

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

Oliver R. A. Dunbar¹, Michael F. Howland^{1,2}, Tapio Schneider¹, Andrew M. Stuart¹

¹California Institute of Technology, Pasadena, USA.

²Civil 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

13 **Abstract**

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

35 **Plain Language Summary**

36 Climate models depend on dynamics across many spatial and temporal scales. It
 37 is infeasible to resolve all of these scales. Instead, the physics at the smallest scales is
 38 represented by parameterization schemes that link what is unresolvable to variables re-
 39 solved on the grid scale. A dominant source of uncertainty in climate predictions comes
 40 from uncertainty in calibrating empirical parameters in such parameterization schemes,
 41 and this uncertainty is generally not quantified. Targeted high-resolution simulations of
 42 small-scale processes in limited areas are one means by which uncertainties in param-
 43 eterizations can be reduced and quantified. Here we demonstrate an algorithm that op-
 44 timizes placement of high-resolution simulations to maximize the information they pro-
 45 vide about uncertain parameters in parameterization schemes. Because the sensitivity
 46 of simulated climate statistics, such as precipitation rates, to parameterizations varies
 47 in space and time, how informative high-resolution simulations are about the parame-
 48 terizations also varies in space and time. In proof-of-concept simulations with an ide-
 49 alized global atmosphere model, we show that our novel algorithm successfully identi-
 50 fies the informative regions and times.

51 **1 Introduction**

52 Parameterizations of subgrid-scale processes, such as the turbulence and convec-
 53 tion controlling clouds, are the principal cause of physical uncertainties in climate pre-
 54 dictions (Cess et al., 1989, 1990; Bony & Dufresne, 2005; Stephens, 2005; Bony et al.,
 55 2006; Vial et al., 2013; Webb et al., 2013; Brient & Schneider, 2016). While these pro-
 56 cesses are too small in scale to become globally resolvable in climate models for the fore-
 57 seeable future, many of them can be resolved in limited-area simulations (Schneider, Teix-
 58 eira, et al., 2017). For example, the turbulence and convection (though currently not the
 59 microphysics) controlling clouds can be resolved in large-eddy simulations (LES) over
 60 limited areas (Siebesma et al., 2003; Stevens et al., 2005; Khairoutdinov et al., 2009; Math-
 61 eou & Chung, 2014; Schalkwijk et al., 2015; Pressel et al., 2015, 2017). High-resolution
 62 simulations have been used to calibrate climate model parameterizations at selected sites,

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

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

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

90 Here we present a mathematical framework for addressing both the experimental
 91 design question of where, and when, to embed high-resolution simulations in a coarser
 92 model. We adopt a Bayesian inverse problem setting (see, e.g., Kaipio and Somersalo
 93 (2006), Tarantola (2005), Stuart (2010), and Dashti and Stuart (2013) for reviews). In
 94 this setting, parameters (or parametric or nonparametric functions) in parameterizations
 95 are treated as having probability distributions. Data (e.g., from high-resolution simu-
 96 lations) are used to reduce the uncertainty reflected by these distributions, balancing con-
 97 tributions of the data with that of prior knowledge about parameters (e.g., physical con-
 98 straints). This setting is well suited for our needs, as it provides the complete joint pos-
 99 terior distribution for parameters, including the correlation structure of uncertainties among
 100 parameters. Distribution information is beneficial, for example, because it enables model-
 101 based predictions of rare events with quantified uncertainties (Dunbar et al., 2021). Anal-
 102 ysis of the posterior distribution also may focus scientific development (e.g., improve-
 103 ment of parameterization schemes, Souza et al. (2020)) on areas where uncertainties can
 104 most effectively be minimized. In this paper, we use the posterior distribution to deter-
 105 mine regions and times where local data (e.g., from high-resolution simulations) are max-
 106 imally effective at reducing parameter uncertainties.

107 Construction of the full posterior distribution of the parameters is well known to
 108 be a computationally intensive task, requiring $\mathcal{O}(10^5)$ evaluations of the model in which
 109 the parameters appear with commonly used Markov chain Monte Carlo (MCMC) meth-
 110 ods (see Geyer (2011) for an overview). The recent development of the calibrate-emulate-
 111 sample (CES) framework accelerates Bayesian learning by a factor of 10^3 (Cleary et al.,
 112 2021a; Dunbar et al., 2021). The calibration stage uses a variant of ensemble Kalman
 113 inversion (Iglesias et al., 2013) building on (Chen & Oliver, 2012; Emerick & Reynolds,
 114 2013; Reich, 2011) to obtain a collection of samples of the model about an optimal set
 115 of parameters. The emulation stage features the training of a Gaussian process (Williams

116 & Rasmussen, 2006; Kennedy & O’Hagan, 2000, 2001a) to emulate the model using the
 117 samples from the calibration stage. The sample stage then samples a posterior distri-
 118 bution with MCMC methods, replacing the computationally expensive model with the
 119 cheap emulator. We build on the CES framework and show how Bayesian experimen-
 120 tal design approaches can be incorporated within CES at negligible additional compu-
 121 tational expense. In particular, we do not require additional forward model evaluations
 122 over what is already required in CES to perform uncertainty quantification.

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

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

140 In Section 2, we define the inverse problems for parameter calibration and the op-
 141 timal design algorithm; details of efficient uncertainty quantification (CES) are left to
 142 Appendix A. In Section 3, we briefly describe the GCM used for demonstrating the al-
 143 gorithm. Results of the optimal design algorithm are described in section 4, first in a set-
 144 ting in which the GCM statistics are statistically stationary, then with seasonally vary-
 145 ing GCM statistics. We end with discussion and future directions in Section 5.

146 2 Methodology

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

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

161 2.1 Inverse problem

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

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

For any given k , assume we have local data \mathbf{z}_k available. In the application of inter-
 est, \mathbf{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

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

where δ_k is a stochastic term to capture discrepancies between model $\mathcal{S}_T(\cdot; k, \cdot)$ and data
 \mathbf{z}_k , (e.g., Kennedy & O’Hagan, 2001a). The initial condition $\mathbf{v}^{(0)}$ appears in this formu-
 lation but is treated a nuisance variable. This view is justified in the context of learn-
 ing about atmospheric parameterizations for climate models, where lower frequency in-
 formation 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 sys-
 tem’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. Follow-
 ing (Dunbar et al., 2021), we write $\mathcal{S}_T(\theta; k, \mathbf{v}^{(0)}) \approx \mathcal{S}_\infty(\theta; k) + \sigma_k$, where $\sigma_k \sim N(0, \Sigma(\theta))$
 is normal noise, independent from δ_k , with mean zero and with a covariance matrix $\Sigma(\theta)$
 reflecting chaotic internal variability. The inverse problem then becomes

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

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

174 Solving (2) involves finding the posterior distribution of θ given the data \mathbf{z}_k , de-
 175 noted $(\theta | \mathbf{z}_k)$. Although we cannot evaluate \mathcal{S}_∞ directly, we use the emulate phase of
 176 the calibrate-emulate-sample (CES) algorithm (Cleary et al., 2021b) to construct a sur-
 177rogate of \mathcal{S}_∞ from carefully chosen evaluations of \mathcal{S}_T . This has been shown to be effi-
 178cient with respect to the required number of evaluations of \mathcal{S}_T (Cleary et al., 2021b; Dun-
 179bar et al., 2021). Details of this algorithm are provided in Appendix A.

180 2.2 Experimental design

181 We imagine a situation where evaluating \mathbf{z}_k has a large computational cost. In the
 182 relevant application of targeted high-resolution simulations, \mathbf{z}_k is data obtained by run-
 183ning a high-resolution simulation at design point k . Our starting point is to assume that
 184 a limited computational budget restricts us to evaluate \mathbf{z}_k at a single design point k at
 185 a time. We want to choose the k that leads to the most informative inverse problem (2).
 186 We take a Bayesian point of view, namely, the optimal k is the one for which the pos-
 187terior distribution of $(\theta | \mathbf{z}_k)$ learned from the inverse problem (2) has the smallest un-
 188certainty.

