Machine learning and differentiable modeling for Geosciences

Chaopeng Shen

[1]Civil and Environmental Engineering
Penn State University

cshen@engr.psu.edu

https://github.com/mhpi
Hydroml.org

Hydroml.org
HydroML Symposium, May 22-26, 2022, Penn State
HydroML 2, May 2023, Berkeley, CA

Dapeng Feng, Farshid Rahmani, Tadd Bindas, Yalan Song, Jiangtao Liu, Doaa Aboelyazeed, Kamlesh Sawadekar

About me

- Ph.D. Michigan State in Env. Engr.
- Postdoc Lawrence Berkeley National Lab
- Associate Editor, Water Resources Research
  Specialty Chief Editor, Frontiers in Water: Water and AI.
- "Grew up" as a process-based modeler, solving PDEs. See both sides of the story.
- Got into ML since 2016.

Overview

- ***What*** is the fundamental strengths of ML models compared to process-based models?
- ***What*** is *differentiable modeling (DM) in geosciences*?
- ***What*** can DM bring into global hydrology?

Shen et al., 2023 Nature Reviews Earth & Environment https://t.co/qyuAzYPA6Y

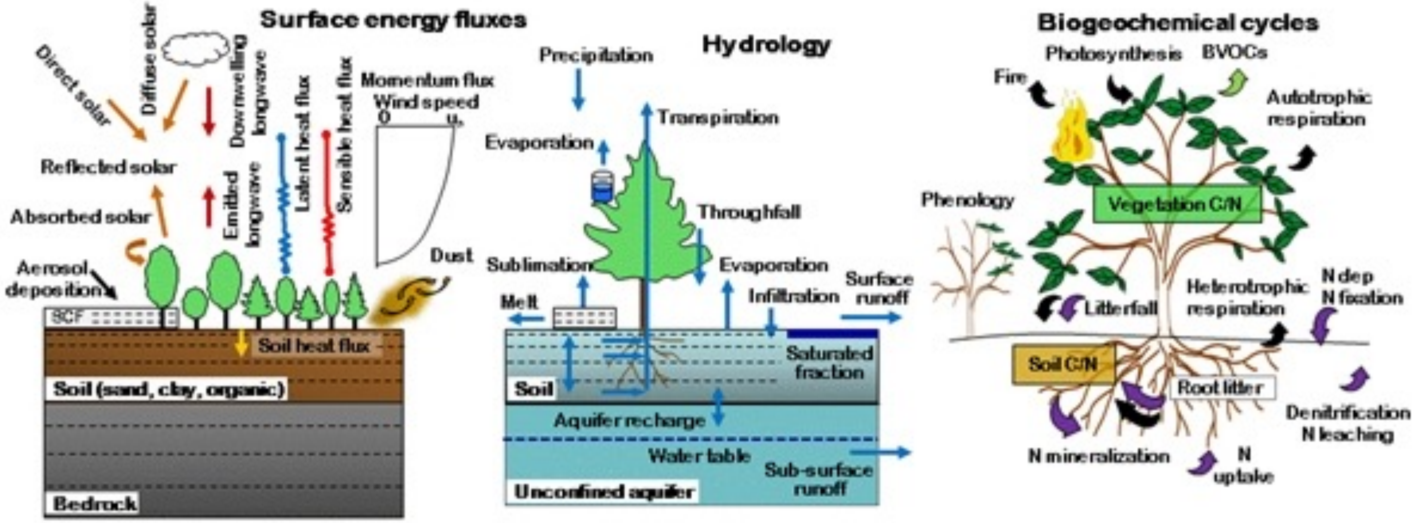
nature reviews earth & environment https://doi.org/10.1038/s43017-023-00450-9

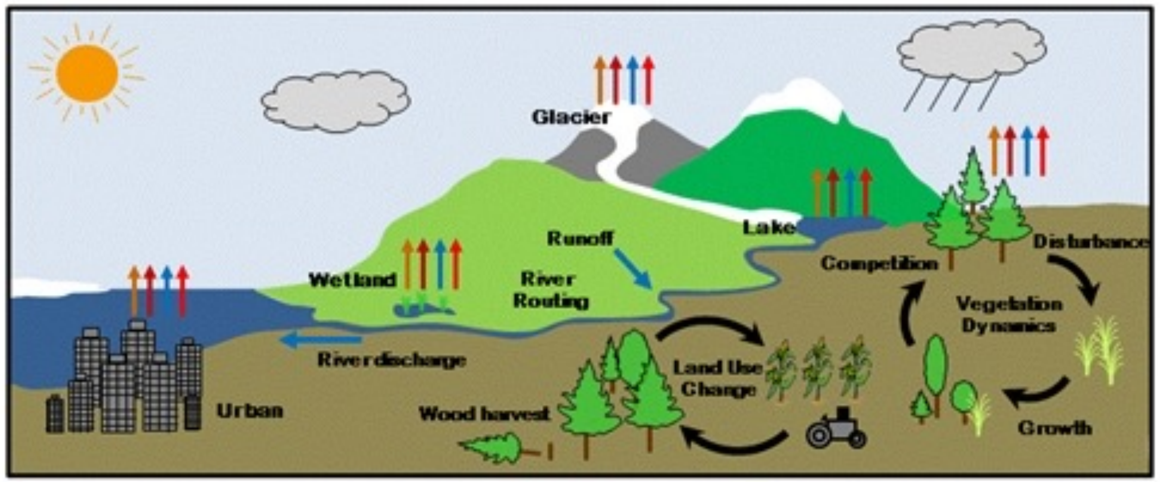
Perspective Check for updates

Differentiable modelling to unify machine learning and physical models for geosciences

A list of authors and their affiliations appears at the end of the paper

Process-based Earth-system models were highly valuable but some challenges emerged…





- Increasing complexity
- Difficult to evolve quickly with more big data.
- May contain problematic assumptions.
- Influenced by human intuition & biases

What is DL and why DL?

a rebranding of neural networks featuring

(i) Large capacity

(ii) Hidden layers that automatically extract features

(iii) Improved architecture/regularization

(iv) Working directly with data

Water Resources Research
AN AGU JOURNAL
Review Article | Open Access

A trans-disciplinary review of deep learning research and its relevance for water resources scientists

Chaopeng Shen

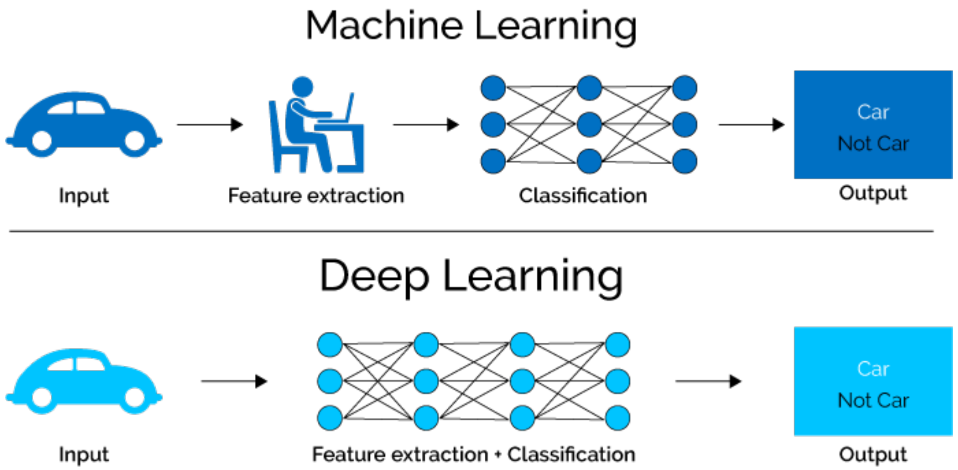
First published: 30 August 2018 | https://doi.org/10.1029/2018WR022643

$X \rightarrow Y$

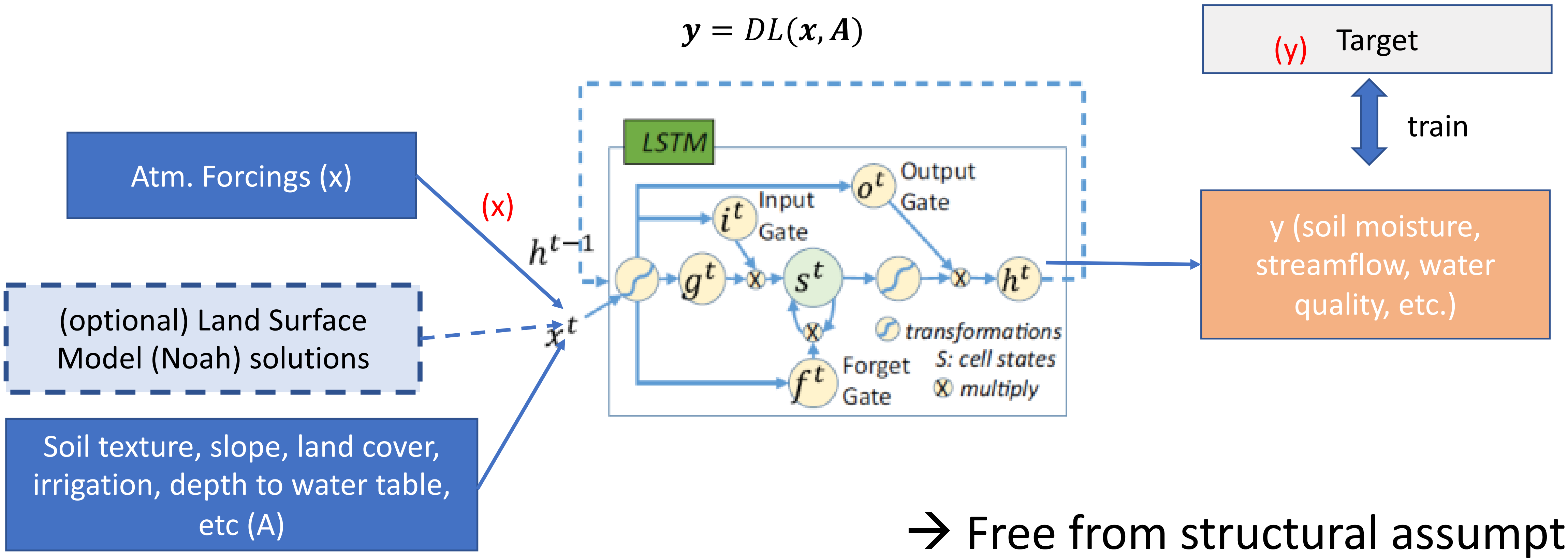
a primary value proposition is the avoidance of expertise!

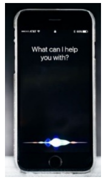
Three phases

1. Use ML to learn where the limit is.
2. Understand the gaps in our knowledge.
3. Using ML to unify across domains.



Hydrologic DL phase 1.
A hydrologic model w/o structural assumptions…





→ Free from structural assumptions

→ A chance to start anew!

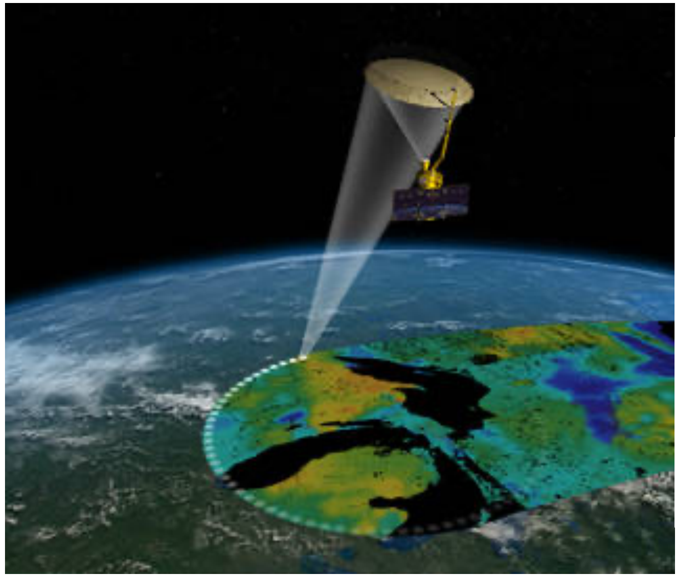
→ A chance to see where the limit is!

Case studies– first phase of DL in water

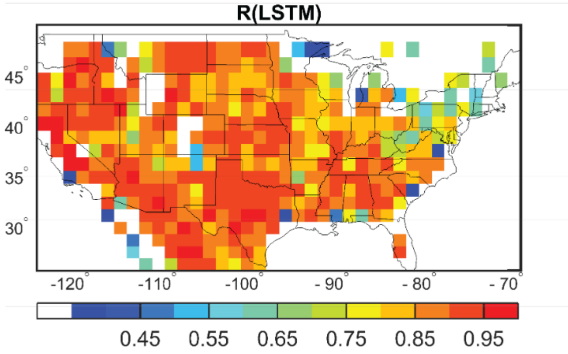
- Soil Moisture Active Passive (SMAP)
    - Launched recently (2015/04)
    - 2~3 days revisit time
    - Senses moisture-dependent top surface soil

- Streamflow modeling
    - Daily data
    - Accompanying attributes
    - With reservoirs, in data-sparse regions

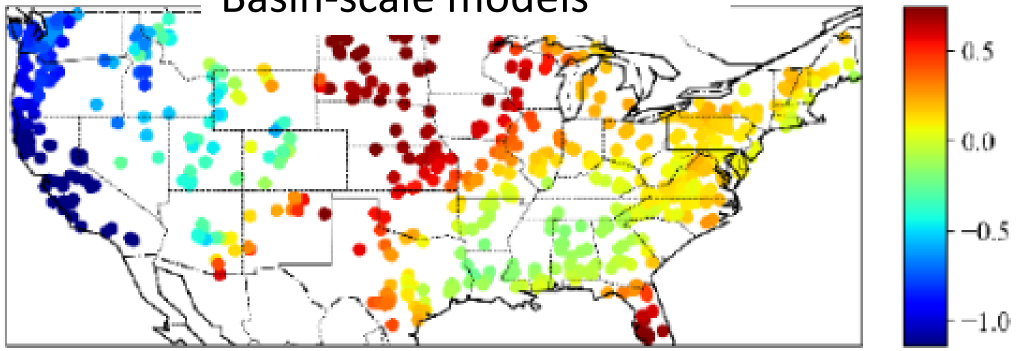
- Dissolved oxygen
- Water temperature
- Sediment
- Snow water equivalent



Gridded models



Basin-scale models



Long-term projections (first-phase of DL in hydrology)

- Examined comparison with in-situ data & long-term projections

Geophysical Research Letters

Research Letter | Full Access

Prolongation of SMAP to Spatiotemporally Seamless Coverage of Continental U.S. Using a Deep Learning Neural Network

Kuai Fang, Chaopeng Shen, Daniel Kifer, Xiao Yang

First published: 16 October 2017 | https://doi.org/10.1002/2017GL075619 | Cited by: 3

Water Resources Research

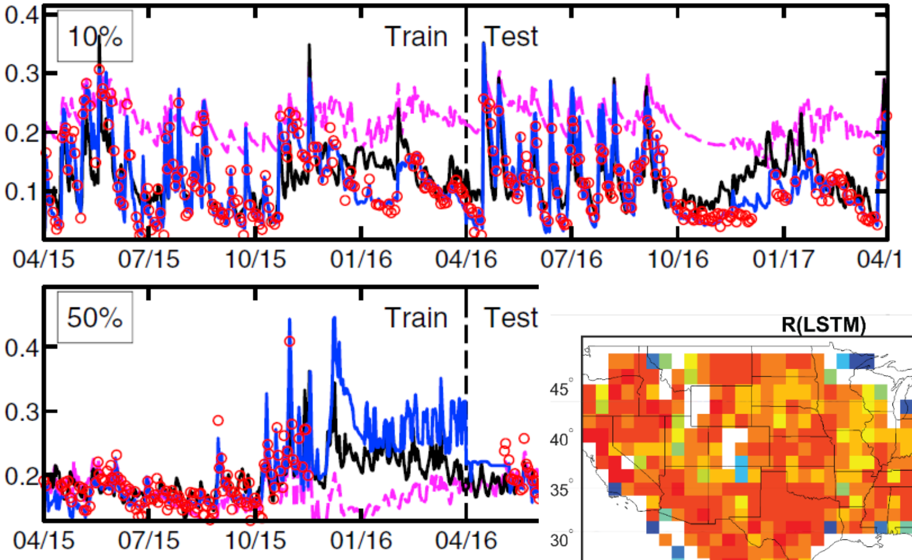
RESEARCH ARTICLE
10.1029/2019WR026793

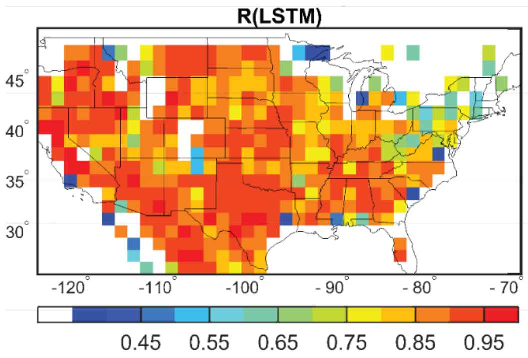
Special Section:
Big Data & Machine Learning in Water Sciences: Recent Progress and Their Use in Advancing Science

