The data synergy and scaling effects of large-sample multiphysics catchment modeling with deep learning

CHAOPENG SHEN, WEN-PING TSAI, DAPENG FENG, KUAI FANG

[1]CIVIL AND ENVIRONMENTAL ENGINEERING, **PENN STATE UNIVERSITY**

CSHEN@ENGR.PSU.EDU

https://github.com/mhpi

@ChaopengShen

AI was a better player in very complex games

- AlphaGo 4:1 Lee Sedol

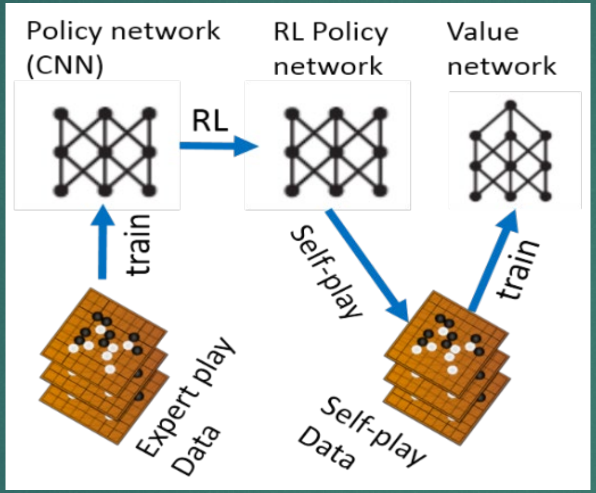

nature
International journal of science

Article | Published: 18 October 2017

Mastering the game of Go without human knowledge

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel & Demis Hassabis

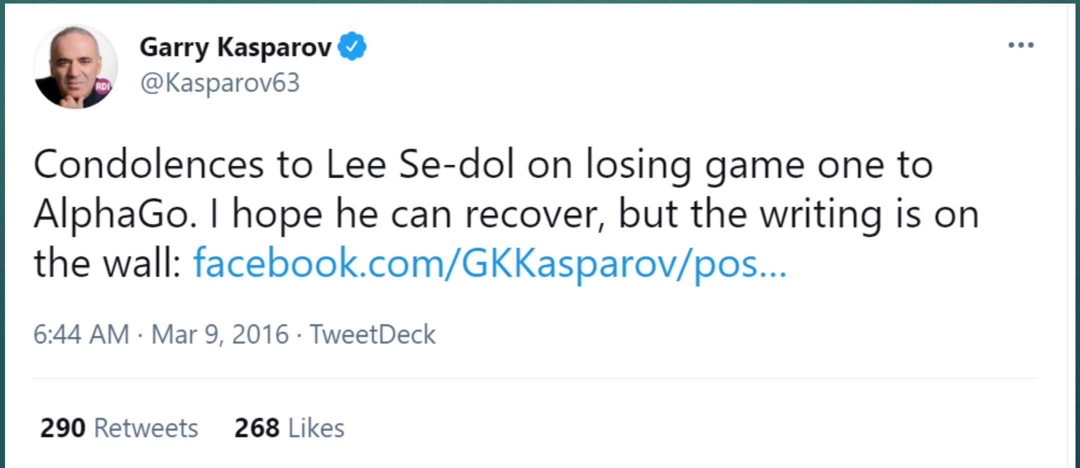
Nature **550**, 354–359 (19 October 2017) | Download Citation

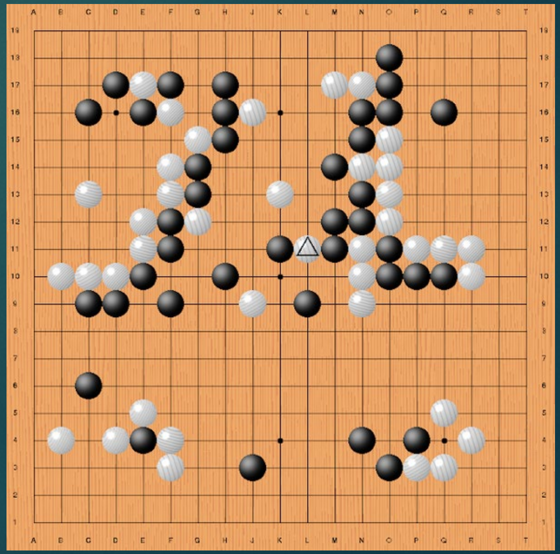


$N_{\text{atoms, universe}} \approx 10^{82}$.

$N_{go} = 10^{360}$

- Abstract
- Extremely complex mapping
- Once AI figures *it* out, we are a goner





Andrej Karpathy
@karpathy

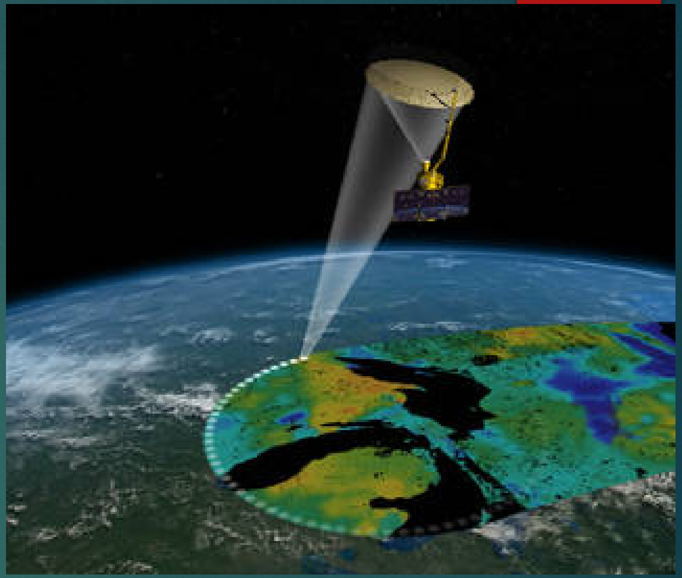
Gradient descent can write code better than you. I'm sorry.

4:56 PM · Aug 4, 2017 · Twitter Web Client

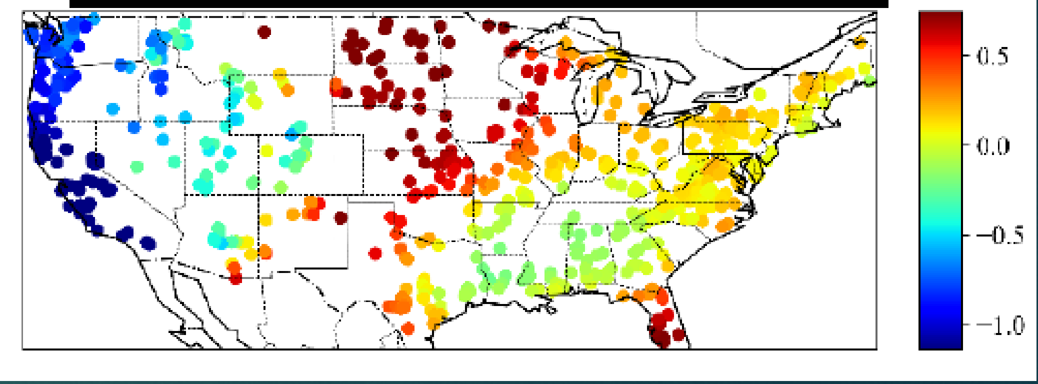
352 Retweets **52** Quote Tweets **1,217** Likes

Case studies

- 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



Rainfall seasonality for USA basins



A hydrologic model w/o structural assumptions…

LSTM model

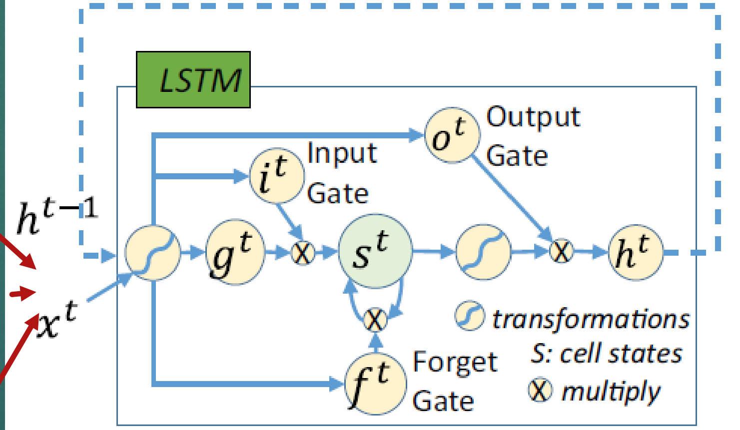
$$y = f(x, A)$$

Atm. Forcings (x)

(optional) Land Surface Model (Noah) solutions

Soil texture, slope, land cover, irrigation, depth to water table, etc (A)

(x)



(y)

