CALIBRATING EARTH SYSTEM MODELS WITH BAYESIAN OPTIMAL EXPERIMENTAL DESIGN

Tim Reichelt
University of Oxford

Shahine Bouabid
University of Oxford

Luke Ong
Nanyang Technological University

Duncan Watson-Parris
University of California San Diego

Tom Rainforth
University of Oxford

ABSTRACT

Earth system models (ESMs) are complex climate simulations that are critical for projecting future climate change and its impacts. However, running ESMs is extremely computationally expensive, limiting the number of simulations that can be performed. This results in significant uncertainty in key climate metrics estimated from ESM ensembles. We propose a Bayesian optimal experimental design (BOED) approach to efficiently calibrate ESM simulations to observational data by actively selecting the most informative input parameters. BOED optimises the expected information gain (EIG) to select the ESM input parameter to reduce the final uncertainty estimates in the climate metrics of interest. Initial results on a synthetic benchmark demonstrate our approach can more efficiently reduce uncertainty compared to common sampling schemes like Latin hypercube sampling.

1 INTRODUCTION

Earth system models (ESMs) (Flato, 2011) constitute an extensive simulation of the Earth's atmosphere and ocean fluid dynamics, based on energy, mass, and momentum conservation equations. As such, they stand as key tools to examine the climate system, project its response to anthropogenic emissions, and study its underlying drivers. An ESM run generates a rich range of spatially resolved climate variables, including surface temperatures, precipitations and winds, which we shall refer to as *simulated variables*. These simulated variables are then used to diagnose important characteristic properties of the climate system, such as the top-of-atmosphere effective radiative forcing, or timescales and equilibrium responses to changes in greenhouse gas and aerosol emissions (Allen et al., 2009; Collins et al., 2014; Levy et al., 2013), which we shall refer to as *latent variables*.

Accurately diagnosing these latent variables is critical to improve our understanding of how the climate system reacts to changes in emissions and better constrain our idea of what an uncertain future warming may look like. For example, diagnosing the so called *climate feedback parameter* allows to quantify the magnitude of the Earth's feedback response to a given change in global mean surface temperature. Diagnosing the *aerosol effective radiative forcing* allows to quantify how much of the current warming is masked by the aerosol cooling effect. Better understanding and constraining these effects is crucial for planning and informing adaption strategies.

The latent variable diagnosis is typically achieved by running multiple ESM simulations for perturbed parameter ensembles, and then using the ensemble of simulated outputs to compute an estimate of the latent variables. However, each simulation is extremely computationally expensive—for reference, running the CESM2 model (Danabasoglu et al., 2020) for a single year ahead takes about 2000 core hours on a supercomputer. Using large ensembles of simulations is therefore computationally prohibitive. As a result, a large uncertainty remains on the estimates of these latent variables. In the case of the the aerosol effective radiative forcing mentioned earlier, the uncertainty bounds are so wide that the forcing could offset global warming or double its effects (Boucher et al., 2013).

A common practice to create ensemble members is to perturb the input parameters using Latin hypercube sampling (LHS) (McKay et al., 2000) or some other form of random sampling (Lee et al., 2011; Qian et al., 2015; Watson-Parris et al., 2020). However, this is potentially wasteful because

it does not take into account the observed data and therefore potentially evaluates the simulators in parameter regions which are unlikely to produce simulations which match the observed data.

As an alternative we propose to select the simulator inputs in a more active manner, directly accounting for the observed data. To achieve this we leverage tools from Bayesian optimal experimental design (BOED) (Lindley, 1956; Chaloner & Verdinelli, 1995; Ryan et al., 2016; Rainforth et al., 2023). In order to leverage BOED approaches, we postulate a surrogate model in the form of a Gaussian process (GP) for the simulator which is trained on the real-world observed data and the simulated data. The input parameters to the simulator are then selected based on optimizing the expected information gain (EIG) which measures the expected reduction in Shannon entropy after having observed a given data point (Lindley, 1956).

In this proposal we are going to outline: 1) how to frame the problem of constraining the uncertainty on ESMs latent variables as a probabilistic inference problem, 2) show how this formulation naturally leads to a BOED mechanism to select the simulator input parameters, and 3) present promising initial results on a synthetic benchmark.

