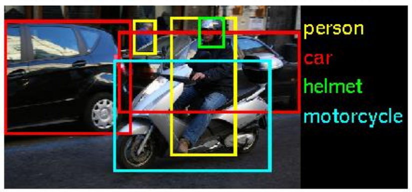
Machine Learning for Scientific Knowledge Discovery

Xiaowei Jia
University of Pittsburgh
Xiaowei@pitt.edu

Promise of Machine Learning in Transforming Scientific Knowledge Discovery

- Success of machine learning in commercial applications

IMAGENET



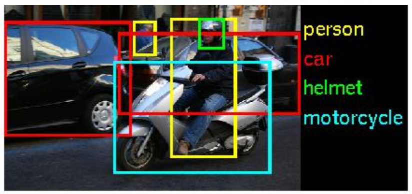
facebook

NETFLIX

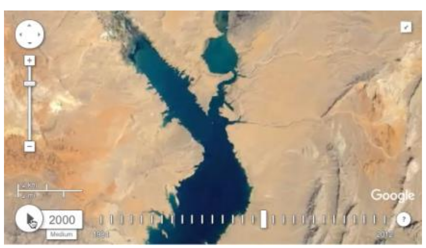
Google Ads

Promise of Machine Learning in Transforming Scientific Knowledge Discovery

- Success of machine learning in commercial applications

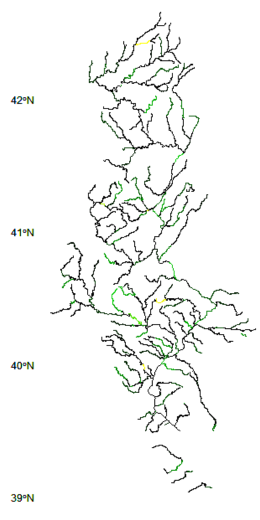


- Applications with scientific and societal relevance

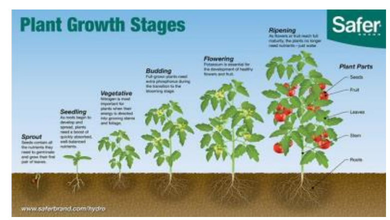


Management of water resources

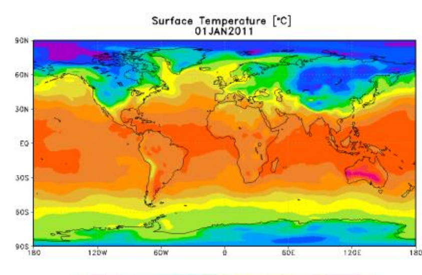
Phosphorus modeling



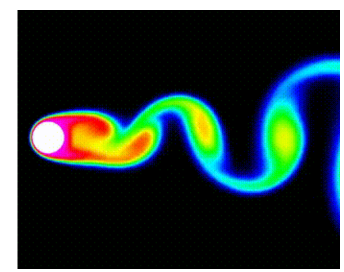
Ungauged basin



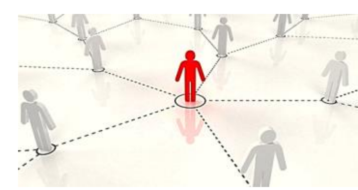
Food supply



Climate science



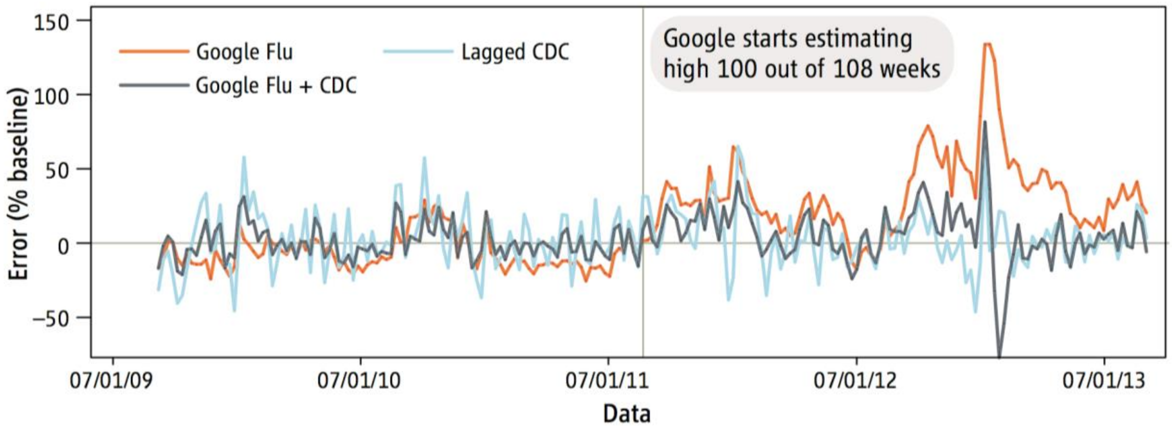
Turbulent flow



Epidemiology

Limits of “Black-box” Machine Learning Methods

- Rise and Fall of Google Flu Trends
  - Predicted flu occurrences using Google search queries
  - Overestimated by a factor of two in later years



  - Lazer, David, et al. "The Parable of Google Flu: Traps in Big Data Analysis." **Science** 343 (2014): 1203-1205.

- Similar observations in other scientific domains:

Climate Science:

Caldwell et al. "Statistical significance of climate sensitivity predictors obtained by data mining.” **Geophysical Research Letters** (2014)

The New York Times

The Opinion Pages | OP-ED CONTRIBUTORS

Eight (No, Nine!) Problems With Big Data

By GARY MARCUS and ERNEST DAVIS APRIL 6, 2014

" . . . you will always need to start with an analysis that relies on an understanding of physics and biochemistry."

Why Do "Black-box" Methods Fail?

- Scientific problems are often under-constrained
  - Scientific problems involve large amount of variables and relationships between variables are "non-stationary"

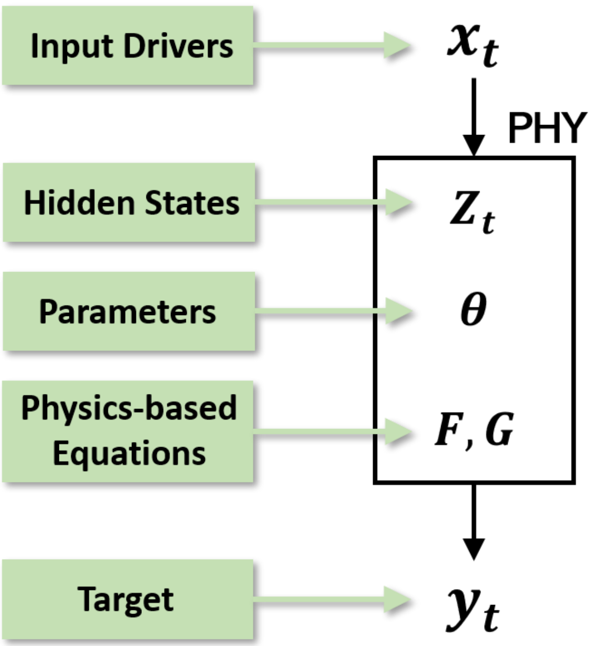
- Black-box methods can only learn from *examples*
  - Results are inconsistent with known physics (*e.g.*, conservation of energy or mass)
  - Available data sets are not always fully representative and black-box models are easy to find spurious patterns in data that do not generalize

- Paucity of labeled data

Huge number of samples is critical to success of deep learning

Physics-based Models of Dynamical Systems

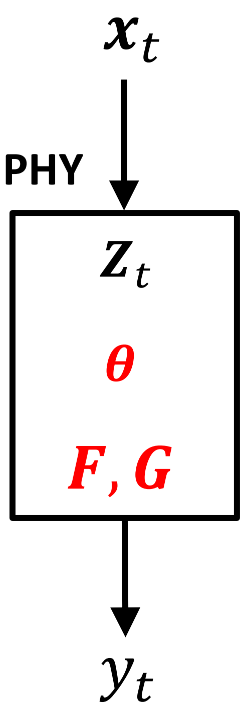
- Relationships b/w input & output variables governed by physics-based partial differential equations (PDEs)



Examples from Hydrology, Limnology, Fluid Dynamics, …

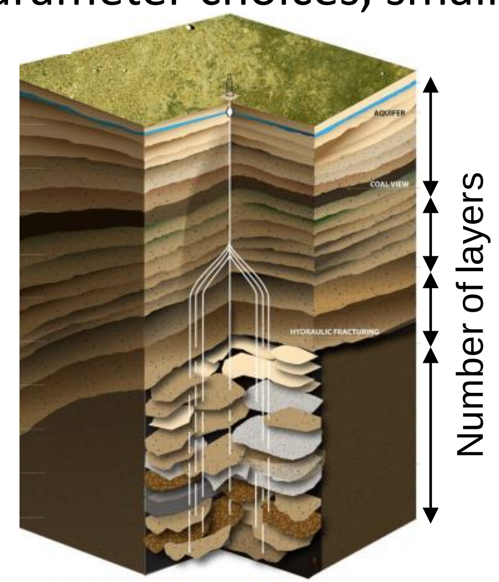
Input	Output	Parameters
Rainfall, topography, land use, river width	River discharge	Soil conductivity, channel flow
Solar radiation, air temp, wind speed	Lake quality	Lake bathymetry, water clarity
Pressure, strain rate tensor, kinetic energy	Velocity field, lift, drag	Reynolds stress, flow geometry

