National Aeronautics and Space Administration

Acceleration of satellite data integration into the JPL CARDAMOM framework using supervised machine learning

Principal Investigator: Nicholas Lahaye (398); Co-Investigators: Steffen Mauceri (398), Alexis Bloom (329)

Program: FY22 R&TD Innovative Spontaneous Concepts

Objectives:
Accelerate the Bayesian integration of observations into CARDAMOM by accelerating the iterative MCMC estimation and uncertainty quantification algorithm. Instead of emulating the highly non-linear model itself we reduce the complexity by directly predicting the likelihood that a set of ecosystem parameters agrees with observations. Given a set of existing CARDAMOM input parameters and their associated model-observation-differences (namely, the CARDAMOM cost function value), we trained multiple machine learning (ML) regressors to approximate this relationship and find the one that provides the best approximation while minimizing computation time.

Background:
The JPL CARDAMOM framework is a Bayesian model-data fusion algorithm designed to integrate satellite observations of the terrestrial carbon and water cycles, including MODIS vegetation indices, GRACE terrestrial water storage anomalies, and satellite-based solar-induced fluorescence among other observations, into a mechanistic carbon-water cycle model of land ecosystems. Extending the use of CARDAMOM for resolving higher spatial resolutions and multiple decades is a priority for making maximal use of the ever-increasing volume of satellite-based Earth observations. One of the primary technical challenges is the computationally intensive MCMC algorithm used to estimate ecosystem parameters, states and fluxes. Accelerating this algorithm is therefore key for integrating satellite observations into ecosystem models at finer scales and longer timespans.

Approach and Results:
1. Evaluating various regression models for our use case. This step was completed using a static MCMC chain generated with CARDAMOM containing 5M elements. Given that the cost function within CARDAMOM is highly nonlinear and the need to optimize for performance and compute time, we evaluated random forests (RFs) and shallow multi-layer perceptron (MLP) architectures. We varied the model training size and sampling technique, using the first N elements of the MCMC chain, or subsampling of the first N, varying the size of N from 1000 to 400000.
Table 1 shows the computation time and performance (R^2), for each model type. The RF proved to be comparable in performance to the MLPs, but significantly faster, so we chose to continue evaluation with only the RF. The RF regression was able to decrease the current computation time of 1000 predictions by 2 orders of magnitude.

Model	Train Size	Test Size	Train Adj. R^2	Test Adj. R^2	Train Seconds	Prediction Seconds (full)	N Estimators	Max Depth	Epoch Limit	Hidden Layers	Hidden Nodes
Random Forest	1000	499000	0.99	0.95	0.3	0.67	15	30	N/A	N/A	N/A
MLP 1	1000	499000	0.94	0.9	86.05	0.2	N/A	N/A	50	2	64, 32

Table 1. Performance of each model during initial testing.

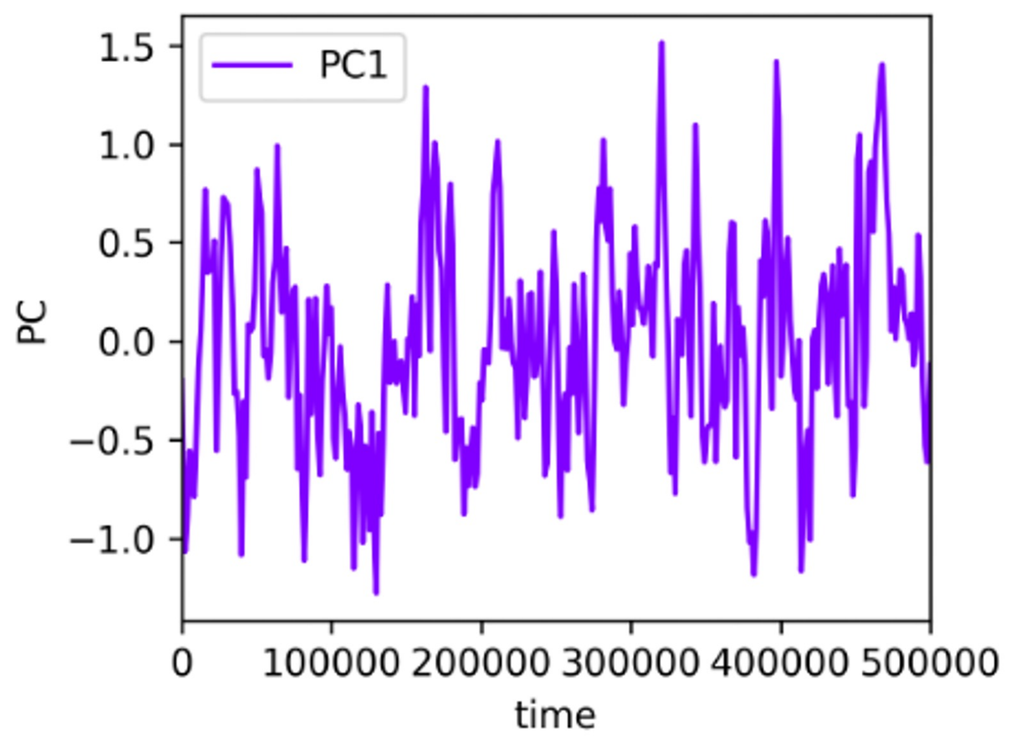

Figure 1. Visualization of first principal component of feature space explored by the CARDAMOM MCMC algorithm.

