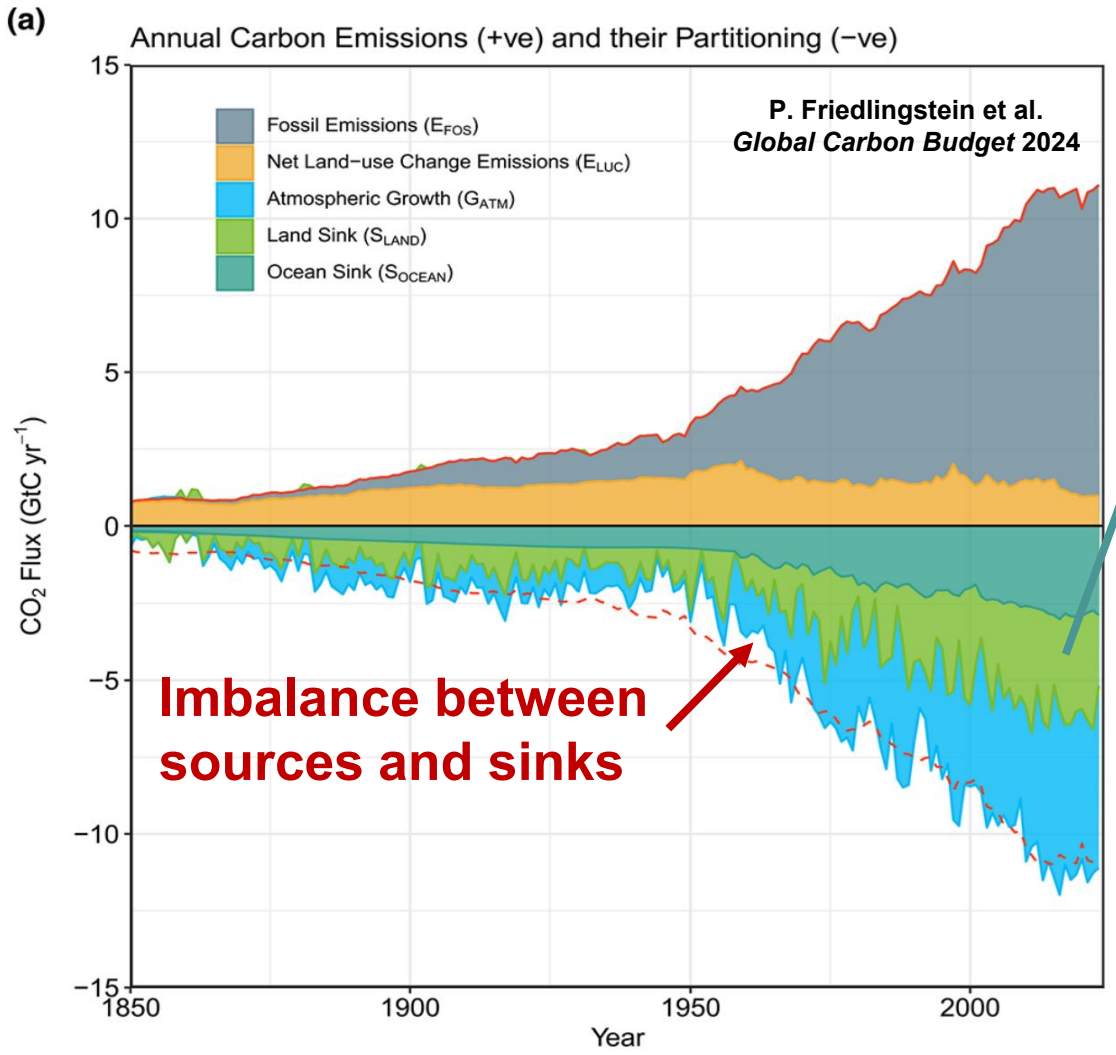
Artificial Intelligence to Advance Modeling and Understanding of Soil Organic Carbon

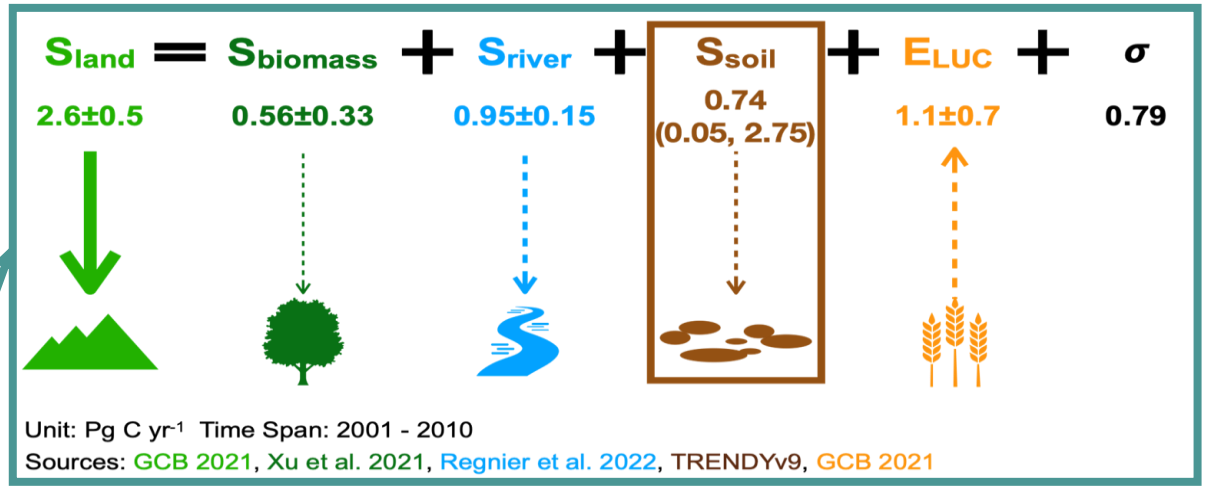
Haodi Xu[1], Joshua Fan[2], Feng Tao[3], Md Nasim[2], Carla P. Gomes[2], Yiqi Luo[1]

Cornell University, Department of [1]Soil & Crop Science, [2]Computer Science, [3]Ecology & Evolutionary Biology

Cornell University

Soil is Critical in Understanding Global Carbon Cycle

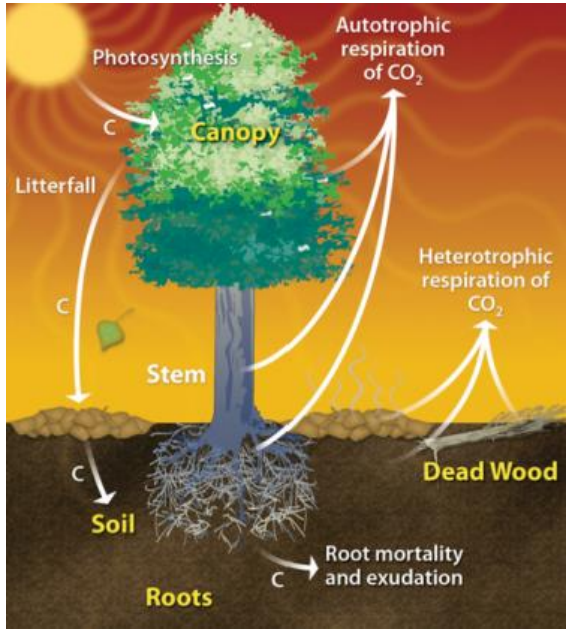


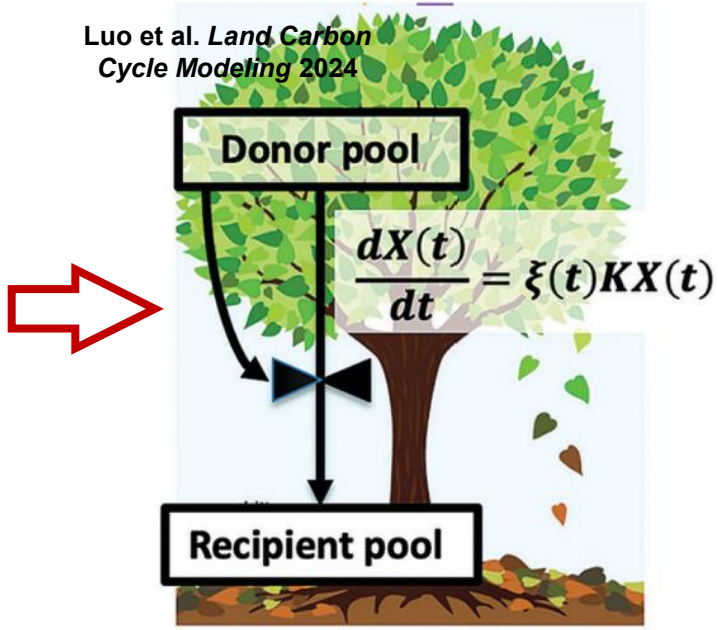


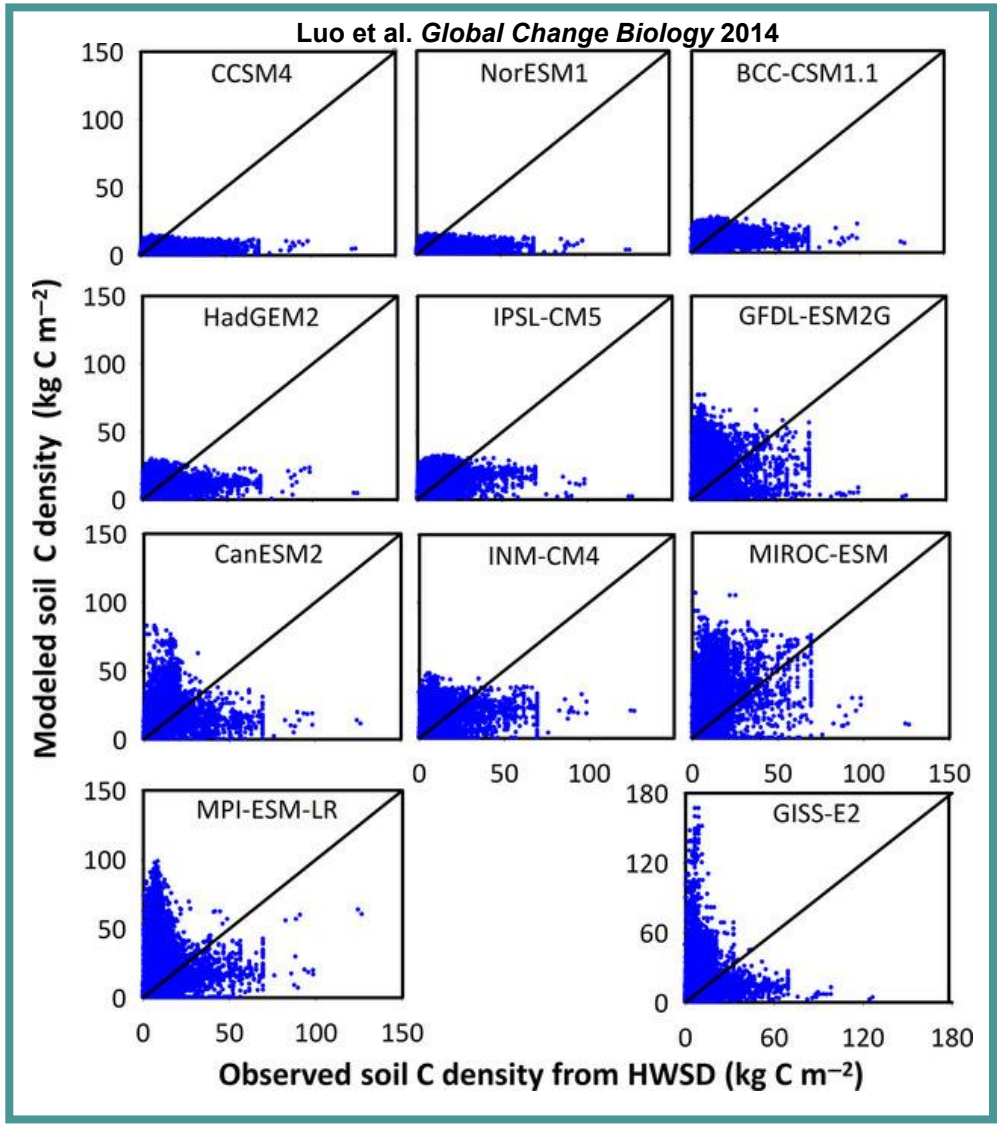
- Imbalance within global carbon cycle can partially explained by the uncertainties in land carbon sinks
- Soils hold a large part of uncertainties in land carbon sinks

Haodi Xu [hx293@cornell.edu]