2 Optimal Experimental Design for Earth System Models

We will use $\xi_t \in \Xi$ to denote the *simulator inputs* of the tth run, consistent with terminology in the experimental design literature we will sometimes refer to them as *designs*. The input parameter space Ξ is often constrained to be the unit hypercube. The *simulated* and *latent variables* produced by the ESM are denoted as $y_{2,t} \in \mathbb{R}$ and $z_{2,t} \in \mathbb{R}$, respectively. We will use $y_1 \in \mathbb{R}$ to denote the actual observed values for the simulated variables and $z_1 \in \mathbb{R}$ to denote the corresponding *latent variables*.[1] We also define the random variable $\theta \in \Xi$ which is meant to represent the setting of input parameters which most closely reproduces the observed data.

In summary, y_1 is observed data that is fixed; $\mathbf{y}_2 = [y_{2,1}, \ldots, y_{2,T}]$ and $\mathbf{z}_2 = [z_{2,1}, \ldots, z_{2,T}]$ are values we observe as we execute the ESM; θ and z_1 are random variables that we need to infer; and $\xi = [\xi_1, \ldots, \xi_T]$ are the input parameters that we get to choose. Our overall goal is then to select the input parameters ξ which are maximally informative about constraining our beliefs about θ and z_1.

2.1 Joint Probabilistic Model

In a order to make an informed decision about how to select the input parameters we need to postulate a joint probabilistic model which encapsulates all the variables of interest. The key modelling decision that we will make is that $\mathbf{y}_2, \mathbf{z}_2, y_1, z_1$ are all jointly modeled by a Gaussian Process (GP):

$$p(y_1, \mathbf{y}_2, z_1, \mathbf{z}_2 \mid \theta, \xi) = \mathcal{N}\left(\begin{bmatrix} y_1 \\ \mathbf{y}_2 \\ z_1 \\ \mathbf{z}_2 \end{bmatrix}; \mathbf{0}, \begin{bmatrix} K(\theta,\theta) & K(\theta,\xi) & \mathbf{0} & \mathbf{0} \\ K(\xi,\theta) & K(\xi,\xi) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & K(\theta,\theta) & K(\theta,\xi) \\ \mathbf{0} & \mathbf{0} & K(\xi,\theta) & K(\xi,\xi) \end{bmatrix} \right) \tag{1}$$

where we follow the notation of Williams & Rasmussen (2006) and use $K(\cdot,\cdot)$ to denote the covariance matrix generated by the GP kernel and $\mathbf{0}$ is an appropriately shaped matrix of zeros. Note that we are modelling the y and z space completely independently. This means that the proxy observations y_1 only influence our inference on z_1 through the inferred simulator parameters θ. Furthermore, due to the properties of a GP, marginals and conditionals can be computed analytically.

By specifying a prior on θ this model gives us a prior distribution $p(z_1, \theta) = p(z_1 \mid \theta)p(\theta)$ on all our unknown random variables in this model; $p(z_1 \mid \theta) = \mathcal{N}(z_1; \mathbf{0}, K(\theta,\theta))$ is defined implicitly through Eq. (1). After observing data this leads to the posterior distribution

$$p(z_1, \theta \mid y_1, \mathbf{y}_2, \mathbf{z}_2, \xi) \propto p(y_1, \mathbf{y}_2, \mathbf{z}_2 \mid z_1, \theta, \xi)p(z_1, \theta). \tag{2}$$

The likelihood $p(y_1, \mathbf{y}_2, \mathbf{z}_2 \mid z_1, \theta, \xi)$ can be derived from the full GP model in Eq. (1). However, the posterior will in general not be analytically tractable so we need to run approximate inference algorithms to estimate it (Robert & Casella, 1999; Brooks et al., 2011; Blei et al., 2017).

[1] Here we assume scalar outputs but the approach generalizes to vector outputs and spatial grids as well.

