Advancing Predictive Understanding of Terrestrial Ecosystem through Machine Learning

Dan Lu; Daniel Ricciuto

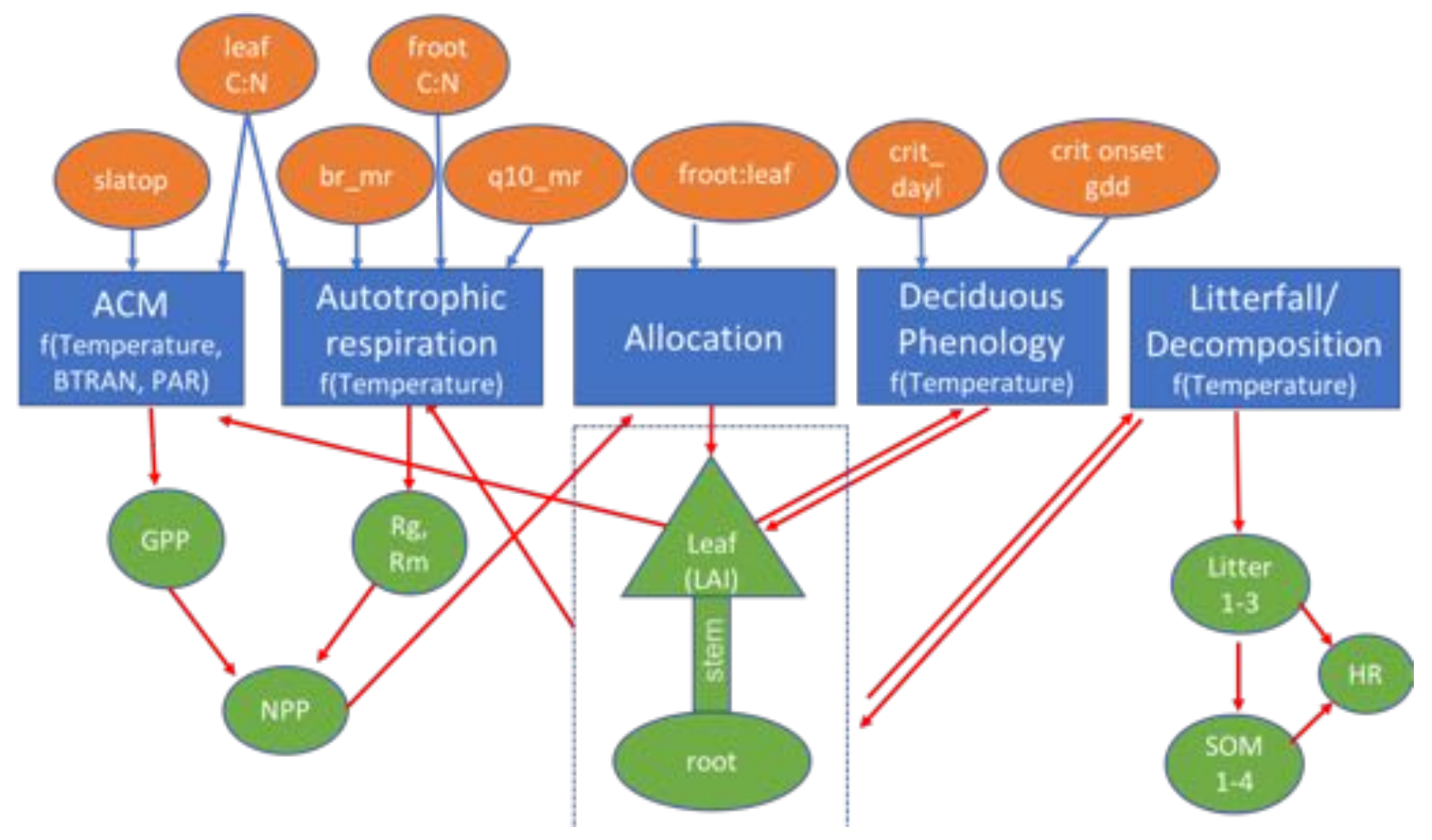
Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831 USA

