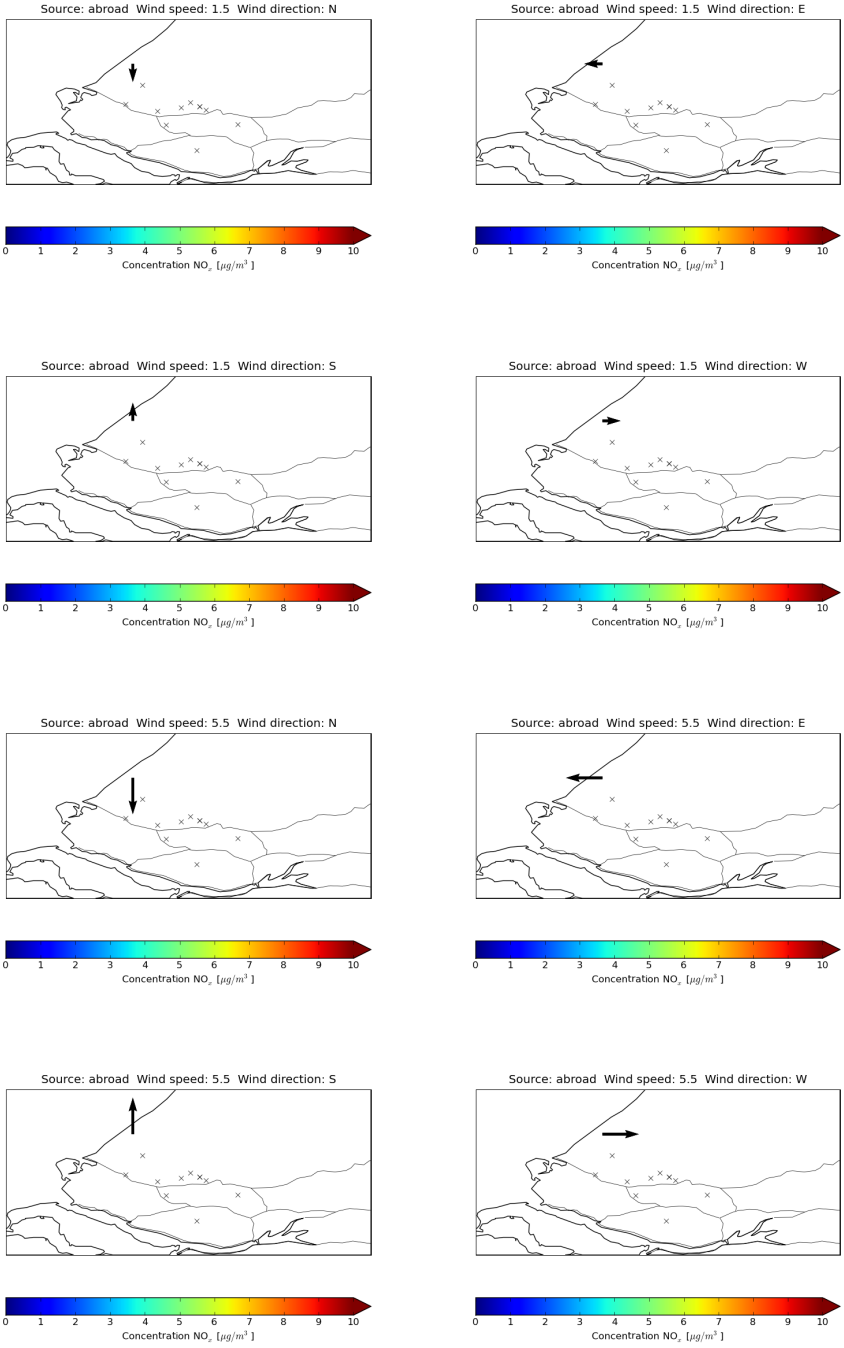


Figure A.1: *Emission Source: Abroad*

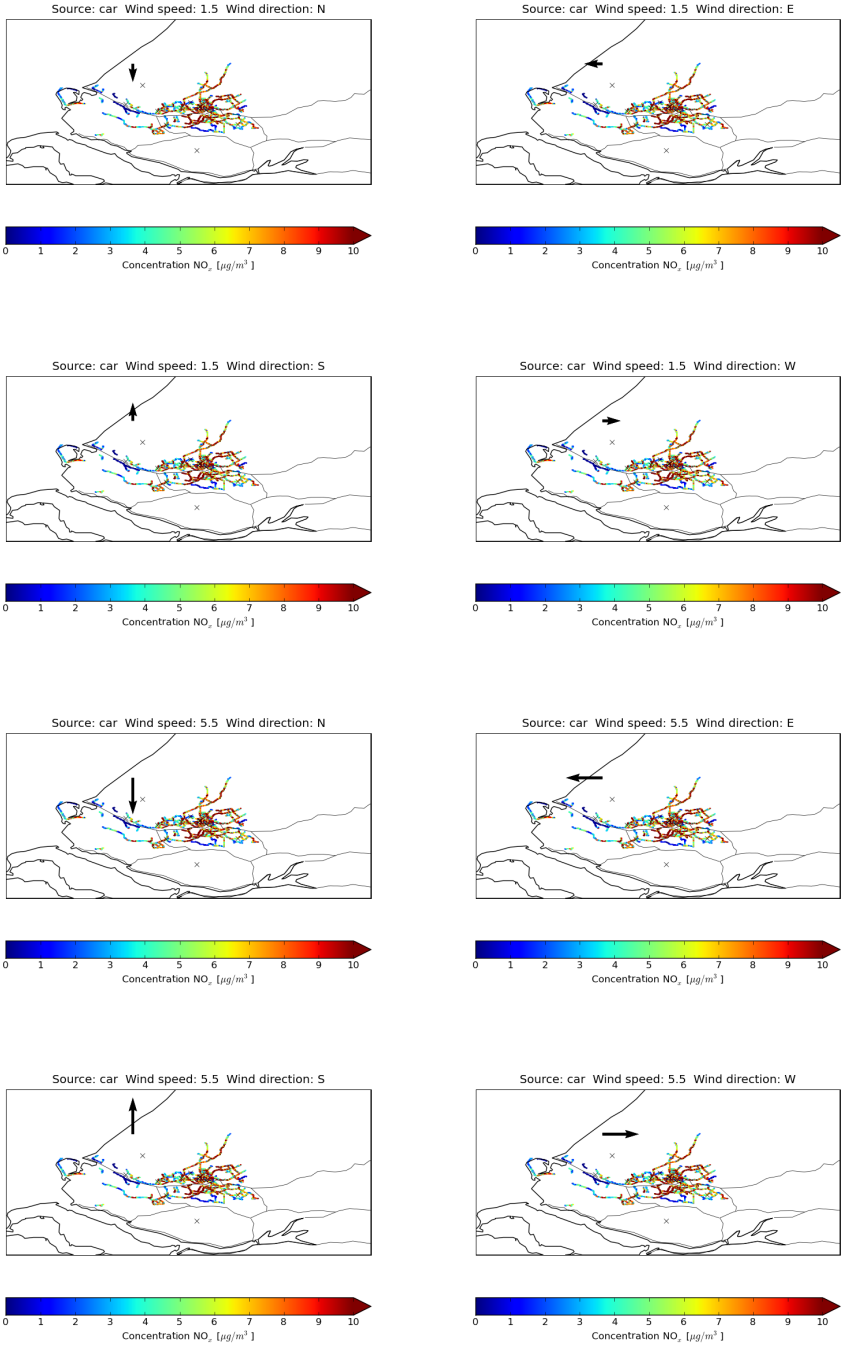


