Responses by author to interactive comments by RC1 on “Calibrating a global atmospheric chemistry transport model using Gaussian process emulation and ground-level concentrations of ozone and carbon monoxide" by Edmund Ryan and Oliver Wild

(1) Reviewer’s comment: The authors use reanalysis data from Flemming et al. (2017). It would help the reader if the authors make a clearer reference to this dataset by referring to it as the “CAMS interim Reanalysis”, considering that ECMWF has published various atmospheric composition reanalyses. Also, even though atmospheric ozone in this reanalysis is constrained by satellite observations, the constraints on surface ozone concentrations are typically less, and are more governed by the model assumptions. To the least, it would be worthwhile to provide insight in the quality of surface ozone and carbon monoxide, e.g. as documented in Huijnen et al. (2020), and/or Flemming et al. (2017), and to what extent these aspects may alter your analysis. Also along these lines, the reference to “measurement data” (pp 15, line 7) may be confusing in this context.

Author’s response: Thank you for pointing out the potential confusion in naming. We have amended the text to refer to this dataset in the manner suggested. We have also added a sentence to note the strengths and weaknesses of this dataset referring to these papers to give a more complete picture of the reanalysis data. We have replaced the reference to “measurement data” with “surface concentration data” to avoid confusion.

Changes to manuscript: We now state: “This reanalysis reproduces observed O_3 and CO distributions relatively well, and biases at surface measurement stations are generally small (Huijnen et al., 2020). The dataset also has the benefit of complete global coverage, allowing us to test the importance of measurement coverage directly.” We now specifically state “we use CAMS interim reanalysis data” in Section 2.2.

(2) Reviewer’s comment: Page 10, line 7. The authors write a little cryptically: “We included p as one of the parameters to estimate for the reanalysis data and found values in the range 0.16-0.19.” Not being a specialist in mathematics, could you give some interpretation of this analysis? Why is it in this range, and not much smaller (or much larger)?

Author’s response: Thank you for highlighting this. The 0.16-0.19 range was derived from estimating the parameters using the reanalysis, prior to carrying out the model calibration involving the synthetic datasets. However, we appreciate that the way this sentence is phrased is not entirely clear. In the revised manuscript we have rewritten this and added extra detail to improve clarity.

Changes to manuscript: We now state “where p is a scaling factor that provides a measure of the representation error. We used the reanalysis data to estimate p alongside the model parameters, and found posterior values of p that were in the range 0.16-0.19. We therefore selected four values of p (0.01, 0.1, 0.2 and 0.3) to explore the importance of representation error when using the synthetic data.”

(3) Reviewer's comment: Page 11, line 8: "In contrast, methods based on neural networks can require thousands of training runs.": This statement reads a bit unfounded. Could you please either add a reference (e.g. a study which actually uses a neural network approach to model atmospheric chemistry), or further clarification? If I understand correctly, one of the key aspects compared to published use of machine learning in atmospheric chemistry is that the full CTM is replaced by the Gaussian Process Emulator, rather than 'just' the chemistry solver code as I am aware of (Keller et al., 2019)

Author's response: Thank you for this comment. The point we make here is that more traditional machine learning methods for mapping model inputs to model outputs typically have hundreds of parameters, and for this reason thousands of training points are required. Neural network methods have not been applied yet to full CTMs for this reason, although they have been used for specific components of atmospheric models, as the reviewer notes. A key benefit of Gaussian Process emulation is the greatly reduced number of training runs that permit the methods to be applied to more complex models. We have rephrased the sentence to make this clearer. We do not cite the Keller et al. paper here, as the point we are making is not specific to atmospheric models, and because they used a random forest approach, but we agree that it is a good example of application of machine learning methods in the discipline.

Changes to manuscript: We now state "This is in contrast to methods based on neural networks, which often have a large number of parameters that necessitate thousands of training runs."

(4) Reviewer's comment: The results from the sensitivity analysis (Sec. 3.1) indicate that surface ozone deposition is the largest driver to explain biases in modeled surface ozone concentration. While I completely understand from mathematical perspective, and experience with old-fashioned tuning in a chemistry transport model, that by changing (modifying) the dry deposition velocity for ozone has indeed a direct impact on surface concentrations, I find this result also a little worrying, in light of what would be a reasonable range of the ozone deposition flux. To get a handle on this, would it be easy and useful to quantify the change in annual total surface ozone deposition as suggested by this optimization method? In fact, the authors also rightly discuss the issues with this sensitivity analysis when including the simultaneous optimization of CO, pp22, line 6-9. When reading this manuscript, I had found it helpful if this aspect was already alluded to in Sec. 3.1 . Indeed, I agree that the use of this synthetic modeling can be very useful, but different optimization factors obtained using (in this case) different combinations of control variables also stress the danger of a false impression of a physically well-constrained parameter.