2.2 Optimizing the Expected Information Gain

The intuition that we want to maximize the information gained from running the simulator can be formalised using the methods of Bayesian optimal experimental design (BOED) (Lindley, 1956; Ryan et al., 2016). BOED uses the reduction in Shannon entropy Shannon (1948) between the prior and the posterior to define the *information gain*:

$$\mathrm{I}(\mathbf{y}_2, \mathbf{z}_2, \xi) := H\left[p(\theta, z_1 \mid y_1)\right] - H\left[p(\theta, z_1 \mid y_1, \mathbf{y}_2, \mathbf{z}_2, \xi)\right]. \tag{3}$$

This quantity depends on the simulator outputs $\mathbf{y}_2, \mathbf{z}_2$ and can therefore not be directly used to find the optimal input parameters. Hence, BOED optimises the *expected information gain* (EIG)

$$\mathcal{I}(\xi) := \mathbb{E}_{p(\mathbf{y}_2, \mathbf{z}_2 | y_1, \xi)}\left[\mathrm{I}(\mathbf{y}_2, \mathbf{z}_2, \xi)\right] = \mathbb{E}_{p(\theta, z_1 | y_1) p(\mathbf{y}_2, \mathbf{z}_2 | z_1, y_1, \theta, \xi)}\left[\log \frac{p(\mathbf{y}_2, \mathbf{z}_2 \mid y_1, z_1, \theta, \xi)}{p(\mathbf{y}_2, \mathbf{z}_2 \mid y_1, \xi)}\right].$$

The marginal $p(\mathbf{y}_2, \mathbf{z}_2 \mid y_1, \xi) = \mathbb{E}_{p(\theta, z_1 | y_1)}\left[p(\mathbf{y}_2, \mathbf{z}_2 \mid y_1, z_1, \theta, \xi)\right]$ is generally intractable, hindering the use of conventional Monte Carlo (MC) estimation to evaluate and optimise the EIG. Rather, estimating the EIG naively would lead to a form of nested MC estimator; these are often biased and have poor computational scaling properties (Rainforth et al., 2018; 2023).

Instead, Foster et al. (2020) introduced bounds for the EIG that can be used to simultaneously estimate the EIG and optimize the designs ξ. We will follow their approach and leverage their prior-contrastive estimation (PCE) bound

$$\mathcal{I}(\xi) \geq \mathcal{L}(\xi, L) := \mathbb{E}\left[\log \frac{p(\mathbf{y}_2, \mathbf{z}_2 \mid y_1, z_{1,0}, \theta_0, \xi)}{\frac{1}{L+1}\sum_{l=0}^{L} p(\mathbf{y}_2, \mathbf{z}_2 \mid y_1, z_{1,l}, \theta_l, \xi)}\right] \tag{4}$$

where the expectation is taken w.r.t. $p(\theta_{0:L}, z_{1,0:L}) p(\mathbf{y}_2, \mathbf{z}_2 \mid y_1, z_{1,0}, \theta_0, \xi)$. The designs are then chosen by finding $\operatorname{argmax}_\xi \mathcal{L}(\xi, L)$ using standard stochastic optimization techniques.

Adaptive Designs. The bound in Eq. (4) gives us a mechanism for *static designs*, i.e. choosing all the input parameters before collecting any simulator runs. However, the real power of BOED comes to fruition in settings in which we want to adaptively select designs as we collect more data. In order to produce adaptive designs we will use the recently introduced *deep adaptive design (DAD)* (Foster et al., 2021) method which trains a policy network that learns to propose a new design ξ_t at step t given the previous observations and designs. More details on this approach can be found in App. A.

3 Initial Results and Ongoing Work

We have validated our methodology in a proof of concept in which we use a GP model as a drop-in replacement for the expensive climate simulator. In this experiment we simulate some ground-truth values y_1 and z_1 and compare different methodologies to select the input parameters ξ to produce simulator outputs $\mathbf{y}_2, \mathbf{z}_2$: deep adaptive design (*DAD*, Foster et al. (2021)); static designs of optimizing Eq. (4) (*Static*); Latin hypercube sampling (*LHS*) (McKay et al., 2000); generating random samples from the prior $p(\theta)$ (*Random*); and evenly spacing the inputs in the constrained input space (*Even*). Note, all the data $y_1, z_1, \mathbf{y}_2$, and $\mathbf{z}_2$ is generated from a GP but at time step t the adaptive design method only gets access to the observed simulated variables y_1 and the data generated from the "simulator" $y_{1:t}, z_{1:t}$.

Table 1: EIG lower bounds. Errors show ± 1 standard error, computed over 1024 rollouts.

Method	Lower Bound (↑)
DAD	**2.86 ± 0.06**
Static	2.30 ± 0.05
LHS	2.16 ± 0.05
Random	1.93 ± 0.05
Even	1.83 ± 0.05

Tab. 1 compares the BOED based design strategies against the baselines based on their achieved EIG lower bound. In our case, the EIG is not only a useful objective to select the designs but is also the metric that we care about at evaluation time. As a reminder, our goal is to reduce our uncertainty in our estimates of the latent variables and the EIG is a natural mechanism to measure this reduction in uncertainty. Our results show that the DAD method is by far the best at achieving a high EIG, clearly outperforming all other baselines.

