Calibrating Earth System Models with Bayesian Optimal Experimental Design

Tim Reichelt[1], Shahine Bouabid[1], Luke Ong[1,2], Duncan Watson-Parris[3], Tom Rainforth[1]

[1] University of Oxford [2] Nanyang Technological University, Singapore [3] University of California San Diego

Overview

- Earth system models (ESMs) are complex climate simulations that are critical for projecting future climate change and its impacts.
- Running ESMs is extremely computationally expensive, limiting the number of simulations that can be performed.
- 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.
- Initial results on a synthetic benchmark demonstrate our approach can more efficiently reduce uncertainty compared to common sampling schemes like Latin hypercube sampling.

Figure: High-level schema of calibration process. Our key innovation lies in using BOED to select the next simulator input parameters.

Probabilistic Surrogate Model

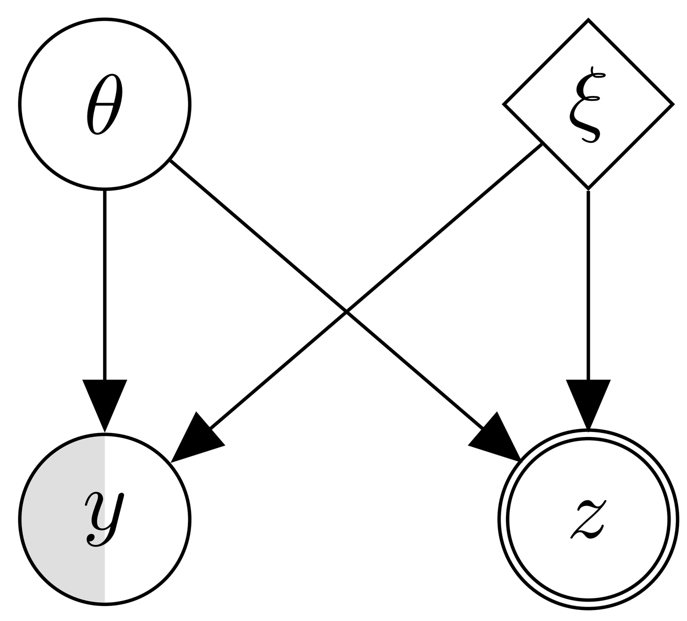

Figure: Graphical representation of Gaussian Process surrogate model.

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.

- $\xi_t \in \Xi$ denote the *simulator inputs* of the tth run
- $y_{2,t} \in \mathbb{R}$ and $z_{2,t} \in \mathbb{R}$ denote the *simulated* and *latent variables* produced by the ESM, respectively
- $y_1 \in \mathbb{R}$ denote the actual observed values for the simulated variables
- $z_1 \in \mathbb{R}$ to denote the corresponding *latent variables*.
- $\theta \in \Xi$ represents the setting of input parameters which most closely reproduces the observed data

The key modelling decision that we will make is that all variables are jointly modeled by a Gaussian Process (GP):

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

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.

Optimizing the Expected Information Gain

BOED uses the reduction in Shannon entropy between the prior and the posterior to define the ***information gain***:

$$\mathrm{I}(\mathbf{y}_2, \mathbf{z}_2, \xi) := H\left[p(\theta, z_1 \mid y_1)\right] - H\left[p(\theta, z_1 \mid y_1, \mathbf{y}_2, \mathbf{z}_2, \xi)\right]. \quad (1)$$

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)***

$$\mathcal{I}(\xi) := \mathbb{E}_{p(\mathbf{y}_2,\mathbf{z}_2|y_1,\xi)}\left[\mathrm{I}(\mathbf{y}_2, \mathbf{z}_2, \xi)\right] \quad (2)$$

$$= \mathbb{E}_{p(\theta,z_1|y_1)p(\mathbf{y}_2,\mathbf{z}_2|z_1,y_1,\theta,\xi)}\left[\log \frac{p(\mathbf{y}_2, \mathbf{z}_2 \mid y_1, z_1, \theta, \xi)}{p(\mathbf{y}_2, \mathbf{z}_2 \mid y_1, \xi)}\right]. \quad (3)$$

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 [1].
Instead, [2] 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

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

The designs are then chosen by finding $\mathrm{argmax}_\xi \mathcal{L}(\xi, L)$ using standard stochastic optimization techniques. Designs can be selected in a ***static***, i.e. all designs determined before any simulator runs, or an ***adaptive***, i.e. after each simulator run a new design computed based on the output of previous runs.

Initial Results and Future Work

Table: EIG lower bounds for initial experiments in which we use our postulated GP model as the simulator. Errors show ± 1 standard error, computed over 1024 rollouts.

Method	EIG Lower Bound (↑)
Adaptive BOED	**2.86 ± 0.06**
Static BOED	2.30 ± 0.05
LHS	2.16 ± 0.05
Random	1.93 ± 0.05
Even Spacing	1.83 ± 0.05

- BOED approaches provide higher expected information gain, i.e. a larger reduction in posterior uncertainty, compared to approaches based on random sampling

Future work

- Validating our approach on the Single Column Atmosphere Model Version 6 [3]
- Comparing against more active learning baselines

References

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

[2] 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*, pages 2959–2969. PMLR, 2020.

[3] 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.