Figure A.2: *Emission Source: Car*

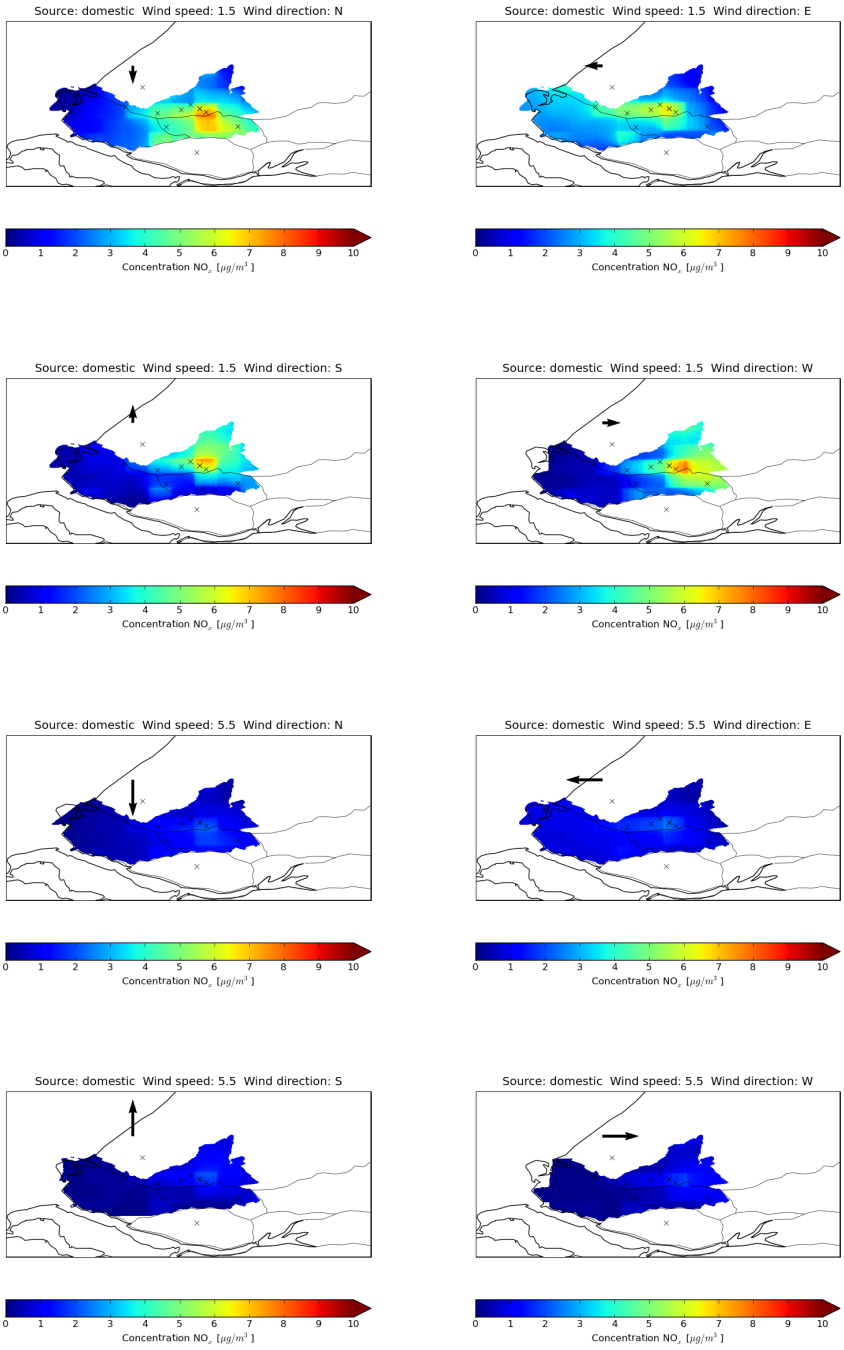


Figure A.3: *Emission Source: Domestic Rijnmond*