189 To answer this conclusively, one would need to evaluate \mathbf{z}_k at all design points k ,
 190 which here is too computationally expensive. Instead, we investigate only the sensitiv-
 191ity of the forward model statistics \mathcal{G}_T to its parameters θ to assess the marginal infor-
 192mation provided at each design point k . This marginal information at k is used as a proxy
 193 for the information content that would exist when learning from data \mathbf{z}_k . The benefits
 194 of this approach are that (i) we do not require any evaluations of \mathbf{z}_k to select the opti-
 195mal location; (ii) the measure of information content is naturally constructed from the

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

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

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

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

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

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 \mathbf{y} . Given this realization of \mathbf{y} , we temporarily forget $\boldsymbol{\theta}^*$, and for any design point k , we consider a restriction of an inverse problem to k

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

205 The posterior distributions of $\boldsymbol{\theta} | W_k \mathbf{y}$ for all k obtained by solving (5) informs us about
 206 the sensitivities of \mathcal{G}_∞ with respect to parameters, when only data at different k is avail-
 207 able. To simplify solution of the inverse problem, we approximate the internal variabil-
 208 ity covariance matrix $\Sigma(\boldsymbol{\theta})$ by a fixed covariance matrix $\Sigma(\boldsymbol{\theta}^*)$. This covariance matrix
 209 can be obtained by running a collection of control simulations with parameters fixed to
 210 (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 $(\boldsymbol{\theta} | W_k \mathbf{y})$ 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),

$$U(W_k) = \left(\det(\text{Cov}(\boldsymbol{\theta} | W_k \mathbf{y})) \right)^{-1}. \tag{6}$$

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

222 2.3 Targeted uncertainty quantification algorithm

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

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

- 228 1. The design stage consists of the following steps:
- 229 (a) Generate a sample of GCM simulated data $\mathbf{y} = \mathcal{G}_T(\boldsymbol{\theta}^*; \mathbf{v}^{(0)})$, and estimate the
 230 internal variability covariance matrix $\Sigma(\boldsymbol{\theta}^*)$. We approximate $\Sigma(\boldsymbol{\theta})$ as $\Sigma(\boldsymbol{\theta}^*)$.
- 231 (b) For each $k \in D$, solve (5), in parallel, for the posterior of $(\boldsymbol{\theta} \mid W_k \mathbf{y})$, using
 232 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

$$\tilde{k} = \arg \max_{k \in D} U(W_k).$$

- 233 2. The uncertainty quantification stage consists of the following steps:
- 234 (a) At the optimal design point \tilde{k} , obtain a sample $\mathbf{z}_{\tilde{k}}$.
- 235 (b) Solve the inverse problem (2) for the posterior distribution of $(\boldsymbol{\theta} \mid \mathbf{z}_{\tilde{k}})$.

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

242 3 Idealized GCM and Experimental Setup

243 3.1 Idealized GCM, parameters, and priors

244 To demonstrate the algorithm in a simplified setting, we use the idealized aqua-
 245 planet GCM described by D. M. W. Frierson et al. (2006) and O’Gorman and Schnei-
 246 der (2008b). The aquaplanet is a climate model with atmosphere and a simplified slab
 247 ocean covering the entire planet surface. Without topography, it exhibits symmetries in
 248 the longitudinal directions. The aquaplanet can produce statistically stationary climates
 249 by prescribing fixed insolation. It can also cyclostationary statistics over seasons through
 250 seasonally varying insolation (Bordoni & Schneider, 2008a; Howland et al., 2021). It has
 251 been shown in Dunbar et al. (2021); Howland et al. (2021) that the parameters of a sim-
 252 ple quasi-equilibrium moist convection parameterization can be calibrated within this
 253 GCM in the stationary and cyclostationary regimes. The quasi-equilibrium moist con-
 254 vection scheme relaxes temperature and specific humidity toward moist-adiabatic refer-
 255 ence profiles with a fixed relative humidity RH (D. M. W. Frierson, 2007). The timescale
 256 with which the temperature and specific humidity relax to their respective reference pro-
 257 files is given by the parameter τ . The two parameters RH and τ are the key parameters
 258 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 dis-
 tributions, $\text{RH} \sim \text{Logitnormal}(0, 1)$ and $\tau \sim \text{Lognormal}(12 \text{ h}, (12 \text{ h})^2)$. That is, we
 define the invertible transformation

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

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

266 **3.2 Objective function for parameter learning**

267 We learn from climate statistics that are known to be sensitive to the parameters.
 268 We have knowledge about these sensitivities from a body of previous studies of large-
 269 scale atmosphere dynamics and mechanisms of climate changes which used this ideal-
 270 ized GCM (e.g., O’Gorman & Schneider, 2008b, 2008a; Bordoni & Schneider, 2008b; O’Gorman
 271 & Schneider, 2009b; Schneider et al., 2010; Merlis & Schneider, 2011; O’Gorman, 2011;
 272 Kaspi & Schneider, 2011, 2013; X. Levine & Schneider, 2015; Bischoff & Schneider, 2014;
 273 Wills et al., 2017; Wei & Bordoni, 2018). We know, for example, that the convection scheme
 274 primarily affects the atmospheric thermal stratification in the tropics, with weaker ef-
 275 fects in the extratropics (Schneider & O’Gorman, 2008). We also know that the relative
 276 humidity parameter RH in the convection scheme controls the humidity of the tropical
 277 free troposphere but has a weaker effect on the humidity of the extratropical free tro-
 278 posphere (O’Gorman et al., 2011). Thus, we expect tropical circulation statistics to be
 279 especially informative about the parameters in the convection scheme. However, convec-
 280 tion plays a central role in intense precipitation events at all latitudes (O’Gorman & Schnei-
 281 der, 2009b, 2009a), so we expect statistics of precipitation intensity to be informative
 282 about convective parameters, and in particular to contain information about the relax-
 283 ation timescale τ .

284 As statistics to learn from, we therefore choose averages of the free-tropospheric
 285 relative humidity, of the precipitation rate, and of a measure of the frequency of intense
 286 precipitation. We use averages over $T = 30$ days in statistically stationary simulations
 287 (Dunbar et al., 2021) and over $T = 90$ days in simulations of the seasonal cycle (Howland
 288 et al., 2021). We exploit the symmetry in the GCM by taking zonal averages in addi-
 289 tion to the time averages. The relative humidity data are evaluated at $\sigma = 0.5$ (where
 290 $\sigma = p/p_s$ is pressure p normalized by the local surface pressure p_s), the precipitation
 291 rate is taken daily, and as a measure of the frequency of intense precipitation, we use the
 292 frequency with which daily precipitation exceeds the latitude-dependent 90th percentile
 293 of precipitation rates in a long (18000 days) control simulation. We run the GCM at the
 294 coarse horizontal spectral resolution of T21, implying 32 discrete latitudes on the spec-
 295 tral transform grid. Hence, we have 3 statistics, each a function of 32 latitude points,
 296 resulting in a 96-dimensional processed output, defined as \mathcal{H}_T . In the statistically sta-
 297 tionary 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

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

298 over a one year cycle (360 days), where $\mathbf{v}^{(i)}$ is the model state at the beginning of each
 299 90-day long season labelled $i = 0, 1, 2, 3$. With this batching, we have now constructed
 300 stationary statistics for the stacked data. The theory of Section 2 applies, and our in-
 301 verse problems can be formulated in the seasonally varying case.

302 **3.3 Design choices**

303 In the stationary GCM setting, we aggregate statistics temporally and zonally. Thus,
 304 a local design implies a restriction to certain latitudes. Recall our discretization has 32
 305 discrete latitudes. We therefore choose a design space that contains sets of ℓ consecu-
 306 tive discrete latitudes, indexed from south to north poles with the design points $k = 1, \dots, 32-$
 307 $(\ell-1)$. In the stationary experiments, we choose $\ell = 3$, indexing designs $k = 1, \dots, 30$,
 308 unless otherwise specified. The choice of ℓ is discussed in Section 4.1.

