MCMC-Bayesian Calibration of the Community Land Model for the US-ARM site

Z. Hou[1], J. Ray[2] and M. Huang[1],

[1]Pacific Northwest National Laboratory, Richland, WA and [2]Sandia National Laboratories, Livermore, CA,

OBJECTIVE

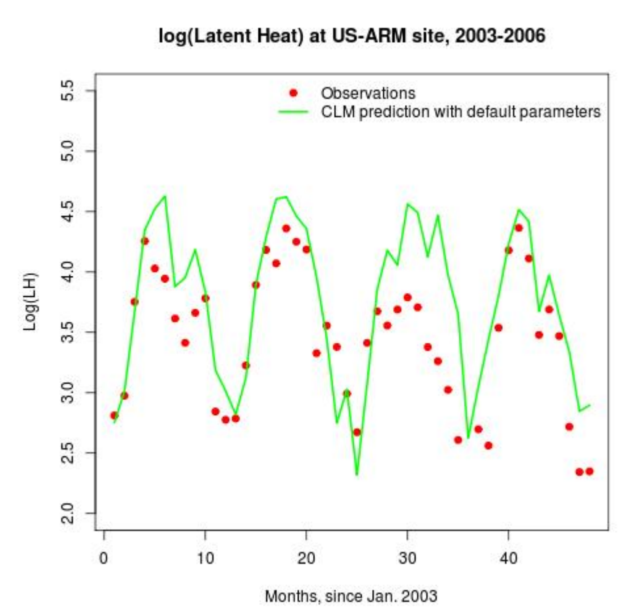
Perform a Bayesian calibration of three hydrological parameters in the Community Land Model (CLM)

- Use monthly-averaged observations of latent heat fluxes (LH) at the US-ARM site (Oklahoma) collected during 2003-2006 (48 months, total)
- Compute the uncertainty in the parameter estimates
- Model and compute the structural error in CLM; investigate sensitivity of the calibration to the choice of structural error model
- Investigate sensitivity of calibration to climatological averaging

TECHNICAL APPROACH

Inverse problem

- CLM comes with parameters set at certain default values; often not very predictive
- We seek to calibrate Θ = {F_{drai}, log(Q_{dm}), B}
  - F_{drai}: Decay factor for subsurface runoff with depth
  - Qdm: max subsurface drainage
  - B: Clapp and Hornberger exponent in the soil water retention curve



LH observation and CLM predictions with parameters set at default values

- Let y^{obs} be log(LH) observations and M(Θ) be the model
- $y^{obs} = M(\Theta) + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \Gamma)$ is a model-data mismatch modeled as a multivariate Gaussian because of time-correlated errors
- Our prior beliefs regarding parameters
  - $F_{drai} \sim \mathcal{U}(0.1, 5.0)$; default: 2.5
  - log(Q_{dm}) ~ $\mathcal{U}$(log(10^{-6}), log(10^{-2})); default : log(5.5 x 10^{-3})
  - B ~ $\mathcal{U}$(0.1, 15)

Bayesian formulation of the inverse problem

- Involves using Bayes' theorem to derive an expression for the probability density of Θ, conditioned on y^{obs}
- Using $\Pi(\Theta)$ to denote the prior distribution on Θ

$$P(\Theta \mid y^{obs}) \propto P(y^{obs} \mid \Theta)\Pi(\Theta)$$
$$= \left|\Gamma\right|^{-1/2} \exp\left[-\left(y^{obs} - M(\Theta)\right)^T \Gamma^{-1}\left(y^{obs} - M(\Theta)\right)\right]\Pi(\Theta)$$

- This is a 3-parameter estimation; higher dimension if we estimate Γ
- Will use an adaptive Markov chain Monte Carlo (MCMC) method [1] to compute estimates as a multidimensional posterior distribution.
- Will require O(10^5) evaluations of CLM
  - We need to make a surrogate of CLM if we desire a converged posterior distribution

