acp-2018-690: Advanced methods for uncertainty assessment and global sensitivity analysis of a Eulerian atmospheric chemistry transport model

by Aleksankina et al.

Response to reviewer #1

General Comments

The authors have made a commendable effort to apply uncertainty and sensitivity analysis methods which have a long theoretical history in the stats literature but have only in recent years begun to be applied to complex models such as this.

Response: We thank the reviewer for this supportive comment.

1) *Given that they only look at sensitivity to emissions, and conclude that the model is not particularly sensitive to those inputs, it is perhaps a shame that they didn't attempt to include more input variables into the analysis, as there have been a number of published studies which demonstrate that these methods can be used with significantly larger numbers of inputs.*

Response: There is still a practical trade-off between the desirability of investigating a large number of inputs and the computational costs associated with running a complex atmospheric chemistry transport model (ACTM) for that number of inputs. Hence in this study we concentrated on the input emissions, because input emissions have been reported to strongly affect uncertainty in the modelled surface concentrations of various air pollutants and because simulation of the effects of changes in emissions is a principal application for an ACTM (Introduction, p2 L17). It is only by deploying formal sensitivity methods as we have done here that robust statements can ultimately be made concerning the sensitivities of a given ACTM – in our illustrative case here, the sensitivities of the EMEP4UK model to input emissions.

2) *There are a few concerns regarding the implementation of the methods and the effect that this may have on the validity of the results. In particular please see the points below concerning sample size and emulator validation, which should be addressed before the paper is recommended for publication.*

Response: We address these points where they are raised in more detail below.

Specific Comments

3) *P5,L2 How many emissions inputs does the model have, and why were the ones used chosen?*

Response: The model uses both anthropogenic and biogenic emissions as inputs – a full description of the model inputs can be found in Simpson et al. (2012). In this study, emissions of all of the major primary anthropogenic pollutant compounds were investigated,

as described in the paper. The biogenic emission sources were not perturbed, as these are computed within the model itself and are not therefore classed as input datasets. This decision was made based on the fact that one of the main applications of the EMEP model is in providing scientific support for policy-making regarding impacts of interventions leading to anthropogenic emissions reductions (see Introduction). In the short to medium term at least, the potential future changes in emissions driven by environmental and climate change policies are not likely to affect biogenic emissions as much as anthropogenic emissions, hence it was decided to investigate model response to the changes in anthropogenic emissions.

4) *P5,L5 Why were the shipping emissions not split by pollutant type when the other emissions were? Does this not make the results harder to interpret?*

Response: This goes back to the trade-off between how many input variables should be used in the analysis to obtain the most benefit from it and the computational cost of the analysis. The decision was made to concentrate on the land-based input emissions. The interpretation of sensitivities to shipping emissions was made harder by our approach only in places particularly impacted by shipping. The majority of this impact takes place in the sea-based grid cells where the assessment of the impact of air pollutants on human health and ecosystems is not relevant.

5) *P6,L19 Could the authors include their reasoning for choosing only 84 design points. The normal recommended minimum number for constructing Gaussian process emulators is 10 per input variable, which in this case would be 130. See, for example Loeppky et al, 2009, Choosing the Sample Size of a Computer Experiment: A Practical Guide, Technometrics.*

Response: We were aware of this paper cited by the reviewer. The paper states that an empirical $n = 10d$ rule (where n is the sample size and d is the number of variables under investigation) provides reasonable accuracy when approximating model response with a Gaussian process emulator. This rule is applicable to the model response of any complexity. The paper indicates this is an empirical rule and does not state that $10d$ is the minimum sample size for a Gaussian process emulator to perform well. In our particular study, the input-output response function is expected to be smooth, hence there is no need for a very large number of sample points as it will not reduce error of the emulator.

Again, it is a balance on return for computational resource. Ideally, in the situation where model runs are computationally expensive a sequential sampling technique can be applied to track the improvement of emulator performance with increase in the sample size.

