Machine learning approaches for surrogate modeling

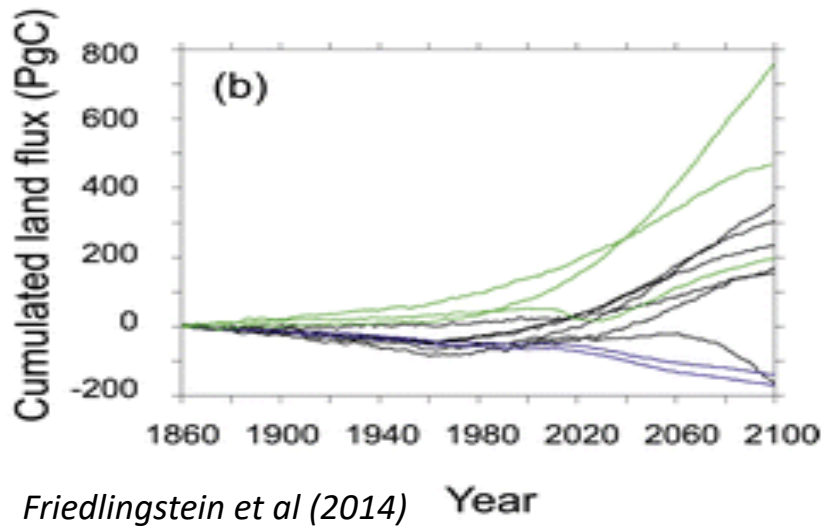
Daniel M. Ricciuto (ORNL) and Khachik Sargsyan (SNL-CA)

Cosmin Safta, Vishagan Ratnaswamy (SNL-CA)
Dan Lu, Peter Thornton, Anthony King (ORNL)
Jayanth Jagalur Mohan, Youssef Marzouk (MIT)

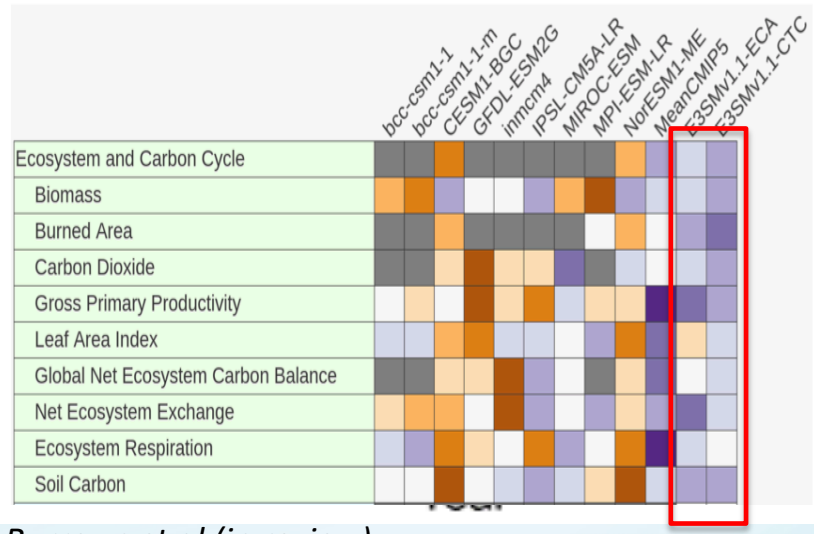
E3SM all hands meeting
November 21st, 2019

Overview and motivation: CBGC models

- Multi-model ensembles standard
  - Large spread in outputs
  - Many differences in parameters and structure
  - Difficult to pinpoint causes of differences
- Within-model ensembles limited
  - Expensive model evaluation
  - High dimensionality
- Key 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?

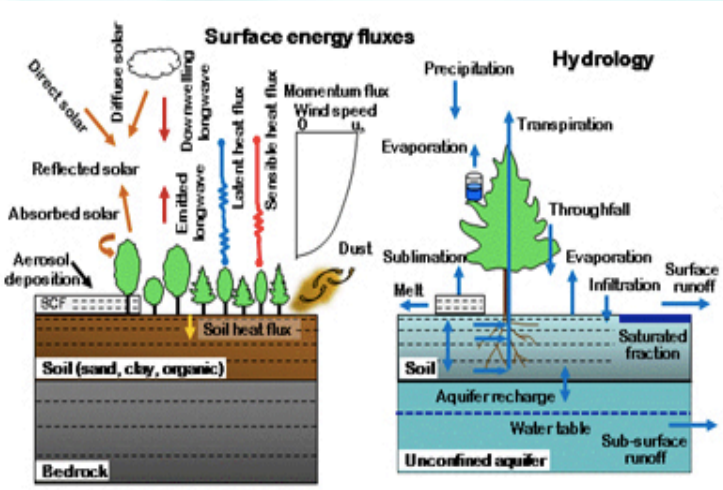


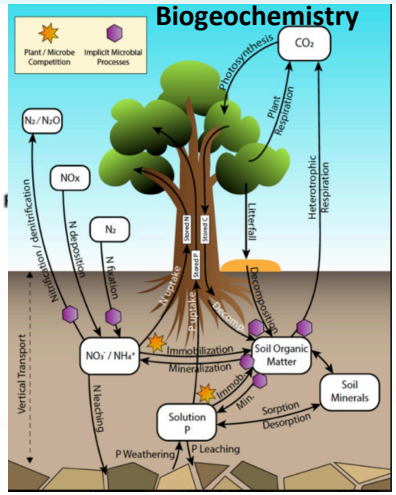
Friedlingstein et al (2014)