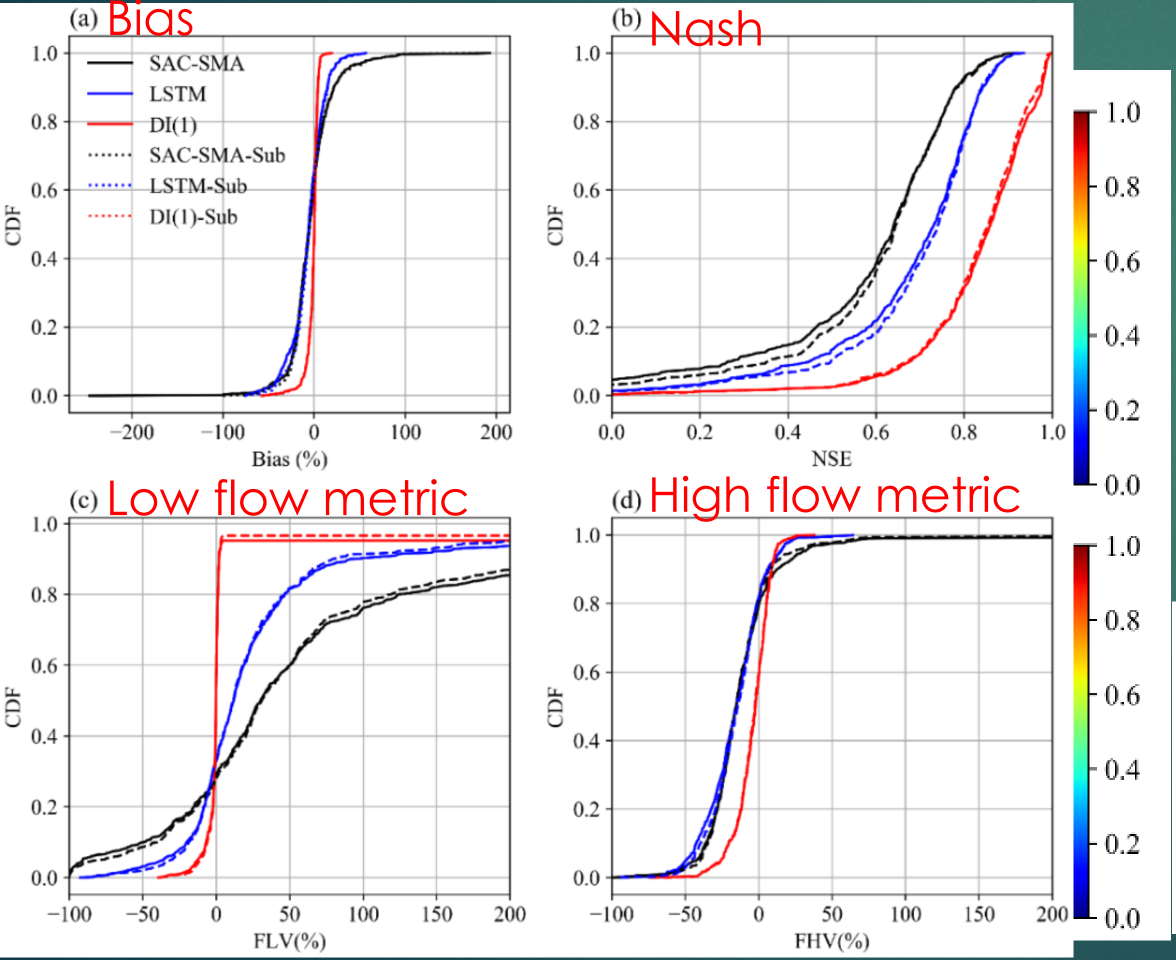
Target

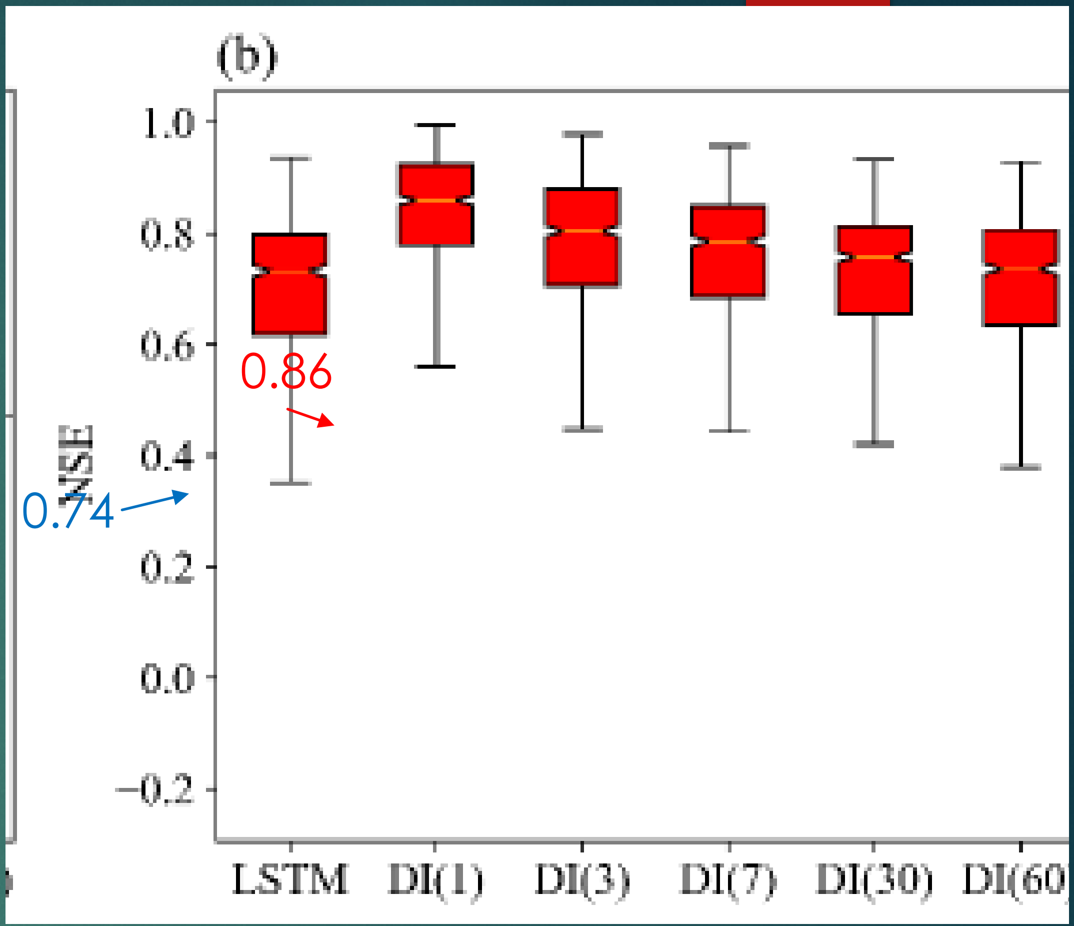
train

y (soil moisture, streamflow, water quality, etc.)

Application flythrough

Forecast for streamflow





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

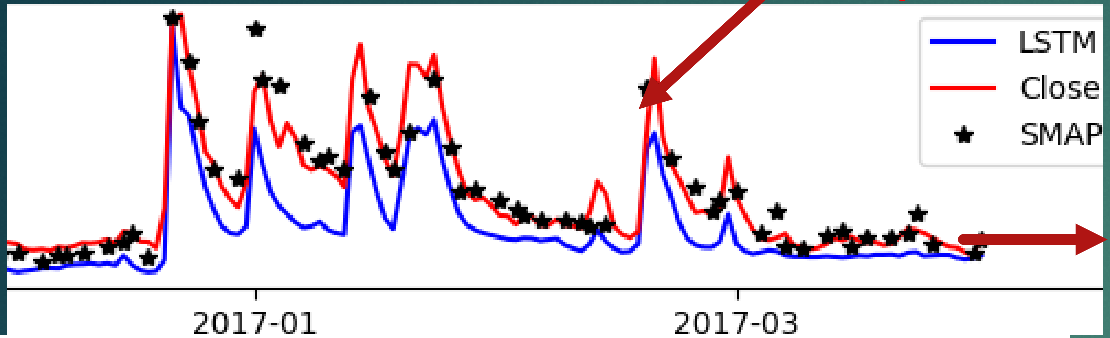
Enhancing Streamflow Forecast and Extracting Insights Using Long-Short Term Memory Networks With Data Integration at Continental Scales

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

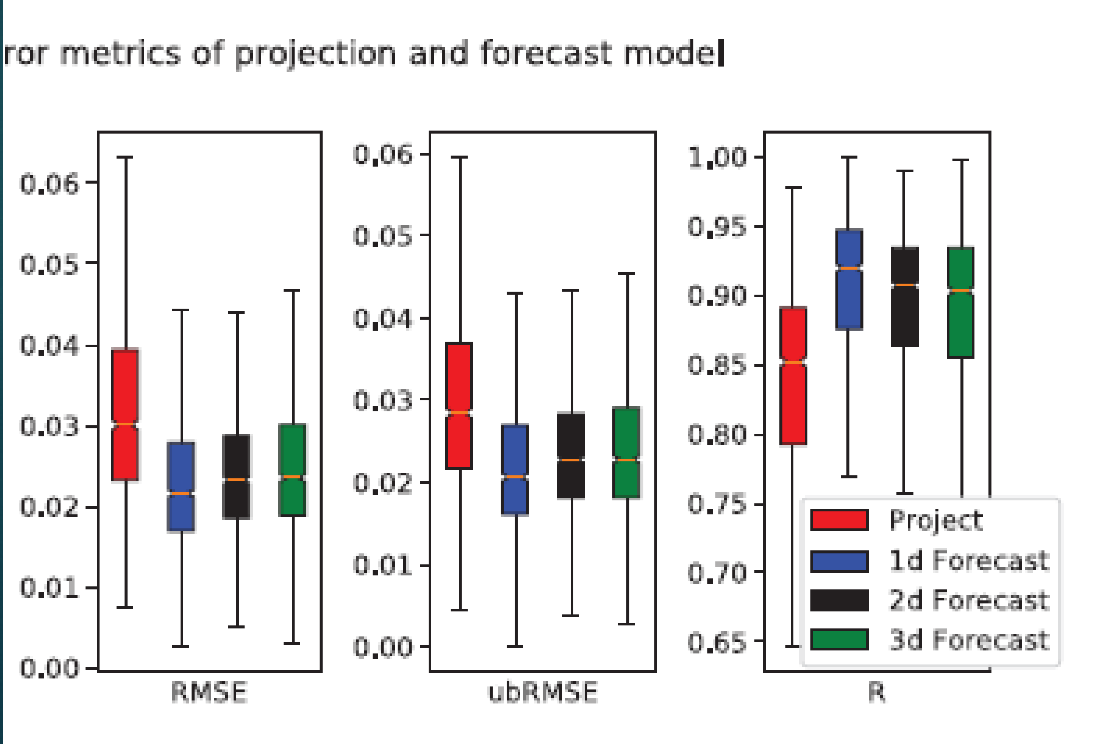
[1]Civil and Environmental Engineering, Pennsylvania State University, State College, PA, USA, [2]Now at: Earth System Science, Stanford University, Stanford, CA, USA

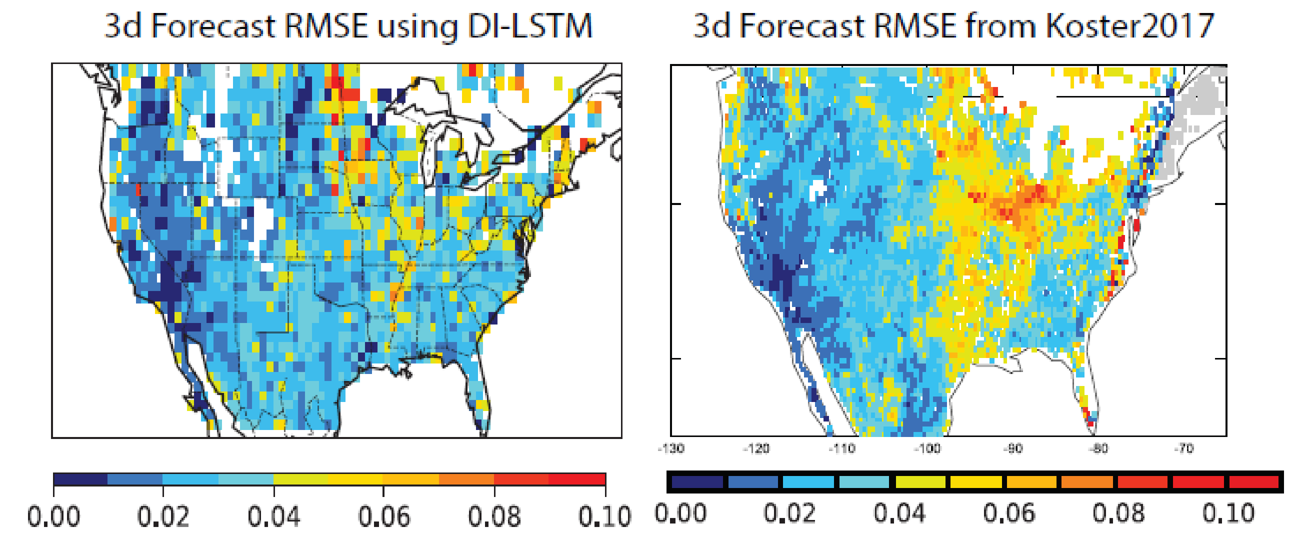
Application flythrough
Forecast for (i) soil moisture (ii) streamflow