6) *P6,L30 The authors state that the choice of a linear mean function incorporates "prior beliefs" – could they explain what prior beliefs motivated this choice.*

Response: We are referring here to the fact that whilst one can choose the form of $\boldsymbol{h(.)}$ to use, the choice incorporates assumptions, i.e. prior beliefs, as to the most appropriate form to use. In this study the choice was a linear mean function. We assumed that the response of the

surface concentration of air pollutants to the changes in the input emissions is likely to be smooth (i.e. with no discontinuities and no sharp fluctuations). Therefore, a prior linear trend was chosen to be more suitable compared to using a constant (mean) or a quadratic (or higher polynomial) function as a trend. In the revised paper we have now expanded the relevant sentence as follow (p7, L12) to explain the choice:

"In this study, the mean function was chosen to have a linear form $\beta_o + \sum_{i=1}^{13} \beta_i x_i$, on the basis that the response of the surface concentration to changes in input emissions is expected to be smooth with no discontinuities or fluctuations."

7) *P7,L10 The authors state that the emulator error was estimated using cross validation and this is presented in the SI. However very little detail is given there except a reference to a paper describing the Matlab package used to construct the emulators (Lataniotis, 2017), which says that the package uses cross validation for parameter estimation, not emulator validation. The authors also say in the paragraph above that they used cross validation for parameter estimation. A clear statement is required as to whether or not the same cross validation was used for both parameter estimation and emulator validation. The accuracy of the emulators is of such key importance to everything that follows that summary statistics of either a separate cross validation, or a validation with a held-out data set, must be presented in the main text.*

Response: We have checked the cross-validation error values by performing a separate cross-validation calculation. The results of explicitly-performed cross-validation do not differ from that which was originally presented in the supplementary material. We accept that the reference we originally cited may have caused some confusion, as the formula for cross-validation is reported in the parameter estimation section. The same formula was used for calculating the cross-validation error for the emulator. We now instead explicitly state in the supplementary information the cross-validation equation used. The three figures in the supplementary information have now been replaced with the updated cross-validation versions. The code written to perform these independent cross-validations is also added to the data repository for this paper.

8) *P8,L2 Could the authors comment on the validity of describing the 5km grid square containing Marylebone Rd as 'urban traffic' and the one containing N. Kensington as 'urban background'. Surely at this resolution both grid squares must be considered as urban background – this is demonstrated by the almost identical sensitivity of NO_2 and O_3 concentrations to NO_x emissions from road transport shown in figure 8.*

Response: The descriptors assigned to the grid squares are the descriptors used by Defra and the UK air quality community for the national network air quality monitoring stations in those grid squares. Whilst we agree that the 5 km resolution of the model causes some averaging of modelled pollutant output compared with a point measurement at those monitoring stations, we do not agree that our results are almost identical for the two grids. The modelled concentrations of NO_2 and O_3 as well as the uncertainty values are different for the two grid squares, and the sensitivity indices (fig. 8) are similar but not identical. It is not

claimed that the "Marylebone Rd" patterns of sensitivity are applicable to all 'urban traffic' sites.

9) *P8,L5 The authors state that the "sensitivity indices were estimated". There are a numbered of published methods for doing this so could they say which method they used and why.*

Response: The references provided in the paper describe the numerical methods used for the estimation of the first and total order indices. These methods for estimation of the sensitivity indices were chosen as they are widely used and well-established methods. However, to clarify explicitly the methods we used for these indices we have now added the following sentence to section 2.4.2: "The first and total-order sensitivity indices were estimated following the methods described by Sobol' (1993) and Janon et al. (2014) respectively."

10) *P14,L9 Given that O_3 concentration is known to be highly non-linear and nonmonotonic in response to changes in NOx and VOC emissions, the contention that these emulators could be used to test emission reduction scenarios is questionable as the small number of training runs used to construct the emulators means they would be unlikely to be able to accurately predict the emissions levels at which the chemical regime changes between being NOx sensitive and VOC sensitive.*

Response: The response of O_3 to changes in NO_x and VOC concentrations although non-linear is still smooth. The effect on O_3 is analysed and presented on annual and monthly timescales, hence additional hourly and daily fluctuations are smoothed out.

