16th International Conference on
Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes
8-11 September 2014, Varna, Bulgaria

EXPERIENCES FOM THE APPLICATION OF A PARAMETER ESTIMATION AND IDENTIFIABILITY ANALYSIS METHODOLOGY TO THE OPERATIONAL STREET POLLUTION MODEL (OSPM)

Thor-Bjørn Ottosen[1,2], Matthias Ketzel[3], Ole Hertel[3],
Henrik Skov[2,3], Jørgen Brandt[3], Ruwim Berkowicz[3] and Konstantinos Kakosimos[1]

[1]Department of Chemical Engineering, Texas A&M University at Qatar, Doha, Qatar
[2]Department of Chemical Engineering, Biotechnology and Environmental Technology, University of Southern Denmark, Odense, Denmark
[3]Department of Environmental Science, Aarhus University, Roskilde, Denmark

Abstract: Uncertainty and sensitivity analysis can potentially increase the transparency in the modelling process and guide research in the relationship between model and data. The uncertainty and sensitivity of the Operational Street Pollution Model (OSPM), being an example of a semi-parameterised air quality model, have not been studied before, and it is therefore the aim to explore the potential advantages of this type of analyses on atmospheric models. An iterative parameter estimation and identifiability analysis methodology along with two different data splitting methodologies were chosen for the present study. The results show that this type of methodology can be informative applied to an atmospheric model, in that the methodology successfully balances the model-measurement errors among the different streets and the different species. Moreover, the results indicate where future research effort in model improvement should be directed, with respect to parameterisations and model parameter uncertainty.

Key words: *Uncertainty Analysis, Sensitivity Analysis, OSPM, Data splitting, Exploratory data analysis*

INTRODUCTION

Over the last decades the use of air quality models for forecasting and scenario studies have become increasingly popular. When using air quality models for planning and decision making, the transparency and the reliability of the results can be enhanced through the analysis of model uncertainty. The use of uncertainty analysis has also been fostered by the rapid development in computational power, due to the large computational requirements of this type of analysis.

The Operational Street Pollution Model (OSPM) has been frequently used over the last two decades, with emphasis in recent years moving towards forecasting and scenario studies (Kakosimos et al. 2010). This development has happened alongside the large increase in measurements resulting from national measurement programmes, which is now available for model validation studies. This means that uncertainty analysis based on non-linear regression of semi-parameterized models has become feasible and could potentially determine hitherto uncertain parameter values or guide research efforts with respect to parameter values and parameterisations.

It is therefore the aim of the present study to examine the model parameter uncertainty of OSPM through the application of a parameter estimation (uncertainty analysis) and identifiability analysis (sensitivity analysis) methodology and, in this way, gain insight into the potential advantages of application of this type of analysis within atmospheric science.

MODEL DESCRIPTION

OSPM is a semi-parameterised model for pollutant concentrations in a street canyon, where the measured concentrations are modelled as a sum of a direct contribution and a recirculating contribution, as illustrated on Figure 1, both calculated through algebraic expressions. The length of the recirculation zone is in the model determined by the wind speed and the upwind building height. This means that, as a rule, the leeward receptor is exposed to emissions from inside the recirculation zone, and the windward receptor is exposed to emissions from outside the recirculation zone (Hertel and Berkowicz 1989b, Berkowicz et al. 1997, Ottosen et al. 2014).

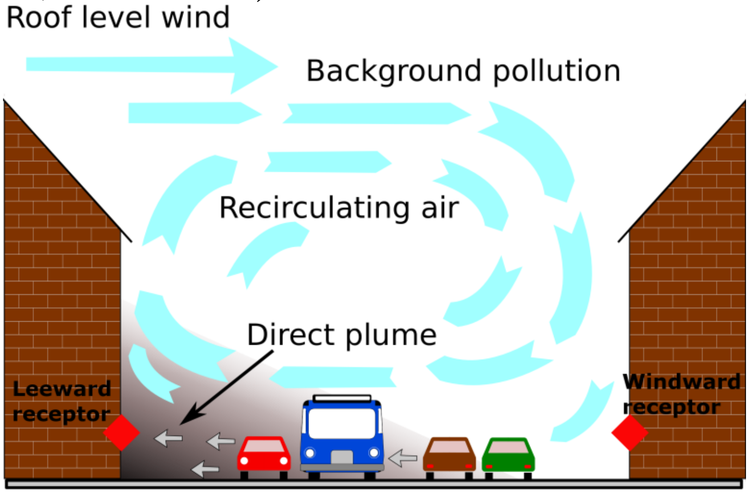

Figure 1. Schematic illustration of the direct and recirculating component of the concentration in OSPM. Figure modified from (Silver, Ketzel, and Brandt 2013).

