Artificial Intelligence to Advance Modeling and Understanding of Soil Organic Carbon

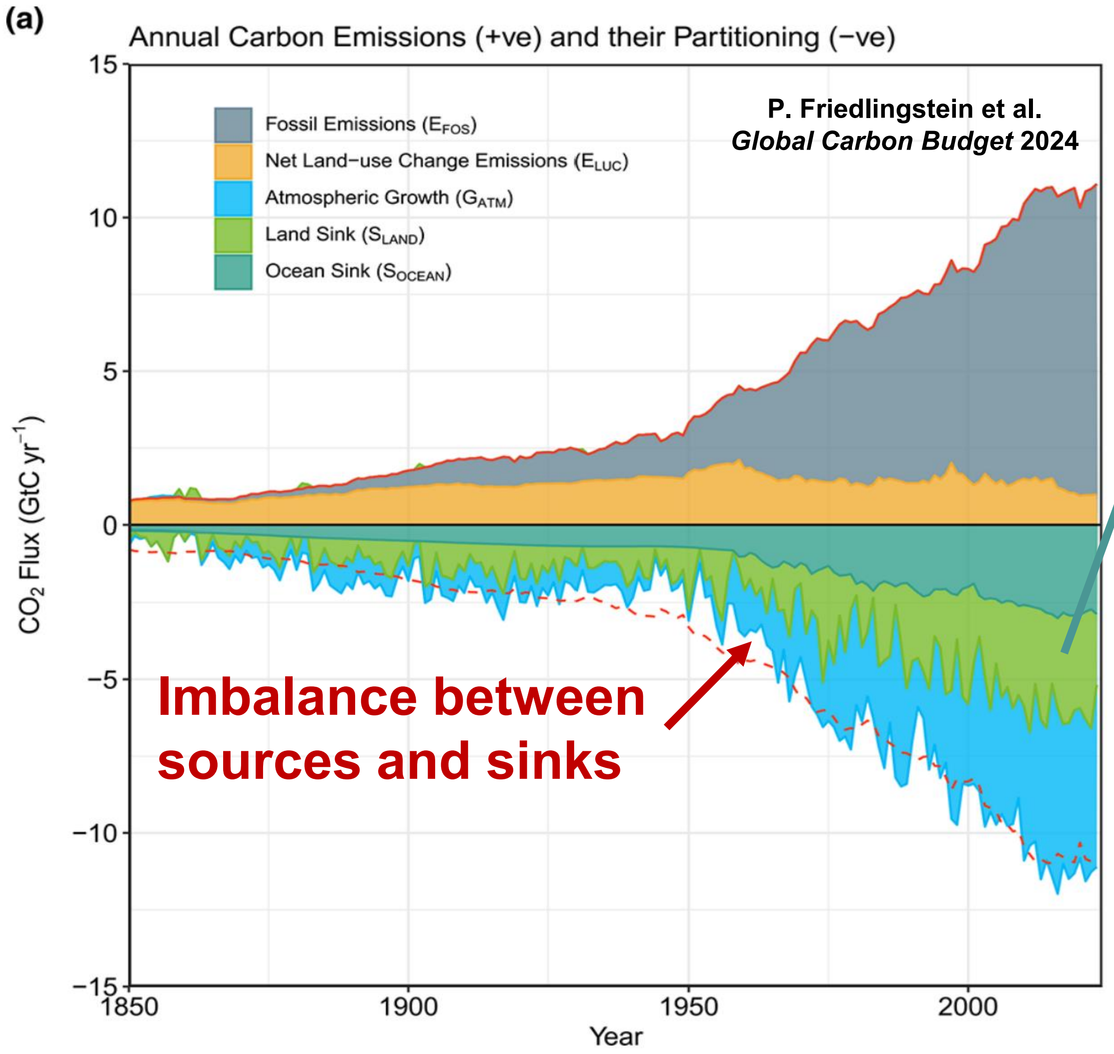
Haodi Xu[1], Joshua Fan[2], Feng Tao[3], Md Nasim[2], Carla P. Gomes[2], Yiqi Luo[1]

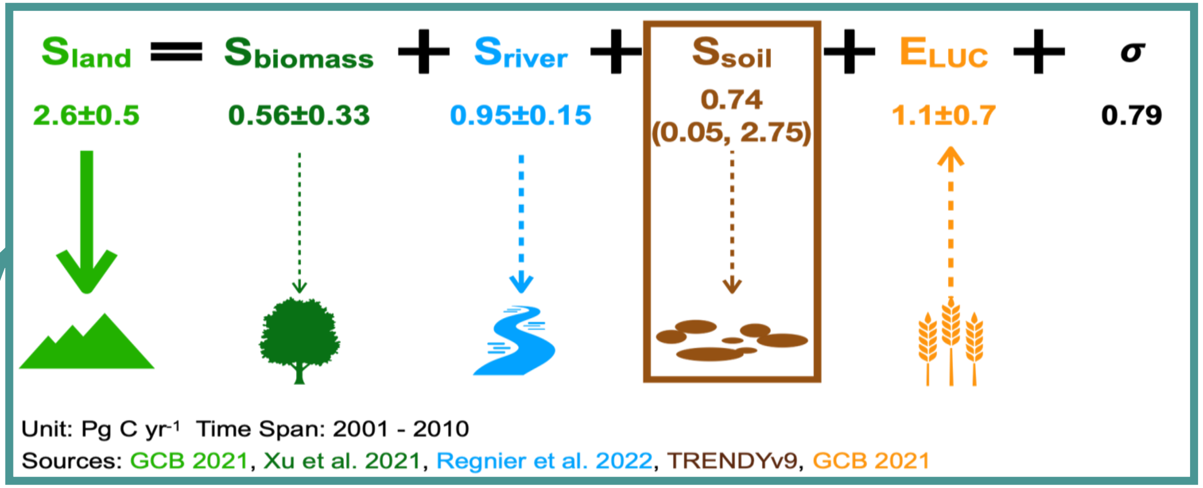
Cornell University, Department of [1]Soil & Crop Science, [2]Computer Science, [3]Ecology & Evolutionary Biology

Cornell University

Haodi Xu [hx293@cornell.edu]

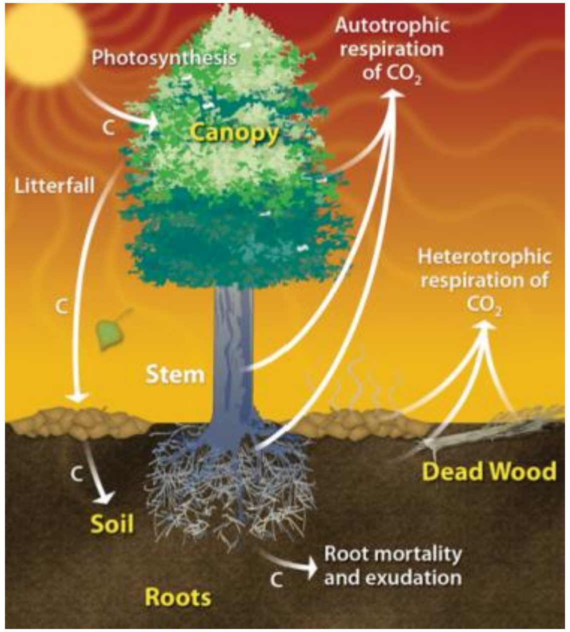
Soil is Critical in Understanding Global Carbon Cycle

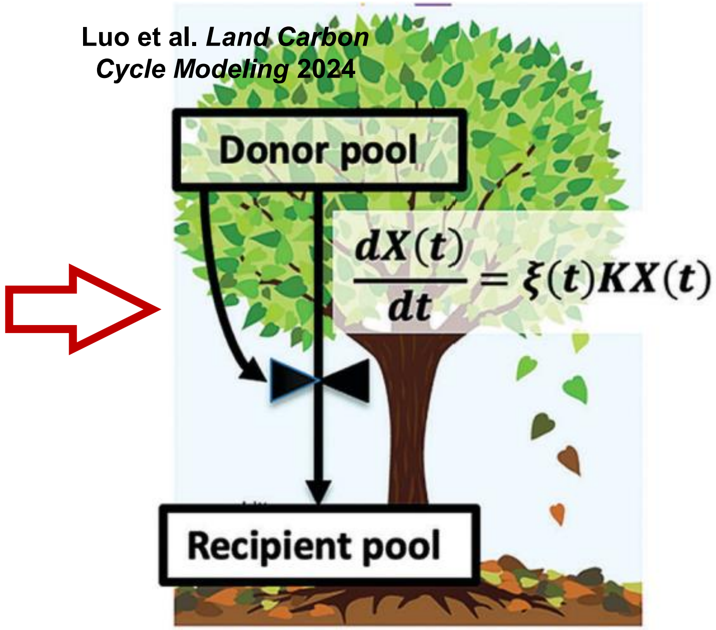


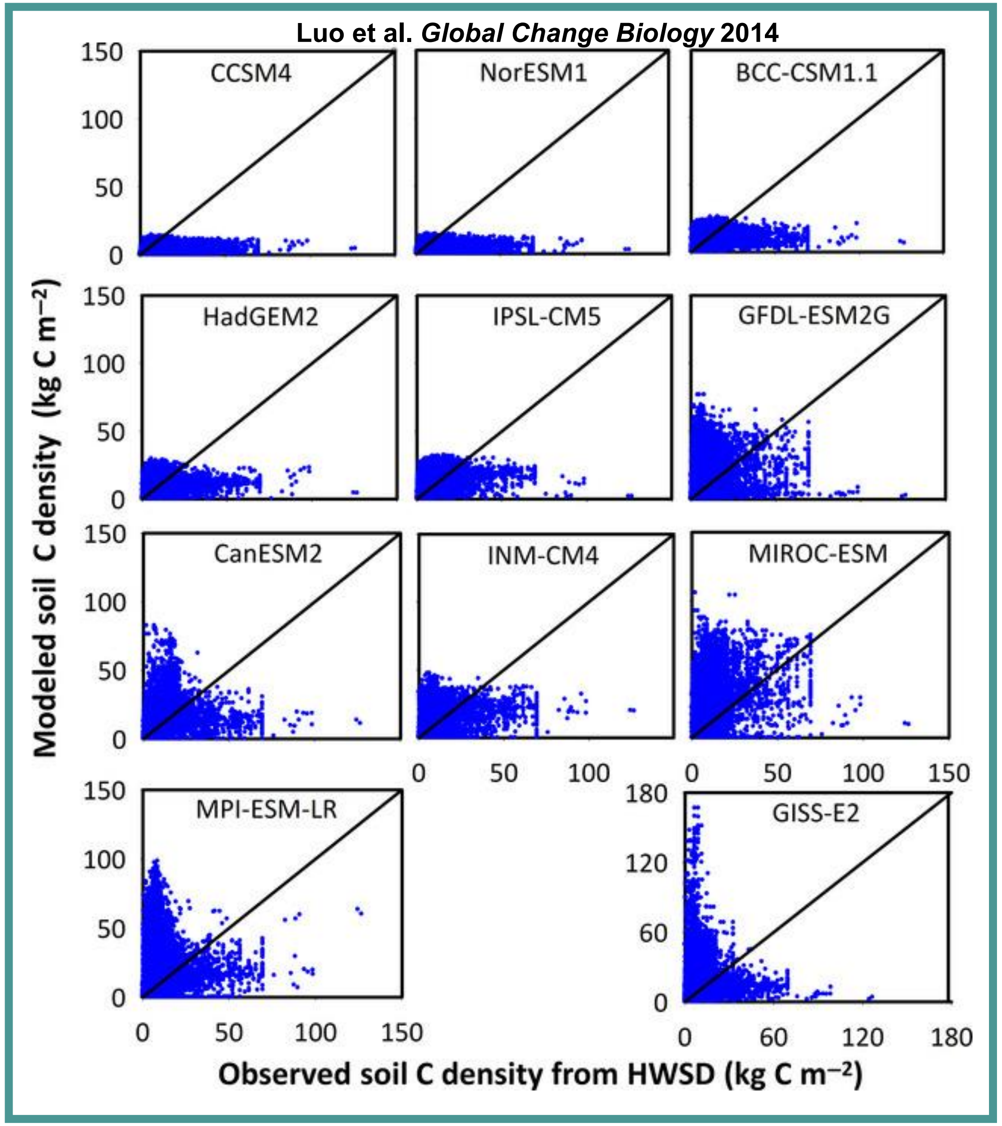


- Imbalance within global carbon cycle can partially explained by the uncertainties in land carbon sinks
- Soils hold a large part of uncertainties in land carbon sinks