Problem Statement

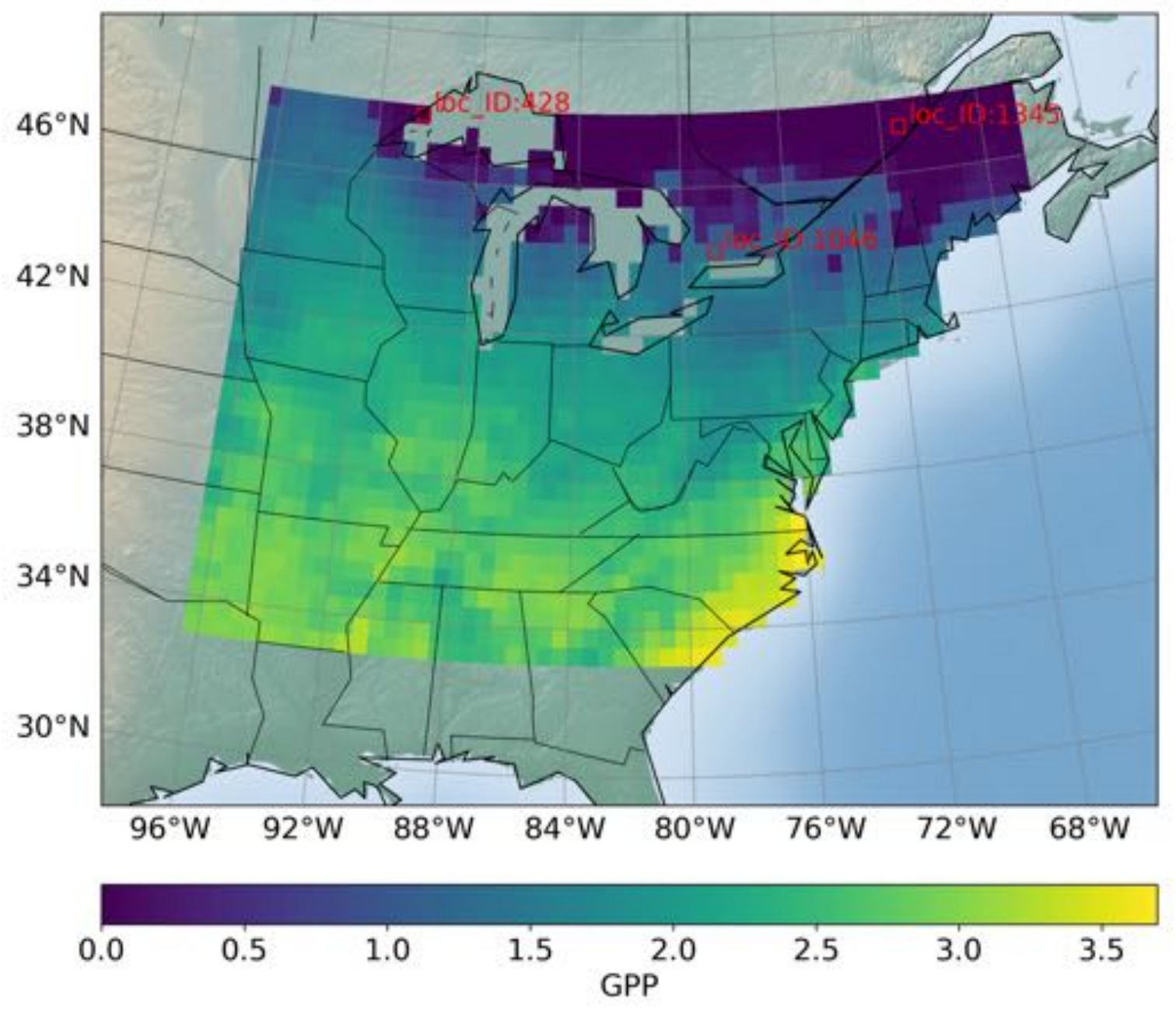
- ❖ Improving predictive understanding of terrestrial ecosystem variability and change requires data-model integration.
- ❖ Efficient data-model integration for complex models requires surrogate modeling to reduce model evaluation time.
- ❖ However, building a surrogate of a large-scale terrestrial ecosystem model (TEM) with many output variables is computationally intensive because it involves a large number of expensive TEM simulations.
- ❖ In this effort, we propose an efficient surrogate method capable of using a few TEM runs to build an accurate and fast-to-evaluate surrogate system of model outputs over large spatial and temporal domains.
- ❖ Procedure: (1) first use singular value decomposition (SVD) to reduce output dimensions, and (2) use Bayesian optimization techniques to generate an accurate neural network (NN) surrogate model based on limited TEM simulation samples.
- ❖ Impact: our machine learning based surrogate methods can build and evaluate a large surrogate system of many variables quickly. Thus, whenever the quantities of interest change, e.g., a new site and a longer simulation time, we can simply extract the information of interest from the system without rebuilding new surrogates, which significantly advances data-model integration analysis and improves predictive understanding of terrestrial ecosystem.