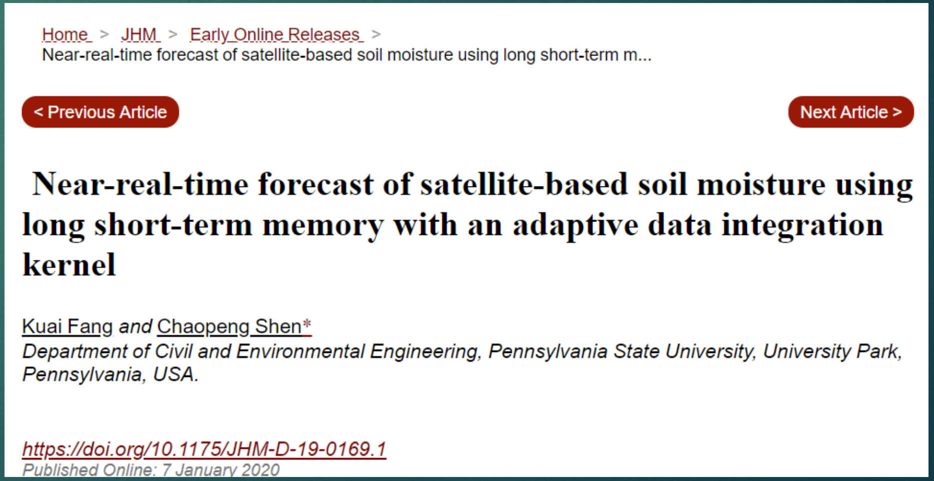
Gaps?



?







Home > JHM > Early Online Releases >
Near-real-time forecast of satellite-based soil moisture using long short-term m...

< Previous Article

Next Article >

Near-real-time forecast of satellite-based soil moisture using long short-term memory with an adaptive data integration kernel

Kuai Fang *and* Chaopeng Shen*
Department of Civil and Environmental Engineering, Pennsylvania State University, University Park, Pennsylvania, USA.

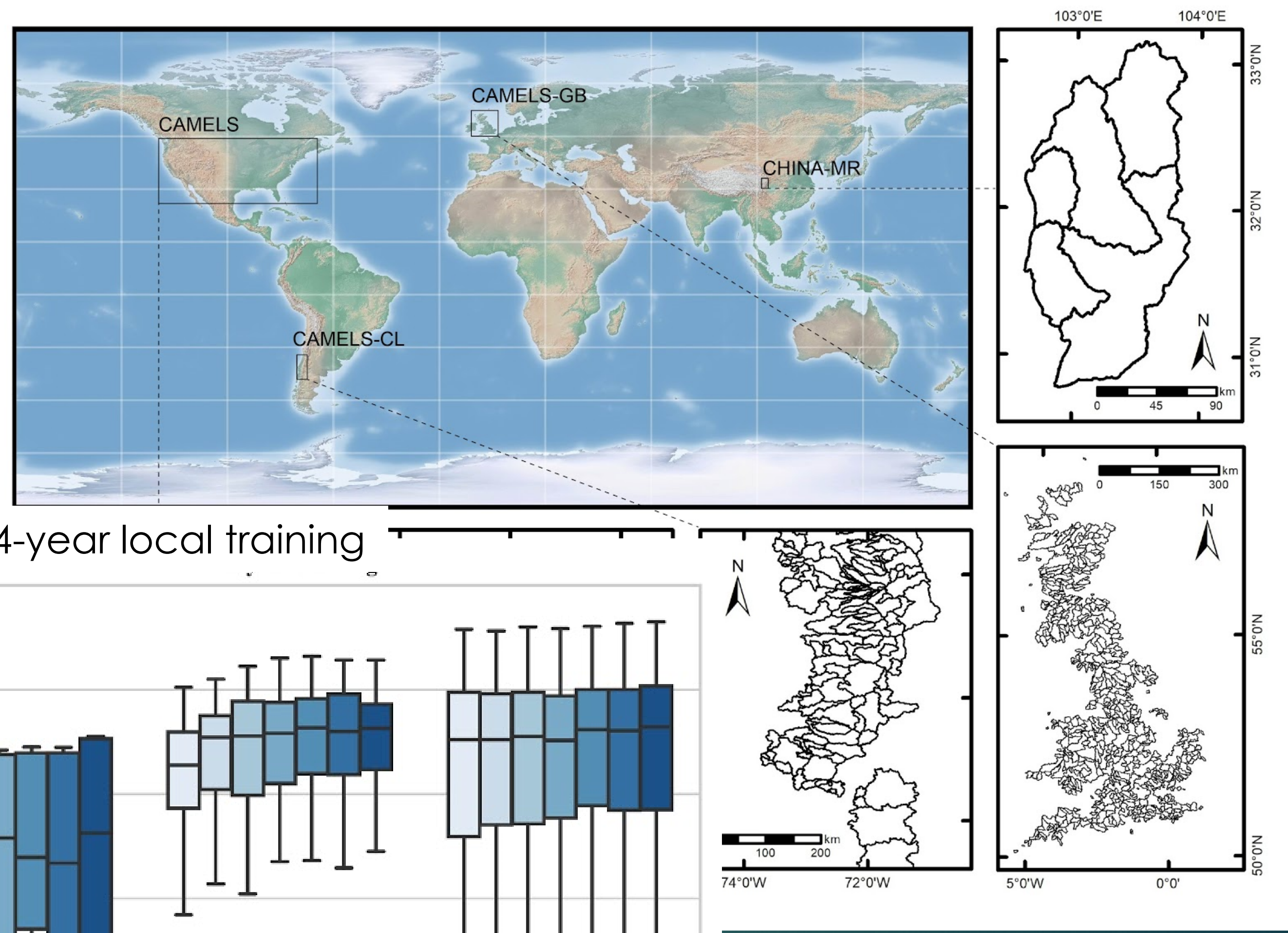
https://doi.org/10.1175/JHM-D-19-0169.1
Published Online: 7 January 2020

