# Physics Guided Machine Learning: A New Framework 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. Khandelwal, J. Willard, M. Steinbach, G. Hansen, University of Minnesota
X. Jia, University of Pittsburgh
P. Hanson, University of Wisconsin
C. Duffy, Penn State
A. Karpatne, Virginia Tech

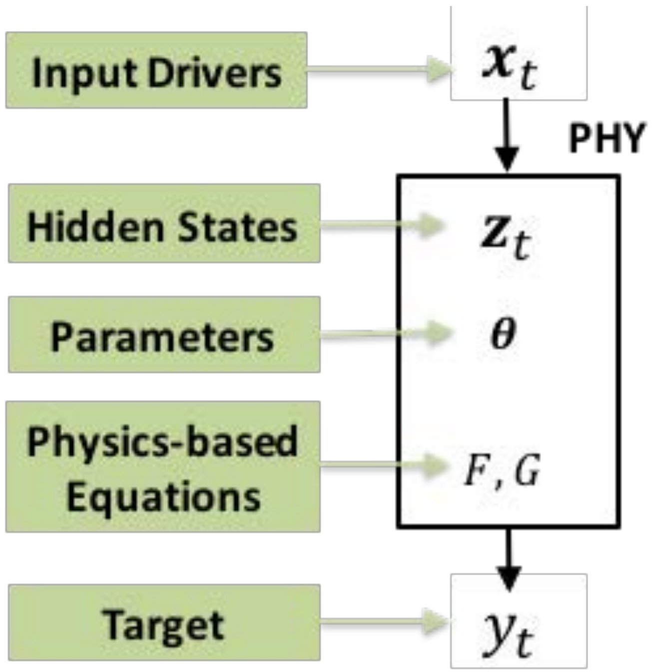

USGS science for a changing world

DARPA

UNIVERSITY OF MINNESOTA
Driven to Discover℠

# Physics-based Models of Dynamical Systems

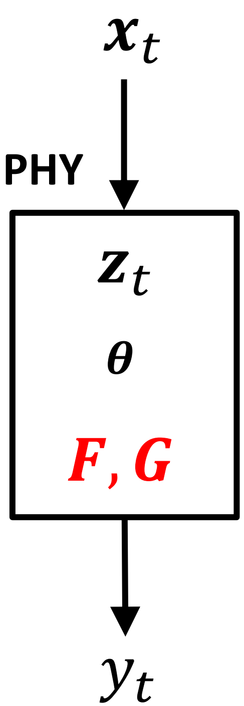
- Relationships b/w input & output variables governed by physics-based partial differential equations (PDEs)



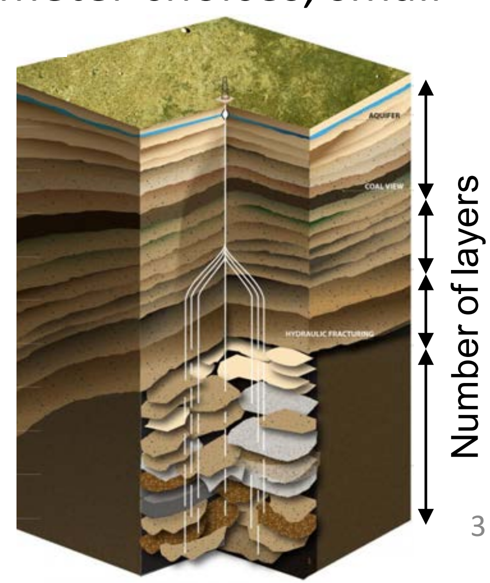

*Examples from Hydrology, Limnology, Fluid Dynamics, …*

| 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 |

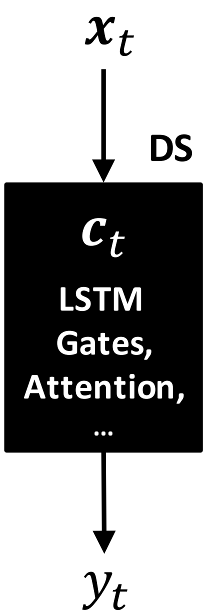
# Limitations of Physics-based Models




- Incomplete or missing physics ($F, G$)
  - Physics-based models often use approximate forms to meet "scale-speed/accuracy" trade-off
  - Results in *inherent model bias*
- Unknown parameters ($\theta$) need to be "calibrated"
  - *Computationally Expensive*
  - *Easy to overfit*: large number of parameter choices, small number of samples, heterogeneity