Application

- ❖ We apply the method to a TEM which simulates several carbon state and flux variables and we consider GPP here.
- ❖ We build a surrogate model of the TEM to predict annual GPP in deciduous forest systems in the eastern region of the US for 30 years between 1981-2010.
- ❖ One TEM run takes about 24 hours on a single processor.


leaf C:N
froot C:N
slatop
br_mr
q10_mr
froot:leaf
crit_dayl
crit onset gdd
ACM f(Temperature, BTRAN, PAR)
Autotrophic respiration f(Temperature)
Allocation
Deciduous Phenology f(Temperature)
Litterfall/ Decomposition f(Temperature)
GPP
Rg, Rm
NPP
Leaf (LAI)
stem
root
Litter 1-3
HR
SOM 1-4

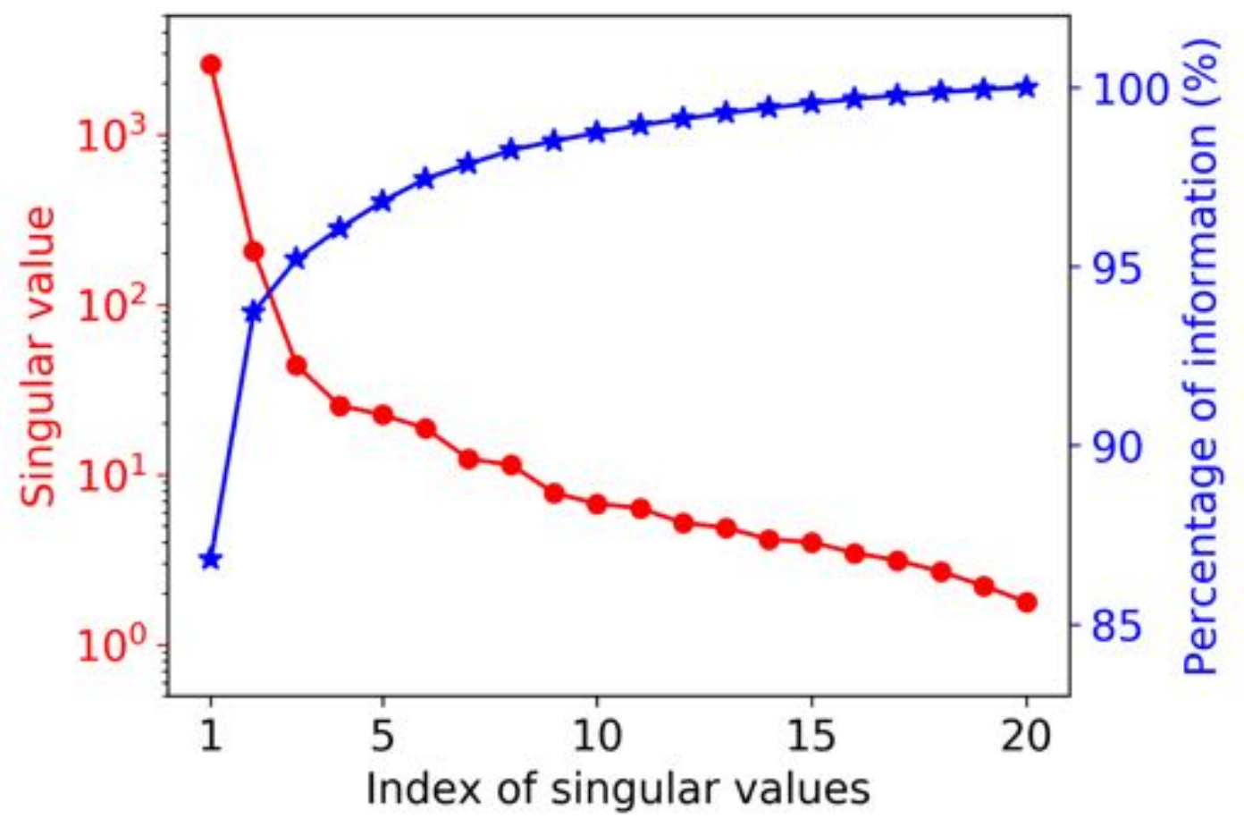
- Orange ovals: 8 uncertain parameter inputs
- Blue boxes: 5 processes
- Green shapes: model state variables
- Blue arrows: parameters are input to a certain processes
- Red arrows: model state variables can be outputs for some processes and input for other processes



