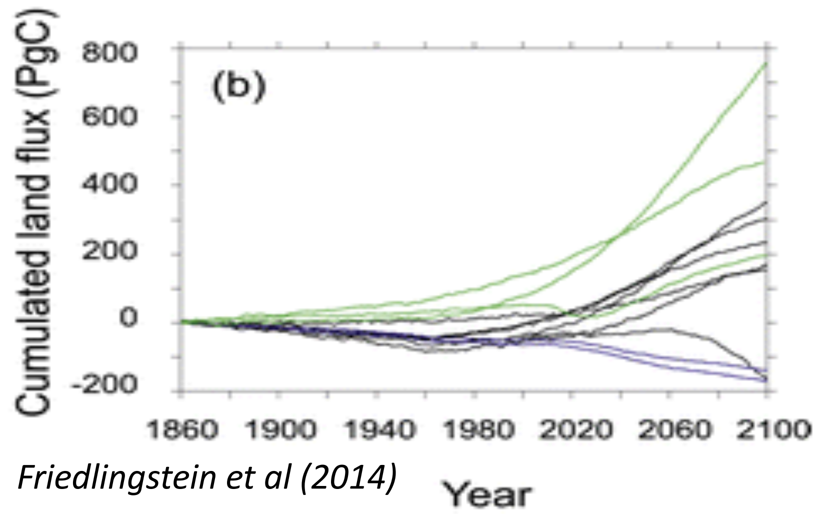
Quantifying and reducing uncertainty in the E3SM land model using surrogate modeling

Daniel M. Ricciuto (ORNL)
Khachik Sargsyan (SNL-CA)
Peter Thornton (ORNL)
Dan Lu (ORNL)

E3SM all hands presentation
May 27, 2021

Overview and motivation: LSMs

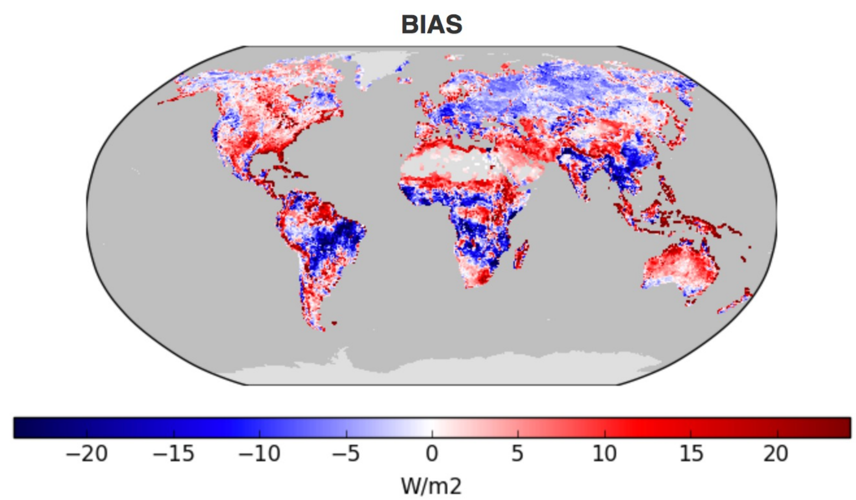
- Uncertainty from Multi-model ensembles
- Large spread in outputs
- Many quantities of interest
- Little formal uncertainty quantification (UQ)
  - Expensive model evaluation
  - High dimensionality



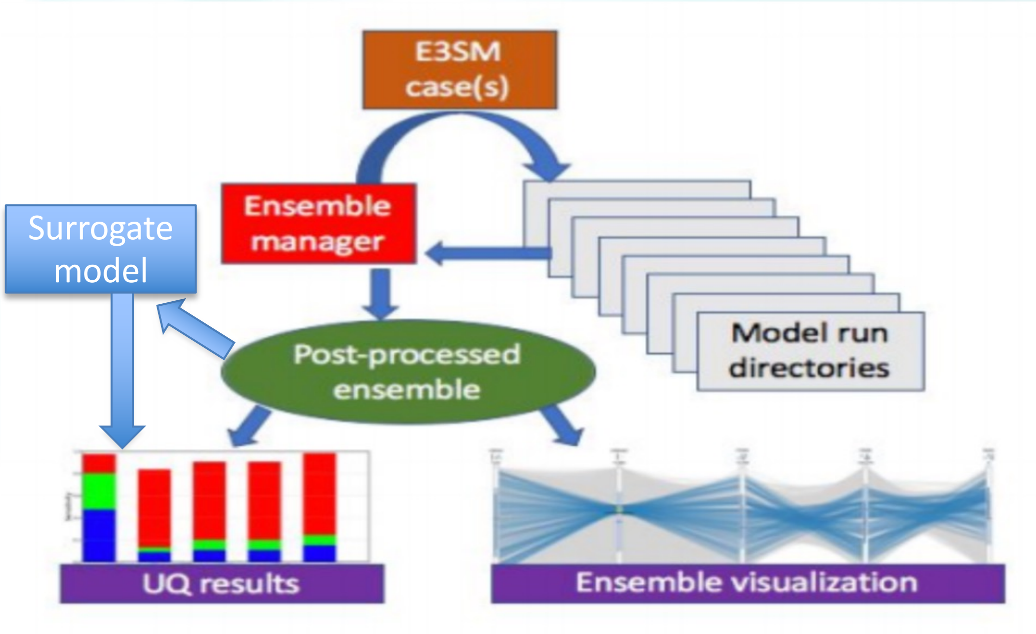
Friedlingstein et al (2014)

UQ challenges in E3SM :

- What processes drive uncertainty?
- What accounts for the key differences among models?
- Can model calibration using observations (e.g. satellite data) reduce uncertainty?



The case for large ELM ensembles



- Needed to understand parametric uncertainty
- High dimensionality (uncertain parameters)
- Can be used to construct surrogate models, which enable UQ methods
  - Sensitivity analysis
  - Parameter calibration

- Single gridcell UQ: Uses mpi-serial version of ELM with mpi4py in the offline land-model testbed (OLMT) to run up to 10k ensemble members
  - Ongoing tasks: ELM-FATES, crop, default ELM

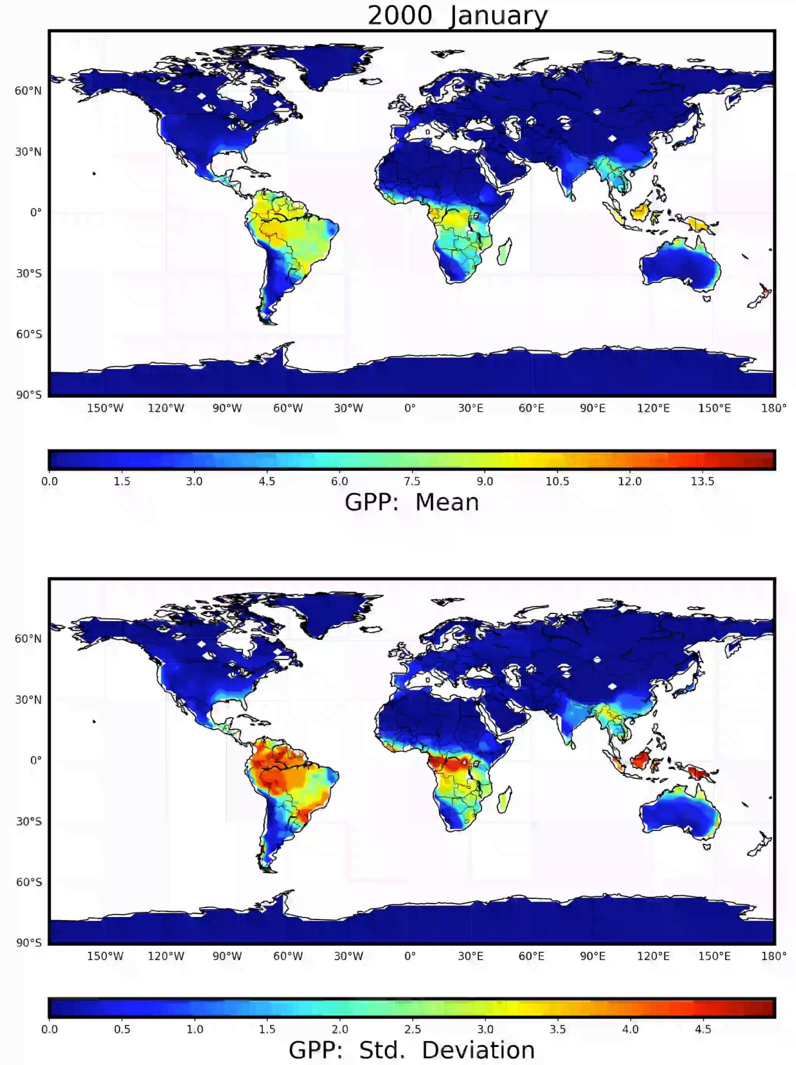
- Global UQ: Generally smaller ensembles (100-200), low resolution