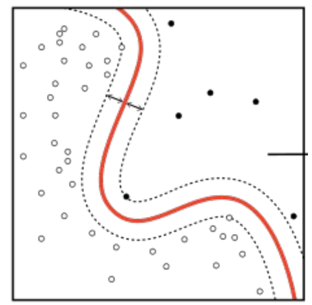


# "Black-box" Data Science Models

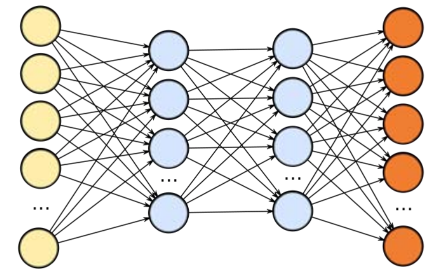



An alternative to modeling dynamical systems?

Choice of model family not governed by physics



Support Vector Machine



Deep Learning

- Hugely successful in commercial applications

DeepMind

Google Ads

NETFLIX

IMAGENET

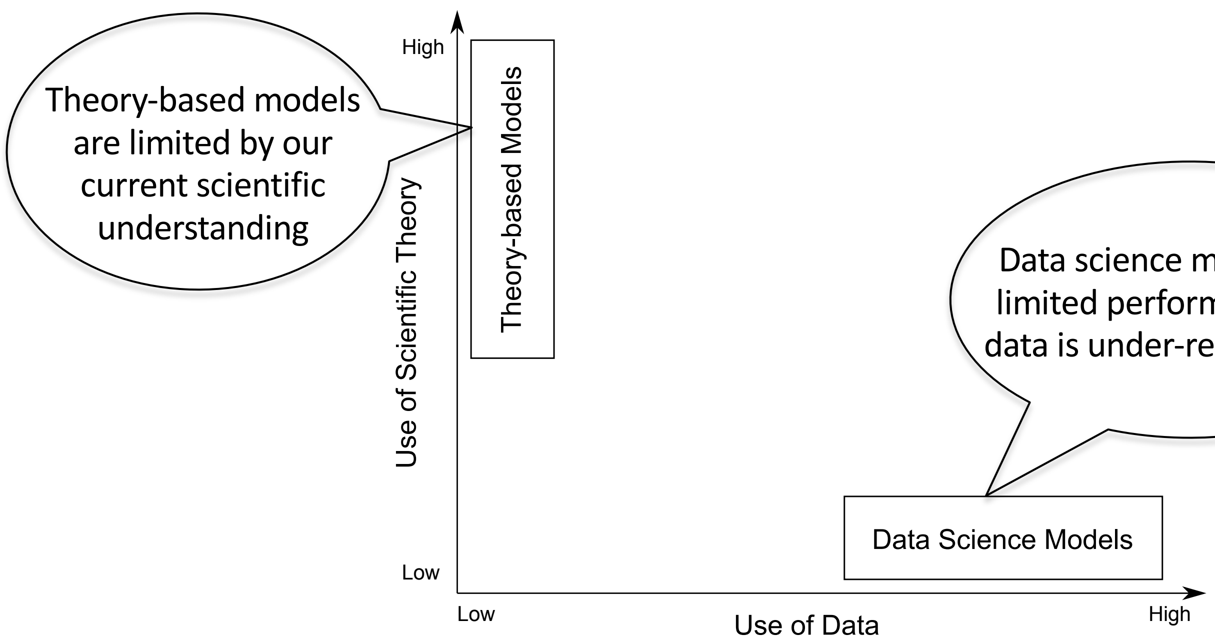
facebook

- But 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

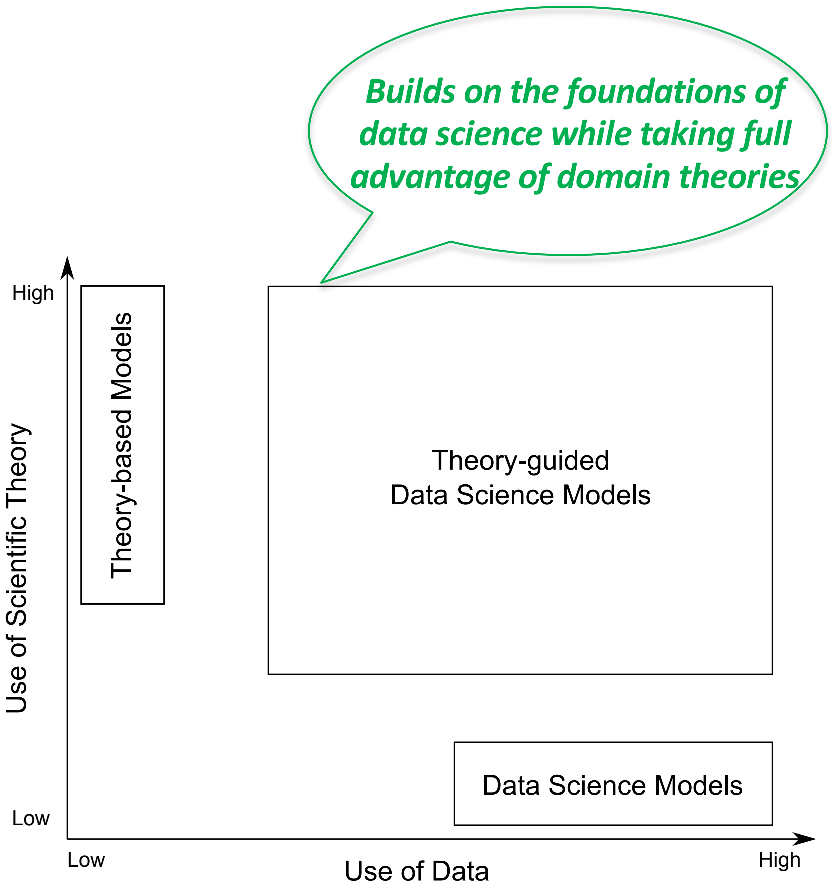
The Parable of Google Flu: Traps in Big Data Analysis

# Dichotomy b/w Scientific Theory-based and Data Science Models



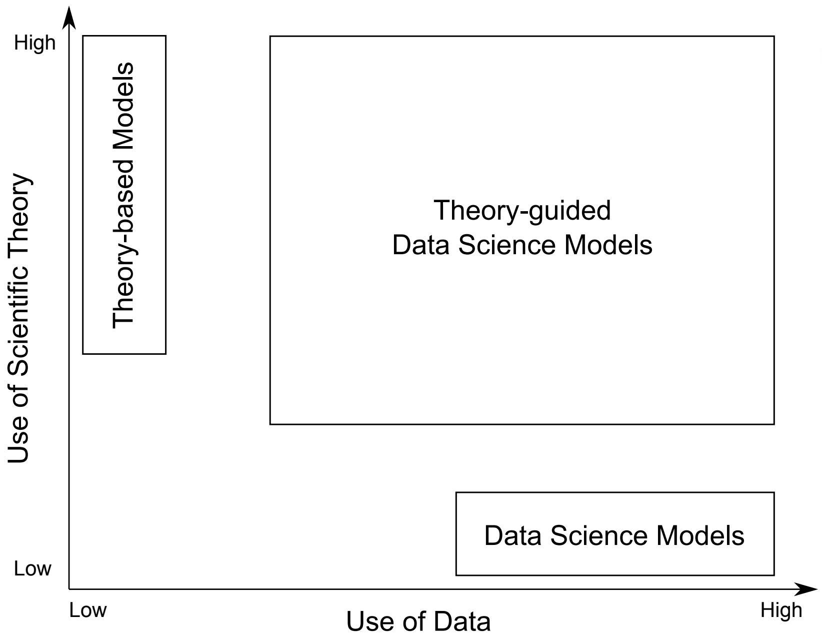