Enhancing Streamflow Forecast and Extracting Insights Using Long-Short Term Memory Networks With Data Integration at Continental Scales

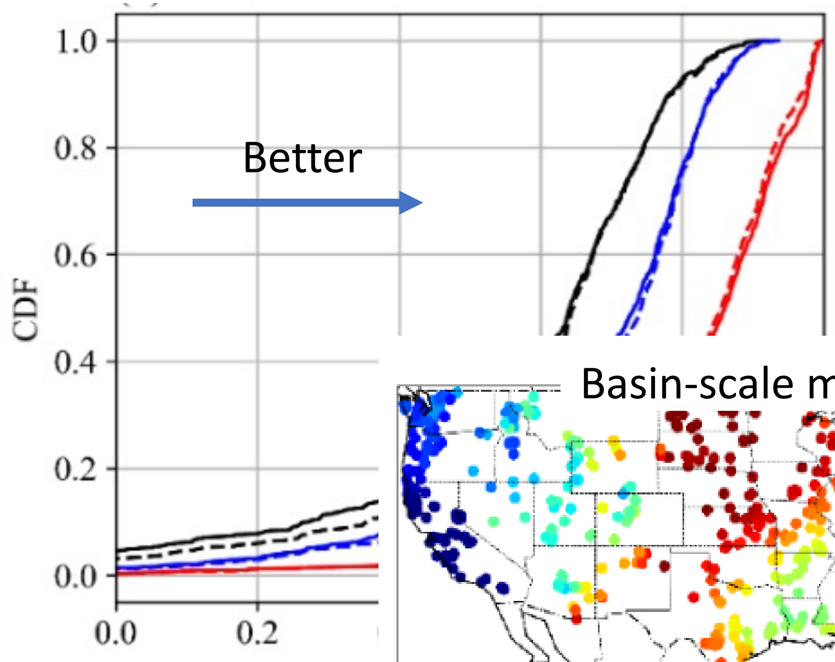
Dapeng Feng[1], Kuai Fang[1,2], and Chaopeng Shen[1]

[1]Civil and Environmental Engineering, Pennsylvania State University, State College, PA, USA, [2]Now at: Earth System Science, Stanford University, Stanford, CA, USA

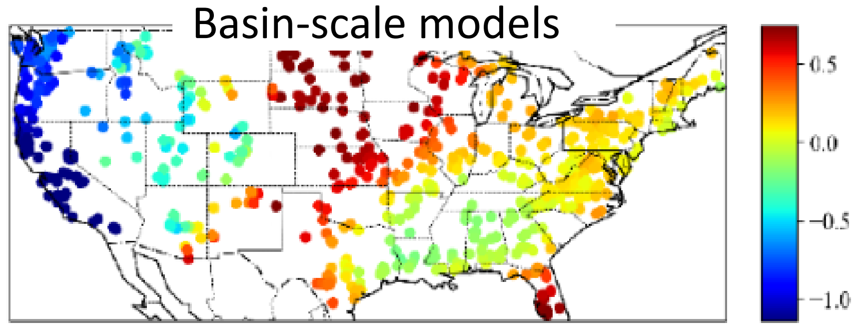




Fang et al., 2017

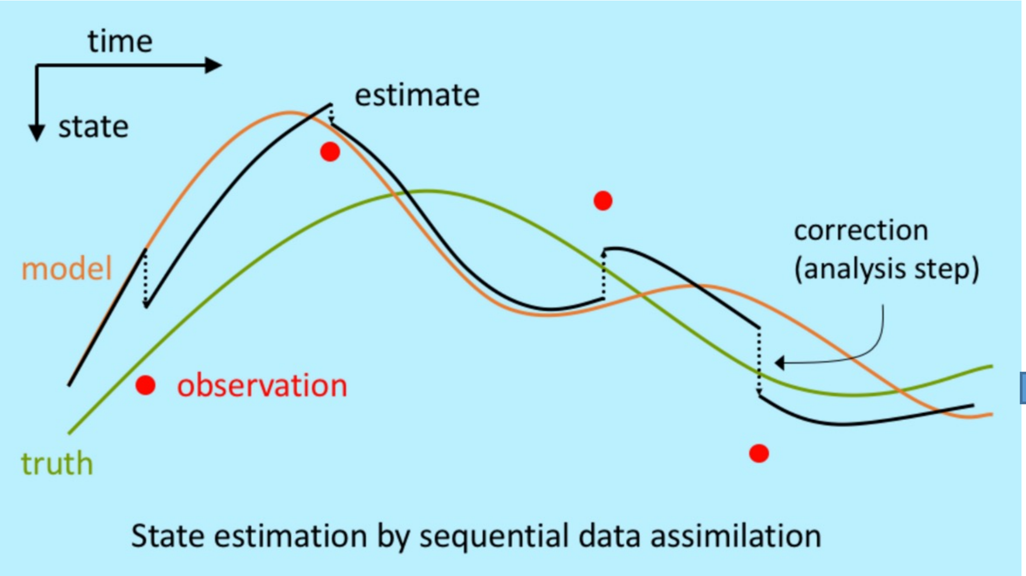


Shows us the information limit!



Short-term forecast

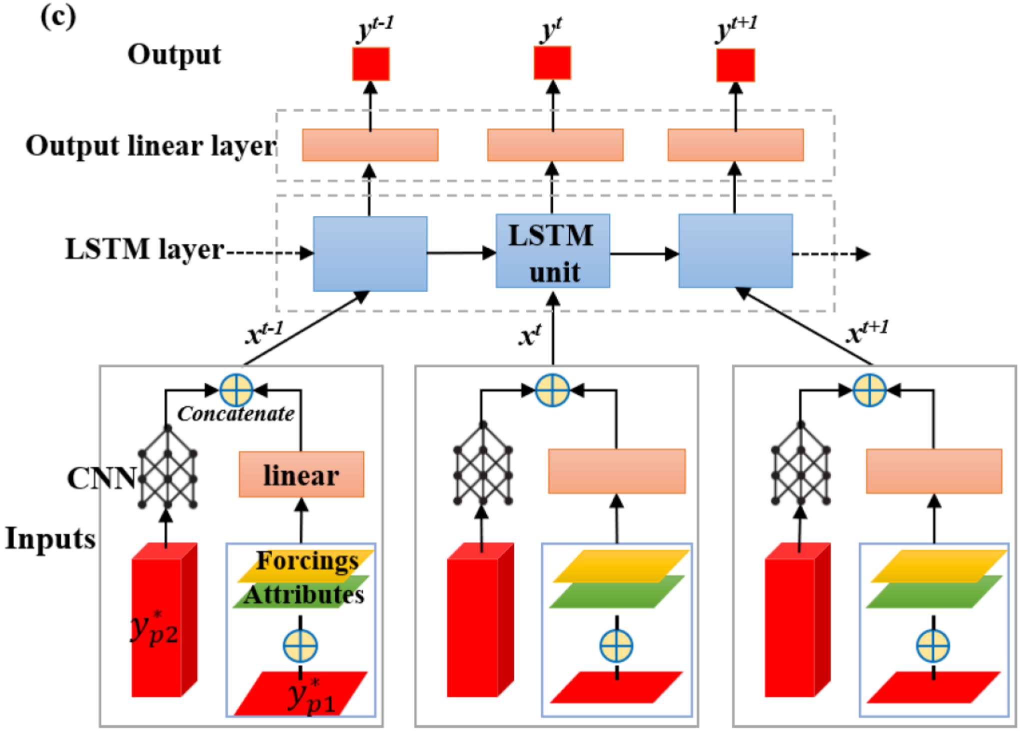
- Traditional “data assimilation” scheme



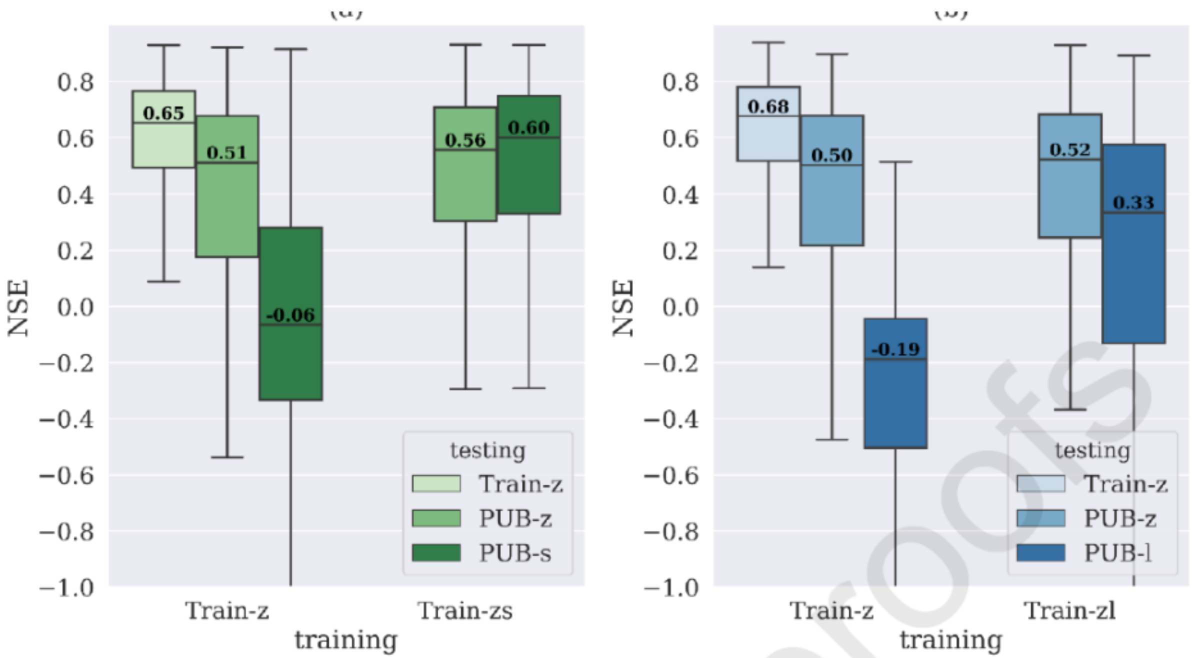
Simulation → Observation – (ENKF) → Correction

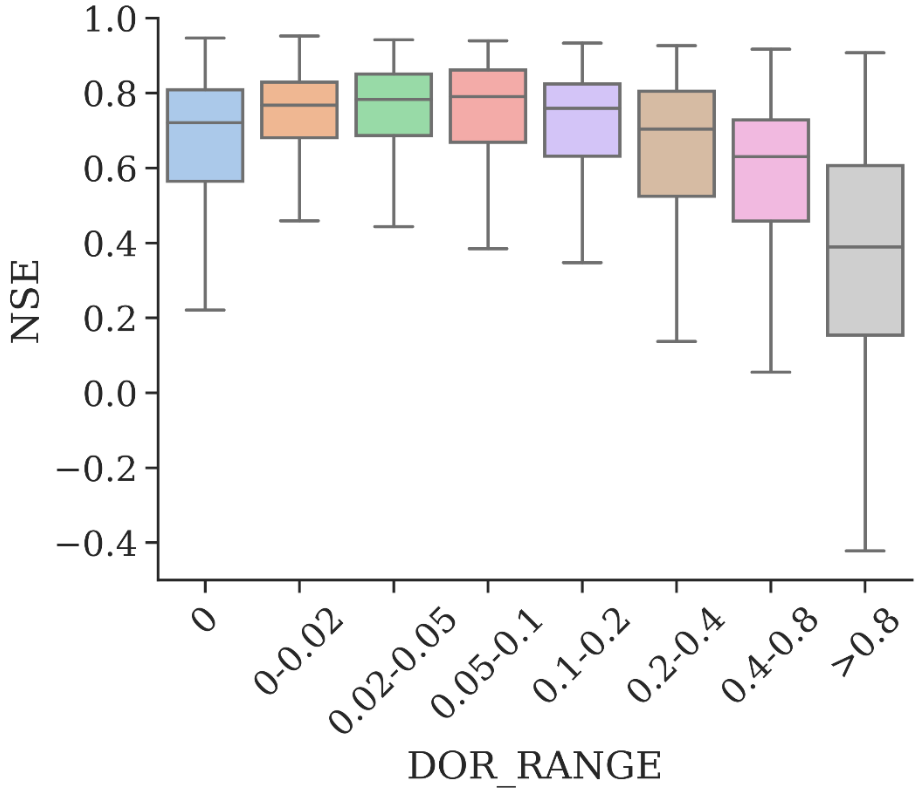
Choices: covariance matrix, what to include, how to solve, bias correction, etc.

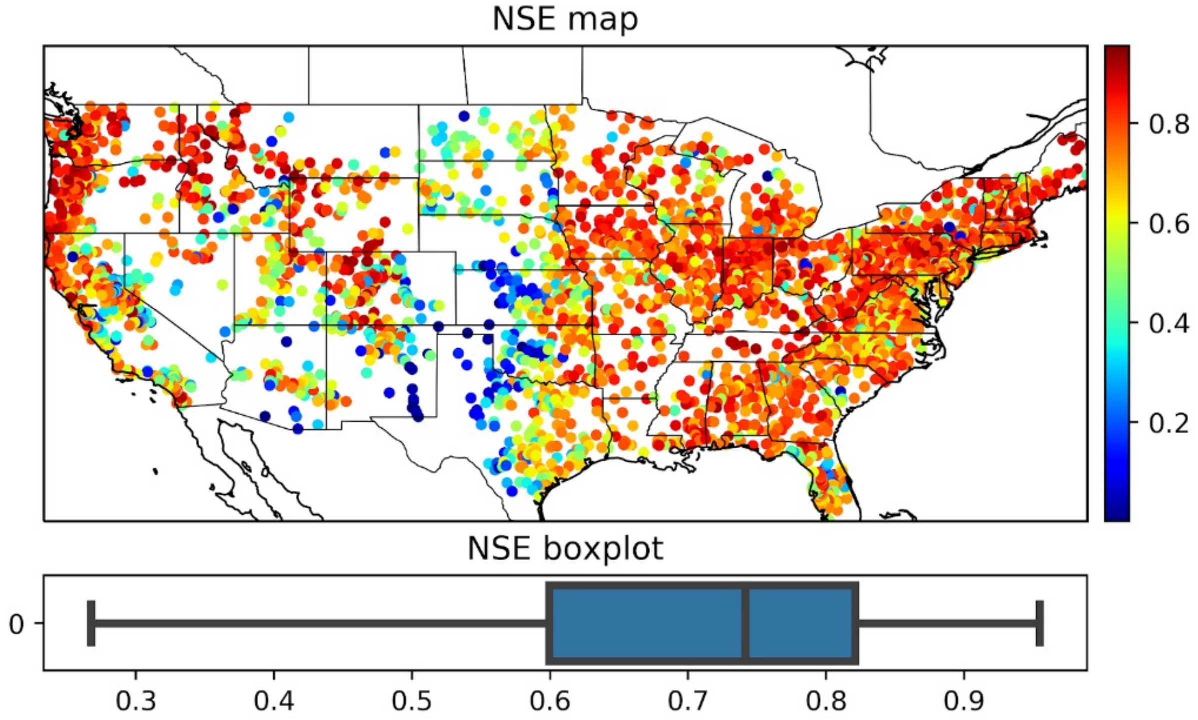
Observation → Corrected Simulation



Q^{t-1}, Q^{t-2}, etc







ELSEVIER

Journal of Hydrology

Available online 16 May 2021, 126455

In Press, Journal Pre-proof

Research papers

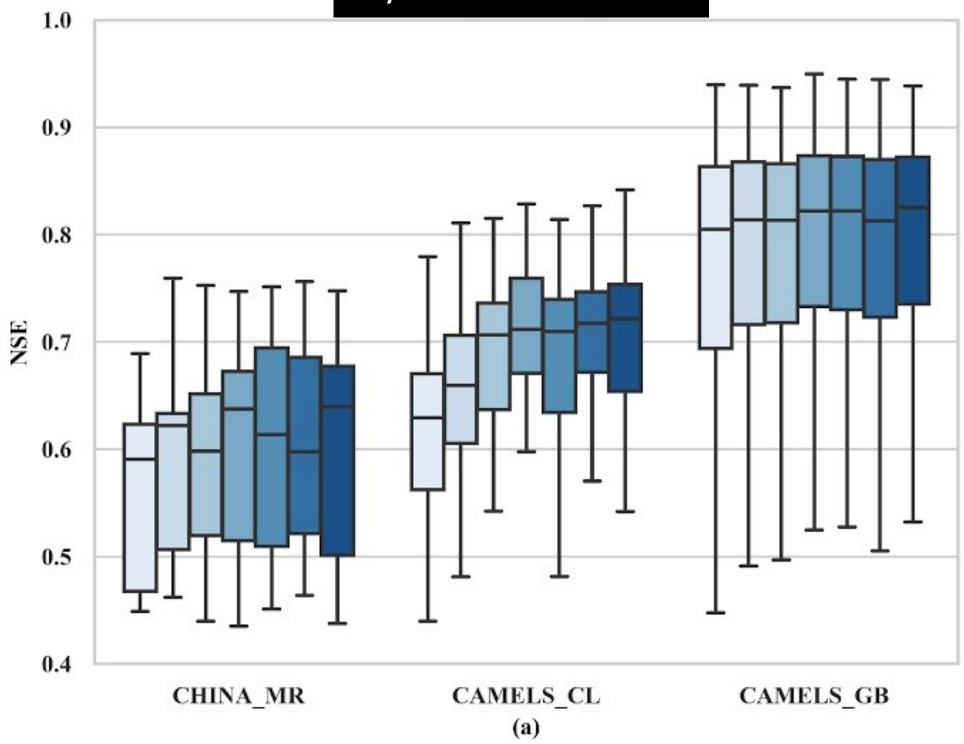
Continental-scale streamflow modeling of basins with reservoirs: towards a coherent deep-learning-based strategy

