Royal Netherlands
Meteorological Institute
Ministry of Infrastructure and the Environment

Statistical post processing of model output from the air quality model LOTOS-EUROS

Annemiek Pijnappel

De Bilt, 2011 | Stageverslag

Koninklijk Nederlands
Meteorologisch Instituut
Ministerie van Infrastructuur en Milieu

Statistical post processing of model output from the air quality model LOTOS-EUROS

Versie 1.0

Datum 24 november 2011
Status Definitief

Universiteit Utrecht

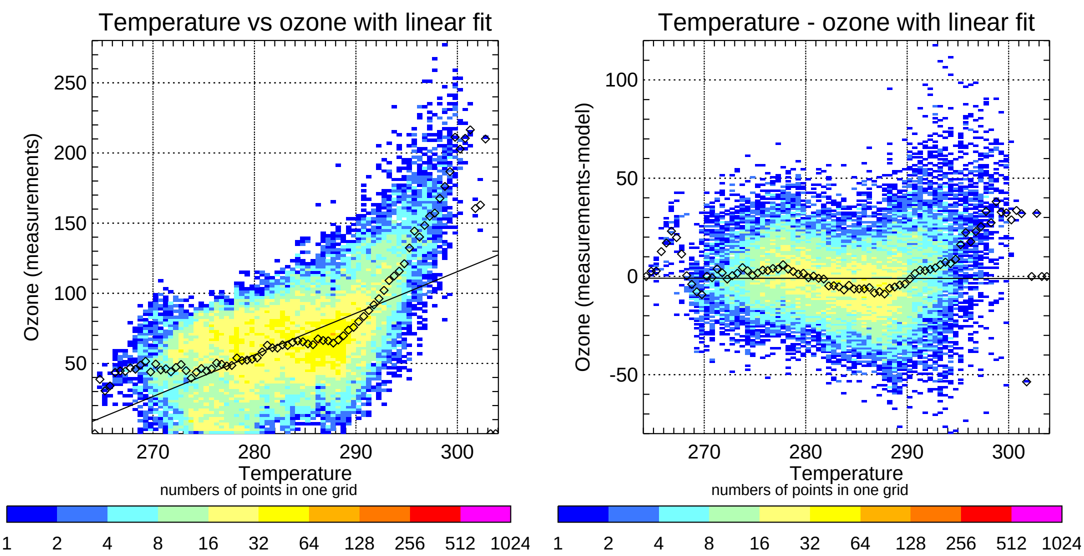
Royal Netherlands Meteorological Institute
Ministry of Infrastructure and the Environment

Statistical post processing of model output from the air quality model LOTOS-EUROS

Author:
Annemiek Pijnappel

Supervisor at KNMI:
Henk Eskes

Supervisor at UU:
Maarten Krol



November 24, 2011

Executive summary

Air quality forecasts are produced routinely, focusing on concentrations of polluting gases and particles up to three days ahead. These air quality forecasts are useful for people which are sensitive for pollution, e.g. COPD and Asthma patients. Long-term exposure to air pollution is harmful for our health, but also for animals and crop plants. Also the EU legislation, which include rules for air pollution e.g. calls for air quality forecasts. In The Netherlands the LOTOS-EUROS model is used to forecast the concentration of PM_{10} and ozone. This model is based on a main equation which describes processes such as transport, diffusion, chemistry, dry and wet deposition and emissions.

The air quality forecasts contain a simple bias correction. For PM_{10} a simple scaling factor is applied and for ozone the bias correction is temperature dependent. The results of this research project are given at four rural stations in The Netherlands, Kollumerwaard, Eibergen, Vredepeel and De Zilk.

The aim of my research is to improve the LOTOS-EUROS model output for daily mean concentration of PM_{10} and daily maximum concentration of ozone through statistical post processing. The approach is based on the Model Output Statistics (MOS) method, which consists of a multiple linear regression between a dependent variable and a couple of independent variables, thereby the error is minimized. In this study the dependent variable is the measured data of PM_{10}/ozone provided by the National Air Quality Monitoring Network of The Netherlands (LML). The independent variables are the modeled data of PM_{10}/ozone provided by LOTOS-EUROS, the meteorological data provide by European Center for Medium-Range Weather Forecasts (ECMWF), components of PM_{10} and precursor trace gases. The regressions were performed based on a multi-years run of the LOTOS-EUROS model for 2003-2005 and the results were checked for data of the year 2006. The regression coefficients of the multiple linear regression are calculated with R, a language for statistical computing. The routine STEP in R is used to remove variables from the regression, which are not significant enough. STEP is based on the Akaike Information criterion (AIC). The performances of the forecasts were evaluated by the correlation (R^2), the root mean square error (RMSE) and the missing, false and well modelled alarms. The two main research questions and corresponding results are:

Can the LOTOS-EUROS model output be improved for PM_{10} through statistical post processing?

For PM_{10} the LOTOS-EUROS model output is improved substantially. The R^2 increases from 0.50-0.64 to 0.69-0.75 for the years 2003-2005. These highest R^2 are reached with configuration H, which include the model, persistence, a couple of meteorological parameters, components of PM_{10} and precursor trace gases. This configuration H gives also the lowest RMSE, between 6.9-7.9 $\mu g\ m^{-3}$ for the years 2003-2005, compared to values between 10.0-10.9 $\mu g\ m^{-3}$ for the model. These results are similar compared with a study presented by Konovalov (2009), were R^2 also increases and the RMSE decreases. Konovalov showed also the more variables are in the regression the better the improvement, this is also shown in this study. The LOTOS-EUROS model shows better scores for the stations in The Netherlands than the CHIMERE model for the European stations used by Konovalov. The model, persistence and sea salt are kept in each regression. At 75% of the stations the boundary layer height, temperature, nitrogen dioxide and sulfur dioxide are kept in the regression.

In 2006 the difference between with or without including CTM tracers in the regression are small. If CTM tracers are included higher R^2s and lower RMSEs are reached. For instance , the maximum decrease of the RMSE is 8% between with or without the use of CTM tracers.

Dividing of the data into summer and winter or based on the boundary layer height and the wind speed did not lead to a significant improvement compared to configuration H.

Overall the regression with the largest improvement of the LOTOS-EUROS model output for PM_{10} is configuration H.

Can the LOTOS-EUROS model output be improved for ozone through statistical post processing?

For ozone the LOTOS-EUROS model output is also improved substantially. The R^2 increases from 0.64-0.76 to 0.81-0.86 for the years 2003-2005. The RMSE decreases from 15.6-19.6 μg m^{-3} to 12.2-15.0 μg m^{-3}, this is reached by configuration H. The variables kept in each regression are the model, persistence and the temperature. The total cloud cover, boundary layer height, wind speed, nitric acid and sea salt are kept in the regression at 75% of the stations. For the year 2006, the difference between with or without including CTM tracers in the regression are small, only the RMSE shows significant lower values if CTM tracers are used.

Dividing the data into summer and winter or a division based on the boundary layer height and the wind speed did not lead to a significant improvement compared to configuration H. Dividing data based on the temperature at the hour of ozone maximum shows better results for the years 2003-2005. The data is divided into two groups, group 1 include the data where the temperature at the ozone maximum is equal or greater than 20 degrees Celsius and group 2 include the data where the temperature is lower than 20 degrees Celsius. In 2006 this does not improve the results for all stations. For instance, at station De Zilk, the RMSE is increasing and there appear peaks at points where it should not occur. For station Eibergen it has a positive impact, the RMSE is lower and the missing exceedances are decreased compared to configuration H. Also the missing exceedances at station Vredepeel and Kollumerwaard are decreased compared to H.

Overall the best results were obtained when the LOTOS-EUROS model output for ozone is to divide the data based on temperature, followed by a multiple linear regression with configuration H for each group.

Further studies are needed to deal with the variability and space between stations. The multiple linear regression is done for each station separately. As final result regression coefficients are needed for all locations, not only at the stations. Without results in between stations these results can not correct the model.

But this is also needed to decide if persistence is needed. If persistence is used, variability will be reduced in the final result. The inclusion of persistence improves the model output significantly, at all stations over the years 2003-2005 and also for the year 2006.

And further studies are also needed to optimize the exceedances. It is important to catch the exceedances for PM_{10}, concentration above 50 μg m^{-3}, and for ozone, maximum 8-hourly mean concentration above 120 μg m^{-3} or hourly concentrations above 180 μg m^{-3}, related to air quality regulations.

Contents

1 Introduction

Air quality research studies describe the concentration of air pollutants in the air we breathe. Air quality is dependent on the presence of pollutants that are harmful for human health, animals and crop plants. Examples of pollutants are sulfur dioxide (SO_2), nitrogen oxides (NO_x), ammonia (NH_3), Ozone (O_3), volatile organic compounds (VOCs) and Particulate matter (PM). Sources of these pollutants are for instance industry, traffic or agriculture [*Jacob, 1999*].

World wide, air pollution is very high in a number of Asian cities (Karachi, New Delhi, Katmandu and Bejing), in Latin American cities (Lima, Arequipa), and in Africa (Cairo) [*World Health Organization*]. In China a third of 113 cities studied failed to meet national air quality standards in 2009. According to the World Bank 16 of the 20 cities of the world with the worst air pollution are located in China. A fifth of urban Chinese breath heavily polluted air. Many places smell like high-sulfur coal. Only a third of the 340 Chinese cities that are monitored meet China's own pollution standards [*China's Ministry of Environmental Protection*].

The issue of air pollution is still a major concern for many European citizens. As well it is one of the areas in which the European Union has been most active. Since the early 1970s, the EU has been working to improve air quality by controlling emissions of harmful substances into the atmosphere, by improving fuel quality and integrating environmental protection requirements into the transport and energy sectors, such as particle filter or a catalyst. But there are also rules for concentration of O_3, PM and NO_2 in the air. As the result of EU legislation, much progress has been made in tackling air pollutants such as sulphur dioxide, lead, nitrogen oxides, carbon monoxide and benzene. However, despite a reduction in some harmful emissions, air pollution continues to cause problems. Summer smog, originating in potentially harmful ground level ozone, regularly exceeds safe limits. Fine particulates present a major health risk which is of increasing concern. Clearly, more needs to be done at local, national, European and international level [*European Commission Environment and WHO rapport (2005)*].

The problem of air pollution in the Netherlands were recognized before the last war, but was not considered to be of great importance. There was some interest in the nuisance caused by fly-ash and soot and this resulted in recommendations regarding the height of industrial chimneys. But the interest in the effects of pollutants increased and offers a sound basis for studies on the effects of separate pollutants. Nowadays in the Netherlands the largest limit exceedances occur for NO_x and PM_{10} [*PBL reports*].

Air quality is influenced not only by how much pollution is emitted into the air, but also by meteorological factors. Weather and climate determine how air circulates and the degree to which pollution builds up and determines most of the day-to-day variability. During winter, high pressure systems lead to cold temperatures, stagnant air and a build up of pollutants in the air near the surface. Low pressure systems bring winds and/or precipitation, which disperse air pollutants and increases deposition of air pollutants. High temperatures lead to an enhanced ozone production [*Clean Air Agency of Puget Sound*]. In the Netherlands a northwestern wind direction carries along relatively clean sea air. Temperature and wind direction are not the only parameters that influence the air quality. The boundary layer height, relative humidity, rainfall, cloud cover, wind speed, temperature at the surface, relative humidity on the surface and wind speed at the surface also influence the air quality.

Aerosols and their precursor trace gases have a negative impact on air quality. Also chemical reactions occur at the surface and in the bulk of solid and liquid aerosol particles which has a negative impact. It can change the property of the particles but also their effects on climate ant human health [*Pöschl (2005)*]. When meteorological and emission data are available for a region, its air pollution levels can be modeled. Models for air quality are also important for RIVM which inform the population about the air pollution levels. Air quality models, or chemistry transport models, use mathematical and numerical techniques to simulate the physical, chemical and meteorological processes that affect air pollutants as they disperse and react in the atmosphere. Based on inputs of meteorological data and information about emission rates, these models are designed to characterize primary pollutants, that are emitted directly into the atmosphere, and secondary pollutants, that are formed as a result of complex chemical reactions within the atmosphere [*U.S. EPA*]. Worldwide there are many different air quality models, including statistical models and deterministic models. Statistical models are based on time series of past measurements in order to define associations between meteorological conditions and an air pollutant. Deterministic models are the chemistry transport models (CTM) mentioned above. Nowadays, the performance CTMs has increased significantly compared to the early nineties due to a better understanding of ozone and PM formation, sinks, meteorology, emissions and the ability to include more complex process descriptions at higher spatial resolution due to the growth in computer power [*Honoré et al., 2008*]. Examples of CTMs are MOZART (developed in the U.S.), GEM-MACH (Canada), CHIMÈRE and MOCAGE (France) and EURAD (German). The air quality model for the Netherlands is LOTOS-EUROS [*Stern et al. (2008)*].

To improve the air quality models, the simulations of the model are evaluated with air pollution measurements. In the Netherlands, the National Institute for Public Health and the Environment (RIVM) monitors the air quality every hour at several regional, street and city locations. The RIVM also publishes three day forecasts of Ozone and PM_{10} for civil authorities and the public based on the LOTOS-EUROS model [*LML*].

The aim of this research project is to improve model output for PM_{10} and ozone, through statistical post processing with the Model Output Statistic method. The evaluation is done over the years 2003 untill 2006, were 2003 untill 2005 are used as trainings set, so all the calculations are done on this set, and 2006 is used as control set.
In this study the following research questions are answered:

- Can the LOTOS-EUROS model output be improved for PM_{10} through statistical post processing?
- Can the LOTOS-EUROS model output be improved for Ozone through statistical post proces-sing?

This report is organized as follow. In section 2 the measurements of the air quality is introduced and detailed information about PM_{10} and ozone is given. In section 3 and 4 describe the history of the Lotos-Euros model and the Lotos-Euros model itself. After that, in section 5, a summary is given of research that has already been done on this subject. Section 6 describes the NMDC project, in the context of which this research was done. The tools which were used for this research project are described in section 7 and the method description is found in section 8. After these introductory sections, sections 9 and 10 contain the results for respectively PM_{10} and ozone. In section 11, the assumptions and results will be discussed. The conclusions are given in the last section, section 12.

2 Air pollution measurements

To monitor the air pollution levels and control the standards in the Netherlands monitoring data is used for 2003-2006 from the Landelijk Meetnet Luchtkwaliteit (LML, National Air Quality Monitoring Network) as operated by the RIVM. The LML stations provide hourly data for ozone and daily mean concentration for PM_{10}. For continuous monitoring there are about 60 permanent stations, spread over the entire country. Some stations play an international role for e.g. determining the long- distance transport of species. There are 57 stations in the Netherlands who measured the concentrations of PM_{10} over the years 2003-2006. For ozone this is smaller, only 34 stations in the Netherlands have measured the concentrations of ozone for the data set 2003-2006 [*LML*].

There are three kind of stations, rural, suburban or urban stations. For the evaluation of the Lotos-Euros model only rural stations are used, these are at some distances from towns and highways. Figure 1 shows which stations were used for our study.

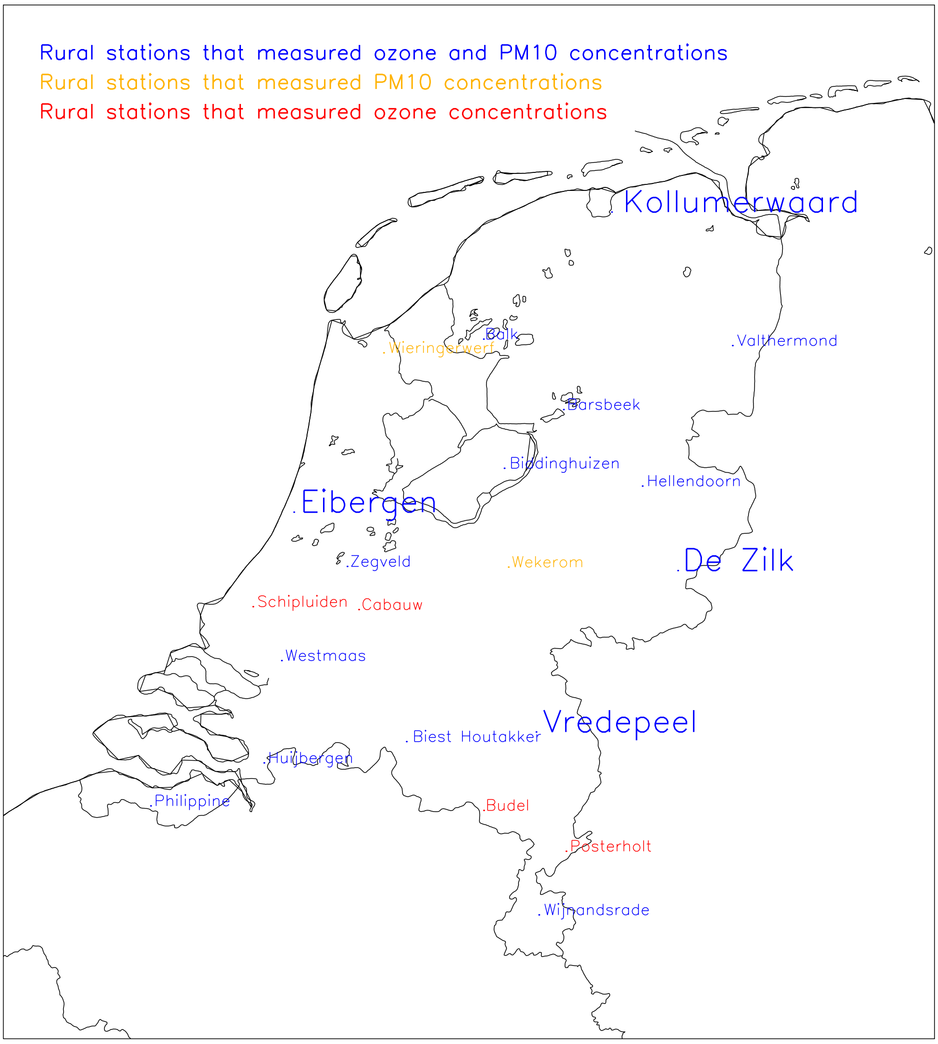


Figure 1: Rural stations that appear in both the measured and modeled data

2.1 PM_{10}

Particles or particulate matter are materials in the air, which are part of air pollution. Particles can occur in almost any shape or size, and can be solid particles or liquid droplets. There are two groups, PM_{10}, all particles smaller then 10 μg m^{-3} in diameter, and $PM_{2.5}$, all particles smaller then 2.5 μg m^{-3}. In this study PM_{10} is used, which also include $PM_{2.5}$. [*Air Info Now*].