Modeling SOC Dynamics



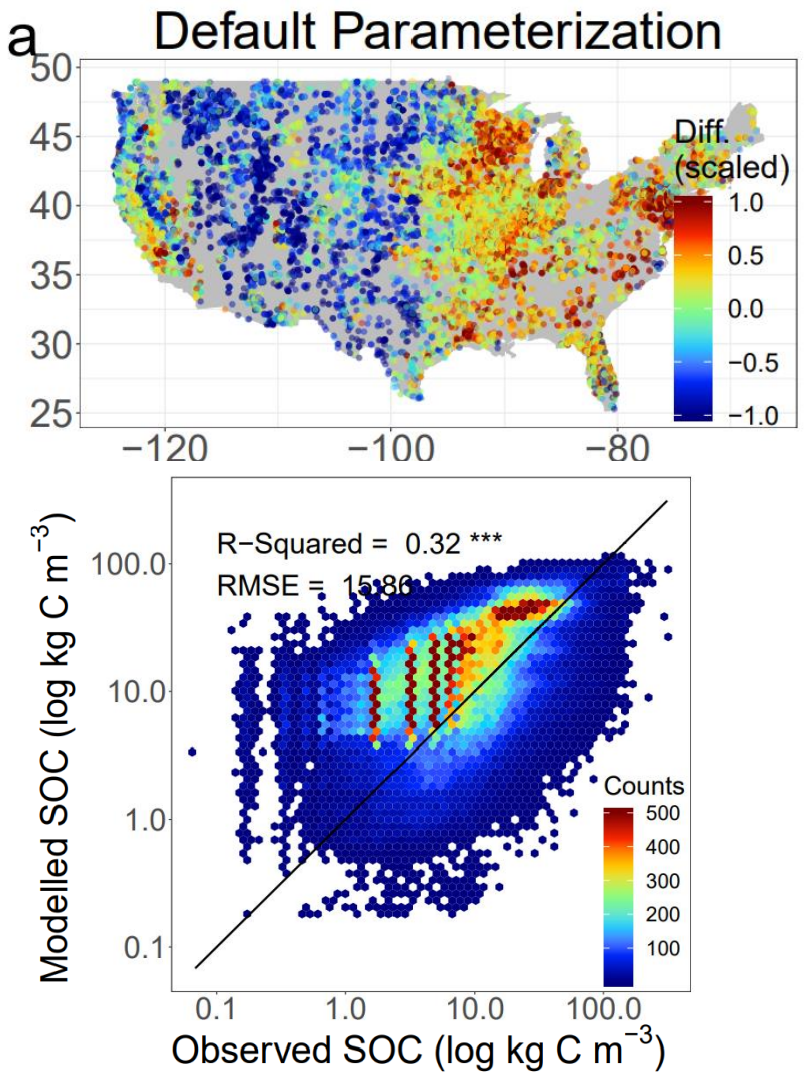


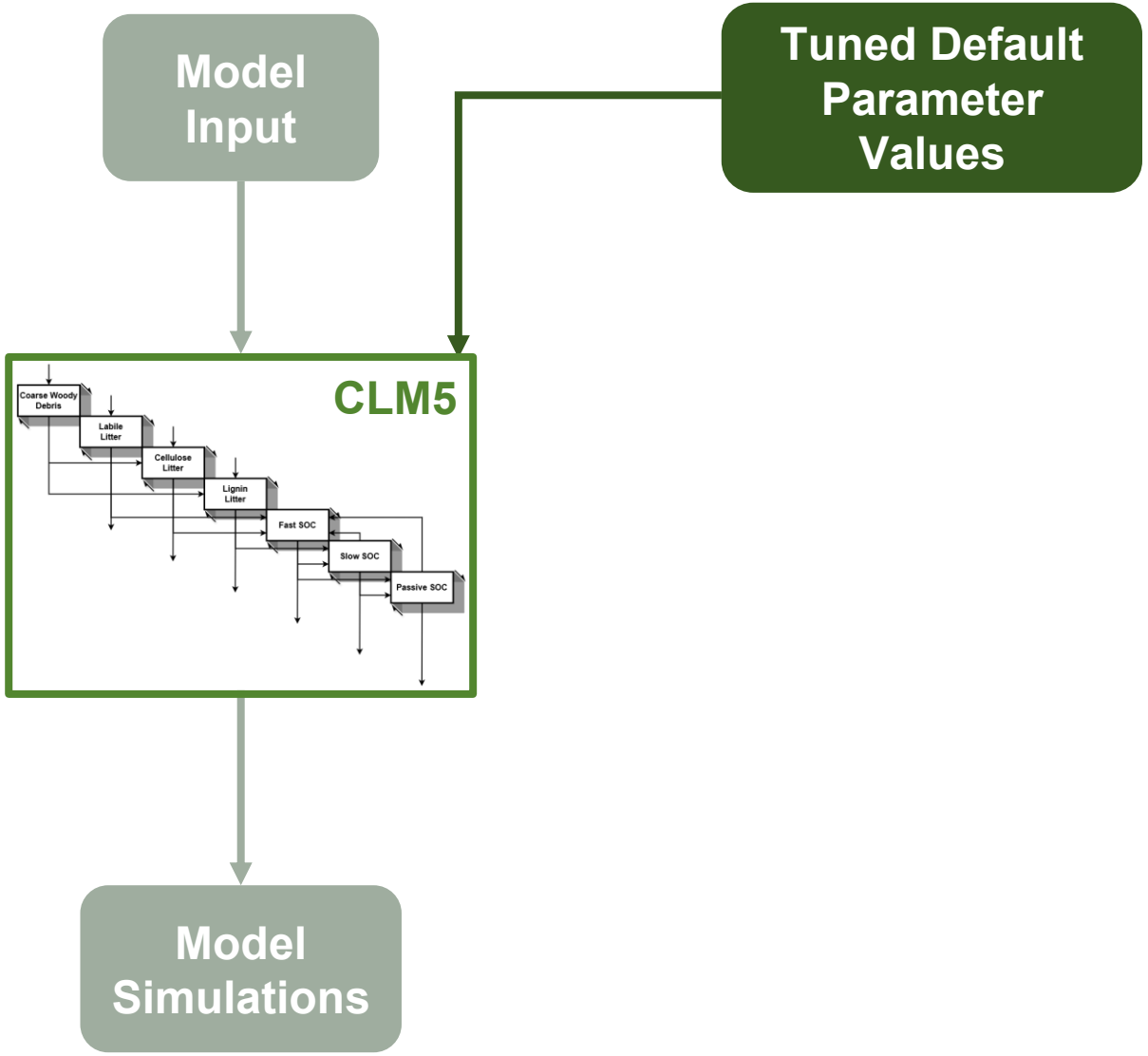


- Our knowledge of SOC processes can be incorporated into process-based models
- Yet high uncertainties remain in capturing the global SOC cycle

Haodi Xu [hx293@cornell.edu]

Modeled SOC across Conterminous US

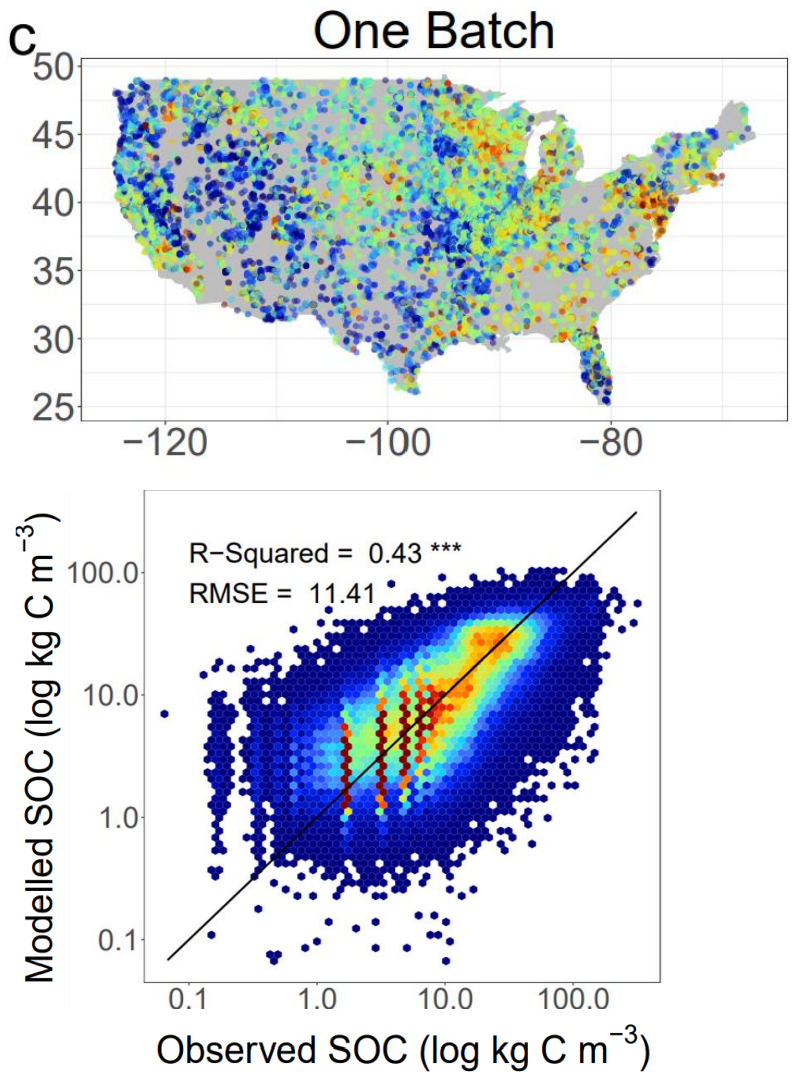


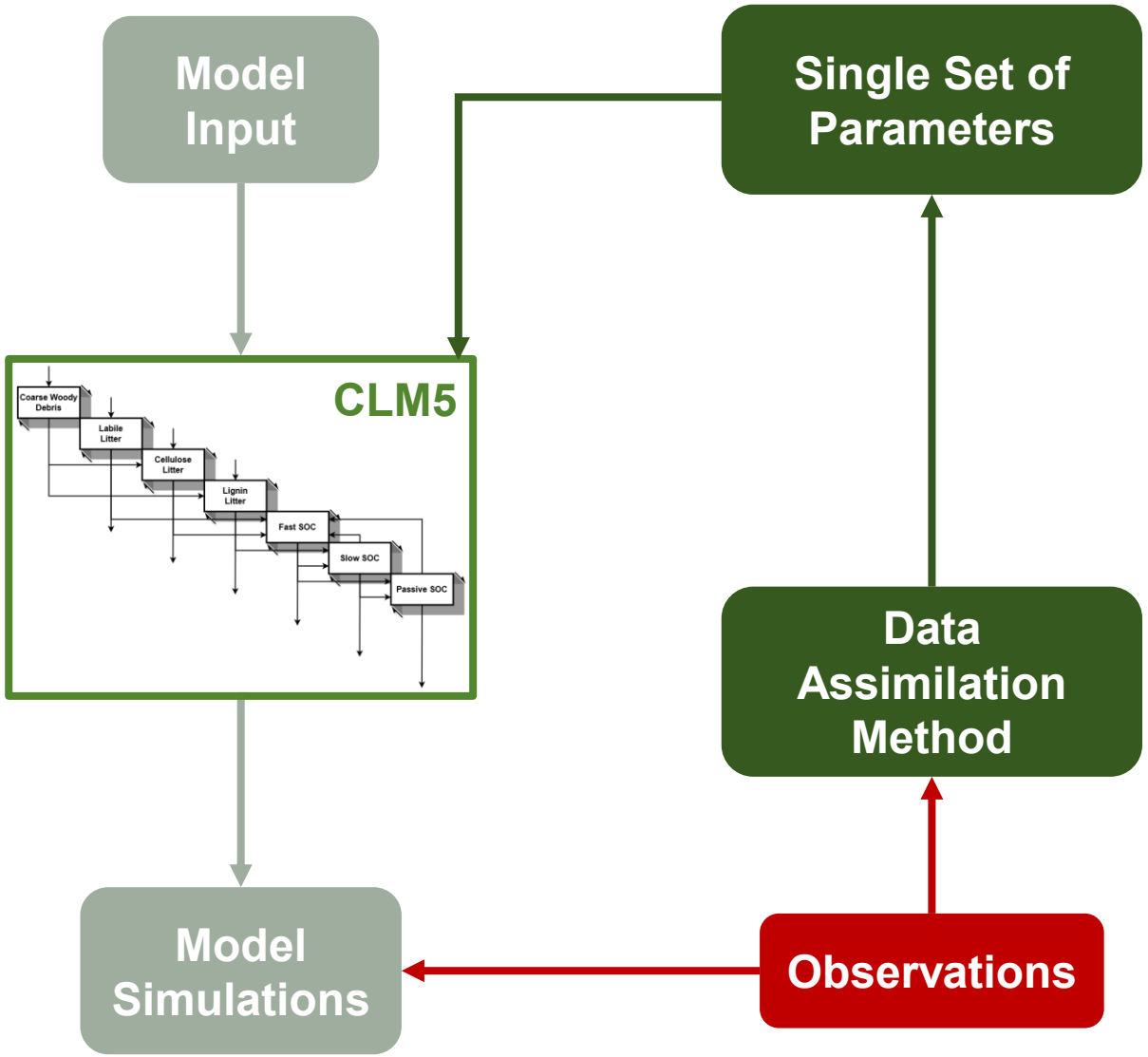


Tao et al. *Frontiers in Big Data* 2020

Haodi Xu [hx293@cornell.edu]