Wenyu Ouyang [a], Kathryn Lawson [b], Dapeng Feng [b], Lei Ye [a], Chi Zhang [a], Chaopeng Shen [b]

Sparse-data region

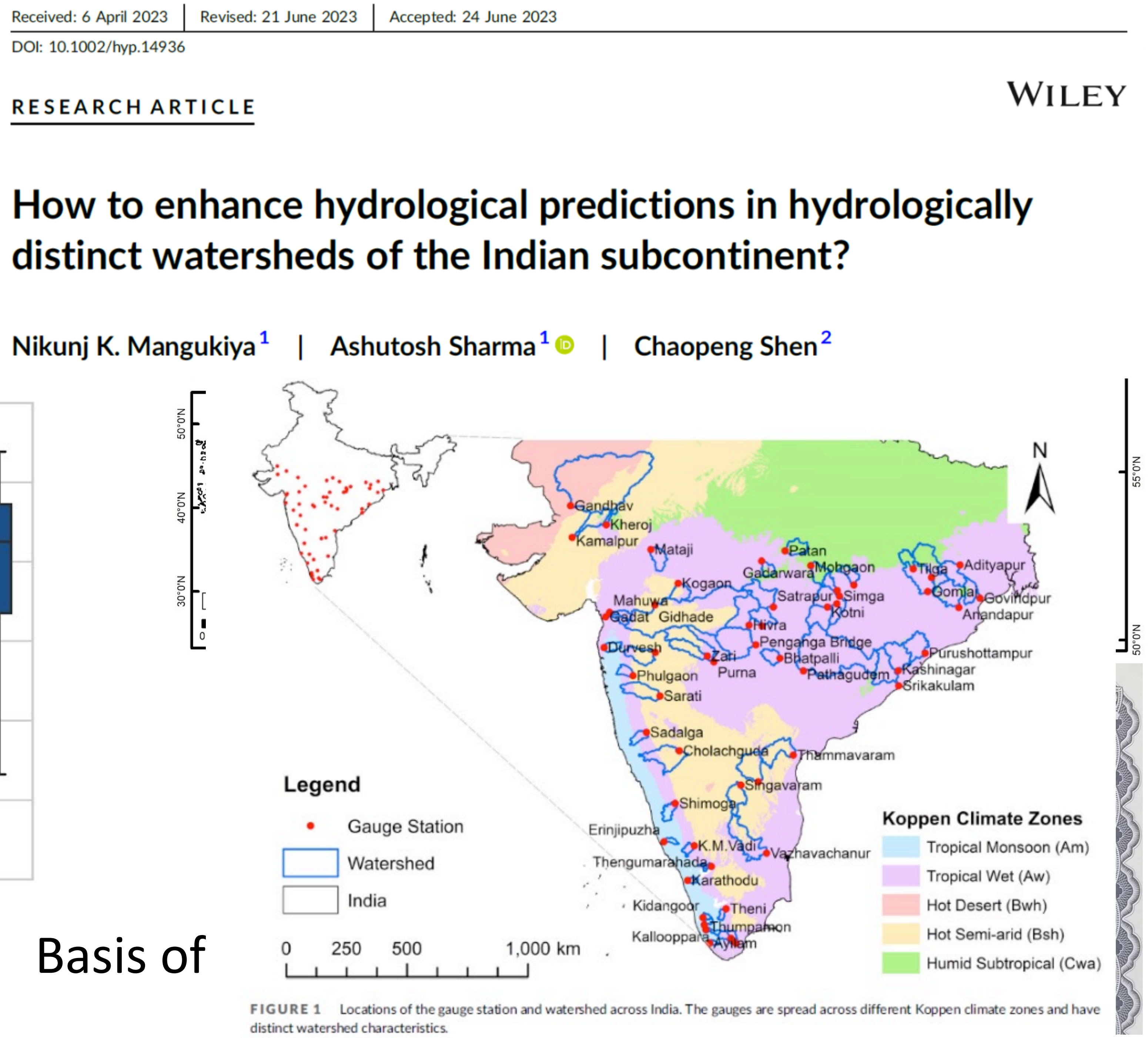
- Transfer learning

1 year local data



Ma et al., WRR
https://doi.org/10.1029/2020WR028600

Basis of

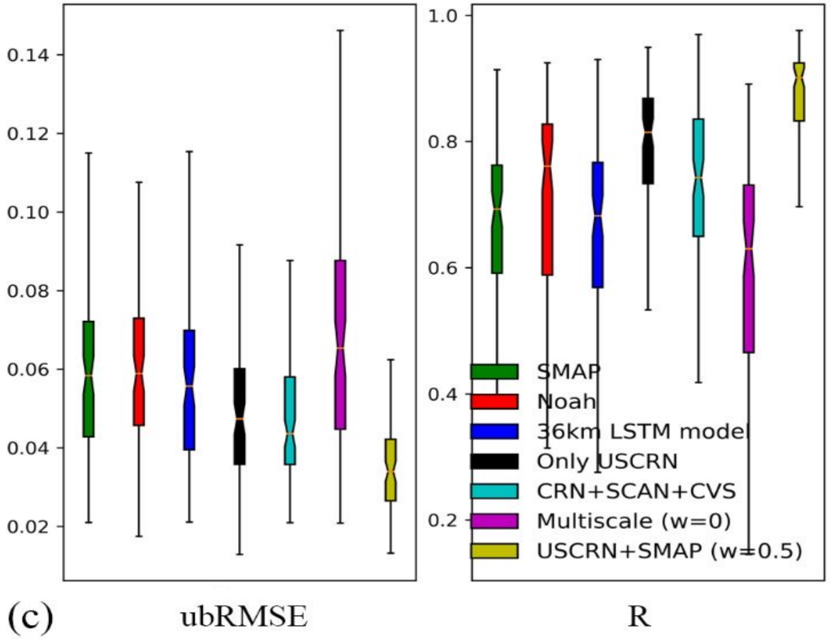


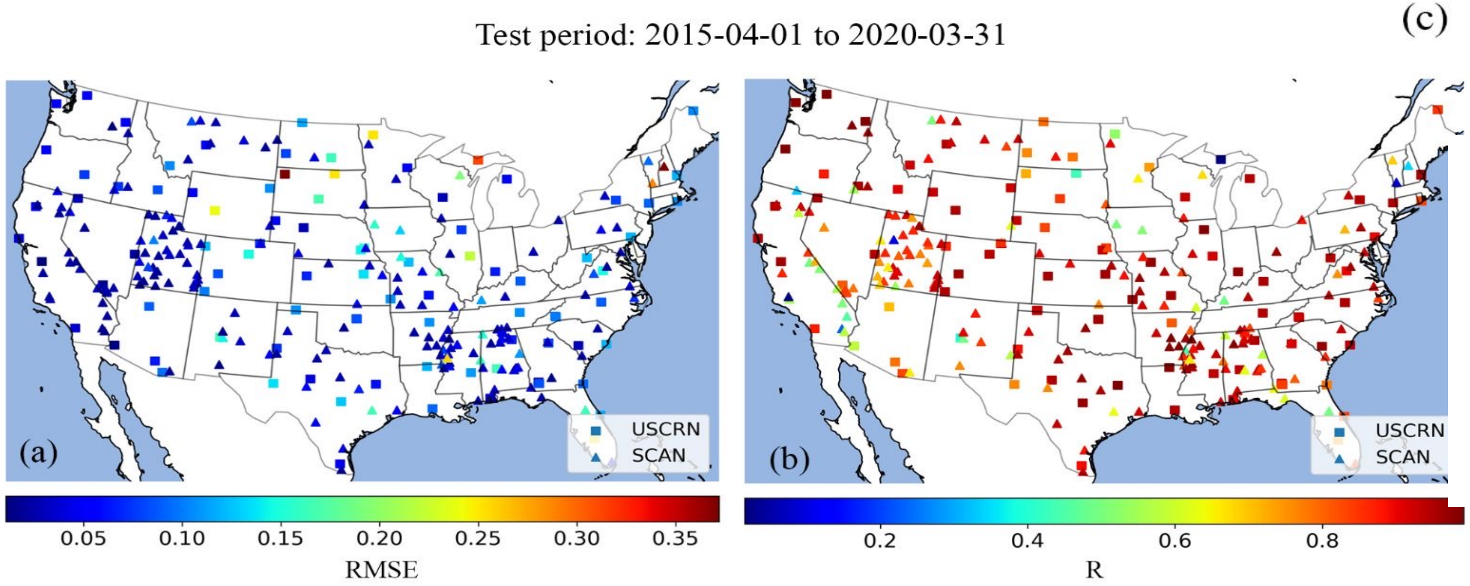
Multiscale soil moisture – learning from two teachers

High-res sim - - - - - In-situ data

Coarsen ↓

Low-res sim - - - - - Satellite data





Geophysical Research Letters

Research Letter | Full Access

A multiscale deep learning model for soil moisture integrating satellite and in-situ data

Jiangtao Liu, Farshid Rahmani, Kathryn Lawson, Chaopeng Shen

First published: 14 March 2022 | https://doi.org/10.1029/2021GL096847

Water quality

nature water

Explore content ⌄ About the journal ⌄ Publish with us ⌄ Subscribe

nature > nature water > articles > article

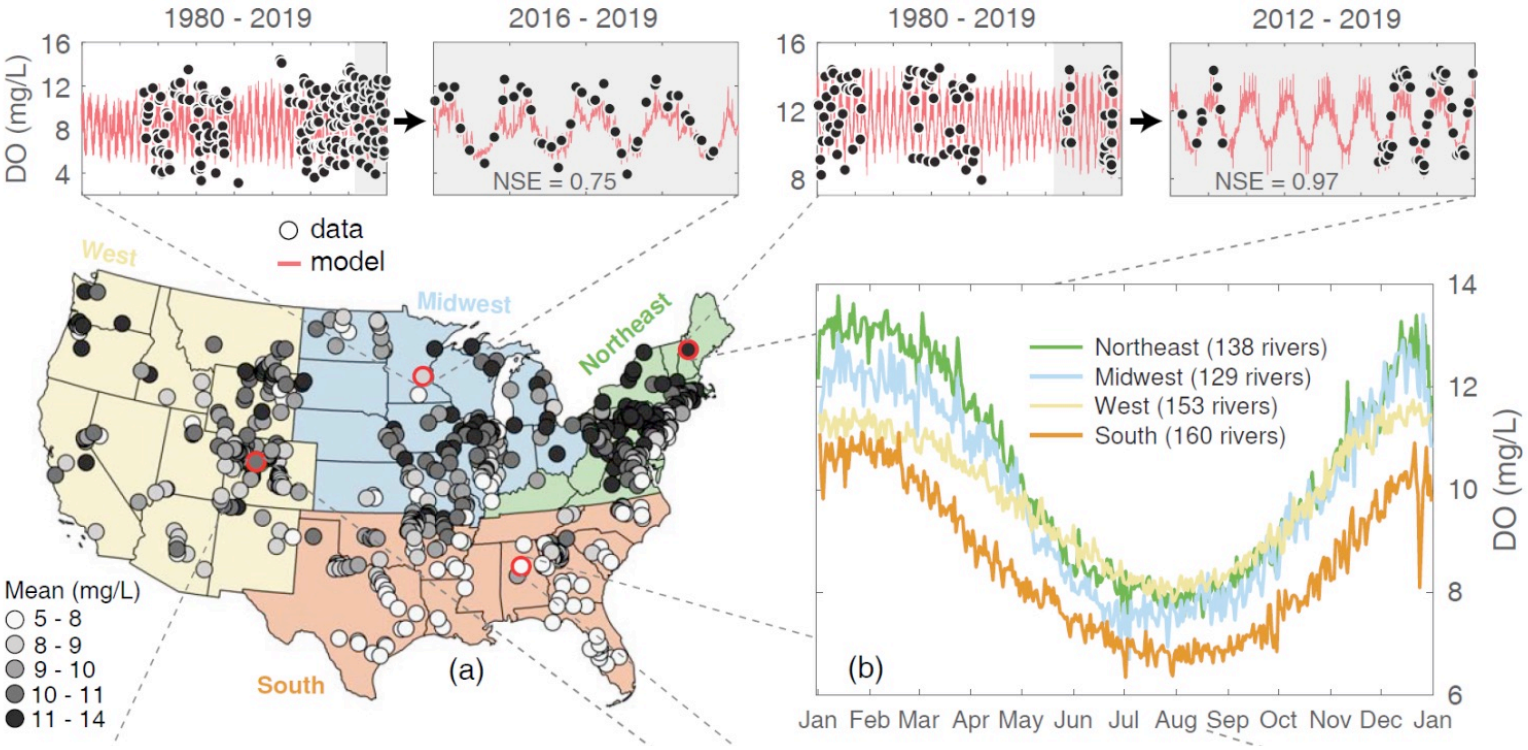
Article | Published: 09 March 2023

Temperature outweighs light and flow as the predominant driver of dissolved oxygen in US rivers

Wei Zhi, Wenyu Ouyang, Chaopeng Shen & Li Li

Nature Water **1**, 249–260 (2023) | Cite this article

Dissolved Oxygen



Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

A deep learning-based novel approach to generate continuous daily stream nitrate concentration for nitrate data-sparse watersheds

Gourab Kumer Saha [a], Farshid Rahmani [b], Chaopeng Shen [b], Li Li [b], Raj Cibin [a,b,*]

[a] Department of Agricultural and Biological Engineering, The Pennsylvania State University, United States of America
[b] Department of Civil and Environmental Engineering, The Pennsylvania State University, United States of America

Nitrate

ENVIRONMENTAL RESEARCH LETTERS

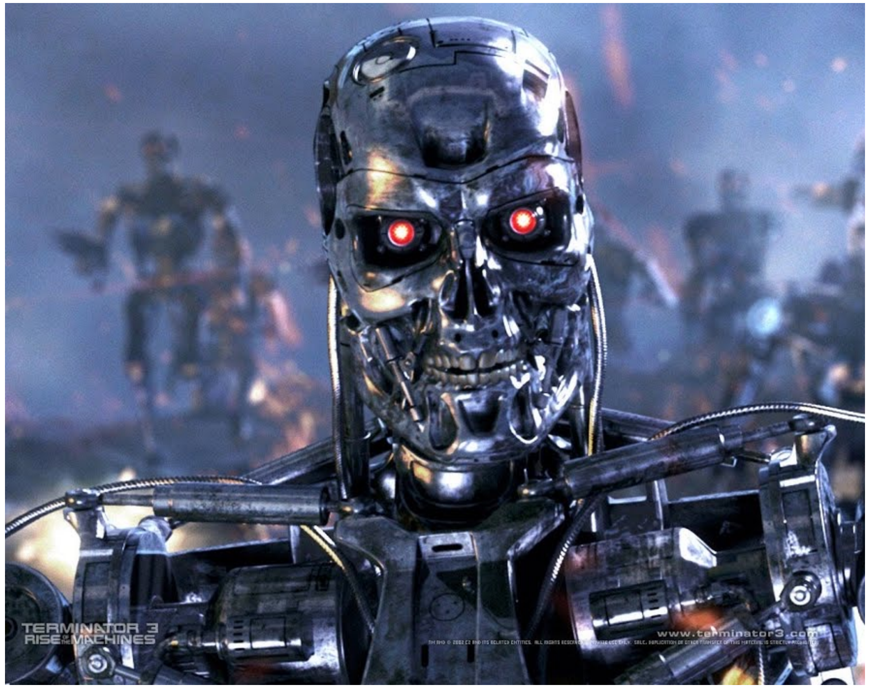
LETTER

Exploring the exceptional performance of a deep learning stream temperature model and the value of streamflow data

Farshid Rahmani[1], Kathryn Lawson[1], Wenyu Ouyang[2], Alison Appling[3], Samantha Oliver[4] and Chaopeng Shen[1]

[1] Civil and Environmental Engineering, Pennsylvania State University, University Park, State College, PA, United States of America
[2] School of Hydraulic Engineering, Dalian University of Technology, Dalian, People's Republic of China
[3] US Geological Survey, Reston, VA, United States of America
[4] US Geological Survey, Upper Midwest Water Science Center, Middleton, WI, United States of America

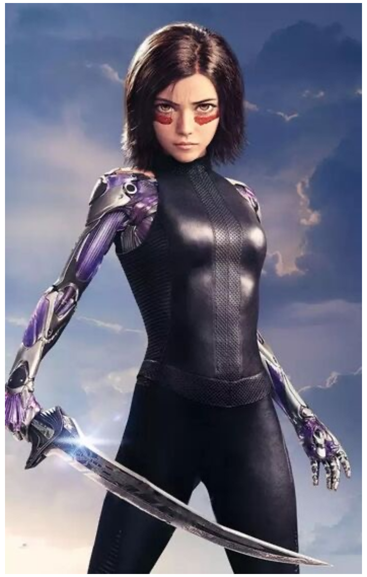
Water temperature



pure DL models

Human modelers

We need hybrids



@ChaopengShen

Phase 2: How to surpass the teacher (training data)

Training data often have limitations:
Resolution, accuracy, time interval, availability (unobserved variables), geographical imbalance, not enough extremes, not capturing nonstationarity…

How to overcome such limitations?