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

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

316 **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 θ^\dagger , adding Gaussian noise δ with zero mean and covariance matrix Δ as in (2). The implication is that we generate z_k with the restricted idealized GCM $\mathcal{S}_T(\theta^\dagger; k)$, 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

$$z_k = \mathcal{S}_\infty(\theta; k) + \gamma_k, \quad \gamma_k \sim N(0, W_k(\Sigma(\theta) + \Delta)W_k^T). \quad (7)$$

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,

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

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:

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

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

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

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

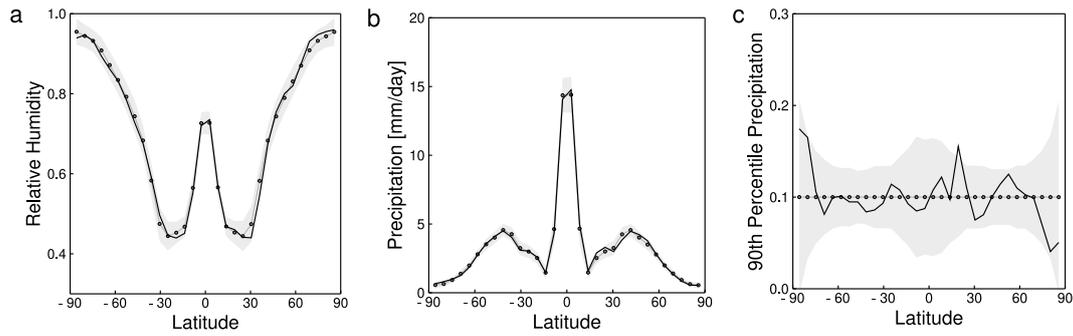


Figure 1. Aggregated climate statistics in the statistically stationary control simulation, with parameters set to the mean of the prior θ^* . The mean (grey lines) and 95% confidence intervals (shading) of the data are plotted against latitude. One realization of the data is shown (black line). No noise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

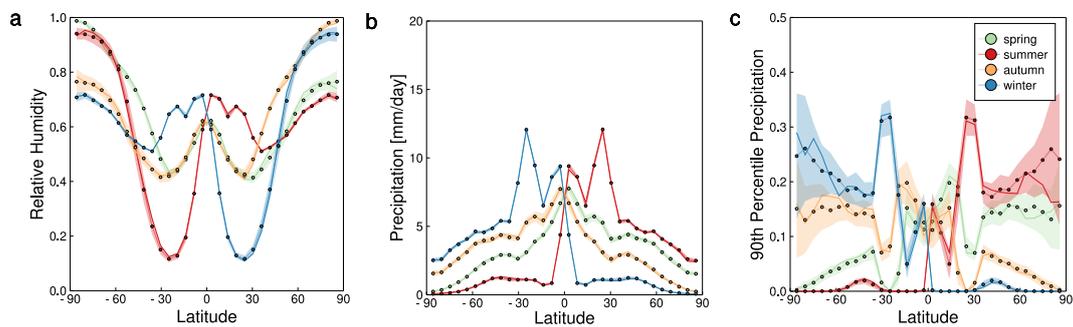


Figure 2. Aggregated climate statistics in the seasonally varying control simulation, with parameters set to the mean of the prior θ^* . The mean (solid lines) and 95% confidence intervals (shading) of the data are plotted against latitude, with the colors indicating the seasons. No noise is added here.

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

348 4 Results

349 4.1 Stationary statistics

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

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

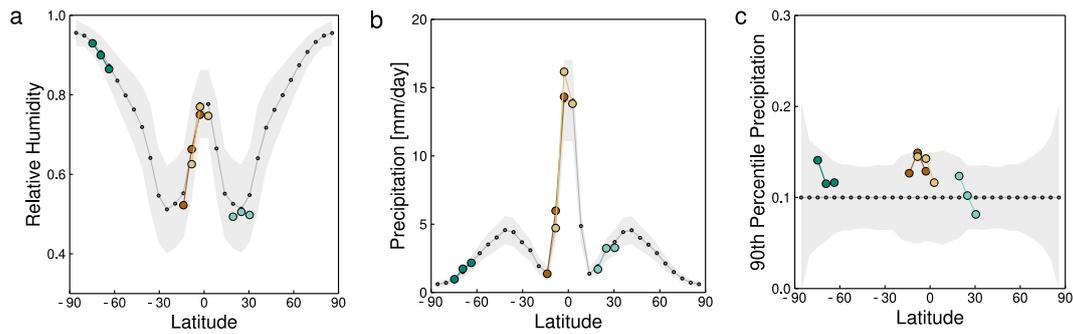


Figure 3. Aggregated climate statistics in the statistically stationary control simulation using the ground truth parameters. Mean (grey lines) and 95% confidence intervals (shading) of the data are plotted against latitude. Additional inflation noise is added. Each set of colored discs represents a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

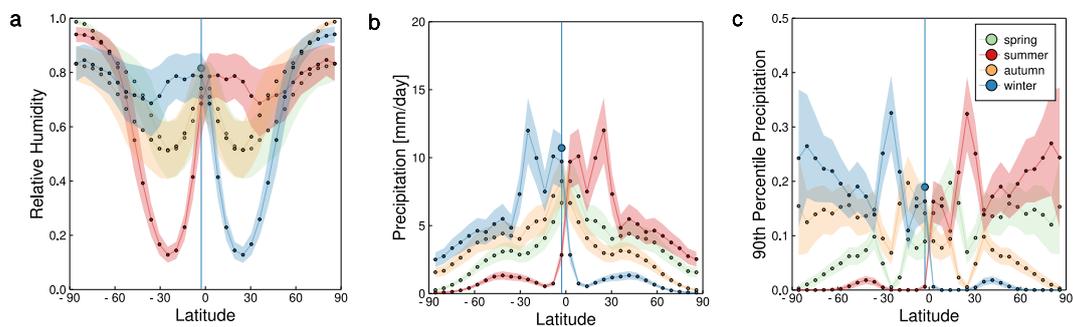


Figure 4. Aggregated climate statistics produced from the seasonally varying control simulation using the ground truth parameters, and with additional inflation. The mean (solid lines) and 95% confidence intervals (shading) of the data are plotted against latitude, with the colors indicating the seasons. 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.

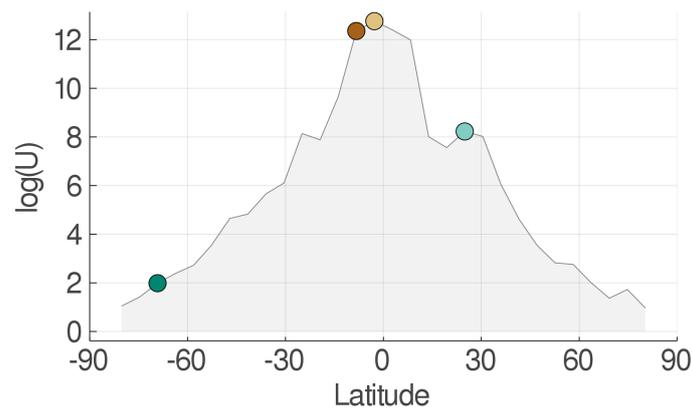


Figure 5. Logarithm of the data utility as a function of latitude, with designs represented by a node at the center of each stencil (comprised of three neighboring latitudes). The colored discs signify the four representative designs indicated in Fig. 3, which are used in the uncertainty quantification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

