Physics Guided Machine Learning: A New Paradigm for Accelerating Scientific Discovery

Vipin Kumar
University of Minnesota
kumar001@umn.edu
www.cs.umn.edu/~kumar

Joint work with

J. Read, A. Appling, J. Zwart, S. Oliver, W. Watkins, USGS
X. Jia, J. Willard, M. Steinbach, G. Hansen, University of Minnesota
P. Hanson, University of Wisconsin
A. Karpatne, Virginia Tech

Harnessing the Data Revolution for Scientific Discovery

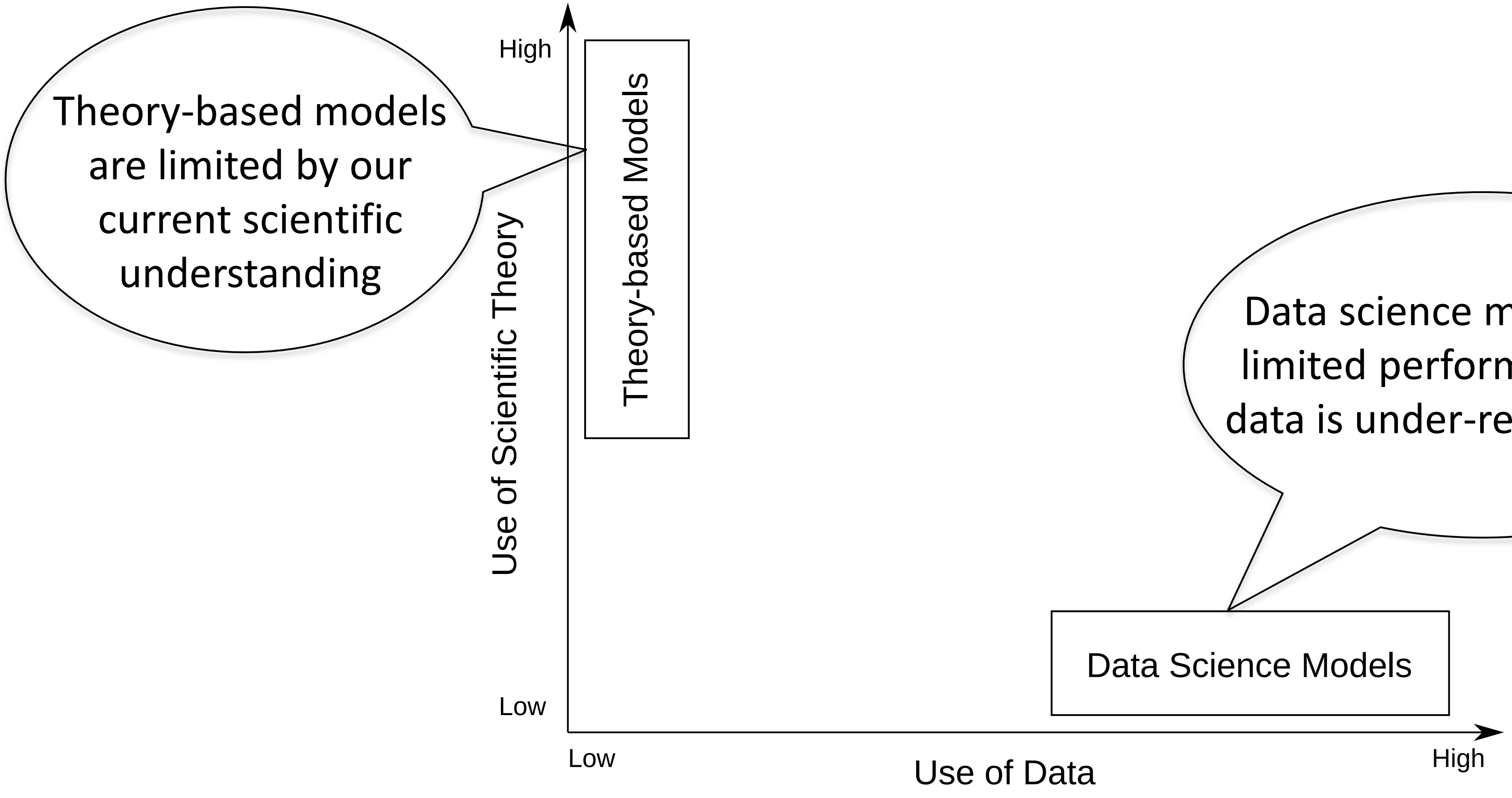
- Big data and ML are Increasingly being considered as an alternative to the traditional scientific discovery paradigm based on domain theories

- However, black box methods have shown disappointing results in scientific domains!
  - Require lots of data
  - Can generate physically inconsistent results
  - Unable to generalize to unseen scenarios
  - Unable to provide valuable physical insights

Science

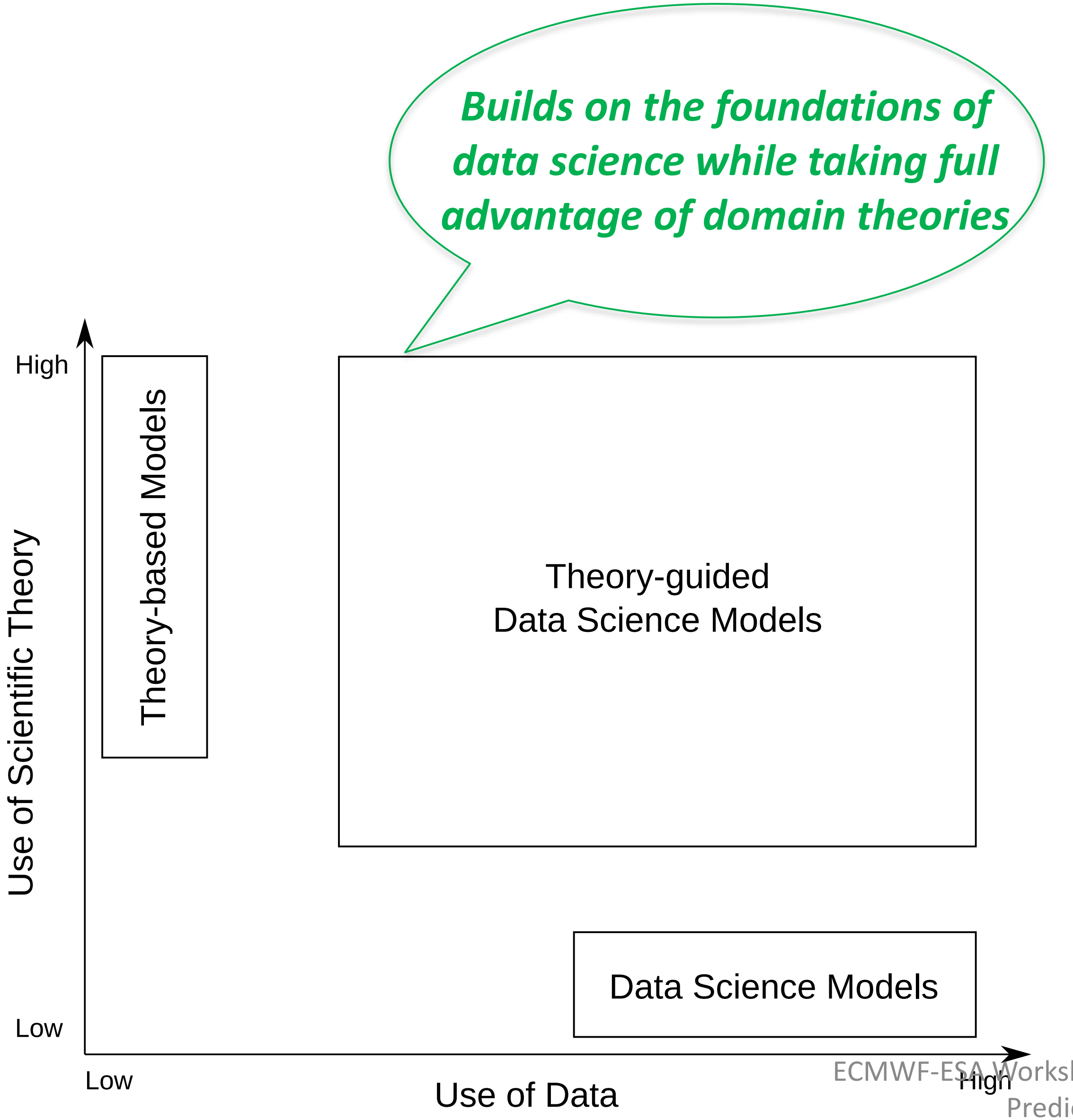
The Parable of Google Flu: Traps in Big Data Analysis

