



Royal Netherlands
Meteorological Institute
Ministry of Infrastructure and the
Environment

Statistical post-processing of the Lotos-Euros model through Model Output Statistics

A. Pijnappel, H. Eskes and M. Krol





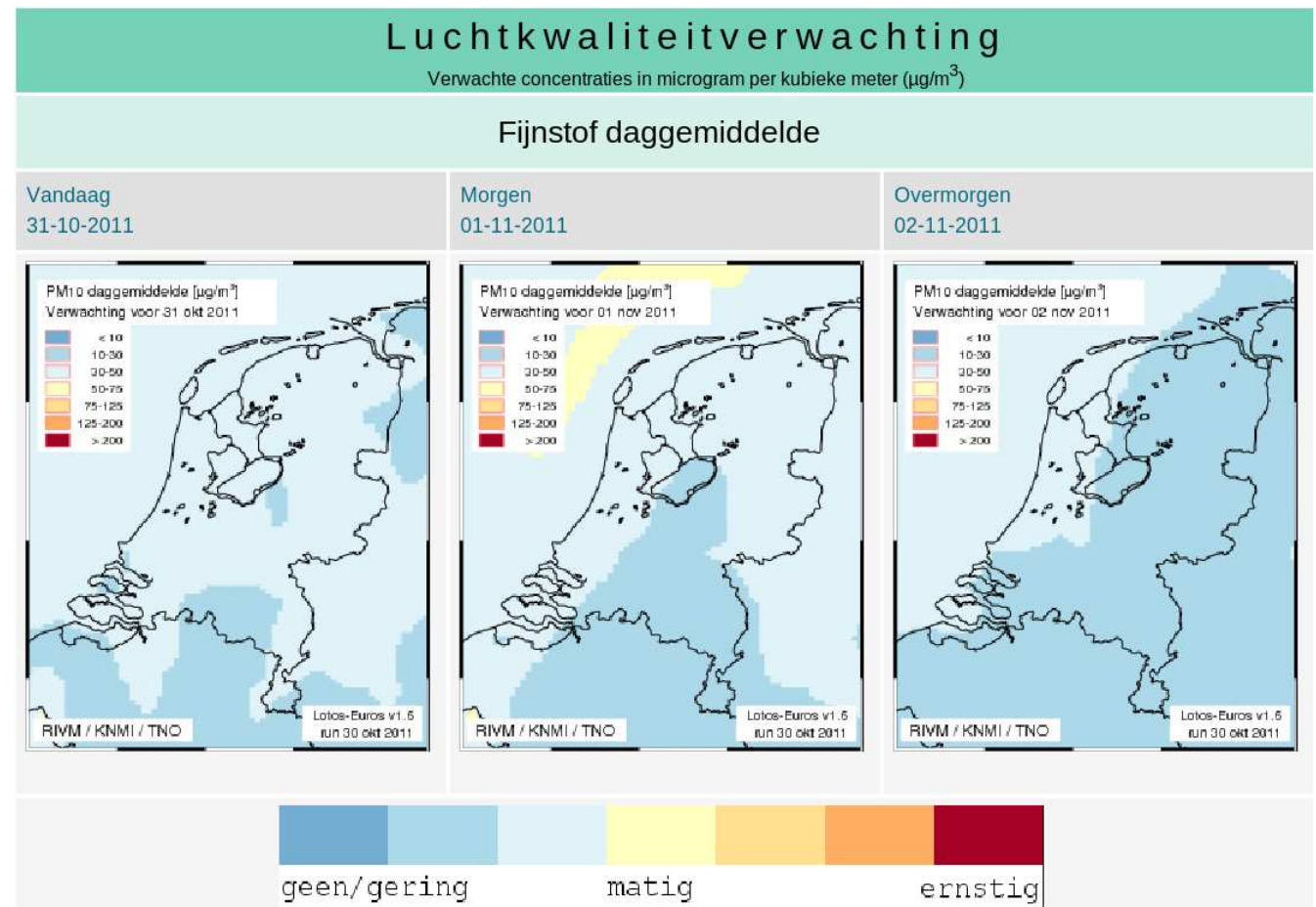
Air quality forecasts

Hourly forecasts for PM10 and ozone for three days.

Cooperated by TNO, RIVM and KNMI

These forecasts are operational at KNMI.

We want to improve these forecast.





Why do we want air quality forecasts?

Very harmful for our health, but also the health of animals and crop plants.

It is one of the areas in which the European Union has been most active. In the EU legislation there are rules for air pollution.

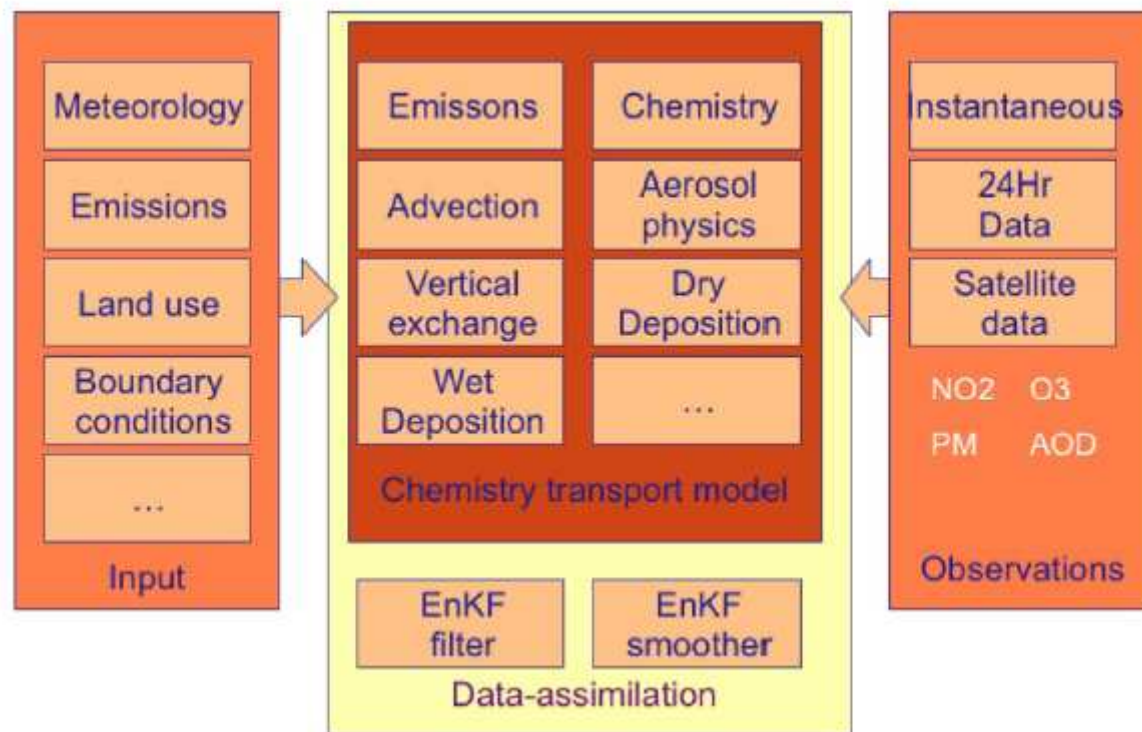


China, Beijing

Pollutant	Period	Limit value
PM10	One day	50 $\mu\text{g}/\text{m}^3$, not to be exceeded, more than 35 times a year.
PM10	Calendar year	40 $\mu\text{g}/\text{m}^3$
Ozone	Maximum daily eight-hour mean	120 $\mu\text{g}/\text{m}^3$ not to be exceeded, more than 25 times a year.
Ozone	Max. hour of the day	180 $\mu\text{g}/\text{m}^3$, government has to inform the population.
Ozone	Max. hour of the day	240 $\mu\text{g}/\text{m}^3$, government gives an alert.