2. Integrate the RF model into the pre-existing MCMC sample generation. Given the nature of the cost function, and the way MCMC operates, the feature space greatly changes across the chain of samples (see Figure 1). Because of the moving feature space and the burn-in period, we tested multiple approaches. Figure 2 illustrates the different approaches:

I. Burn-in and full chain generated by MCMC with physical model (for baseline comparison)
II. Burn-in and first small subset of samples generated by MCMC with a physical model, and the remaining samples are generated by the emulator RF, trained on the samples generated by the physical model. This will be referred to as the ‘fully emulated’ approach.
III. Instead of using the same emulator RF to complete the whole chain, we toggle between the RF emulator and the physical model. The physical model is used intermittently to encourage parameter space exploration and provide new data to retrain the RF. This will be referred to as the ‘windowed’ approach.

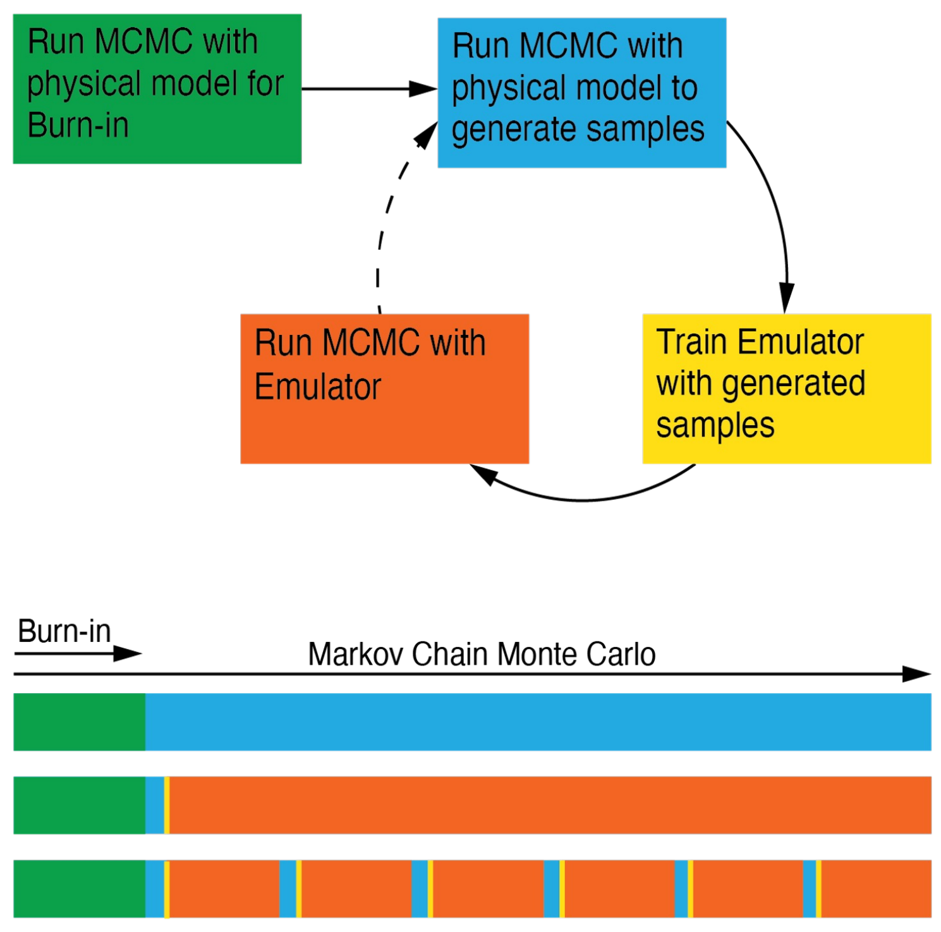
Run MCMC with physical model for Burn-in
Run MCMC with physical model to generate samples
Train Emulator with generated samples
Run MCMC with Emulator
Burn-in
Markov Chain Monte Carlo
a)
b)
c)

Figure 2. A depiction of the three approaches tested when using both the physical model and emulator ML model for sample generation within MCMC. (a) is the traditional method of using the physical model for all sample generation. (b), the ‘fully emulated’ approach, uses the physical model only for initial burn-in and training sample generation, followed by the emulator ML model, used to generate the rest of the chain. (c), the ‘windowed’ approach, begins the same as (b), but toggles between the emulator and the physical model.

