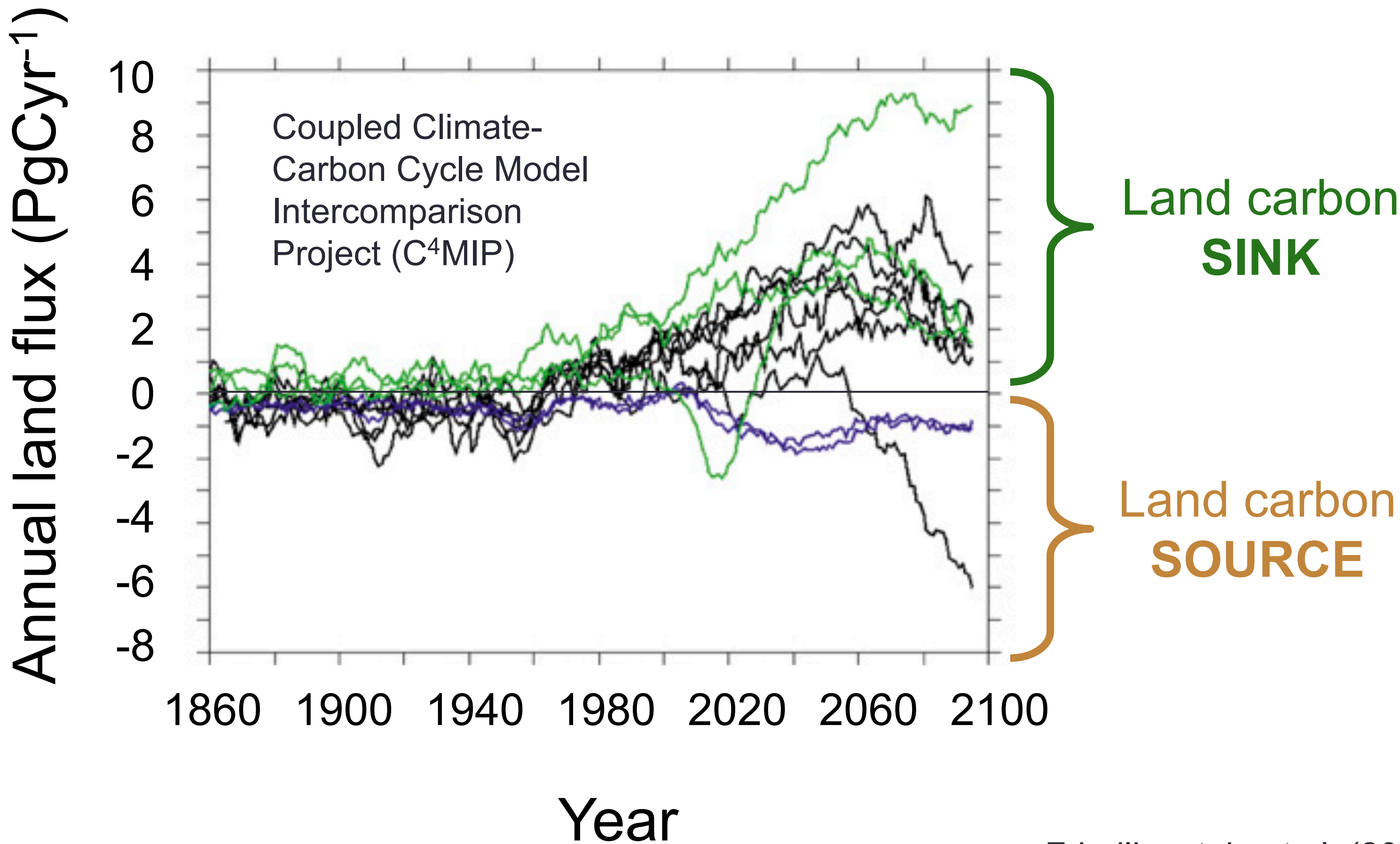
Machine Learning for Parameter Estimation in CLM5

Katie Dagon
NCAR ASP Postdoc

Land Model Working Group Meeting
February 11, 2019

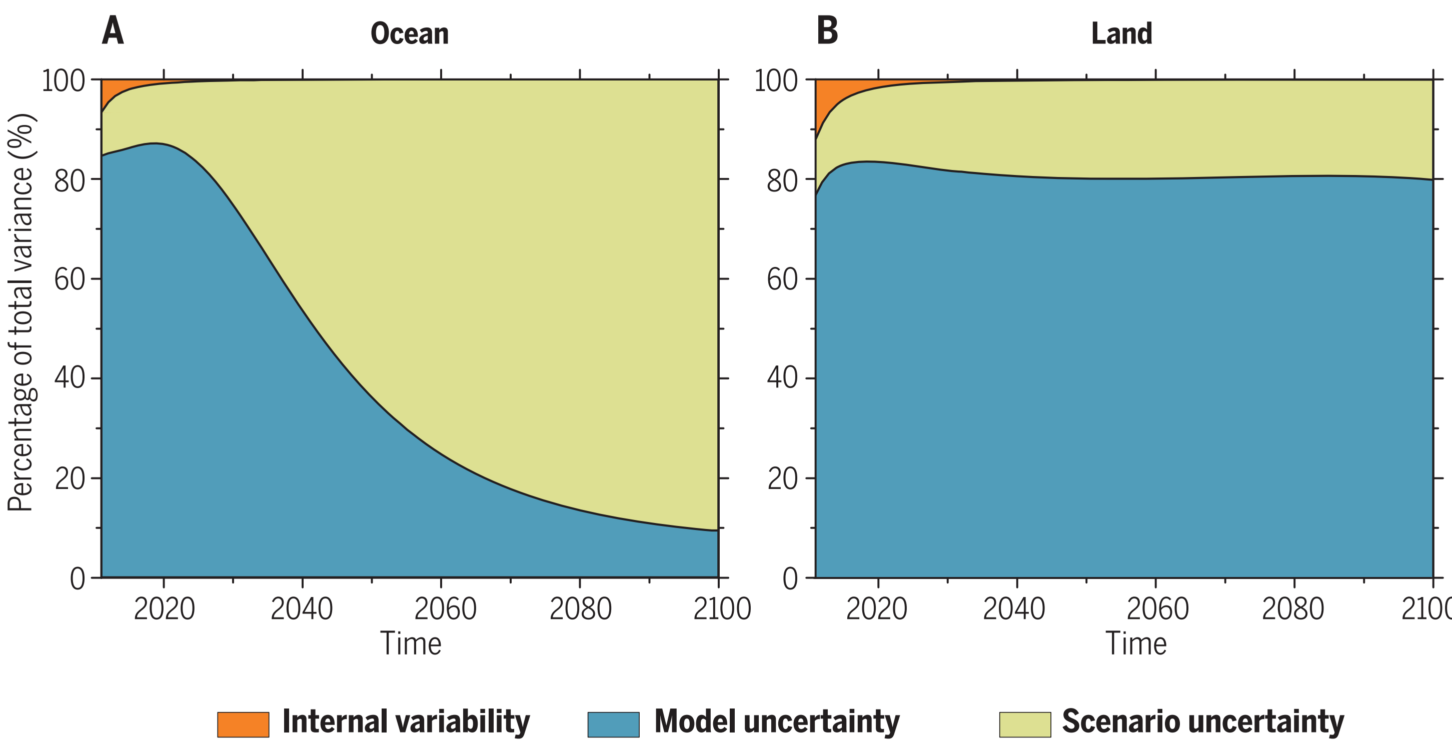
With input and assistance from: Rosie Fisher, Dave Lawrence, Ben Sanderson, and the LMWG

What is driving uncertainty in land surface model projections of climate change?



Friedlingstein et al. (2014)

What are the sources of uncertainty in land surface models?



Uncertainty in model structure and **parameters**

Bonan and Doney (2018), based on Lovenduski and Bonan (2017)

Can we use machine learning to emulate land surface model behavior?

Hand-tuning parameter values takes a long time (many model runs, trial and error).

How can we speed this process up?

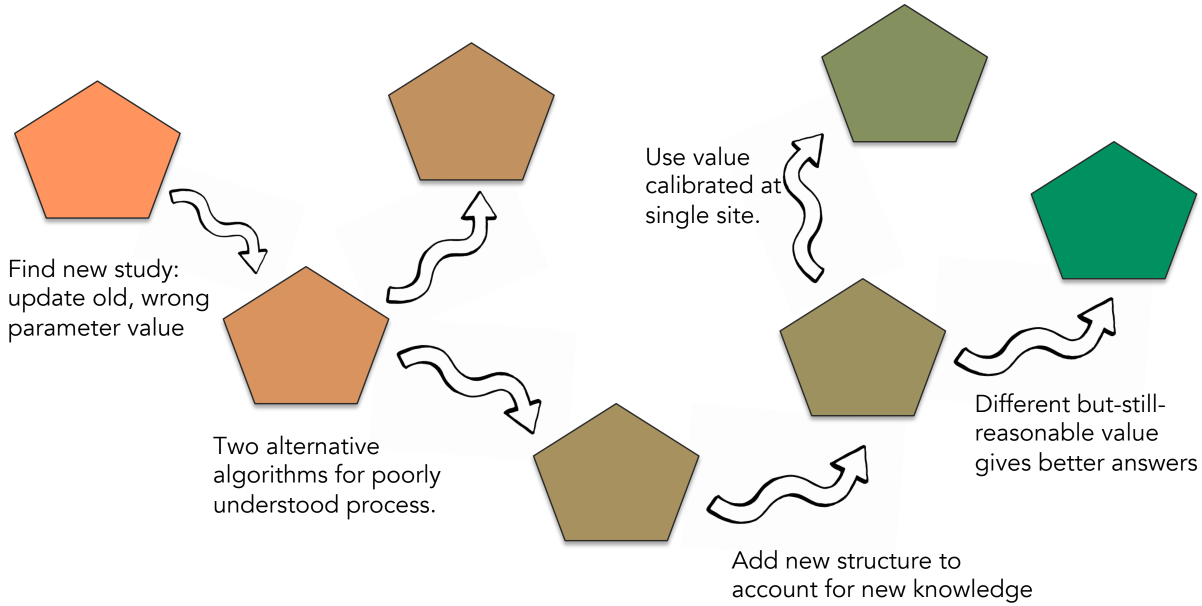


Figure from Rosie Fisher

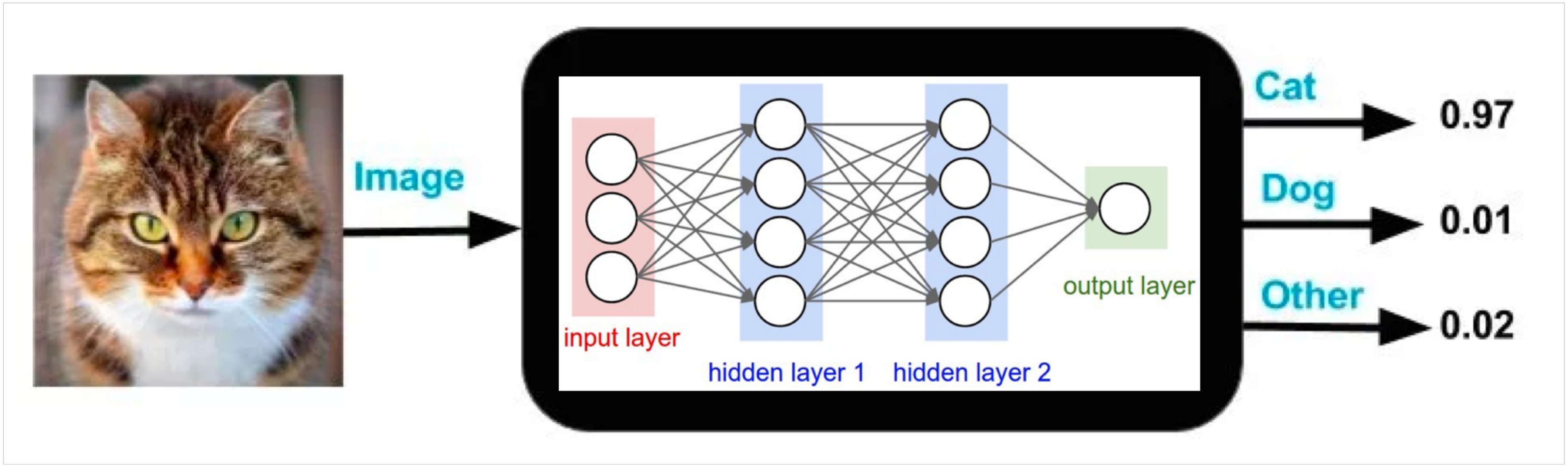
Machine Learning Goals

1. Build and train a series of neural networks to predict CLM output, given parameter values as input.
2. Inflate ensemble size of possible parameter combinations using trained networks.
3. Compare network predictions with observations to estimate best fit parameter values.

