Uncertainty quantification in CLM: Comprehensive Parameter sensitivity analysis

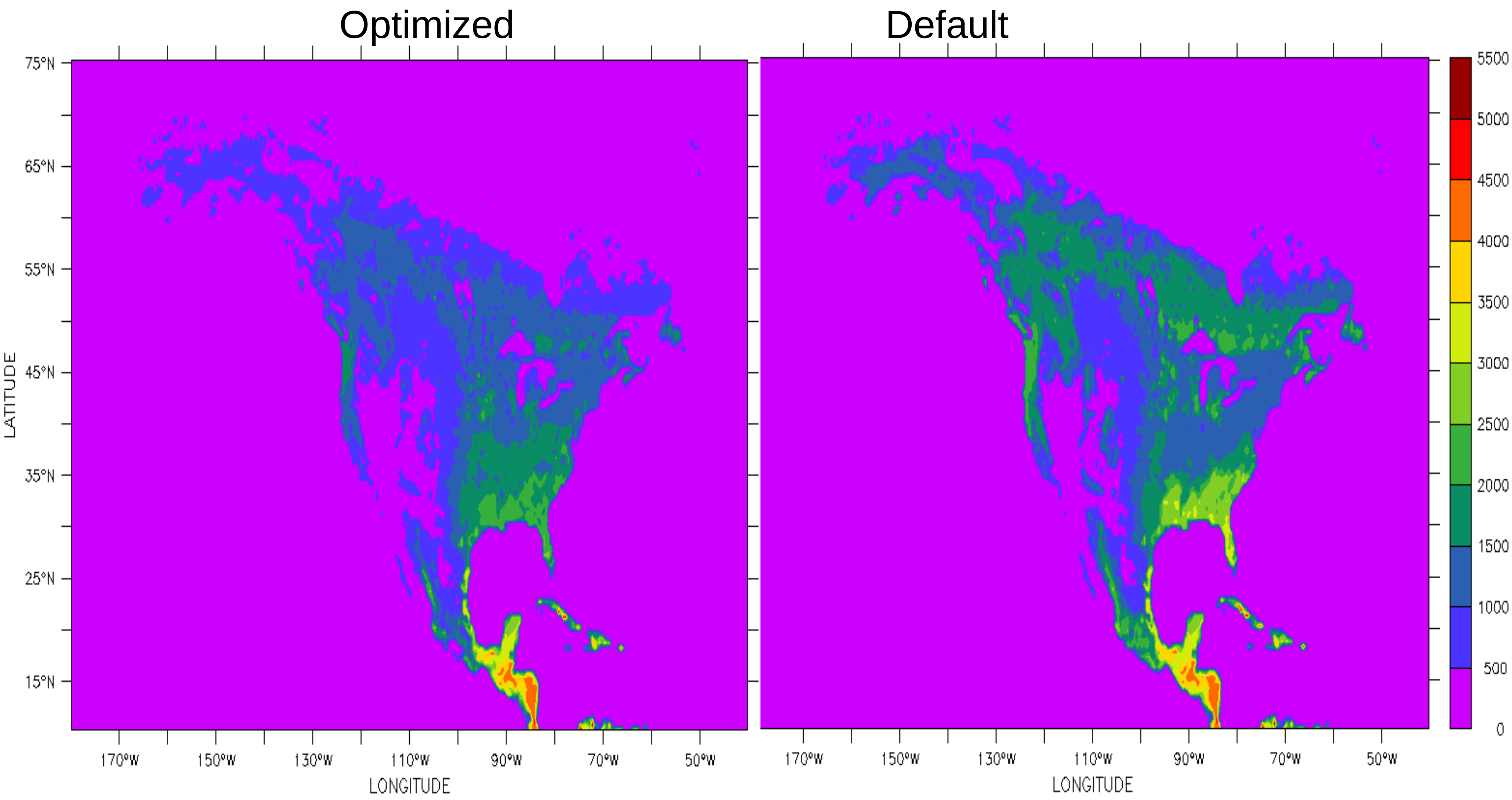
Daniel Ricciuto, Peter Thornton,
Dali Wang, Marcia Branstetter, Jitendra Kumar, Forrest Hoffman
Oak Ridge National Laboratory

