Ensemble-Based Experimental Design for Targeted High-Resolution Simulations to Inform Climate Models

Oliver Dunbar[1], Michael F. Howland[2], Tapio Schneider[1], and Andrew Stuart[1]

[1]California Institute of Technology
[2]Massachusetts Institute of Technology

November 22, 2022

Abstract

Targeted high-resolution simulations driven by a general circulation model (GCM) can be used to calibrate GCM parameterizations of processes that are globally unresolvable but can be resolved in limited-area simulations. This raises the question of where to place high-resolution simulations to be maximally informative about the uncertain parameterizations in the global model. Here we construct an ensemble-based parallel algorithm to locate regions that maximize the uncertainty reduction, or information gain, in the uncertainty quantification of GCM parameters with regional data. The algorithm is based on a Bayesian framework that exploits a quantified posterior distribution on GCM parameters as a measure of uncertainty. The algorithm is embedded in the recently developed calibrate-emulate-sample (CES) framework, which performs efficient model calibration and uncertainty quantification with only O(10^2) forward model evaluations, compared with O(10^5) forward model evaluations typically needed for traditional approaches to Bayesian calibration. We demonstrate the algorithm with an idealized GCM, with which we generate surrogates of high-resolution data. In this setting, we calibrate parameters and quantify uncertainties in a quasi-equilibrium convection scheme. We consider (i) localization in space for a statistically stationary problem, and (ii) localization in space and time for a seasonally varying problem. In these proof-of-concept applications, the calculated information gain reflects the reduction in parametric uncertainty obtained from Bayesian inference when harnessing a targeted sample of data. The largest information gain results from regions near the intertropical convergence zone (ITCZ) and indeed the algorithm automatically targets these regions for data collection.

Ensemble-Based Experimental Design for Targeted High-Resolution Simulations to Inform Climate Models

Oliver R. A. Dunbar[1], Michael F. Howland[1,2],Tapio Schneider[1], Andrew M. Stuart[1]

[1]California Institute of Technology, Pasadena, USA.
[2]Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

Key Points:

- Climate models can be calibrated with limited-area high-resolution simulations; we address their optimal placement in space and time.
- We propose an algorithm that places high-resolution simulations so that they are maximally informative about climate model parameters.
- The algorithm is benchmarked in a idealized aquaplanet general circulation model.

Corresponding author: Oliver Dunbar, `odunbar@caltech.edu`

Abstract

Targeted high-resolution simulations driven by a general circulation model (GCM) can be used to calibrate GCM parameterizations of processes that are globally unresolvable but can be resolved in limited-area simulations. This raises the question of where to place high-resolution simulations to be maximally informative about the uncertain parameterizations in the global model. Here we construct an ensemble-based parallel algorithm to locate regions that maximize the uncertainty reduction, or information gain, in the uncertainty quantification of GCM parameters with regional data. The algorithm is based on a Bayesian framework that exploits a quantified posterior distribution on GCM parameters as a measure of uncertainty. The algorithm is embedded in the recently developed calibrate-emulate-sample (CES) framework, which performs efficient model calibration and uncertainty quantification with only $\mathcal{O}(10^2)$ forward model evaluations, compared with $\mathcal{O}(10^5)$ forward model evaluations typically needed for traditional approaches to Bayesian calibration. We demonstrate the algorithm with an idealized GCM, with which we generate surrogates of high-resolution data. In this setting, we calibrate parameters and quantify uncertainties in a quasi-equilibrium convection scheme. We consider (i) localization in space for a statistically stationary problem, and (ii) localization in space and time for a seasonally varying problem. In these proof-of-concept applications, the calculated information gain reflects the reduction in parametric uncertainty obtained from Bayesian inference when harnessing a targeted sample of data. The largest information gain results from regions near the intertropical convergence zone (ITCZ) and indeed the algorithm automatically targets these regions for data collection.

Plain Language Summary

Climate models depend on dynamics across many spatial and temporal scales. It is infeasible to resolve all of these scales. Instead, the physics at the smallest scales is represented by parameterization schemes that link what is unresolvable to variables resolved on the grid scale. A dominant source of uncertainty in climate predictions comes from uncertainty in calibrating empirical parameters in such parameterization schemes, and this uncertainty is generally not quantified. Targeted high-resolution simulations of small-scale processes in limited areas are one means by which uncertainties in parameterizations can be reduced and quantified. Here we demonstrate an algorithm that optimizes placement of high-resolution simulations to maximize the information they provide about uncertain parameters in parameterization schemes. Because the sensitivity of simulated climate statistics, such as precipitation rates, to parameterizations varies in space and time, how informative high-resolution simulations are about the parameterizations also varies in space and time. In proof-of-concept simulations with an idealized global atmosphere model, we show that our novel algorithm successfully identifies the informative regions and times.

1 Introduction

Parameterizations of subgrid-scale processes, such as the turbulence and convection controlling clouds, are the principal cause of physical uncertainties in climate predictions (Cess et al., 1989, 1990; Bony & Dufresne, 2005; Stephens, 2005; Bony et al., 2006; Vial et al., 2013; Webb et al., 2013; Brient & Schneider, 2016). While these processes are too small in scale to become globally resolvable in climate models for the foreseeable future, many of them can be resolved in limited-area simulations (Schneider, Teixeira, et al., 2017). For example, the turbulence and convection (though currently not the microphysics) controlling clouds can be resolved in large-eddy simulations (LES) over limited areas (Siebesma et al., 2003; Stevens et al., 2005; Khairoutdinov et al., 2009; Matheou & Chung, 2014; Schalkwijk et al., 2015; Pressel et al., 2015, 2017). High-resolution simulations have been used to calibrate climate model parameterizations at selected sites,

primarily in low latitudes (Liu et al., 2001; Siebesma et al., 2003; Stevens et al., 2005; Siebesma et al., 2007; Hohenegger & Bretherton, 2011; M. Zhang et al., 2013; de Rooy et al., 2013; Romps, 2016; Tan et al., 2018; Smalley et al., 2019; Couvreux et al., 2021; Hourdin et al., 2021). Similarly, limited-area high-resolution simulations (e.g., Wang et al., 1996; Fox-Kemper & Menemenlis, 2013) have been used to calibrate subgrid-scale parameterizations of upper-ocean turbulence (Souza et al., 2020; Li & Fox-Kemper, 2017; Van Roekel et al., 2018; Reichl et al., 2016; Li et al., 2019; Campin et al., 2011; Reichl & Hallberg, 2018).

However, high-resolution simulations can be used more systematically, by driving them with a large-scale GCM and using the mismatch between statistics of the high-resolution simulations and GCM output to calibrate parameterizations (Schneider, Lan, et al., 2017). For example, Shen et al. (2020, 2021) drive atmospheric LES in domains $\mathcal{O}(10$ km$)$ in the horizontal with output from climate models, to produce simulated local data with which climate model parameterizations can be calibrated. In principle, LES can be embedded into coarse-resolution GCMs to sample thousands of sites across the globe. Parameterizations in the GCM can then learn automatically from the high-resolution simulations, and the high-resolution simulations can even be spun off on-the-fly during the integration of a coarse-resolution model (Schneider, Lan, et al., 2017). This allows systematic calibration and uncertainty quantification of parameterizations.

Pursuing such automatic calibration and uncertainty quantification of parameterizations raises the question of where to embed the high-resolution simulations so that they are maximally informative about the parameterizations. This is a question of experimental design (Chaloner & Verdinelli, 1995), akin to the question of how to choose sites for supplementary weather observations to optimally improve weather forecasts (Lorenz & Emanuel, 1998; Bishop & Toth, 1999; Emanuel et al., 1995). When climate statistics are non-stationary, there is a related question of what time periods should the expensive high-resolution simulation be integrated over.

Here we present a mathematical framework for addressing both the experimental design question of where, and when, to embed high-resolution simulations in a coarser model. We adopt a Bayesian inverse problem setting (see, e.g., Kaipio and Somersalo (2006), Tarantola (2005), Stuart (2010), and Dashti and Stuart (2013) for reviews). In this setting, parameters (or parametric or nonparametric functions) in parameterizations are treated as having probability distributions. Data (e.g., from high-resolution simulations) are used to reduce the uncertainty reflected by these distributions, balancing contributions of the data with that of prior knowledge about parameters (e.g., physical constraints). This setting is well suited for our needs, as it provides the complete joint posterior distribution for parameters, including the correlation structure of uncertainties among parameters. Distribution information is beneficial, for example, because it enables model-based predictions of rare events with quantified uncertainties (Dunbar et al., 2021). Analysis of the posterior distribution also may focus scientific development (e.g., improvement of parameterization schemes, Souza et al. (2020)) on areas where uncertainties can most effectively be minimized. In this paper, we use the posterior distribution to determine regions and times where local data (e.g., from high-resolution simulations) are maximally effective at reducing parameter uncertainties.

Construction of the full posterior distribution of the parameters is well known to be a computationally intensive task, requiring $\mathcal{O}(10^5)$ evaluations of the model in which the parameters appear with commonly used Markov chain Monte Carlo (MCMC) methods (see Geyer (2011) for an overview). The recent development of the calibrate-emulate-sample (CES) framework accelerates Bayesian learning by a factor of 10^3 (Cleary et al., 2021a; Dunbar et al., 2021). The calibration stage uses a variant of ensemble Kalman inversion (Iglesias et al., 2013) building on (Chen & Oliver, 2012; Emerick & Reynolds, 2013; Reich, 2011) to obtain a collection of samples of the model about an optimal set of parameters. The emulation stage features the training of a Gaussian process (Williams

& Rasmussen, 2006; Kennedy & O'Hagan, 2000, 2001a) to emulate the model using the samples from the calibration stage. The sample stage then samples a posterior distribution with MCMC methods, replacing the computationally expensive model with the cheap emulator. We build on the CES framework and show how Bayesian experimental design approaches can be incorporated within CES at negligible additional computational expense. In particular, we do not require additional forward model evaluations over what is already required in CES to perform uncertainty quantification.

To target high-resolution simulations, we use tools from experimental design, which provides methods for assessing the efficacy of learning about parameters from different designs (e.g., data from different locations or time periods) (Ryan et al., 2016). We determine the optimal designs where the model is most sensitive to the parameters by using the posterior distribution for the parameters as a utility to be optimized. We choose a utility function that assigns a score of the information entropy loss between posterior and prior for each design (Chaloner & Verdinelli, 1995; Alexanderian et al., 2014; Alexanderian & Saibaba, 2018); it is scale invariant and scalable to problems with high-dimensional input parameter spaces. The region with maximal utility determines where to acquire high-resolution data, and hence where to divert scarce computational resources for maximal effect.

We demonstrate the effectiveness of this targeted learning approach through simulations with an idealized moist GCM, with which we generate surrogates for high-resolution data and in which we are calibrating parameters in a quasi-equilibrium convection scheme (D. M. Frierson, 2007; O'Gorman & Schneider, 2008). We showcase our algorithms by showing that the recovered posterior distributions are reflective of the utility of information at different points and different times along the annual cycle.

In Section 2, we define the inverse problems for parameter calibration and the optimal design algorithm; details of efficient uncertainty quantification (CES) are left to Appendix A. In Section 3, we briefly describe the GCM used for demonstrating the algorithm. Results of the optimal design algorithm are described in section 4, first in a setting in which the GCM statistics are statistically stationary, then with seasonally varying GCM statistics. We end with discussion and future directions in Section 5.