Dichotomy b/w Theory-based and Data Science Models

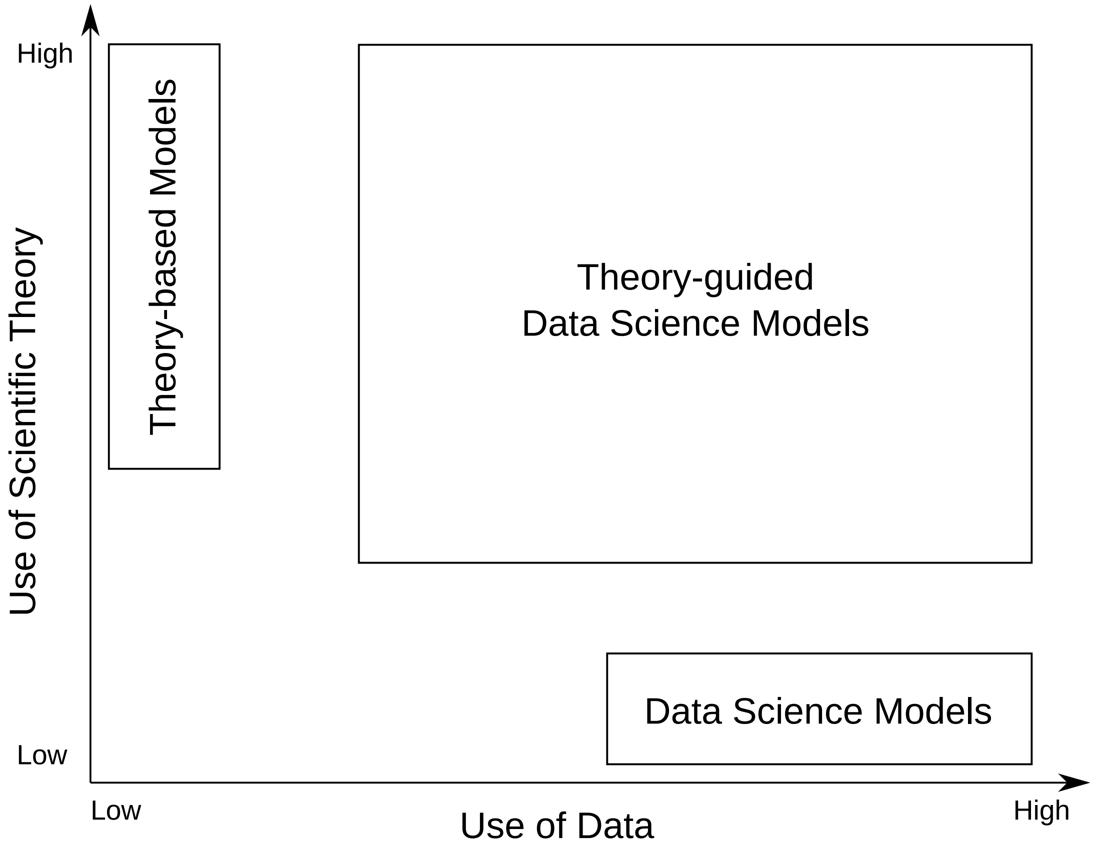


Both use incomplete sources of information about the two key components of knowledge discovery: *scientific theory* and *data*

Theory-guided Data Science (TGDS)



Theory-guided Data Science (TGDS)



Home / Journals / IEEE Transactions on Knowledge and Data Engineering / 2017.10

Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data

Oct. 2017, pp. 2318-2331, vol. 29

Anuj Karpatne, Gowtham Atluri, James H. Faghmous, Michael Steinbach, Arindam Banerjee, Auroop Ganguly, Shashi Shekhar, Nagiza Samatova, and Vipin Kumar

A Big Data Guide to Understanding Climate Change: The Case for Theory-Guided Data Science

James H. Faghmous and Vipin Kumar

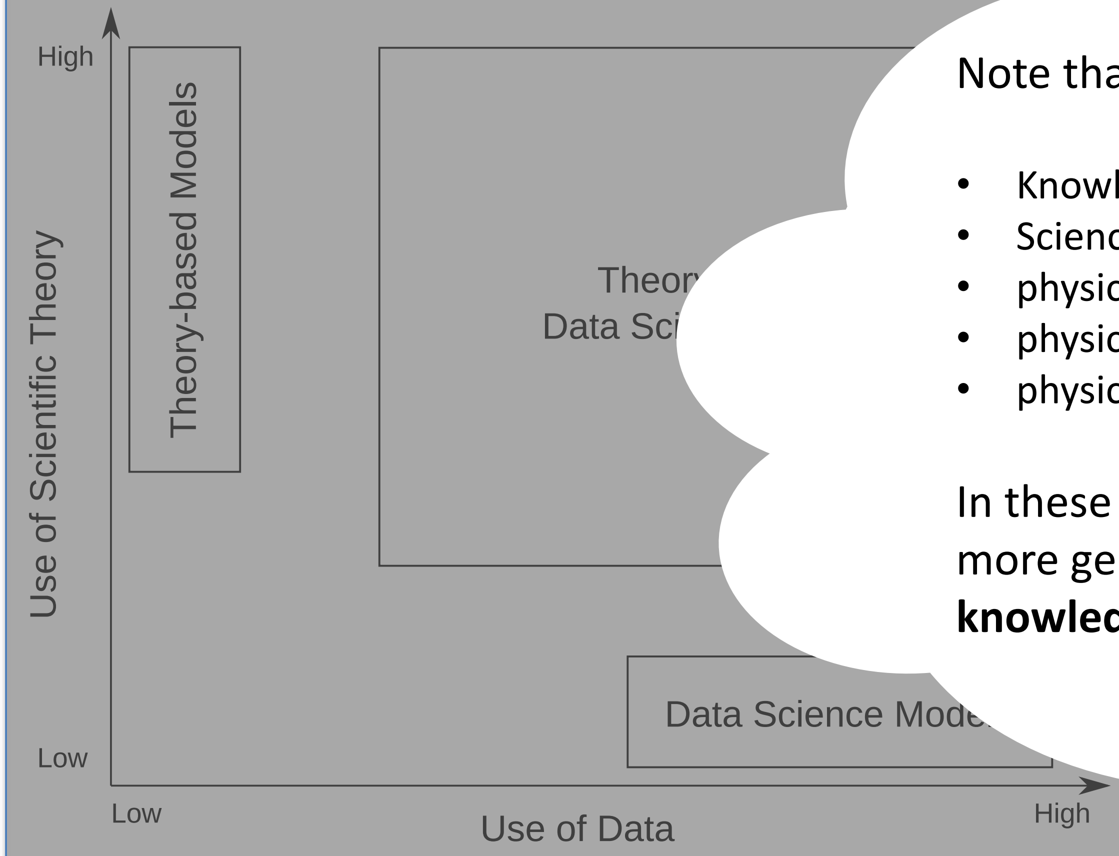
Published Online: 15 Sep 2014 | https://doi.org/10.1089,

Theory-Guided Data Science for Climate Change

James H. Faghmous, Arindam Banerjee, Shashi Shekhar, Michael Steinbach, and Vipin Kumar, *University of Minnesota, Twin Cities*
Auroop R. Ganguly, *Northeastern University*
Nagiza Samatova, *North Carolina State University*

To adequately address climate change, we need novel data-science methods that account for the spatiotemporal and physical nature of climate phenomena. Only then will we be able to move from statistical analysis to scientific insights.