**Both use incomplete sources of information about the two key components of knowledge discovery: *scientific theory* and *data***

# Theory-guided Data Science (TGDS)




6

# 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/

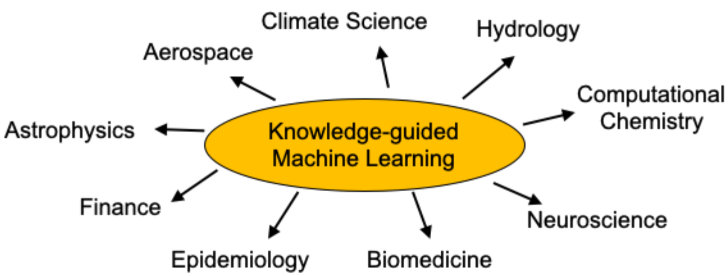
Theory-Guided Data Science for Climate Change

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

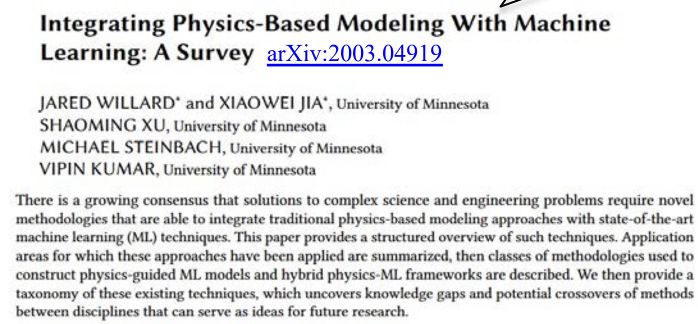
To adequately address climate change, we need novel data-science methods that account for the spatiotemporal and physical nature of climate phenomena. Only then will we be able to move from statistical analysis to scientific insights.