- Inclusion of physics
- Learning about physics.

Similarity & Differences between deep learning (DL) and process-based models (PBM)?

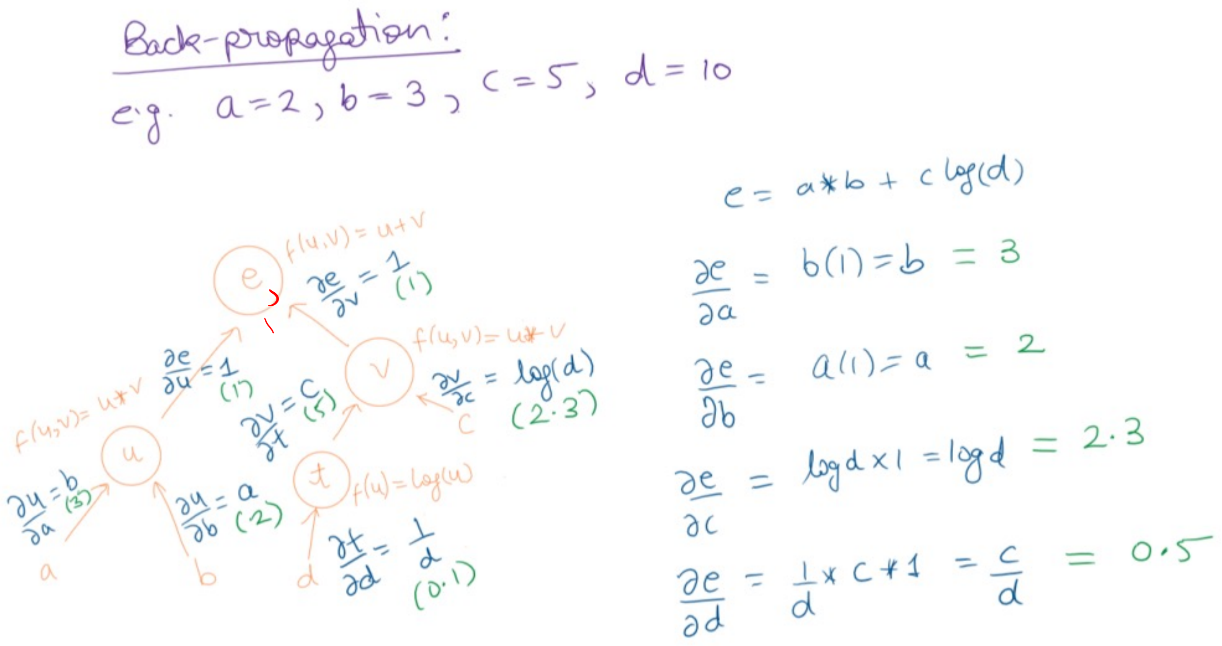
Purely data-driven NNs	Purely process-based models
Similarities	
$y = g^{W}(u, x, A)$ $W = argmin(L(y, y^*))$	$y = f^{\theta}(u, x, A)$ $\theta = argmin(L(y, y^*))$

The secret? Differentiable programming!

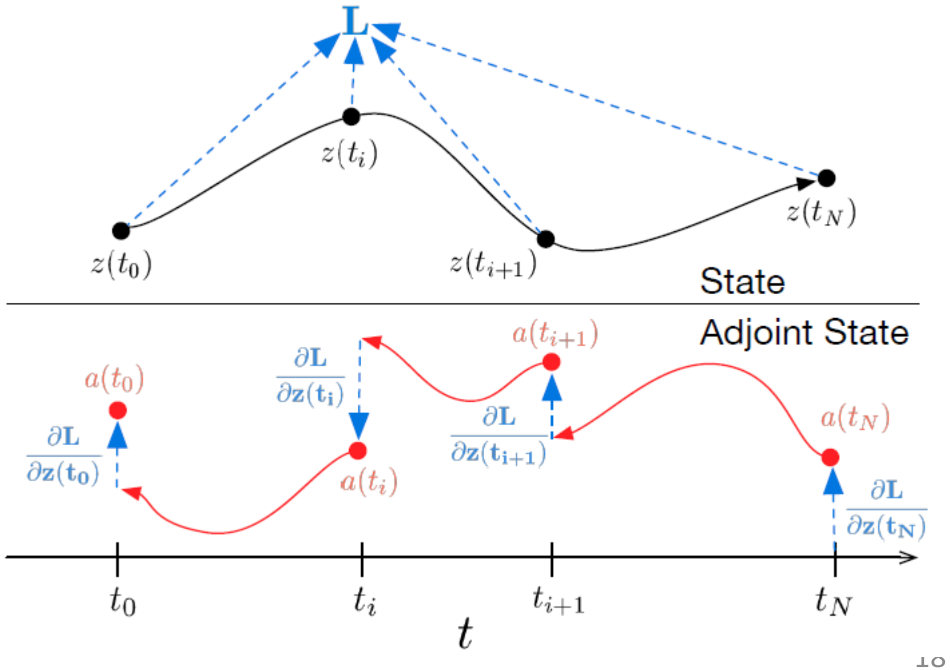
What does “Differentiable” mean?

- The ability to rapidly compute gradients $\frac{dL}{d\theta}$
- Enabling training by gradient descent

Automatic differentiation



Adjoint State method



Differentiable parameter learning

nature COMMUNICATIONS

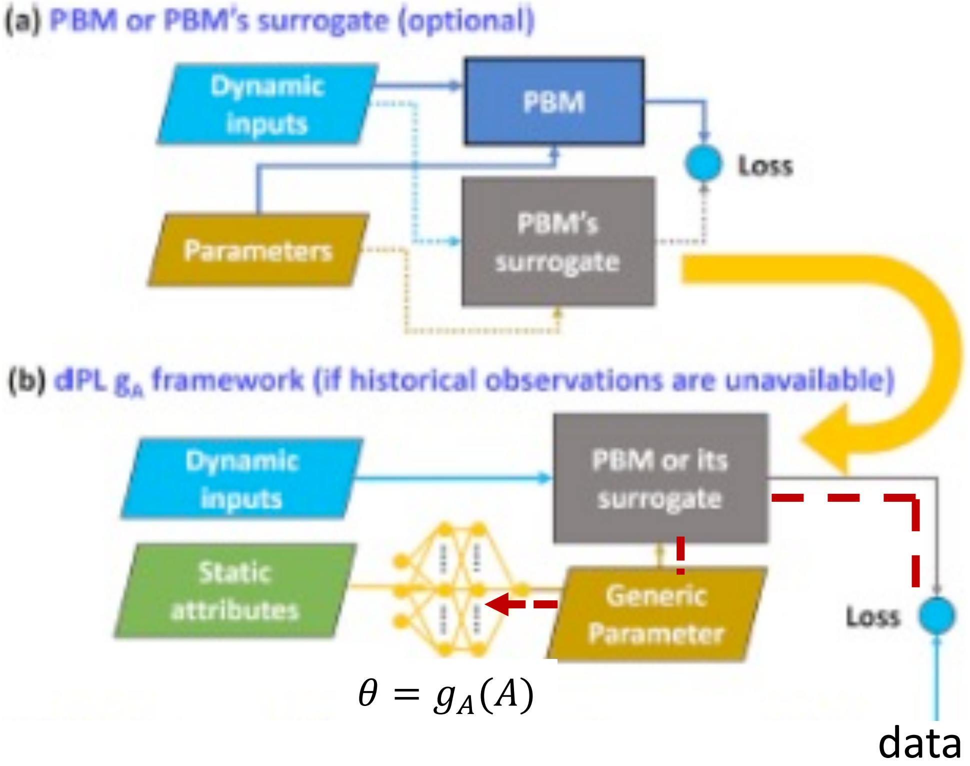
ARTICLE

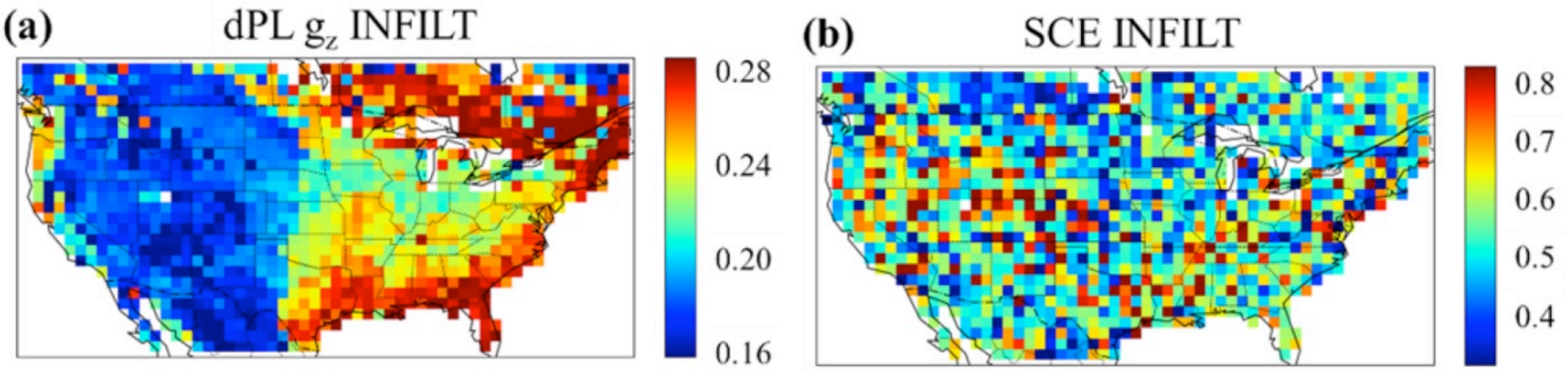
Check for updates

https://doi.org/10.1038/s41467-021-26107-z OPEN

From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling

Wen-Ping Tsai[1], Dapeng Feng[1], Ming Pan[2,3], Hylke Beck[4], Kathryn Lawson[1,5], Yuan Yang[6,7], Jiangtao Liu[1] & Chaopeng Shen[1,5]✉

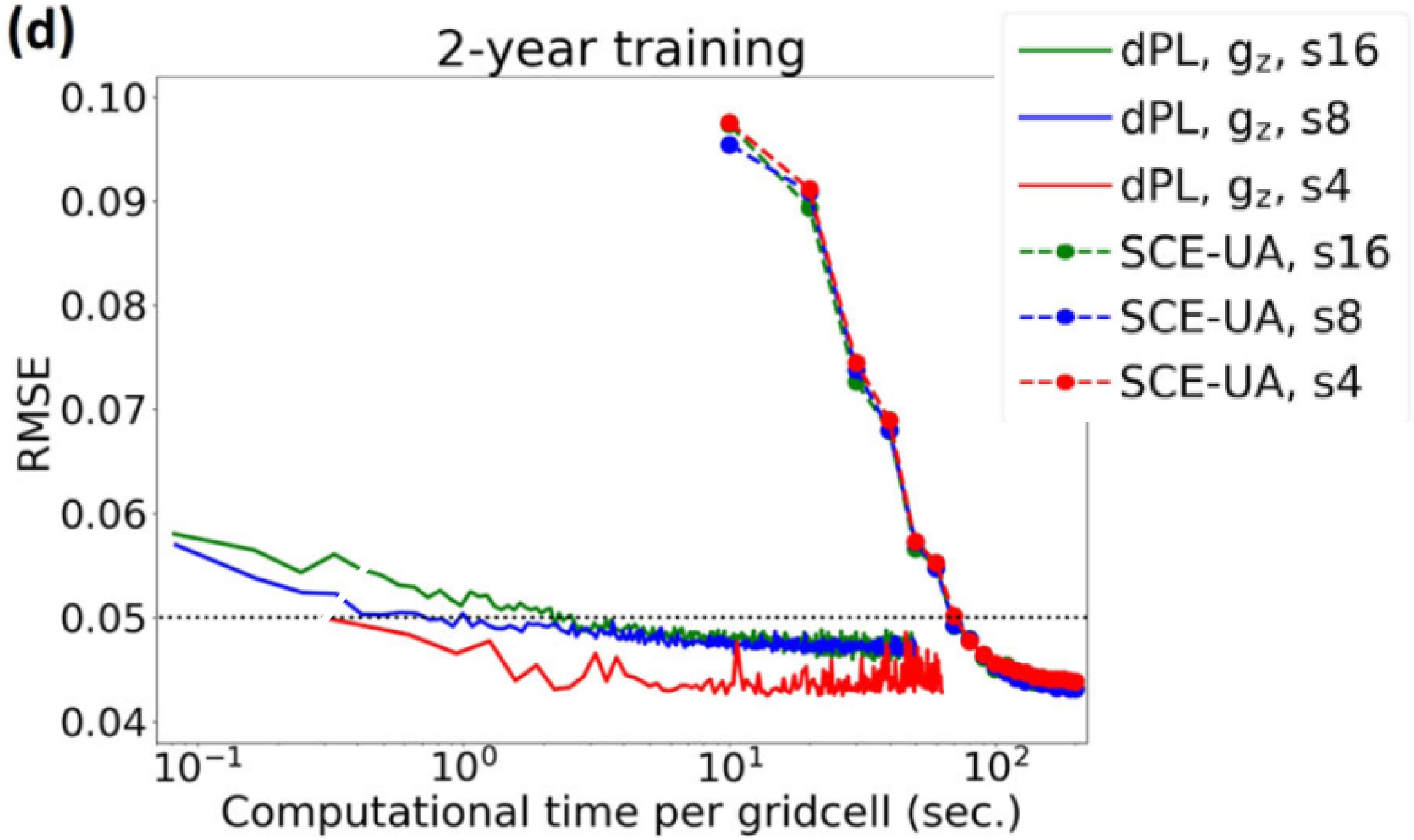




Point #1. Data scaling relationships (network effect?)

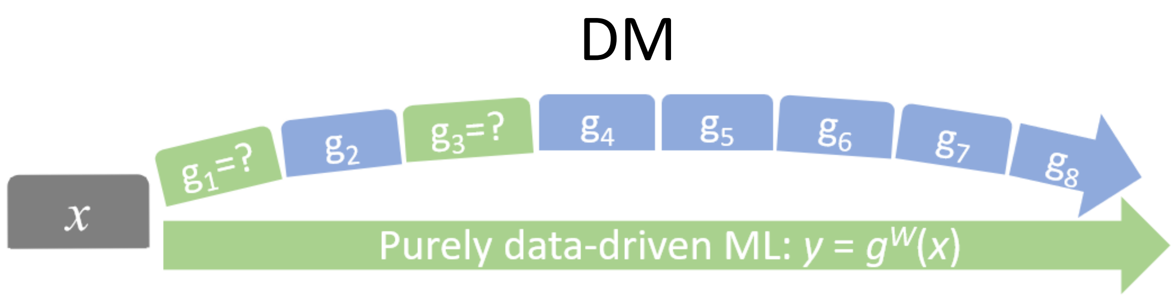
1. dPL = SCEUA for lowest RMSE
2. dPL scales better with more data
3. Orders of magnitude more efficient
4. (not shown) better results for **untrained** variables and better **spatial generalization** than traditional approach!