2 Methodology

Our goal is to target data acquisition to regions and times at which uncertainty reduction (information gain) is maximized. We do so by first learning the temporally and spatially varying sensitivities of the model statistics with respect to model parameters. We then use this knowledge to target data acquisition to regions and times at which the model is maximally sensitive to new data. We work in a framework similar to Dunbar et al. (2021), which focuses on accelerated uncertainty quantification within a GCM.

Our point of departure in Section 2.1 is to specify the inverse problem for uncertainty quantification of parameters from data at a specific design. Related to application, this can be seen as the stage of learning parameter uncertainties from high-resolution simulation data at a certain region or time. Treating such sdata as computationally expensive to obtain, in Section 2.2 we investigate how to efficiently choose which region or time to learn from. To do this we formulate a set of related inverse problems, whose solutions allow us to assess the quality of different choices. In Section 2.3 we connect these two stages to form the targeted uncertainty quantification algorithm .

2.1 Inverse problem

We study calibration of parameters in a GCM by formulating parameter learning as a Bayesian inverse problem. Define $\mathcal{G}_T(\boldsymbol{\theta}; \boldsymbol{v}^{(0)})$ to be the forward map sending the pa-

rameters $\boldsymbol{\theta}$ to time-aggregated simulated climate statistics (averaged over a window of length $T > 0$) from an initial state $\boldsymbol{v}^{(0)}$. We assume that the aggregation $\mathcal{G}_T(\boldsymbol{\theta}, \cdot)$ is statistically stationary, and samples of such aggregated climate statistics are referred to as *data* throughout. We consider a situation in which data are only locally available, at a particular spatial or spatio-temporal location, indexed by k, which we refer to as the design point. This is relevant to our application of targeted high-resolution simulations with limited spatial and temporal extent. We make use of a restriction operation W_k to a point k, and define the local forward map, $\mathcal{S}_T(\boldsymbol{\theta}; k, \boldsymbol{v}^{(0)}) = W_k \mathcal{G}_T(\boldsymbol{\theta}; \boldsymbol{v}^{(0)})$.

For any given k, assume we have local data $\boldsymbol{z}_k$ available. In the application of interest, $\boldsymbol{z}_k$ is produced by a high-resolution simulation run. We can construct the forward map $\mathcal{S}_T(\cdot; k, \cdot)$, to form an inverse problem for the GCM learning from the local data as

$$\boldsymbol{z}_k = \mathcal{S}_T(\boldsymbol{\theta}; k, \boldsymbol{v}^{(0)}) + \delta_k, \tag{1}$$

where δ_k is a stochastic term to capture discrepancies between model $\mathcal{S}_T(\cdot; k, \cdot)$ and data $\boldsymbol{z}_k$, (e.g., Kennedy & O'Hagan, 2001a). The initial condition $\boldsymbol{v}^{(0)}$ appears in this formulation but is treated a nuisance variable. This view is justified in the context of learning about atmospheric parameterizations for climate models, where lower frequency information is informative (Schneider, Lan, et al., 2017). We use time-averaged data to filter out the high frequency information, and take T is larger than the dynamical system's Lyapunov timescale (for the atmosphere, this equates to $T \gtrsim 15$ days (F. Zhang et al., 2019)). To deal with the initial condition, one can view finite-time averaging as a perturbation of an infinite-time average by means of a central limit theorem. Following (Dunbar et al., 2021), we write $\mathcal{S}_T(\boldsymbol{\theta}; k, \boldsymbol{v}^{(0)}) \approx \mathcal{S}_\infty(\boldsymbol{\theta}; k) + \sigma_k$, where $\sigma_k \sim N(0, \Sigma(\boldsymbol{\theta}))$ is normal noise, independent from δ_k, with mean zero and with a covariance matrix $\Sigma(\boldsymbol{\theta})$ reflecting chaotic internal variability. The inverse problem then becomes

$$\boldsymbol{z}_k = \mathcal{S}_\infty(\boldsymbol{\theta}; k) + \delta_k + \sigma_k, \qquad \sigma \sim N(0, W_k \Sigma(\boldsymbol{\theta}) W_k^T). \tag{2}$$

This is now a desirable form of the inverse problem since the dependence on the initial condition has been removed.

Solving (2) involves finding the posterior distribution of $\boldsymbol{\theta}$ given the data $\boldsymbol{z}_k$, denoted $(\boldsymbol{\theta} \mid \boldsymbol{z}_k)$. Although we cannot evaluate $\mathcal{S}_\infty$ directly, we use the emulate phase of the calibrate-emulate-sample (CES) algorithm (Cleary et al., 2021b) to construct a surrogate of $\mathcal{S}_\infty$ from carefully chosen evaluations of $\mathcal{S}_T$. This has been shown to be efficient with respect to the required number of evaluations of $\mathcal{S}_T$ (Cleary et al., 2021b; Dunbar et al., 2021). Details of this algorithm are provided in Appendix A.

2.2 Experimental design

We imagine a situation where evaluating $\boldsymbol{z}_k$ has a large computational cost. In the relevant application of targeted high-resolution simulations, $\boldsymbol{z}_k$ is data obtained by running a high-resolution simulation at design point k. Our starting point is to assume that a limited computational budget restricts us to evaluate $\boldsymbol{z}_k$ at a single design point k at a time. We want to choose the k that leads to the most informative inverse problem (2). We take a Bayesian point of view, namely, the optimal k is the one for which the posterior distribution of $(\boldsymbol{\theta} \mid \boldsymbol{z}_k)$ learned from the inverse problem (2) has the smallest uncertainty.

To answer this conclusively, one would need to evaluate $\boldsymbol{z}_k$ at all design points k, which here is too computationally expensive. Instead, we investigate only the sensitivity of the forward model statistics $\mathcal{G}_T$ to its parameters $\boldsymbol{\theta}$ to assess the marginal information provided at each design point k. This marginal information at k is used as a proxy for the information content that would exist when learning from data $\boldsymbol{z}_k$. The benefits of this approach are that (i) we do not require any evaluations of $\boldsymbol{z}_k$ to select the optimal location; (ii) the measure of information content is naturally constructed from the

uncertainty reflected by the Bayesian posterior distribution; and (iii) we can perform this efficiently, and in a embarrassingly parallel fashion, requiring only $O(100)$ GCM runs, determined by the product of the ensemble size and the number of iterations typically needed in the calibration stage of the CES algorithm (see Appendix A). The approach necessarily will contain a bias from the prior distribution of the parameters, and it implicitly assumes unbiased model statistics $\mathcal{G}_T$. The latter in practice requires the inclusion of models for structural model error within $\mathcal{G}_T$, for example, learned error models that enforce conservation laws and sparsity (M. E. Levine & Stuart, 2021; Schneider et al., 2021).

Each evaluation of the forward map involves a simulation with the GCM and thus depends on an initial condition $\boldsymbol{v}^{(0)}$ and parameters $\boldsymbol{\theta}$. Together this gives rise to the definition of time-aggregated model statistics $\boldsymbol{y}$,

$$\boldsymbol{y} = \mathcal{G}_T(\boldsymbol{\theta}; \boldsymbol{v}^{(0)}). \tag{3}$$

Using the central limit theorem as before, we may write this relationship as

$$\boldsymbol{y} = \mathcal{G}_\infty(\boldsymbol{\theta}) + \sigma, \qquad \sigma \sim N\left(0, \Sigma(\boldsymbol{\theta})\right), \tag{4}$$

where $\Sigma(\boldsymbol{\theta})$ is the internal variability covariance matrix for parameters $\boldsymbol{\theta}$. To proceed, we choose a control value $\boldsymbol{\theta}^*$, for example we take the mean of the prior distribution, and, fixing $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, we generate a realization of $\boldsymbol{y}$. Given this realization of $\boldsymbol{y}$, we temporarily forget $\boldsymbol{\theta}^*$, and for any design point k, we consider a restriction of an inverse problem to k

$$W_k\boldsymbol{y} = W_k\mathcal{G}_\infty(\boldsymbol{\theta}) + \sigma_k, \qquad \sigma_k \sim N(0, W_k\Sigma(\boldsymbol{\theta})W_k^T). \tag{5}$$

The posterior distributions of $\boldsymbol{\theta} \mid W_k\boldsymbol{y}$ for all k obtained by solving (5) informs us about the sensitivities of $\mathcal{G}_\infty$ with respect to parameters, when only data at different k is available. To simplify solution of the inverse problem, we approximate the internal variability covariance matrix $\Sigma(\boldsymbol{\theta})$ by a fixed covariance matrix $\Sigma(\boldsymbol{\theta^*})$. This covariance matrix can be obtained by running a collection of control simulations with parameters fixed to (the known) $\boldsymbol{\theta}^*$ but with different initial conditions.

The *utility* U of a design W_k is a scalar function reflecting the quality of a given design. The design that maximizes the utility function is known as the optimal design. We choose a utility function by measuring information gain (or uncertainty reduction) in $(\boldsymbol{\theta} \mid W_k\boldsymbol{y})$ relative to the prior, in a form of Bayesian optimal design. We use the utility function arising from the linear Bayesian design (Chaloner & Verdinelli, 1995), which is the determinant of the information matrix (inverse posterior covariance matrix),

$$U(W_k) = \Big(\det\big(\mathrm{Cov}(\boldsymbol{\theta} \mid W_k\boldsymbol{y})\big)\Big)^{-1}. \tag{6}$$

The posterior covariance matrix $\mathrm{Cov}(\boldsymbol{\theta} \mid W_k\boldsymbol{y})$ can be estimated as the empirical covariance matrix of samples drawn from the posterior distribution $(\boldsymbol{\theta} \mid W_k\boldsymbol{y})$ for a design W_k. We refer to (6) as the D-utility because it fulfills the so-called D-optimality criterion. It is invariant under arbitrary linear transformations of the parameters, for example, when parameters are on different dimensional scales, unlike trace-based measures (e.g., A-optimal utility functions). For linear forward maps and Gaussian priors, maximization of this D-utility is equivalent to maximization of the expected Kullback-Leibler divergence (KLD), a relative entropy measure (Ryan et al., 2014; Huan & Marzouk, 2013; Cook et al., 2008; Kim et al., 2014). While KLD has beneficial mathematical properties, especially for highly non-Gaussian posteriors (Paninski, 2005), in practice it is difficult to evaluate, especially in high-dimensional problems (e.g., Huan & Marzouk, 2013).

2.3 Targeted uncertainty quantification algorithm

The combined algorithm for targeted uncertainty quantification consists of two stages: first, finding an optimal design point $\tilde{k}$ in a design stage and, second, evaluating parameter uncertainty with data from $\tilde{k}$ in an uncertainty quantification stage. Let D be the

finite index set for the set of design points, and define W_k to be the restriction map for any $k \in D$. The two stages then are as follows:

1. The design stage consists of the following steps:
   (a) Generate a sample of GCM simulated data $\boldsymbol{y} = \mathcal{G}_T(\boldsymbol{\theta}^*; \boldsymbol{v}^{(0)})$, and estimate the internal variability covariance matrix $\Sigma(\boldsymbol{\theta}^*)$. We approximate $\Sigma(\boldsymbol{\theta})$ as $\Sigma(\boldsymbol{\theta}^*)$.
   (b) For each $k \in D$, solve (5), in parallel, for the posterior of $(\boldsymbol{\theta} \mid W_k\boldsymbol{y})$, using the CES-type algorithm described in Appendix A.
   (c) For each $k \in D$, calculate the D-utility $U(W_k)$ from (6) and choose
   $$\tilde{k} = \arg\max_{k \in D} U(W_k).$$
2. The uncertainty quantification stage consists of the following steps:
   (a) At the optimal design point $\tilde{k}$, obtain a sample $\boldsymbol{z}_{\tilde{k}}$.
   (b) Solve the inverse problem (2) for the posterior distribution of $(\boldsymbol{\theta} \mid \boldsymbol{z}_{\tilde{k}})$.