The Lotos-Euros model

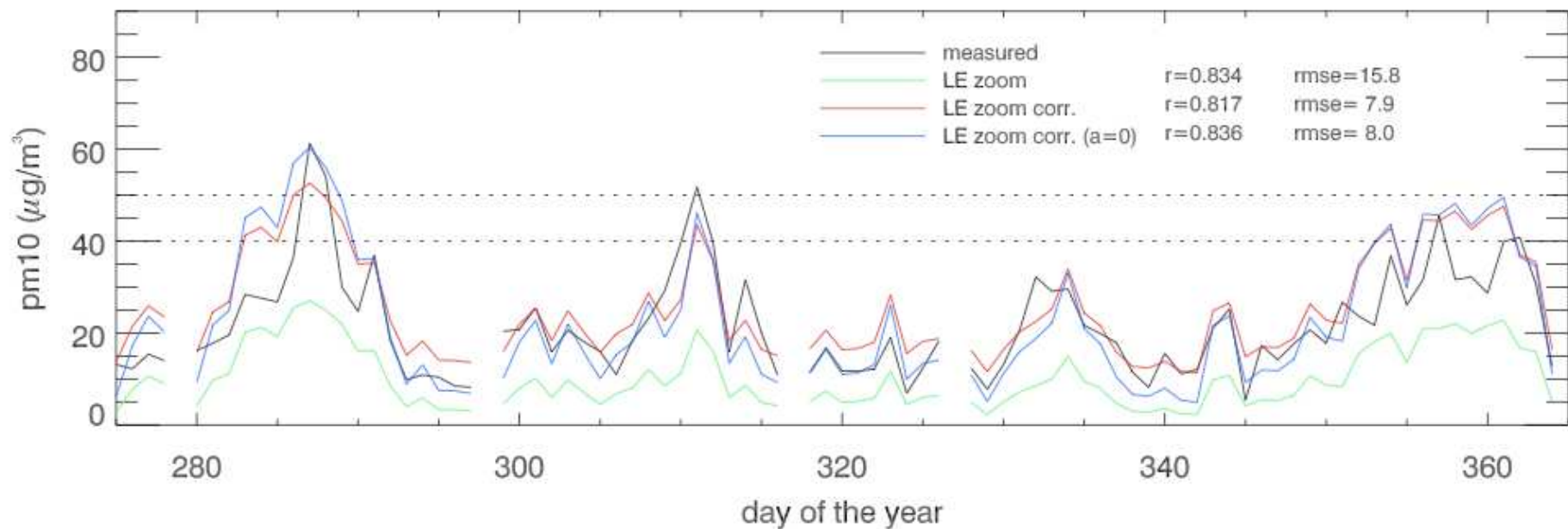




Bias corrections for PM10

$$\mathbf{PM10}_{\text{corr}} = \mathbf{F} * \mathbf{PM10}_{\text{mod}}$$

$$\text{with } \mathbf{F} = 2.11 + 0.291 * \sin(2\pi(d-319.8)/365)$$



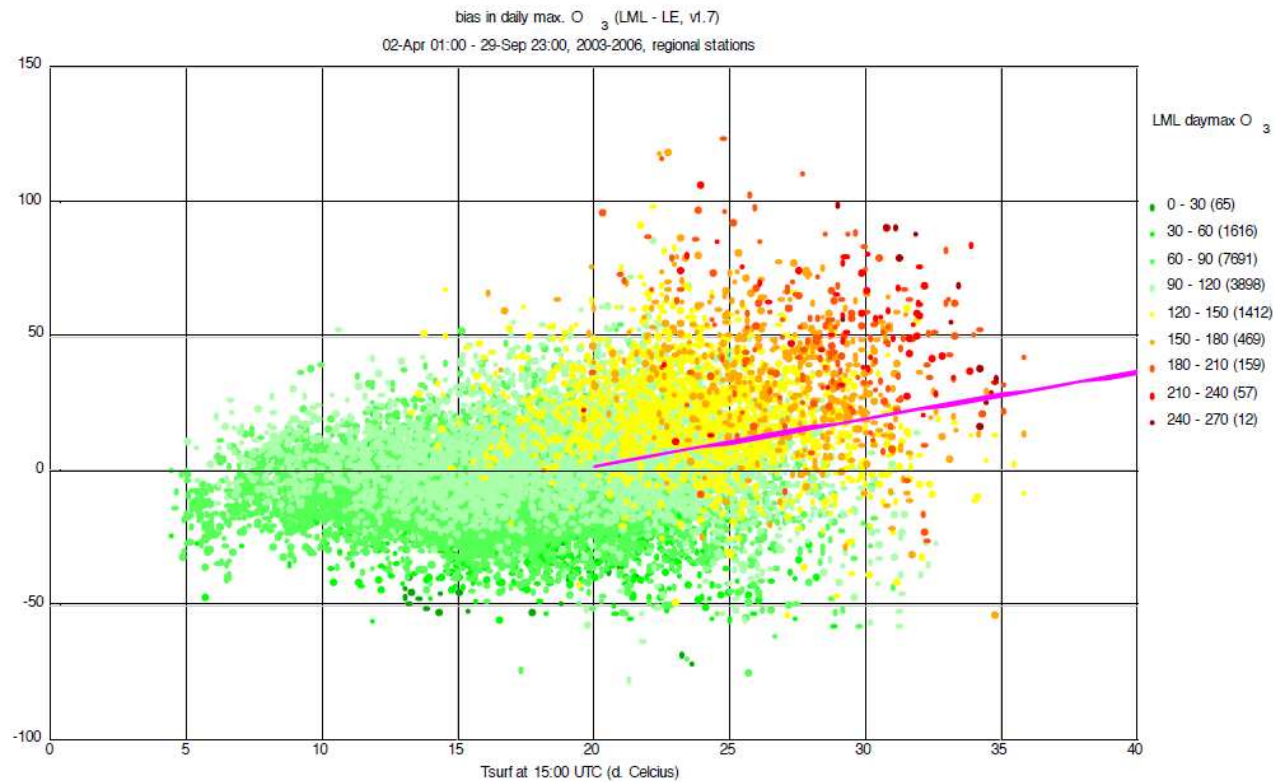
M. de Ruyter de Wildt



Bias corrections for Ozone

$$\text{Ozone}_{\text{corr}} = \mathbf{G} + \text{Ozone}_{\text{mod}},$$

with $\mathbf{G} = -0.00194883 * T^2 + 1.86295 * T - 35.1348$



F. Sauter



What is post processing of model forecasts?

We use Model output Statistics (MOS) to improve the skill of the model forecasts.

MOS is based on a multiple linear regression. Multiple linear regression (MLR) is a method, which model the linear relationship between a dependent variable and a couple of independent variables.

$$PM_{10}(i) = p_0 + \sum_{k=1}^n p_k x_{i,k} + \varepsilon_i$$

MLR is based on least squares: the \mathbf{p} in the equation are fit such that the sum of squares of error is minimized.

The MOS method to post processing the model forecasts is already used in the weather forecasts.



Can we improve the LE-model output for PM10 and ozone through statistical post-processing?

What information do we have:

- **measurements, from Dutch stations (LML)**
- **Model results for PM10 and ozone**
- **meteorological situation (ECMWF)**
- **other compounds simulated by the model (gases and aerosol components)**
- **persistence – measurements of yesterday**

Model results:

Lotos-Euros v1.7, 4-year run, European + zoom run, 7 km resolution.