Modeling SOC Dynamics

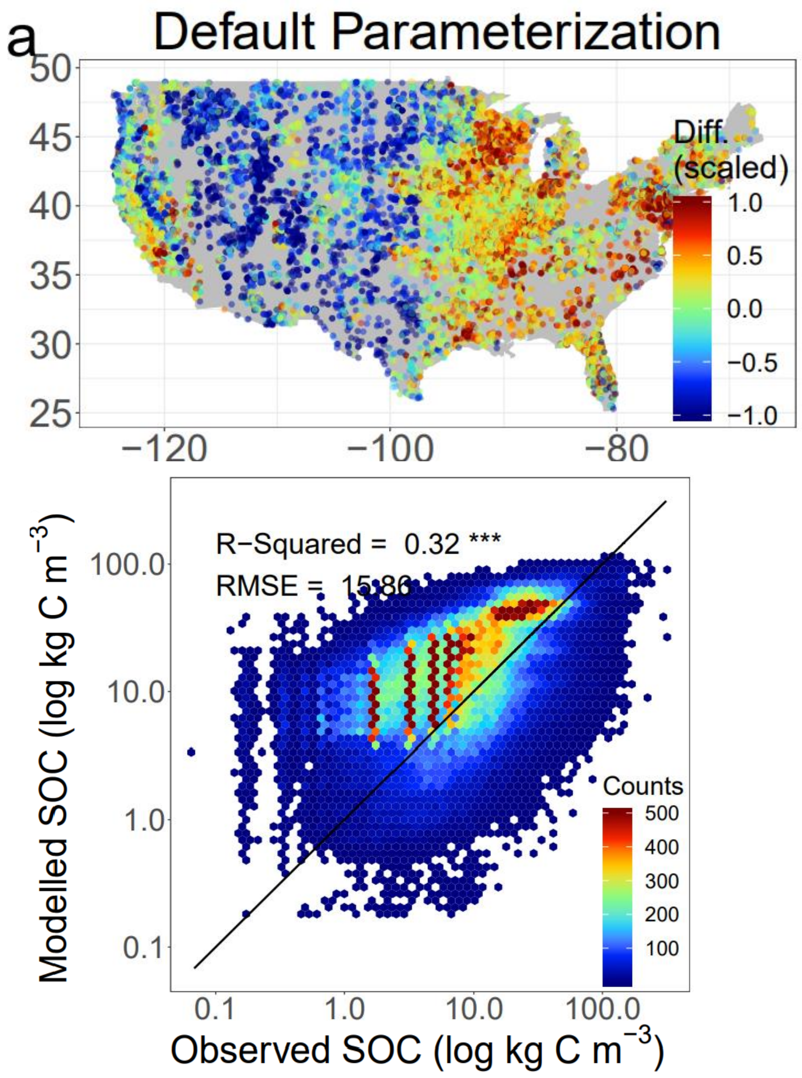


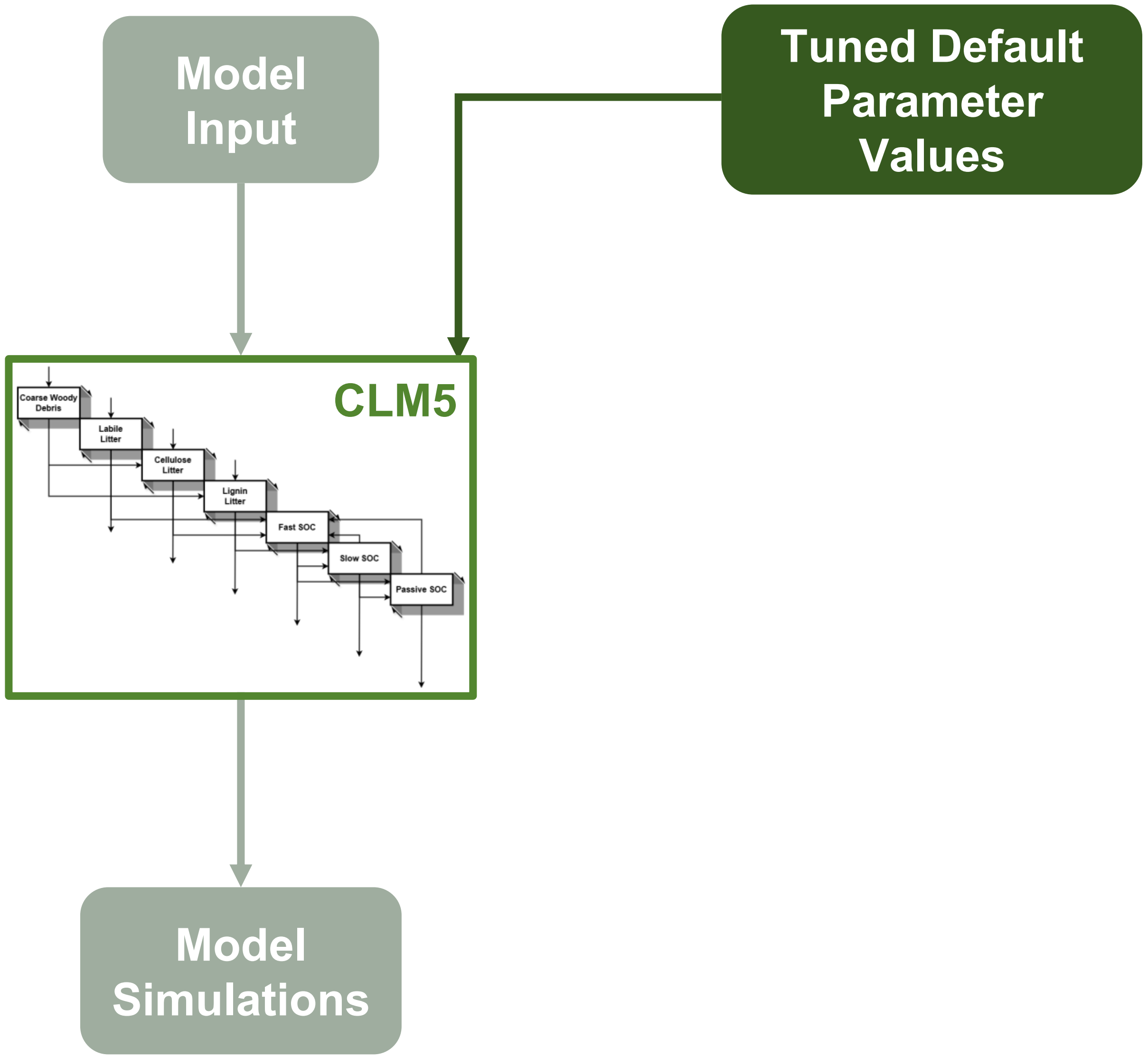




- Our knowledge of SOC processes can be incorporated into process-based models
- Yet high uncertainties remain in capturing the global SOC cycle