# Scientific Theory Guided Machine Learning:

## A Paradigm Shift in Scientific Discovery



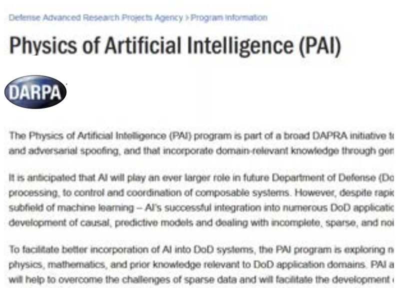

Surveys more than 300 papers



**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 DAPRA initiative t
and adversarial spoofing, and that incorporate domain-relevant knowledge through gen

It is anticipated that AI will play an ever larger role in future Department of Defense (Do
processing, to control and coordination of composable systems. However, despite rapi
subfield of machine learning – AI's successful integration into numerous DoD applicati
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**

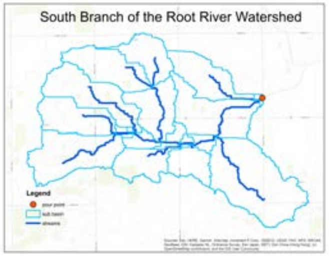
- 2020 AAAI Spring Symposium on ML in Physical Sciences
- 2020 AAAI Fall Symposium on Physics-Guided AI
- 2020 SIAM MDS Mini-symposium on Physics-guided AI
- 2020 Physics-informed Machine Learning Workshop at LANL,
- 2020 Physics-Informed Learning Machines for Multiscale and Multiphysics Problems at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

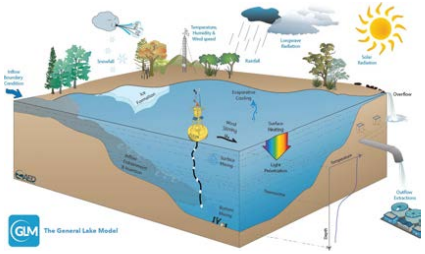
# Questions

- Can physics-guided machine learning (PGML) models
  - outperform pure physics based/mechanistic  models?
    - provide better accuracy with limited observation data?
    - produce results that are physically consistent?
    - generalize to novel testing scenarios
  - model a collection of processes that are unfolding at different scales?
  - dynamically assimilate new information/data?
  - create data at high resolution (super-resolution)?

