Bayesian calibration of the Community Land Model using a multi-chain Markov chain Monte Carlo method

Laura Swiler*, J. Ray†, M. Huang‡, Z. Hou‡
*Sandia National Laboratories, Albuquerque, NM; †Sandia National Laboratories, Livermore, CA; and ‡Pacific Northwest National Laboratory, Richland, WA

OBJECTIVE
Perform Bayesian calibration of three hydrological parameters in the Community Land Model (CLM) using observed latent heat fluxes (LH) from the US-ARM site.

BACKGROUND
- Community Land Model (CLM) is the land component of the Community Earth System Model. It simulates bio-geophysical processes such as energy and water fluxes from canopy and soil; heat transfer in soil and snow; hydrology of soil, canopy, and snow; and stomatal and photosynthesis.
- Sensitivity analysis of LH simulated by CLM to 10 parameters was performed: the top three parameters contributing to the variance of LH are \( F_{\text{drai}} \) (subsurface runoff decaying factor), \( Q_{\text{sm}} \) (maximum subsurface drainage), and \( b \) (Clapp-Hornberger exponent in soil water retention curve).
- MCMC requires O(10^5) evaluations of CLM version 4. The expense of running CLM becomes prohibitive for these calculations, especially when one needs to run CLM for multiple years to compare latent heat fluxes.
- Surrogates have been explored, but the accuracy of the surrogates could be questionable depending on the variables and/or conditions.
- Our solution is to examine parallel MCMC chains. This will allow exploration of the parameter space using multiple communicating chains so that surrogates could be avoided.

Bayesian Inverse Problems
- Let \( y(t) \) be the observed monthly latent heat at month \( t \); let \( M(t; \Theta) \) be the CLM model prediction for the same month.
- Observation model: \( y(t) = M(t; \Theta) + e + N(0, \sigma^2) \), where \( e \) is a model-data mismatch model and \( \sigma^2 \) is a model-data mismatch parameter.
- Calibration problems involves developing an expression for the posterior probability distribution \( P(\theta|y(t)) \)

MPI Implementation
- SACHES is implemented using one-sided MPI 2 communication.
- Done to ensure scalability, given that chains are loosely coupled.
- Also, for resilience – MPI 2 allows for re-starting of dead processes.
- Useful if chains have to be restarted.
- DREAM requires chain-to-chain communications between arbitrary, mutually disjoint chain pairs during differential evolution.
- Implemented using lock synchronization (MPI_Win_SyncLockUnlock).
- DREAM requires all-to-all communications between processors to construct proposal covariances.
- DREAM is the preferred implementation.
- Alternative implementation with active target synchronization also exists.
- Actual data transfers done using MPI_Send.

SaCHES
Scalable Adaptive Chain-Ensemble Sampling
- DREAM is a hybrid method that incorporates:
  - DREAM to utilize multiple chains to obtain high-quality proposal densities; useful in the early part of the sampling epoch.
  - DRAM to obtain posterior distributions efficiently; useful only after a few samples have been collected by DREAM.
- Parallel chains to accelerate computations.

DREAM: DiffeRential Evolution Adaptive Metropolis
- DREAM [1] uses information from multiple chains to construct a new chain position that is a weighted value of previous chain positions.
- DREAM chains communicate in pairs every MCMC step but these communication are unconnected to the communication of other pairs. There is no need for synchronization.
- DREAM performs best if there are a few informative samples to begin with. This is difficult to do in practice. DREAM can perform the task of collecting the initial informative samples.

Test Problem Results
- Test #1: Sampling a 2D Rosenbrock distribution.
  - Analytical expression of a very twisted probability distribution; sampling is a challenge.
  - Figure (top left) shows that our chains are exploring the correct distribution.
  - Figure (bottom right) shows that our chains (different colors) are sampling uniformly i.e., not stuck in one part of the distribution.
  - This is also shown in the marginal PDFs.

CLM Inversion Results
- We modeled four years of monthly latent heat (LH) data, from the US-ARM flux tower site. The four years of data were climatologically averaged, to obtain 12 months of average LH flux. These 12 monthly averaged observations were then compared to the same four years of CLM predictions which were also climatologically averaged.

CLM Posterior Densities
- Range of the parameters:
  - \( 8.1 \leq F_{\text{drai}} \leq 8.5 \); default: 2.5
  - \( 4 \leq \log(Q_{\text{sm}}) \leq 2 \); default: \( \log(5.5) = 1.096 \)
  - \( 1 \leq b \leq 15 \); default: \( 3.979 \)
- The parallel 8-chain MCMC is still in progress, but early chain results show that the parameter \( b \) appears to be converging. \( F_{\text{drai}} \) and \( Q_{\text{sm}} \) are not yet.

CLM Posterior Traces from 3 Chains
- The LH predictions based on posterior (not yet final) are substantially better than those based on default CLM parameters.
- Dashed line = true value.

Posterior Densities
- We have a version of SaCHES working to estimate CLM parameters.
- The estimated parameters are more predictive than the default values.
- The \( b \) parameter appears to have converged for our CLM problem, still need more runs to determine convergence for \( F_{\text{drai}} \) and \( Q_{\text{sm}} \).
- Parameter intuitions using CLM have been compared to those obtained using surrogates. They agree for \( b \).
- Next step: Investigate efficiency of parallel methods with coordinating chains vs. multiple independent chains.

Summary

For additional information, please contact:
J. Ray, Sandia National Laboratories, jray@sandia.gov

SAND2015-10732C