https://sites.google.com/view/mhpi/locust

Application flythrough

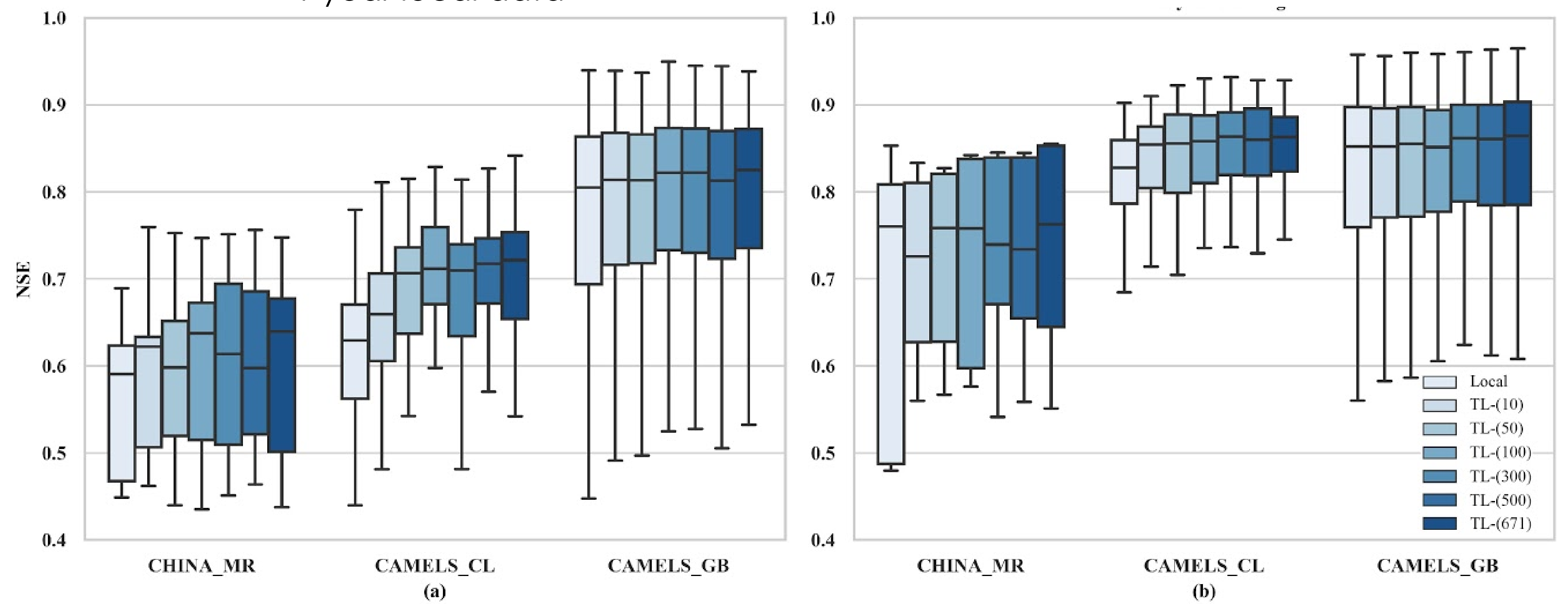
Sparse-data region

- Transfer learning



1 year local data

4-year local training



Ma et al., WRR, accepted, preprint doi: 10.1002/essoar.10504132.1

Application flythrough
Reservoir

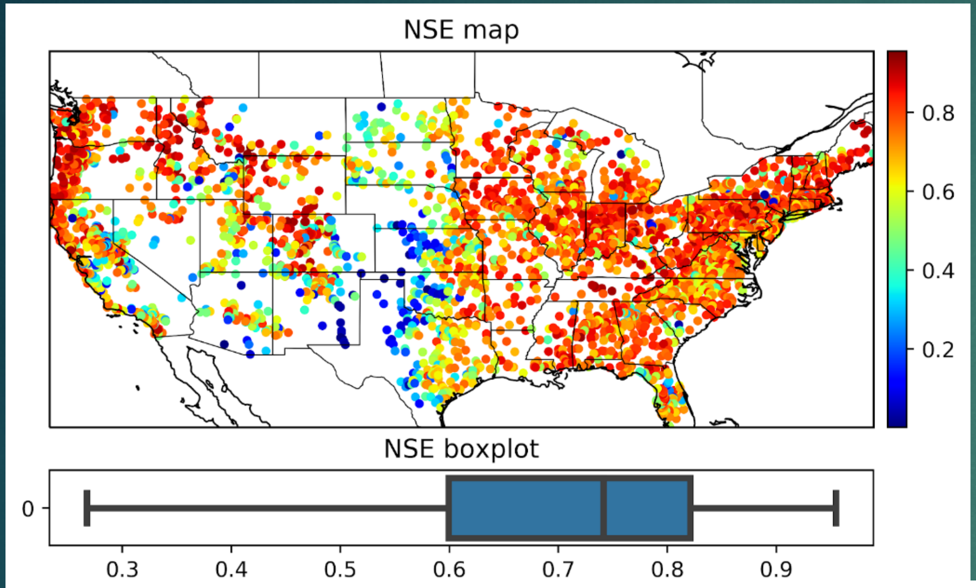
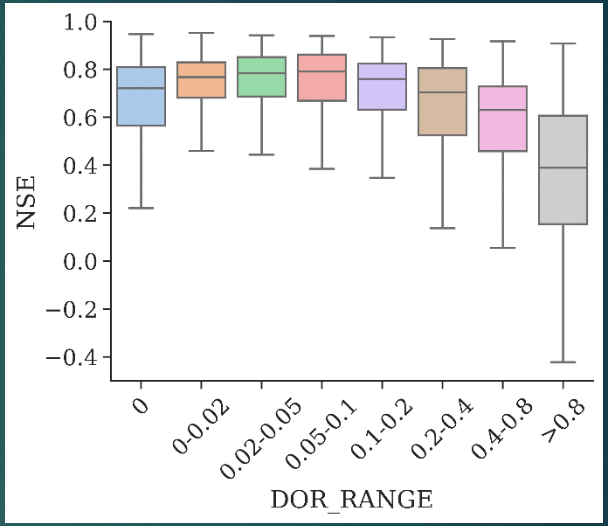
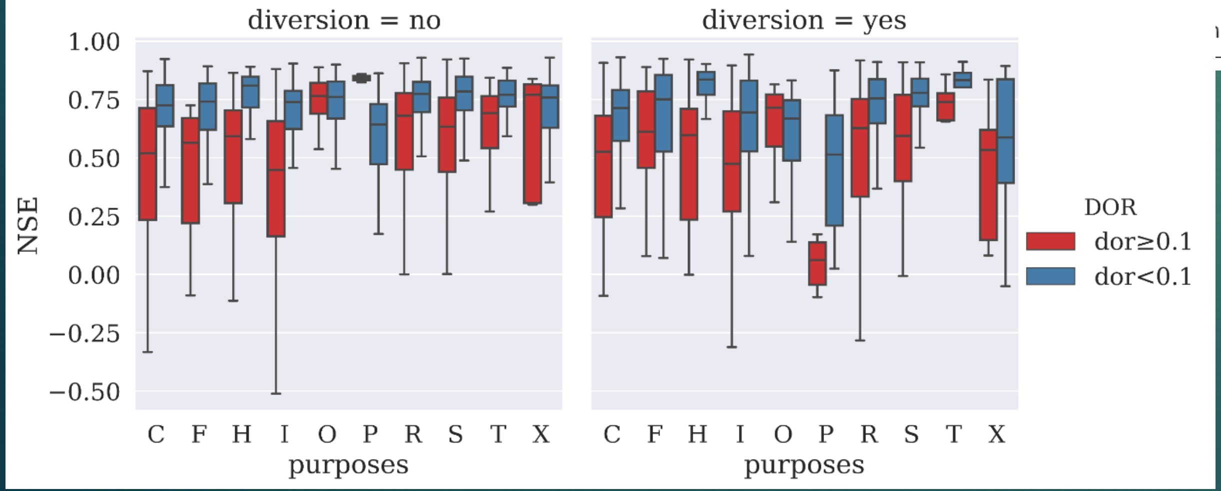


Table 1. The types of main reservoir purpose in a basin

Type	Purpose	Number
C	Flood Control and Storm Water Management	313
F	Fish and Wildlife Pond	94
H	Hydroelectric	196
I	Irrigation	328
O	Other	163
P	Fire Protection, Stock, or Small Farm Pond	66
R	Recreation	1207
S	Water Supply	426
T	Tailings	52
[illegible]	[illegible]	66

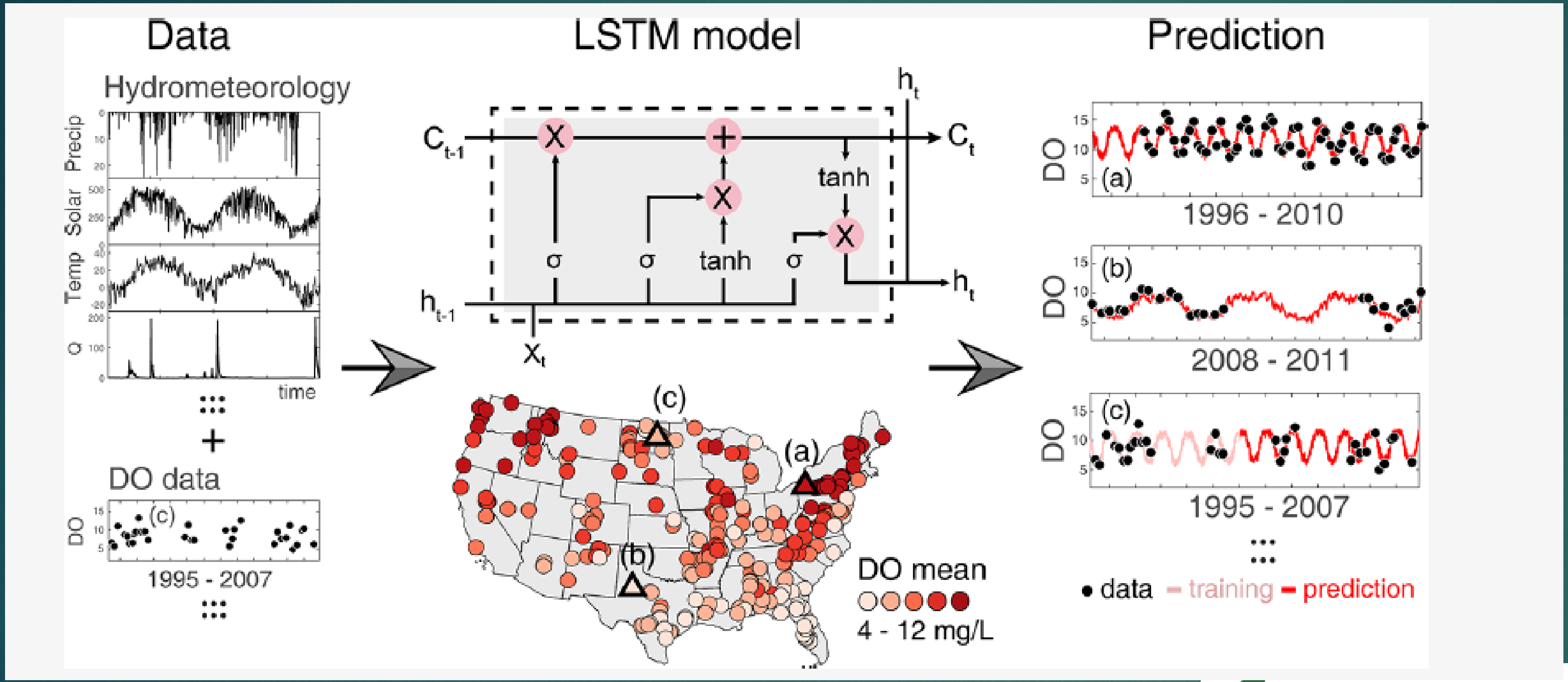




- Modeling reservoir over CONUS
- Basins with small reservoirs (<a month of flow) can be directly simulated.
- They are different from reference basins!

Ouyang WY et al., Journal of Hydrology, accepted, preprint: https://arxiv.org/abs/2101.04423

Application flythrough
Dissolved oxygen



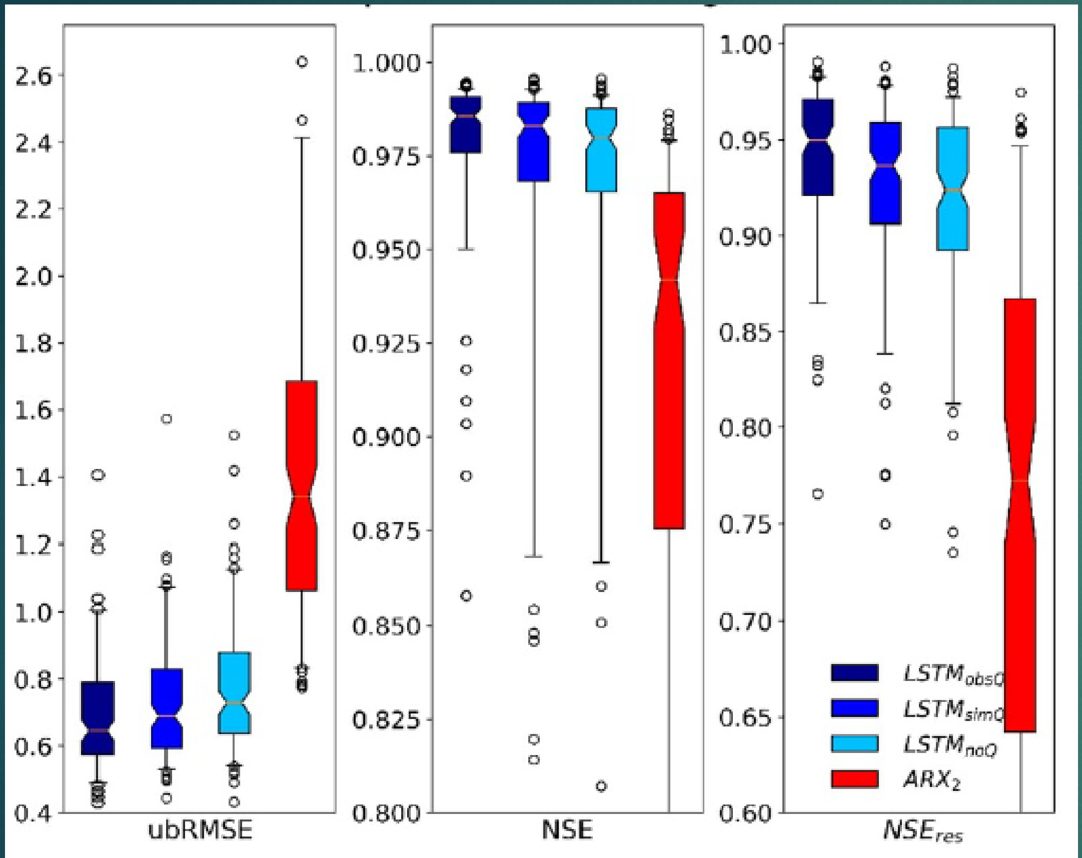
Environmental Science & Technology
pubs.acs.org/est
Article

From Hydrometeorology to River Water Quality: Can a Deep Learning Model Predict Dissolved Oxygen at the Continental Scale?