The complexity of the first stage grows linearly with the candidate design points k because we only consider a point at a time. However, if one wishes to choose a design composed of K simultaneous points from a set D, a combinatorial problem arises, with complexity growing like $|D|!/((|D| - |K|)!|K|!)$. This will become prohibitively costly to solve by brute force, even in parallel. We focus on the algorithm for single design points k for now, addressing scaling questions in the discussion section.

3 Idealized GCM and Experimental Setup

3.1 Idealized GCM, parameters, and priors

To demonstrate the algorithm in a simplified setting, we use the idealized aquaplanet GCM described by D. M. W. Frierson et al. (2006) and O'Gorman and Schneider (2008b). The aquaplanet is a climate model with atmosphere and a simplified slab ocean covering the entire planet surface. Without topography, it exhibits symmetries in the longitudinal directions. The aquaplanet can produce statistically stationary climates by prescribing fixed insolation. It can also cyclostationary statistics over seasons through seasonally varying insolation (Bordoni & Schneider, 2008a; Howland et al., 2021). It has been shown in Dunbar et al. (2021); Howland et al. (2021) that the parameters of a simple quasi-equilibrium moist convection parameterization can be calibrated within this GCM in the stationary and cyclostationary regimes. The quasi-equilibrium moist convection scheme relaxes temperature and specific humidity toward moist-adiabatic reference profiles with a fixed relative humidity RH (D. M. W. Frierson, 2007). The timescale with which the temperature and specific humidity relax to their respective reference profiles is given by the parameter τ. The two parameters RH and τ are the key parameters to be calibrated and whose uncertainties we want to determine and minimize.

The priors for these parameters are taken to be logit-normal and lognormal distributions, $\mathrm{RH} \sim \mathrm{Logitnormal}(0, 1)$ and $\tau \sim \mathrm{Lognormal}(12\ \mathrm{h}, (12\ \mathrm{h})^2)$. That is, we define the invertible transformation

$$\mathcal{T}(\mathrm{RH}, \tau) = \left(\mathrm{logit}(\mathrm{RH}),\ \ln\left(\frac{\tau}{1\ \mathrm{s}}\right)\right),$$

which transforms each parameter to values along the real axis. We label the transformed (or computational) parameters as $\boldsymbol{\theta} = \mathcal{T}(\mathrm{RH}, \tau)$, and the untransformed (or physical) parameters (relative humidity and timescale) are uniquely defined by $\mathcal{T}^{-1}(\boldsymbol{\theta})$. We apply our calibration methods in the space of the transformed parameters $\boldsymbol{\theta}$, where priors are unit-free, normally distributed, and unbounded; meanwhile, the climate model uses the physical parameters $\mathcal{T}^{-1}(\boldsymbol{\theta})$, with $\mathrm{RH} \in [0, 1]$ and $\tau \in [0, \infty)$. In this way, the prior distributions enforce physical constraints on the parameters.

3.2 Objective function for parameter learning

We learn from climate statistics that are known to be sensitive to the parameters. We have knowledge about these sensitivities from a body of previous studies of large-scale atmosphere dynamics and mechanisms of climate changes which used this idealized GCM (e.g., O'Gorman & Schneider, 2008b, 2008a; Bordoni & Schneider, 2008b; O'Gorman & Schneider, 2009b; Schneider et al., 2010; Merlis & Schneider, 2011; O'Gorman, 2011; Kaspi & Schneider, 2011, 2013; X. Levine & Schneider, 2015; Bischoff & Schneider, 2014; Wills et al., 2017; Wei & Bordoni, 2018). We know, for example, that the convection scheme primarily affects the atmospheric thermal stratification in the tropics, with weaker effects in the extratropics (Schneider & O'Gorman, 2008). We also know that the relative humidity parameter RH in the convection scheme controls the humidity of the tropical free troposphere but has a weaker effect on the humidity of the extratropical free troposphere (O'Gorman et al., 2011). Thus, we expect tropical circulation statistics to be especially informative about the parameters in the convection scheme. However, convection plays a central role in intense precipitation events at all latitudes (O'Gorman & Schneider, 2009b, 2009a), so we expect statistics of precipitation intensity to be informative about convective parameters, and in particular to contain information about the relaxation timescale τ.

As statistics to learn from, we therefore choose averages of the free-tropospheric relative humidity, of the precipitation rate, and of a measure of the frequency of intense precipitation. We use averages over $T = 30$ days in statistically stationary simulations (Dunbar et al., 2021) and over $T = 90$ days in simulations of the seasonal cycle (Howland et al., 2021). We exploit the symmetry in the GCM by taking zonal averages in addition to the time averages. The relative humidity data are evaluated at $\sigma = 0.5$ (where $\sigma = p/p_s$ is pressure p normalized by the local surface pressure p_s), the precipitation rate is taken daily, and as a measure of the frequency of intense precipitation, we use the frequency with which daily precipitation exceeds the latitude-dependent 90th percentile of precipitation rates in a long (18000 days) control simulation. We run the GCM at the coarse horizontal spectral resolution of T21, implying 32 discrete latitudes on the spectral transform grid. Hence, we have 3 statistics, each a function of 32 latitude points, resulting in a 96-dimensional processed output, defined as $\mathcal{H}_T$. In the statistically stationary case, we take the forward map $\mathcal{G}_T = \mathcal{H}_T$.

For the simulations with a seasonal cycle, $\mathcal{H}_T$ is not statistically stationary but is cyclostationary over multiples of a year. The year length in the GCM is 360 days. We stack four 90-day seasons of data together (Howland et al., 2021) and define the forward map

$$\mathcal{G}_T(\boldsymbol{\theta}; \boldsymbol{v}^{(0)}) = [\mathcal{H}_T(\boldsymbol{\theta}; \boldsymbol{v}^{(0)}), \dots, \mathcal{H}_T(\boldsymbol{\theta}; \boldsymbol{v}^{(3)})],$$

over a one year cycle (360 days), where $\boldsymbol{v}^{(i)}$ is the model state at the beginning of each 90-day long season labelled $i = 0, 1, 2, 3$. With this batching, we have now constructed stationary statistics for the stacked data. The theory of Section 2 applies, and our inverse problems can be formulated in the seasonally varying case.

3.3 Design choices

In the stationary GCM setting, we aggregate statistics temporally and zonally. Thus, a local design implies a restriction to certain latitudes. Recall our discretization has 32 discrete latitudes. We therefore choose a design space that contains sets of ℓ consecutive discrete latitudes, indexed from south to north poles with the design points $k = 1, \dots, 32-(\ell-1)$. In the stationary experiments, we choose $\ell = 3$, indexing designs $k = 1, \dots, 30$, unless otherwise specified. The choice of ℓ is discussed in Section 4.1.

In the seasonal GCM setting, we still aggregate in time and longitudinal directions, but we also stack the seasons in a vector. We define a local design by indexing both a

restriction to a season and a restriction to certain latitudes. We choose a design space that contains sets of ℓ consecutive discrete latitudes, collected season by season, indexed from south to north poles as $1, \dots, 32-(\ell-1)$, and from spring to winter as $0, \dots, 3$, all collected as $k = (\text{season, latitude})$. In the seasonal experiments, we choose $\ell = 1$, which indexes the designs $k = (0,1), \dots, (3,32)$.

3.4 Synthetic data and noise

To generate surrogates of locally available data from high-resolution simulations, we generate data with the idealized GCM itself at a fixed parameter vector $\boldsymbol{\theta}^\dagger$, adding Gaussian noise δ with zero mean and covariance matrix Δ as in (2). The implication is that we generate $\boldsymbol{z}_k$ with the restricted idealized GCM $\mathcal{S}_T(\boldsymbol{\theta}^\dagger; k)$, corrupted by noise to describe model error (Kennedy & O'Hagan, 2001b). In this way, the inverse problem (2) can be written in the compact form

$$\boldsymbol{z}_k = \mathcal{S}_\infty(\boldsymbol{\theta}; k) + \gamma_k, \qquad \gamma_k \sim N(0, W_k(\Sigma(\boldsymbol{\theta}) + \Delta)W_k^T). \tag{7}$$

We construct the measurement error covariance matrix Δ to be diagonal with entries $d_i^2 = \Delta_{ii} > 0$, where i indexes over data type (three observed quantities) and over the number of discrete latitudes,

$$\Sigma + \text{diag}(d_i^2) = \Sigma + \Delta. \tag{8}$$

We choose d_i so that it is proportional to the mean μ_i of the variable in question, with a proportionality factor $C_{\max} = 0.1$. To prevent the noise from becoming so large that the variables can cross a physical boundary $\partial\Omega_i$ (e.g., relative humidity becoming negative), we limit the noise standard deviation to a factor $C = 0.2$ times the distance between the approximate 95% noise confidence interval and the physical boundary:

$$d_i = \min\left(C \min\left(\text{dist}(\mu_i + 2\sqrt{\Sigma_{ii}}, \partial\Omega_i), \text{dist}(\mu_i - 2\sqrt{\Sigma_{ii}}, \partial\Omega_i)\right), C_{\max}\mu_i\right).$$

We carry out a set of control simulations, with the parameters fixed to standard values $\boldsymbol{\theta}^\dagger$, where $\mathcal{T}^{-1}(\boldsymbol{\theta}^\dagger) = (0.7, 2\text{ h})$ are standard values used in previous studies (O'Gorman & Schneider, 2008b). We use this set of control simulations to estimate the restricted covariance matrix $W_k\Sigma(\boldsymbol{\theta})W_k^T \approx W_k\Sigma(\boldsymbol{\theta}^\dagger)W_k^T$ for performing uncertainty quantification with local data $\boldsymbol{z}_k$ (stage 2 in Section 2.3). In the statistically stationary case, we carry out control simulations for 650 windows of length $T = 30$ days, discarding the first 50 months for spin-up, and calculate the sample covariance matrix $\Sigma(\boldsymbol{\theta}^\dagger)$ from the latter 600 samples. Here, $W_k\Sigma(\boldsymbol{\theta}^\dagger)W_k^T$ is a symmetric matrix whose size depends on the design space; it represents noise from internal variability in 30-day time averages. In the seasonally varying case, we carry out a control simulation for 150 years, discarding the first 4 years for spin-up, and obtain the sample covariance matrix $\Sigma(\boldsymbol{\theta}^\dagger)$ from the stacked seasonal ($T = 90$ days) averages. In the seasonal case, it is a symmetric matrix whose size depends on 4 times the design space and represents noise from internal variability in the 90-day time averages. In practical implementations of this methodology, good estimates of the local variability that we represent with $W_k\Sigma(\boldsymbol{\theta}^\dagger)W_k^T$ can be made from the observed climatology of the statistics of interest.