Machine learning by neural networks

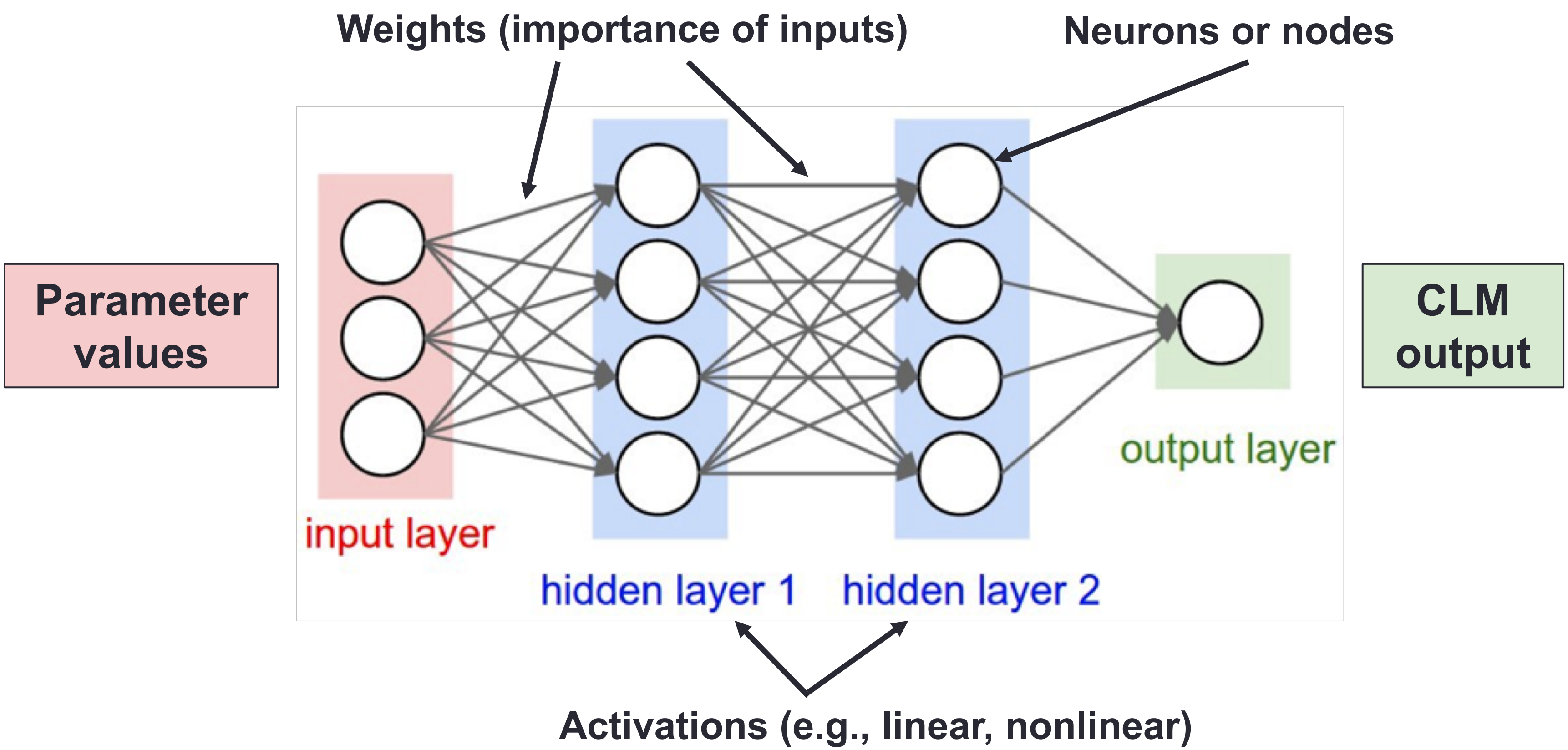
Network image: https://www.learnopencv.com/neural-networks-a-30000-feet-view-for-beginners/

Machine learning by neural networks

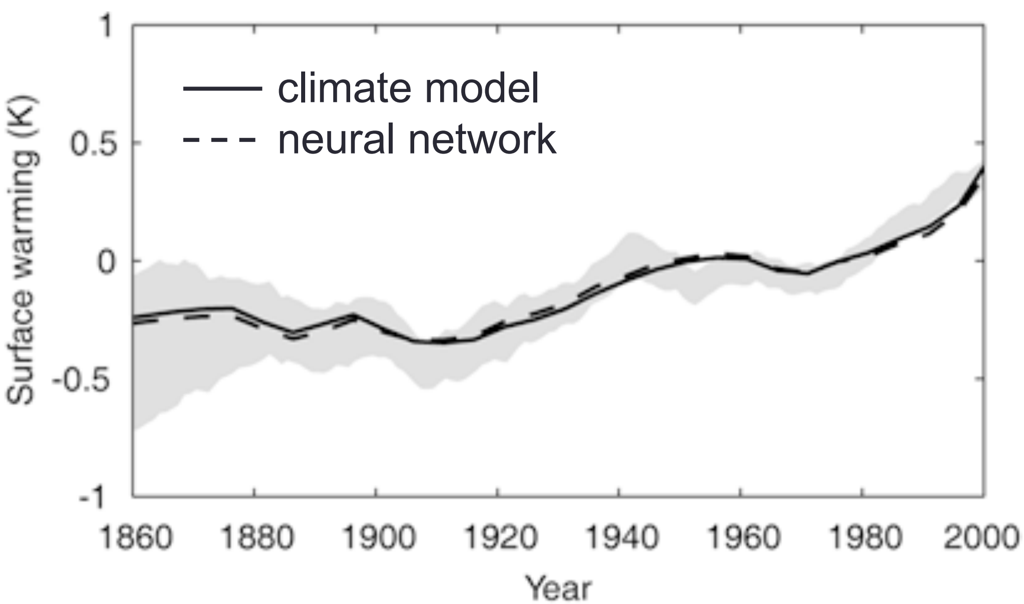


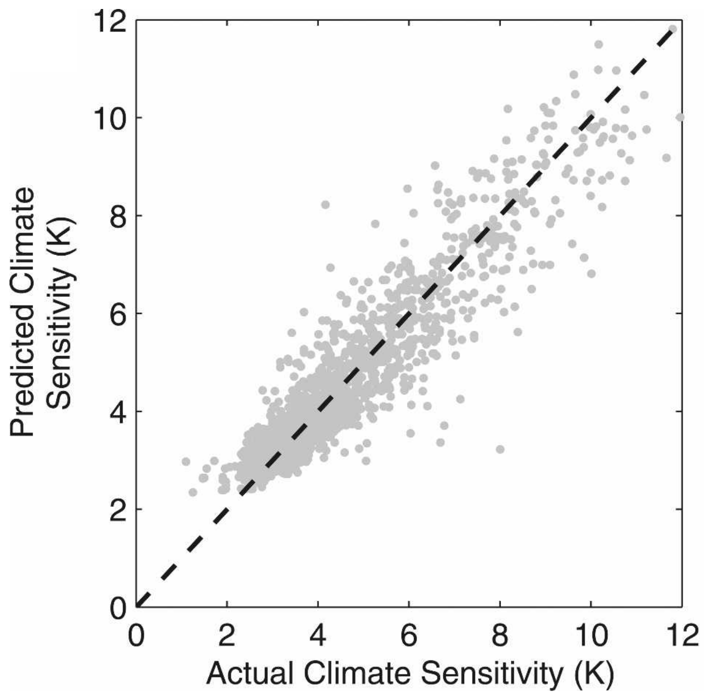
Network image: http://cs231n.github.io/neural-networks-1/

Machine learning by neural networks



Neural networks for climate emulation





Knutti et al. (2003)

Sanderson et al. (2008)

CLM5 Perturbed Parameter Ensemble

- Perturbed parameter ensemble (PPE) using 100 randomly sampled values from uncertainty ranges for 6 biophysical parameters

Name	Biophysical parameter description
medlynslope	Slope of stomatal conductance-photosynthesis relationship
dleaf	Leaf boundary layer resistance parameter
kmax	Plant hydraulic stress parameter
dsl	Soil evaporation parameter
f_over	Surface runoff parameter
baseflow_scalar	Sub-surface runoff parameter

CLM5 Perturbed Parameter Ensemble

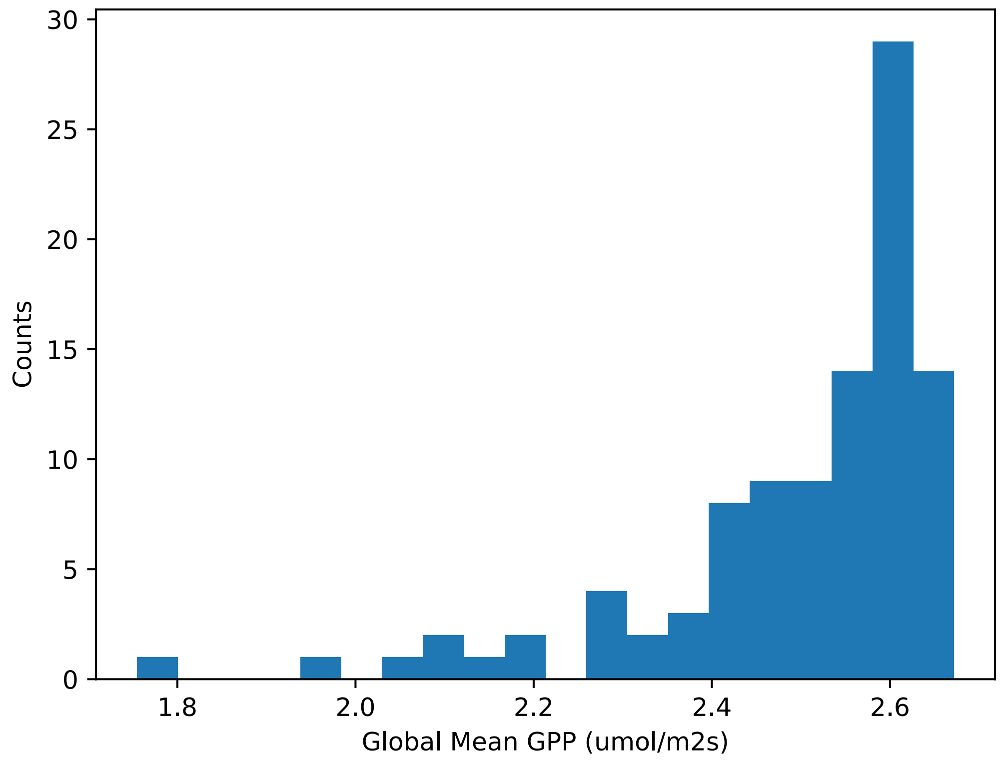
- Perturbed parameter ensemble (PPE) using 100 randomly sampled values from uncertainty ranges for 6 biophysical parameters
- CLM5SP, 4°x5° resolution, 20 year runs (sample last 5 years)
  - Land model forced by GSWP meteorological data