Theory-guided Data Science (TGDS)



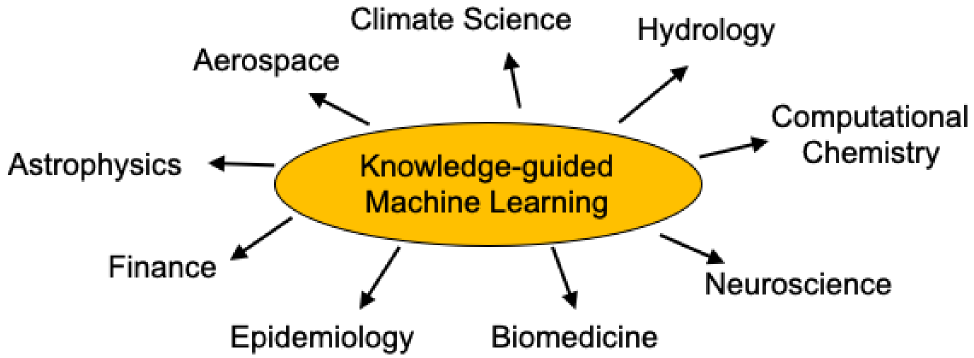
Note that work on this topic has been referred to as

- Knowledge-guided ML
- Science-guided ML
- physics-guided ML
- physics-informed ML
- physics-aware AI

In these works, “**physics**” or "**physics-guided**" should be more generally interpreted as “**science”** or “**scientific knowledge”.**

Knowledge Guided Machine Learning:

A Paradigm Shift in Scientific Discovery

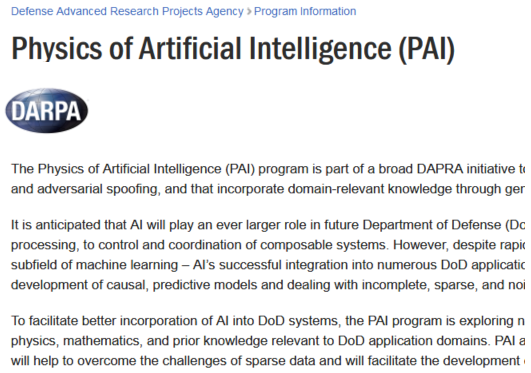


Surveys more than 300 papers

Integrating Physics-Based Modeling With Machine Learning: A Survey arXiv:2003.04919

JARED WILLARD* and XIAOWEI JIA*, University of Minnesota
SHAOMING XU, University of Minnesota
MICHAEL STEINBACH, University of Minnesota
VIPIN KUMAR, University of Minnesota

There is a growing consensus that solutions to complex science and engineering problems require novel methodologies that are able to integrate traditional physics-based modeling approaches with state-of-the-art machine learning (ML) techniques. This paper provides a structured overview of such techniques. Application areas for which these approaches have been applied are summarized, then classes of methodologies used to construct physics-guided ML models and hybrid physics-ML frameworks are described. We then provide a taxonomy of these existing techniques, which uncovers knowledge gaps and potential crossovers of methods between disciplines that can serve as ideas for future research.


Defense Advanced Research Projects Agency › Program Information

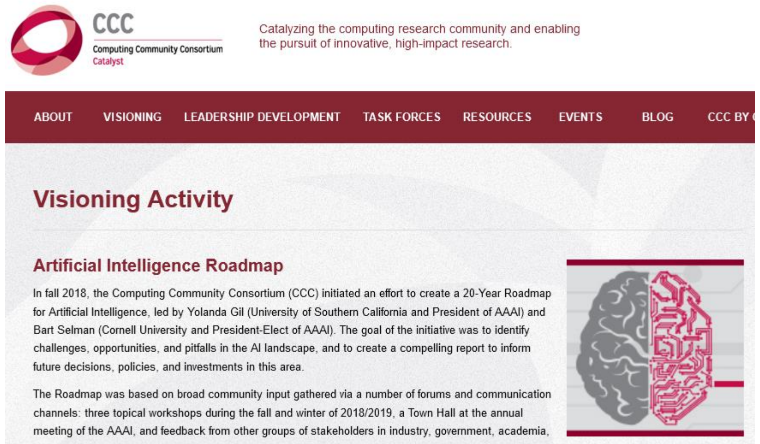
Physics of Artificial Intelligence (PAI)

DARPA

The Physics of Artificial Intelligence (PAI) program is part of a broad DARPA initiative t
and adversarial spoofing, and that incorporate domain-relevant knowledge through ger

It is anticipated that AI will play an ever larger role in future Department of Defense (Dc
processing, to control and coordination of composable systems. However, despite rapi
subfield of machine learning – AI's successful integration into numerous DoD applicatic
development of causal, predictive models and dealing with incomplete, sparse, and noi

To facilitate better incorporation of AI into DoD systems, the PAI program is exploring n
physics, mathematics, and prior knowledge relevant to DoD application domains. PAI a
will help to overcome the challenges of sparse data and will facilitate the development


CCC Computing Community Consortium Catalyst

Catalyzing the computing research community and enabling the pursuit of innovative, high-impact research.

ABOUT VISIONING LEADERSHIP DEVELOPMENT TASK FORCES RESOURCES EVENTS BLOG CCC BY

Visioning Activity

Artificial Intelligence Roadmap

In fall 2018, the Computing Community Consortium (CCC) initiated an effort to create a 20-Year Roadmap for Artificial Intelligence, led by Yolanda Gil (University of Southern California and President of AAAI) and Bart Selman (Cornell University and President-Elect of AAAI). The goal of the initiative was to identify challenges, opportunities, and pitfalls in the AI landscape, and to create a compelling report to inform future decisions, policies, and investments in this area.

The Roadmap was based on broad community input gathered via a number of forums and communication channels: three topical workshops during the fall and winter of 2018/2019, a Town Hall at the annual meeting of the AAAI, and feedback from other groups of stakeholders in industry, government, academia,

Many conferences/workshops

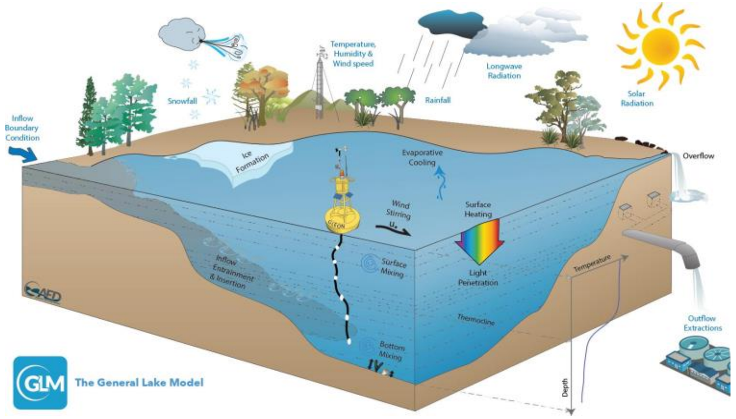
- 2020 Knowledge-guided Machine Workshop
- 2020 AAAI Spring Symposium on ML in Physical Sciences
- 2020 AAAI Fall Symposium on Physics-Guided AI
- 2020 SIAM MDS Mini-symposium on Physics-guided AI
- 2020 Physics-informed Machine Learning Workshop at LANL,
- 2020 Physics-Informed Learning Machines for Multiscale and Multiphysics Problems at PNNL

Questions

- Can machine learning (ML) models outperform physics based models given sufficient data?
- Can KGML models
  - learn with limited observation data?
  - generalize to novel testing scenarios?
  - produce results that are physically consistent?
- Can KGML models provide novel insights?
- Can KGML models be useful in absence of observation data?