Limitations of Physics-based Models

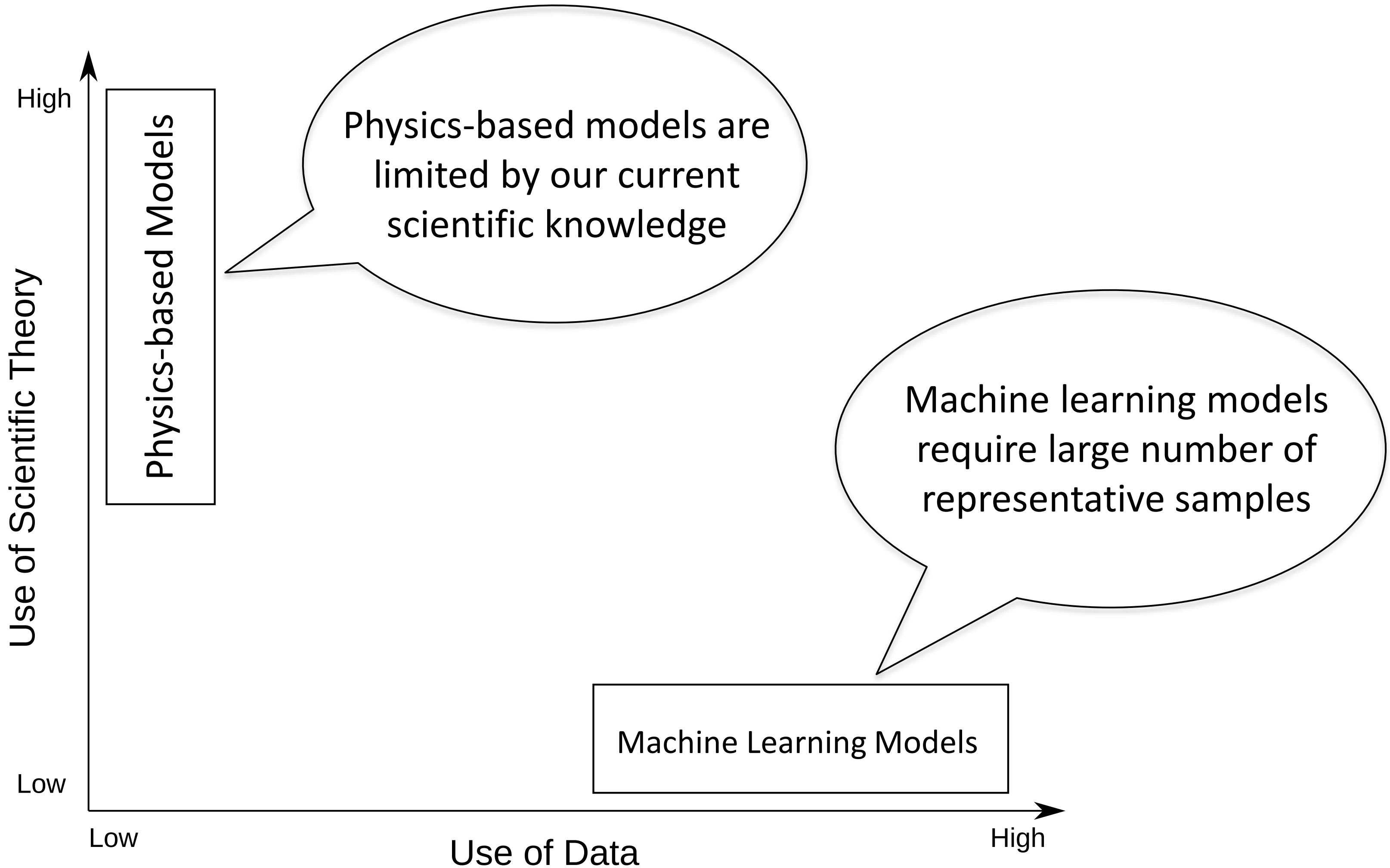


- Incomplete or missing physics (F, G)
  - Physics-based models often use approximate forms to meet "scale-accuracy" trade-off
  - Results in *inherent model bias*

- Unknown parameters (θ) need to be "calibrated"
  - *Computationally Expensive*
  - *Easy to overfit*: large number of parameter choices, small number of samples



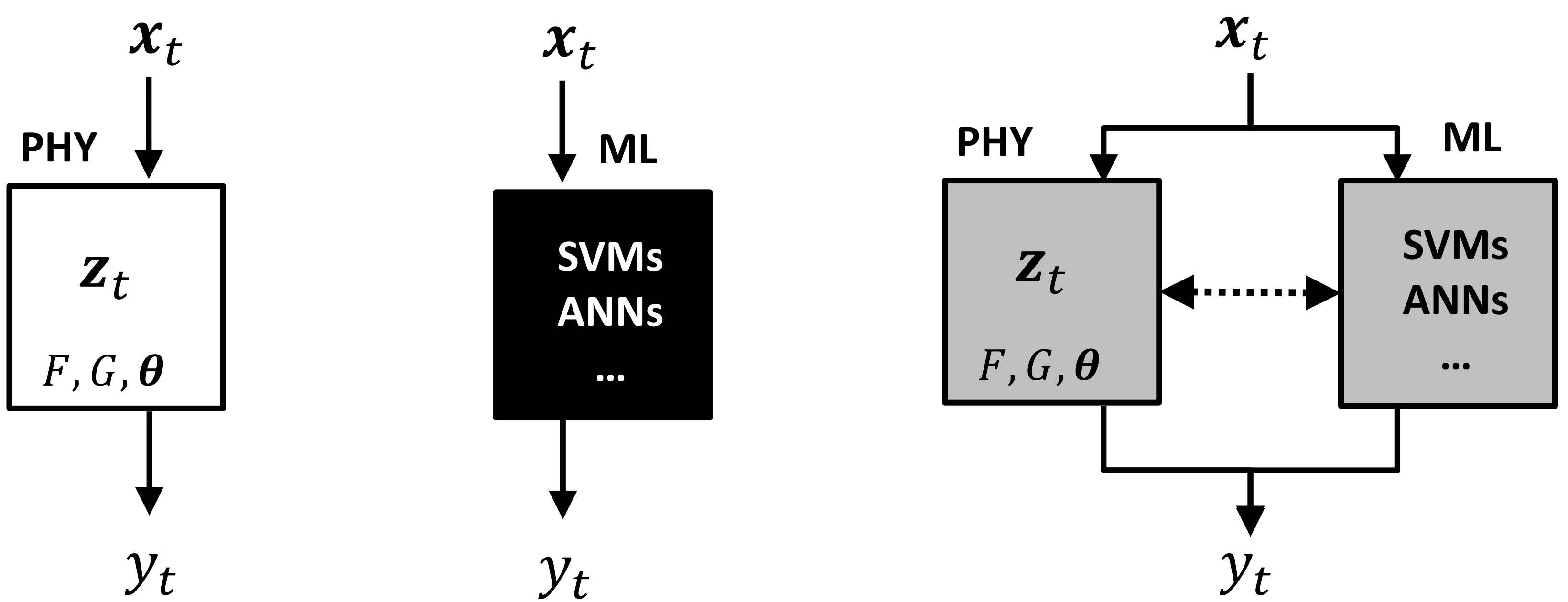
Physics-based Models vs. Machine Learning Models



Both use incomplete sources of information about the two key components of knowledge discovery: *scientific theory* and *data*

Physics-Guided Machine Learning:

A Paradigm Shift in Machine Learning



Physics-based Models | Machine Learning Models | HPD Models

Contain knowledge gaps in describing certain processes

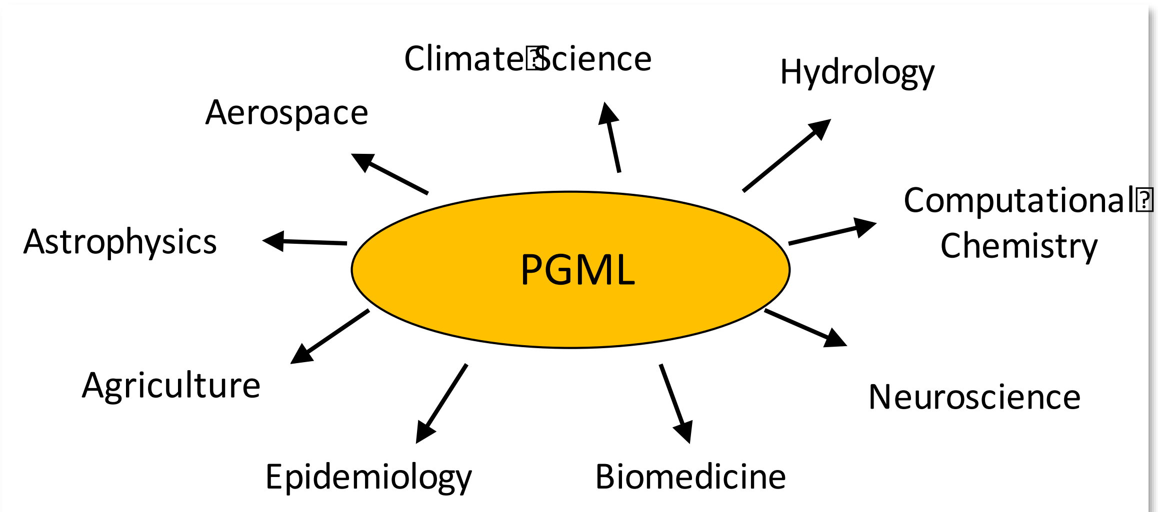
Require large number of representative samples

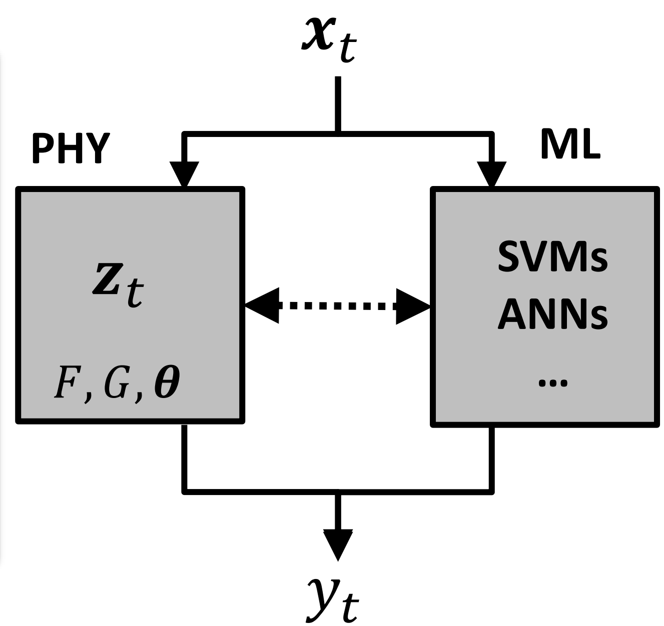
Overcome complementary weaknesses of both by combining PHY and ML in novel ways

Karpatne et al. "**Theory-guided data science: A new paradigm for scientific discovery**," TKDE 2017

Physics-Guided Machine Learning:

A Paradigm Shift in Machine Learning



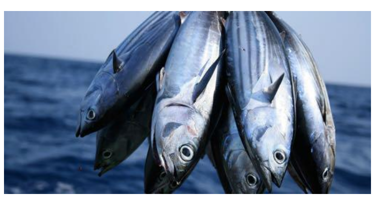


Physics-based Models	Machine Learning Models	HPD Models
Contain knowledge gaps in describing certain processes	Require large number of representative samples	Overcome complementary weaknesses of both by combining PHY and ML in novel ways

Karpatne et al. "**Theory-guided data science: A new paradigm for scientific discovery**," TKDE 2017
Willard et al. "**Integrating Physics-Based Modeling with Machine Learning: A Survey**", 2019

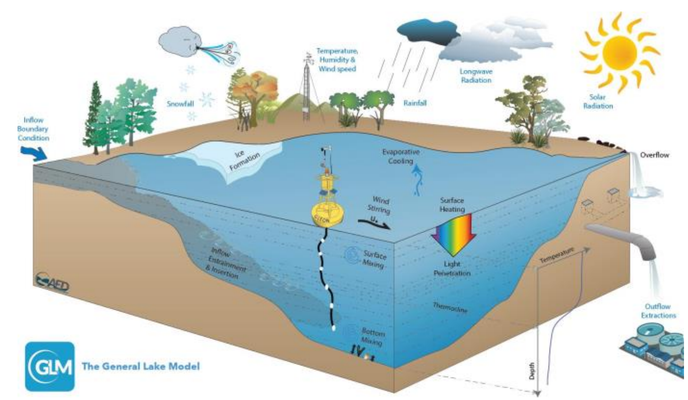
Questions

- Can machine learning (ML) models outperform physics based models given sufficient data?
- Can ML models leverage physics
  - to produce results that are physically consistent?
  - to learn with limited observation data?
  - To generalize to unseen scenarios
- Can physics guided ML models provide novel insights?
- *Illustrative example: Modeling Lake Water Temperature*

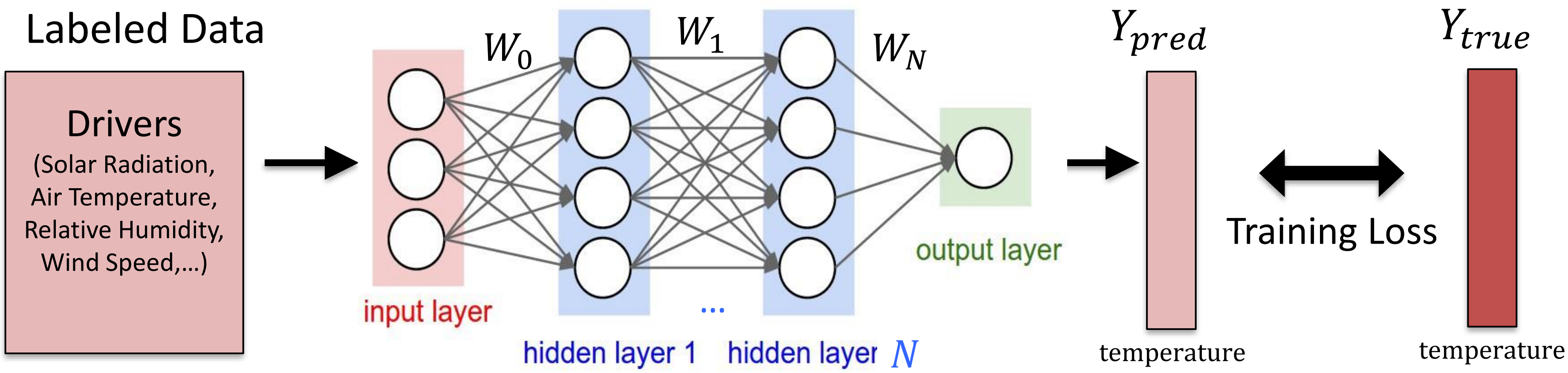

Growth and Survival of fisheries


Chemical Constituents: N, C, O_2

Algal Blooms



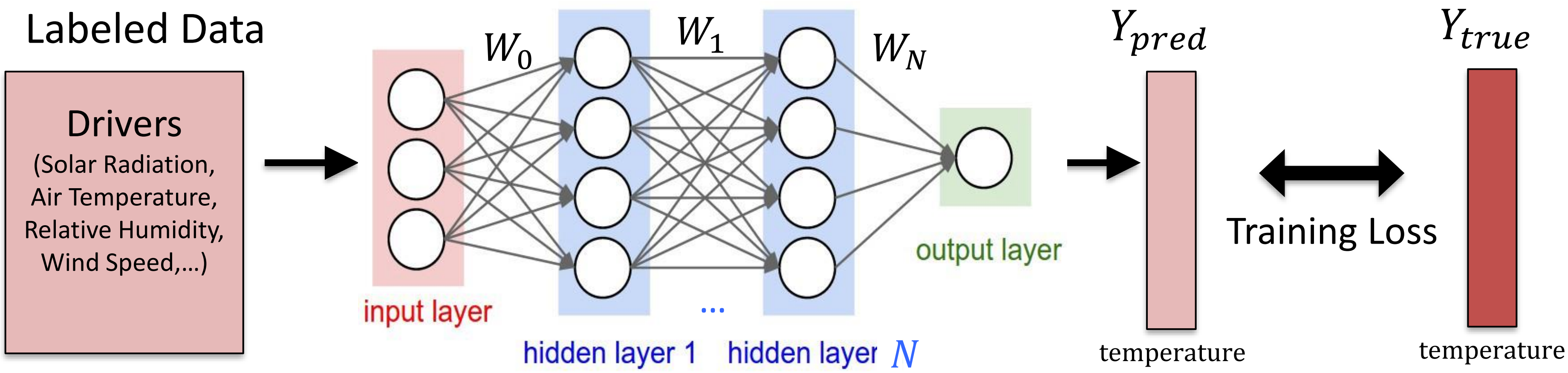
Training Machine Learning Models



$$\text{Objective} := \text{Supervised Loss}(Y_{true}, Y_{pred}) + \lambda\, \mathrm{R}(W)$$