In models, the PM_{10} concentration is computed as the sum of several components, such as sulphate, nitrate, ammonia, primary particulate matter (PPM), black carbon (BC) and sea salt. Sources of PM_{10} are very diverse, half of the concentration are of natural origin such as forest and grassland fires, living vegetation, mineral dust, water and sea salt. But also human activities such as the burning of fossil fuels in vehicles, power plants and various industrial processes generate a significant amounts of particulates. One third of the concentration of PM_{10} consist of particles which are chemically formed in the air from gaseous precursors, such as SO_2, NH_3, NO_x and organic compounds. So the PM_{10} concentrations is the result of many different emissions and processes and also the contribution from long-range transport plays an important role [*Velders G.J.M. (2009)*].

The EU regulations include limit values for several components. The most stringent limit values for the Netherlands are for PM_{10} and NO_2. The limit value of PM_{10} is 50 μg m^{-3} for the daily average concentration, not to be exceeded more than 35 times a year. There is also an annual limit value of 40 μg m^{-3} which also not may be exceeded [*ECE*]. In most parts of the Netherlands, the concentrations of PM_{10} are below the EU limit. The exceedances occur at specific locations, mostly close to a number of industrial sites and stock farms [*Velders (2009)*].

2.2 Ozone

Ozone (O_3) occurs both in the stratosphere, troposphere and at ground-level. Ozone in the stratosphere, between 10 to 50 kilometer, is good for our health: it protects us and other life from the sun's harmful ultraviolet (UV) rays. Ozone in the troposphere, extends up to a level of 10 kilometers, is a greenhouse gas and ozone near the surface is a harmful air pollutant. Ozone is not directly emitted in the air but it formed under influence of sunlight through chemical reactions involving NO_x and volatile organic compounds (VOCs). Some stratospheric ozone is transported into the troposphere, and some VOC and NO_x occur naturally, but the majority of ozone at ground-level is the result of reactions of VOC and NO_x. Examples of sources of VOC are chemical plants, automobile emissions, gasoline pump, oil-based paints, forests, grasslands and swamps. NO_x result primarily from high temperature combustion. Examples of sources are power plants, industrial furnaces and boilers and automobiles [*U.S. EPA*].

The concentrations of ground-level ozone are higher during periods with high temperatures. Also the concentration over the day shows a pattern, with a maximum later in the afternoon and a minimum around sunrise. High ozone concentrations generally occur in the Netherlands if the weather situation causes the air in the Netherlands to be transported from continental Europe. This often occurs at stagnant weather systems during the summer, where under the influence of sunshine and high temperatures much ozone is formed. The ozone accumulates and because the residence time of ozone in the lower parts of the troposphere is a few days, ozone can be transported over long distances to and from the Netherlands. During these episodes, ozone contributes to smog or haze. In the south and east of the Netherlands occur most of smog days. This is due to higher average temperatures and the supply of ozone-rich air from the southern and eastern parts of Europe [*Velders G.J.M. (2009)*].

The limit value of ozone is 120 μg m^{-3} for maximum daily 8 hour average concentration, not to be exceeded more than 25 days per year. There is also an information and alert threshold, at respectively 180 and 240 μg m^{-3} for one hour average [*ECE*].

3 PROZON and PROPART

Before 2009 the RIVM used two statistical models for forecasting air quality, called PROZON [*Noordijk, H. (2003)*] and PROPART [*Noordijk, H. (2003)*].

PROZON is used since 1992 and is a forecast model for ozone. This model gives a forecast of the maximum hourly average of the ozone concentration for the next day or some days after that day. The model is based on the following factors to divide the statistics into classes: station type (rural, suburban or urban), season, concentration level and temperature. The statistics shows an increase in ozone concentration with increasing temperature [*Noordijk, H. (2003)*].

PROPART is used since 1996 for the operational 1 day PM_{10} forecast in the Netherlands. This model was developed at RIVM. The goal of this model is to forecast the daily mean concentration of PM_{10} for tomorrow. This concentration can be constructed from today's observed concentration by multiplying it by a factor. This factor was constructed from meas-measurements from the past and depends on today's observed concentration, the station type and today's meteorological conditions and their forecast. The meteorological variables that are used in PROPART are wind speed, wind direction, temperature, rainfall and rain duration. Each variable is subdivided into classes to model its impact on the forecast. In this way a decision tree is made, the contribution of each variable and subclass is determined on a statistical basis. For the meteorological variables, one value per meteorological variable per day is used for the whole country [*Noordijk, H. (2003)*].

4 The LOTOS-EUROS model

4.1 History

At the same time as PROZON and PROPART there were also two operational models, the LOTOS model, stands for LOng Term Ozone Simulation, and the EUROS model, EURopean Ozone Simulation. They were independently developed by respectively TNO and RIVM.

The first version of the LOTOS model was available in 1988 and was focussing on ozone. In 1999 aerosols and data assimilation schemes were included in the model.

The EUROS model was developed in 1990 for winter smog periods. Later ozone, persistent organic pollutants and data assimilation were included in the model.

Both models cover Europe with the same grid resolution and treat the chemistry with a modified version of the Carbon Bond Mechanism IV (CBM-IV) in combination with a thermodynamic equilibrium module for semi-volatile aerosol species. The projection, meteorological data and the technical structure were different in the two models. RIVM and TNO came to an agreement and so both models were integrated into one common LOTOS-EUROS model version 1.1 [*Schaap, M. (2005)*]. In January 2011 the latest model version 1.7 was released, the results of which are used in this study [*Segers, A. (2011)*].

4.2 The LOTOS-EUROS modelling system

Below an overview of the LOTOS-EUROS modelling system is given[*Schaap, M. (2008)*].
The master domain of the LOTOS-EUROS model is from 35° till 70° North and from 10° West till 60° East. The standard grid resolution is 0.50° longitude by 0.25° latitude, this is approximately 30 by 30 km. It is possible to increase or decrease the resolution up to respectively a factor 8 or 2.

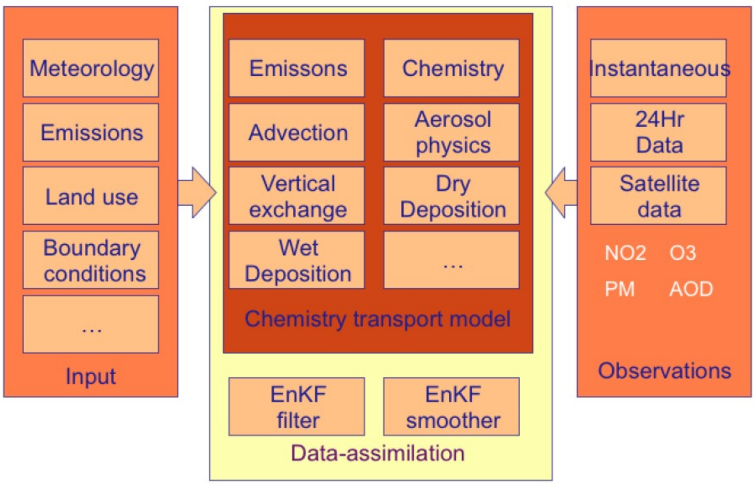


Figure 2: Schematic overview of the main building blocks of the LOTOS-EUROS model.

The block in the middle in figure 4.2 contains the main prognostic equation which describes de change in time of the concentration of a component as a result of transport and diffusion, chemistry, dry and wet deposition, emissions and entrainment,

$$\frac{\delta C}{\delta t}+U\frac{\delta C}{\delta x}+V\frac{\delta C}{\delta y}+W\frac{\delta C}{\delta z}=\frac{\delta}{\delta x}\left(K_h\frac{\delta C}{\delta x}\right)+\frac{\delta}{\delta y}\left(K_h\frac{\delta C}{\delta y}\right)+\frac{\delta}{\delta z}\left(K_z\frac{\delta C}{\delta z}\right)+E+R+Q-D-W \tag{1}$$

C is the concentration of a pollutant, U,V and W are the large scale wind components in respectively west-east direction, in south-north direction and in vertical direction. K_h and K_z are the horizontal and vertical turbulent diffusion coefficients. E is the entrainment or detrainment due to variations in layer height, Q is the contribution by emissions and D and W are loss terms due to processes of dry and wet deposition respectively.

Transport occurs in three dimensions, consisting of horizontal and vertical diffusion and entrainment processes. Entrainment is caused by the growth of the mixing layer during the day. Chemistry is described with two chemical mechanisms, the TNO CBM-IV scheme and the EB90. The TNO CBM-IV is a modified version of the original CBM-IV. The TNO scheme includes 28 species and 66 reactions, including 12 photolytic reactions. The dry deposition is parametrized following the resistance approach and the wet deposition is described using simple scavenging coefficients for gases and for particles.

The LOTOS-EUROS model is equipped with a data assimilation package. For this observations (box on the right in figure 4.2) are needed. In data assimilation, the modeled concentration of yesterday is compared with the observed value of yesterday and the processes in the model are changed to bring the modeled value in better agreement with the observed one.

The left block in figure 4.2 shows the input data. The meteorological input data of the LOTOS-EUROS system includes 3D fields of wind direction, wind speed, temperature, humidity and density and 2D fields of mixing layer height, precipitation rates, cloud cover, several boundary layer and surface variables. The data sets studied are produced with meteorological data obtained from ECMWF (European Center for Medium-Range Weather Forecasts).

The input data also contains the anthropogenic emissions, a combination of the TNO emission database and the CAFE baseline emissions for 2000. The land use data in LOTOS-EUROS are derived from land use database PELINDA and the IIASA database for Russia. The boundary conditions for trace gases are obtained from the MOZART global model used in the MACC (Monitoring Atmospheric Composition and Climate) project [*MACC*]. Model top boundary concentrations were set to 0.8 μgm^{-3} for sulphate and ammonium was set to neutralize the sulphate. Other aerosol species were set to zero.
The LOTOS-EUROS modelling system contains the following components:

- Oxidants: O_3, VOCs, NO_x, Nitric Acid (HNO_3), etc.
- Secondary Inorganic Aerosol (SIA): Sulphate (SO_4), Nitrate (NO_3) and Ammonium (NH_4)
- Secondary Organic Aerosol (SOA) from terpenes
- Primary aerosol: $PM_{2.5}$, PM_{10}, Black Carbon (BC) and sea salt
- Heavy metals: Cadmium (Cd), Lead (Pb) and other non-volatile metals
- Persistent Organic Pollutants

4.3 The LOTOS-EUROS model output

The output of modeled species is in NetCDF format and can be fully specified in the control file. Also the starting date of the output, the number of layers and the number of components can be set. For this study a model run is done with version 1.7. First on European scale, $\frac{1}{2}^\circ$ longitude by $\frac{1}{4}^\circ$ latitude, and then a zoom scale, $\frac{1}{8}^\circ$ longitude by $\frac{1}{16}^\circ$ latitude. This is done over the years 2003-2006. The modeled concentrations are extracted for 45 stations in the Netherlands.

The model output, used in this study, contains the following compounds: O_3, NO_2, NO, SO_2, NH_3, HNO_3 and aerosol components, such as BC, primary $PM_{2.5}$ and primary PM_{10}, SO_4, NO_3, NH_4, sea salt, total $PM_{2.5}$ and total PM_{10}.

4.4 Operational LOTOS-EUROS forecast

In the operational LOTOS-EUROS forecast a post-processing is applied in the form of a bias correction for PM_{10} and ozone. This bias correction brings the model in better agreement with the measurements.

The bias correction for PM_{10} is based on a simple scaling factor. So the corrected value of PM_{10} is equal to a function F times the modeled concentration of PM_{10}. Were $F = 2.11 + 0.291 * sin(2\pi(d - 319.8)/365$ and d the day of the year [*Ruyter de Wildt, M. de* (2011)]. When the modeled concentration of PM_{10} is used the bias correction is included in the modeled concentration. So whenever "model" is used, the model includes the bias correction.

The bias correction for ozone is based on the temperature. Ozone simulations perform well at medium and low temperatures, so for the modeled temperature at the surface below 20 degrees of Celsius, nothing is done. For the modeled temperature above 20 degrees of Celsius the corrected ozone concentration is equal to a function G plus the modeled ozone concentration. Were $G = -0.00194883 * T_s^2 + 1.86295 * T_s - 35.1348$ and T_s the temperature at the surface [Sauter, F. (2011)]. This bias correction is not included in this study, because it is based on a regression.

5 Literature study

My research is mainly based on two published studies. I will summarize the results of these articles, the results which are relevant for my research. The results for my study were compared with the results of these articles. The first article, 'Combining deterministic and statistical approaches for PM_{10} forecasting in Europe', is from Konovalov et al. written in 2009. The second, 'Predictability of European air quality: Assessment of 3 years of operational forecasts and analysis by the PREV'AIR system', is written by Honoré et al. in 2008.

5.1 Konovalov et al.

The goal of this study was to investigate the prospects of the combined use of deterministic and statistical methods for PM_{10} forecasting in Europe. The deterministic forecasts were produced by CHIMERE, a chemistry transport model (CTM), and the daily PM_{10} data is obtained from the AirBase air quality database of the European Environment Agency (EEA). To realize this combination the Model Output statistics (MOS) method is used. The MOS method uses time series of past measurements in order to define associations between meteorological conditions and PM_{10} concentration.

In the statistical model Konovalov used 7 meteorological parameters: near surface temperature, horizontal wind speed (two components), specific humidity, boundary layer height, optical attenuation due to clouds and precipitation. In his paper he only reported results obtained with classical linear regressions and he used only the background monitors.

Konovalov did this study for the years 2003-2006, he divided the years in a cold season (November-March) and a warm season (May-September) and evaluated the seasons separately. To evaluated the forecasting skills he used the root mean squared error (RMSE) and the coefficient of determination (R^2). Konovalov divided his data randomly in a training and a validation subset and repeated this experiment 10 times. He averaged the 10 validation experiments to evaluate the result.

The CHIMERE model was found to underestimate the observed concentrations both in summer and winter and the best comparison was obtained with rural monitors.

Konovalov has tested various configurations of input variables of the statistical model, because not all the results of these configurations could be presented in his paper, he selected the five most representative configurations of statistical models:

A 7 meteo parameters (D+1)

B 7 meteo parameters (D+1) + PM_{10}^{obs}(D+0)

C PM_{10}^{CTM}(D+1)

D PM_{10}^{CTM}(D+1) + 7 meteo parameters (D+1)

E PM_{10}^{CTM}(D+1) + 7 meteo parameters (D+1) + PM_{10}^{obs}(D+0)

$D+1$ means the forecasts for the day after the current day, so $D+0$ is the current day. obs stands for the observed values and CTM are the output values of the chemical transport model. These parameter subsets are used to describe the observations, $PM_{10}^{obs}(D+1)$, the value that will be observed the day after the current day.

All configurations yield significantly better results than the CTM simulations and combined forecasts (D and E) show better performances than "pure" statistical forecasts (A and B). The best performances are obtained by E, were also $PM_{10}^{obs}(D+0)$ is used. This means that the concentration of the next day e.g. correlated with the concentrations of the current day.

The major part of reduction of the RMSE can be achieved by debiasing and scaling the raw simulations, in configuration C. But to increase the correlation more complicated models are needed. The maximum reduction of RMSE reaches 50% and the R^2 increases from 0.32 to 0.6. In the summer the most important meteorological parameter is temperature and in the winter the boundary layer height.

The performance of forecasts were evaluated by means of the cross-validation method. It was found that the MOS method enables significant improvements of the deterministic forecasts. Also an important finding of Konovalov is that overall the post processed forecasts are better than the raw forecasts not only for a given monitoring site, but also for territories of similar type of environment (rural, suburban or urban) within several hundreds kilometers away from the considered site.