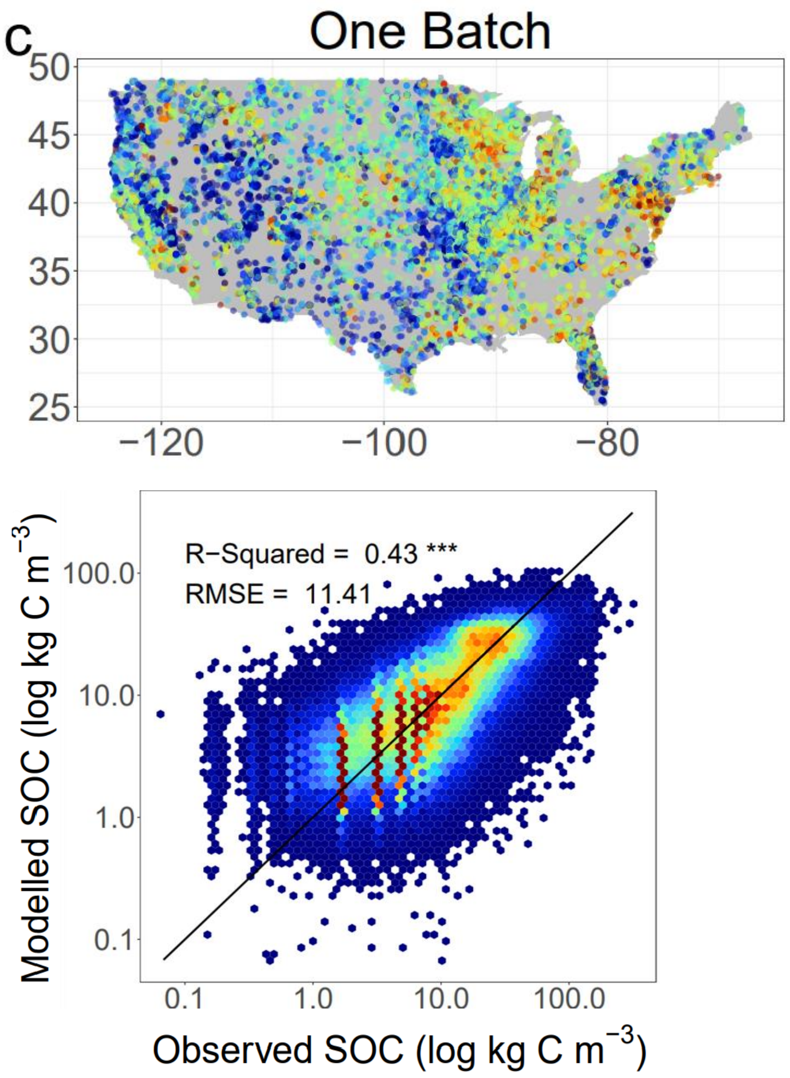
Modeled SOC across Conterminous US

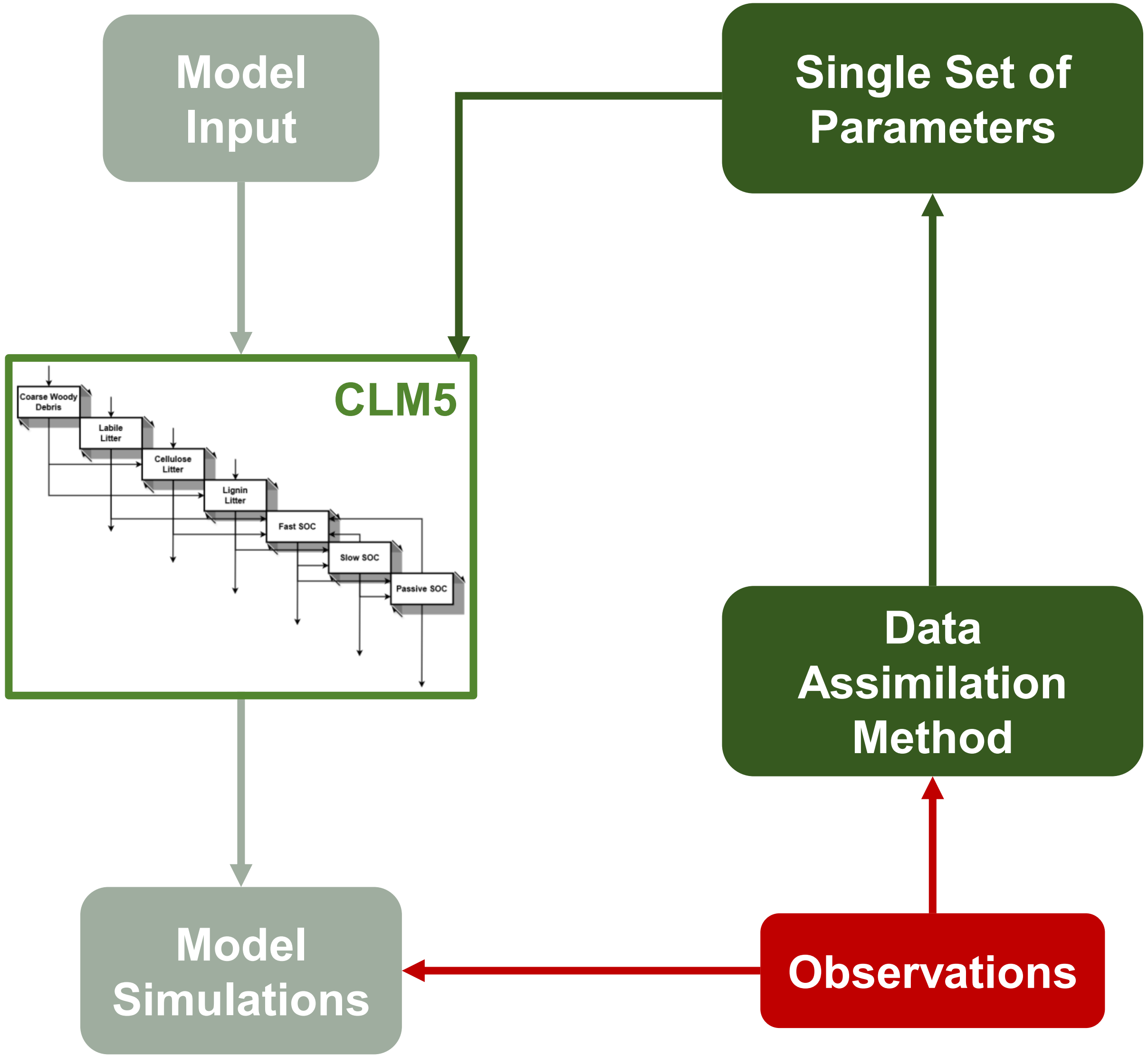




Tao et al. *Frontiers in Big Data* 2020

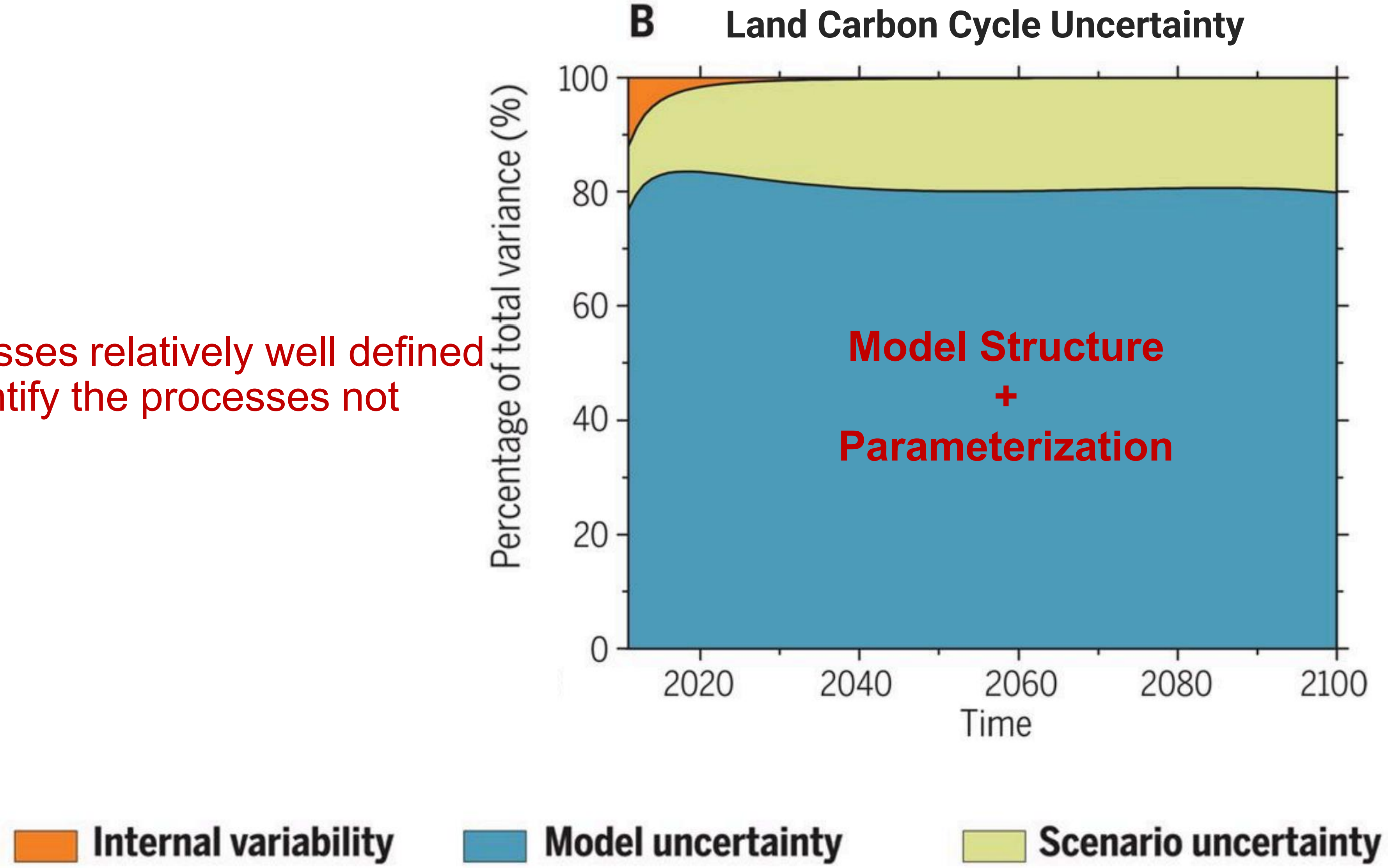
Data Assimilation with One Set of Parameters





Tao et al. *Frontiers in Big Data* 2020

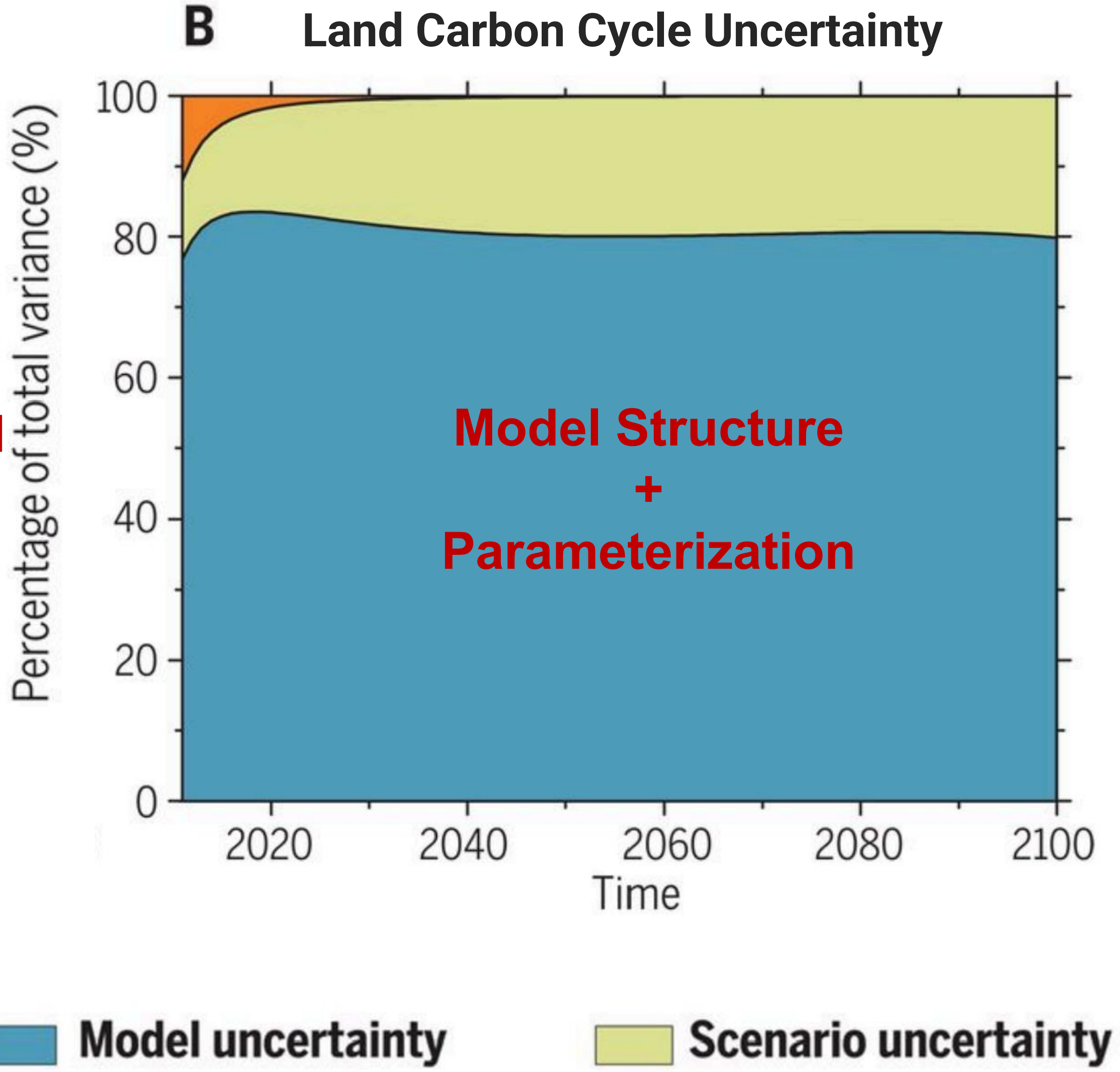
Model Structure: processes relatively well defined
Parameterization: quantify the processes not explicitly represented



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

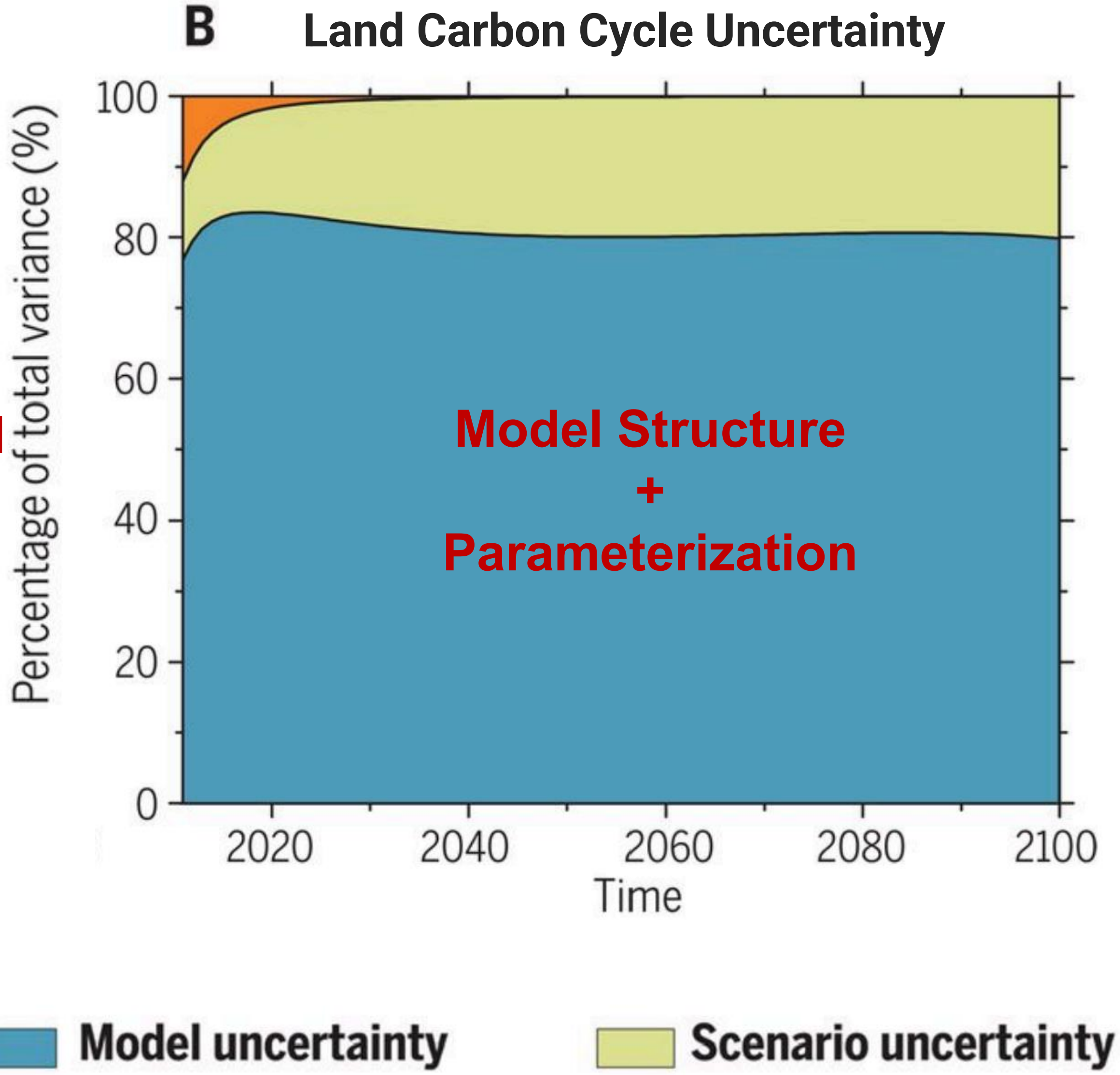


Bonan et al., *Science*, 2018

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

Environmental Information

Climate

Soil pH

Lon, Lat, Elevation

Soil Texture

Bulk Density

Land Cover

⋮

Environmental Information → ? → **Process-based Model Parameters**

Climate

Soil pH

Lon, Lat, Elevation

Soil Texture

Bulk Density

Land Cover

⋮

Artificial Intelligence (Deep Learning)