CLM5 Perturbed Parameter Ensemble

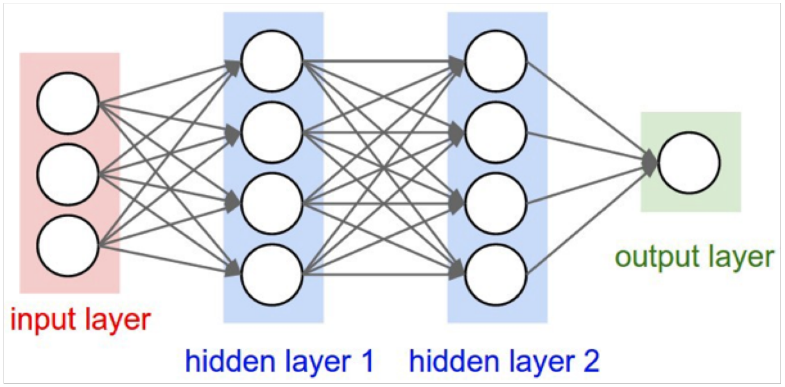
- Perturbed parameter ensemble (PPE) using 100 randomly sampled values from uncertainty ranges for 6 biophysical parameters
- CLM5SP, 4°x5° resolution, 20 year runs (sample last 5 years)
  - Land model forced by GSWP meteorological data
- Begin training neural network on output from 100 PPE simulations

Begin training on simple global mean metric

Distribution of global mean gross primary productivity (GPP, μmol $m^{-2}s^{-1}$)



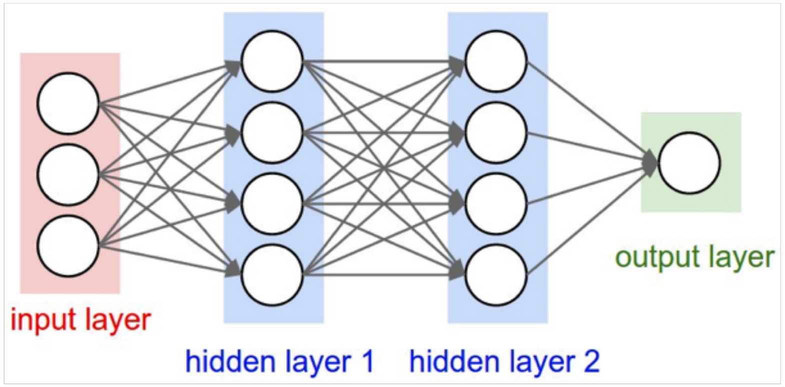
Build and train a neural network to predict land model output based on parameter values



Build and train a neural network to predict land model output based on parameter values

Input (100 parameter sets; 6 parameters)

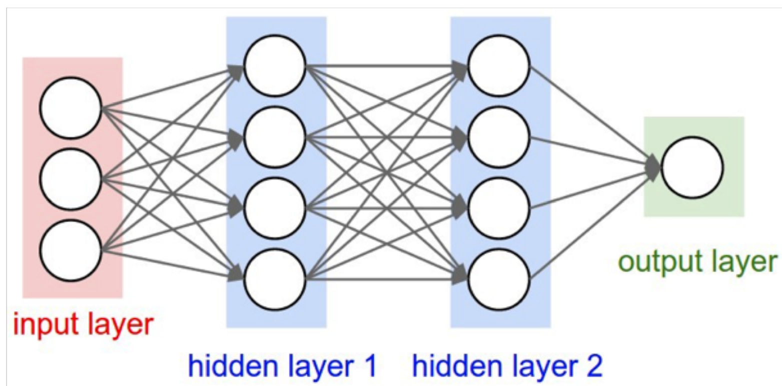
	P1	P2	P3	P4	P5	P6
S1	x1,1	x1,2	x1,3	x1,4	x1,5	x1,6
S2	x2,1	x2,2	x2,3	x2,4	x2,5	x2,6
S3	x3,1	x3,2	x3,3	x3,4	x3,5	x3,6
…	…	…	…	…	…	…
S100	x100,1	x100,2	x100,3	x100,4	x100,5	x100,6



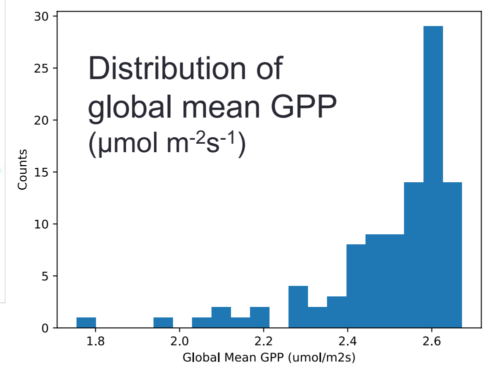
Build and train a neural network to predict land model output based on parameter values

Input (100 parameter sets; 6 parameters)

	P1	P2	P3	P4	P5	P6
S1	x1,1	x1,2	x1,3	x1,4	x1,5	x1,6
S2	x2,1	x2,2	x2,3	x2,4	x2,5	x2,6
S3	x3,1	x3,2	x3,3	x3,4	x3,5	x3,6
…	…	…	…	…	…	…
S100	x100,1	x100,2	x100,3	x100,4	x100,5	x100,6



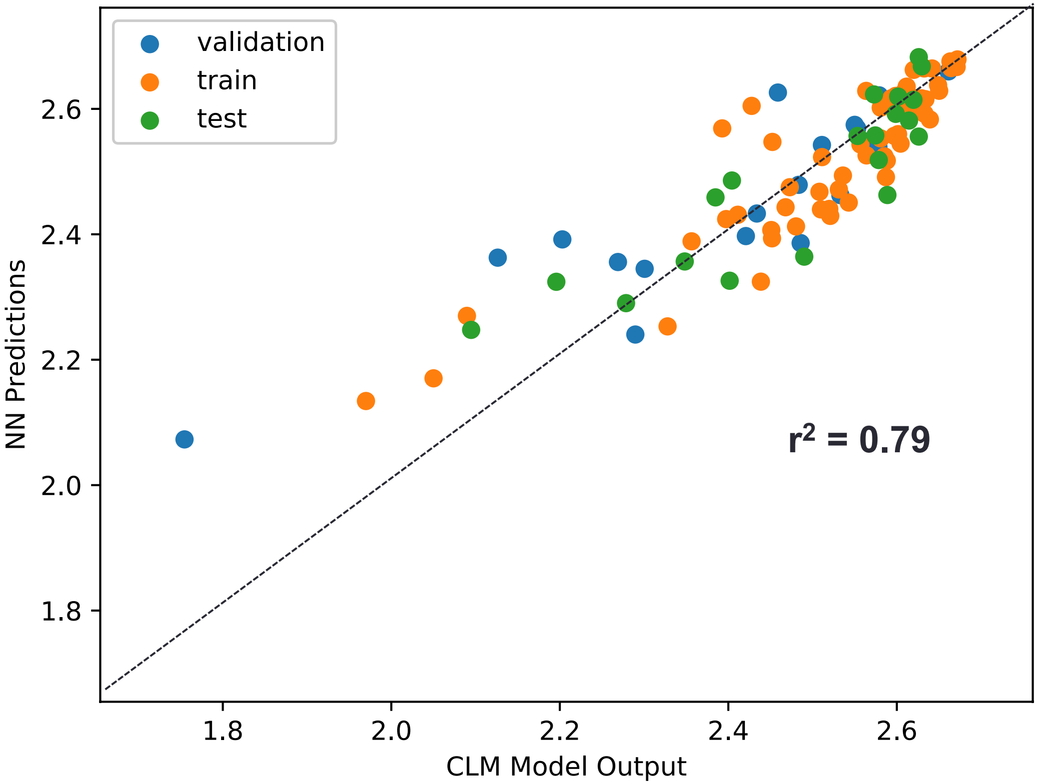
Output (100 CLM simulations)



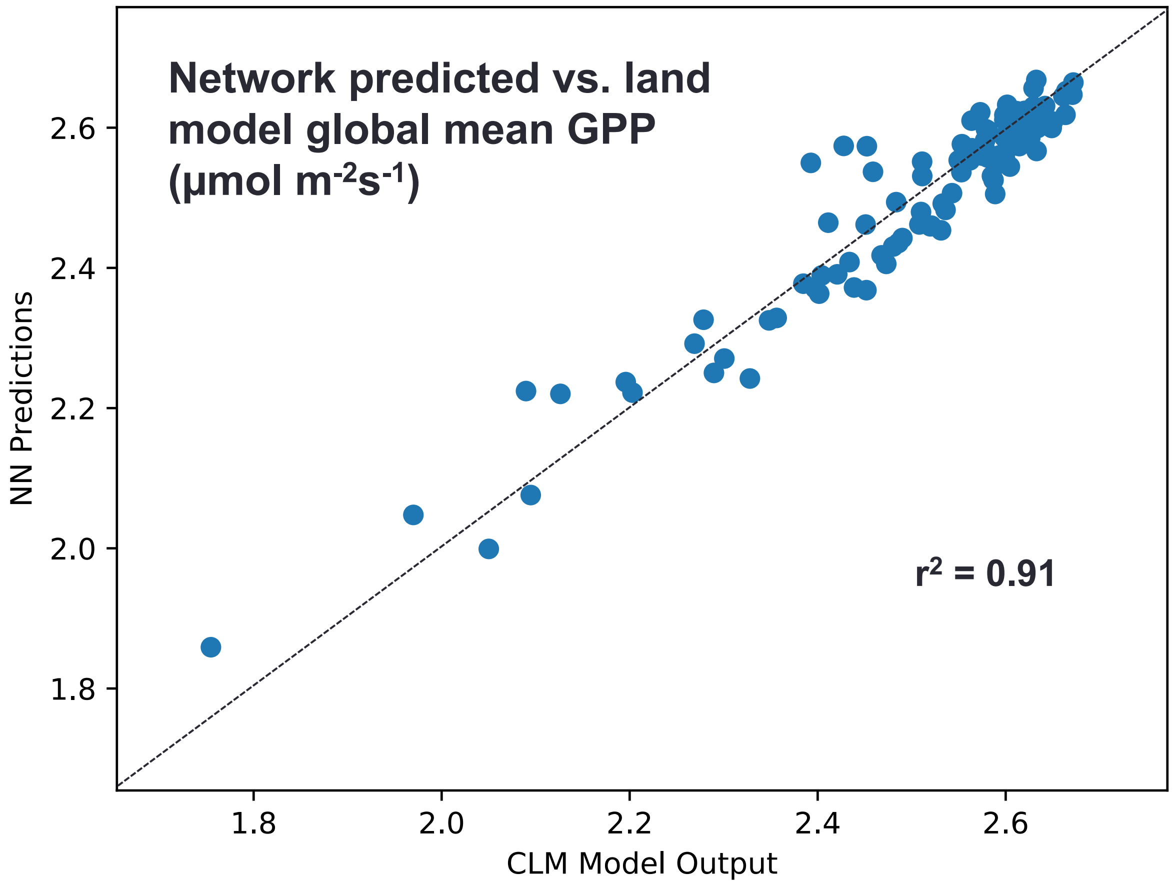
Training Process

- **Subset** the data into training and validation sets
- **Test** different network configurations (neurons, layers, activations)
- **Resample** training data to avoid overfitting
- **Assess performance** based on overall fit