5.2 Honoré et al.

The purpose of this paper is to give a quantitative assessment of the French national air quality forecasting and monitoring system (PREV'AIR). The evaluation was carried out over the years 2004, 2005 and 2006 for three pollutants: O_3, PM_{10} and NO_2.

PREV'AIR is designed to provide forecasts of air quality up to three days ahead. The system ingests input data from various origins and uses numerical models to produce a forecast. The CHIMERE model and the MOCAGE model are run every day and produce routine forecasts within PREV'AIR.

The MOS method was applied in PREV'AIR using a training period, the summers of 2003 to 2005. The forecast error $O_{3,obs}(s) - O_{3,mod}(s)$ is regressed, at each monitoring site s, from the predicted 2m temperature $T_{2m,mod}(s)$ and the predicted ozone daily maximum $O_{3,mod}(s)$. So the estimated forecast error (EFE(s)) will be:

$$O_{3,obs}(s) - O_{3,mod}(s) \approx EFE(s) = \alpha \cdot T_{2m,mod}(s) + \beta \cdot O_{3,mod}(s) + \gamma \tag{2}$$

A different set of multiple regression coefficients is calculated for each site and each forecast lead time. After that MOS daily maximum ozone, $O_{3,MOS}(s)$, is calculated for each monitoring station, as the sum of the ozone forecast and the estimated forecast error:

$$O_{3,MOS}(s) = O_{3,mod}(s) + EFE(s) \tag{3}$$

The MOS forecasts issued for each monitoring station are then interpolated over the whole modeling domain.

The model overestimates the ozone daily maximum over coastal areas and underestimates them over continental, central areas. The agreement is better for rural and for suburban sites than for urban sites. The RMSE averaged over Europe varies between 16.8 and 19.4 μg m^{-3} at rural sites depending on the lead time. The skill smoothly decreases with lead time. It is found that only 6% of the square of the RMSE is due to the meteorological forecast error. The mean correlation varies between 0.76 to 0.84, with decreasing values as the forecast lead time increases. The impact of the MOS procedure is higher when ozone episodes take place: all skill scores are improved whatever the type of station considered.

As future outlook two promising avenues may be mentioned. First, the system performance could be improved by taking observations more fully into account: analyzed concentrations should be integrated in the forecasting chain at the initialization state. Second, ensemble forecasting is a rather promising approach for substantially increasing the performance of air quality forecasting system. This approach is adopted in the MACC project.

6 The NMDC-project

NMDC stands for national models and data center. It is a strategic partnership between KNMI, RIVM, TNO, Deltares, PBL en Alterra. An important motivation for this partnership was the grow of the complexity of issues, the increasing demand for efficient research and development.

The NMDC is a virtual organization in which experts from the six partner institutions are collaborating. In the first phase of the NMDC, fourteen research projects in the areas of Innovation, Integration and Infrastructure are setup. One of the research projects is the learning project LOTOS-EUROS. LOTOS-EUROS describes complex processes such as atmospheric chemistry,transport and the exchange of substances with the surface. Partly because of this complexity and large input/output volumes, the model is computing intensive.

The goal of this project is to provide better and clearer frameworks for research models regarding operationalisation, maintenance, management and policy analysis. With this the management and quality of research models will be more efficient and the LOTOS-EUROS model will be improved [*NMDC*]. This research project will lead to an improvement of the operational LOTOS-EUROS model.

7 Tools for programming

Before moving to description of the method the tools are described which are used for these methods. Three programming languages are used: Fortran 90, IDL and R.

Fortran 90 is a programming language that is especially suited to numeric computation and scientific computing. Originally developed by IBM at their campus in south San Jose, California in the 1950s for scientific and engineering applications. Fortran came to dominate this area of programming early on and has been in use for over half a century in computationally intensive areas such as numerical weather prediction and computational chemistry. It is one of the most popular languages in the area of high-performance computing and is the language used for programs that benchmark and rank the world's fastest supercomputers. Fortran 90 is used mainly for reading, writing and manipulating the LOTOS-EUROS output data, to change hourly values into daily mean data for PM_{10} and into daily maximum for ozone, but also for the meteorological data, components of PM_{10} or ozone and precursor trace gases. [*Fortran*].

IDL(Interactive Data Language) is a programming language used for data analysis and is developed in the 1970s at the Laboratory for Atmospheric and Space Physics (LASP) at the University of Colorado at Boulder. It is popular in particular areas of science, such as astronomy and medical imaging. IDL shares a common syntax with PV-Wave and originated from the same codebase, though the languages have subsequently diverged in detail. IDL is vectorized, numerical, and interactive, and is commonly used for interactive processing of large amounts of data (including image processing). The syntax includes many constructs from Fortran and some from C. IDL is used for the graphics in this report [*IDL*].

R is a language and environment for statistical computing and graphics. R is created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand. R is named partly after the first names of the first two R authors. Now it is developed by the R Development Core Team. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time series analysis, classification, etc.), graphical techniques (can produce well designed publication-quality plots, including mathematical symbols and formulas) and is highly extensible. For this project the routine LM for linear modelling and the routine STEP are used, which will be explained in section 8.1 [*R*].

8 Method description

8.1 Model Output Statistics

The Model Output Statistics (MOS) [Wilks, D.S., (2006)] method is a classical procedure in meteorology used to improve the skill of model forecasts based on multiple linear regression. Multiple linear regression (MLR) is a method used to model the linear relationship between a dependent variable and a couple of independent variables. In our observed concentration of PM_{10} or ozone are needed as accurate as possible:

$$Y(i) = p_0 + \sum_{k=1}^{n} p_k x_{i,k} + \varepsilon_i \tag{4}$$

$Y(i)$ is the dependent variable for the i^{th} day. The dependent variable is sometimes also called the predictand. In this study it is the 24h average concentration of PM_{10} or the daily maximum concentration of ozone for the i^{th} day. $\mathbf{x}$ is a vector of independent variables. The independent variables are called the predictors, which are in this study the modeled concentration of PM_{10} or ozone for i^{th} day, but also the meteorological parameters for the i^{th} day and the measured concentration of PM_{10} or ozone for the $(i-1)^{th}$ day. p_i are the regression constant and coefficients, n is the number of predictors and ε_i is the forecast error of the i^{th} day.

$$\varepsilon_i = Y^{obs}(i) - Y(i) \tag{5}$$

$Y^{obs}(i)$ is the observed concentration of PM_{10} or ozone for the i^{th} day.

MLR is based on least squares: the p_i's in equation (4) are fit such that the sum of squares of ε_i is minimized. In this study the period is three years (2003, 2004 and 2005), or n is equal to 1096 days.

$$\sum_{i=0}^{n} \varepsilon_i^2 = minimum \tag{6}$$

In the process of fitting or estimating the model, statistics are computed that summarize the accuracy of the regression model. The MLR is calculated in R, a language and environment for statistical computing. To know with variables are significant enough for the multiple regression, the routine STEP is used. STEP selects a suitable model by dropping terms with the Akaike Information criterion (AIC). The AIC rewards variables for good fit, but imposes a penalty for unnecessary variables. AIC is calculated by (7) or (8).

$$AIC = -2\ln(L) + 2 \cdot K \tag{7}$$

$$AIC = n \cdot \ln(RSS/n) + 2 \cdot K \tag{8}$$

With $\ln(L)$ the log likelihood function, where the estimaters are the most probable values for the parameters, given the observed data. The parameters of interest are chosen in such a way that the data is most likely. K is the number of parameters in the multiple linear regression. $RSS = \sum(\epsilon_i^2)$, the sum of squares of the estimated residuals from the fitted model and n is the number of measurements. So the routine STEP provides a list of the variables with its coefficients which are significant enough to take into account [*Wilks*].

8.2 Nonlinear regressions

Nonlinear regressions are alternatives for statistical post processing and is used for optimalisation of the thresholds. The logistic regression is one of them. The logistic regression produces probability forecasts. First the predictand, in our case the measured concentration of PM_{10} or ozone, should be transformed in a binary variable, taking on the values zero and one. If the measured concentration is above a certain threshold value the concentration will be set to one and if the concentration is below it will be set to zero. Then the linear regression could be used as described previously.

The major difference between the linear regression and the logistic regression is the output. The logistic regression gives a probability that a measured value will be above a certain limit, while the linear regression what the measured value will be.

If the value of the concentration in the logistic regression is set to the limit value, described in de EU legislation, the logistic regression gives the probability that the limit value will be exceeded. So a better description of when there are peaks occurs should be given. This is an advantage compared to the linear regression, where the focus is on the bulk values. So the peaks will be not necessarily improved.

8.3 Evaluation of the forecasting skills

To Evaluate the MOS output statistical skill scores, root mean square error (RMSE) and correlation (R) are used. Bias indicates if the forecasts are under- or overestimated. Here, negative values indicate underestimation and positive values means overestimation. RMSE gives the skill in predicting the overall magnitude of the observations,

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - Y_i\right)^2}{n}} \tag{9}$$

X_i and Y_i correspond in our case to PM_{10}^{obs} and PM_{10} or O_3^{obs} and O_3 (see eq. (5)). n is number of days.

Correlation is a measure of whether forecasts and observation change in the same way. The closer the correlation is to one, the better the forecasts variability is in agreement with the observations.

$$R = \frac{\sum_{i=1}^{n} \left[\left(X_i - X_i \frac{1}{n} \sum_{i=1}^{n} X_i\right)\left(Y_i - \frac{1}{n} \sum_{i=1}^{n} Y_i\right)\right]}{\sqrt{\sum_{i=1}^{n} \left(X_i - \frac{1}{n} \sum_{i=1}^{n} X_i\right)^2} \sqrt{\sum_{i=1}^{n} \left(Y_i - \frac{1}{n} \sum_{i=1}^{n} Y_i\right)^2}} \tag{10}$$

Here the variables are the same as in (9) [*Wilks*].

The last thing which is checked is the skill of predicting the high concentration of ozone (above 120 μg m^{-3}) or PM_{10} (above 40 μg m^{-3}). This is done by calculating the percentage of modeled days that are also measured, the percentage of measured days that are also modeled, the percentage of false alarms (modeled but not measured) or the percentage of missed events (measured but not modeled). As I said before, the focus of the MOS method is not on the exceedances, so these alarms are not necessarily improved by the MOS method.

9 Results for PM_{10}

9.1 Measurements against meteorological variables

The following meteorological variables are available: temperature, boundary layer height (blh), wind at the surface, temperature at the surface, cloud cover, relative humidity, rainfall, meridional wind speed, zonal wind speed, relative humidity at the surface, wind speed and wind direction. Figure 3 shows scatter density plots of the meteorological variables against PM_{10} measurements. The black line is a linear fit between the meteorological variable and PM_{10} measurements. The diamonds are the mean value of each column.

Figure 3a shows a relation between the boundary layer height and PM_{10}, a reciprocal function. Between wind direction and PM_{10}, figure 3l, the diamonds suggest a sine-like dependence, this is also found by Manders et al (2009) and De Ruyter de Wildt (2011).

The correlation between a meteorological variable and PM_{10} is calculated for each station. The boundary layer height and wind at the surface have a good correlation against PM_{10} measurements, the correlation values are ranging between -0.40 and -0.56. The correlation values for wind at the surface are lower ranging between -0.15 and -0.48.

9.2 Measurements - model against meteorological variables

More interesting is the behavior of the meteorological variables against measurements - model (see figure 4), because improvement of the model output is needed.

The relations shows in figure 4 e.g. the correlations are much weaker. So the dependence on the meteorological variables are well described by the model. The improvement of the model output will be very small. So the model will be used also as independent variable in the multiple linear regression.

Figure 4l suggest also a sine-like dependence with an amplitude much smaller than in figure 3l. If the wind comes from the southwest, 64% of the measured concentration above 40 μg m^{-3} is modeled and if the wind comes from southwest or southeast, 70% of the measured concentration above 40 μg m^{-3} is modeled. For northwestern or northeastern wind this is much lower, 57% of the measured concentration above 40 μg m^{-3} is modeled.

9.3 Model output compared with the measurements

Before the bias correction the correlation between the model and measurements was ranging between 0.62 and 0.76, the RMSE was ranging between 15.5 and 28.4 μg m^{-3}. With the bias correction, the correlation shows a small increase, ranging between 0.71 an 0.80, it is small because the bias correction does not change the variability so much. This is different for the RMSE, ranging between 10.0 and 10.9 μg m^{-3}, which shows a significantly decrease. Figure 5 shows how the modeled PM_{10} concentrations (with bias correction) behave compared to the PM_{10} measurements. Most of the time the LOTOS-EUROS model, with bias correction, underestimates the PM_{10} concentration.

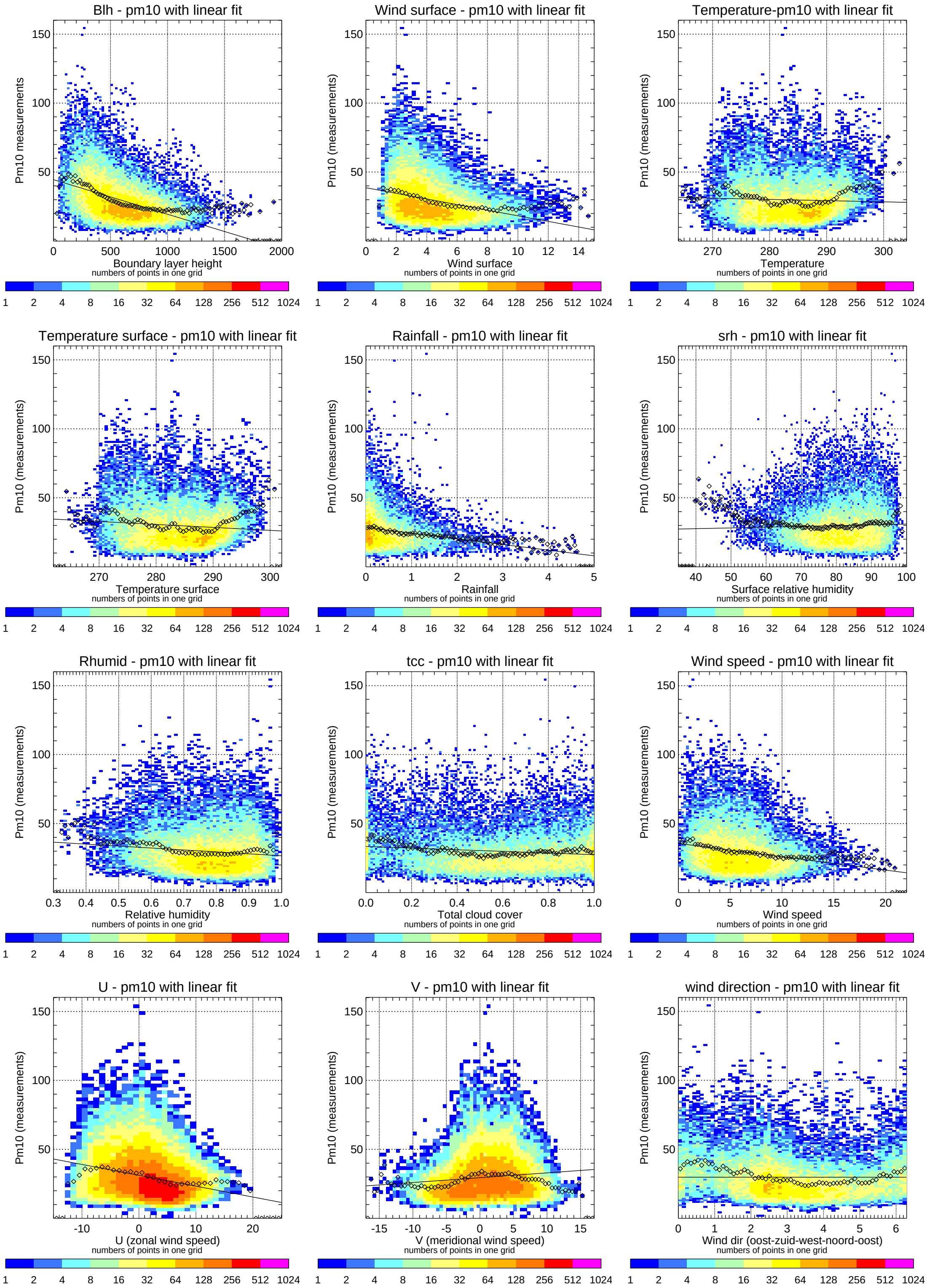


Figure 3: The scatter density plot of PM_{10} concentration in μg m^{-3} against a) boundary layer height in m; b) wind at the surface in m/s; c) temperature in Kelvin; d) temperature at the surface in Kelvin; e) rainfall in mm; f) relative humidity at the surface in %; g) relative humidity, were 1.0 is really high humidity and 0 is zero humidity; h) total cloud cover, were 0 is clear sky and 1.0 is overcast sky; i) wind speed in m/s; j) zonal wind speed in m/s; k) meridional wind speed in m/s; l) wind direction, 0 is east, π is west and 2π is east. The color indicates the density.