Data Assimilation with One Set of Parameters





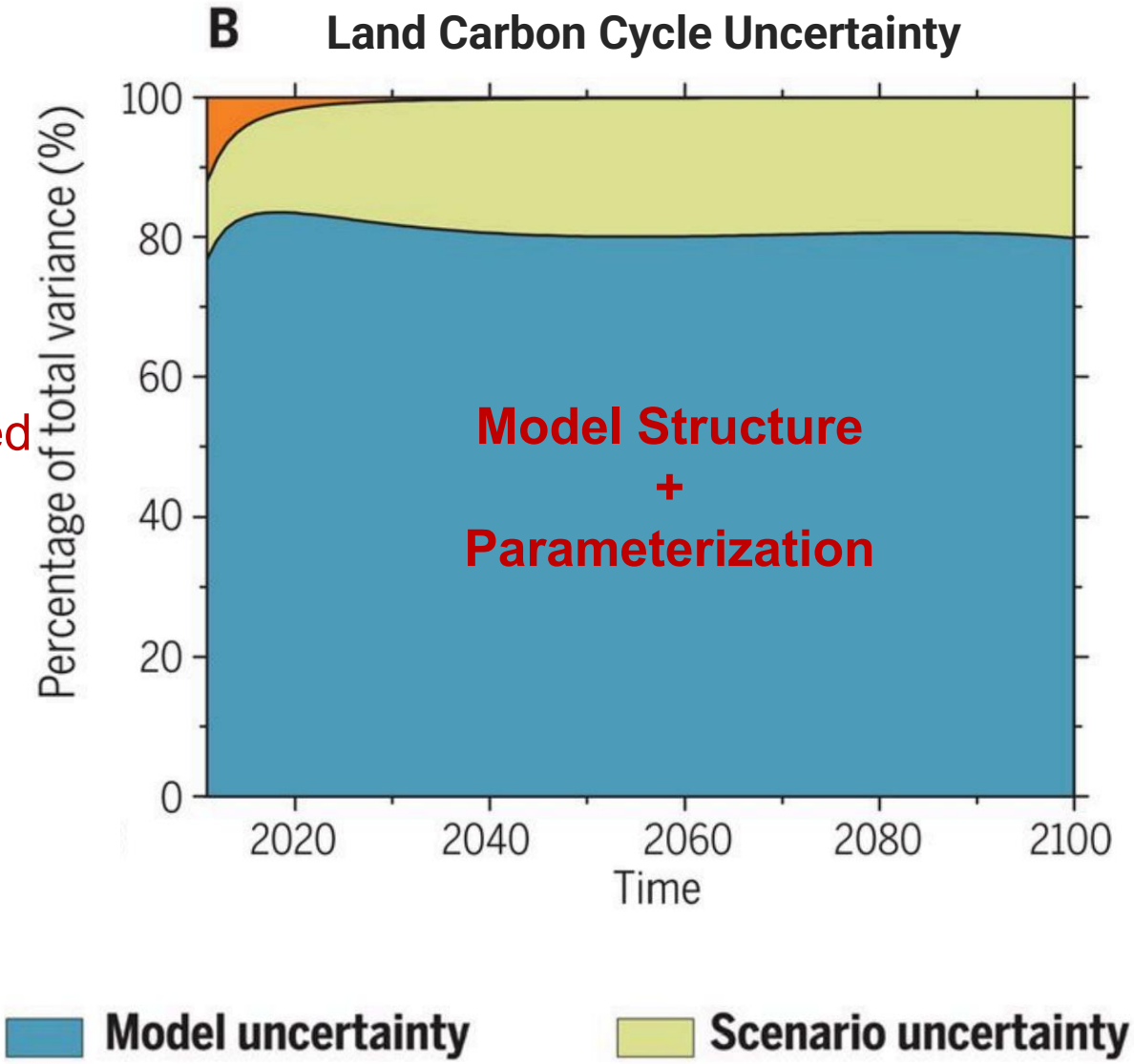
Tao et al. *Frontiers in Big Data* 2020

Cornell CALS College of Agriculture and Life Sciences

Haodi Xu [hx293@cornell.edu]

ECOLAB OF DR. YIQI LUO

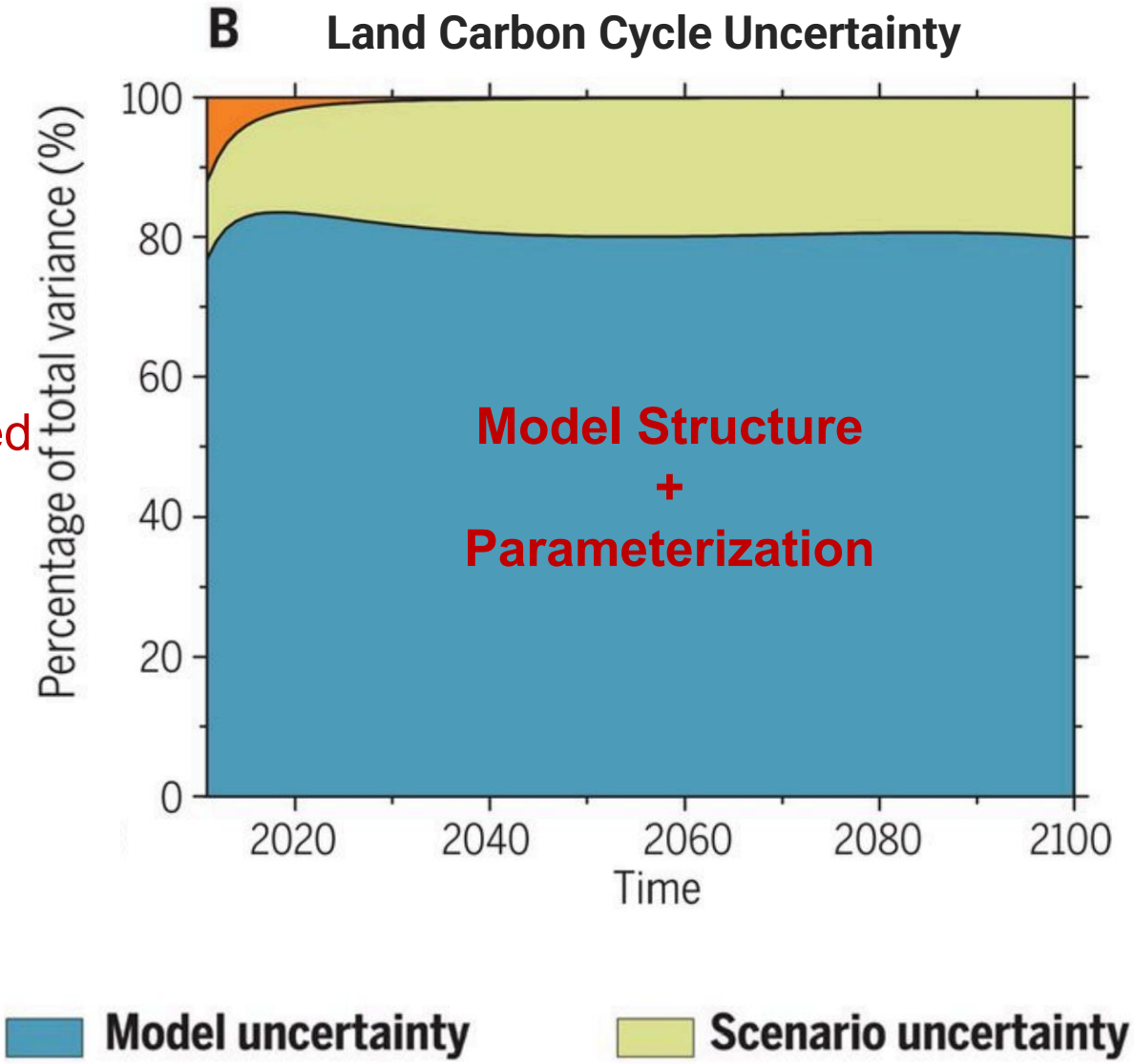
Model Structure: processes relatively well defined
Parameterization: quantify the processes not explicitly represented



Bonan et al., *Science*, 2018

Heterogeneity of soils is one of the major sources of uncertainty, can hardly be represented by a single set of parameters

Model Structure: processes relatively well defined
Parameterization: quantify the processes not explicitly represented



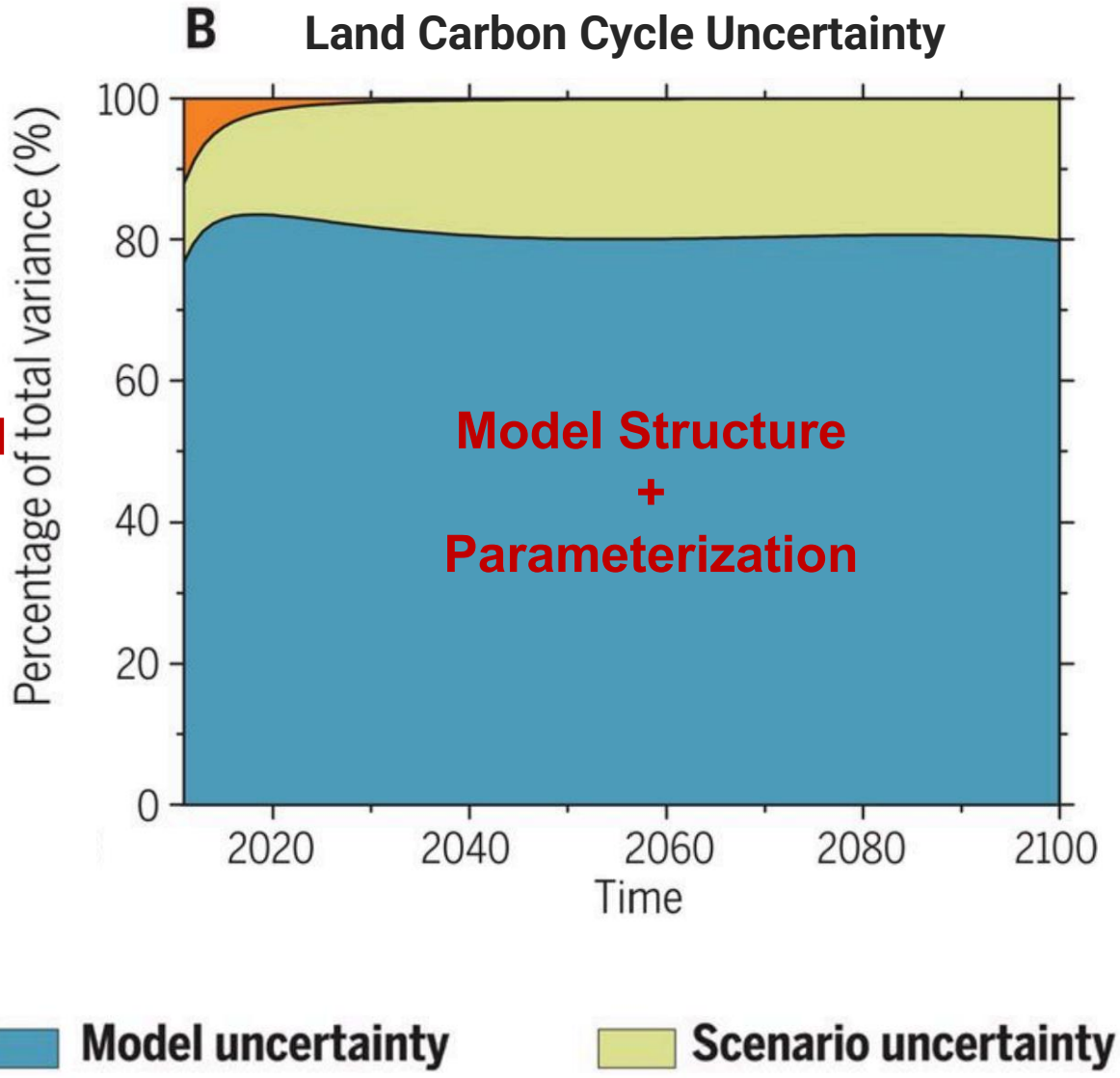
Bonan et al., *Science*, 2018

Haodi Xu [hx293@cornell.edu]

Heterogeneity of soils is one of the major sources of uncertainty, can hardly be represented by a single set of parameters

Model Structure: processes relatively well defined
Parameterization: quantify the processes not explicitly represented

Identify parameter sets that accurately simulate SOC storage, allowing the model to more effectively represent the complexity of real soil systems.



Bonan et al., *Science*, 2018

Haodi Xu [hx293@cornell.edu]

Environmental Information

Climate

Soil pH

Lon, Lat, Elevation

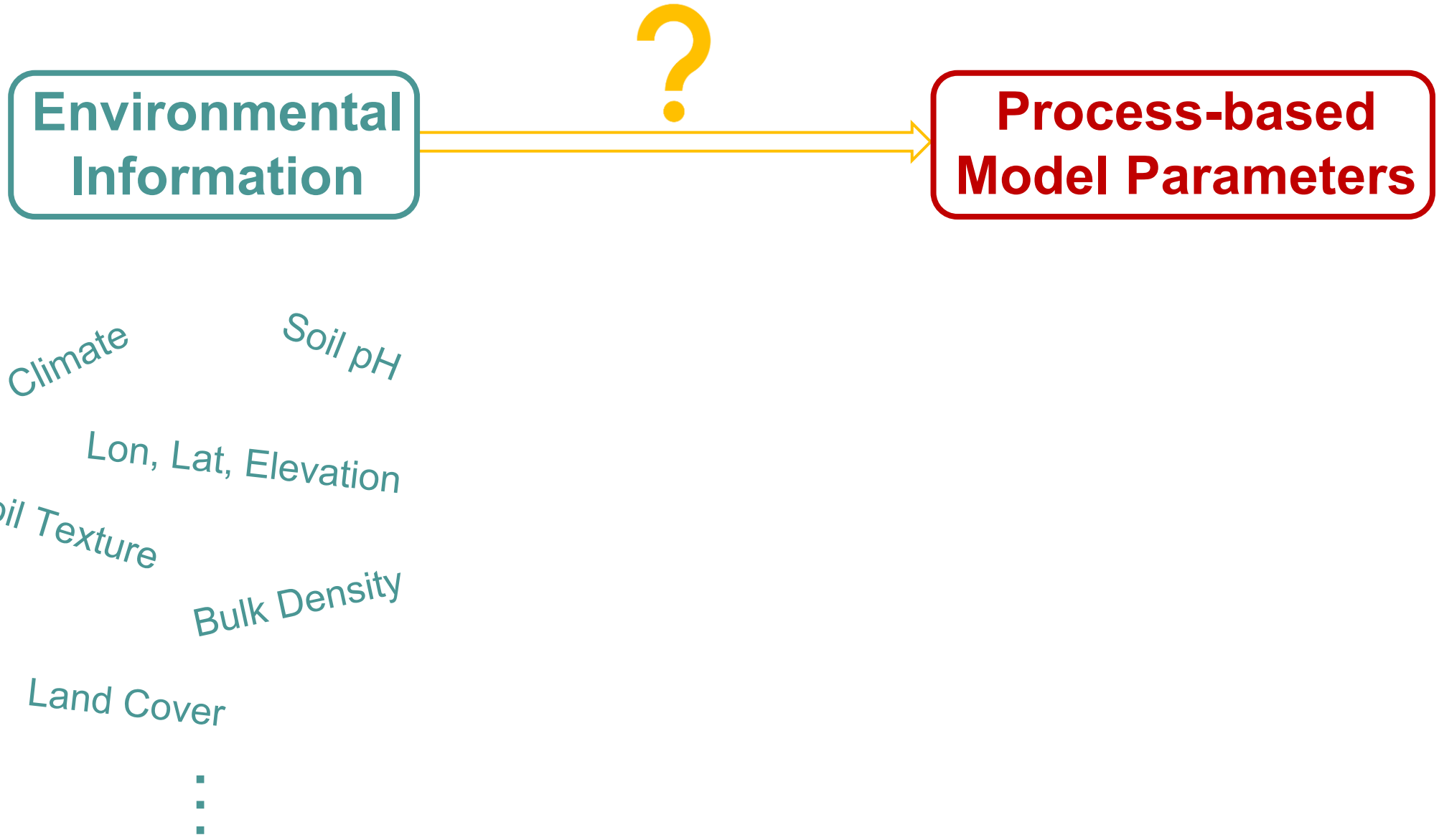
Soil Texture

Bulk Density

Land Cover

⋮

Haodi Xu [hx293@cornell.edu]



