Environmental Data Science (2023), xx: 1–11
doi:10.1017/xxxx

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Statistical constraints on climate model parameters using a scalable cloud-based inference framework

James Carzon[1*], Bruno Abreu[2], Leighton Regayre[3,4], Kenneth Carslaw[3], Lucia Deaconu[5,6], Philip Stier[5], Hamish Gordon[7,8] and Mikael Kuusela[1,9]

[1]Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
[2]National Center for Supercomputing Applications, University of Illinois Urbana-Champaign, Urbana-Champaign, Illinois, USA
[3]Institute for Climate and Atmospheric Science, School of Earth and Environment, University of Leeds, Leeds, UK
[4]Met Office Hadley Centre, Exeter, UK
[5]Atmospheric, Oceanic and Planetary Physics Department, University of Oxford, Oxford, UK
[6]Faculty of Environmental Science and Engineering, Babes-Bolyai University, Cluj, Romania
[7]Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
[8]Center for Atmospheric Particle Studies, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
[9]NSF AI Planning Institute for Data-Driven Discovery in Physics, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
*Corresponding author. Email: jcarzon@andrew.cmu.edu

Received: 5 April 2023

Keywords: UKESM1; perturbed parameter ensemble; Gaussian process emulator; inverse problem; strict bounds; model discrepancy

Abstract

Atmospheric aerosols influence the Earth's climate, primarily by affecting cloud formation and scattering visible radiation. However, aerosol-related physical processes in climate simulations are highly uncertain. Constraining these processes could help improve model-based climate predictions. We propose a scalable statistical framework for constraining parameters in expensive climate models by comparing model outputs with observations. Using the C3.ai Suite, a cloud computing platform, we use a perturbed parameter ensemble of the UKESM1 climate model to efficiently train a surrogate model. A method for estimating a data-driven model discrepancy term is described. The strict bounds method is applied to quantify parametric uncertainty in a principled way. We demonstrate the scalability of this framework with two weeks' worth of simulated aerosol optical depth data over the South Atlantic and Central African region, written from the model every three hours and matched in time to twice-daily MODIS satellite observations. When constraining the model using real satellite observations, we establish constraints on combinations of two model parameters using much higher time-resolution outputs from the climate model than previous studies. This result suggests that, within the limits imposed by an imperfect climate model, potentially very powerful constraints may be achieved when our framework is scaled to the analysis of more observations and for longer time periods.

Impact Statement

Atmospheric aerosols influence the amount of solar radiation reflected by Earth, but the magnitude of the effect is highly uncertain, and this is one of the key reasons why climate predictions are highly uncertain. We propose a framework for reducing uncertainty in aerosol effects on radiation by comparing simulations from complex climate models to satellite observations. This framework uses parallel computing and statistical theory to ensure efficient computations and valid inferences.

1. Introduction

Atmospheric aerosols affect the formation of clouds and scatter and absorb visible radiation, thereby influencing Earth's climate. Improved estimates of the change in the total effect of aerosols on the climate over the industrial era – a highly uncertain quantity termed the aerosol *effective radiative forcing* (ERF) – have the potential to reduce uncertainty in the sensitivity of the climate to these aerosols. Recent efforts to constrain ERF have involved first reducing uncertainty in the distributions of more basic aerosol-related physical parameters and then studying the effects of these constraints on ERF. This has been notably done by comparing simulated data with observations collected on various global aircraft and ship campaigns as well as from satellites and ground stations (Johnson et al, 2020, 2018; Regayre et al, 2020, 2018). Toward constraining the aerosol radiative forcing, Regayre et al (2023) employs simulated outputs from the UKESM1 climate model that are averaged over month-long periods. In contrast, the present work employs three-hourly simulation outputs from the same model, requiring an inferential framework capable of handling this increase in the resolution and quantity of the data.

Establishing constraints on the input parameters of expensive computer models by comparing their outputs with observational data is an area of active research (Biegler et al, 2010). It is often imperative that a surrogate model be trained from an ensemble of model input-output pairs and used in place of the simulator to ensure tractable computations, but this step contributes an additional source of uncertainty. If the outputs, or surrogate outputs, do not match observations within some tolerance, then those parameter values are deemed implausible. In Johnson et al (2020); Regayre et al (2020), the method of *history matching* is used, a technique from oil reservoir engineering which has been adapted to the evaluation of computer models more generally in recent decades (Verly et al, 1984; Craig et al, 1997; Johansen, 2008; Bower et al, 2010). However, the aim of history matching is to constrain parameter spaces and not necessarily to provide well-understood probabilistic guarantees on those constraints.

In contrast with previous work on constraining climate model parameters, our framework draws on a recent surge of interest in simulator-based inference (Dalmasso et al, 2023; Cranmer et al, 2020; Schafer and Stark, 2009) to produce parameter constraints that provide rigorous statistical guarantees of frequentist coverage. Specifically, our work deals with a special case of simulator-based inference where the observations are given by a deterministic simulator and an additive noise model. Patil et al (2022); Stanley et al (2022) use a *strict bounds* method (Stark, 1992) to construct efficient confidence sets for the model parameters in closely related inverse problems in remote sensing and high energy physics. Unlike in these works where the forward models of interest are linear and known exactly, the present problem features a forward model (UKESM1) which is nonlinear and estimated using an emulator. We take advantage of the strict bounds method while inverting the emulated forward model numerically and accounting for emulation uncertainty.

Our framework also offers a novel means of accounting for the systematic disagreement, known as the *model discrepancy*, between a simulator and the physical system which it purports to model. A number of approaches to accounting for model discrepancy in computer model calibration or simulator-based inference have been developed (McNeall et al, 2016; Higdon et al, 2008; Kennedy and O'Hagan, 2001). We propose a new data-driven procedure for incorporating model discrepancy (and other sources of error that we cannot separately quantify, such as representation errors; Schutgens et al (2017)) into the strict bounds inversion framework. Cloud-based computing resources are leveraged to make each step in the framework computationally scalable.