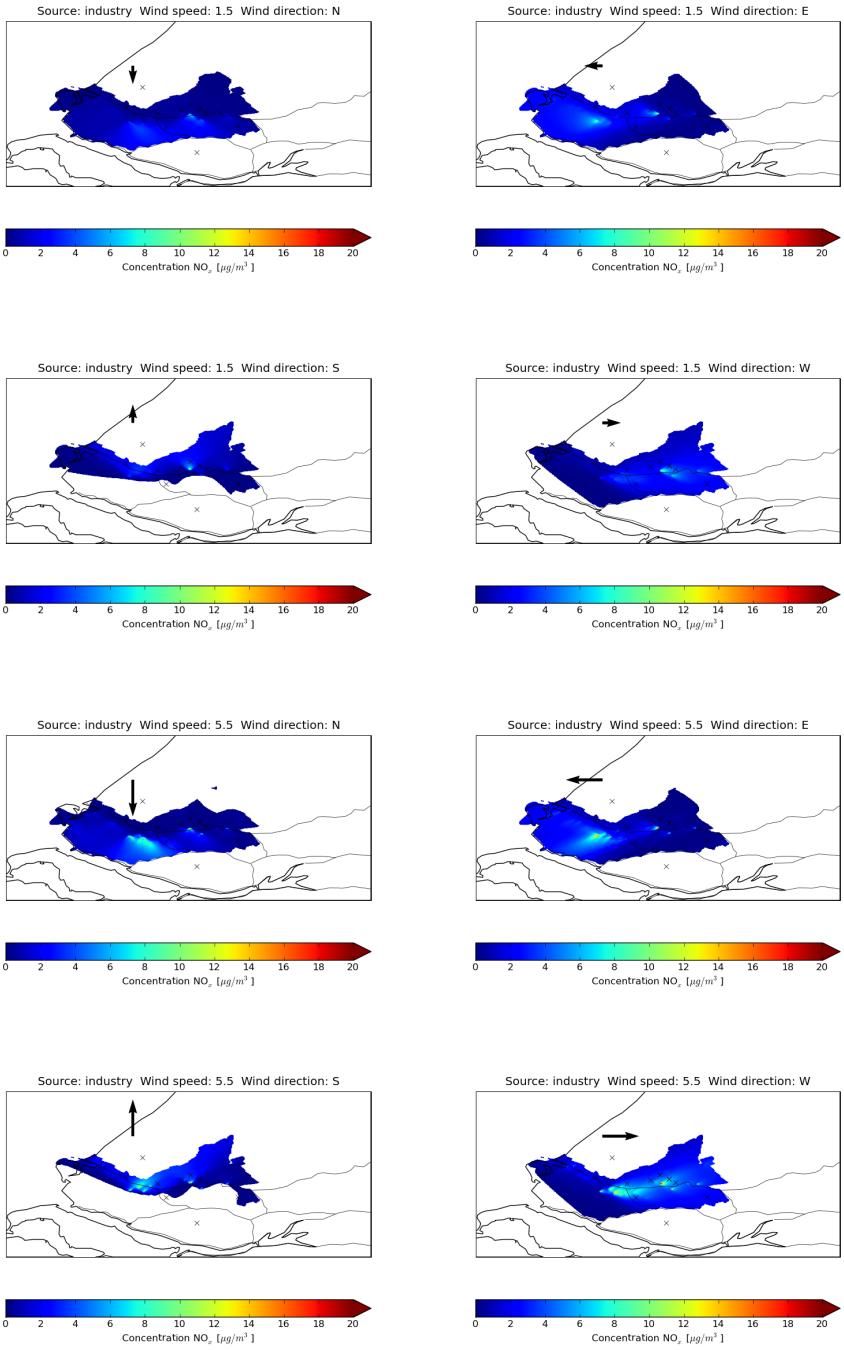


Figure A.4: *Emission Source: Industry*