Environmental Information → **Process-based Model Parameters**

Climate

Soil pH

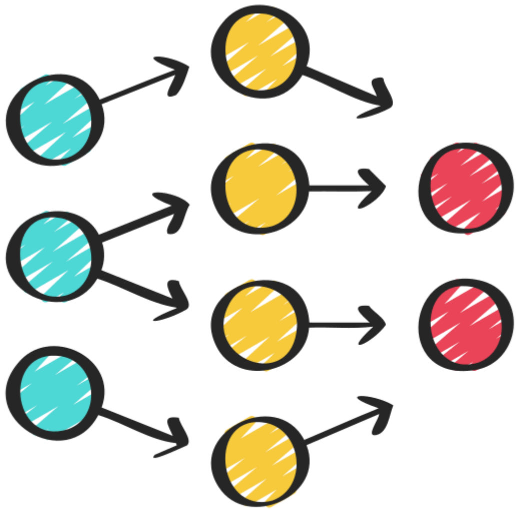
Lon, Lat, Elevation

Soil Texture

Bulk Density

Land Cover

⋮



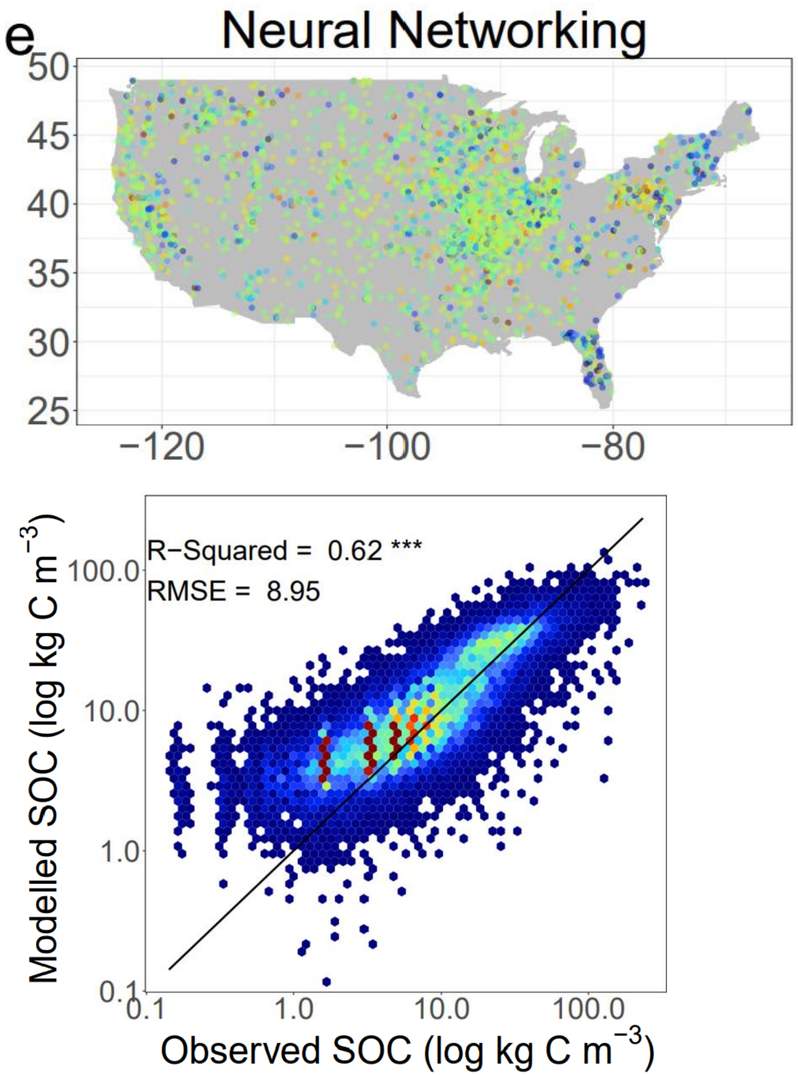
Microbial Carbon Use Efficiency

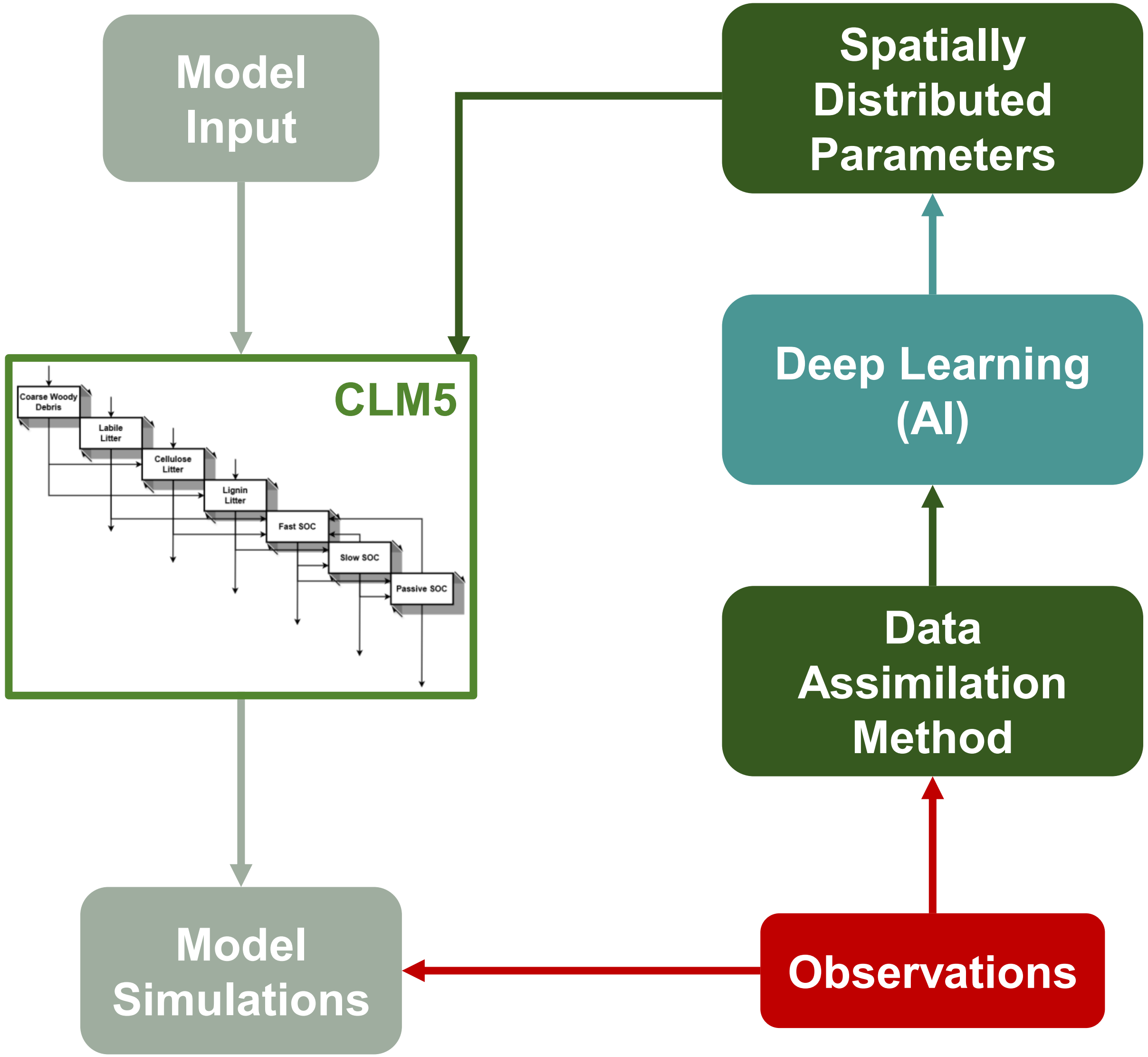
Substrate Decomposability

Temperature Sensitivity

⋮

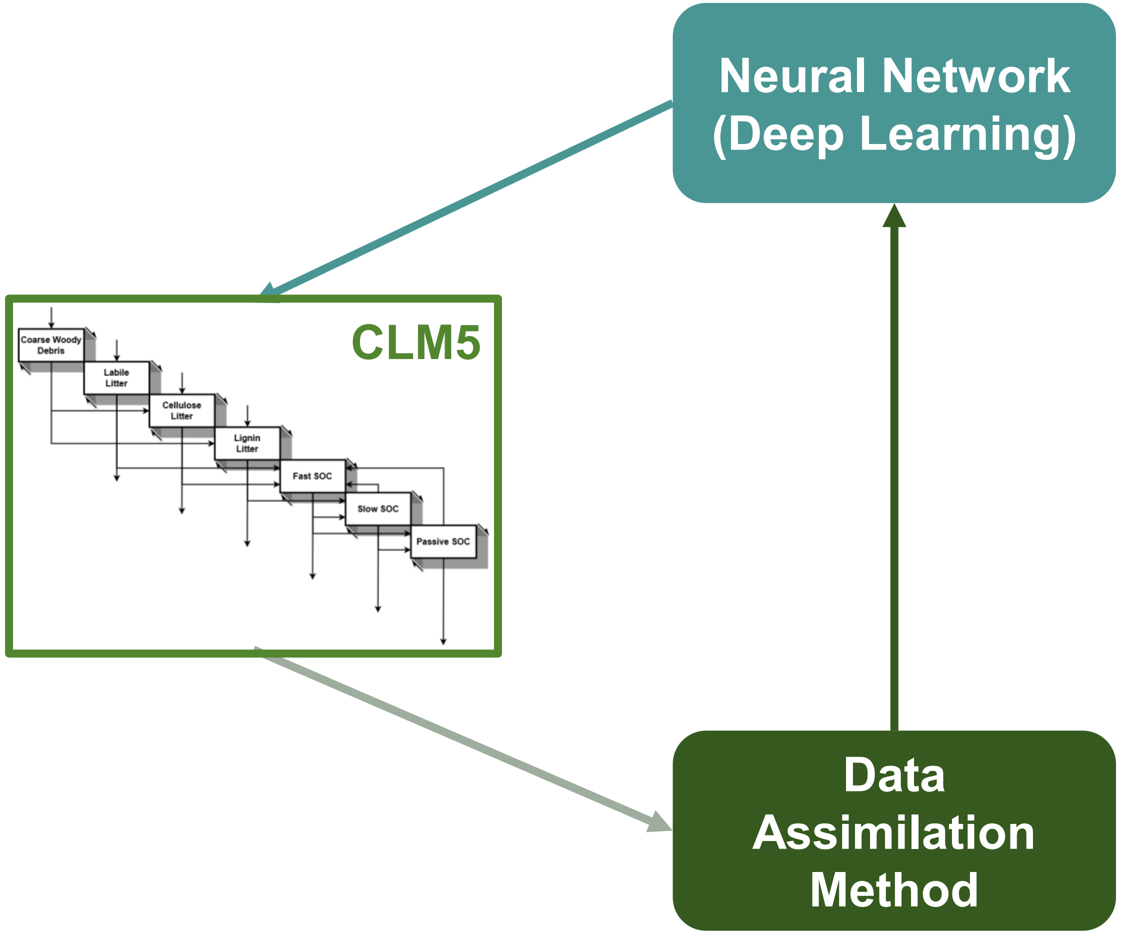
PROcess guided deep learning and DAta driven modeling (PRODA)





Tao et al. *Frontiers in Big Data* 2020

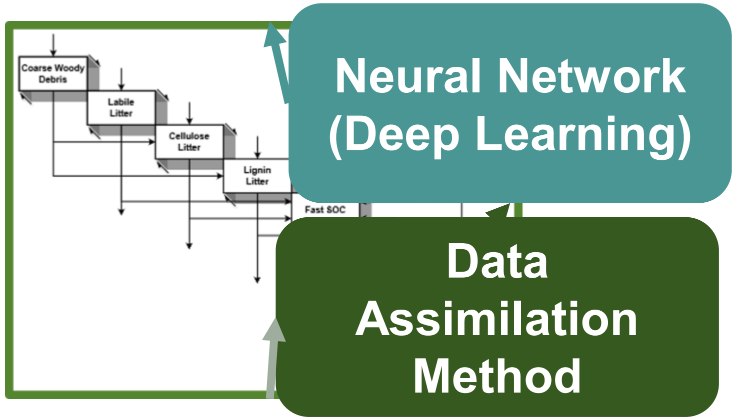
The need to further improve PRODA



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

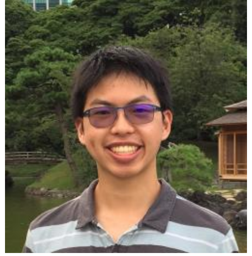
The need to further improve PRODA



Combine these altogether?

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



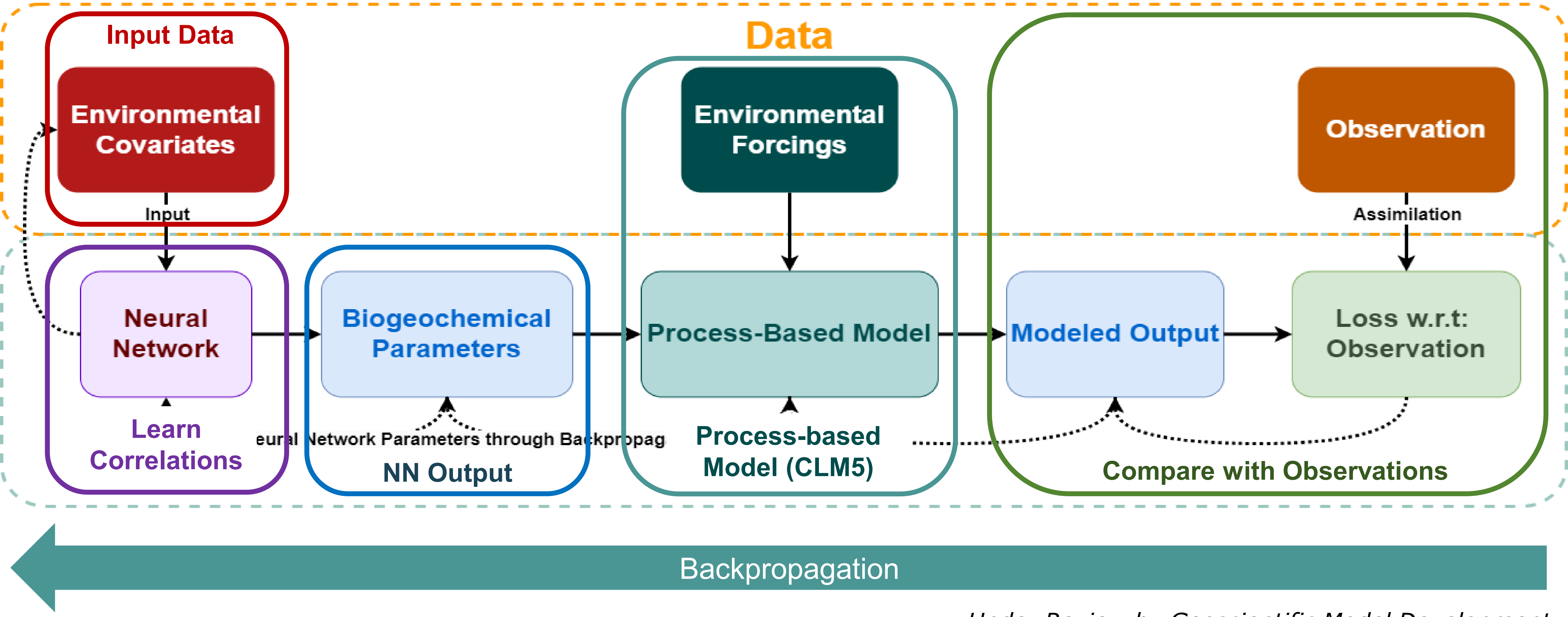
A Biogeochemistry-Informed Neural Network (BINN)

Cornell CALS College of Agriculture and Life Sciences

Haodi Xu [hx293@cornell.edu]

