Response to second review comments

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

by Aleksankina et al.

Comments made on our revised paper by the co-editor and the reviewer are given in italics.

Author responses to Co-editor's non-public comments to the authors

I consulted with Reviewer #1 who opines that his/her comments have led to only a few minor revisions in the main text of the manuscript itself. Please consider whether revisions to your manuscript would be appropriate in response to Reviewer #1 comments #3, 4, 5, 8, 10, 11, 12, and 13. You have been much more responsive to comments from Reviewer #2 in terms of making revisions to the manuscript. With respect to comments from Reviewer #2, the addition of a qualifier ("so-called" compensation of errors) in response to comment #9 is not sufficient or helpful to readers. Please also consider whether revisions to the manuscript are warranted in response to Reviewer #2, comment #10.

I am fine with the meta-model evaluation being presented in supporting information as you have done if that is your preference.

Author response: The following revisions to the manuscript have been made in response to the comments 3, 4, 5, 8, 10, 11, 12, and 13 made by Reviewer #1 in their first review.

Relating to Reviewer #1's comment no. 3: The following text has now been added to Section 2.2: "In this study, emissions of all of the major primary anthropogenic pollutant compounds were investigated. The decision to concentrate on the anthropogenic emissions was made based on the fact that one of the main applications of the EMEP4UK model is in providing scientific support for policy-making regarding impacts of interventions leading to anthropogenic emissions reductions. Hence 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. Therefore, it was decided to investigate model response to the changes in anthropogenic emissions."

Relating to Reviewer #1's comment no. 4: The following text has now been added to Section 2.2: "The shipping emissions were not split by pollutant because the inclusion of this split for this source comprised too great a computational cost for the analyses. Most shipping emissions do not impact on land-based population and ecosystem exposure, which was our focus, compared with the terrestrial emissions."

Relating to Reviewer #1's comment no. 5: The following text has now been added to Section 2.3: "In this study, 84 sampling points were found to be sufficient to create an adequately performing emulator as the input-output response function for the EMEP4UK model was expected to be smooth on monthly and annually averaged timescales. More generally, however, the case where model runs are computationally expensive and input-output relationship is less predictable a sequential sampling technique can be applied to track the improvement of emulator performance with increase in the sample size."

Relating to Reviewer #1's comment no. 8: As we noted in our original response, we disagree that our results show no difference between the London Marylebone Rd and London N. Kensington grid cells. We described and discussed the differences in the 'Results and discussions' section. On the matter of grid-averaging and 'site type', it is not within the scope of this paper to start discussion on issues associated with comparing model volume outputs with point source measurements; however, in order to emphasise that the names of 'urban background' and 'roadside' are not our designations but those assigned by the measurement network operators we have now amended text at the start of Section 3.3 to read: "for five different grid cells that were assigned the following environment types based on the site-type attributed to the national-network monitoring site…"

Relating to Reviewer #1's comment no. 10: The text now added to Section 2.2 in response to this reviewer's comment no. 5 (see above) also covers this comment, as there are no expected discontinuities or sharp peaks and troughs in the O_3 response to the input emissions on the annual or monthly timescale.

Relating to Reviewer #1's comment no. 11: We do not believe there is any additional benefit to our paper of adding a statement somewhere that the annual average 8-hour maximum O_3 concentration metric is well correlated to the annual average concentration metric. We have responded to the reviewer's question on this in our online responses to the comment when it was originally made.

Relating to Reviewer #1's comment no. 12: Our text already includes a statement that the surface concentrations of the modelled pollutants in the UK may be dominated by the precursor emissions and long-range transport from outside the UK and are therefore relatively insensitive to changes in the UK emissions. We have now also added additional text to the paper relating to the reviewer's comments nos. 3 (as above) and 13 (as below) that further acknowledges that our work focused only on investigations of model uncertainty via land-based emissions.

Relating to Reviewer #1's comment no. 13: The following text has now been added to Section 3.4: "Finally, in this study the overall model output uncertainty is likely to be lower than the theoretical total model output uncertainty, as in addition to the input emissions there is a variety of other uncertain model inputs. Assessing the effect of variation in every model input and parameter on the model output is a laborious task, hence ideally sensitivity analysis should be incorporated as a part of the model development process. By using this approach, the effect of all uncertain inputs and parameters could be assessed without having to do it retrospectively."

Author response: The following revisions to the manuscript have been made in response to the comments no. 9 and 10 made by Reviewer #2 in their first review.

Relating to Reviewer #2's comment no. *9:*
Original comment *Results Section 3.1: You refer to the 'compensation of errors' as one explanation why the surface response is weak given the input uncertainties. Can you point to the literature for evidence of this statement? I've only seen "compensation of errors" only referred to in the context of process representation in models.*
Our original Response*: The phrase comes from Skeffington et al. 2007. In that paper the reason for narrowing of confidence limits for critical loads compared to those of the input*

parameters was explained to be due to a "compensation of errors" mechanism, but no further explanation was provided. Here, by compensation of errors we mean a situation when the variation in the output is less than expected. This could be caused by multiple inputs having an opposite effect on the magnitude of change in the output of interest.
We have changed the phrasing in our paper to "so-called compensation of errors".