Modeling stream flow in a watershed

**SWAT**: physics based model used by hydrological community

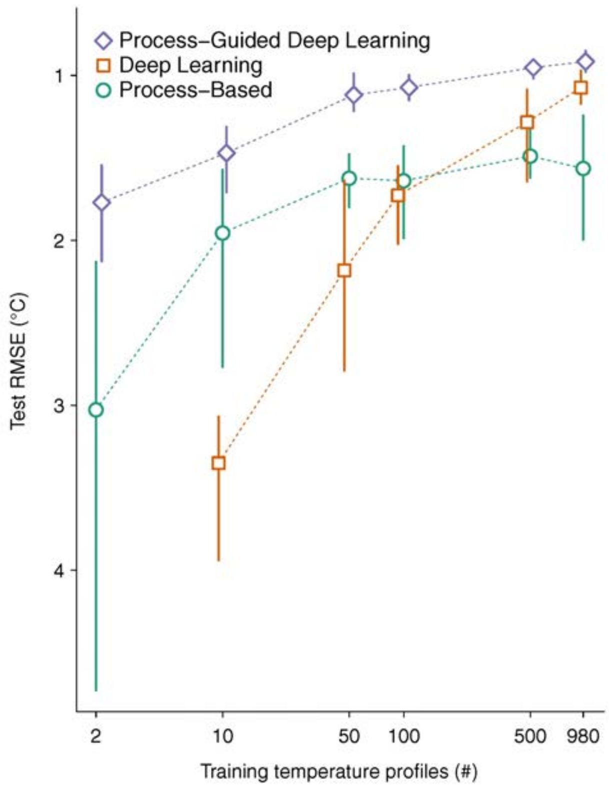


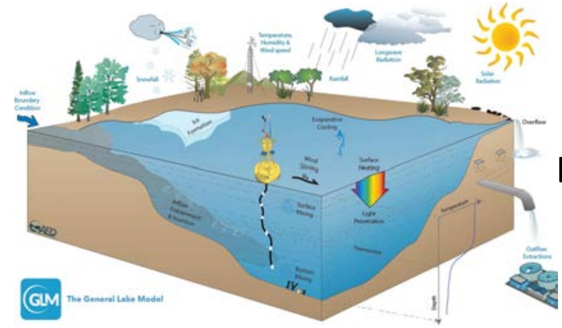

Modeling Lake Water Temperature dynamics

**GLM**: physics based model used by USGS

# PGML for Modeling Lake Water Temperature: Performance under varying # of observations




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



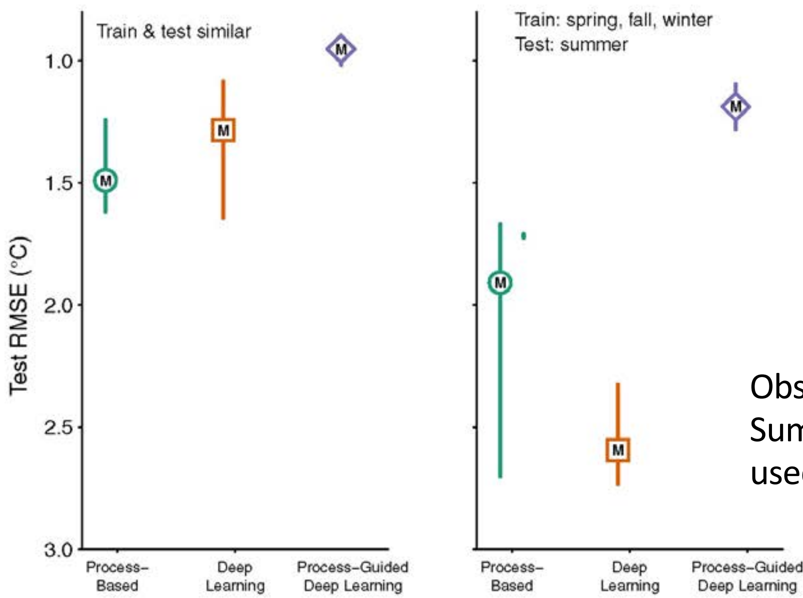

Labeled Data

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

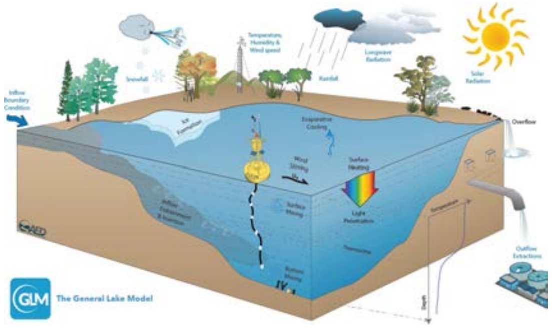
**RNN**: A black-box machine learning model that can incorporate time

**PGRNN**: A machine learning framework that leverages physics

# PGML 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

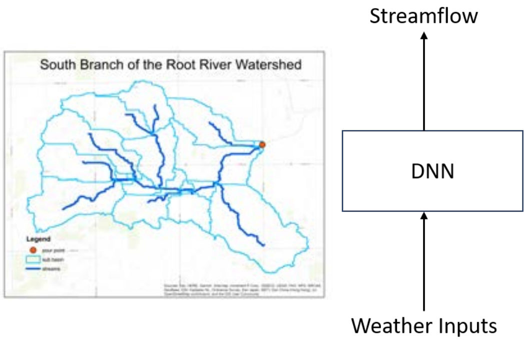
**PGRNN**: A machine learning framework that leverages physics

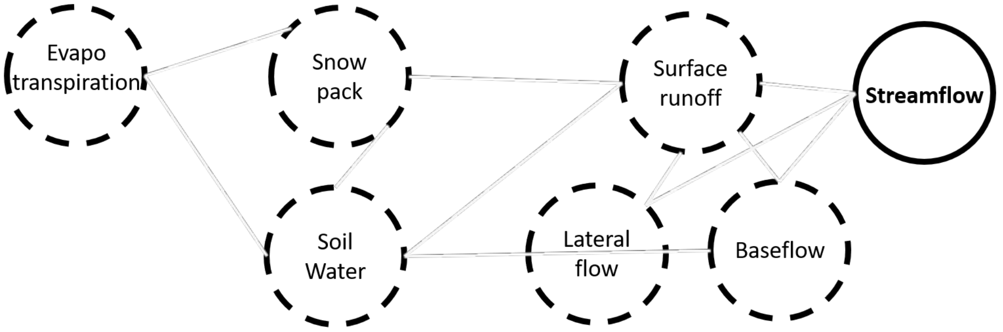
Process-Guided Deep Learning Predictions of Lake Water Temperature, Read et.al. WRR, Nov. 2019.