CONSTRUCTING CLM SURROGATES

Surrogate models

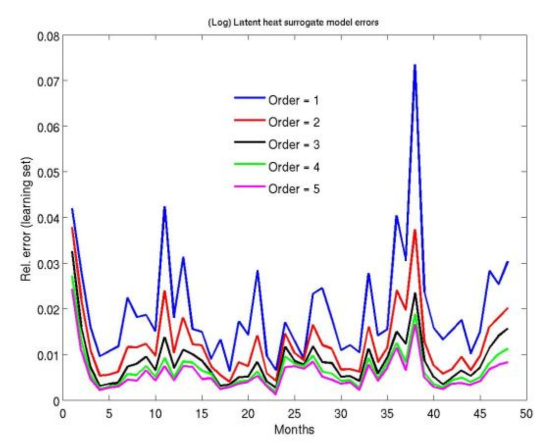
- Surrogate model – an inexpensive "curve fit" which approximates the input-output mapping by CLM of LH and Θ
- We propose a polynomial form

$$\log(LH) = \sum_{p=0}^{M} \sum_{q=0}^{M} \sum_{i=1}^{3} \sum_{j=1}^{3} w_{i,j}^{(p,q)} \theta_i^p \theta_j^q$$

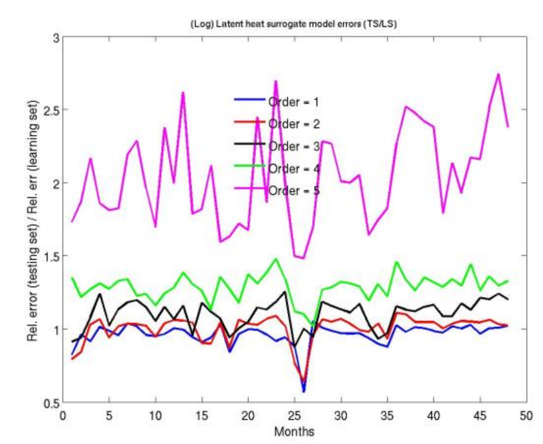
- $\Theta = \{\theta_1, \theta_2, \theta_3\}$ = {F_{drai}, log(Q_{dm}), B}; (p+q) < M
- Tested M = 1 … 5 to explore linear to 5^{th}-order models
- $w^{(p,q)}_{i,j}$ are estimated by linear regression to a training set of CLM runs

Selecting and configuring surrogates

- Sampled 256 points in the Θ-space (via space-filling quasi-Monte Carlo sampling) and generated 48-month time-series of log(LH) predictions
- Constructed polynomial models for each month
- Segregated 85% of the runs into a learning set (LS); estimated $w^{(p,q)}_{i,j}$ using linear regression
  - Used AIC to remove superfluous polynomial terms and prevent overfitting
- Used the remaining 15% of the runs as the testing set (TS); computed the RMS prediction error for the fitted polynomial model
- Repeated for 100 (LS/TS) pairs
- A good model should:
  - Have similar accuracy for LS and TS – else, we have overfitting
  - Have a small error (<10%) for LS



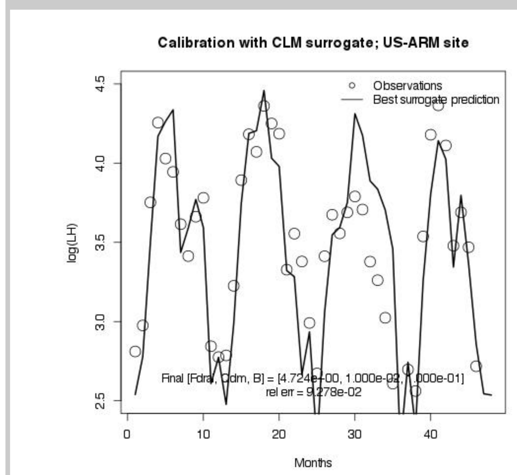
Learning set errors for surrogate models of order 1 to 5