1.1. Data sources

Aerosol optical depth (AOD) is a measure for how much aerosol there is in the atmosphere. It is measured by *MODerate-resolution Imaging Spectroradiometer* (MODIS) which is found on board

Table 1. *Seventeen of the 37 UKESM1 parameters (Regayre et al, 2023) used to build the surrogate model, selected based on relevance for predicting AOD. N_d is the cloud droplet number concentration. SF is short for scale factor*

Parameter name	Physical name	Min.	Max.
sea_spray	Sea spray emission flux SF	0.25	4.00
bl_nuc	Boundary layer nucleation rate SF	0.1	10
ait_width	Aitken mode width (nm)	1.2	1.8
cloud_ph	Cloud droplet pH	4.6	7
prim_so4_diam	Median diameter of primary ultrafine anthropogenic sulfate particles (nm)	3	100
anth_so2_r	Anthropogenic SO_2 emissions flux SF outside of Europe, Asia and North America	0.6	1.5
bvoc_soa	Biogenic secondary organic aerosol from volatile organic compounds SF	0.32	3.68
dms	Dimethyl sulfide emission flux SF	0.33	3.0
dry_dep_ait	Aitken mode aerosol dry deposition velocity SF	0.5	2.0
dry_dep_acc	Accumulation mode aerosol dry deposition velocity SF	0.1	10.0
dry_dep_so2	SO_2 dry deposition velocity SF	0.2	5.0
bc_ri	Imaginary part of the black carbon refractive index	0.2	0.8
a_ent_1_rp	Cloud top entrainment rate SF	0	0.5
autoconv_exp_nd	Exponent of N_d in power law for initiating autoconversion of cloud drops to rain drops	-3	-1
dbsdtbs_turb_0	Cloud erosion rate (s^{-1})	0	0.001
bparam	Coefficient of the spectral shape parameter (beta) for effective radius	-0.15	-0.13
carb_bb_diam	Carbonaceous biomass burning primary particle median diameter (nm)	90	300

Table 2. *A notational reference table.*

Notation	Meaning	Notation	Meaning
u	Vector of parameter values from $\mathbb{R}^p$	z	AOD observations
x	Location in space and time of a measurement	ζ	True climate system
$\mathcal{M}_{\text{sim}}$	`SimulatedGrid` (as in Figure 1)	η	Climate model
$\mathcal{M}_{\text{sat}}$	`SatelliteGrid`	$\tilde{\eta}$	Emulator for the climate model
$\mathcal{M}$	`MatchedGrid`	D_{train}	Triplets $(u, x, \eta(x, u))$
$\mathcal{M}^*$	`MatchedGrid`, excluding outliers	D_{test}	Tuples $(u, x, \mathbb{E}[\tilde{\eta}_x(u) \mid D_{\text{train}}], \text{Var}[\tilde{\eta}_x(u) \mid D_{\text{train}}])$

the NASA-launched Terra and Aqua satellites that offer near-global coverage twice daily and provide easily readable open-access data sets. In the flow chart in Figure 1, this data set is the `SatelliteTimeseries`. For the present application, we focus on MODIS retrievals from the South Atlantic and Central African region over July 1–14, 2017. This domain and period of study is selected for its known biomass burning-related atmospheric aerosol activity.

We compare these data with climate model outputs taken from the UKESM1 model (Sellar et al, 2019). We use simulations that were nudged to the observed meteorology following the method of Telford et al (2008), and therefore the simulated weather conditions will be sufficiently realistic that we can examine the ability of the model to represent the observed aerosols. We use the perturbed parameter ensemble (PPE) of 221 atmosphere-only simulations documented by Regayre et al (2023), which have a configuration that closely matches the atmosphere component of the UKESM1 model used in the CMIP6 experiments (Sellar et al, 2019). For each ensemble member, let u be a vector of parameter inputs which determine climatic aerosol-related processes of practical importance, and let x be a vector of control variables which define the specific output of the climate model, denoted $\eta(x, u)$, representing the climate observable $\zeta(x)$. In our setting, x denotes a latitude-longitude-time triple in the climate model output's spatiotemporal grid, denoted $\mathcal{M}_{\text{sim}}$ and called the `SimulatedGrid` in Figure 1. The parameters u are listed in Table 1. The ensemble is a set of simulations, in notation

$$D_{\text{train}} = \{(x, u^j, \eta(x, u^j)) : x \in \mathcal{M}_{\text{sim}}, j = 1, 2, \ldots, 221\}.$$

For reference, this and some later notation used throughout this paper is summarized in Table 2.

1.2. Problem setup

In order to emulate the model output for unobserved parameter values, we assume as in Johnson et al (2020) that at $x \in \mathcal{M}_{\text{sim}}$, the model output is a realization of a Gaussian process,

$$\tilde{\eta}_x(u) \sim \mathcal{GP}[m_x(\cdot), k_x(\cdot,\cdot)]. \qquad (u \in \mathbb{R}^p) \tag{1}$$

For each x, we train the surrogate model $\tilde{\eta}_x$ as described in Section 2.1.