For the design stage (stage 1 in Section 2.3) we estimate $\Sigma(\boldsymbol{\theta^*})$ from a second set of control simulations of the GCM in which we fix the parameters to the prior mean $\boldsymbol{\theta}^*$, equivalent to the physical values $\mathcal{T}^{-1}(\boldsymbol{\theta}^*) = (0.5, 12\text{ h})$. In the stationary case, the 3 latitude-dependent fields evaluated at 32 latitude points produce a 96×96 symmetric matrix $\Sigma(\boldsymbol{\theta}^*)$, representing noise from internal variability in 30-day averages; in the seasonal case, the stacked statistics produce a 384×384 symmetric matrix $\Sigma(\boldsymbol{\theta}^*)$, representing noise from variability of 90-day averages. In either case, we take $\Sigma(\boldsymbol{\theta}) = \Sigma(\boldsymbol{\theta}^*)$ in the optimal design stage of the algorithm.

The mean and 95% confidence interval of the data at $\boldsymbol{\theta}^*$, with covariance constructed from $\Sigma(\boldsymbol{\theta}^*)$, are shown in Figure 1 for the statistically stationary case and in Figure 2

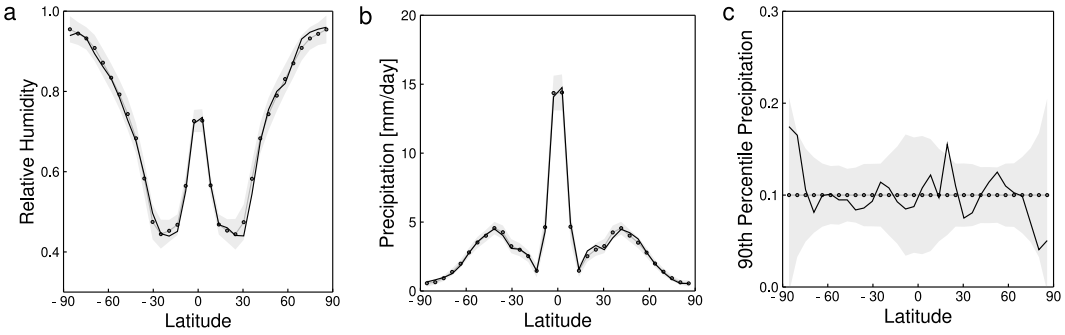

Figure 1. Aggregated climate statistics in the statistically stationary control simulation, with parameters set to the mean of the prior $\boldsymbol{\theta}^*$. The mean (grey lines) and 95% confidence intervals (shading) of the data are plotted against latitude. One realization of the data is shown (black line). No noise is added here.

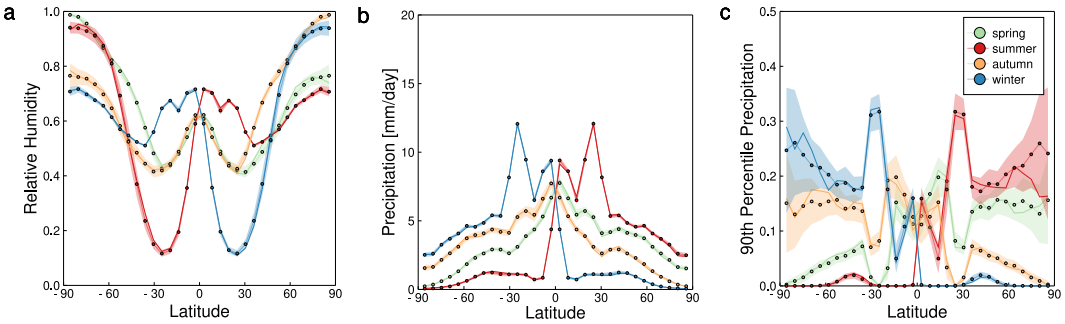

Figure 2. Aggregated climate statistics in the seasonally varying control simulation, with parameters set to the mean of the prior $\boldsymbol{\theta}^*$. The mean (solid lines) and 95% confidence intervals (shading) of the data are plotted against latitude, with the colors indicating the seasons. No noise is added here.

for the seasonally varying case. The black (stationary) and colored (seasonal) solid lines illustrate a realization of the data for one initial condition. Similarly, the mean and 95% confidence interval of the data at $\boldsymbol{\theta}^\dagger$, with noise added with covariance matrix $\Delta+\Sigma(\boldsymbol{\theta}^\dagger)$ (over all designs for illustration), are shown in Figure 3 for the stationary and in Figure 4 for the seasonal case.

4 Results

4.1 Stationary statistics

We first apply the optimal design algorithm to the statistically stationary GCM. The logarithm of the utility function is shown in Figure 5, with four representative samples shown by the colored discs (specifically, these are the design points $k = 15$, 14, 20, and 3, in decreasing order of utility). The extent to which hemispheric symmetry of the statistics is broken in Figure 5 is an indication of sampling variability, as the infinite-time GCM statistics are hemispherically symmetric.

The distribution of the inflated climate statistics produced at the true parameters $\boldsymbol{\theta}^\dagger$ are represented by the mean and 95% confidence interval in grey in Figure 3, which also shows the data samples for each three-latitude design stencil as colored discs for four representative design locations. We apply the uncertainty quantification stage in Sec-

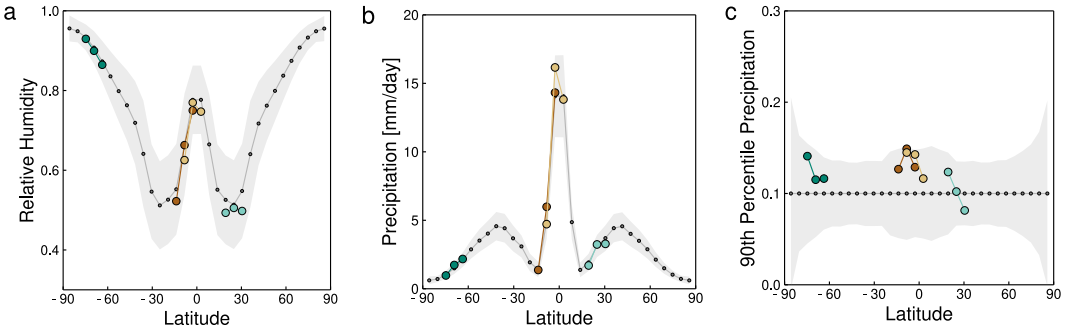

Figure 3. Aggregated climate statistics in the statistically stationary control simulation using the ground truth parameters. Mean (grey lines) and 95% confidence intervals (shading) of the data are plotted against latitude. Additional inflation noise is added. Each set of colored discs represents a 30-day realization of inflated GCM data coming from a different 3-latitude design used in the experiment.

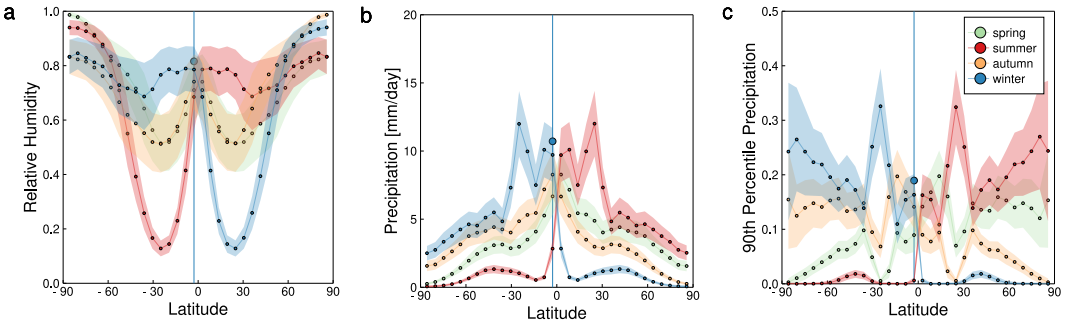

Figure 4. Aggregated climate statistics produced from the seasonally varying control simulation using the ground truth parameters, and with additional inflation. The mean (solid lines) and 95% confidence intervals (shading) of the data are plotted against latitude, with the colors indicating the seasons. The blue vertical line indicates the location and season (northern winter) in which we observe the data for uncertainty quantification; the specific 90-day realization of inflated GCM data for the 1-latitude design is given by the blue disc.

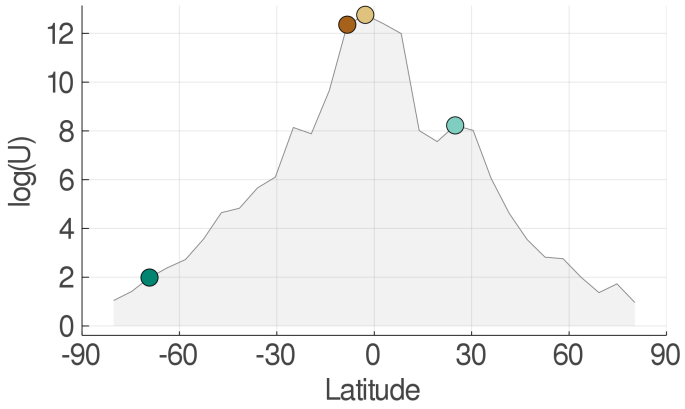

Figure 5. Logarithm of the data utility as a function of latitude, with designs represented by a node at the center of each stencil (comprised of three neighboring latitudes). The colored discs signify the four representative designs indicated in Fig. 3, which are used in the uncertainty quantification experiment.

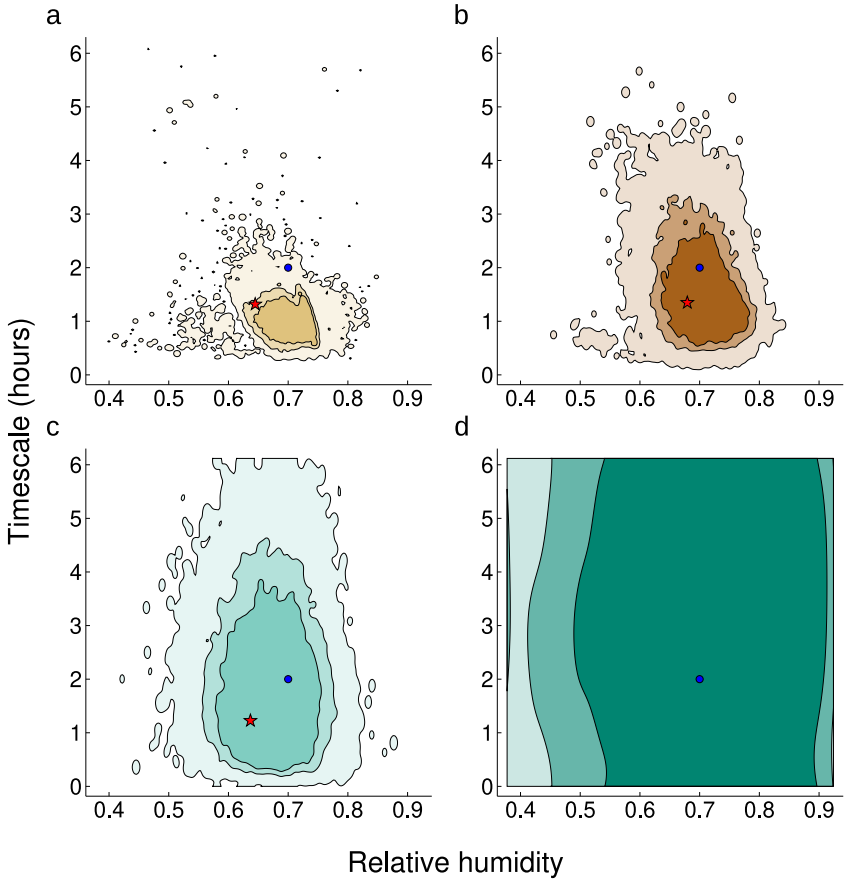

Figure 6. Posterior distributions for convection parameters learned from data restricted to different design points. The drawn contours bound 50%, 75% and 99% of the distribution. Panels (a, b, c, d) correspond to designs $k = (51, 14, 20, 3)$, ordered to express learning from data at decreasingly informative design points (i.e., points of decreasing utility). The true parameter values in the control simulation are given by the blue circle. The parameters found to be optimal in the calibration scheme (given a single random realization of data) are given by the red star in each case (in panel (d) this is outside the plotting region).