Related studies done in the past

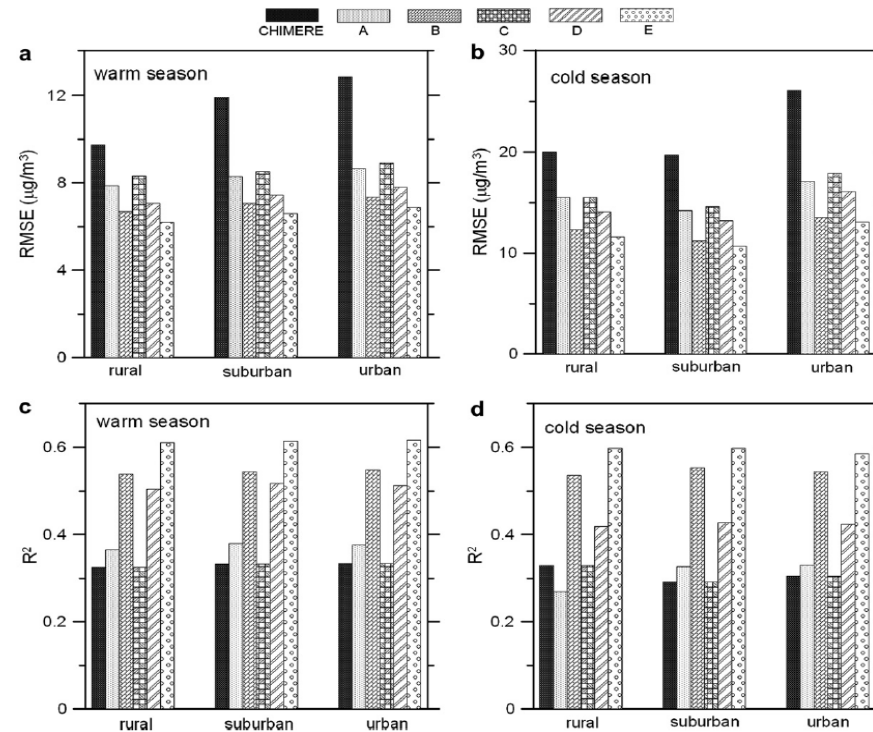
My research project is partly based on the study done by Konovalov.

He did a multiple linear regression for PM10 on CHIMERE model output at stations in Europe.

Table 1

The considered configurations of input variables of statistical models.

Configuration index	Predictors
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 ($D + 1$) + $PM_{10}^{obs}(D + 0)$





How can we use this information?

- **Search for mutual correlation:** Linear regression and choice of variables/predictors
 - Fit measurements against all others
 - Fit against subsets
 - (obs-model) against other predictors
- **Training years are 2003 till 2005**
- **Control year is 2006**



Configurations for multiple regression:

A Meteo parameters (D+1)

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

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

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

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

F Components of total $PM_{10}(D+1)$ + $PM_{10}^{CTM}(D+1)$

G Components of total $PM_{10}(D+1)$ + $PM_{10}^{CTM}(D+1)$ + meteo parameters (D+1)

H Components of total $PM_{10}(D+1)$ + $PM_{10}^{CTM}(D+1)$ + meteo parameters (D+1) + $PM_{10}^{obs}(D+0)$

Compounds and precursor trace gases: SO₂, NH₃, NO_x, HNO₃, ppm₁₀, ppm₂₅, SO_{4a}, NO_{3a}, NH_{4a}, BC, Na_f and Na_c.

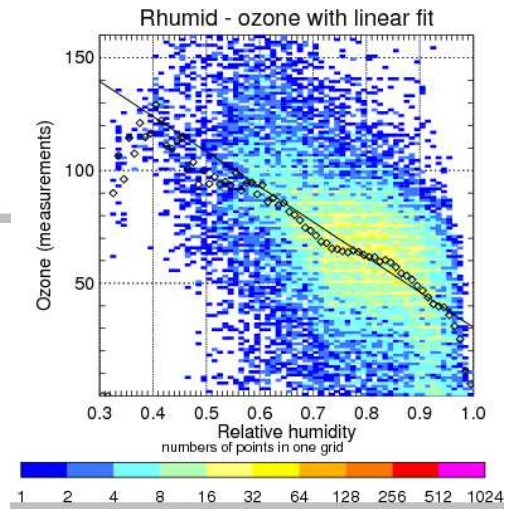
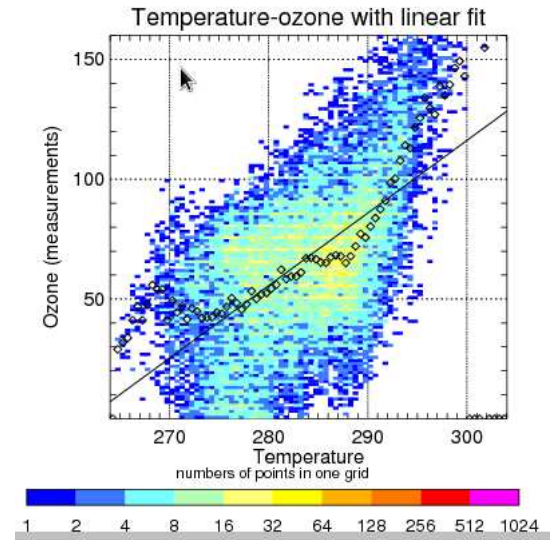
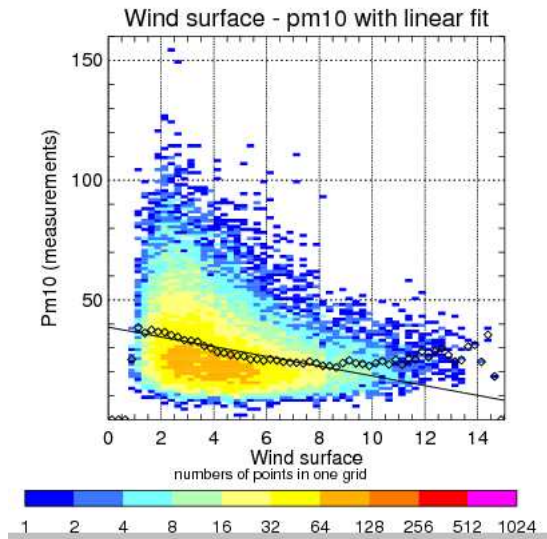
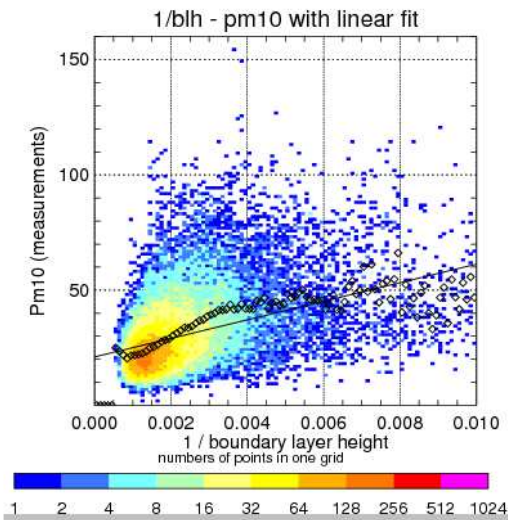


How can we do a multiple linear regression

- **R (routines "lm" and "step")**
a language for statistical computing, the routines linear modeling (lm) and step are used to fit the regression coefficients.
step select a suitable model by dropping terms and preserving hierarchies. Step gives us the variables with there coefficients which are significant enough.
- **IDL**
- **Fortran**