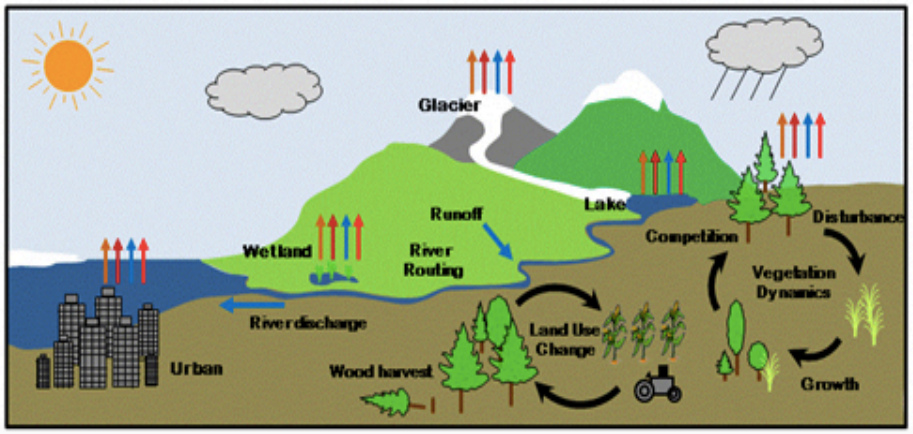


Burrows et al (in review)

Overview and motivation: Land BGC

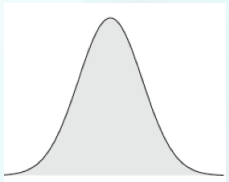






- ELM is an increasingly complex model with many processes
- Large ensembles are needed for UQ, which are expensive (even for land)
- Land model is a good testbed for new approaches
- Surrogate models increase UQ efficiency

Uncertainty Propagation … enabled by Surrogate Models

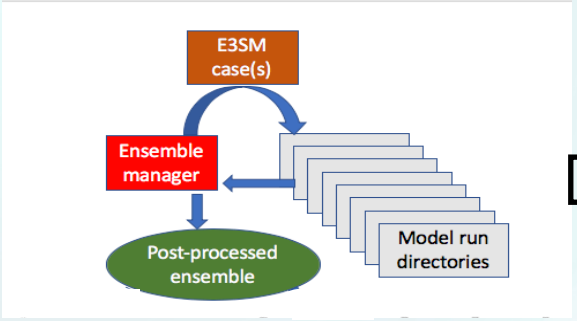


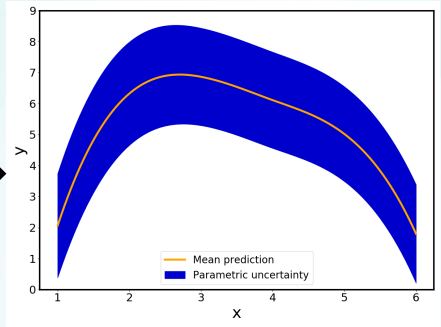
λ

Input parameters

model

$$f(x;\lambda)$$





Output prediction

Work with the existing model as a black-box

non-intrusive: No change to model code, but significant workflow challenge

create an ensemble of simulations with varying/perturbing parameters

Ensemble is used for training and/or validation samples of our **surrogate:**
(proxy, metamodel, emulator, response sfc, etc.)
Enables uncertainty propagation, sensitivity analysis, efficient parameter calibration

$$f(x;\lambda) \approx f_s(x;\lambda)$$

Various surrogate types explored

UQ

ML

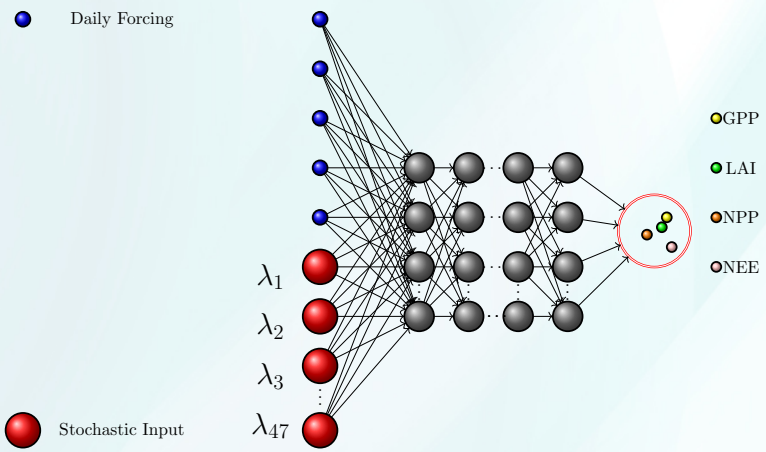
- **Polynomial chaos (PC):**
  - Not dynamical chaos! Essentially a polynomial fit/regression to a model
  - Extremely convenient for uncertainty propagation, moment estimation, global sensitivity analysis
    - e.g., PC surrogate allows extraction of sensitivity indices 'for free'
  - Can deal with highly non-linear models, but assumes some level of smoothness
  - Used successfully for past site-level ELM sensitivity analysis (Sargsyan et al., 2014; Ricciuto et al, 2018), calibration (Lu et al., 2018)