- The region of interest covers 1422 land grid cells and each simulates 30 years, so total 42660 output variables.
- The model uses one PFT and the phenological drivers e.g., air temperature, solar radiation, vapor pressure deficit, and CO_2 concentration are used as boundary conditions.

Dimension Reduction

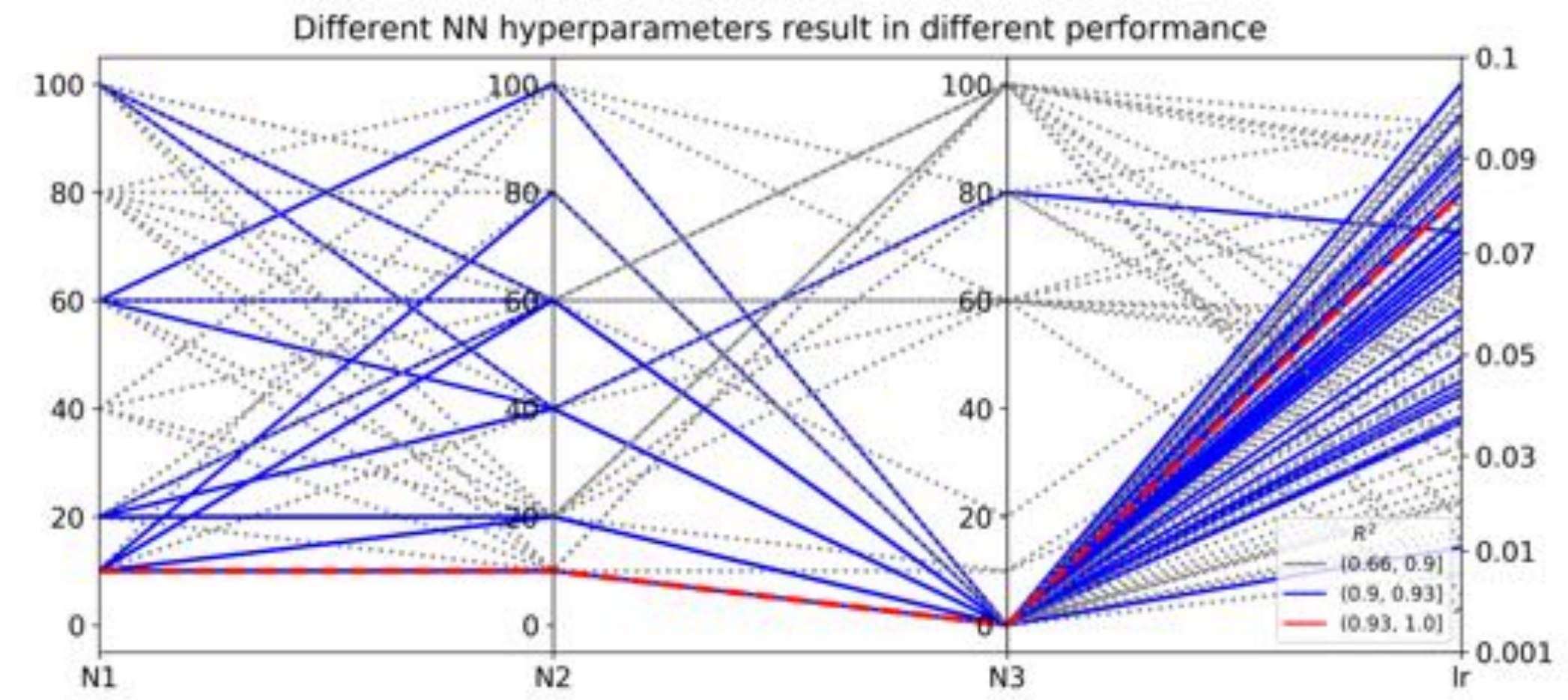
- ❖ We use SVD to reduce output sample dimensions to 5.
- ❖ We build the surrogate model in the reduced dimension space to reduce the required number of training data.



