Calibrating a global atmospheric chemistry transport model using Gaussian process emulation and ground-level concentrations of ozone and carbon monoxide

Edmund Ryan[1,2], Oliver Wild[1]*

[1]Lancaster Environment Centre, Lancaster University, UK.

[2]Now at: Corndel, London, UK.

*Corresponding author:

Lancaster Environment Centre,

Lancaster University,

Lancaster,

LA1 4YW

o.wild@lancaster.ac.uk

Keywords: atmospheric chemistry transport model, model calibration, surface ozone, measurement representativeness, Gaussian Process emulation, Markov Chain Monte Carlo, Just Another Gibbs Sampler.

For submission to: Geoscientific Model Development

Abstract

Atmospheric chemistry transport models are important tools to investigate the local, regional and global controls on atmospheric composition and air quality. To ensure that these models represent the atmosphere adequately it is important to compare their outputs with measurements. However, ground based measurements of atmospheric composition are typically sparsely distributed and representative of much smaller spatial scales than those resolved in models, and thus direct comparison incurs uncertainty. In this study, we investigate the feasibility of using observations of one or more atmospheric constituents to estimate parameters in chemistry transport models and to explore how these estimates and their uncertainties depend upon representation errors and the level of spatial coverage of the measurements. We apply Gaussian process emulation to explore the model parameter space and use monthly averaged ground-level concentrations of ozone (O_3) and carbon monoxide (CO) from across Europe and the US. Using synthetic observations we find that the estimates of parameters with greatest influence on O_3 and CO are unbiased, and the associated parameter uncertainties are low even at low spatial coverage or with high representation error. Using reanalysis data, we find that estimates of the most influential parameter - corresponding to the dry deposition process - are closer to its expected value using both O_3 and CO data than using O_3 alone. This is remarkable because it shows that while CO is largely unaffected by dry deposition, the additional constraints it provides are valuable for achieving unbiased estimates of the dry deposition parameter. In summary, these findings identify the level of spatial representation error and coverage needed to achieve good parameter estimates and highlight the benefits of using multiple constraints to calibrate atmospheric chemistry models.

Introduction

Changes in atmospheric composition due to human activities make an important contribution to Earth's changing climate (Stocker et al., 2013) and to outdoor air pollution, which is currently responsible for about 4.2 million deaths worldwide each year (Cohen et al., 2017), with 365,000 deaths due to surface ozone (DeLang et al., 2021). Chemistry transport models (CTMs) simulate the production, transport, and removal of key atmospheric constituents, and are important tools for understanding variations in atmospheric composition across space and time. They permit investigation of future climate and emission scenarios that fully account for the interactions and feedbacks that characterise physical, chemical and dynamical processes in the atmosphere. For practical application, CTMs need to reproduce the magnitude and variation in pollutant concentrations observed at a wide range of measurement locations. Where biases occur, these can often be reduced by improving process representation through adjusting model parameters so that the CTM matches the measurements to a sufficient level of accuracy (e.g. Menut et al., 2014). 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. Instead, data assimilation has become a standard method for ensuring that model states are consistent with measurements, usually treating model parameters as fixed (Khattatov et al., 2000, Bocquet et al., 2015, van Loon et al., 2000, Emili et al., 2014).

In this study, we explore computationally efficient ways of estimating parameters in chemistry transport models, focusing on two important tropospheric constituents, ozone (O_3) and carbon monoxide (CO). Ozone is a major pollutant that is produced in the troposphere by oxidation of precursors such as CO and hydrocarbons, which are emitted during combustion processes from vehicular, industrial and residential sources. Ozone is harmful to human health and has been shown to damage vegetation and reduce crop yields (Goldsmith and Landaw, 1968, Kampa and Castanas, 2008, Van Dingenen et al., 2009, van Zelm et al., 2008). A recent assessment of surface O_3 was carried out for the Tropospheric Ozone Assessment Report (TOAR) based on measurements from an extensive network of 10,000 sites around the world (Schultz et al., 2017). A simple statistical model of changes in surface O_3 between 2000 and 2014 showed that significant decreases of 28% and 6% have occurred in Eastern North America and Europe, respectively, but increases of 20% and 45% in south-east and east Asia (Chang et al., 2017). In recent decades, a similar pattern of decreases in CO in Europe and North America and increases over parts of Asia has also been observed (Granier et al., 2011). To fully explain and attribute these changes, a thorough understanding of the processes controlling these pollutants is needed.

To assess the performance of CTMs, it is essential to compare simulations of tropospheric chemical composition with measurements. A comprehensive evaluation of 15 global models found that they broadly matched measured O_3, but that modelled O_3 was biased high in the northern hemisphere and biased low in the southern hemisphere (Young et al., 2018). The models were unable to capture the long-term trends in tropospheric O_3 observed at different altitudes. Similar biases were found in an independent study of long-term trends involving three chemistry climate models (Parrish et al., 2014). While identification of these model biases is informative, correcting the deficiencies is challenging because it is often unclear why different models perform well at certain times and for certain places, but poorly elsewhere (Young et al., 2018). A practical solution is to perform global sensitivity analysis to identify the parameters or processes that influence the model results most and then to calibrate the model to estimate these parameters and their uncertainties by comparing model predictions with measurements in a statistically rigorous way. This provides insight into the physical processes causing model biases that is typically unavailable from simpler approaches.

The principal challenge with performing global sensitivity analysis and model calibration is that they may require thousands of model runs, and this is infeasible for a typical global CTM that may require 12-24 hours to simulate a year on high performance computing facilities. This can be overcome by replacing the model with a surrogate function such as a Gaussian process emulator that is computationally much faster to run (Johnson et al., 2018, Ryan et al., 2018, Lee et al., 2013). Sensitivity analysis and model calibration can then be performed based on thousands of runs with the emulator rather than the CTM. Since the first application of emulation methods for model calibration (Kennedy and O'Hagan, 2001), these approaches have been extended to models with highly multivariate output (Higdon et al., 2008). Examples include an earth system model (Wilkinson, 2010), an aerosol model (Johnson et al., 2015), an ice sheet model (Chang et al., 2016) and a climate model (Salter et al., 2018). In this study, we apply these approaches to models of tropospheric ozone for the first time to demonstrate the feasibility of parameter estimation.

We identify three issues that need to be addressed for successful atmospheric model calibration. 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. G~~global chemistry transport models typically have~~ a~~ ~~spatial~~ 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. These e~~E~~rrors associated with spatial representativeness may be important even for satellite measurements which provide information at a 10 km scale (Boersma et al., 2016, Schultz et al., 2017). This representation error is distinct from instrument error, which is often relatively narrow and better understood. The effect of representation errors was explored in simple terrestrial Carbon model by Hill et al. (2012), who found that as these errors decreased, the accuracy of parameter estimates improved.

Secondly, the spatial coverage of atmospheric composition measurements is typically relatively poor, and this limits our ability to estimate parameters accurately. Thus, it is important to explore how the spatial coverage of measurements affects estimates of model parameters and their associated uncertainties.

Thirdly, evaluation of atmospheric chemistry models is typically performed for different variables independently (e.g., Stevenson et al., 2006, Fiore et al., 2009). However, atmospheric constituents such as O_3, CO, NOx, and VOC are often closely coupled through interrelated chemical, physical and dynamical processes. Evaluation of a model with measurements of a single species neglects the additional process information available from accounting for species relationships. Lee et al. (2016) highlight the limitation of using a single observational constraint on modelled aerosol concentrations, finding that this resulted in reduced uncertainty in concentrations but not in the associated radiative forcing. The benefits of using multiple constraints have been highlighted previously. For example Miyazaki et al. (2012) used the Ensemble Kalman Filter and satellite measurements of NO_2, O_3, CO and HNO_3 to constrain a CTM, resulting in a significant reduction in model bias in NO_2 column, O_3 and CO concentrations simultaneously. Nicely et al. (2016) used aircraft measurements of O_3, H_2O and

NO to constrain a photochemical box model, and found estimates of column OH that were 12-40% higher than those from unconstrained CTMs. They also found that although the CTMs simulated O_3 well, they underestimated NOx by a factor of two, explaining the discrepancy in column OH.

To address these gaps in knowledge, we estimate the probability distributions of eight parameters from a CTM, given surface O_3 and CO concentrations from the USA and Europe. We focus on model calibration with a limited number of parameters as a proof of concept, but show how this could be expanded to a much wider range of parameters in future. To overcome the excessive computational burden of running the model a large number of times, we replace the model with a fast surrogate using Gaussian process emulation. After evaluation of the emulator to ensure that it is an accurate representation of the input-output relationship of the CTM, we investigate how well model parameters can be estimated from chemical measurement data. We quantify the impacts of measurement representation error and spatial coverage on the bias and uncertainty in the estimated model parameters and highlight the extent to which parameter estimates can be improved using measurements of different variables simultaneously.

2. Materials and methods

2.1 Atmospheric Chemical Transport Model

Chemistry transport models simulate the changes in concentration of a range of atmospheric constituents (e.g. O_3, CO, NO_x, CH_4) with time over a specified three-dimensional domain. They represent many of the physical and chemical processes involved, usually in a simplified form, but a detailed understanding is often incomplete. Key processes include the emission of trace gases into the atmosphere, photochemical reactions that result in chemical transformations, transport by the winds, convection and turbulence, and removal of trace gases from the atmosphere through deposition processes. In this study, we apply the Frontier Research System for Global Change version of the University of California, Irvine chemical transport model, the FRSGC/UCI CTM (Wild and Prather, 2000; Wild et al, 2004). We focus on eight important processes affecting tropospheric oxidants that were chosen based on one-at-a-time sensitivity studies with the model (Wild, 2007) and that have been used in previous global sensitivity analyses of tropospheric ozone burden and methane lifetime (Ryan et al., 2018; Wild et al., 2020). These processes include the surface emissions of nitrogen oxides (NOx), lightning emissions of NO, biogenic emissions of isoprene, wet and dry deposition of atmospheric constituents, atmospheric humidity, cloud optical depth and the efficiency of turbulent mixing in the boundary layer, see Table 1. These do not encompass all sources of uncertainty in the model, but are broadly representative of major uncertainties across a range of different processes. 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. ~~in each process, and~~ ~~t~~These scaling factors form the parameters that we aim to calibrate. The choice of parameters and uncertainty ranges are described in more detail in Wild et al. (2020). For this study, we focus on monthly-mean surface O_3 and CO distributions at the model native grid resolution of 2.8°×2.8° and compare with observations over North America and Europe for model calibration (Fig. 1). The model uses meteorological driving data for 2001, a relatively typical meteorological year without strong climate phenomena such as El Nino (Fiore et al. 2009).

2.2 Surface O_3 and CO data

Ground-based observations of O_3 are relatively abundant in Europe and North America, where there are ~1800 individual sites that have continuous long-term measurements of O_3 (~~Chang et~~

~~al., 2016~~Chang et al., 2017, Schultz et al., 2017). Measurements of CO are made at fewer locations, but reliable long-term data are available from 57 sites that are part of the Global Atmospheric Watch network (Schultz et al., 2015). To allow more thorough testing of the effects of spatial coverage over these regions, we use ~~model~~ CAMS interim reanalysis data of surface O_3 and CO from the European Centre for Medium-Range Weather Forecasts (ECMWF) which has been tuned to match measurements using 4D-Var data assimilation (Flemming et al. 2017). This reanalysis ~~data closely resembles~~reproduces observed O_3 and CO distributions relatively well, and biases at surface measurement stations are generally small (Huijnen et al., 2020). ~~where measurements are available and~~ The dataset also has the benefit of complete global coverage, allowing us to test the importance of measurement coverage directly.