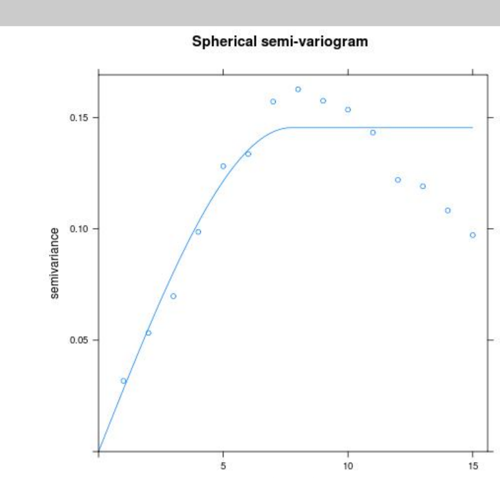
Ratio of LS / TS RMS errors for surrogate models of order 1 to 5

We chose quadratic models. Errors are < 4%

STRUCTURAL ERROR MODEL



Deterministic fit of CLM surrogates

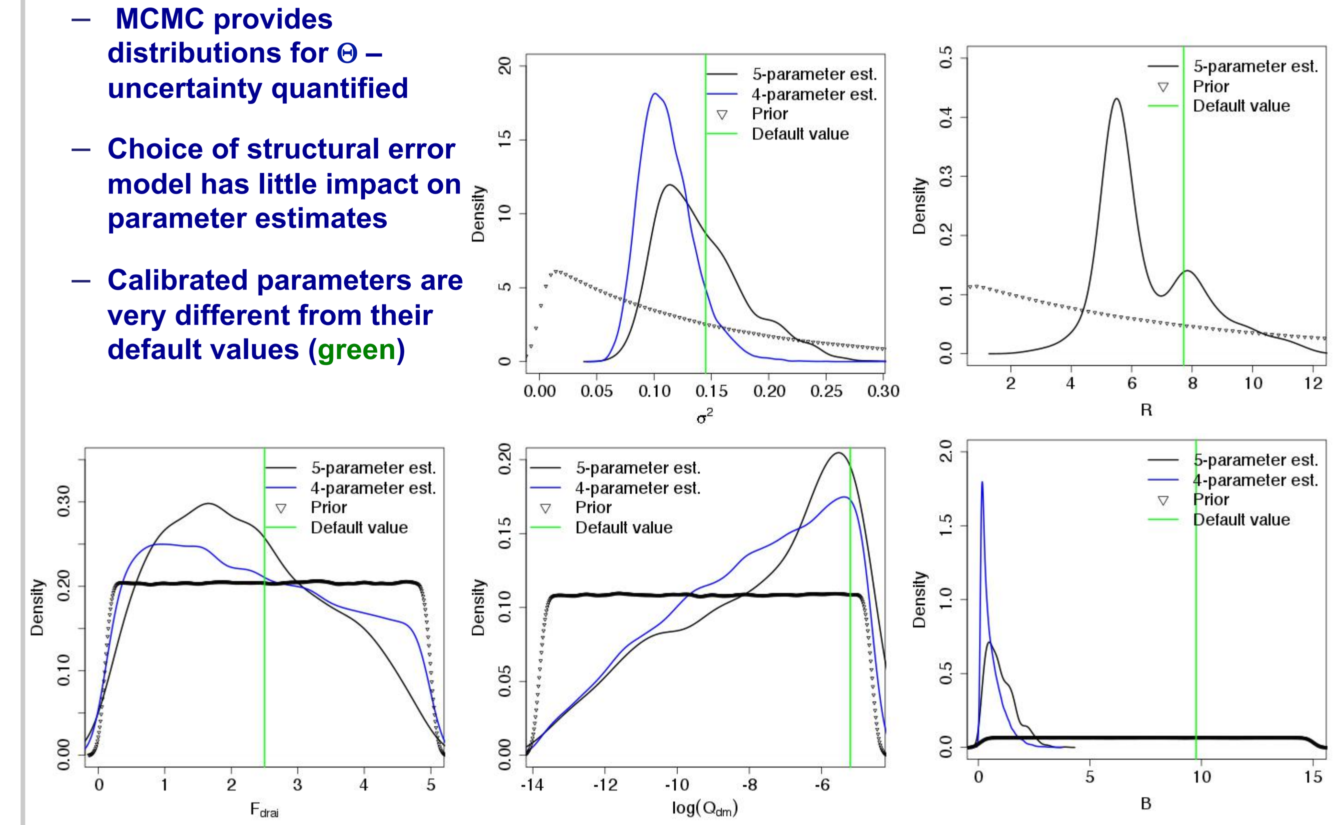


Spherical semi-variogram modeling

- A deterministic calibration of CLM surrogates revealed that the model-data mismatch may be correlated in time
- Correlation was modeled with a spherical semi-variogram; sill σ^2 = 0.145, range R = 7.72 months
  - Will henceforth use a spherical semi-variogram to model Γ

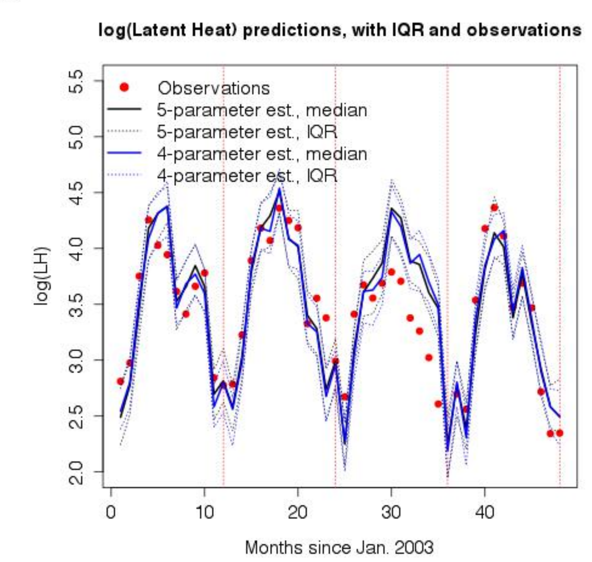
THE ESTIMATION PROBLEM

- 5-parameter estimation for {Θ, σ^2, R} [plotted in black]
  - Priors: σ^2 ~ Exp (mean = 0.145); R ~ Exp (mean = 7.72)
- A competing model: Assume data-model mismatch is NOT correlated in time (errors are i.i.d. Gaussians); estimate a 4-dimension problem for {Θ, σ^2} [plotted in blue]
  - Prior: σ^2 modeled with conjugate prior (inverse Gamma distribution)
- MCMC provides distributions for Θ – uncertainty quantified
- Choice of structural error model has little impact on parameter estimates
- Calibrated parameters are very different from their default values (green)