Author's response: Our aim in this paper is to demonstrate the feasibility of calibrating a full CTM, and we have intentionally chosen to use a simplified system with only 8 scaling parameters to demonstrate the method. As there are many sources of uncertainty that we do not consider here, we do not expect the calibration to generate the same results that it might when including all sources. However, we still expect it to provide useful insight. Surface O_3 is a little high over land masses in the model, and thus in the absence of consideration of uncertainty in chemical processes, the calibration suggests that dry deposition fluxes should be about 40% greater than the a priori value. Deposition accounts for 930 Tg(O_3)/yr in the model, so this would constitute a sink of 1300 Tg/yr (neglecting feedbacks); while this is at the high end of recent estimates, it remains within the spread of published model results. However, in the absence of consideration of all sources of uncertainty, we choose to highlight the ability to constrain processes rather than the value of the scaling factors that arise, for the reasons that the reviewer identifies. For the simple system we consider, the parameters are constrained well for this parameter as shown by the small posterior uncertainty. A more complete study addressing uncertainty in a much wider range of processes is needed to generate a more robust assessment of the scaling factors needed for a particular process. In Section 3.1 we already acknowledge the limited range of processes considered (“of the eight considered here”) but to address the reviewer’s concerns we add a statement at the end of Section 2.8 that alludes to the sensitivity of the calibration to the system considered.

Changes to manuscript: We have now added: "We discuss the implication of these results and the limitations of considering a simple eight-parameter system rather than all sources of model uncertainty in Section 4.”

(5) Reviewer’s comment: Page 18, line 19: “along with a reduction in associated uncertainty”: Could you be more explicit on this? I don’t directly see such decrease in uncertainty.

Author’s response: Thank you for spotting this error. The uncertainty is reduced compared to using CO alone, but is comparable to that using O_3 alone, and therefore this phrase isn’t needed. We have now removed this part of the sentence.

Changes to manuscript: this phrase has been removed.

(6) Reviewer’s comment: Figure 10, panels c/f/i : In almost any of the test configurations the parameter estimate for the Boundary Layer Mixing is approaching the maximum range that is given. Can you provide further interpretation in this aspect? Is the given uncertainty range for boundary layer mixing sufficient? Now it is written on page 21, l. 11-13 that this process “may not be represented well in the model”. Or could this be an artifact of other (missing) processes not considered in this sensitivity analysis, or that this process is just treated differently between datasets.

Author’s response: The range of the Boundary Layer Mixing scaling parameter is already large, spanning a factor of 100. The fact that the parameter estimates lie very close to the bounds suggests that the process is not well represented in the model. However, the reviewer is right to point out that other processes may be important; we are considering a simplified system here with only 8 scaling parameters and the boundary layer mixing parameter may thus be acting as a surrogate for uncertainty in processes not considered here, most notably chemical processes. We already acknowledge this point in Section 4.4 of the paper. Our aim in the paper is to demonstrate how well the calibration method works in a simplified system, and more complete coverage of uncertain parameters would be needed to provide a more robust assessment of specific processes. However, we have altered the text at the end of section 4.3 to acknowledge the influence of other processes.

Changes to manuscript: We now state: “In contrast, our estimates of the boundary layer mixing scaling parameter are substantially larger than those from the model, suggesting that this process is not represented well in the model, or that other processes not considered here may be influencing the result.”

(7) Reviewer’s comment: pp 21, l 13: “Our results suggest that dry deposition and isoprene emissions are represented relatively well in the FRSGC/UCI CTM “ : Good to add phrase “with respect to the independent reanalysis data”?

Author’s response: We agree that this should be clarified and have rephrased this sentence in the manuscript as suggested.

Changes to manuscript: We have rephrased the sentence as suggested so that it now reads “Overall, our estimates of the dry deposition and isoprene emission scaling parameters are close to a priori values from the FRSGC/UCI CTM, with respect to the independent reanalysis data.”

(8) Reviewer’s comment: pp 23, l9: “…while it *is* effective …”

Author’s response: Thank you for spotting this. We have corrected this in the revised manuscript.