- **Neural networks (MLP, RNN, LSTM):**
  - Can potentially better deal with non-smooth behaviors
  - Can potentially deal with more complex outputs
  - Cons: harder to train and interpret

Neural Network surrogates allow more flexibility

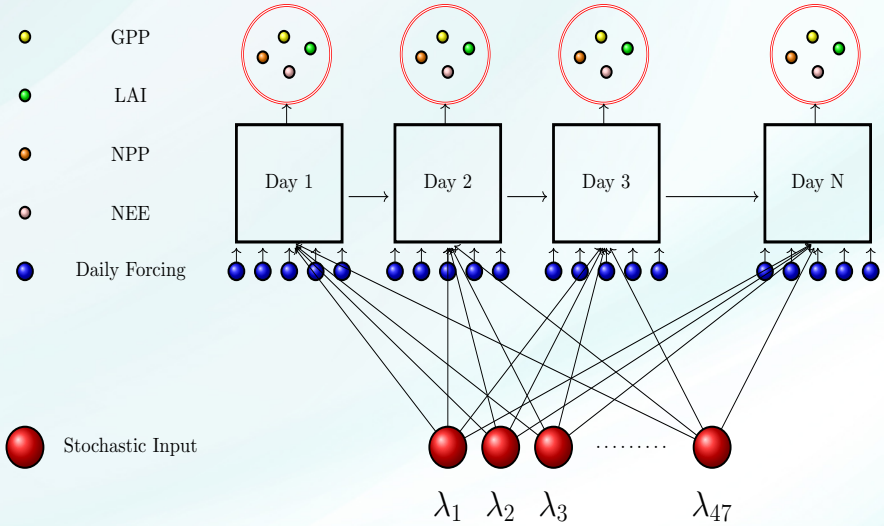
Multilayer Perceptron (MLP)

- Feedforward artificial neural network
- 3 or more layers (input, output, hidden)
- Scalar quantities of interest (QoIs), e.g. long-term means at one point



Recurrent Neural Network (RNN)

- Connections between nodes along a temporal sequence
- Current day affected by history
- Useful for timeseries

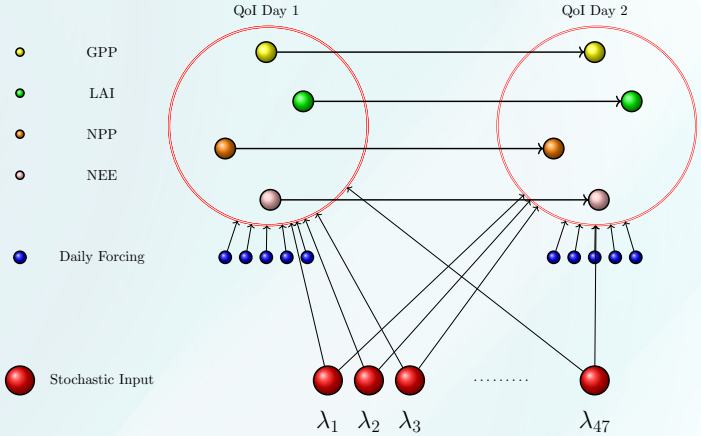


3 applications of ML surrogates

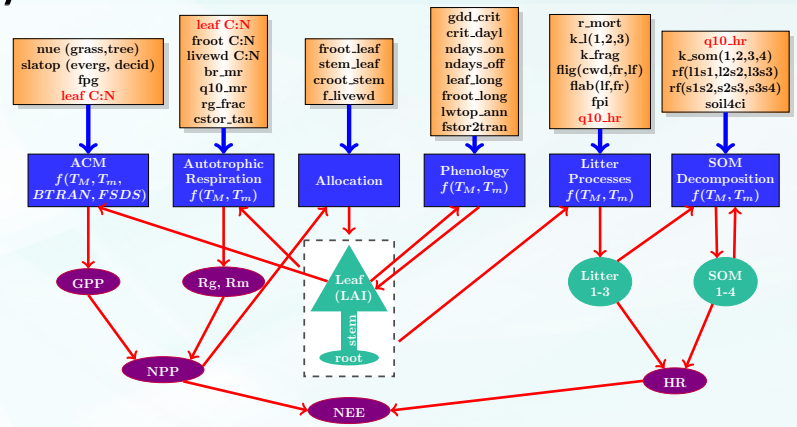
- Creating an LSTM (type of RNN) surrogate model to explore parametric uncertainty effect on a land model timeseries.
- Using dimension reduction approaches combined with MLP to create an accurate surrogate model for a spatiotemporally varying output
- Replacing a land model sub-component with an MLP surrogate model to improve computational efficiency and explore parametric uncertainty

We have created specialized RNN architecture knowing the connections between processes