A global ELM ensemble

- Focused on GPP (gross primary productivity)
  - Primary input for land C-cycle,
  - strong coupling with transpiration
- 10 parameters analyzed using 275 ensemble members (1.9x2.5 resolution), satellite phenology

Parameter	Description	Min	Max	Default range
flnr	Fraction of leaf in in RuBisCO	0	0.25	[0.042,0.176]
mbbopt	Stomatal slope (Ball-Berry)	2	13	[4,9]
bbbopt	Stomatal intercept (Ball-Berry)	1000	40000	[10000,40000]
roota_par	Rooting depth distribution	1	10	[3,10]
vcmaxha	Activation energy for Vcmax	50000	90000	72000
vcmaxse	Engropy for Vcmax	640	700	670
jmaxha	Activation energy for jmax	50000	90000	72000
dayl_scaling	Day length factor	0	2.5	2
dleaf	Characteristic leaf dimension	0.01	0.1	[0.01,0.1]
xl	Leaf/stem orientation index	-0.6	0.8	[-0.5,0.65]

GPP (gC m^{-2} day^{-1}) ensemble



E3SM Energy Exascale Earth System Model

U.S. DEPARTMENT OF ENERGY

Goal: create a ***surrogate*** model

Surrogate model is a "good-enough" approximation of the full model over a range of parameter variability.

... otherwise called

- Metamodels
- Response surfaces
- Emulators
- Low-fidelity model

Black Box

$$Y = f(\lambda; \underbrace{x, t, \ldots}_{})$$

Uncertain input parameters

Operating conditions, forcings, indexed QoIs

Surrogate models are needed for computationally intensive tasks:

- Parameter estimation
- Optimization
- Experimental/computational design
- Forward uncertainty propagation

$$f(\lambda; x, t, \ldots) \approx f_{surr}(\lambda; x, t, \ldots)$$

Curse of dimensionality hits twice!

Black Box

$$\boxed{Y = f(\lambda; \underbrace{x, t, \ldots}_{})}$$

Uncertain input parameters

Operating conditions, forcings, indexed QoIs

Challenge:

High-d input: large number of uncertain parameters

High-d output: large number of QoIs over high-res grid

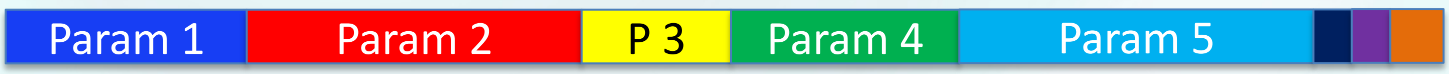
Fix:

Sensitivity analysis to select the most important parameters

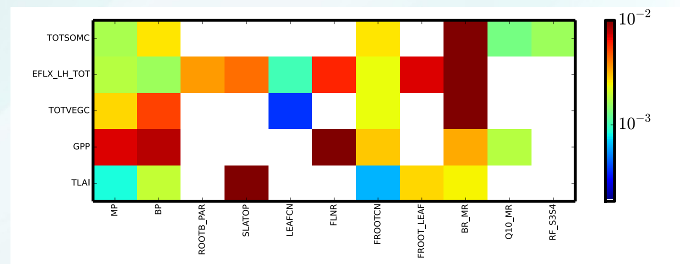
Principal component analysis to reduce dimensionality

Global Sensitivity Analysis (GSA) enables parameter selection

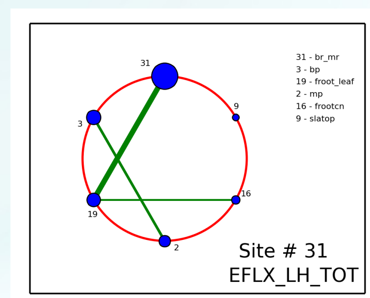
… otherwise called Sobol indices, variance-based decomposition
Attribute fractions of output variance to input parameters



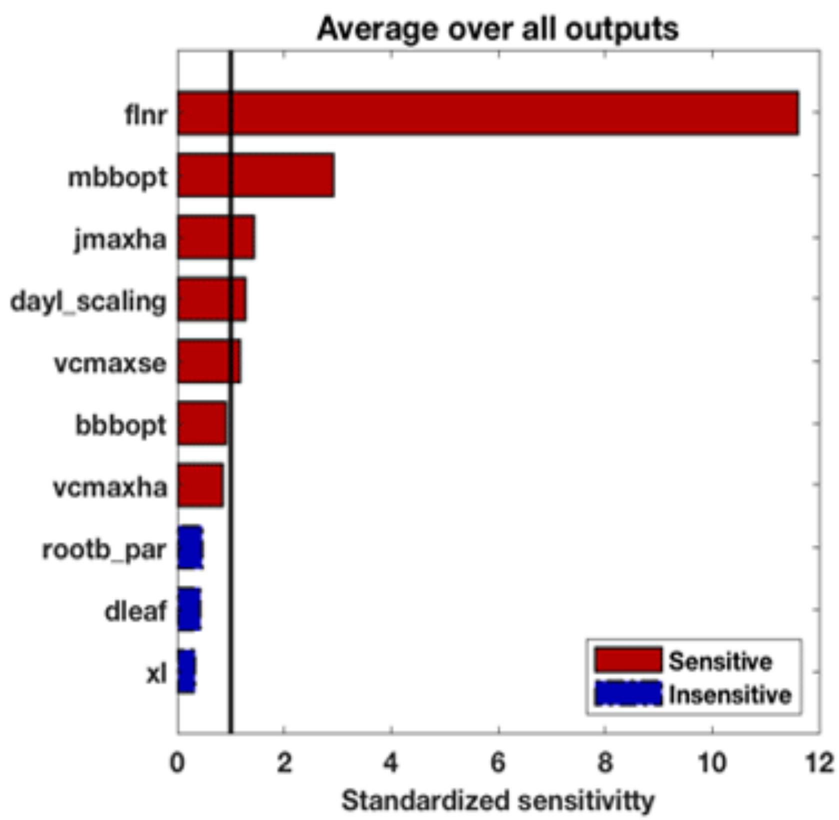
- i.e., how much output variance would reduce if a given parameter is fixed to its nominal value

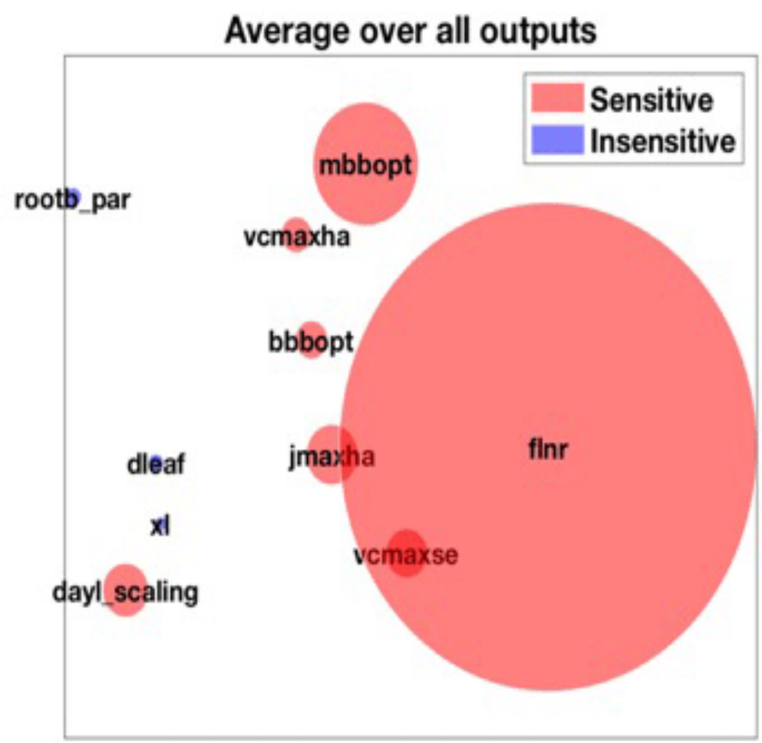


- also generalizes to joint sensitivities: joint parameter impact to a given QoI



Sensitivities of global average GPP





Spatio-temporal surrogate model via Karhunen-Loève and Polynomial Chaos

$$f(\lambda; x, t) \overset{\textbf{KL}}{\approx} \sum_{k=0}^{K} f_k(\lambda)\, \varphi_k(x,t) \overset{\textbf{PC}}{\approx} \underbrace{\sum_{k=0}^{K} \sum_{j=0}^{J} f_{kj}\, \phi_j(\lambda)\;\; \varphi_k(x,t)}_{f_{surr}(\lambda; x, t)}$$