Questions

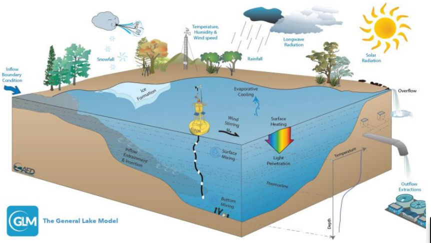
- Can machine learning (ML) models outperform physics based models given sufficient data?
- Can KGML models
  - learn with limited observation data?
  - generalize to novel testing scenarios?
  - produce results that are physically consistent?
- Can KGML models provide novel insights?
- Can KGML models be useful in absence of observation data?



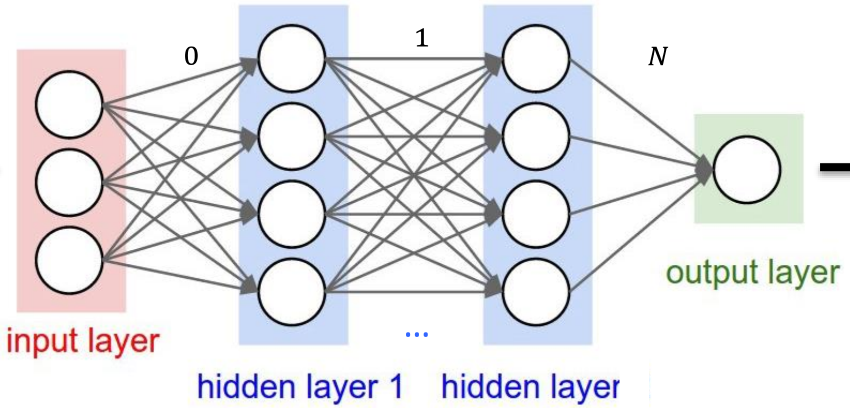
Modeling Lake Water Temperature dynamics

GLM: State of the Art physics based model used by USGS

Machine Learning Model for Lake Temperature Dynamics



Drivers



Y_{pred}

Training Loss

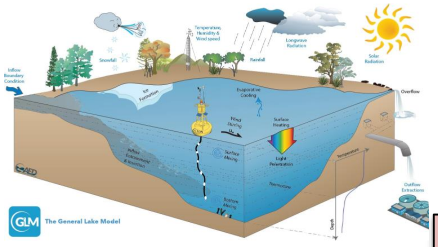
Y_{true}

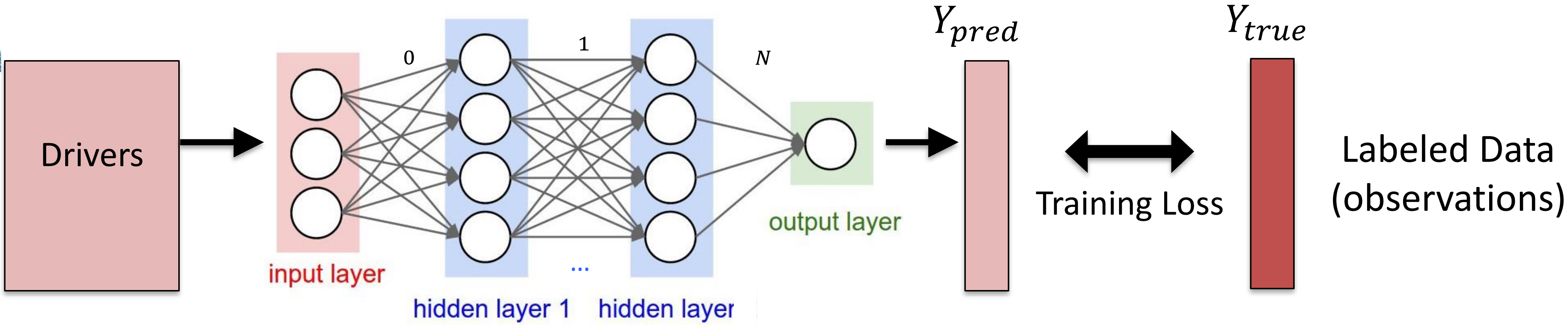
Labeled Data (observations)

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

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

Machine Learning Model for Lake Temperature Dynamics





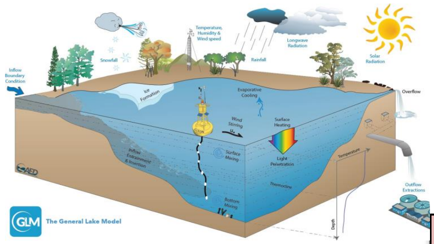
$$\text{Objective} := \text{Training Loss}\left(Y_{true}, Y_{pred}\right) + \lambda\, \mathrm{R}(W)$$

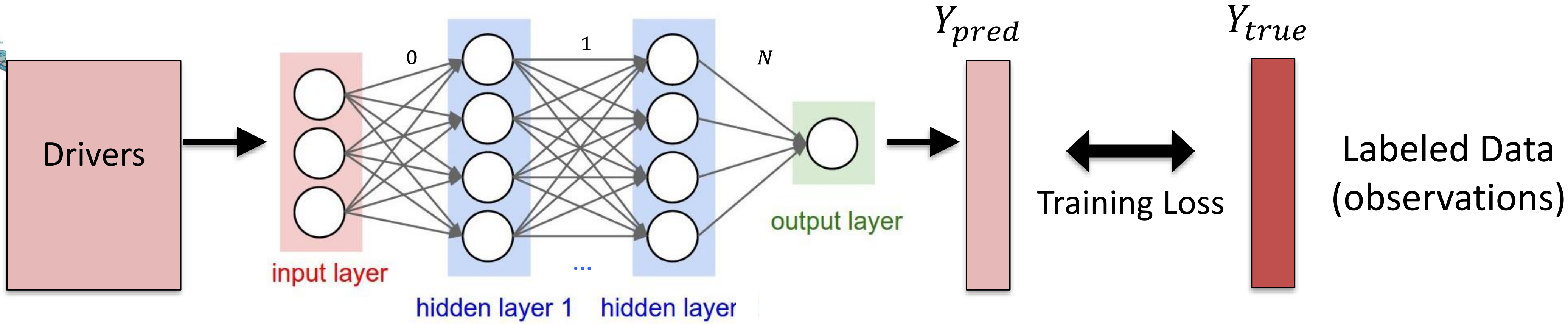
Regularization *(e.g., L1/L2-norm)*

Challenges:

1. Labels (Y_{true}) are scarce
   - Difficult to train models with sufficient complexity
   - Standard methods for assessing generalization performance break down

Machine Learning Model for Lake Temperature Dynamics





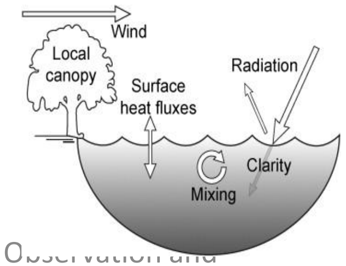
$$\text{Objective} := \text{Training Loss}\left(Y_{true}, Y_{pred}\right) + \lambda\, \mathrm{R}(W)$$