Relies on differentiable programming!



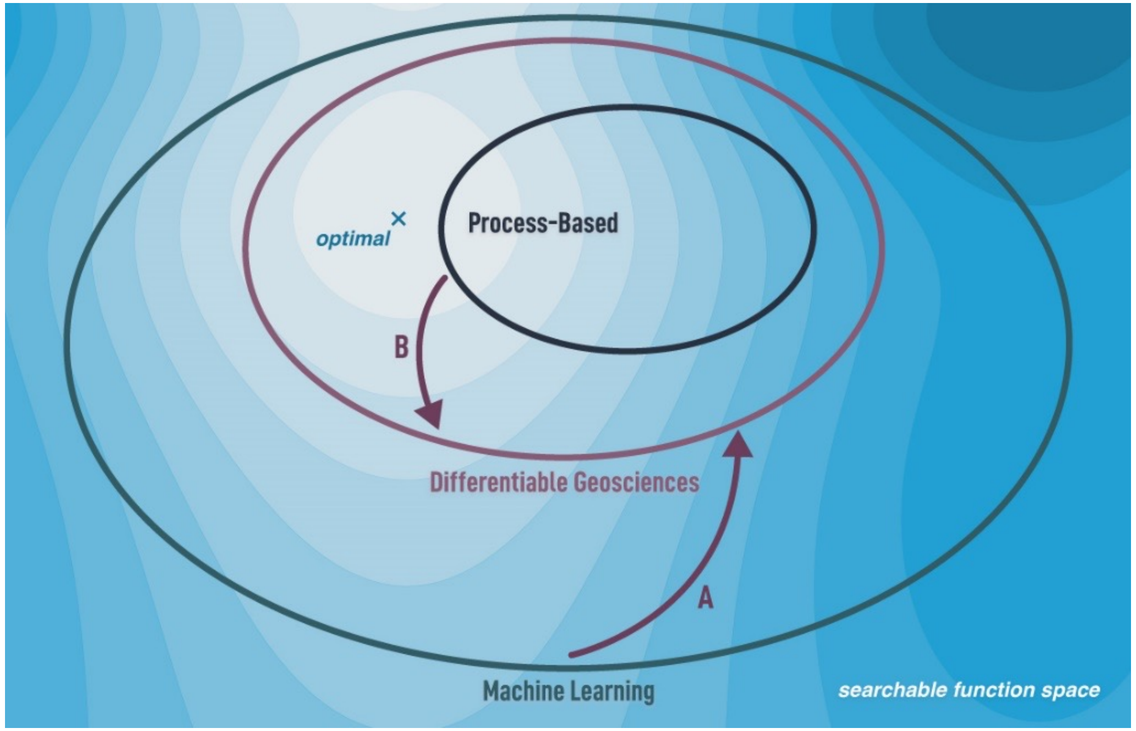
Tsai et al. 2021, Nature Communications

What is Differentiable Modeling (DM) in Geosciences?



- NNs mixed w/ process-based equations (priors)
- The priors constrain the learning to an interpretable scope.
- intermediate physical variables.
- Update our knowledge and learn unrecognized relationships from data.

2 perspectives



Differentiable, learnable models to learn functions

Hydrol. Earth Syst. Sci., 27, 2357–2373, 2023
https://doi.org/10.5194/hess-27-2357-2023

Hydrology and Earth System Sciences

The suitability of differentiable, physics-informed machine learning hydrologic models for ungauged regions and climate change impact assessment

Dapeng Feng[1], Hylke Beck[2], Kathryn Lawson[1], and Chaopeng Shen[1]
[1]Civil and Environmental Engineering, The Pennsylvania State University, University Park, PA, USA
[2]Physical Science and Engineering, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia
Correspondence: Chaopeng Shen (cshen@engr.psu.edu)

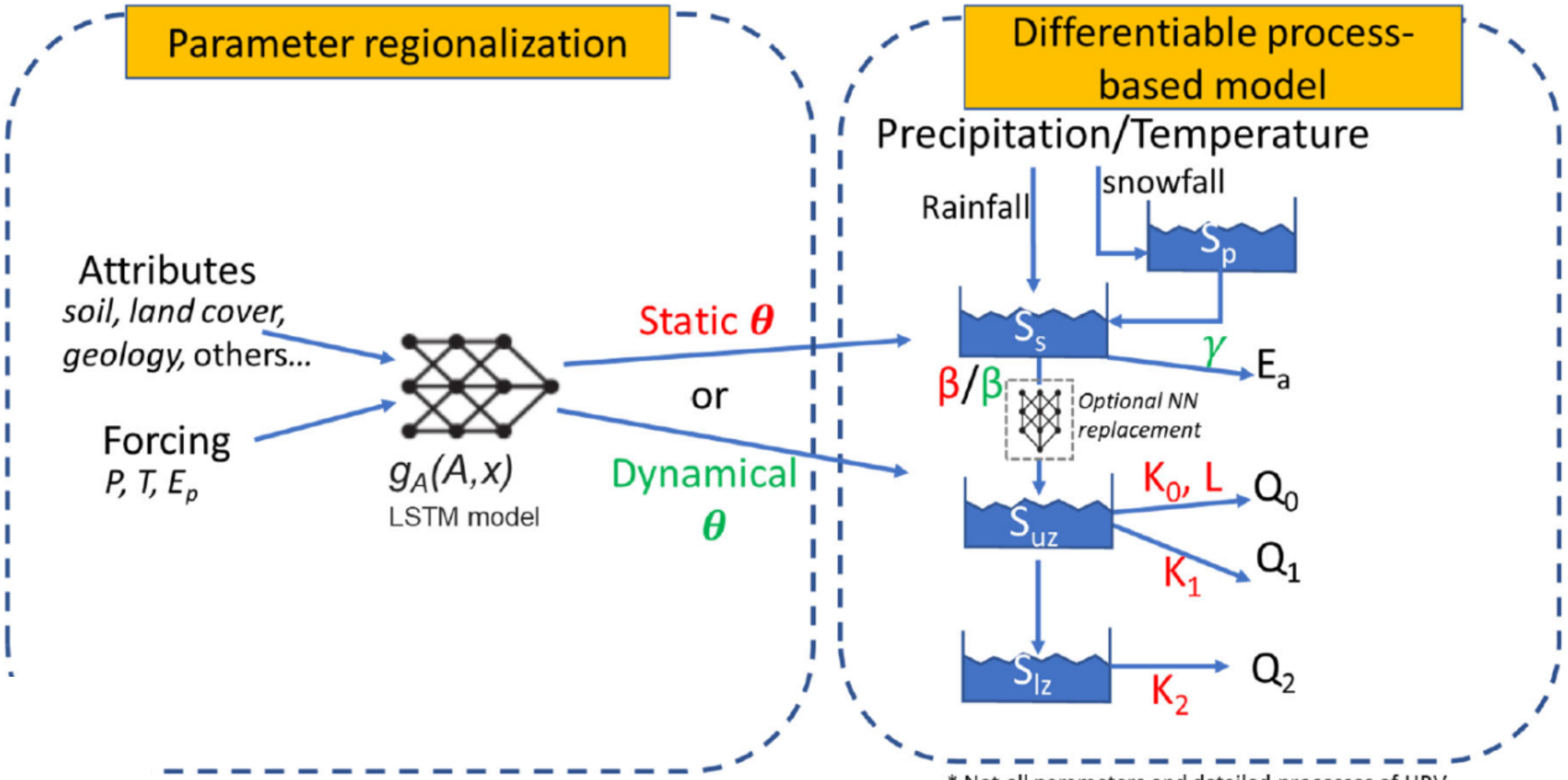
Water Resources Research

Research Article | Full Access

Differentiable, learnable, regionalized process-based models with multiphysical outputs can approach state-of-the-art hydrologic prediction accuracy

Dapeng Feng, Jiangtao Liu, Kathryn Lawson, Chaopeng Shen

First published: 19 September 2022 | https://doi.org/10.1029/2022WR032404



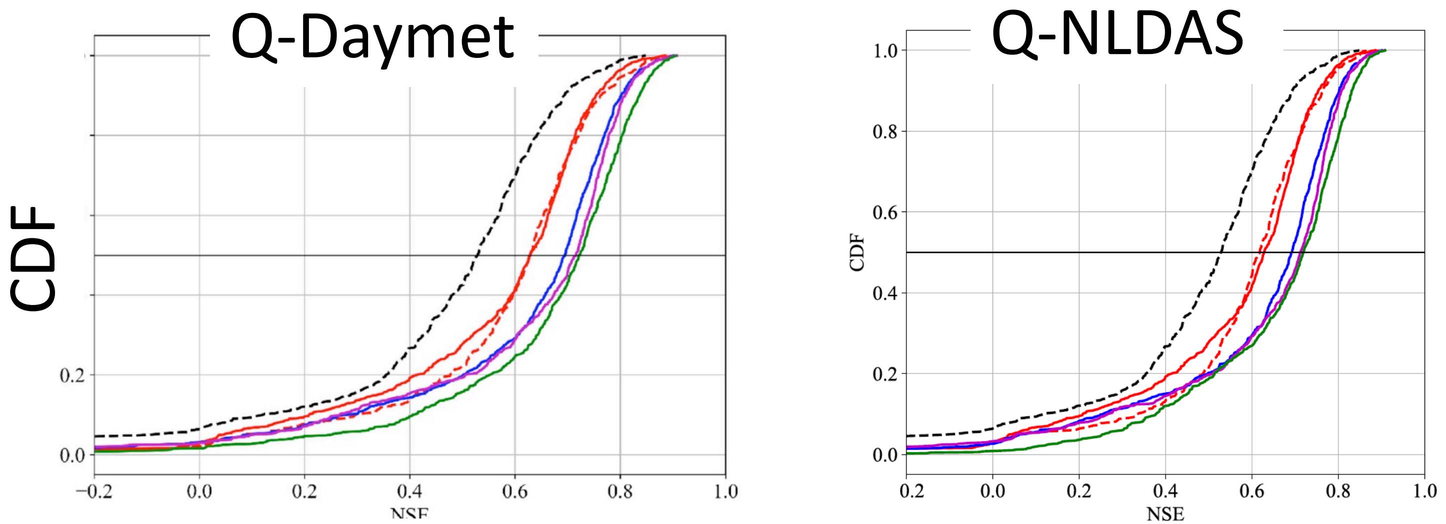
* Not all parameters and detailed processes of HBV sketched here for the sake of simplicity.

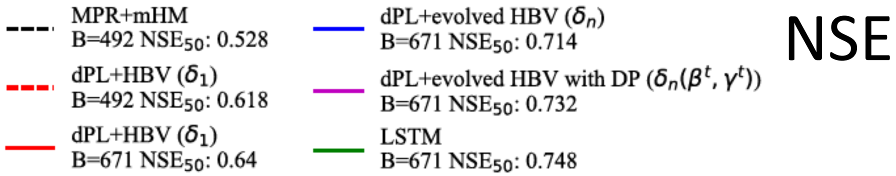
Rewritten in PyTorch

Evolve model structure

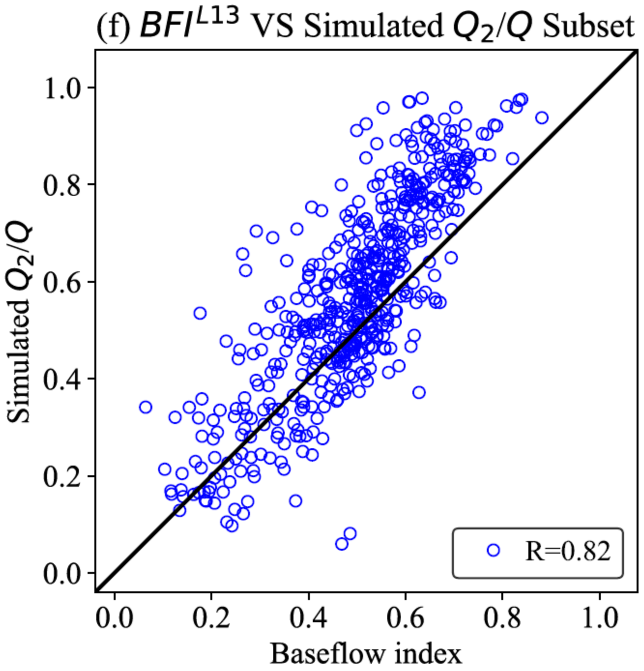
Approaching LSTM!
But….

- Output untrained variables.
- Multivariate constraints.
- It can help us answer questions!

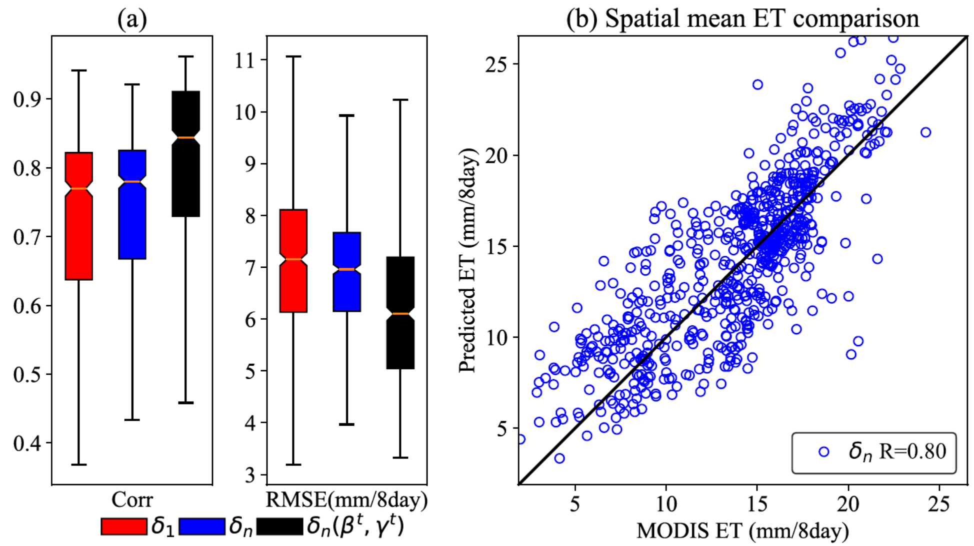




Baseflow



Evapotranspiration



Caveat: not using the ensemble
-- first iteration. Priors do matter.