Karhunen-Loève (KL) expansion is essentially a continuous version of principal component analysis

Polynomial chaos (PC) expansion is essentially a polynomial regression with respect to uncertain parameters

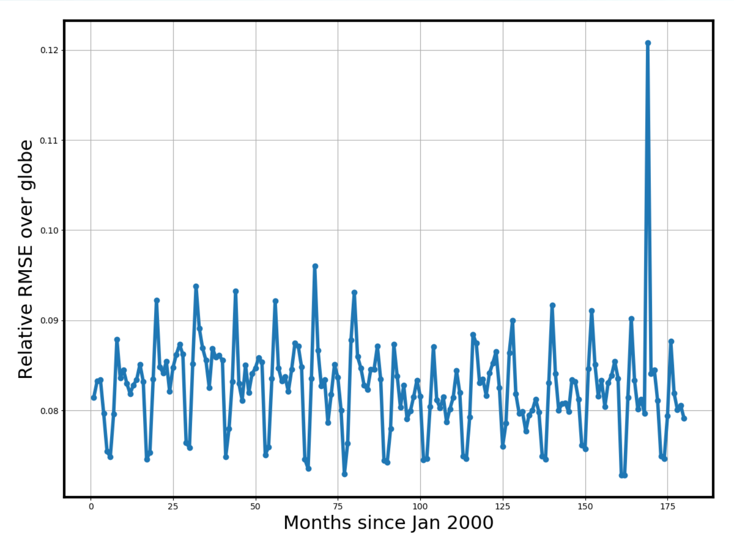
E3SM Energy Exascale Earth System Model

U.S. DEPARTMENT OF ENERGY

Spatio-temporal KL-PC surrogate is globally within 10% accuracy

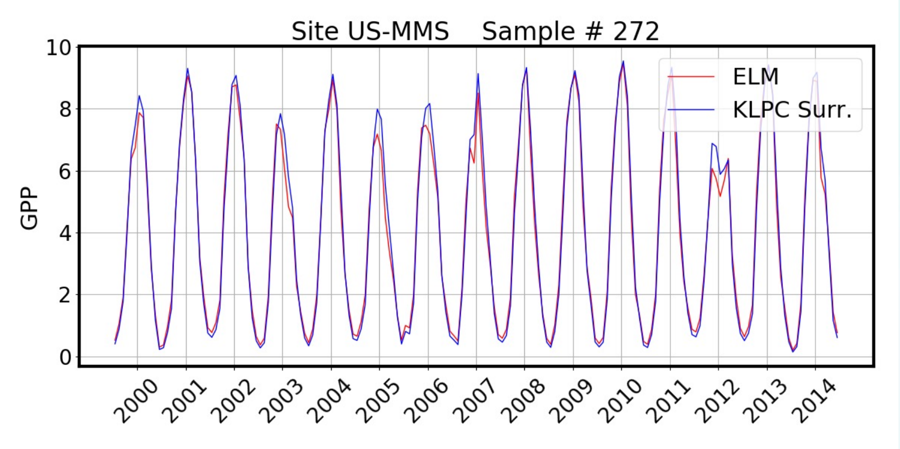
- 3183 land cells over 180 months is > 500,000 outputs
- Instead of 500K surrogates, we build about 2K surrogates, one for each eigen-component

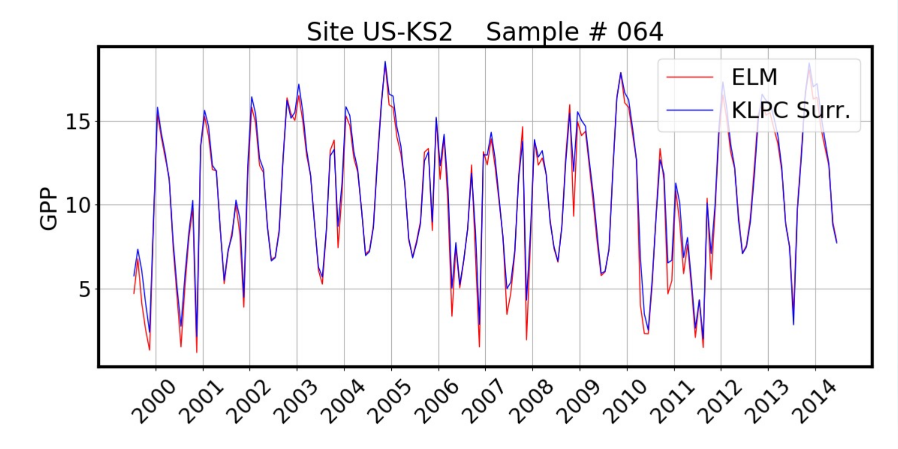
- End result: a single surrogate, resolved in space and time, with about 10% relative error compared to true ELM
- Surrogate ELM is extremely cheap to evaluate and is being used online to calibrate the parameters
- Room to improve: neural networks!!



ELM vs Surrogate:

accuracy can be improved with higher order KL or PC

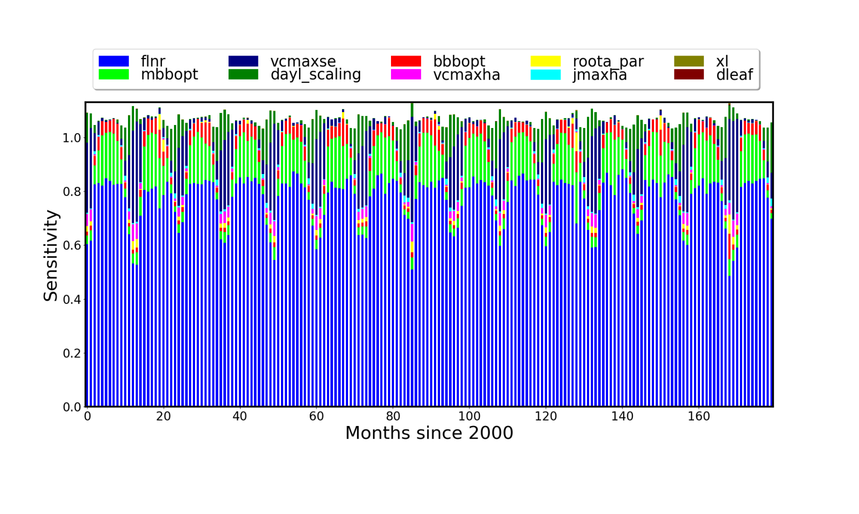




Gridcell-level sensitivities

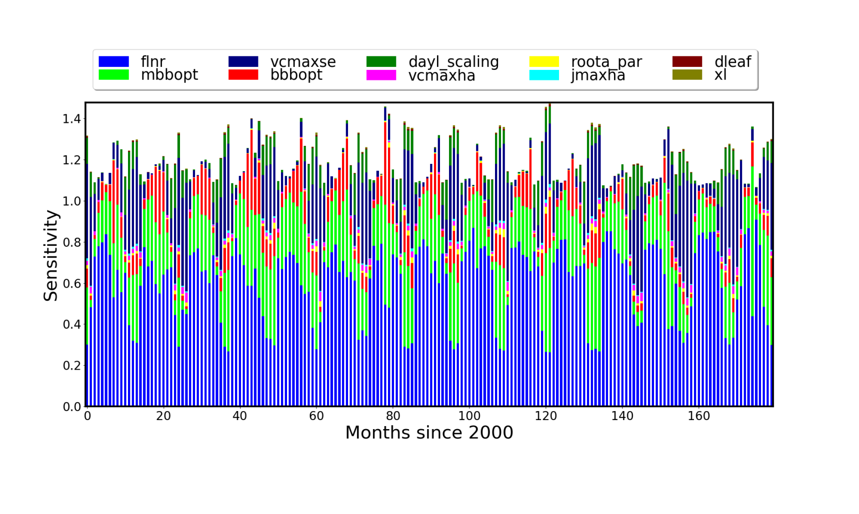
Site US-Ha1

Deciduous forest
Massachusetts, USA

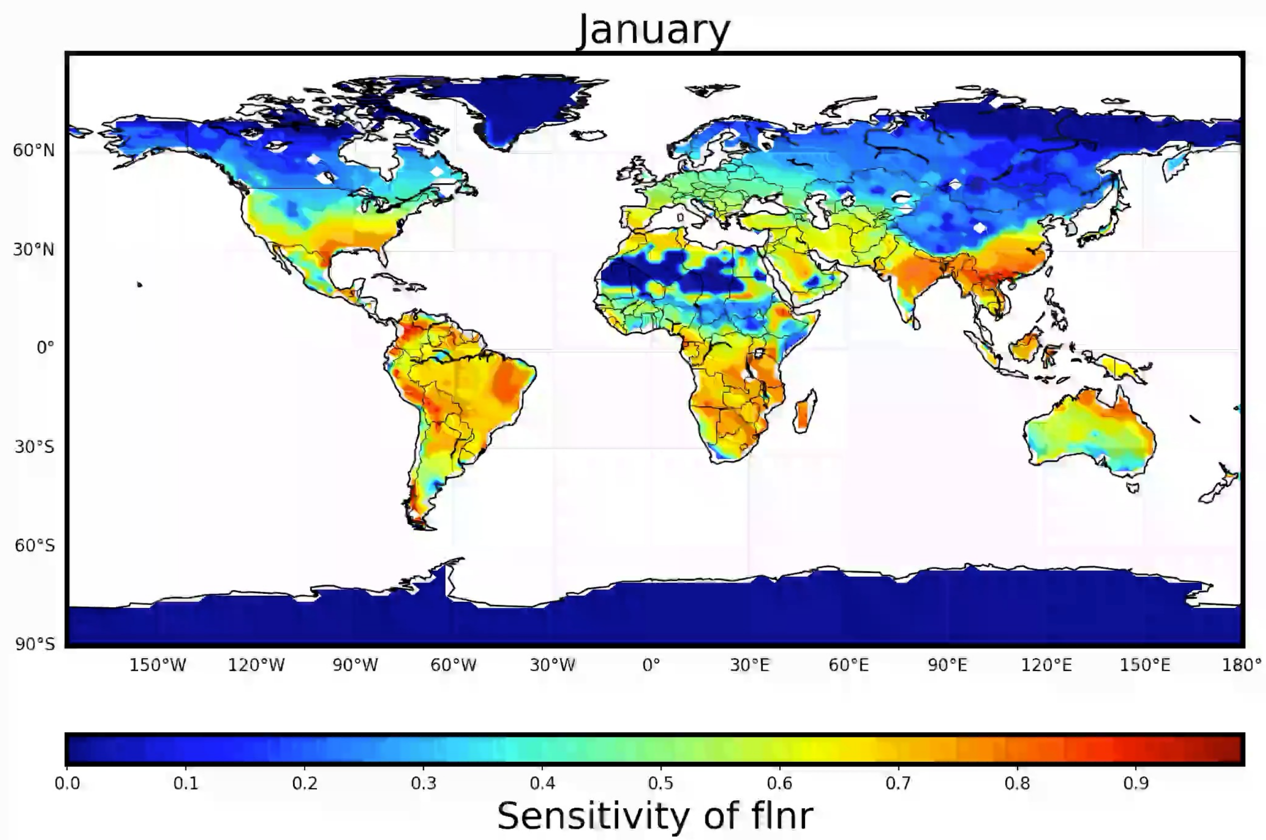


Site US-Fpe

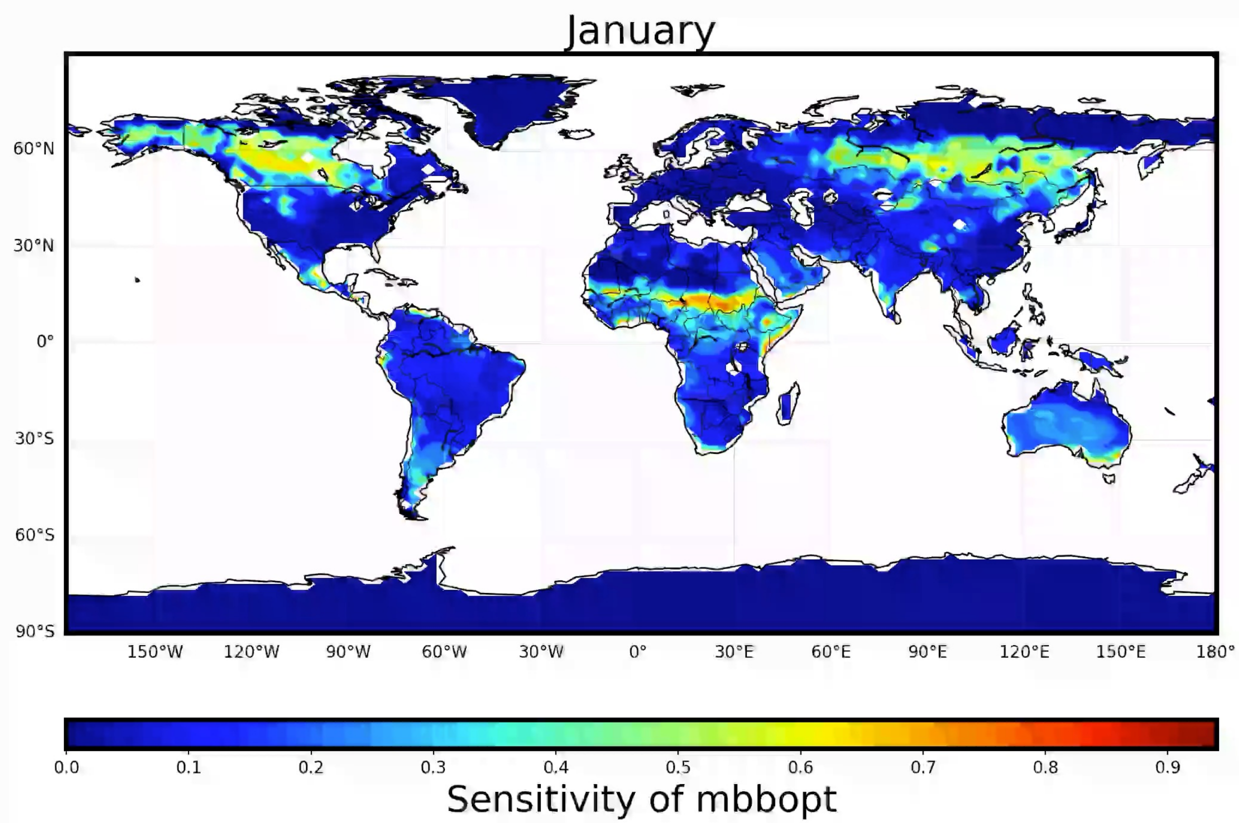
Grassland
Montana, USA



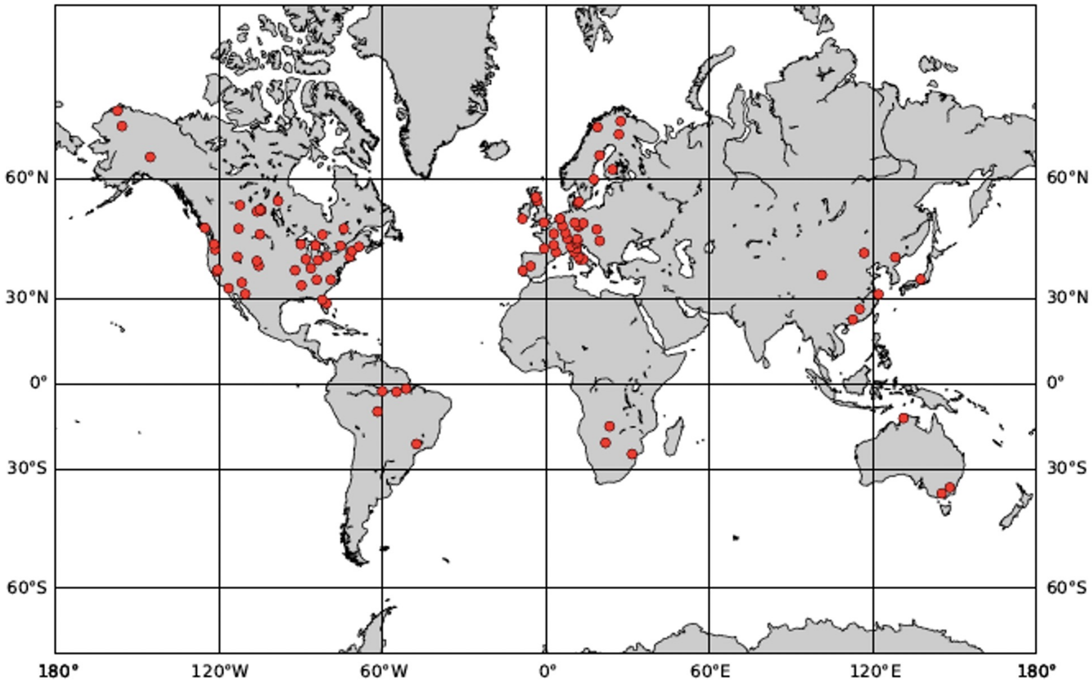
Sensitivity to flnr
(fraction of leaf N in RuBisCO)



Sensitivity to m_{BBopt} (stomatal slope)