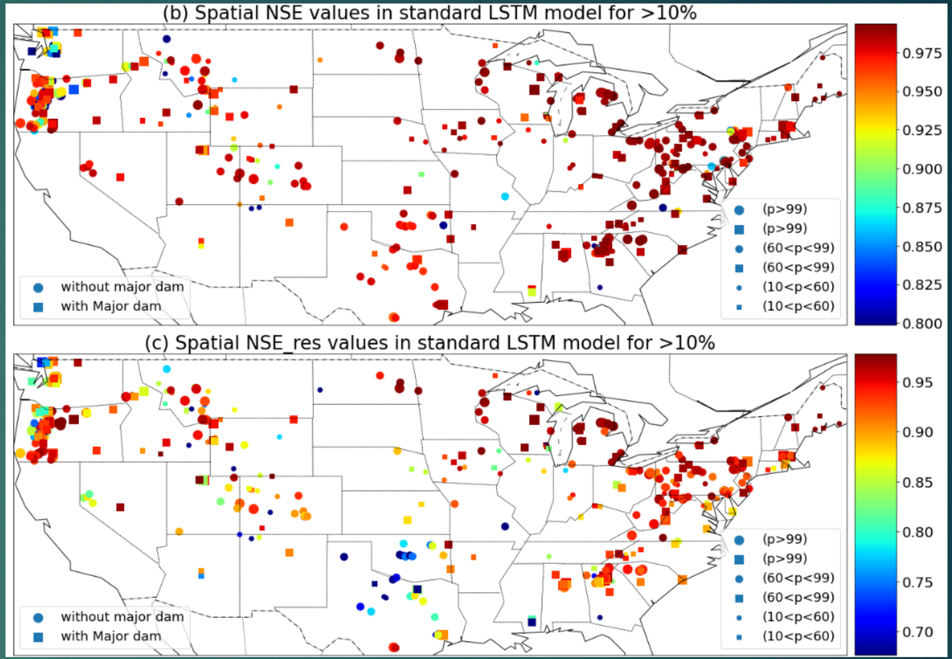
Wei Zhi, Dapeng Feng, Wen-Ping Tsai, Gary Sterle, Adrian Harpold, Chaopeng Shen, and Li Li*

Cite This: *Environ. Sci. Technol.* 2021, 55, 2357−2368
Read Online

Application flythrough
Stream water temperature model



n temperature models for the test period. $LSTM_{obsQ}$ incorporated observed
rmation, while $LSTM_{simQ}$ incorporated simulated streamflow (Q_{sim}). ARX_2 is
uts. The lower whisker, lower box edge, center bar, upper box edge and upper
lata, respectively.



ENVIRONMENTAL RESEARCH
LETTERS

LETTER • OPEN ACCESS

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]

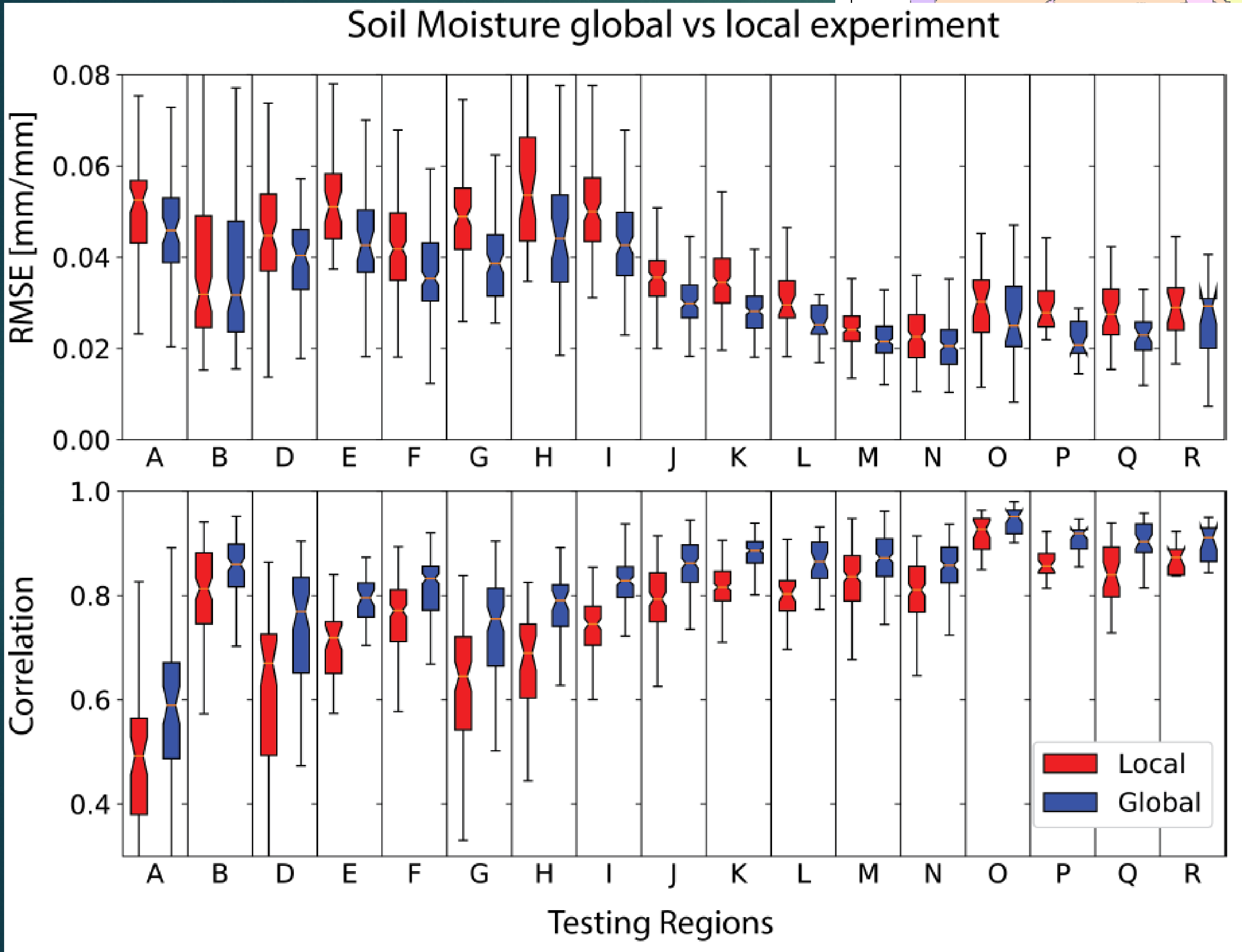
Published 27 January 2021 • © 2021 The Author(s). Published by IOP Publishing Ltd
Environmental Research Letters, Volume 16, Number 2
Citation Farshid Rahmani *et al* 2021 *Environ. Res. Lett.* **16** 024025

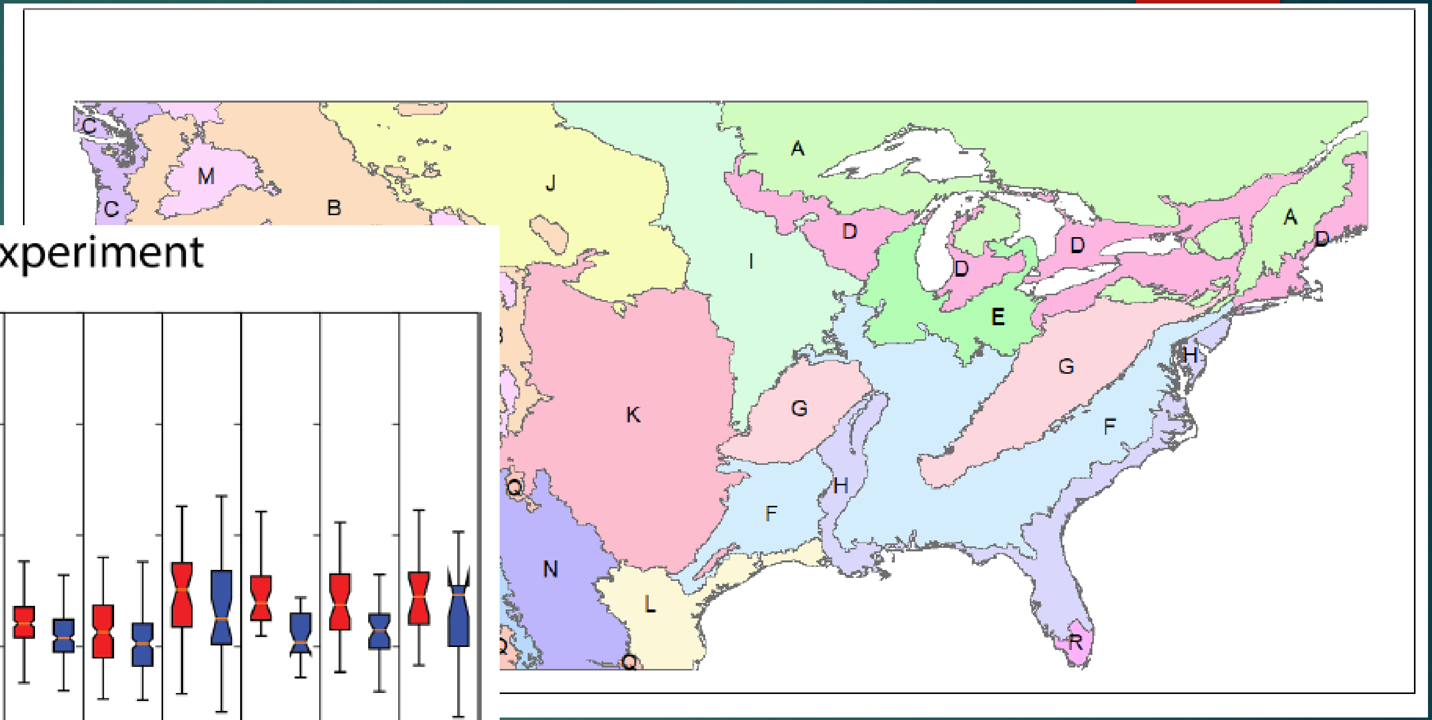
Data synergy and data scaling

- Deep learning (DL) inherently works better with bigger data
- Here we demonstrate the virtuous scaling of DL with big data
- Examples:
  (1) Direct data-driven LSTM for (i) soil moisture; (ii) streamflow

  (2) differentiable Parameter learning for (i) soil moisture; (ii) CAMELS streamflow; (iii) global prediction in ungauged basins.