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

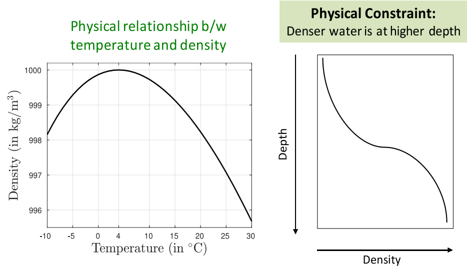
Challenges:

1. Labels (Y_{true}) are scarce
   - Difficult to train models with sufficient complexity
   - Standard methods for assessing generalization performance break down

2. Y_{pred} may be physically inconsistent

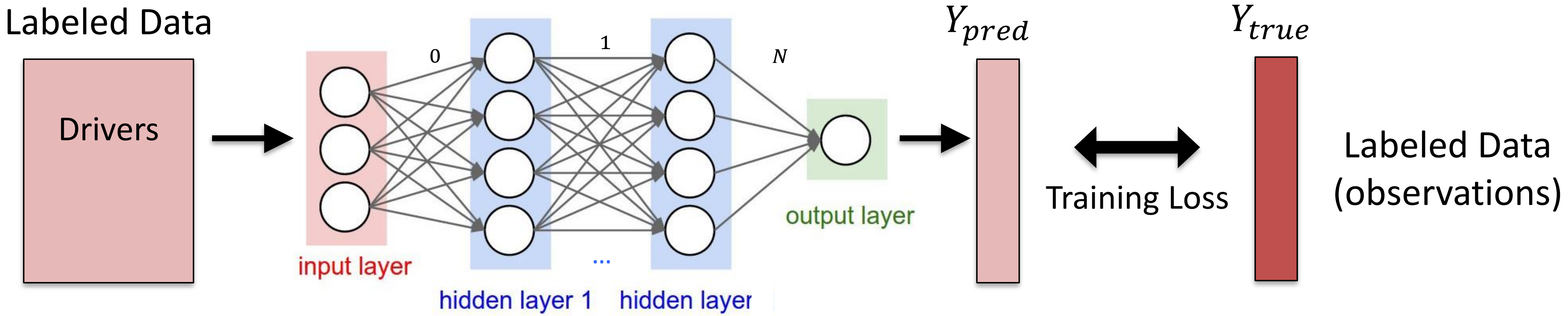


Energy Conservation



Density-depth constraint

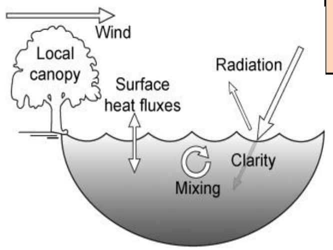
Incorporating Physics in ML Models



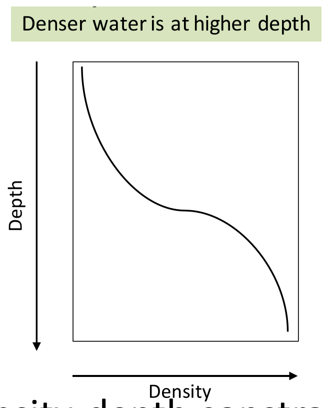
Objective Function :=

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

Physics-based Loss $\left(Y_{pred}\right)$



Energy Conservation

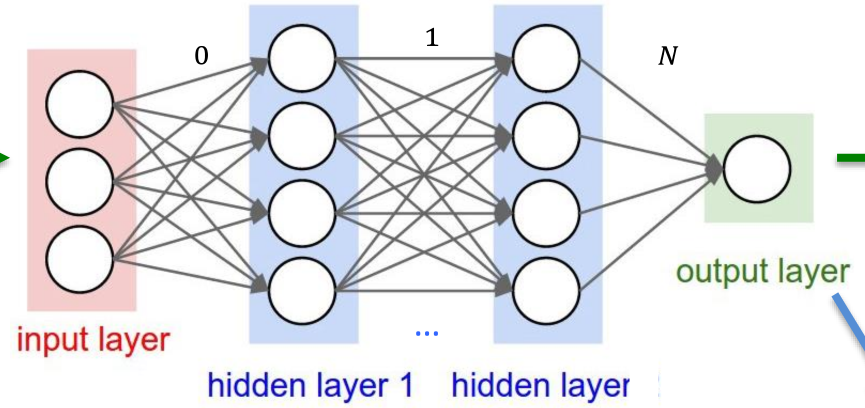


Density-depth constraint

KGML Models Can Learn from Unlabeled Data

Labeled Data

Drivers



Y_{pred}

Training Loss $\left(Y_{true}, Y_{pred}\right)$

Unlabeled Data

Drivers

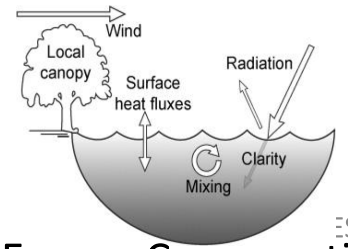
Y_{pred}

Objective Function :=

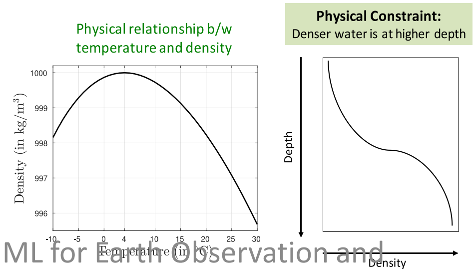
Training Loss $\left(Y_{true}, Y_{pred}\right) + \lambda\, \mathrm{R}(W) +$

Physics-based Loss $\left(Y_{pred}\right)$

Physics-based Loss $\left(Y_{pred}\right)$

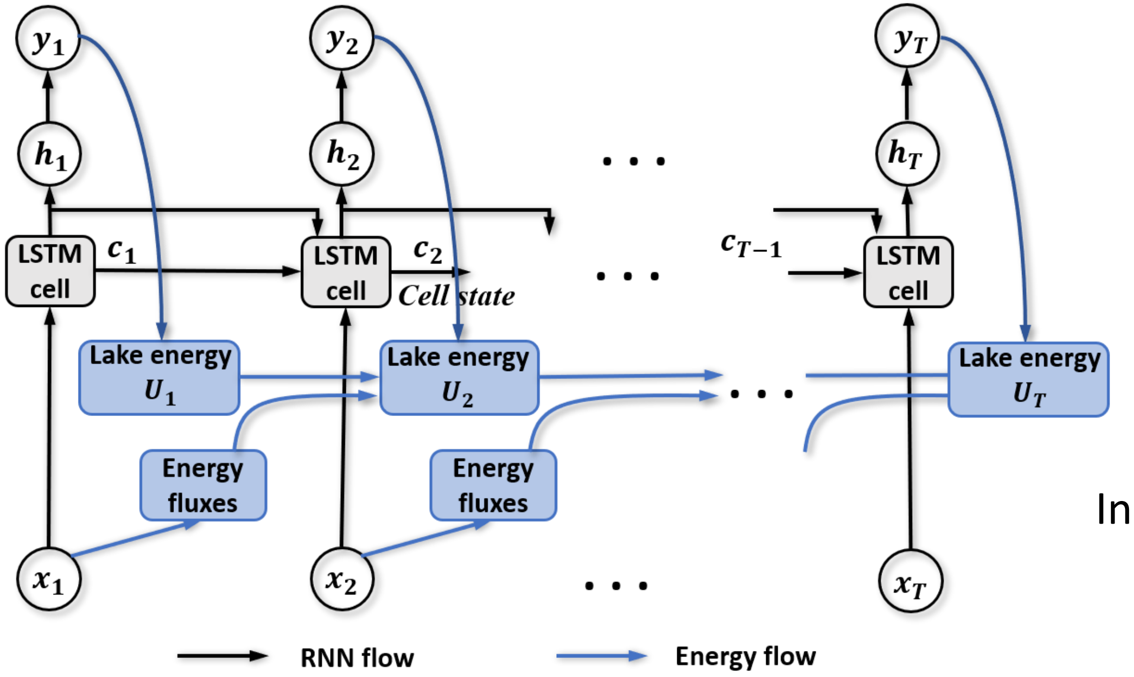


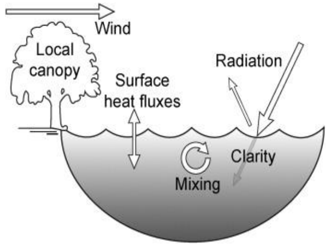
Energy Conservation



Density-depth constraint

Physics-Guided Recurrent Neural Network (PGRNN)