Constraining ELM with FLUXNET

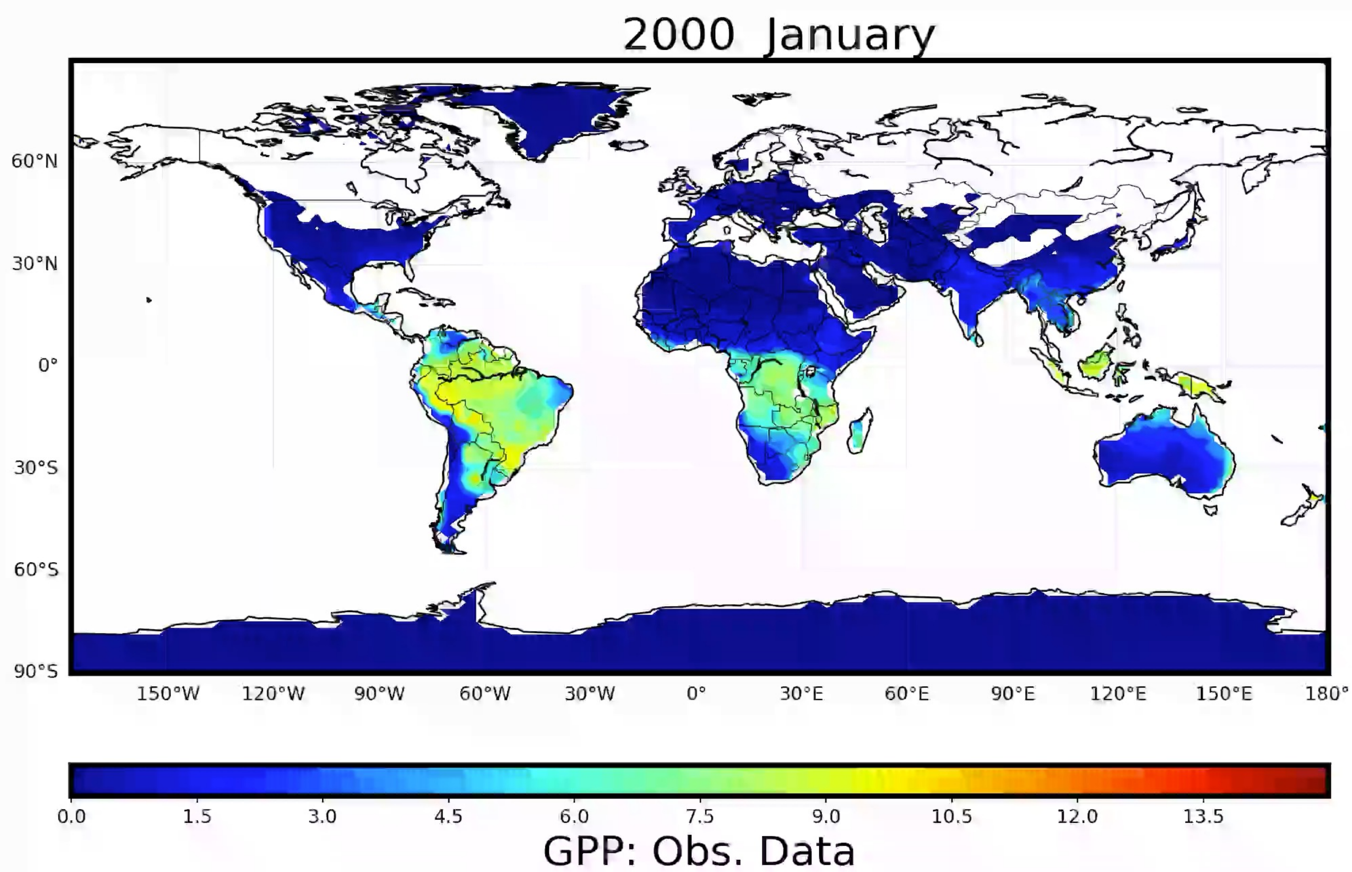


FLUXNET sites

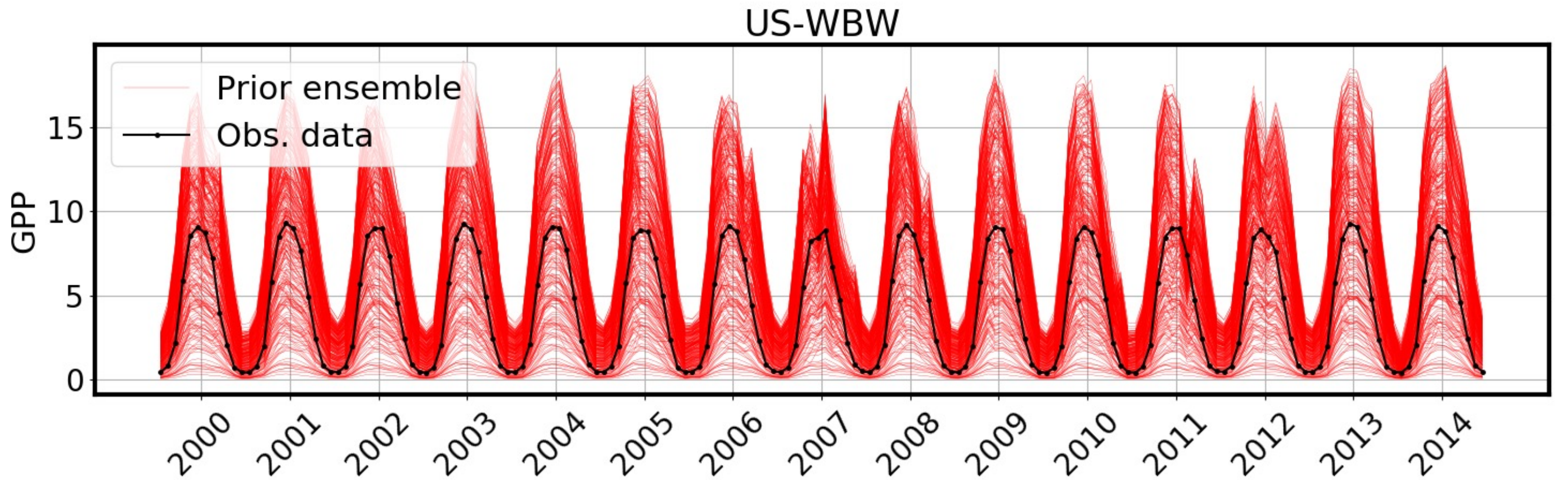
- FLUXNET towers measure CO_2, water, energy fluxes
- FLUXCOM: A gridded GPP benchmark upscaled from FLUXNET network using meteorology, remote sensing
- In this study: 96 high-quality sites selected for calibration

- This analysis focuses on calibration at multiple gridcells, but methodology can be used to calibrate gridded observations/benchmarks (e.g. FLUXCOM)