Predicted vs. Actual Global Mean GPP (μmol $m^{-2}s^{-1}$)



Assessing network performance



Climate model emulation

Input (1000 parameter sets; 6 parameters)

	P1	P2	P3	P4	P5	P6
S1	x1,1	x1,2	x1,3	x1,4	x1,5	x1,6
S2	x2,1	x2,2	x2,3	x2,4	x2,5	x2,6
S3	x3,1	x3,2	x3,3	x3,4	x3,5	x3,6
…	…	…	…	…	…	…
…	…	…	…	…	…	…
S1000	x1000,1	x1000,2	x1000,3	x1000,4	x1000,5	x1000,6

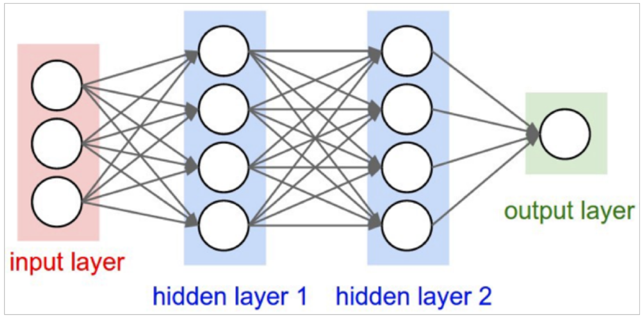
Increase the ensemble size from 100 to 1000 parameter values.

Climate model emulation

Input (1000 parameter sets; 6 parameters)

	P1	P2	P3	P4	P5	P6
S1	x1,1	x1,2	x1,3	x1,4	x1,5	x1,6
S2	x2,1	x2,2	x2,3	x2,4	x2,5	x2,6
S3	x3,1	x3,2	x3,3	x3,4	x3,5	x3,6
…	…	…	…	…	…	…
…	…	…	…	…	…	…
S1000	x1000,1	x1000,2	x1000,3	x1000,4	x1000,5	x1000,6

Run through trained neural network



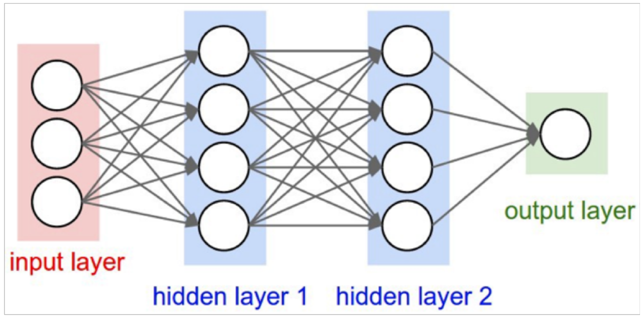
Climate model emulation

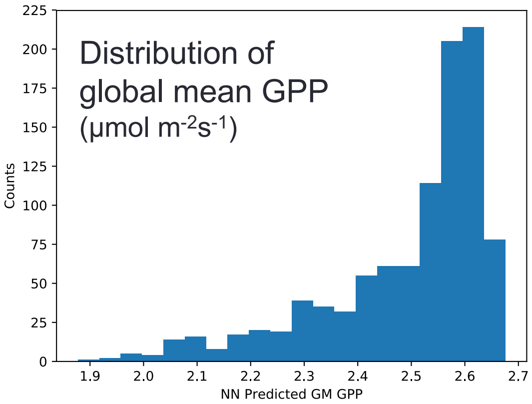
Input (1000 parameter sets; 6 parameters)

Run through trained neural network

Output (1000 neural network predictions)

	P1	P2	P3	P4	P5	P6
S1	x1,1	x1,2	x1,3	x1,4	x1,5	x1,6
S2	x2,1	x2,2	x2,3	x2,4	x2,5	x2,6
S3	x3,1	x3,2	x3,3	x3,4	x3,5	x3,6
…	…	…	…	…	…	…
…	…	…	…	…	…	…
S1000	x1000,1	x1000,2	x1000,3	x1000,4	x1000,5	x1000,6





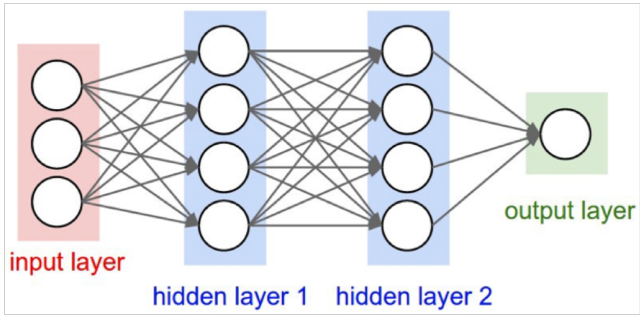
Climate model emulation

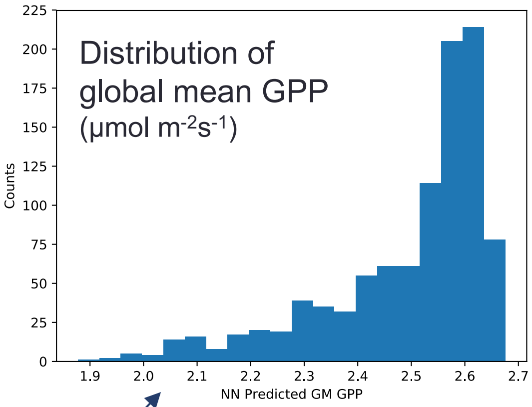
Input (1000 parameter sets; 6 parameters)

Run through trained neural network

Output (1000 neural network predictions)

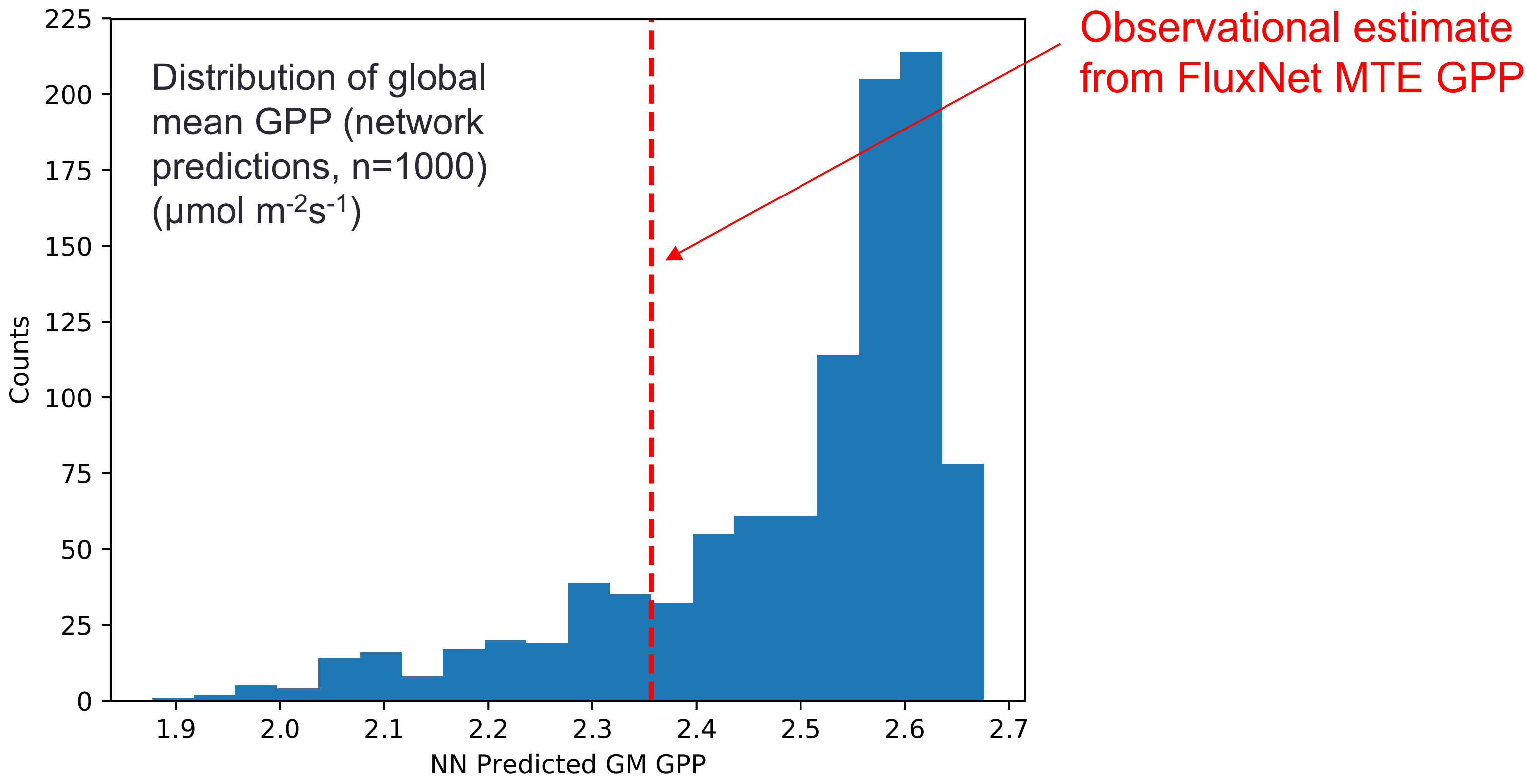
	P1	P2	P3	P4	P5	P6
S1	x1,1	x1,2	x1,3	x1,4	x1,5	x1,6
S2	x2,1	x2,2	x2,3	x2,4	x2,5	x2,6
S3	x3,1	x3,2	x3,3	x3,4	x3,5	x3,6
…	…	…	…	…	…	…
…	…	…	…	…	…	…
S1000	x1000,1	x1000,2	x1000,3	x1000,4	x1000,5	x1000,6