Let $z(x)$ denote the AOD observations and u^* the true parameter value. Assuming that the emulated climate model is unbiased, these observations and parameters are related according to the equations

$$z(x) = \zeta(x) + \epsilon_{\text{meas},x} = \mathbb{E}\left[\tilde{\eta}_x(u^*)|D_{\text{train}}\right] + \epsilon_{\text{emu},x}(u^*) + \epsilon_{\text{meas},x} + \epsilon_{\text{other},x}.$$

The different sources of variance in the measurements – namely, the measurement uncertainty ($\epsilon_{\text{meas},x}$), the emulation uncertainty ($\epsilon_{\text{emu},x}$), and any other sources ($\epsilon_{\text{other},x}$) which are not analyzed uniquely, including uncertainty due to model discrepancy, including erroneously simulated meteorology, error in representativeness of measurements, etc. – are assumed to be mean zero and independent across x. This is a simplifying assumption in that, in reality, these terms might be correlated across x. By further assuming that $\epsilon_{\text{meas},x}$ and $\epsilon_{\text{other},x}$ are Gaussian, the observations $z(x)$ are jointly normally distributed across the x on the spatiotemporal grid $\mathcal{M}_{\text{sat}}$ (`SatelliteGrid`) on which the MODIS retrivals are gridded. In particular,

$$z = (z(x))_{x\in\mathcal{M}_{\text{sat}}} \sim N\left(\mathbb{E}\left[(\tilde{\eta}_x(u^*))_{x\in\mathcal{M}_{\text{sat}}}|D_{\text{train}}\right], \Sigma_{\text{meas}} + \Sigma_{\text{emu}} + \delta^2 I_{|\mathcal{M}_{\text{sat}}|}\right), \tag{2}$$

where Σ_{meas} and Σ_{emu} are covariance matrices such that entry $\Sigma_{\text{meas},i,j} = \text{Var}(\epsilon_{\text{meas},x^i})$ is the measurement uncertainty at location x^i when $i = j$, otherwise zero (i.e., it is a diagonal matrix and the measurement errors are assumed to be uncorrelated between locations); $\Sigma_{\text{emu},i,j} = \text{Var}(\epsilon_{\text{emu},x^i}(u^*))$ is the emulation uncertainty when $i = j$, otherwise zero; and $\delta^2 = \text{Var}(\epsilon_{\text{other, x}})$ is a homoscedastic variance term standing in for all other unaccounted-for errors. The matrix $I_{|\mathcal{M}_{\text{sat}}|}$ is an identity matrix with number of rows equal to the size of the set $\mathcal{M}_{\text{sat}}$. The disagreement between grids $\mathcal{M}_{\text{sim}}$ and $\mathcal{M}_{\text{sat}}$ is addressed in the following section. The $\Sigma_{\text{emu},i,i}$ are modeled by the surrogate model (see Section 2.1). The $\Sigma_{\text{meas},i,i}$ are the published MODIS uncertainties, which may not account for all possible problems with the retrievals, but these unaccounted-for uncertainties will be captured by δ^2, which is estimated from the observations (see Section 2.4).

2. Inference framework

We use the C3.AI Suite, a cloud computing platform for data analytics workflows deployed to Microsoft Azure infrastructure (C3 Enterprise AI, 2022). The platform combines databases, open-source packages, and proprietary machine-learning workflows optimized for working with large-scale, data-intensive applications. We built new data structures and methods for processing NETCDF4 files containing high-dimensional time-series datasets. We also developed a scalable inference pipeline for training and predicting through several thousands of Gaussian process models using asynchronous processing such as parallel batch and map-reduce jobs. This pipeline is summarized in Figure 1, and complements other recently published workflows for similar tasks (e.g., Watson-Parris et al, 2021).

The grid of satellite measurements is finer than the grid of simulations. The raw MODIS retrievals are pre-processed algorithmically by the MODIS team before they are provided on a $1^\circ \times 1^\circ$ spatial grid in the Level-3 Global Gridded Atmosphere Product (Hubank et al, 2020). However, the model outputs are on a $1.35^\circ \times 1.875^\circ$ grid. To reconcile these differences, we match simulated grid cells to their nearest neighbor on the Level-3 MODIS product grid in space and time. Differences are computed on the resulting $1.35^\circ \times 1.875^\circ$ resolution `MatchedGrid`, denoted $\mathcal{M}$.

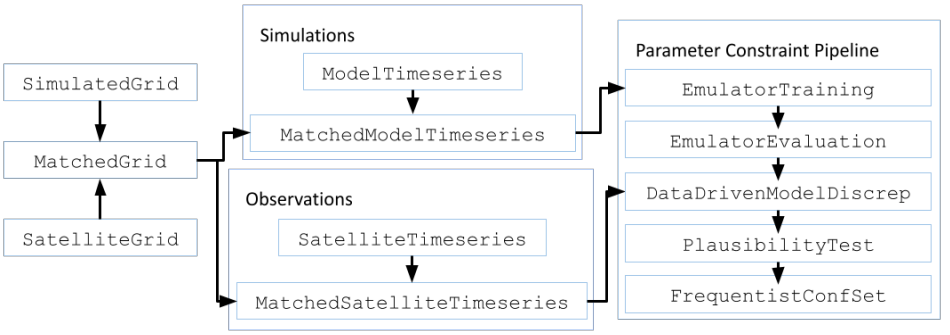

Figure 1. *The flow chart for our pipeline for building frequentist confidence sets on climate model parameters. After matching both the satellite observation and model output grids, five steps of processing follow. The* `EmulatorTraining` *and* `EmulatorEvaluation` *pipes provide scalability to the framework by leveraging parallel computing in these most expensive steps. The* `DataDrivenModelDiscrep`, `PlausibilityTest`, *and* `FrequentistConfSet` *pipes implement the strict bounds based method to ensure principled uncertainty quantification.*

2.1. Emulate: Inside the `EmulatorTraining` *pipe*

As indicated in Eq. (1), we assume the climate model is a realization of a Gaussian process where for each $x \in \mathcal{M}$ the mean and anisotropic exponential covariance functions are

$$m_x(u) = \beta_{0,x}, \qquad k_x(u,u') = \beta_{1,x}^2 \exp\left(-\sqrt{\sum_{i=1}^{p} \frac{(u_i - u_i')^2}{\ell_{i,x}^2}}\right).$$

We make this choice for k_x for the reason that a relatively rough process appears reasonable in this problem. We place an anisotropy assumption on the model by fitting a different length scale parameter $\ell_{i,x}$ for each model parameter u_i. By fitting $\ell_{i,x}$ separately for each x, we let the parameters' effects on the model output vary geographically.