Haodi Xu [hx293@cornell.edu]

Artificial Intelligence (Deep Learning)

Environmental Information

Process-based Model Parameters

Climate

Soil pH

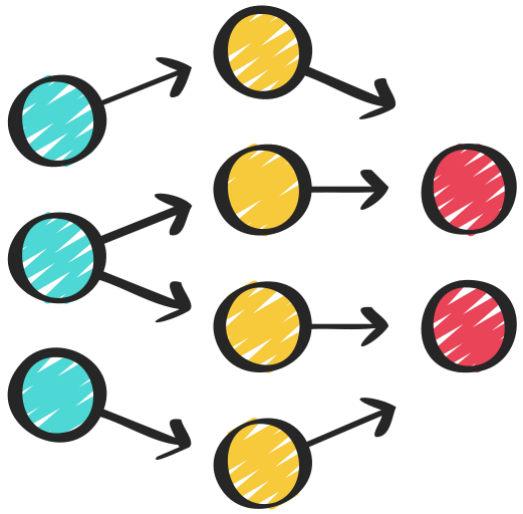
Lon, Lat, Elevation

Soil Texture

Bulk Density

Land Cover

⋮



Microbial Carbon Use Efficiency

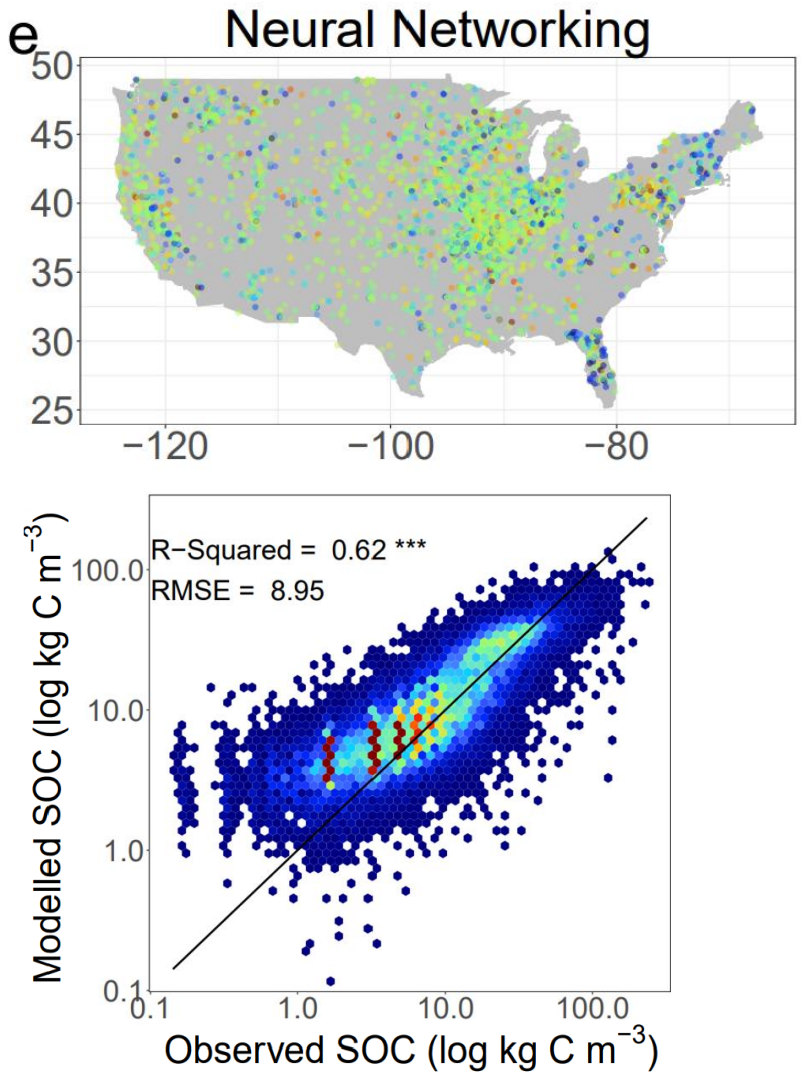
Substrate Decomposability

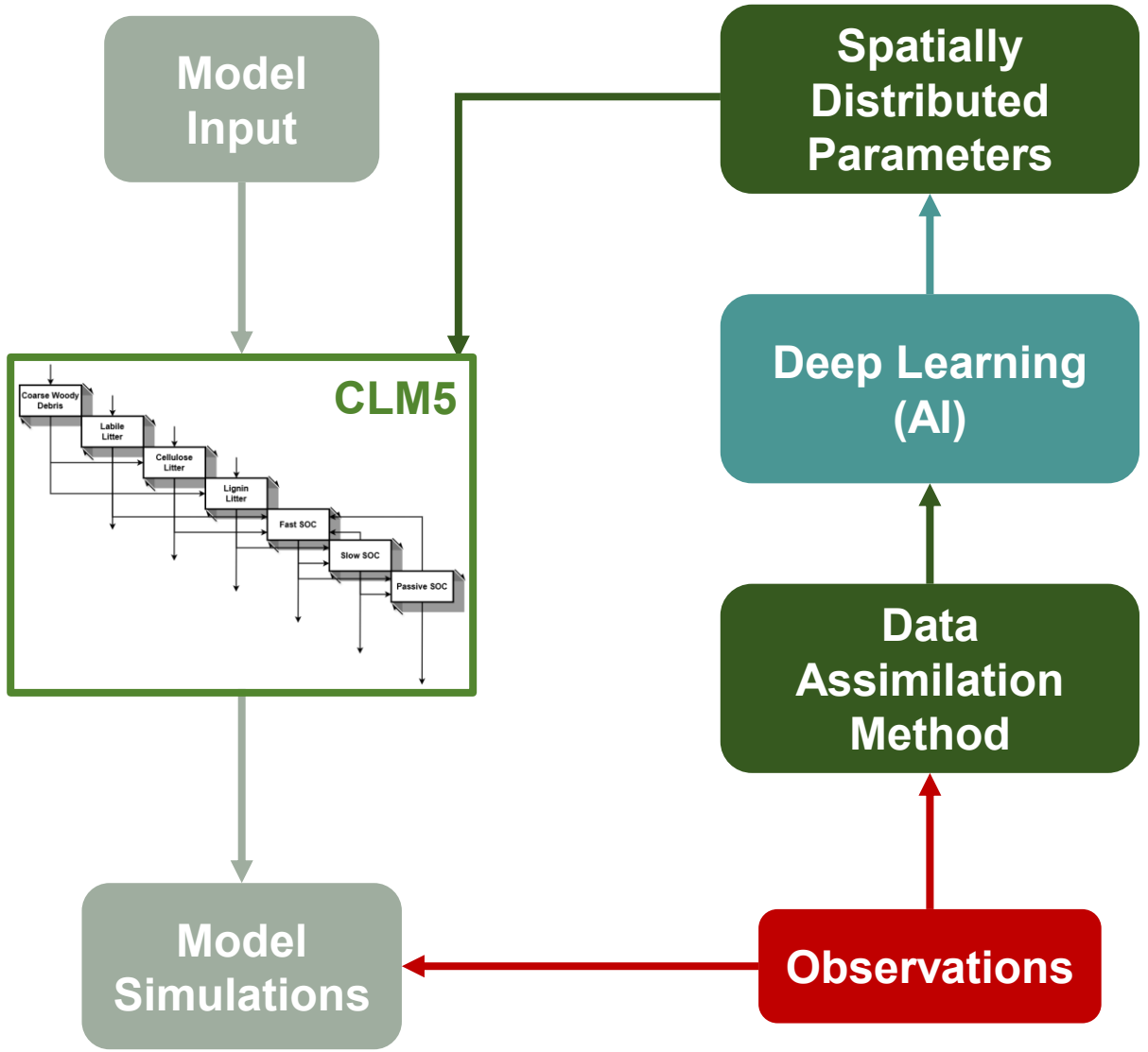
Temperature Sensitivity

⋮

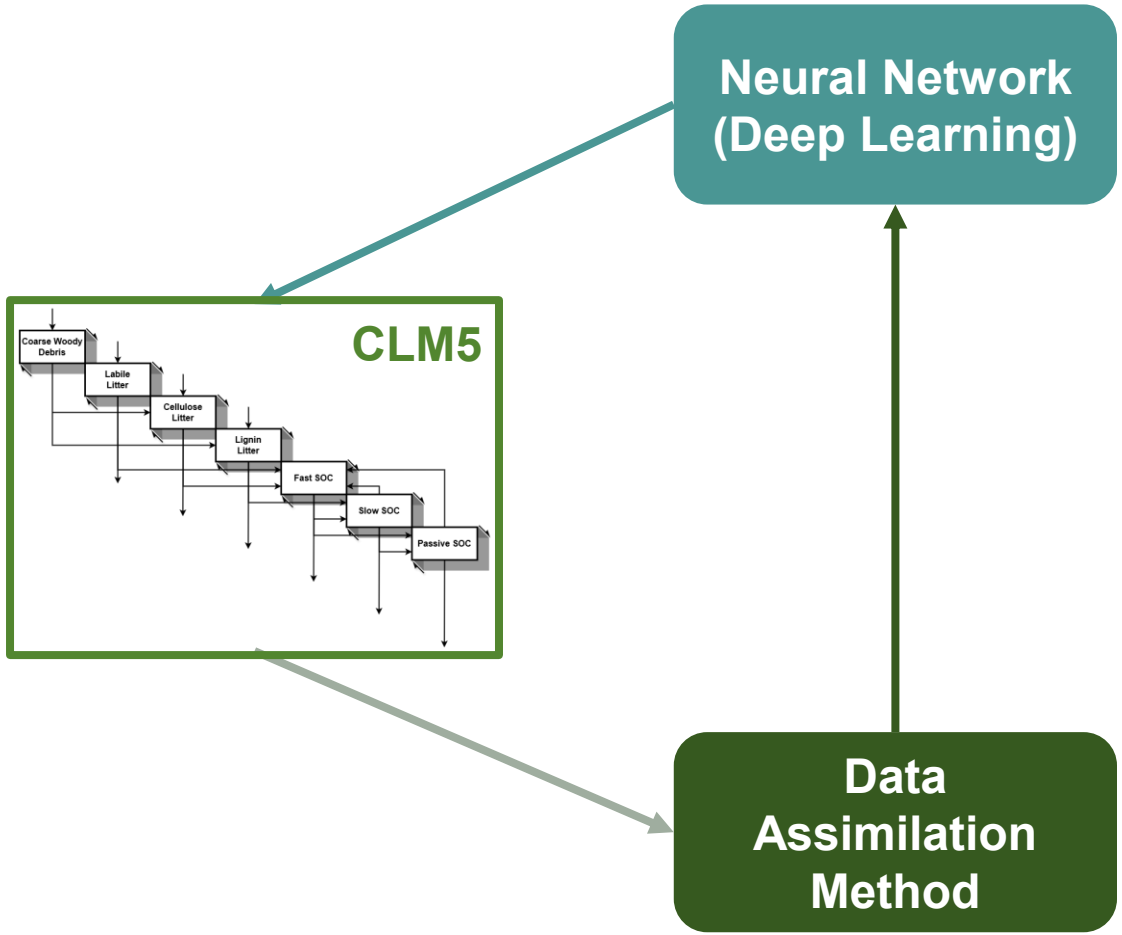
Haodi Xu [hx293@cornell.edu]

PROcess guided deep learning and DAta driven modeling (PRODA)





The need to further improve PRODA

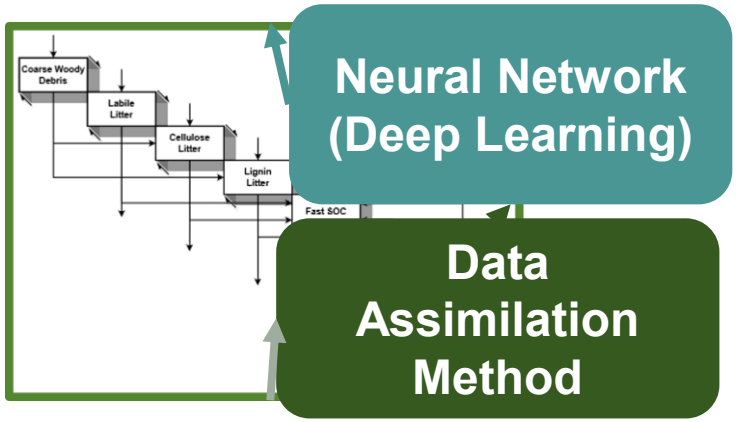


PRODA includes three components:

- High computational cost due to **Bayesian-based** data assimilation method (e.g. MCMC)
- Limited flexibility with observed data
- Hard to understand and use

Haodi Xu [hx293@cornell.edu]

The need to further improve PRODA

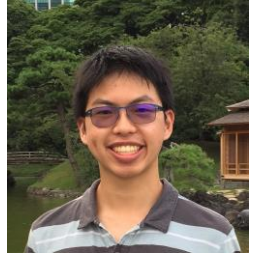


PRODA includes three components:

- High computational cost due to **Bayesian-based** data assimilation method (e.g. MCMC)
- Limited flexibility with observed data
- Hard to understand and use

Combine these altogether?