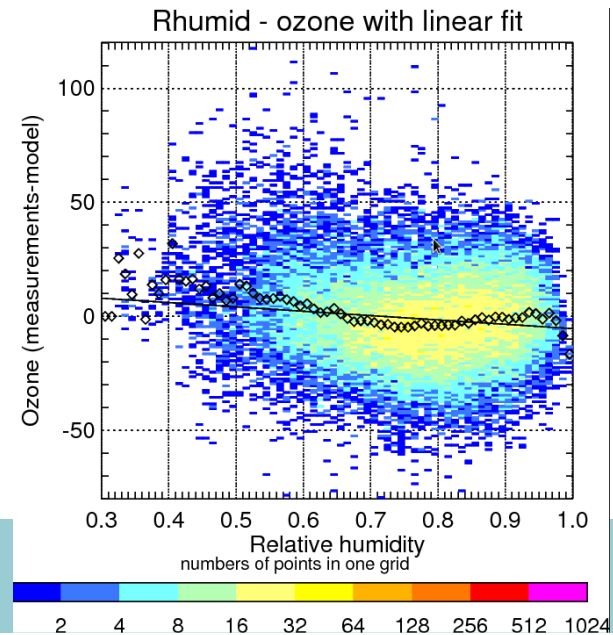
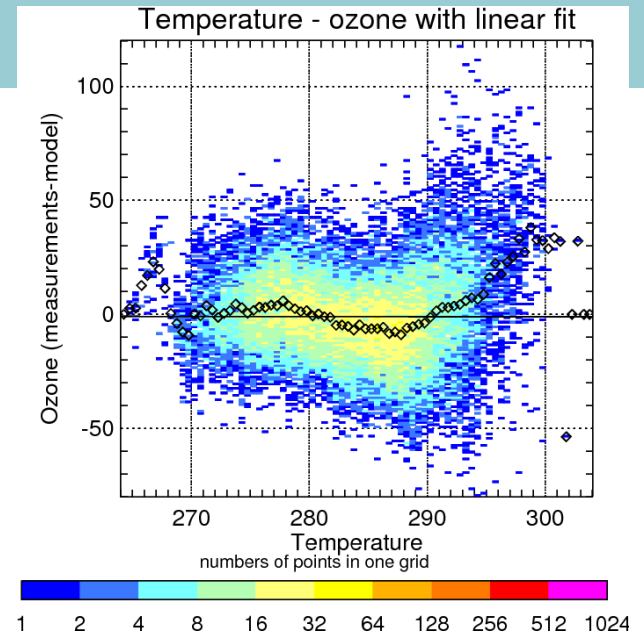
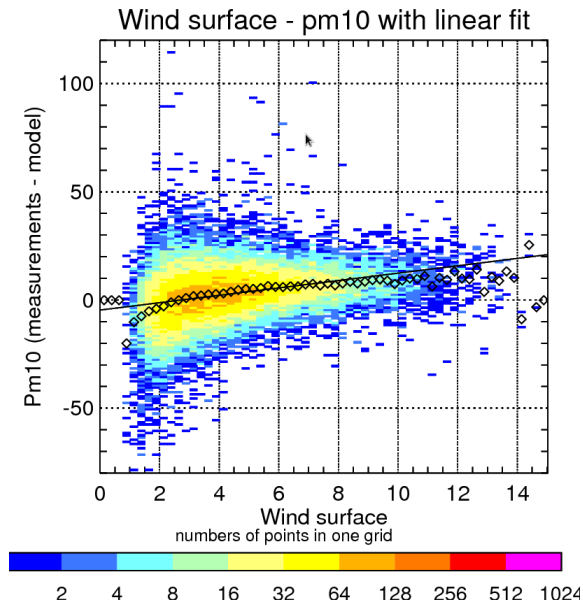
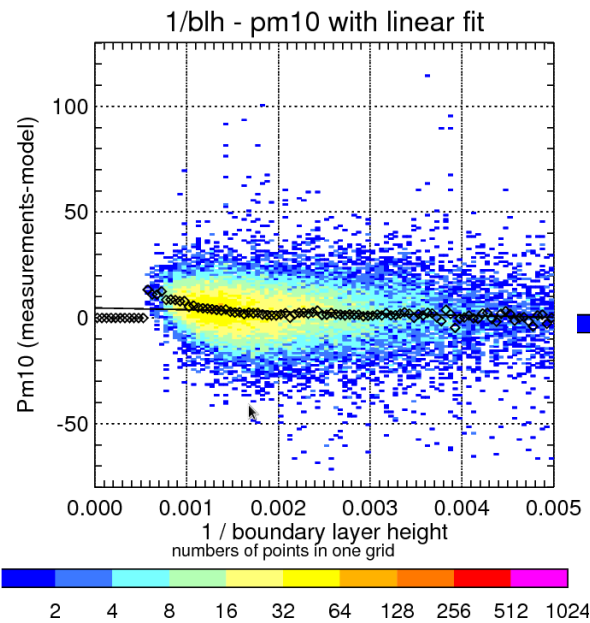


Meteorological variables: Temperature, wind speed, wind direction, rain, relative humidity, boundary layer height and cloud cover