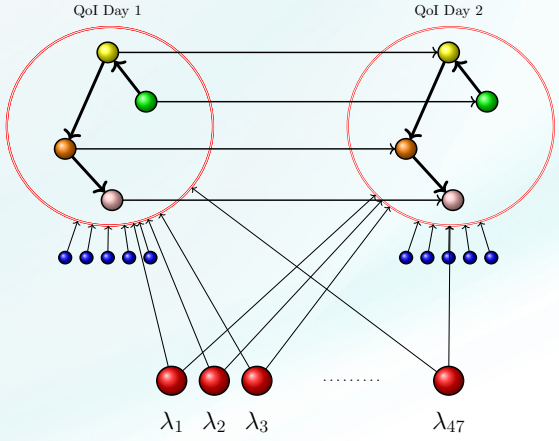
Vanilla long short-term memory (LSTM) network



Physics-informed LSTM

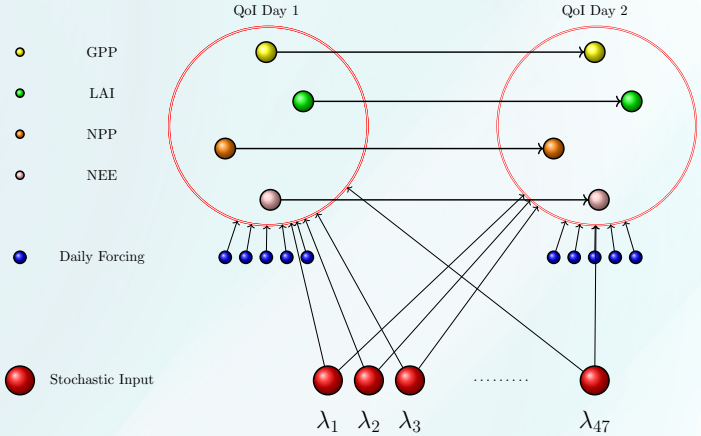


sELM (ELM carbon only)

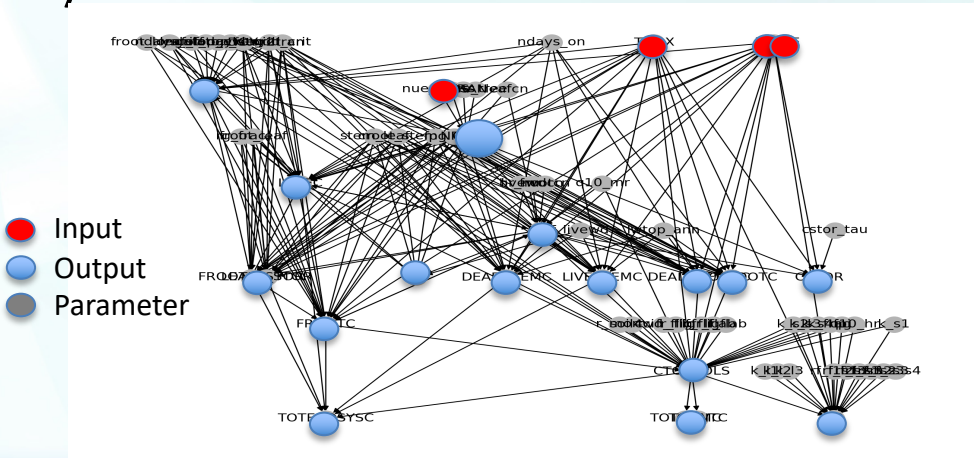


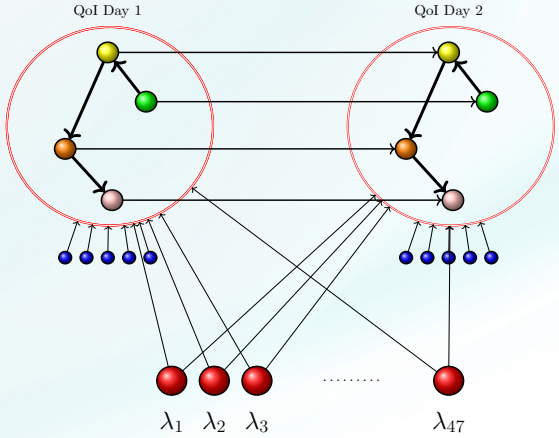
We have created specialized RNN architecture knowing the connections between processes