Reanalysis data for O_3 and CO are available for 2003–2015, and we average the data by month across this period to provide a climatological comparison. The control run of the FRSGC/UCI model matches CO from the reanalysis data reasonably well (Fig. 2), but overestimates surface O_3. Overestimation of O_3 in continental regions has been noted in previous studies and is partly a consequence of rapid photochemical formation from fresh emissions that is magnified at coarse model resolution (Wild and Prather, 2006). For this exploratory study we bias-correct the modelled surface O_3 by reducing it by 25%, following the approach taken by Shindell et al. (2018), so that it matches the reanalysis data (Fig. 2a). This adjustment accounts for the effect of chemical processes and model resolution which are not explored in this study, and provides a firmer foundation for investigating the effects of other processes.

2.3 Representation error

The "representation error" describes how well measurements made at a single location represent a wider region at the spatial scale of the model (2.8°×2.8° for this study). The error may be reduced by averaging measurements made at different stations within a model grid box, although atmospheric measurements may be too sparse to permit this (Schultz, 2016). The representation error is sometimes taken as the mean of the spatial standard deviation of different measurements within a grid-box (Sofen et al. 2016). However, this measure quantifies the spatial variability of measured O_3 within a grid-box and may not match the representation error

To test the effect on parameter estimates of varying this representation error, we use synthetic data from the control run of the model using parameters set to their nominal default values. Synthetic O_3 and CO data were generated by adding different levels of representation error for each level of spatial coverage. In mathematical terms:

$$data_i = m_i(x_{control}) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_i^2) \tag{1}$$

where for the *i*th point in space or time, $data_i$ refers to the synthetic data for O_3 or CO, $m_i(x_{control})$ is the O_3 or CO from the model control run, and ε_i is generated from a Normal distribution with mean of zero and standard deviation σ_i that is directly proportional to the magnitude of $m_i(x_{control})$. In this case, $\sigma_i = p \times m_i(x_{control})$ where p is a scaling factor that provides a measure of the representation error. ~~scaling factor that we varied.~~ 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 ~~We included *p* as one of the parameters to estimate for the reanalysis data and found values in the range 0.16–0.19. Thus, when using the synthetic data we set the representation error scaling factor for these variables to at four level:~~ *p* (~~=~~0.01, 0.1, 0.2 and 0.3) to explore the importance of representation error when using the synthetic data~~. These four levels were chosen because when we used the reanalysis data to estimate the model parameters including *p*, we found that the posterior values of p were in the range 0.16-0.19.~~

2.4 Global sensitivity analysis

Sensitivity analysis was carried out to determine the sensitivity of the simulated surface O_3 and CO to changes in each of the eight parameters. This allows us to identify which of the parameters are most important in governing surface O_3 and CO. We use global sensitivity analysis (GSA), varying each input while averaging over the other inputs. This provides a more integrated assessment of uncertainty than the traditional one-at-a-time approach varying each input in turn while fixing the other inputs at nominal values. We use the extended FAST method (Saltelli et al., 1999), a common and robust approach to GSA in which the sensitivity indices are quantified by partitioning the total variance in the model output (i.e. modelled surface O_3 or CO) into different sources of contribution from each input. Like most sensitivity analysis methods, this approach requires several thousand executions of the model, which would be computationally expensive for the CTM used here. This is overcome by replacing the CTM with a Gaussian process (GP) emulator. Further details of the implementation of GSA are described in Ryan et al. et al. (2018).

2.5 Gaussian Process Emulation - theory

We replace the CTM with a surrogate model that maps the inputs of the CTM (the eight parameters listed in Table 1) with its outputs (surface O_3 and CO). We employ a surrogate model based on Gaussian process (GP) emulation for three reasons. Firstly, due to the attractive mathematical properties of a GP, the emulator needs very few runs of the computationally expensive model to train it, typically less than 100. This is ~~I~~in contrast to~~,~~ methods based on neural networks, which often have a large number of parameters that necessitate ~~can require~~ thousands of training runs. Secondly, a GP emulator is an interpolator and so predicts the output of the model with no uncertainty at the input points it is trained at. Thirdly, it gives a complete probability distribution, as a measure of uncertainty, for estimates of the model output at points it is not trained at.

A GP is an extension of the multivariate Gaussian distribution, where instead of a mean vector μ and covariance matrix $\mathbf{\Sigma}$, mean and covariance functions given by $E(f(x))$ and $\mathrm{cov}(f(x), f(x'))$ are used (Rasmussen, 2006). Here, $f(\cdot): \chi \in \mathbb{R}^q \to \mathbb{R}^{q'}$ represents the computationally expensive model and χ denotes the input space given by $\mathrm{x} = (x_1, \ldots, x_q) \in \chi_1 \times \ldots \times \chi_q = \chi \subset \mathbb{R}^q$, and q is the number of input variables. GP emulators within a Bayesian framework were first developed in the 1990s and early 2000s (O'Hagan, 2006, Oakley and O'Hagan, 2004, Kennedy and O'Hagan, 2000, Currin et al., 1991). The simplest and most common GP emulator is one where the outputs to be emulated are scalar. Thus, if the computationally expensive model is given by $f(\cdot)$, then the one-dimensional output y is calculated by $y = f(x)$. This means that if the model output is multidimensional – e.g. a global map or a time-series – then we need to build a separate emulator for each point in the output space. To build the emulator requires training runs from the expensive model. In general, we choose n training inputs, denoted by $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, based on a space filling design such as a Maximin Latin Hypercube design (Morris and Mitchell, 1995). The number of training points is based on the rule of thumb $n = 10 \times q$ (Loeppky et al., 2012).

Denoting the scalar outputs by $y_1 = f(\mathbf{x}_1)$, $y_2 = f(\mathbf{x}_2)$, …, $y_n = f(\mathbf{x}_n)$, we then build an emulator $\hat{f}(\cdot)$ given by $\hat{y} = \hat{f}(x)$, where $\hat{y}$ is the estimated output from the emulator. If x represents one of the training inputs (i.e. $x = \mathbf{x}_i, 1 \leq i \leq n$), then $\hat{y}$ is equal to the output from $f(\cdot)$ with no uncertainty (i.e. $\hat{y} = y$). If x represents an input the emulator is not trained at, then $\hat{y}$ has a probability distribution represented by a mean function $m(x)$ and a covariance function $V(x, x')$, where x' is a different input. The mean function is given by:

$$m(x) = h(x)^T\hat{\beta} + t(x)^T\mathbf{A}^{-1}(\mathbf{y} - H\hat{\beta}), \tag{2}$$

where $h(x)^T$ is a 1×(q+1) vector given by $(1,\ x^T)$, $\hat{\beta}$ is a vector of coefficients determined by $\hat{\beta} = (H^T\mathbf{A}^{-1}H)^{-1}H^T\mathbf{A}^{-1}\mathbf{y}$, $t(x)^T = \big(C(x, x_1; \psi), \dots, C(x, x_n; \psi)\big)$, and $\mathbf{A}$ is a matrix whose elements are determined by $\mathbf{A_{i,j}} = C(\mathbf{x_i}, \mathbf{x_j}; \psi)$, $\mathbf{y} = [f(\mathbf{x_1}), \dots, f(\mathbf{x_n})]^T$, $H = [h(\mathbf{x_1}), \dots, h(\mathbf{x}_n)]^T$. Here, $C(x, x'; \psi)$ is a correlation function that represents our prior belief about how the inputs x and x' are correlated. A common choice is a Gaussian correlation function which takes the form: $C(x, x'; \psi) = exp\big(-(x - x')^T\mathbf{B}(x - x')\big)$, where $\mathbf{B}$ is a $p \times p$ matrix with zeros in the off-diagonals and diagonal elements given by the roughness parameters $\psi = (\psi_1, \dots, \psi_q)$. The~~se~~ roughness parameters give an indication of whether the input-output relationship for each input variable, given the training data, should be linear. Low values reflect a linear (or smooth) relationship, whereas high values (e.g. > 20) suggest a non-linear (or non-smooth) response surface. For implementation purposes we express the correlation function as $C(x, x'; \psi) = \sum_{j=1}^{q+1} \exp\left(-\psi_j\big(x_j - x_j{}'\big)^2\right)$, where $x = (x_1, \dots, x_q)$ and $x' = (x_1{}', \dots, x_q{}')$. The formula for the covariance function $V(x, x')$ is given in appendix A.

A final ~~modelling~~ issue to resolve is how to estimate the roughness parameter since the posterior distribution of $f(\cdot)$ is conditional on these emulator parameters. A Bayesian approach would be to integrate out these emulator parameters in the formulation of the GP emulator. This would require highly informative priors, but in most cases such informative priors do not exist. ~~Hence,~~ Kennedy and~~&~~ O'Hagan (2001) propose using maximum likelihood to provide a point estimate of the emulator parameters and to use these in the formulae for the mean and covariance functions of the GP emulator. We adopt this approach in this study.

2.6 Gaussian Process Emulation - implementation

Using the Loeppky rule we choose $n = 80$ different training inputs for our eight-parameter calibration study. In total, we emulate two variables (surface O_3 and CO) over 12 months at 272 spatial locations, and so require 6528 different GP emulators. To estimate the model parameters we evaluate each of the GP emulators tens of thousands of times. Although emulation is computationally fast, this presents a substantial computational burden, even for more computationally efficient versions of the emulator (Marrel et al., 2011, Roustant et al., 2012). We overcome this by computing parts of equation (2) prior to these evaluations. Specifically, we compute the vectors $\hat{\beta}$, m_{LP} and ψ for all points in the output space, where m_{LP} denotes $\mathbf{A}^{-1}(\mathbf{y} - H\hat{\beta})$, the last part of *m(x)* from equation (2). We store these three objects as three matrices $\hat{\beta}_{ALL}$, $m_{LP.ALL}$ and ψ_{ALL}. Evaluated at a new input x_{new}, the mean function of the emulator (equation 1) can now be expressed as:

$$\begin{gathered} m_i(x_{new}) = h(x_{new})^T \hat{\beta}_{ALL}[i,:] + t_i(x_{new})^T m_{LP.ALL}[i,:], \\ t_i(x_{new})^T = \left(C(x_{new}, x_1; \psi_{ALL}[i,:]), \dots, C(x_{new}, x_n; \psi_{ALL}[i,:])\right), \end{gathered} \tag{3}$$

where i ($1 \leq i \leq 6528$) denotes the ith point in the output space. The equivalent formula for $V(x, x')$ is given in appendix A.

To the test the accuracy of GP emulation, we ran each of the 6528 emulators at 20 sets of parameters which were not used for training the emulators. The estimated O_3 and CO values from the emulators for all spatial locations and months closely match the simulated O_3 and CO output from the FRSGC/UCI model for these validation runs, with $R^2 > 0.995$ for each variable, see Fig. 3.

Finally, we recognise that ~~using~~ principal component analysis (PCA) could be used to reduce the dimensionality of the output space ~~is a viable option for reducing~~and hence the number of emulators required (Higdon et al., 2008). ~~For example,~~ In a previous study we found that ~~the~~a PCA-emulator hybrid approach resulted in similar performance compared to using separate emulators for each ~~dimension of~~point in the output space, and reduced the number of emulators required from 2000 to 40 or fewer~~despite the PCA-emulator approach requiring only 5-40 emulators in contrast to 2000 emulators for the emulator-only approach~~ (Ryan et al., 2018). However, ~~F~~for this study, we ~~opted for the~~choose an emulator-only approach because it is much simpler to -demonstrate~~code up particularly given slight complexity of reorganising the formula for the mean function so that certain parts could be evaluated prior to the calibration run~~. Nonetheless, future ~~MCMC based~~ emulation-calibration studies could ~~certainly~~ benefit from the ~~potential~~ computational savings ~~with~~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).

2.7 Parameter Estimation

We estimate the eight model parameters using Bayesian statistics via the software package Just Another Gibbs Sampler (Plummer, 2003). This uses Gibbs sampling, which is an approach based on ~~a~~ Markov Chain Monte Carlo (MCMC) ~~approach to~~ that we use to determine ~~sample from the~~ multi-dimensional posterior probability distribution of the model parameters (~~Berg, 2005~~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.