Habib Najm, Cosmin Safta, Khachik Sargsyan
Sandia National Laboratory

CESM land model working group
March 1st, 2012

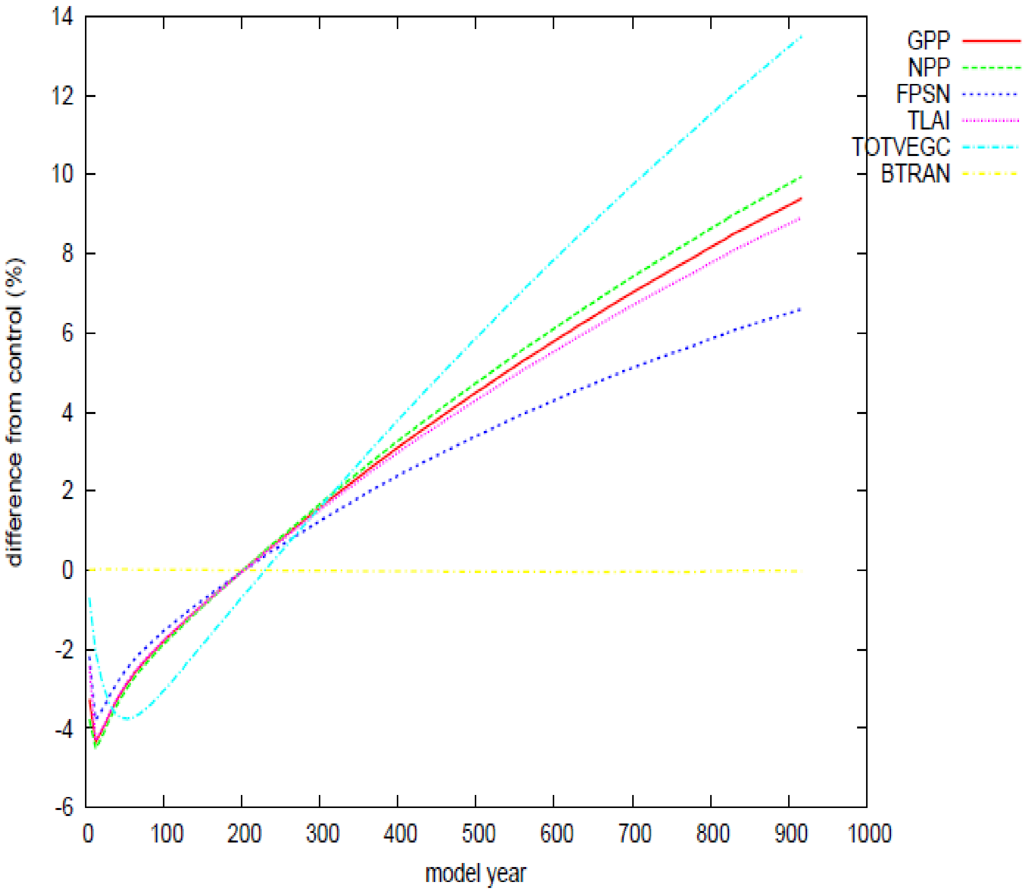
Motivation: Improved prediction at global scales

Use of FLUXNET observations in **LoTEC** to optimize PFT-level parameters results in a more realistic distribution of GPP in North America