Conclusion. We have outlined a mechanism for actively selecting the input parameters of climate simulators to calibrate them to observational data and validated our approach in a synthetic setting showing that the approach promises increased uncertainty reduction. Currently, we are working on further scaling up our approach, validating it on an actual ESM, namely the Single Column Atmosphere Model Version 6 (Gettelman et al., 2019), and comparing it against more sophisticated calibration methods that go beyond random sampling (Cleary et al., 2021; Jiang & Willett, 2022).

ACKNOWLEDGEMENTS

This project is partly supported by UK EPSRC with the grants EP/S024050/1 and EP/Y037200/1, and by the National Research Foundation, Singapore, under its RSS Scheme (NRF-RSS2022-009).

REFERENCES

Myles R Allen, David J Frame, Chris Huntingford, Chris D Jones, Jason A Lowe, Malte Meinshausen, and Nicolai Meinshausen. Warming caused by cumulative carbon emissions towards the trillionth tonne. *Nature*, 2009.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Olivier Boucher, David Randall, Paulo Artaxo, Christopher Bretherton, Christopher Feingold, Piers Forster, Veli-Matti Kerminen, Yutaka Kondo, Hong Liao, Ulrike Lohmann, Philip Rasch, S.K. Satheesh, Steven Sherwood, Bjorn Stevens, and Xiao-Ye Zhang. Clouds and aerosols. pp. 571 – 657, 2013.

Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of markov chain monte carlo*. CRC press, 2011.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical science*, pp. 273–304, 1995.

Emmet Cleary, Alfredo Garbuno-Inigo, Shiwei Lan, Tapio Schneider, and Andrew M Stuart. Calibrate, emulate, sample. *Journal of Computational Physics*, 424:109716, 2021.

M. Collins, R. Knutti, J. Arblaster, J.-L. Dufresne, T. Fichefet, P. Friedlingstein, X. Gao, W. J. Gutowski, Krinner G. Johns, T., M. Shongwe, C. Tebaldi, A. J. Weaver, and M. Wehner. *Long-term Climate Change: Projections, Commitments and Irreversibility Pages 1029 to 1076*. 2014.

Gokhan Danabasoglu, J-F Lamarque, J Bacmeister, DA Bailey, AK DuVivier, Jim Edwards, LK Emmons, John Fasullo, R Garcia, Andrew Gettelman, et al. The community earth system model version 2 (cesm2). *Journal of Advances in Modeling Earth Systems*, 12(2):e2019MS001916, 2020.

Gregory M. Flato. Earth system models: an overview. *WIREs Climate Change*, 2011.

Adam Foster, Martin Jankowiak, Matthew O'Meara, Yee Whye Teh, and Tom Rainforth. A unified stochastic gradient approach to designing bayesian-optimal experiments. In *International Conference on Artificial Intelligence and Statistics*, pp. 2959–2969. PMLR, 2020.

Adam Foster, Desi R Ivanova, Ilyas Malik, and Tom Rainforth. Deep adaptive design: Amortizing sequential bayesian experimental design. In *International Conference on Machine Learning*, pp. 3384–3395. PMLR, 2021.

A Gettelman, JE Truesdale, JT Bacmeister, PM Caldwell, RB Neale, PA Bogenschutz, and IR Simpson. The single column atmosphere model version 6 (scam6): Not a scam but a tool for model evaluation and development. *Journal of Advances in Modeling Earth Systems*, 11(5):1381–1401, 2019.

Ruoxi Jiang and Rebecca Willett. Embed and emulate: Learning to estimate parameters of dynamical systems with uncertainty quantification. *Advances in Neural Information Processing Systems*, 35:11918–11933, 2022.

LA Lee, KS Carslaw, KJ Pringle, GW Mann, and DV Spracklen. Emulation of a complex global aerosol model to quantify sensitivity to uncertain parameters. *Atmospheric Chemistry and Physics*, 11(23):12253–12273, 2011.

Hiram Levy, Larry W Horowitz, M Daniel Schwarzkopf, Yi Ming, Jean-Christophe Golaz, Vaishali Naik, and V Ramaswamy. The roles of aerosol direct and indirect effects in past and future climate change. *Journal of Geophysical Research: Atmospheres*, 2013.