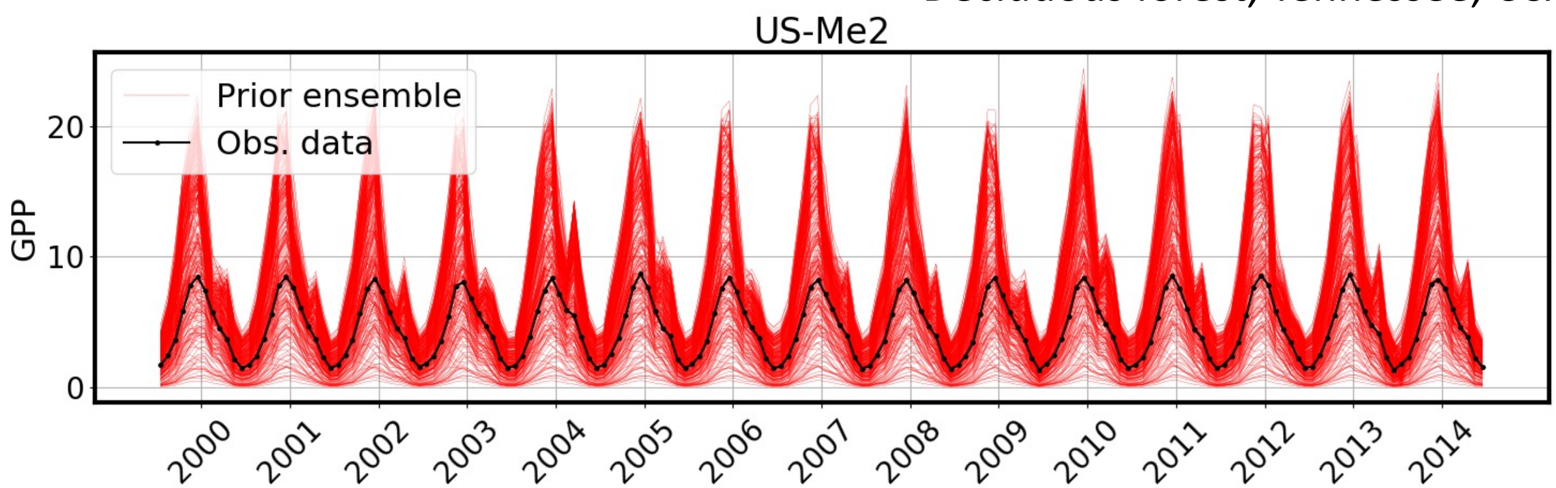
FLUXCOM data, visualized



Model ensemble with GPP data



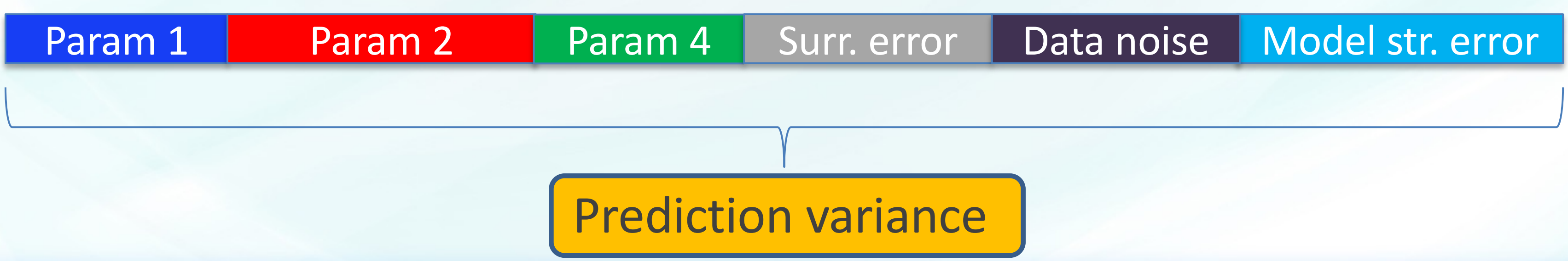
Deciduous forest, Tennessee, USA



Evergreen forest, Oregon, USA

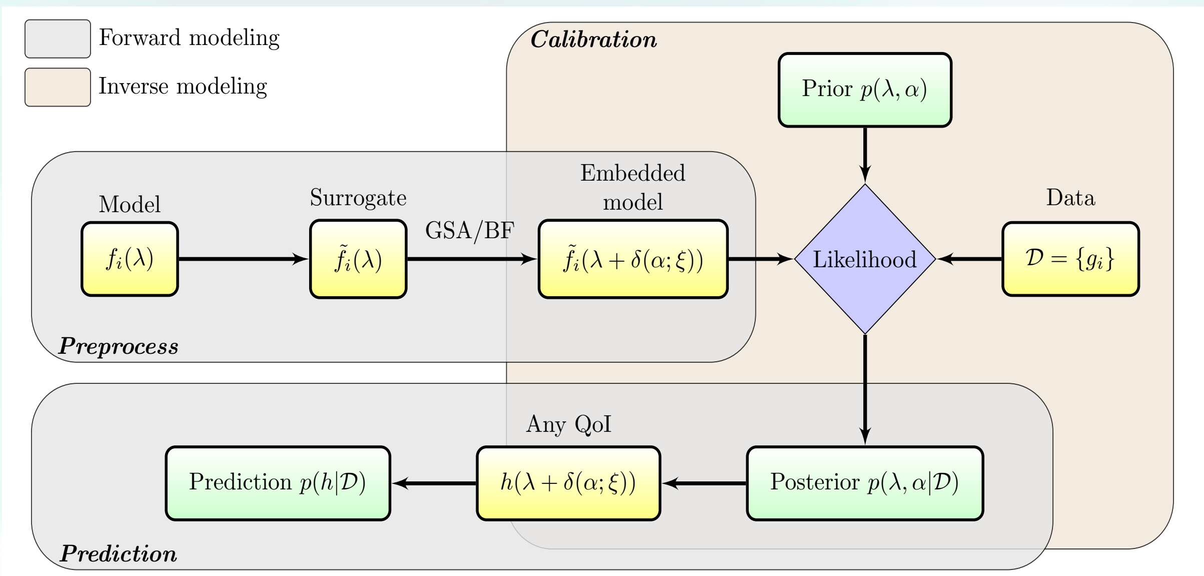
Bayesian approach is main tool for parameter calibration

- Bayesian inference allows incorporation of various sources of uncertainty
- Markov chain Monte Carlo (MCMC) for building posterior PDFs
  - Ugly high-dimensional parameter PDFs, but advanced MCMC methods are available
- Requires many online evaluations of the model
  - **This is why we needed the surrogate!**
- Predictive uncertainty decomposition augmented with surrogate error and observational noise, and *model structural error*



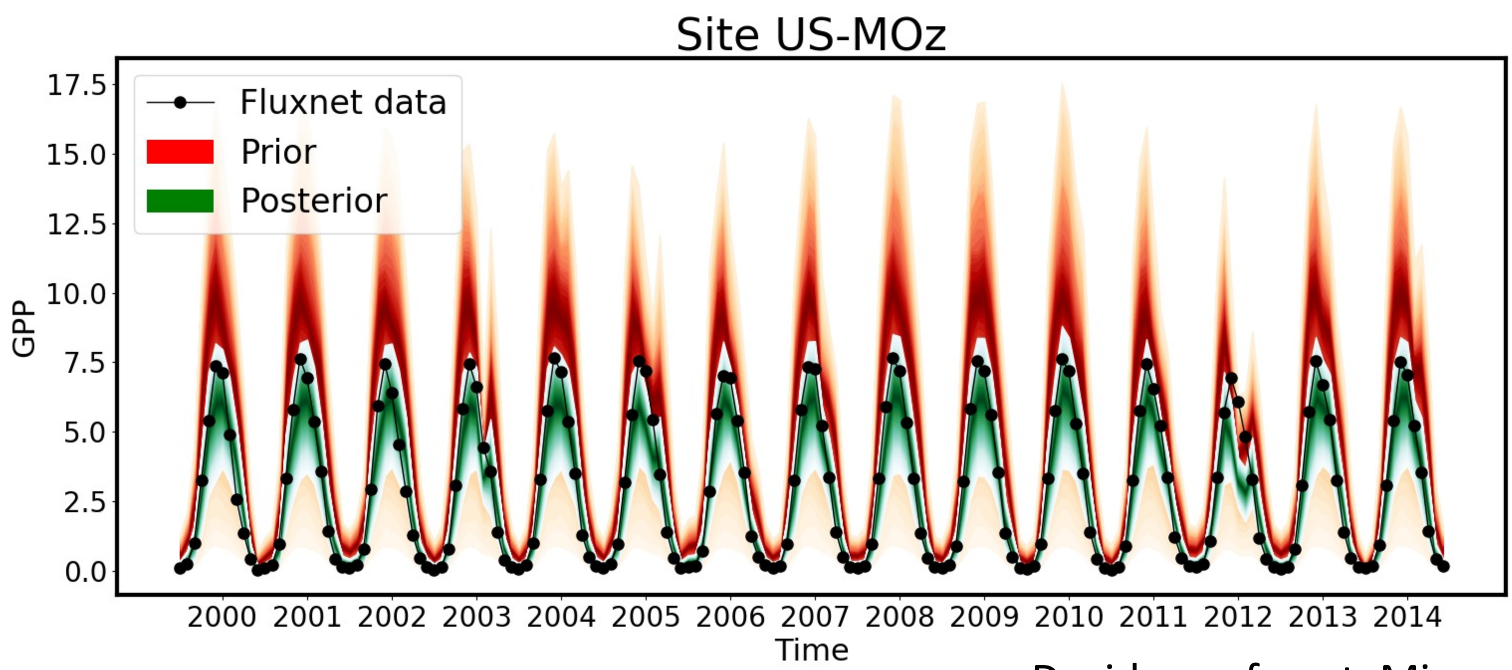
Calibration with *Embedded* Model Structural Error

- Model structural error embedding approach *[Sargsyan et. al., 2015, 2018]*
  - Embedded, but not intrusive, i.e. black-box
  - Meaningful extrapolation to full set of QoI predictions
  - Disambiguation between model error and data noise
  - Removes parameter biases and overfitting