Regularization *(e.g., L1/L2-norm)*

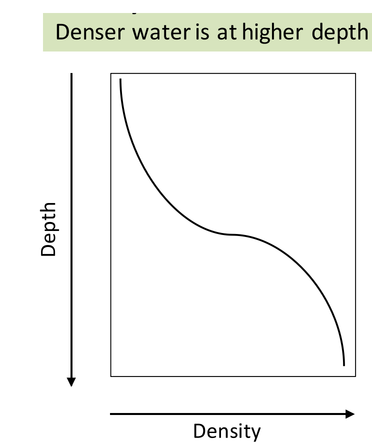
Incorporating Physics in ML Models



Objective Function :=

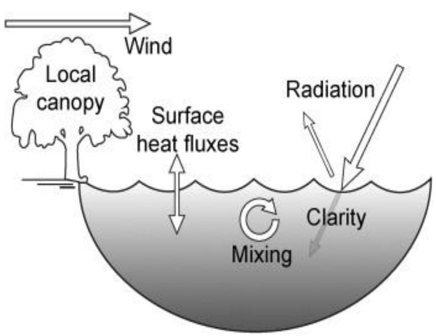
$$\text{Supervised Loss}\left(Y_{true}, Y_{pred}\right) + \lambda\, \mathrm{R}(W)$$

$$+\ \textbf{Physics-based Loss}\ \left(Y_{pred}, V_{PHY}\right)$$

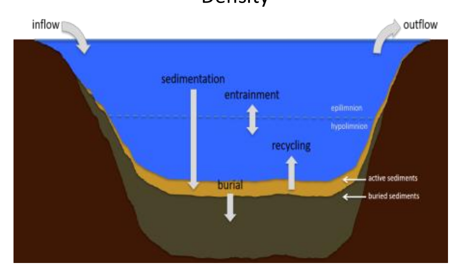


Physical variables

(extracted by new architectures)



Energy Conservation

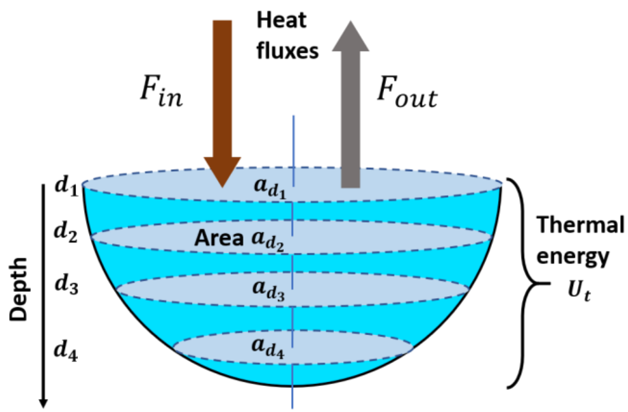


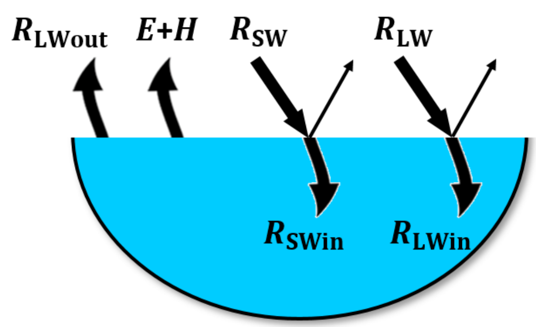
Mass Conservation

Incorporating Energy Conservation

- **Lake energy budget** - a balance between incoming energy fluxes and heat losses from the lake.
- A mismatch in losses and gains results in a temperature change.
- Thermal energy change $\mathrm{d}U_t/\mathrm{d}t = F_{in} - F_{out}$

▪ *Energy fluxes F_{in} and F_{out} include long-term and short-term radiation, sensible and latent heat fluxes, etc.*

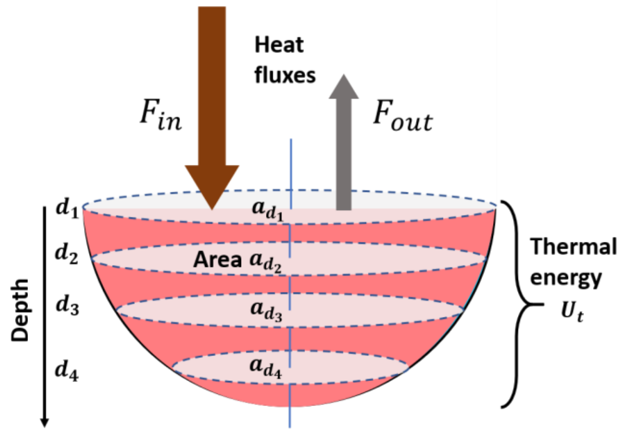


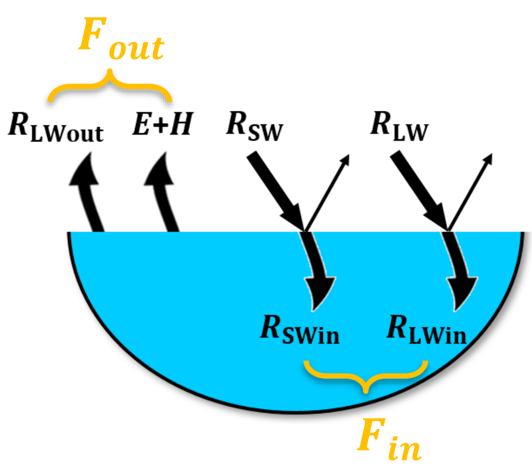


- Energy conservation is also generalizable to other dynamical systems.

Incorporating Energy Conservation

- **Lake energy budget** - a balance between incoming energy fluxes and heat losses from the lake.
- A mismatch in losses and gains results in a temperature change.
- Thermal energy change $\mathrm{d}U_t/\mathrm{d}t = F_{in} - F_{out}$
  - *Energy fluxes F_{in} and F_{out} include long-term and short-term radiation, sensible and latent heat fluxes, etc.*

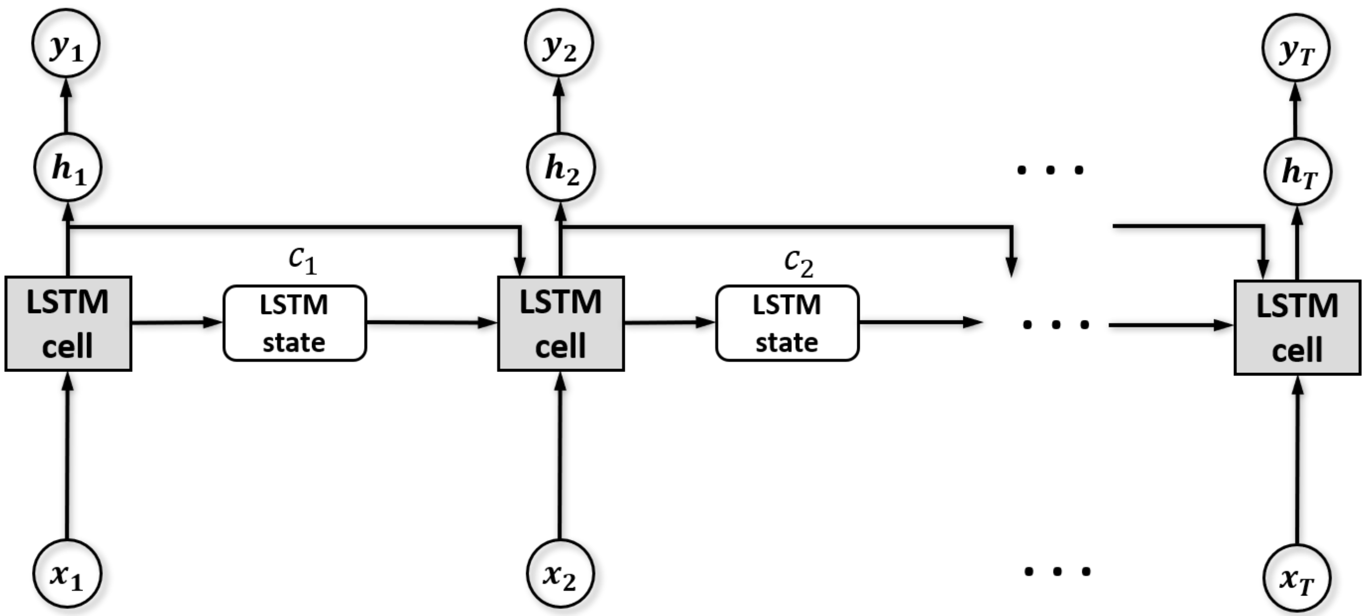




- Energy conservation is also generalizable to other dynamical systems.

Recurrent Neural Networks

- Given a sequence of input drivers $\{x_1, x_2, \dots, x_T\}$, we aim to predict the outputs at each time step $\{y_1, y_2, \dots, y_T\}$.