CLM parameter sensitivity analysis: First steps

Q_{10} for heterotrophic respiration

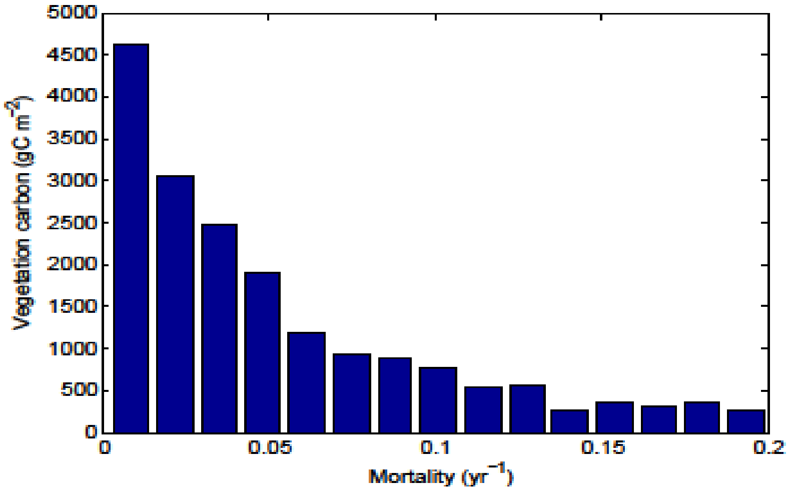


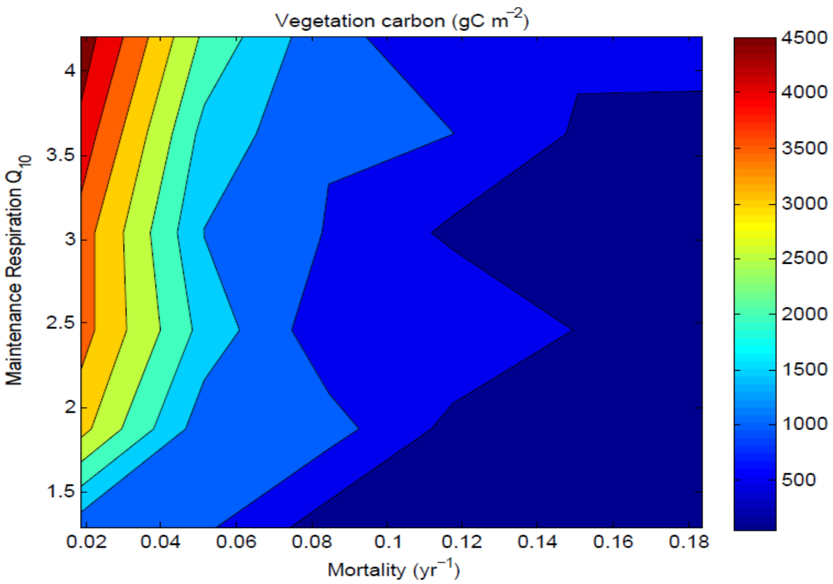
- Using CLM4-CN with modification for plant nitrogen pool
- 45 “non-PFT” parameters identified and pulled out of code into pft-phys file to enable sensitivity analysis
- Sensitivity of key variables to parameter perturbations (+5% for 81 parameters) for Niwot Ridge flux site
- Key points:

 - reequilibration after parameter perturbation takes millennia

- Response depends upon timescale

Monte Carlo parameter sensitivity analysis





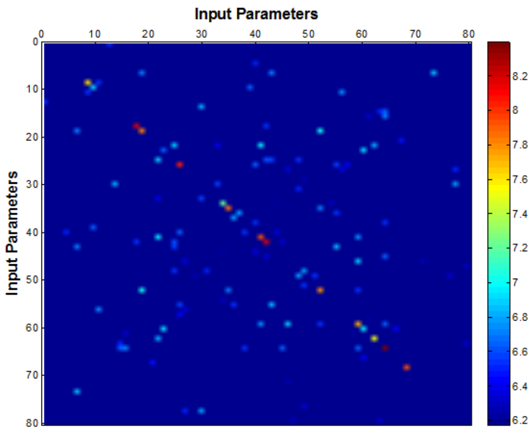
- We must also consider the interactions among the parameters
- 1000 samples randomly chosen from 81 parameters using uniform prior ranges (literature-based survey) for Niwot Ridge
- Analyze marginal PDFs of outputs

- reveals dominant parameters (e.g., mortality is the dominant control on TotvegC), parameter interactions (e.g. mortaility and Q_{10MR})

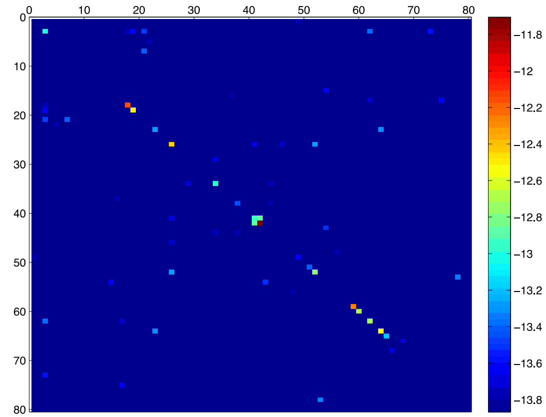
model samples can also be used to build an emulator, or "model of the model" which can interpolate response variables in parameter space and speed up DA

- Requires at least 10^3 – 10^4 simulations

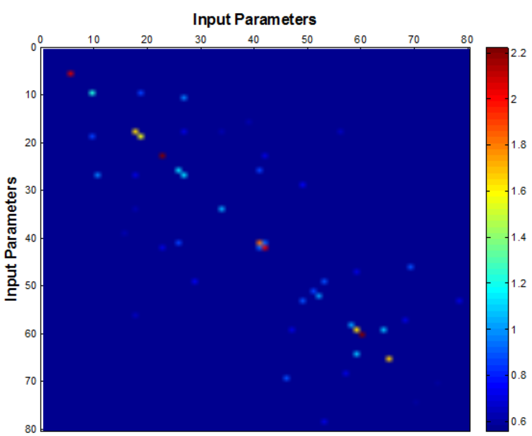
Determining key parameter interactions for emulator construction