Haodi Xu [hx293@cornell.edu]



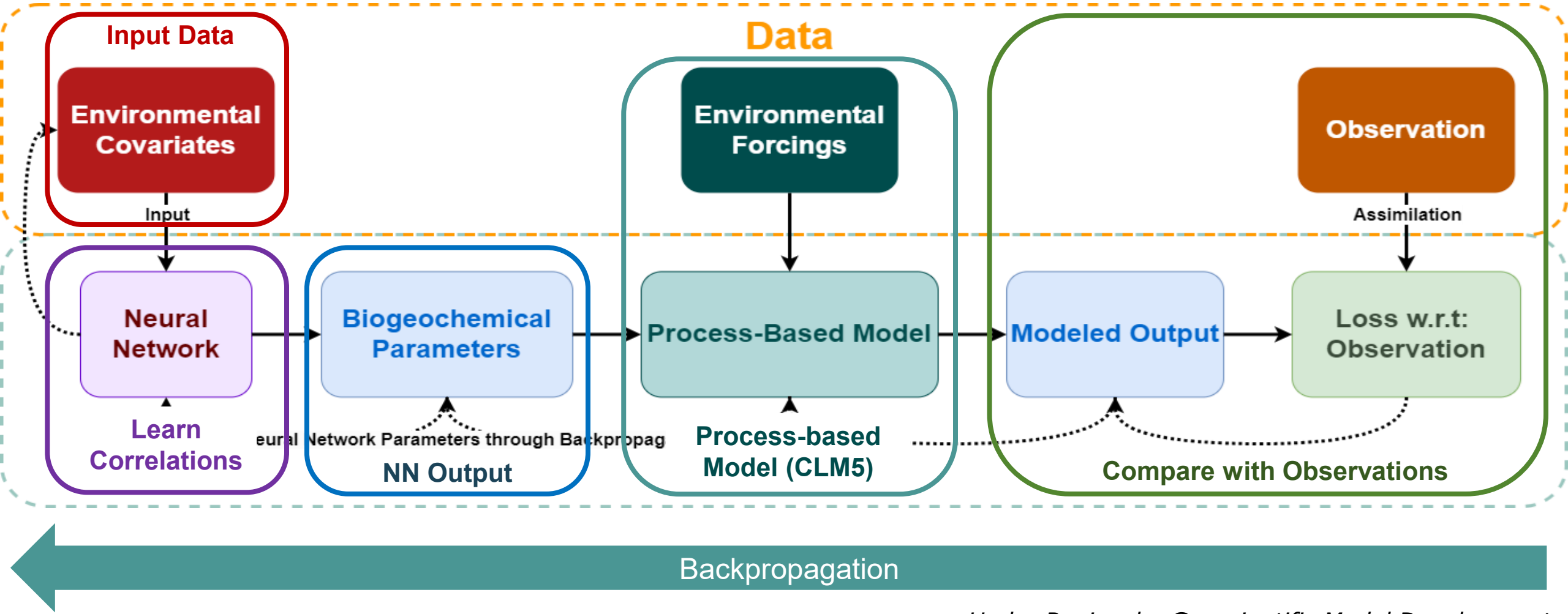
A Biogeochemistry-Informed Neural Network (BINN)

CornellCALS College of Agriculture and Life Sciences

Haodi Xu [hx293@cornell.edu]

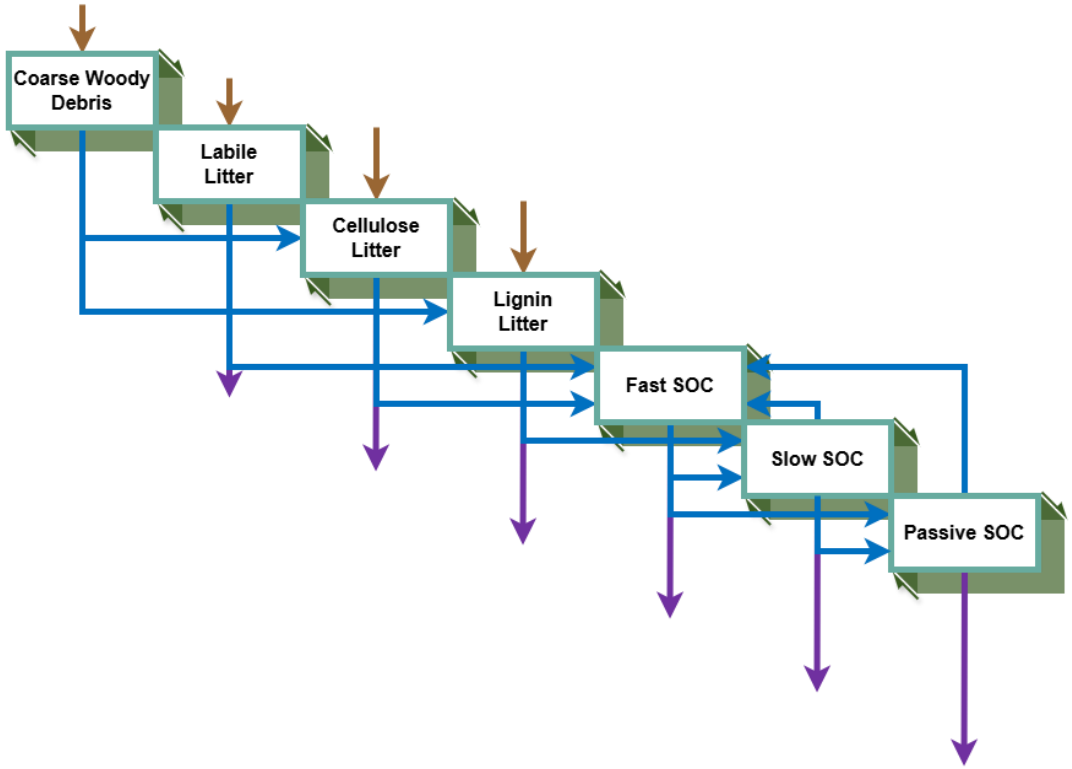
ECOLAB OF DR. YIQI LUO

Workflow of BINN



Under Review by Geoscientific Model Development

Matrix Representation of CLM5



Changes in Carbon in each pool through out the time

C movement among different pools

Carbon Dynamics through Diffusion

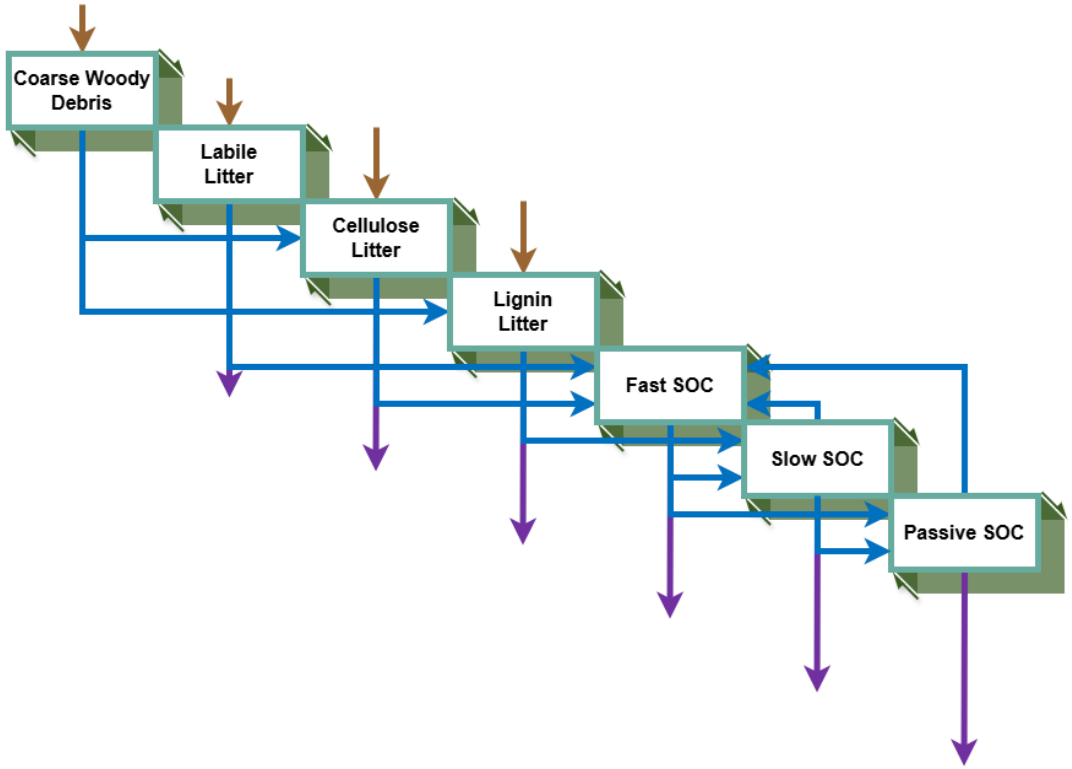
$$\frac{dX(t)}{dt} = B(t)I(t) - A\,\xi(t)KX(t) - V(t)X(t)$$

Carbon Input

The intrinsic decomposition rate of each C pool, modified by the environmental scalar

Each matrix is constructed from one or more unique parameters (21 parameters in total)

Matrix Representation of CLM5



Changes in Carbon in each pool through out the time

C movement among different pools

Carbon Dynamics through Diffusion

$$\frac{dX(t)}{dt} = B(t)I(t) - A\xi(t)KX(t) - V(t)X(t)$$

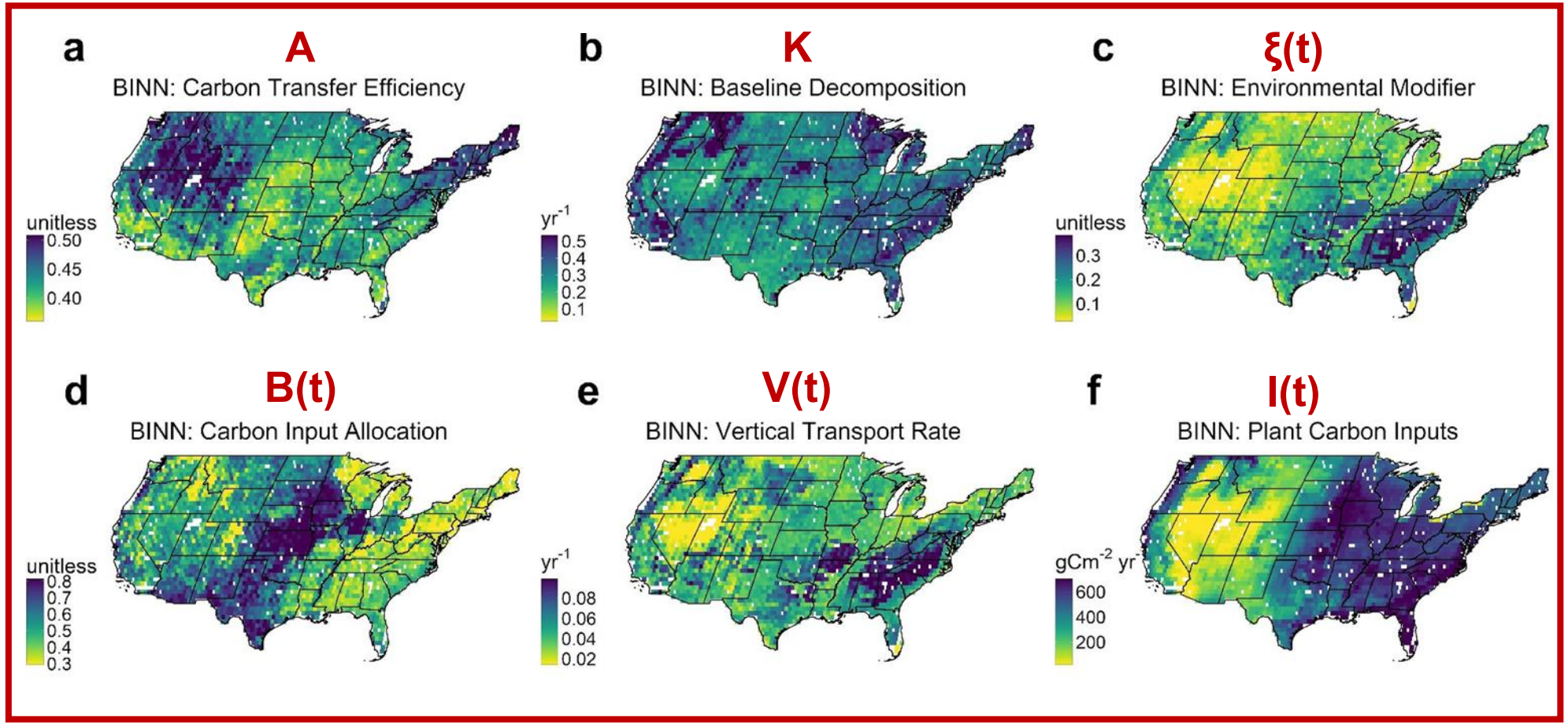
Carbon Input

The intrinsic decomposition rate of each C pool, modified by the environmental scalar