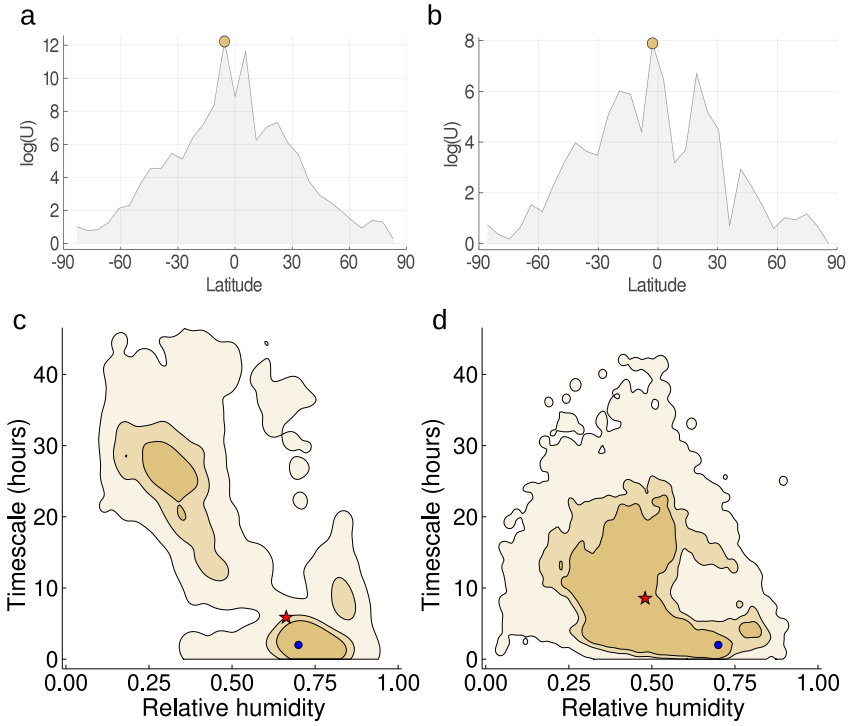

Figure 7. Performance for different optimal design selections at smaller stencil sizes. The contours bound 50%, 75%, and 99% of the distribution. Panels (a,c) is a two-latitude stencil, and panels (b,d) is a one-latitude stencil. The top row displays the logarithm of the utility plot, and the bottom row the corresponding posterior from a sample at the optimal latitude, marked by a disc at the top.

tion 2.3 to each location, resulting in the posterior distributions shown in Figure 6. Each panel shows the density contours bounding 50%, 75%, and 99% of the posterior distribution, shaded dark to light; the priors are largely uninformative and have been excluded from the plots. The panels are ordered (a – d) by decreasing D-utility, a predictor of information content based on uncertainty at the prior mean parameter $\boldsymbol{\theta}^*$. We see this monotonicity is preserved when considering data produced from the true parameter $\boldsymbol{\theta}^\dagger$ in this example. In particular, this implies that the design optimizing the chosen utility produces minimal uncertainty in the uncertainty quantification stage. As observed in other investigations (Dunbar et al., 2021), the posterior distributions are subject to variability due to the finite-time sampling and the inflation. However, all distributions capture the true parameter values within 99% of the posterior mass.

For the statistically stationary case, we investigate the choice of ℓ, a measure of the design sparsity. To this end we repeat the experiment, choosing $\ell = 2$ or 1 in Section 3.3). For each, Figure 7 shows the utility function against the latitude at the center of the stencil and the posterior distribution at the respective optimal designs. We see that in both cases, the optimal design remains robust, coinciding with the three-stencil case. Peak utility is consistently at a design near the equator. The posterior distributions are seen to be far broader than in the three-latitude case, offering only marginal improvement over the prior distribution in the one-latitude stencil case. They are non-Gaussian and multimodal but nevertheless capture the true parameters (blue disc) with high probability. They provide insight into the correlation structure between the parameters at the optimal design location. We observe that for these sparser designs, non-identifiability (mul-

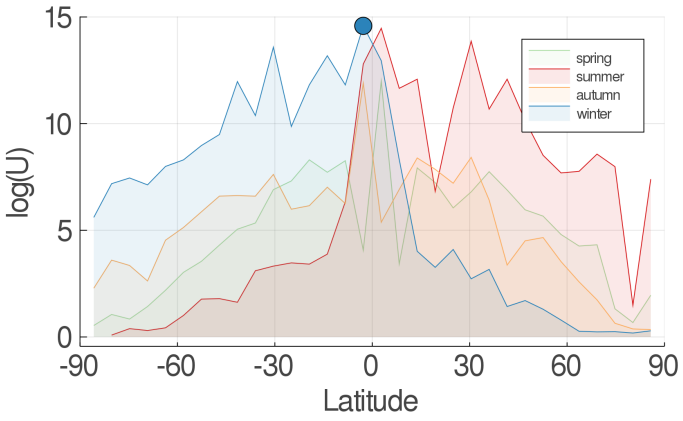

Figure 8. Logarithm of the data utility plotted against latitude (1 design per latitude and season). The blue disc signifies that a latitude in northern winter maximizes the utility function.

timodality) appears only at data from $\boldsymbol{\theta}^{\dagger}$, but not at $\boldsymbol{\theta}^{*}$. As a result, the optimal uncertainty is not guaranteed to be found at the location of optimal utility. This is remedied by having a better initial guess through the prior, or (as demonstrated in the three-latitude set) a less sparse data set from which the parameters are more identifiable.

4.2 Seasonally varying statistics

In the seasonally varying case, we choose the optimal design with the algorithm in Section 2.3 applied to the stacked data. Figure 8 shows the logarithm of the utility function. Hemispheric and seasonal asymmetries are evident here. In northern winter, latitudes just south of the the equator ($k = (3, 16)$) optimize the design, in the vicinity of the ITCZ. Conversely, in northern summer, latitudes just north of the equator ($k = (1, 17)$) optimize the design, again in the vicinity of the seasonally migrating ITCZ; additional peaks at around 30 degrees can be seen. The equinox seasons have less utility at the optimal designs ($k = (0, 17)$ and $k = (2, 16)$). Because the equinoctial Hadley cells and ascent regions in the ITCZ are less pronounced than the solstitial Hadley cells (Schneider et al., 2010), utility is more spread out across the latitudes.

We solve the analogue inverse problem (7) as in the nonseasonal case with a data sample at the optimal spatial design location for each season. The posterior distributions are collected in Figure 9, colored by season. In general, the true parameter values lie in regions of high posterior density in each case. We see qualitatively that the utility of the different designs predicts the the size of support of the corresponding posterior distribution, in particular the design with highest utility (northern winter) also has the smallest support. This indicates that the utility is still a good predictor of data quality for learning the convection parameters in the cyclo-stationary settings.

5 Conclusions and Discussion

We have presented a novel framework for automated optimal placement of high-resolution simulations embedded in lower-resolution models. The framework can be used with computationally expensive and chaotic (noisy) low-resolution models, whose derivatives may not be available. Given low-resolution simulations, we use parameter uncertainty information provided by the CES algorithm to guide our choice of design. We have demonstrated the efficacy of the algorithm for finding optimally informative locations in perfect-model settings in which we generated surrogates of embedded high-resolution

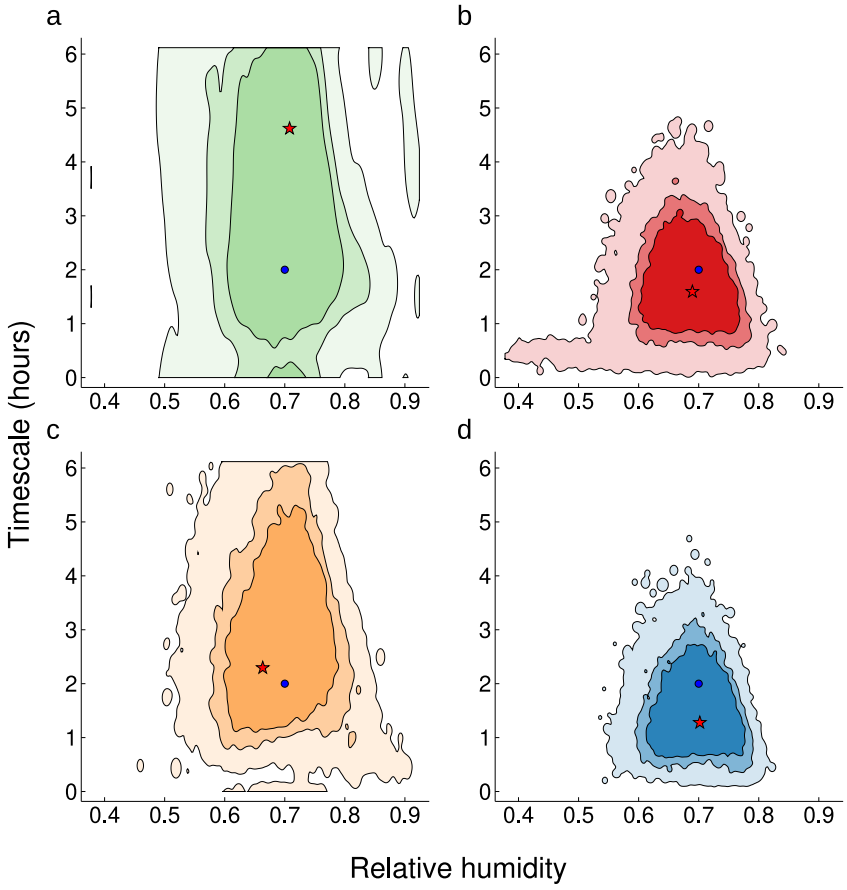

Figure 9. Posterior distribution obtained from using data at the optimal latitude from each season. Contours bound 50%, 75% and 99% of the distribution (in decreasing color saturation). Panels (a, b, c, d) correspond to designs $k = ((0, 17), (1, 17), (2, 16), (3, 16))$, ordered by season. The true parameter values in the control simulation are given by the blue circle. The parameters found to be optimal in the calibration scheme (given a single random realization of data) are given by the red star in each case.

simulations directly with an idealized GCM, with statistically stationary or seasonally varying statistics. In these settings, we have demonstrated learning about parameters in a convection parameterization, exploring both spatial and spatio-temporal designs. Our design framework can also be used more broadly, to automate selection of optimally informative climate statistics from libraries of high-resolution simulations (Shen et al., 2021).