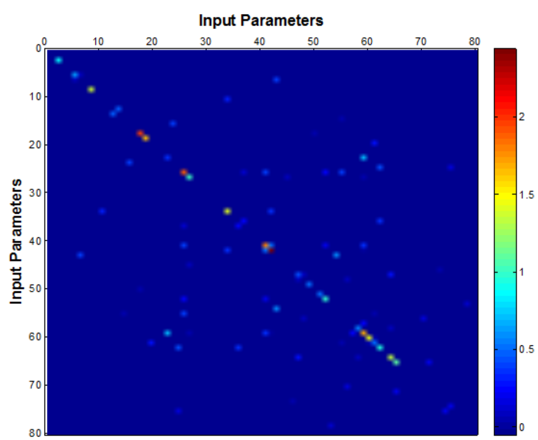
TOTSOMC



GPP



FSH

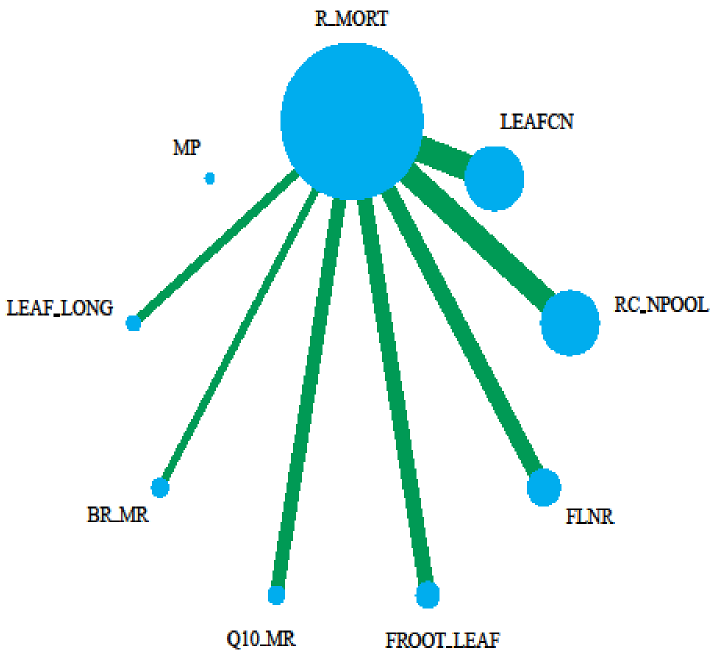


EFLX_LH_TOT

- Bayesian Compressive sensing (BCS) identifies key parameter and interaction terms
- These terms vary considerably as a function of output variable (e.g. TOTSOMC has more important terms)
- Limits number of terms required to create polynomial chaos based emulator

Key parameters and interactions

Parameters controlling TOTVEGC

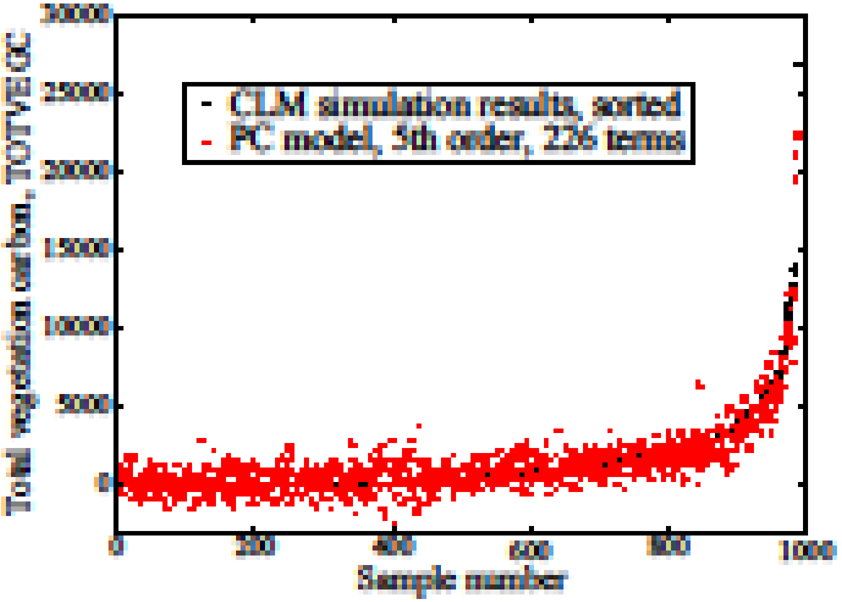


Courtesy of Habib Najm, Sandia National Laboratory

- Another way of visualizing the key parameters
- Based on analysis of 10^4 MC samples over 81 parameters
- Majority of variance is controlled by a handful of parameters and key interactions
  - Size of circle and thickness of line denotes contribution of variance

Building a CLM emulator

- $N = 987$ training runs based on uniformly distributed parameter values
- Outputs: steady-state, 10-year averages of 7 quantities



Name	Units	Description
TOTVEGC	gC/m^2	Total vegetation carbon
TOTSOMC	gC/m^2	Total soil carbon
GPP	$gC/m^2/s$	Gross primary production
ERR	W/m^2	Energy conservation error
TLAI	none	Total leaf area index
EFLX_LH_TOT	W/m^2	Total latent heat flux
FSH	W/m^2	Sensible heat flux