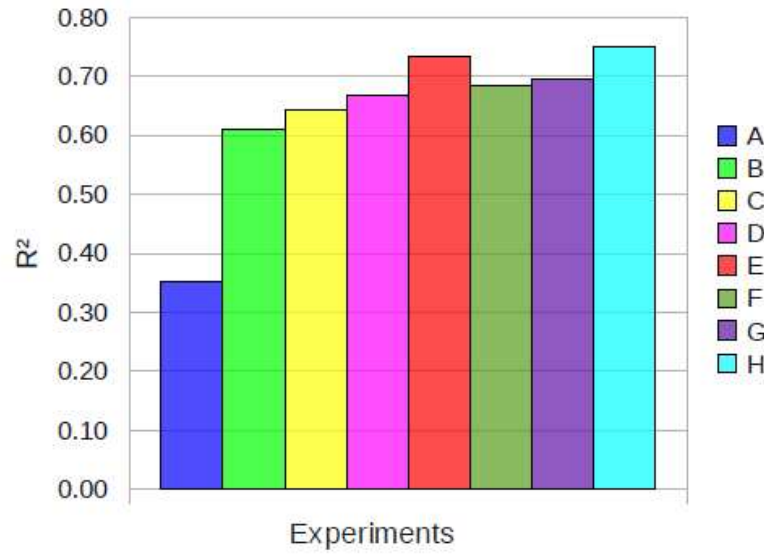


Model - measurements

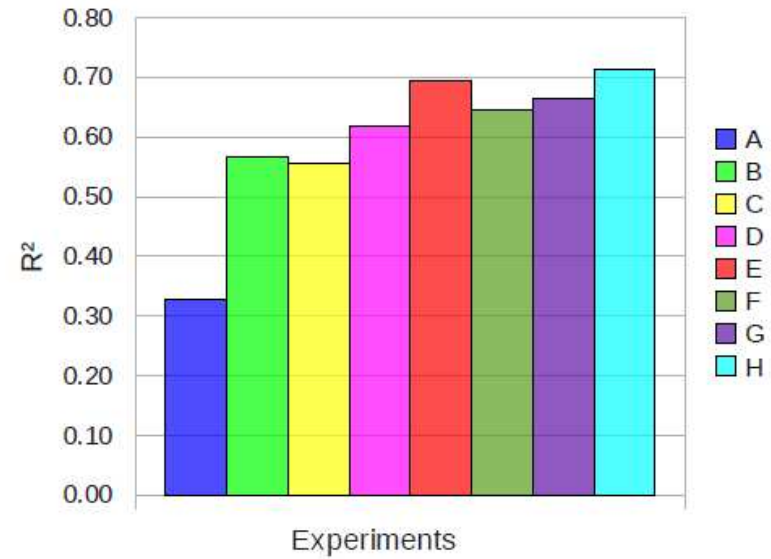




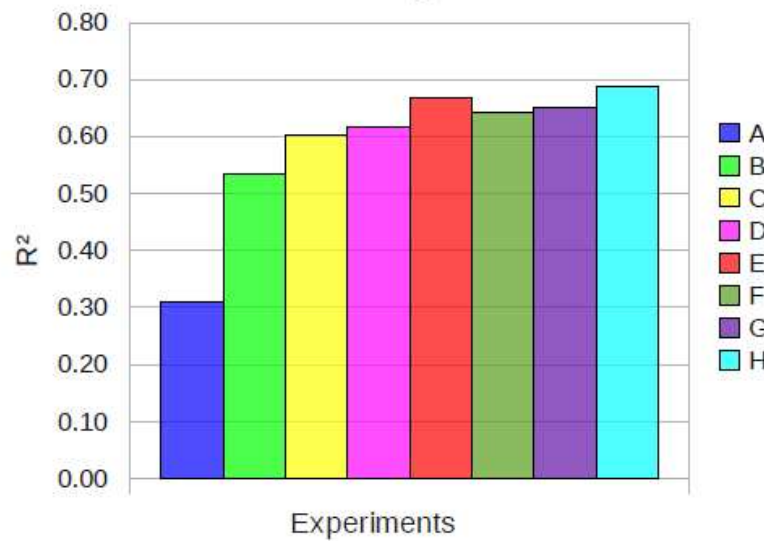
Vredepeel



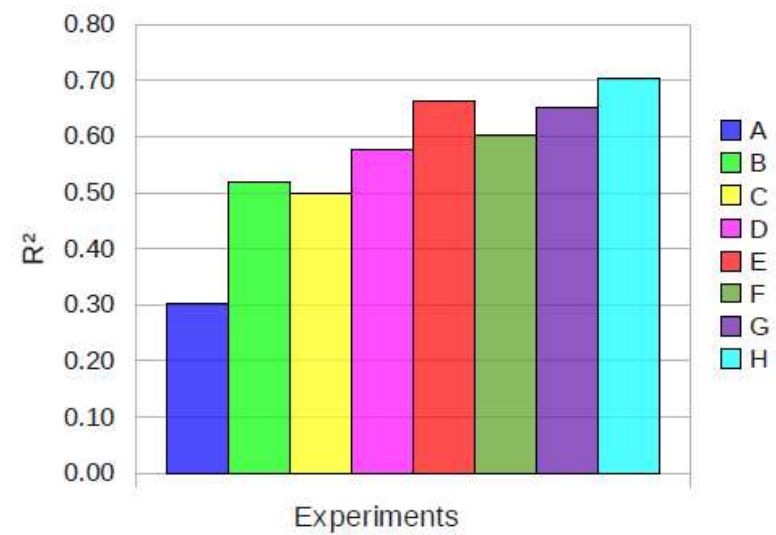
De Zilk



Eibergen

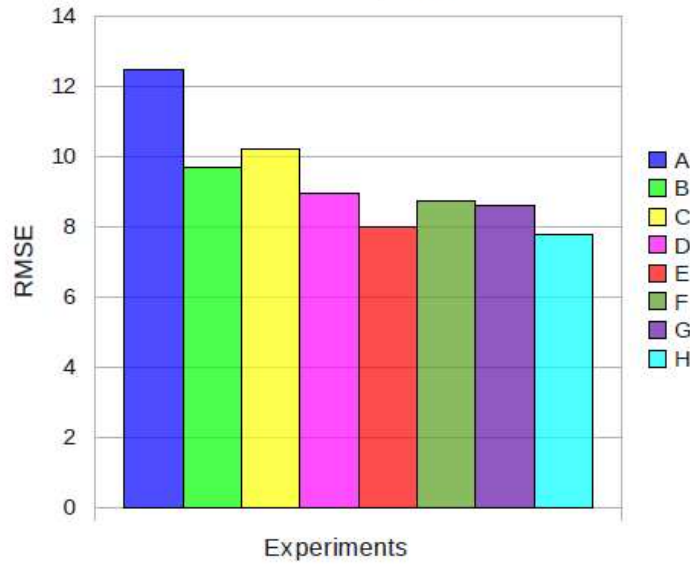


Kollumerwaard

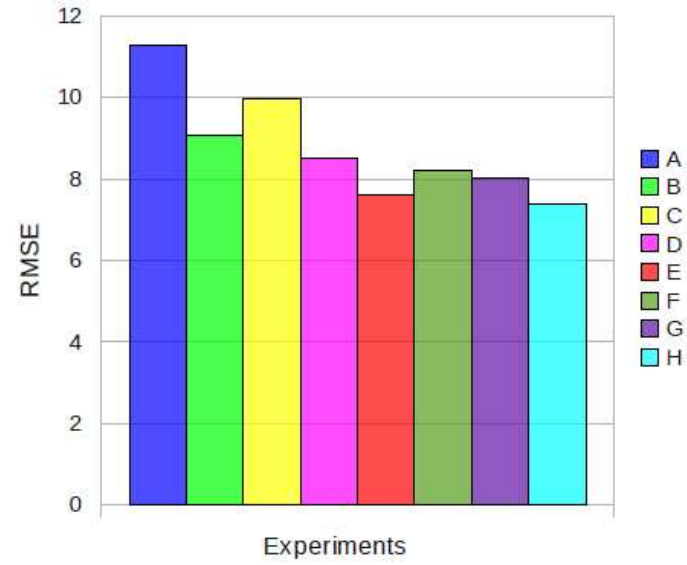




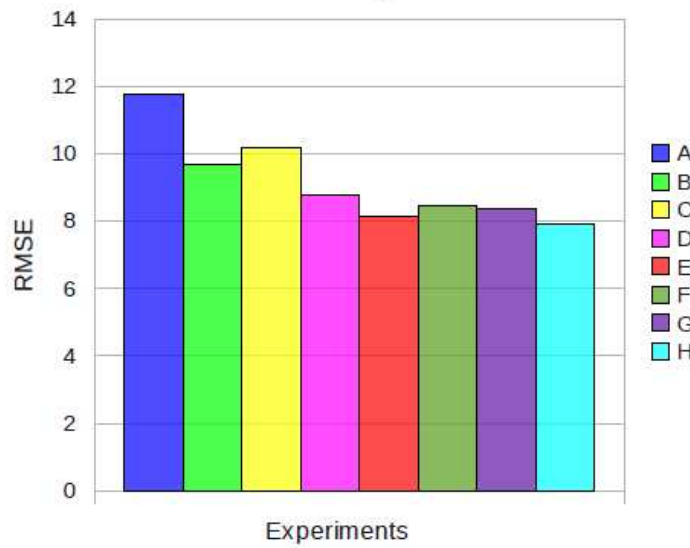
Vredepeel