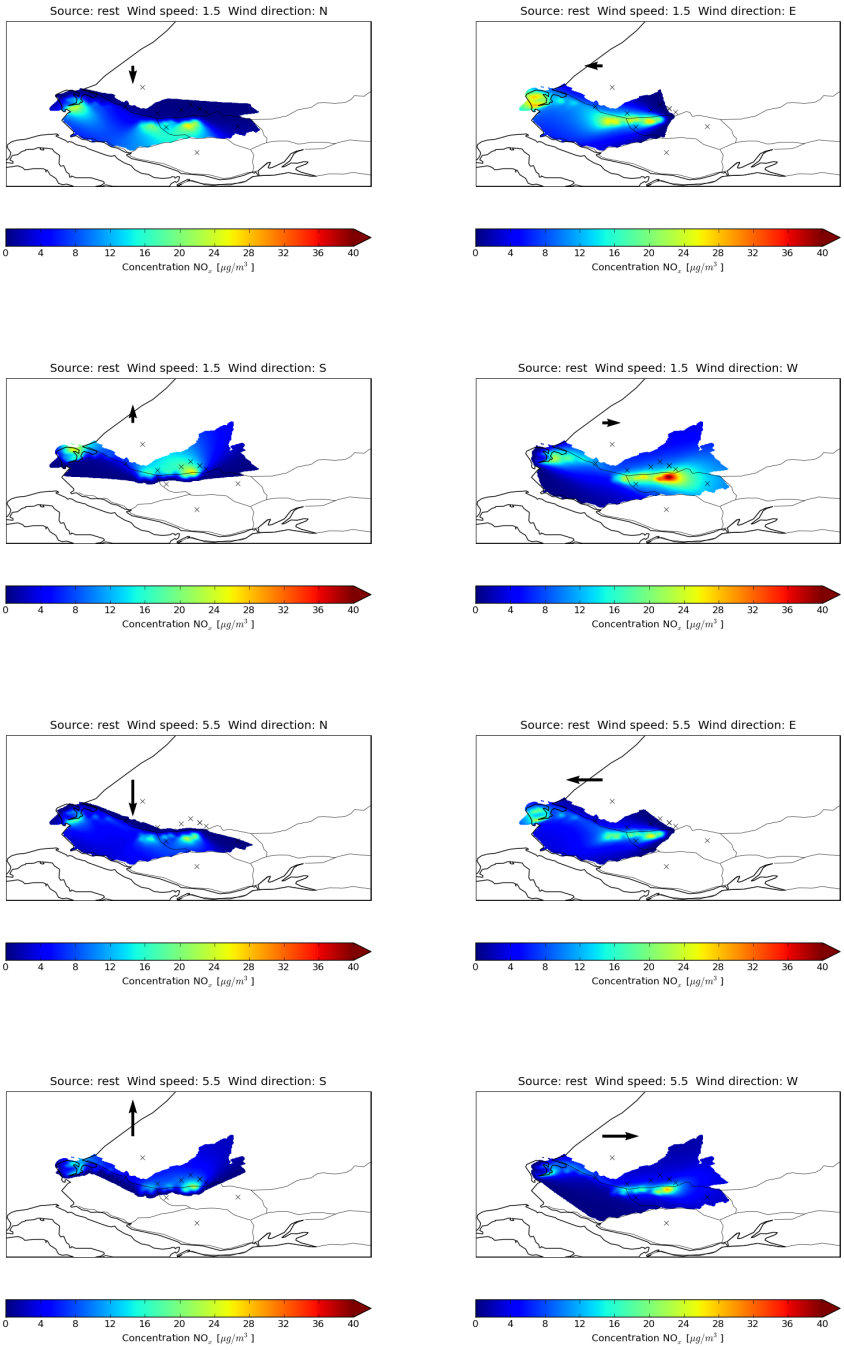


Figure A.5: *Emission Source: Rest*

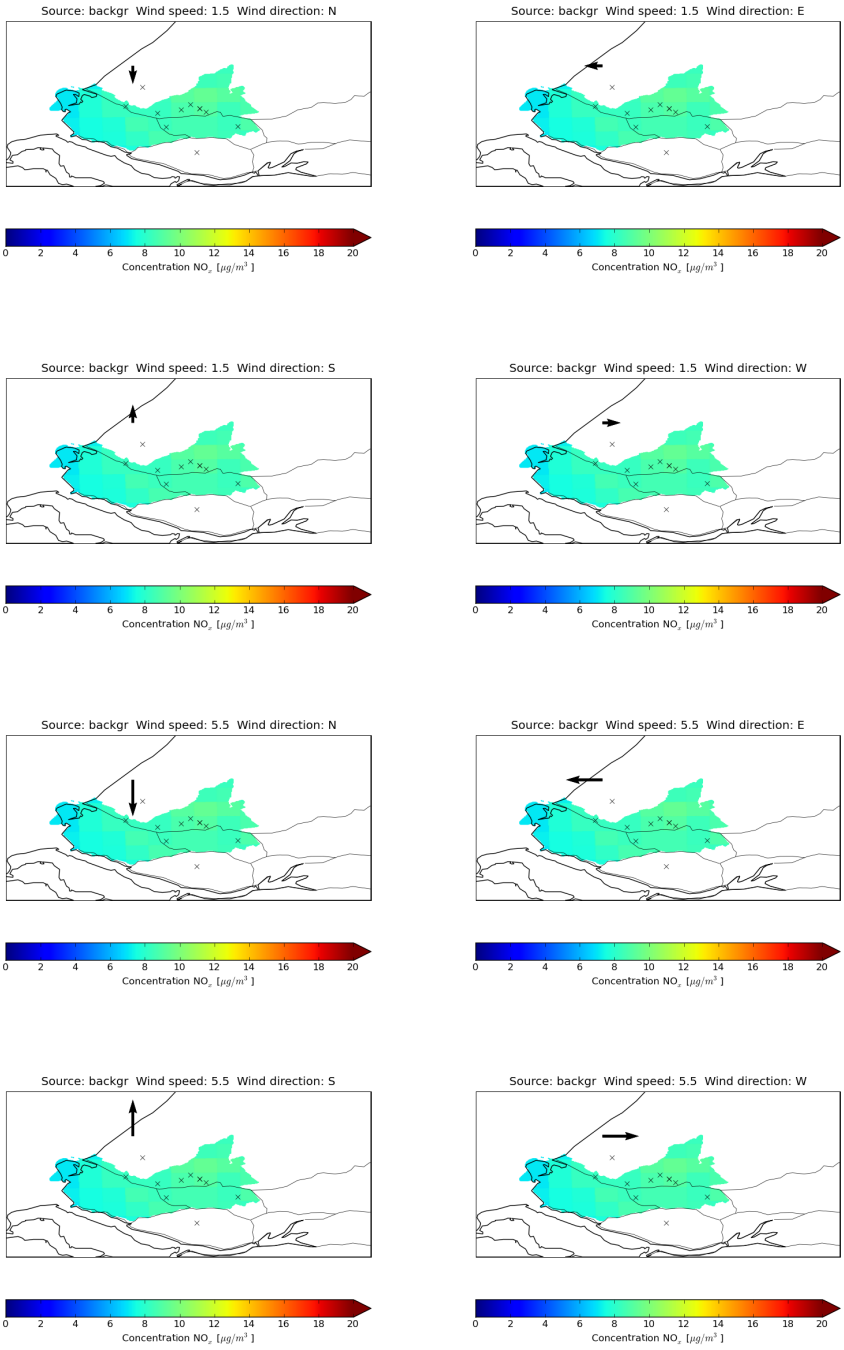