: Model Components, quantified by 21 model parameters

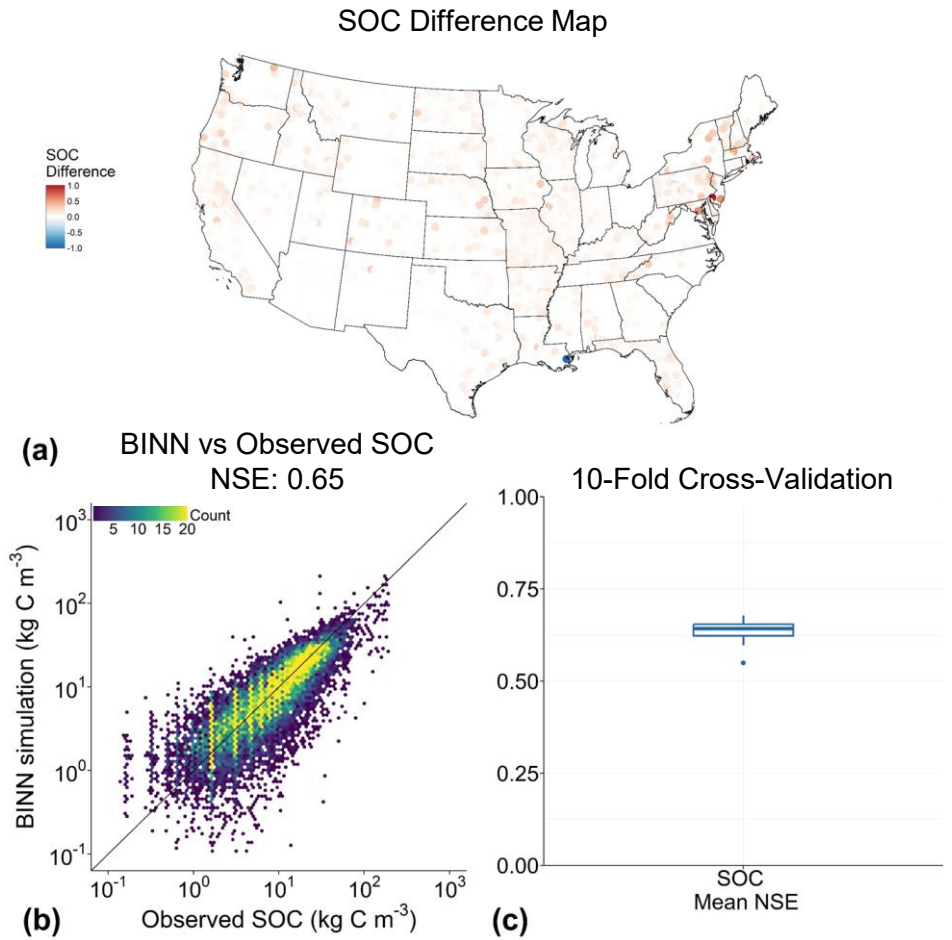
Haodi Xu [hx293@cornell.edu]

BINN discovers the spatial distribution of underlying mechanisms from SOC observations and environmental covariates (without any data for these mechanisms)

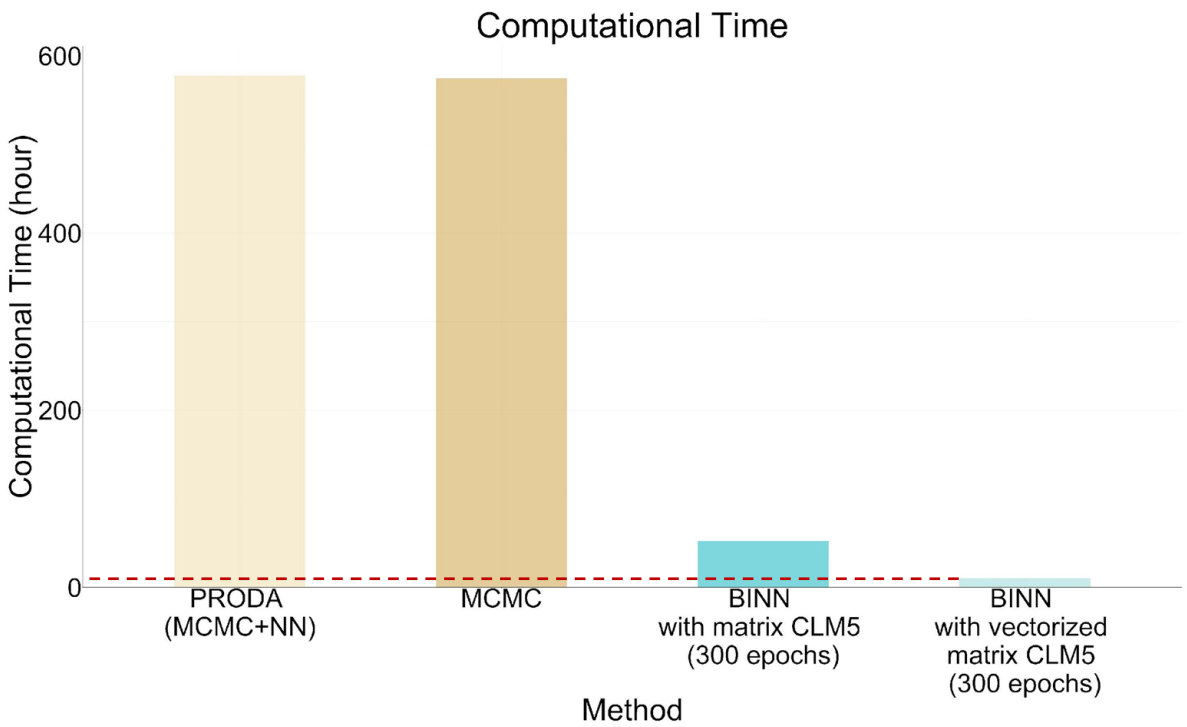


Haodi Xu [hx293@cornell.edu]

More Realistic Representations of SOC Distribution



Way Less Computational Requirements (More than 50 times)



CornellCALS College of Agriculture and Life Sciences

Haodi Xu [hx293@cornell.edu]

ECOLAB OF DR. YIQI LUO

Data

Environmental Covariates

Environmental Forcings

Observation

Input

Assimilation

Neural Network

Biogeochemical Parameters

Process-Based Model

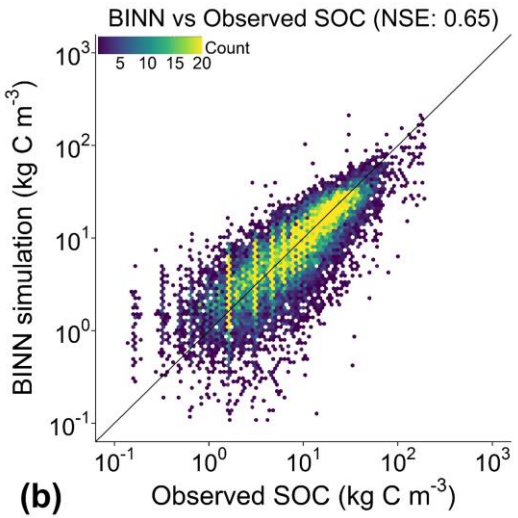
Modeled Output

Loss w.r.t: Observation

Update Neural Network Parameters through Backpropagation

BINN

SOC Model



(b)

Data

Environmental Covariates

Environmental Forcings

Observation

Input

Assimilation

Neural Network

Biogeochemical Parameters

Process-Based Model

Modeled Output

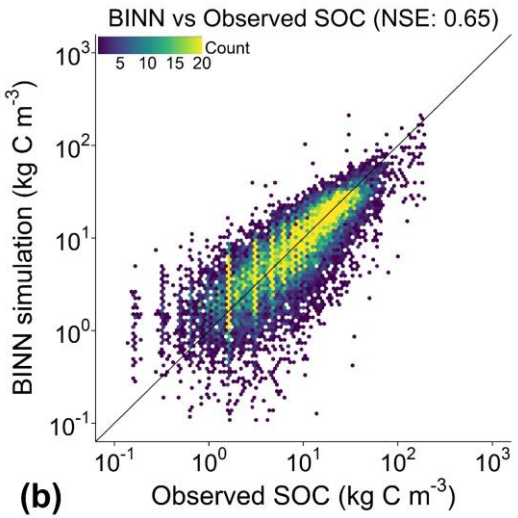
Loss w.r.t: Observation

Update Neural Network Parameters through Backpropagation

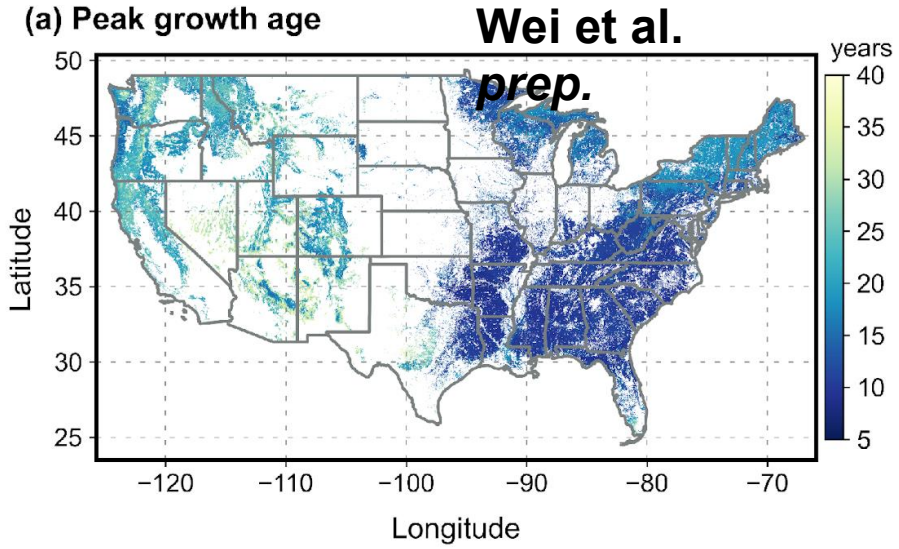
BINN

SOC Model

Forest Growth Model



(b)



CornellCALS College of Agriculture and Life Sciences

Haodi Xu [hx293@cornell.edu]

ECOLAB OF DR. YIQI LUO

Data

Environmental Covariates

Environmental Forcings

Observation

Input

Assimilation

Neural Network

Biogeochemical Parameters

?

Process-Based Model

Modeled Output

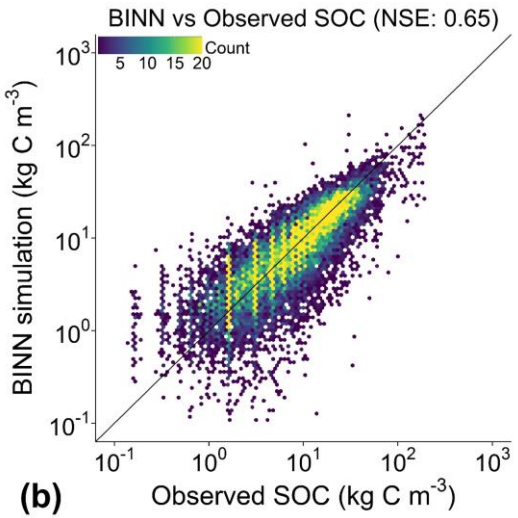
Loss w.r.t: Observation

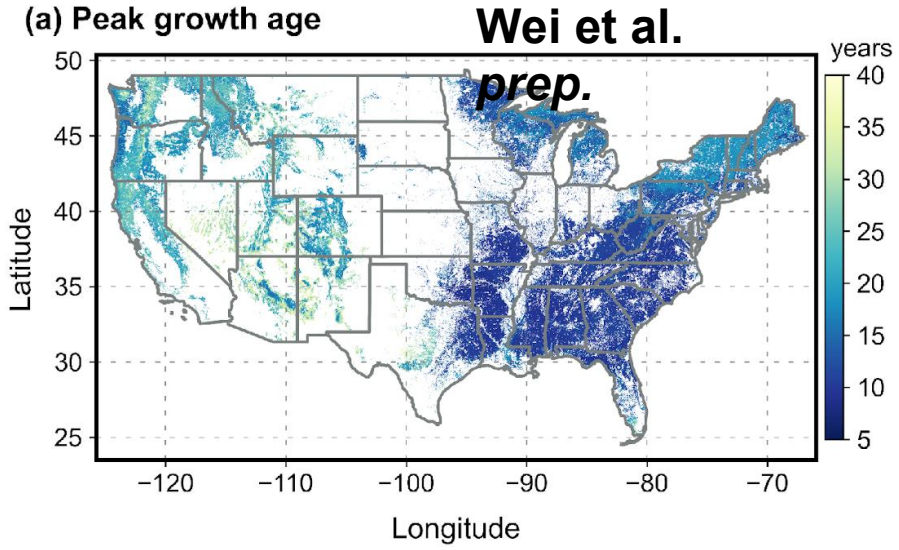
Update Neural Network Parameters through Backpropagation

