Royal Netherlands
Meteorological Institute
*Ministry of Infrastructure and the
Environment*

The influence of climate change on air quality in the Netherlands

A statistical analysis

Lennard Jansen

De Bilt, 2013 | Scientific report; WR 2013-01

The influence of climate change on air quality in the Netherlands

Versie	1.0

Date	January 31, 2013
Status	Final

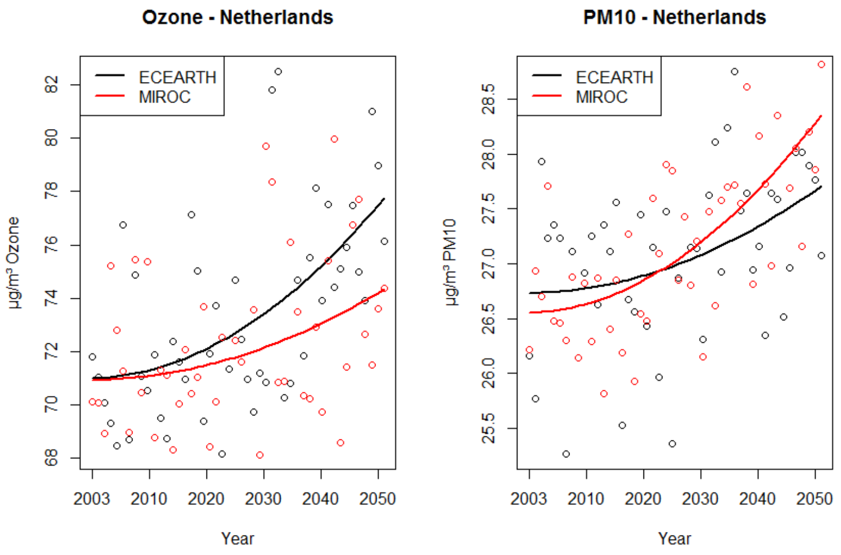
KONINKLIJK NEDERLANDS METEOROLOGISCH INSTITUUT

The influence of climate change on air quality in the Netherlands

A statistical analysis

Lennard Jansen

31-1-2013



Preface

This report is the result of an internship at the KNMI from November 2012 to January 2013 under the guidance of dr. Michiel van Weele and dr. Henk Eskes of the Chemistry and Climate division, department of Climate and Seismology. The internship is part of the master program Statistical Science of Life and Behaviour Sciences of the Mathematical Institute at Leiden University.

Contents

1. Introduction

Recent studies show that in the upcoming decades the likelihood of weather conditions changing is high (IPCC, 2007). How exactly these conditions are going to change is hard to predict, but one can be certain that some areas of the world will undergo more change than others. Because this change in climate can have influence on the environment in the Netherlands, this climate research area is one of most important of Royal Netherlands Meteorological Institute (KNMI).

As stated above, predicting the future climate is difficult. Accurate prediction models require the inclusion of most of the variables influencing the outcome. As one can imagine, the world climate is extremely complex so a significant part of the total variance can't be explained by the current models yet. Therefore researchers at KNMI use climate scenarios together with an ensemble of models and climate runs instead of one prediction.

One of the environmental aspects that can be influenced by a changing climate is air quality. Research shows that air quality changes as a function of meteorological variables like temperature and boundary layer height (EEA[1], 2012). It is because of these relations that it could be possible to estimate the change in air quality due to the climate changing over time. In this report a statistical analysis will be done to quantify this dependency and to estimate the changes of air quality in the future in the Netherlands based on different regional climate runs.

Air quality is often described in terms of concentration of air pollutants in the air. Examples of air pollutants are sulfur dioxide, nitrogen oxides, ammonia, ozone and particulate matter. Because these substances are harmful for human health, the National Institute for Public Health and Environment (RIVM) monitors the concentrations continuously at multiple locations in the Netherlands. Two of these substances, ozone (O3) and particulate matter (PM10), will be the focus of this report.

Ozone can be found in multiple layers of the atmosphere. In the stratosphere its presence is good for human skin as it filters UV rays. Near the surface, however, it is harmful for human health and is considered an air pollutant. Ozone, in contrast to other substances like PM10, is not emitted directly in the air but is formed through chemical reactions under influence of sunlight. PM10 refers to particle matter of which the diameter is smaller than 10 micrometer and is harmful for human health when inhaled. Traffic, power plants but also mineral dust and sea salt are sources of PM10.The most important sources of anthropogenic PM10 emissions in 2010 in Europe were the sectors 'Commercial, institutional and households' (42% of total emissions), 'Industrial processes' (15%), 'Road transport' (14%) and 'Agriculture' (10%). The 'Commercial, institutional and households' sector includes combustion-related emissions from sources such as heating of residential and commercial properties (EEA[2], 2012). Also natural aerosols such as mineral dust and sea salt are sources of PM10 (EEA[2], 2012).

The aim of the current research is to project both ozone and PM10 concentrations for the Netherlands until 2050 using regional climate model output, based on relations between modeled meteorological variables and observed ozone and PM10 values in the past.

In order to achieve this goal, a statistical model is developed based on training data which consists of the above mentioned meteorological variables and the air pollutants in the period 2003 to 2006. The output of the regional climate model will be used as new data in order to give an indication of future pollution concentrations.

This report is organized in the following way. Section two describes related work done in the past at KNMI on which this current report is based. Section three covers data analysis which shows the distributions of the data and what transformations have been applied. Section four describes the methods used in this study. After this in section five, the results of the study will be shown. The conclusions and discussion can be found in section six.

2. Previous work

In 2011 an intern at KNMI, A. Pijnappel, did a research project in which the aim was to improve the accuracy of the chemistry transport model LOTOS-EUROS through statistical post processing for multiple locations in the Netherlands (Pijnappel, 2011). For this analysis the daily mean concentration of PM10 and daily maximum concentration ozone was used. Multiple linear regression had been carried out whereby the model parameters were estimated by ordinary least squares. The dependent variables were ozone and PM10 provided by the National Air Quality Monitoring Network of the Netherlands (LML). The independent variables were the modeled data of PM10 and ozone provided by LOTOS-EUROS, the meteorological data provided by the European Center for Medium-Range Weather Forecasts (ECMWF), components of PM10 and precursor trace gasses. Data was available for the years 2003-2006. The years 2003-2005 were used as training set for estimating the regression coefficients, whereas the year 2006 was used as a test set.

Because Pijnappel already built a statistical model in order to be able to make predictions of air quality based on meteorological variables, the first idea for the current research was to use the estimated regression coefficients of her model, and predict future ozone and PM10 concentrations by providing the meteorological regional climate scenario data as new input.

There were, however, some aspects that could possibly be improved. First of all for the current research, not all LML stations that Pijnappel used were to be preferred right now. Consequently, the regression models for the stations that would differ from her choice had to be fitted anyway. Secondly, when using the linear model to make predictions for the future with satisfying accuracy, we are assuming that this model predicts well enough. Even though there is no doubt about the models used in her analysis predicting reasonably well, it could always be useful to see whether it is possible to do it even better (e.g. by adding first order interaction effects). It is for this reason that we tried to find aspects of the analysis that could be improved.

The scatter plots of boundary layer height and temperature versus ozone concentration (Figure 1) show a first potential point of improvement.

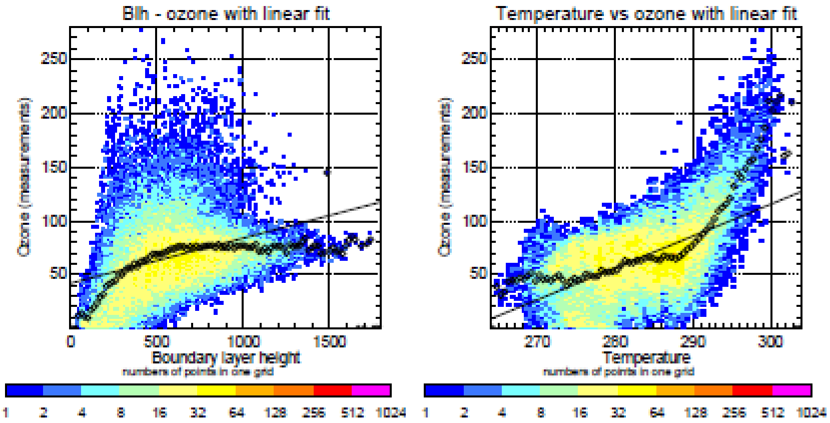


Figure 1. Results derived from Pijnappel 2011 showing the limitations of a linear approach to fit ozone as a function boundary layer height (left) and temperature (right).

The linear model performs well when the underlying relationship is approximately linear. Looking at Figure 1, one can see that this is not a good approximation for these two variables. Especially the temperature plot is problematic. When predicting air pollution, one is interested in the cases when the pollution is maximum and for instance exceeds a certain health norm. The linear fit is clearly underestimating the ozone concentration when temperatures are high and therefore not quite satisfying.

Another aspect that could be improved is the reliability of the parameter estimates by using cross-validation instead of separating the dataset in a training and a test part. Weather conditions can vary by a vast amount, even between years (i.e. some summers can be relatively hot and dry, whereas others can be cold and wet). Therefore instead of picking certain years as training set, cross-validation will ensure that this randomness doesn't influence the results as much.

With these points of improvement in mind we thought it was worthwhile to start the analysis from scratch in order to be able to make the most accurate predictions of ozone and PM10.

3. Data analysis

The data used in this report can be split in three different parts, namely the air pollution measurements, the meteorological measurements and the model data coming from the regional climate runs.

3.1 Air pollution measurements

The National Air Quality Monitoring Network (LML) measures concentrations of various air pollutants including ozone and PM10 at more than 50 stations in the Netherlands. In this study data from the period 2003-2006 is used from LML stations located in Vredepeel, De Zilk, Kollumerwaard, Wieringerwerf, Eibergen, Philippine and Balk. Reasons for choosing these stations include availability of ozone and PM10 data for the given period, reasonably equal spread across the Netherlands and included only the stations that were not influenced by very local pollution sources. Figure 2 shows the locations on the map.

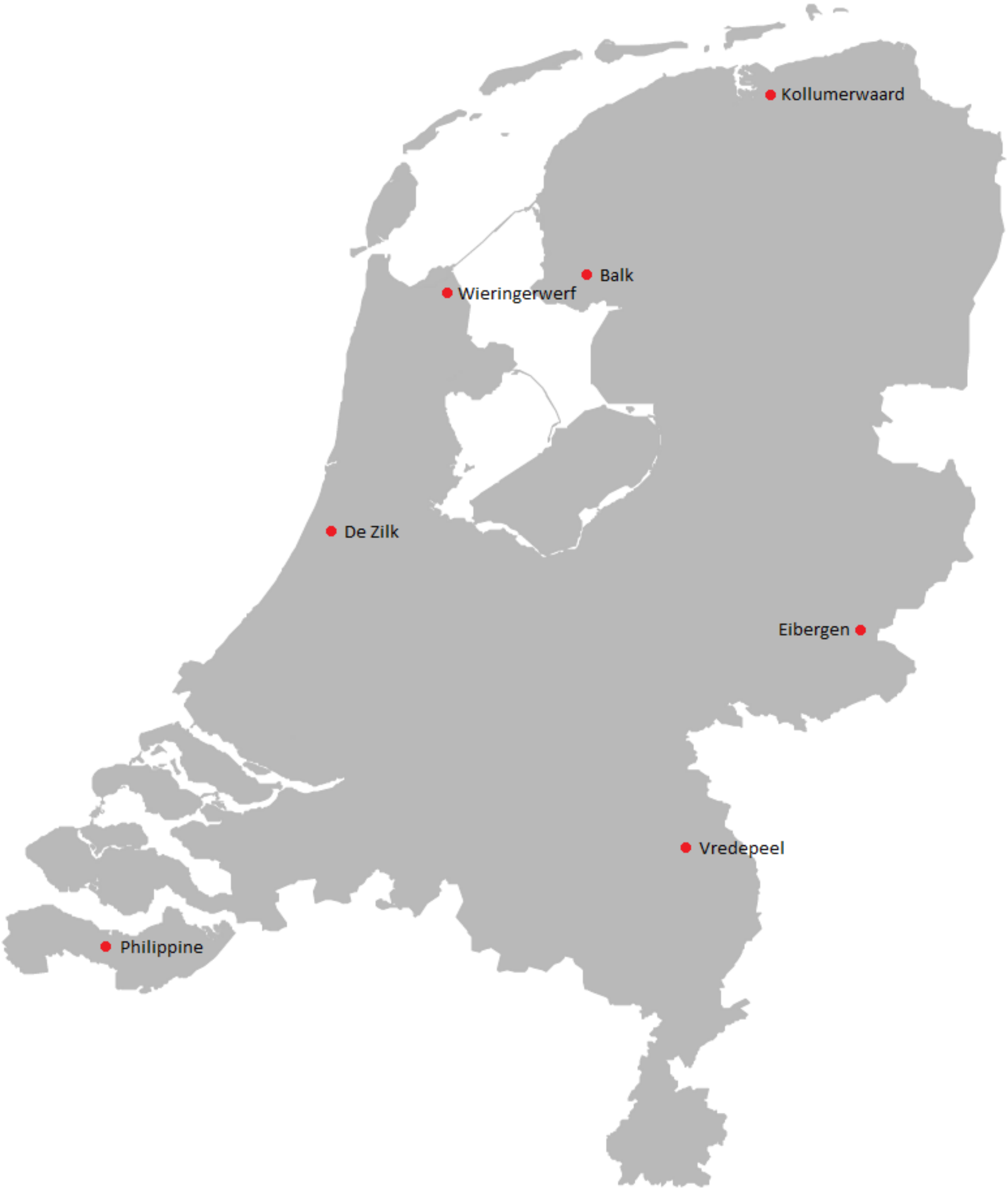


Figure 2. Map of the Netherlands showing the spread of the selected locations of the LML stations. In the north west Balk has been selected for the ozone observations and Wieringerwerf for the PM10 observations. All other stations measured both ozone and PM10.

Because of European air quality norms, maximum daily ozone and average daily PM10 concentration expressed as micrograms per cubic meter air were used. Figure 3 shows the distribution of ozone per station.