- Since the spatiotemporal GPPs have strong correlation, we can use a few principal components to capture the most information.
- The top 5 singular values contain 97% information of training data matrix with 42660 GPP variables and 20 samples.

Neural Network

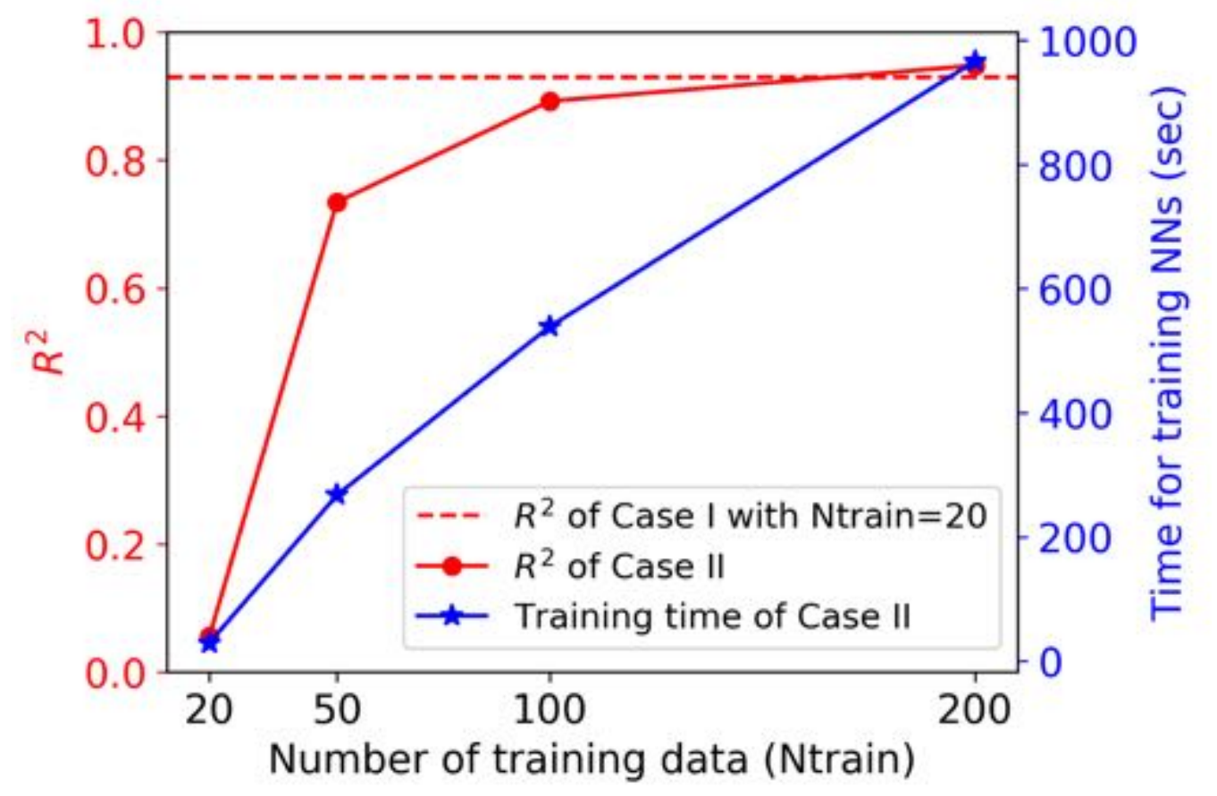
- ❖ We use neural network (NN) to build the surrogate model.
- ❖ We use Bayesian optimization to design the NN architecture and optimize its hyperparameters.
- ❖ Different sets of hyperparameters result in different NN performance and hyperparameter tuning is necessary to mitigate under- and over-fitting.
- ❖ Dimension reduction can simply the NN architecture and accelerate its hyperparameter optimization.



N*l* is the number of nodes in hidden layer *l*, where *l*=1, 2, and 3. lr is the learning rate of Adam algorithm for NN parameter optimization.