To find the posterior distribution, the MCMC algorithm searches the parameter space using multiple sets of independent chains. Here, a chain refers to a sequence of steps in the parameter space that the algorithm takes. A new proposed parameter set in this search is accepted on two conditions: (1) the set is consistent with the prior probability distribution, which for our study was a set of Uniform distributions with the lower and upper bounds given by the defined ranges in Table 1; and (2) the resulting modelled values using the proposed set of parameters are consistent with measurements, which is assessed using the following Gaussian likelihood function:

$$L(\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i}} exp\left(\frac{f_i(\theta) - m_i}{\sigma_i^2}\right)^2, \tag{4}$$

where N is the number of measurements used, $f_i(\theta)$ is the ith model output ($1 \leq i \leq N$) using the proposed parameter set θ, m_i is the measurement corresponding to the ith model output and σ_i is the representation error for measurement m_i. We ~~wish to~~ note that although separate emulators are used for each of the spatial and temporal locations in the model output ~~(due to the constraint that emulators are required to have univariate outputs)~~, there is still only ~~one~~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.

We ran three parallel chains for 10,000 iterations each. After discarding the first half of these iterations as 'burn in', we thinned the chains by a factor of five to reduce within-chain autocorrelation. Convergence was assessed using the Brooks-Gelman-Rubin diagnostic tool (Gelman et al., 2013). This produced 3000 independent samples from the posterior distribution for each parameter, which we summarize using their posterior means and 95% credible intervals (CIs) defined by the 2.5th and 97.5th percentiles (Gelman et al., 2013). We used the R language to code up our configuration of the MCMC algorithm.

2.8 Model discrepancy

It has been suggested that a model discrepancy term should be included ~~in~~ 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 ~~specifically~~ chosen not to include a discrepancy term for two reasons. ~~: (1) f~~Firstly, 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. ~~; (2) f~~Secondly, for ~~the~~ scenarios involving reanalysis data, there is no simple and defensible method to estimate the ~~discrepancy~~ term. ~~The discrepancy represents the missing processes in the model. However we often do not know what these missing processes are or how to estimate them.~~ When performing ~~a regular~~ model calibration ~~without an emulator (e.g.,~~by applying MCMC ~~on the original model)~~directly, ~~we would not include~~ a discrepancy term would not be included. Since the ~~main function~~purpose of the emulator here is to estimate the output of the ~~expensive~~ model for a given set of ~~values of the model~~ parameter values, ~~it could therefore be~~we argue~~d~~ that ~~there~~ it is not necessary to include a discrepancy term into the calibration formulation~~ when using an emulator~~. ~~Moreover, in Brynjarsdóttir & O'Hagan (2014), the authors state: *"The challenge with incorporating model discrepancy in statistical inverse problems is being confounded with calibration parameters, which will only be resolved with meaningful priors"*. However, Brynjarsdóttir & O'Hagan (2014) nor others do not address what to do one does not have highly informative priors, as is the case in study. We acknowledge~~ However, we agree that including ~~a discrepancy~~such a term may be helpful in ~~certain~~ situations where~~n~~ there is good prior information~~, but we also wish to highlight that Brynjarsdóttir & O'Hagan (2014) or Kennedy & O'Hagan (2001) do not give any practical instructions for how one might estimate this term~~.

~~In order~~ T~~t~~o investigate the importance of ~~this~~a discrepancy 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~~ repeat~~ed~~ the experiment to estimate the eight scaling parameters using surface ozone reanalysis data ~~at 2.5% spatial coverage~~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~~s in the formulation~~ compared with when we omit ~~them~~it (see Figure S16 in the supplemental material). ~~These results clearly demonstrate that including a discrepancy term results in almost no difference to the derived posterior distributions for the situation that we are considering here. Due to this result and the other arguments we have given, we have not included such a~~We therefore choose to omit the ~~discrepancy~~ term ~~as part of~~for our study.

2.~~8~~9 Experimental approach

We first perform a global sensitivity analysis to identify the parameters which have the greatest influence on the two variables we consider. We then perform parameter estimation using ~~measurement~~ surface concentration data over the regions of North America and Europe shown in Fig 1 and focus our analysis on the parameters which have the greatest influence. To provide a demonstration of the approach we first use "synthetic" measurement data drawn from the control run of the CTM which was not used to train the emulators, adding increasing levels of noise to represent measurement representation errors of 1, 10, 20 and 30% (p = 0.01, 0.1, 0.2 and 0.3), and varying the spatial coverage of these measurements over the regions considered over a wide range: 2.5, 5, 10, 20, 40 and 100%. We focus on surface O_3 only, surface CO only and then both variables together. We then use the reanalysis data to represent the measurements, focussing on the effects of spatial coverage alone, and estimating the representation error p from this independent dataset. The 90 different scenarios we consider are summarised in Table 2. 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.

3. Results

3.1 Global sensitivity analysis

Results from global sensitivity analysis reveal that over the continental regions of Europe and North America considered here, the simulated monthly mean concentrations of surface O_3 are most sensitive to dry deposition and, to a lesser extent, to isoprene emissions (Fig. 4). This is not unexpected, given the importance of direct deposition of ozone to the Earth's surface, and the role of isoprene as a natural source of ozone in continental regions. The simulated surface CO is most sensitive to isoprene emissions, which represent a source of CO, and to boundary layer mixing, which influences the transport of CO from polluted emission regions. We thus identify the scaling parameters corresponding to dry deposition, isoprene emissions and boundary layer mixing as the most important of the eight considered here to estimate accurately to reduce the bias in modelled surface O_3 and CO. For completeness, we show the geographical distribution of sensitivity indices in Figs 5 and 6, which reveal the importance of humidity in governing O_3 over oceanic regions and highlight the very different responses of surface O_3 and CO to the major driving processes.

3.2 Estimation of scaling parameters using synthetic data

We next use synthetic observation data to calibrate the model and estimate scaling parameters. For synthetic data we use the model control run with a specified level of representation error (Table 2), and the default model parameters define the true scaling that we aim to retrieve. Prescribing surface O_3 with very little error ($p = 0.01$) gives an estimate of the dry deposition scaling parameter, which has the largest influence on modelled surface O_3, close to its true value and the uncertainty is small even when the spatial coverage of measurements is only 2.5% (Fig. 7, column 1). As the representation error is increased to $p = 0.1$, the parameter uncertainty is larger at low spatial coverage but the mean estimate remains unbiased (Fig. 7, column 2). The uncertainty at all levels of spatial coverage becomes larger as p increases to 0.2 and 0.3, but the means remain very close to the true values (Fig. 7, columns 3 and 4). Surface CO is largely unaffected by dry deposition, and thus provides very little constraint on the scaling parameter. The effect of prescribing surface CO and O_3 together is very similar to that of using surface O_3 alone.

Using surface CO alone with very little representation error ($p = 0.01$), the mean estimate of the isoprene emission scaling parameter is equal to the true value with very little uncertainty, regardless of the spatial coverage (Fig. 8, column 1). When the representation error is increased to $p = 0.1$, the estimate remains very close to the true value, but the uncertainty is substantially higher at low spatial coverage (2.5% and 5%) than at higher coverage (40% and 100%) (Fig. 8, column 2). The estimates deviate further from the truth at higher levels of representation error ($p = 0.2$ and 0.3) and the uncertainty is greater (Fig. 8, columns 3 and 4). Estimates of the isoprene scaling parameter are less accurate than those of the dry deposition scaling parameter as the posterior means are further from the true value of the parameter and the uncertainty intervals are wider (Fig. 8 vs Fig. 7). As with our findings for dry deposition, the posterior means and the lengths of the uncertainty intervals for the isoprene scaling parameter remain relatively unchanged when surface O_3 data is prescribed at the same time.

Our findings for the boundary layer mixing scaling parameter follow a similar pattern to the other two parameters (Fig. 9). In all combinations of representation error and spatial coverage, we find that the mean estimates are unbiased. Furthermore, we find that the parameter uncertainty is significantly smaller when the spatial coverage is 10% or higher when p = 0.1, 20% or higher when p = 0.2, and 40% or higher when p = 0.3 (Fig. 9, Table 2). It is clear from these results that the scalings for these three model parameters can be successfully estimated from synthetic data with low uncertainty when the representation error is low, and that the estimates remain good, albeit with higher uncertainty, at higher representation error if the spatial coverage is relatively good.

3.3 Estimation of scaling parameters using reanalysis data

We consider next the CAMS interim reanalysis data for surface O_3 and CO which are based on assimilated concentrations from the ECMWF model and are thus independent of the FRSGC/UCI model. The reanalysis is representative of similar spatial scales to the FRSGC/UCI model, and thus we ignore the representation error and vary the spatial coverage only. However, we are able to estimate the representation error factor p by treating it as a parameter to estimate. With 100% spatial coverage, this error term is estimated with the MCMC algorithm to be p = 0.168 ± 0.004 and p = 0.191 ± 0.005 for surface O_3 and CO, respectively. Although we do not know the true values of the parameters in this case, the good agreement between the control run of the FRSGC/UCI model and the reanalysis data suggests that they lie close to their true values.

Using the reanalysis data for surface O_3 alone, we find that the posterior means and uncertainty for the dry deposition parameter are in the upper half of the range defined, indicating that the real dry deposition flux is greater than that calculated with the FRSGC/UCI model. This is largely as expected, as the FRSGC/UCI model overestimates surface O_3 at these continental sites and greater deposition would bring the model into better agreement with the reanalysis. As the spatial coverage is increased, the estimate of the scaling factor increases to around 1.4 and the uncertainty is reduced (Fig. 10a). In contrast, using surface O_3 and CO together results in an estimate closer to 1 and an additional reduction in uncertainty (Fig. 10g). Inclusion of surface CO measurements, as an additional constraint to surface O_3, results in an estimate of the dry deposition parameter closer to that modelled~~ along with a reduction in the associated uncertainty~~.

Using surface CO alone, estimates of the isoprene scaling parameter lie in the central part of the defined range, whilst estimates of the boundary layer mixing scaling parameter lie in the upper half of the defined range (Fig 10e,f). For both parameters, increasing the spatial coverage leads to a reduction in uncertainty. Unlike for dry deposition, inclusion of surface O_3 when estimating either of these parameters results in very little difference in the magnitude of the estimate or in the associated uncertainty (Fig. 10e vs 10h; Fig. 10f vs 10i).

3.4 Evaluation of surface O_3 following calibration

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.

4. Discussion

4.1 Representation error

Our results show the impact of the size of the representation error on the accuracy of estimated model parameters. The parametric uncertainty (i.e. the size of the credible intervals in Figs 7-9) increases at an approximately linear rate as the representation error increases from $p = 0.01$ to $p = 0.3$. This is consistent with Hill et al. (2012) who estimated the parameters and uncertainties of a simple terrestrial carbon model under varying levels of measurement error.

For the reanalysis data, we treat the representation error as a parameter for the MCMC algorithm to estimate along with the eight model parameters. This is possible because we assume that the measured value of O_3 is proportional to the simulated value from a forward run of the FRSGC/UCI model, although such an assumption may not be possible in other situations. An alternative approach to estimate the representation error would be to carry out an intensive measurement campaign to determine whether the average O_3 from different measuring stations within a grid-square is representative of the true average. Satellite products of the terrestrial biosphere are checked for accuracy using this type of approach (De Kauwe et al., 2011). Although measurement campaigns at these large spatial and temporal scales would be challenging and costly, they may not be need to continue for long periods of time since we might expect representation error to decrease as the temporal scale increases (Schutgens et al., 2016).