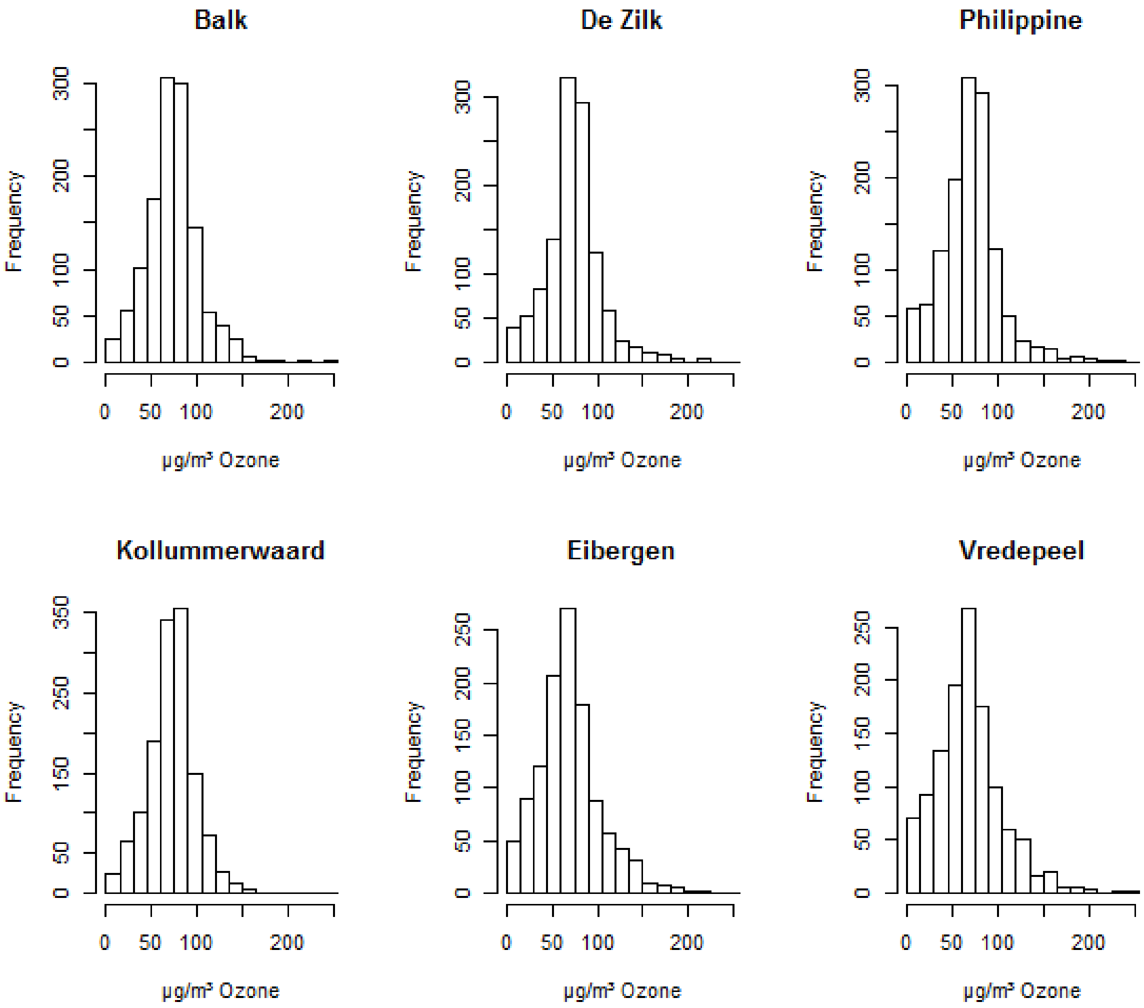


Figure 3. Ozone frequency distribution in bins of 15 µg/m^3 for the daily maximum ozone concentrations during the period 2003-2006 at the six selected LML stations.

It can be seen that the distribution of ozone is reasonably similar across stations. Both means and variances do not seem to differ much. The most common concentrations lie between 50 and 100 µg/m^3. Figure 4 shows the distribution of PM10 per station.

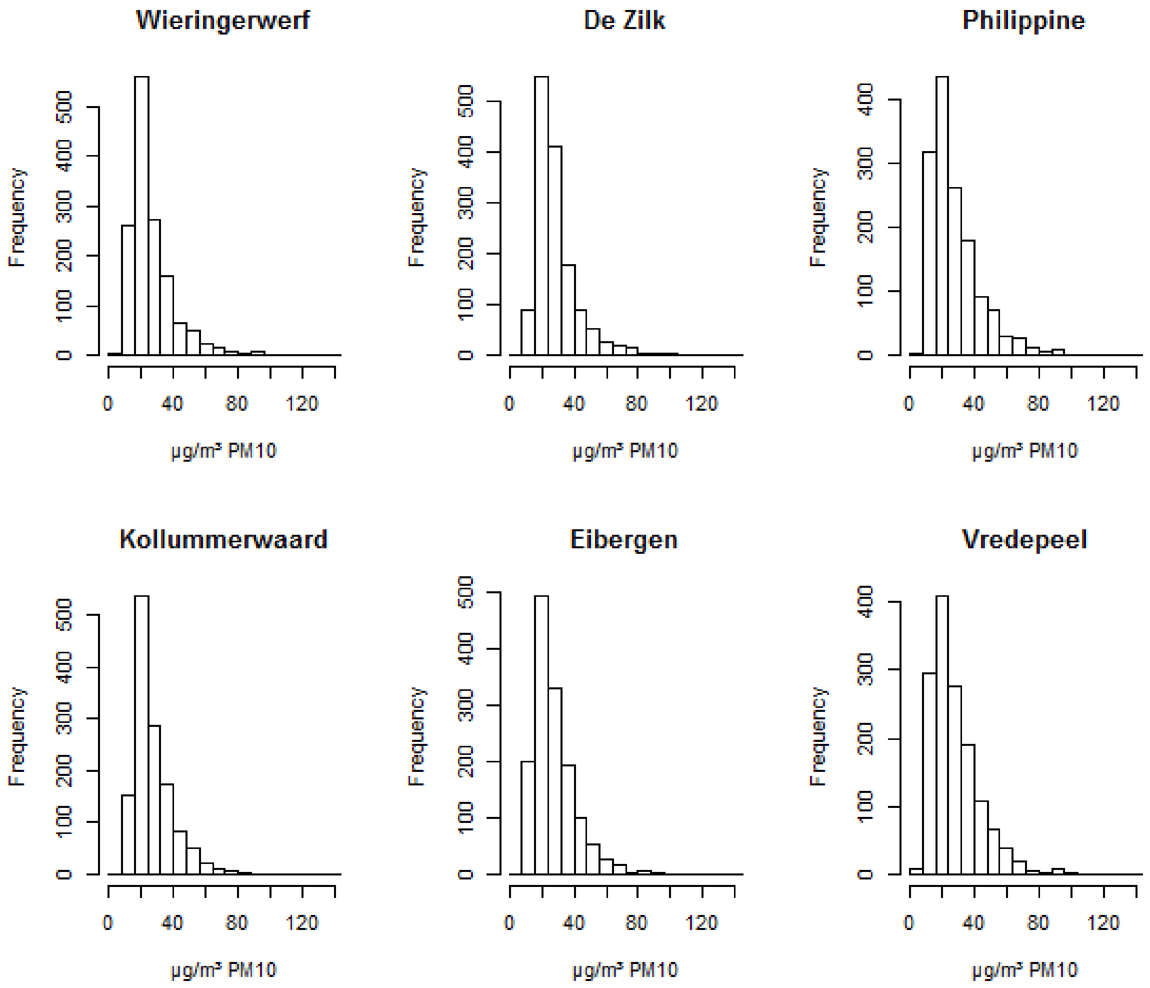


Figure 4. PM10 frequency distribution in bins of 8 µg/m^3 for the daily average PM10 concentrations during the period 2003-2006 at the six selected LML stations.

The graphs concerning PM10 show the same phenomenon. The shape and location of the distributions look similar regardless of the measurement site. Compared to the ozone concentrations, PM10 tends to be less spread out and more peaked.

To get even a better idea of what the data looks like, Table 1 and 2 show the characteristics of respectively the ozone and PM10 distributions in more detail.

	Median	**Mean**	**Max.**	**Exceedences**	**Nr. Obs**	**Missing Obs**
Balk	67	69.18	244	105	1198	263
De Zilk	71	71.00	235	73	1286	175
Philippine	74	74.49	224	71	1175	286
Kollummerwaard	67	69.57	221	98	1158	303
Eibergen	73	73.83	250	79	1237	224
Vredepeel	74	72.52	176	45	1342	119

Table 1. Summary statistics regarding the ozone concentrations for the period 2003-2006. Apart from median and mean, also the observed maximum concentration is shown. Exceedences refer to the total number of days in the four year period wherein the maximum concentration is higher than the threshold of 120 µg/m^3. The total number of observations and the amount of missing days are also shown.

	Median	Mean	Max.	Exceedences	Nr. Obs	Missing Obs
Wieringerwerf	24.05	27.88	104.3	122	1429	32
De Zilk	23.12	27.65	109.1	127	1439	22
Philippine	25.11	28.68	108.0	102	1434	27
Kollummerwaard	22.02	25.99	114.2	83	1420	41
Eibergen	24.41	28.03	125.6	99	1432	29
Vredepeel	23.60	27.26	104.2	78	1328	133

Table 2. Summary statistics regarding PM10 concentrations. The same statistics are shown as in table 1. Exceedences refer to the total number of days the PM10 threshold of 50 µg/m^3 is exceeded.

Although the shapes of the distributions seem to be the similar, Tables 1 and 2 show that there are differences. The most striking finding seems to be that although from research we would expect station Vredepeel to have relatively high PM10 concentrations compared to the others (because of the region it is in), but this is not that obvious from our LML data. One possible reason for this could be that local pollutant sources are small.

The European Union introduced norms regarding ozone and PM10 concentrations to which the Netherlands has to adhere. With respect to ozone the norm says that the maximum daily eight hour mean may exceed 120 µg/m^3 25 times over three years. In this study, ozone data were only available in the form of maximum daily concentrations. From research we know that maximum ozone concentration is not far above the maximum eight hourly average. Although this difference is not that large, in this report, when addressing exceedences, we will refer to whether the daily maximum exceeds the 120 µg/m^3. On that account, although in Table 1 the total number of exceedences are given for the period of four year, it can be seen that on average Balk doesn't meet the norm and that there is a high chance that Kollumerwaard is not able to either.

Regarding PM10 the EU norm states that the average daily concentration should not exceed 50 µg/m^3 more than seven times per year. Table 2 clearly shows that on average this norm isn't met at any station.

3.2 ECMWF Meteorological data

The meteorological data used in this study is based on ECMWF operational data (OD) for the model grid cell in which the LML measurements stations are situated. The data are available for the same LML stations as the measured ozone and PM10. We will refer to these model data as observed meteorology, because the ECMWF analysis is based on many weather observations.

The original variables available in the meteorological dataset that were relevant for the analysis are denoted in Table 3 and are averages per day.

	Measurement Unit
Temperature	Degree Kelvin at 10m
Zonal wind speed (U)	Meter per second at 10m
Meridional wind speed (V)	Meter per second at 10m
Relative humidity	0 (min) to 1 (max) at 10m
Boundary layer height	Meter from earth surface
Total cloud cover	0 (clear) to 1 (overcast)

Rain	Millimeter per hour

Table 3. Meteorological variables that have been selected from ECMWF at the grid cells of the LML station and their corresponding units.

Because we were not only interested in the zonal and meridional wind speed, but also would like to say something about wind direction and wind speed separately, we decided to calculate and use the latter in the analysis as well (see Table 4).

	Measurement Unit
Wind speed	Meter per second at 10m
Wind direction	0 (east) to 2π, were ½π equals south at 10m

Table 4. Alternative variables calculated from the zonal and meridional wind speed.

The distributions of the meteorological variables averaged over the stations are shown in Figure 5.

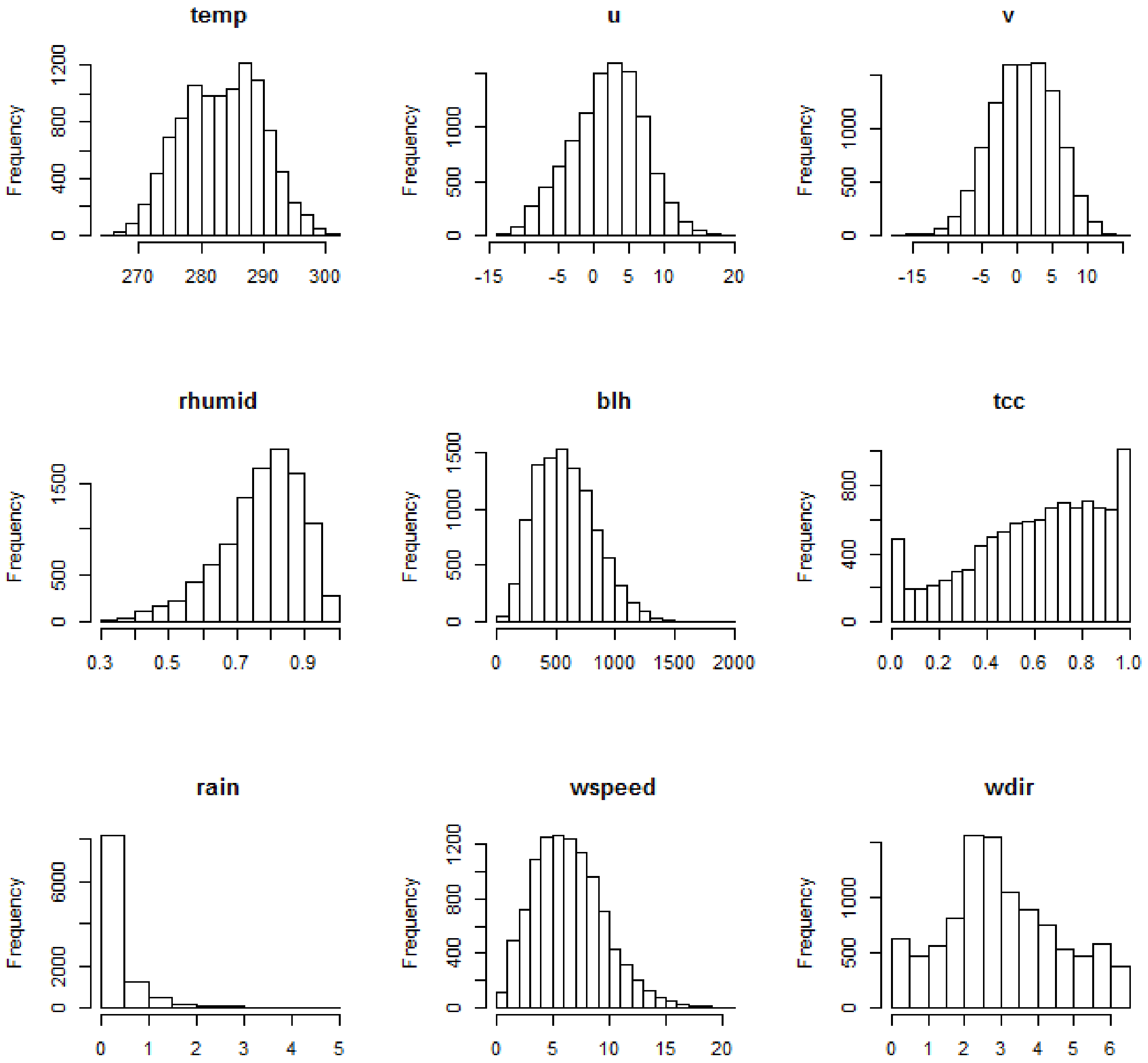


Figure 5. Frequency distributions of the ECMWF meteorological variables for the period 2003-2006 averaged over the LML stations. Units on the x-axis as given in Table 3 and 4.