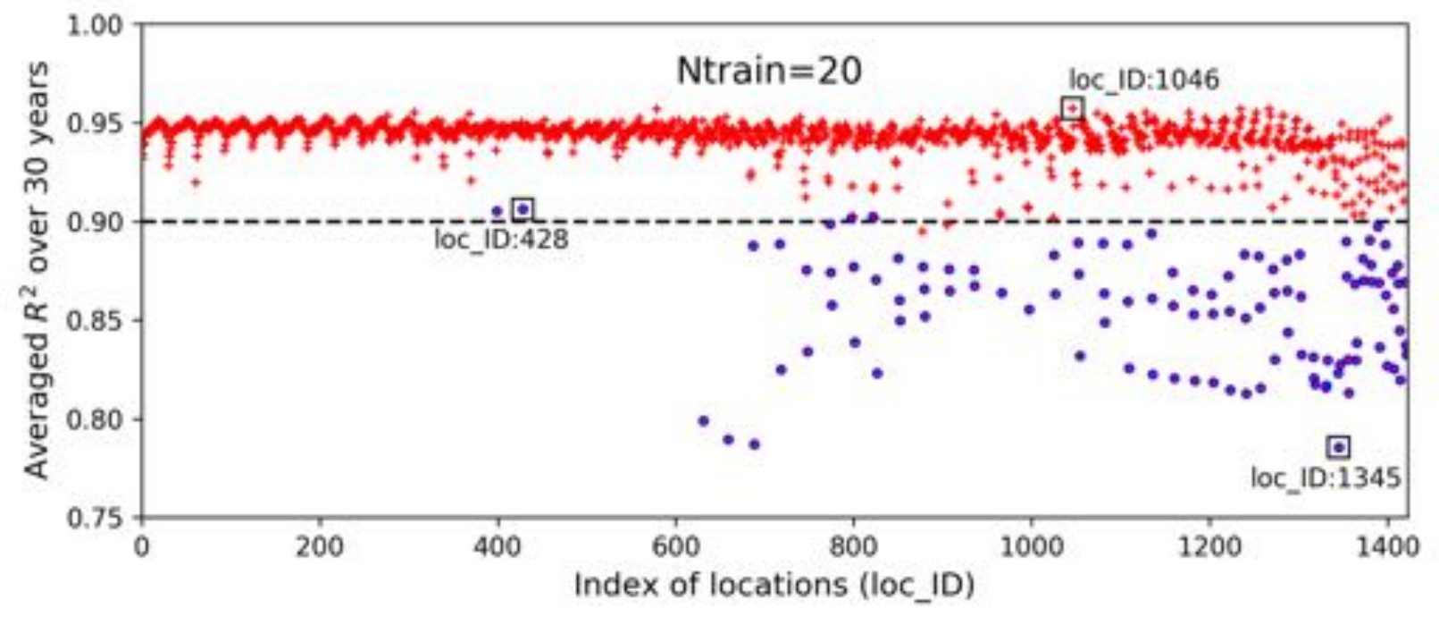
Results

- ❖ Our method can use 20 samples to produce accurate NN predictions, otherwise 200 samples are needed for the similar accuracy.