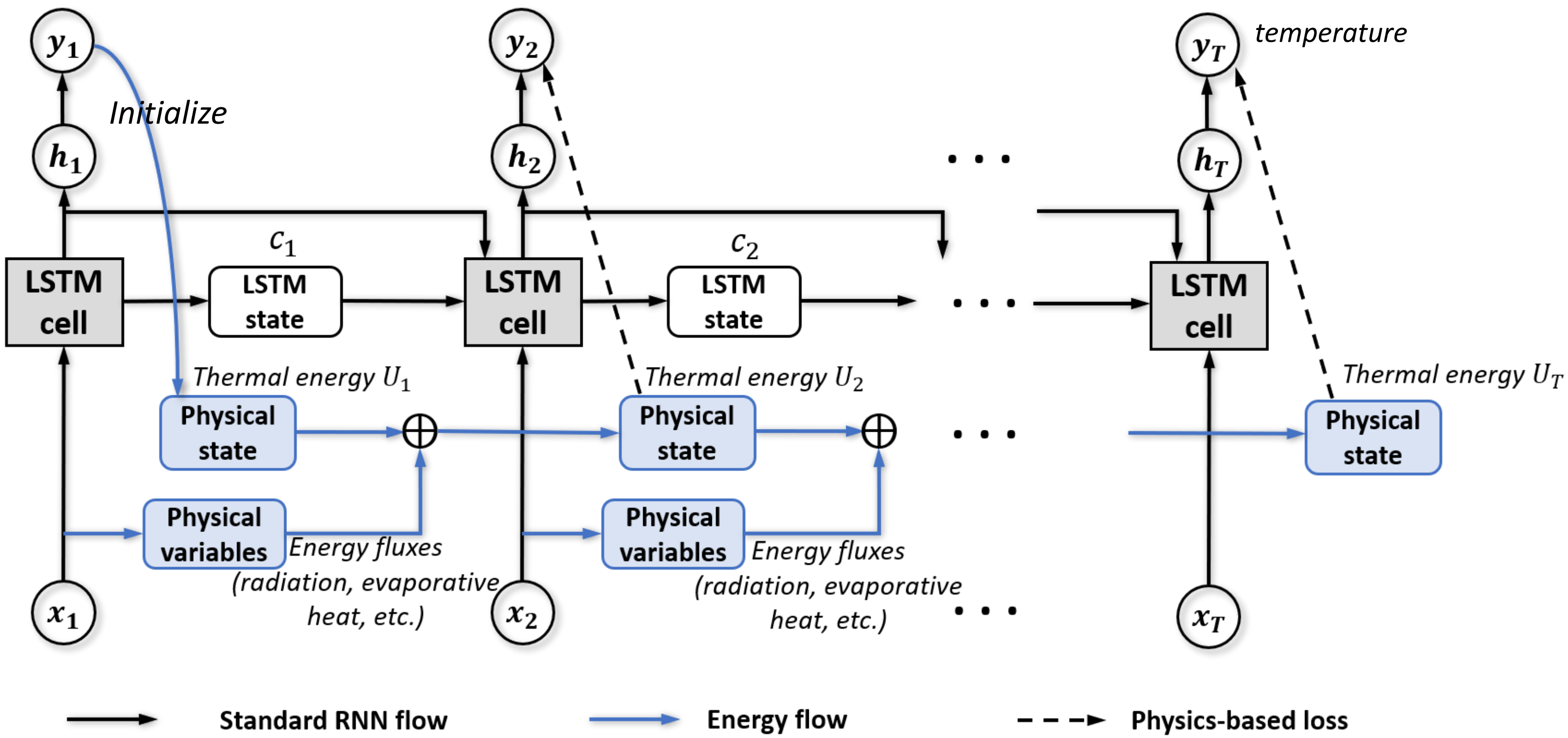


$$i^t = \sigma(W_h^i h^{t-1} + W_x^i x^t)$$
$$\tilde{c}^t = tanh(W_h^c h^{t-1} + W_x^c x^t)$$
$$f^t = \sigma(W_h^f h^{t-1} + W_x^f x^t)$$

$$\longrightarrow \quad c^t = f^t \otimes c^{t-1} + i^t \otimes \tilde{c}^t$$
$$o^t = \sigma(W_h^o h^{t-1} + W_x^o x^t)$$

$$\longrightarrow \quad h^t = o^t \otimes tanh(c^t)$$
$$p^t = \sigma(U h^t)$$

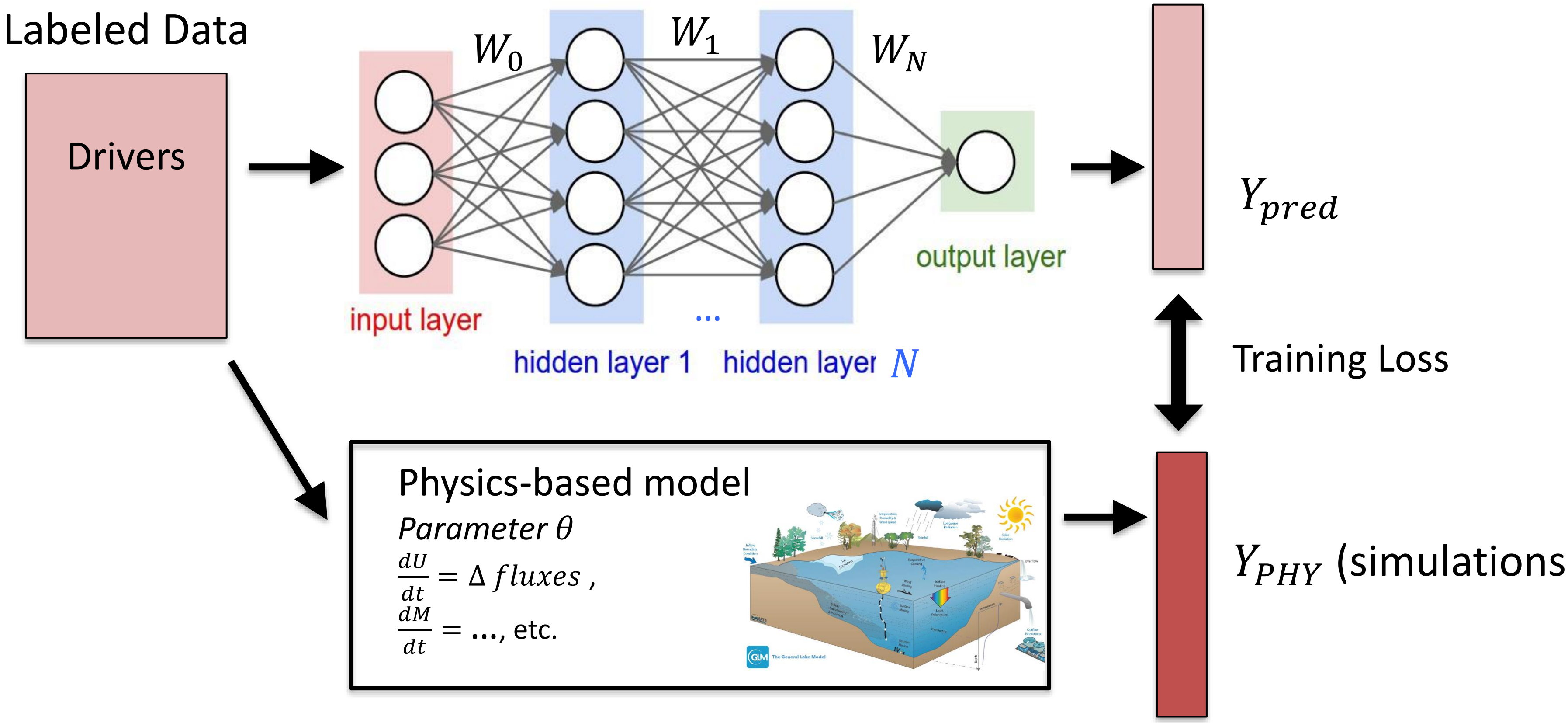
Physics-Guided Recurrent Neural Networks (PGRNN)



(Lake thermal energy U_t is proportional to the volume-average of temperatures)

Jia et al. Physics Guided RNNs for Modeling Dynamical Systems: A Case Study in Simulating Lake Temperature Profiles, 2019

Can we leverage knowledge hidden in physics based models via pre-training?