All variables appear to have distributions that are to be expected, although rain fall is notoriously difficult to simulate with meteorological models and the representativeness of the ECMWF precipitation for the station location might be a limitation, especially in the case of showers which may or may not have passed over the measurement station. The distribution of total cloud cover seems to have rather high density at the tails, but this is a common phenomenon.

Figure 6 and 7 show scatter plots of the meteorological variables against ozone and PM10 respectively.

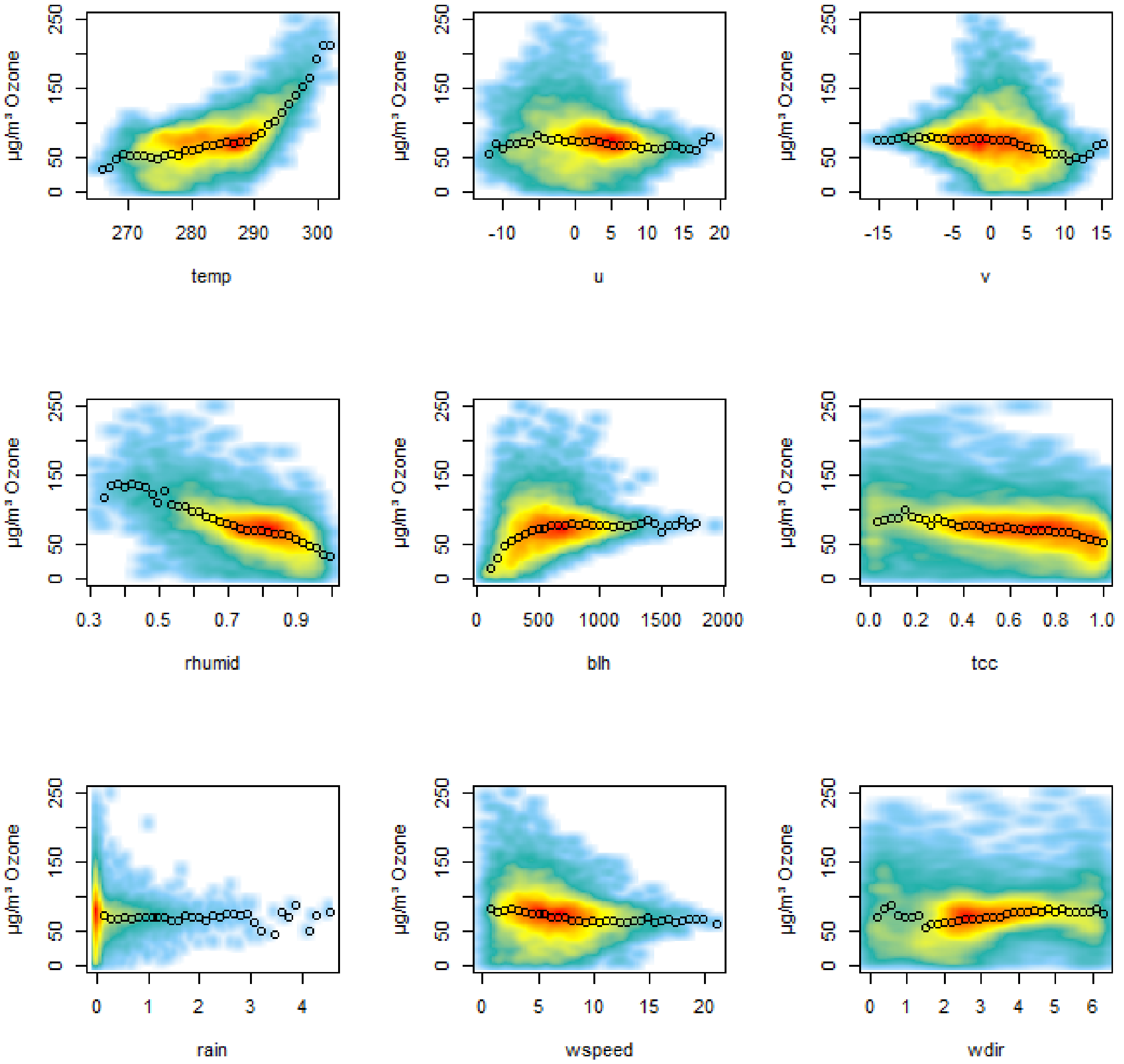


Figure 6. Scatter plots of the observed ozone concentrations as a function of the selected ECMWF meteorological variables. The circles denote the mean ozone concentration per bin (35 bins in each panel). The density of the observations is shown in a color scale ranging from red (high) to blue (low).

As can be seen from the graphs, the relationship between ozone and some of the variables is non-linear. Especially temperature and boundary layer height show nonlinear relationships with ozone concentration.

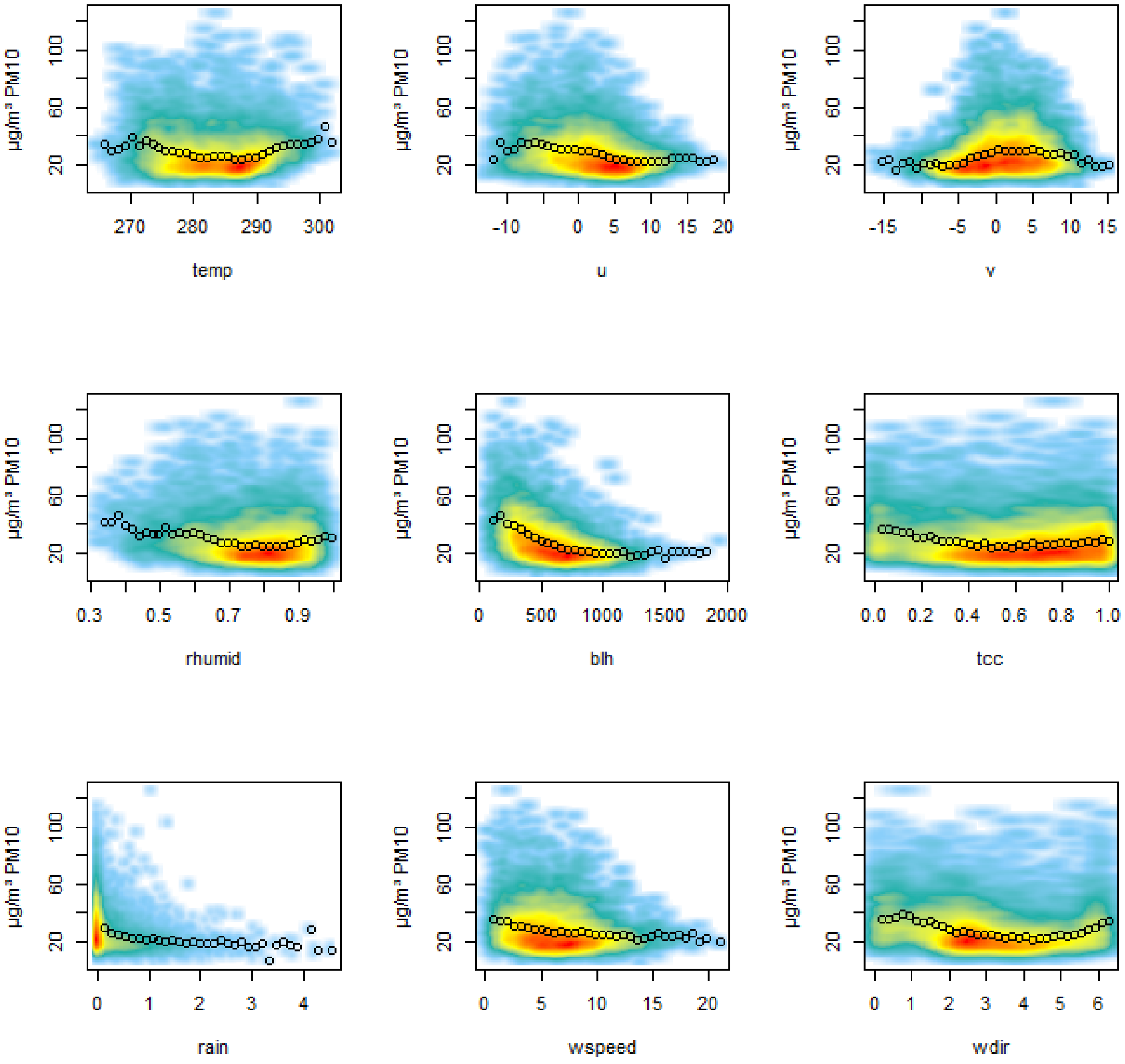


Figure 7. Same as Figure 6 but for the PM10 observations.

The graphs regarding the relationship between PM10 and the meteorological variables show the same phenomenon. Understandably, we shouldn't restrict ourselves to ordinary linear regression models when looking for a model that predicts future data best.

3.3 Regional climate model time series

The Regional Atmospheric Climate Model (RACMO), operational at KNMI, is the regional climate model that is used to generate climate scenarios for the future. A regional climate model is typically used to downscale global climate model results to the regional (Europe) and local scale (Netherlands), and to address physical phenomena at a greater detail than is possible with resolutions of global climate models. RACMO, however, requires input from a global climate model at its lateral boundaries.

To express some amount of uncertainty about the future climate, two RACMO runs have been done with different boundary conditions from global climate model input. One run is based on boundary conditions from a simulation with ECEARTH (model experiment 'wCS6-v420-fECEARTH-Member01'), whereas the other is based on a MIROC simulation (model experiment 'wCS6-v420-fMIROC') . Both runs have been performed in the context of the simulations in preparation of the upcoming IPCC 5th assessment report and provide hourly meteorological data on the same variables that were discussed in the previous section over the period 1950 to 2050. Because daily averaged meteorological data was used, the RACMO climate model data was naturally averaged by day as well.

In order to make use of the RACMO climate model data and make predictions based upon it by applying a statistical model that was fit on the observed pollution and meteorological data, one has to make sure that the distributions of the RACMO data are comparable to the observed data. A reasonable assumption is that the meteorological data discussed previously resemble more accurately the actual weather conditions at each time point than RACMO output. So, it is preferable to transform the RACMO model data in such a way that the distributions of its variables are approximately the same as the observed meteorological distributions. More specifically, both the means, variances and extreme values should correspond for present day. Looking at the relation between temperature and ozone, it is likely that high temperatures in general lead to high ozone concentrations. Therefore it is important for RACMO to be able to predict these extreme temperatures. Hence it is significant not only to have the means of the distributions correspond but also the variances and extreme values. The bias of RACMO becomes clear when looking at Figure 8.

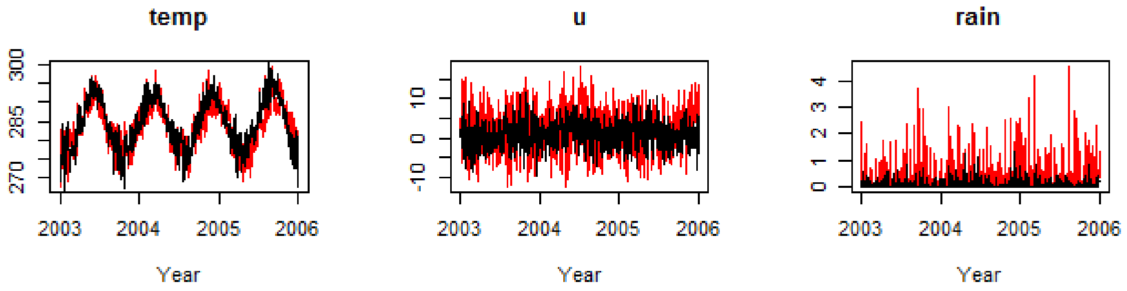


Figure 8. Comparison of the 2003-2006 time series of observed (red) and RACMO-ECEARTH (black) temperature, zonal wind speed and rain. Units on the y-axis as given in table 3 and 4.

Figure 8 shows both RACMO-ECEARTH predictions (black) and the more accurate meteorological values (red) for station Cabauw on temperature, zonal wind speed and rain. As can be seen, RACMO does a reasonable job regarding temperature in terms of variance. Looking at the zonal wind speed (u), however, it fails as it is having problems producing the same spread (winds close to the surface are difficult to model because it depends to a large extend on local conditions). With respect to rain RACMO has trouble regarding both mean and variance.

This, of course, isn't satisfying. Therefore we opted to transform the RACMO in such a way that it resembles the meteorological data more closely.

Let x_{Rji} be variable *i* from station *j* given by RACMO with data in the years 2003 to 2006 and x_{Mij} be variable *i* from the corresponding station by the meteorological dataset in the same period. Then for all *i* and *j*

$$\underset{c1\,c2}{\operatorname{argmin}}\left\{w_{var}\left|var\left(c1\left(x_{Rij}-\bar{x}_{Rij}\right)\right)-var\left(x_{Mij}\right)\right|+\ w_{mean}\left|mean\left(c1\left(x_{Rij}-\bar{x}_{Rij}\right)+c2+\bar{x}_{Rij}\right)\right.\right.$$
$$\left.-\,mean\left(x_{Mij}\right)\right|+\ w_{max}\left|max\left(c1\left(x_{Rij}-\bar{x}_{Rij}\right)+c2+\bar{x}_{Rij}\right)-max\left(x_{Mij}\right)\right|$$
$$+\ w_{min}\left|min\left(c1\left(x_{Rij}-\bar{x}_{Rij}\right)+c2+\bar{x}_{Rij}\right)-min\left(x_{Mij}\right)\right|\}$$

will give the optimal values for *c1* and *c2* for each station and variable for certain arbitrary weights *w*. The inclusion of the differences between minima and differences between maxima makes sure that the variance shouldn't be as close as possible at the cost of producing extremely unusual high or low values. Weights are introduced in order to lead the algorithm to a more satisfying solution; having approximately equal variance is more important after all. In this case w_{var} and w_{mean} received a weight ten times as high as the other two. Figure 9 shows the same three variables at station Cabauw but now after the transformation.