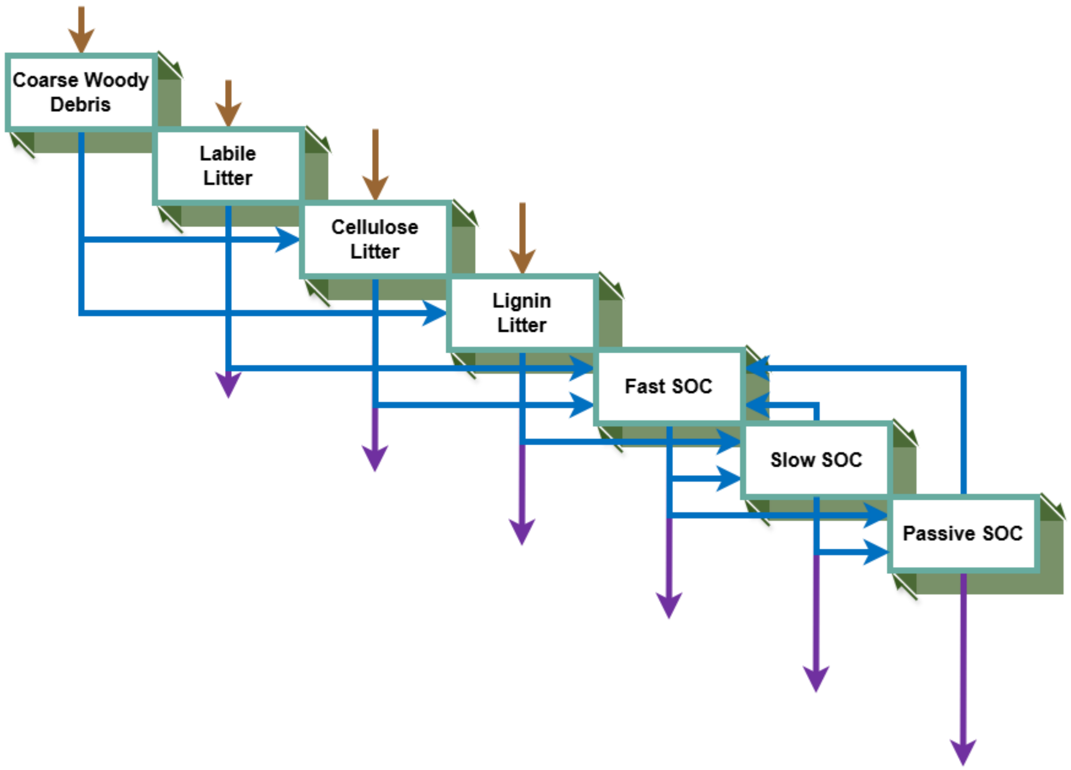
ECOLAB OF DR. YIQI LUO

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

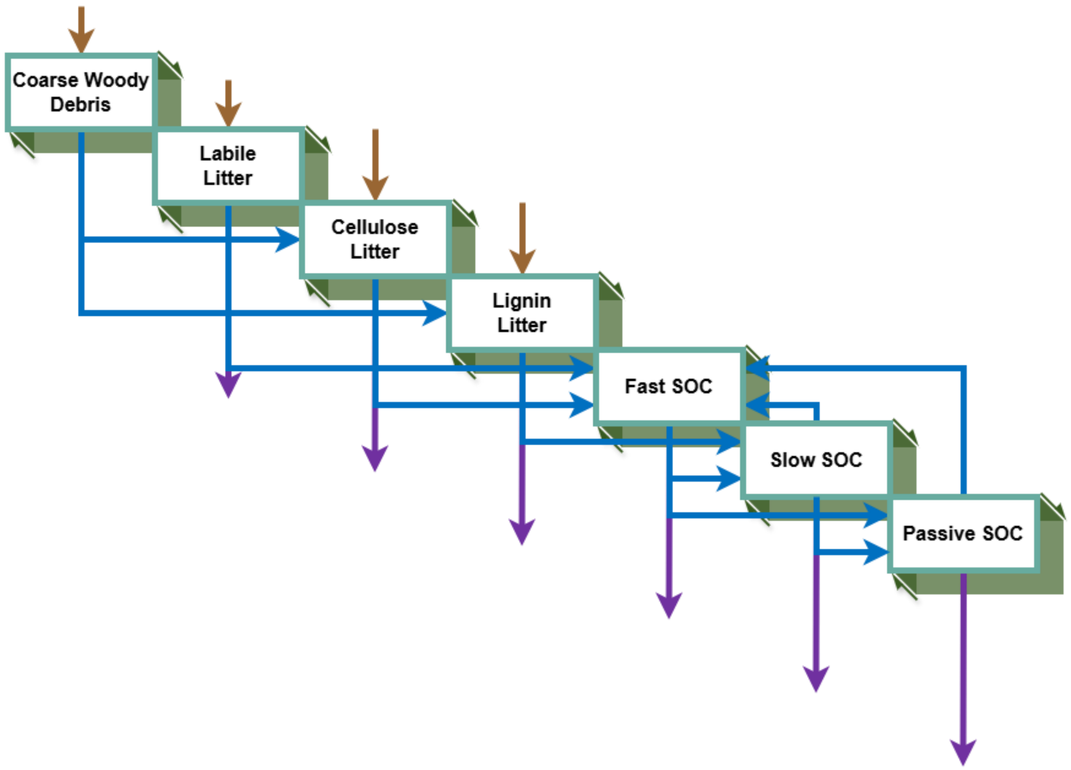
$$\boxed{\frac{dX(t)}{dt}} = \boxed{B(t)I(t)} - \boxed{A}\boxed{\xi(t)KX(t)} - \boxed{V(t)X(t)}$$

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

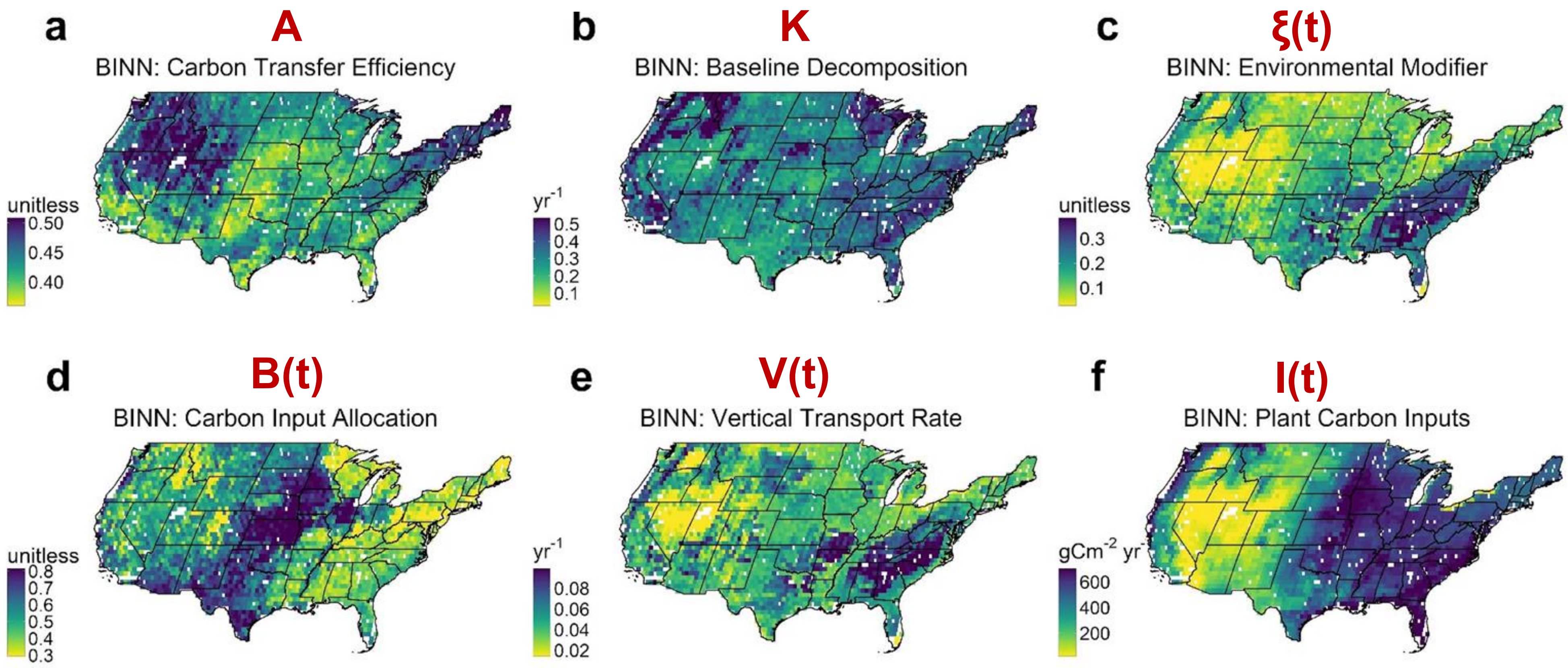
$$\frac{dX(t)}{dt} = B(t)I(t) - A\xi(t)KX(t) - V(t)X(t)$$

Carbon Input

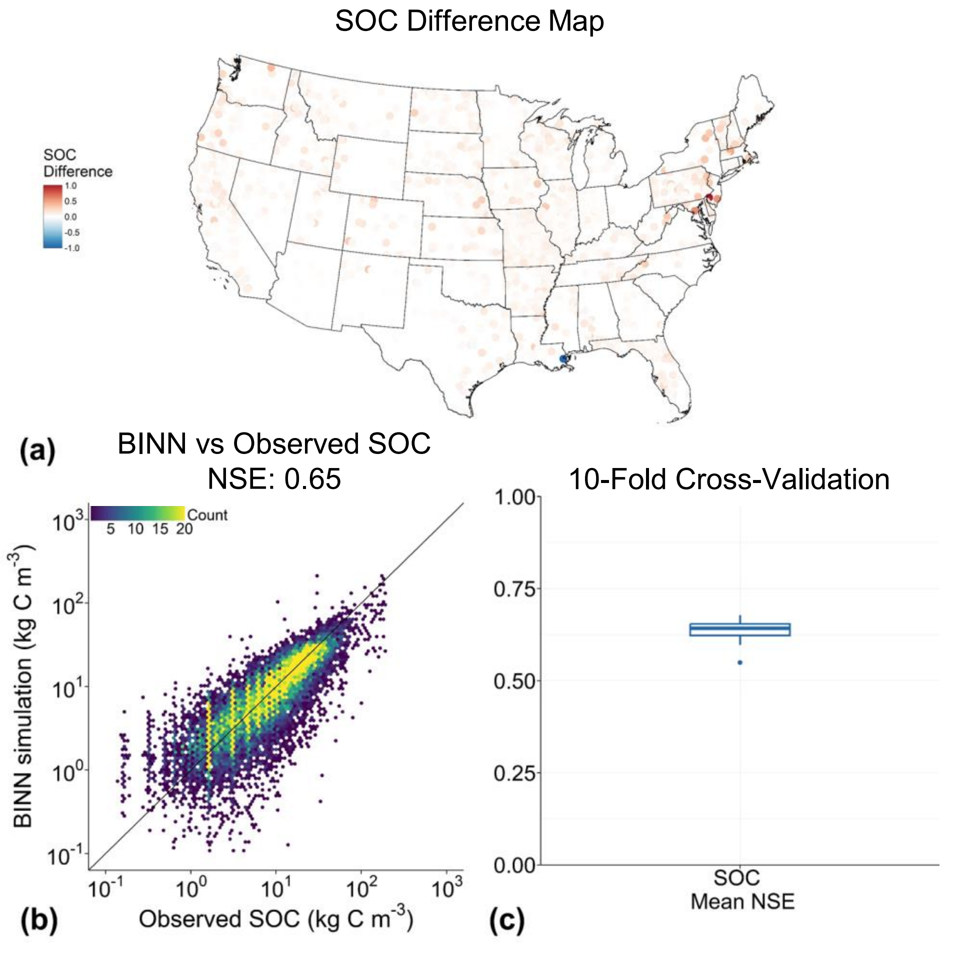
The intrinsic decomposition rate of each C pool, modified by the environmental scalar

: Model Components, quantified by 21 model parameters

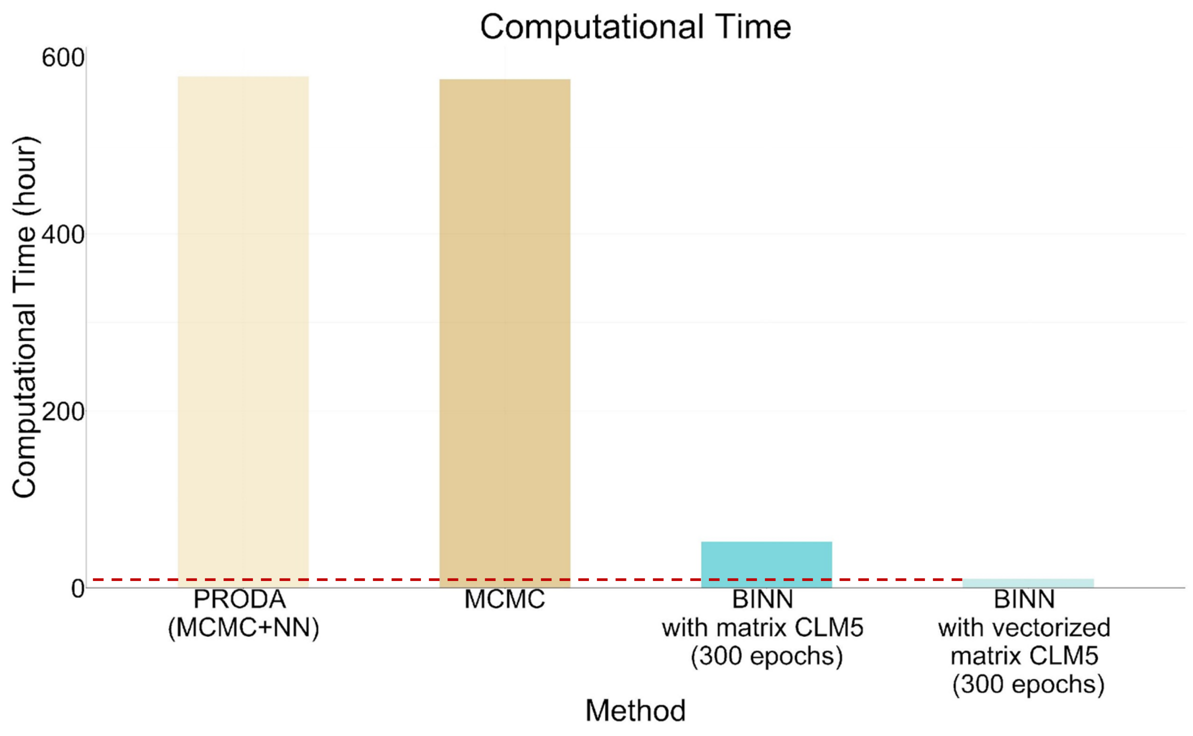
BINN discovers the spatial distribution of underlying mechanisms from SOC observations and environmental covariates (without any data for these mechanisms)