The specific characteristics of the model are:

- The direct contribution is modelled as a Gaussian plume model, where the emissions are homogeneously distributed in the full length and width of the canyon. The Gaussian plume uses a top hat distribution for the vertical diffusion and assume that horizontal diffusion can be neglected.
- The recirculating contribution is modelled as a trapezium shaped box model with the fundamental assumption that the inflow of pollutants equals the outflow of pollutants. This is justified based on the temporal resolution of the model of one hour.
- Moreover, the model contains algebraic expressions for traffic produced turbulence and a numerical averaging procedure to account for wind direction meandering especially pronounced for low wind speeds (Hertel and Berkowicz 1989c).
- The emissions are modelled using the COPERT IV emission model (EEA 2009). The parameters of the emission model is not included in the subsequent parameter estimation to limit the scope of the study.
- The model contains an algebraic expression for the conversion of NO to NO_2 in the presence of Ozone (Hertel and Berkowicz 1989a). To limit the scope of the present study, the parameters of this conversion scheme has been left out of the subsequent analyses.

METHODOLOGY

The iterative parameter estimation and identifiability analysis as presented by Brun, Reichert, and Künsch (2001) has been adopted for the present study due to its widespread use in other scientific disciplines (cited 157 times in Web of Science as per July, 2014). The methodology consists of running parameter estimation and identifiability analysis in an iterative cycle until convergence between the estimated parameters and the identifiability of the parameters is achieved.

The parameter estimation is performed using standard weighted non-linear regression techniques (Seber and Wild 1989).

The identifiability analysis is based on the calculation of two measures: The sensitivity measure and the collinearity index. The sensitivity measure is calculated as the root mean square of the non-dimensional sensitivity, which again is the non-dimensional local sensitivity of the model output with respect to a change in a model parameter, all other model parameters kept constant. As such, the sensitivity measure represents an average of how much the output of the model changes with the individual parameter in the neighbourhood of the original model parameter. The collinearity index is a measure of to what extend it is possible to cancel a change in one model parameter by linearly adjusting the other model parameters. If the sensitivity of the model to changes in two or more model parameters are linearly independent, the collinearity index will be equal to unity, otherwise it will go towards infinity. The sensitivity measure and the collinearity index are used together to find a parameter combination that is only slightly influenced by collinearity and with a reasonable sensitivity. Such a parameter combination will be less prone to overfitting, and thus yield less uncertain parameter estimates. The above methodology is based on local sensitivity analysis, and if the estimated parameters are far from the original parameters the analysis have to be repeated again.

Before the above analysis was initiated the data were split into an estimation and a prediction set, to assess both the replicative and the predictive validity of the model, using two different approaches that have been in use in the literature: The DUPLEX data splitting procedure (Snee 1977) splits the data into an estimation- and a prediction data set based on the Euclidean distance to the data point from the data points already split, thus creating two non-identical dataset with similar statistical properties. The seasonal data splitting procedure (used among others by Silver, Ketzel, and Brandt (2013)), splits the data into an estimation- and prediction data set such that the first six months of the year are used for estimation, and the last six months of the year are used for prediction. The performance of the two data splitting approaches were subsequently compared.

RESULTS AND DISCUSSION

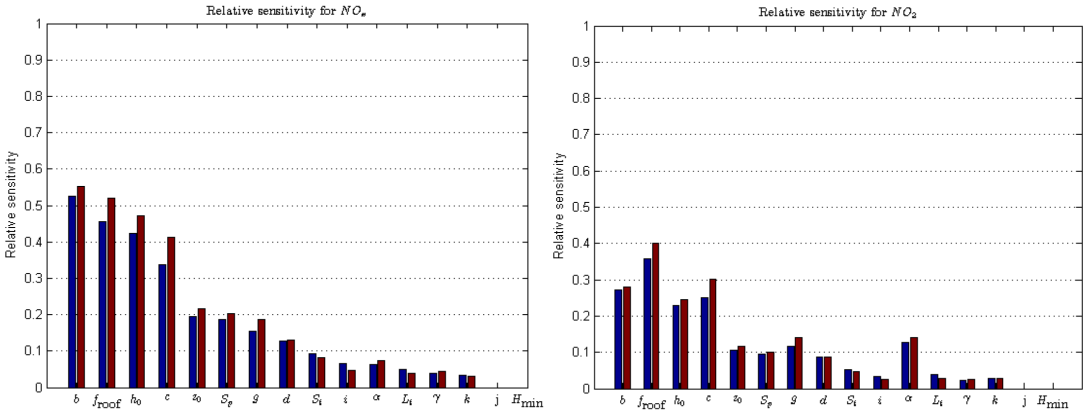

Figure 2. Root mean square local sensitivity for the DUPLEX estimation set (blue) and the Seasonal estimation set (red) for respectively NO_x (left) and NO_2 (right).