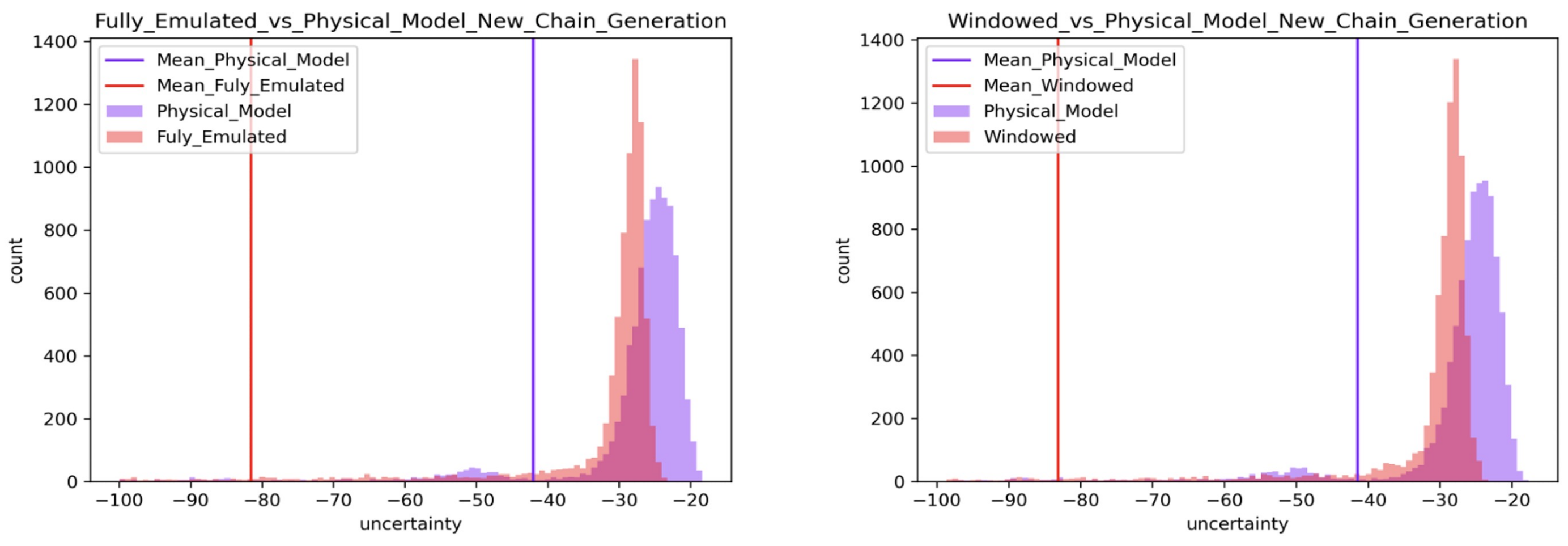

Figure 3. Plots of distribution comparisons for samples generated via the traditional approach vs. the fully emulated approach [left], and the traditional approach vs. the windowed approach [right], generated separately in distinct MCMC chains. Both newly generated distributions majoritively map to a similar space as that of the distribution generated by the physical model.

Our first test of these approaches involved using the same initial dataset, and comparing the distributions generated. We then integrated both options into the CARDAMOM MCMC software, and generated new chains with each approach. Figure 3 shows the distribution comparisons done after running this test. Both approaches generate data whose distributions are similar to the initial one with the majority of the values for both appear within a similar range as that of the newly generated physical-model-only chain.

Significance/Benefits to JPL and NASA: This pilot effort explored the feasibility to speed up MCMC by orders of magnitude using a machine learning surrogate model. We demonstrated the overall feasibility of the approach but more research is necessary to quantify the benefits of our approach. Through future funding, we aim to further develop our approach that could then be applied to many NASA / JPL challenges. For example, it would allow for the speed up of carbon-water cycle models. This in turn would allow such models to be run at higher spatial and temporal resolution and revolutionize our understanding of the carbon-water cycle. Additionally, similar approaches could be used to speed up MCMC for other retrievals and atmospheric inversion.

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

www.nasa.gov

Clearance Number: CL#
Poster Number: RPC#

Publications:

Nick LaHaye, Steffen Mauceri, Anthony Bloom, “Using Machine Learning to Accelerate MCMC for the CARbon DAta-MOdel fraMework,” Abstract submitted to AGU Fall Meeting, Chicago, Il, 2022.

PI/Task Mgr. Contact Information:

Email: Nicholas.J.Lahaye@jpl.nasa.gov