We train the emulating Gaussian processes by estimating their parameters on D_{train} using maximum likelihood (Rasmussen and Williams, 2006) with the scikit-learn Python library and L-BFGS-B optimization algorithm, and thus we obtain a collection of models $\tilde{\eta}_x$, $x \in \mathcal{M}$. Figure 2 illustrates the quality of these emulators by verifying that the emulated AOD varies with respect to active parameters exactly as the training data set would suggest. In terms of scalability, an ordinary Gaussian process model on the entire `ModelTimeseries` would compute in $O\left((|\mathcal{M}|n)^3\right)$ time, n being the number of members in the PPE, whereas the `EmulatorTraining` pipe trains in $O\left(|\mathcal{M}|n^3\right)$ time. This routine is parallelized by distributing batches of training jobs to independent worker nodes on the C3.AI suite cluster, making the physical wait time much shorter.

2.2. Predict: Inside the `EmulatorEvaluation` *pipe*

We uniformly sample within the ranges given in Table 1 a collection of 5,000 new parameter vectors u^k (in contrast with 221 training vectors) to obtain a testing sample

$$D_{\text{test}} = \{(x, u^k, \mathbb{E}[\tilde{\eta}_x(u^k) \mid D_{\text{train}}], \text{Var}[\tilde{\eta}_x(u^k) \mid D_{\text{train}}]) : x \in \mathcal{M}, k = 1, 2, \ldots, 5000\}.$$

This set pairs points in the spatiotemporal-parametric space with the emulated mean and variance of the AOD response, specifying a distribution closely mimicking the model response surface $\eta(x,u)$. We expect that this number of sampling points u^k is sufficient to reliably constrain a small number of

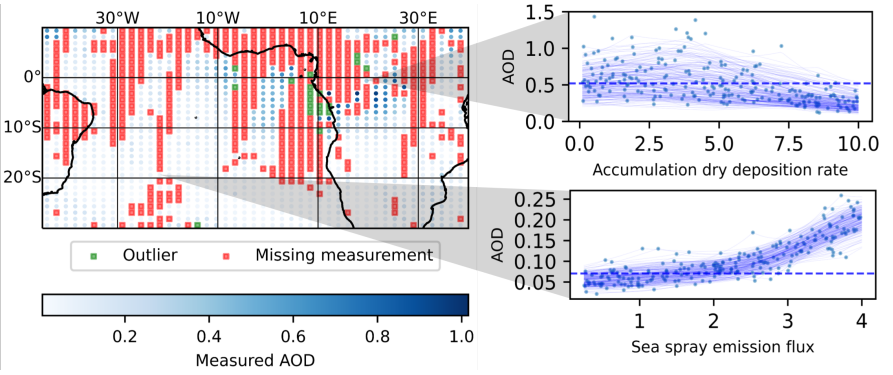

Figure 2. *Sample curves of the emulated response $\mathbb{E}[\tilde{\eta}_x(u) \mid D_{train}]$ averaged over two MODIS observing times on July 1, 2017 for two locations x. (Left) Red gridpoints are missing MODIS AOD retrievals. Green gridpoints are ruled out as outliers per Section 2.2. (Top right) The scattered points are from D_{train}, and the 221 curves are slices of the trained emulator response surface where all of the parameters are fixed to their training values from D_{train} except the parameter labeling the x axis of each subplot, which is varied within its range given in Table 1. Near $(0°, 20°)$, AOD decreases as the accumulation dry deposition rate increases. The average MODIS measurement is given by the dashed line. (Bottom right) At $(-20°, -20°)$, emulated AOD responds positively to the sea spray emission flux.*

parameters. These predictions are performed efficiently by a map-reduce job across the collection of models $\tilde{\eta}_x$.

The surrogate model appears to perform well at most locations in the spatio-temporal domain. However, gross discrepancies between the observations and the model output arise in a small fraction of the grid points which cannot be accounted for using our model discrepancy term (described later). In particular, when we consider the distance metric

$$J_{x,\delta}(u^k) = \frac{|\mathbb{E}[\tilde{\eta}_x(u^k) \mid D_{\text{train}}] - z(x)|}{\sqrt{\text{Var}[\tilde{\eta}_x(u^k) \mid D_{\text{train}}] + \text{Var}[\epsilon_{\text{meas},x}] + \gamma^2}},$$

we identify about 2% of the grid points for which $J_{x,\delta}(u^k)$ is a visible outlier for all u^k, $k = 1, \ldots, 5000$. The parameter γ here was tuned visually based on QQ plots since the other variance parameter δ is yet unestimated. Let $\mathcal{M}^*$ be the remaining coordinates in the spatiotemporal grid which have not been excluded either as outliers or due to missingness.

2.3. Discrepancy: Inside the `DataDrivenModelDiscrep` *pipe*

The quantity δ^2 as seen in (2) is the variance of the unaccounted-for uncertainty $\epsilon_{\text{other},x}$ in our model, which we estimate using maximum likelihood. We write down the likelihood for unknowns u and δ^2 from our Gaussian assumption,

$$L(u, \delta^2; D_{\text{train}}) = \prod_{x \in \mathcal{M}^*} \frac{1}{\sqrt{2\pi}} \left(\sigma^2_{x,\text{emu}}(u) + \sigma^2_{\text{meas},x} + \delta^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2} \frac{[\hat{\eta}_x(u) - z(x)]^2}{\sigma^2_{x,\text{emu}}(u) + \sigma^2_{\text{meas},x} + \delta^2}\right\},$$

where

$$\sigma^2_{x,\text{emu}}(u) = \text{Var}[\tilde{\eta}_x(u) \mid D_{\text{train}}], \quad \sigma^2_{\text{meas},x} = \text{Var}[\epsilon_{\text{meas},x}], \quad \hat{\eta}_x(u) = \mathbb{E}[\tilde{\eta}_x(u) \mid D_{\text{train}}].$$

To numerically obtain the maximum likelihood estimate for δ^2, we compute the maximizing value $\hat{\delta}^2_k$ of $\log L$ for each of the test parameters u^k using the scipy Python library's implementation of Brent's

algorithm (Press et al, 1992); then among these maximizing values we select the one which gives the overall maximum likelihood over $k = 1, 2, \ldots, 5000$. The resulting estimate $\hat{\delta}^2_{\text{MLE}} = \hat{\delta}^2_{\hat{k}}$, where $\hat{k} = \arg\max_k \log L(u^k, \hat{\delta}^2_k; D_{\text{train}})$, is an approximate estimator, where the approximation is due to the search over the parameter space being finite.