The results of the local sensitivity analysis are shown in Figure 2. It can be seen that the sensitivity of the model to changes in the individual parameter decreases approximately linearly for both species and that there are only marginal differences in the sensitivity among the two data split. The reason behind this is the large dataset used for the present study (five street canyons and years 1994-2010), which means that the parameters will be sensitive proportionally to their influence in the dataset.

The sensitivity of the model to changes in the individual parameters is lower for NO_2 than for NO_x, which is caused by the fact that the NO_2 concentrations are calculated as a function of the NO_x concentrations. The different order of parameter sensitivity of NO_2 compared to NO_x is caused by the model structure.

What can be seen from Figure 2 is that the most influential model parameters are the ones controlling the street level wind speed (b, f_{roof}, and h_0), and the parameter controlling the length of the recirculation zone (c). This is not surprising given that the wind speed is a very important factor in all Gaussian plume models, and since the length of the recirculation zone controls the emission that the leeward receptor is exposed to, the sensitivity of the model to this parameter is neither surprising.

As can also be seen from Figure 2 there is a number of parameters with very low sensitivity. This means that these parameters can be set at any value (locally), without changing the model output significantly. In a modelling context this should be avoided, since these parameters have large uncertainties, and thus a model improvement approach could be to try to remove some of these parameters or replace them with parameters of high sensitivity through a change in parameterisations. Moreover, the above analysis serves to guide as to which parameters future research attention should be given.

The collinearity analysis shows that 12 model parameters can be estimated out of a total of 16 model parameters.

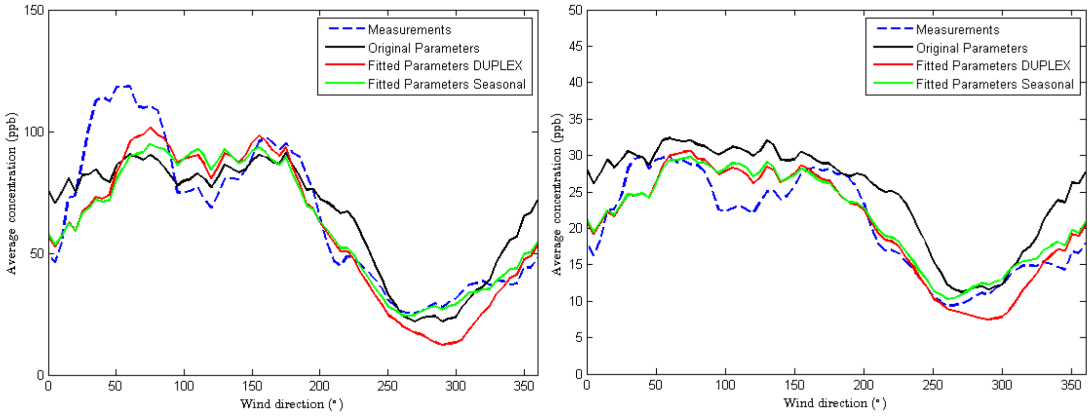

Figure 3. Average concentration as a function of wind direction for all years for Vesterbro street in Aalborg, Denmark. Results are for respectively NO_x (left) and NO_2 (right).

The parameter estimation results in two sets of parameters (one for each data split), that are significantly different (results not shown). Exploring the performance of the two parameter sets in terms of correlation coefficient (R^2), fractional bias, and normalized mean square error however, show very little difference. An example of this is shown on the wind direction plot of Figure 3. It is evident from the figure, that the two fitted parameters have a better mean performance than the original model parameters, however it seems like the Seasonal fit have slightly better performance than the DUPLEX fit. An inspection of the histograms of the different dimensions of the two data split show very little difference. The difference in fitted model parameters can thus be explained by that whereas the Seasonal data split is more representative for the whole data set, since the representation of situations in the data set are proportional to the representations in the full data set, the DUPLEX data split has a better coverage of the different situations the model can be exposed to. This could be interpreted as that the Seasonal data split is more relevant in a forecasting setting, whereas the DUPLEX data split is more relevant in a model design setting.

It can as well be seen from Figure 3 that the parameter estimation procedure have balanced the errors such that the performance is approximately similar for the concentrations of NO_x and NO_2. This is as well the case for the other streets where the parameter estimation procedure have balanced the errors among the included species and among the streets. This is desirable for a model like OSPM, which is designed to have an equal performance for all streets, meaning that the parameters should be interpreted as constants and not be recalibrated for each street.