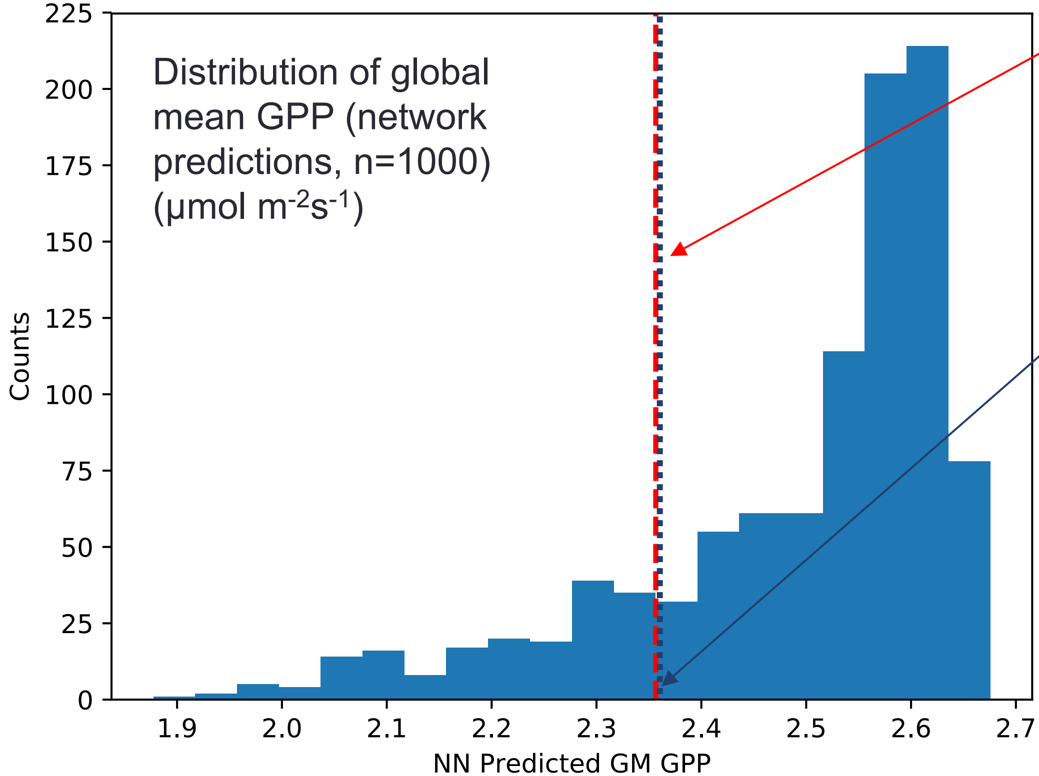


Emulated CLM output! How good are these predictions? Can we use them to constrain parameter values?

Comparing with observations



Parameter estimation

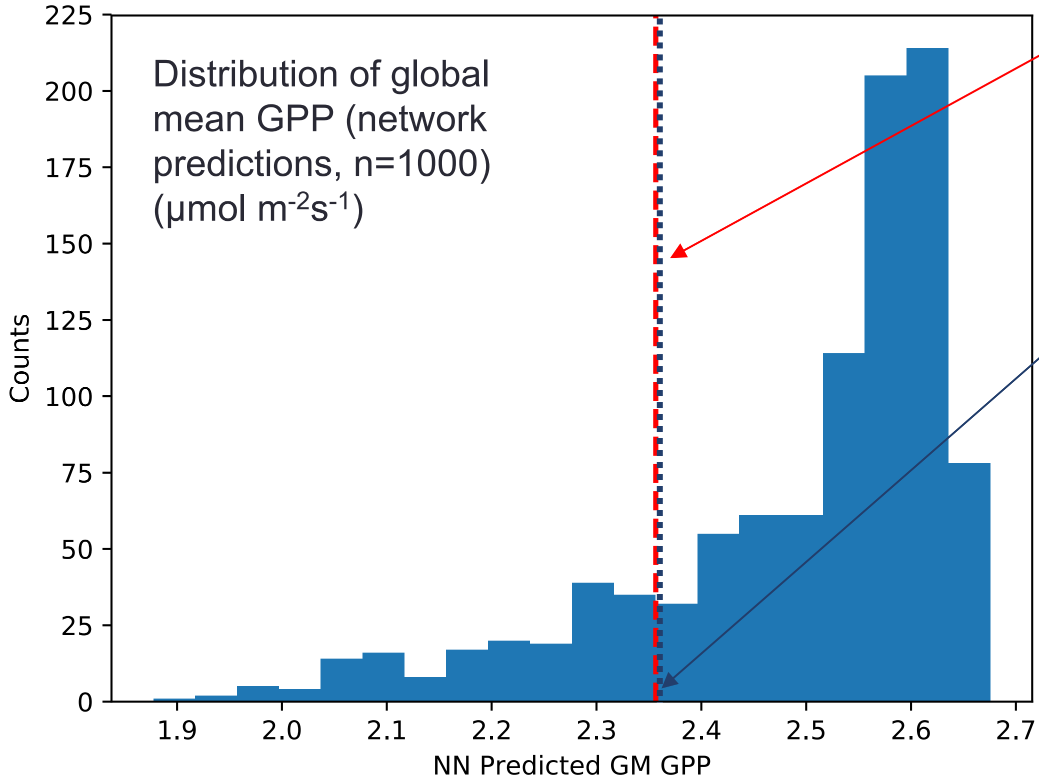


Observational estimate from FluxNet MTE GPP

Isolate the emulator prediction closest to observations.

This gives an estimate of "best fit" parameter values.

Parameter estimation



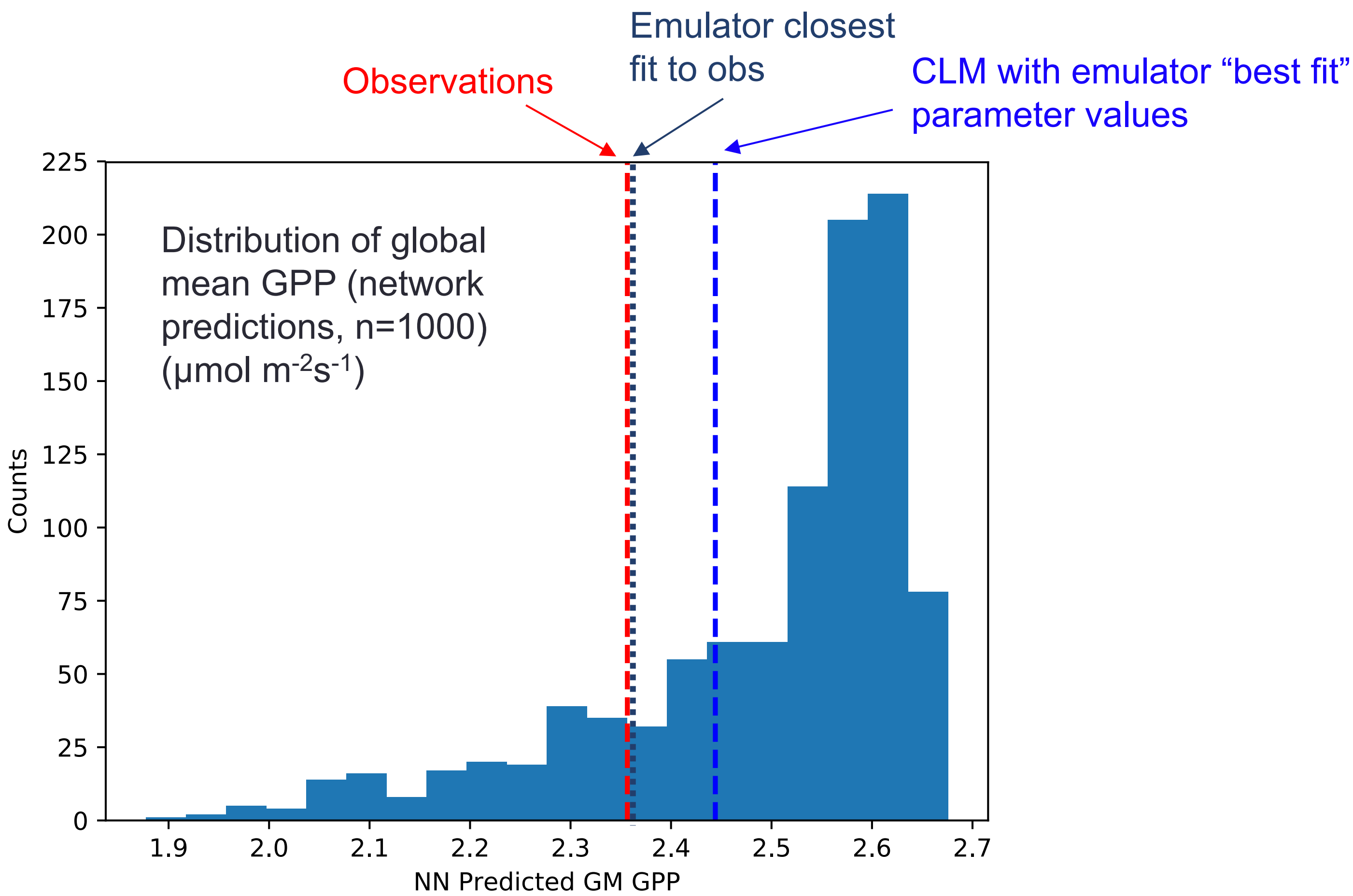
Observational estimate from FluxNet MTE GPP

Isolate the emulator prediction closest to observations.

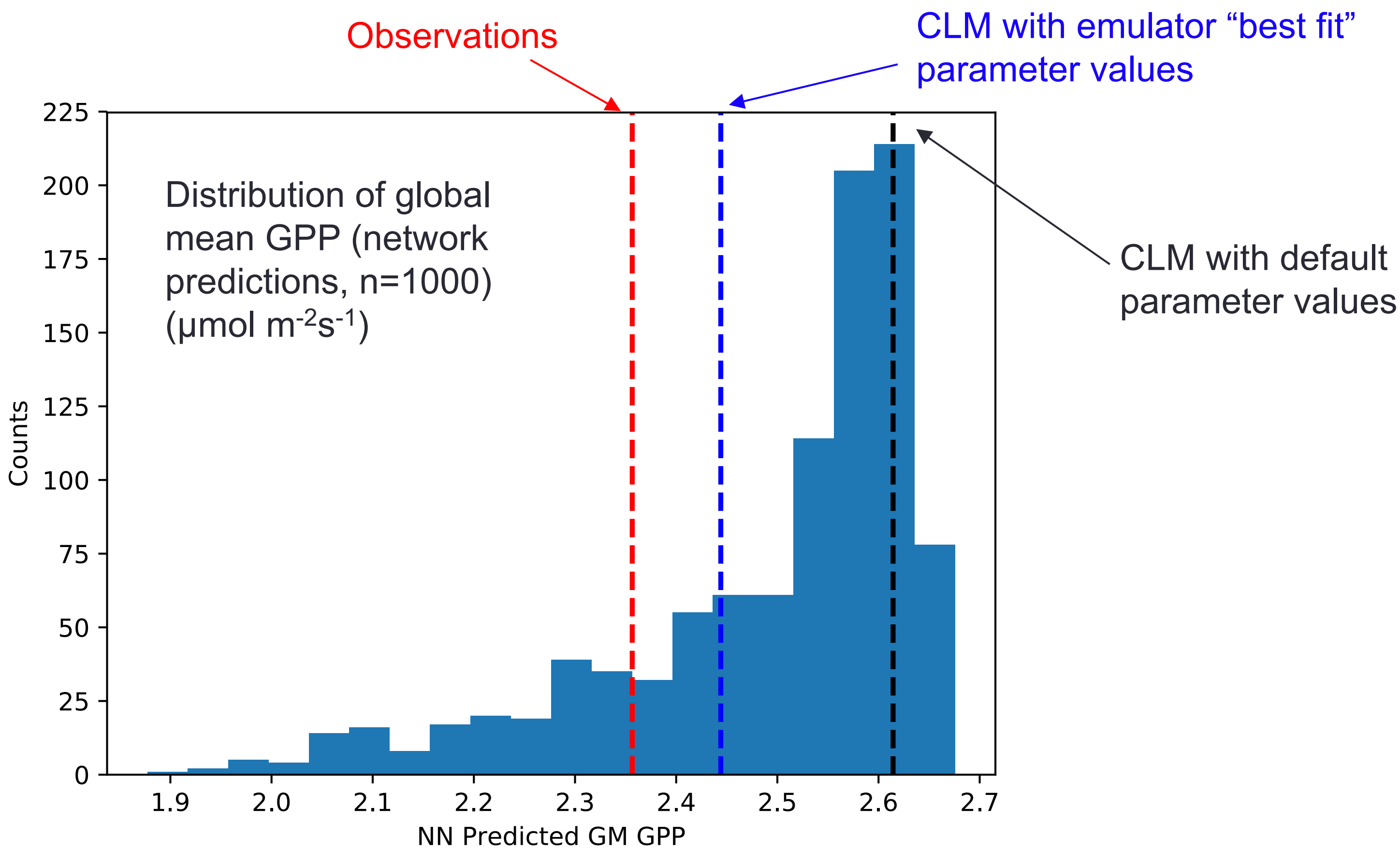
This gives an estimate of “best fit” parameter values.