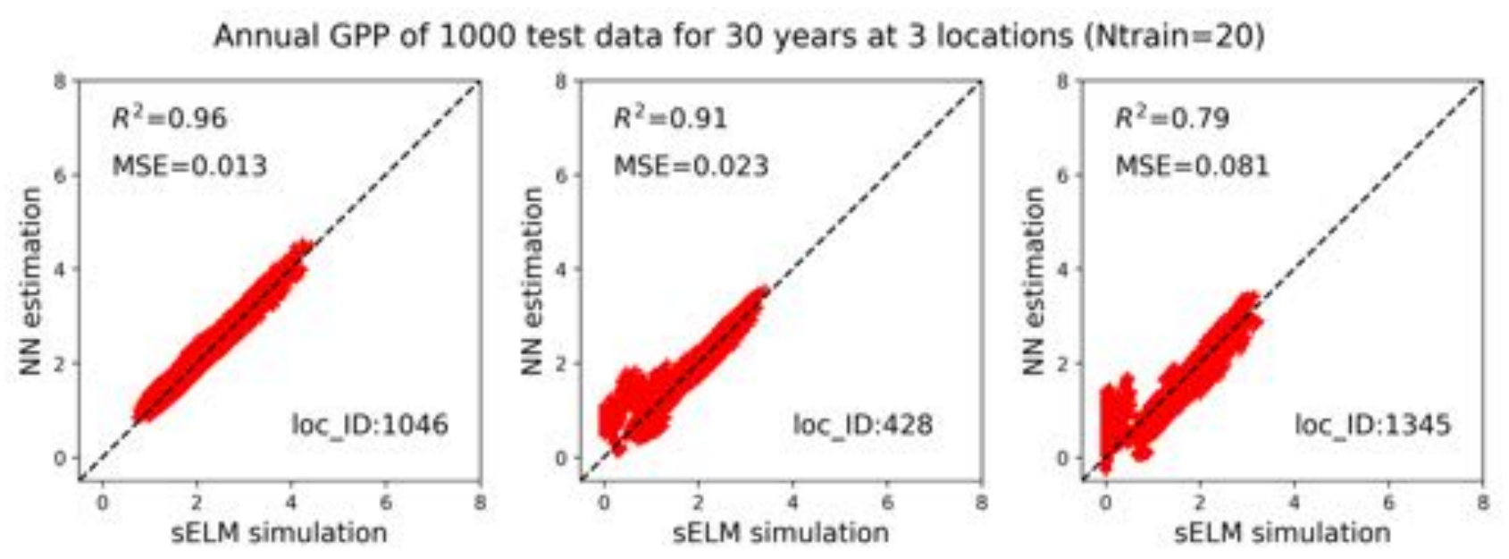
Comparison of NN performance between Case I (building surrogates in reduced dimension using our method) and Case II (building surrogates of all outputs directly). The right y-axis shows training time of Case II. The training time of Case I is 4 seconds.

- ❖ The surrogate accuracy is high expect at locations with 0 GPP simulations.