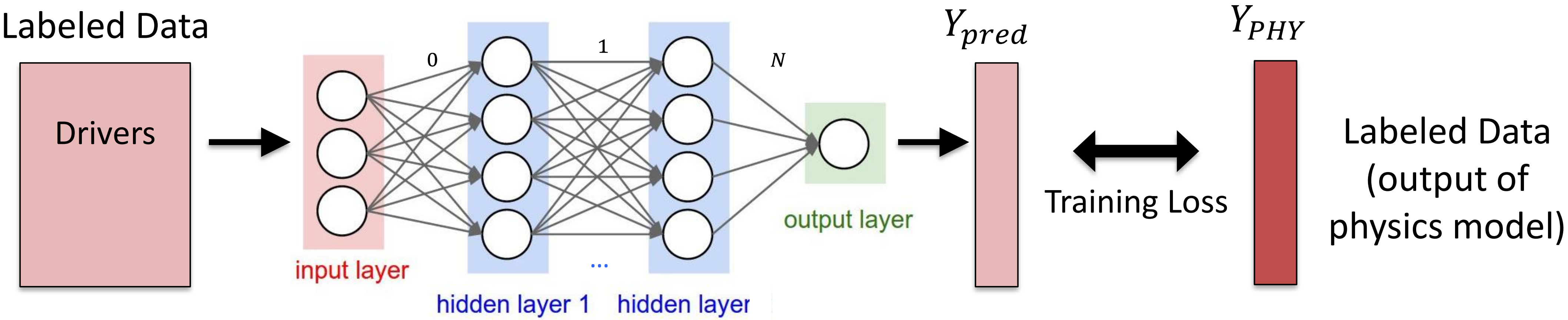




Incorporating Energy Consistency

Jia et.al SDM2019

Overcoming the Data Sparsity Challenge by Pre-training with Physics-based Models



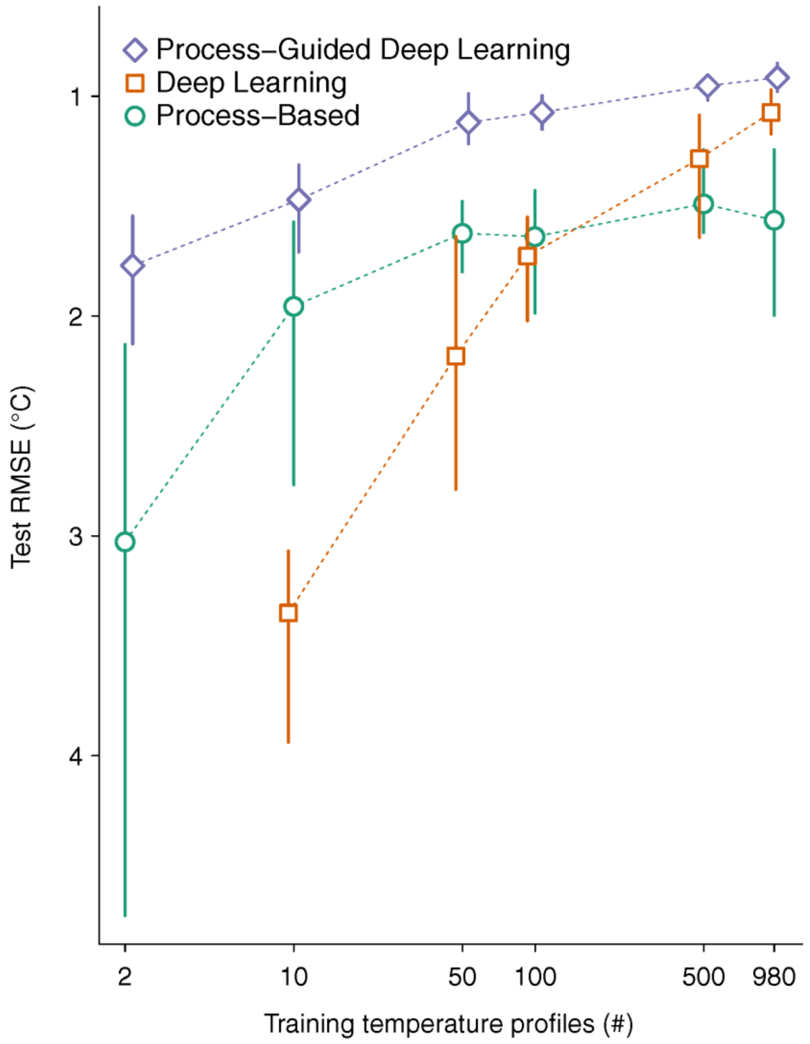
Objective Function :=

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

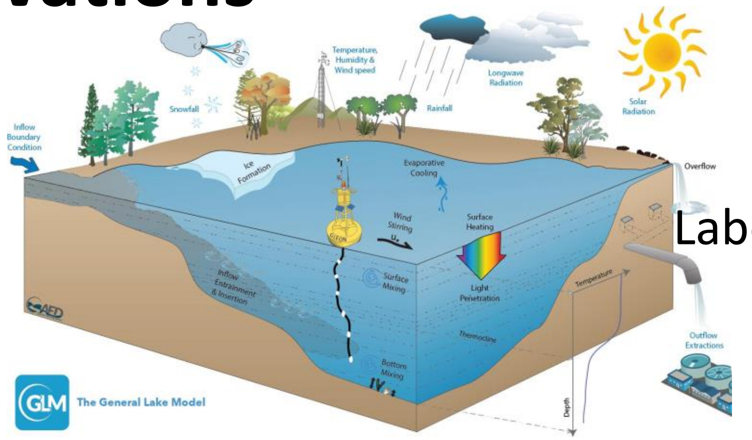
Physics-based Loss $\left(Y_{pred}\right)$

KGML for Modeling Lake Water Temperature:

Performance under varving # of observations



Process-Guided Deep Learning Predictions of Lake Water Temperature, Read et.al. WRR, Nov. 2019.