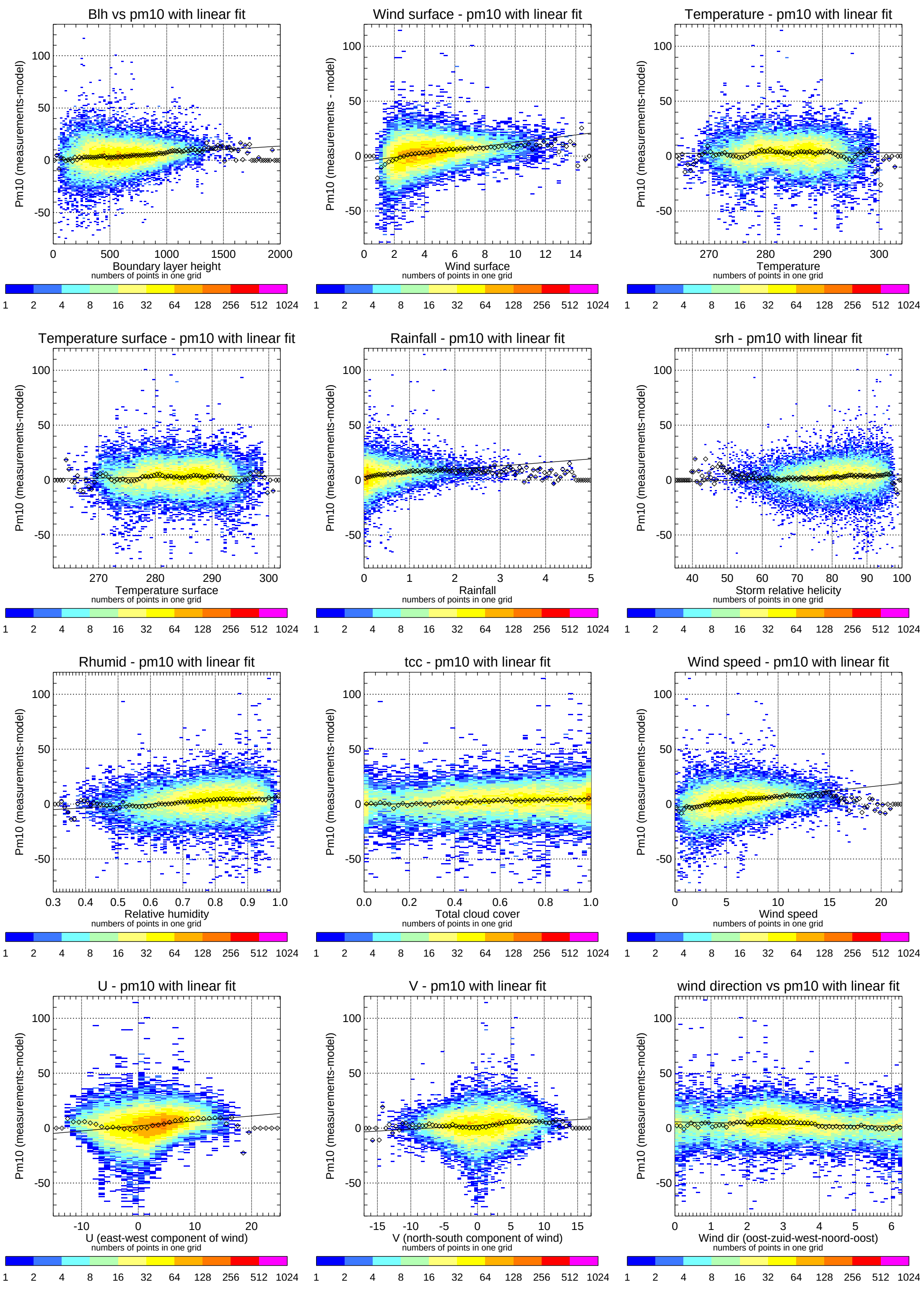


Figure 4: The scatter density plot of measured - modeled PM_{10} concentration, in μg m^{-3}, against a) boundary layer height in m; b) wind at the surface in m/s; c) temperature in Kelvin; d) temperature at the surface in Kelvin; e) rainfall in mm; f) relative humidity at the surface in %; g) relative humidity, were 1.0 is really high humidity and 0 is zero humidity; h) total cloud cover, were 0 is clear sky and 1.0 is overcast sky; i) wind speed in m/s; j) zonal wind speed in m/s; k) meridional wind speed in m/s; l) wind direction, 0 is east, π is west and 2π is east. The color indicates the density.

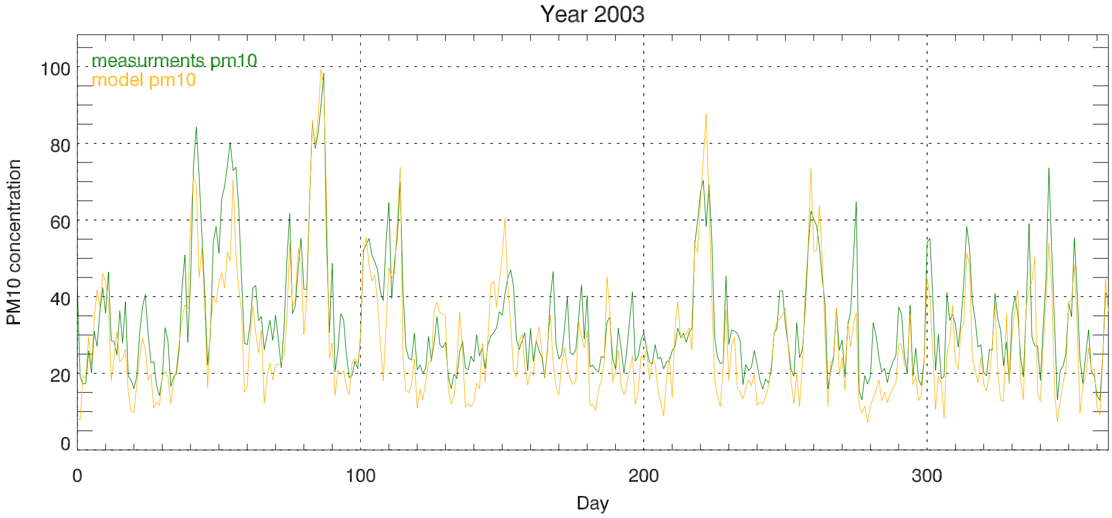


Figure 5: The behavior of the modeled PM_{10} concentrations vs the measured PM_{10} concentrations, in μg m^{-3}

9.4 The parameter configurations

To improve the PM_{10} simulations, eight parameter configurations were selected:

A Meteorological parameters (D+1)

B Meteorological parameters (D+1) + PM_{10}^{obs}(D+0)

D PM_{10}^{CTM}(D+1) + meteorological parameters (D+1)

E PM_{10}^{CTM}(D+1) + meteorological parameters (D+1) + PM_{10}^{obs}(D+0)

F PM_{10}^{CTM}(D+1) + CTM tracers (D+1)

G PM_{10}^{CTM}(D+1) + CTM tracers (D+1) + meteorological parameters (D+1)

H PM_{10}^{CTM}(D+1) + CTM tracers (D+1) + meteorological parameters (D+1) + PM_{10}^{obs}(D+0)

Configuration A and B are the standard statistical models which do not involve CTM simulations. Both employ meteorological parameters and configuration B also involves the PM_{10} daily-mean concentration observed on the past day. C is not a configuration, it involves the forecasts from the LOTOS-EUROS model. Configuration D and E correspond to the combined forecasts. Configuration F,G and H are also combined forecasts, but they include CTM tracers, which include components of PM_{10} and aerosol precursor trace gases.

Which parameters are selected for each configuration depends on the AIK criterion. The terms which are removed by STEP differ for every station and configuration. By normalizing the variables, a dimensionless regression is obtained, e.g. $T_{norm}(k) = \frac{T(k) - mean(T)}{\sqrt{variance(T)}}$, with $T(k)$ the temperature at day k. Appendix 1 contains information about the regression coefficients for each variable in the eight configurations at station Vredepeel, De Zilk, Eibergen and Kollumerwaard.

9.5 Performance of the multiple linear regression

In figure 6 the performance of the regression for the different configurations is shown. Not every configuration shows a better result than the model, configuration C, because the model is not included in each configuration. This happens for configuration A and also almost every time with configuration B. So the model is better than the regression for configurations based on meteorological parameters and measurements of yesterday only.

Figure 6 shows that, the more variables in the regression, the better the improvement will be. If persistence, measurements of yesterday is used a 5-8% better correlation is obtained. The best performance is obtained in configuration H, when statistical models involve the set of meteorological variables, the observed PM_{10} concentration and CTM tracers for a current day. The maximum increase of R^2 goes from 0.5 to 0.71 at station Kollumerwaard, this is a big change.

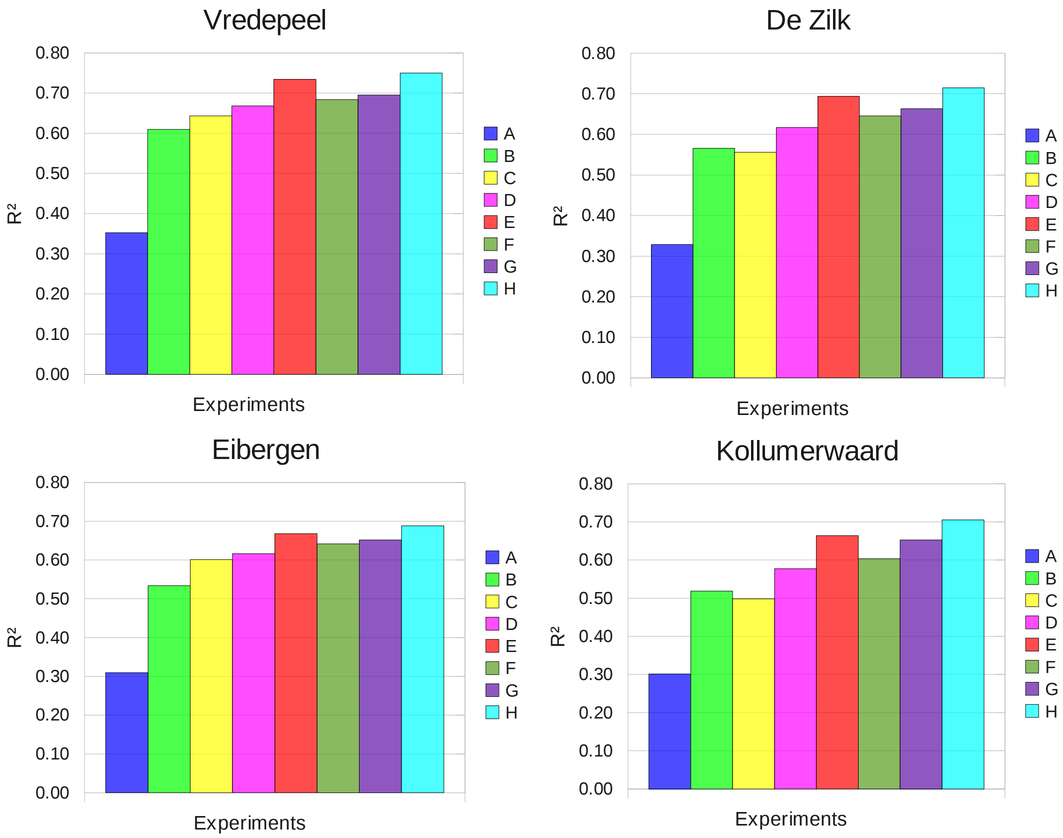


Figure 6: R^2 for the different configurations at station Vredepeel, De Zilk, Eibergen and Kollumerwaard.

In figure 7 the RMSE is given for the different configurations and the same four stations. The maximum reduction of RMSE reaches 29% also for configuration H and station Kollumerwaard. This is also the lowest error with a RMSE of 6.9 μg m^{-3}. At station Eibergen the RMSE is only 12% lower than the model and has the largest error of 7.9 μg m^{-3}.

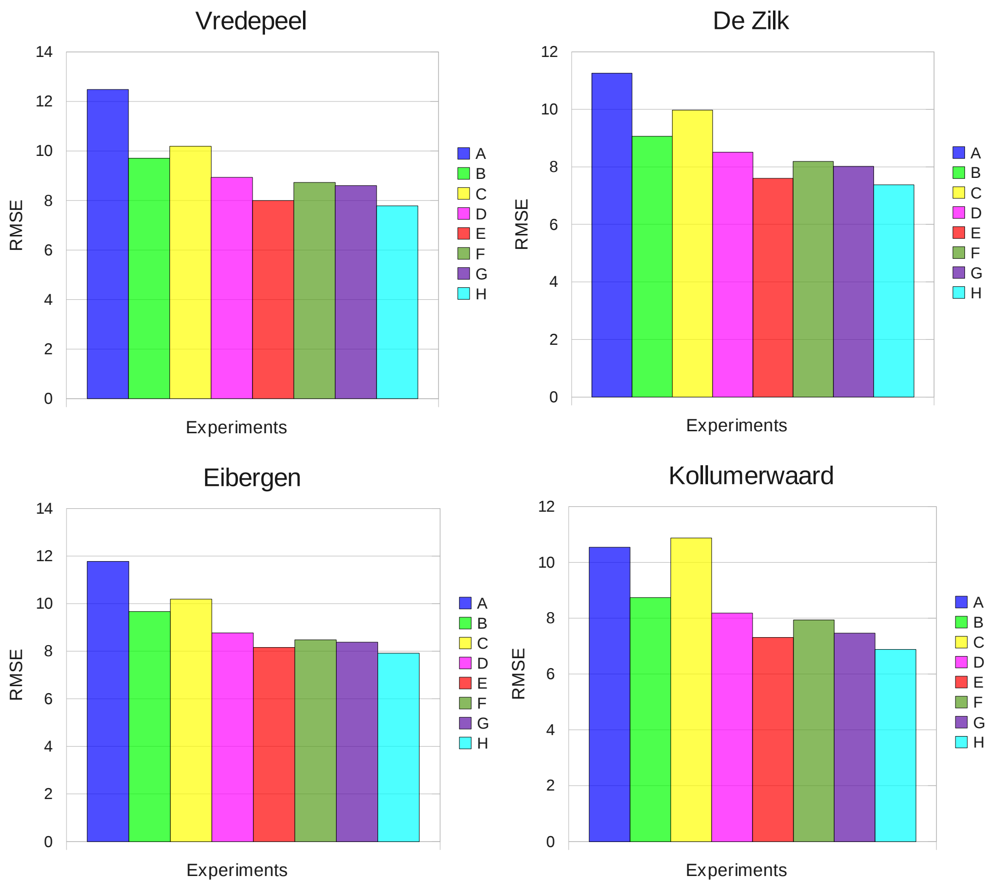


Figure 7: RMSE, in $\mu g\ m^{-3}$, for the different configurations at station Vredepeel, De Zilk, Eibergen and Kollumerwaard.

In STEP some of the parameters in configuration H will be used at each station, such as the model forecast, the measurements of the day before and sea salt. The boundary layer height, temperature, temperature at the surface, nitrogen dioxide and sulfur dioxide are parameters used in the regression at more than 75% of the stations.

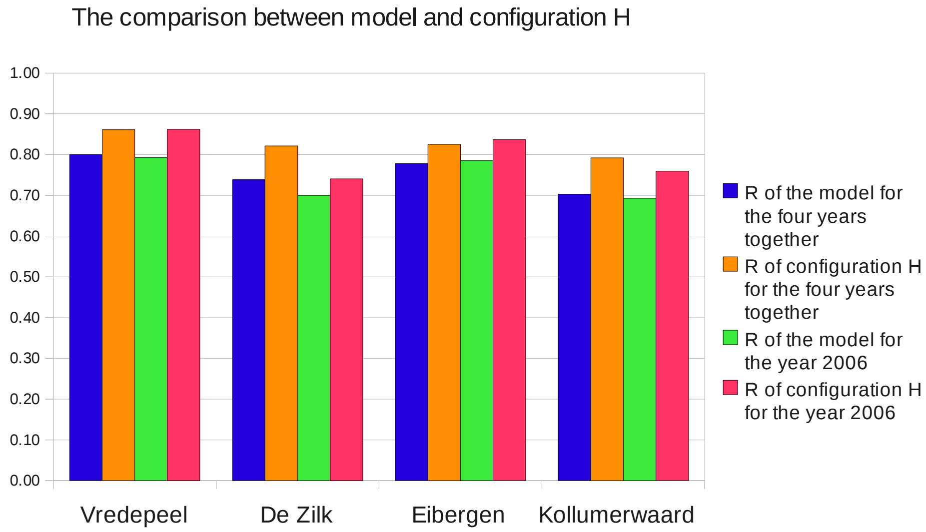
The variables left over by STEP are so different in configuration E and H and also different at each station, see Appendix 1. Therefore it is difficult to say which variable influence at most and which give the largest change.

9.6 Checking the best configuration for the year 2006

The best configuration, configuration H, shows that it improves the model output substantially for 2003-2005. In this section the behavior of configuration H in the year 2006 is checked. But in figure 6 and 7 there is not a big difference between the configuration which does not use CTM tracers (configuration E) and the configuration which use CTM tracers (configuration H). So also the configuration without CTM tracers is checked for the year 2006. These configurations shows the largest improvement in R^2 and the RMSE for 2003 untill 2005. The improvement for the year 2006 is also the largest for configuration E and H.