Vanilla long short-term memory (LSTM) network

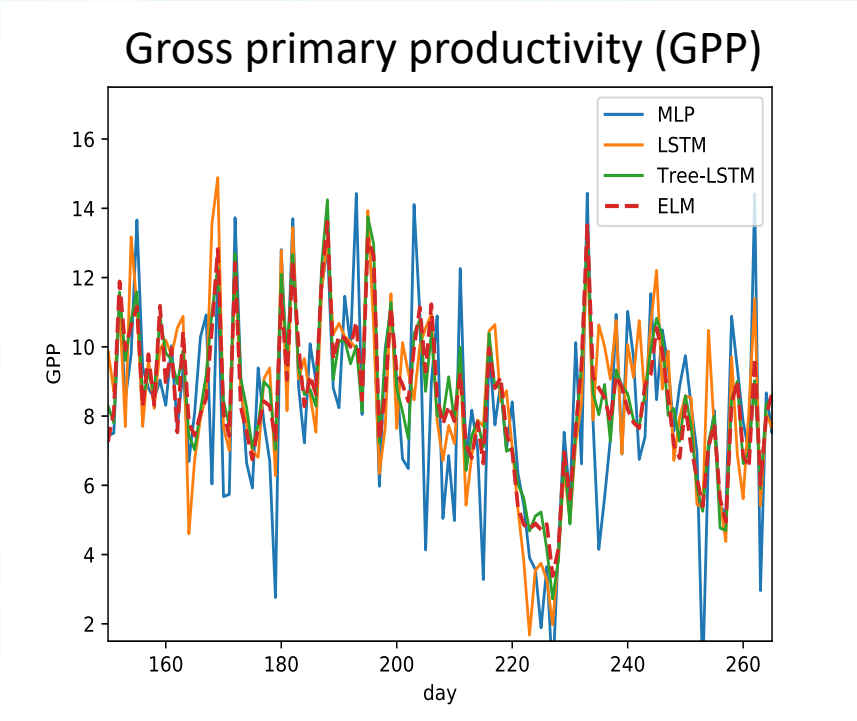


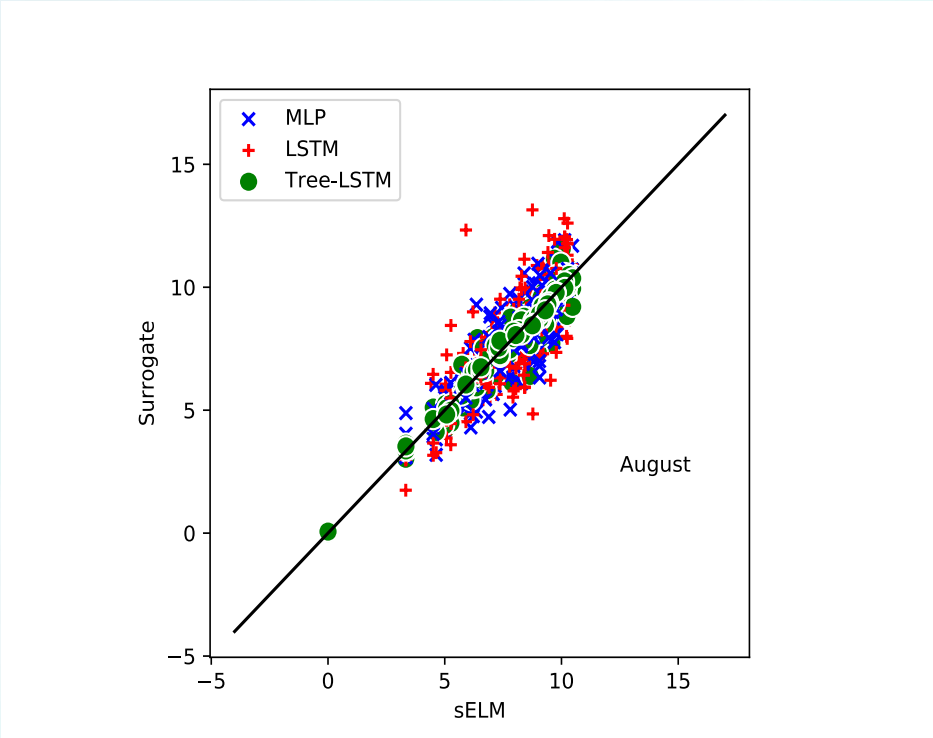
Physics-informed LSTM





Physics-informed RNN architecture captures daily dynamics well with a fraction of the cost