Figure A.6: *Emission Source: Rest of the Netherlands*

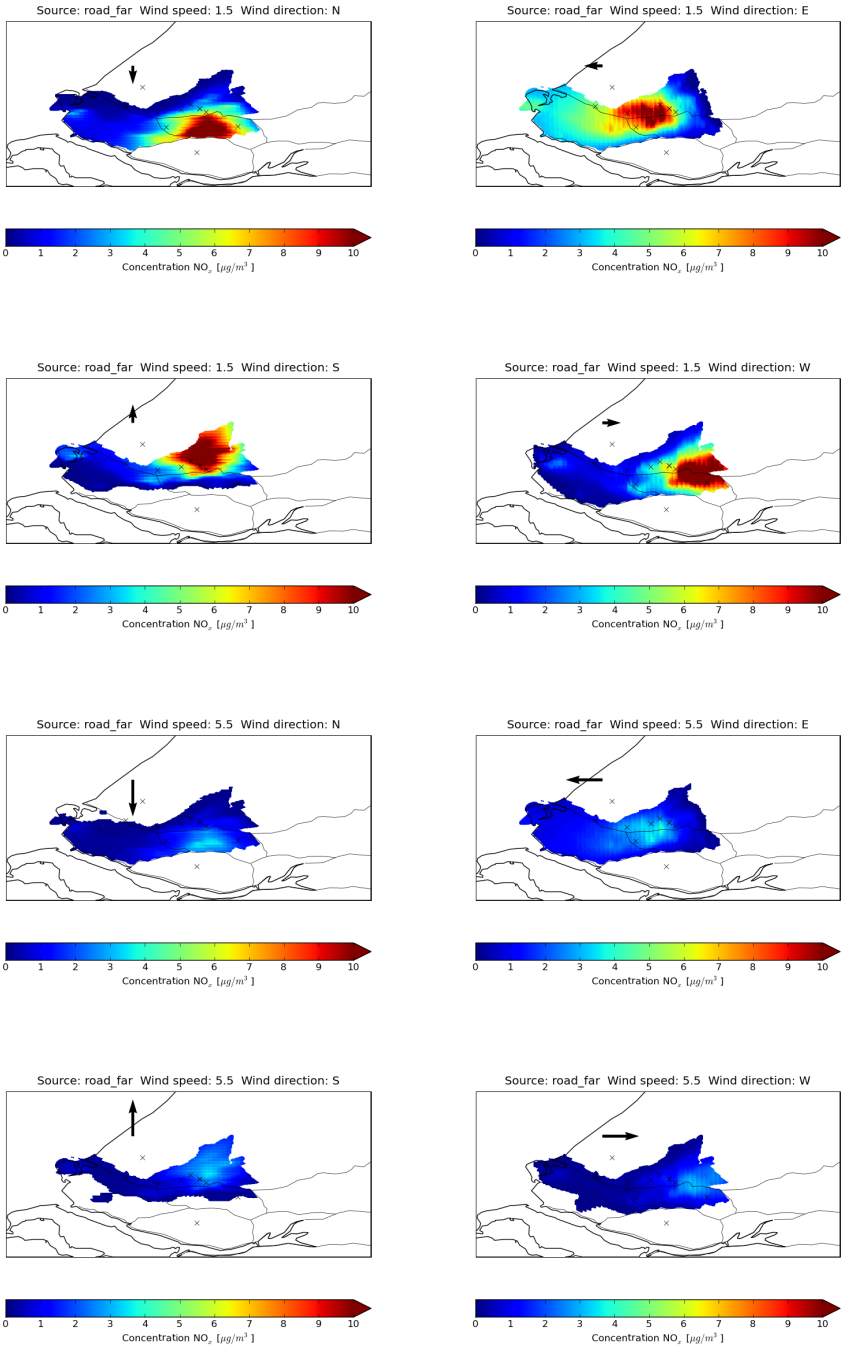


Figure A.7: *Emission Source: Road far*