9.6.1 R^2 and RMSE

When R^2 and the RMSE are calculated for configuration with or without CTM tracers (resp. configuration E and H), the configuration with CTM tracers will be the best configuration for 2006 at most stations. For instance at station Vredepeel the error will decrease with 8% compared to the configuration without CTM tracers. In 2003-2005 the error decrease with 3%. The correlation between with or without CTM tracers does not change significantly at all stations.



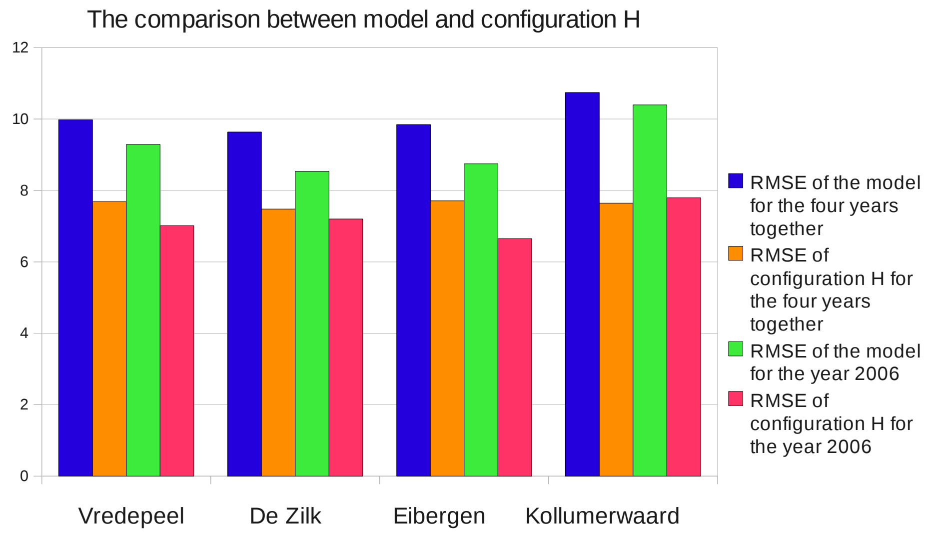


Figure 8: The behavior of the correlation and RMSE for the model and configuration H

Figure 8 shows the largest improvement at station Kollumerwaard: the correlation for 2006 is 0.76 compared to 0.69 for the model. Also for the four years together the correlation at this station has the largest increase, from 0.70 to 0.79. Figure 8 shows that the RMSE for station Kollumerwaard is much better, where it decreases from 10.7 to 7.6 for the four years together and from 10.4 to 7.8 for 2006. The highest correlation for configuration H appears at station Vredepeel, where it increases from 0.80 to 0.86 for the four years together and from 0.79 to 0.84 for 2006. The lowest RMSE for configuration H appears at station Eibergen for 2006, where it decreases from 8.7 to 6.7 μg m^{-3}. While it has the largest RMSE for configuration H in

2003-2005, 7.9 μg m^{-3} (see figure 7). So the regression is robust from year to year.

Figure 9 shows how configuration H behaves compared to the measurements and the model in the year 2006.

9.6.2 Exceedances

Not only the correlation and the error is studied, but also the skill of predicting the high concentration. For PM_{10} these are the concentrations above 40 μg m^{-3}. This is done by calculating the percentage of false alarms or missed alarms.

At station Kollumerwaard the missing alarms for configuration H for the four years together are decreasing with 43% compared to the model. For this station it is not true that the counted false alarms are decreasing, because this configuration models a lot more high concentration values than the model. For that reason more false alarms appear. First the model was to low, it underestimate the measured values. The model gives 101 times high concentration values compared to 178 times of measured high concentration values. Configuration H modeled 182 times high concentration values, much closer to the 178 that were measured.

At station Vredepeel the false alarms for configuration H for the four years together are decreasing with 35% and the missing alarms shows 11% improvement.

Overall configuration H improve the model output at all four stations. For the complete overview of counted alarms see Appendix 2.

9.7 Regressions for data sub-sets

Also the performance is studied of the multiple linear regression by dividing the data sets into groups where a different behavior is expected. In one test the data set is divided into summer (March till September) and winter (September till March), where for April and October there were different options. These months are not clear defined as winter or summer months, so all option are tried. One time April goes with the summer and October with the winter, the other time April with the winter and October with the summer, April and October with the winter or April and October with the summer.

In an other test the data is divided based on the boundary layer height and the wind speed. The data was dividing into three groups, group 1 the data were the boundary layer is higher than 400 meter, group 2 the boundary layer below 400 meter and wind speed above 4 m/s and group 3 where the boundary layer was below 400 meter and the wind speed below 4 m/s.

By checking the correlation, the root mean square error and the figures for 2006 no significant improvement is found compared to the best configuration H. So dividing is not a recommendation.

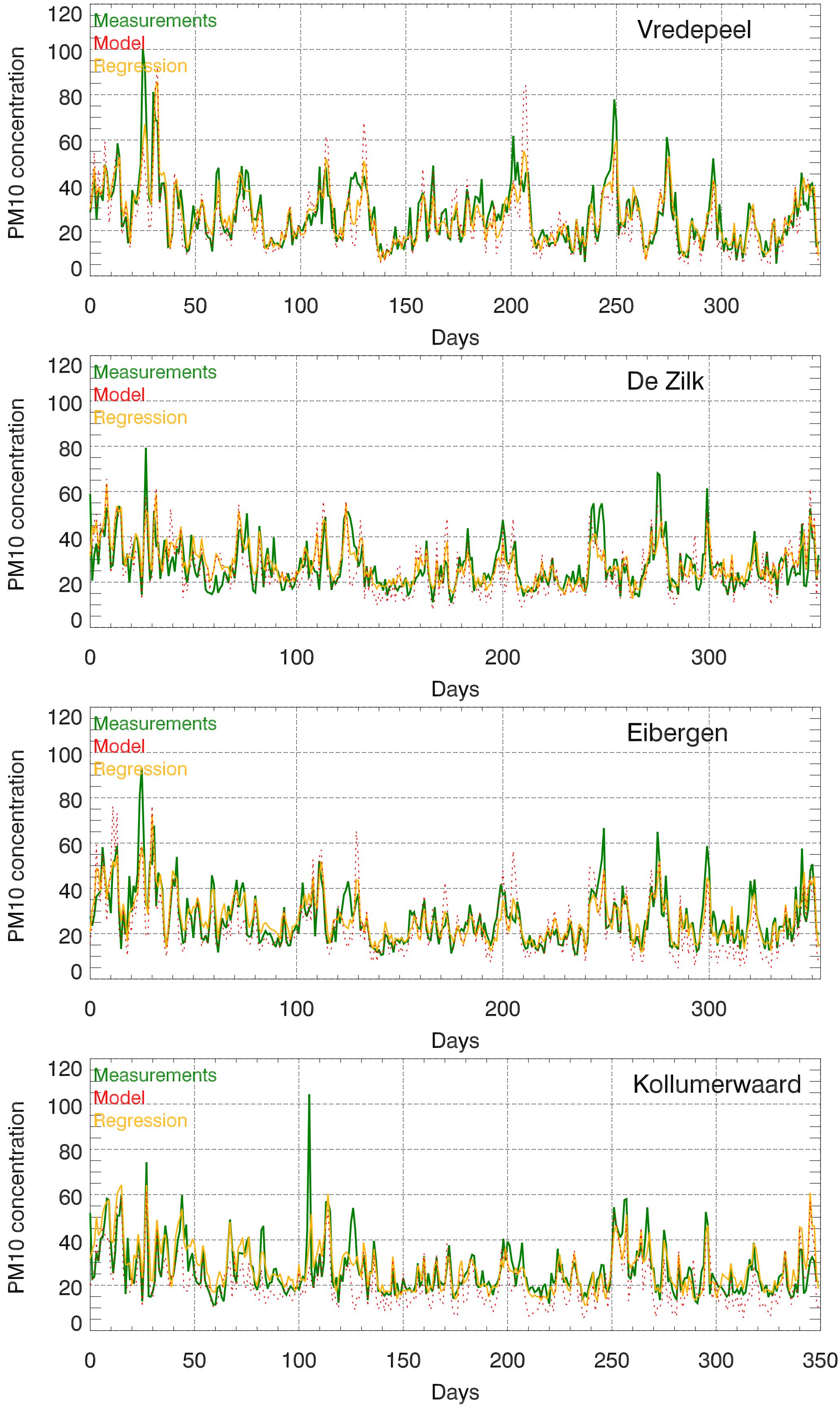


Figure 9: The performance of the model and configuration H in the year 2006

10 Results for ozone

The approach for ozone is similar to that of PM_{10}, but ozone behaves different because other variables influence ozone. For ozone the focus is on the daily maximum concentration, because this varies more strongly from day to day than the daily mean. Secondly the EU legislation has limit values for the concentration at the hour of ozone maximum. The bias correction, see section 4.4, for ozone is not used in this study.

10.1 Measurements against meteorological variables

The meteorological variables are plotted against ozone measurements, to see which variables shows a relation with the measured values. As for PM_{10}, the parameters used are: temperature, boundary layer height (blh), wind at the surface, temperature at the surface, cloud cover, relative humidity, rainfall, meridional wind speed, zonal wind speed, relative humidity at the surface, wind speed and wind direction.

Figure 10.1 shows scatter density plots of meteorological variables against ozone measurements. Here the black line is a linear fit between the meteorological variable and ozone measurements and the diamonds are the mean value of ozone measurements for a given value of the meteorological parameter.

Ozone shows a clear dependence on temperature and relative humidity. The temperature gives correlation values ranging between 0.43 and 0.70 for each station. The correlation values for relative humidity are ranging between -0.43 and -0.69. Because the temperature at the surface and the relative humidity at the surface are also the same as temperature and relative humidity, they give also a clear dependence.

10.2 Measurements - model against meteorological variables

More interesting is the behavior of the meteorological variables against measurements - model (see figure 11), because improvement of the model is needed.

The dependence shown in figure 11 are also much smaller than in figure 10.1. Figure 11c shows a dependence, how larger the temperatures how more the model underestimate the measured values. Also for ozone the variance of the meteorological variables are well defined in the model, therefore the dependence is small. The improvement will be very small, so the model is used as independent variable in the regression.

Figure 11l shows a relation between measurements - model concentrations of ozone against wind direction. If the wind comes from south- or northwest 54% of the measured concentration values above 100 μg m^{-3} is also modeled. For north- and southeastern wind the modeled ozone concentration values are closer to the measured ozone concentration values. Respectively 67% and 74% of the measured concentration values above 100 μg m^{-3} is also modeled.

10.3 Model output compared with the measurements

Figure 12 shows how the LOTOS-EUROS model, without bias correction, behaves compared to the ozone measurements. The LE model preforms quite well for lower ozone concentrations. For higher ozone concentrations the model underestimate the concentration of ozone some what.

10.4 The parameter configurations

To improve this model output for ozone the same configurations as for PM_{10} are used. The modeled ozone concentration, the measured ozone concentration and the ozone concentration of yesterday are changed into daily maximum concentration, the meteorological parameters

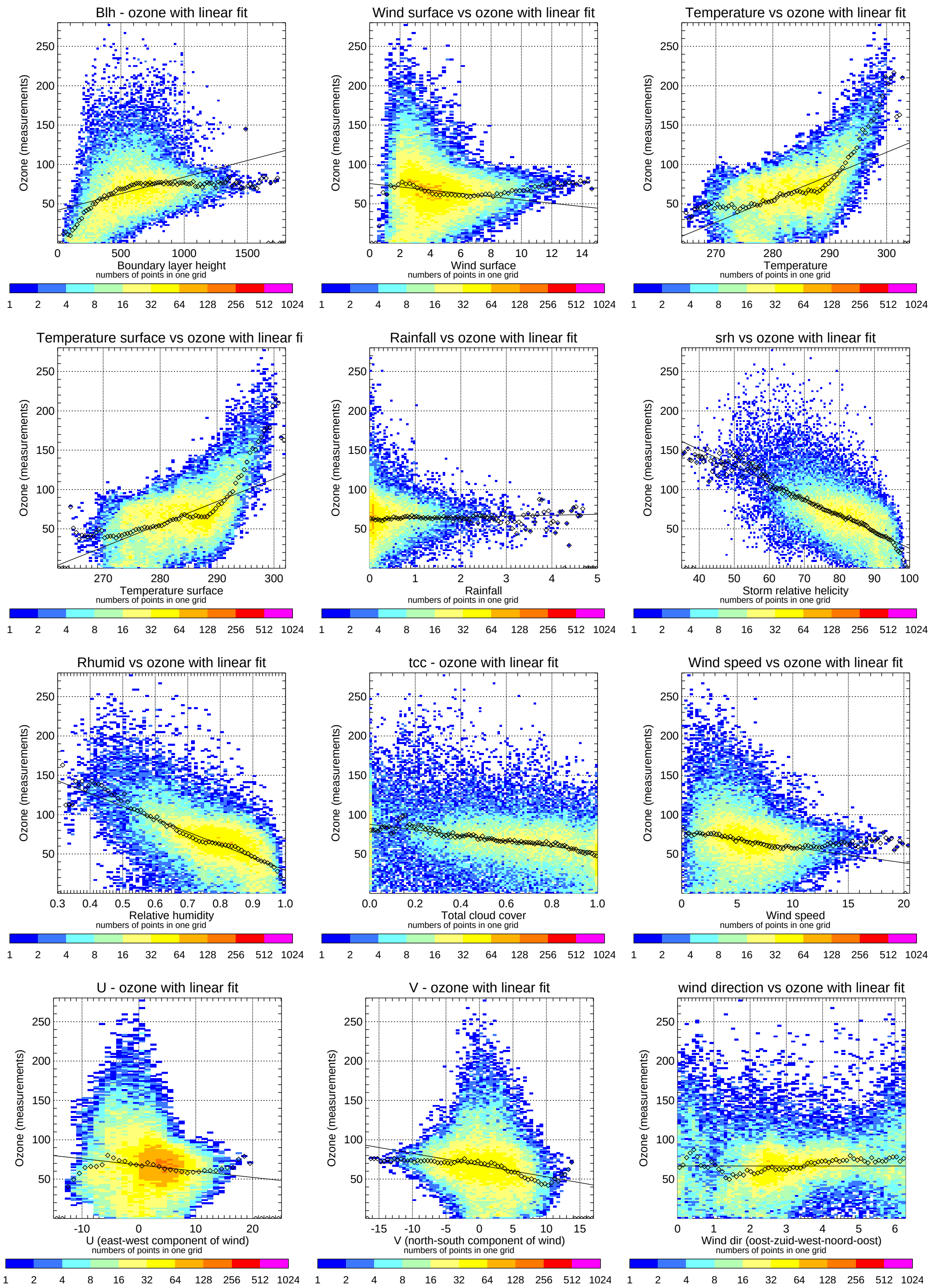


Figure 10: The scatter density plot of measured daily maximum ozone concentration in μg m^{-3} against a) boundary layer height in m; b) wind at the surface in m/s; c) temperature in Kelvin; d) temperature at the surface in Kelvin; e) rainfall in mm; f) relative humidity at the surface in %; g) relative humidity, were 1.0 is really high humidity and 0 is zero humidity; h) total cloud cover, were 0 is clear sky and 1.0 is overcast sky; i) wind speed in m/s; j) zonal wind speed in m/s; k) meridional wind speed in m/s; l) wind direction, 0 is east, π is west and 2π is east. The color indicates the density.