Courtesy of Cosmin Safta, Sandia National Laboratory

Key challenges to building an emulator

- Many parameter combinations fail to grow any vegetation
  - 40% at Niwot Ridge site
  - Hard to fit polynomial functions to these “flat regions” of parameter space
- Spinup requirements
  - We must run a full spinup for each parameter perturbation to avoid transient effects
  - 1-2 days of processing time per ensemble member for a single point
- Parallelization
  - We can run samples in parallel, but does not scale well above 200-400 simultaneous point simulations. Computing resource requirements for regional/global runs are extreme.
- Solutions
  - Identify flat regions and use different functions within the emulator to fit
  - Methods to accelerate spinup – some promising ideas have been presented
  - Reduced global grid using spatial clustering

Parameter optimization technique

- Markov Chain Monte Carlo (MCMC)
  - Joint parameter PDF estimated from observations
  - Bayesian technique: uses prior knowledge (e.g. $\eta > 0$, $Q_{10}>1$, $\beta > 0$)
  - Too slow to run directly on CLM, will run on **emulator**
  - Step 1: compute likelihood L_0 given initial parameter guesses

$$L(\mathbf{x}_i|\boldsymbol{\theta}_\mathbf{k}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{1}{2}\left[\frac{f(\boldsymbol{\theta}_\mathbf{k}, t_i) - x_i}{\sigma_i}\right]^2\right)$$

x_i: ith observation

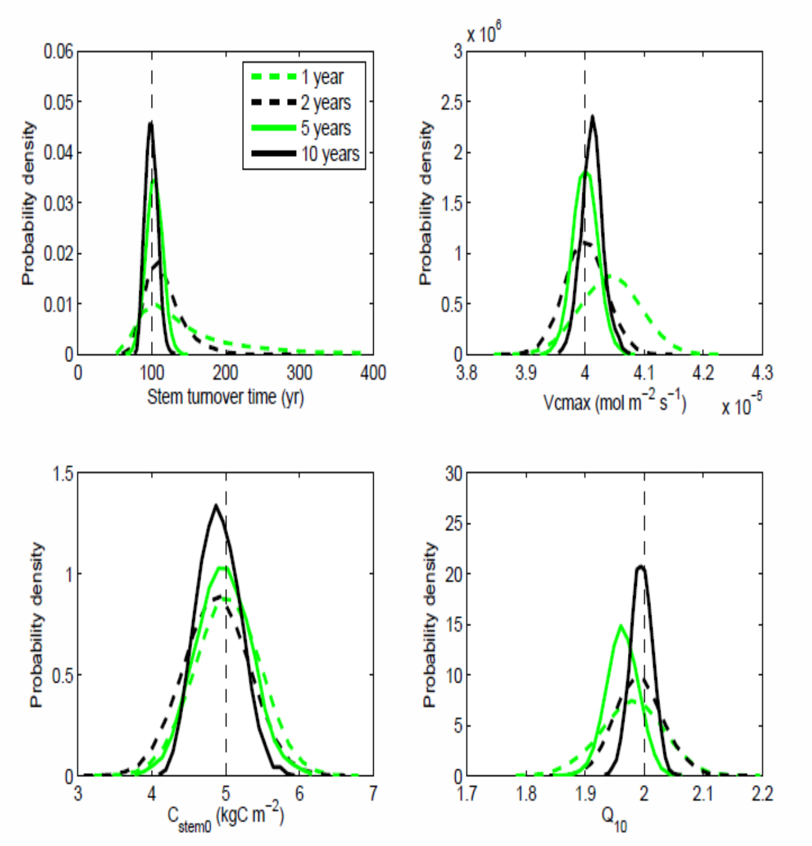
Θ_{k}: kth parameter set

σ: obs error (normal dist)

  - Can be adjusted to account for autocorrelation in data
  - Perturb parameter set, compute new likelihood L_1
    - If $L_1/L_0 > U[0,1]$, accept
    - If $L_1/L_0 < U[0,1]$, reject, perturb again
  - Repeat until chain is stationary (10^5 **emulator** evaluations)
  - Remove burn-in (dependence on initial conditions)

Parameter and prediction uncertainties

LoTEC parameter PDFs



From Ricciuto et al. (2011)

- MCMC provides the full joint posterior parameter PDF (means, variances, covariances)
- This is a measure of the uncertainty about model parameters.
  - Uncertainty can be quantified as a function of observation type, error, or record length
  - Have not done this yet with CLM – example from LoTEC shows how uncertainty is reduced as more observations are added
- Parameter uncertainty → prediction uncertainty
  - We can sample from the parameter PDF and run an ensemble of forward predicitons to generate confidence intervals