Data synergy

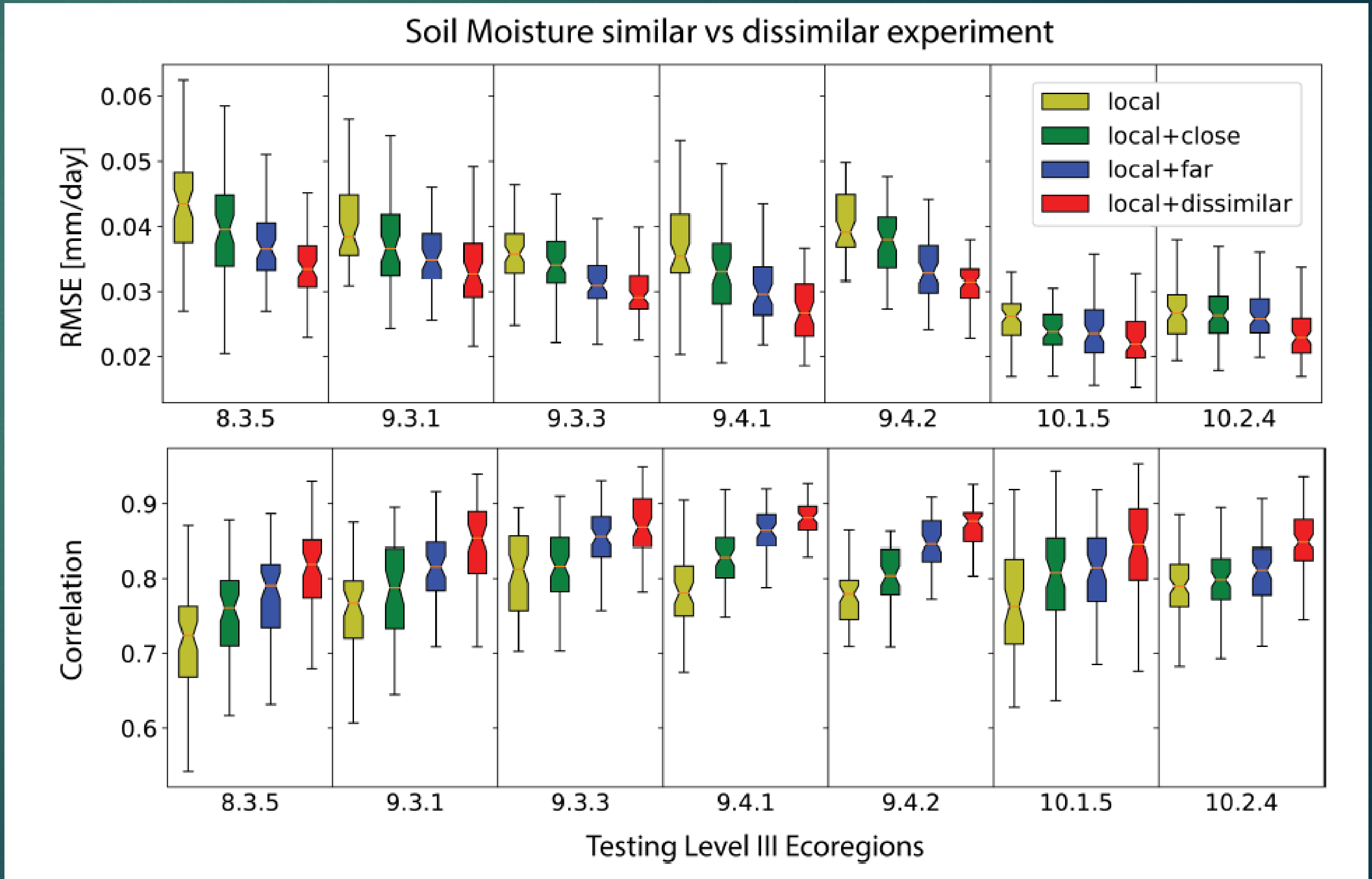




- Global model > Local model
- Global model w/ more diverse data is an inherent advantage of using DL

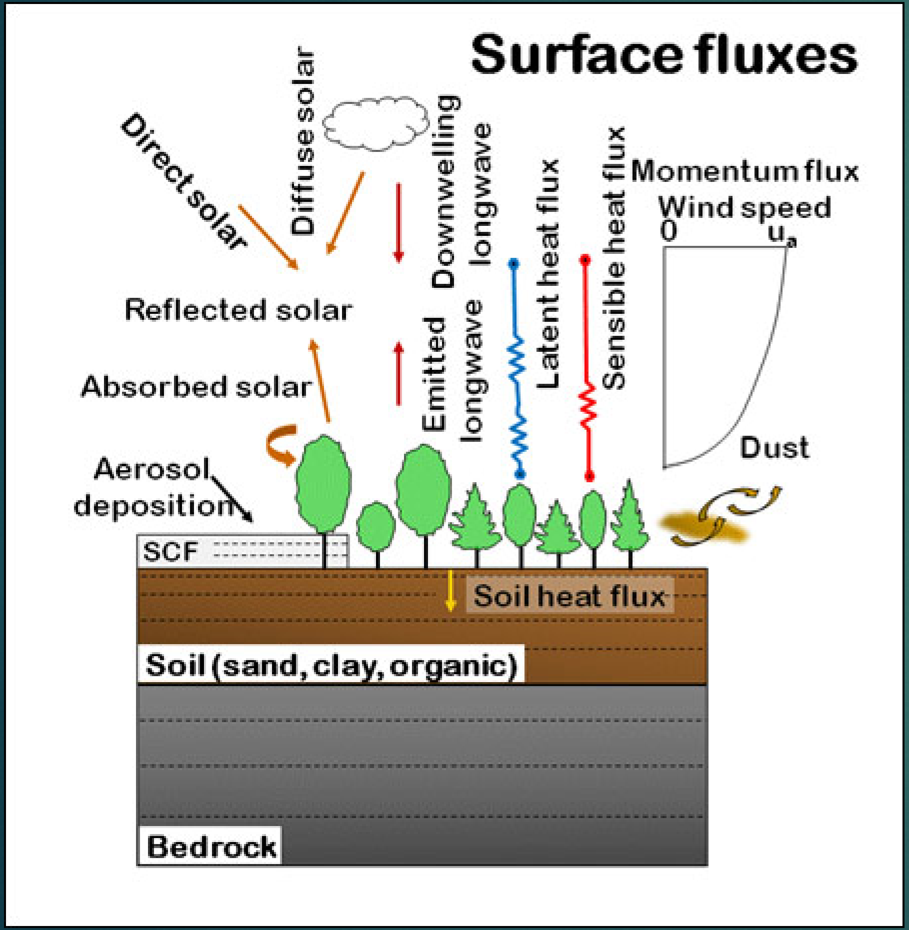
Having dissimilar examples are good!

- Even for the same amount of training data, DL models prefer more dissimilar training examples.



Fang et al.
https://arxiv.org/abs/2101.01876

Interactions w/ ecosystem/biogeochemistry



How about variables we cannot observe accurately on large scales?
→ ET, Groundwater, deeper soil moisture?

From parameter calibration to →

differentiable parameter learning (dPL)

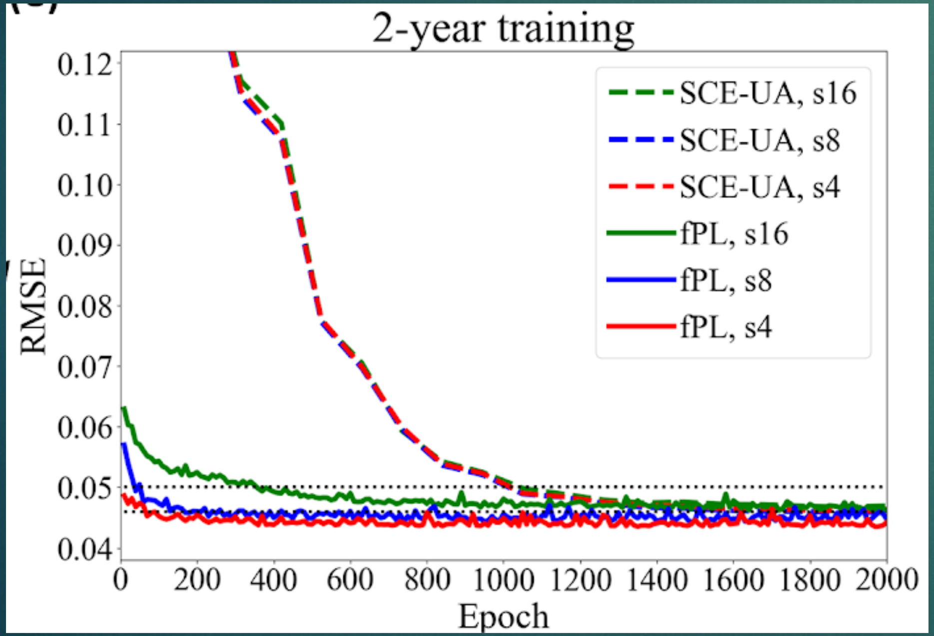
Network → Parameters → Hydrologic model e.g., VIC → Outputs

We calibrated VIC model using SMAP data and the DL-based dPL scheme.