This part of the pipeline runs quickly. Evaluating the expression for the likelihood and performing the optimization routine took about five minutes in real time in our case and can be performed in local memory. Our value for the estimator on the remaining data was $\hat{\delta}^2_{\text{MLE}} = 0.025$, which is of similar magnitude as the average measurement variance of $\bar{\sigma}^2_{\text{meas}} = 0.027$ and emulation variance of $\bar{\sigma}^2_{\text{emu}}(\hat{u}) = 0.038$.

2.4. Test: Inside the `PlausibilityTest` *pipe*

To obtain a confidence set for the underlying atmospheric parameters, we perform a test for parameter plausibility using MODIS AOD observations on the `MatchedGrid`. Following the terminology used in Johnson et al (2020), we write down the *implausibility metric*

$$I(u) = \sqrt{\sum_{x \in \mathcal{M}^*} \left(\frac{\mathbb{E}[\tilde{\eta}_x(u) \mid D_{\text{train}}] - z(x)}{\sqrt{\text{Var}[\tilde{\eta}_x(u) \mid D_{\text{train}}] + \text{Var}[\epsilon_{\text{meas},x}] + \hat{\delta}^2_{\text{MLE}}}} \right)^2}.$$

For each u in D_{test}, we compare our observed implausibility measure against its approximate distribution under the null hypothesis that u is the correct parameter, $H_0 : I(u) \sim \sqrt{\chi^2(df = |\mathcal{M}^*|)}$. Here we use the facts that the sum of n squared independent standard Gaussian random variables is distributed as $\chi^2(df = n)$ and that n is assumed to be large enough that we can ignore the variation in the model-discrepancy variance estimate $\hat{\delta}^2_{\text{MLE}}$. In other words, testing at the 0.05 significance level, a parameter vector u will be deemed implausible if $I(u)$ exceeds the 95th percentile of the above null distribution.

A method for obtaining confidence sets inspired by the application of history matching in Johnson et al (2020) can also be derived. We find that inference based on this method is sensitive to the choice of tolerance level that is explicit in their choice of implausibility statistic. For a discussion of how our test differs from this instance of history matching, see the Appendix.

2.5. Infer: Inside the `FrequentistConfSet` *pipe*

Having obtained a collection of test results for each u^k, $k = 1, \ldots, 5000$, we approximate the Neyman inversion (Dalmasso et al, 2023) of the test for plausibility described above by retaining all those parameters u^k for which we do not reject the null at 0.05 significance level to obtain an approximate 95% confidence set.[1] This is the region of the 17-dimensional parameter space which exclusively contains non-implausible parameter vectors. A 2-dimensional projection of this set is on the right of Figure 3.

3. Results

Using our strict bounds-based test for parameter plausibility, we obtain a simultaneous 95% confidence set on the selected climate model parameters. We find that large values of the sea spray emission flux parameter are on the verge of being constrained with just two weeks of AOD data, shown in the upper left panel of Figure 3. However, the two plots on the bottom left show that for no level of either BVOC SOA or the accumulation mode dry deposition rate does our test for plausibility always fail, and so formal constraints on these cannot be obtained. To illustrate this point, suppose that the value of the accumulation dry deposition rate parameter (bottom left of Figure 3) was wrongly set to 1 when the

[1] Note that the repetition of this singular hypothesis test is for inversion purposes. Since we are not inverting more than one distinct test, we do not face multiple testing issues of controlling Type I error.

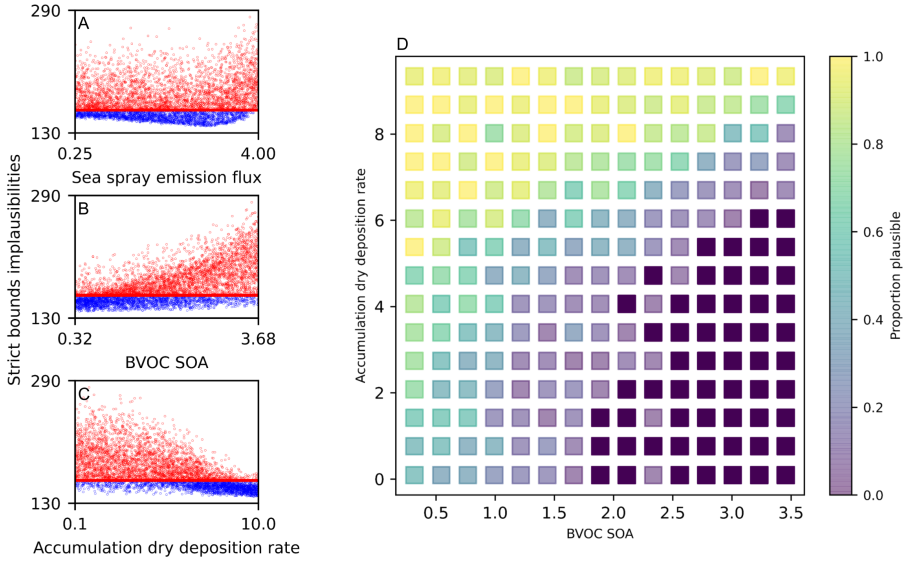

Figure 3. *Parameter constraints at 95% confidence level. (A-C) One-dimensional projections of the* `FrequentistConfSet` *described in Section 2.5. The* 95*th percentile of the approximate null distribution H_0 is indicated by the horizontal red lines. The sea spray emission flux parameter appears to be on the verge of being constrained on its own from only two weeks of data. (D) The space spanned by the BVOC SOA and accumulation mode dry deposition rate parameters is binned, and the color of each bin shows the proportion of plausible parameter values inside. Dark purple indicates a proportion of zero—evidently, the lower right corner of this space is ruled out as implausible.*