With the idealized GCM, we showed how to optimally target a location at which additional data will produce parameter estimates that minimize uncertainty. In our proof-of-concept in which we calibrated parameters in a simple convection scheme, the automatically targeted optimal location for new data was consistently near the equator, in the vicinity of the seasonally migrating ITCZ. This is consistent with the fact that the convection scheme in the idealized GCM is most important near the ITCZ (O'Gorman & Schneider, 2008). We showed that the optimal targeting is limited in its effectiveness in settings of very sparse data, where parameter posteriors can be multimodal. However, with access to the posterior distributions of the parameters, the behavior is both diagnosable a posteriori, and actionable with successive iterations of the optimal design process (simply take the current posterior as the prior for a subsequent iteration with additional data).

The design algorithm is very efficient with respect to evaluations of the GCM and the high-resolution model. Due to our integration of the design framework within the CES algorithm (detailed in Appendix A), only a relatively modest $\mathcal{O}(100)$ forward model evaluations of the GCM are required for the design selection process; no evaluations of the high-resolution model (or, in our proof-of-concept, of the surrogate for it) are required. The calculation of the utility function can be performed in an embarrassingly parallel fashion. Thus, for moderately sized design spaces, the computational cost is dominated by the cost of evaluation of the GCM.

Despite being efficient, the current algorithm relies on evaluating utilities naively at all design points. In a practical climate model application, where we may have 10^2 LES that are computationally affordable to be placed optimally within 10^6 or more possible locations, such naive approaches are infeasible. Instead, one can use more sophisticated optimization algorithms. For determinant based (i.e., D-optimal) utilities, this typically requires accelerating the determinant evaluation (and its gradients). Various methods have been developed to do so, e.g., using Laplace approximations (Long et al., 2013; Beck et al., 2018; Rue et al., 2009), polynomial chaos surrogates (Huan & Marzouk, 2014), optimization of criteria bounds (Tsilifis et al., 2017), fast random determinant approximation (Alexanderian et al., 2014; Alexanderian & Saibaba, 2018), and Gaussian process surrogates (Buathong et al., 2020; Paglia et al., 2020). The latter, kernel-based approaches are particularly amenable to our setting, as they give sparse representations of the utility function that are independent of the underlying computational grid. They may offer a way forward in the climate modeling setting.

Acknowledgments

We gratefully acknowledge the generous support of Eric and Wendy Schmidt (by recommendation of Schmidt Futures) and the National Science Foundation (grant AGS-1835860). The simulations were performed on Caltech's High Performance Cluster, which is partially supported by a grant from the Gordon and Betty Moore Foundation. AMS is also supported by the Office of Naval Research (grant N00014-17-1-2079).

Data Availability. All computer code used in this paper is open source. The code for the idealized GCM, the Julia code for the optimal design algorithm, the plot tools, and the slurm/bash scripts to run both GCM and design algorithms are available at:

https://doi.org/10.5281/zenodo.5835269.

References

Alexanderian, A., Petra, N., Stadler, G., & Ghattas, O. (2014). A-optimal design of experiments for infinite-dimensional bayesian linear inverse problems with regularized \ell_0-sparsification. *SIAM Journal on Scientific Computing*, *36*(5), A2122–A2148.

Alexanderian, A., & Saibaba, A. K. (2018). Efficient d-optimal design of experiments for infinite-dimensional bayesian linear inverse problems. *SIAM Journal on Scientific Computing*, *40*(5), A2956–A2985.

Beck, J., Dia, B. M., Espath, L. F., Long, Q., & Tempone, R. (2018). Fast bayesian experimental design: Laplace-based importance sampling for the expected information gain. *Computer Methods in Applied Mechanics and Engineering*, *334*, 523 - 553.

Bischoff, T., & Schneider, T. (2014). Energetic constraints on the position of the Intertropical Convergence Zone. *J. Climate*, *27*, 4937–4951. doi: 10.1175/JCLI-D-13-00650.1

Bishop, C. H., & Toth, Z. (1999). Ensemble transformation and adaptive observations. *Journal of the atmospheric sciences*, *56*(11), 1748–1765.

Bony, S., Colman, R., Kattsov, V. M., Allan, R. P., Bretherton, C. S., Dufresne, J.-L., ... Webb, M. J. (2006). How well do we understand and evaluate climate change feedback processes? *J. Climate*, *19*, 3445–3482. doi: 10.1175/JCLI3819.1

Bony, S., & Dufresne, J. L. (2005). Marine boundary layer clouds at the heart of tropical cloud feedback uncertainties in climate models. *Geophys. Res. Lett.*, *32*, L20806.

Bordoni, S., & Schneider, T. (2008a). Monsoons as eddy-mediated regime transitions of the tropical overturning circulation. *Nature Geoscience*, *1*(8), 515–519.

Bordoni, S., & Schneider, T. (2008b). Monsoons as eddy-mediated regime transitions of the tropical overturning circulation. *Nature Geosci.*, *1*, 515–519. doi: 10.1038/ngeo248

Brient, F., & Schneider, T. (2016). Constraints on climate sensitivity from space-based measurements of low-cloud reflection. *J. Climate*, *29*, 5821–5835. doi: 10.1175/JCLI-D-15-0897.1

Buathong, P., Ginsbourger, D., & Krityakierne, T. (2020). Kernels over sets of finite sets using RKHS embeddings, with application tobBayesian (combinatorial) optimization. In *International conference on artificial intelligence and statistics* (pp. 2731–2741).

Campin, J.-M., Hill, C., Jones, H., & Marshall, J. (2011). Super-parameterization in ocean modeling: Application to deep convection. *Ocean Modelling*, *36*(1), 90 - 101. doi: https://doi.org/10.1016/j.ocemod.2010.10.003

Cess, R. D., Potter, G., Blanchet, J., Boer, G., Ghan, S., Kiehl, J., ... others (1989). Interpretation of cloud-climate feedback as produced by 14 atmospheric general circulation models. *Science*, *245*, 513–516.

Cess, R. D., Potter, G. L., Blanchet, J. P., Boer, G. J., Del Genio, A. D., Déqué, M., ... Zhang, M.-H. (1990). Intercomparison and interpretation of climate feedback processes in 19 atmospheric general circulation models. *J. Geophys. Res.*, *95*, 16601–16615. doi: 10.1029/JD095iD10p16601

Chaloner, K., & Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, *10*(3), 273–304.

Chen, Y., & Oliver, D. S. (2012). Ensemble randomized maximum likelihood method as an iterative ensemble smoother. *Mathematical Geosciences*, *44*(1), 1–26.

Cleary, E., Garbuno-Inigo, A., Lan, S., Schneider, T., & Stuart, A. M. (2021a). Calibrate, emulate, sample. *Journal of Computational Physics*, *424*, 109716. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0021999120304903` doi: https://doi.org/10.1016/j.jcp.2020.109716

Cleary, E., Garbuno-Inigo, A., Lan, S., Schneider, T., & Stuart, A. M. (2021b). Calibrate, emulate, sample. *J. Comp. Phys.*, *424*, 109716. Retrieved from 10.1016/j.jcp.2020.109716

Cook, A. R., Gibson, G. J., & Gilligan, C. A. (2008). Optimal observation times in experimental epidemic processes. *Biometrics*, *64*(3), 860-868.

Couvreux, F., Hourdin, F., Williamson, D., Roehrig, R., Volodina, V., Villefranque, N., ... others (2021). Process-based climate model development harnessing machine learning: I. A calibration tool for parameterization improvement. *J. Adv. Model. Earth Sys.*, *13*, e2020MS002217. doi: 10.1029/2020MS002217

Dashti, M., & Stuart, A. M. (2013). The Bayesian approach to inverse problems. *arXiv preprint arXiv:1302.6989*.

de Rooy, W. C., Bechtold, P., Fröhlich, K., Hohenegger, C., Jonker, H., Mironov, D., ... Yano, J.-I. (2013). Entrainment and detrainment in cumulus convection: an overview. *Quart. J. Roy. Meteor. Soc.*, *139*, 1–19.

Dunbar, O. R. A., Garbuno-Inigo, A., Schneider, T., & Stuart, A. M. (2021). Calibration and uncertainty quantification of convective parameters in an idealized GCM. *J. Adv. Model. Earth Sys.*, *13*, e2020MS002454. doi: 10.1029/2020MS002454

Duncan, A. B., Stuart, A. M., & Wolfram, M.-T. (2021). Ensemble inference methods for models with noisy and expensive likelihoods. *arXiv preprint arXiv:2104.03384*.

Emanuel, K., Raymond, D., Betts, A., Bosart, L., Bretherton, C., Droegemeier, K., ... others (1995). Report of the first prospectus development team of the us weather research program to noaa and the nsf. *Bulletin of the American Meteorological Society*, 1194–1208.

Emerick, A. A., & Reynolds, A. C. (2013). Ensemble smoother with multiple data assimilation. *Computers & Geosciences*, *55*, 3–15.

Fox-Kemper, B., & Menemenlis, D. (2013). Can large eddy simulation techniques improve mesoscale rich ocean models? In *Ocean modeling in an eddying regime* (p. 319-337). American Geophysical Union (AGU). doi: 10.1029/177GM19

Frierson, D. M. (2007). The dynamics of idealized convection schemes and their effect on the zonally averaged tropical circulation. *Journal of the Atmospheric Sciences*, *64*(6), 1959–1976.

Frierson, D. M. W. (2007). The dynamics of idealized convection schemes and their effect on the zonally averaged tropical circulation. *J. Atmos. Sci.*, *64*, 1959–1976.

Frierson, D. M. W., Held, I. M., & Zurita-Gotor, P. (2006). A gray-radiation aquaplanet moist GCM. Part I: Static stability and eddy scale. *J. Atmos. Sci.*, *63*, 2548–2566.

Geyer, C. J. (2011). Introduction to markov chain monte carlo. In S. Brooks, A. Gelman, G. L. Jones, & X.-L. Meng (Eds.), *Handbook of markov chain monte carlo* (pp. 3–48). Chapman and Hall/CRC.

Hohenegger, C., & Bretherton, C. S. (2011). Simulating deep convection with a shallow convection scheme. *Atmos. Chem. Phys.*, *11*, 10389–10406. doi: 10.5194/acp-11-10389-2011

Hourdin, F., Williamson, D., Rio, C., Couvreux, F., Roehrig, R., Villefranque, N., ... Volodina, V. (2021). Process-based climate model development harnessing machine learning: II. model calibration from single column to global. *J. Adv. Model. Earth Sys.*, *13*, e2020MS002225. doi: 10.1029/2020MS002225

Howland, M. F., Dunbar, O. R. A., & Schneider, T. (2021). Parameter uncertainty quantification in an idealized gcm with a seasonal cycle. *arXiv preprint arXiv:2108.00827*.

Huan, X., & Marzouk, Y. (2014). Gradient-based stochastic optimization methods in bayesian experimental design. *International Journal for Uncertainty Quantification*, *4*(6).

Huan, X., & Marzouk, Y. M. (2013, 1). Simulation-based optimal bayesian experimental design for nonlinear systems. *Journal of Computational Physics*, *232*(1).