Response by author to interactive comment by RC2 on “Calibrating a global atmospheric chemistry transport model using Gaussian process emulation and ground-level concentrations of ozone and carbon monoxide" by Edmund Ryan and Oliver Wild

Major Comments

(1) Reviewer’s comment: When I read the abstract and introduction, I have an impression that the author will apply the calibration framework developed by Kenndy and O’Hagan (2001), but it turns out that it is not the case. The author should have explained in the beginning why they do not include model discrepancy in the methodology or GP model. It is well known that model discrepancy is very important for model calibration (Brynjarsdóttir and O’Hagan, 2014), and it is obvious there is no CTM that can perfectly predict the true process, even with the optimised inputs.
Brynjarsdóttir, J. and O'Hagan, A.: Learning about physical parameters: The importance of model discrepancy, Inverse Problem, 30, 114007, 2014

Author’s response: Many thanks for this comment. We have specifically chosen not to include a discrepancy term for two reasons:

- For the scenarios where we use synthetic data, no discrepancy term is required because the synthetic data is generated by adding noise and spatial gaps to the emulator output for the control run.
- For the scenarios involving reanalysis data, there is no simple and defensible method to estimate the discrepancy term.

Neither of the papers cited in the reviewer’s comment explain the need for the term for anything other than the most simple model system.

The discrepancy represents the missing processes in the model. However we often don’t know what these missing processes are or how to estimate them. When performing a regular model calibration without an emulator (e.g., applying MCMC on the original model) we would not include a discrepancy term. It is therefore not clear why we need to include a discrepancy term into the calibration formulation when using an emulator. We highlight that in the abstract of the Brynjarsdóttir and O’Hagan paper it states: *“The challenge with incorporating model discrepancy in statistical inverse problems is being confounded with calibration parameters, which will only be resolved with meaningful priors”.* If we wish to estimate the discrepancy term as part of the calibration process, then this confounds the parameters we are trying to estimate. Hence, the paper makes the point that you need highly informative priors. The paper does not address what to do if you do not have these, as is the case here.

However, to investigate the importance of this term we adopt the simple rule of thumb that the discrepancy term is 10% of the magnitude of the observation (Jeremy Oakley, personal communication). We repeated the experiment to estimate the eight scaling parameters and the SD term using surface ozone reanalysis data at 2.5% spatial coverage. We find that there is almost no difference in the marginal posterior distribution when we include the discrepancy terms in the formulation compared with when we omit them, see Figure 1.

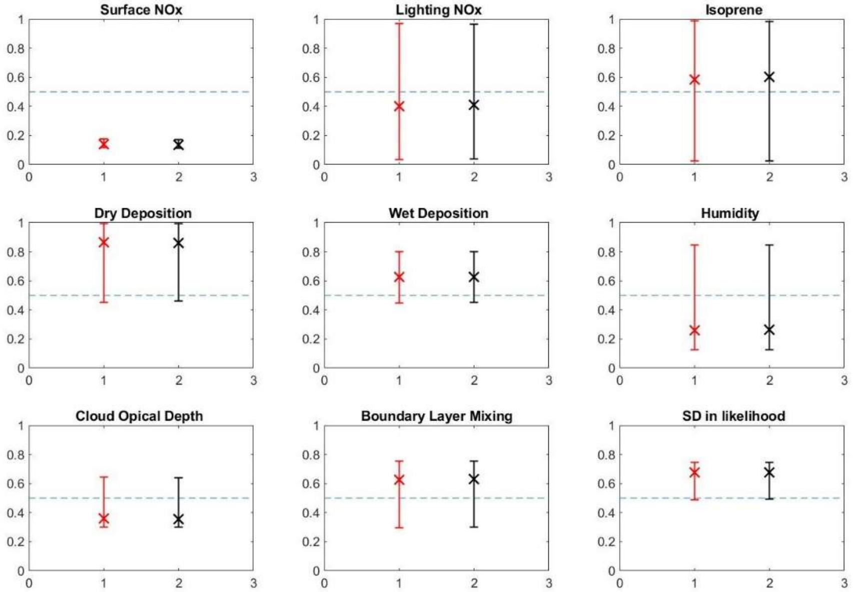

Figure 1: Mean and 95% credible interval of the marginal posterior distributions of the eight scaling parameters and the SD term, corresponding to the MCMC scenario involving surface ozone reanalysis with 2.5% spatial coverage. The red crosses and error bars correspond to the results using the discrepancy term in the emulator formulation, while the black crosses show the standard results without.

These results clearly demonstrate that including a discrepancy term makes no substantial difference to the derived posterior distributions for the situation that we are considering here. We therefore feel that in the absence of clear justification for the use of the term and how it might be estimated, our results stand as they are.