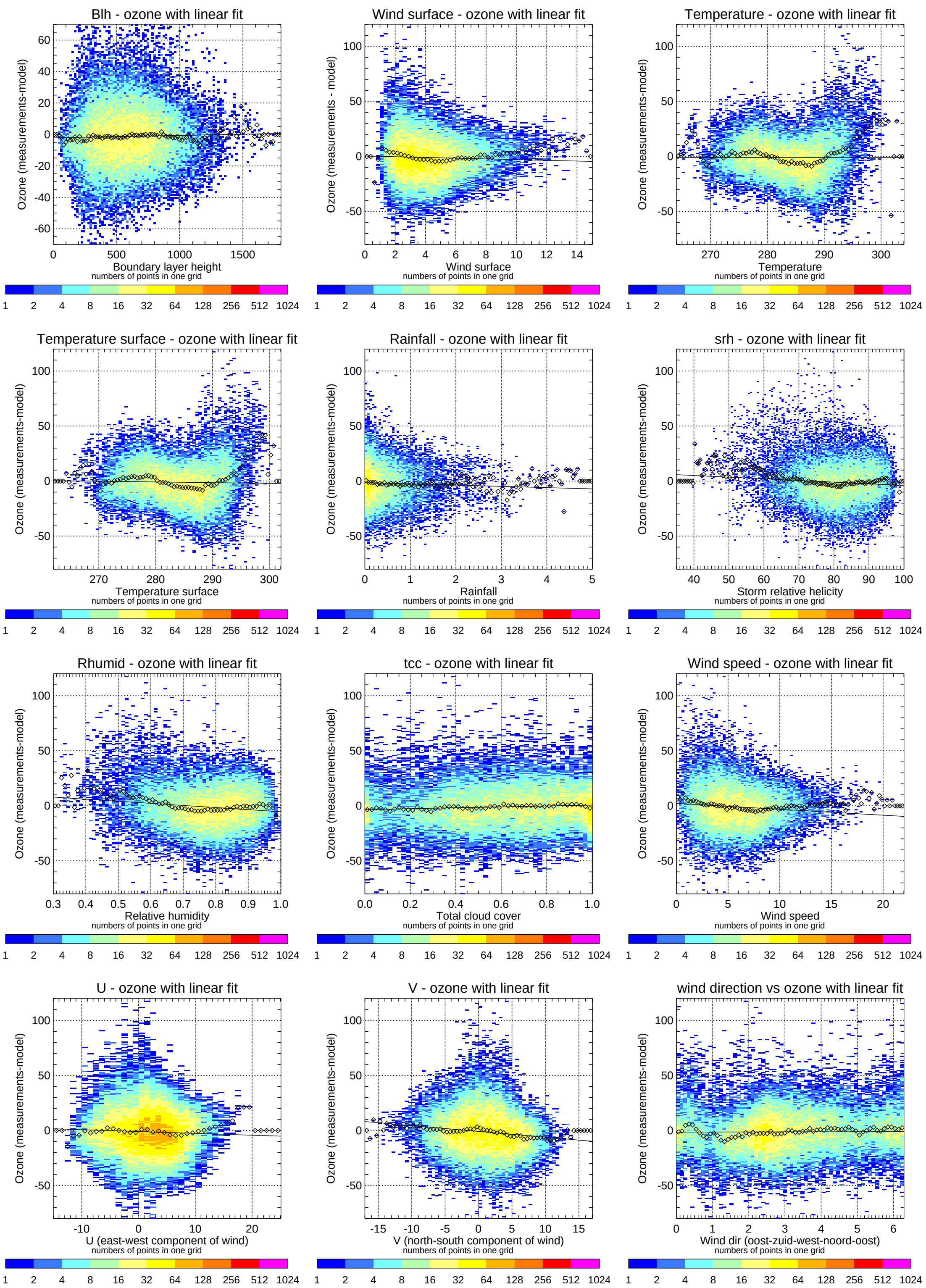


Figure 11: The scatter density plot of measured - modeled daily maximum ozone concentration, in μg m^{-3}, against a) boundary layer height in m; b) wind at the surface in m/s; c) temperature in Kelvin; d) temperature at the surface in Kelvin; e) rainfall in mm; f) relative humidity at the surface in %; g) relative humidity, were 1.0 is really high humidity and 0 is zero humidity; h) total cloud cover, were 0 is clear sky and 1.0 is overcast sky; i) wind speed in m/s; j) zonal wind speed in m/s; k) meridional wind speed in m/s; l) wind direction, 0 is east, π is west and 2π is east. The color indicates the density.

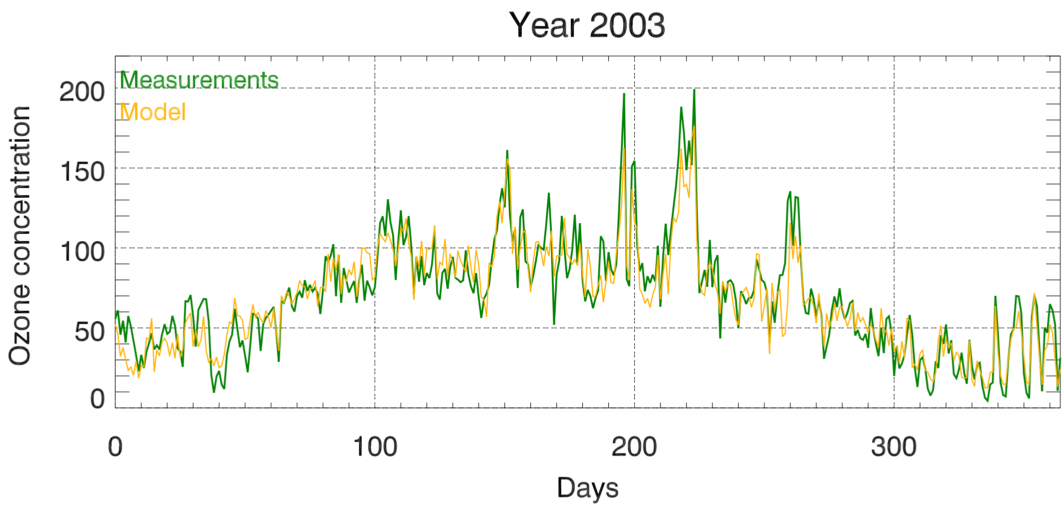


Figure 12: The Ozone daily maximum model values vs. the measurements with concentrations in μg m^{-3}

including temperature, the emissions who influence ozone and precursor trace gases are daily mean values.

Also for ozone the routine STEP is used to select which variables are important. At each station it differs which variable is removed by STEP. Appendix 2 contains information about the regression coefficients for each variable in the eight configurations at station Vredepeel, De Zilk, Eibergen and Kollumerwaard.

10.5 Performance of the multiple regression

Figure 13 shows the performance of the regression for the different configurations over the years 2003-2005. The model is almost for each station better than the regression for configurations based on meteorological parameters and measurements of yesterday, configurations A and B.

The improvements of R^2 for ozone are smaller than for PM_{10}, because the R^2 values for the model are higher for the ozone daily maximum. Configuration H shows also for ozone the best correlation. The maximum increase of R^2 appears at station De Zilk from 0.64 to 0.76. Without persistence, configuration G, the maximum increase of R^2 appears also at station De Zilk from 0.64 to 0.75. When persistence is used the correlation is 0.8-3% better than without using persistence.

In figure 14 the RMSE is given for the different configurations at the same four stations over the years 2003-2005. Not surprising the maximum reduction of RMSE appears for configuration H. The maximum decrease is found for station De Zilk, with a decrease of 28.7% from 19.59 to 15.22 μg m^{-3}. Without persistence the maximum decrease of RMSE is also for station De Zilk, with a decrease of 26.6%. The lowest RMSE for configuration H is 12.34 μg m^{-3} found at station Kollumerwaard.

The model forecast, the measured value of the day before and temperature are variables in configuration H that are chosen by STEP for every station. There are also parameters used at more than 75% of the stations, such as total cloud cover, boundary layer height, wind speed and the components nitric acid and sea salt.

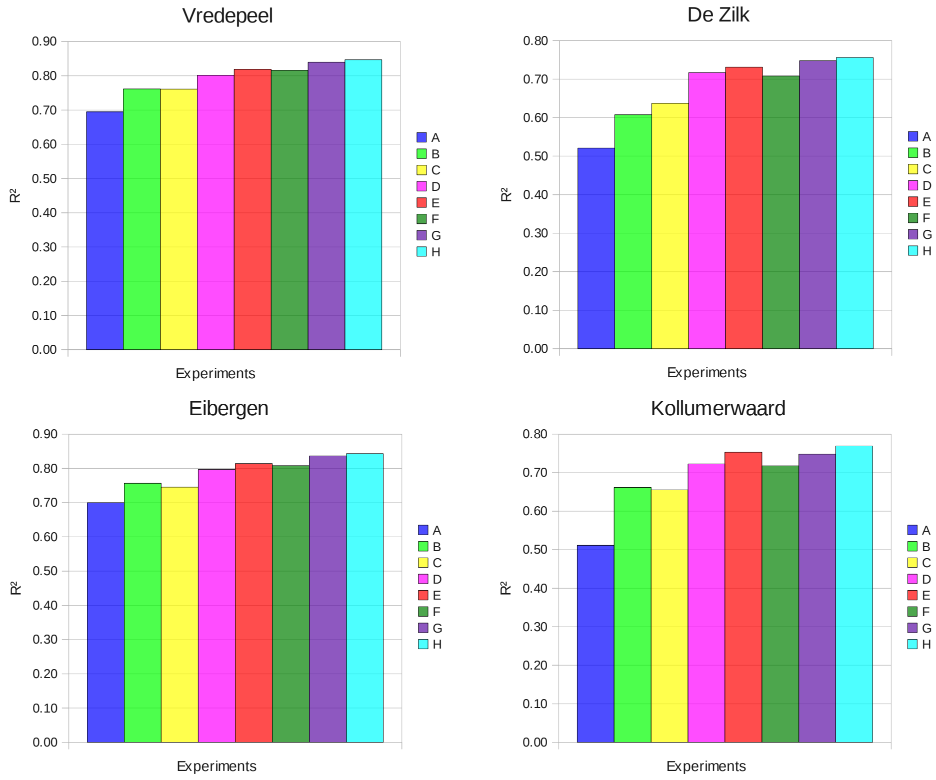


Figure 13: The R^2 for the different configurations at station Vredepeel, De Zilk, Eibergen and Kollumerwaard.

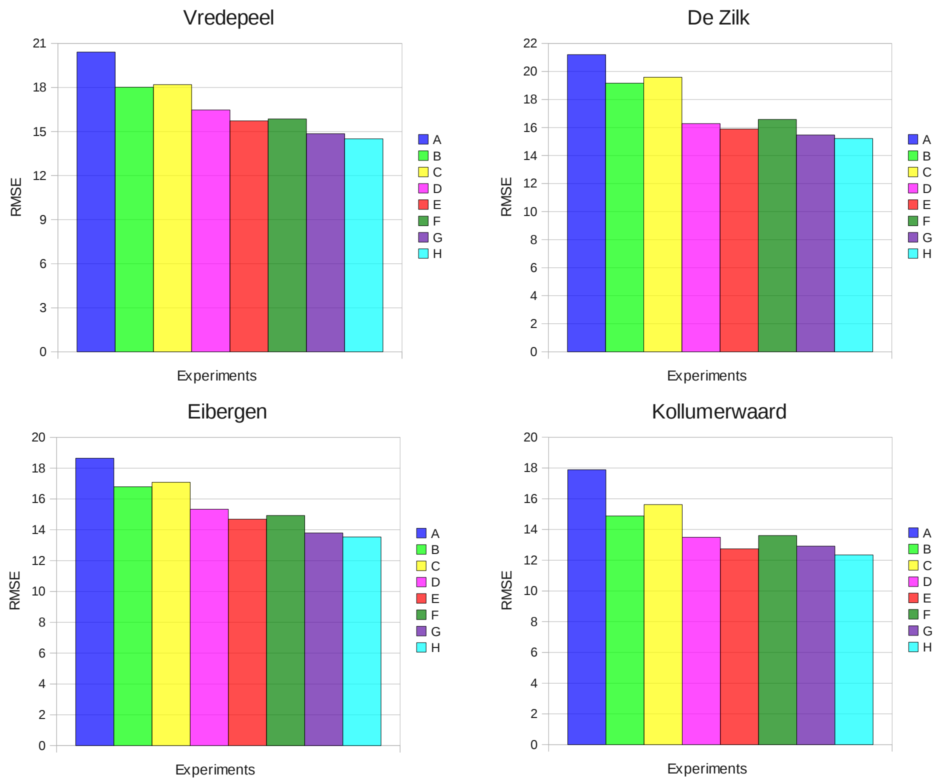


Figure 14: The RMSE for the different configurations at station Vredepeel, De Zilk, Eibergen and Kollumerwaard.

10.6 Checking the best configuration for the year 2006

10.6.1 R^2 and RMSE

At all the stations the correlation for configuration H is larger than configuration E and G over the year 2006. The RMSE is also lower for 2006. This regression, Configuration H, works also for 2006. But these improvements are small: at all stations the correlation will increase not more than 2% compared to configuration E and G. For the RMSE the improvement is larger, but not more than 8% compared to configuration E and G. In 2003-2005 the RMSE decreases between 4 and 8%.

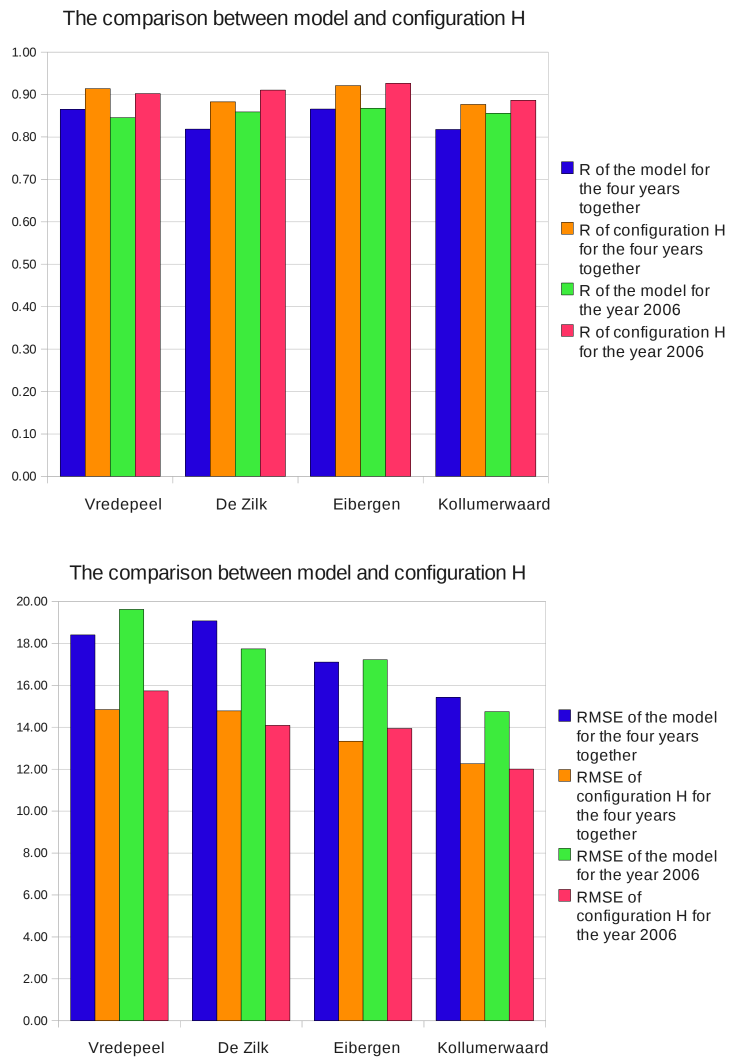


Figure 15: The behavior of the correlation and RMSE for the model and configuration H

As shown in figure 15, for 2006 the correlation are similar for all stations, ranging between 0.89 and 0.93. For the four years together the highest correlation for configuration H appear at station Eibergen, it increase with 6% compared to the model. The RMSE shows large improvement for the year 2006, it decreases with 20% at each station. The lowest RMSE occurs at station Kollumerwaard, the RMSE decreases from 14.7 to 12.0. The largest RMSE occurs at station Vredepeel, where it decrease from 19.6 to 15.7.

Figure 16 shows the time series of configuration H compared to the measurements and the model output in the year 2006 for the four stations. The higher peaks in figure 16 are not well represented anymore: at most of the stations the model describes the peaks in 2006 better than configuration H. This happens because reducing extremes is a general feature of multiple linear regression, which optimizes the RMSE.

10.6.2 Exceedances

The skill of predicting the high concentration, for ozone the concentrations above 120 μg m^{-3}, is also used, by calculating the percentage of false alarms or missed alarms.

At station Kollumerwaard the missing exceedances are increased from 66.7% to 82.2% and the measured exceedances that are also modeled are decreased, from 33.3% to 17.8%. Kollumerwaard shows a large reduction of the peaks, the modeled values above 120 μg m^{-3} for the four years dropped down from 37 to 13 compared to the 45 times this is measured. At the other stations there is an improvement, for instance at station Eibergen the missing exceedances decreased from 66.7% to 35.4%. The false exceedances for station Vredepeel is increasing from 16.7% to 18.0%, at station De Zilk and Eibergen the false exceedances are decreasing, respectively from 18.0% to 11.3% and from 23.2% to 13.5%.

Overall configuration H improve configuration C at all four stations, for the exceedances, R^2 and the RMSE. A complete overview of the counted exceedances is given in Appendix 3.

10.7 Regressions for data sub-sets

Also for ozone the full data set has been divided in subsets where a different behavior is expected. For ozone two divisions are tried, a division into summer and winter and a divisions based on the boundary layer height and the wind speed. The same as for PM_{10} is concluded, these divisions do not improve the model significantly more than configuration H.

For ozone another dividing of the data is tried, based on the temperature at the hour of the ozone maximum. The first group include the data where the temperature of the time where the highest ozone concentration occurs is 20 degrees Celsius or more and the second group where this temperature is lower than 20 degrees Celsius. This is motivated by figure 11c, where the model begins to show a negative bias above 20 degrees. Also the bias correction is based on the temperature at 20 degrees Celsius [Sauter, F. (2011)]. For the years 2003-2005 this way of dividing the data the results. Especially the RMSE decreases between 10-13% compared to configuration H for the four stations. There is a small increase in the correlation, between 1-3% compared to configuration H.

For 2006, not all the stations show better results, compared to configuration H. For Eibergen the temperature division is better than the model and configuration H, the RMSE decreases to 12.9 μg m^{-3} compare to 17.2 μg m^{-3} for the model and 13.9 μg m^{-3} for configuration H. R^2 is 0.93, the same as for configuration H. Also the well modeled exceedances are increasing at station Eibergen, from 65% to 71%. In figure 17, configuration H and the temperature division are shown for station Eibergen.