true value is 10. Then there are combinations of the other 16 selected parameters that enable us to fit the model outputs to the observed AOD data within the modeled uncertainties. Hence we cannot rule out value 1 for the accumulation dry deposition rate parameter, and likewise for the other values for each parameter. If evaluated at a lower confidence level, our results seem broadly consistent with Regayre et al (2023), who constrain the dry deposition rate toward large values, and Regayre et al (2020) and Johnson et al (2020), who also constrain the sea spray emission parameter.

We are able to obtain a constraint at 95% confidence level on the combination of the deposition rate of dry aerosols in the accumulation mode and biogenic secondary organic aerosol from volatile organic compounds. See the right panel of Figure 3 for a binned projection of the resulting confidence set onto their span. The resulting constraint can be understood as physically meaning the following: If one hypothesizes that there are a lot of aerosols emitted by vegetation while the deposition rate is low for relatively large particles, then one will overestimate AOD so much that even when controlling for the uncertainty of the MODIS retrieval estimates due to instrumental error, the imperfection of our surrogate model, and any other sources of model discrepancy that we estimate for our climate model, a significant region of the subspace of these parameters can be ruled as implausible at the 95% confidence level.

4. Conclusion

To our knowledge, this is the first use of simulated AOD from a climate model at a time resolution as high as three-hourly to obtain observation-based constraints on input parameters likely to regulate

AOD. That a salient and meaningful constraint has been gleaned from just two weeks' worth of data is suggestive of promising uses of high time resolution data in the future. Our framework is well suited for problem settings where a perturbed parameter ensemble is available for one's climate simulator and where Gaussian process emulation is appropriate. Notably, any unquantified sources of uncertainty in this setting are accounted for by the data-driven model discrepancy built into the presented pipeline, an aspect which differs from other recent scalable frameworks for model calibration, such as ESEm (Watson-Parris et al, 2021). Our approach assumes that the model discrepancy can be captured by an additive Gaussian error that is independent across space and time so our constraints rely on these assumptions being at least approximately satisfied. A fundamental difference between our approach and the previous history matching approach as done by Johnson et al (2020) is that our method provides frequentist confidence sets with well-defined probabilistic guarantees. In addition, as described in the Appendix, our method has a potential advantage over Johnson et al (2020) in that the latter is sensitive to the tuning of its tolerance level. The computational cost of our pipeline is dominated by the Gaussian process computations, so the approach is computationally feasible as long as enough parallel processing resources are available for training and evaluating the pixelwise emulators.

There are limitations to what can be achieved: with an imperfect and over-parameterized model, constraints can become inconsistent, or different parameter combinations can yield the same simulated ERF (Lee et al, 2016). However, employing more quantities for which we have observational and simulated data would nonetheless allow us to constrain a larger number of these parameters and get the most stringent constraints on aerosol radiative forcing we can. Other observable atmospheric quantities, such as sulfates or organic carbon, are sensitive to different sets of atmospheric parameters than those to which aerosol optical depth is sensitive, yielding potential opportunities to further constrain the parameter space using larger, more diverse observational data sets.

Acknowledgements. We are grateful to the two anonymous referees for the insightful comments we received through the peer review process for this manuscript. We also thank the Carnegie Mellon University Statistical Methods for the Physical Sciences (STAMPS) Research Group for helpful discussions and feedback at various stages throughout this work.

Author Contributions. Conceptualization: H.G., M.K.; Formal analysis: J.C.; Methodology: J.C., M.K., K.C., L.R., H.G.; Climate simulations: L.R., K.C., L.D., P.S.; Analysis software: B.A., J.C.; Writing–original draft: J.C. Writing–review and editing: J.C., M.K., H.G., L.R., P.S.

Competing Interests. The authors declare that no competing interests exist.

Data Availability Statement. The MODIS AOD measurements supporting the results of this work can be accessed through the MODIS web page: https://modis.gsfc.nasa.gov/data/dataprod/mod04.php (last access: 17 January 2023). Output from the A-CURE PPE is available on the CEDA archive (Regayre et al, 2023). The code to reproduce results from this paper can be found on GitHub: https://github.com/c3aidti/smoke/tree/main/climateInformatics2023 (last version: 31 March 2023).

Funding Statement. J.C., M.K. and H.G. acknowledge grant funding from the C3.ai Digital Transformation Institute. M.K. was also supported in part by NSF grants DMS-2053804 and PHY-2020295, and H.G. by NASA grant 80NSSC21K1344. L.R. and K.S. acknowledge funding from NERC under grants AEROS, ACID-PRUF, GASSP and A-CURE (NE/G006172/1, NE/I020059/1, NE/J024252/1 and NE/P013406/1). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Appendix: Comparison with the method of history matching. It is possible to provide frequentist confidence sets on parameters using a method inspired by history matching. Below, we describe an adaptation of the version of the method employed in Johnson et al (2020). Principally, the typical use of history matching in that work and others is to exclude implausible parameter values in a systematic way, not to produce a plausible constraint set with fixed probabilistic properties.