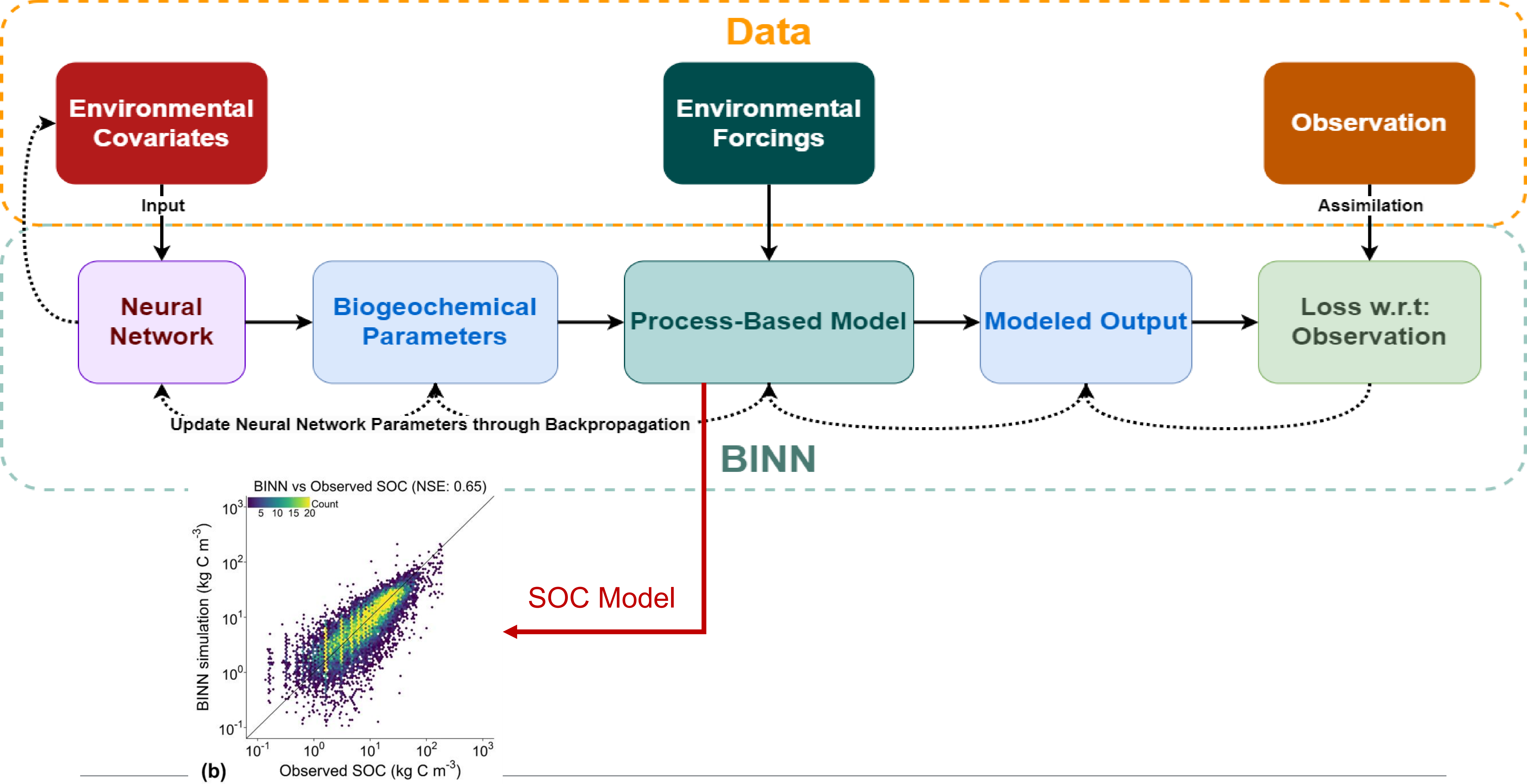


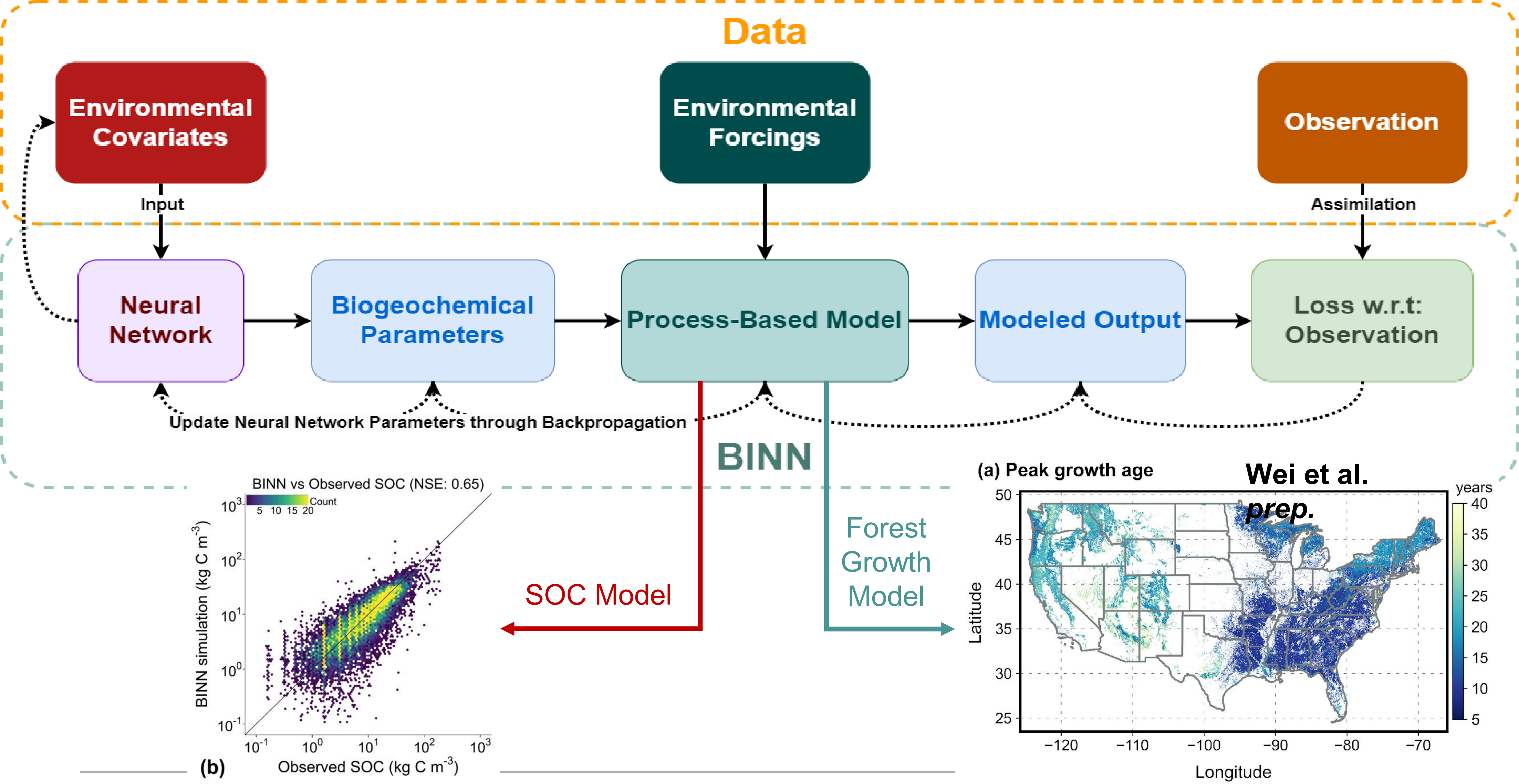
More Realistic Representations of SOC Distribution

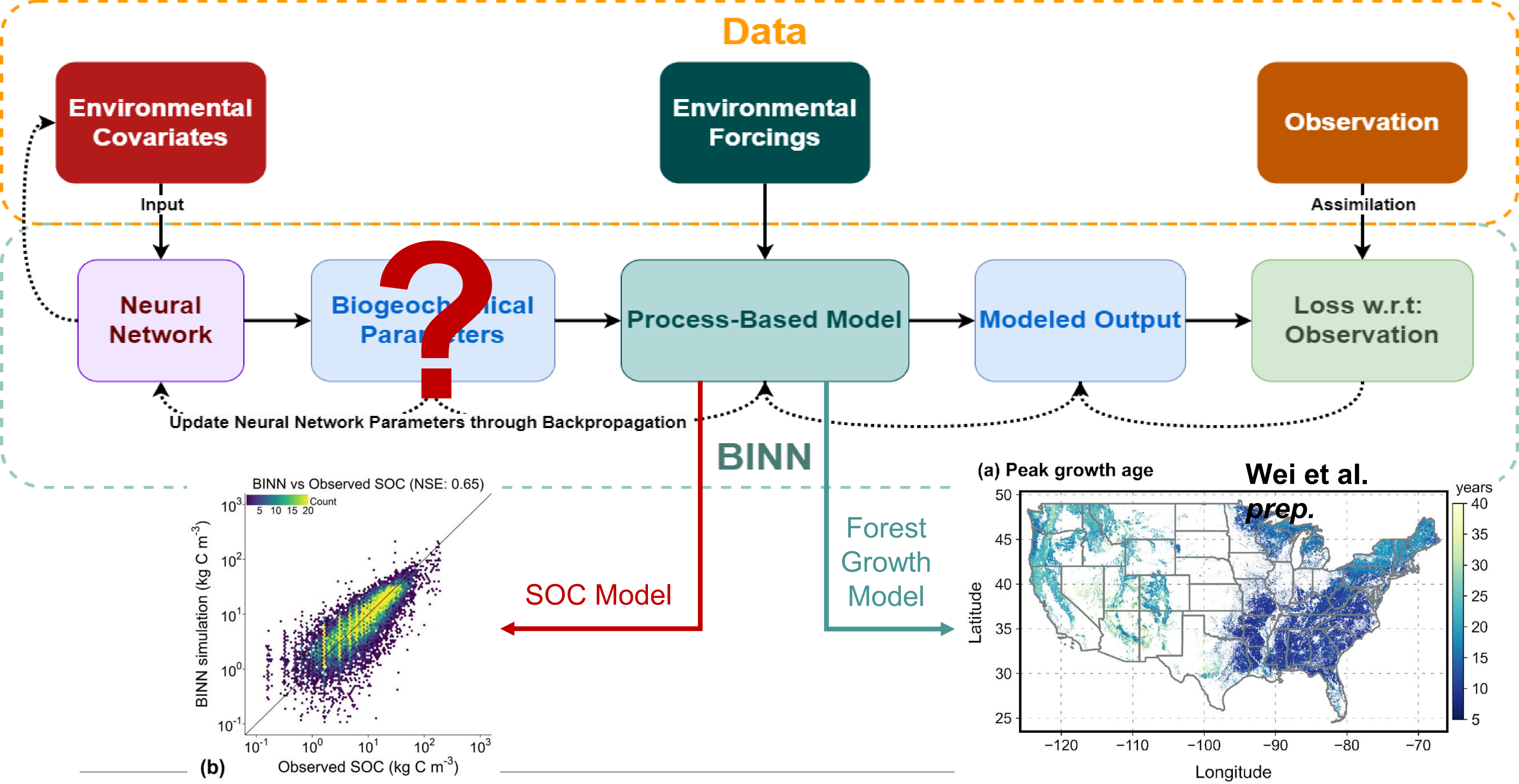


Way Less Computational Requirements (More than 50 times)