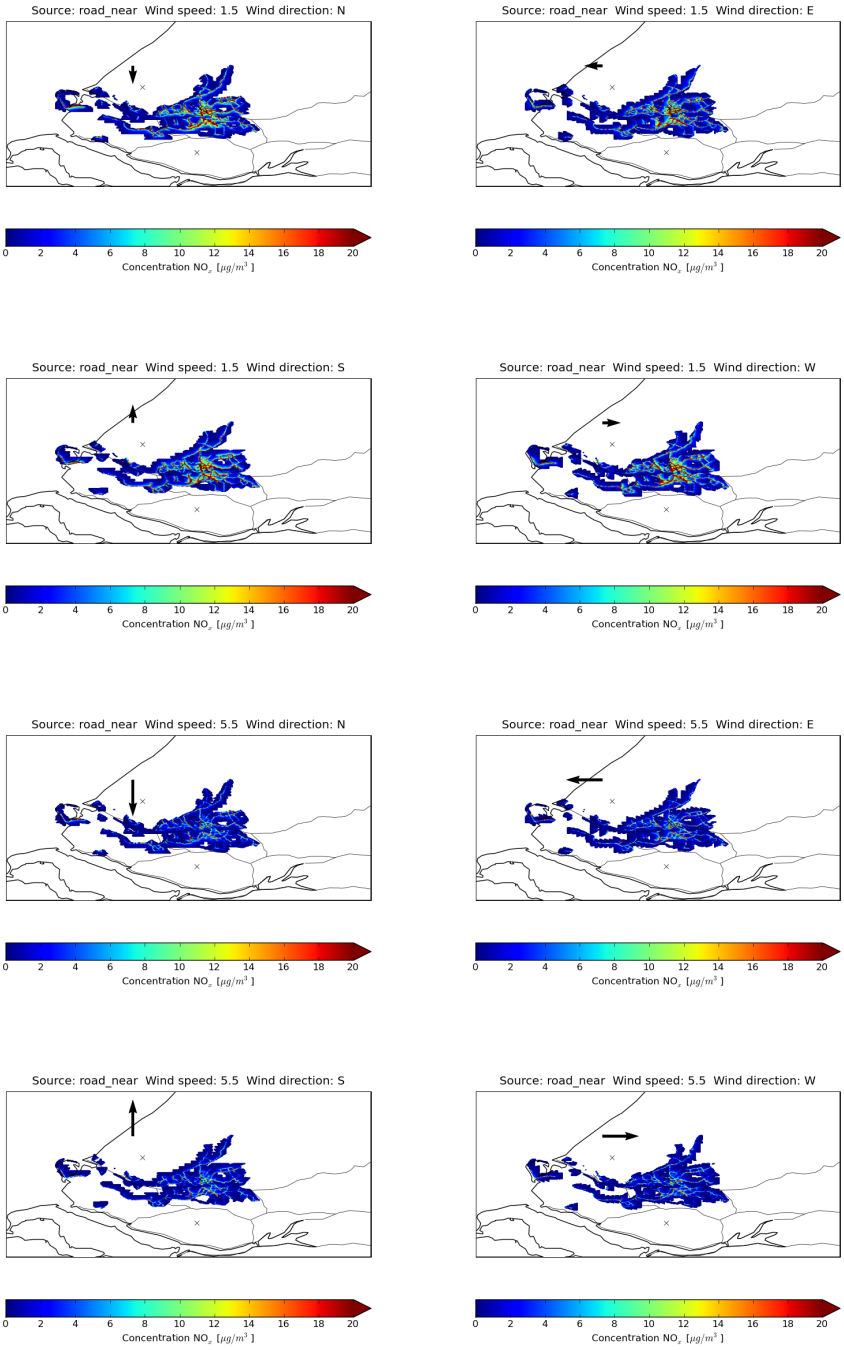


Figure A.8: *Emission Source: Road nearby*

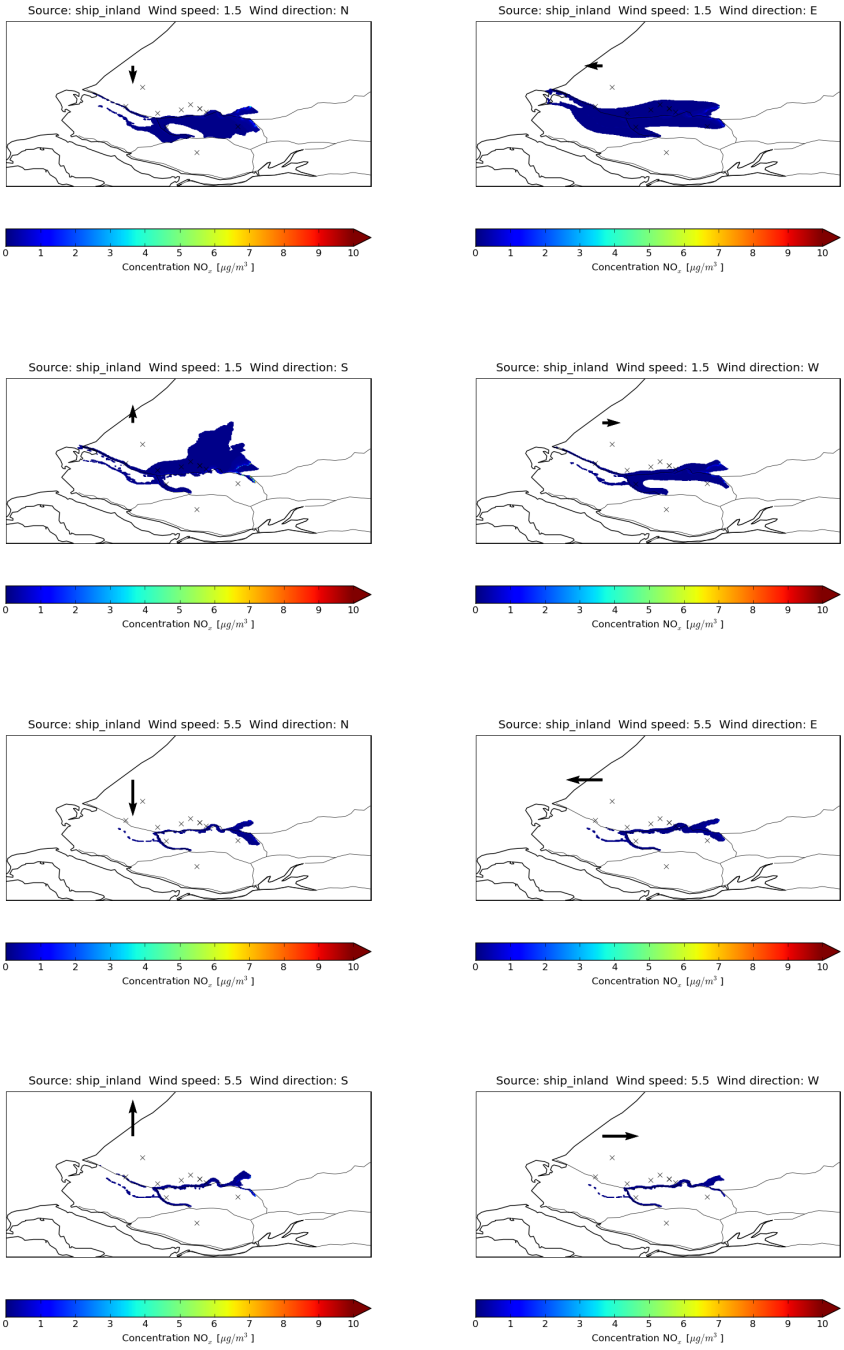