➢ What happens if we run CLM with these parameter values?

Testing the emulator



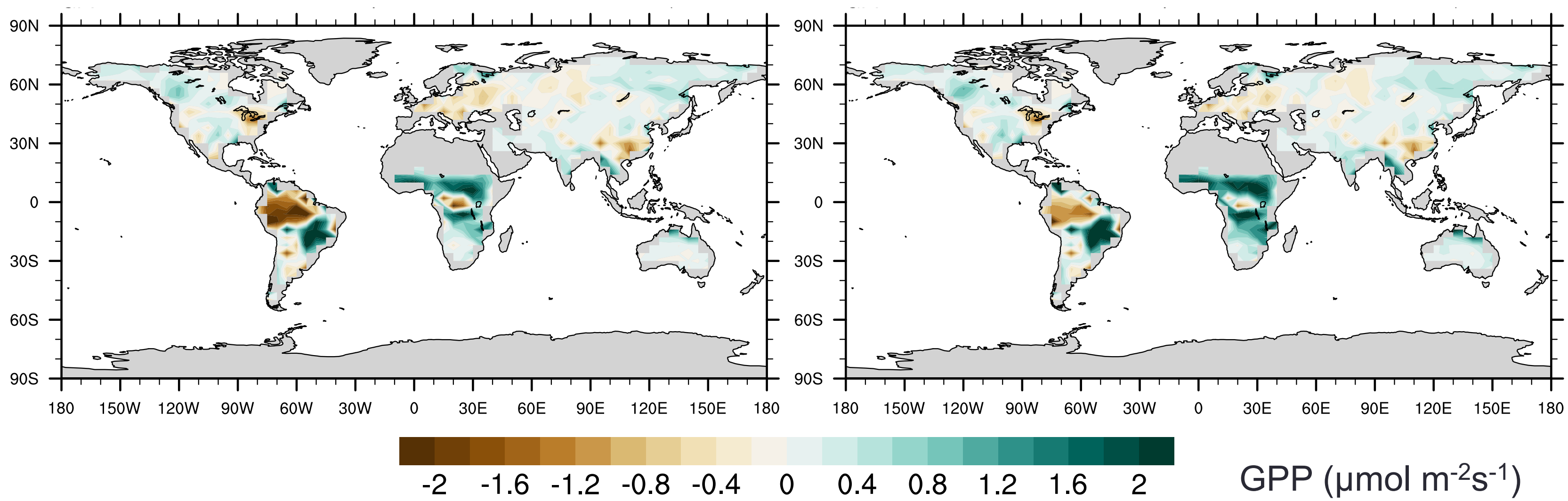
Making progress!



Regional performance needs improvement

CLM with emulator “best fit” parameters
minus Observations

CLM with default parameters
minus Observations

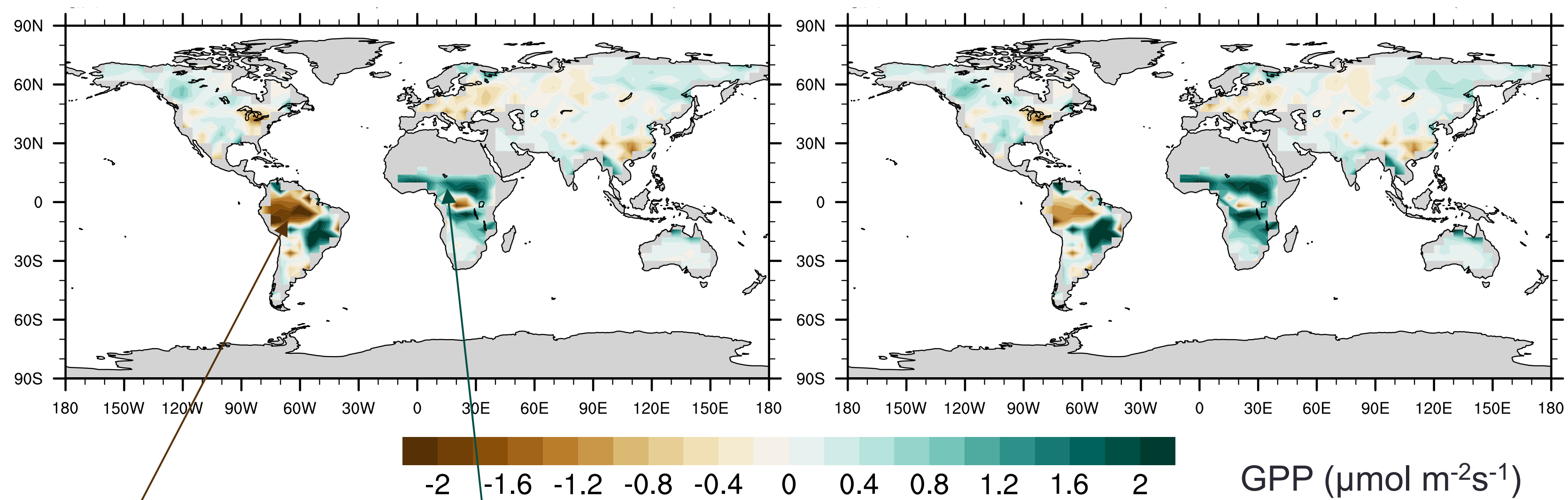


GPP ($\mu mol\ m^{-2}s^{-1}$)

Regional performance needs improvement

CLM with emulator "best fit" parameters *minus* Observations

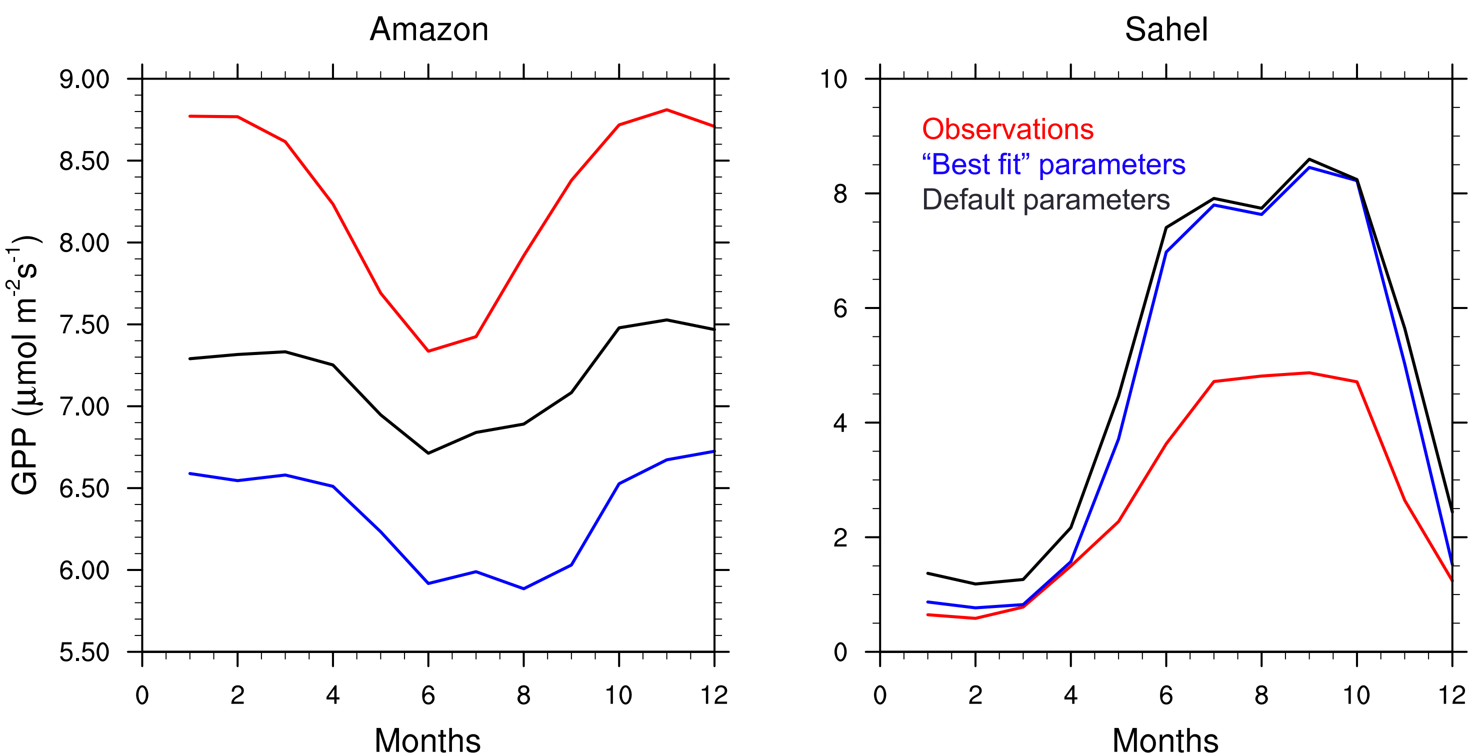
CLM with default parameters *minus* Observations