4.2 Spatial coverage

We find that as the volume of measurements increase, the estimates of the model parameters are closer to the truth and the width of the credible intervals decrease. This is particularly clear for the dry deposition and isoprene emission scaling parameters when using both O_3 and CO concentrations (Figs 8 and 9). While this highlights the value of good spatial coverage, we note that the benefits are greatly reduced if the representation error is relatively high. For the boundary layer mixing parameter, we find little decrease in the credible intervals using synthetic

CO data with the highest representation error ($p = 0.3$), where the spatial coverage is less than 20% (Fig. 9, row 2). In contrast, at the $p = 0.1$ level, a large decrease in uncertainty is seen between the 2.5% and 20% coverage. Similar effects are seen, to a lesser extent, for the dry deposition and isoprene scaling parameters as the spatial coverage increases.

Our results using synthetic data show that while the size of the uncertainty intervals vary substantially depending on the spatial coverage or representation error, the posterior means are for the most part very close to the true values. Deviation from these typically occurs when the measurements contain less information either due to low spatial coverage or high representation error. However, the uncertainty intervals include the true values of the parameters for all the experimental scenarios considered here, unlike in Hill et al. (2012). This gives strong confidence in the reliability of the MCMC method used to estimate the parameters.

4.3 Applying multiple constraints

The importance of multiple constraints was most apparent for scenarios involving the reanalysis data. For the dry deposition scaling parameter, which explains much of the variance in surface O_3 (Fig. 4), we ~~iund~~ found that using O_3 data alone results in mean estimates that are in the upper half of the range of possible values (Fig 10a). However, including CO data brought the mean estimates into the central part of the range where we would expect the true value to lie (Fig. 10g). This is remarkable given that dry deposition is not an important process for controlling CO, and highlights the coupling between processes that permits constraints on one process from one variable to influence those on another. However, it is consistent with previous studies exploring the uncertainty in estimates of key parameters in an aerosol-chemistry-climate model (Johnson et al., 2018). For the isoprene emission and boundary layer mixing scaling parameters, there was little difference in the mean estimates or the size of the uncertainty intervals when using O_3 and CO together rather than a single constraint. This reveals that the importance of using multiple constraints is dependent on the process and on the variable constrained. A judicious choice of these could allow a particular process to be targeted. 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. 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.

4.4 Towards constraint with real surface measurements

Our results have demonstrated the feasibility of using measurement data to constrain model parameters under the right conditions. We have chosen to use synthetic data as they have allowed us to vary the spatial coverage and to investigate the effects of representation error which is poorly characterised when using real measurements data. Quantifying this type of error for real measurements is difficult because measurement sites are relatively sparse and are often representative of a limited area rather than the larger area typical of a model grid-square. However, this study has allowed us to estimate the representation error associated with the reanalysis data, and in the absence of more information these values could be used as a guide when applying surface measurements as a constraint.

The reanalysis data provide a more critical test, as they are independent of the FRSGC/UCI CTM used here. Although we do not know the true values of the scaling parameters, we expect them to lie close to those used in the control run given the relatively good agreement for O_3 and CO concentrations. For the dry deposition parameter, we expect scaling values to be close to 1, but using surface O_3 reanalysis data alone we found posterior mean scaling parameters approaching 1.4, with credible intervals that did not include 1 (Fig. 10a). This likely reflects overestimation of surface O_3 in continental regions in the CTM and may reflect uncertainties and biases in other processes not considered here, most notably in the chemical formation and destruction of O_3 and in model transport processes. In the absence of consideration of the uncertainty in these processes in this feasibility study, the dry deposition parameter is used as a proxy process to reduce O_3 concentrations. This is an example of equifinality, where different sets of parameters can result in model predictions that give equally good agreement with observations (Beven et al., 2001). Applying simultaneous constraints to CO goes some way to addressing this, but does not remove the problem. Before applying real surface measurements to constrain the CTM, we propose a more comprehensive assessment of model uncertainties with a wider range of parameters so that the constraints can more directly inform process understanding and model development.

Conclusion

We have demonstrated the use of surface O_3 and CO concentrations to constrain a global atmospheric chemical transport model and generate accurate and robust estimates of model parameters. This would normally be prohibitive for such a model given that thousands of model runs are required. Our approach is to replace the CTM with a surrogate model using Gaussian process emulation and then estimate the parameters using the emulator in place of the CTM. In this feasibility study we have shown that surface O_3 has a large sensitivity to dry deposition, and that surface CO is most sensitive to isoprene emissions and boundary layer mixing processes, as expected. We find that estimates of the scaling parameters for these processes are dependent on the spatial coverage and representation error of the surface O_3 and CO data. Our parameter estimates become less uncertain as coverage increases and as the representation error decreases, whilst remaining unbiased. Furthermore, we show that using two separate data constraints, in this case surface O_3 and CO, instead of a single one can result in mean parameter estimates that are much closer to their likely true values. However, this is dependent on the processes considered and constraints applied, and while it is effective for dry deposition here, we find relatively little improvement in the estimates or uncertainties for isoprene emission or boundary layer mixing processes that are also considered here.

The approach we adopt here provides a means of constraining atmospheric models with observations and identifying sources of model error at a process level. Our results based on the independent reanalysis data suggest that dry deposition and isoprene emissions are represented relatively well in the FRSGC/UCI CTM but that boundary layer mixing processes may be somewhat underestimated. However, we have explored the effect of only eight parameters in this study and consideration of a more complete set of processes, including those governing photochemistry and dynamics, is needed to generate more realistic constraints for key pollutants such as O_3. We aim to expand this study to investigate a more extensive range of parameters and processes and to constrain with a wider range of observation data. The emulator-based approach for estimating parameters that we have successfully demonstrated here can be applied to any model where evaluating the model the required number of times is too computationally demanding.

Code and data availability

The R code used for building and validating the emulators and estimating the posterior distribution of the model parameters using the Markov Chain Monte Carlo algorithm is available from the Zenodo data repository via the link: https://zenodo.org/record/4537614. The

FRSGC/UCI model output used for training the emulators is available from the CEDA data repository via the link: https://catalogue.ceda.ac.uk/uuid/d5afa10e50b44229b079c7c5a036e660.

Appendix A

The formula for the covariance function $V(x, x')$ from §2.2 is given by:

$$V(x, x') = \sigma^2[C(x, x'; \psi) - t(x)^T \mathbf{A}^{-1} t(x) \\ + (h(x)^T + t(x)^T \mathbf{A}^{-1} H)(H^{\mathrm{T}} \mathbf{A}^{-1} H)^{-1}(h(x')^T + t(x')^T \mathbf{A}^{-1} H)^T]$$

where,

$$\sigma^2 = \frac{\mathbf{y}^{\mathrm{T}}(\mathbf{A}^{-1} - \mathbf{A}^{-1} H (H^{\mathrm{T}} \mathbf{A}^{-1} H)^{-1} H^{\mathrm{T}} \mathbf{A}^{-1})\mathbf{y}}{n - q - 1}$$

To compute the variance or uncertainty of a prediction x we use the formula for $V(x, x')$ with $x' = x$, which results in $C(x, x; \psi) = 1$. Since we need to evaluate a large number of emulators for each MCMC iteration step (because we have a separate emulator for every dimension of the model output), it is more computationally efficient to compute the parts of the above formula prior to using the emulator. Hence, the above formula can be replaced with:

$$V_i(x_{new}, x_{new}) = \sigma^2_{ALL}[i, 1]\big[\big(1 - t_i(x_{new})^T V_{i,1} t_i(x_{new}) \\ + \big(h(x_{new})^T + t(x_{new})^T V_{i,2}\big) V_{i,3} \big(h(x_{new})^T + t(x_{new})^T V_{i,2}\big)^T\big]$$

where:

- i ($1 \leq i \leq r$) denoted the ith point in the r-dimensional simulator output.
- σ^2_{ALL} is a r×1 vector that stores the values of σ^2 for all r outputs.
- $V_{i,1}$ is the $n \times n$ matrix $\mathbf{A}^{-1}$ corresponding to the ith point in the simulator's output. It is stored as the ith block of the $nr \times n$ matrix V_1 defined by:

$$V_1 = \begin{pmatrix} V_{1,1} \\ V_{2,1} \\ \vdots \\ V_{r,1} \end{pmatrix}$$

- $V_{i,2}$ is the $n \times q$ matrix $\mathbf{A}^{-1}H$ corresponding to the ith point in the simulator's output. It is stored as the ith block of the $nr \times q$ matrix V_2 defined by:

$$V_2 = \begin{pmatrix} V_{1,2} \\ V_{2,2} \\ \vdots \\ V_{r,2} \end{pmatrix}$$

- $V_{i,3}$ is the $q \times q$ matrix $(H^{\mathrm{T}}\mathbf{A}^{-1}H)^{-1}$ corresponding to the ith point in the simulator's output. It is stored as the ith block of the $qr \times q$ matrix V_3 defined by:

$$V_3 = \begin{pmatrix} V_{1,3} \\ V_{2,3} \\ \vdots \\ V_{r,3} \end{pmatrix}$$

Author contributions

ER and OW designed the study. ER carried out the statistical analyses, and OW ran the FRSGC/UCI model and provided the outputs that were used to train and validate the emulators. ER wrote the paper with input from OW.

Acknowledgements

This work was supported by the Natural Environment Research Council [grant number NE/N003411/1]. We thank Karl Hennermann at the ECMWF for making CAMS interim reanalysis data for O_3 and CO ~~from CAMS~~ available. We also thank Lindsay Lee from Sheffield Hallam University for her feedback and comments on an early version of this paper.

References

BARET, F., WEISS, M., ALLARD, D., GARRIGUES, S., LEROY, M., JEANJEAN, H., FERNANDES, R., MYNENI, R., PRIVETTE, J. & MORISETTE, J. 2005. VALERI: a network of sites and a methodology for the validation of medium spatial resolution land satellite products. *Remote Sensing of Environment,* 76**,** 36-39.

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

BERG, B. A. 2005. Introduction to Markov chain Monte Carlo simulations and their statistical analysis. *Markov Chain Monte Carlo Lect Notes Ser Inst Math Sci Natl Univ Singap*, *7*, 1-52.

BEVEN, K. and FREER, J., 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *Journal of hydrology*, *249*(1-4), pp.11-29.

BOCQUET, M., ELBERN, H., ESKES, H., HIRTL, M., ŽABKAR, R., CARMICHAEL, G., FLEMMING, J., INNESS, A., PAGOWSKI, M. & PÉREZ CAMAÑO, J. 2015. Data assimilation in atmospheric chemistry models: current status and future prospects for coupled chemistry meteorology models. *Atmospheric Chemistry and Physics,* 15**,** 5325-5358.

Brynjarsdóttir, J., & O'Hagan, A. (2014). Learning about physical parameters: The importance of model discrepancy. *Inverse problems*, *30*(11), 114007.

BOERSMA, K., VINKEN, G. & ESKES, H. 2016. Representativeness errors in comparing chemistry transport and chemistry climate models with satellite UV–Vis tropospheric column retrievals. *Geoscientific model development,* 9**,** 875.

CHANG, K.-L., PETROPAVLOVSKIKH, I., COOPER, O. R., SCHULTZ, M. G. & WANG, T. 2017. Regional trend analysis of surface ozone observations from monitoring networks in eastern North America, Europe and East Asia. *Elem Sci Anth,* 5.

CHANG, W., HARAN, M., APPLEGATE, P. & POLLARD, D. 2016. Calibrating an ice sheet model using high-dimensional binary spatial data. *Journal of the American Statistical Association,* 111**,** 57-72.

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.

COHEN, A. J., BRAUER, M., BURNETT, R., ANDERSON, H. R., FROSTAD, J., ESTEP, K., ... & FEIGIN, V. 2017. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. *The Lancet*, *389*(10082), 1907-1918.

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.