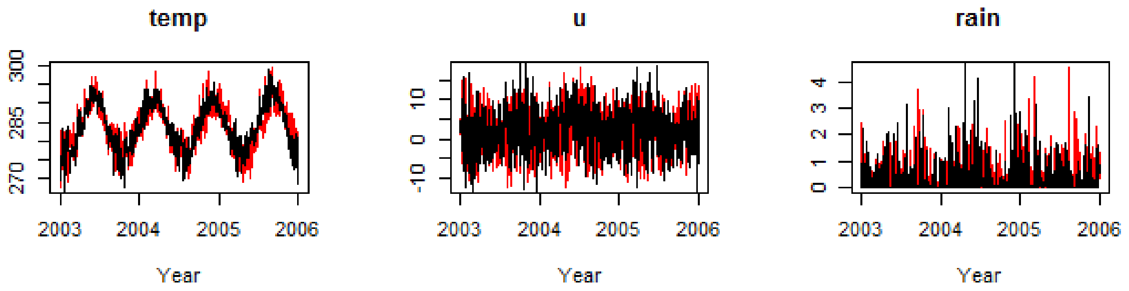


Figure 9. Same as Figure 8, but now for the comparison of the observed (red) and transformed RACMO-ECEARTH (black) temperature, zonal wind speed and rain.

As can be seen, the means and especially the variances are much more equal after the transformation. For a complete comparison of the observed and RACMO-transformed distributions of temperature, zonal and meridional wind speed, see Figure 10, 11 and 12 respectively.

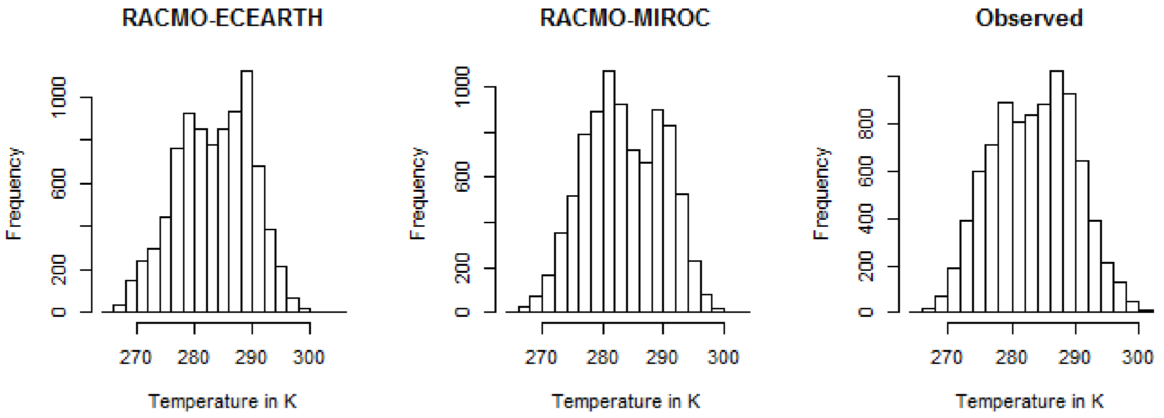


Figure 10. Comparison of the transformed RACMO-ECEARTH, transformed RACMO-MIROC and observed frequency distributions of temperature.

The distributions of temperature seem to be the same. This is not unexpected as the untransformed RACMO temperature data already were close to the observed ones.

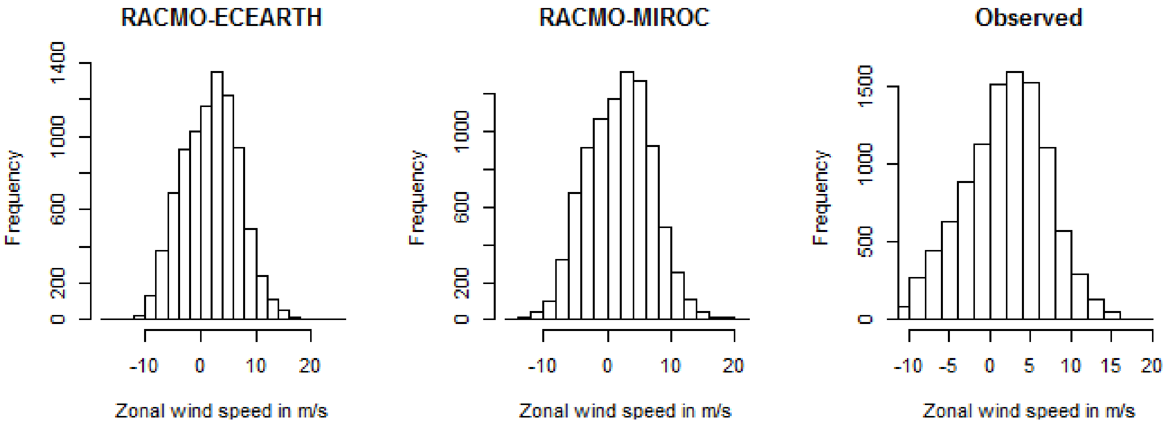


Figure 11. Comparison of the transformed RACMO-ECEARTH, transformed RACMO-MIROC and observed frequency distributions of zonal wind speed.

Figure 11 shows that the distribution of zonal wind speed is quite similar after the transformation. The same can be said about rain (Figure 12).

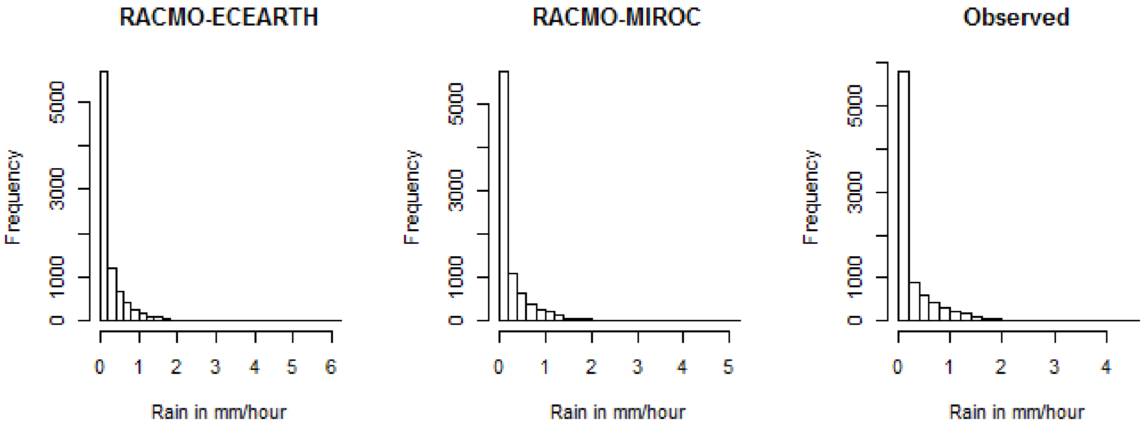


Figure 12. Comparison of the transformed RACMO-ECEARTH, transformed RACMO-MIROC and observed frequency distributions of rain.

Figure 13 shows the distributions of temperature, zonal wind speed and rain of the untransformed RACMO-MIROC output. The necessity of transforming the data becomes clear after seeing the difference in distribution between the observed and RACMO-modeled histograms.

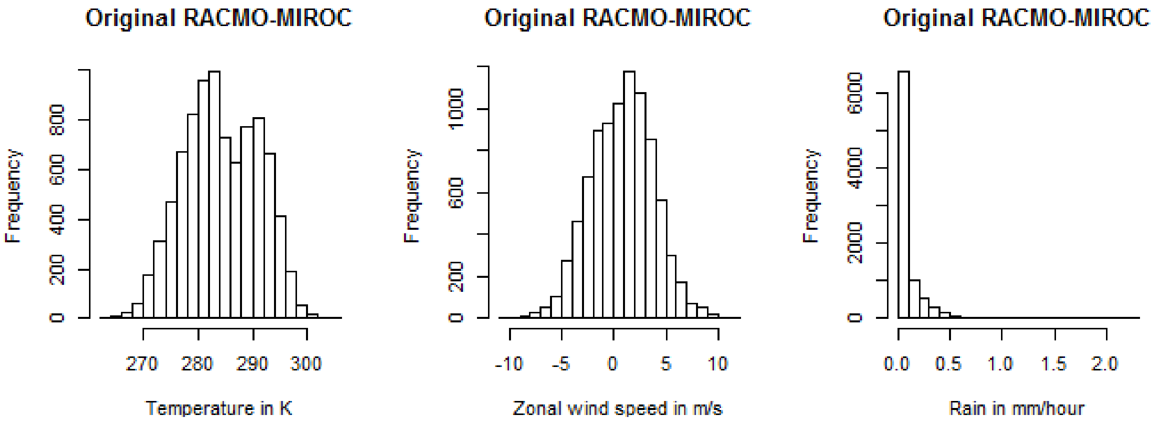


Figure 13. The original RACMO-MIROC distributions of temperature, zonal wind speed and rain.

As can be seen in Figure 13, especially the rain variable is problematic when untransformed. An indication of how the meteorological variables change over time in the RACMO-ECEARTH and RACMO-MIROC simulations is shown in Figure 14.

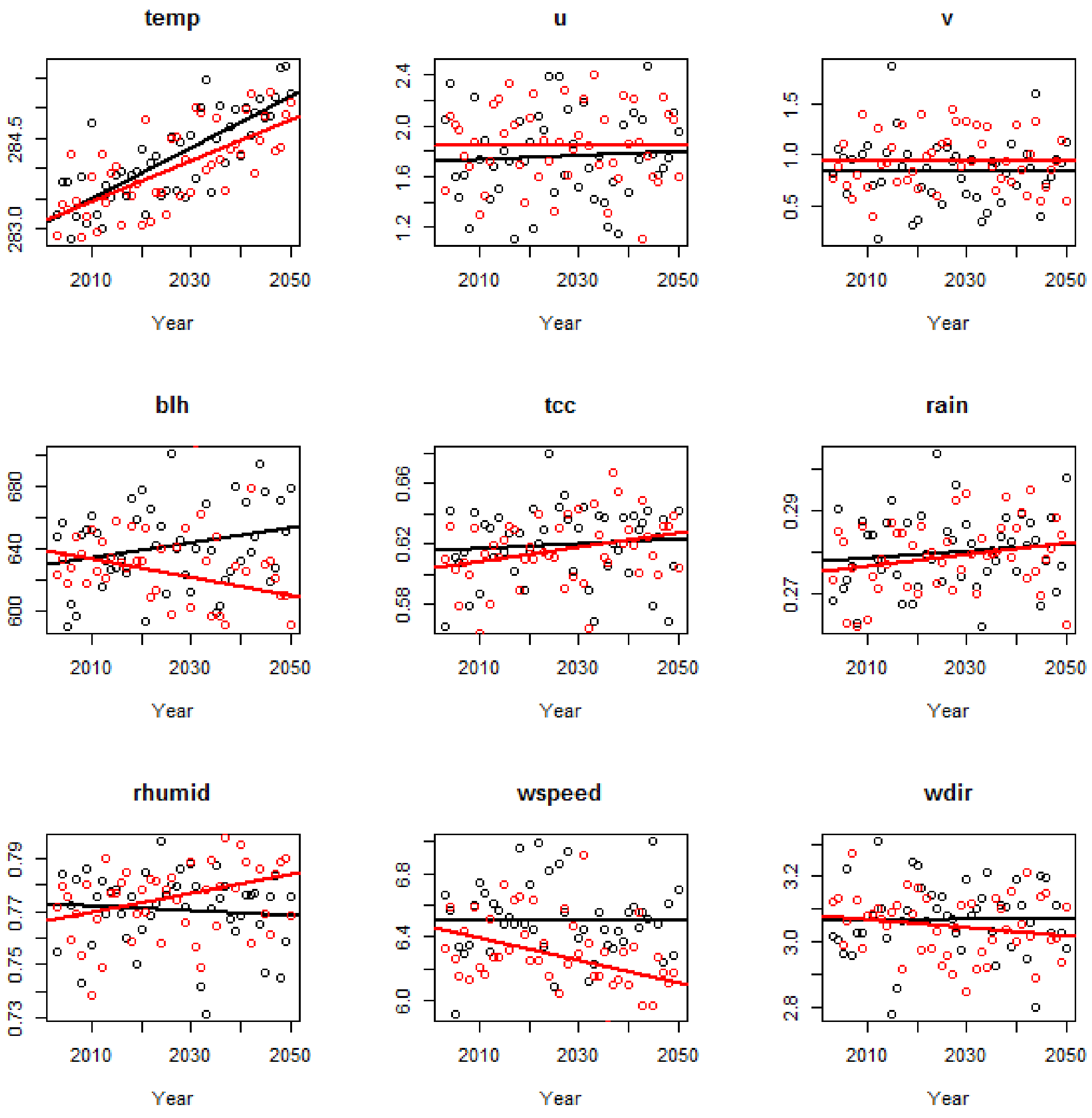


Figure 14. Evolution of the annual means of the selected meteorological variables from 2003-2050 as projected by RACMO-ECEARTH (black) and RACMO-MIROC (red).

Figure 14 shows that especially temperature is likely to change in the coming decades according to RACMO-ECEARTH and RACMO-MIROC. Both version of RACMO also agree on an increase in total cloud cover (tcc) and rain.

4. Methods

As the goal of the current study is to estimate future concentrations of ozone and PM10 based on the relation between these air pollutants with meteorological variables, a statistical model has to be chosen in order to optimally determine this relation. In previous work multiple linear regression models were applied for both ozone and PM10 for each station. As shown earlier in this report, this linear model probably is not preferred because of the possible nonlinear relationship between the meteorological variables and the pollutants. In this study we analyzed a couple of alternatives.

A second possible solution is to apply a multiple linear regression model to transformed data. Because it is not directly clear what common transformation has to be applied to make the data linear, a better way could be to apply a function that maps the 'nonlinear variable' into a variable that has a linear relationship with the dependent variable. Figure 15 shows the application of such a function in the form of a polynomial.

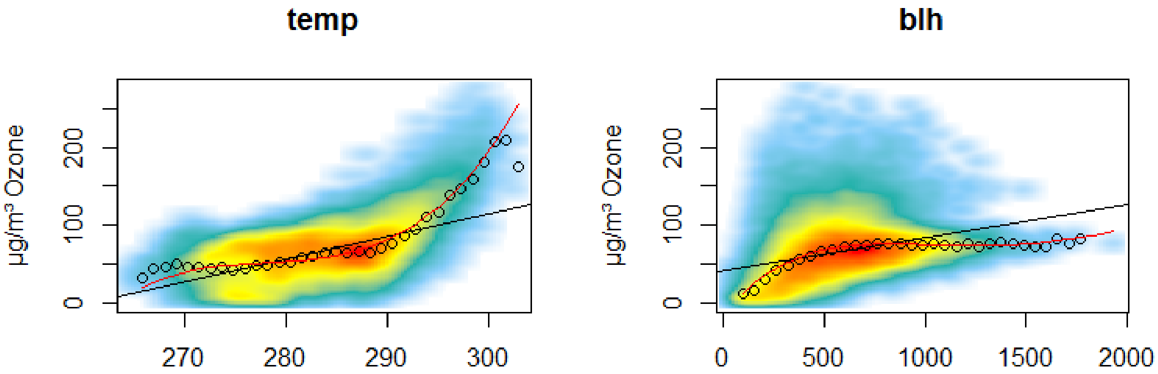


Figure 15. Linear regression (black) and polynomial fit (red) of the 2003-2006 ozone concentrations to temperature (left) and boundary layer height (right).