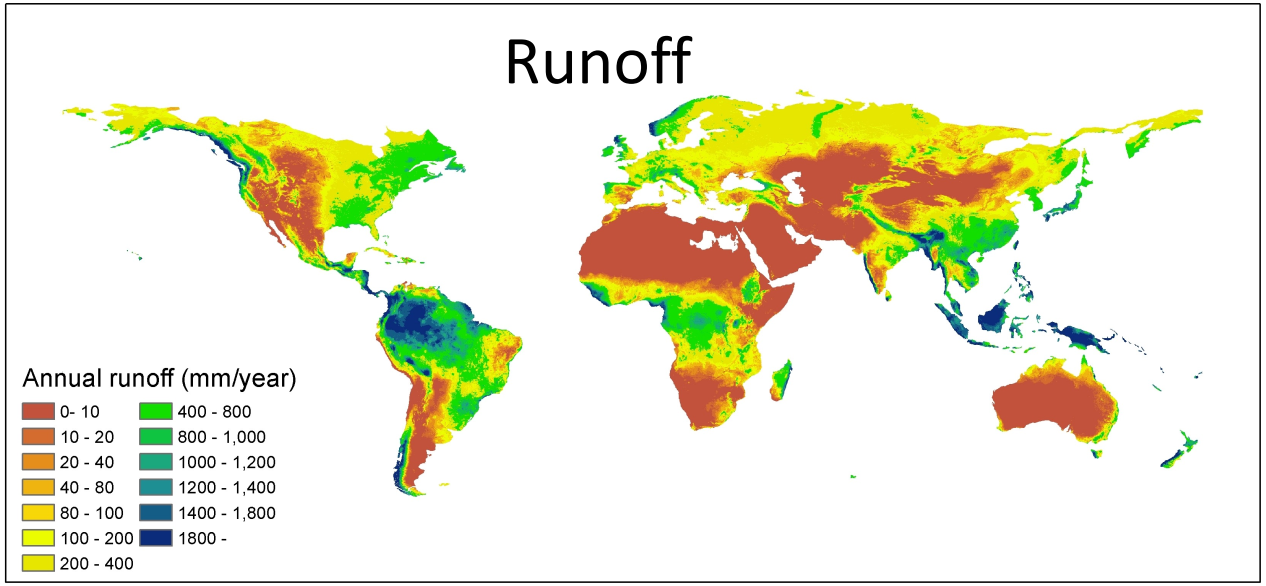

PBM

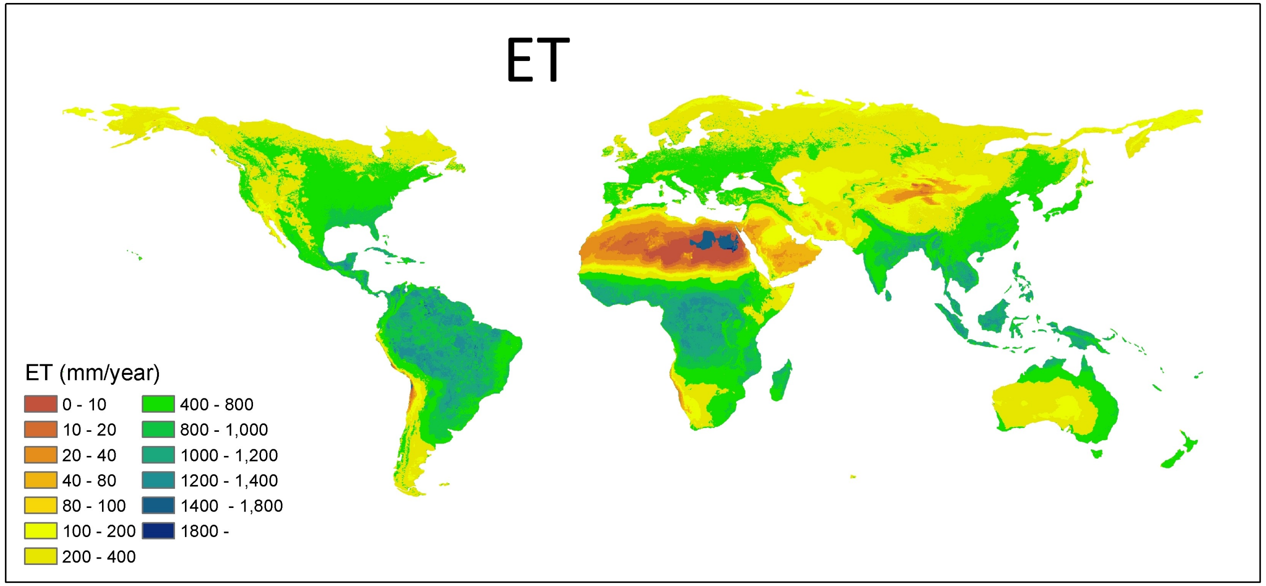

PBM+dPL

What can DM bring to global hydrology?

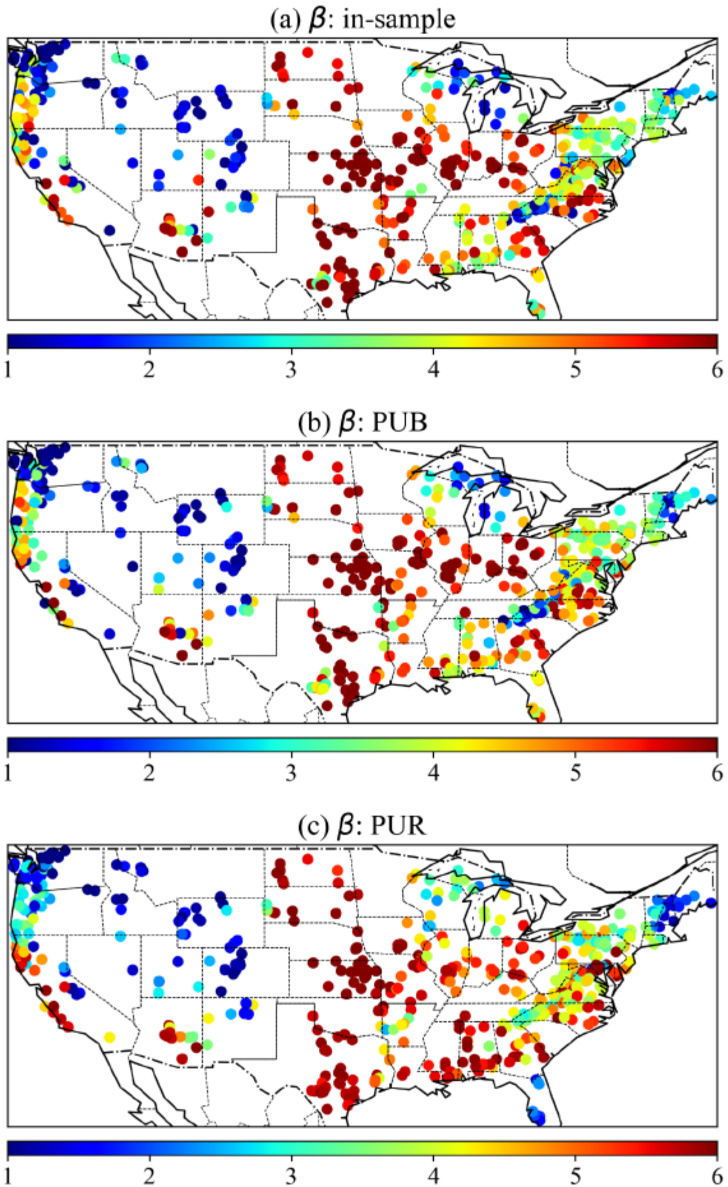
- Spatial extrapolation in data-sparse regions
- Extremes
- Learn robust unknown functions
- Human dynamics or unknown physcs
- Correct forcings

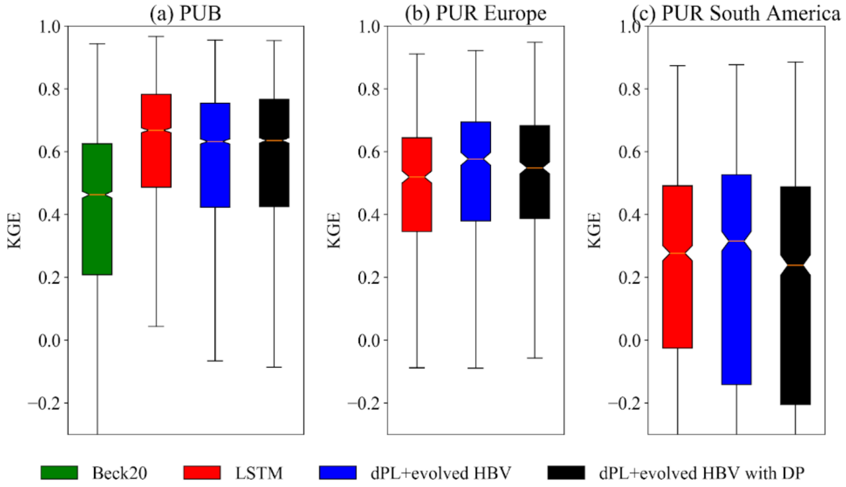
Produced by differentiable models

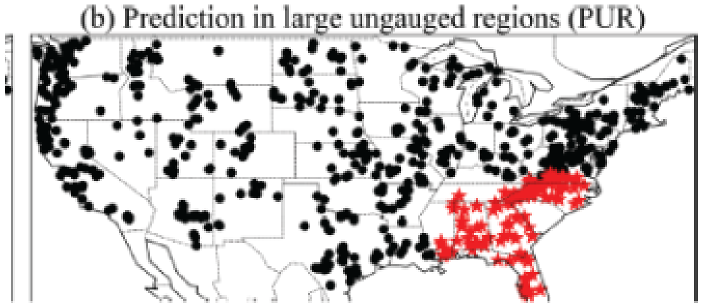




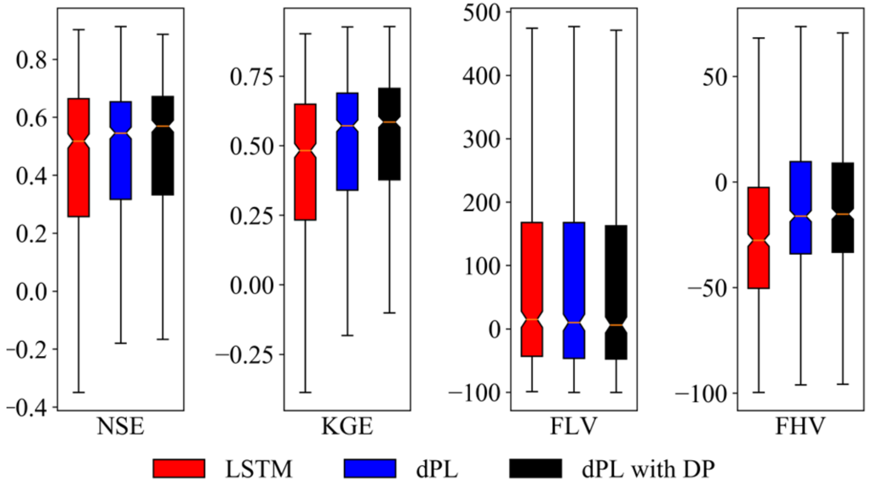
Differentiable models extrapolation better

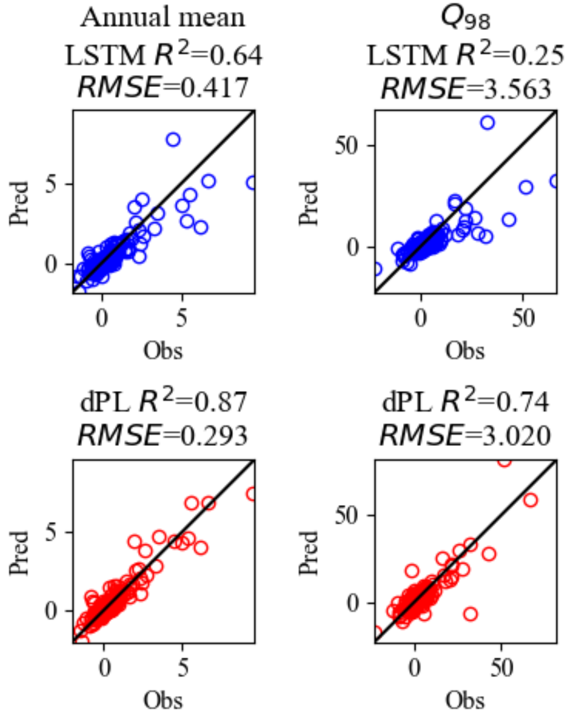






For PUR





New model

Water Resources Research

Research Article | Open Access

Differentiable, Learnable, Regionalized Process-Based Models With Multiphysical Outputs can Approach State-Of-The-Art Hydrologic Prediction Accuracy

Dapeng Feng, Jiangtao Liu, Kathryn Lawson, Chaopeng Shen

First published: 19 September 2022 | https://doi.org/10.1029/2022WR032404 | Citations: 18

https://doi.org/10.5194/hess-27-2357-2023

Article | Assets | Peer review | Metrics | Related articles

Research article | 30 Jun 2023

The suitability of differentiable, physics-informed machine learning hydrologic models for ungauged regions and climate change impact assessment

Dapeng Feng, Hylke Beck, Kathryn Lawson, and Chaopeng Shen

https://doi.org/10.5194/gmd-2023-190

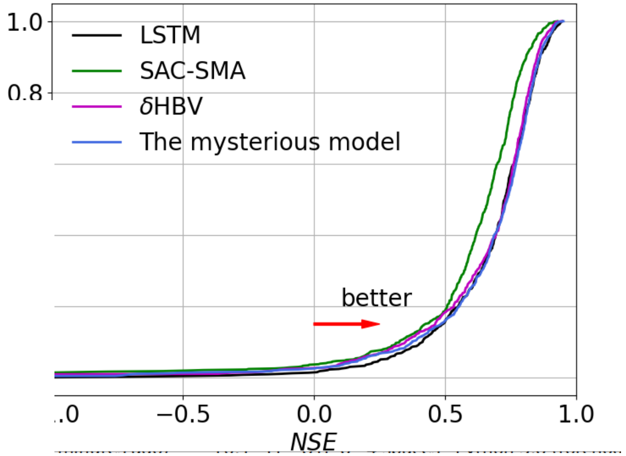
Abstract | Discussion | Metrics

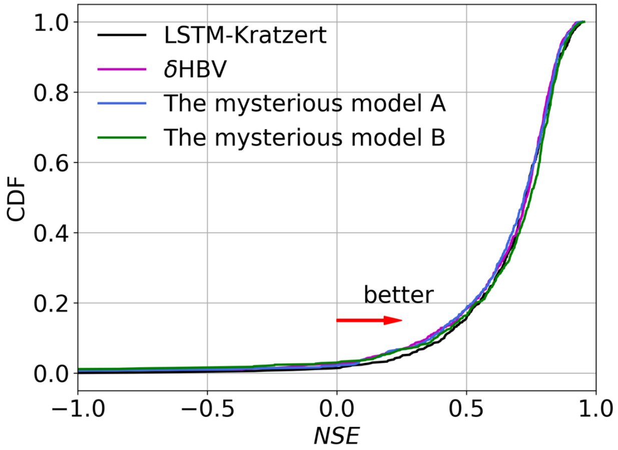
Submitted as: model evaluation paper | 05 Oct 2023

Status: this preprint is currently under review for the journal GMD.

Deep Dive into Global Hydrologic Simulations: Harnessing the Power of Deep Learning and Physics-informed Differentiable Models (δHBV-globe1.0-hydroDL)

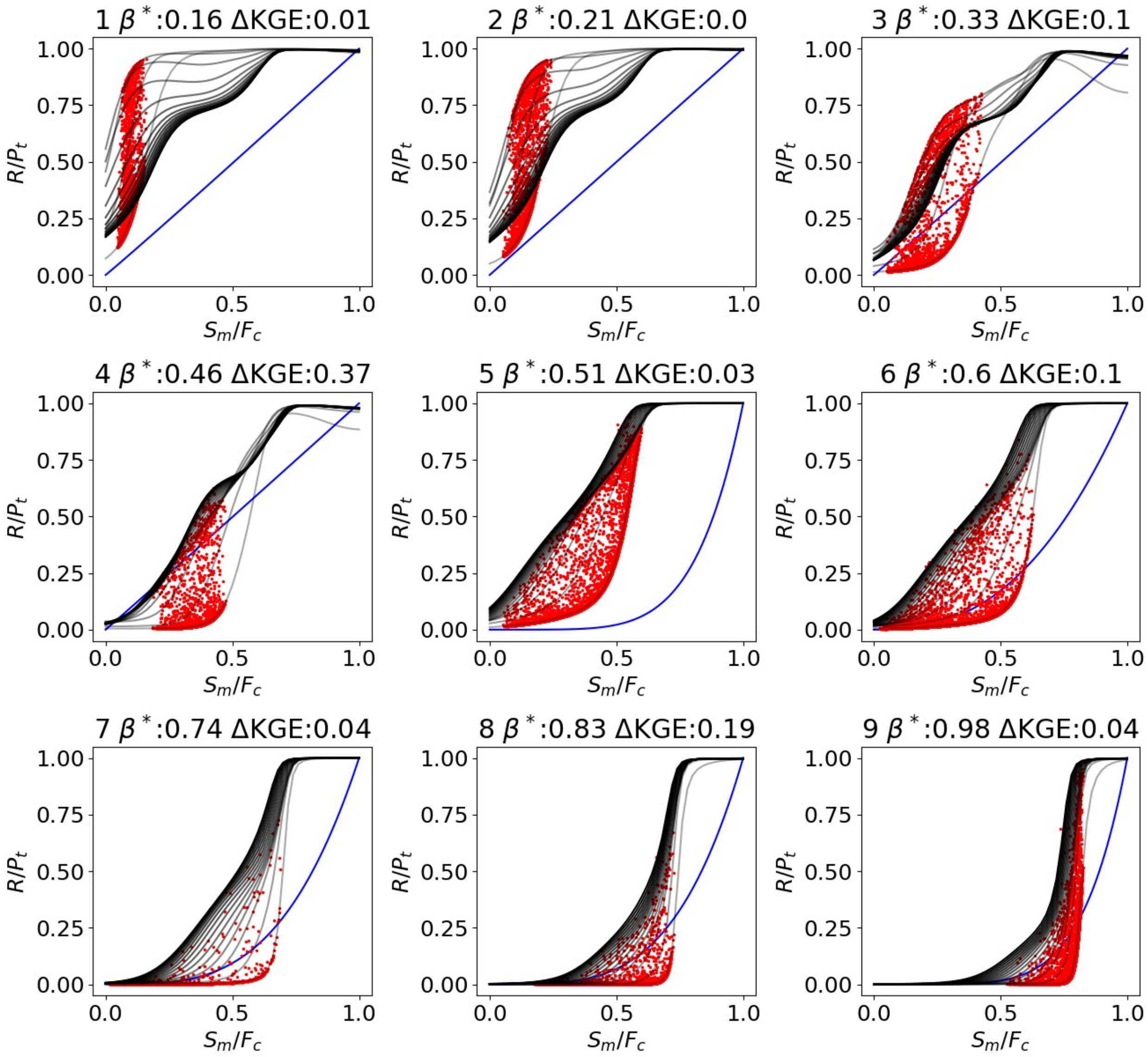
Dapeng Feng, Hylke Beck, Jens de Bruijn, Reetik Kumar Sahu, Yusuke Satoh, Yoshihide Wada, Jiangtao Liu, Ming Pan, Kathryn Lawson, and Chaopeng Shen