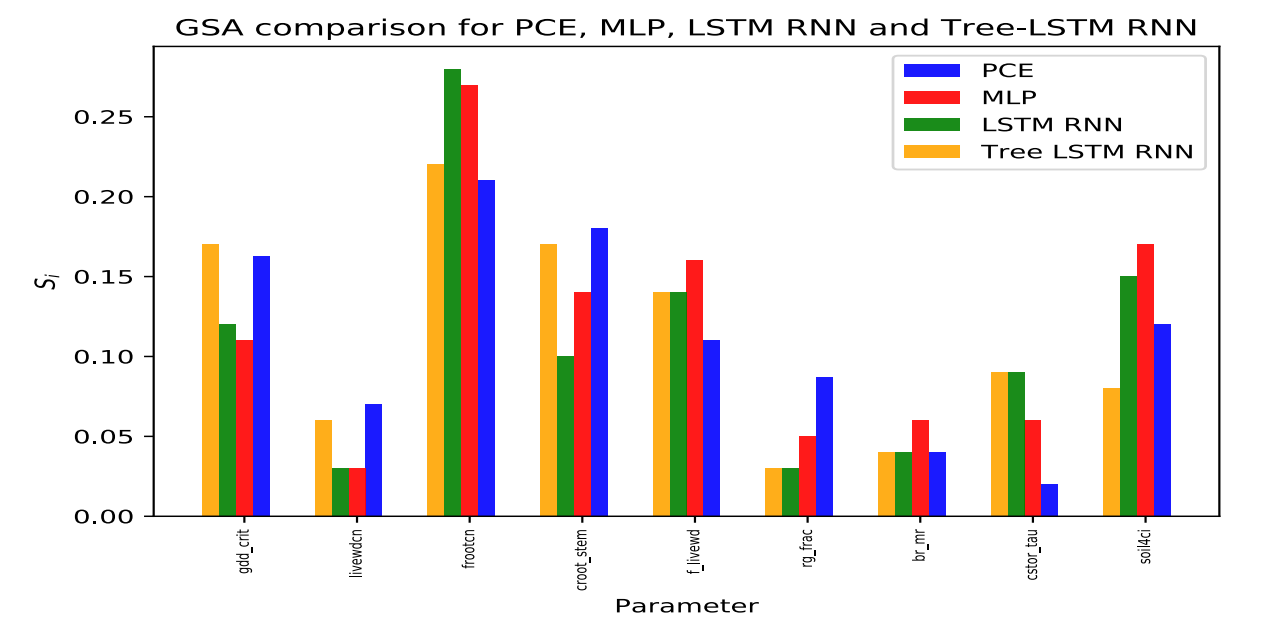




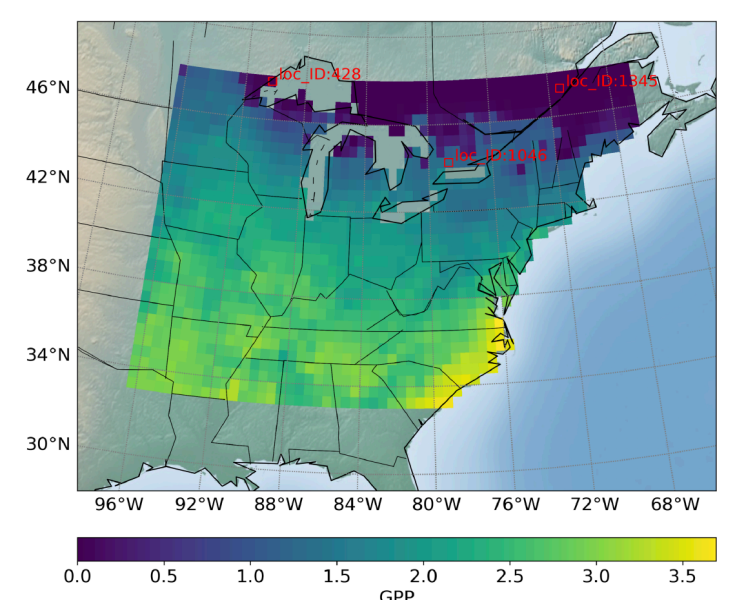
US-UMB flux site, northern Michigan

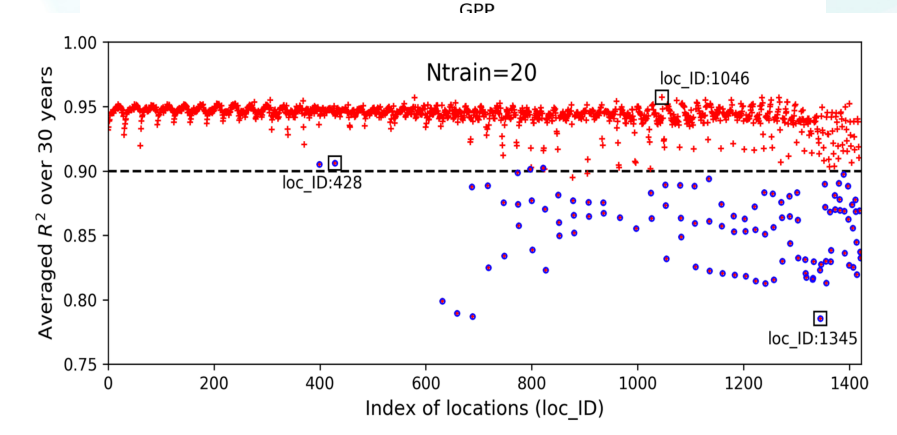
Physics-informed RNN architecture captures daily dynamics well with a fraction of the cost of ELM

- For GSA, some disadvantages compared to PC:
  a) GSA requires extensive sampling of the RNN surrogates.
     * Not a big deal if the limiting factor is cost of ELM simulations
  b) Does not come with uncertainties
- Surrogate accuracy much more important for calibration