The phenomenon that two (or more) models (or model parameter sets, model parameterisations, etc.) have approximately equal performance is known as equifinality (Beven 2006). Equifinality in the model parameters arise in the interplay between the model structure and the data used for parameter estimation. The underlying assumption in all parameter estimation procedures is that the model structure is correct. Since this will seldom be the case in environmental science, the estimated parameters will compensate for deficiencies in model structure, and parameters will change with the data used for fitting. This serves as a kind of identifiability problem not accounted for in the identifiability analysis, since the identifiability analysis applied here inherently is of local nature. One could argue that the number of identifiable model parameters should be reduced until the model parameter uncertainty would be within acceptable limits, however, this poses several problems: First there is no procedure to decide which model parameters to determine using parameter estimation and which to determine by other means. Second, reducing the number of estimated model parameters would make the estimated parameters dependent on other (poorly defined) parameters, and the parameter estimate might thus not be more reliable. Third, leaving parameters with high uncertainty out of the analysis will make the analysis less informative. Statistical parameter estimation should thus not be rejected in atmospheric modelling due to model parameter equifinality, since other methods of parameter estimation might suffer from the same deficiencies.

From a modelling point of view model parameter equifinality should be minimised, since this constitutes uncertainty in the model. This can either be done by making the model more suitable to the existing data, through e.g. removal of non-sensitive parameters, creation of more accurate parameterisations etc. or through acquiring better data for parameter estimation.

The identifiability analysis on the estimated parameters for the DUPLEX and the seasonal data split show that both sets of parameters are identifiable.

CONCLUSION

A number of experiences have been gained from the application of parameter estimation and identifiability analysis to the Operational Street Pollution Model:

- The sensitivity measure presented in Figure 2 serves to guide the future research efforts into which aspects of the model to improve. Moreover, the results indicate that, if possible, some parameters should be eliminated from the model.
- The results of the parameter estimation show that it is possible to apply this methodology to a semi-parameterised air quality model, and that the procedure will balance the model performance among individual streets and individual species.
- Through data splitting it was shown that the combination of model and data result in model parameter equifinality. Future research efforts should serve to quantify this phenomenon and through improved modelling efforts reduce this phenomenon.

Summing up, the parameter estimation and identifiability analysis methodology applied in the present study have not provided accurate estimates of the model parameters in OSPM, however, the methodology has increased the transparency of the relationship between the model and the results and provided guidance of further research in improving the model.

REFERENCES

Berkowicz, R., O. Hertel, S.E. Larsen, N. N. Sørensen, and M. Nielsen. 1997. Modelling traffic pollution in streets. Ministry of Environment and Energy, National Environmental Research Institute.

Beven, Keith. 2006. "A manifesto for the equifinality thesis." *Journal of Hydrology* no. 320 (1-2):18-36.

Brun, Roland, Peter Reichert, and Hans R. Künsch. 2001. "Practical identifiability analysis of large environmental simulation models." *Water Resources Research* no. 37 (4):1015-1030.

EEA. 2009. EMEP/CORINAIR Atmospheric Emissions Inventory Guidebook, Chapter on Exhaust Emissions from Road Transport. Methodology for COPERT 4.: European Environmental Agency.

Hertel, O., and R. Berkowicz. 1989a. Nodelling NO_2 Concentrations in a Street Canyon. In *DMU LUFT*.

Hertel, Ole, and Ruvim Berkowicz. 1989b. Modelling Pollution from Traffic in a Street Canyon. Evaluation of Data and Model Development.: National Environmental Research Institute.

Hertel, Ole, and Ruwim Berkowicz. 1989c. Operational Street Pollution Model (OSPM) Evaluation of the Model on Data from St1 Olavs Street in Oslo. National Environmental Research Institute.

Kakosimos, Konstantinos E., Ole Hertel, Matthias Ketzel, and Ruwim Berkowicz. 2010. "Operational Street Pollution Model OSPM) - a review of performed validation studies, and future prospects." *Environmental Chemistry* no. 7 (6):485-503.

Ottosen, Thor-Bjørn, Konstantinos E. Kakosimos, Christer Johansson, Ole Hertel, Jørgen Brandt, Henrik Skov, Ruwim Berkowicz, and Matthias Ketzel. 2014. "Implementation and Validation of an Inhomogeneous Emission Geometry Scheme in a Semi-Parametrized Street Canyon Model." *In preparation*.

Seber, G. A. F., and C. J. Wild. 1989. *Nonlinear Regression*: John Wiley & Sons.

Silver, J. D., M. Ketzel, and J. Brandt. 2013. "Dynamic parameter estimation for a street canyon air quality model." *Environmental Modelling & Software* no. 47:235-252.

Snee, Ronald D. 1977. "Validation of Regression Models: Methods and Examples." *Technometrics* no. 19 (4):415-428.