# KGML for Streamflow Prediction in a Watershed

- A traditional black box approach models streamflow directly from weather inputs

- Aspects of the physical system that can guide the ML models:
  1. Simultaneous modeling of inter-related variables

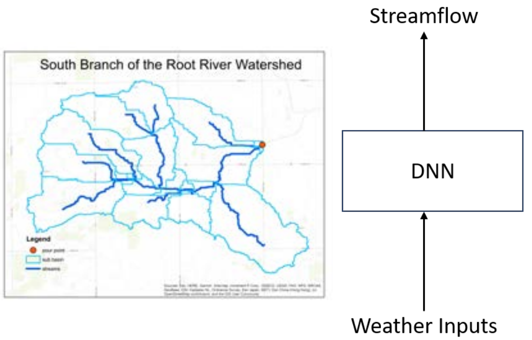


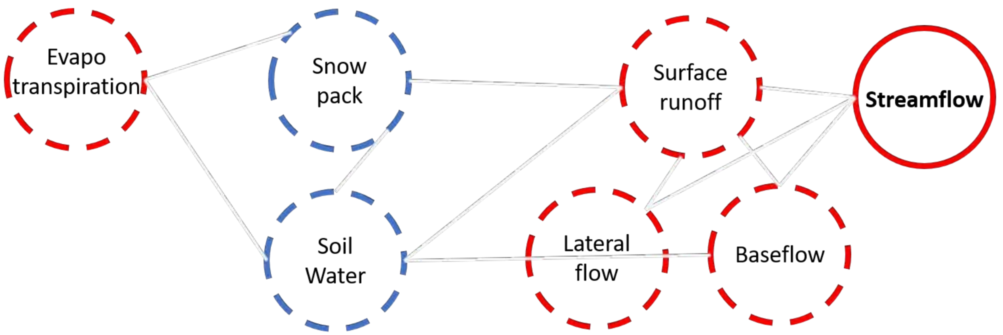



# KGML for Streamflow Prediction in a Watershed

- A traditional black box approach models streamflow directly from weather inputs

- Aspects of the physical system that can guide the ML models:
  1. Simultaneous modeling of inter-related variables
  2. Nature of variables (e.g., states vs fluxes)

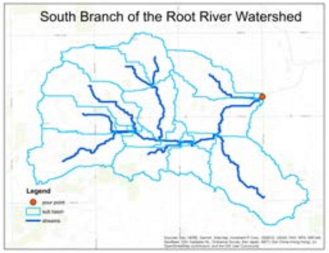


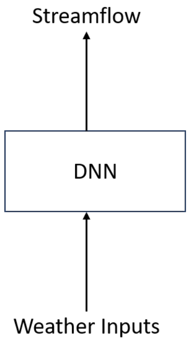



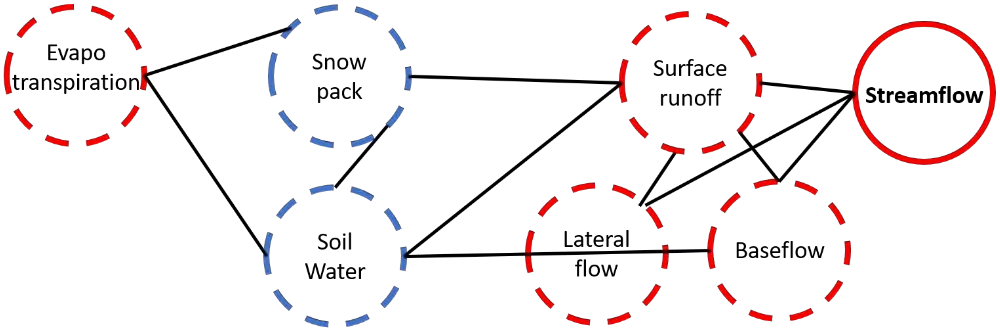
# KGML for Streamflow Prediction in a Watershed

- A traditional black box approach models streamflow directly from weather inputs

- Aspects of the physical system that can guide the ML models:
  1. Simultaneous modeling of inter-related variables
  2. Nature of variables (e.g., states vs fluxes)
  3. Dependency structure between variables