Changes to manuscript: The results relating to the reanalysis data remain unchanged, but we have included a new subsection in the methods (2.8) which gives reasons why we not included a discrepancy term in this study, but also that recognises that including such a term may be useful in other calibration studies. This includes the following text:
"It has been suggested that a model discrepancy term should be included when carrying out model calibration involving Gaussian process emulators (e.g. Kennedy and O'Hagan, 2001; Brynjarsdóttir and O'Hagan, 2014). The discrepancy term represents the processes missing in the model. However, in this demonstration study we have chosen not to include a discrepancy term for two reasons. Firstly, for scenarios where we use synthetic data, no discrepancy term is required because the synthetic data is generated by adding noise and spatial gaps to the emulator output for the control run. Secondly, for scenarios involving reanalysis data, there is no simple and defensible method to estimate the term. When performing model calibration by applying MCMC directly, a discrepancy term would not be included. Since the purpose of the emulator here is to estimate the output of the model for a given set of parameter values, we argue that it is not necessary to include a discrepancy term into the calibration formulation. However, we agree that including such a term may be helpful in situations where there is good prior information.

"To investigate the importance of a discrepancy term, we repeat the experiment to estimate the eight scaling parameters using surface ozone reanalysis data and assuming a discrepancy term that is 10% of the magnitude of the observation. We find that there is almost no difference in the marginal posterior distribution when we include the discrepancy term compared with when we omit it (see Figure S16 in the supplemental material). We therefore choose to omit the term for our study."

We now include the figure above as Figure S16 in the Supplementary material.

(2) Reviewer's comment: I used to think the calibration won't work on tropospheric ozone, since accurate tropospheric ozone simulation relies on accurate regional emission inventory and meteorology. These inputs need to be continuously updated, rather than calibrated. As far as I am aware, the existing calibration approaches assume the parameters are "fixed initial condition", rather than time-varying or spatial-varying settings. Would it be making more sense to consider such input, e.g. emission inventory, as a time series?
Morris, M. D. (2012). Gaussian surrogates for computer models with time-varying inputs and outputs. Technometrics, 54(1), 42-50.

Author's response: The parameters we are estimating are global scaling parameters applied uniformly to major input variables, e.g. global surface NOx emissions. These variables already vary in space and time, and our parameters merely scale the global annual magnitude. It would certainly be interesting in a future study to explore the effects of seasonality through temporally varying factors, as the reviewer suggests, although this comes at the cost of a greatly increased number of parameters to calibrate. However, for the current study our aim is to demonstrate the feasibility of calibration and thus we have attempted to keep the problem as simple as possible, and focus only on a single, globally uniform scaling parameter for a given input. We anticipate that using the approaches successfully demonstrated here, atmospheric modellers will be able to explore the more complex spatial and temporal sensitivities for different variables in future.

Changes to manuscript: We have rephrased our introduction to the scaling factors in section 2.1 to highlight that they are global factors applied uniformly to the particular process considered. We now state: "To provide a simple and easily interpretable approach to calibration, we define a global scaling factor for each process that spans the range of uncertainty in the process and that is applied uniformly in space and time."

(3) Reviewer's comment: The authors split the total output into 6528 separated GP emulators (p13). Since each emulator is treated separately, how can they be certain all the emulators yield the same calibrated inputs?
- what if the calibrated inputs from different emulators are diverse?
- even if the calibrated inputs from different emulators are similar, how the optimised inputs are determined?

Author's response: The estimates and associated uncertainties of the eight scaling parameters are determined using a common type of Markov Chain Monte Carlo (MCMC) algorithm called Gibbs sampling. In its basic form, Gibbs sampling is a special case of the Metropolis-Hastings algorithm. A key component of the algorithm is a likelihood function which quantifies the mismatch between the measurements and the model predictions for a given set of values for the parameters. The 6528 separate emulators correspond to 12 monthly measurements at each of the 272 spatially varying grid cells, for each of the 2 variables (ozone and carbon monoxide). In short, the number of emulators corresponds to the dimension of the output space. For the scenario involving all 6528 emulators, the likelihood function incorporates all 6528 modelled values (where the emulator is used in place of the chemistry model) and the corresponding 6528 measurement values. Thus while the emulators are treated separately, only a single likelihood function is considered, not 6528 separate likelihood functions as the reviewer's comment suggests. We thus end up with a single set of calibrated inputs that has already been optimised over all times and locations. We have made changes to our description of the approach used to make this clearer.

Changes to manuscript: We recognise that in the current description of the MCMC algorithm is brief and in particular we do not mention Gibbs sampling.
We now provide more detail at the start of Section 2.7: "This uses Gibbs sampling, which is an approach based on Markov Chain Monte Carlo (MCMC) that we use to determine multi-dimensional posterior probability distribution of the model parameters (Gelman et al., 2013). Gibbs sampling is an extension of the more traditional Metropolis-Hasting variant of MCMC, and uses conditional probability to sample from the marginal distribution when moving around the multi-dimensional parameter space."