Figure 16: The time series of the ozone daily maximum of the model and configuration H in the year 2006

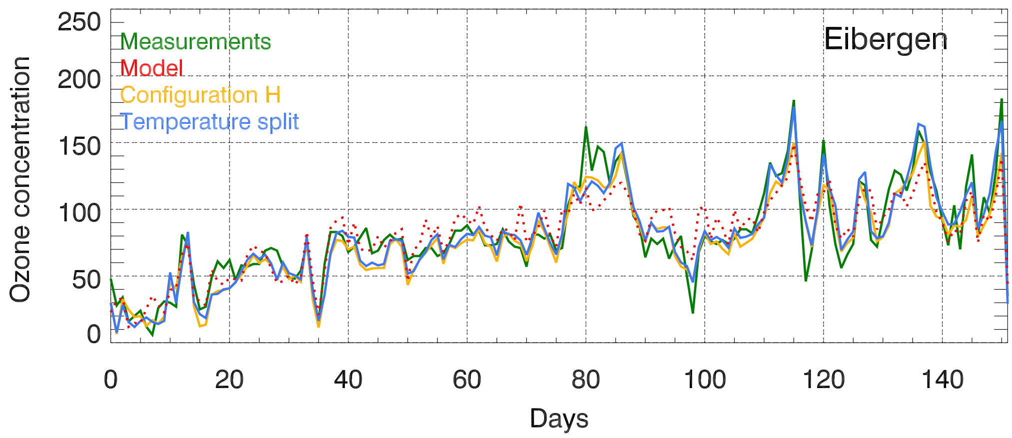


Figure 17: The time series of the regression which the data is divide based on temperature (blue) compared to the model, observations and configuration H in the year 2006 at station Eibergen.

Especially the peaks between day 110 and 140 are captured better than in configuration H and in the model. At station Vredepeel en Kollumerwaard the correlation for the year 2006 are the same as configuration H, but the error shows a small increase. On the other hand the well modeled exceedances are increasing respectively from 62% to 80% and from 18% to 53% of the measured exceedances were modeled. As mentioned before configuration H at station Kollumerwaard showed a decrease in well modeled exceedances. The temperature division reproduces the exceedances better compared to the model and configuration H, 53% of the measured exceedances were modeled compared to 33% in the model and 18% in configuration H.

At station De Zilk the error increases. This is because the amplitude of the peaks is often too high. Figure 18 shows configuration H and the temperature division for station De Zilk. Between day 210 and 240 there appear peaks for the temperature division, which should not be there.

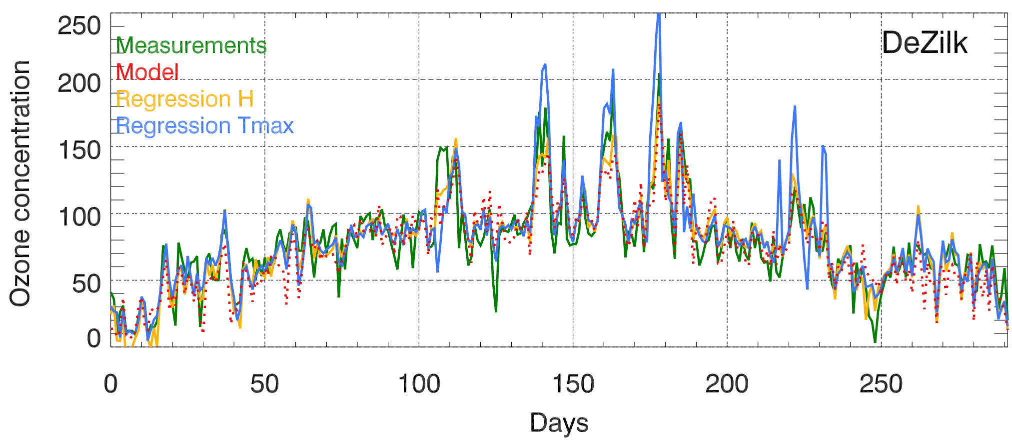


Figure 18: The time series of the regression which the data is divide based on temperature (blue) compared to the model, observations and configuration H in the year 2006 at station De Zilk.

11 Discussion

Before summarizing the results a couple of assumptions and issues are discussed.

Multiple linear regression
In this study only a linear regression with linear functions were used, but it is also possible to use linear regression with non-linear functions of the variables. The use of non linear functions, such as one divided by the boundary layer height or temperature square, does not influence the regression significant, so only linear functions are used.

How to deal with the variability and space between stations?
The multiple linear regression is done for each station separately. For each station an other set off regression coefficients is found. Secondly, the STEP routine in R drops terms in the multiple linear regression which differs for each station. A multiple linear regression for the space between stations is not possible, because there is no data in between stations. As final result regression coefficients are needed for all locations, not only at the stations. Without results in between stations, the results can not be used to correct the model. Fortunately, there are techniques to interpolate the results, see for instance Konovalov et al.

Is the use of persistence needed?
If persistence is used, which means the measurements of yesterday, the variability will be lost in the final result. The inclusion of persistency improves the model output significantly, at all stations over the years 2003-2005 and also for the year 2006. This is shown by comparing configuration D and G with E and H. But maybe the variability in the model is better modeled.

Is there a way to optimize the exceedances?
It is important to catch the exceedances for PM_{10}, concentration above 50 μg m^{-3}, and for ozone, maximum 8-hourly mean concentration above 120 μg m^{-3} or hourly concentrations above 180 μg m^{-3}, related to air quality regulations. A multiple linear regression does not always optimize the exceedances, because a linear regression minimizes the RMSE. A regression is more focused on the values which are appearing frequently than on the exceedances. Several options exist to optimize the exceedances:

1. One of the options to optimize the exceedances is a logistic regression. The measured values are changed into one or zero, one is for the concentrations above a certain value and zero for the concentrations below that value. The regression result is the probability that the value will be above that certain threshold value.

2. An other option is thinning the lower concentration values, take for instance 1/3 of the lower concentration values and all the higher concentration values into the regression. The higher concentration values get more weight.

3. Also dividing the data in subsets where a different behavior is expected may improve the exceedances results. The division into summer and winter and the division based on the boundary layer height and wind speed has no impact for the improvement of the model output. The temperature division for the ozone maximum is one example where a positive impact was found. It could be that other division based on the knowledge of the occurrence of air pollution or based on model uncertainties have also a positive impact on the model output.

References

Articles:

Barmpadimos, I., Hüglin, C., Keller, J., Henne, S., Prévôt, A.S.H., (2010) Influence of meteorology on PM_{10} trends and variability in Switzerland form 1991 to 2008. Atmospheric Chemistry and Physics Discussions, 10, 26961-27014.

Denby, B., Schaap, M., Segers, A., Builtjes, P., Horálek, J., (2008) Comparison of two data assimilation methods for assessing PM_{10} exceedances on the European scale. Atmospheric Environment 42, 7122-7134.

Glahn, H.R., Lowry, D.A., (1972) The use of Model Output Statistics (MOS) in Objective Weather Forecasting. Journal of Applied Meteorology 11, 1203-1211.

Honoré, C., Rouïl, L., Vautard, R., et al., (2008) Predictability of European air quality: assessment of 3 years of operational forecasts and analyses by the PREV'AIR system. Journal of Geophysical Research 113, D04301. doi: 10.1029/2007JD008761.

Konovalov, I.B., Beekmann, M., Meleux, F., Dutot, A., Foret, G., (2009) Combining deterministic and statistical approaches for PM_{10} forecasting in Europe. Atmospheric Environment 43, 6425-6434.

Manders, A.M.M., Schaap, M., Hoogerbrugge, R., (2009) Testing the capability of the chemistry transport model LOTOS-EUROS to forecast PM10 levels in the Netherlands. Atmospheric Environment 43, 4050-4059.

Noordijk, H., (2003) PROZON EN PROPART; statische modellen voor smogprognose. RIVM rapport.

Pöschl, U. (2005) Atmospheric Aerosols: Composition, Transformation, Climate and Health Effects. Angewandte Chemie 44, 7520-7540.

Ruyter de Wildt, M.S. de, Eskes, H., (2011) Lotos-Euros v1.7: validatierapport voor PM_{10} en bias-correctie. KNMI rapport.

Ruyter de Wildt, M. de, Eskes, H., Manders, A., Sauter, F.,Schaap, M., Swart, D., Van Velthoven, P., (2011) Six-day PM10 air quality forecasts for the Netherlands with the chemistry transport model Lotos-Euros. Atmospheric Environment 45, 5586-5594.

Sauter, F. (2011) LOTOS-EUROS bias in daily maxima ozone for rural stations in the Netherlands.

Schaap, M., Roemer, M., Sauter, F.J., Boersen, G.A.C., Timmermans, R., Builtjes, P.J.H., Vermeulen, A.T., (2005) LOTOS-EUROS: Documentation. TNO-Report B&O-A R 2005/297.

Schaap, M., Timmermans, R.M.A., Roemer, M., Boersen, G.A.C., Builtjes, P.J.H., Sauter, F.J., Velders, G.J.M. and Beck, J.P., (2008) The LOTOS-EUROS model: description, validation and latest developments. *Int. J. Environment and Pollution*, Vol. 32, No. 2, pp.270-290.

Schaap, M., Boersen, G., Timmermans, R., Sauter, F., Noije, T. van, Builtjes, P., (2008) LOTOS-EUROS: User Manual.

Segers, A., (2011) LOTOS-EUROS v1.7 User Guide. LOTOS-EUROS Report.

Stern, R., Builtjes, P., Schaap, M., Timmermans, R., Vautard, R., Hodzic, A., Memmesheimer, M., Feldmann, H., Renner, E., Wolke, R., Kerschbaumer, A., (2008) A model inter-comparison study focussing on episodes with elevated PM_{10} concentrations. Atmospheric Environment 42, 4567-4588.

Vannitsem, S., (2008) Dynamical Properties of MOS Forecasts: Analysis of the ECMWF Operational Forecasting System. Wea. Forecasting, 23, 1032–1043.

Velders, G.J.M., Matthijsen, J., (2009) Meteorological variability in NO_2 and PM_{10} concentrations in the Netherlands and its relation wiht EU limit values. Atmospheric Environment 43, 3858-3866.

Velders G.J.M., Diederen H.S.M.A.(2009) Likelihood of meeting the EU limit values for NO2 and PM10 concentrations in the Netherlands. Atmospheric Environment 43, 3060-3069.

WHO rapport (2005) Air quality guidelines. Global update 2005. Particulate matter, ozone, nitrogen dioxide and sulfur dioxide.

Books:

Daniel, J. Jacob (1999) Introduction to Atmospheric Chemistry.

Glahn, H.R., Murphy, A.H., Wilson, L.J., Jenscnius, J.S, (1991) Programme on short- and medium-range weather prediction research. PSMP Report series no. 34.

Ian Chivers and Jane Sleightholme Introducing Fortran 90.

Ostrom, C.W., Jr., (1990) Time Series Analysis, Regression Techniques, Second Edition: Quantitative Applications in the Social Sciences, v. 07-009: Newbury Park, Sage Publications.

Wilkes, D.S., (2006) Statistical Methods in the Atmospheric Science. Second Edition.

Internet websites:

AIN (Air Info Now) http://www.airinfonow.org/index.asp

China's Ministry of environmental protection http://english.mep.gov.cn/

Clean Air Agency of Puget Sound http://www.pscleanair.org/airq/basics/weather/default.aspx

ECE (European Commission Environment) http://ec.europa.eu/environment/index_en.htm

and http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32008L0050:EN:NOT

Fortran http://www.owlnet.rice.edu/ ceng303/manuals/fortran/index.html

and http://www.cs.mtu.edu/ shene/COURSES/cs201/NOTES/fortran.html

IDL file:///usr/local/appl/installed/idl80/idl/idl80/help/online_ help/IDL/idl.html

and http://www.ittvis.com/language/en-us/productsservices/idl.aspx

Lecture notes of Applied Time Series Analysis by David M. Meko

http://www.ltrr.arizona.edu/∼dmeko/notes_11.pdf

LML (Landelijk Meetnet Luchtkwaliteit) http://www.lml.rivm.nl/

LOTOS-EUROS http://www.lotos-euros.nl/

MACC http://www.gmes-atmosphere.eu/

NL Agency ***Ministry of Infrastructure and the Environment***

http://www.infomil.nl/english/subjects/air/

NMDC project (Nationaal Modellen- en Data Centrum) http://www.nmdc.eu/

PBL reports http://www.pbl.nl/en/dossiers/transboundaryairpollution/policyevaluations

Queensland Government http://www.derm.qld.gov.au/air/monitoring/modelling.html

R programming language http://www.r-project.org/

Rijksoverheid http://www.rijksoverheid.nl/onderwerpen/luchtkwaliteit

RIVM (Rijksinstituut voor Volksgezondheid en Milieu) http://www.rivm.nl/

The World Bank http://www.worldbank.org/

U.K. Department for Environment Food and Rural Affairs

http://uk-air.defra.gov.uk/air-pollution/

U.S. Environmental Protection Agency http://www.epa.gov/airquality/

World Health Organization http://www.who.int/en/

Appendix 1

This Appendix contains the regression coefficients of the multiple lineare regressions for PM_{10}. There are 7 configurations:

A Meteorological parameters (D+1)

B Meteorological parameters (D+1) + PM_{10}^{obs}(D+0)

D PM_{10}^{CTM}(D+1) + meteorological parameters (D+1)

E PM_{10}^{CTM}(D+1) + meteorological parameters (D+1) + PM_{10}^{obs}(D+0)

F PM_{10}^{CTM}(D+1) + CTM tracers (D+1)

G PM_{10}^{CTM}(D+1) + CTM tracers (D+1) + meteorological parameters (D+1)

H PM_{10}^{CTM}(D+1) + CTM tracers (D+1) + meteorological parameters (D+1) + PM_{10}^{obs}(D+0)

Configuration A and B are the standard statistical models which do not involve CTM simulations. Both employ meteorological parameters and configuration B also involves the PM_{10} daily-mean concentration observed on the past day. C is not a configuration, it involves the forecasts from the LOTOS-EUROS model. Configuration D and E correspond to the combined forecasts. Configuration F,G and H are also combined forecasts, but they include CTM tracers, which include components of PM_{10} and aerosol precursor trace gases.

All the variables, exept the model and yesterday, the measurements of the day before, are normalized. The model and yesterday are in μg m^{-3}. The RMSE (root mean squar error) gives the skill in predicting the overall magnitude of the observations, and is also given in μg m^{-3}. The correlation (R) is a measure of whether forecasts and observation change in the same way. The closer the correlation is to one, the better the forecasts variability is in agreement with the observations.

Vredepeel								
	A	B	C	D	E	F	G	H
intercept	28.3	12.8		9.6	5.5	-1.1	-4.1	0.7
model				0.7	0.5	1.1	1.2	0.7
yesterday		0.5			0.3			0.3
temp	37.2	22.7		14.1	9.6		15.7	
u		0.7		0.6	0.8		0.8	1.0
v							-0.9	
rhum	8.2	3.3		6.0	3.1		4.4	
blh	-6.2	-3.6		-3.3	-2.7		-2.8	-2.3
tcc		0.7			0.5			
speed				0.7				-1.0
rain	-1.2	-1.2		0.5				
dir								
wsurf	-2.3	-2.2					1.4	1.4
tsurf	-37.8	-23.2		-14.6	-10.1		-14.5	
srh	-10.7	-5.3		-7.1	-4.4		-5.7	-1.7
so2						-0.9	-1.1	-0.9
nh3						-0.8	-0.8	
nh4a								
ppm10								
ppm25						-2.5	-3.3	-1.5
no2						3.7	3.2	2.6
hno3						1.6	1.1	1.2
no3a						-4.9	-6.8	-3.3
no								
so4a						-1.9	-1.8	
bc								
meano3								
na_f						-3.0	-3.2	-2.2
na_c						0.8	0.8	1.1
RMSE	12.48	9.70	10.19	8.94	8.00	8.73	8.61	7.78
R^2	0.35	0.61	0.64	0.67	0.73	0.68	0.70	0.75

Figure 19: The regression coefficients for station Vredepeel for the 7 configurations.

De Zilk								
	A	B	C	D	E	F	G	H
intercept	29.1	13.9		10.4	5.5	-4.8	-19.5	-11.7
model				0.7	0.5	1.2	1.8	1.2
yesterday		0.5			0.3			0.3
temp	13.4	9.2		-7.8	-4.2		-6.3	-3.1
u				-0.5			0.8	1.1
v	1.0			2.1	1.2			
rhum							-1.2	
blh	-6.3	-3.4		-3.6	-2.0		-1.5	-1.4
tcc								
speed	2.1	0.5						
rain	-1.3	-1.1		0.6				
dir		-1.6		-0.6	-0.8			
wsurf				1.2			1.5	0.7
tsurf	-16.2	-10.8		7.1	3.7		8.5	4.6
srh	-0.7	-0.5					1.2	
so2							-0.8	
nh3								0.7
nh4a								
ppm10						-3.3	-4.1	-2.7
ppm25						5.7	3.2	2.5
no2						0.7	1.6	1.0
hno3						1.6	1.6	1.2
no3a						-5.2	-10.3	-5.8
no							0.9	1.3
so4a						-3.0	-5.4	-3.6
bc						-5.7	-5.5	-4.1
meano3								
na_f						-4.6	-5.8	-3.6
na_c						-2.6	-5.8	-3.5
RMSE	11.26	9.06	9.97	8.51	7.60	8.19	8.02	7.38
R^2	0.33	0.57	0.56	0.62	0.69	0.65	0.66	0.71

Figure 20: The regression coefficients for station De Zilk for the 7 configurations.