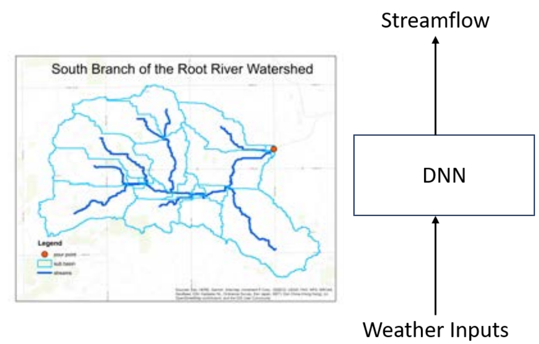


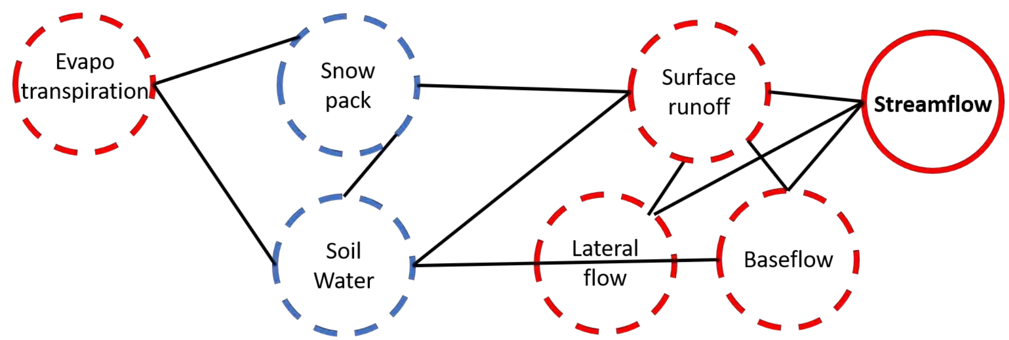



# KGML for Streamflow Prediction in a Watershed

- A traditional black box approach models streamflow directly from weather inputs




- Aspects of the physical system that can guide the ML models:
  1. Simultaneous modeling of inter-related variables
  2. Nature of variables (e.g., states vs fluxes)
  3. Dependency structure between variables
  4. Physical constraints among variables (e.g., mass conservation)



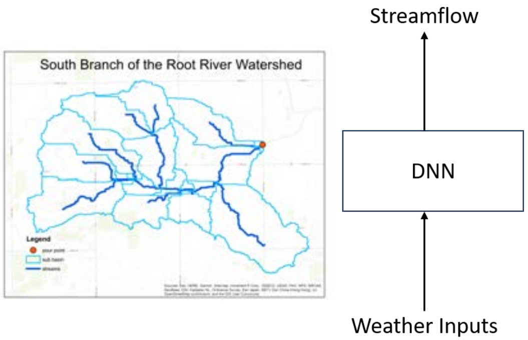

$$P - ET - Q = \sum \Delta S_i$$

$S_i$: *Soil Water, Snowpack, Ground Water* …

$P$: *Precipitation*, $ET$: *Evapotranspiration*, $Q$: *Streamflow*

# KGML for Streamflow Prediction in a Watershed

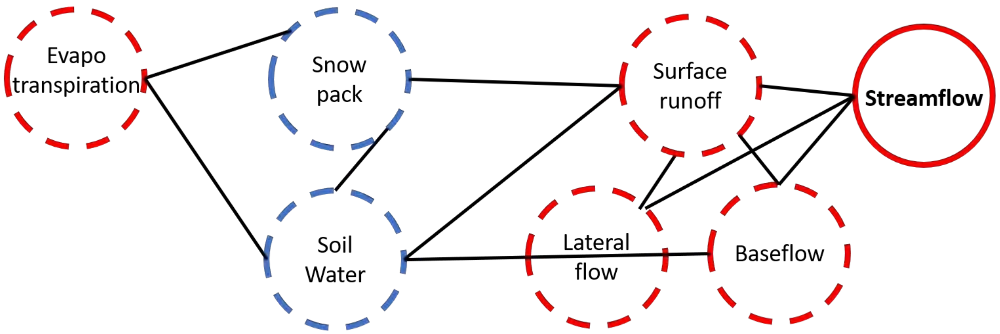
- A traditional black box approach models streamflow directly from weather inputs **(Basic)**




- Aspects of the physical system that can guide the ML models:
  1. Simultaneous modeling of inter-related variables **(Multi-task)**
  2. Nature of variables (e.g., states vs fluxes) **(State-aware)**
  3. Dependency structure between variables **(Dependency-aware)**
  4. Physical constraints among variables (e.g., mass conservation) **(Constraint-aware)**

| Model | RMSE |
|---|---|
| Basic | 0.63 |
| Multi-task | 0.55 |
| Multi-task + State-aware | 0.40 |
| Multi-task + State + Dependency aware | 0.30 |