CURRIN, C., MITCHELL, T., MORRIS, M. & YLVISAKER, D. 1991. Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *Journal of the American Statistical Association,* 86**,** 953-963.

DE KAUWE, M. G., DISNEY, M., QUAIFE, T., LEWIS, P. & WILLIAMS, M. 2011. An assessment of the MODIS collection 5 leaf area index product for a region of mixed coniferous forest. *Remote Sensing of Environment,* 115**,** 767-780.

Delang, M. N., et al., 2021, Mapping Yearly Fine Resolution Global Surface Ozone through the Bayesian Maximum Entropy Data Fusion of Observations and Model Output for 1990–2017, Environ. Sci. Technol. 2021, 55, 8, 4389–4398, doi:10.1021/acs.est.0c07742.

EMILI, E., BARRET, B., MASSART, S., LE FLOCHMOEN, E., PIACENTINI, A., EL AMRAOUI, L., PANNEKOUCKE, O. & CARIOLLE, D. 2014. Combined assimilation of IASI and MLS observations to constrain tropospheric and stratospheric ozone in a global chemical transport model. *Atmospheric Chemistry and Physics,* 14**,** 177-198.

FIORE, A. M., DENTENER, F., WILD, O., CUVELIER, C., SCHULTZ, M., HESS, P., TEXTOR, C., SCHULZ, M., DOHERTY, R. & HOROWITZ, L. 2009. Multimodel estimates of intercontinental source-receptor relationships for ozone pollution. *Journal of Geophysical Research: Atmospheres,* 114.

FLEMMING, J., BENEDETTI, A., INNESS, A., ENGELEN, R. J., JONES, L., HUIJNEN, V., REMY, S., PARRINGTON, M., SUTTIE, M. & BOZZO, A. 2017. The CAMS interim Reanalysis of Carbon Monoxide, Ozone and Aerosol for 2003-2015. *Atmospheric Chemistry and Physics,* 17**,** 1945.

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. & RUBIN, D. B. 2013. *Bayesian data analysis*, CRC press.

GOLDSMITH, J. R. & LANDAW, S. A. 1968. Carbon monoxide and human health. *Science,* 162**,** 1352-1359.

GRANIER, C., BESSAGNET, B., BOND, T., D'ANGIOLA, A., VAN DER GON, H. D., FROST, G. J., HEIL, A., KAISER, J. W., KINNE, S. & KLIMONT, Z. 2011. Evolution of anthropogenic and biomass burning emissions of air pollutants at global and regional scales during the 1980–2010 period. *Climatic Change,* 109**,** 163.

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.

HILL, T. C., RYAN, E. & WILLIAMS, M. 2012. The use of CO2 flux time series for parameter and carbon stock estimation in carbon cycle research. *Global Change Biology,* 18**,** 179-193.

Huijnen, V., Miyazaki, K., Flemming, J., Inness, A., Sekiya, T., and Schultz, M. G.: An intercomparison of tropospheric ozone reanalysis products from CAMS, CAMS interim, TCR-1, and TCR-2, *Geosci. Model Dev.*, 13, 1513–1544, https://doi.org/10.5194/gmd-13-1513-2020, 2020.

JOHNSON, J. S., REGAYRE, L. A., YOSHIOKA, M., PRINGLE, K. J., LEE, L. A., SEXTON, D. M., ROSTRON, J. W., BOOTH, B. B. & CARSLAW, K. S. 2018. The importance of comprehensive parameter sampling and multiple observations for robust constraint of aerosol radiative forcing. *Atmospheric Chemistry and Physics,* 18**,** 13031-13053.

JOHNSON, J.S., CUI, Z., LEE, L.A., GOSLING, J.P., BLYTH, A.M. and CARSLAW, K.S., 2015. Evaluating uncertainty in convective cloud microphysics using statistical emulation. *Journal of Advances in Modeling Earth Systems*, *7*(1), pp.162-187.

KAMPA, M. & CASTANAS, E. 2008. Human health effects of air pollution. *Environmental pollution,* 151**,** 362-367.

KENNEDY, M. C. & O'HAGAN, A. 2000. Predicting the output from a complex computer code when fast approximations are available. *Biometrika,* 87**,** 1-13.
KENNEDY, M. C. & O'HAGAN, A. 2001. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 63**,** 425-464.
KHATTATOV, B. V., LAMARQUE, J. F., LYJAK, L. V., MENARD, R., LEVELT, P., TIE, X., BRASSEUR, G. P. & GILLE, J. C. 2000. Assimilation of satellite observations of long-lived chemical species in global chemistry transport models. *Journal of Geophysical Research: Atmospheres,* 105**,** 29135-29144.
LANDRIGAN, P. J., FULLER, R., ACOSTA, N. J., ADEYI, O., ARNOLD, R., BALDÉ, A. B., BERTOLLINI, R., BOSE-O'REILLY, S., BOUFFORD, J. I. & BREYSSE, P. N. 2017. The Lancet Commission on pollution and health. *The Lancet*.
LEE, L., PRINGLE, K., REDDINGTON, C., MANN, G., STIER, P., SPRACKLEN, D., PIERCE, J. & CARSLAW, K. 2013. The magnitude and causes of uncertainty in global model simulations of cloud condensation nuclei. *Atmos. Chem. Phys,* 13**,** 8879-8914.
LEE, L. A., REDDINGTON, C. L. & CARSLAW, K. S. 2016. On the relationship between aerosol model uncertainty and radiative forcing uncertainty. *Proceedings of the National Academy of Sciences,* 113**,** 5820-5827.
LOEPPKY, J. L., SACKS, J. & WELCH, W. J. 2012. Choosing the sample size of a computer experiment: A practical guide. *Technometrics*.
MALLEY, C. S., HENZE, D. K., KUYLENSTIERNA, J. C., VALLACK, H. W., DAVILA, Y., ANENBERG, S. C., TURNER, M. C. & ASHMORE, M. R. 2017. Updated global estimates of respiratory mortality in adults>/= 30Years of age attributable to long-term ozone exposure. *Environmental Health Perspectives, 2017, vol. 125, num. 8, p. 087021*.
MARREL, A., IOOSS, B., JULLIEN, M., LAURENT, B. & VOLKOVA, E. 2011. Global sensitivity analysis for models with spatially dependent outputs. *Environmetrics,* 22**,** 383-397.
MENUT, L., BESSAGNET, B., KHVOROSTYANOV, D., BEEKMANN, M., BLOND, N., COLETTE, A., COLL, I., CURCI, G., FORET, G. & HODZIC, A. 2014. CHIMERE 2013: a model for regional atmospheric composition modelling. *Geoscientific model development,* 6**,** 981-1028.
MIYAZAKI, K., ESKES, H., SUDO, K., TAKIGAWA, M., VAN WEELE, M. & BOERSMA, K. 2012. Simultaneous assimilation of satellite NO2, O3, CO, and HNO3 data for the analysis of tropospheric chemical composition and emissions. *Atmos. Chem. Phys,* 12**,** 9545-9579.
MORRIS, M. D. & MITCHELL, T. J. 1995. Exploratory designs for computational experiments. *Journal of statistical planning and inference,* 43**,** 381-402.
NICELY, J. M., ANDERSON, D. C., CANTY, T. P., SALAWITCH, R. J., WOLFE, G. M., APEL, E. C., ARNOLD, S. R., ATLAS, E. L., BLAKE, N. J. & BRESCH, J. F. 2016. An observationally constrained evaluation of the oxidative capacity in the tropical western Pacific troposphere. *Journal of Geophysical Research: Atmospheres,* 121**,** 7461-7488.
O'HAGAN, A. 2006. Bayesian analysis of computer code outputs: a tutorial. *Reliability Engineering & System Safety,* 91**,** 1290-1300.
OAKLEY, J. E. & O'HAGAN, A. 2004. Probabilistic sensitivity analysis of complex models: a Bayesian approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 66**,** 751-769.
PARRISH, D., LAMARQUE, J. F., NAIK, V., HOROWITZ, L., SHINDELL, D., STAEHELIN, J., DERWENT, R., COOPER, O., TANIMOTO, H. & VOLZ-THOMAS, A. 2014. Long-term changes in lower

tropospheric baseline ozone concentrations: Comparing chemistry-climate models and observations at northern midlatitudes. *Journal of Geophysical Research: Atmospheres,* 119**,** 5719-5736.
PLUMMER, M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. Proceedings of the 3rd international workshop on distributed statistical computing, 2003. Technische Universit at Wien, 125.
RASMUSSEN, C. E. 2006. Gaussian processes for machine learning.
RICHARDSON, A. D., WILLIAMS, M., HOLLINGER, D. Y., MOORE, D. J., DAIL, D. B., DAVIDSON, E. A., SCOTT, N. A., EVANS, R. S., HUGHES, H. & LEE, J. T. 2010. Estimating parameters of a forest ecosystem C model with measurements of stocks and fluxes as joint constraints. *Oecologia,* 164**,** 25-40.
ROUSTANT, O., GINSBOURGER, D. & DEVILLE, Y. 2012. DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization.
RYAN, E., WILD, O., VOUGARAKIS, A. & LEE, L. 2018. Fast sensitivity analysis methods for computationally expensive models with multi-dimensional output. *Geoscientific model development,* 11**,** 3131-3146.
SALTELLI, A., TARANTOLA, S. & CHAN, K.-S. 1999. A quantitative model-independent method for global sensitivity analysis of model output. *Technometrics,* 41**,** 39-56.
SALTER, J. M., WILLIAMSON, D. B., SCINOCCA, J. & KHARIN, V. 2018. Uncertainty quantification for spatio-temporal computer models with calibration-optimal bases. *arXiv preprint arXiv:1801.08184*.
SCHULTZ, M. G. 2016. Cluster analysis of European surface ozone observations for evaluation of MACC reanalysis data. *Atmospheric Chemistry and Physics,* 16**,** 6863.
SCHULTZ, M. G., AKIMOTO, H., BOTTENHEIM, J., BUCHMANN, B., GALBALLY, I. E., GILGE, S., HELMIG, D., KOIDE, H., LEWIS, A. C. & NOVELLI, P. C. 2015. The Global Atmosphere Watch reactive gases measurement network. *Elem Sci Anth,* 3.
SCHULTZ, M. G., SCHRÖDER, S., LYAPINA, O., COOPER, O., GALBALLY, I., PETROPAVLOVSKIKH, I., VON SCHNEIDEMESSER, E., TANIMOTO, H., ELSHORBANY, Y. & NAJA, M. 2017. Tropospheric Ozone Assessment Report: Database and metrics data of global surface ozone observations. *Elem Sci Anth,* 5.
SCHUTGENS, N. A., GRYSPEERDT, E., WEIGUM, N., TSYRO, S., GOTO, D., SCHULZ, M. & STIER, P. 2016. Will a perfect model agree with perfect observations? The impact of spatial sampling. *Atmospheric Chemistry and Physics,* 16**,** 6335-6353.
SHINDELL, D., FALUVEGI, G., SELTZER, K. and SHINDELL, C., 2018. Quantified, localized health benefits of accelerated carbon dioxide emissions reductions. *Nature climate change*, *8*(4), pp.291-295.
SOFEN, E., BOWDALO, D., EVANS, M., APADULA, F., BONASONI, P., CUPEIRO, M., ELLUL, R., GALBALLY, I., GIRGZDIENE, R. & LUPPO, S. 2016. Gridded global surface ozone metrics for atmospheric chemistry model evaluation. *Earth System Science Data,* 8**,** 41.
STEVENSON, D., DENTENER, F., SCHULTZ, M., ELLINGSEN, K., VAN NOIJE, T., WILD, O., ZENG, G., AMANN, M., ATHERTON, C. & BELL, N. 2006. Multimodel ensemble simulations of present-day and near-future tropospheric ozone. *Journal of Geophysical Research: Atmospheres,* 111.