One difference between the method of Johnson et al (2020) and our PLAUSIBILITYTEST pipe of Section 2.4 is the choice of implausibility statistic. Johnson et al (2020) use the statistic

$$I_{\mathrm{HM},N}(u) = \left\{ \frac{|\mathbb{E}[\,\tilde{\eta}_x(u) \mid D_{\mathrm{train}}\,] - z(x)|}{\sqrt{\mathrm{Var}[\,\tilde{\eta}_x(u) \mid D_{\mathrm{train}}\,] + \mathrm{Var}[\,\epsilon_{\mathrm{meas},x}\,] + \mathrm{Var}[\,\epsilon_{\mathrm{other},x}\,]}} : x \in \mathcal{M}^* \right\}_{(N)},$$

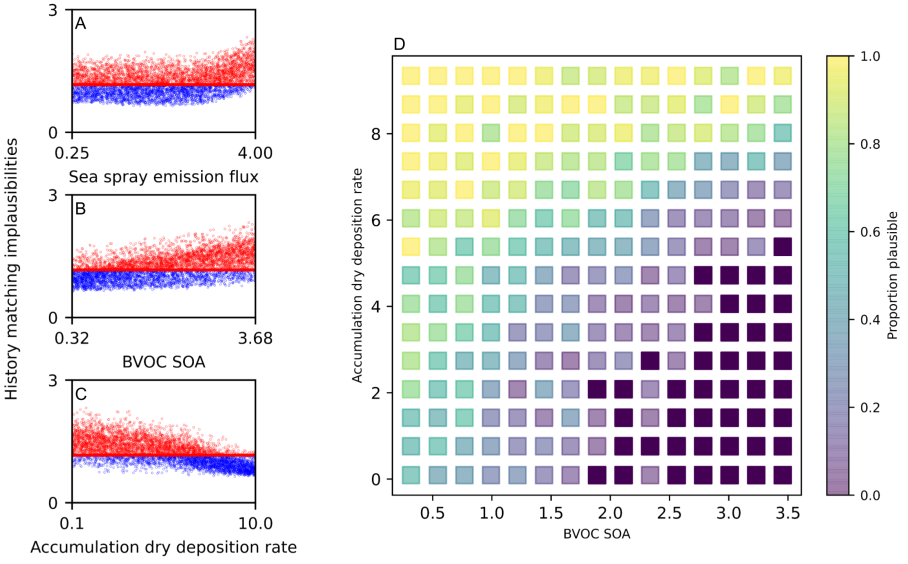

Figure 4. *Constraints resulting from an adaptation of the history matching method at* 95% *confidence level. Referring to the implausibility statistic given by Eq.* (3)*, we use* $q = 0.25$.

where $S_{(N)}$ denotes the Nth largest element of a set S. For instance,

$$I_{\mathrm{HM},1}(u) = \max\left\{ \frac{|\mathbb{E}[\,\tilde{\eta}_x(u) \mid D_{\mathrm{train}}\,] - z(x)|}{\sqrt{\mathrm{Var}[\,\tilde{\eta}_x(u) \mid D_{\mathrm{train}}\,] + \mathrm{Var}[\,\epsilon_{\mathrm{meas},x}\,] + \mathrm{Var}[\,\epsilon_{\mathrm{other},x}\,]}} : x \in \mathcal{M}^* \right\}.$$

The authors then tune the "tolerance level" (which corresponds to the number $(|\mathcal{M}^*| - N)$ in our expression for the statistic) and the "exceedence threshold" (which, by analogy to our test, corresponds to an implausibility cutoff or critical value; in most cases set to 3.5) until a fixed proportion (say, 40%) of test parameters u are retained as plausible under the statistic. By this account, a constraint on the parameters is necessarily yielded, but the confidence level at which this constraint holds is undetermined.

Alternatively, we can set a confidence level first and compare the resulting constraints obtained by the strict bounds approach or this history matching-inspired approach, where the former approach yields the constraints shown in Figure 3. In particular, consider the implausibility statistic

$$I_{1-q}(u) = \mathrm{quantile}_{1-q}\left\{ \frac{|\mathbb{E}[\,\tilde{\eta}_x(u) \mid D_{\mathrm{train}}\,] - z(x)|}{\sqrt{\mathrm{Var}[\,\tilde{\eta}_x(u) \mid D_{\mathrm{train}}\,] + \mathrm{Var}[\,\epsilon_{\mathrm{meas},x}\,] + \hat{\delta}^2_{\mathrm{MLE}}}} : x \in \mathcal{M}^* \right\}, \tag{3}$$

where $\mathrm{quantile}_{1-q}$ returns the $(1-q)$th quantile of the given set. For example, $I_1(u)$ returns the maximum absolute normalized discrepancy for parameter u and $I_{0.5}(u)$ is the median. This is a close analog to the statistic seen in Johnson et al (2020) when $q \approx N/|\mathcal{M}^*|$. Naturally the approximate null distribution for this new statistic is that of the $(1-q)$th percentile of a sample of $|\mathcal{M}^*|$ half-normal random variables. A critical value for the plausibility test at the 5% significance level can therefore be estimated quickly by simulating a collection of samples of $|\mathcal{M}^*|$ half-normal random variables, drawing the $(1-q)$th percentile from each sample, and then selecting the 95th percentile from that collection. In Figure 4, we show the 95% confidence level constraints on the aerosol parameters that are obtained from the above-described history matching method using parameter $q = 0.25$.

As this example illustrates, similar non-trivial and principled constraints on aerosol parameters are possible. However, the history matching-inspired approach requires choosing the tuning parameter q which substantially affects the final constraints. In Figure 4, we chose q to obtain constraints similar to those given by our method in Figure 3. If we choose $q = 0.5$ (i.e., use the median), we find that the constraints (not shown) become looser than those provided by our method. On the other hand, if we choose q close to zero, we find that this history matching approach becomes sensitive to non-Gaussianities in the tails of the error distributions, leading to overly discriminating plausiblity test results. If history matching was calibrated like this to obtain confidence sets at a prescribed confidence level (which, we emphasize, is not currently done), it seems difficult to choose q optimally to balance the power of the tests with robustness to mismodeling of the error distribution tails.

References

C3 Enterprise AI https://c3.ai/, Accessed: 2022-12-13.