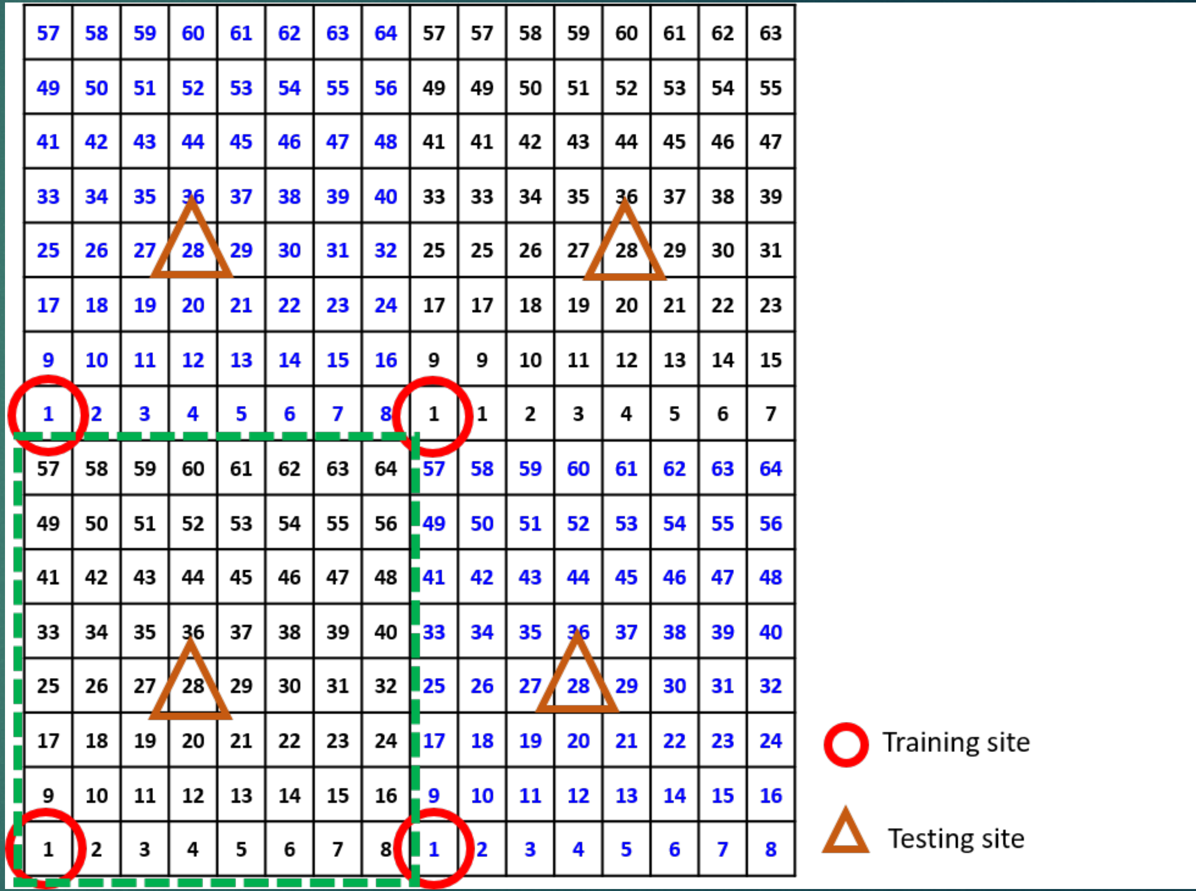
A knowledge learning opportunity!

Parameter learning (*dPL*) -- results

- Stronger than SCE-UA!
- Saves $O(10^4)$ computation!
- Now capable of modeling other variables such as ET and/or streamflow

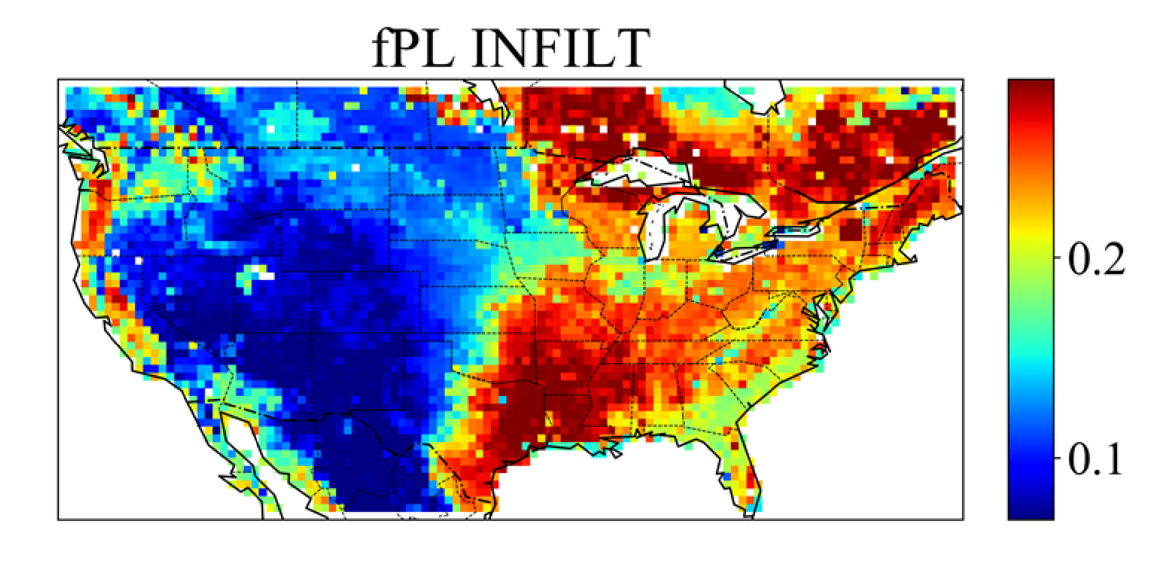


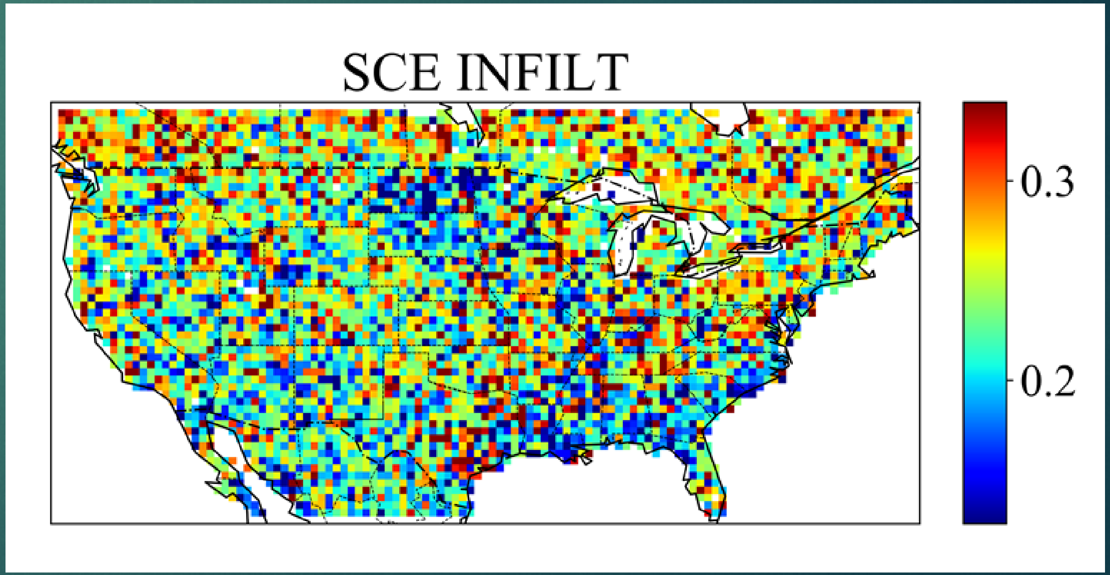
More efficient and scales better with more data!



Parameter learning (*fPL*) -- results

- Stronger than SCE-UA!
- Saves $O(10^4)$ computation!
- Now capable of modeling other variables such as ET and/or streamflow