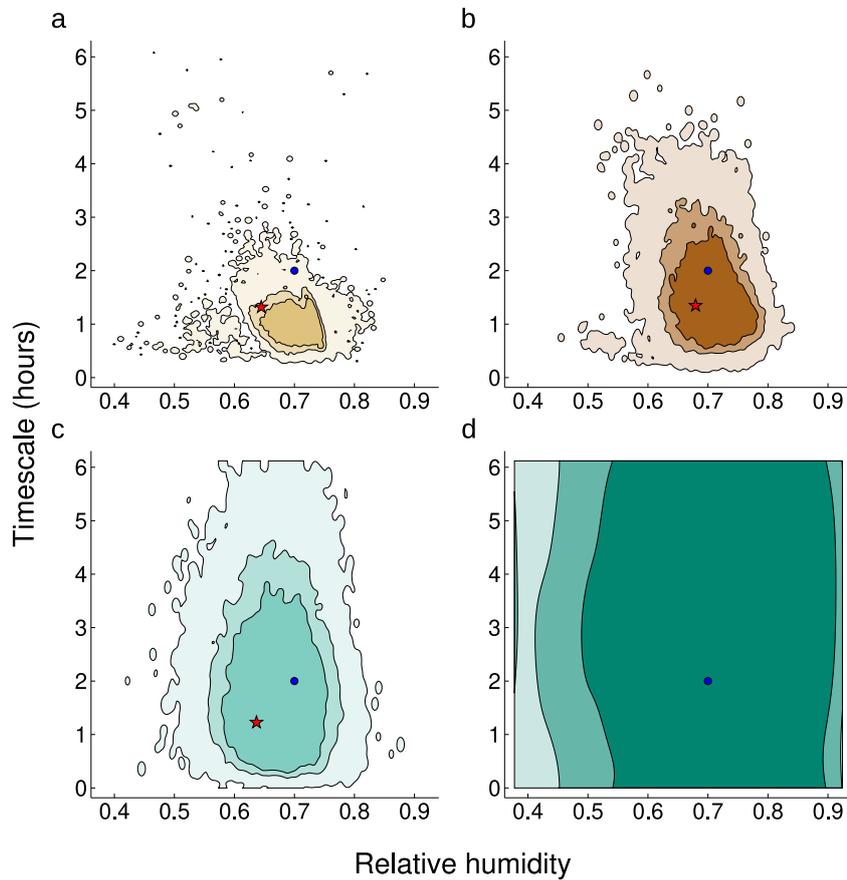


Figure 6. Posterior distributions for convection parameters learned from data restricted to different design points. The drawn contours bound 50%, 75% and 99% of the distribution. Panels (a, b, c, d) correspond to designs $k = (51, 14, 20, 3)$, ordered to express learning from data at decreasingly informative 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).

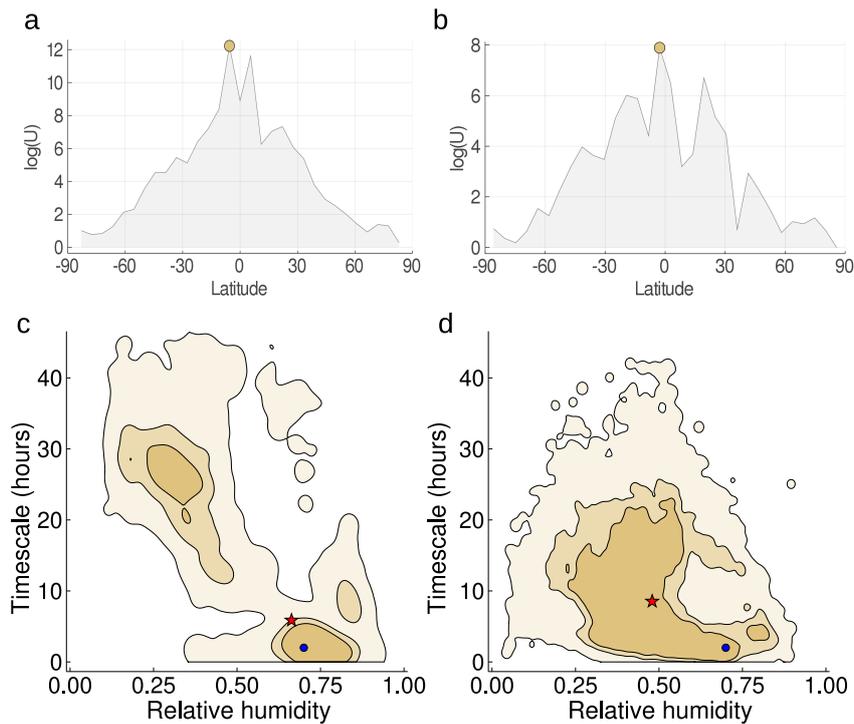


Figure 7. Performance for different optimal design selections at smaller stencil sizes. The contours bound 50%, 75%, and 99% of the distribution. Panels (a,c) is a two-latitude stencil, and panels (b,d) is a one-latitude stencil. The top row displays the logarithm of the utility plot, and the bottom 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 θ^* . We see this monotonicity is preserved when considering data produced from the true parameter θ^\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-

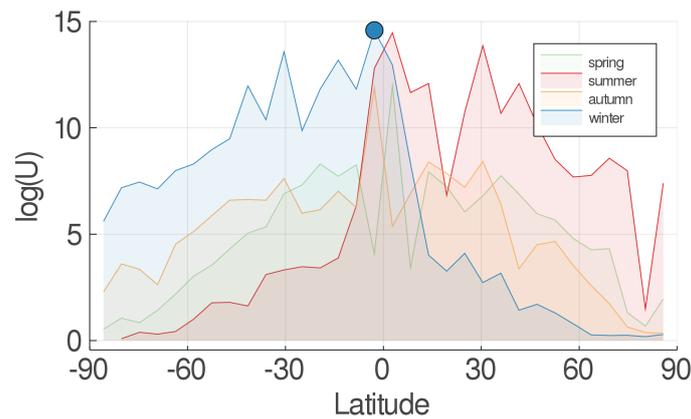


Figure 8. Logarithm of the data utility plotted against latitude (1 design per latitude and season). The blue disc signifies that a latitude in northern winter maximizes the utility function.

382 timodality) appears only at data from θ^\dagger , but not at θ^* . As a result, the optimal un-
 383 certainty is not guaranteed to be found at the location of optimal utility. This is reme-
 384 died by having a better initial guess through the prior, or (as demonstrated in the three-
 385 latitude set) a less sparse data set from which the parameters are more identifiable.

386 4.2 Seasonally varying statistics

387 In the seasonally varying case, we choose the optimal design with the algorithm
 388 in Section 2.3 applied to the stacked data. Figure 8 shows the logarithm of the utility
 389 function. Hemispheric and seasonal asymmetries are evident here. In northern winter,
 390 latitudes just south of the the equator ($k = (3, 16)$) optimize the design, in the vicin-
 391 ity of the ITCZ. Conversely, in northern summer, latitudes just north of the equator ($k =$
 392 $(1, 17)$) optimize the design, again in the vicinity of the seasonally migrating ITCZ; ad-
 393 ditional peaks at around 30 degrees can be seen. The equinox seasons have less utility
 394 at the optimal designs ($k = (0, 17)$ and $k = (2, 16)$). Because the equinoctial Hadley
 395 cells and ascent regions in the ITCZ are less pronounced than the solstitial Hadley cells
 396 (Schneider et al., 2010), utility is more spread out across the latitudes.