As shown in Figure 15, the polynomial function (red) follows the mean ozone concentration per bin (circles) as a function of temperature(polynomial degree three) and boundary layer height (polynomial degree 4) more closely than the linear fit (black). The degree of the polynomials are chosen based on cross-validated R^2 to avoid overfitting. Four-fold cross-validation was applied in blocks of a year to account for the correlation of observations in time. Figure 16 demonstrates the point of using a polynomial to linearize the data.

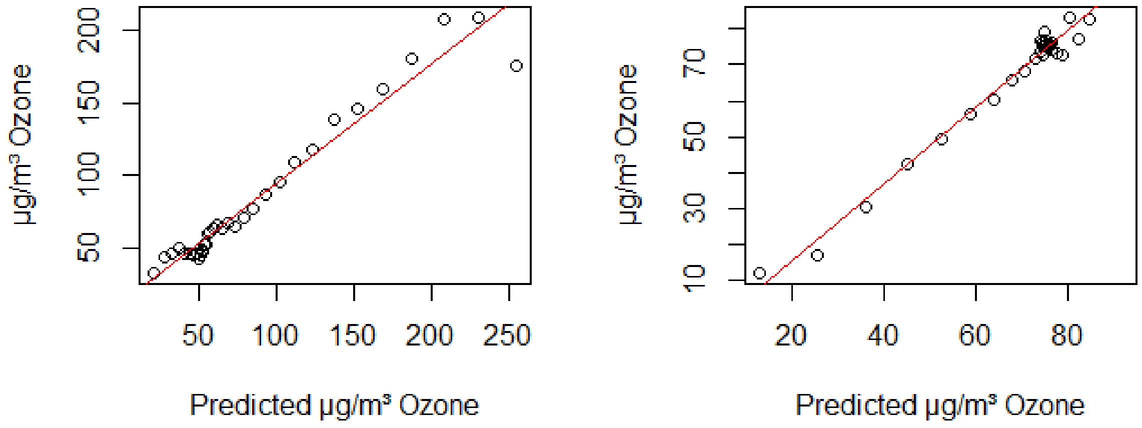


Figure 16. Predicted ozone concentrations based on the polynomial fit to temperature (left) and boundary layer height (right) as a function of the measured ozone concentrations.

A third possibility is to use a model that does not make assumptions about the underlying functional relationship between the dependent and independent variables. Multivariate adaptive regression splines (MARS) is such a model. MARS models that yield accurate predictions can be derived even in situations where the relationship between the predictors and the dependent variables is non-monotone and difficult to approximate with parametric models. It is a nonparametric regression procedure with the main purpose to predict the values of a continuous dependent variable from a set of independent variables.

MARS implements the model in a two stage process: forward and backward. In forward process, it partitions the input space into regions, each with its own regression equation. MARS determines the intervals that show different functional patterns between the variables. In the backward process, some of the intervals are removed from the model as they contribute too little to the model in terms of variance that is explained.

This backward process makes MARS particularly suitable for this study, because it minimizes problems when having a high dimensional parameter space. MARS allows us therefore to also add all first order interaction effects (i.e. adding parameters for the possibility that the some of the independent variables don't have an additive effect on the dependent variable) without worrying of the curse of high dimensionality. For more information on MARS, see Friedman (1991). See Figure 17 to get an idea of what an univariate MARS fit looks like.

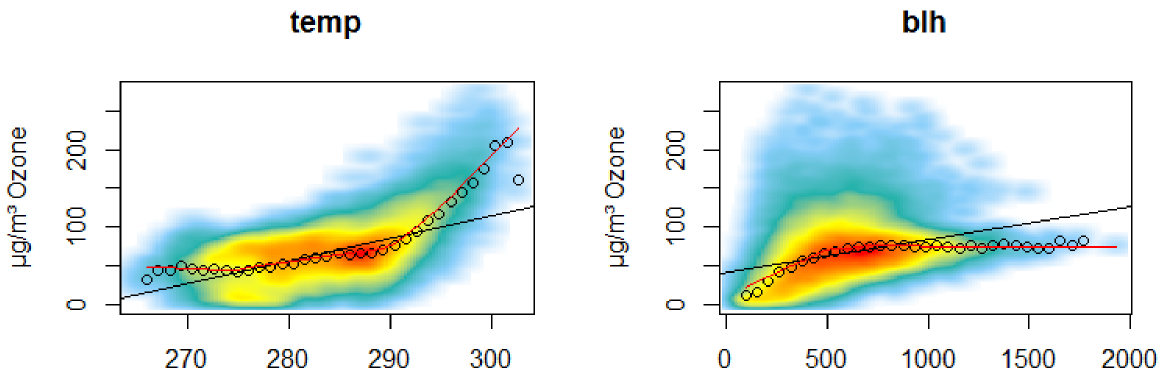


Figure 17. Linear regression (black) and univariate MARS (red) of the 2003-2006 ozone concentrations to temperature (left) and boundary layer height (right).

Although MARS should in theory determine the optimal amount of parameters by pruning (e.g. making the model less complex by removing partitions) the model during the backward step, it turned out in our case that in practice it doesn't always remove the most appropriate partitions. This is because in general terms, when testing a model one wants to test generalization performance. That is, the performance of the model that matters is not the performance on the training set, but on an independent test set. In other words one wants to measure error on independent data and not on the training data itself. Cross-validation is one way to estimate generalization performance, but can be rather slow. Therefore the creators of MARS used an alternative method which is faster and approximates the cross-validation error, namely generalized cross-validation (GCV). The GCV increases the training residual sum of squares (RSS) to take into account the flexibility of the model, and therefore approximates the RSS that would be measured on independent data [Friedman]. It is stated that even when the approximation is not that good, it is usually good enough for comparing models during pruning.

However, using GCV in order to find the optimal model did not work well on our data. For example, some models with all first order interactions performed worse in terms of R^2 when cross-validated, than models without interactions. In theory this could never be the case because overfitting should be prevented by MARS by estimating the GCV. Apparently the GCV approximation of cross-validation doesn't work all that well, which could be explained by our auto correlated data (observations correlated in time). To solve this problem, instead of first adding all the first order interactions and letting MARS prune the model till it decides the model is optimal, the first order interactions are added in a forward stepwise fashion, whereby the inclusion of a term is based on improvement of cross-validated R^2. In this way it is assured that the probability of overfitting is minimized.

The above mentioned three approaches, namely linear regression on untransformed variables, linear regression on transformed variables, and multivariate adaptive regression splines based on the mentioned variables and their first order interactions were compared to each other in terms of variance explained. Four-fold cross-validation was used, with each year in the period 2003-2006 being a block. This approach is preferred over leave-one-out cross-validation when the data are auto correlated. The model with the highest R^2 is the preferred method and will be used to predict ozone and PM10 based on the RACMO scenarios.

It is well known that ozone and PM10 concentrations also depend heavily on the time of the year. It is proposed to split the data in a summer and a winter part and fit models on both parts separately. Another way of including a time of year effect is to add a day-of-year variable to the models, which is simply the day number (1 to 365 for regular years). To find out what is best we will in this study check which way of accounting for the time of year is optimal.

5. Results

First of all the results Pijnappel found in her study were reproduced. The coefficients of her regression models agreed to a large extend with the coefficients we found, confirming the use of the same data.

5.1 Statistical model evaluation and selection

In this section the three approaches discussed in the methods section are tested by using cross-validated R^2 averaged over the stations in order to measure prediction accuracy. Table 5 shows the results of the models.

	R^2 ozone	R^2 PM10
Linear regression untransformed without interactions	0.56	0.19
Linear regression untransformed with interactions	0.72	0.26
Linear regression transformed with interactions	0.73	0.28
MARSwith interactions	0.76	0.32

Table 5. Proportion explained variance in terms of R^2 averaged over the stations, for both ozone and PM10, for the different approaches introduced in Section 4.

As can be seen in Table 5, the MARS model performs best, explaining the most variance when modeling ozone and PM10. It is therefore our model of choice for predicting ozone and PM10 based on RACMO.

Regarding the method of choice with respect to accounting for the time of the year, the results indicated that adding the day-of-year variable was superior to splitting the dataset in a winter and summer part.

It is by the way striking to see that linear regression on untransformed variables with the first order interaction effects performs really well compared to MARS and the regression on transformed data. The difference in R^2 between the linear regression model without and with significant first order interactions is 0.16. While the difference between the latter and the linear regression model based on transformed data seems to be insignificant. Therefore one could conclude that the inclusion of the interaction effects account for the nonlinear relationship between the air pollutants and the meteorological variables.

The R^2 of all the models including the MARS model with respect to PM10 is quite disappointing. Apparently the meteorological variables play only a marginal role in the presence of PM10. The final results as discussed in the next section regarding PM10 should therefore be interpreted cautiously.

5.2 Predicted vs. observed distributions

In section three we showed the distributions of the observed concentrations of ozone and PM10 per station. It is interesting to compare these measured distributions with the predicted distributions based on the MARS model. Figure 18 shows the mentioned distributions regarding ozone for the period 2003 to 2006.

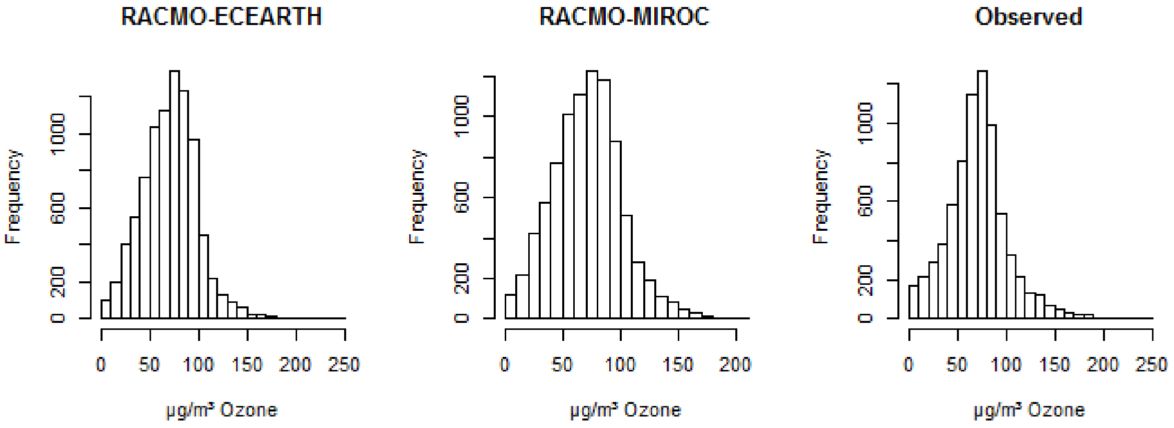


Figure 18. Comparison of the frequency distribution for the ozone concentrations based on RACMO-ECEARTH and RACMO-MIROC with the observed frequency distribution of ozone in the period 2003 to 2006.

It can be seen that while the means of the distributions are similar, the observed ozone distribution has slightly more extreme values. A possible explanation for this could be that the MARS model can only account for 76% of the total variance. This isn't a problem when estimating annual mean ozone concentrations, but it is problematic when trying to estimate exceedences.

Figure 19 shows the same set of distributions, but now regarding PM10.

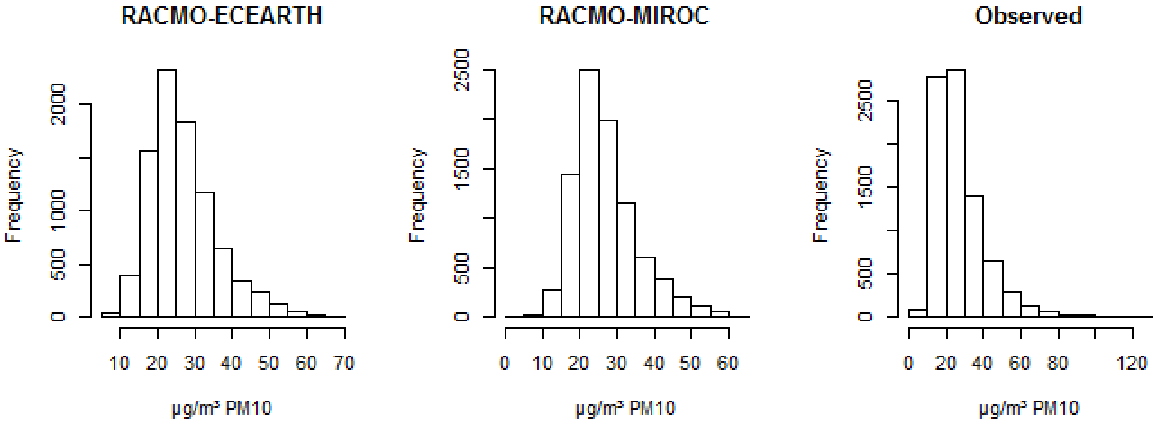


Figure 19. Same as Figure 18 but for PM10.

As could be expected based on the amount of variance explained by the model, the model PM10 concentrations above around 70 µg/m³ are nonexistent.

5.3 Relative importance of the meteorological variables

Now that the MARS model has been applied to the data, it is interesting to see which variables are the most important when predicting ozone and PM10. A variable's importance is a measure of the effect that observed changes to the variable have on the observed response. Estimating predictor importance is in general a tricky problem and there is usually no completely reliable way to estimate the importance of the variables in a standard MARS model. The *evimp* function in the R package *earth* that is used to give this information, makes an educated estimate based upon the number of model subsets that include the variable, the GCV and the RSS (see Friedman 1991 for more information).

Table 6 shows the relative importance of the meteorological variables when modeling ozone concentrations.

Rank	
1	Temperature
2	Day-of-year
3	Boundary layer height
4	Interaction Temperature : Relative humidity
5	Interaction Relative humidity : Wind direction
6	Interaction Temperature : Boundary layer height
7	Wind direction
8	Interaction Temperature : Day-of-year
9	Interaction Wind speed : Wind direction
10	Interaction Rain : Wind speed
11	Relative humidity
12	Interaction Boundary layer height : Wind direction
13	Total cloud cover
14	Interaction Boundary layer height : Day-of-year
15	Interaction Temperature : Wind speed
16	Interaction Boundary layer height : Total cloud cover

Table 6. Relative importance of the terms in the MARS model predicting ozone concentrations following the number of model subsets that include the variable, the GCV and the RSS.

Temperature seems to be the most important predictor, followed by day-of-year, boundary layer height and the interaction between relative humidity and temperature.

Variable importance regarding PM10 differs from ozone, as can be seen in Table 7.