Running global CLM ensembles

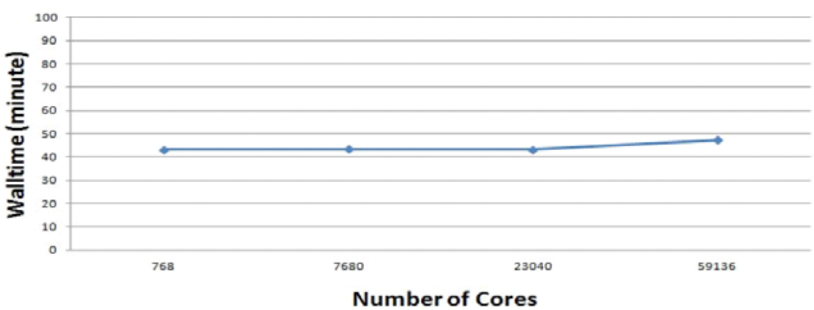
Assimilation of gridded datasets will require regional to global simulations
An ensemble of global runs scales well on Jaguar up to 75 simultaneous simulations
Can use > 50k cores

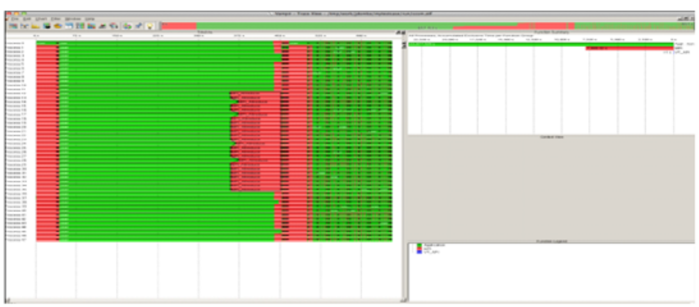
Global Parameter Sensitivity Simulation on OLCF

Workflow Design

New CLM Case (I1850CN) using the Restart file AD_Spinup simulation (1290yr, CRUNECP 0.5 degree dataset) → **Reconfigure CLM Case** (reduce software dependencies) → **Parallel Utility to Setup Ensemble Simulations** (77 parameters with 5% variation) → **Global Simulations Utility** (manage simulation, restart and archive)

