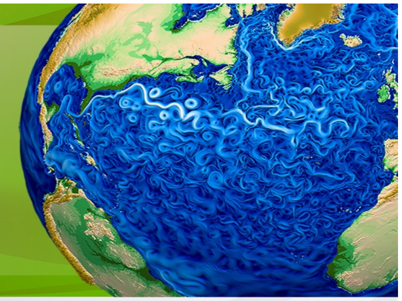


Optimization of sensor networks for climate models:

Overview, progress and future plans

Daniel Ricciuto[1], Khachik Sargsyan[2], Youssef Marzouk[3], Dan Lu[1], Cosmin Safta[2], Miroslav Stoyanov[1], Kate Evans[1], Jayanth Jagalur Mohan[3], and Fengyi Li[3]

[1] Oak Ridge National Laboratory; [2] Sandia National Laboratories; [3] Massachusetts Institute of Technology

Background:

Earth System models like E3SM are computationally expensive to perform even a single forward simulation, but such models contain hundreds of uncertain parameters and algorithms. Uncertainty quantification in or calibration of an ESM requires ensembles, which may need to be quite large given the dimensionality of the problem. Many of the relevant land and atmospheric observations are at "point" scale (1 model grid cell or smaller) – how can we best use these?

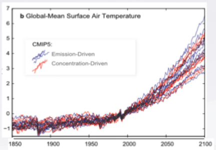
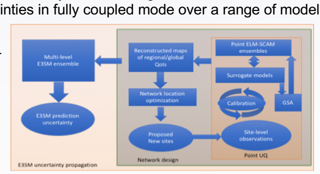


Figure 1: Climate model uncertainty is estimated using intercomparisons, in which each modeling center performs a single simulation or small ensemble. Structural or parametric uncertainty is usually not considered in individual models. A variety of approaches among centers make it hard to attribute biases to specific differences. Figure from Friedlingstein et al., 2014.

In this project, we are conducting global sensitivity analysis (GSA) to identify the most important parameters and processes in the E3SM land and atmosphere models (ELM and EAM), and then calibrating a network of single-column coupled land-atmosphere models using point scale data using surrogate approaches. We then reconstruct global maps of quantities of interest with posterior uncertainties. This new framework is being used to optimize placement of new observations for maximum uncertainty reduction. Finally, we are using a Multi-level Monte Carlo (MLMC) to propagate uncertainties in fully coupled mode over a range of model fidelities and resolutions.

Figure 2: Project workflow. Site-level observations for calibration include land-atmosphere flux measurements, temperature and precipitation. The network optimization approach will suggested new measurement locations.



Surrogate modeling

Surrogate modeling can greatly reduce the computational costs in global sensitivity analysis and model calibration which involve large ensemble model simulations.

Bayesian neural networks (BNN) can build accurate surrogate models based on a small number of model simulation results with consideration of model uncertainty. We used 900 ELM samples to build an accurate surrogate of 49260 annual GPP for 30 years at 1642 locations. The mean-squared-error between the surrogate and ELM simulations of the 49260 GPP at 100 validation samples is 0.3.

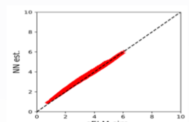


Figure 3: NN and ELM simulation of 49260 GPP outputs.

Global sensitivity analysis

GSA is a variance-based decomposition to measure the fractional contributions of each parameter towards the total variance of selected quantities of interest (QoIs):

- Total effect Sobol indices

$$S_i^T = \frac{\mathrm{E}_{\boldsymbol{x}_{-i}}\left[\mathrm{V}_{x_i}[f(\boldsymbol{x})|\boldsymbol{x}_{-i}]\right]}{\mathrm{V}\left[f(\boldsymbol{x})\right]} = 1 - \frac{\mathrm{V}_{\boldsymbol{x}_{-i}}\left[\mathrm{E}_{x_i}[f(\boldsymbol{x})|\boldsymbol{x}_{-i}]\right]}{\mathrm{V}\left[f(\boldsymbol{x})\right]}$$

- Sobol indices estimates
  - Polynomial Chaos Expansions – exploit orthogonality of basis terms
  - Random Sampling – need computationally cheap models

Here we compare the skill of machine-learning models (e.g. deep neural networks) with sparse learning techniques to identify the optimal number of parameters that are influential for the selected QoIs.

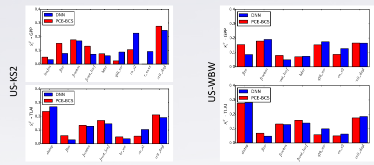


Figure 4: Comparison of GSA using polynomial chaos-based expansion (PCE) vs. a deep neural network (DNN) for ELM at 2 flux measurement sites.

Model calibration:

After key parameters are identified using GSA, we construct a surrogate model to be used for calibration using observations. Because of the computational efficiency of the surrogate model, we can use Markov Chain Monte Carlo (MCMC) approaches for calibration. In addition to parameter errors, model structural error is a major challenge for predictive science. Predictive skills of climate models do not improve due to substantial structural error. Conventional statistical for bias correction do not allow meaningful predictions with corrections, and entangle model error with data noise. Here, we pursue model error embedding approach [Sargsyan, et. al. 2015, 2018] which has the advantages of:

- Physics-driven model correction
- Meaningful extrapolation to full set of QoI predictions
- Disambiguation between model error and data noise

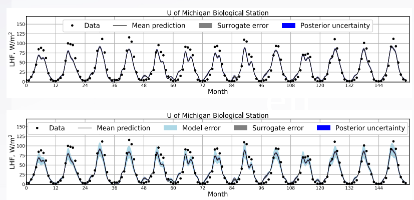


Figure 5: Calibration of ELM with latent heat flux (LHF) observations at the UMBS flux site. With the embedded model error approach (lower figure), posterior uncertainties are more realistic as compared to the traditional calibration approach (upper figure).