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

405 5 Conclusions and Discussion

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

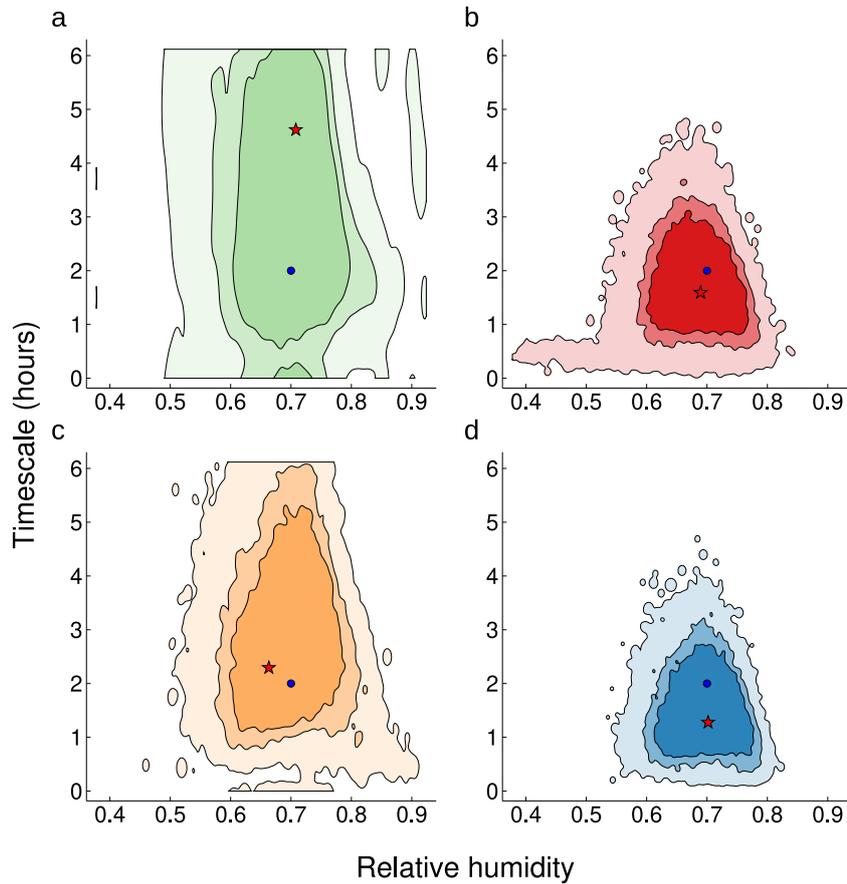


Figure 9. Posterior distribution obtained from using data at the optimal latitude from each season. Contours bound 50%, 75% and 99% of the distribution (in decreasing color saturation). Panels (a, b, c, d) correspond to designs $k = ((0, 17), (1, 17), (2, 16), (3, 16))$, ordered by season. The true parameter 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.

413 simulations directly with an idealized GCM, with statistically stationary or seasonally
 414 varying statistics. In these settings, we have demonstrated learning about parameters
 415 in a convection parameterization, exploring both spatial and spatio-temporal designs.
 416 Our design framework can also be used more broadly, to automate selection of optimally
 417 informative climate statistics from libraries of high-resolution simulations (Shen et al.,
 418 2021).

419 With the idealized GCM, we showed how to optimally target a location at which
 420 additional data will produce parameter estimates that minimize uncertainty. In our proof-
 421 of-concept in which we calibrated parameters in a simple convection scheme, the auto-
 422 matically targeted optimal location for new data was consistently near the equator, in
 423 the vicinity of the seasonally migrating ITCZ. This is consistent with the fact that the
 424 convection scheme in the idealized GCM is most important near the ITCZ (O’Gorman
 425 & Schneider, 2008). We showed that the optimal targeting is limited in its effectiveness
 426 in settings of very sparse data, where parameter posteriors can be multimodal. However,
 427 with access to the posterior distributions of the parameters, the behavior is both diag-
 428 nosable a posteriori, and actionable with successive iterations of the optimal design pro-
 429 cess (simply take the current posterior as the prior for a subsequent iteration with ad-
 430 ditional data).

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

439 Despite being efficient, the current algorithm relies on evaluating utilities naively
 440 at all design points. In a practical climate model application, where we may have 10^2
 441 LES that are computationally affordable to be placed optimally within 10^6 or more pos-
 442 sible locations, such naive approaches are infeasible. Instead, one can use more sophis-
 443 ticated optimization algorithms. For determinant based (i.e., D -optimal) utilities, this
 444 typically requires accelerating the determinant evaluation (and its gradients). Various
 445 methods have been developed to do so, e.g., using Laplace approximations (Long et al.,
 446 2013; Beck et al., 2018; Rue et al., 2009), polynomial chaos surrogates (Huan & Mar-
 447 zouk, 2014), optimization of criteria bounds (Tsilifis et al., 2017), fast random determi-
 448 nant approximation (Alexanderian et al., 2014; Alexanderian & Saibaba, 2018), and Gaus-
 449 sian process surrogates (Buathong et al., 2020; Paglia et al., 2020). The latter, kernel-
 450 based approaches are particularly amenable to our setting, as they give sparse represen-
 451 tations of the utility function that are independent of the underlying computational grid.
 452 They may offer a way forward in the climate modeling setting.

453 Acknowledgments

454 We gratefully acknowledge the generous support of Eric and Wendy Schmidt (by rec-
 455 ommendation of Schmidt Futures) and the National Science Foundation (grant AGS-1835860).
 456 The simulations were performed on Caltech’s High Performance Cluster, which is par-
 457 tially supported by a grant from the Gordon and Betty Moore Foundation. AMS is also
 458 supported by the Office of Naval Research (grant N00014-17-1-2079).

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

462 <https://doi.org/10.5281/zenodo.5835269>.

463

References

464

Alexanderian, A., Petra, N., Stadler, G., & Ghattas, O. (2014). A-optimal design of experiments for infinite-dimensional bayesian linear inverse problems with regularized ℓ_0 -sparsification. *SIAM Journal on Scientific Computing*, *36*(5), A2122–A2148.

468

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.

469

470

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.

475

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

476

477

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

478

479

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

480

481

482

483

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.

484

485

486

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

487

488

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

489

490

491

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

492

493

494

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

495

496

497

498

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>

499

500

501

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.

502

503

504

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

505

506

507

508

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

509

510

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

511

512

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>

513

514

515

516

- 517 Cleary, E., Garbuno-Inigo, A., Lan, S., Schneider, T., & Stuart, A. M. (2021b). Cal-
 518 ibrate, emulate, sample. *J. Comp. Phys.*, *424*, 109716. Retrieved from 10
 519 .1016/j.jcp.2020.109716
- 520 Cook, A. R., Gibson, G. J., & Gilligan, C. A. (2008). Optimal observation times in
 521 experimental epidemic processes. *Biometrics*, *64*(3), 860-868.
- 522 Couvreur, F., Hourdin, F., Williamson, D., Roehrig, R., Volodina, V., Villefranque,
 523 N., ... others (2021). Process-based climate model development harnessing
 524 machine learning: I. A calibration tool for parameterization improvement. *J.*
 525 *Adv. Model. Earth Sys.*, *13*, e2020MS002217. doi: 10.1029/2020MS002217
- 526 Dashti, M., & Stuart, A. M. (2013). The Bayesian approach to inverse problems.
 527 *arXiv preprint arXiv:1302.6989*.
- 528 de Rooy, W. C., Bechtold, P., Fröhlich, K., Hohenegger, C., Jonker, H., Mironov, D.,
 529 ... Yano, J.-I. (2013). Entrainment and detrainment in cumulus convection:
 530 an overview. *Quart. J. Roy. Meteor. Soc.*, *139*, 1–19.
- 531 Dunbar, O. R. A., Garbuno-Inigo, A., Schneider, T., & Stuart, A. M. (2021).
 532 Calibration and uncertainty quantification of convective parameters in an
 533 idealized GCM. *J. Adv. Model. Earth Sys.*, *13*, e2020MS002454. doi:
 534 10.1029/2020MS002454
- 535 Duncan, A. B., Stuart, A. M., & Wolfram, M.-T. (2021). Ensemble inference
 536 methods for models with noisy and expensive likelihoods. *arXiv preprint*
 537 *arXiv:2104.03384*.
- 538 Emanuel, K., Raymond, D., Betts, A., Bosart, L., Bretherton, C., Droegeleier, K.,
 539 ... others (1995). Report of the first prospectus development team of the
 540 us weather research program to noaa and the nsf. *Bulletin of the American*
 541 *Meteorological Society*, 1194–1208.
- 542 Emerick, A. A., & Reynolds, A. C. (2013). Ensemble smoother with multiple data
 543 assimilation. *Computers & Geosciences*, *55*, 3–15.
- 544 Fox-Kemper, B., & Menemenlis, D. (2013). Can large eddy simulation techniques
 545 improve mesoscale rich ocean models? In *Ocean modeling in an eddying regime*
 546 (p. 319-337). American Geophysical Union (AGU). doi: 10.1029/177GM19
- 547 Frierson, D. M. (2007). The dynamics of idealized convection schemes and their
 548 effect on the zonally averaged tropical circulation. *Journal of the Atmospheric*
 549 *Sciences*, *64*(6), 1959–1976.
- 550 Frierson, D. M. W. (2007). The dynamics of idealized convection schemes and their
 551 effect on the zonally averaged tropical circulation. *J. Atmos. Sci.*, *64*, 1959–
 552 1976.
- 553 Frierson, D. M. W., Held, I. M., & Zurita-Gotor, P. (2006). A gray-radiation aqua-
 554 planet moist GCM. Part I: Static stability and eddy scale. *J. Atmos. Sci.*, *63*,
 555 2548–2566.
- 556 Geyer, C. J. (2011). Introduction to markov chain monte carlo. In S. Brooks,
 557 A. Gelman, G. L. Jones, & X.-L. Meng (Eds.), *Handbook of markov chain*
 558 *monte carlo* (pp. 3–48). Chapman and Hall/CRC.
- 559 Hohenegger, C., & Bretherton, C. S. (2011). Simulating deep convection with a shal-
 560 low convection scheme. *Atmos. Chem. Phys.*, *11*, 10389–10406. doi: 10.5194/
 561 acp-11-10389-2011
- 562 Hourdin, F., Williamson, D., Rio, C., Couvreur, F., Roehrig, R., Villefranque, N.,
 563 ... Volodina, V. (2021). Process-based climate model development harnessing
 564 machine learning: II. model calibration from single column to global. *J. Adv.*
 565 *Model. Earth Sys.*, *13*, e2020MS002225. doi: 10.1029/2020MS002225
- 566 Howland, M. F., Dunbar, O. R. A., & Schneider, T. (2021). Parameter uncer-
 567 tainty quantification in an idealized gcm with a seasonal cycle. *arXiv preprint*
 568 *arXiv:2108.00827*.
- 569 Huan, X., & Marzouk, Y. (2014). Gradient-based stochastic optimization methods in
 570 bayesian experimental design. *International Journal for Uncertainty Quantifi-*
 571 *cation*, *4*(6).