Iglesias, M. A., Law, K. J., & Stuart, A. M. (2013). Ensemble kalman methods for inverse problems. *Inverse Problems*, *29*(4), 045001.

Kaipio, J., & Somersalo, E. (2006). *Statistical and computational inverse problems* (Vol. 160). Springer Science & Business Media.

Kalnay, E. (2003). *Atmospheric modeling, data assimilation and predictability.* Cambridge, UK: Cambridge Univ. Press.

Kaspi, Y., & Schneider, T. (2011). Winter cold of eastern continental boundaries induced by warm ocean waters. *Nature*, *471*, 621–624.

Kaspi, Y., & Schneider, T. (2013). The role of stationary eddies in shaping midlatitude storm tracks. *J. Atmos. Sci.*, *70*, 2596–2613.

Kennedy, M. C., & O'Hagan, A. (2000). Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, *87*(1), 1–13.

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

Kennedy, M. C., & O'Hagan, A. (2001b). Bayesian calibration of computer models. *J. Roy. Statist. Soc. B*, *63*, 425–464. doi: 10.1111/1467-9868.00294

Khairoutdinov, M. F., Krueger, S. K., Moeng, C.-H., Bogenschutz, P. A., & Randall, D. A. (2009). Large-eddy simulation of maritime deep tropical convection. *J. Adv. Model. Earth Sys.*, *1*, Art. #15, 13 pp. doi: 10.3894/JAMES.2009.1.15

Kim, W., Pitt, M. A., Lu, Z.-L., Steyvers, M., & Myung, J. I. (2014). A hierarchical adaptive approach to optimal experimental design. *Neural Computation*, *26*(11), 2465-2492.

Levine, M. E., & Stuart, A. M. (2021). A framework for machine learning of model error in dynamical systems. (https://arxiv.org/abs/2107.06658)

Levine, X., & Schneider, T. (2015). Baroclinic eddies and the extent of the Hadley circulation: An idealized GCM study. *J. Atmos. Sci.*, *72*, 2744–2761. doi: 10.1175/JAS-D-14-0152.1

Li, Q., & Fox-Kemper, B. (2017). Assessing the effects of langmuir turbulence on the entrainment buoyancy flux in the ocean surface boundary layer. *Journal of Physical Oceanography*, *47*(12), 2863–2886.

Li, Q., Reichl, B. G., Fox-Kemper, B., Adcroft, A. J., Belcher, S. E., Danabasoglu, G., ... others (2019). Comparing ocean surface boundary vertical mixing schemes including langmuir turbulence. *Journal of Advances in Modeling Earth Systems*, *11*(11), 3545–3592.

Liu, C., Moncrieff, M. W., & Grabowski, W. W. (2001). Hierarchical modelling of tropical convective systems using explicit and parametrized approaches. *Quart. J. Roy. Meteor. Soc.*, *127*, 493–515.

Long, Q., Scavino, M., Tempone, R., & Wang, S. (2013). Fast estimation of expected information gains for bayesian experimental designs based on laplace approximations. *Computer Methods in Applied Mechanics and Engineering*, *259*, 24–39.

Lorenz, E. N., & Emanuel, K. A. (1998). Optimal sites for supplementary weather observations: Simulation with a small model. *J. Atmos. Sci.*, *55*, 399–414. doi: 10.1175/1520-0469(1998)055⟨0399:OSFSWO⟩2.0.CO;2

Matheou, G., & Chung, D. (2014). Large-eddy simulation of stratified turbulence. Part II: Application of the stretched-vortex model to the atmospheric boundary layer. *J. Atmos. Sci.*, *71*, 4439–4460. doi: 10.1175/JAS-D-13-0306.1

Merlis, T. M., & Schneider, T. (2011). Changes in zonal surface temperature gradients and walker circulations in a wide range of climates. *J. Climate*, *24*, 4757–4768.

Notz, W. I., Santner, T. J., & Williams, B. J. (2018). *The design and analysis of computer experiments* (2nd ed. ed.). Springer.

O'Gorman, P. A. (2011). The effective static stability experienced by eddies in a moist atmosphere. *J. Atmos. Sci.*, *68*, 75–90.

O'Gorman, P. A., Lamquin, N., Schneider, T., & Singh, M. S. (2011). The relative humidity in an isentropic advection–condensation model: Limited poleward influence and properties of subtropical minima. *J. Atmos. Sci.*, *68*, 3079–3093.

O'Gorman, P. A., & Schneider, T. (2008a). Energy of midlatitude transient eddies in idealized simulations of changed climates. *J. Climate*, *21*, 5797–5806.

O'Gorman, P. A., & Schneider, T. (2008b). The hydrological cycle over a wide range of climates simulated with an idealized GCM. *J. Climate*, *21*, 3815–3832.

O'Gorman, P. A., & Schneider, T. (2009a). The physical basis for increases in precipitation extremes in simulations of 21st-century climate change. *Proc. Natl. Acad. Sci.*, *106*, 14773–14777.

O'Gorman, P. A., & Schneider, T. (2009b). Scaling of precipitation extremes over a wide range of climates simulated with an idealized GCM. *J. Climate*, *22*, 5676–5685.

Oliver, D. S., Reynolds, A. C., & Liu, N. (2008). *Inverse theory for petroleum reservoir characterization and history matching.* Cambridge Univ. Press.

O'Gorman, P. A., & Schneider, T. (2008). The hydrological cycle over a wide range of climates simulated with an idealized gcm. *Journal of Climate*, *21*(15), 3815-3832. doi: 10.1175/2007JCLI2065.1

Paglia, J., Eidsvik, J., & Karvanen, J. (2020). Efficient spatial designs using hausdorff distances and bayesian optimisation. *Statistical modeling for safer drilling operations*, 77.

Paninski, L. (2005). Asymptotic theory of information-theoretic experimental design. *Neural Computation*, *17*(7), 1480-1507.

Pressel, K. G., Kaul, C. M., Schneider, T., Tan, Z., & Mishra, S. (2015). Large-eddy simulation in an anelastic framework with closed water and entropy balances. *J. Adv. Model. Earth Sys.*, *7*, 1425–1456. doi: 10.1002/2015MS000496

Pressel, K. G., Mishra, S., Schneider, T., Kaul, C. M., & Tan, Z. (2017). Numerics and subgrid-scale modeling in large eddy simulations of stratocumulus clouds. *J. Adv. Model. Earth Sys.*, *9*, 1342–1365. doi: 10.1002/2016MS000778

Reich, S. (2011). A dynamical systems framework for intermittent data assimilation. *BIT Numerical Mathematics*, *51*(1), 235–249.

Reichl, B. G., & Hallberg, R. (2018). A simplified energetics based planetary boundary layer (epbl) approach for ocean climate simulations. *Ocean Modelling*, *132*, 112 - 129. doi: https://doi.org/10.1016/j.ocemod.2018.10.004

Reichl, B. G., Wang, D., Hara, T., Ginis, I., & Kukulka, T. (2016). Langmuir turbulence parameterization in tropical cyclone conditions. *Journal of Physical Oceanography*, *46*(3), 863–886.

Romps, D. M. (2016). The Stochastic Parcel Model: A deterministic parameterization of stochastically entraining convection. *J. Adv. Model. Earth Sys.*, *8*, 319–344. doi: 10.1002/2015MS000537

Rue, H., Martino, S., & Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*(2), 319-392.

Ryan, E. G., Drovandi, C. C., McGree, J. M., & Pettitt, A. N. (2016). A review of modern computational algorithms for bayesian optimal design. *International Statistical Review*, *84*(1), 128–154.

Ryan, E. G., Drovandi, C. C., Thompson, M. H., & Pettitt, A. N. (2014). Towards bayesian experimental design for nonlinear models that require a large number of sampling times. *Computational Statistics & Data Analysis*, *70*, 45 - 60.

Schalkwijk, J., Jonker, H. J. J., Siebesma, A. P., & Van Meijgaard, E. (2015). Weather forecasting using GPU-based large-eddy simulations. *Bull. Amer. Meteor. Soc.*, *96*, 715–723. doi: 10.1175/BAMS-D-14-00114.1

Schillings, C., & Stuart, A. M. (2017). Analysis of the ensemble kalman filter for inverse problems. *SIAM Journal on Numerical Analysis*, *55*(3), 1264–1290.

Schneider, T., Lan, S., Stuart, A., & Teixeira, J. (2017). Earth system modeling 2.0: A blueprint for models that learn from observations and targeted high-resolution simulations. *Geophys. Res. Lett.*, *44*, 12396–12417. doi: 10.1002/2017GL076101

Schneider, T., & O'Gorman, P. A. (2008). Moist convection and the thermal stratification of the extratropical troposphere. *J. Atmos. Sci.*, *65*, 3571–3583.

Schneider, T., O'Gorman, P. A., & Levine, X. J. (2010). Water vapor and the dynamics of climate changes. *Rev. Geophys.*, *48*, RG3001. (doi:10.1029/2009RG000302)

Schneider, T., Stuart, A. M., & Wu, J. (2021). Imposing sparsity within ensemble Kalman inversion.

Schneider, T., Teixeira, J., Bretherton, C. S., Brient, F., Pressel, K. G., Schär, C., & Siebesma, A. P. (2017). Climate goals and computing the future of clouds. *Nature Climate Change*, *7*, 3–5. doi: 10.1038/nclimate3190

Shen, Z., Pressel, K. G., Tan, Z., & Schneider, T. (2020). Statistically steady state large-eddy simulations forced by an idealized GCM: 1. forcing framework and simulation characteristics. *J. Adv. Model. Earth Sys.*, *12*. doi: 10.1029/2019MS001814

Shen, Z., Sridhar, A., Tan, Z., Jaruga, A., & Schneider, T. (2021). A library of large-eddy simulations for calibrating cloud parameterizations. *https://essoar.org*. doi: https://doi.org/10.1002/essoar.10507112.1

Siebesma, A. P., Bretherton, C. S., Brown, A., Chlond, A., Cuxart, J., Duynkerke, P. G., ... Stevens, D. E. (2003). A large eddy simulation intercomparison study of shallow cumulus convection. *J. Atmos. Sci.*, *60*, 1201–1219.

Siebesma, A. P., Soares, P. M. M., & Teixeira, J. (2007). A combined eddy-diffusivity mass-flux approach for the convective boundary layer. *J. Atmos. Sci.*, *64*, 1230–1248. doi: 10.1175/JAS3888.1

Smalley, M., Suselj, K., Lebsock, M., & Teixeira1, J. (2019). A novel framework for evaluating and improving parameterized subtropical marine boundary layer cloudiness. *Mon. Wea. Rev.*, *147*, 3241–3260.

Souza, A. N., Wagner, G. L., Ramadhan, A., Allen, B., Churavy, V., Schloss, J., ... Ferrari, R. (2020). Uncertainty quantification of ocean parameterizations: Application to the k-profile-parameterization for penetrative convection. *Journal of Advances in Modeling Earth Systems*, *12*(12), e2020MS002108. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020MS002108` (e2020MS002108 10.1029/2020MS002108) doi: https://doi.org/10.1029/2020MS002108

Stephens, G. L. (2005). Cloud feedbacks in the climate system: A critical review. *J. Climate*, *18*, 237–273. doi: 10.1175/JCLI-3243.1

Stevens, B., Moeng, C.-H., Ackerman, A. S., Bretherton, C. S., Chlond, A., de Roode, S., ... Zhu, P. (2005). Evaluation of large-eddy simulations via observations of nocturnal marine stratocumulus. *Mon. Wea. Rev.*, *133*, 1443–1462. doi: 10.1175/MWR2930.1

Stuart, A. M. (2010). Inverse problems: a Bayesian perspective. *Acta Numerica*, *19*, 451–559.