GPP too low in the Amazon

GPP too high in the Sahel

Regional performance needs improvement



GPP too low in the Amazon; emulator overcompensates

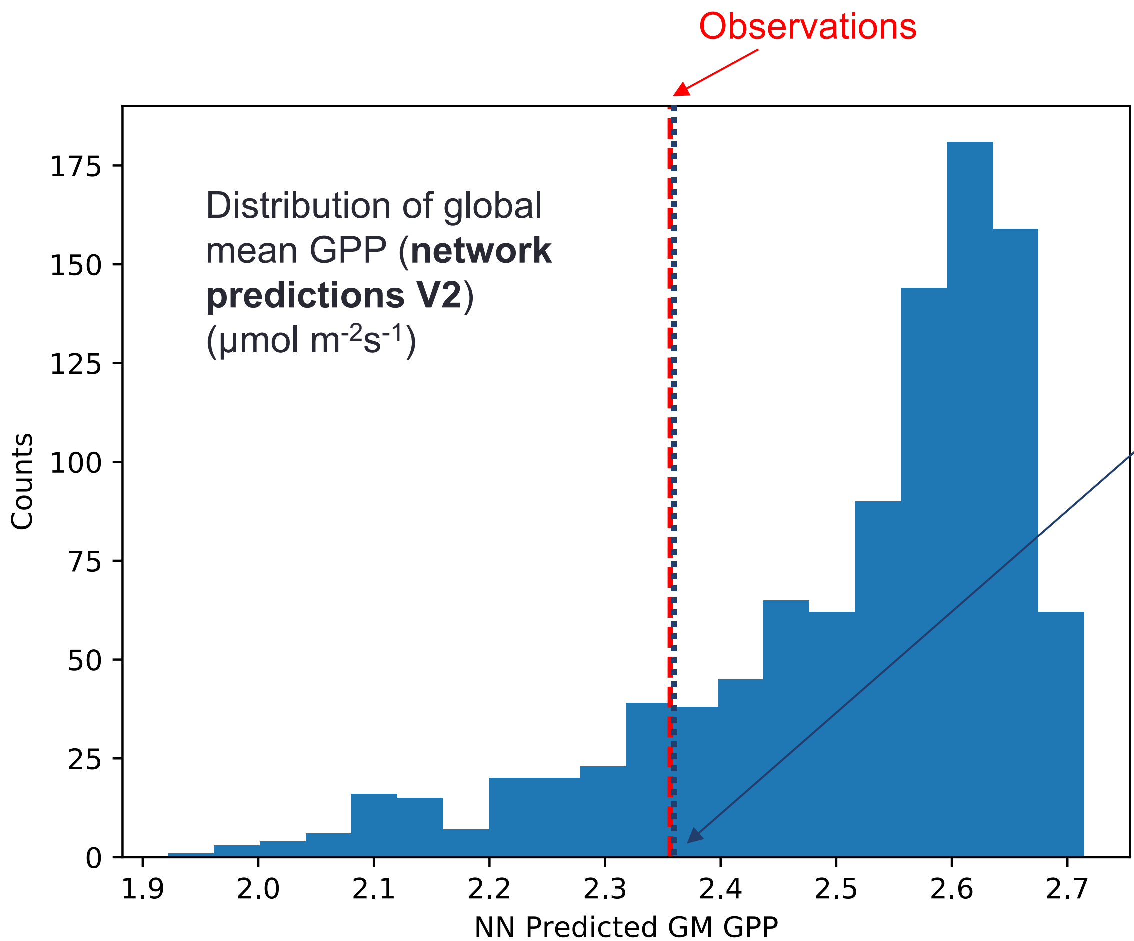
GPP too high in the Sahel; emulator has small effect but in the right direction

What are the parameters doing?

Parameter	**Scaling factor**	**Difference from default value**
Stomatal conductance slope	0.94	All PFTs higher than default (avg 56% increase)
Leaf resistance	0.86	Generally higher than default (avg 68% increase)
Plant hydraulic stress	0.32	32% decrease
Soil evaporation	0.55	3% decrease
Surface runoff	0.53	429% increase
Sub-surface runoff	0.06	575% increase

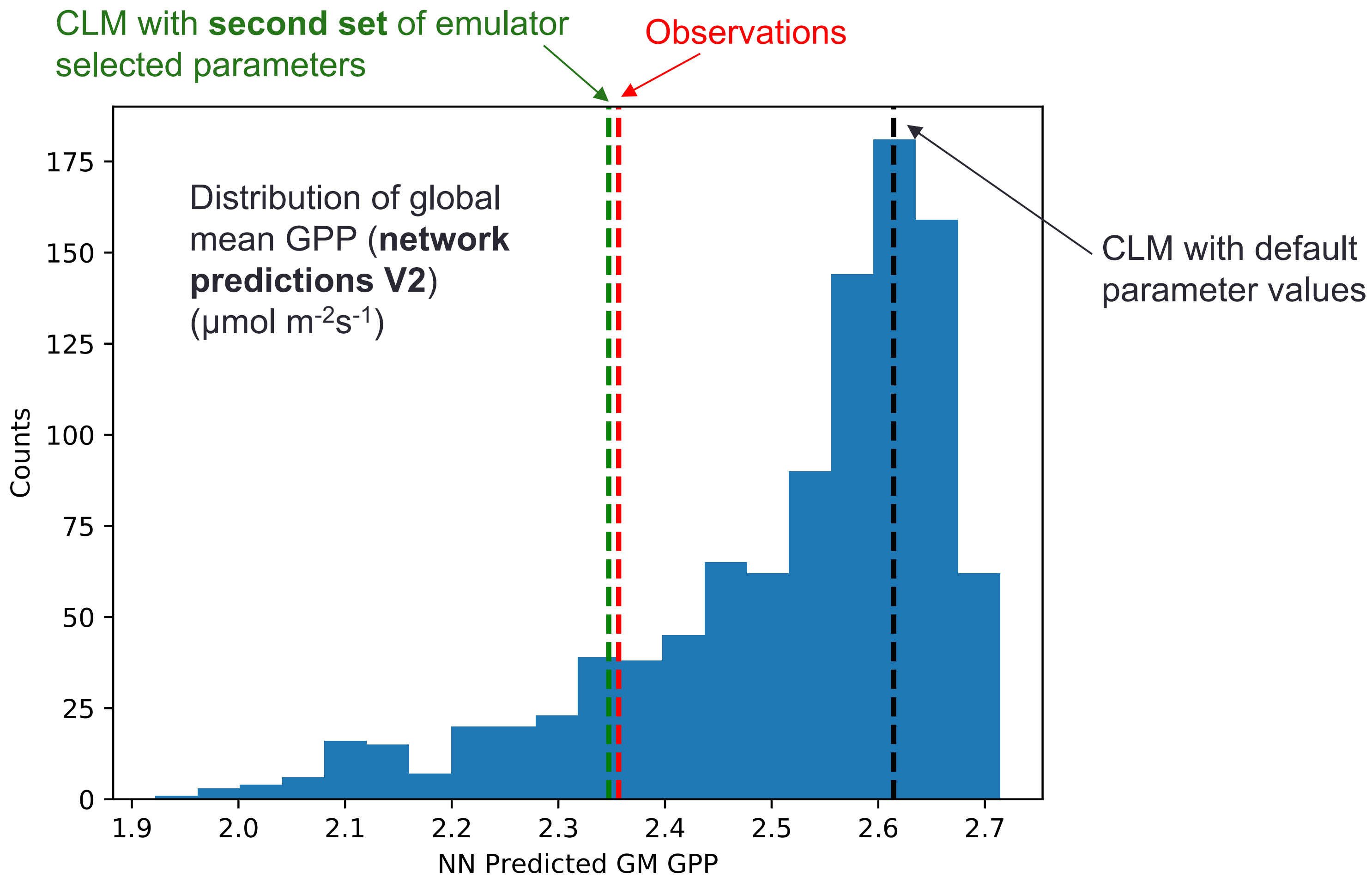
Default value is not necessarily centered!

Try a different set of parameters



Isolate the emulator prediction closest to observations **based on this distribution**.

Try a different set of parameters



Try a different set of parameters

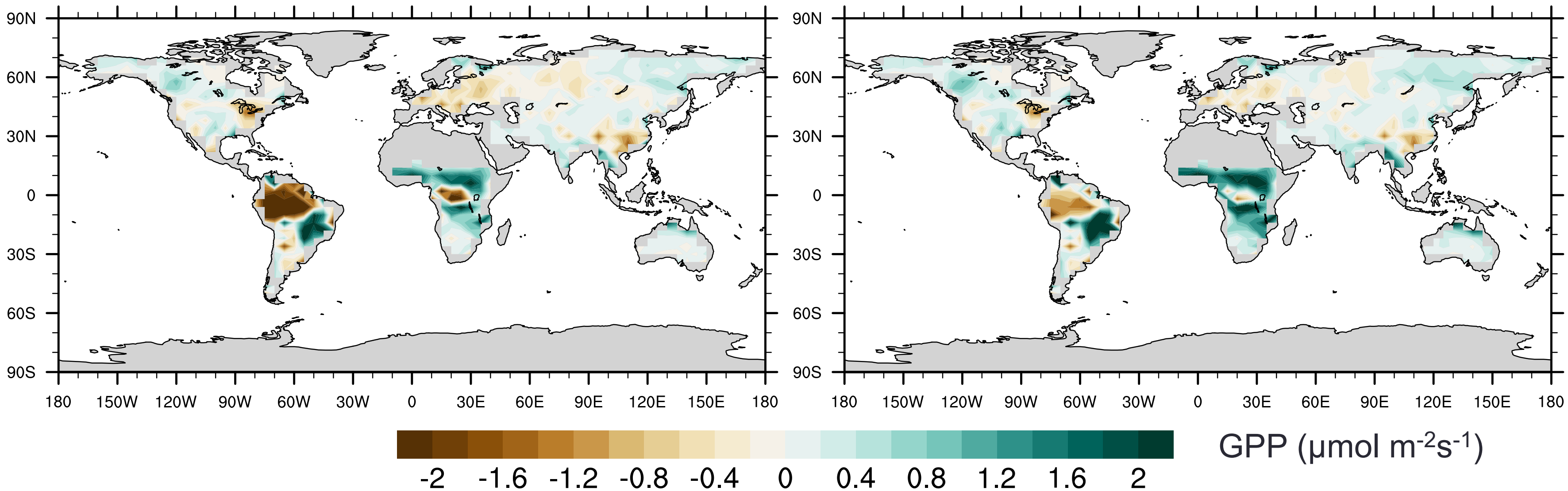
Parameter	Scaling factor (Set 1)	Scaling factor (Set 2)
Stomatal conductance slope	0.94	0.75
Leaf resistance	0.86	0.05
Plant hydraulic stress	0.32	0.18
Soil evaporation	0.55	0.67
Surface runoff	0.53	0.19
Sub-surface runoff	0.06	0.19

Second parameter set is sampling from very different parts of the uncertainty ranges.