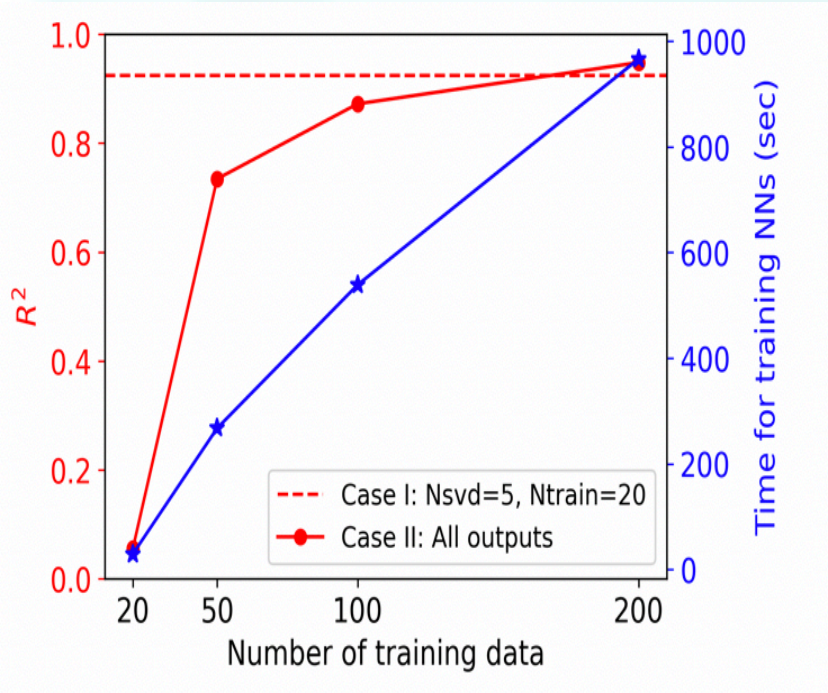
Surrogates for spatially varying outputs





- In this example, we have 42660 GPP outputs (30 years * 1422 gridcells)
- 8 model parameters → 2000 ensembles
- Singular value decomposition (SVD) can be applied to reduce the dimensionality of our output
- Retaining first 5 singular values captures > 97% of output variance
- For validation samples, surrogate model has strong correlation with original model output. Exception: northern areas with marginal GPP

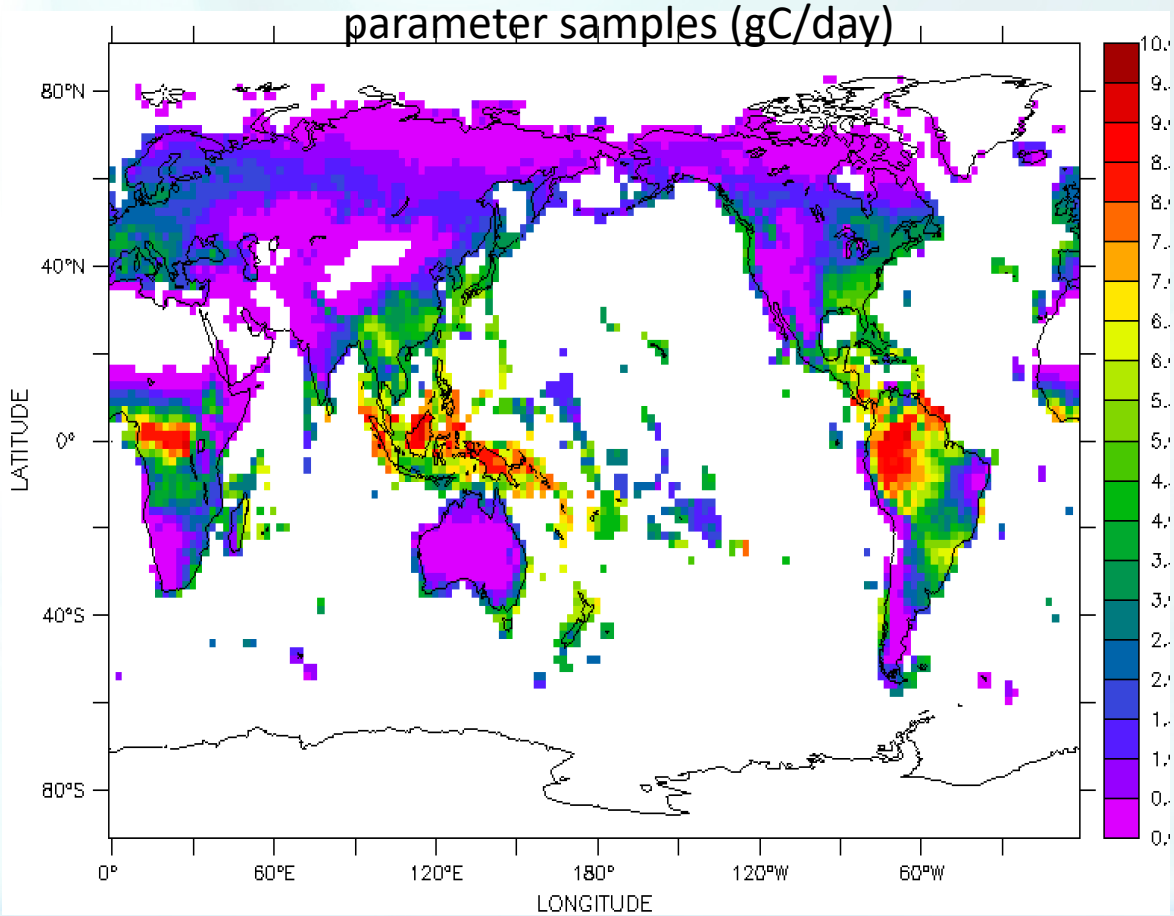
Surrogates for spatially varying outputs



- When we train 46,600 surrogates independently, we need more samples and more time for the same level of accuracy..
- NN with 5 singular values trained in 4 seconds – fewer samples and far less time than standard approach
- Only 20 training samples (model simulations) are necessary for good ELM surrogate accuracy at most locations.
- If this holds for coupled simulations, feasible approach for model tuning