Model	Median NSE	Median KGE	Median absolute (non-absolute) FLV (%)	Median absolute (non-absolute) FHV (%)	Median low flow RMSE (mm/day)	Median peak flow RMSE (mm/day)	Baseflow index spatial correlation	Median NSE of temporal ET simulation
LSTM	**0.73**	**0.77**	40.59 (29.70)	13.46 (-4.19)	0.055	2.56	-	-
SAC-SMA	0.66	0.73	59.40 (46.96)	17.55 (-9.79)	0.081	3.19	-	-
HBV	**0.73**	0.73	56.53 (50.93)	15.29 (-8.89)	0.074	2.56	0.76	0.59
The mysterious model	0.72	0.75	43.29 (37.61)	**13.25 (-4.33)**	**0.048**	**2.47**	**0.83**	**0.61**

Learning unknown relationships from data (in preparation)



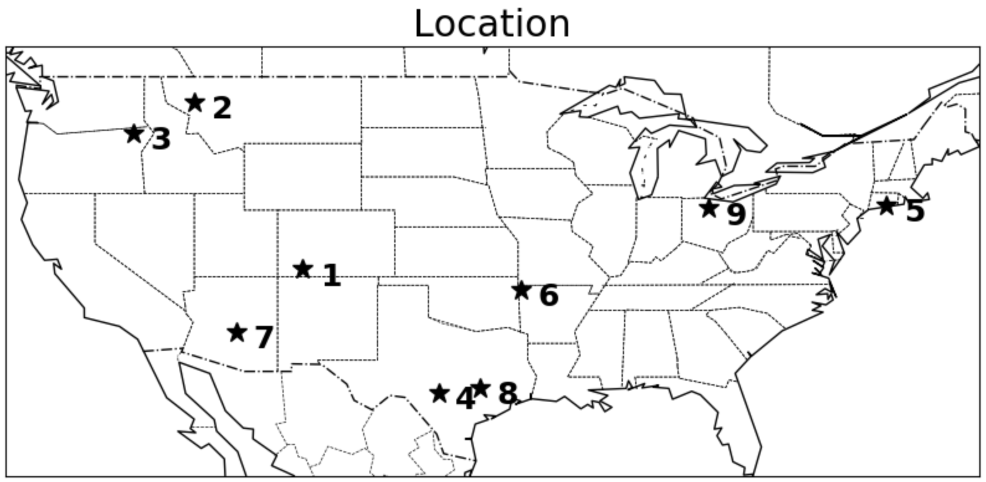
$$R/P_t = (S_m/F_c)^{\beta}$$

$$\boxed{R/P_t = ANN(\beta^*, F_c, S_m, S_m/F_c, P_t)}$$

Blue line: original power law relation

Red dots: ANN simulations

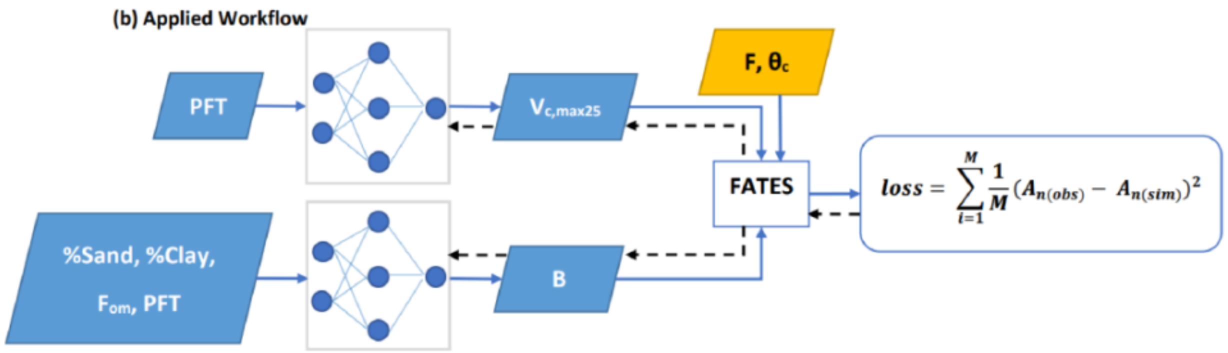
Black lines: continuous plotting of ANN functions



How is differentiable modeling different from physics-guided ML?

	Physics-guided ML	Differentiable modeling
Goal	Use physics to constrain ML → Improve ML generalization	Use ML to learn unknown relationships and improve simulation quality → Advance our process understanding
Approach	May not be differentiable; Various approaches like modifying the loss function	Differentiable; end-to-end training
Philosophy	Physical laws is treated as ground truth	Constantly seeking to improve our equations

Example 4. Ecosystem modeling: photosynthesis



Biogeosciences, 20, 2671–2692, 2023
https://doi.org/10.5194/bg-20-2671-2023

Biogeosciences

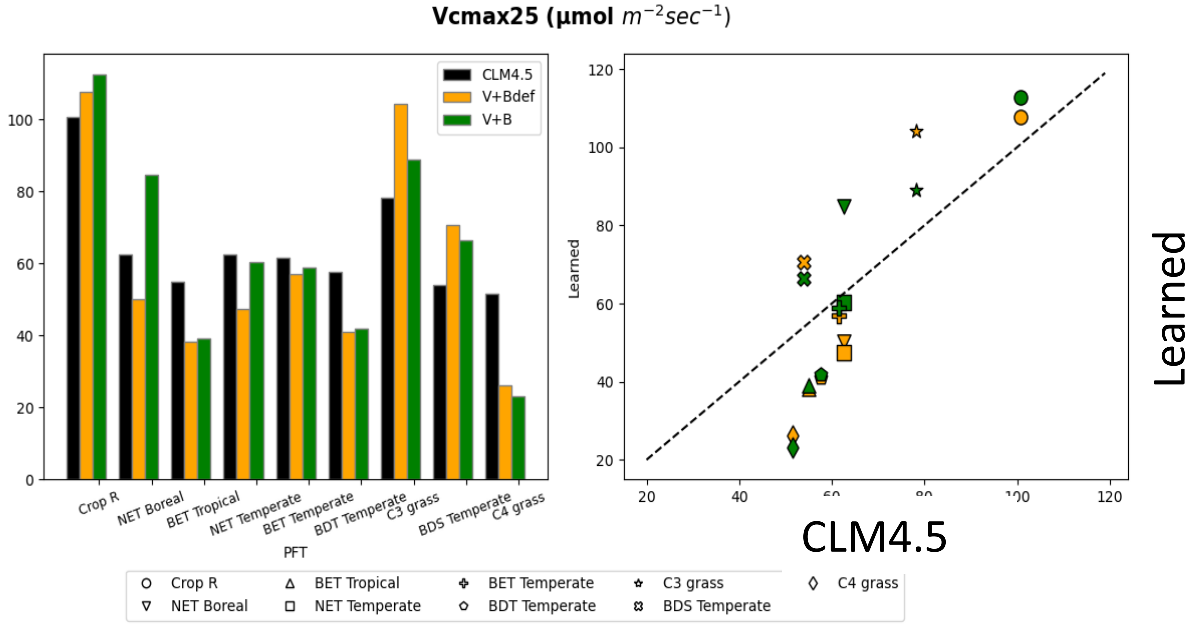
A differentiable, physics-informed ecosystem modeling and learning framework for large-scale inverse problems: demonstration with photosynthesis simulations

Doaa Aboelyazeed[1], Chonggang Xu[2], Forrest M. Hoffman[3,4], Jiangtao Liu[1], Alex W. Jones[5], Chris Rackauckas[6], Kathryn Lawson[1], and Chaopeng Shen[1]

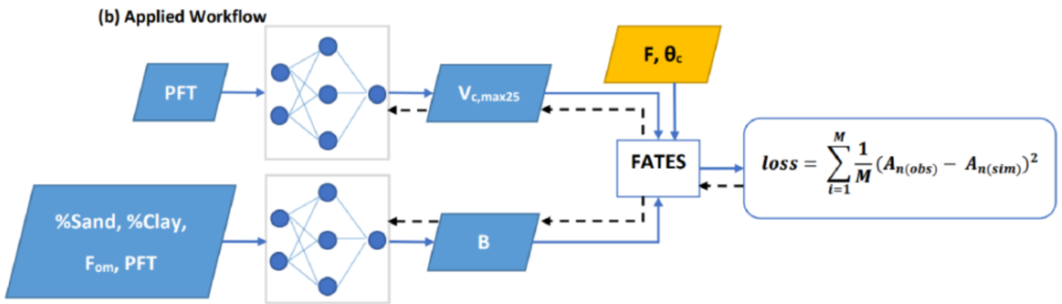
(a) Temporal holdout test for the following system

Runs	Corr		RMSE (μmol m^{-2} s^{-1})		Bias (μmol m^{-2} s^{-1})		NSE	
	Train	Test	Train	Test	Train	Test	Train	Test
V_{def}+B_{def}	0.565		6.780		1.476		0.041	
V_{def}+B_{def} **	0.592		5.488		1.034		0.318	
V_{def}+B	0.678	0.547	5.887	6.730	1.353	1.754	0.321	-0.084
V+B_{def}	0.769	0.593	4.595	5.677	-0.129	-1.368	0.587	0.229
V+B	0.800	0.748	4.299	4.421	0.037	0.347	0.638	0.532
V+B **	0.774	0.768	4.269	4.198	0.056	0.092	0.597	0.581

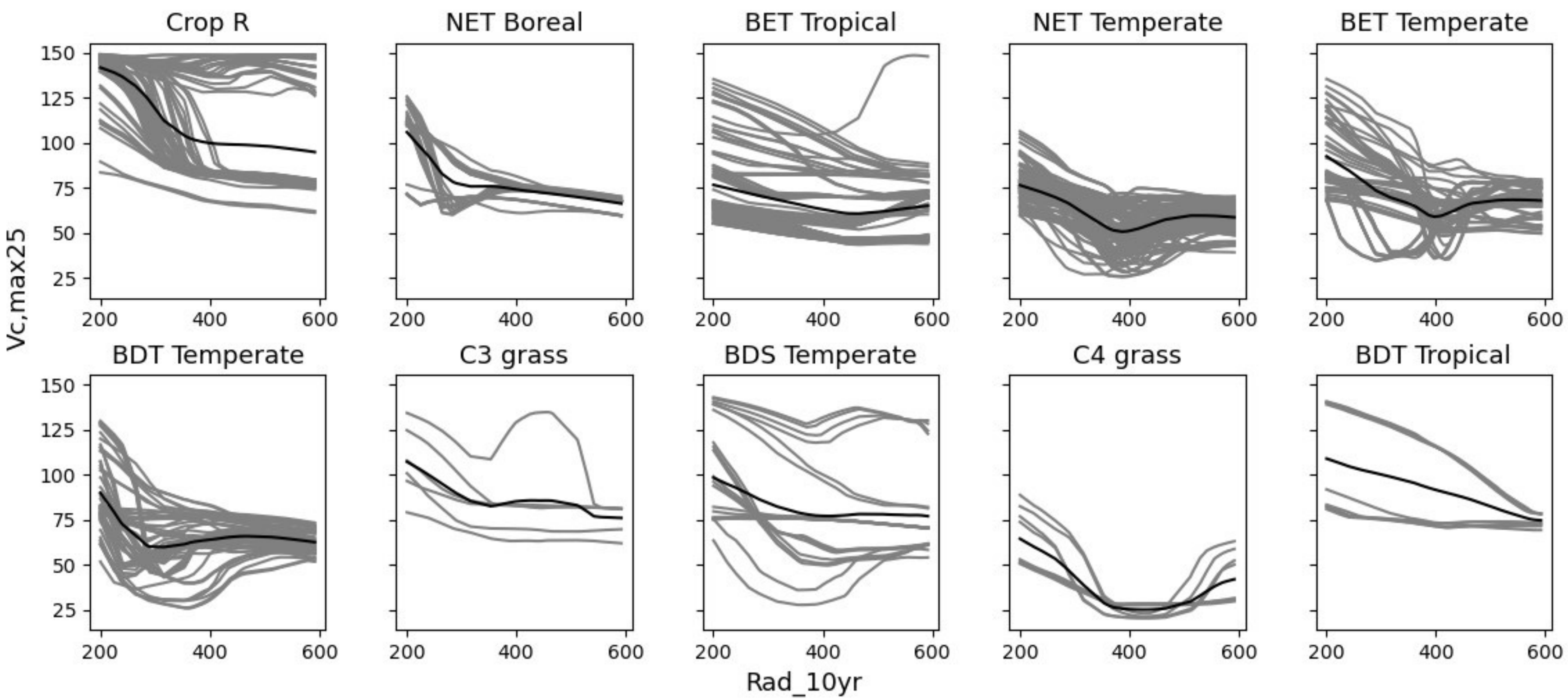
*** refers to using C3_only plants in dataset*



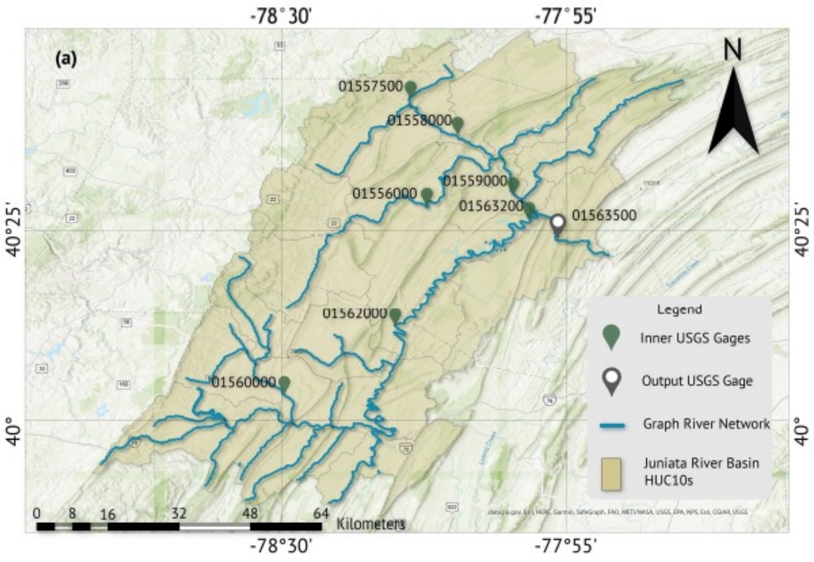
Example 4. *Ecosystem modeling: photosynthesis*

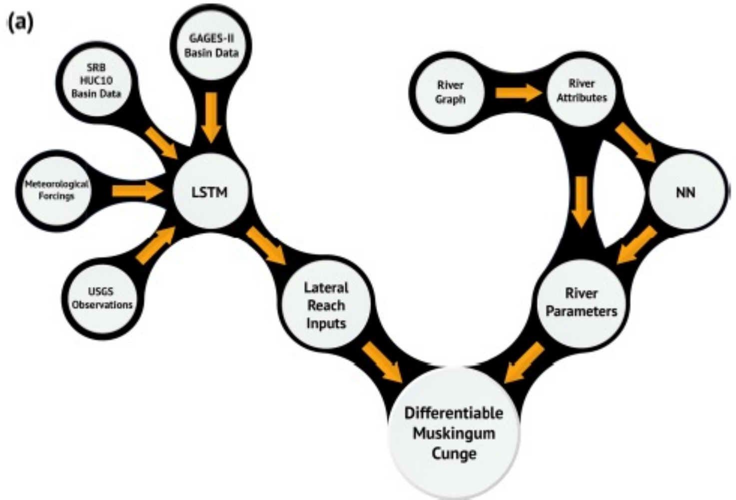


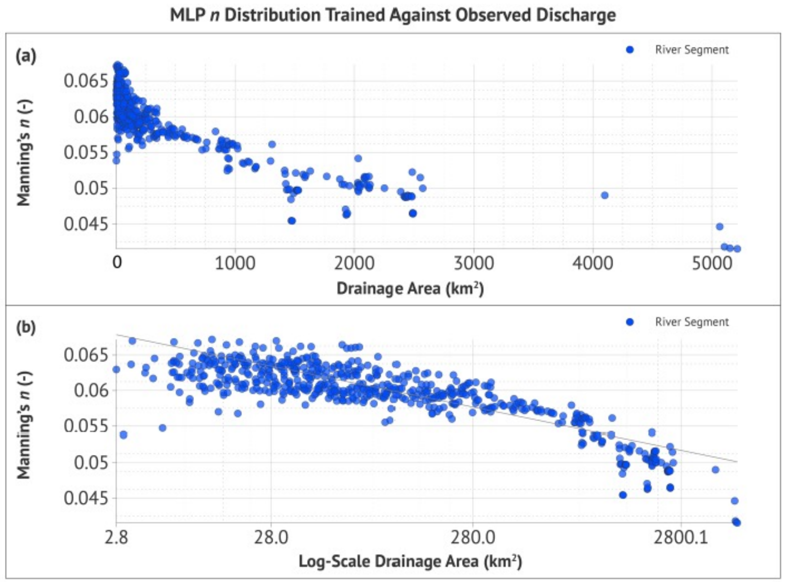
Discovering environmental dependencies of previously PFT-dependent parameter