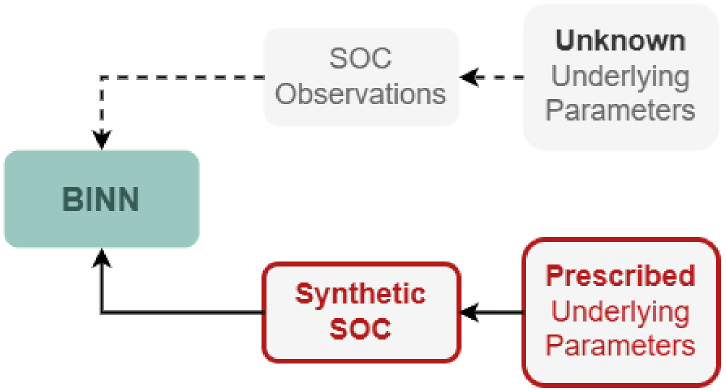


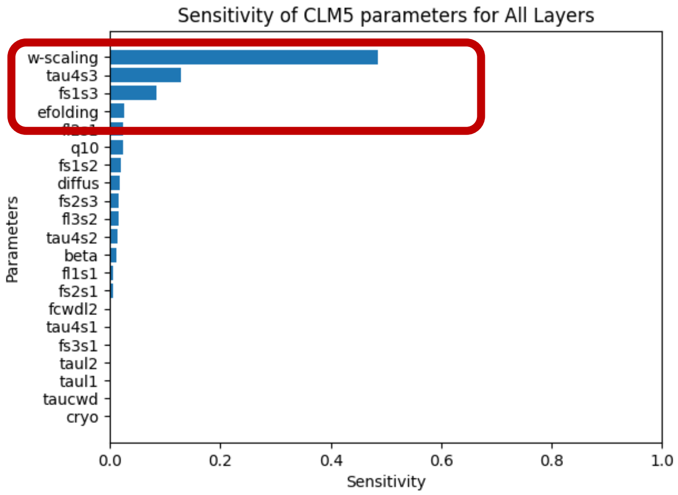

Data
Environmental Covariates
Environmental Forcings
Observation
Input
Assimilation
Neural Network
Biogeochemical Parameters
Process-Based Model
Modeled Output
Loss w.r.t: Observation
Update Neural Network Parameters through Backpropagation
BINN
SOC Model
BINN vs Observed SOC (NSE: 0.65)
Count
5 10 15 20
BINN simulation (kg C m-3)
Observed SOC (kg C m-3)
(b)

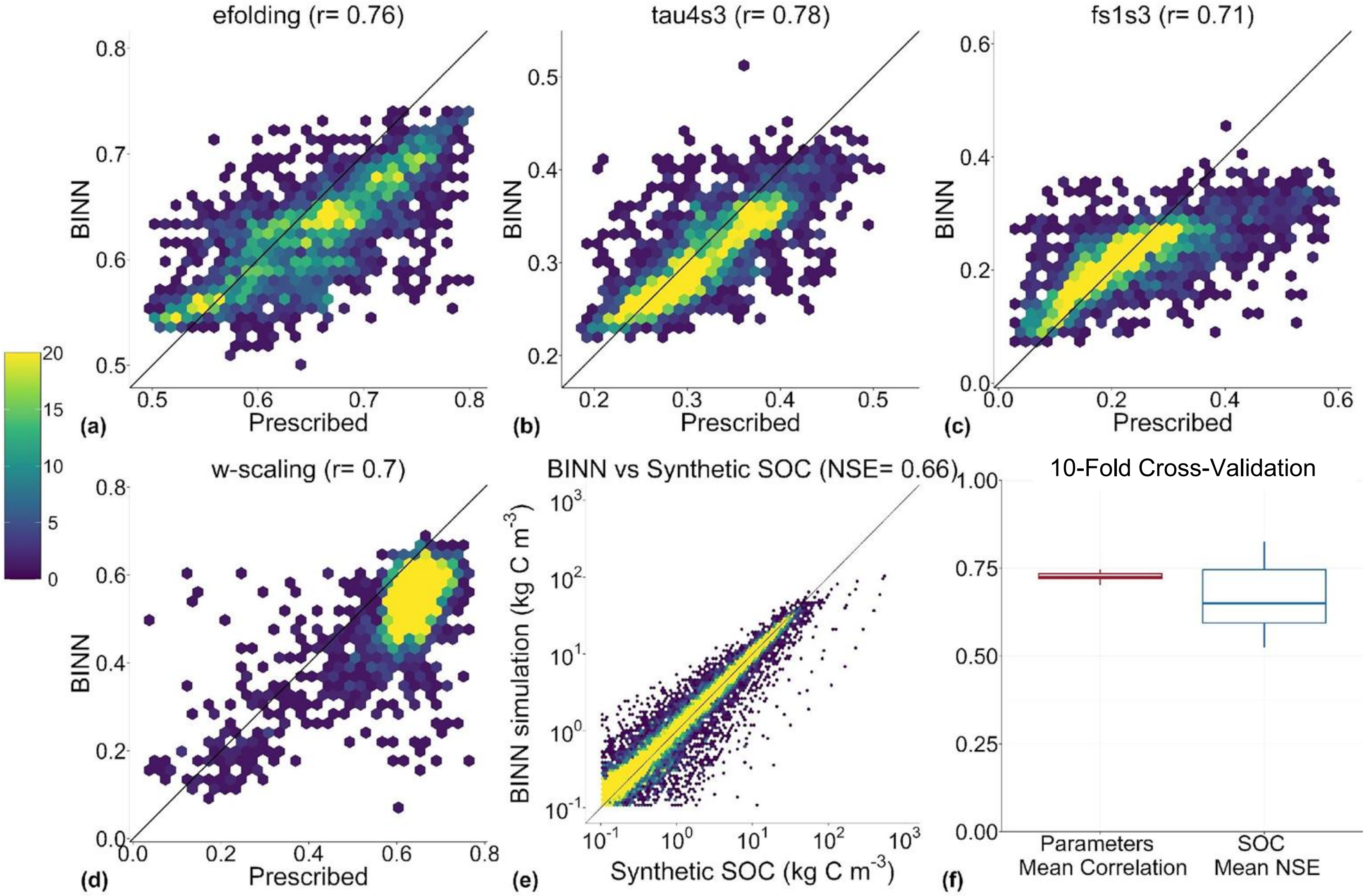

Data
Environmental Covariates
Environmental Forcings
Observation
Input
Assimilation
Neural Network
Biogeochemical Parameters
Process-Based Model
Modeled Output
Loss w.r.t: Observation
Update Neural Network Parameters through Backpropagation
BINN
SOC Model
Forest Growth Model
BINN vs Observed SOC (NSE: 0.65)
Count
5 10 15 20
BINN simulation (kg C m^{-3})
Observed SOC (kg C m^{-3})
10^{-1} 10^{0} 10^{1} 10^{2} 10^{3}
(b)
(a) Peak growth age
Wei et al. prep.
years
40 35 30 25 20 15 10 5
Latitude
50 45 40 35 30 25
Longitude
−120 −110 −100 −90 −80 −70

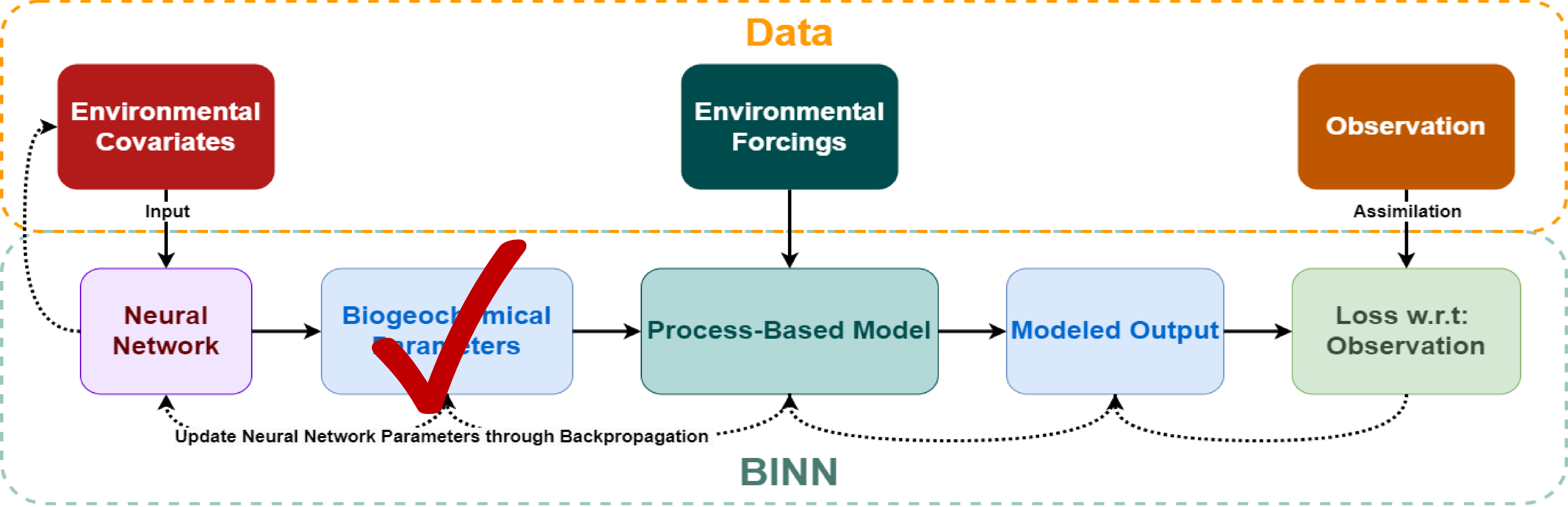

Data
Environmental Covariates
Environmental Forcings
Observation
Input
Assimilation
Neural Network
Biogeochemical Parameters
?
Process-Based Model
Modeled Output
Loss w.r.t: Observation
Update Neural Network Parameters through Backpropagation
BINN
SOC Model
Forest Growth Model
BINN vs Observed SOC (NSE: 0.65)
Count
5 10 15 20
BINN simulation (kg C m-3)
Observed SOC (kg C m-3)
(b)
(a) Peak growth age
Wei et al. prep.
years
Latitude
Longitude

Recovery Experiment with 4 Most Influential Parameters



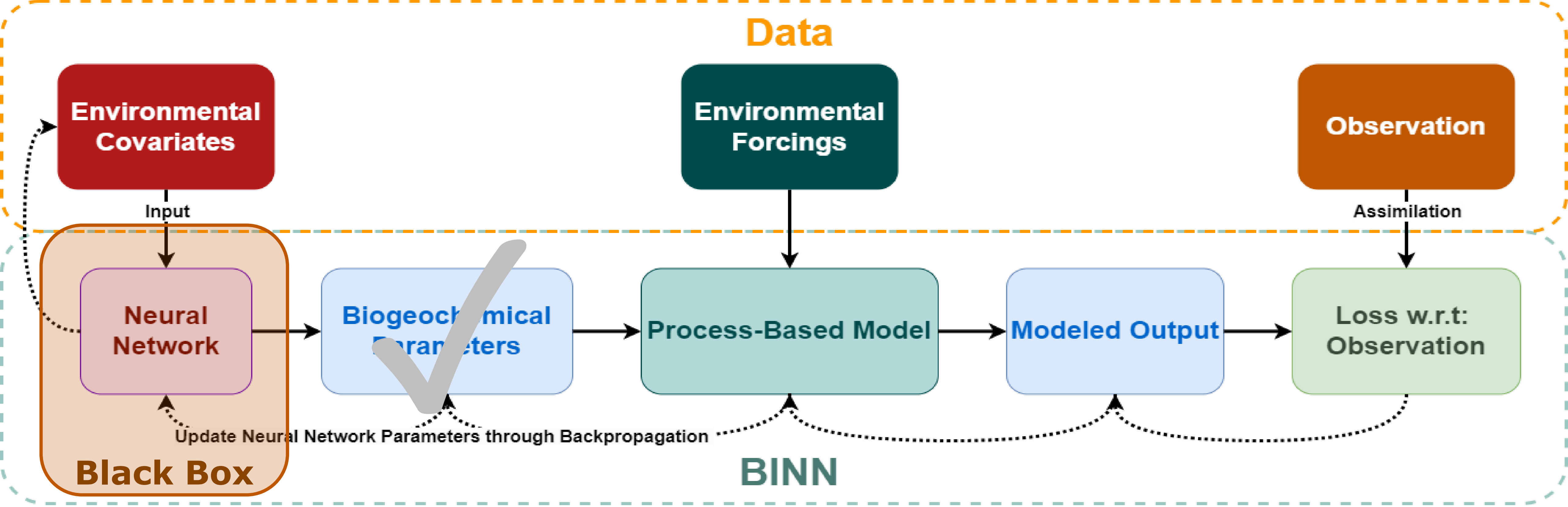


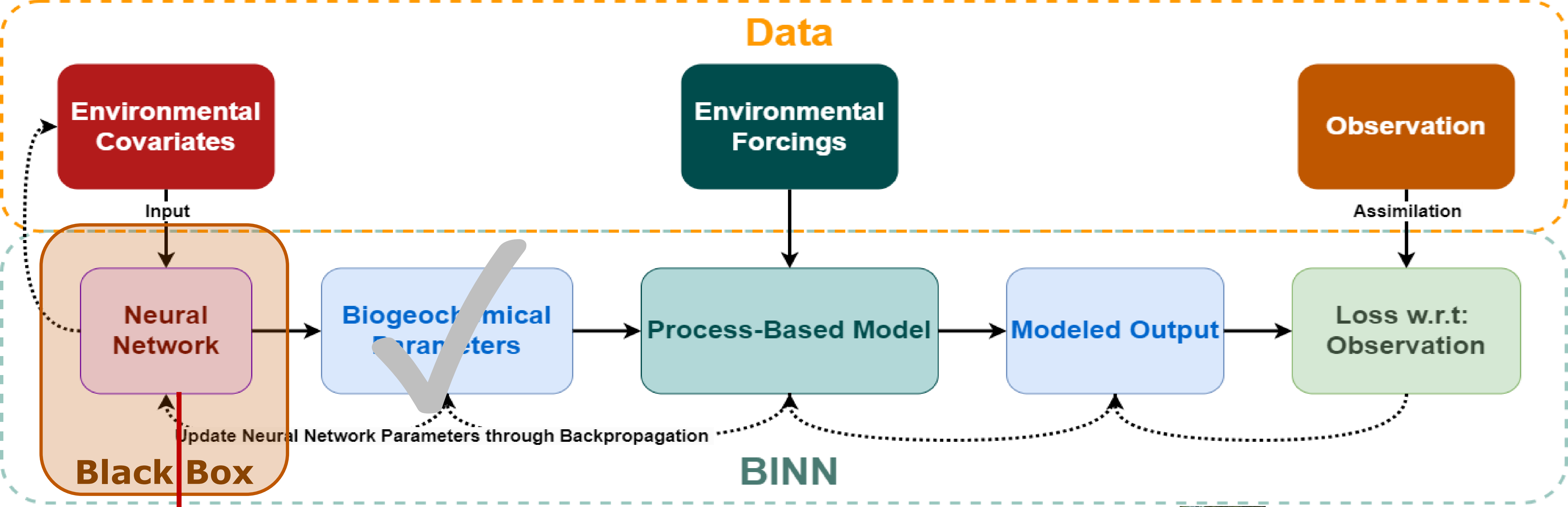




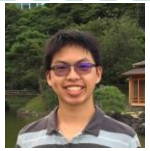
Recover causes from data (biogeochemical parameters)