(4) Reviewer's comment: Given the fact that the authors build an emulator separately for each location and month, I am actually concerned whether this study is actually useful. If the inputs are only calibrated at a specific month/location, and might not be valid for other months/locations, then this message is not really useful for modelers. If they cannot run the model at a single setting of the input parameters, the calibration does not consider to be working. It is also contradictory to its title "Calibrating a global atmospheric chemistry transport model", because the authors have not taken any spatial and temporal correlations into account in their GP model.

Author's response: As noted in our previous response, the likelihood function is applied across all independent emulators so that we derive a single set of calibrated inputs across all times and places. We agree that independent calibration of each emulator would not be useful, as the reviewer suggests, but we hope that we have now addressed this misunderstanding through the changes we have made to the text in response to the comment above. Our approach generates a single set of scaling parameters and thus allows us to effectively calibrate the model.

Changes to manuscript: To address this point, we have added two sentences following equation 4 in section 2.7: "We note that although separate emulators are used for each of the spatial and temporal locations in the model output, there is still only a single likelihood function. Hence, evaluating all of the emulators for a specific set of values of the scaling parameters is equivalent to evaluating the CTM once at those values of the parameters."

(5) Reviewer's comment: A collection of 272 locations is not strictly prohibited for GP emulation/calibration, why building 24 emulators is not an option?

Author's response: As noted above, we build separate emulators for each variable (ozone and CO) for each month at each of the 272 grid cells. It would be possible to reduce the number of emulators needed through application of principal component analysis methods, which we have demonstrated in a previous study (Ryan et al., 2018), but we have chosen to generate separate emulators here both for reasons of simplicity and because emulator generation is not the most computationally demanding aspect of this study.

Changes to manuscript: To address this point we have added a new paragraph to the end of section 2.6: "Finally, we recognise that principal component analysis (PCA) could be used to reduce the dimensionality of the output space and hence the number of emulators required (Higdon et al., 2008). In a previous study we found that a PCA-emulator hybrid approach resulted in similar performance compared to using separate emulators for each point in the output space, and reduced the number of emulators required from 2000 to 40 or fewer (Ryan et al., 2018). However, for this study, we choose an emulator-only approach because it is much simpler to demonstrate. Nonetheless, future emulation-calibration studies could benefit from the computational savings of applying a PCA-emulator hybrid approach. Other approaches for dealing with high dimensional output are also available, such as low rank approximations (Bayerri et al.,2007)."

(6) Reviewer's comment: Calibration is not just about estimating the optimised inputs, but also about estimating (under reasonable assumptions) the potential simulator misspecification, or discrepancy. However, this latter estimation is not explored in the studies. This should be explored more carefully, even if the model discrepancy is not accounted for. For example, does the final calibrated model outperform than the ensemble mean output from their Latin hypercube design with respect to synthetic or real observations?

Author's response: The reviewer makes a good point here about estimating the model misspecification. To address this point, we have created a new figure that compares the emulator predictions of the measurements using the prior and posterior values for the inputs (figure 2). The left-hand plot shows the mean and 95% prediction interval of the surface O_3 as predicted by the emulators, using 1000 samples of the inputs from the prior distribution. For the right-hand plot, we took 1000 samples of the inputs from posterior distribution, based on the calibration run involving only surface O_3 data, 20% spatial coverage, and a representation error factor of p=0.2 (the third level of representation error). We then ran all 1000 posterior samples through each emulator. For each panel there are 272*12 points, which are made up of 272 spatial pixels and 12 months for each pixel. Note that although the calibration run involved 20% of the data, these predictions involve 100% of the data. We can clearly see that the prior distribution values of the predicted surface ozone are unbiased and have large uncertainty (more so at high values). In contrast the posterior distribution values of the predicted ozone have a very small value of the median absolute difference (MAD).

Note that we have updated this figure and above text from our initial response to the reviewer's comments, due to a typo in the code used to create the figure that had resulted in bias in the prior distribution.