STOCKER, T. F., QIN, D., PLATTER, G. K., TIGNOR, M., ALLEN, S. K., BOSCHUNG, J., ... & MIDGLEY, P. M. 2013. Climate change 2013: The physical science basis. *Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change*, *1535*.

VAN DINGENEN, R., DENTENER, F. J., RAES, F., KROL, M. C., EMBERSON, L. & COFALA, J. 2009. The global impact of ozone on agricultural crop yields under current and future air quality legislation. *Atmospheric Environment,* 43**,** 604-618.

VAN LOON, M., BUILTJES, P. J. & SEGERS, A. 2000. Data assimilation of ozone in the atmospheric transport chemistry model LOTOS. *Environmental Modelling & Software,* 15**,** 603-609.

VAN ZELM, R., HUIJBREGTS, M. A., DEN HOLLANDER, H. A., VAN JAARSVELD, H. A., SAUTER, F. J., STRUIJS, J., VAN WIJNEN, H. J. & VAN DE MEENT, D. 2008. European characterization factors for human health damage of PM10 and ozone in life cycle impact assessment. *Atmospheric Environment,* 42**,** 441-453.

WILD, O. 2007. Modelling the global tropospheric ozone budget: exploring the variability in current models. *Atmospheric Chemistry and Physics Discussions,*7 (1), pp.1995- 2035. ffhal-00302576f

WILD, O. and PRATHER, M.J., 2006. Global tropospheric ozone modeling: Quantifying errors due to grid resolution. *Journal of Geophysical Research: Atmospheres*, *111*(D11).

WILD, O., Voulgarakis, A., O'Connor, F., Lamarque, J. F., Ryan, E., & Lee, L. 2020. Global sensitivity analysis of chemistry-climate model budgets of tropospheric ozone and OH: Exploring model diversity. *Atmospheric Chemistry and Physics*, *20*, 4047-4058.

WILKINSON, R. D. 2010. Bayesian calibration of expensive multivariate computer experiments. John Wiley & Sons.

WILLIAMS, M., RICHARDSON, A., REICHSTEIN, M., STOY, P., PEYLIN, P., VERBEECK, H., CARVALHAIS, N., JUNG, M., HOLLINGER, D. & KATTGE, J. 2009. Improving land surface models with FLUXNET data. *Biogeosciences,* 6**,** 1341-1359.

YOUNG, P. J., NAIK, V., FIORE, A. M., GAUDEL, A., GUO, J., LIN, M., NEU, J., PARRISH, D., RIEDER, H. & SCHNELL, J. 2018. Tropospheric Ozone Assessment Report: Assessment of global-scale model performance for global and regional ozone distributions, variability, and trends. *Elem Sci Anth,* 6.

Figures and Tables for

'Calibrating a global atmospheric chemistry transport model using Gaussian process emulation and ground-level concentrations of ozone and carbon monoxide'

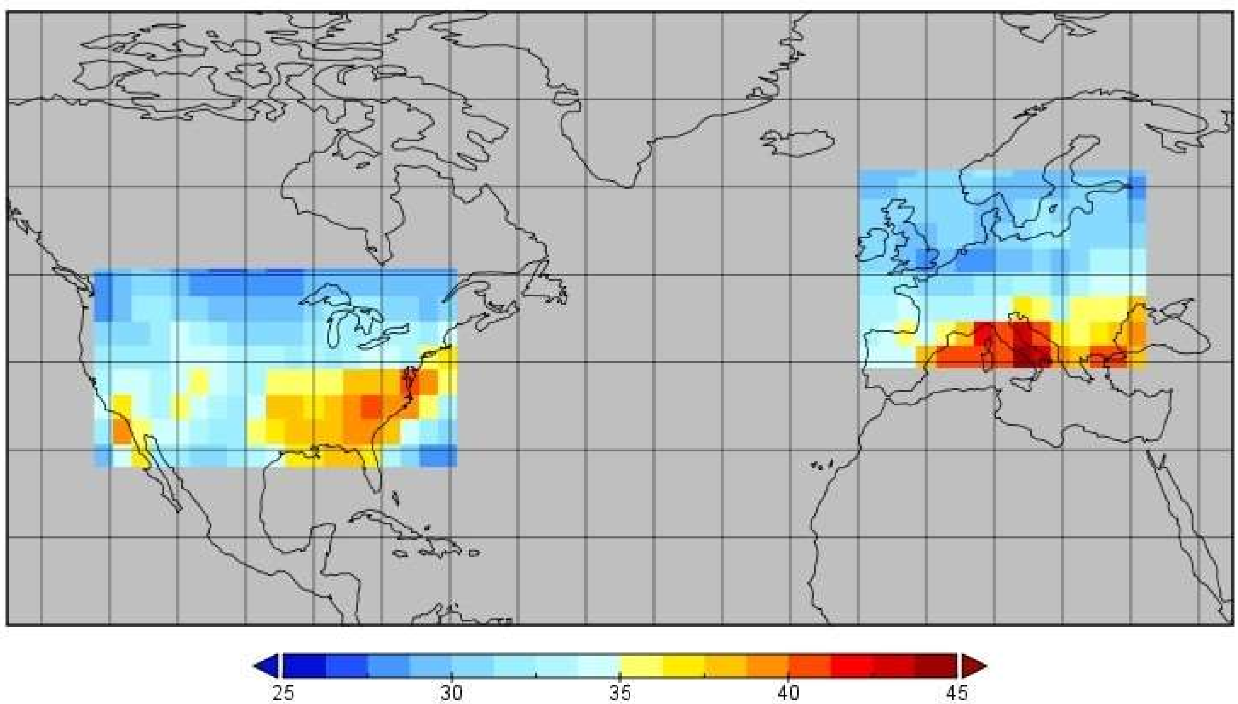

Figure 1. Annual mean surface ozone mixing ratio (in ppb) from the FRSGC/UCI CTM showing the regions considered here and the 272 grid cells used for model calibration.

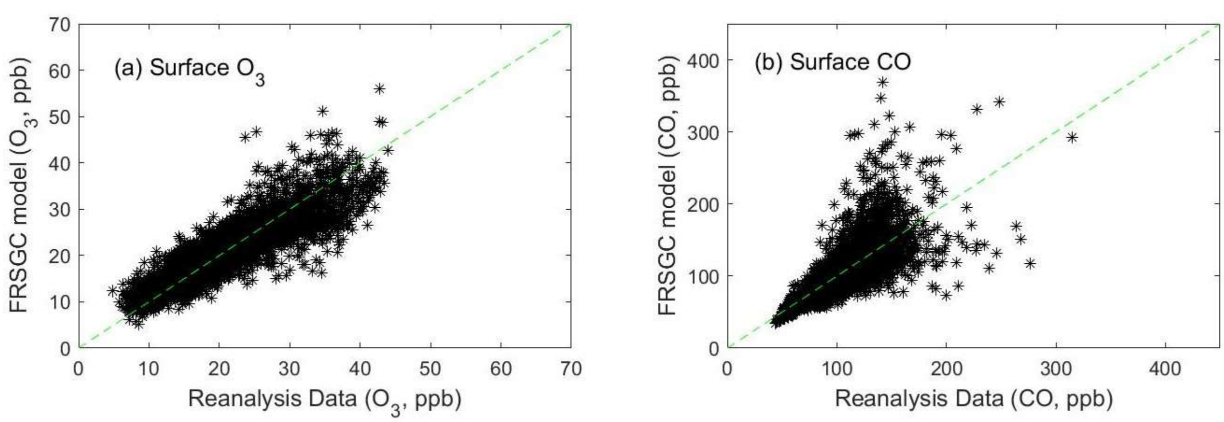

Figure 2. Monthly mean surface O_3 (panel a) and surface CO (panel b) over Europe and North America simulated with the FRSGC/UCI CTM compared with ECMWF reanalysis data.

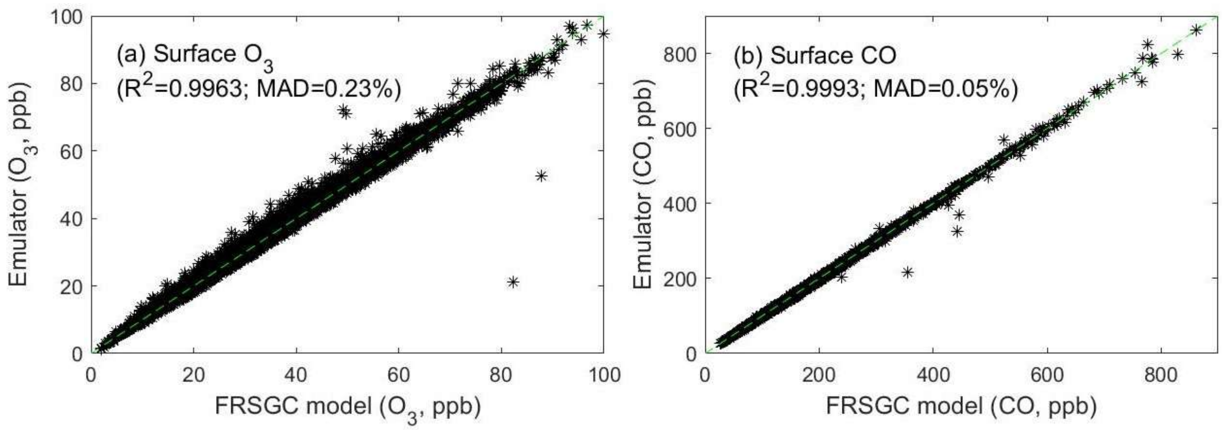

Figure 3. Simulated surface O_3 (panel a) and surface CO (panel b) from the FRSGC/UCI CTM versus those predicted from the Gaussian process emulators. The simulated and emulated concentrations were generated using 20 sets of model parameters that were not used for training the emulators.

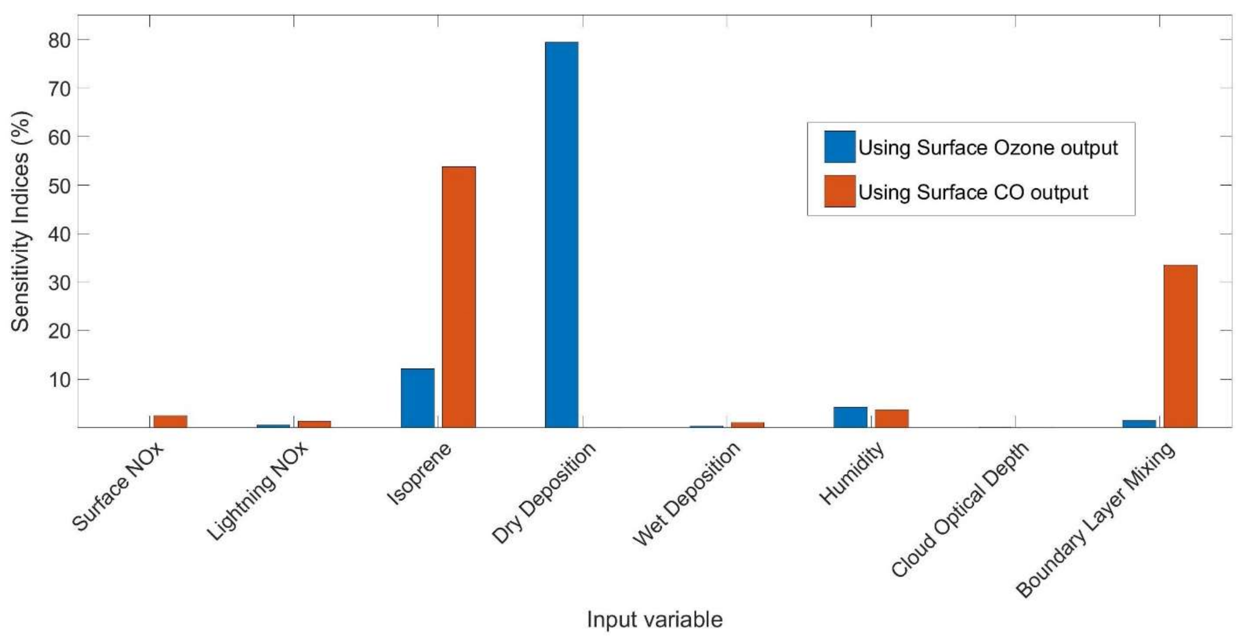

Figure 4. Sensitivity indices representing the percentage of the variance in surface O_3 and CO over the USA and Europe in the FRSGC/UCI model output due to changes in each parameter.