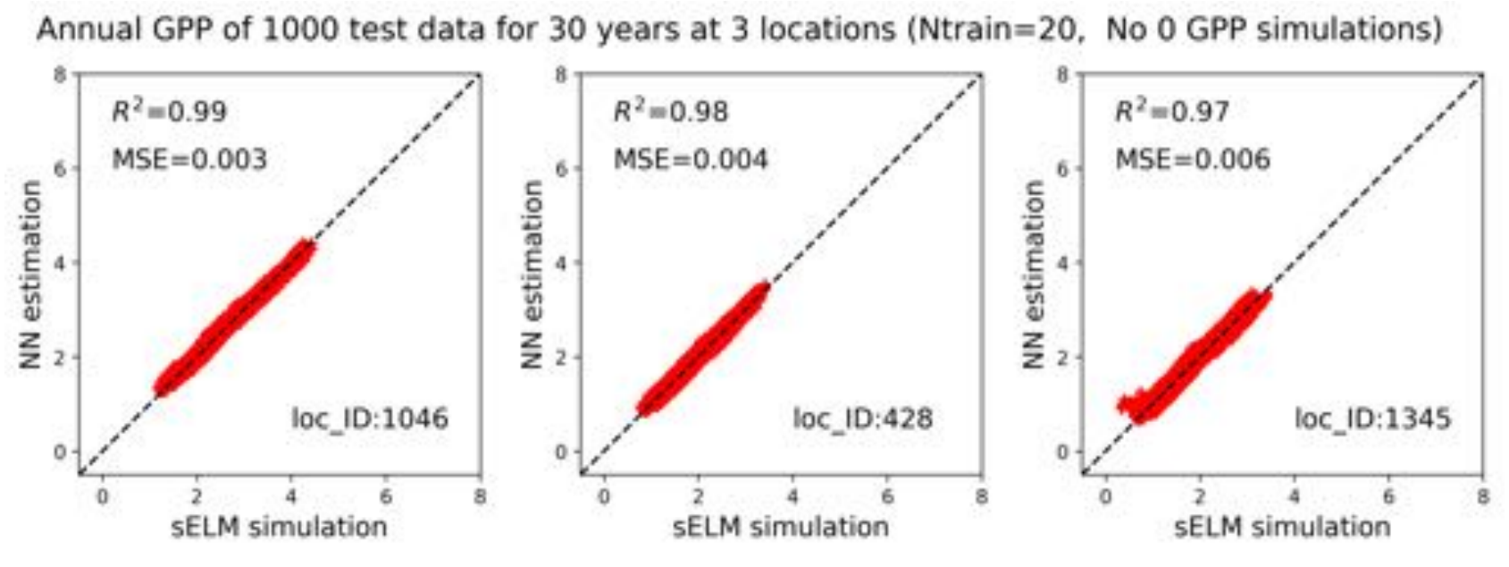


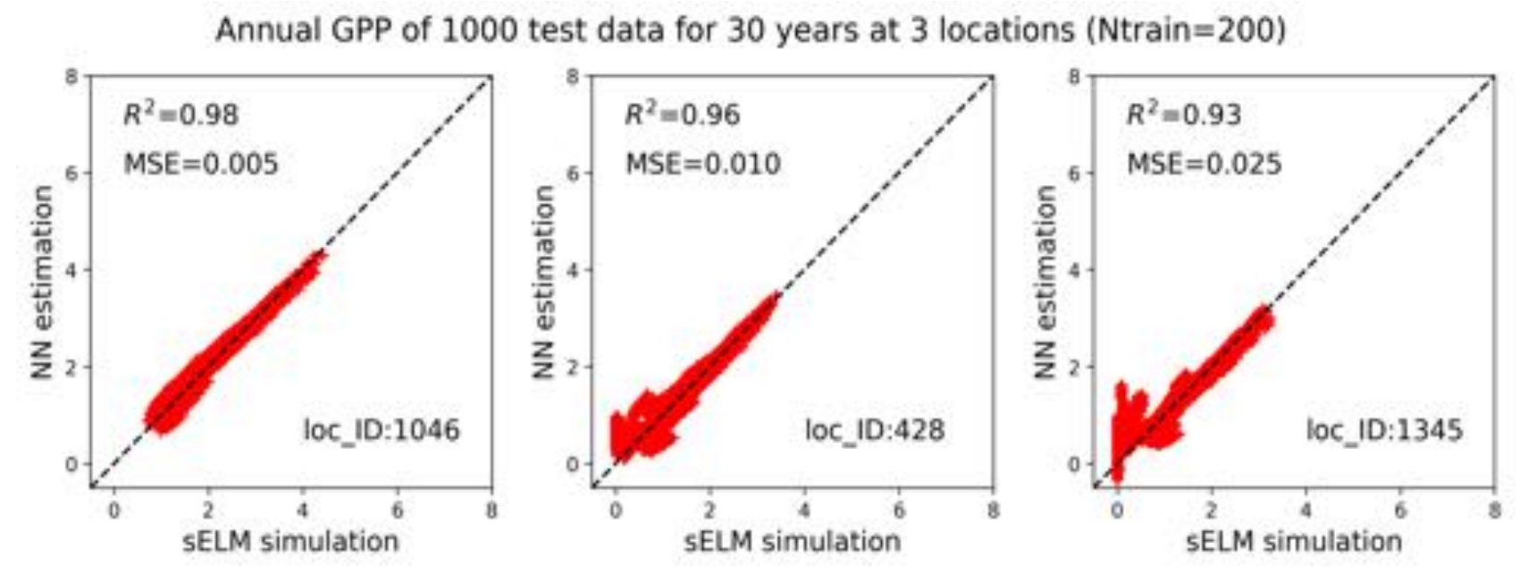
Averaged R^2 scores over 30 years at 1422 locations in evaluating the 1000 test data based on 20 training samples, where the blue circles identify the locations having zero GPP simulations.

- ❖ For locations with low vegetation growth for some parameter samples and large variation in atmospheric drivers that cause discontinuous response surfaces, using physics-informed domain decomposition or the increase of training samples, our method can produce accurate predictions with the R^2 score of 0.97 and 0.96, respectively.



- ❖ For locations with robust vegetation growth across the ensemble, our method can almost perfectly predict the model simulations with R^2 score of 0.96.





Conclusions

- ❖ Based on only 20 model runs, our method can build an accurate surrogate system of 42660 variables; the consistency between surrogate prediction and actual model simulation is 0.93 and the mean squared error is 0.02.
- ❖ This highly-accurate and fast-to-evaluate surrogate system will greatly enhance computational efficiency in data-model integration to improve predictions and advance our understanding of terrestrial ecosystems.

Reference: Lu, D. and Ricciuto, D.: Efficient surrogate modeling methods for large-scale Earth system models based on machine learning techniques, Geosci. Model Dev., accepted, 2019.