Tan, Z., Kaul, C. M., Pressel, K. G., Cohen, Y., Schneider, T., & Teixeira, J. (2018). An extended eddy-diffusivity mass-flux scheme for unified representation of subgrid-scale turbulence and convection. *J. Adv. Model. Earth Sys.*, *10*, 770-800. doi: 10.1002/2017MS001162

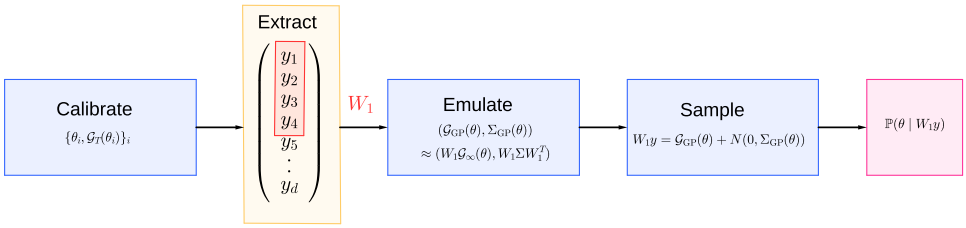

Figure A1. Procedure of the uncertainty quantification framework (blue), to produce output (pink). A restriction operator W_1 extracting a subset of the GCM output (yellow); the subsequent emulate and sample stages may be performed in parallel for different W_i.

Tarantola, A. (2005). *Inverse problem theory and methods for model parameter estimation* (Vol. 89). siam.

Tsilifis, P., Ghanem, R. G., & Hajali, P. (2017). Efficient bayesian experimentation using an expected information gain lower bound. *SIAM/ASA Journal on Uncertainty Quantification*, *5*(1), 30–62.

Van Roekel, L., Adcroft, A. J., Danabasoglu, G., Griffies, S. M., Kauffman, B., Large, W., ... Schmidt, M. (2018). The kpp boundary layer scheme for the ocean: Revisiting its formulation and benchmarking one-dimensional simulations relative to les. *Journal of Advances in Modeling Earth Systems*, *10*(11), 2647–2685.

Vial, J., Dufresne, J.-L., & Bony, S. (2013). On the interpretation of inter-model spread in CMIP5 climate sensitivity estimates. *Clim. Dyn.*, *41*, 3339–3362. doi: 10.1007/s00382-013-1725-9

Wang, D., Large, W. G., & McWilliams, J. C. (1996). Large-eddy simulation of the equatorial ocean boundary layer: Diurnal cycling, eddy viscosity, and horizontal rotation. *Journal of Geophysical Research: Oceans*, *101*(C2), 3649–3662.

Webb, M. J., Lambert, F. H., & Gregory, J. M. (2013). Origins of differences in climate sensitivity, forcing and feedback in climate models. *Clim. Dyn.*, *40*, 677–707. doi: 10.1007/s00382-012-1336-x

Wei, H.-H., & Bordoni, S. (2018). Energetic constraints on the ITCZ position in idealized simulations with a seasonal cycle. *J. Adv. Model. Earth Sys.*, *10*. doi: 10.1029/2018MS001313

Williams, C. K., & Rasmussen, C. E. (2006). *Gaussian processes for machine learning* (Vol. 2) (No. 3). MIT press Cambridge, MA.

Wills, R. C., Levine, X. J., & Schneider, T. (2017). Local energetic constraints on Walker circulation strength. *J. Atmos. Sci.*, *74*, 1907–1922. doi: 10.1175/JAS-D-16-0219.1

Zhang, F., Sun, Y. Q., Magnusson, L., Buizza, R., Lin, S.-J., Chen, J.-H., & Emanuel, K. (2019). What is the predictability limit of midlatitude weather? *Journal of the Atmospheric Sciences*, *76*(4), 1077 - 1091. Retrieved from `https://journals.ametsoc.org/view/journals/atsc/76/4/jas-d-18-0269.1.xml` doi: 10.1175/JAS-D-18-0269.1

Zhang, M., Bretherton, C. S., Blossey, P. N., Austin, P. H., Bacmeister, J. T., Bony, S., ... others (2013). CGILS: Results from the first phase of an international project to understand the physical mechanisms of low cloud feedbacks in general circulation models. *J. Adv. Model. Earth Sys.*, *5*, 826–842. doi: 10.1002/2013MS000246

Appendix A Calibrate-Emulate-Sample with design

One fundamental aspect of this work, is the ability to efficiently calculate the the posterior distribution (in particular the covariance), which is needed to calculate the utility function (6) at all designs. We present a methodology: calibrate-extract-emulate-sample, (CEES) which allows for the calculation of posterior covariance for all designs with just $\mathcal{O}(100)$ evaluations of our forward model.

The methodology is based on the calibrate-emulate-sample (CES) algorithm, for full details of the individual stages see Cleary et al. (2021a); Dunbar et al. (2021), here we present an overview and motivation. The core purpose of CES is to form a computationally cheap statistical emulator of $\mathcal{G}_\infty$ from intelligently chosen samples of $\mathcal{G}_T$; then one is able to solve the Bayesian inverse problem for the emulated $\mathcal{G}_\infty$ with a sampling method. We achieve this by using Gaussian process emulators, trained on the samples of the (noisy and expensive) forward map. The Gaussian process mean function is naturally smoother than the data it is trained on (Kennedy & O'Hagan, 2001a; Notz et al., 2018), and is capable of representing the the noise of the forward model within the covariance function, leading to a smooth likelihood function that is quick to evaluate. The training points for the Gaussian Process are given by applying an optimization scheme, EKI (Ensemble Kalman Inversion), (Iglesias et al., 2013; Schillings & Stuart, 2017) to the inverse problem in its finite-time averaged form (3). Theoretical work shows that noisy continuous-time versions of EKI exhibit an averaging effect that skips over fluctuations superimposed onto the ergodic averaged forward model (Duncan et al., 2021), and similar effects are observed in practice for EKI, thus it is highly suited to optimization of parameters coming from a noisy, expensive model without derivatives available. Ensemble Kalman methods are scalable to very high dimensional problems (Kalnay, 2003; Oliver et al., 2008) with use of localization and regularization.

Let D index a finite space of designs. Given a time $T > 0$, and prior on $\boldsymbol{\theta}$ with prior mean $\boldsymbol{\theta}^*$. Draw a sample $\boldsymbol{y} = \mathcal{G}_T(\boldsymbol{\theta}^*, \boldsymbol{v}^{(0)})$, from any initial condition $\boldsymbol{v}^{(0)}$:

1. **Calibrate:** We solve (3) with $\boldsymbol{y}$ using evaluations of $\mathcal{G}_T$ in an optimization sense, where we minimize the functional.
$$\Phi_T(\boldsymbol{\theta}, \boldsymbol{y}) = \|\boldsymbol{y} - \mathcal{G}_T(\boldsymbol{\theta}; \boldsymbol{v}^{(0)})\|^2_{2\Sigma}. \tag{A1}$$
The notation $\|\cdot\|_\Sigma = \|\Sigma^{-\frac{1}{2}} \cdot \|_2$ is the Mahalanobis distance. The weight 2Σ is the sum of internal variability of $\mathcal{G}_T$ and of $\boldsymbol{y}$. The optimization is performed using several iterations the Ensemble Kalman Inversion algorithm. This leads to $\{\boldsymbol{\theta}_i, \mathcal{G}_j(\boldsymbol{\theta}_j)\}_{j=1}^J$ of input-output pairs that are localized around the optimal parameter value.
2. **Extract:** For each design $k \in D$, we apply the restriction mapping W_k to the forward map, $\{\boldsymbol{\theta}_j, W_k\mathcal{G}_T(\boldsymbol{\theta}_j)\}_{j=1}^J$, and apply the following **Emulate(k)** and **Sample(k)** stages.
3. **Emulate(k):** We decorrelate the data space with an SVD on the internal variability covariance Σ, yielding a change-of-basis matrix V. We train Gaussian process emulators, on the pairs $\{\boldsymbol{\theta}_j, VW_k\mathcal{G}_T(\boldsymbol{\theta}_j)\}_{j=1}^J$, yielding $(\mathcal{G}_{\text{GP}}(\boldsymbol{\theta}), \Sigma_{\text{GP}}(\boldsymbol{\theta}))$, where $\mathcal{G}_{\text{GP}} \approx VW_k\mathcal{G}_\infty(\boldsymbol{\theta})$ (crucially $\mathcal{G}_\infty$ and not $\mathcal{G}_T$) and $\Sigma_{\text{GP}}(\boldsymbol{\theta}) \approx VW_k\Sigma W_k^T V^T$.
4. **Sample(k):** We now solve the inverse problem (5), This is feasible as the emulator provides us with an approximation of $\mathcal{G}_\infty$ (not just $\mathcal{G}_T$). The posterior distribution associated with (5) is proportional to a product of prior and likelihood contribution from Bayes theorem. Explicitly, for a Gaussian prior $N(\boldsymbol{m}, C)$ on the computational parameters, and the likelihood dependent on the emulator, we write the MCMC objective function (also known as the log-posterior) as
$$\begin{aligned}\Phi_{\text{MCMC}}(\boldsymbol{\theta}, VW_k\boldsymbol{y}) =& \frac{1}{2}\|VW_k\boldsymbol{y} - \mathcal{G}_{\text{GP}}(\boldsymbol{\theta})\|^2_{\Sigma_{\text{GP}}(\boldsymbol{\theta})} + \frac{1}{2}\log\det\Sigma_{\text{GP}}(\boldsymbol{\theta}) \\ &+ \frac{1}{2}\|\boldsymbol{\theta} - \boldsymbol{m}\|^2_C\,.\end{aligned}$$

The posterior is then given by

$$\mathbb{P}(\boldsymbol{\theta} \mid VW_k\boldsymbol{y}) \propto \exp(-\Phi_{MCMC}(\boldsymbol{\theta}, VW_k\boldsymbol{y})).$$

This can be sampled with a standard random walk metropolis sampling algorithm.

We then collect the posterior distributions $\{\boldsymbol{\theta} \mid W_k\boldsymbol{y}\}_k,\ \forall k \in D$ and calculate the utility function using (6). This CEES algorithm in Figure A1. In particular, note that the subsampling occurs after the J model evaluations, therefore all posterior distributions can be performed in an embarassingly parallel fashion, and all use the same forward model evaluations.

The CEES algorithm is also used to solve (2) at a given design $\tilde{k}$, with this algorithm using the model $\mathcal{S}_T(\cdot;\tilde{k},\cdot)$, and data sample $\boldsymbol{z}_k = \mathcal{S}_T(\boldsymbol{\theta};\tilde{k},\boldsymbol{v}^{(0)})+\delta$, and weighting the data misfit norm with the additional contribution from δ. We then perform **Emulate($\tilde{k}$)**, and **Sample($\tilde{k}$)** at the chosen design.