- 572 Huan, X., & Marzouk, Y. M. (2013, 1). Simulation-based optimal bayesian ex-
 573 perimental design for nonlinear systems. *Journal of Computational Physics*,
 574 *232*(1).
- 575 Iglesias, M. A., Law, K. J., & Stuart, A. M. (2013). Ensemble kalman methods for
 576 inverse problems. *Inverse Problems*, *29*(4), 045001.
- 577 Kaipio, J., & Somersalo, E. (2006). *Statistical and computational inverse problems*
 578 (Vol. 160). Springer Science & Business Media.
- 579 Kalnay, E. (2003). *Atmospheric modeling, data assimilation and predictability*. Cam-
 580 bridge, UK: Cambridge Univ. Press.
- 581 Kaspi, Y., & Schneider, T. (2011). Winter cold of eastern continental boundaries in-
 582 duced by warm ocean waters. *Nature*, *471*, 621–624.
- 583 Kaspi, Y., & Schneider, T. (2013). The role of stationary eddies in shaping midlati-
 584 tude storm tracks. *J. Atmos. Sci.*, *70*, 2596–2613.
- 585 Kennedy, M. C., & O’Hagan, A. (2000). Predicting the output from a complex com-
 586 puter code when fast approximations are available. *Biometrika*, *87*(1), 1–13.
- 587 Kennedy, M. C., & O’Hagan, A. (2001a). Bayesian calibration of computer mod-
 588 els. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*,
 589 *63*(3), 425–464.
- 590 Kennedy, M. C., & O’Hagan, A. (2001b). Bayesian calibration of computer models.
 591 *J. Roy. Statist. Soc. B*, *63*, 425–464. doi: 10.1111/1467-9868.00294
- 592 Khairoutdinov, M. F., Krueger, S. K., Moeng, C.-H., Bogenschutz, P. A., & Randall,
 593 D. A. (2009). Large-eddy simulation of maritime deep tropical convection. *J.*
 594 *Adv. Model. Earth Sys.*, *1*, Art. #15, 13 pp. doi: 10.3894/JAMES.2009.1.15
- 595 Kim, W., Pitt, M. A., Lu, Z.-L., Steyvers, M., & Myung, J. I. (2014). A hierar-
 596 chical adaptive approach to optimal experimental design. *Neural Computation*,
 597 *26*(11), 2465–2492.
- 598 Levine, M. E., & Stuart, A. M. (2021). A framework for machine learning of model
 599 error in dynamical systems.
 600 (<https://arxiv.org/abs/2107.06658>)
- 601 Levine, X., & Schneider, T. (2015). Baroclinic eddies and the extent of the Hadley
 602 circulation: An idealized GCM study. *J. Atmos. Sci.*, *72*, 2744–2761. doi: 10
 603 .1175/JAS-D-14-0152.1
- 604 Li, Q., & Fox-Kemper, B. (2017). Assessing the effects of langmuir turbulence on
 605 the entrainment buoyancy flux in the ocean surface boundary layer. *Journal of*
 606 *Physical Oceanography*, *47*(12), 2863–2886.
- 607 Li, Q., Reichl, B. G., Fox-Kemper, B., Adcroft, A. J., Belcher, S. E., Danabasoglu,
 608 G., ... others (2019). Comparing ocean surface boundary vertical mixing
 609 schemes including langmuir turbulence. *Journal of Advances in Modeling Earth*
 610 *Systems*, *11*(11), 3545–3592.
- 611 Liu, C., Moncrieff, M. W., & Grabowski, W. W. (2001). Hierarchical modelling of
 612 tropical convective systems using explicit and parametrized approaches. *Quart.*
 613 *J. Roy. Meteor. Soc.*, *127*, 493–515.
- 614 Long, Q., Scavino, M., Tempone, R., & Wang, S. (2013). Fast estimation of ex-
 615 pected information gains for bayesian experimental designs based on laplace
 616 approximations. *Computer Methods in Applied Mechanics and Engineering*,
 617 *259*, 24–39.
- 618 Lorenz, E. N., & Emanuel, K. A. (1998). Optimal sites for supplementary weather
 619 observations: Simulation with a small model. *J. Atmos. Sci.*, *55*, 399–414. doi:
 620 10.1175/1520-0469(1998)055<0399:OSFSWO>2.0.CO;2
- 621 Matheou, G., & Chung, D. (2014). Large-eddy simulation of stratified turbulence.
 622 Part II: Application of the stretched-vortex model to the atmospheric bound-
 623 ary layer. *J. Atmos. Sci.*, *71*, 4439–4460. doi: 10.1175/JAS-D-13-0306.1
- 624 Merlis, T. M., & Schneider, T. (2011). Changes in zonal surface temperature gradi-
 625 ents and walker circulations in a wide range of climates. *J. Climate*, *24*, 4757–
 626 4768.