Rank	
1	Boundary layer height
2	Day-of-year
3	Rain
4	Temperature
5	Wind direction
6	Interaction Temperature : Wind direction
7	Interaction Temperature : Boundary layer height
8	Interaction Wind speed : Wind direction
9	Interaction Wind speed : Day-of-year

Table 7. Same as Table 6 but for predicting PM10 concentrations.

Boundary layer height is the most important prediction, followed by day-of-year, rain and temperature. Note that in Table 6 and 7 not all variables and interactions are listed. This is due to the fact that MARS (in combination with the previously mentioned approach with respect to the inclusion of interaction terms) has determined that the optimal model given the data, does not include the variables and interactions that are not in the tables.

5.4 Evolution of ozone and PM10

In this section the effect of climate change on ozone and PM10 will be studied. Our focus is initially on the pollutant concentrations over time in the Netherlands in general. Because different seasons show different effects, results will be stratified by winter and summer as well. Hereafter the predictions will be shown for the stations individually.

The results in this section show changes in concentrations given that all other factors, that are not included in the analysis like emissions, stay constant. In other words, the change in air pollutant concentrations is purely based on change in weather conditions.

5.4.1 Evolution of ozone and PM10 for 2003-2050

Because RACMO data was only available for the six previously mentioned stations, the concentrations of the air pollutants in the Netherlands are averages over these stations. Figure 20 shows the ozone and PM10 concentration in the Netherlands based on RACMO for the period 2003 to 2050. For illustrative purposes, trend lines have been added to the panels.

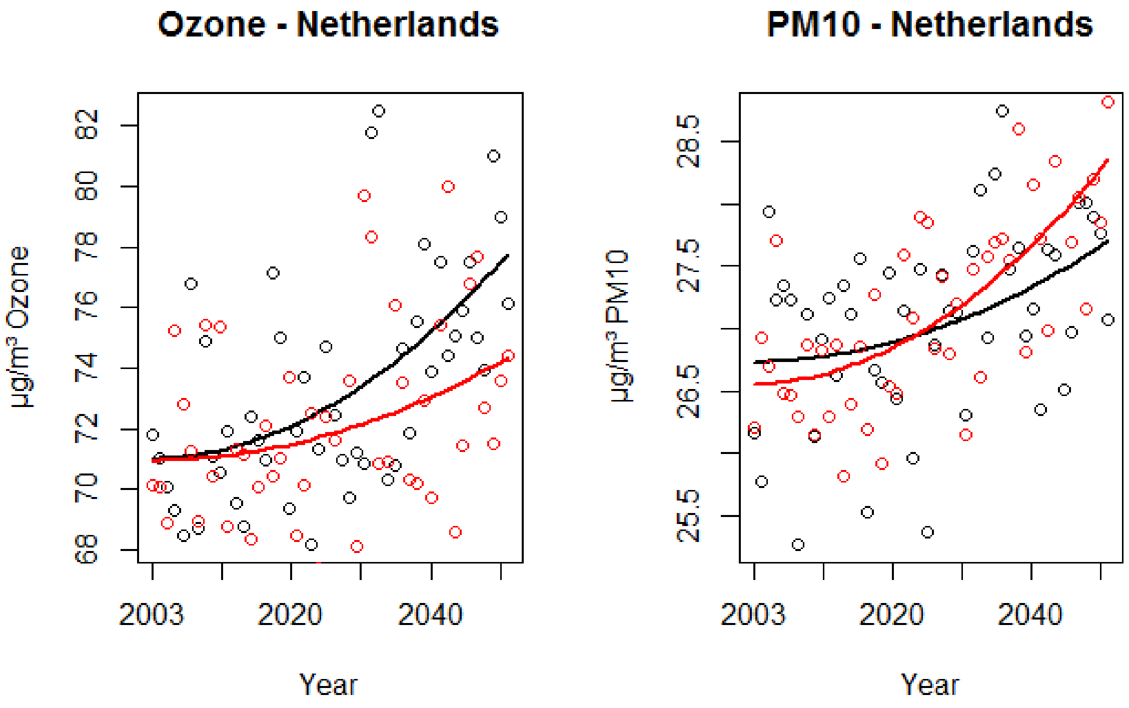


Figure 20. Evolution of the MARS-predicted ozone (left) and PM10 (right) concentrations for the period 2003-2050 based on RACMO-ECEARTH (black) and RACMO-MIROC (red).

The circles represent the yearly average concentration while the lines show the trend over time, with black denoting RACMO-ECEARTH and red RACMO-MIROC. The graphs show an upward trend for both ozone and PM10. Keep in mind while interpreting these graphs regarding PM10 that the model isn't that good in terms of R^2. Table 8 and 9 tell the same story as the above shown graphs, but now the average ozone and PM10 concentrations are quantified per decade.

	2003-2009	2010-2019	2020-2029	2030-2039	2040-2050
RACMO-ECEARTH	70.88	71.87	71.74	74.76	76.30 (+7.6%)
RACMO-MIROC	71.05	71.33	70.73	73.09	73.80 (+3.9%)

Table 8. Evolution of ozone concentration (in μg/m^3) in terms of average concentrations per decade. The figures between brackets shows the change in percentages with respect to the pollutant concentration in the 2003-2009 period.

	2003-2009	2010-2019	2020-2029	2030-2039	2040-2050
RACMO-ECEARTH	26.71	26.83	26.78	27.52	27.36 (+2.4%)
RACMO-MIROC	26.69	26.56	27.05	27.34	28.00 (+4.9%)

Table 9. Same as Table 8 but for the evolution of PM10 concentration (in µg/m^3).

Figure 21 shows the distributions of ozone for the period 2003 to 2006 and the period 2040 to 2050 for both RACMO-ECEARTH and RACMO-MIROC.

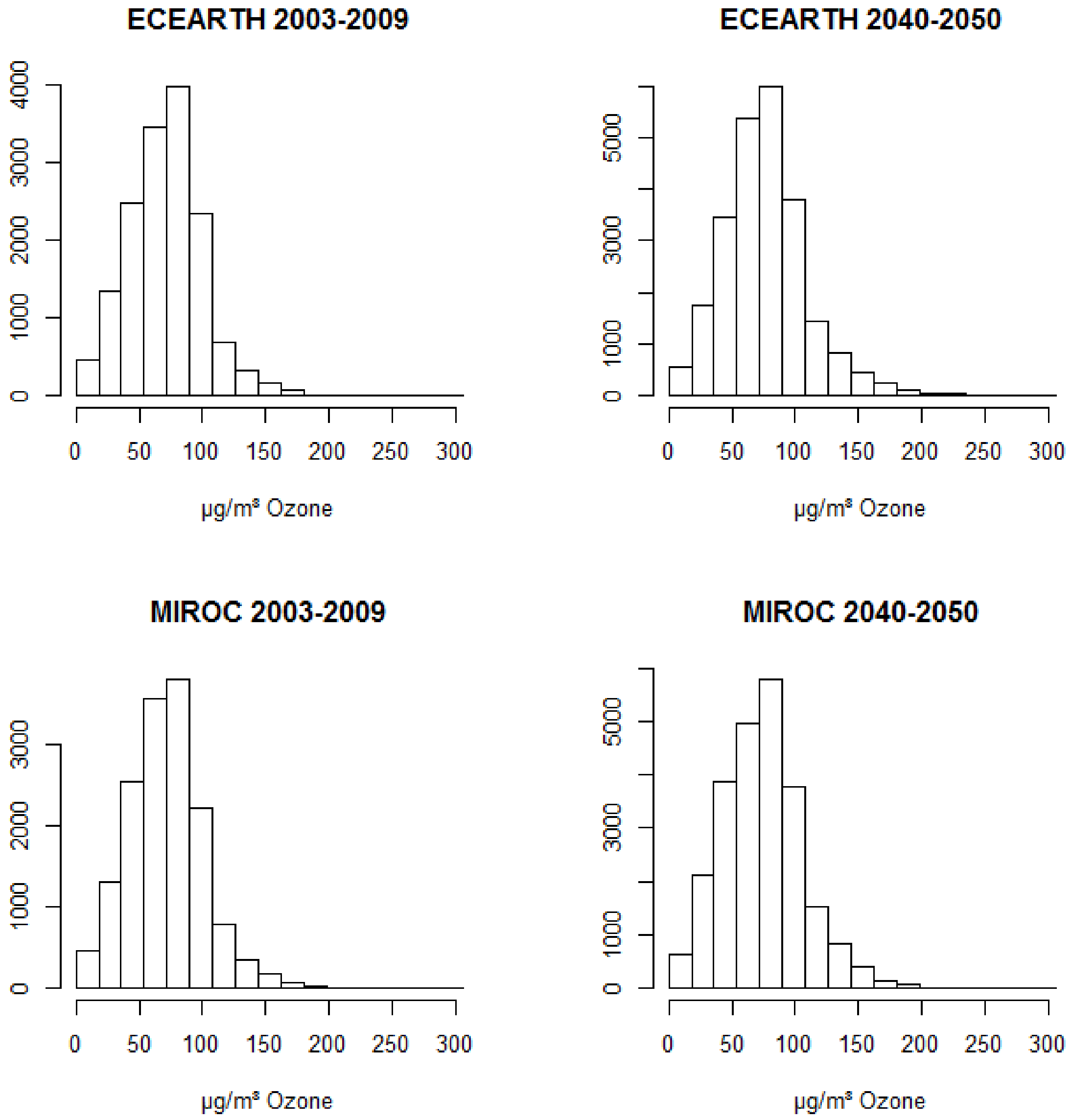


Figure 21. Ozone frequency distributions for RACMO-ECEARTH (top) and RACMO-MIROC (bottom) for the periods 2003 to 2009 (left) and 2040 to 2050 (right).

As can be seen, apart from the shift in ozone concentrations, the shape of the distribution stays the same. Figure 22 shows that this is also the case for PM10.

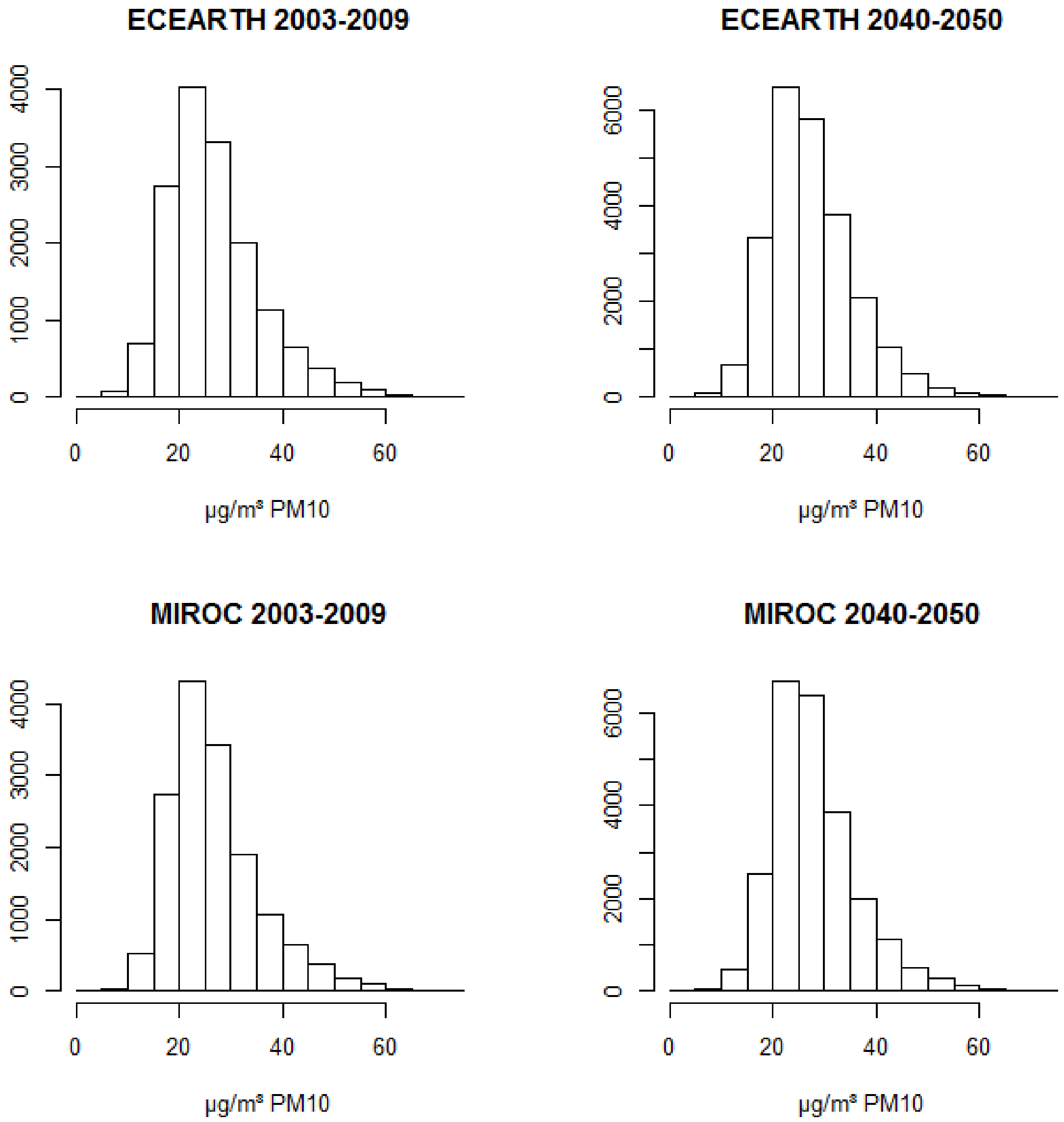


Figure 22. Same as Figure 21 but for the frequency distributions of PM10.

5.4.2 Summer and winter analysis

Because air pollution concentrations are known to show different effects in summer and winter, it is interesting to see how the predicted evolution of ozone and PM10 depend on the season. Figure 23 shows the ozone concentrations per summer and winter over time.