Discover emerging mechanisms governing SOC from big data (observations)


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
Black Box
BINN



Kolmogorov Arnold Network (KAN)



An alternative to neural networks that is easier to interpret

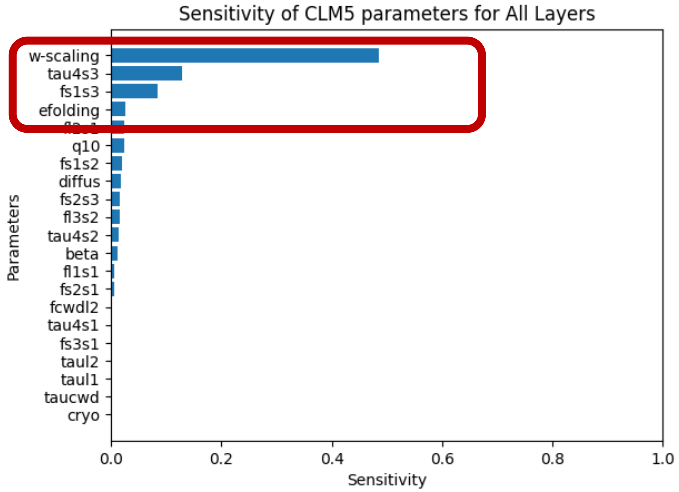
Scientifically-Interpretable Reasoning Network(ScIReN)

Submitted to NeurIPS 2025

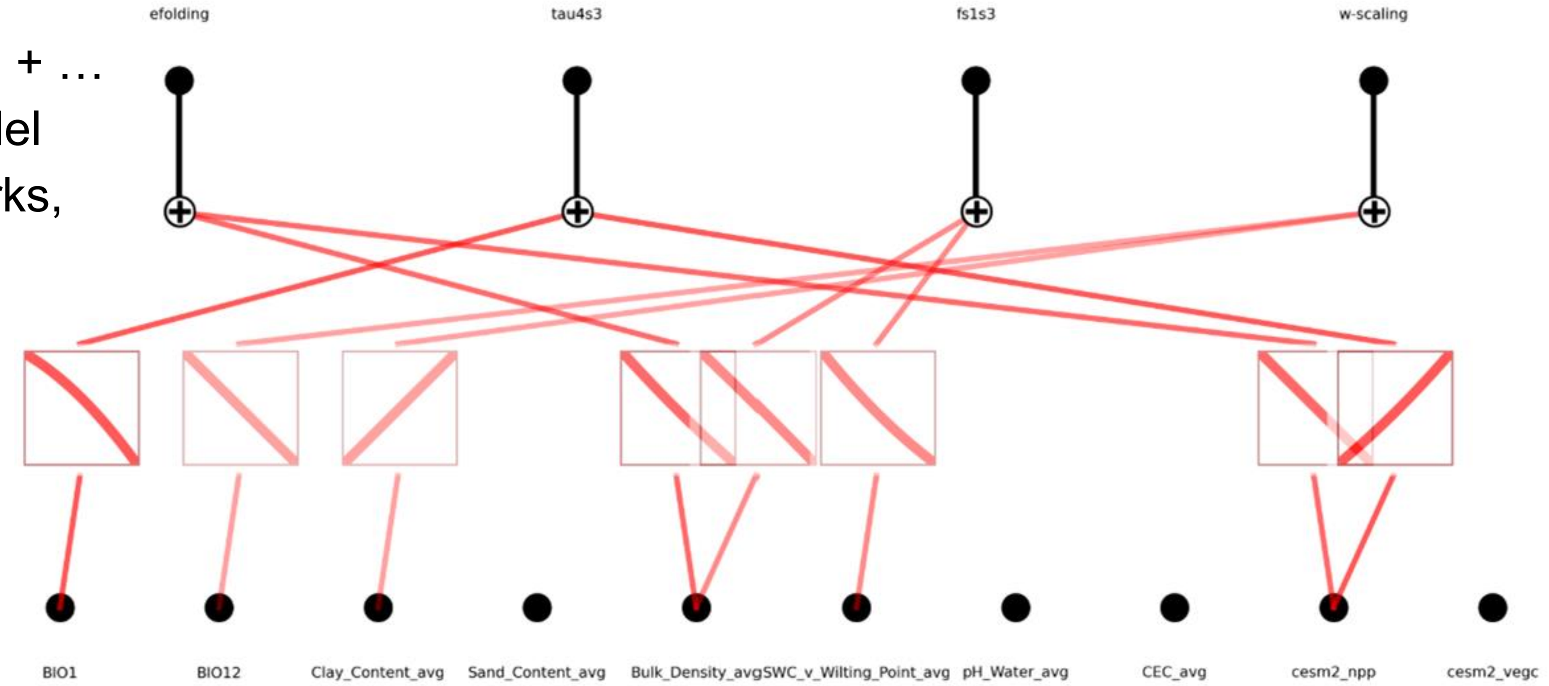
Haodi Xu [hx293@cornell.edu]

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 = $(\text{input1})^2 + \exp(\text{input6}) + \ldots$
- 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

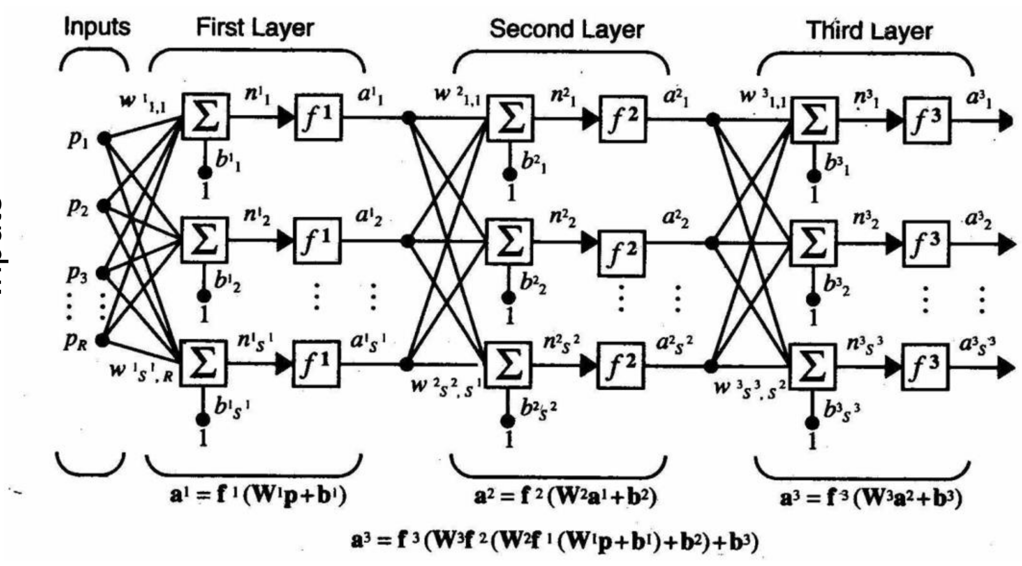


Inputs (Environmental Covariates) - observed

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)

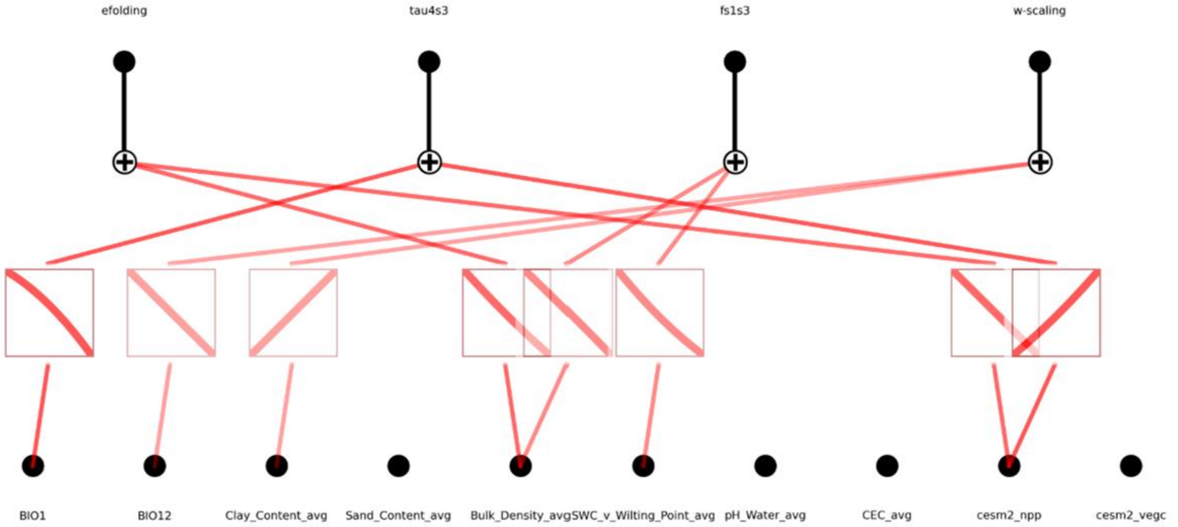
Inputs



Outputs

- **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

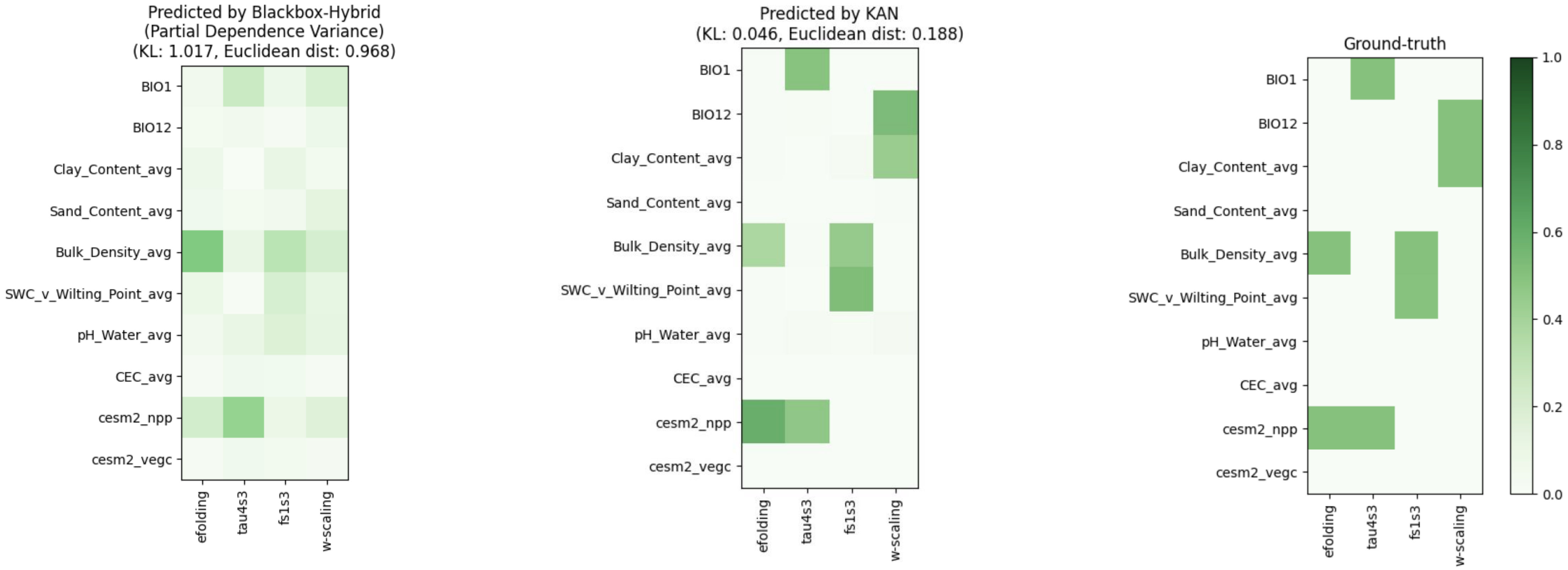
Outputs



Inputs

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.



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

Cornell CALS College of Agriculture and Life Sciences

ECOLAB OF DR. YIQI LUO

BINN Tutorial

- Access: Open ***binns_KGML_workshop.ipynb*** in the *BINN_Demo* GitHub repo
- Recommended Run Environment: **Google Colab** (simply by clicking click the **Open in Colab** button in the notebook Open in Colab)
- When running the first setup cell to setup the environment, you may be asked to install one or more package. If you are using Google Colab, you will get prompts to install the required packages.
- Try to work through the notebook to learn BINN's structure. Then consider how you'd swap in your own process-based model (or process-based model in your field). What changes you might need to make?