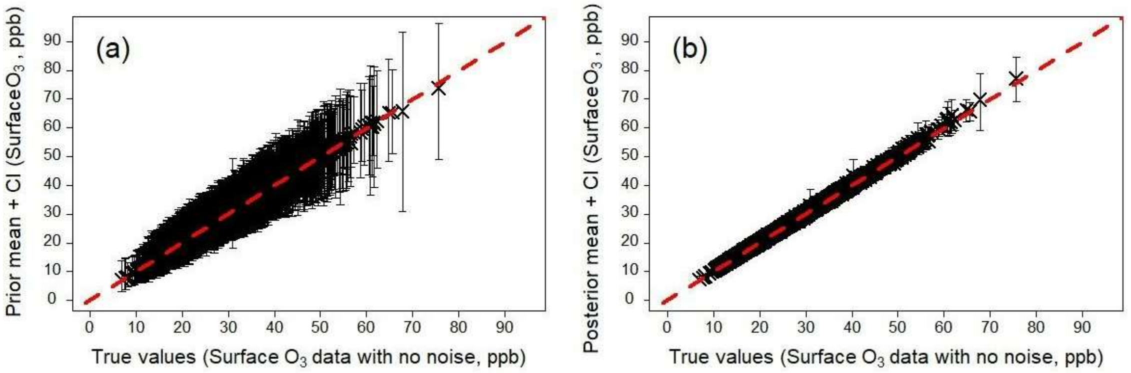

Figure 2

Changes to manuscript: A new section 3.4 has been created and the above figure has also been included in the revised manuscript as figure 11. Section 3.4 states: "We demonstrate the benefit of the calibration by evaluating the emulators using the values of the scaling parameters sampled from the prior and posterior distributions. As an example, we show surface O$_3$ before and after calibration using the calibration runs involving synthetic data at 20% spatial coverage and a representation error of p=0.2 (Figure 11). Despite the calibration involving only 20% spatial coverage, we apply the resulting parameter values to all grid squares. We can clearly see that the prior surface O$_3$ concentrations are unbiased but have large uncertainty, especially at high values. In contrast the calibrated O$_3$ concentrations have a small uncertainty, demonstrating that even with 20% spatial coverage in the calibration data we are able to achieve improved predictions for all model grid boxes."

(7) Reviewer's comment: How do the authors choose the settings of their prior distributions for the hyperparameters in GP mean and covariance functions? This is not discussed anywhere.

Author's response: The hyperparameters for each emulator are estimated by maximum likelihood using the DiceKriging R package (Roustant et al., 2012). As Kennedy & O'Hagan (2001) point out, in order to integrate out the hyperparameters in the formulation of the GP emulator, we would require highly informative priors. In most cases, such informative priors do not exist. Hence, Kennedy & O'Hagan (2001) propose to provide a point estimate of the hyperparameters and to use these in the formulae for the mean and covariance functions of the GP emulator.

Changes to manuscript: To address this point, a new paragraph has been added to the end of section 2.5: "A final issue to resolve is how to estimate the roughness parameter since the posterior distribution of f(·) is conditional on these hyperparameters. A Bayesian approach would be to integrate out these hyperparameters in the formulation of the GP emulator. This would require highly informative priors, but in most cases such informative priors do not exist. Kennedy and O'Hagan (2001) propose using maximum likelihood to provide a point estimate of the hyperparameters and to use these in the formulae for the mean and covariance functions of the GP emulator. We adopt this approach in this study."

(8) Reviewer's comment: The authors often attribute the variation between irregularly distributed measurements and gridded output to representative error throughout the paper. This is unfair and inadequate, because model simulated output at gridded points does not imply model output is more representative at a broader scale, but simply because the model is too limited or too coarse to represent all the fine structures.

Author's response: The model output represents concentrations averaged over a coarse model grid square (not at a "gridded point"), while measured concentrations are indeed at specific points. There is a fundamental incommensurability in this comparison, given that the model cannot resolve fine structures, while the measurements may or may not sample them. To calibrate the model we need a measurement based assessment of concentrations at the model grid scale, ideally, but there are insufficient measurement sites per grid square to generate this. Our comparison is thus dependent on how well the available measurement sites represent the wider area, and we have defined this as the representation error in the paper. We note that this is a function of the grid scale considered as much as a property of the measurement due to its siting.

Changes to manuscript: We have rephrased our definition of representation error on page 5 in the introduction to make this clearer: "Firstly, global chemistry transport models typically have grid scales of the order of 100 km which is insufficient to resolve spatial variability in many atmospheric constituents. Surface measurements made at a single location may not be representative of the spatial scales resolved in the model." See also our response to the specific comments about this point, below.

Minor Comments

Reviewer's comment: P2, l5 and P6, l4, I understand that this is a standard statement to say poor spatial coverage of atmospheric composition measurements, but in your case the surface ozone measurements are very dense in Europe and most of the US, compared to the 2.8 degree resolution of their model output.

Author's response: While there are certainly more surface measurements in Europe and the US than elsewhere around the globe, the adequacy of the coverage must be judged on the spatial variability of the variable of interest, not the model resolution. While the lifetime of ozone is long in the free troposphere, the short timescales for chemical and dynamical processes controlling surface ozone drive much greater spatial variability. Given that the current network is still far from sufficient to capture this, we feel fully justified in describing the measurements as sparse.

Changes to manuscript: No changes were made.

Reviewer's comment: p3, l4, this reference should be updated to the most recent GBD 2019 study. 4.2m premature deaths are referring to the pm2.5 estimates. This study is about ozone, so the latest estimate of 365 thousand premature deaths is more appropriate.