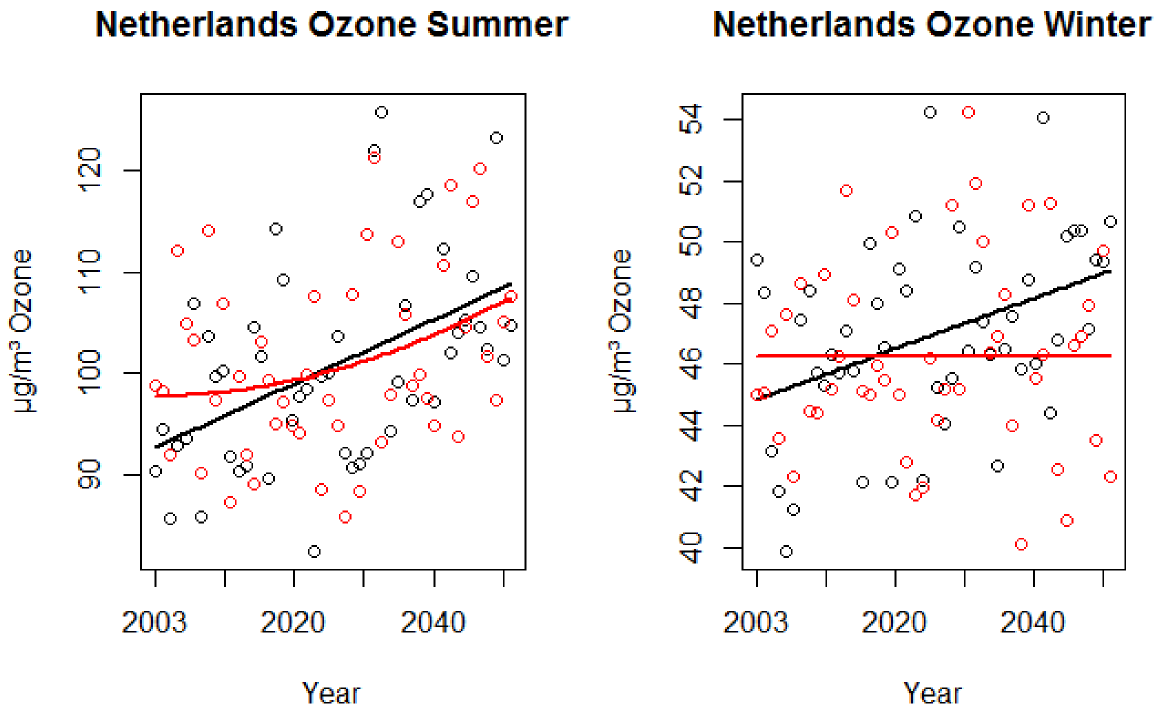


Figure 23. Evolution of the MARS-predicted ozone concentrations in summer (left) and winter (right) for the period 2003-2050 based on RACMO-ECEARTH (black) and RACMO-MIROC (red).

When we look at the graphs of ozone concentrations in summer and winter, we see that the increase depends on the season. The ozone concentrations will increase the most in summer, whereas in winter it is not clear whether there will be any increase at all.

Because temperature is the most important predictor for ozone, it is interesting to see how temperature changes in the summer and winter over the years. Figure 24 shows these time series.

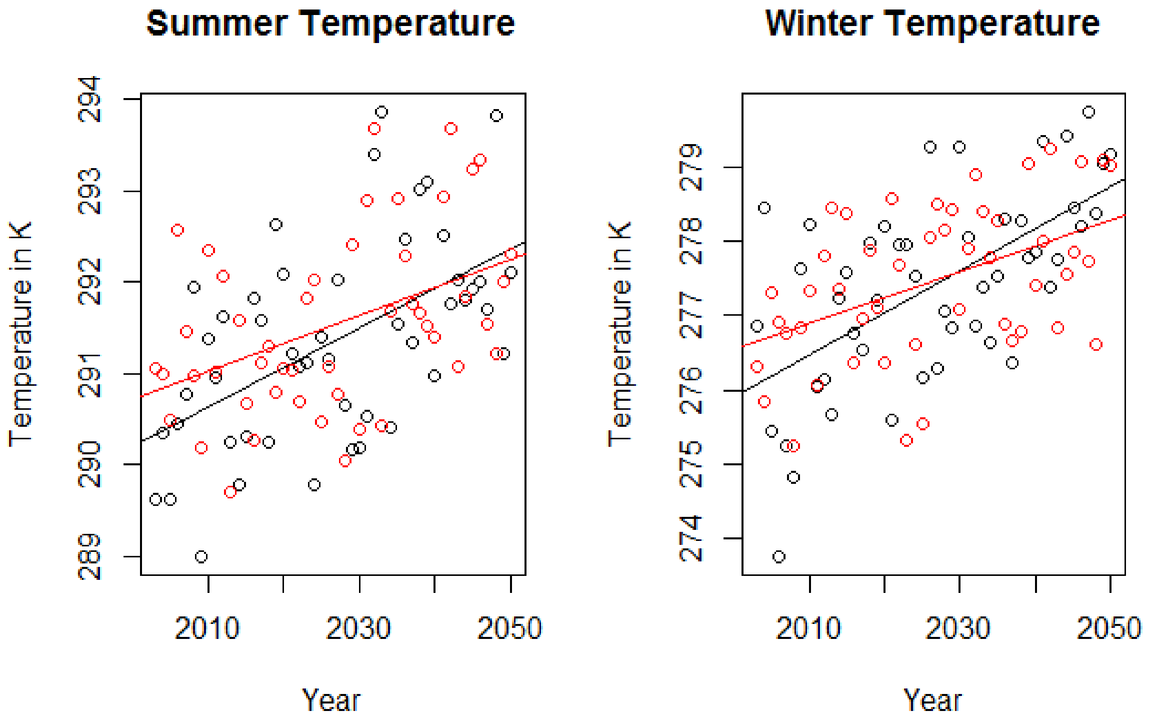


Figure 24. Evolution of temperature in summer (left) and winter (right) for the period 2003-2050 based on RACMO-ECEARTH (black) and RACMO-MIROC (red).

As can be seen, the absolute increase in temperature is independent of winter and summer. However, in section three and four we have shown that ozone doesn't have a linear relationship with temperature. The increase in ozone depends on where the increase in temperature takes place; increases at higher temperatures have far more impact on ozone than increases at lower temperatures.

Figure 25 shows the summer and winter graphs regarding PM10. As we can see, the concentrations of PM10 seem to increase in the summers, whereas they remain the same in the winters.

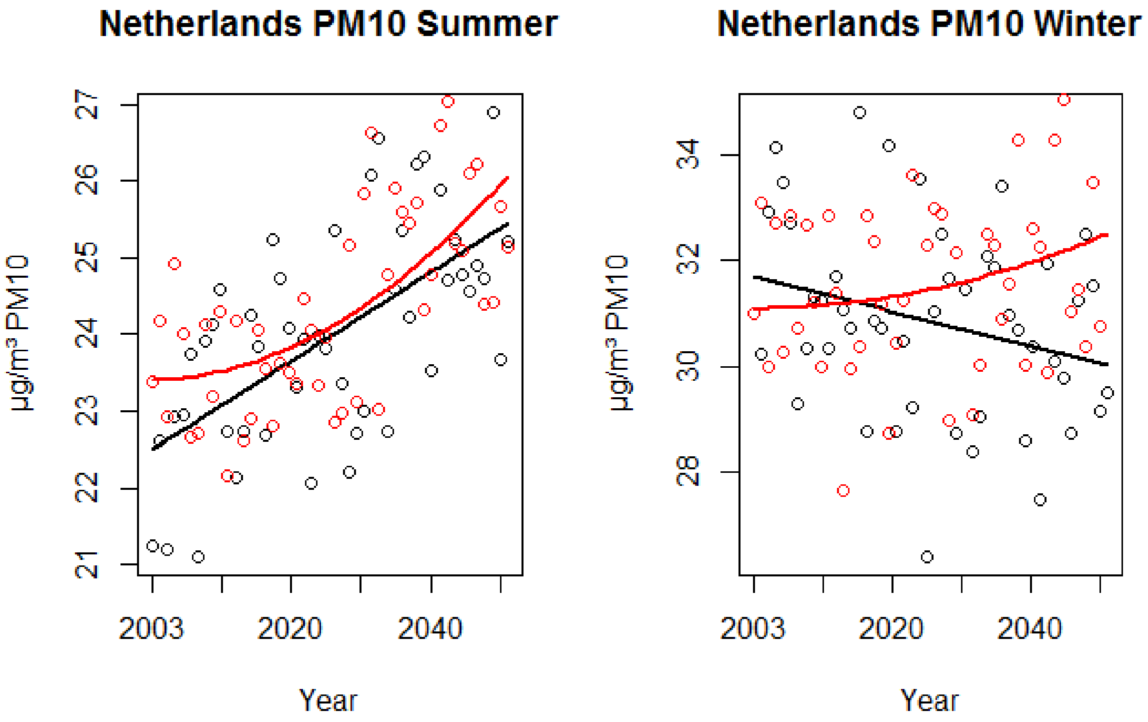


Figure 25. Evolution of the MARS-predicted PM10 concentrations in summer (left) and winter (right) for the period 2003-2050 based on RACMO-ECEARTH (black) and RACMO-MIROC (red).

The most important predictor with respect to PM10 is boundary layer height. However, the evolution of boundary layer height doesn't seem to differ much between winter and summer (Figure 26).

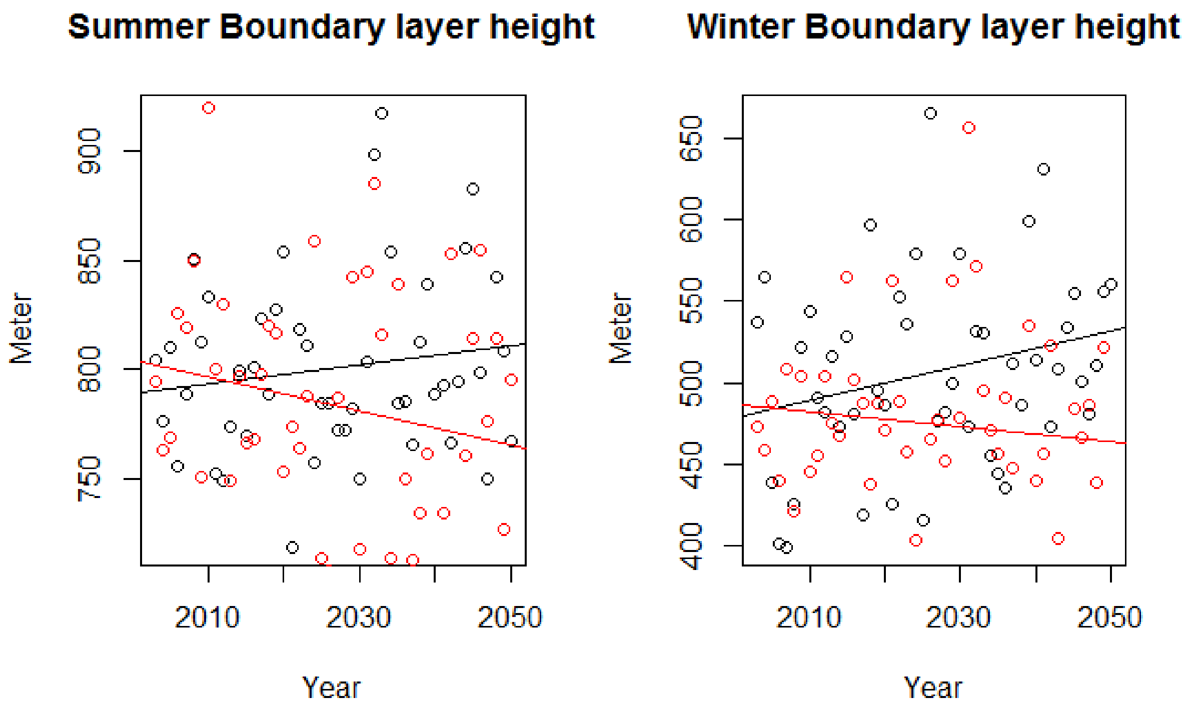


Figure 26. Evolution of boundary layer height in summer (left) and winter (right) for the period 2003-2050 based on RACMO-ECEARTH (black) and RACMO-MIROC (red).

Rain and temperature are the second and third most important meteorological variable, and seem to be responsible for the difference between PM10 evolution in summer and winter (Figure 27 and 24 respectively).

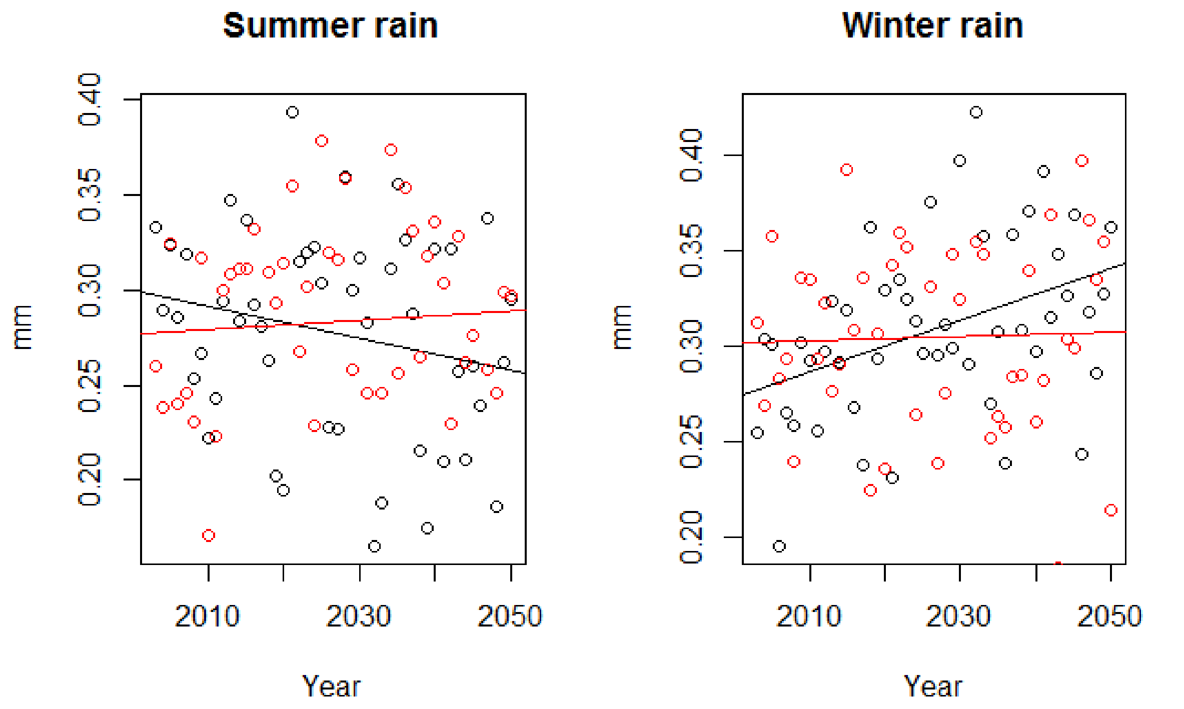


Figure 27. Same as Figure 26 but for rain.

The RACMO scenarios indicate that there will be slightly less rainfall during summers as years go by, whereas in winters it is exactly the opposite. As section three shows that PM10 increases as rain decreases, it could explain the summer increase. The relationship between temperature and PM10 is less obvious. Figure 7 in section three shows a convex relationship. Because temperatures increase both in summer and winter by an approximately equal amount, PM10 decreases in winter and increases in summer due to this predictor.

5.4.3 Evolution at the individual stations

Figure 28 shows the ozone concentrations for the different stations over time. There doesn't seem to be large differences in the amount of increase of the pollutant per station. All stations show an increase in ozone over time.