BINN

SOC Model

Forest Growth Model



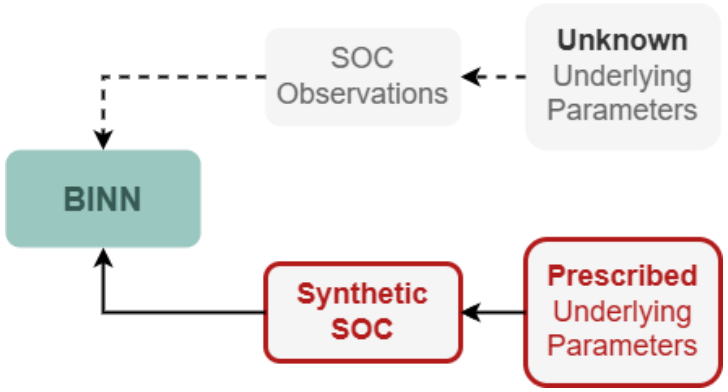


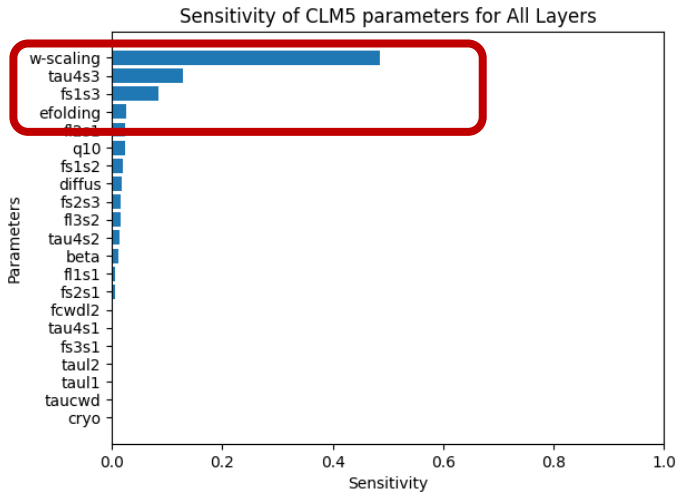
CornellCALS College of Agriculture and Life Sciences

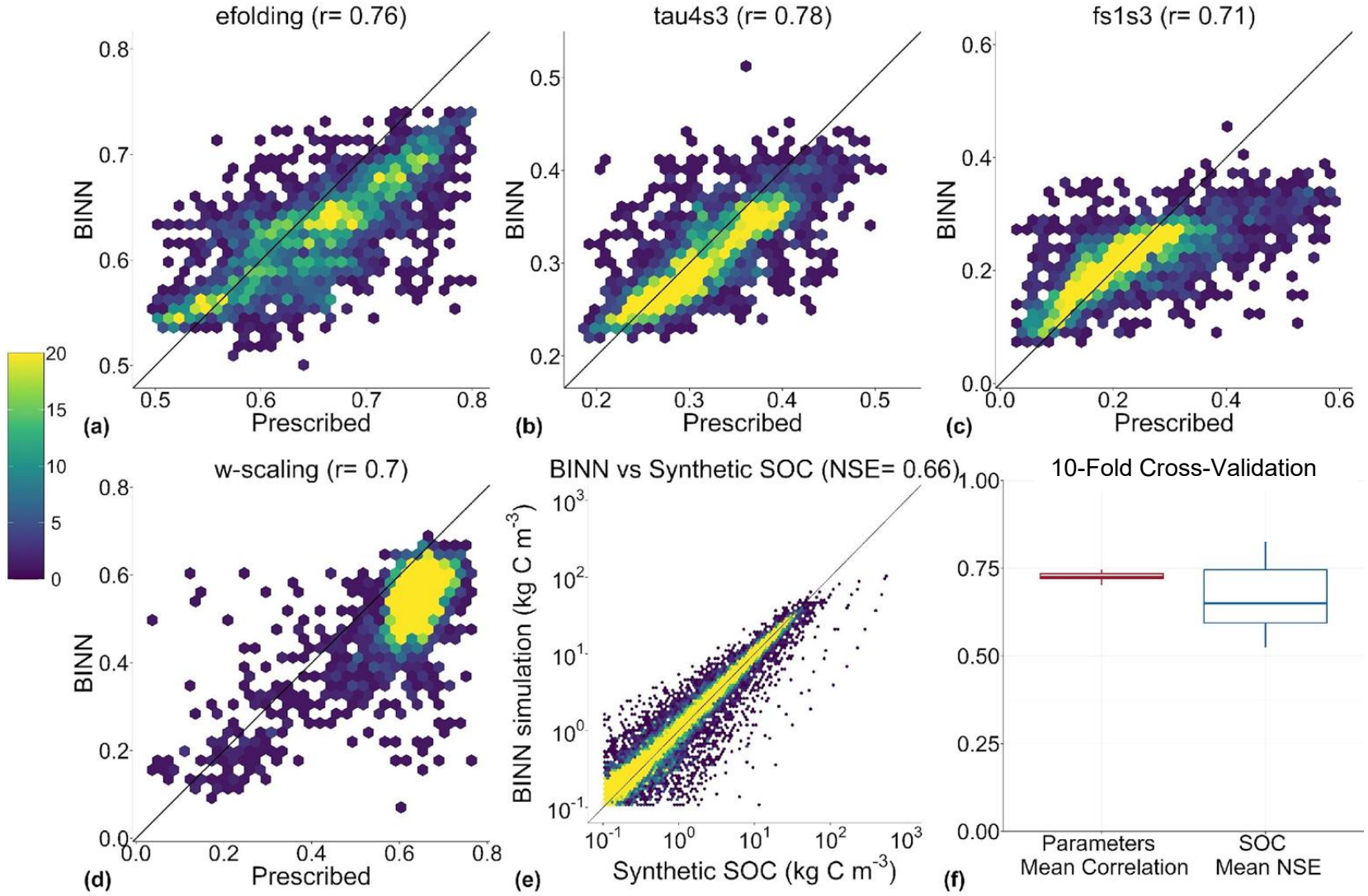
Haodi Xu [hx293@cornell.edu]

ECOLAB OF DR. YIQI LUO

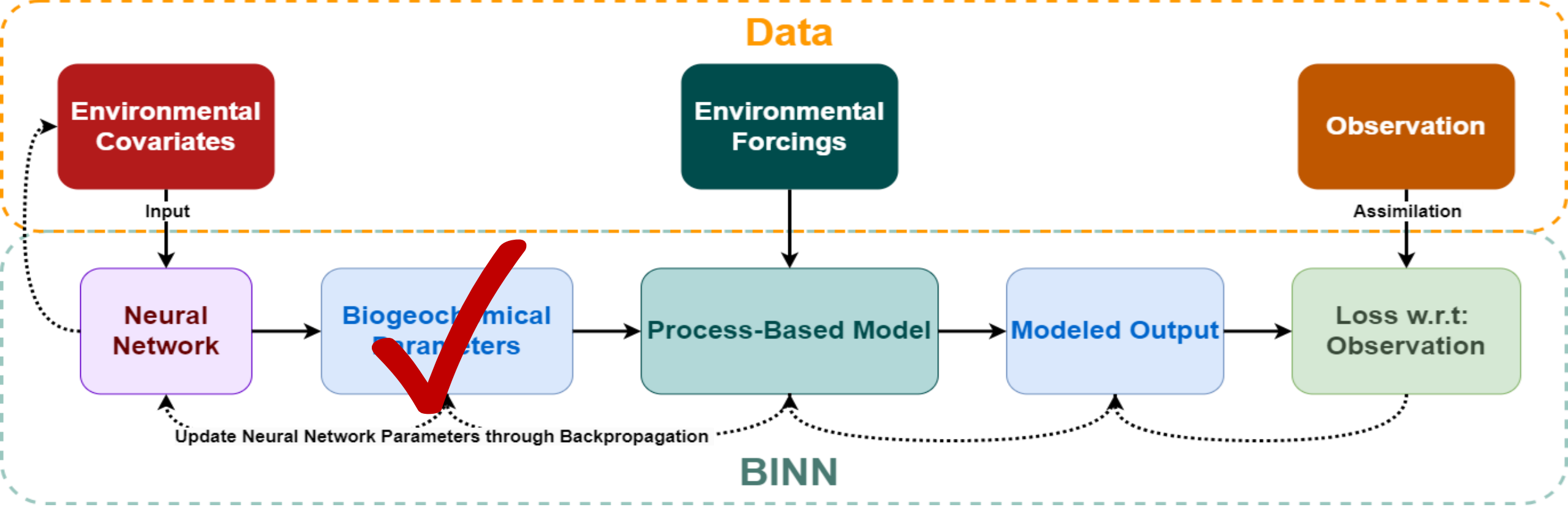
Recovery Experiment with 4 Most Influential Parameters





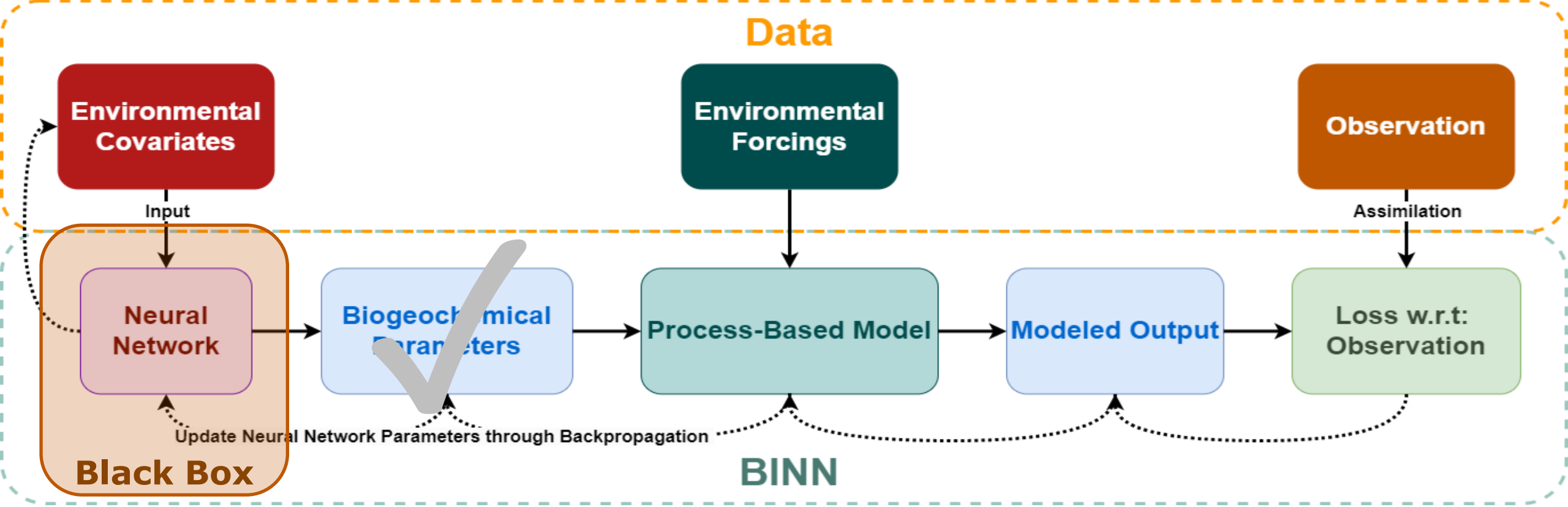


Haodi Xu [hx293@cornell.edu]

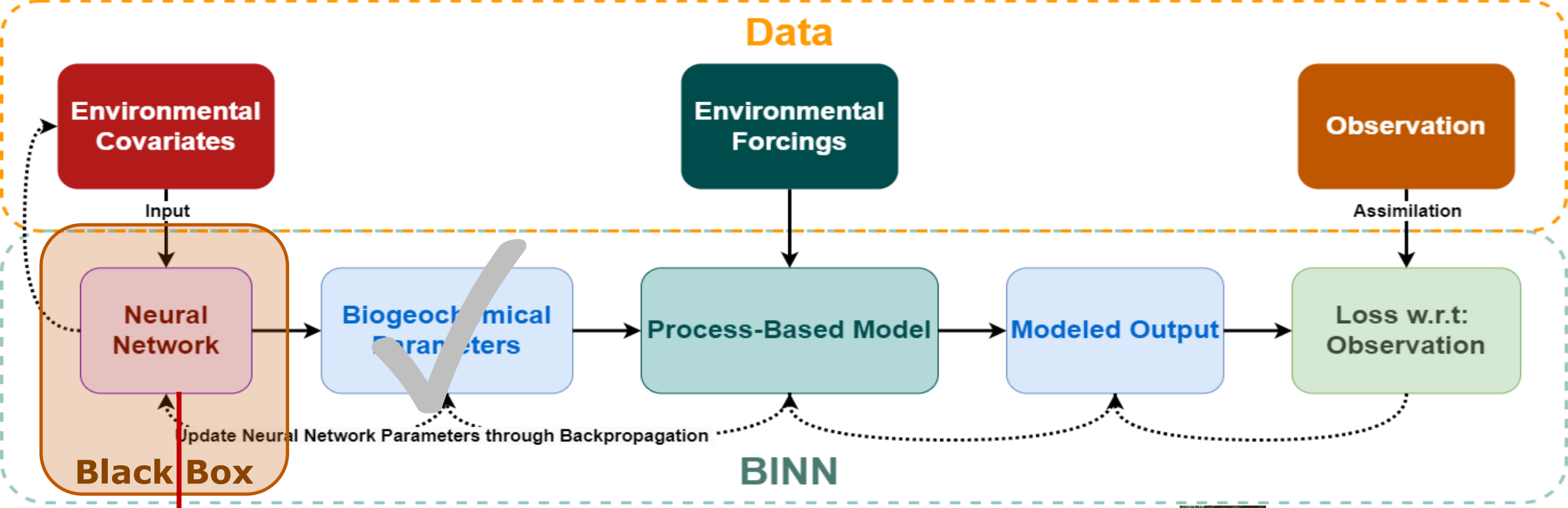


Recover causes from data (biogeochemical parameters)

Discover emerging mechanisms governing SOC from big data (observations)



Haodi Xu [hx293@cornell.edu]



Kolmogorov Arnold Network (KAN)

An alternative to neural networks that is easier to interpret

Scientifically-Interpretable Reasoning Network(ScIReN)