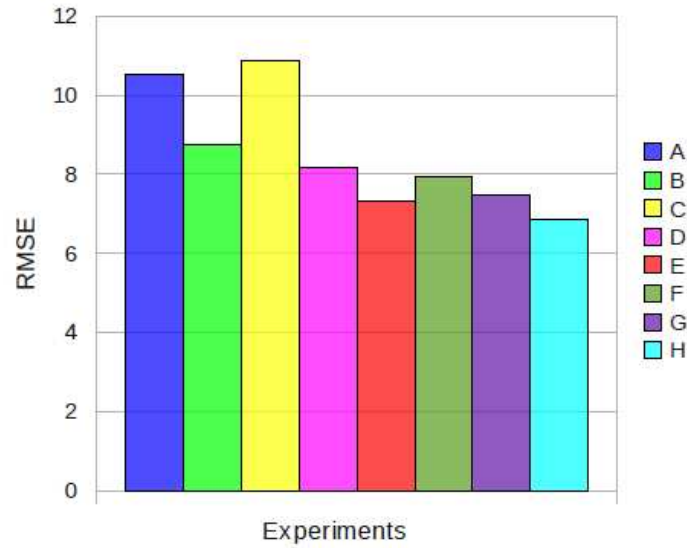
De Zilk

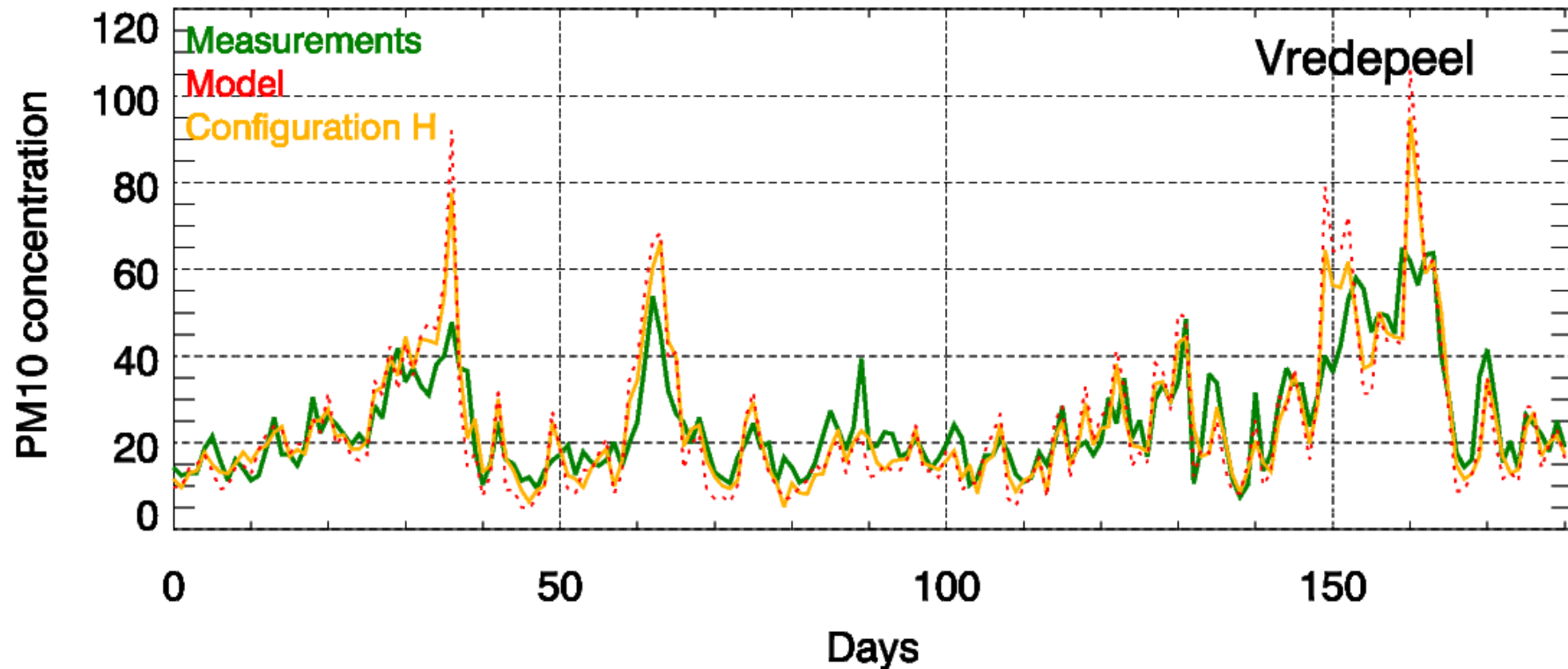


Eibergen



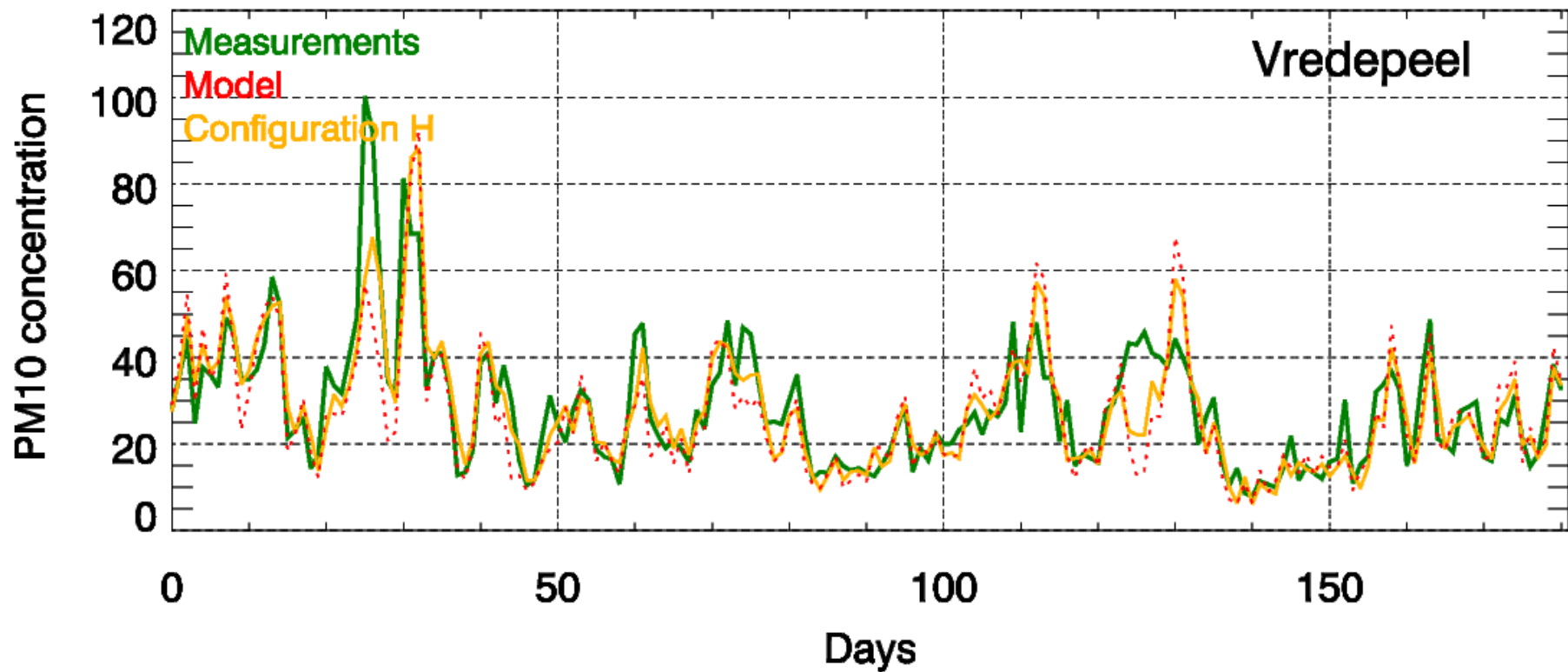
Kollumerwaard





Multiple regression, July-December 2004

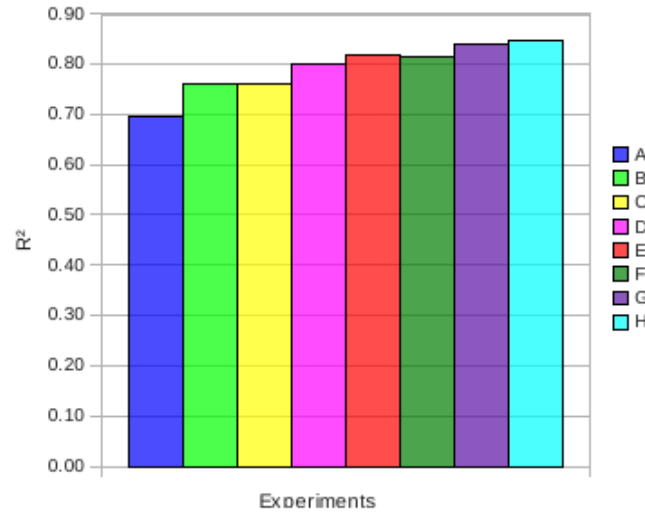
The model, persistence, wind speed, boundary layer height, the relative humidity and wind speed on the surface, so₂, ppm₂₅, no₂, hno₃, no_{3a}, na_f and na_c.