11) *P14,L19 Could the very low sensitivity of annual mean ozone to emissions be a result of the strong diurnal variation and photochemical nature of the production of this pollutant? Could the authors comment on whether annual average 8-hour maximum might have been a more informative metric to emulate.*

Response: The annual average 8-hour metric and the annual average are very strongly correlated metrics (see, for example, page 34 of AQEG (2009)); so whilst the absolute values will change we do not believe the nature of the O_3 response to the input emission perturbations will change.

12) *P14,L20 As the authors findings that variation in emissions does not cause substantial variation in the outputs contrasts with their statement in the introduction that it has been previously found that uncertainties in input emissions are major contributors to the uncertainty in the ACTM outputs, could they comment on why this might be – do they think it is a feature of the EMEP model or a result of the analysis methods?*

Response: As was noted in the paper, the previous studies we refer to were performed with different models. The lack of sensitivity here could be due both to the fact that annual and monthly averaged surface concentrations were investigated and to the possible strong effect

of background concentrations of pollutants as well as pollutants transported from outside the UK.

The point of this study was not to show how good or not the EMEP4UK model is at reproducing observed pollutant concentrations but to demonstrate a method that can be used to improve the understanding of model response and/or to potentially point out the aspects that may need attention or improvement in future model development.

13) *P14,L25 Given the authors assertion in the previous paragraph that the uncertainty in model output is likely to be driven mainly by variables that they have not included in their analysis, would they concede that their uncertainty estimates are likely to underestimate by a large degree the real uncertainty in the model output, i.e. that caused by uncertainty in all of the input variables plus the model discrepancy.*

Response: Yes, we agree that the total uncertainty in the estimated concentration of the pollutants is higher than the uncertainty propagated from the input emissions only as there are other uncertain model inputs and parameters. Ideally sensitivity analysis should be incorporated as a part of the model development process (which could aid in both model simplification/reduction and calibration). In that approach the effect of all uncertain inputs and parameters could be assessed without having to do it retrospectively. In addition, screening techniques, e.g. the Morris method (Morris, 1991), could be applied to identify the inputs and parameters that most drive variation in the model outputs and which need to be investigated further. However, here we concentrate on presenting the application of the method itself and on a subset of inputs which previously was shown to drive uncertainty in the model output values.

References:

AQEG: Ozone in the United Kingdom. Fifth report of the Air Quality Expert Group., UK Department for Environment, Food and Rural Affairs, London. PB13216. ISBN 978-0-85521-184-4. http://uk-air.defra.gov.uk/library/reports?report_id=544, 2009.

Janon, A., Klein, T., Lagnoux, A., Nodet, M. and Prieur, C.: Asymptotic normality and efficiency of two Sobol index estimators, ESAIM Probab. Stat., 18, 342–364, doi:10.1051/ps/2013040, 2014.

Morris, M. D.: Factorial Sampling Plans for Preliminary Computational Experiments, Technometrics, 33(2), 161, doi:10.2307/1269043, 1991.

Simpson, D., Benedictow, a., Berge, H., Bergström, R., Emberson, L. D., Fagerli, H., Flechard, C. R., Hayman, G. D., Gauss, M., Jonson, J. E., Jenkin, M. E., Nyíri, a., Richter, C., Semeena, V. S., Tsyro, S., Tuovinen, J.-P., Valdebenito, Á. and Wind, P.: The EMEP MSC-W chemical transport model – technical description, Atmos. Chem. Phys., 12(16), 7825–7865, doi:10.5194/acp-12-7825-2012, 2012

Sobol', I. M.: Sensitivity estimates for nonlinear mathematical models, Math. Model. Comput. Exp., 1(4), 407–414, 1993.)