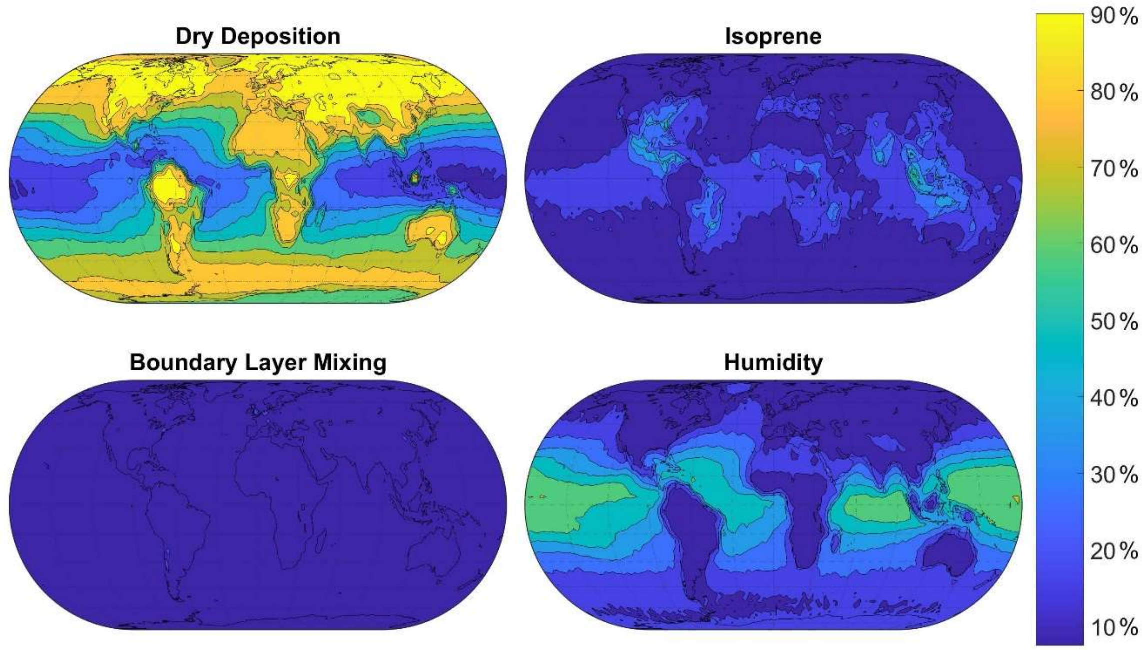

Figure 5. Sensitivity indices representing the percentage of the variance in surface O_3 in the FRSGC/UCI model output due to changes in each input parameter. The four parameters displayed here have the highest sensitivity indices and the largest effect on simulated surface O_3. Maps of sensitivity indices corresponding to the other four parameters are shown in Figure S2 of the supplementary material.

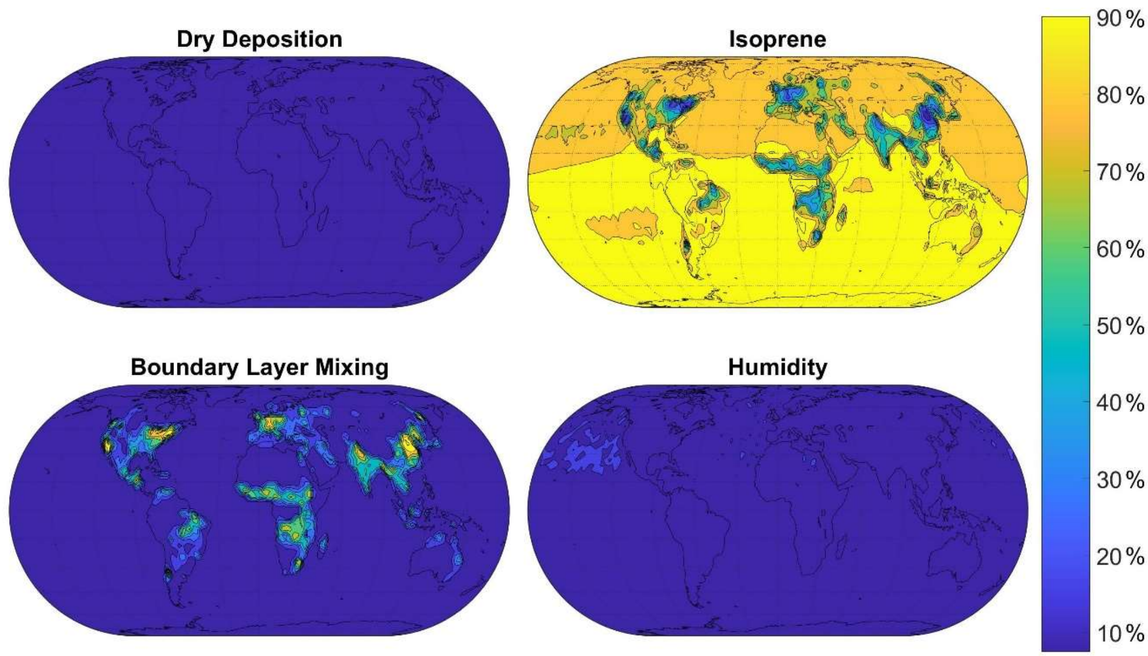

Figure 6. Sensitivity indices representing the percentage of the variance in surface CO in the FRSGC/UCI model output due to changes in each input parameter. Maps of sensitivity indices for the other four parameters are shown in Figure S3 of the supplementary material.

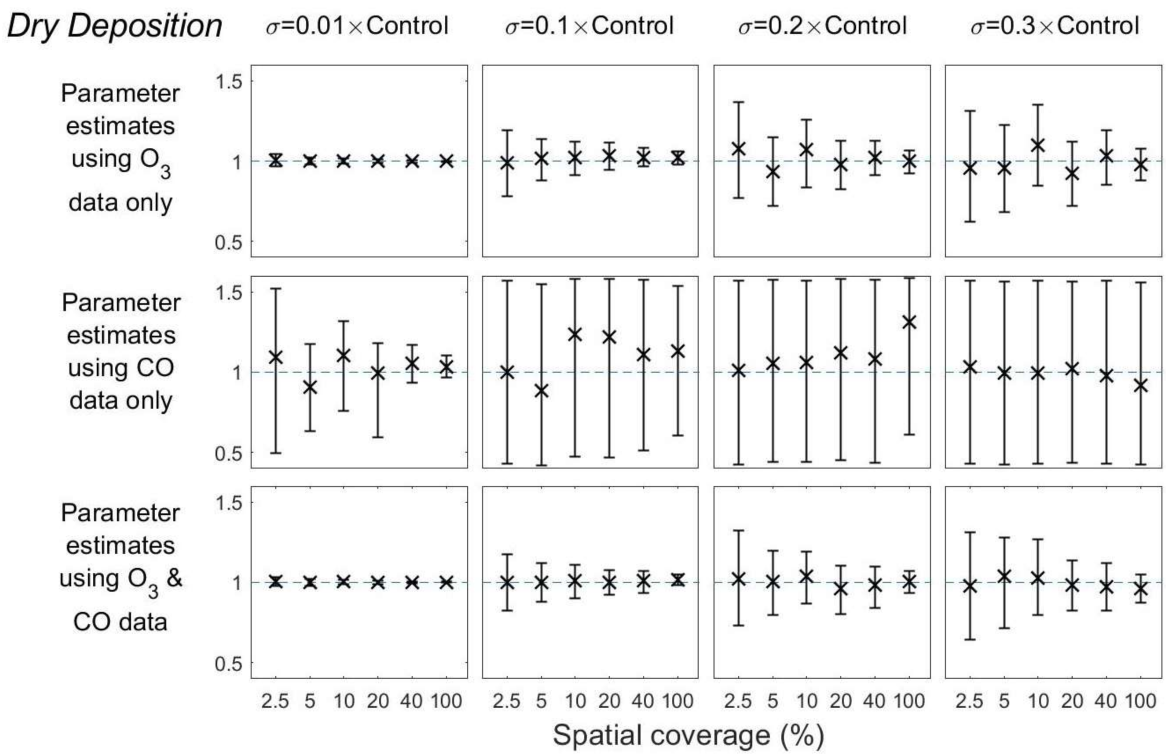

Figure 7. Means and 95% credible intervals of 3000 samples of the **Dry Deposition** scaling parameter from posterior distributions using the MCMC algorithm based on synthetic datasets from scenarios 1-72 (table 1). *Control* refers to the FRSGC/UCI model control run surface concentration for each output point.

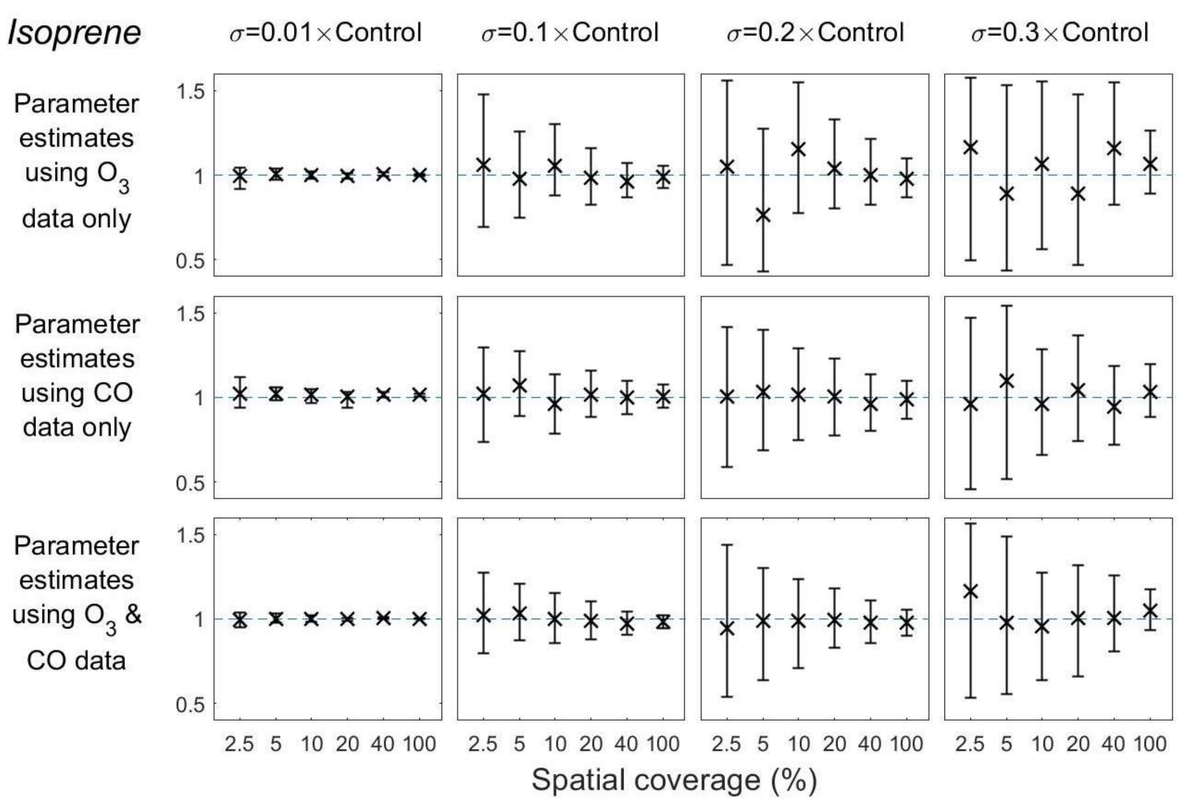

Figure 8. Means and 95% credible intervals of 3000 samples of the **Isoprene** emission scaling parameter from posterior distributions using the MCMC algorithm based on synthetic datasets from scenarios 1-72 (table 1). *Control* refers to the FRSGC/UCI model control run surface concentration for each output point.