Dennis V Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

Michael D McKay, Richard J Beckman, and William J Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 42(1):55–61, 2000.

Yun Qian, Huiping Yan, Zhangshuan Hou, Gardar Johannesson, Stephen Klein, Donald Lucas, Richard Neale, Philip Rasch, Laura Swiler, John Tannahill, et al. Parametric sensitivity analysis of precipitation at global and local scales in the community atmosphere model cam5. *Journal of Advances in Modeling Earth Systems*, 7(2):382–411, 2015.

Tom Rainforth, Rob Cornish, Hongseok Yang, Andrew Warrington, and Frank Wood. On nesting monte carlo estimators. In *International Conference on Machine Learning*. PMLR, 2018.

Tom Rainforth, Adam Foster, Desi R Ivanova, and Freddie Bickford Smith. Modern bayesian experimental design. *arXiv preprint arXiv:2302.14545*, 2023.

Christian P Robert and George Casella. *Monte Carlo statistical methods*, volume 2. Springer, 1999.

Elizabeth G Ryan, Christopher C Drovandi, James M McGree, and Anthony N Pettitt. A review of modern computational algorithms for bayesian optimal design. *International Statistical Review*, 84(1):128–154, 2016.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Duncan Watson-Parris, Nicolas Bellouin, LT Deaconu, Nick AJ Schutgens, Masaru Yoshioka, Leighton Anunda Regayre, Kirsty J Pringle, Jill S Johnson, CJ Smith, KS Carslaw, et al. Constraining uncertainty in aerosol direct forcing. *Geophysical Research Letters*, 2020.

Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

A Adaptive Designs

Due to the consistency properties of Bayesian inference, an appealing property of Bayesian optimal experimental design is that it naturally prescribes a way to compute designs *adaptively*, i.e. instead of computing T fixed designs before collecting any data we can choose a new design based on the previously observed data points. After having collected $t-1$ simulator runs we have access to the simulator input parameters and the generated outputs which together make up the *history* $h_{t-1} = \{(\xi_k, y_{2,k}, z_{2,k})\}_{k=1}^{t-1}$.

Foster et al. (2021) introduced deep adaptive design (DAD) which trains an amortized policy network, π, that learns to optimize the EIG. At each step the policy takes in all the previous history and generates a new design, i.e. $\xi_t = \pi(h_{t-1})$. This defines a new likelihood $p(\mathbf{y}_2, \mathbf{z}_2 \mid y_1, \theta, z_1, \pi)$ which is defined by using the policy to autoregressively propose new designs at each step. Crucially, due to the additive properties of Shannon entropy, the total expected information gain for the policy is analogous to the EIG for the static designs

$$\mathcal{I}(\pi) = \mathbb{E}_{p(\theta,z_1|y_1)p(\mathbf{y}_2,\mathbf{z}_2|z_1,y_1,\theta,\pi)}\left[\log \frac{p(\mathbf{y}_2, \mathbf{z}_2 \mid y_1, z_1, \theta, \pi)}{p(\mathbf{y}_2, \mathbf{z}_2 \mid y_1, \pi)}\right]. \tag{5}$$

Similarly, the policy network can be trained using the same lower bound as the static designs:

$$\mathcal{L}(\pi, L) := \mathbb{E}\left[\log \frac{p(\mathbf{y}_2, \mathbf{z}_2 \mid y_1, z_{1,0}, \theta_0, \pi)}{\frac{1}{L+1}\sum_{l=0}^{L} p(\mathbf{y}_2, \mathbf{z}_2 \mid y_1, z_{1,l}, \theta_l, \pi)}\right] \tag{6}$$

with expectation taken w.r.t. $p(\theta_{0:L}, z_{1,0:L})p(\mathbf{y}_2, \mathbf{z}_2 \mid y_1, z_{1,0}, \theta_0, \pi)$. We follow the convention of Foster et al. (2021) to parameterize out policy network with a neural network (using the same architecture choices) and train the parameters using stochastic gradient ascent on Eq. (6). While this might conceptually seem like a simple shift, it significantly increases our ability to learn good designs because we are now able to *learn to adapt our designs to previously observed data*. As we saw in our initial results that can lead to significant reductions in uncertainty for the estimates of our latent variables.