Reconstruction and network optimization:

The goal is to reproduce model spatial variations using a network of single column simulations. A simple version of ELM is running at MIT and an interface with an optimal experimental design (OED) system is being developed.

Current strategy and next steps:

- OED problem using Bayesian linear-Gaussian setting, with "greedy" selection strategy for some simple synthetic problems.
- Provide performance guarantees for the greedy strategy that bound how much worse it could be than the combinatorial optimum

The example below illustrates how sites can be selected to best reconstruct the original high-resolution model spatial field of a single simulation using a simple bilinear interpolation. In the future, the full OED system will also consider model uncertainties.

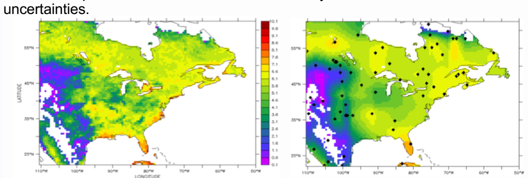


Figure 6: Gross primary productivity at 0.5x0.5 degree resolution from ELM (left), and reconstructed GPP using 64 selected grid cells (right). The locations of the 64 points are optimized using a genetic algorithm to minimize the root mean squared error between the model and interpolated GPP fields. Here the model is treated as "truth", but the selected points are sensitive to the reconstruction (interpolation) method and model configuration.

Multi-level Monte Carlo approach:

Multi-level Monte Carlo (MLMC) method can solve high-dimensional uncertainty quantification (UQ) problems efficiently.

Standard MC estimation conducts repeated sampling on a single numerical model with high resolution. MLMC estimation conducts repeated sampling on a sequence of numerical models with different resolutions. MLMC saves computational costs in the way that it conducts a large number of simulations on the computationally cheap low-resolution models and a very few simulations on the computationally expensive high-resolution models.

In a stochastic subsurface problem with >1000 parameters, the MLMC method greatly improve the computational efficiency than the standard MC in forward UQ.

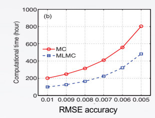


Figure 7: Performance of standard MC and MLMC in UQ of QoI of a stochastic subsurface model with >1000 parameters.

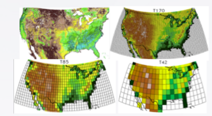


Figure 8: To simulate the reality, we have a sequence of E3SM with different spatial resolutions. MLMC can evaluate ~10 T170 models, ~100 T85 models, and ~1000 T42 models to get accurate prediction uncertainty.

Future plans and partnerships:

Since the release of E3SM version 1, additional progress has been made towards efficient ensemble simulations using the single column atmosphere model (SCM). We are currently testing this capability and investigating the resources necessary to scale from 10s to 1000s of simulations for UQ efforts.

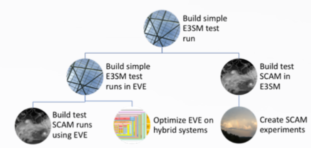


Figure 9: Workflow for the development of a E3SM single column model within the Extended V&V for ESM kit (EVE). The end result will be a script to enable users to run SCM ensembles with minimal interaction, benefitting multiple projects. We are also exploring how to best take advantage of new computational architectures.

We are currently collaborating with the FASTMath and RAPIDS institutes to accelerate algorithmic development for our project. With FASTMath, we are exploring improved methodologies for surrogate modeling, GSA and calibration. With RAPIDS, we are currently exploring improving visualization capabilities and artificial intelligence (AI)-based approaches.

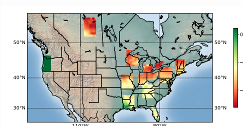
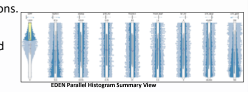


Figure 10: Using results from simple ELM, we are engaging in a partnership with FASTMath Institute to explore low-rank functional tensor train representations to discover model structure in the combined spatial-stochastic spaces. We are using sensitivity analysis and model error embedding to calibrate model components that are responsible the most for discrepancies between predictions and observations.

Figure 11: Visualization of model ensembles using the Exploratory Data analysis ENvironment (EDEN), developed by the RAPIDS institute. Efficient visualization of UQ results is highly beneficial to this work.



This work will inform E3SM about the key drivers of uncertainty in the land-atmosphere system. We envision that this approach can be used to reduce biases in future simulations, and to help DOE design new observation systems targeted towards reducing model uncertainty.

For additional information, contact:

Daniel Ricciuto

Climate Change Science Institute

(865) 574-7067

ricciutodm@ornl.gov

E3SM Energy Exascale Earth System Model

SciDAC Scientific Discovery through Advanced Computing

Oak Ridge National Laboratory

Sandia National Laboratories

Massachusetts Institute of Technology

U.S. DEPARTMENT OF ENERGY