Author's response: We agree that this would be more relevant here and thank the reviewer for pointing this out.

Changes to manuscript: We have updated the reference and amended the text to include this figure, as suggested.

Reviewer's comment: p3, l14-16, this statement seems unfair, since several attempts were already made for calibrating/emulating this type of model output. See following references Chang, K. L., & Guillas, S. (2019). Computer model calibration with large non-stationary spatial outputs: application to the calibration of a climate model. Journal of the Royal Statistical Society: Series C (Applied Statistics), 68(1), 51-78.Couvreux, F., Hourdin, F., Williamson, D., Roehrig, R., Volodina, V., Villefranque, N. et al. (2021). Process-based climate model development harnessing machine learning: I. a calibration tool for parameterization improvement. Journal of Advances in Modeling Earth Systems, 13(3), e2020MS002217. Guan, Y., Sampson, C., Tucker, J. D., Chang, W., Mondal, A., Haran, M., & Sulsky, D. (2019). Computer model calibration based on image warping metrics: an application for sea ice deformation. Journal of Agricultural, Biological and Environmental Statistics, 24(3), 444-463. Karagiannis, G., Konomi, B. A., & Lin, G. (2019). On the Bayesian calibration of expensive computer models with input dependent parameters. Spatial Statistics, 34, 100258. Salter, J. M., Williamson, D. B., Scinocca, J., & Kharin, V. (2019). Uncertainty quantification for computer models with spatial output using calibration-optimal bases. Journal of the American Statistical Association.

Author's response: The point we are making here refers specifically to atmospheric chemistry models, as stated in the text. However, we acknowledge that calibration has been successfully applied to climate models, as the reviewer highlights, and we have therefore modified the text to note this.

Changes to manuscript: We have updated text and cited some of the papers mentioned by the reviewer: "While estimation of model parameters is common in many fields of science, and has successfully been applied to climate models (e.g. Chang and Guillas, 2019; Couvreux et al., 2021), it is rarely attempted with atmospheric chemistry models because they are computationally expensive to run and it is thus burdensome to perform the large number of model runs required to explore model parameter space."

Reviewer's comment: P5, l5 and p22, l21 "require thousands of model runs" appear to be exaggerating, since the authors are aware that will be dependent on how many parameters to be turned (and they only use 80 runs).

Author's response: The point that we are making here is that global sensitivity analysis (e.g. extended FAST) and model calibration (e.g. MCMC) require thousands of model evaluations. This is true whether we use a computationally expensive model or a surrogate model. The process is currently only feasible with a surrogate model, and this is our motivation for using Gaussian Process emulation. Generating the emulator requires only a small number of executions of the expensive model (80 in our case), but we still need to carry out 1000s of runs with the emulator to conduct sensitivity analysis and model calibration. However, we acknowledge the reviewer's concern and have rephrased the first occurrence to avoid the appearance of exaggeration.

Changes to manuscript: We have updated the text to make the point clearer, noting specifically that sensitivity analysis and model calibration “may require thousands of model runs”.

Reviewer’s comment: P5, l10, “Since the first application of emulation methods for model calibration (Kennedy and O'Hagan, 2001), [...] In this study, we apply these approaches to models of tropospheric ozone for the first time to demonstrate the feasibility of parameter estimation.” This is inappropriate because the authors do not consider model discrepancy and measurement uncertainty in their emulator component (see my major comment 1).

Author’s response: As noted in our responses above, we have chosen to omit the discrepancy term for this demonstration of feasibility, as our exploratory tests showed that the effect is small, but measurement uncertainty is included as a component of representation error. We have applied the same emulation methods as Kennedy and O’Hagan, despite considering a very different system, and therefore we feel that it is fully appropriate to credit them with introducing it.

Changes to manuscript: We have implemented changes in response to major comment 1 by including a new section 2.8, which addresses this comment as well.

Reviewer’s comment: P5, l12, Higdon et al. (2008) should be cited, since this is the first paper successfully extending the calibration framework into the “highly multivariate output”. Higdon, D., Gattiker, J., Williams, B., & Rightley, M. (2008). Computer model calibration using high-dimensional output. Journal of the American Statistical Association, 103(482), 570-583.

Author’s response: Thank you for this reference; we now cite this here as suggested.

Changes to manuscript: This paper is now cited in the manuscript on page 5.

Reviewer’s comment: P5, l18, “Firstly, ground-level composition measurements are usually made at a single location which may not be representative of a wider region at the grid-scale of the model. Global chemistry transport models typically have a spatial scale of the order of 100 km.” This statement is somehow misleading, because dense ground based measurements (especially in Europe) reflect the local fine variations that can not be solved by coarse model resolution.