- 1000-year simulation from the SWAT model for South Branch of the Root River at Garden Meadow (1,112 ha.) in SE Minnesota.
- Experiment Setting:
  - First 600 years for training, last 400 years for testing
  - Sequence length for LSTM 180 days
  - Hidden features = 64



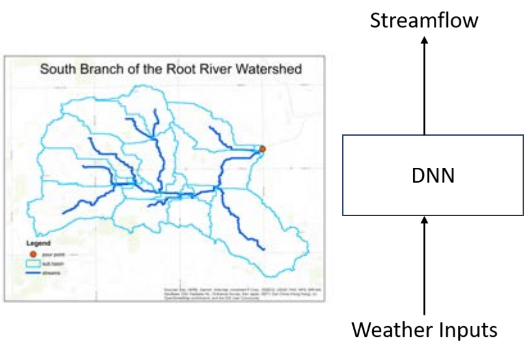

$$\boldsymbol{P - ET - Q = \sum \Delta S_i}$$

$\boldsymbol{S_i}$: ***Soil Water, Snowpack, Ground Water*** …

$\boldsymbol{P}$: ***Precipitation***, $\boldsymbol{ET}$: ***Evapotranspiration***, $\boldsymbol{Q}$: ***Streamflow***

# KGML for Streamflow Prediction in a Watershed

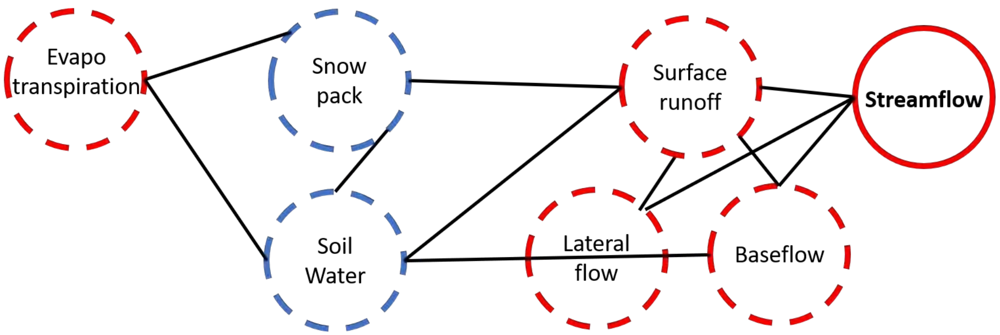
- A traditional black box approach models streamflow directly from weather inputs **(Basic)**




- Aspects of the physical system that can guide the ML models:
  1. Simultaneous modeling of inter-related variables **(Multi-task)**
  2. Nature of variables (e.g., states vs fluxes) **(State-aware)**
  3. Dependency structure between variables **(Dependency-aware)**
  4. Physical constraints among variables (e.g., mass conservation) **(Constraint-aware)**

| Model | RMSE |
|---|---|
| Basic | 0.63 |
| Multi-task | 0.55 |
| Multi-task + State-aware | 0.40 |
| Multi-task + State + Dependency aware | 0.30 |

- 1000-year simulation from the SWAT model for South Branch of the Root River at Garden Meadow (1,112 ha.) in SE Minnesota.
- Experiment Setting:
  - First 600 years for training, last 400 years for testing
  - Sequence length for LSTM 180 days
  - Hidden features = 64




$$\boldsymbol{P - ET - Q = \sum \Delta S_i}$$

$\boldsymbol{S_i: Soil\ Water, Snowpack, Ground\ Water}$ ...

$\boldsymbol{P: Precipitation, ET: Evapotranspiration,\ Q: Streamflow}$

# Concluding Remarks

- PGML 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
- Khandelwal, Ankush, Shaoming Xu, Xiang Li, Xiaowei Jia, Michael Stienbach, Christopher Duffy, John Nieber, and Vipin Kumar. "**Physics Guided Machine Learning Methods for Hydrology.**" *arXiv preprint arXiv:2012.02854* (2020).

# Acknowledgements

- Collaborators and Team Members

Anuj Karpatne
Virginia Tech

Jordan Read
USGS

Jacob Zwart
USGS

Xiaowei Jia
UMN

Jared Willard
UMN

Alison Appling (USGS), Samantha Oliver (USGS), Gretchen Hansen (UMN), Paul Hanson (U Wisconsin), William Watkins (USGS)