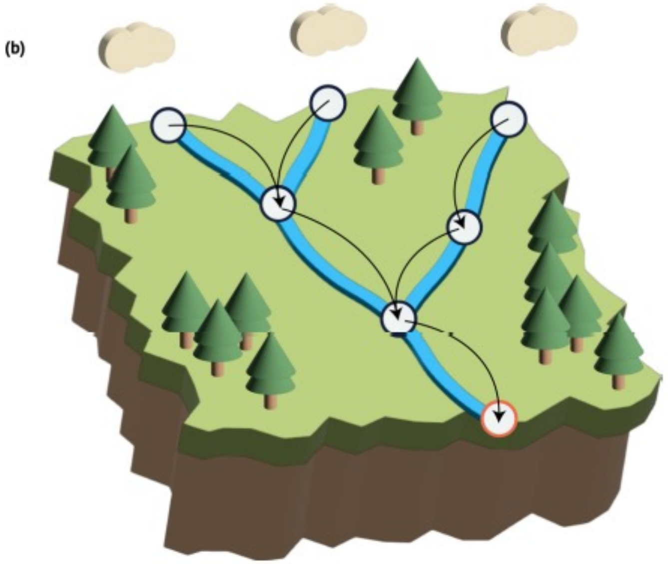


Example 4. Differentiable routing model



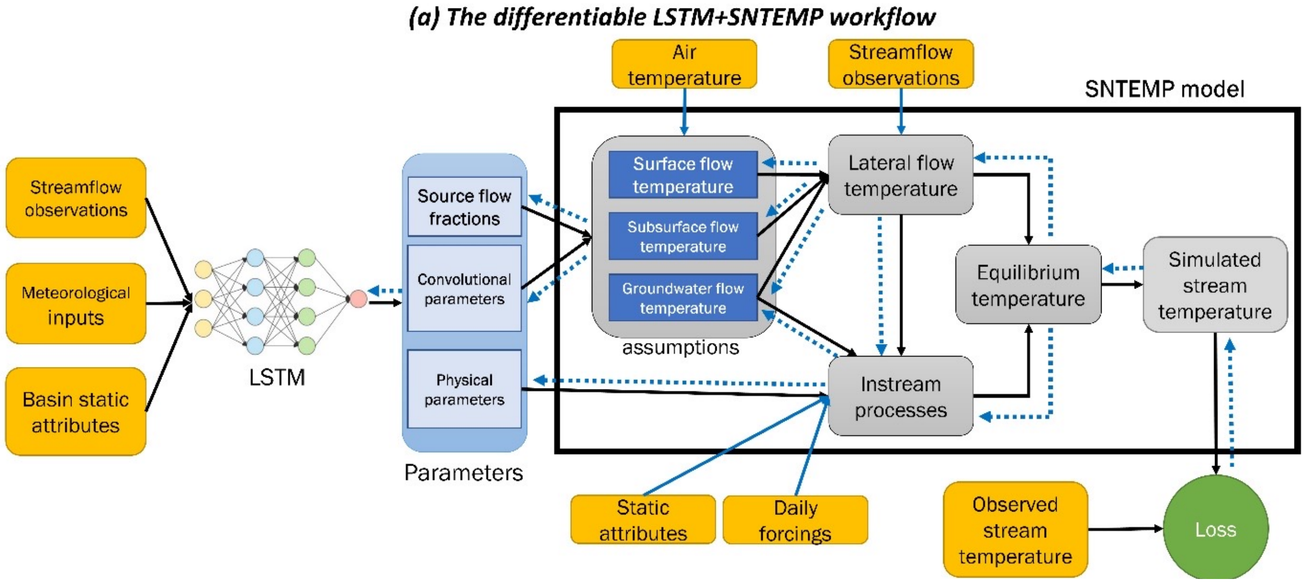


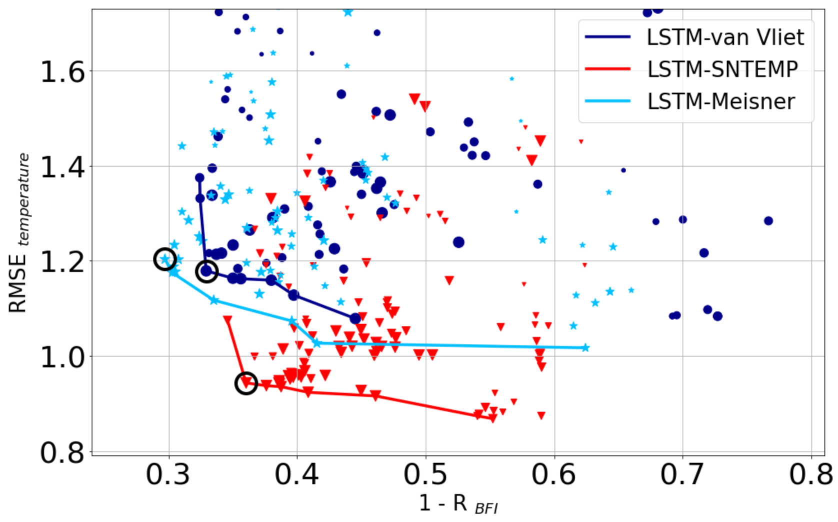


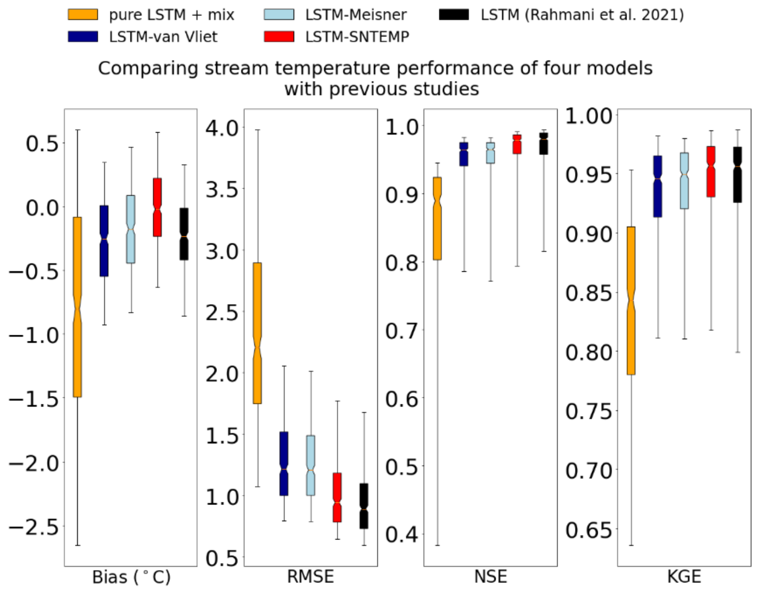


Example 5. *Water temperature modeling*

Prior assumptions matter!

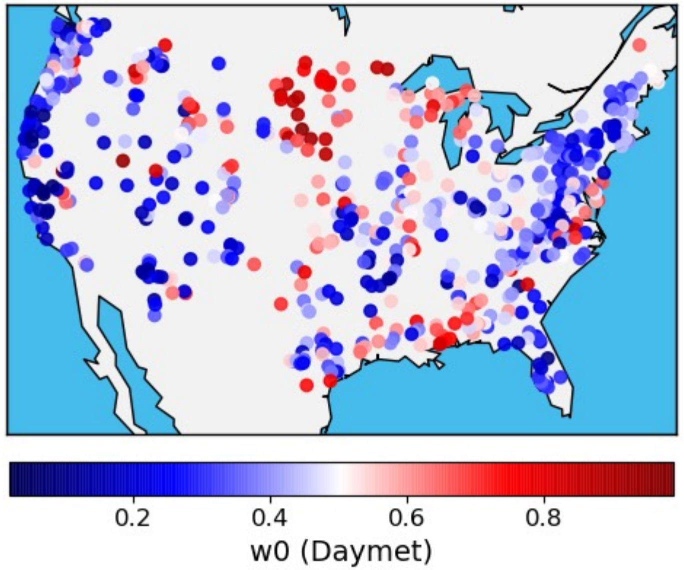


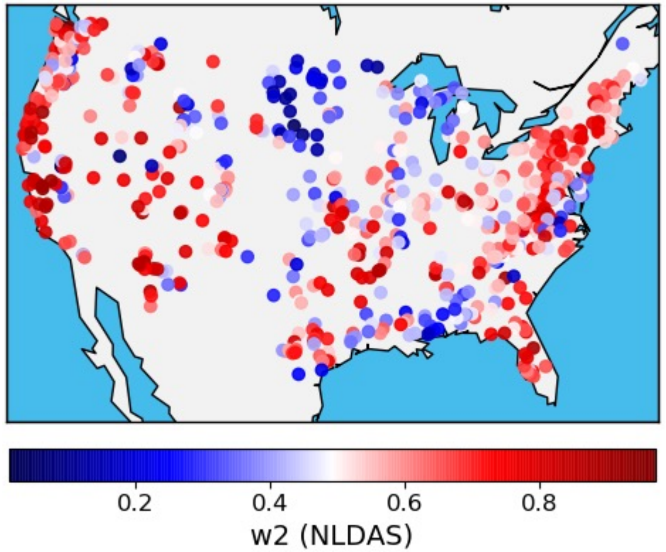


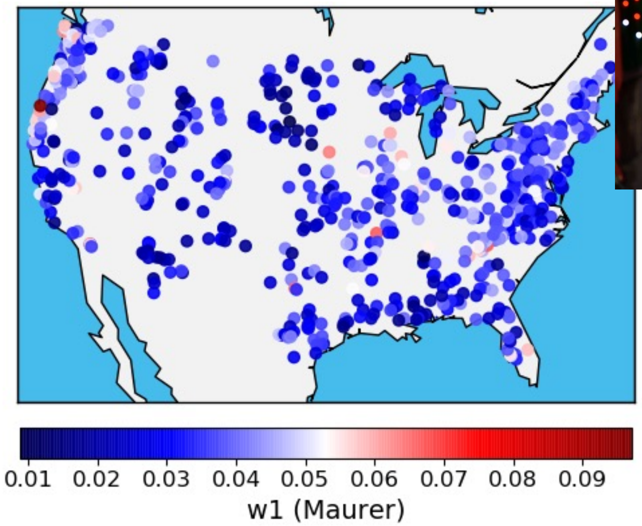


Example 6. Fusion of forcings (in preparation)

NLDAS (0.56) > Daymet (0.41) > Maurer (0.03)

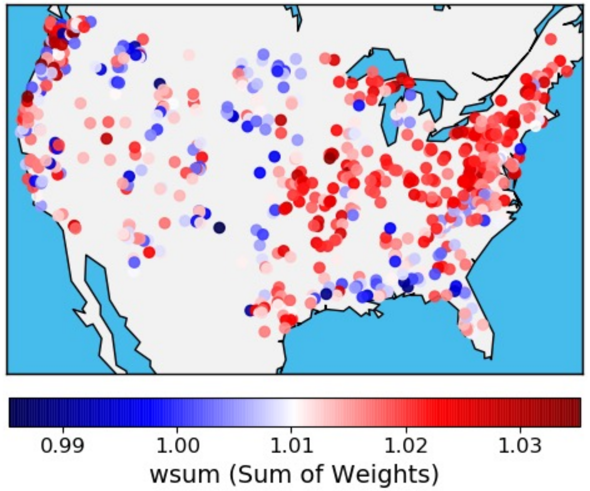






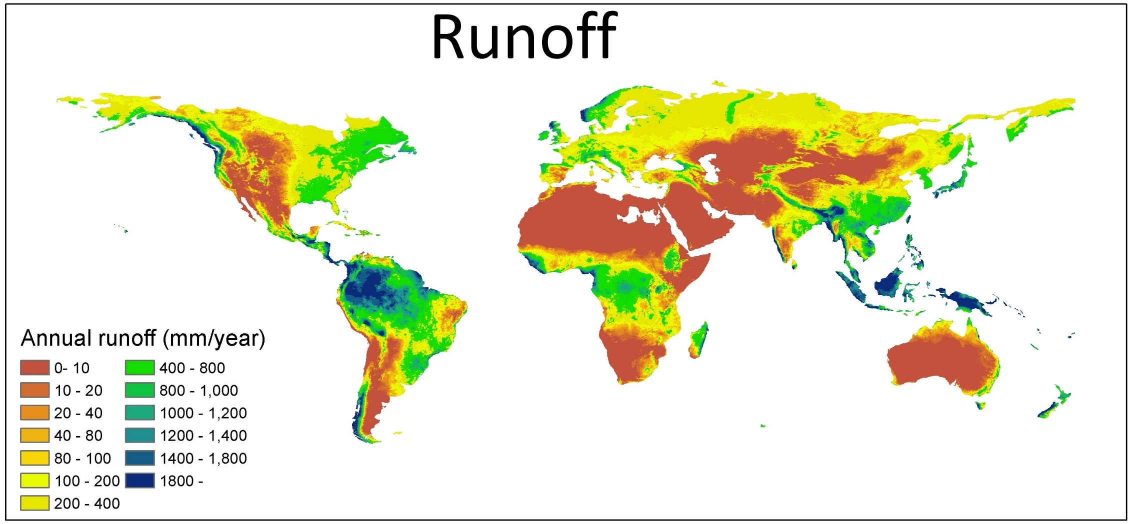
Simulation	Forcings	Median NSE	Median KGE	Low flow RMSE (mm/day)	High Flow RMSE (mm/day)
Single forcing w/o bias corr	Daymet	0.737	0.728	0.134	3.990
Multiforcing with bias correction	Daymet, Maurer, NLDAS	**0.770**	**0.780**	**0.082**	**3.414**

Low bias



Future

- All kinds of models will be differentiable
- Climate change impact assessment will be done using high-quality models that have absorbed big data
- Many theories will be rewritten
- WaterGPT?



Thank you!

@ChaopengShen
cshen@engr.psu.edu
Hydroml.org
https://github.com/mhpi

http://water.engr.psu.edu/shen/hydroDL.html

CUAHSI cyberseminar series on BDML

WRR special issue on BDML

AGU Editor’s review

Hydrol. Earth Syst. Sci., 22, 5639–5656, 2018
https://doi.org/10.5194/hess-22-5639-2018

Hydrology and Earth System Sciences
Open Access
EGU

HESS Opinions: Incubating deep-learning-powered hydrologic science advances as a community

Chaopeng Shen[1], Eric Laloy[2], Amin Elshorbagy[3], Adrian Albert[4], Jerad Bales[5], Fi-John Chang[6], Sangram Ganguly[7], Kuo-Lin Hsu[8], Daniel Kifer[9], Zheng Fang[10], Kuai Fang[1], Dongfeng Li[10], Xiaodong Li[11], and Wen-Ping Tsai[1]

Water Resources Research

REVIEW ARTICLE
10.1029/2018WR022643

Special Section:
Big Data & Machine Learning in Water Sciences: Recent Progress and Their Use in Advancing Science

A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists

Chaopeng Shen[1]

[1]Civil and Environmental Engineering, Pennsylvania State University, University Park, PA, USA

nature COMMUNICATIONS

ARTICLE
https://doi.org/10.1038/s41467-021-26107-z
OPEN
Check for updates

From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling

Wen-Ping Tsai[1], Dapeng Feng[1], Ming Pan[2,3], Hylke Beck[4], Kathryn Lawson[1,5], Yuan Yang[6,7], Jiangtao Liu[1] & Chaopeng Shen[1,5]✉