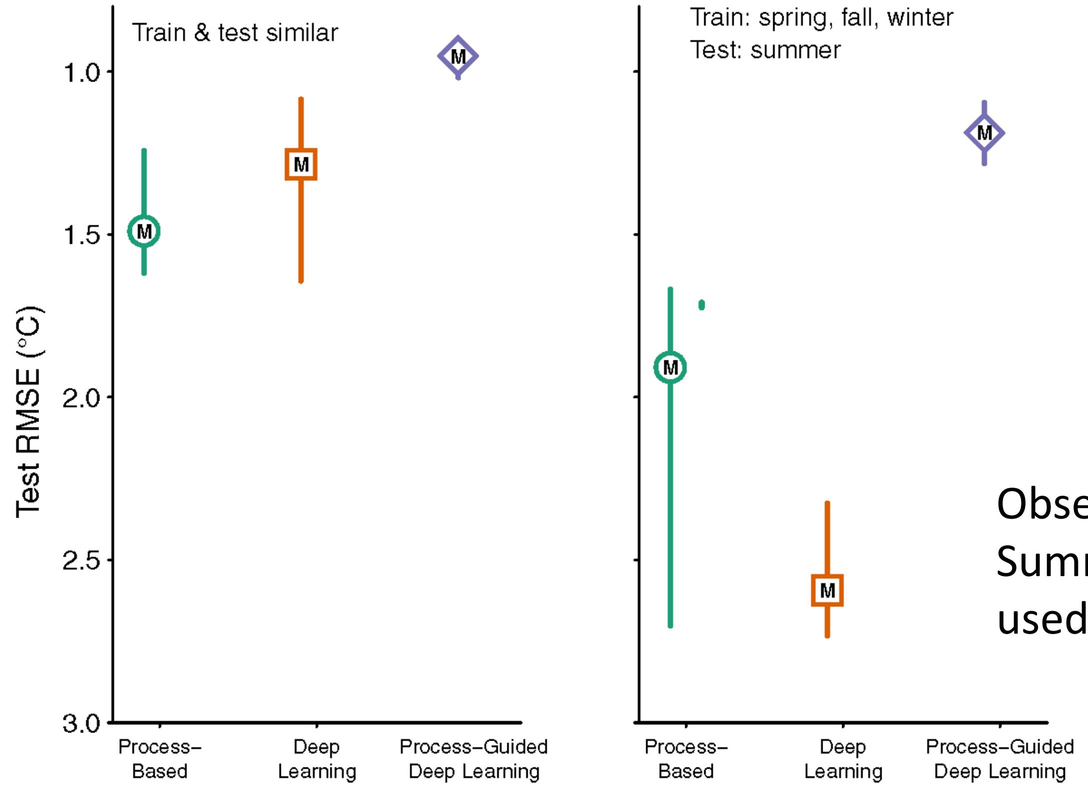
Labeled Data

GLM: State of the Art physics-based model used by USGS

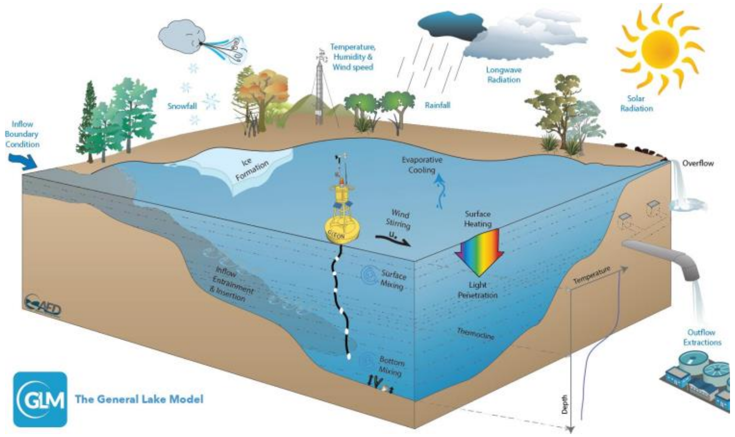
RNN: A black-box machine learning model that can incorporate time

PGRNN: A machine learning framework that leverages physics

KGML for Modeling Lake Water Temperature:
Performance in Novel Testing Scenarios



Observations from Summer seasons are used only during test



GLM: State of the Art physics-based model used by USGS

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

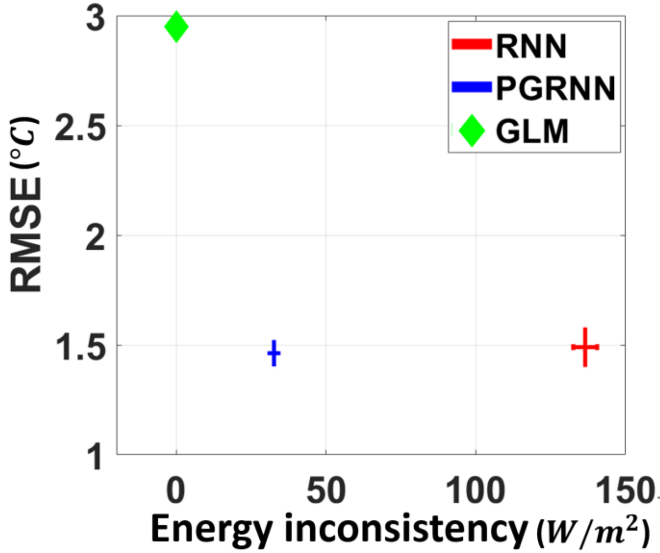
PGRNN: A machine learning framework that leverages physics

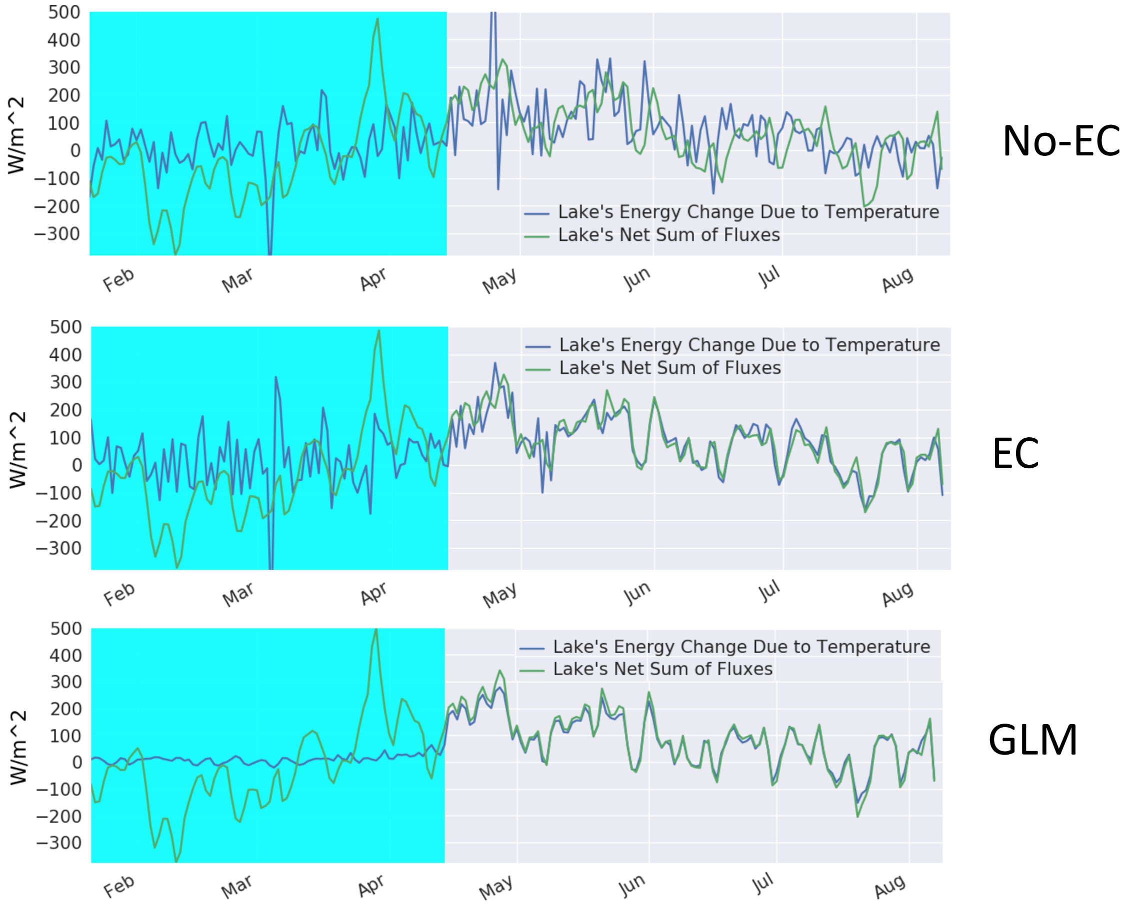
Process-Guided Deep Learning Predictions of Lake Water Temperature, Read et.al. WRR, Nov. 2019.