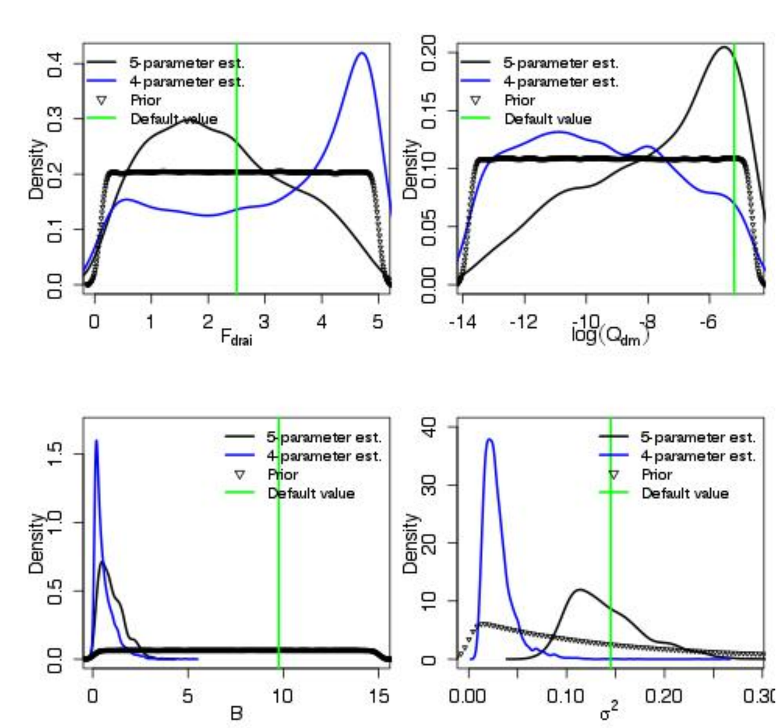
Predictive skill of calibrated model

- Both models (calibrated with and without time-correlated structural errors) are equally predictive
- And a big improvement over default values of the parameters
- The time-uncorrelated errors are preferable – much simpler



IMPACT OF CLIMATOLOGICAL AVERAGING

- Compute monthly average of LH for observations and CLM predictions
- Remake surrogates for 12 months; re-calibrate
- Figure (right) plots the PDFs for Θ
- Climatological averaging makes an enormous difference in F_{drai} and Q_{dm} calibrations
- Also B is still very different from default value
- Finally, the calibration error (σ^2) is far smaller than when calibrating to a 48-month time-series
  - We should expect this calibration to be far more predictive

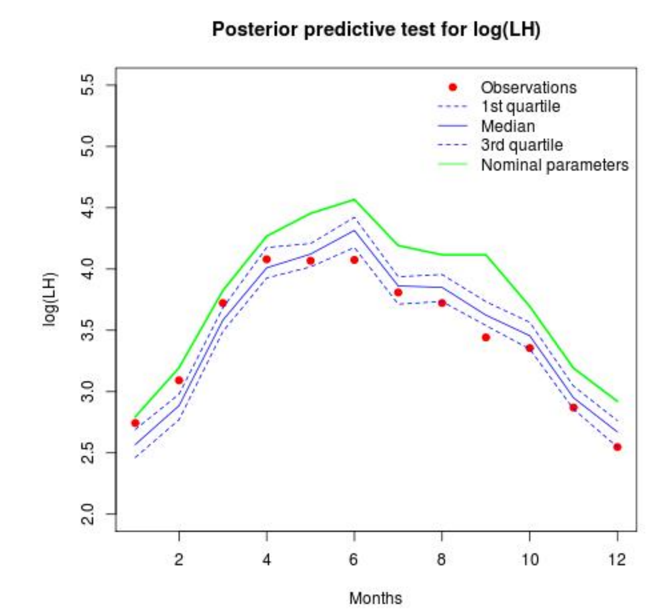


Comparison of PDFs of Q generated before (black) and after (blue) climatological averaging. The default values are in green

POSTERIOR PREDICTIVE TEST

Predictive skill of calibrated model after climatological averaging

- The calibration is far more predictive than before
- Reflected in the calibrated value of σ^2
- Not clear if this is due to a smaller σ^2 or better/sharper calibration of Θ
- But a huge improvement over the predictive skill of the uncalibrated model



CONCLUSIONS

- Surrogate models can allow efficient Bayesian calibration of CLM parameters
- The calibration is done with MCMC – it provides parameters as a joint probability density distribution
  - Automatically quantifies uncertainty in the estimates
  - Allows inference of structural error, magnitude and form
- For the US-ARM site, we find
  - 2 out of 3 hydrological parameters are very different from their default values
  - Choice of structural error model has very little impact
  - Climatological averaging has a huge impact on the calibration; it becomes more predictive as annual variations are averaged out
  - But climatological averaging still doesn't get the calibrated parameter values closer to the default ones

Acknowledgements

The project was funded by the Department of Energy Office of Science, via the Office of Advanced Scientific Computing Research (OASCR) and Biological and Environmental Research (BER) office.

References

1. H. Haario, M. Laine and A. Mira, "DRAM: Efficient adaptive MCMC", Statistical Computing, 16:339-354, 2006.

For additional information, please contact:

J. Ray, Sandia National Laboratories, jairay@sandia.gov

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

PNNL is operated by Battelle Memorial Institute for the U.S. Department of Energy under contract DE-AC05-76RLO1830.