Figure A.9: *Emission Source: Ships inland*

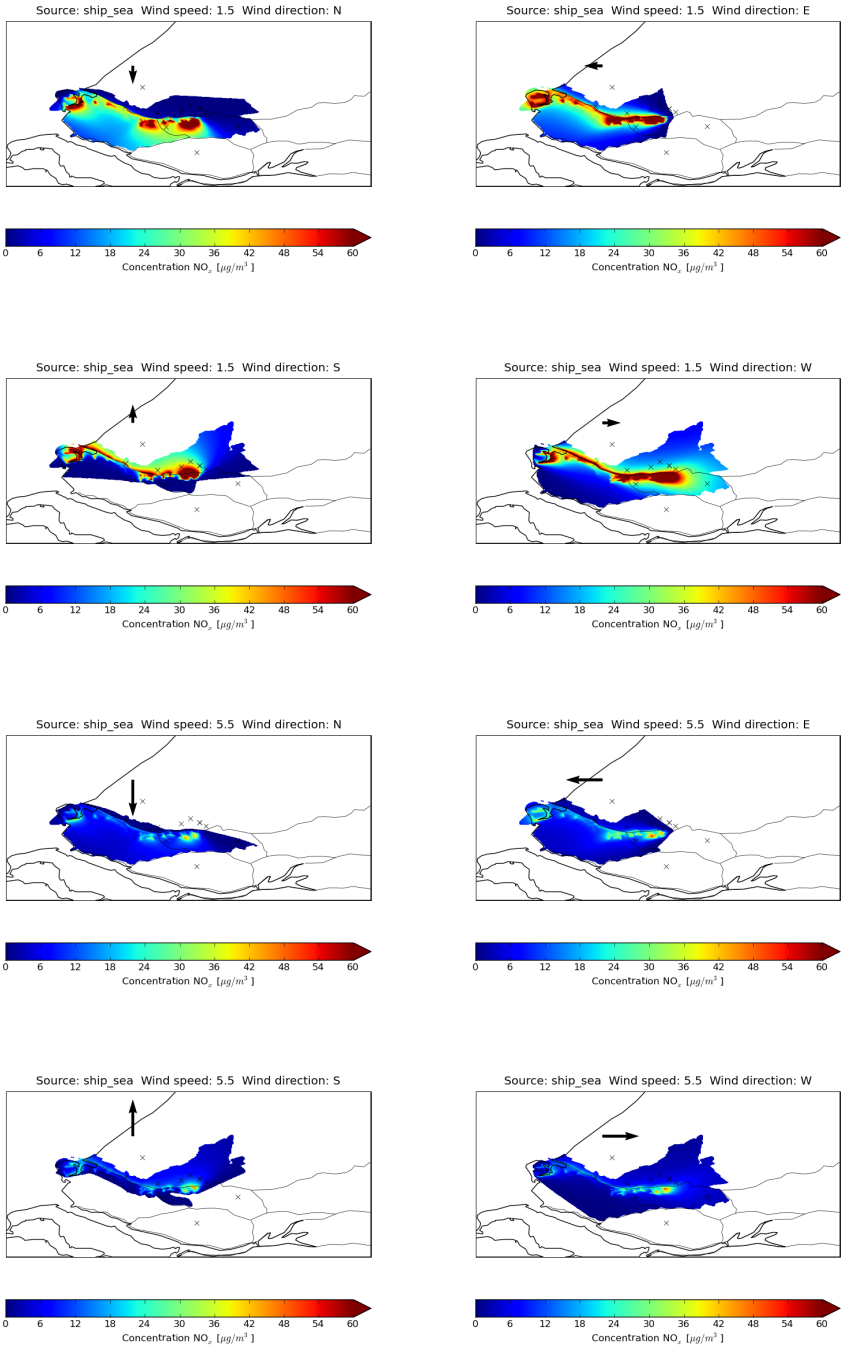