Can KGML Models Maintain Energy Consistency?

$$dU_t/dt = R_{SW}(1-\alpha_{SW}) + R_{LWin}(1-\alpha_{LW}) - R_{LWout} - E - H$$

- Energy conservation loss penalizes if the model predicts an “impossible” energy change on a daily scale

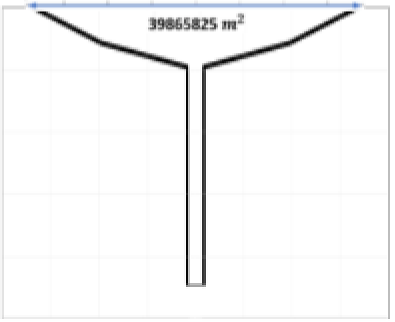




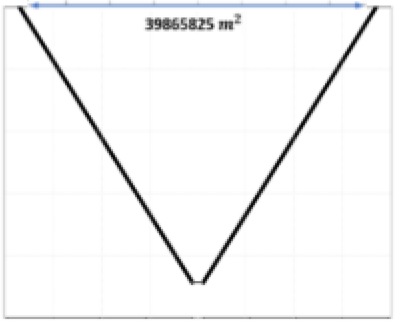
Can we pre-train PGRNN using lakes that are very different from the target lakes?

- Key Parameters
  - Area Depth Profile
  - Lake Clarity
  - Climate Conditions

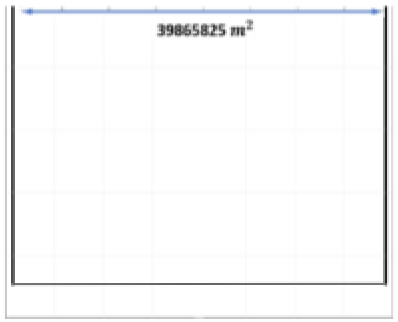
Impact of Training with Incorrect Geometries



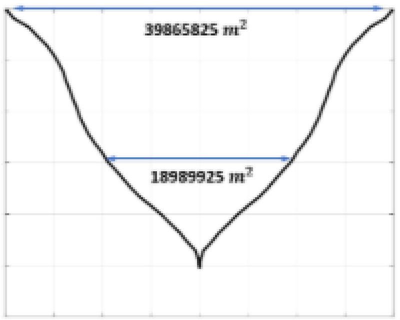
Martini



Cone



Barrel



Mendota

Method	0%	0.2%	2%	20%	100%
RNN	-	4.615(±0.173)	2.311(±0.240)	1.531(±0.083)	1.489(±0.091)
RNN^{EC}	-	4.107(±0.181)	2.149(±0.163)	1.489(±0.115)	1.471(±0.077)
$\text{RNN}^{EC,p(\text{cone})}$	2.469(±0.168)	2.056(±0.184)	1.595(±0.097)	1.452(±0.113)	1.374(±0.074)
$\text{RNN}^{EC,p(\text{barrel})}$	3.239(±0.098)	2.060(±0.144)	1.617(±0.090)	1.401(±0.098)	1.383(±0.078)
$\text{RNN}^{EC,p(\text{martini})}$	5.340(±0.110)	3.033(±0.104)	2.216(±0.141)	1.485(±0.092)	1.459(±0.059)

Concluding Remarks

- KGML offer a promising approach for addressing limitations of pure ML and pure process guided approaches.

- Future Directions:
  - How to incorporate complex physical knowledge into model learning and model architecture
  - How to model a system with multiple components (e.g., network of river streams, a complex hydrological system).
  - How to make use of real time observation data (i.e., data assimilation in KGML setting)?

Publications

- Anuj Karpatne, Gowtham Atluri, James H. Faghmous, Michael Steinbach, Arindam Banerjee, Auroop Ganguly, Shashi Shekhar, Nagiza Samatova, Vipin Kumar. **Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data.** IEEE on Knowledge and Data Engineering, vol. 29, no. 10, pp. 2318-2331, 1 October 20Transactions17. https://ieeexplore.ieee.org/document/7959606
- Jared Willard, Xiaowei Jia, Shaoming Xu, Michael Steinbach, Vipin Kumar. **Integrating Physics-Based Modeling with Machine Learning: A Survey.** April 2020. https://arxiv.org/abs/2003.04919
- Xiaowei Jia, Jared Willard, Anuj Karpatne, Jordan Read, Jacob Zwart, Michael Steinbach, Vipin Kumar. **Physics Guided RNNs for Modeling Dynamical Systems: A Case Study in Simulating Lake Temperature Profiles**. Proceedings of the 2019 SIAM International Conference on Data Mining, May 2019. doi: 101137/1.9781611975673.63 Updated, January 2020. https://arxiv.org/pdf/2001.11086.pdf
- Jordan S. Read, Xiaowei Jia, Jared Willard, Alison P. Appling, Jacob A. Zwart, Samantha K. Oliver, Anuj Karpatne, Gretchen J.A. Hansen, Paul C. Hanson, William Watkins, Michael Steinbach, Vipin Kumar. **Process-Guided Deep Learning Predictions of Lake Water Temperature**. 2019. Water Resources Research (55). https://doi.org/10.1029/2019WR024922
- Faghmous, James H., and Vipin Kumar. "**A big data guide to understanding climate change: The case for theory-guided data science.**" *Big data* 2, no. 3 (2014): 155-163. https://www.liebertpub.com/doi/full/10.1089/big.2014.0026
- Faghmous, James H., Arindam Banerjee, Shashi Shekhar, Michael Steinbach, Vipin Kumar, Auroop R. Ganguly, and Nagiza Samatova. "**Theory-guided data science for climate change**." *Computer* 47, no. 11 (2014): 74-78. **DOI:** 10.1109/MC.2014.335

9:40 AM Thu Sep 24

sites.google.com

UNIVERSITY OF MINNESOTA
Driven to Discover

Workshop on Knowledge Guided Machine Learning (KGML):

A Framework for Accelerating Scientific Discovery

Dates: **August 18-20, 2020.** (Daily times: 9:30am-12:30 and 1:30-4:30pm CDT)

Format: 100% virtual, open to public - available as zoom webinar *and* as YouTube live stream.

Contact Workshop

Email Opt-in

Email Opt-in Background Agenda Organizers

Quicklinks to session details: Introduction Aquatic Sciences Hydrology Weather/Climate Translational Biology Closing Panel

LINK TO YOUTUBE CHANNEL TO VIEW RECORDED SESSIONS. *Please note, recorded sessions will be posted approximately 48 hours post recording.

Inaugural Workshop Information:

The inaugural workshop took place August 18-20, 2020 virtually over zoom. Over 1000 people registered from over 30 countries, and a variety of topics and vibrant discussions were recorded and are available to watch here: https://z.umn.edu/kgmlworkshopyoutube.

If you would like to receive emails of future events, please Opt-in for KGML emails here.

Background:

This workshop is part of a 2-year conceptualization project funded by the NSF's Harnessing the Data Revolution (HDR) program involving researchers from the University of Minnesota, University of Wisconsin, Penn State, Colorado State University, and the University of Virginia. This project aims to develop a framework

https://sites.google.com/umn.edu/kgml/workshop