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.

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.

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.