Submitted to NeurIPS 2025

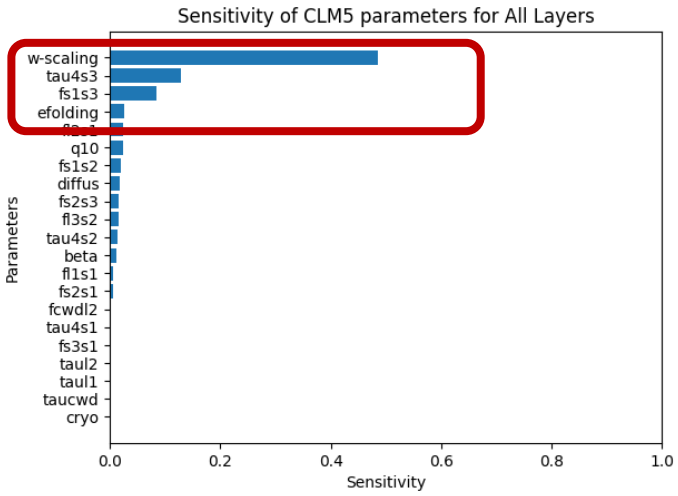
Cornell CALS College of Agriculture and Life Sciences

Haodi Xu [hx293@cornell.edu]

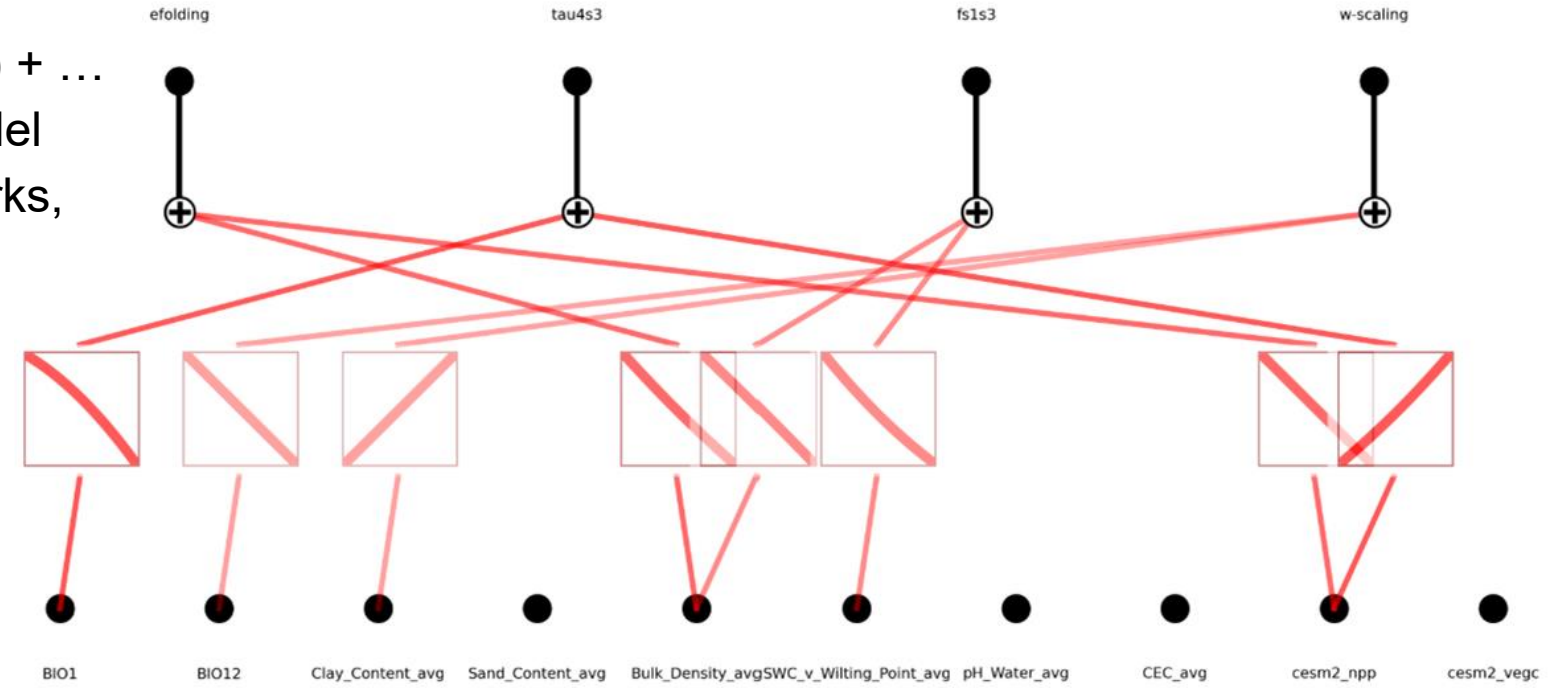
ECOLAB OF DR. YIQI LUO

Structure of KAN

- Curves show how changing one environmental variable will affect a biogeochemical parameter
- Each output = sum of interpretable contributions from each input, e.g.
  - parameter = $(input1)^2$ + exp(input6) + …
- 1-layer KAN = generalized additive model
  - Not as expressive as neural networks, but similar accuracy in our case
  - Can add more layers



Outputs (Biogeochemical Parameters) - not observed

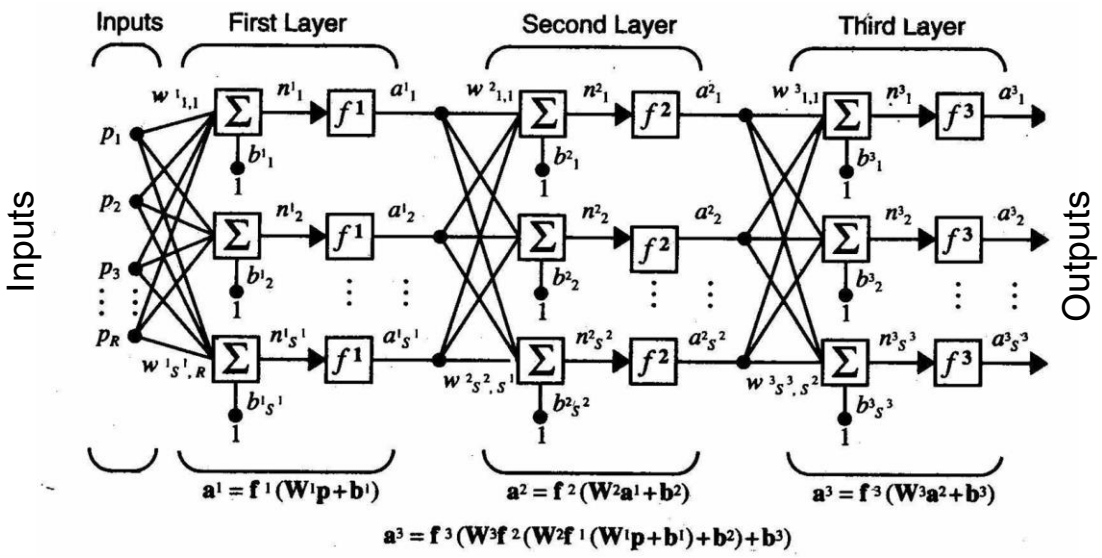


Inputs (Environmental Covariates) - observed

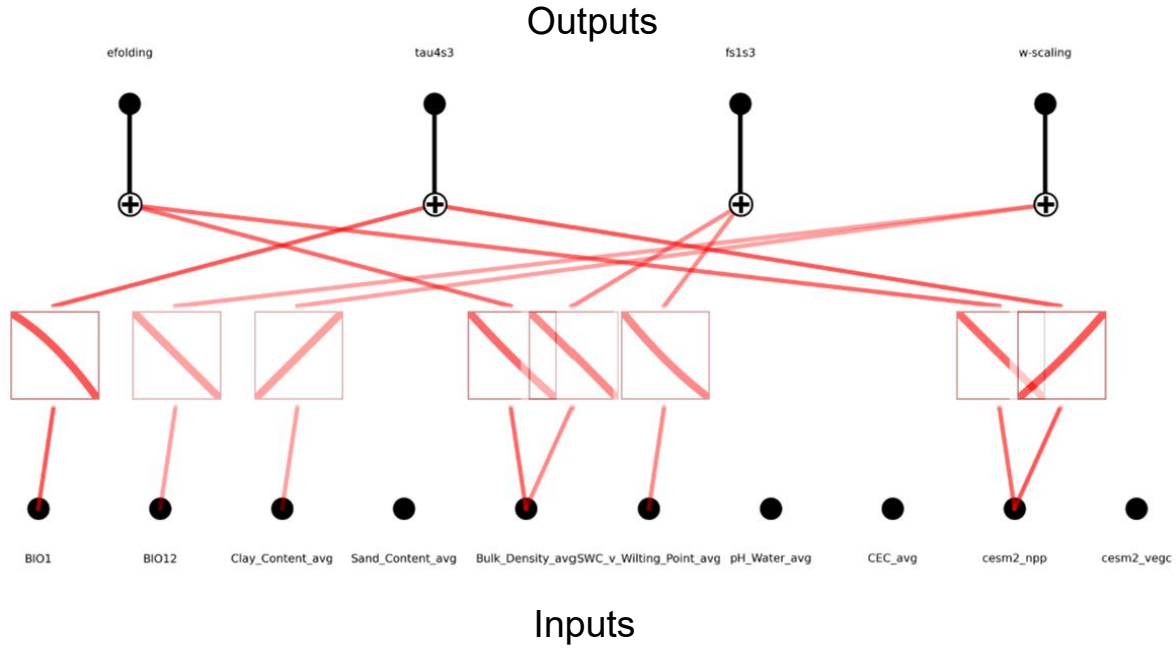
Haodi Xu [hx293@cornell.edu]

Difference between NN and KAN

- **Standard neural networks:** every input variable affects every output
  - Inputs are mixed together and passed through nonlinearities - hard to understand
  - Unclear how each input (environmental variable) influences each output (biogeochemical parameters)



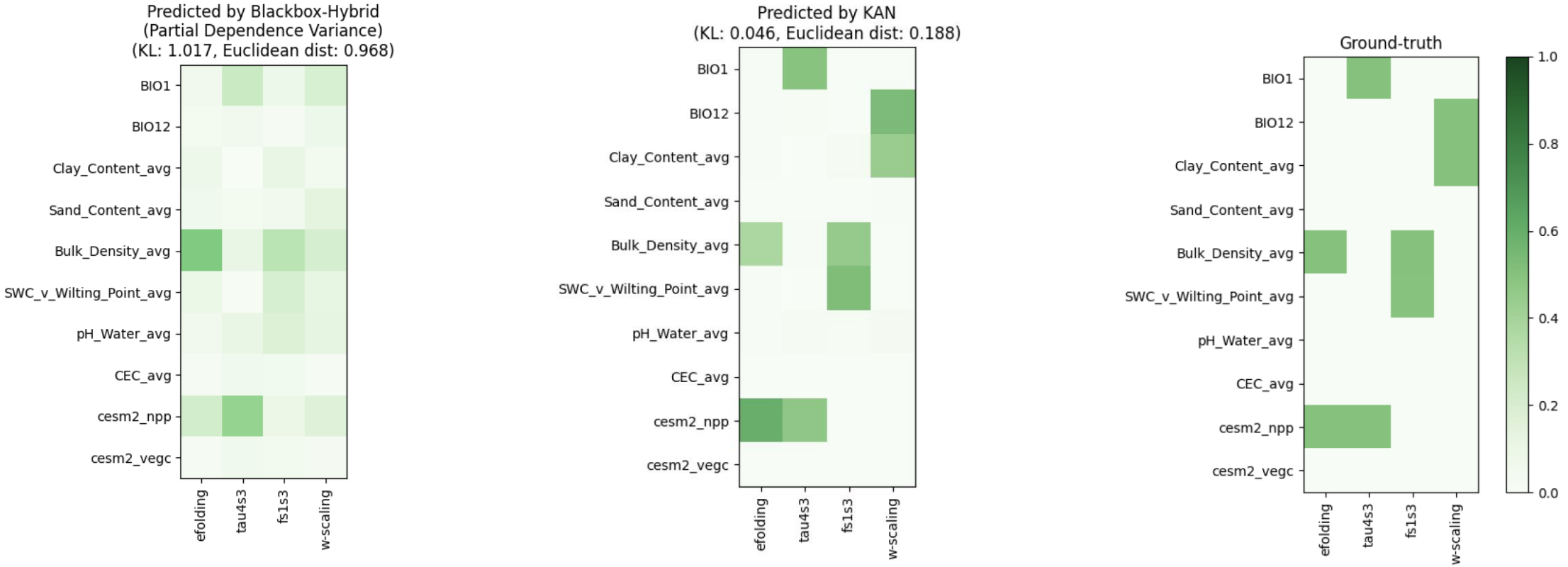
- **Kolmogorov-Arnold Networks:** learn **interpretable** functional relationships between inputs and outputs
  - Apply a 1D function on each edge (link), then add contributions from each input
  - **Sparsity:** focus on a few crucial relationships; remove unnecessary connections



Haodi Xu [hx293@cornell.edu]

Functional Relationship Retrieval Test

- Prescribed functional relationships between 10 environmental inputs and 4 most sensitive biogeochemical parameters in CLM5
- BINN does not reveal functional relationships. Even after applying a post-hoc interpretation method (Partial Dependence Variance), we find it did not even implicitly learn the correct relationships.
- ScIReN (center) revealed the relationships very accurately.



Haodi Xu [hx293@cornell.edu]

In Conclusion, BINN/SciReN can …

Recover causes from data (biogeochemical parameters)

Discover emerging mechanisms governing SOC from big data (observations)

Uncover novel relationships between environmental conditions and underlying mechanisms

Haodi Xu [hx293@cornell.edu]

Thanks for Listening

AI-LEAF USDA NIFA NSF

CornellCALS College of Agriculture and Life Sciences

ECOLAB OF DR. YIQI LUO