Scalability and Preliminary Profiling Information with VampirTrace

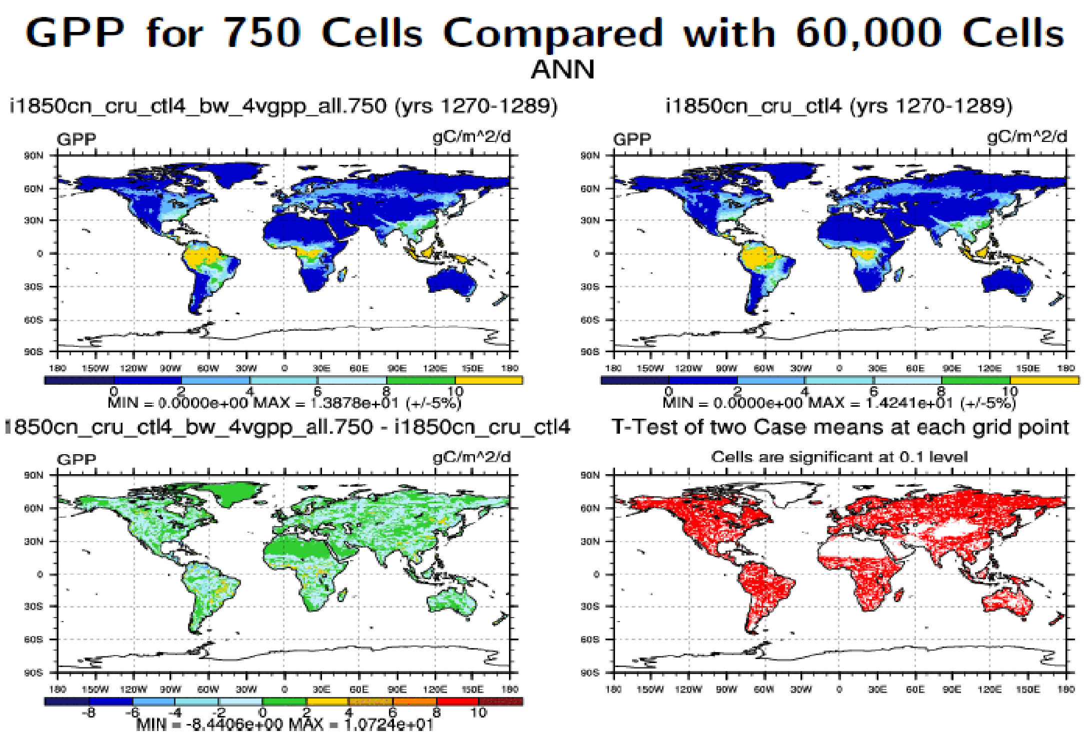


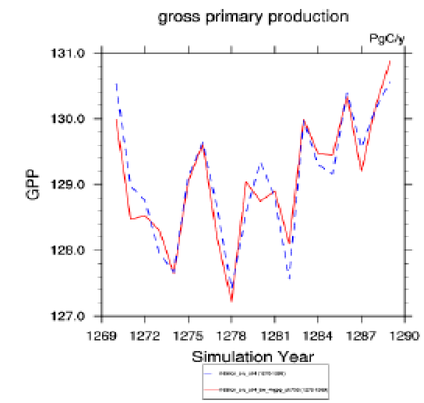


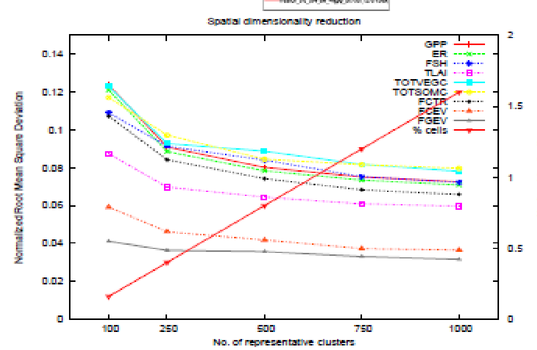
Courtesy of Dali Wang, Oak Ridge National Laboratory

Spatial Dimensionality Reduction for Sensitivity Analysis

- Global CLM simulations at $0.5^\circ \times 0.5^\circ$ have ~60,000 grid cells that must be modeled in hundreds of 100–1000 y simulations, which is computationally untenable.
- Cluster analysis uses the CRU-NCEP climate data, plant functional type (PFT) characteristics, and steady-state modeled quantities.