Author’s response: The reviewer is correct to point out that this is phrased somewhat awkwardly in the paper from the perspective of the model (the observations aren’t representative of the model grid square) rather than of the measurements (the model is unable to represent observed variations below the grid scale). However, even over Europe where there are more observations than in other parts of the world, measurement sites are insufficiently dense to fully characterise the spatial variability that would be needed to integrate them reliably to the model grid scale. To avoid any potential confusion, we have now rephrased this sentence in the manuscript.

Changes to manuscript: We have rephrased these sentences to read: “Firstly, global chemistry transport models typically have grid scales of the order of 100 km which is insufficient to resolve spatial variability in many atmospheric constituents. Surface measurements made at a single location may not be representative of the spatial scales resolved in the model."

Reviewer's comment: P8, l6, As I mentioned earlier, emissions are also dynamic in time and space. So the authors should comment if these parameters only represent the initial conditions.

Author's response: As noted in our response to the earlier comment, we are applying globally uniform scaling factors that do not vary in space and time. These are applied to the processes continuously, and are independent of the initial conditions. As noted above, the introductory text in Section 2.1 has been adjusted to make this clearer.

Changes to manuscript: As noted in our response to point 2 above.

Reviewer's comment: P8, l19, I believe the reference is Chang et al. (2017).

Author's response: Thank you, and apologies for getting the publication date wrong here!

Changes to manuscript: The reference has been corrected in the manuscript.

Reviewer's comment: P11, l23, There are a few alternative approaches, such as principal components (Higdon et al. 2008; Holden et al., 2015) or low rank approximations (Bayerri et al.,2007; Bowman and Woods 2016; Chang and Guillas, 2019), that are proposed to tackle high dimensional output.
Bayarri, M., Berger, J., Cafeo, J., Garcia-Donato, G., Liu, F., Palomo, J., Parthasarathy, R., Paulo, R., Sacks, J., and Walsh, D. (2007). Computer model validation with functional output. Ann. Statist., 35(5):1874–1906 Bowman, V. E. and Woods, D. C. (2016). Emulation of multivariate simulators using thin-plate splines with application to atmospheric dispersion. J. Uncertnty Quant., 4(1):1323–1344. Holden, P. B., Edwards, N. R., Garthwaite, P. H., and Wilkinson, R. D. (2015). Emulation and interpretation of high-dimensional climate model outputs. J. Appl. Statist., 42(9):2038–2055.

Author's response: In a previous paper using the same model, we implemented the principal component approach to carry out global sensitivity analysis (Ryan et al., 2018, cited in the manuscript). This gives a similar answer to building separate scalar-output emulators for each dimension of the output space, and is thus a useful way to reduce the dimensionality. We have considered multivariate outputs, but feel that unless the outputs of the emulator are being used as inputs for another emulator it is fine to use the scalar output emulator approach. However, we agree that it would useful to reference other approaches to dealing with multivariate output when implementing emulators, and have now done this.

Changes to manuscript: We have added references to other approaches of dealing with multivariate output when implementing emulators, including those suggested by the referee, in the final paragraph of section 2.6 quoted in the response to point (5) above.

Reviewer's comment: P12, l16, If "where B is a p × p matrix with zeros in the off diagonals" how the different input parameters can be correlated?

Author's response: As explained in section 2.5, B is a diagonal matrix where the elements are roughness parameters that describe the linearity of the input-output relationship. The matrix does not describe the relationship between the input terms as the reviewer suggests here. This is part of a standard explanation for the description of the Gaussian process emulator (e.g. see papers by Jeremey Oakley and Tony O'Hagan).

Changes to manuscript: We have added the phrase "roughness parameters" to the following sentence to try to make this clearer: "The roughness parameters give an indication of whether the input-output relationship for each input variable, given the training data, should be linear."

Reviewer's comment: P14, l1, The emulators used in this study are not taken into account measurement uncertainty or model discrepancy, so it merely represents the "output interpolator", I do not see why the authors should report R2.

Author's response: The uncertainty in the measurements is typically substantially less that that in their representativeness of the model grid scale, and it is thus effectively included as a small component of the representation error that we do fully consider. We do not include a model discrepancy term as explained above. We implement a Gaussian process emulator as described in O'Hagan (2006), which also quantifies the uncertainty at points in the output spaces where there are no training data. We therefore feel that it is appropriate to quote an R^2 for the comparison.
O'Hagan (2006) Bayesian analysis of computer code outputs: a tutorial. Reliability Engineering & System Safety, 91, 1290-1300

Changes to manuscript: The R^2 term has been retained in the manuscript.