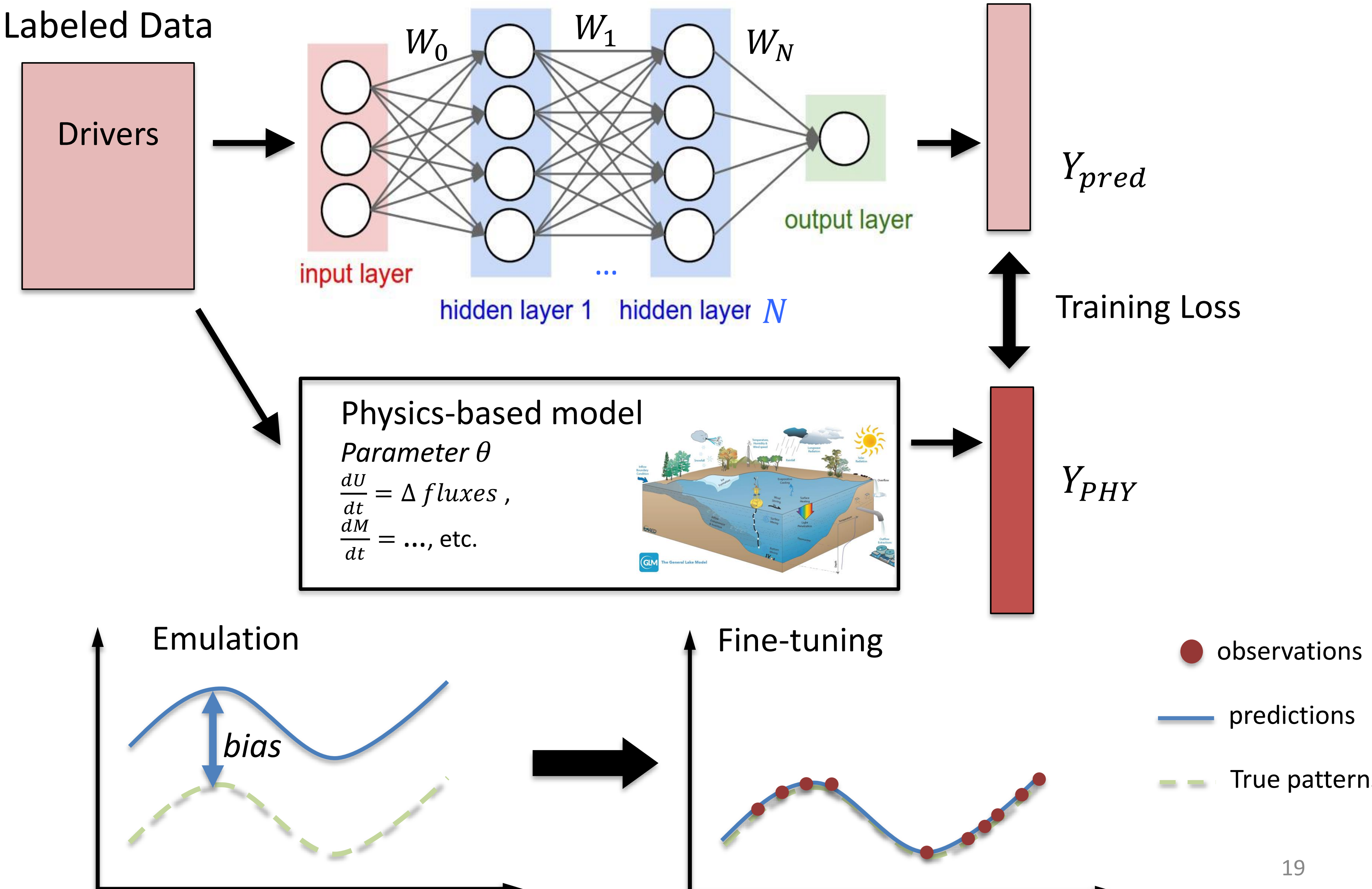


Objective Function :=

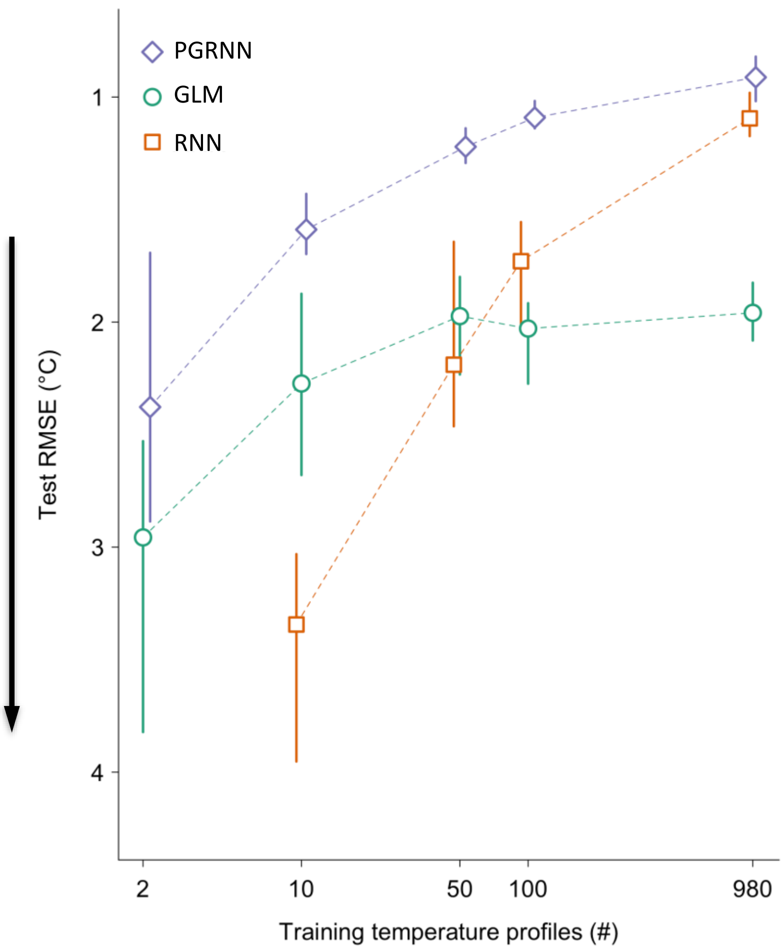
Supervised Loss$\left(Y_{PHY}, Y_{pred}\right) + \lambda\, \mathrm{R}(W)$

$+$ **Physics-based Loss** $\left(Y_{pred}, V_{PHY}\right)$

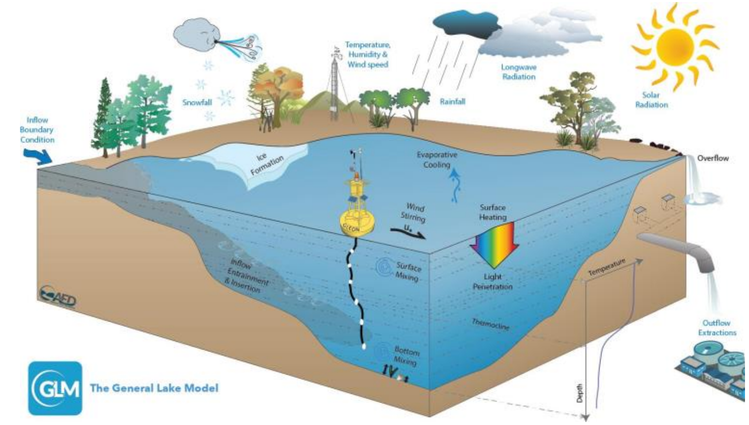
Can we leverage knowledge hidden in physics based models via pre-training?



PGML for Modeling Lake Water Temperature:
Performance Using Limited Observation Data



Read et al. Process-guided deep learning predictions of lake water temperature, 2019



General Lake Model (GLM): State of the Art physics based model used by USGS

RNN: A black-box machine learning model that can incorporate time

PGRNN: A machine learning framework that leverages physics

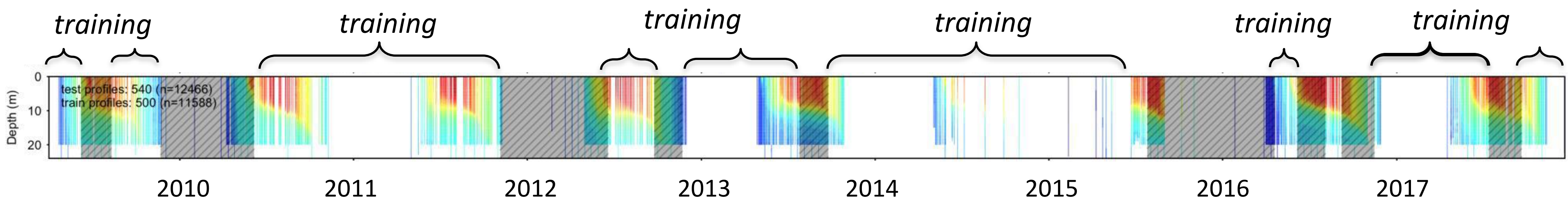
- The training and testing data are randomly selected (repeat 5 times) from a 9-year period 2009-2017.
- The PGRNN is pretrained using the simulated data in past 30 years.