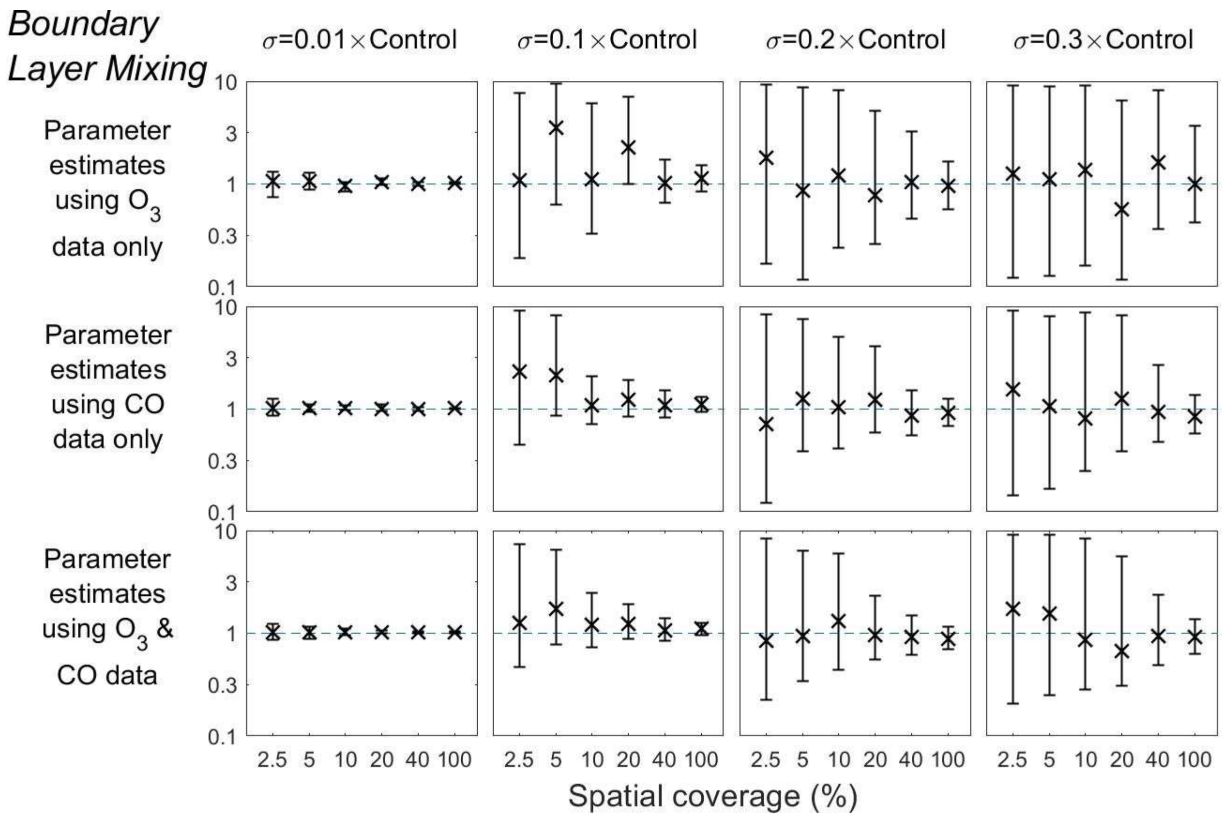

Figure 9. Means and 95% credible intervals of 3000 samples of the **Boundary Layer Mixing** scaling parameter from posterior distributions using the MCMC algorithm based on synthetic datasets from scenarios 1-72 (table 1). *Control* refers to the FRSGC/UCI model control run surface concentration at each output point. The scaling parameter values are given here on ~~the~~ a $\log_{10}$ scale.

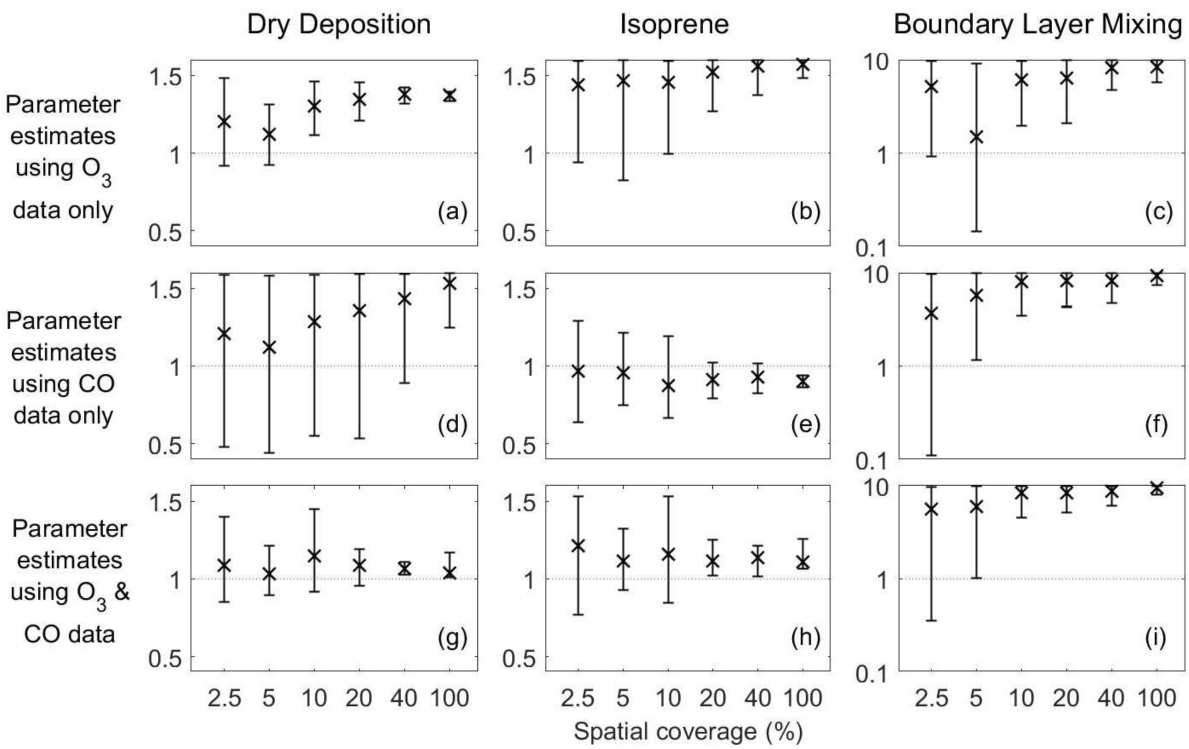

Figure 10. Means and 95% credible intervals of 3000 samples of the Dry Deposition, Isoprene and Boundary Layer Mixing scaling parameters from posterior distributions using the MCMC algorithm based on reanalysis datasets from scenarios 73-90 (table 1). The first and second rows show these parameters estimated using one stream of data (O_3 for the first row and CO for the second row), while the third row shows estimates using two data streams (O_3 and CO).

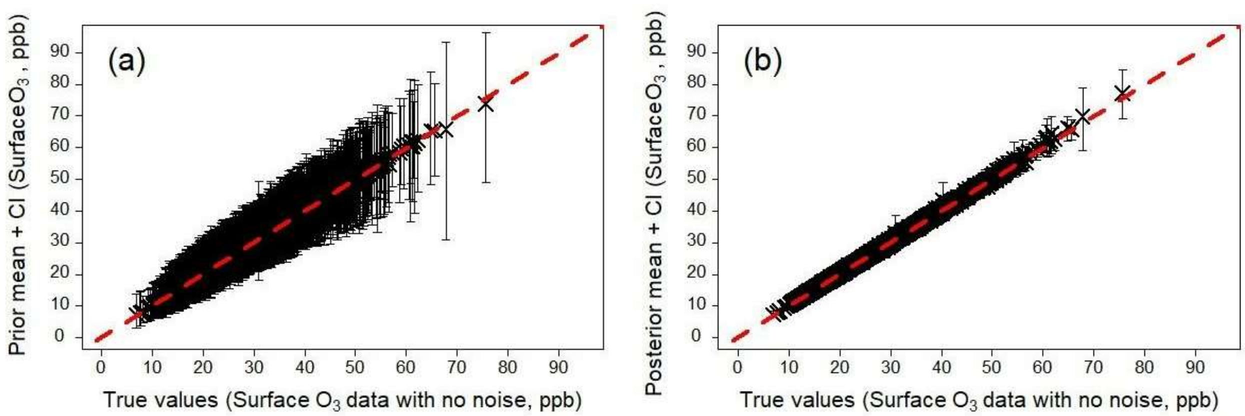

Figure 11. Emulator predictions of surface O_3, evaluated at values of the scaling parameters sampled from the prior distribution (panel a) and posterior distribution (panel b) showing the effects of calibration. In panel b, the outputs correspond to the scenario where the calibration involved synthetic O_3 data, a representation error of $p = 0.2$ and a spatial coverage of 20% (table 2). The predictions shown here are carried out for all model grid boxes, i.e. 100% spatial coverage.

Table 1. Model processes and associated scaling parameter ranges used in this study.

Number	Model process	Control run value	Scaling parameter values
1	Global surface NOx emissions (TgN/year)	40	0.75 – 1.25
2	Global lightning NO emissions (TgN/year)	5	0.40 – 1.60
3	Global isoprene emissions (TgC/year)	500	0.40 – 1.60
4	Dry deposition rates	model value	0.40 – 1.60
5	Wet deposition rates	model value	0.40 – 1.60
6	Humidity	model value	0.80 – 1.20
7	Cloud optical depth	model value	0.33 – 3.00
8	Boundary Layer mixing	model value	0.10 – 10.0

Table 2. Summary of the 90 different MCMC scenarios carried out for this study. The scenarios involved varying: (i) the type of data (synthetic or reanalysis); (ii) the representation error used for the synthetic data (p) where $m_i(x_{control})$ is the control run output of the CTM and σ_i is the amount of statistical noise added; (iii) the percentage coverage of grid-squares in the USA and Europe. For the synthetic data the 24 scenarios correspond to a full factorial combination of four levels of representation error and six levels of spatial coverage, while for the reanalysis data the six scenarios correspond to the six levels of spatial coverage.

Scenarios	**Dataset**	**Representation error, p** $(\sigma_i = p \times m_i(x_{control}))$	**Spatial coverage**
1-24	Synthetic O_3	0.01, 0.1, 0.2, 0.3	2.5%, 5%, 10%, 20%, 40%, 100%
25-48	Synthetic CO	0.01, 0.1, 0.2, 0.3	2.5%, 5%, 10%, 20%, 40%, 100%
49-72	Synthetic O_3 & CO	0.01, 0.1, 0.2, 0.3	2.5%, 5%, 10%, 20%, 40%, 100%
73-78	Reanalysis data (O_3)	Parameter to be estimated	2.5%, 5%, 10%, 20%, 40%, 100%
79-84	Reanalysis data (CO)	Parameter to be estimated	2.5%, 5%, 10%, 20%, 40%, 100%
85-90	Reanalysis data (O_3 & CO)	Parameter to be estimated	2.5%, 5%, 10%, 20%, 40%, 100%