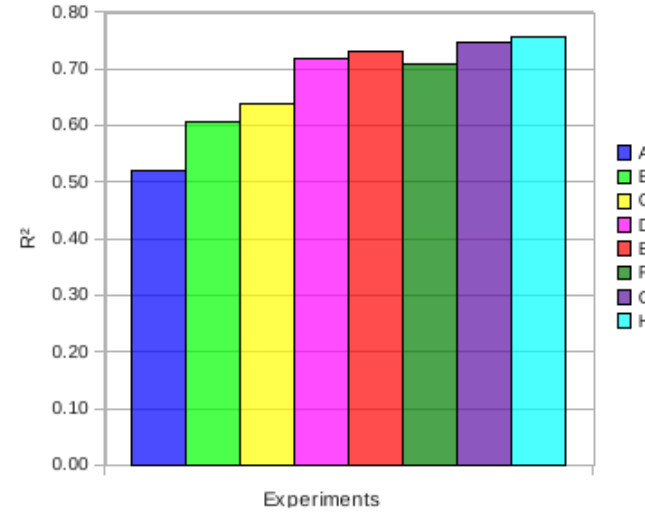
Multiple regression for 2006, also 2th part of the year and with the same predictors as the slide before.



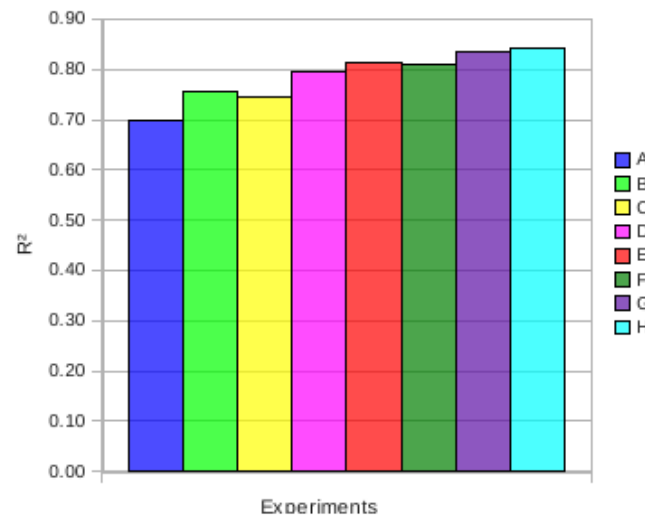
Vredepeel



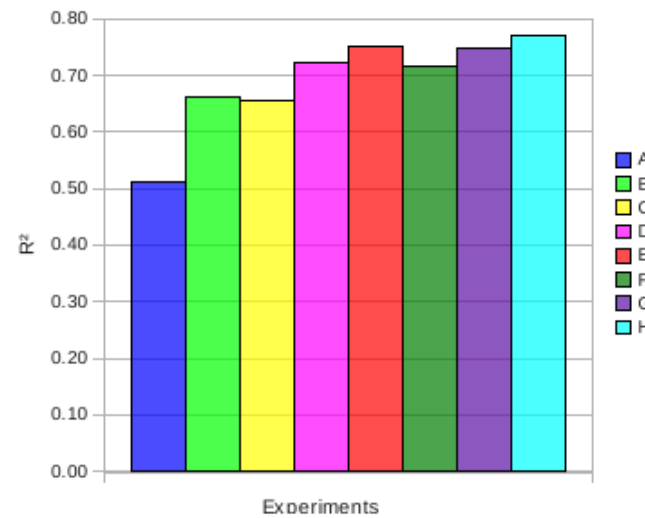
De Zilk



Eibergen

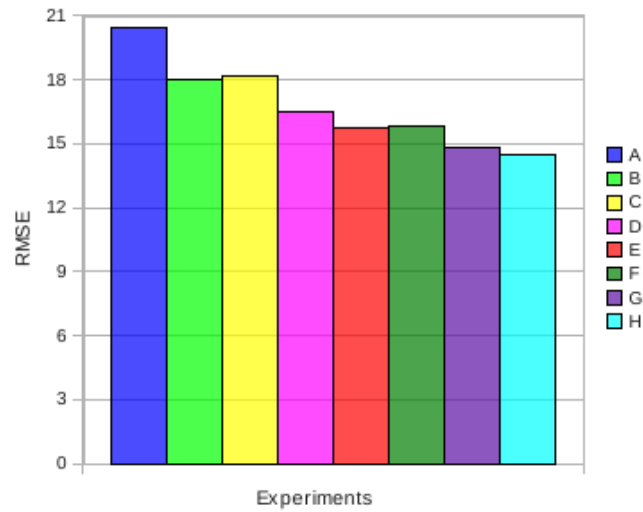


Kollumerwaard

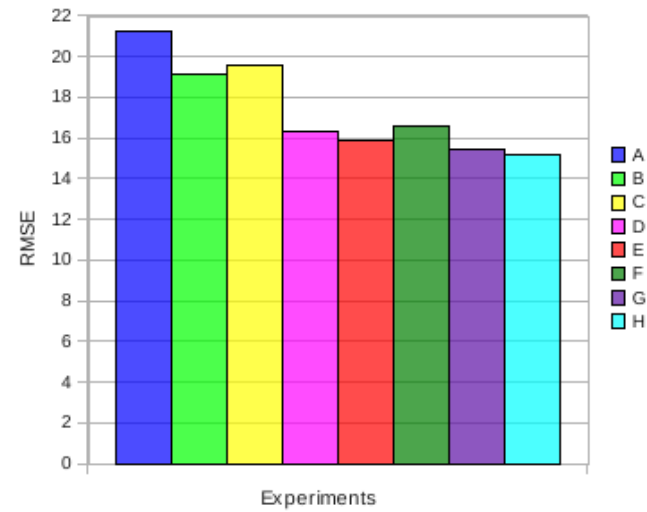




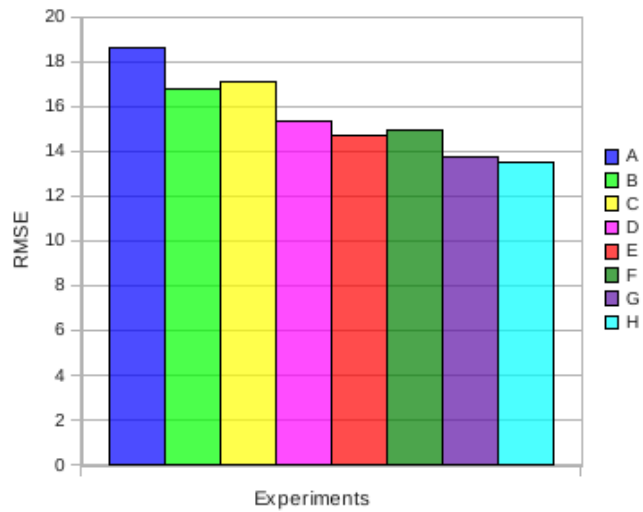
Vredepeel



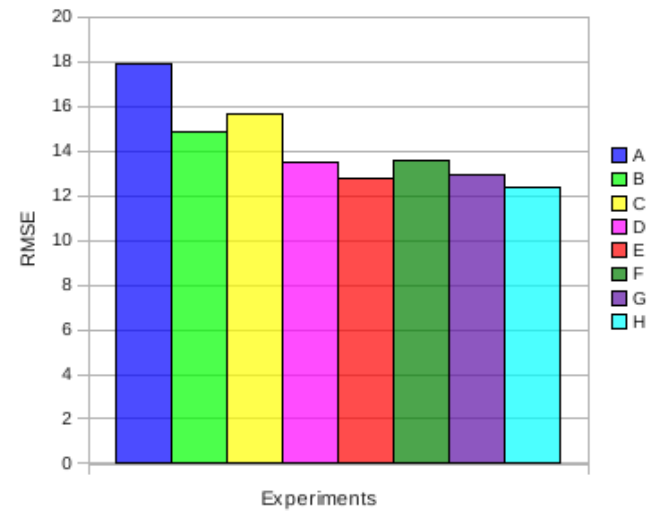
De Zilk

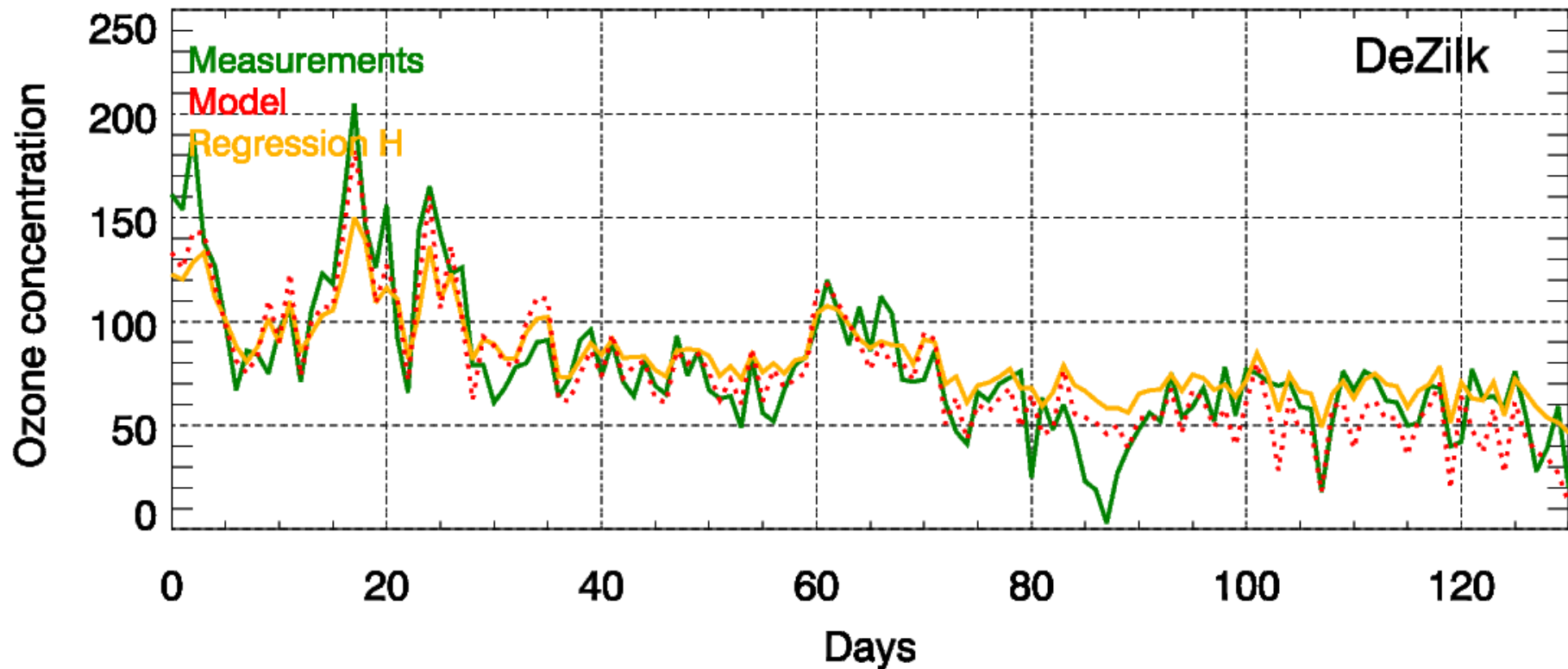


Eibergen



Kollumerwaard





Multiple regression for July till December 2006.

The model, persistence, temperature, meridional wind speed, the relative humidity on the surface, total cloud cover, wind speed, nh4a, nh3, ppm25, no2, hno3, no3a, so4a, bc and na_f.

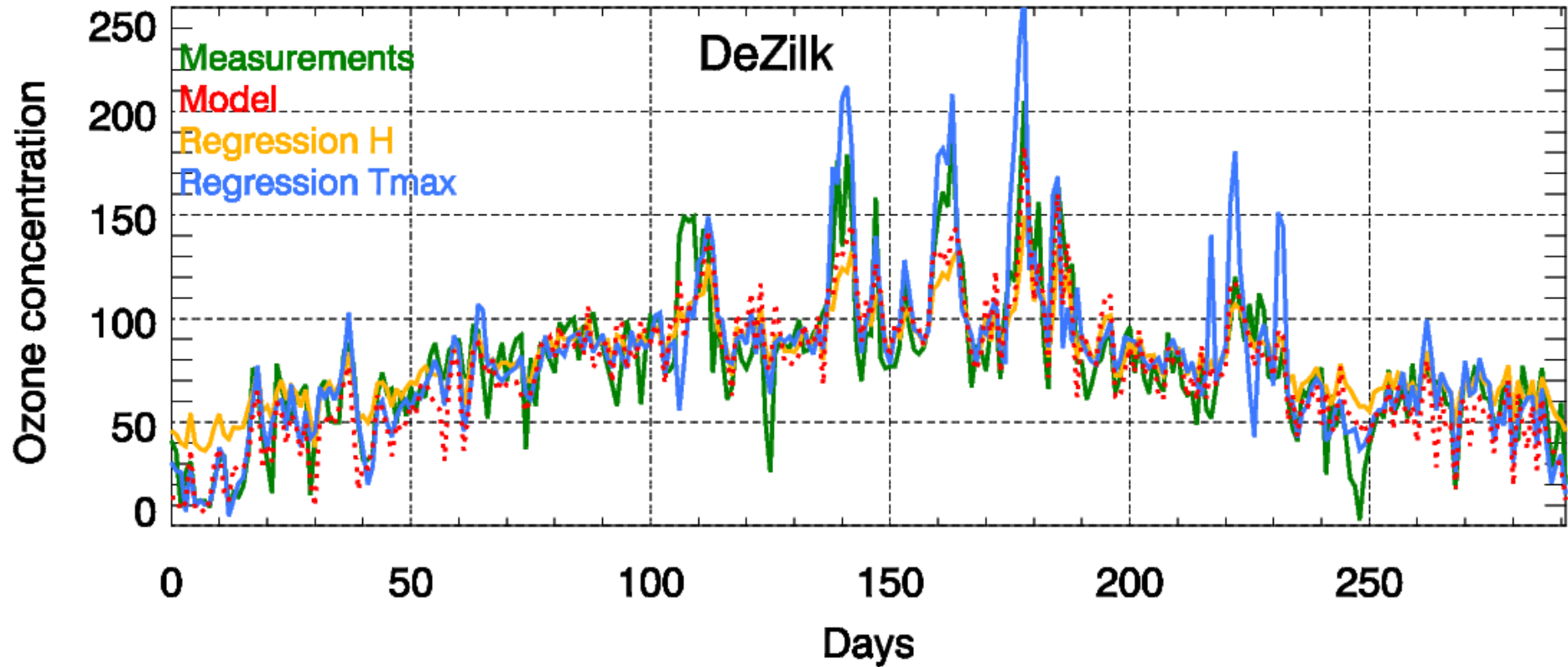


Splitting the data set to improve the model output for pm10 and ozone

- Based on summer and winter:
 - Class 1: summer (April till September)
 - Class 2: winter (October till March)
- Based on the boundary layer height and the wind speed:
 - Class 1: blh above 400m
 - Class 2: blh below 400m and wind speed above 4 m/s
 - Class 3: blh below 400m and wind speed below 4 m/s

Only for ozone:

- Based on the temperature at the hour of the ozone maximum:
 - Class 1: temperature is above 20 degrees of Celsius
 - Class 2: temperature is below 20 degrees of Celsius



Multiple regression for the year 2006.

The model, persistence, temperature, meridional wind speed, the relative humidity on the surface, total cloud cover, wind speed, nh4a, nh3, ppm25, no2, hno3, no3a, so4a, bc and na_f.



Conclusions:

Best multiple linear regression for pm10 is with configuration H, so with the model, persistence, meteorological parameters and other compounds.

For ozone we first split the data set at a temperature of 20 degrees and then we used configuration H as multiple linear regression input.

Further research:

- How to deal with variability among stations, and space in between stations?
- Do we want to use persistency?
- Approach to optimise threshold exceedance?
- Sampling, splitting data in other subsets.