Figure A.10: *Emission Source: Ships sea*

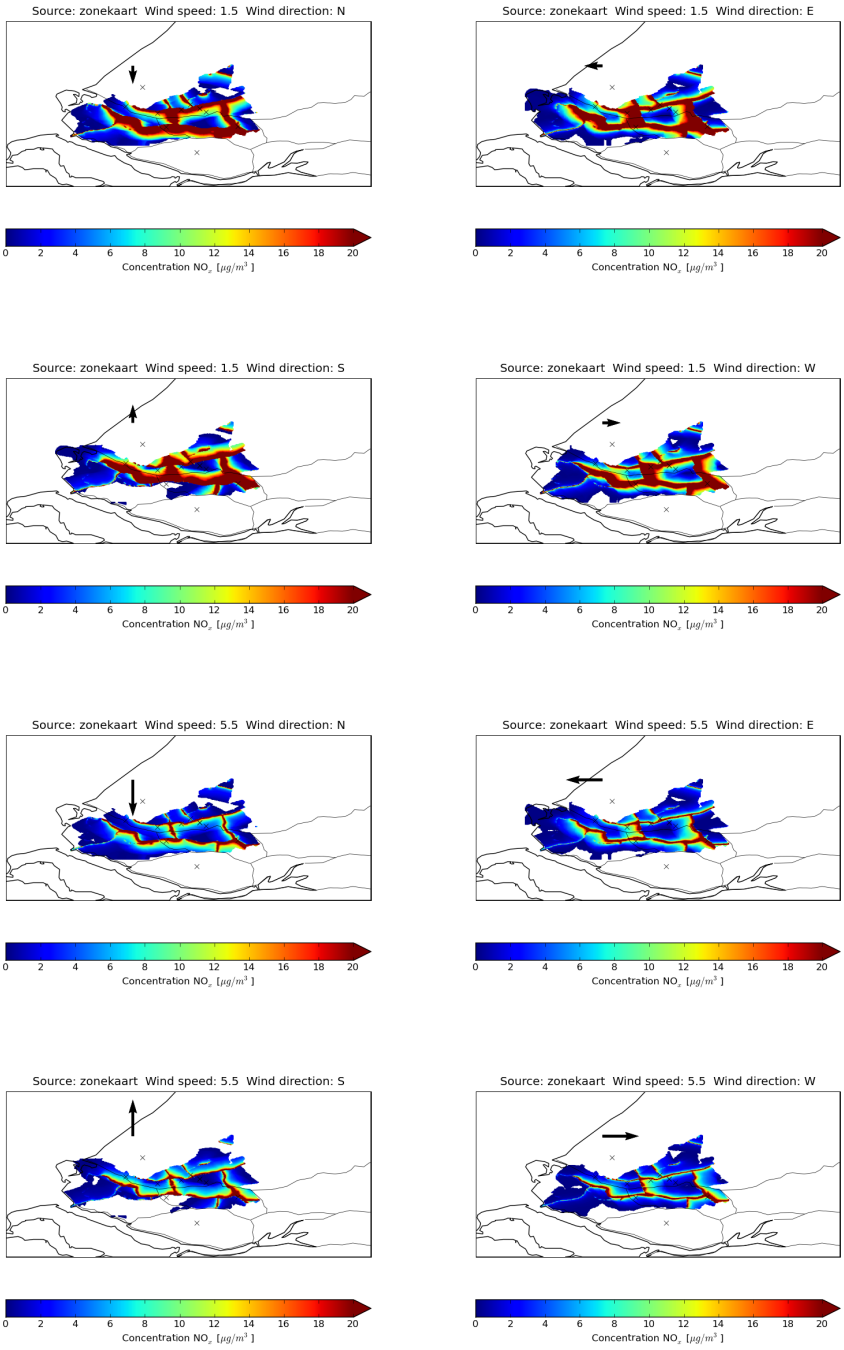


Figure A.11: *Emission Source: Zone Cards*