- 627 Notz, W. I., Santner, T. J., & Williams, B. J. (2018). *The design and analysis of*
 628 *computer experiments* (2nd ed. ed.). Springer.
- 629 O’Gorman, P. A. (2011). The effective static stability experienced by eddies in a
 630 moist atmosphere. *J. Atmos. Sci.*, *68*, 75–90.
- 631 O’Gorman, P. A., Lamquin, N., Schneider, T., & Singh, M. S. (2011). The relative
 632 humidity in an isentropic advection–condensation model: Limited poleward in-
 633 fluence and properties of subtropical minima. *J. Atmos. Sci.*, *68*, 3079–3093.
- 634 O’Gorman, P. A., & Schneider, T. (2008a). Energy of midlatitude transient eddies
 635 in idealized simulations of changed climates. *J. Climate*, *21*, 5797–5806.
- 636 O’Gorman, P. A., & Schneider, T. (2008b). The hydrological cycle over a wide range
 637 of climates simulated with an idealized GCM. *J. Climate*, *21*, 3815–3832.
- 638 O’Gorman, P. A., & Schneider, T. (2009a). The physical basis for increases in pre-
 639 cipitation extremes in simulations of 21st-century climate change. *Proc. Natl.*
 640 *Acad. Sci.*, *106*, 14773–14777.
- 641 O’Gorman, P. A., & Schneider, T. (2009b). Scaling of precipitation extremes over
 642 a wide range of climates simulated with an idealized GCM. *J. Climate*, *22*,
 643 5676–5685.
- 644 Oliver, D. S., Reynolds, A. C., & Liu, N. (2008). *Inverse theory for petroleum reser-*
 645 *voir characterization and history matching*. Cambridge Univ. Press.
- 646 O’Gorman, P. A., & Schneider, T. (2008). The hydrological cycle over a wide range
 647 of climates simulated with an idealized gcm. *Journal of Climate*, *21* (15), 3815-
 648 3832. doi: 10.1175/2007JCLI2065.1
- 649 Paglia, J., Eidsvik, J., & Karvanen, J. (2020). Efficient spatial designs using haus-
 650 dorff distances and bayesian optimisation. *Statistical modeling for safer drilling*
 651 *operations*, 77.
- 652 Paninski, L. (2005). Asymptotic theory of information-theoretic experimental design.
 653 *Neural Computation*, *17*(7), 1480-1507.
- 654 Pressel, K. G., Kaul, C. M., Schneider, T., Tan, Z., & Mishra, S. (2015). Large-eddy
 655 simulation in an anelastic framework with closed water and entropy balances.
 656 *J. Adv. Model. Earth Sys.*, *7*, 1425–1456. doi: 10.1002/2015MS000496
- 657 Pressel, K. G., Mishra, S., Schneider, T., Kaul, C. M., & Tan, Z. (2017). Numerics
 658 and subgrid-scale modeling in large eddy simulations of stratocumulus clouds.
 659 *J. Adv. Model. Earth Sys.*, *9*, 1342–1365. doi: 10.1002/2016MS000778
- 660 Reich, S. (2011). A dynamical systems framework for intermittent data assimilation.
 661 *BIT Numerical Mathematics*, *51*(1), 235–249.
- 662 Reichl, B. G., & Hallberg, R. (2018). A simplified energetics based planetary bound-
 663 ary layer (epbl) approach for ocean climate simulations. *Ocean Modelling*, *132*,
 664 112 - 129. doi: https://doi.org/10.1016/j.ocemod.2018.10.004
- 665 Reichl, B. G., Wang, D., Hara, T., Ginis, I., & Kukulka, T. (2016). Langmuir tur-
 666 bulence parameterization in tropical cyclone conditions. *Journal of Physical*
 667 *Oceanography*, *46*(3), 863–886.
- 668 Romps, D. M. (2016). The Stochastic Parcel Model: A deterministic parameteriza-
 669 tion of stochastically entraining convection. *J. Adv. Model. Earth Sys.*, *8*, 319–
 670 344. doi: 10.1002/2015MS000537
- 671 Rue, H., Martino, S., & Chopin, N. (2009). Approximate bayesian inference for
 672 latent gaussian models by using integrated nested laplace approximations.
 673 *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*,
 674 *71*(2), 319-392.
- 675 Ryan, E. G., Drovandi, C. C., McGree, J. M., & Pettitt, A. N. (2016). A review
 676 of modern computational algorithms for bayesian optimal design. *International*
 677 *Statistical Review*, *84*(1), 128–154.
- 678 Ryan, E. G., Drovandi, C. C., Thompson, M. H., & Pettitt, A. N. (2014). Towards
 679 bayesian experimental design for nonlinear models that require a large number
 680 of sampling times. *Computational Statistics & Data Analysis*, *70*, 45 - 60.

- 681 Schalkwijk, J., Jonker, H. J. J., Siebesma, A. P., & Van Meijgaard, E. (2015).
 682 Weather forecasting using GPU-based large-eddy simulations. *Bull. Amer.*
 683 *Meteor. Soc.*, *96*, 715–723. doi: 10.1175/BAMS-D-14-00114.1
- 684 Schillings, C., & Stuart, A. M. (2017). Analysis of the ensemble kalman filter for in-
 685 verse problems. *SIAM Journal on Numerical Analysis*, *55*(3), 1264–1290.
- 686 Schneider, T., Lan, S., Stuart, A., & Teixeira, J. (2017). Earth system model-
 687 ing 2.0: A blueprint for models that learn from observations and targeted
 688 high-resolution simulations. *Geophys. Res. Lett.*, *44*, 12396–12417. doi:
 689 10.1002/2017GL076101
- 690 Schneider, T., & O’Gorman, P. A. (2008). Moist convection and the thermal stratifi-
 691 cation of the extratropical troposphere. *J. Atmos. Sci.*, *65*, 3571–3583.
- 692 Schneider, T., O’Gorman, P. A., & Levine, X. J. (2010). Water vapor
 693 and the dynamics of climate changes. *Rev. Geophys.*, *48*, RG3001.
 694 (doi:10.1029/2009RG000302)
- 695 Schneider, T., Stuart, A. M., & Wu, J. (2021). Imposing sparsity within ensemble
 696 Kalman inversion.
- 697 Schneider, T., Teixeira, J., Bretherton, C. S., Brient, F., Pressel, K. G., Schär, C.,
 698 & Siebesma, A. P. (2017). Climate goals and computing the future of clouds.
 699 *Nature Climate Change*, *7*, 3–5. doi: 10.1038/nclimate3190
- 700 Shen, Z., Pressel, K. G., Tan, Z., & Schneider, T. (2020). Statistically steady
 701 state large-eddy simulations forced by an idealized GCM: 1. forcing frame-
 702 work and simulation characteristics. *J. Adv. Model. Earth Sys.*, *12*. doi:
 703 10.1029/2019MS001814
- 704 Shen, Z., Sridhar, A., Tan, Z., Jaruga, A., & Schneider, T. (2021). A li-
 705 brary of large-eddy simulations for calibrating cloud parameterizations.
 706 <https://essoar.org>. doi: <https://doi.org/10.1002/essoar.10507112.1>
- 707 Siebesma, A. P., Bretherton, C. S., Brown, A., Chlond, A., Cuxart, J., Duynkerke,
 708 P. G., ... Stevens, D. E. (2003). A large eddy simulation intercomparison
 709 study of shallow cumulus convection. *J. Atmos. Sci.*, *60*, 1201–1219.
- 710 Siebesma, A. P., Soares, P. M. M., & Teixeira, J. (2007). A combined eddy-
 711 diffusivity mass-flux approach for the convective boundary layer. *J. Atmos.*
 712 *Sci.*, *64*, 1230–1248. doi: 10.1175/JAS3888.1
- 713 Smalley, M., Suselj, K., Lebsock, M., & Teixeira, J. (2019). A novel framework
 714 for evaluating and improving parameterized subtropical marine boundary layer
 715 cloudiness. *Mon. Wea. Rev.*, *147*, 3241–3260.
- 716 Souza, A. N., Wagner, G. L., Ramadhan, A., Allen, B., Churavy, V., Schloss, J.,
 717 ... Ferrari, R. (2020). Uncertainty quantification of ocean parameteriza-
 718 tions: Application to the k-profile-parameterization for penetrative convection.
 719 *Journal of Advances in Modeling Earth Systems*, *12*(12), e2020MS002108.
 720 Retrieved from [https://agupubs.onlinelibrary.wiley.com/doi/abs/](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020MS002108)
 721 [10.1029/2020MS002108](https://doi.org/10.1029/2020MS002108) (e2020MS002108 10.1029/2020MS002108) doi:
 722 <https://doi.org/10.1029/2020MS002108>
- 723 Stephens, G. L. (2005). Cloud feedbacks in the climate system: A critical review. *J.*
 724 *Climate*, *18*, 237–273. doi: 10.1175/JCLI-3243.1
- 725 Stevens, B., Moeng, C.-H., Ackerman, A. S., Bretherton, C. S., Chlond, A., de
 726 Roode, S., ... Zhu, P. (2005). Evaluation of large-eddy simulations via obser-
 727 vations of nocturnal marine stratocumulus. *Mon. Wea. Rev.*, *133*, 1443–1462.
 728 doi: 10.1175/MWR2930.1
- 729 Stuart, A. M. (2010). Inverse problems: a Bayesian perspective. *Acta Numerica*, *19*,
 730 451–559.
- 731 Tan, Z., Kaul, C. M., Pressel, K. G., Cohen, Y., Schneider, T., & Teixeira, J. (2018).
 732 An extended eddy-diffusivity mass-flux scheme for unified representation of
 733 subgrid-scale turbulence and convection. *J. Adv. Model. Earth Sys.*, *10*, 770-
 734 800. doi: 10.1002/2017MS001162