PGML for Modeling Lake Water Temperature:

Generalization to Unseen Scenarios

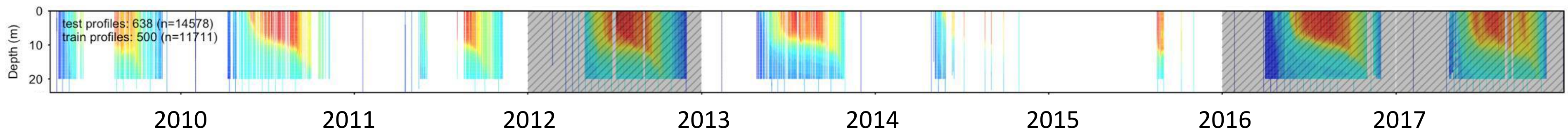
1. Train and test in similar data:

train [max 28.1 min 0.0 avg 14.8], test [max 29.3 min 0.1 avg 15.3])



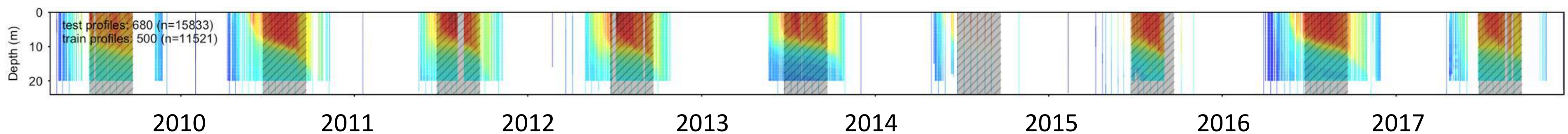
2. Train in coldest years and test in warmest years:

train [max 28.4 min 0.3 avg 14.4], test [max 29.3 min 0.1 avg 15.3]



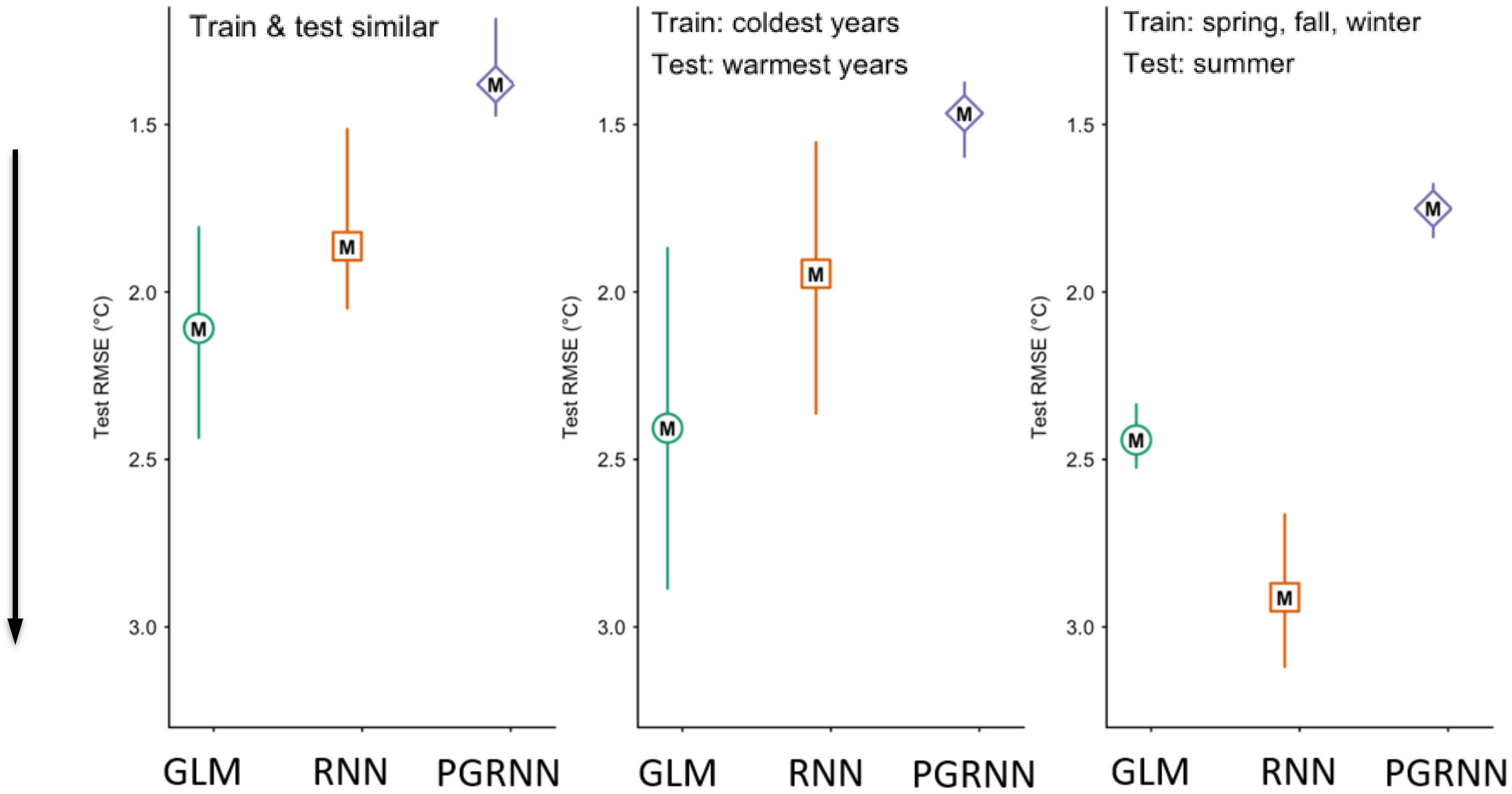
3. Train in coldest seasons and test in warmest seasons:

train [max 24.2 min 0.1 avg 12.5], test [max 29.3 min 8.3 avg 18.3])



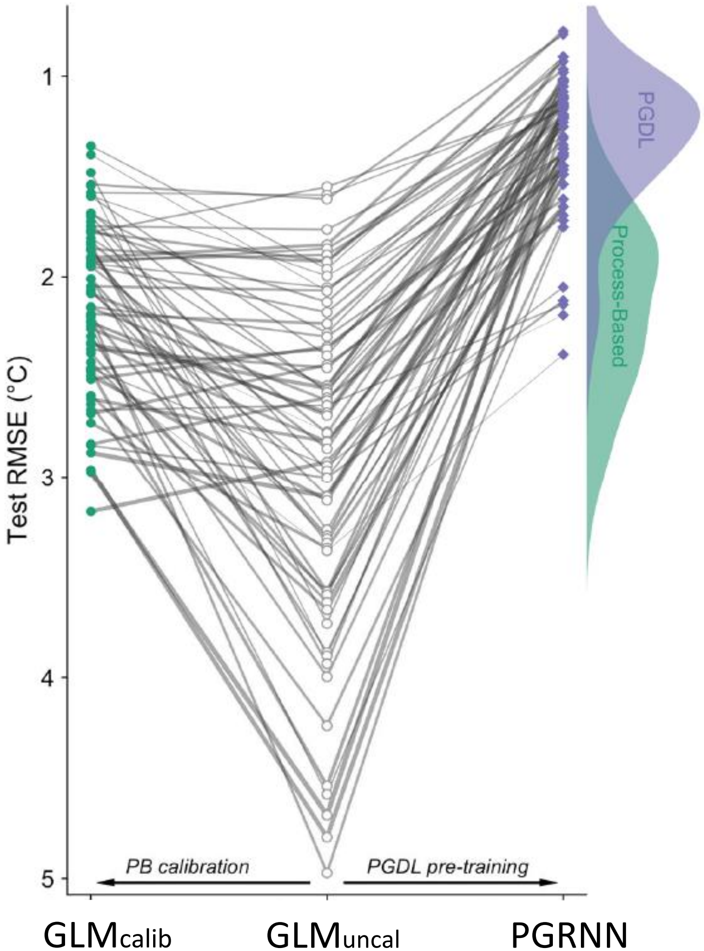
Shaded areas are used for testing.

PGML for Modeling Lake Water Temperature:
Generalization to Unseen Scenarios

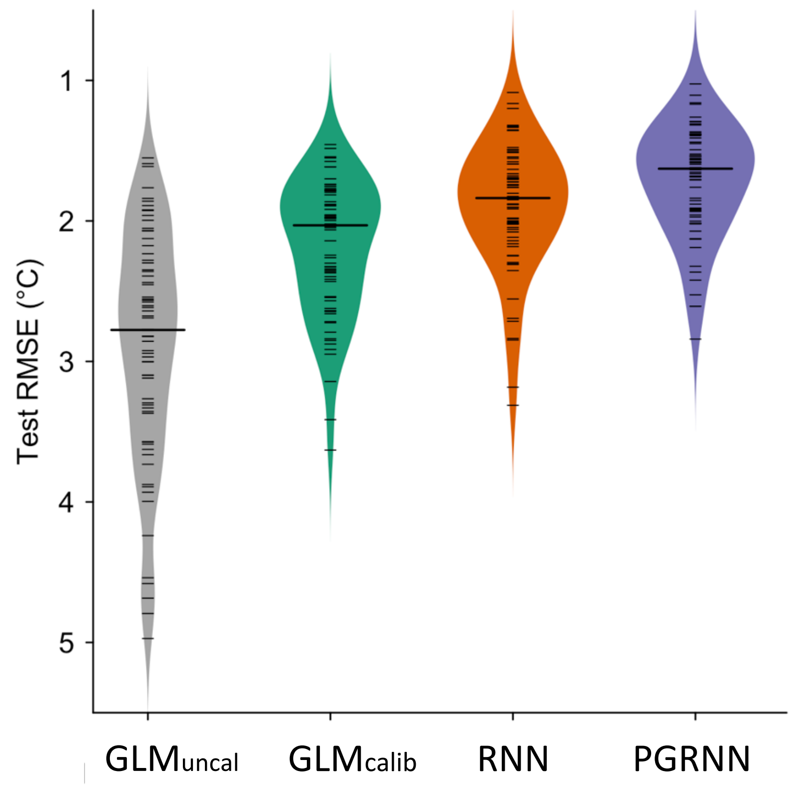


PGML for Modeling Lake Water Temperature:

Performance Across a Variety of Lakes

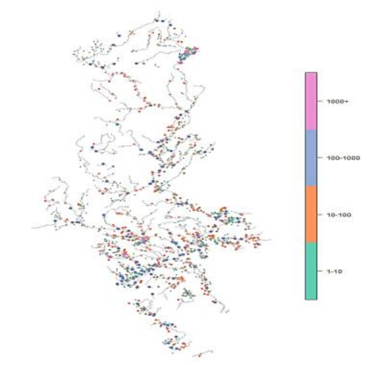


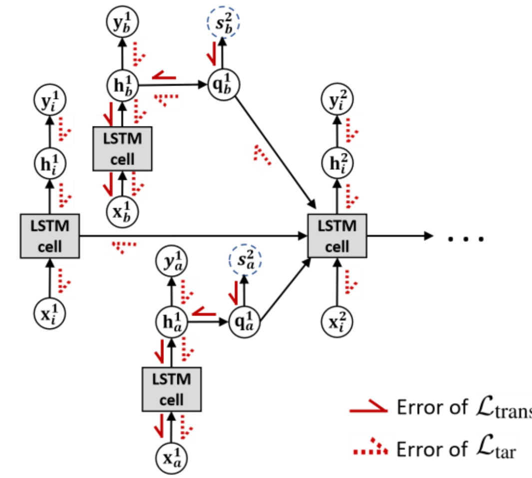
Improvements in water temperature predictions between uncalibrated process-based model (pre-trainer) and physics-guided model for 68 lakes in the Midwest US.



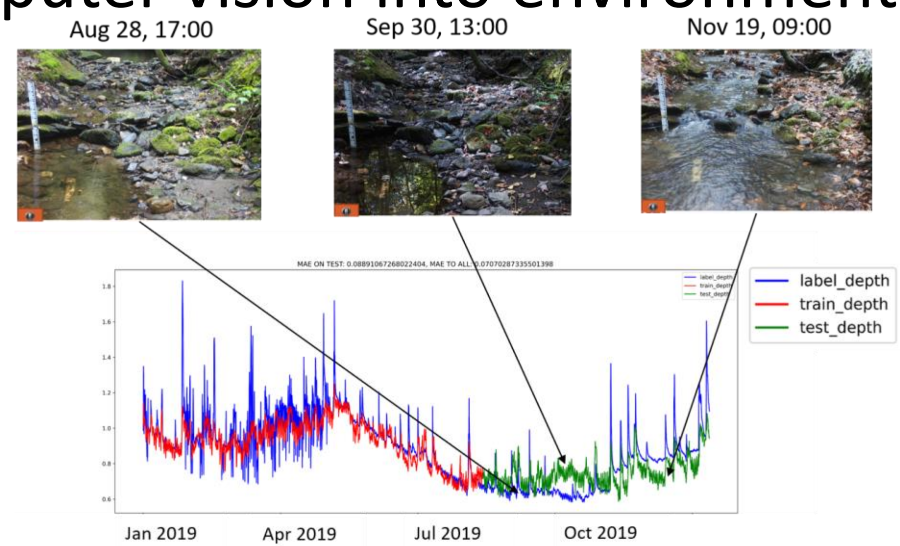
Other projects:
Dynamical systems with interacting processes

- Modeling river networks





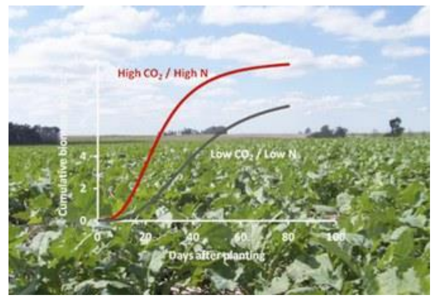
- Bring computer vision into environmental modeling



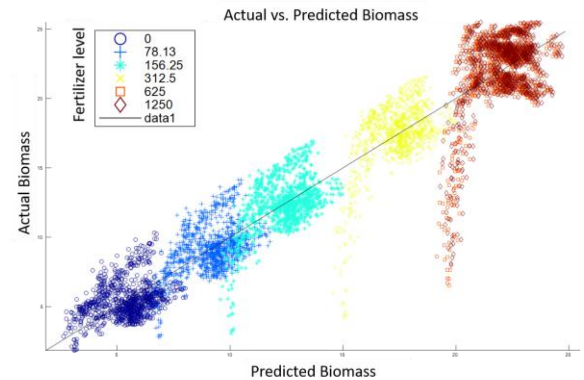
Other projects:
Remote Sensing + Physical Processes

- Crop modeling

Landsat September 2016, Minnesota



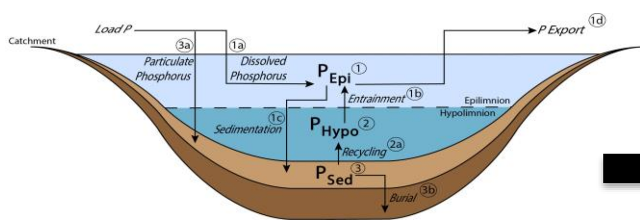
Physics-based models for modeling crop, soil and water (DSSAT, SWAT, CYCLES)



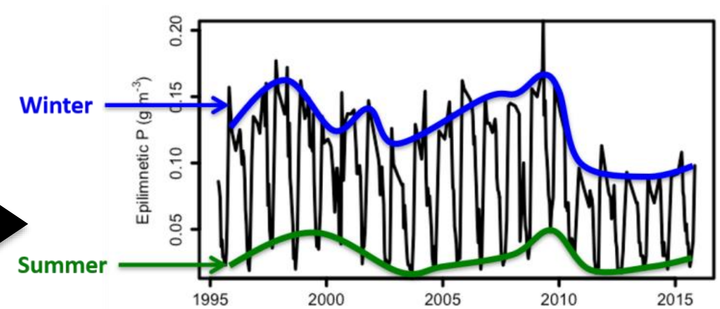
1. Crop yield
2. Better strategy (e.g., fertilizer)

- Lake phosphorus modeling

Algae bloom on Lake Erie in 2011 (NOAA)



(Under submission to Ecological Modelling)



Capture long-term and short-term patterns in predictions

Acknowledgements

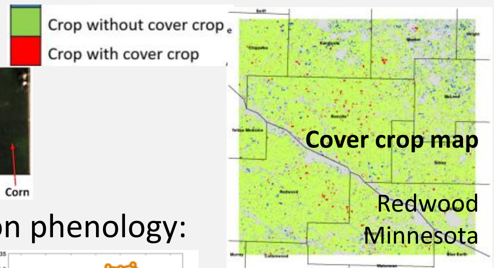
- Colleagues and graduate students:

- Shengyu Chen, Ankush Khandelwal, Anuj Karpatne, Jared Willard, Guruprasad Nayak, Saurabh Agarwal, Rahul Ghosh, Kshitij Tayal, Shaoming Xu

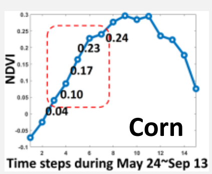
- Collaborators

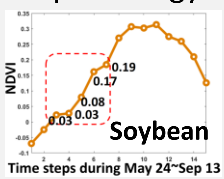
Important global changes

Agriculture



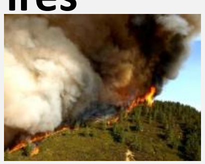
Interpretability based on phenology:

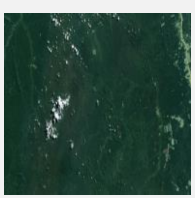


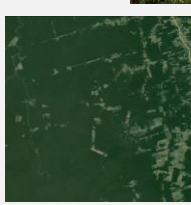


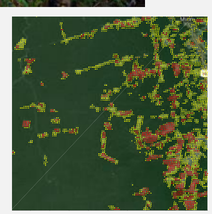
Urbanization

Forest Fires





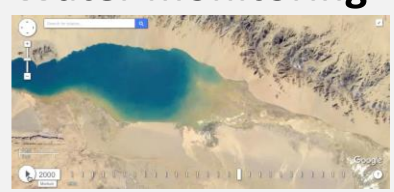




Year 2002 *2015* *Detection 2002-2014*

Google Earth image

Water monitoring



Water dynamics

Water quality