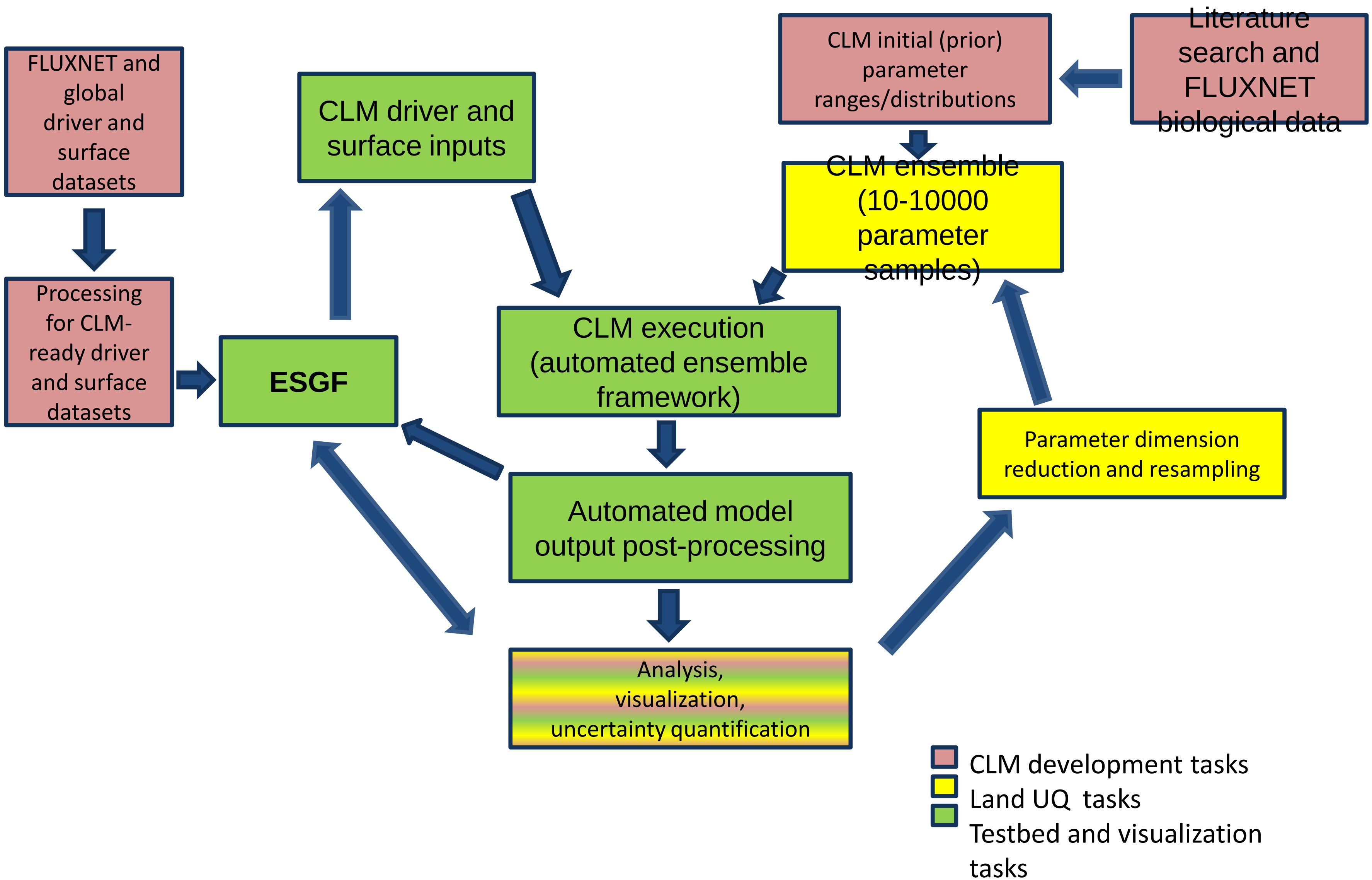




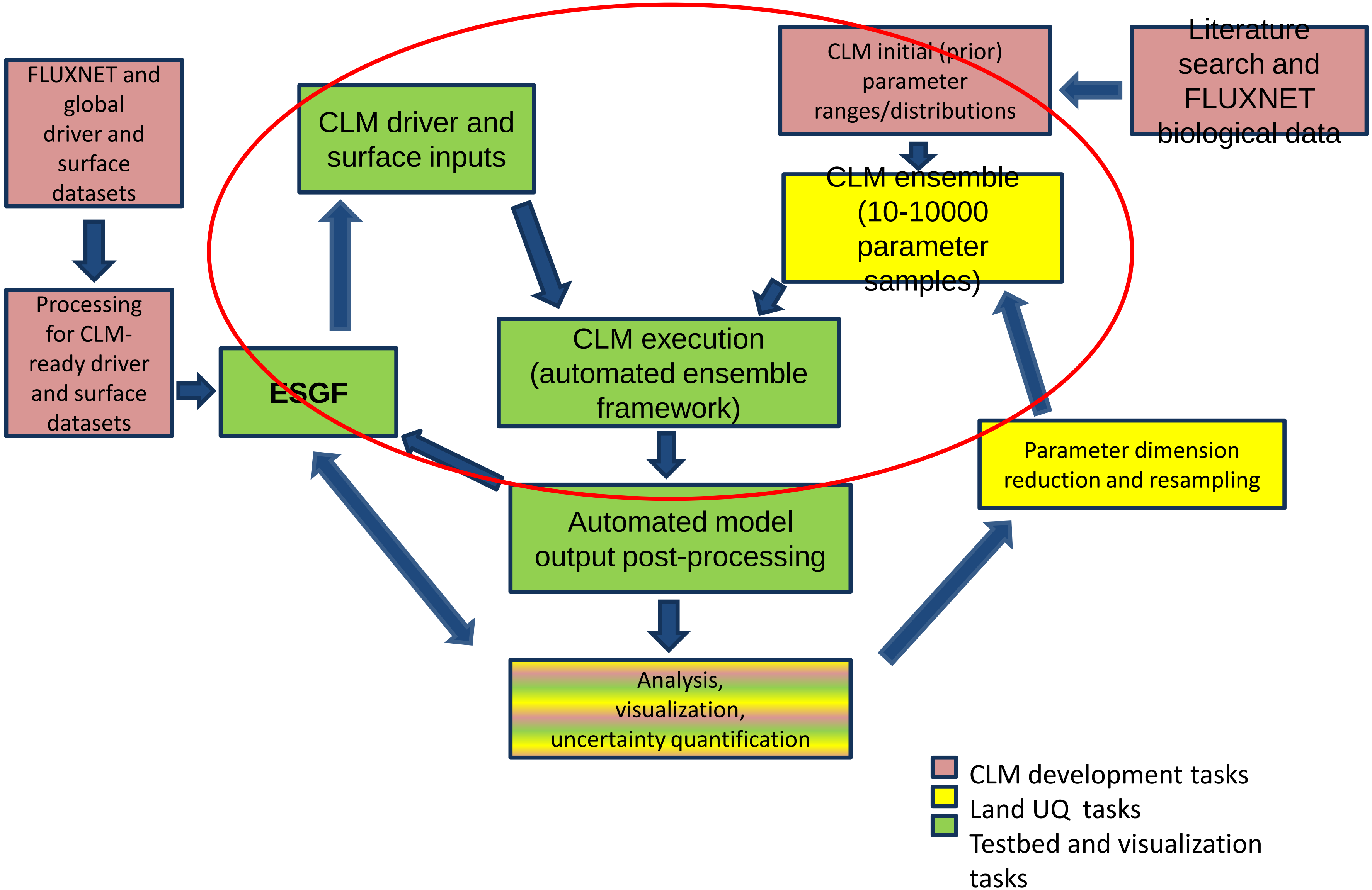


Courtesy of Forrest Hoffman and Jitendra Kumar, ORNL

Proposed workflow for CLM data assimilation



Proposed workflow for CLM data assimilation



Funding Acknowledgements

- Funding for this work is provided by
  – Climate Science for a Sustainable Energy Future (CSSEF)

    Biological and Environmental Sciences (BER), DOE

  -- Terrestrial Ecosystem Science SFA

    Biological and Environmental Sciences (BER), DOE