Using ML surrogates for model subcomponents

Mean surrogate model GPP over 500 parameter samples (gC/day)



- Canopyfluxes: 20-40% of land model computation

- **GPP functional unit**:

GPP = f(met, params, states)

Met data: T, RH, FSDS, wind, Pa

Params: slatop, mbbopt, leaf C:N, flnr

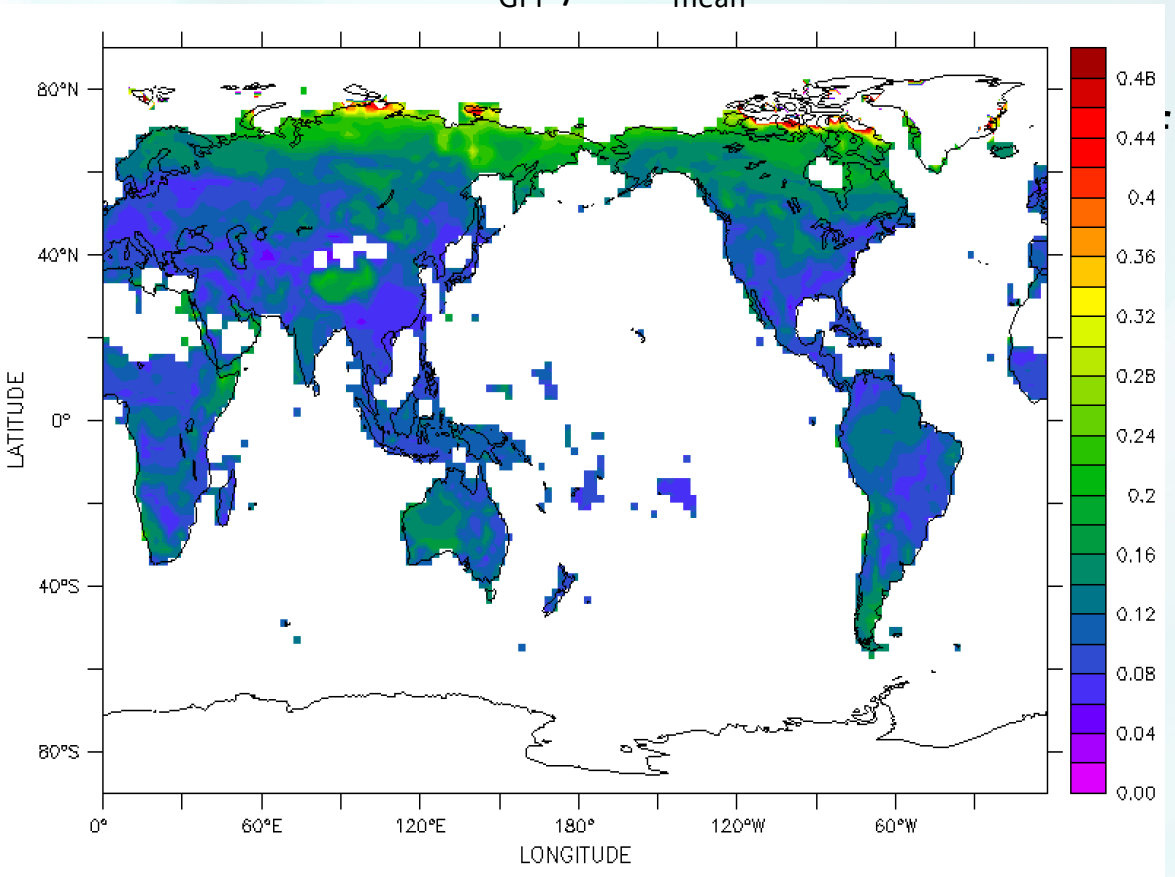
States: LAI, BTRAN (input)

Does not consider feedbacks to LAI, soil moisture

- MLP trained on outputs

Using ML surrogates for model subcomponents

σ_{GPP} / GPP_{mean}



- Neural network trained on 20k daily ensembles randomly selected from forcing and run through functional unit
- Trained network used to predict global GPP uncertainty as a function of input parameter uncertainty (+/- 25%) and global drivers
- Full ELM-SP: 50k core-hours
- Surrogate – 60 core-hours

Summary

- **Forward UQ using ML approaches for surrogate modeling**
  - For point/site ELM simulations, we have **well-developed ensemble workflows** for running model ensembles, post-processing, and some UQ analysis using the UQ toolkit (https://www.sandia.gov/UQToolkit/).
  - Extending to **FATES, crop** versions of ELM, as well as **SCM.**
  - **LSTMs are a promising method** for accurate surrogates of **timeseries** outputs
  - **Dimension reduction** using SVD or other approaches **simplifies surrogate** model training and reduce the computational demand

- **Next steps**
  - Work towards **high accuracy** of surrogates with smallest possible number of simulations using combined approaches, and quantify surrogate errors.
  - Explore **combinations of the 3 approaches** presented for creating large spatiotemporally varying surrogates with large ensembles
  - **Calibrate ELM parameters** using observations (e.g. remotely sensed/synthesis products for GPP, ET, LAI)
  - Explore UQ for **tuning coupled system** (AMIP or fully coupled configurations)