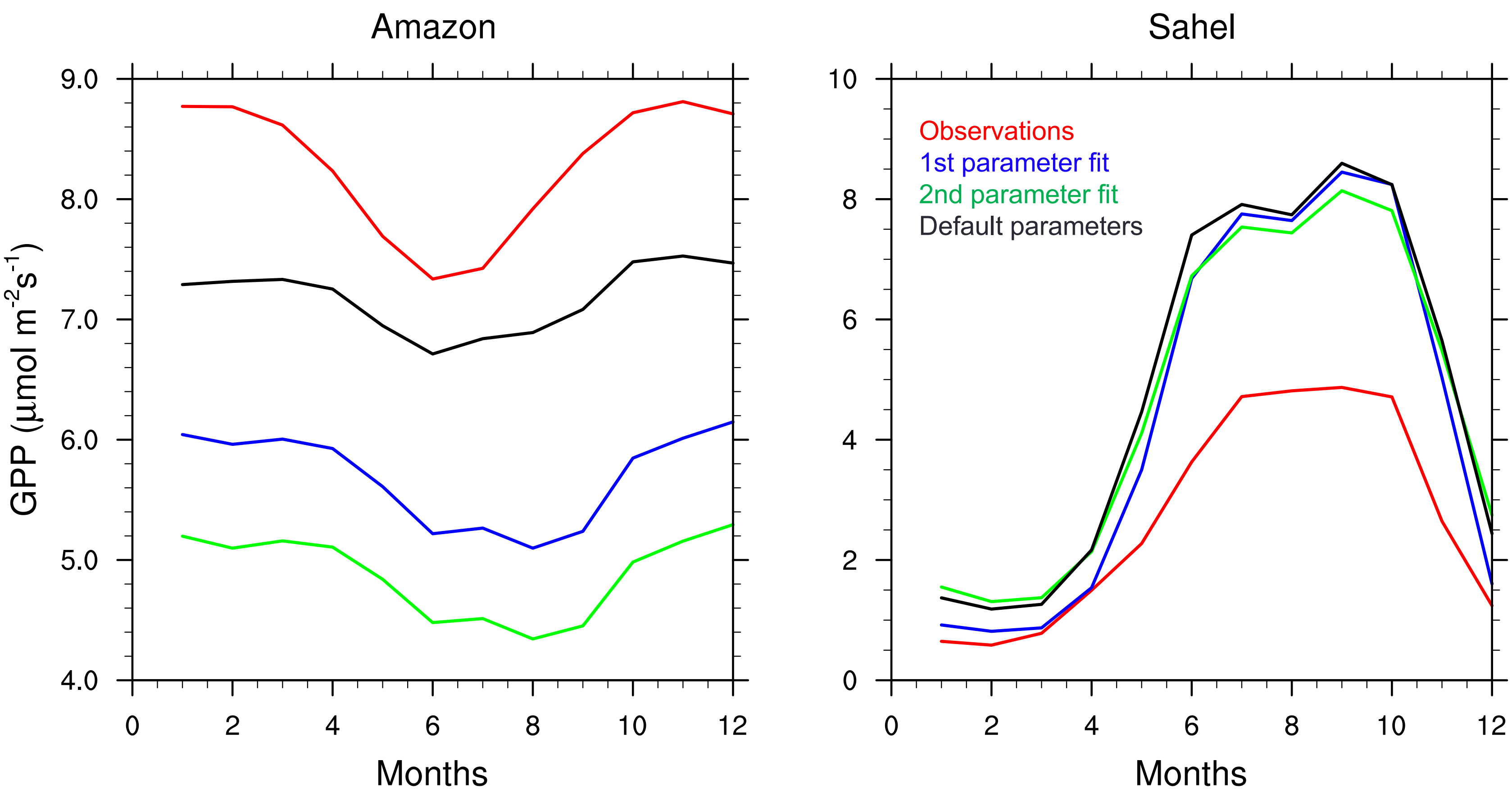
Check in on the regional effect

CLM with emulator-selected parameters *minus* Observations

CLM with default parameters *minus* Observations

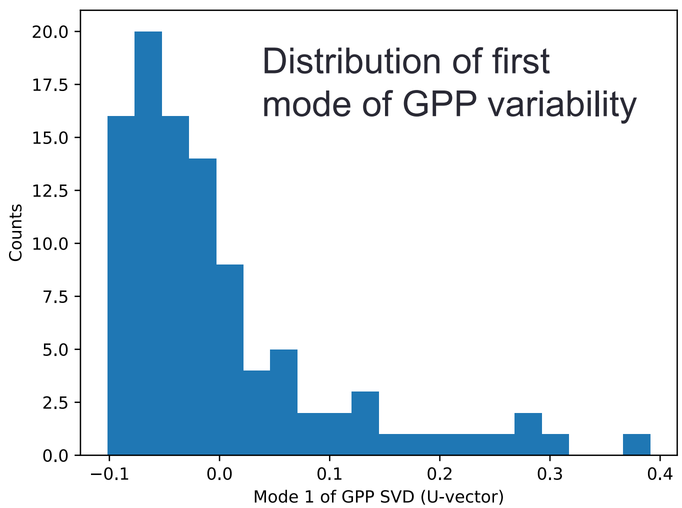


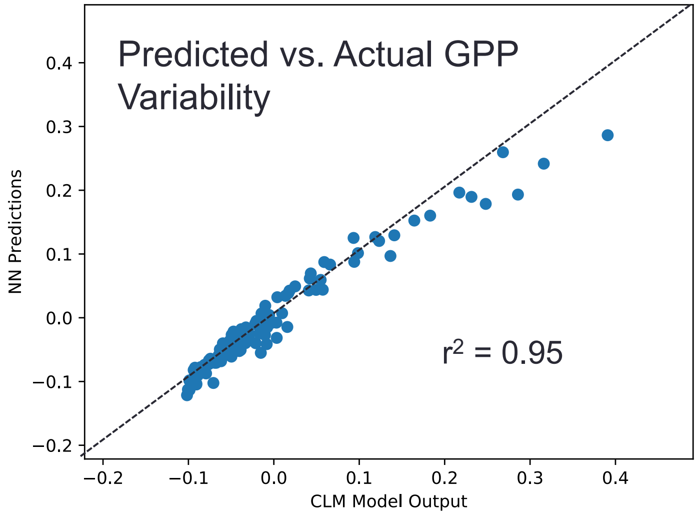
Check in on the regional effect



Next steps

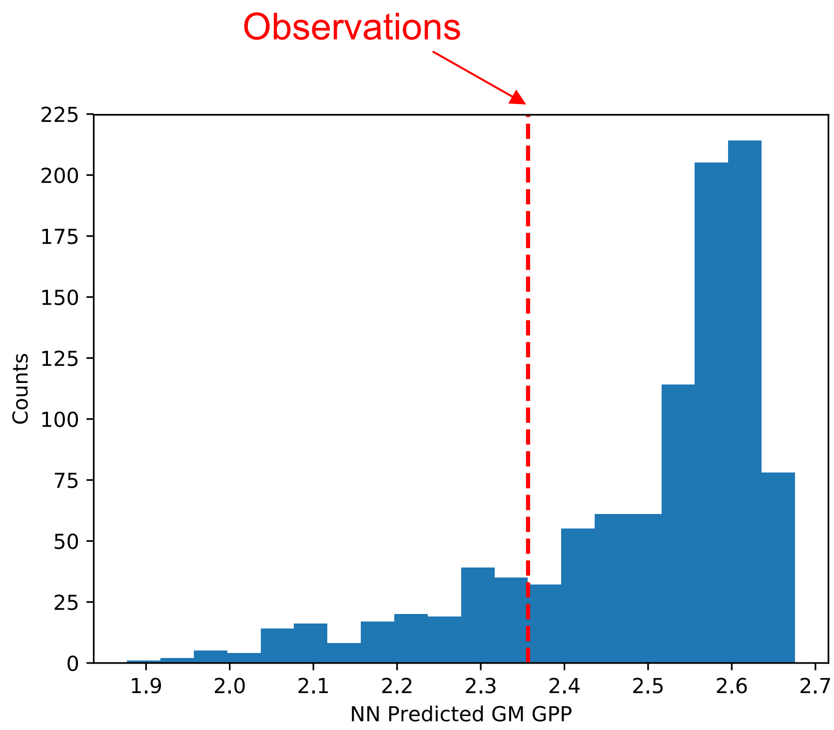
- Assess ***multiple metrics*** (e.g., mean and ***spatial variability***)





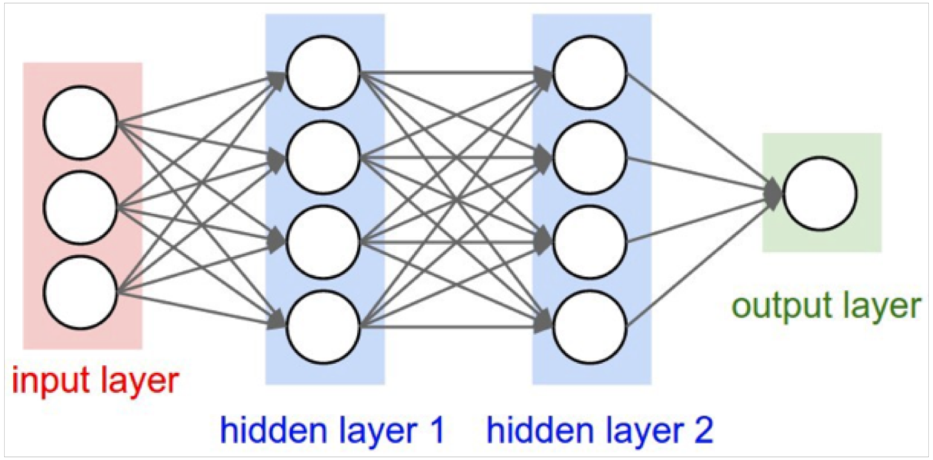
Next steps

- Assess multiple metrics (e.g., mean and spatial variability)
- Use multiple observational datasets to span ***observational uncertainty***



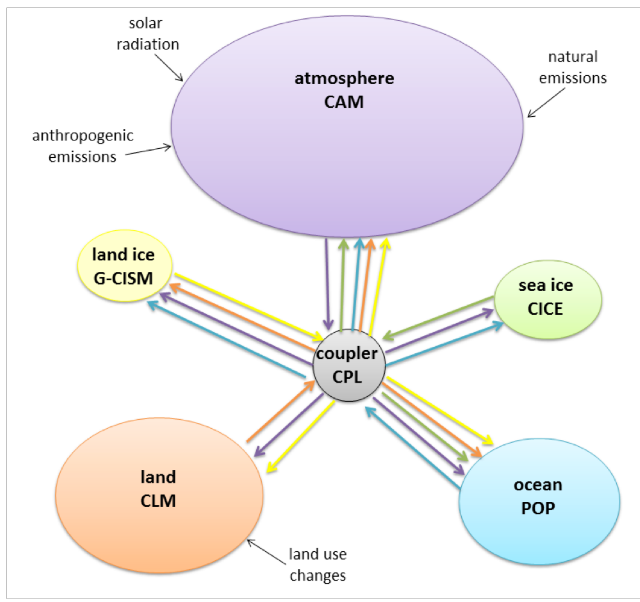
Next steps

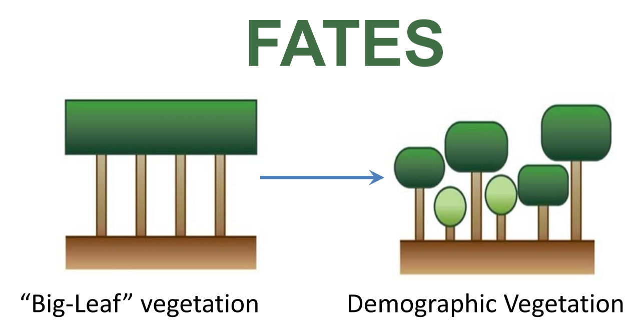
- Assess multiple metrics (e.g., mean and spatial variability)
- Use multiple observational datasets to span observational uncertainty
- Use multiple neural network configurations to span ***prediction uncertainty***



Next steps

- Assess multiple metrics (e.g., mean and spatial variability)
- Use multiple observational datasets to span observational uncertainty
- Use multiple neural network configurations to span prediction uncertainty
- Apply uncertainty framework to ***different models/model configurations***





Summary

- We can reduce uncertainty in land surface models by **studying parameters**.
- Machine learning can help in **climate model emulation**.
- CLM emulator used to **estimate parameter values** by comparing with observations.

Questions?

Thanks!

Contact: *kdagon@ucar.edu*