Prior vs posterior predictions…

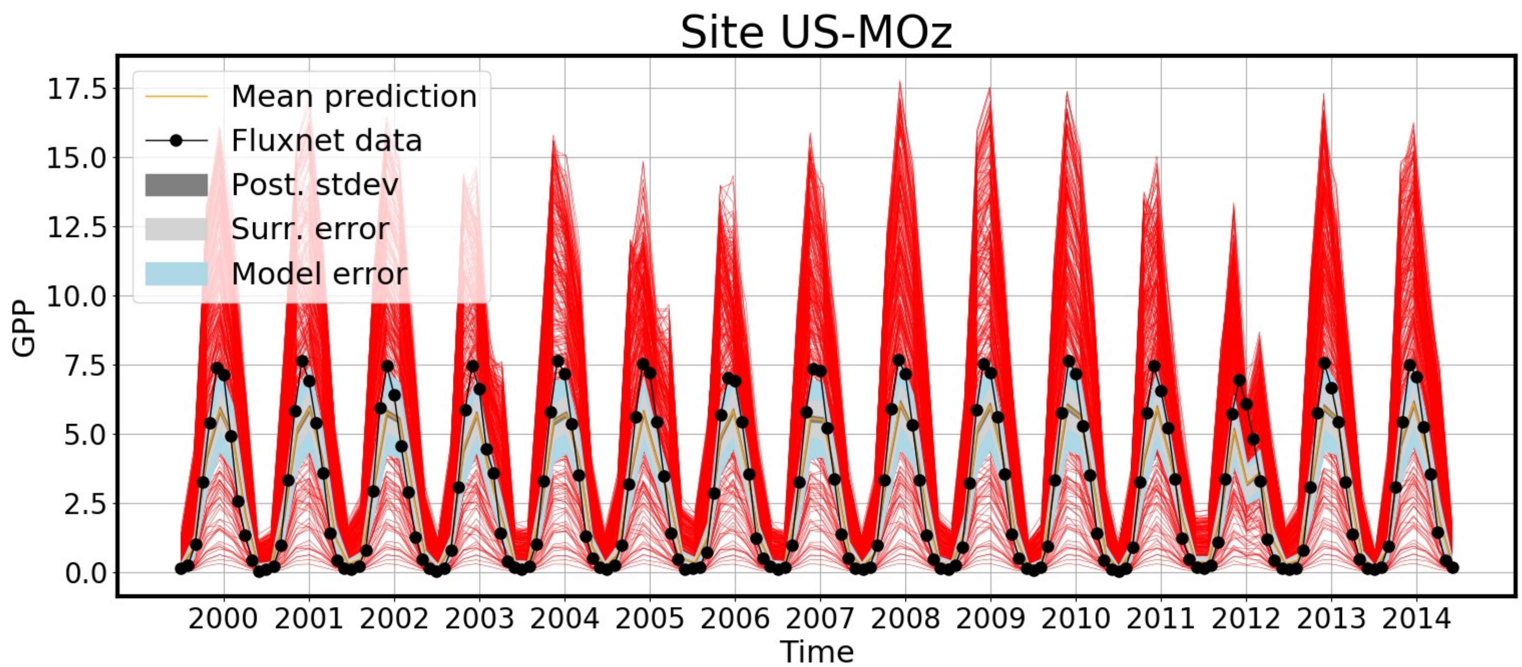
Uncertainty reduction: zoom in the parameter space regions relevant to obs. data