Eibergen

	A	B	C	D	E	F	G	H
intercept	28.4	14.1		10.7	7.0	3.7	-6.8	-7.0
model				0.7	0.5	1.0	1.4	1.1
yesterday		0.5			0.3			0.2
temp	45.4	22.3		16.2	7.5		13.5	6.0
u					0.6			
v	1.4	1.5		0.8	0.7			
rhum	11.1	3.7		6.1	2.8		2.8	
blh	-3.6	-2.5			-1.2		-1.4	-0.8
tcc	0.9	0.8						
speed	2.7			1.7			0.7	
rain		-1.2		0.6				
dir	-0.8							
wsurf	-5.7	-2.3		-2.7	-0.8			
tsurf	-46.5	-22.9		-16.6	-7.7		-11.4	-4.3
srh	-12.1	-4.5		-4.8	-2.7		-2.6	
so2						-2.1	-2.2	-1.6
nh3								
nh4a						-5.7	-11.5	-9.5
ppm10						1.0		
ppm25						-1.6	-3.1	-2.4
no2						3.2	3.0	2.1
hno3							-0.5	
no3a								
no							-0.8	
so4a						1.3	1.8	1.6
bc								
meano3						0.8	1.0	
na_f						-2.9	-3.1	-2.5
na_c						0.6		
RMSE	11.77	9.67	10.19	8.77	8.16	8.48	8.38	7.92
R^2	0.31	0.53	0.60	0.62	0.67	0.64	0.65	0.69

Figure 21: The regression coefficients for station Eibergen for the 7 configuration.

Kollumerwaard								
	A	B	C	D	E	F	G	H
intercept	27.5	13.4		11.7	6.5	2.5	-19.3	-14.1
model				0.7	0.5	1.1	2.1	1.6
yesterday		0.5			0.3			0.3
temp	19.3	13.7		1.4	7.1		12.9	10.9
u							1.6	1.6
v				2.0	1.5		-2.0	-1.7
rhum	5.2	3.3		2.3	3.0		3.3	3.1
blh	-5.7	-4.0		-1.7	-1.3		-1.2	-1.2
tcc								
speed	2.1	1.1			1.0		1.8	1.2
rain	-1.2	-1.2						
dir	-2.3	-1.6		-0.8	-0.6		-1.0	-0.9
wsurf					-1.3			
tsurf	-18.3	-13.1			-6.1		-9.0	-8.0
srh	-5.7	-3.8		-1.8	-2.5		-3.3	-3.2
so2						-1.9	-1.8	-0.9
nh3								
nh4a						-33.8	-17.4	-12.6
ppm10								1.2
ppm25							-2.7	-3.1
no2						2.6	2.5	2.4
hno3								
no3a						20.0		
no						-1.1	-1.1	-0.6
so4a						11.2	1.1	1.1
bc								
meano3						1.7		
na_f						-4.3	-7.7	-5.7
na_c						-1.7	-5.9	-4.0
RMSE	10.54	8.74	10.87	8.18	7.31	7.94	7.46	6.88
R^2	0.30	0.52	0.50	0.58	0.66	0.60	0.65	0.71

Figure 22: The regression coefficients for station Kollumerwaard for the 7 configurations.

Appendix 2

The following tables show the regression coefficients for ozone. An explaination of the tables is found in Appendix 1.

Vredepeel

	A	B	C	D	E	F	G	H
intercept	70.0	42.1		21.8	14.7	27.0	38.2	31.7
model				0.7	0.6	0.6	0.5	0.4
yesterday		0.4			0.2			0.2
temperature	59.0	52.2		41.3	37.3		35.6	33.4
tempmax								
u	4.6	1.7		3.4	2.4		3.0	1.6
v	-8.2	-5.1		-3.9	-3.9		-1.9	
rel. humidity								
blh	2.7	4.3					3.4	3.7
tcc				1.5	1.0		1.5	
speed	-7.8	-5.2		-4.3	-4.1		-2.7	-2.7
rain	2.5				-1.1			
direction				1.2			1.3	1.5
wind surf.	6.3	4.8		4.9	5.6			
temp. surf.	-44.6	-43.9		-37.2	-34.5		-33.9	-33.0
rel. hum. Surf.	-19.4	-11.8		-9.7	-7.7		-8.1	-6.9
so2						5.0	1.5	
nh4a						8.1	7.7	5.4
nh3						4.5	1.7	1.2
ppm10								
ppm25						-2.3		
no2						-6.0		
hno3						8.5	7.8	7.0
no3a								
no						-5.3	-4.9	-4.6
so4a						-6.9	-4.9	-2.3
pm10								
bc						4.1		
na_f						6.2	6.1	5.3
na_c						-2.6	-2.8	-1.5
RMSE	20.41	18.02	18.20	16.47	15.73	15.86	14.85	14.51
R^2	0.69	0.76	0.76	0.80	0.82	0.82	0.84	0.85

Figure 23: The regression coefficients for station Vredepeel for the 7 configurations.

De Zilk

	A	B	C	D	E	F	G	H
intercept	74.0	45.3		27.5	20.7	30.9	34.4	28.4
model				0.7	0.6	0.6	0.6	0.5
yesterday		0.4			0.2			0.1
temperature	65.3	62.2		36.4	36.0		39.4	37.1
u	4.7	1.9		2.3	1.3			
v	-9.5	-6.4		-5.9	-4.7		-2.8	-2.0
rel. humidity	6.2	6.3					3.1	2.7
blh	-5.5				1.8			
tcc				3.4	2.6		3.7	3.1
speed	-6.7			-1.7	-1.9		-1.9	-2.0
rain								
direction	3.1	1.8						
wind surf.	7.9							
temp. surf.	-48.9	-50.8		-31.4	-31.9		-36.1	-34.4
rel. hum. Surf.	-17.8	-12.6		-6.5	-4.9		-7.6	-7.1
so2						4.3	2.1	
nh4a						-8.3	-10.6	-27.8
nh3						6.5	3.5	2.8
ppm10								
ppm25						-18.5	-12.1	-8.6
no2						-7.4	-4.0	-3.5
hno3						4.0	5.0	4.4
no3a						11.8	12.5	24.5
no						-4.2		
so4a								6.5
pm10								
bc						19.0	9.8	7.5
na_f						5.8	3.2	3.1
na_c						-2.6		
RMSE	21.20	19.17	19.59	16.28	15.89	16.58	15.47	15.22
R^2	0.52	0.61	0.64	0.72	0.73	0.71	0.75	0.76

Figure 24: The regression coefficients for station De Zilk for the 7 configurations.

Eibergen

	A	B	C	D	E	F	G	H
intercept	68.3	43.3		22.9	15.7	24.1	34.1	27.8
model				0.7	0.6	0.6	0.5	0.5
yesterday		0.4			0.2			0.1
temperature	34.8	16.1		27.5	16.5		33.0	20.0
u		-1.6						-2.3
v	-6.9	-4.2		-4.4	-3.6			
rel. humidity	-6.4	-11.8		-5.2	-8.1			-7.8
blh	4.1	5.3					1.5	3.2
tcc				2.2	1.7		2.2	1.7
speed	-6.1	-4.7		-3.5	-3.7		-1.4	-1.3
rain	3.4	1.2			-0.9			
direction				-1.1	-1.0			
wind surf.	3.8	3.8		4.5	5.4			
temp. surf.	-24.2	-9.5		-23.3	-13.6		-32.2	-18.6
rel. hum. Surf.	-12.5			-4.8			-9.6	
so2						6.6		
nh4a						-6.5		
nh3						2.6		
ppm10						2.3	2.7	
ppm25						-6.5	-5.1	-5.1
no2						-6.5		
hno3						6.7	5.8	5.1
no3a						9.3	4.9	
no						-2.4		-1.4
so4a								
pm10								4.6
bc						2.8	-2.5	
na_f						6.3	6.8	5.8
na_c						-4.9	-4.1	-3.4
RMSE	18.64	16.79	17.08	15.34	14.69	14.93	13.79	13.53
R^2	0.70	0.76	0.75	0.80	0.81	0.81	0.84	0.84

Figure 25: The regression coefficients for station Eibergen for the 7 configuration.

Kollumerwaard

	A	B	C	D	E	F	G	H
intercept	73.5	37.4		22.2	14.6	29.1	31.8	23.2
model				0.7	0.5	0.6	0.6	0.5
yesterday		0.5			0.3			0.2
temperature				17.4	11.3		18.8	12.4
u	1.1			1.3	0.8			
v	-6.9	-3.9		-4.4	-3.6		-1.3	
rel. humidity	-7.0	-7.2		-2.8	-4.5			-3.6
blh		1.2					-1.8	
tcc				3.0	2.1		2.5	1.4
speed	-11.0	-8.7		-4.9	-5.7		-4.5	-4.8
rain	1.9				-1.1			-1.1
direction								
wind surf.	9.3	8.1		5.3	6.9		5.7	5.0
temp. surf.	9.1	5.1		-16.5	-10.5		-17.1	-11.2
rel. hum. Surf.	-5.8			-3.3			-4.4	
so2						2.4	1.4	
nh4a								
nh3						2.1	1.4	
ppm10								
ppm25						-9.2	-10.1	-6.6
no2						-7.6	-5.9	-5.1
hno3							1.0	
no3a							-5.9	-5.6
no								
so4a						-2.2	-4.5	-2.5
pm10						4.8	11.8	9.8
bc						5.5	5.3	2.9
na_f						3.3		
na_c						-3.6	-4.2	-3.5
RMSE	17.89	14.88	15.63	13.49	12.74	13.61	12.92	12.34
R^2	0.51	0.66	0.66	0.72	0.75	0.72	0.75	0.77

Figure 26: The regression coefficients for station Kollumerwaard for the 7 configurations.

Appendix 3

The following tables contain the alarms for PM_{10}, the concentration above 40 μg m^{-3} and the exceedances for ozone, concentrations above 120 μg m^{-3}. This is calculated for the model and configuration H, which include the model, measurements of the day before, meteorological variables and CTM tracers for the four stations: Vredepeel, De Zilk, Eibergen and Kollumerwaard.

In the columns the total measured exceedances, the total modeled exceedances, the true and false modeled exceedances and the missing exceedances are given. The first column in numbers, the second in percentage of the total modeled exceedances, so how many of the modeled exceedances are true, $\frac{TRUE}{totalmodeled} * 100$ and how many of the modeled exceedances are false, $\frac{False}{totalmodeled} * 100$. In the third column the percentage of the total measured exceedances are given, how many of the measured exceedances are modeled, $\frac{TRUE}{totalmeasured} * 100$ and how many of the measured exceedances are missing $\frac{MISSED}{totalmeasured} * 100$.

Exceedance of PM10 above 40 µg per m3

Vredepeel, 4 years of PM10 data

	Model			Configuration H		
	Total	Modelled (in %)	Measured (in %)	Total	Modelled (in %)	Measured (in %)
total measured	249			249		
total modelled	229			212		
TRUE	150	65.5	60.24	161	75.94	64.66
FALSE	79	34.5		51	24.06	
MISSED	99		39.76	88		35.34

De Zilk, 4 years of PM10 data

	Model			Configuration H		
	Total	Modelled (in %)	Measured (in %)	Total	Modelled (in %)	Measured (in %)
total measured	205			205		
total modelled	211			210		
TRUE	130	61.61	63.41	143	68.1	69.76
FALSE	81	38.39		67	31.9	
MISSED	75		36.59	62		30.24

Eibergen, 4 years of PM10 data

	Model			Configuration H		
	Total	Modelled (in %)	Measured (in %)	Total	Modelled (in %)	Measured (in %)
total measured	214			214		
total modelled	219			200		
TRUE	132	60.27	61.68	143	71.5	66.82
FALSE	87	39.73		57	28.5	
MISSED	82		38.32	71		33.18

Kollumerwaard, 4 years of PM10 data

	Model			Configuration H		
	Total	Modelled (in %)	Measured (in %)	Total	Modelled (in %)	Measured (in %)
total measured	178			178		
total modelled	101			182		
TRUE	75	74.26	42.13	119	65.38	66.85
FALSE	26	25.74		63	34.62	
MISSED	103		57.87	59		33.15

Figure 27: The alarms for PM_{10}

Exceedance of ozone above 120 µg per m3

Vredepeel, 4 years of ozone data

	Model			Configuration H		
	Total	Modelled (in %)	Measured (in %)	Total	Modelled (in %)	Measured (in %)
total measured	104			104		
total modelled	60			78		
TRUE	50	83.33	48.08	64	82.05	61.54
FALSE	10	16.67		14	17.95	
MISSED	54		51.92	40		38.46

De Zilk, 4 years of ozone data

	Model			Configuration H		
	Total	Modelled (in %)	Measured (in %)	Total	Modelled (in %)	Measured (in %)
total measured	64			64		
total modelled	39			53		
TRUE	32	82.05	50	47	88.68	73.44
FALSE	7	17.95		6	11.32	
MISSED	32		50	17		26.56

Eibergen, 4 years of ozone data

	Model			Configuration H		
	Total	Modelled (in %)	Measured (in %)	Total	Modelled (in %)	Measured (in %)
total measured	99			99		
total modelled	43			74		
TRUE	33	76.74	33.33	64	86.49	64.65
FALSE	10	23.26		10	13.51	
MISSED	66		66.67	35		35.35

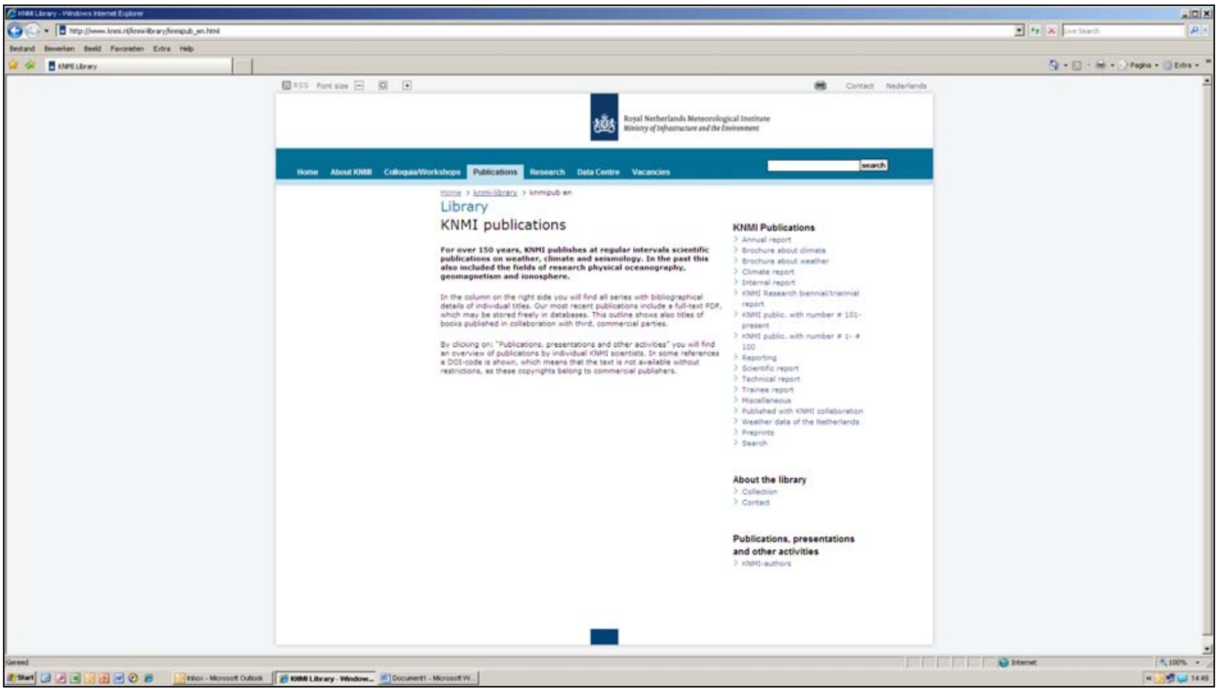
Kollumerwaard, 4 years of ozone data

	Model			Configuration H		
	Total	Modelled (in %)	Measured (in %)	Total	Modelled (in %)	Measured (in %)
total measured	45			45		
total modelled	37			13		
TRUE	15	40.54	33.33	8	61.54	17.78
FALSE	22	59.46		5	38.46	
MISSED	30		66.67	37		82.22

Figure 28: The exceedances for ozone

A complete list of all KNMI-publications (1854 – present) can be found on our website

www.knmi.nl/knmi-library/knmipub_en.html



The most recent reports are available as a PDF on this site.