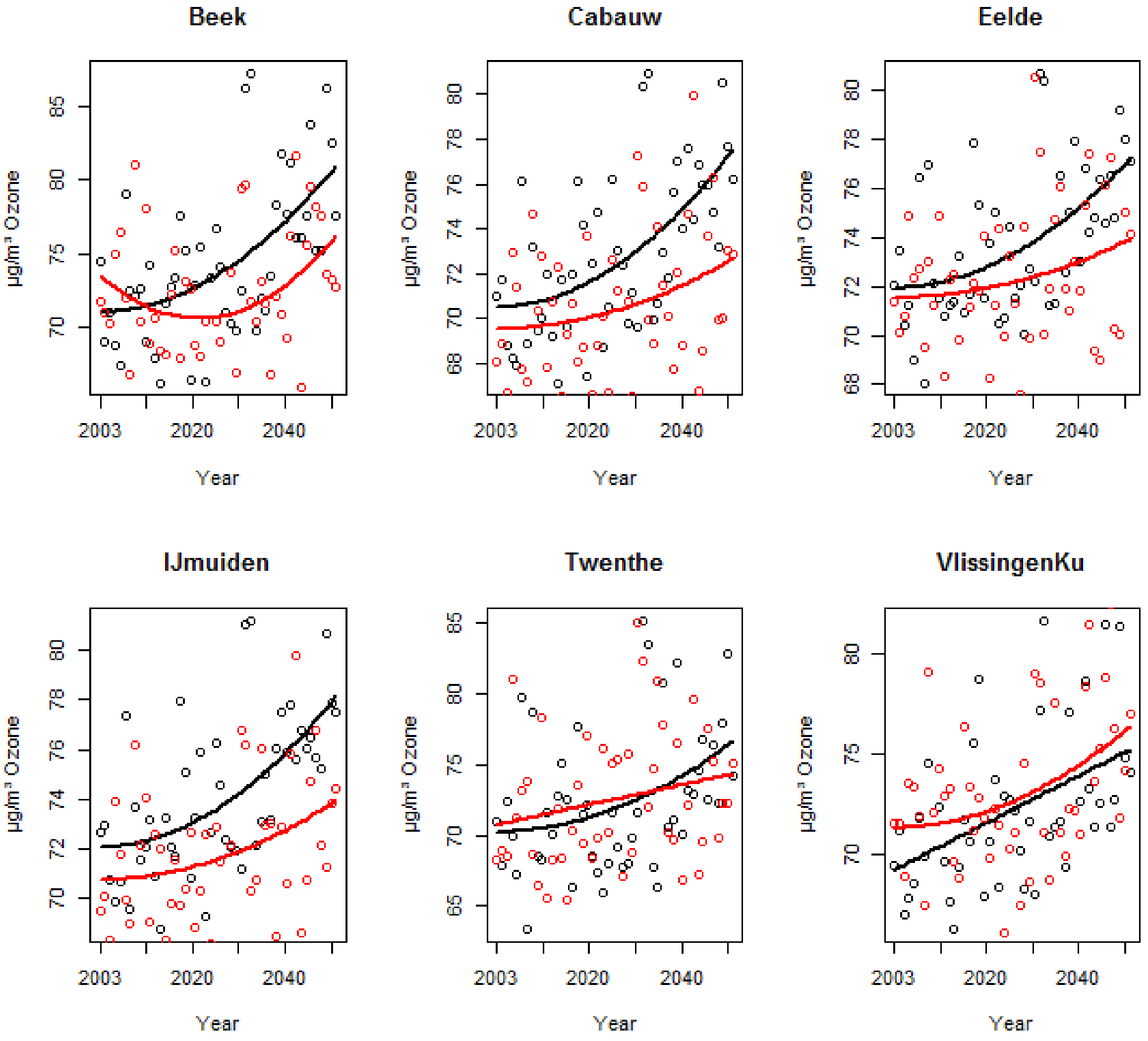


Figure 28. Evolution of the MARS-predicted ozone concentrations per station for the period 2003-2050 based on RACMO-ECEARTH (black) and RACMO-MIROC (red).

The results of PM10 concentration over time per station are not shown because a model that explains such a small amount of variance should not be covered in that much detail.

5.4.4 Ozone and PM10 exceedences

In this report an attempt was made to model ozone and PM10 exceedences. Because MARS isn't created to do classification (i.e. whether a concentration will be above or below a certain threshold), logistic regression has been applied instead. However, although the specificity of the model was very good (classifying actual non-exceedences as non-exceedences), the sensitivity was very poor (classifying actual exceedences as exceedences). The main reason for this is probably the small amount of exceedences (based on the thresholds discussed earlier in this report). More data is needed to properly model this.

Even though MARS isn't the optimal method for predicting exceedences, because it is made to predict means instead, it can still be applied and give a reasonable prediction of the increase in exceedences, especially with respect to ozone. PM10 is harder, because the MARS model fails to accurately predict the distribution of PM10 concentrations (see Figure 19). Figure 29 shows the annual mean number of exceedences in the Netherlands across all six stations with respect to ozone and PM10.

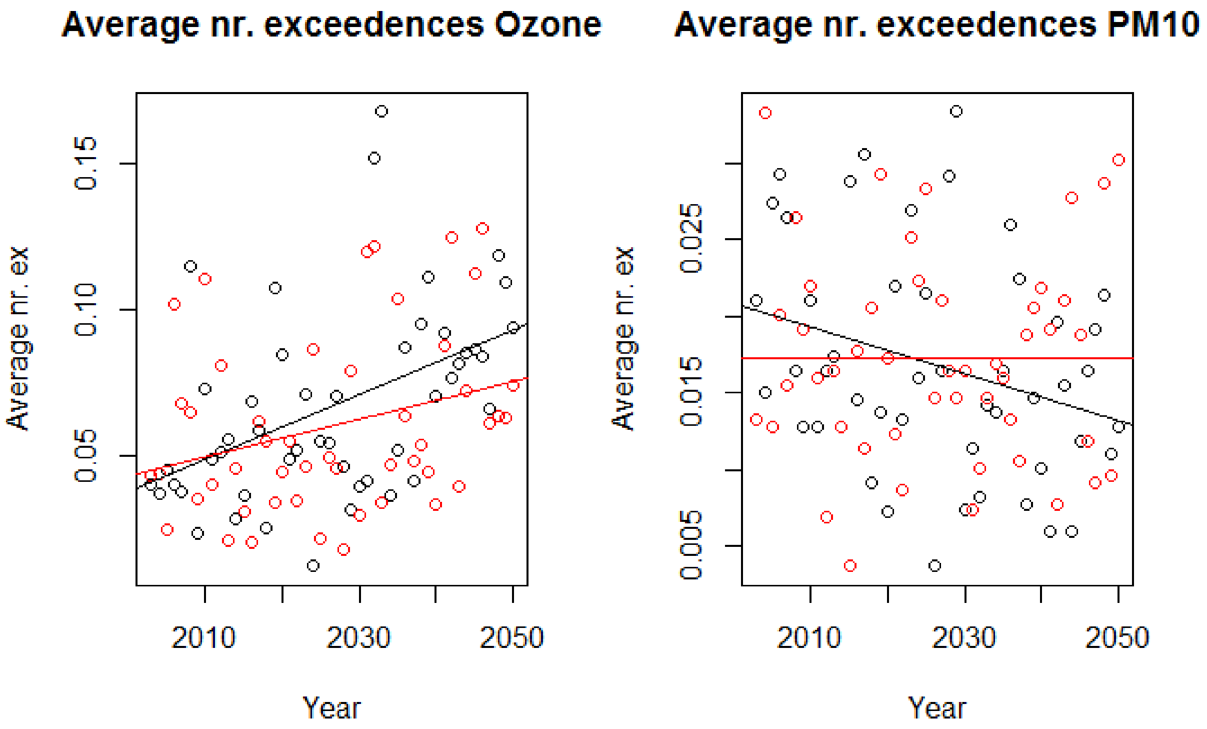


Figure 29. Evolution of the MARS-predicted average number of exceedences for ozone (left, threshold at 120 µg/m^3) and PM10 (right, threshold at 50 µg/m^3) for the period 2003-2050 based on RACMO-ECEARTH (black) and RACMO-MIROC (red).

As expected, the exceedences of PM10 cannot be predicted very well and should not be paid attention to. Exceedences with respect to ozone, on the other hand, can be estimated with more confidence although it is still the case that the absolute average number of exceedences will be underestimated.

6. Discussion and conclusions

In this study a MARS model has been applied to ozone, PM10 and observed meteorological variables in order to project future concentrations of these air pollutants based on regional climate model simulations (RACMO-ECEARTH and RACMO-MIROC). The MARS model outperformed the linear regression model used in the previous study significantly in terms of R^2.

The results of the current study show that both ozone and PM10 concentrations are likely to increase in the coming decades as a result of the changing climate in the Netherlands. It is important to note that this increase in ozone and PM10 is only based on changes in meteorological variables. Due to changes in technology and energy production, emissions and concentrations may go down.

The most important cause of the increase in ozone concentration is the increase in temperature. Because of the nonlinear relationship between temperature and ozone, this results in increasing ozone concentrations during the summers especially. With respect to PM10, the most important meteorological factors contributing to the increase are rain and temperature. Because of the specific relationships of these variables with PM10, the increase of the air pollutant is most probable during summers.

Although the MARS model achieved a reasonable fit regarding ozone, it couldn't model PM10 that well. This implies that there still remains uncertainty about the evolution of the air pollutants over time, especially PM10. The fact that the model with respect to PM10 didn't fit that well should make one reluctant with drawing far reaching conclusions with respect to the impact of changes in meteorology on PM10.

In this report an attempt was made to model ozone and PM10 exceedences. There are however three aspects that can be improved in future analysis. First of all, instead of using daily maximum ozone concentrations, daily 8-hour maximum concentrations should be used in order to be able to draw conclusions with respect to EU norms. Second, MARS is not the most ideal method and should be replaced by a model more suited to this problem. A form of weighted logistic regression could be used in order to balance the sensitivity and specificity. Third, more observational data could be used to better quantify the relationship between meteorological variables and exceedences.

7. Acknowledgments

Erik van Meijgaard (KNMI) kindly provided the RACMO regional climate model time series that have been used in the study.

8. References

EEA[1], 2012: indicators; http://www.eea.europa.eu/data-and-maps/indicators/air-pollution-by-ozone/air-pollution-by-ozone-assessment

EEA[2], 2012: indicators; http://www.eea.europa.eu/data-and-maps/indicators/emissions-of-primary-particles-and-5/assessment-2

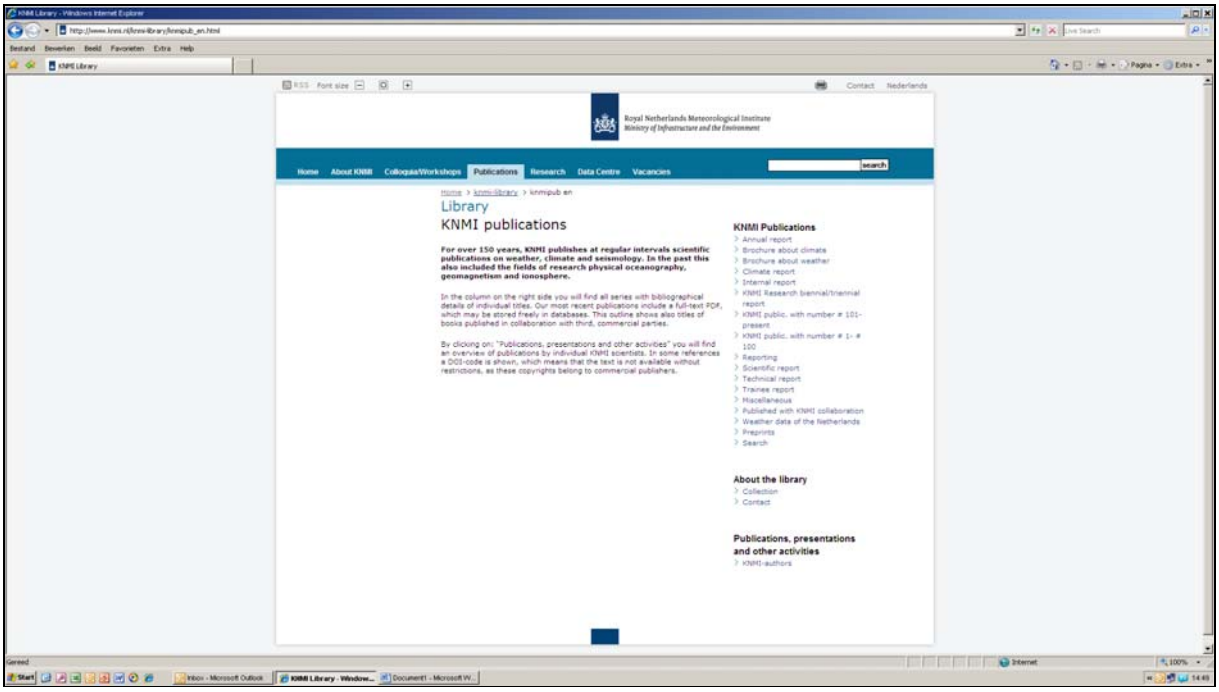
IPCC, 2007: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M.Tignor and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Friedman, J.H. (1991). Multivariate Adaptive Regression Splines (with discussion). Annals of Statistics 19/1, 1U141, 1991. http://www.salfordsystems.com/doc/MARS.pdf.

Pijnappel, A. (2011). Statistical post processing of model output from the air quality model LOTOS-EUROS.

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

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


RSS Font size
Contact Nederlands
Royal Netherlands Meteorological Institute
Ministry of Infrastructure and the Environment
Home About KNMI Colloquia/Workshops Publications Research Data Centre Vacancies
search
Home > knmi-library > knmipub en

Library

KNMI publications

For over 150 years, KNMI publishes at regular intervals scientific publications on weather, climate and seismology. In the past this also included the fields of research physical oceanography, geomagnetism and ionosphere.

In the column on the right side you will find all series with bibliographical details of individual titles. Our most recent publications include a full-text PDF, which may be stored freely in databases. This outline shows also titles of books published in collaboration with third, commercial parties.

By clicking on: "Publications, presentations and other activities" you will find an overview of publications by individual KNMI scientists. In some references a DOI-code is shown, which means that the text is not available without restrictions, as these copyrights belong to commercial publishers.

KNMI Publications
- Annual report
- Brochure about climate
- Brochure about weather
- Climate report
- Internal report
- KNMI Research biennial/triennial report
- KNMI public. with number # 101-present
- KNMI public. with number # 1- # 100
- Reporting
- Scientific report
- Technical report
- Trainee report
- Miscellaneous
- Published with KNMI collaboration
- Weather data of the Netherlands
- Preprints
- Search

About the library
- Collection
- Contact

Publications, presentations and other activities
- KNMI-authors

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