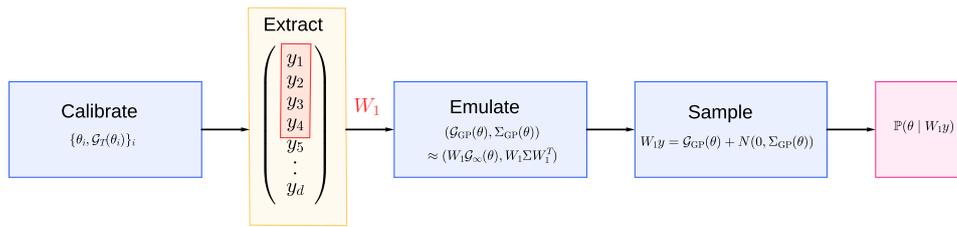


Figure A1. Procedure of the uncertainty quantification framework (blue), to produce output (pink). A restriction operator W_1 extracting a subset of the GCM output (yellow); the subsequent emulate and sample stages may be performed in parallel for different W_i .

735 Tarantola, A. (2005). *Inverse problem theory and methods for model parameter esti-*
 736 *mation* (Vol. 89). siam.

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

740 Van Roekel, L., Adcroft, A. J., Danabasoglu, G., Griffies, S. M., Kauffman, B.,
 741 Large, W., ... Schmidt, M. (2018). The kpp boundary layer scheme for the
 742 ocean: Revisiting its formulation and benchmarking one-dimensional simula-
 743 tions relative to les. *Journal of Advances in Modeling Earth Systems*, 10(11),
 744 2647–2685.

745 Vial, J., Dufresne, J.-L., & Bony, S. (2013). On the interpretation of inter-model
 746 spread in CMIP5 climate sensitivity estimates. *Clim. Dyn.*, 41, 3339–3362. doi:
 747 10.1007/s00382-013-1725-9

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

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

754 Wei, H.-H., & Bordoni, S. (2018). Energetic constraints on the ITCZ position in ide-
 755 alized simulations with a seasonal cycle. *J. Adv. Model. Earth Sys.*, 10. doi: 10
 756 .1029/2018MS001313

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

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

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

767 Zhang, M., Bretherton, C. S., Blossey, P. N., Austin, P. H., Bacmeister, J. T., Bony,
 768 S., ... others (2013). CGILS: Results from the first phase of an interna-
 769 tional project to understand the physical mechanisms of low cloud feedbacks
 770 in general circulation models. *J. Adv. Model. Earth Sys.*, 5, 826–842. doi:
 771 10.1002/2013MS000246

772 **Appendix A Calibrate-Emulate-Sample with design**

773 One fundamental aspect of this work, is the ability to efficiently calculate the the
 774 posterior distribution (in particular the covariance), which is needed to calculate the util-
 775 ity function (6) at all designs. We present a methodology: calibrate-extract-emulate-sample,
 776 (CEES) which allows for the calculation of posterior covariance for all designs with just
 777 $\mathcal{O}(100)$ evaluations of our forward model.

778 The methodology is based on the calibrate-emulate-sample (CES) algorithm, for
 779 full details of the individual stages see Cleary et al. (2021a); Dunbar et al. (2021), here
 780 we present an overview and motivation. The core purpose of CES is to form a compu-
 781 tationally cheap statistical emulator of \mathcal{G}_∞ from intelligently chosen samples of \mathcal{G}_T ; then
 782 one is able to solve the Bayesian inverse problem for the emulated \mathcal{G}_∞ with a sampling
 783 method. We achieve this by using Gaussian process emulators, trained on the samples
 784 of the (noisy and expensive) forward map. The Gaussian process mean function is nat-
 785 urally smoother than the data it is trained on (Kennedy & O’Hagan, 2001a; Notz et al.,
 786 2018), and is capable of representing the the noise of the forward model within the co-
 787 variance function, leading to a smooth likelihood function that is quick to evaluate. The
 788 training points for the Gaussian Process are given by applying an optimization scheme,
 789 EKI (Ensemble Kalman Inversion), (Iglesias et al., 2013; Schillings & Stuart, 2017) to
 790 the inverse problem in its finite-time averaged form (3). Theoretical work shows that noisy
 791 continuous-time versions of EKI exhibit an averaging effect that skips over fluctuations
 792 superimposed onto the ergodic averaged forward model (Duncan et al., 2021), and sim-
 793 ilar effects are observed in practice for EKI, thus it is highly suited to optimization of
 794 parameters coming from a noisy, expensive model without derivatives available. Ensemble
 795 Kalman methods are scalable to very high dimensional problems (Kalnay, 2003; Oliver
 796 et al., 2008) with use of localization and regularization.

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

1. **Calibrate:** We solve (3) with \mathbf{y} using evaluations of \mathcal{G}_T in an optimization sense,
 where we minimize the functional.

$$\Phi_T(\theta, \mathbf{y}) = \|\mathbf{y} - \mathcal{G}_T(\theta; \mathbf{v}^{(0)})\|_{2\Sigma}^2. \tag{A1}$$

799 The notation $\|\cdot\|_\Sigma = \|\Sigma^{-\frac{1}{2}} \cdot\|_2$ is the Mahalanobis distance. The weight 2Σ is
 800 the sum of internal variability of \mathcal{G}_T and of \mathbf{y} . The optimization is performed us-
 801 ing several iterations the Ensemble Kalman Inversion algorithm. This leads to $\{\theta_i, \mathcal{G}_j(\theta_j)\}_{j=1}^J$
 802 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
 803 forward map, $\{\theta_j, W_k \mathcal{G}_T(\theta_j)\}_{j=1}^J$, and apply the following **Emulate(k)** and **Sam-**
 804 **ple(k)** stages.
3. **Emulate(k):** We decorrelate the data space with an SVD on the internal vari-
 805 ability covariance Σ , yielding a change-of-basis matrix V . We train Gaussian pro-
 806 cess emulators, on the pairs $\{\theta_j, VW_k \mathcal{G}_T(\theta_j)\}_{j=1}^J$, yielding $(\mathcal{G}_{GP}(\theta), \Sigma_{GP}(\theta))$, where
 807 $\mathcal{G}_{GP} \approx VW_k \mathcal{G}_\infty(\theta)$ (crucially \mathcal{G}_∞ and not \mathcal{G}_T) and $\Sigma_{GP}(\theta) \approx VW_k \Sigma W_k^T V^T$.
4. **Sample(k):** We now solve the inverse problem (5), This is feasible as the emu-
 808 lator provides us with an approximation of \mathcal{G}_∞ (not just \mathcal{G}_T). The posterior dis-
 809 tribution associated with (5) is proportional to a product of prior and likelihood
 contribution from Bayes theorem. Explicitly, for a Gaussian prior $N(\mathbf{m}, C)$ on the
 computational parameters, and the likelihood dependent on the emulator, we write
 the MCMC objective function (also known as the log-posterior) as

$$\begin{aligned} \Phi_{\text{MCMC}}(\theta, VW_k \mathbf{y}) &= \frac{1}{2} \|VW_k \mathbf{y} - \mathcal{G}_{GP}(\theta)\|_{\Sigma_{GP}(\theta)}^2 + \frac{1}{2} \log \det \Sigma_{GP}(\theta) \\ &\quad + \frac{1}{2} \|\theta - \mathbf{m}\|_C^2. \end{aligned}$$

The posterior is then given by

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

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

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

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