Deciduous forest, Missouri, USA

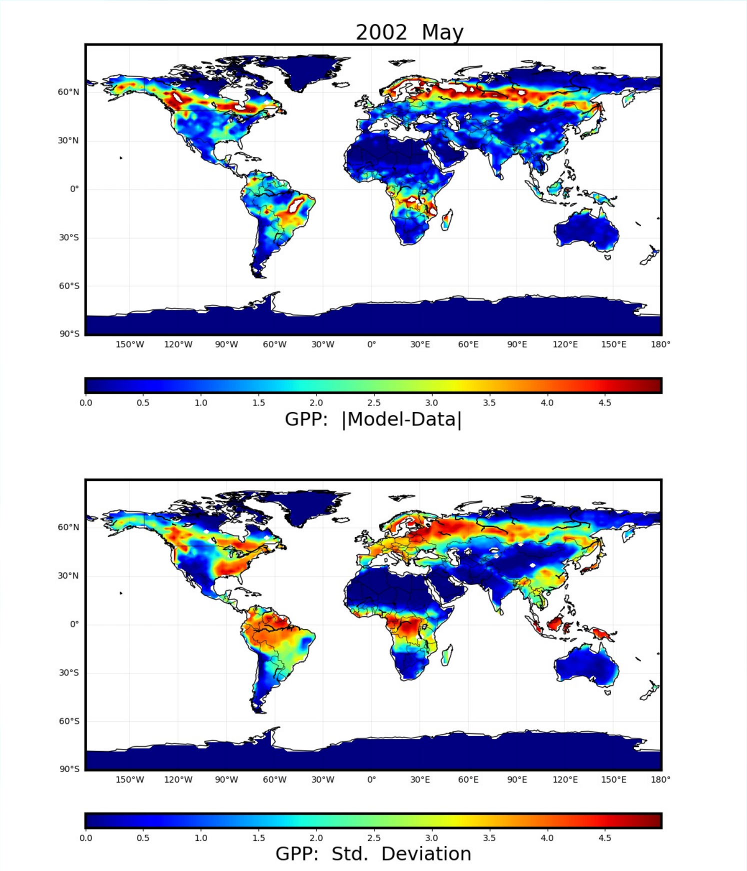
… with uncertainty decomposition

Model structural error is usually the largest contributor of predictive variance

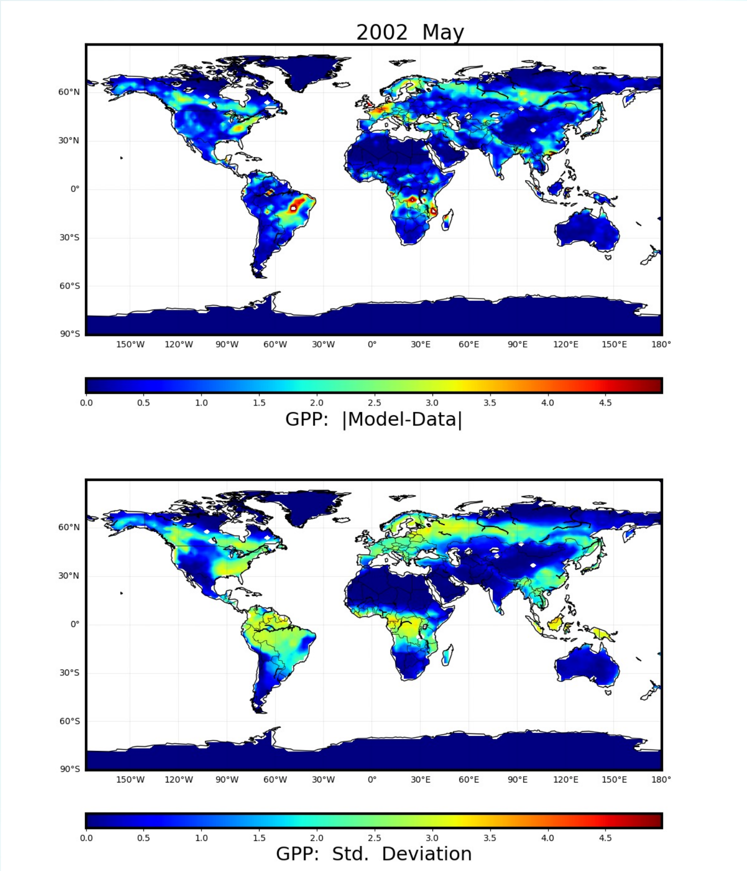


Calibration reduces predictive mean error and predictive standard deviation

Prior

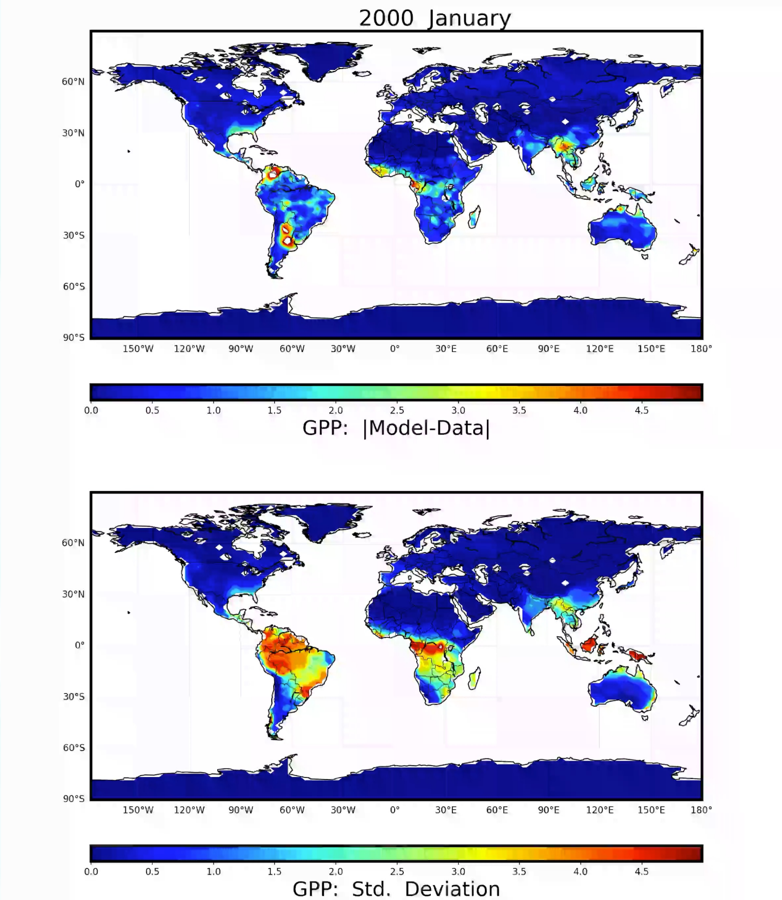


Posterior

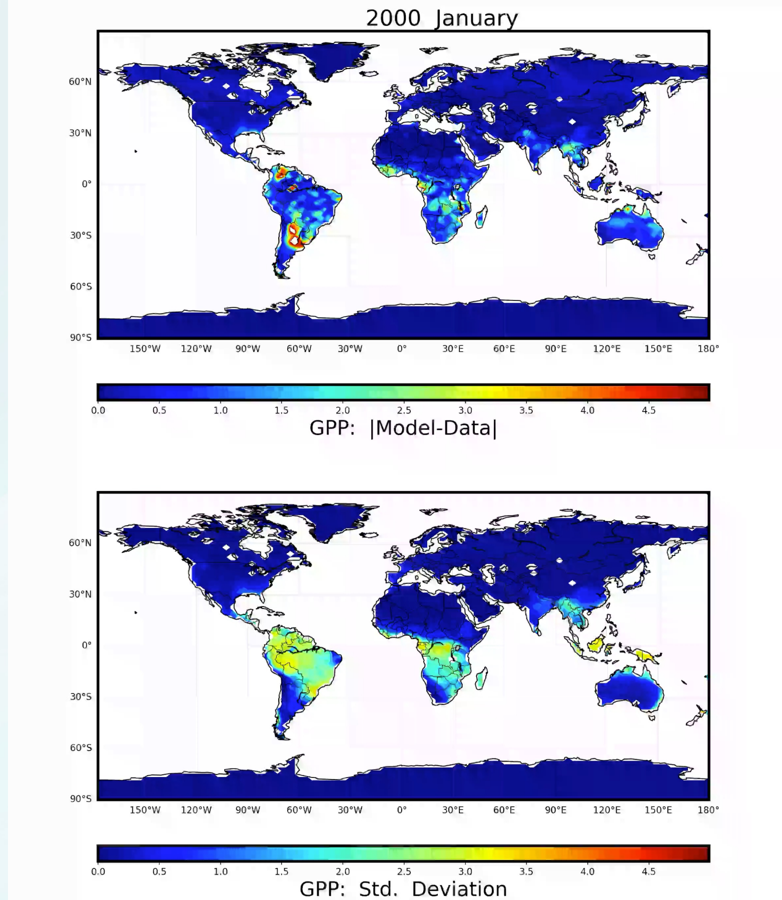


Calibration reduces predictive mean error and predictive standard deviation

Prior

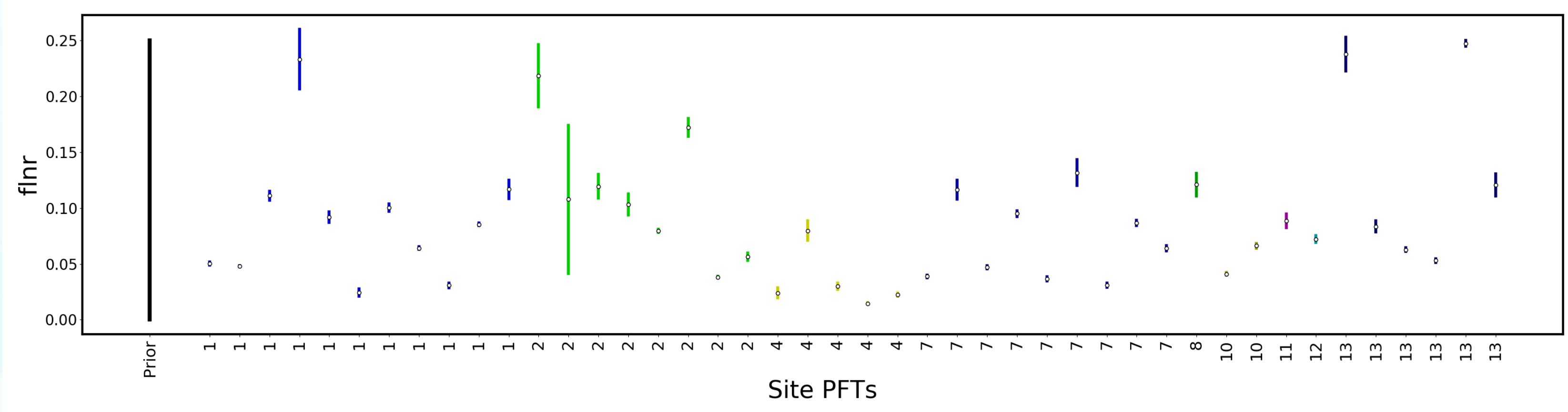


Posterior

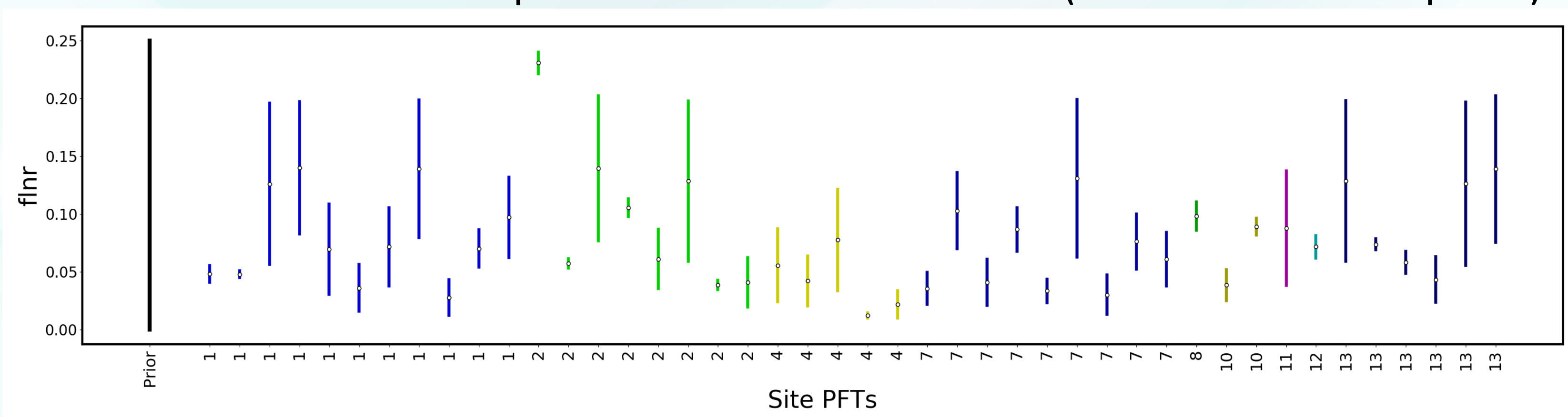


Calibrated parameter values across FLUXNET sites

Without model error: Overfitting, i.e. high variability in flnr across sites within and across PFTs



With model error: fair representation of unknown flnr (still narrower than prior!)



E3SM Energy Exascale Earth System Model

U.S. DEPARTMENT OF ENERGY

Summary

- **Constructed spatio-temporal surrogate to approximate ELM**
  - Karhunen-Loève + Polynomial Chaos expansions
  - Surrogate is orders of magnitude less expensive than ELM
- **Global sensitivity analysis or variance decomposition is a free bi-product**
- **Bayesian calibration using online evaluation of the surrogate**
  - Embedded model structural error provides the missing uncertainty component
  - Reduction of predictive uncertainty in light of FLUXCOM data
  - Full decomposition of predictive uncertainty

Param 1	Param 2	Param 4	Surr. error	Data noise	Model str. error

Next:

- **Build a global land-model calibration framework**
  - Construct ensembles with land biogeochemistry active (higher expense)
  - Determine sensitive parameters for land variables that couple to Earth system
  - Engage with ILAMB to prioritize datasets to be used to integrate with ELM
  - Find best parameters to use in future offline and coupled experiments