Biegler L, Biros G, Ghattas O, Heinkenschloss M, *et al.* (2010) *Large-Scale Inverse Problems and Quantification of Uncertainty*, John Wiley & Sons.

Bower RG, Goldstein M and Vernon I (2010) Galaxy formation: a Bayesian uncertainty analysis, *Bayesian Analysis 5*(4), 619–670.

Craig PS, Goldstein M, Scheult AH and Smith JA (1997) "Pressure matching for hydrocarbon reservoirs: a case study in the use of Bayes linear strategies for large computer experiments." In Gatsonis, C, Hodges, JS, Kass, RE, McCulloch, R, Rossi, P and Singpurwalla, ND (eds.), *Case Studies in Bayesian Statistics*, Springer-Verlag.

Cranmer K, Brehmer J and Louppe G (2020) The frontier of simulation-based inference, *Proceedings of the National Academy of Sciences 117*(48), 30,055–30,062.

Dalmasso N, Masserano L, Zhao D, Izbicki R, and Lee AB (2023) Likelihood-Free Frequentist Inference: Confidence Sets with Correct Conditional Coverage [preprint], https://arxiv.org/abs/2107.03920.

Higdon D, Gattiker J, Williams B and Rightley M (2008) Computer Model Calibration Using High-Dimensional Output, *Journal of the American Statistical Association 103*(482), 570–583.

Hubank P, Platnick S, King M and Ridgway B (2020) MODIS Atmosphere L3 Gridded Product Algorithm Theoretical Basis Document (ATBD) & Users Guide, https://atmosphere-imager.gsfc.nasa.gov/sites/default/files/ModAtmo/documents/L3_ATBD_C6_C61_2020_08_06.pdf

Johansen K (2008) *Statistical Methods for History Matching*, PhD Thesis, Technical University of Denmark, Kongens Lyngby, Denmark.

Johnson JS, Regayre LA, Yoshioka M, Pringle KJ, *et al.* (2018) The importance of comprehensive parameter sampling and multiple observations for robust constraint of aerosol radiative forcing, *Atmospheric Chemistry and Physics 18*(17), 13,031–13,053.

Johnson JS, Regayre LA, Yoshioka M, Pringle KJ, *et al.* (2020) Robust observational constraint of uncertain aerosol processes and emissions in a climate model and the effect on aerosol radiative forcing, *Atmospheric Chemistry and Physics 20*(15), 9,491–9,524.

Kennedy MC and O'Hagan A (2001) Bayesian calibration of computer models, *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63*(3), 425–464.

Lee LA, Reddington CL, Carslaw KS (2016) On the relationship between aerosol model uncertainty and radiative forcing uncertainty, *PNAS 113*(21), 5,820–5,827.

McNeall D, Williams J, Booth B, Betts R, *et al.* (2016) The impact of structural error on parameter constraint in a climate model, *Earth System Dynamics 7*(4), 917–935.

Patil P, Kuusela M and Hobbs J (2022) Objective Frequentist Uncertainty Quantification for Atmospheric CO_2 Retrievals, *SIAM/ASA Journal on Uncertainty Quantification*, *10*(3), 827–859.

Press W, Teukolsky SA, Vetterling WT and Flannery BP (1992) *Numerical Recipes in C*, Cambridge University Press.

Rasmussen CE and Williams CKI (2006) *Gaussian processes for machine learning*, MIT Press.

Regayre LA, Johnson JS, Yoshioka M, Pringle KJ, *et al.* (2018) Aerosol and physical atmosphere model parameters are both important sources of uncertainty in aerosol ERF, *Atmospheric Chemistry and Physics*, *18*(13), 9,975–10,006.

Regayre LA, Schmale J, Johnson JS, Tatzelt C, *et al.* (2020) The value of remote marine aerosol measurements for constraining radiative forcing uncertainty, *Atmospheric Chemistry and Physics*, *20*(16), 10,063–10,072.

Regayre LA, Deaconu L, Grosvenor DP, Sexton DMH, *et al.* (2023) Identifying climate model structural inconsistencies allows for tight constraint of aerosol radiative forcing, *EGUsphere* [preprint], https://egusphere.copernicus.org/preprints/2023/egusphere-2023-77/.

Schafer CM and Stark PB (2009) Constructing Confidence Regions of Optimal Expected Size, *Journal of the American Statistical Association 104*(487), 1,080–1,089.

Schutgens N, Tsyro S, Gryspeerdt E, Goto D, *et al.* (2017) On the spatio-temporal representativeness of observations, *Atmospheric Chemistry and Physics*, *17*(16), 9,761–9,780.

Sellar AA, Jones CG, Mulcahy JP, Tang Y, *et al.* (2019) UKESM1: Description and Evaluation of the U.K. EarthSystem Model, *Journal of Advances in Modeling Earth Systems*, http://dx.doi.org/10.1029/2019MS001739.

Stanley M, Patil P and Kuusela M (2022) Uncertainty quantification for wide-bin unfolding: one-at-a-time strict bounds and prior-optimized confidence intervals, *Journal of Instrumentation*, *17*(10), P10013.

Stark PB (1992) Inference in infinite-dimensional inverse problems: Discretization and duality, *Journal of Geophysical Research*, *97*(B10), 14,055–14,082.

Telford PJ, Braesicke P, Morgenstern O and Pyle JA (2008) Technical Note: Description and assessment of a nudged version of the new dynamics Unified Model, *Atmospheric Chemistry and Physics*, *8*(6), 1,701–1,712.

Verly G, David M, Journel AG, Marechel A (1984) *Geostatistics for Natural Resources Characterization, Part 2*, Geostatistics for Natural Resources Characterization, D. Reidel Publishing Company.

Watson-Parris D, Williams A, Deaconu L and Stier P (2021) Model calibration using ESEm v1.1.0 – an open, scalable Earth system emulator, *Geoscientific Model Developemnt 14*(12) 7,659–7,672.