Updated response: Perhaps the co-editor did not notice that the text in the manuscript already explains what was meant by 'compensation error', viz: "Another explanation is the 'so-called compensation of errors' whereby a positive effect of one or multiple input variables on the output is compensated by a negative effect of another input variable(s). This leads to the narrower confidence intervals associated with the EMEP4UK outputs."
At the last paper revision stage we chose to add the qualifier 'so called' so as to emphasise that we were using the 'compensation error' terminology used by Skeffington et al. (2007) and that it was not our phrasing. We have now added the Skeffington et al. (2007) reference to the relevant sentence in the paper.

Relating to Reviewer #2's comment no. *10:*
Original comment *Results Section 3.3: One potential explanation for the seasonal change in sensitivity at Harwell to shipping emissions is the seasonal change in the wind direction which results in more NOx from shipping emissions being transported to the site. Can this be verified from the WRF meteorology used to drive the model?*
Our original response: *The seasonal wind speed and direction for the year 2012 is shown in Figure A below, using the meteorology supplied from the AURN data as extracted using the openair package. It could be argued that there is some correlation between sensitivity index patterns in Figure 8 of our paper and the wind direction; however most likely the seasonality in NOx sensitivity to shipping emissions is due to a combination of interacting processes within the model.*

Updated response:
The wind rose for Harwell AURN site for year 2012 has been added to the supplementary information and has been referenced in the main text of the manuscript in Section 3.3. "Potential explanation for this is seasonal change in the wind direction which results in more NO_x from shipping emissions being transported to the grid cell during the summer months (the wind rose is presented in Fig. S3)."

Author responses to Reviewer #1's second review report

The authors have chosen to respond to most of the original comments in the online discussion rather than modifying the manuscript. Many of these responses are acceptable, but there are however a couple of important issues which have not been adequately addressed. Those original comments, along with the authors' response and a corresponding assessment are given below.

Original comment: *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.*

Authors' 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.*

Assessment of response: *It is correct that the paper does not state that 10d model runs must be performed in every case, rather that it is a rule of thumb for an initial experiment, after which more runs may be performed if the emulators do not prove to be accurate enough. There are, however, published studies using fewer than 10d model runs, which I am sure the authors will be familiar with, and this is perfectly acceptable as long as a thorough emulator validation is performed – please see the point below.*

Authors' response: We don't think that any further comment is needed here, nor any further modification to our paper. The reviewer is agreeing with us that 10d sample points is a rule of thumb given by one paper and that other studies have used fewer – we respond to the reviewer's comment on emulator validation below.

Original comment: *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.*

Authors' 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.
Assessment of response: *Unfortunately the authors have chosen not to present the results of their emulator validation in the main text, as requested, despite this being one of the most critical parts of the analysis. Every other published study using these methods that I am aware of presents the emulator validation as an important part of the main text because the validity of all of the results of such studies is dependent upon it. The authors have instead made a small modification to the supporting information, but still only give the emulator error as a fraction of the model output variance, so that the reader has no way of knowing what the absolute magnitude of the emulation errors might be.*
What has been presented in the supporting information would be a useful addition to a proper validation exercise, if the emulator variance were given as a fraction of the model output variance, rather than the emulator error, as this would support the use of the emulator mean value in the sensitivity calculations which follow.

Authors' response: As stated in our paper, k-fold cross-validation has been used to evaluate the performance of the chosen emulator; the results of the cross-validation and the code used to perform it are made available. We chose not to overwhelm the manuscript and the reader with extensive detail of the code implementation of the Gaussian Process emulator and its validation as the original MATLAB code for the emulator is a part of a freely available package UQLab. Additionally, the manuscript concentrates on the application of already developed methods (emulation, sensitivity analysis, and uncertainty analysis) rather than the development of those numerical methods.

The co-editor has confirmed that they are happy that our evaluation of the emulator is presented in the supporting information.

We have chosen to present the result of k-fold cross-validation (CV) as a pollutant concentration value (i.e. value of the model output) divided by the overall variance of the model output in order to make the number more meaningful. In the areas of the modelled domain where model is not sensitive to changes in the input emissions the variance around the predicted mean is very small, hence the CV error is going to be small. By presenting CV errors relative to the variance of the output for the corresponding grid cell, we avoid falling into a trap of stating that in these grid cells the emulator performs well because we take into account the sensitivity to the underlying inputs.

Original comment: *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.*
Authors' 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.
Assessment of response: *Given that the authors agree with this comment, it would seem appropriate to modify the manuscript accordingly.*

Authors's response: In addition to the existing manuscript text: "[Overall uncertainty was found to be low]. This indicates that the variation in the input data (i.e. emissions) does not cause a substantial variation in the outputs. Our results indicate, that this can likely be explained by variations in the other model input parameters such as chemical reaction rates, deposition velocities or physical constant values which might cause more variation in the model outputs. Alternatively, surface concentrations of the modelled pollutants in the UK may be dominated by the precursor emissions and long-range transport from outside the UK and are therefore relatively insensitive to changes in the UK emissions."
The following text has now been added to the end of Section 3.4: "Finally, in this study the overall model output uncertainty is likely to be lower than the theoretical total model output uncertainty, as in addition to the input emissions there is a variety of other uncertain model inputs. Assessing the effect of variation in every model input and parameter on the model output is a laborious task, hence ideally sensitivity analysis should be incorporated as a part of the model development process. By using this approach, the effect of all uncertain inputs and parameters could be assessed without having to do it retrospectively."