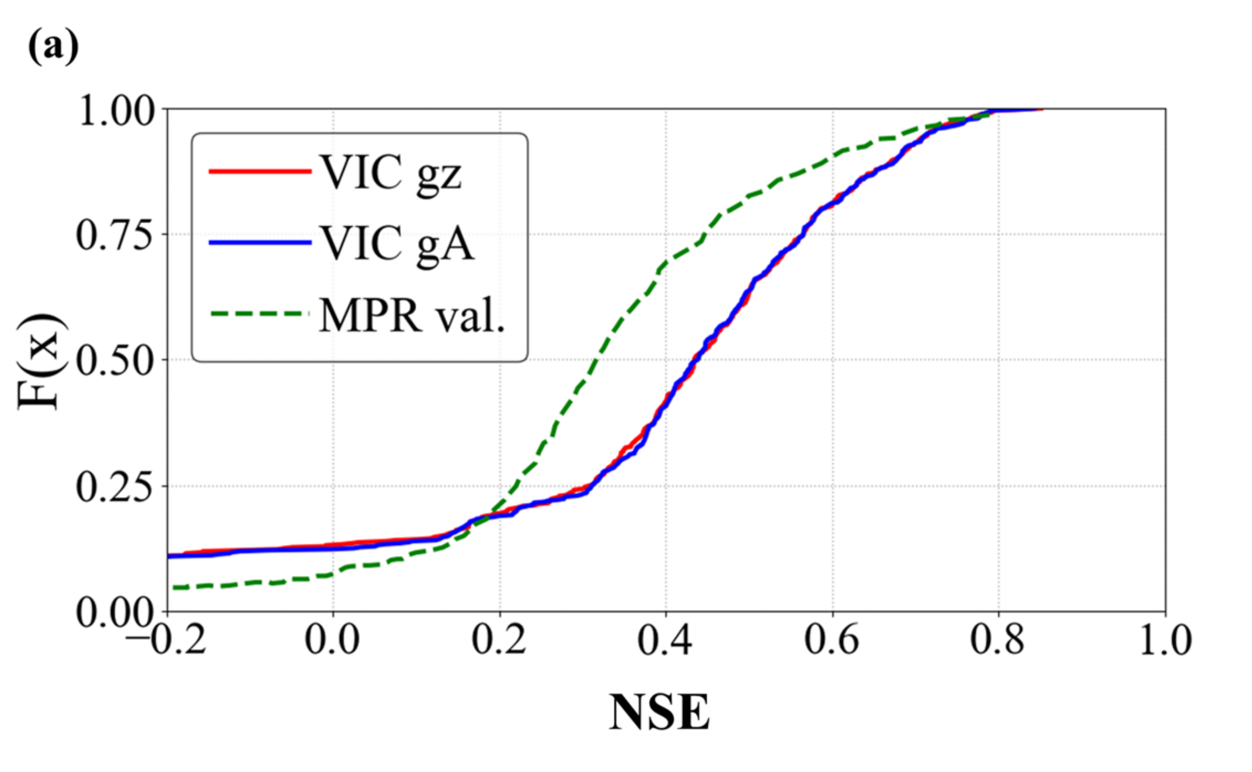




Tsai et al. 2020, half preprint, https://arxiv.org/abs/2007.15751

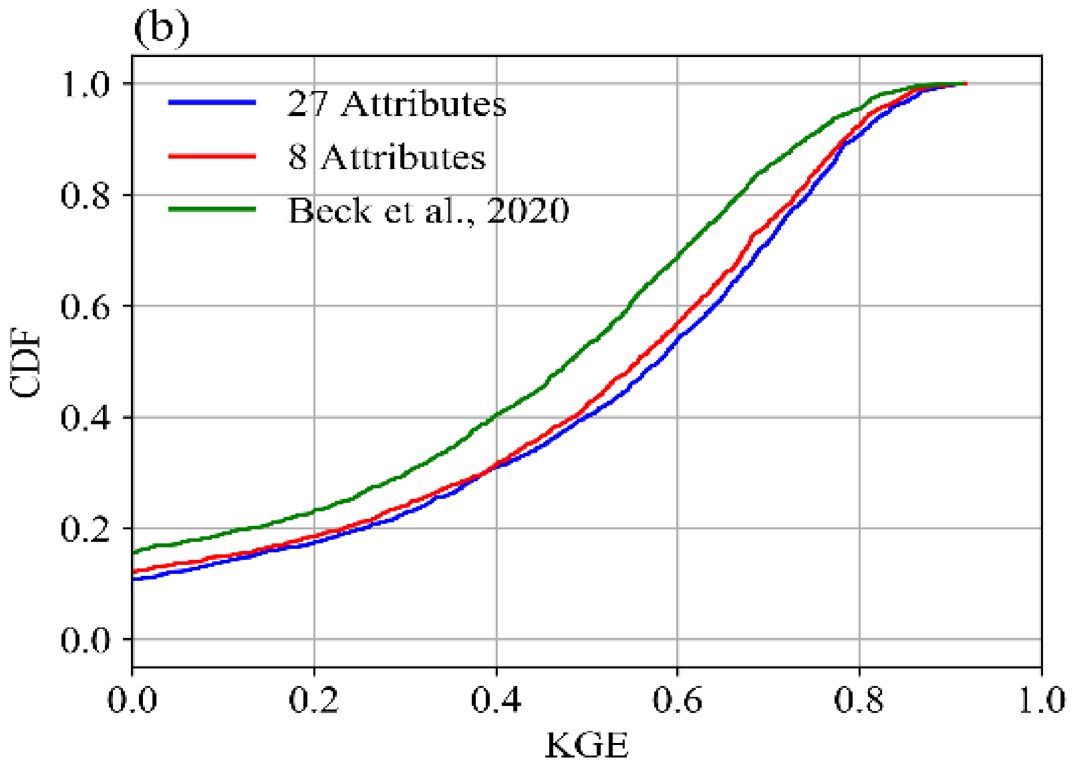
Comparing with regionalization schemes

Comparing w/ MPR



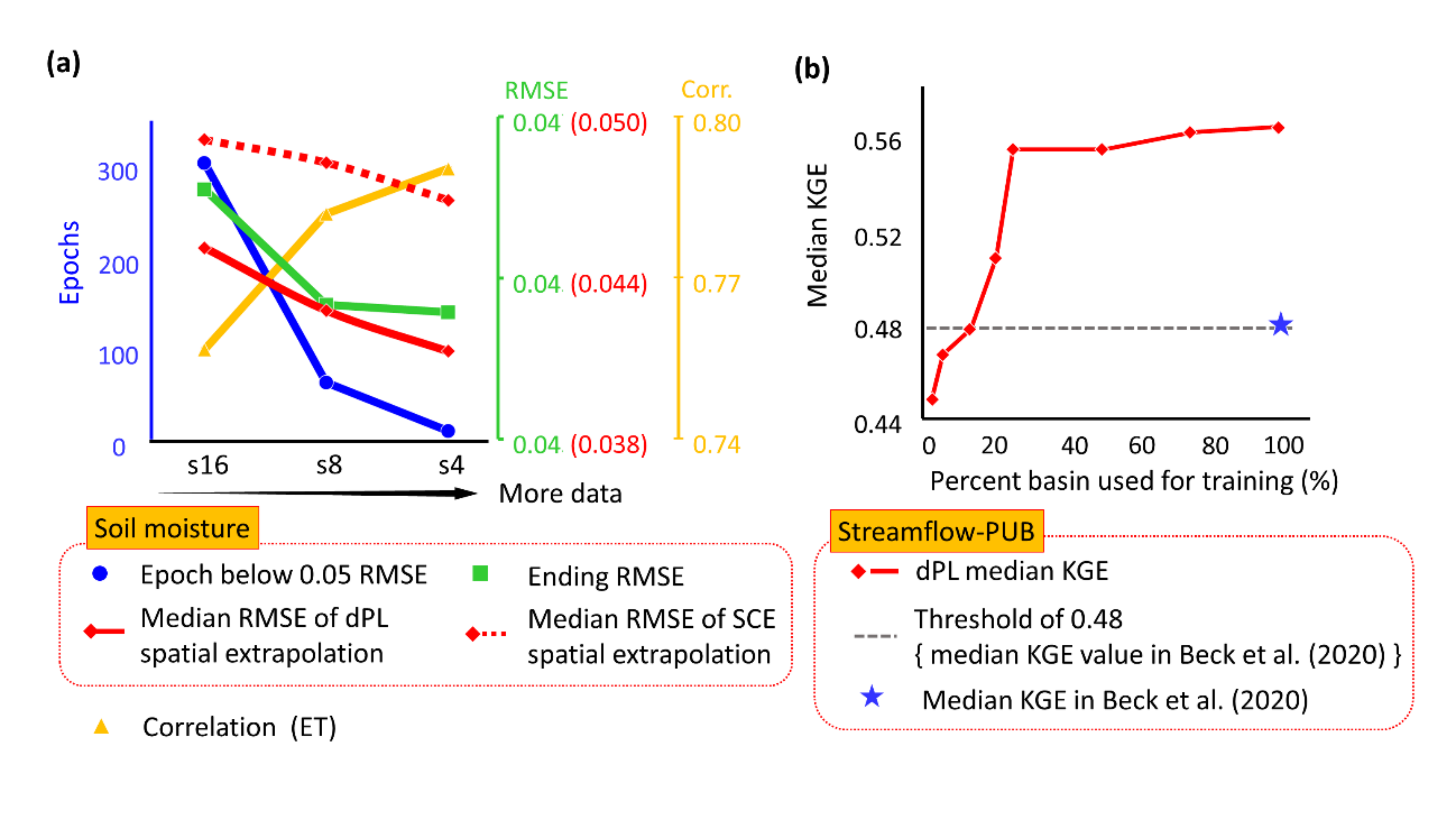
CAMELS – temporal generalization

Comparing with Beck et al. 2020



Global PUB – spatial generalization, using HBV

Virtuous scaling curves of dPL



Each gridcell gets better service+ Economies of scale!

Conclusion

- DL gains its strength from big data, and behaves differently depending on training data quantity.
- DL models prefer to *see* data with diversity and differences (although not irrelevantly different). It does NOT prefer homogeneous training data.
- DL powered schemes like LSTM or parameter learning may have a virtuous scaling curve where everything becomes better with more data.
- This encourages the community to share data and *ride the scaling curve up* together!

References

1. Fang, K.* et al., The data synergy effects of time-series deep learning models in hydrology
2. Tsai, W-P.* et al. From parameter calibration to parameter learning: Revolutionizing large-scale geoscientific modeling with big data
3. Feng, DP*, K. Fang* and **CP. Shen**, Enhancing streamflow forecast and extracting insights using continental-scale long-short term memory networks with data integration, ***Water Resources Research***, doi: 10.1029/2019WR026793 (2020)
4. Feng, DP.* et al., Prediction in ungauged regions with sparse flow duration curves and input-selection ensemble modeling
5. Ouyang WY.** et al., Continental-scale streamflow modeling of basins with reservoirs: a demonstration of effectiveness and a delineation of challenges, Journal of hydrology. Accepted, 2021
6. Ma K.**, DP. Feng*, K. Lawson*, W-P Tsai*, C. Liang, XR. Huang, A. Sharma*, and **CP. Shen**, Transferring hydrologic data across continents –– leveraging data-rich regions to improve hydrologic prediction in data-sparse regions, ***Water Resources Research***, doi: 10.1029/2020WR028600 (2021) pre-review preprint
7. Zhi, W., DP. Feng*, WP Tsai*, G. Sterle, A. Harpold, **CP. Shen** and L. Li, From hydrometeorology to river water quality: can a deep learning model predict dissolved oxygen at the continental scale? ***Environmental Science & Technology*** , doi: 10.1021/acs.est.0c06783 (2021) pre-review preprint
8. Rahmani, F.*, K. Lawson*, WY. Ouyang**, A. Appling, S. Oliver and **CP. Shen**, Exploring the exceptional performance of a deep learning stream temperature model and the value of streamflow data, ***Environmental Research Letters***, doi: 10.1088/1748-9326/abd501 (2020)
9. Fang, K.*, D. Kifer, K. Lawson* and **CP. Shen**, Evaluating the potential and challenges of an uncertainty quantification method for long short-term memory models for soil moisture predictions, ***Water Resources Research***, doi: 10.1029/2020WR028095 (2020) preprint link
10. Fang, K.* and **CP. Shen**, Near-real-time forecast of satellite-based soil moisture using long short-term memory with an adaptive data integration kernel, ***Journal of Hydrometeorology***, JHM-D-19-0169.1, doi:10.1175/JHM-D-19-0169.1 (2020)
11. **Shen, CP.**, Laloy, E., Albert, A., Chang, F.-J., Elshorbagy, Bales, J., A., Ganguly, S., Hsu, K.-L., Kifer, D., Fang, Z., Fang, K.*, Li, D., Li, X., and Tsai, W.-P.*, HESS Opinions: Incubating deep-learning-powered hydrologic science advances as a community, ***Hydrol. Earth Syst. Sci.***, 22, 5639–5656, doi: 10.5194/hess-22-5639-2018, (2018)
12. Fang, K*, M. Pan, and **CP. Shen**, The value of SMAP for long-term soil moisture estimation with the help of deep learning, ***Transactions on Geoscience and Remote Sensing***, 57(4), 2221-2233, doi: 10.1109/TGRS.2018.2872131 (2018)
13. **Shen, CP.**, A trans-disciplinary review of deep learning research and its relevance for water resources scientists, ***Water Resources Research***. 54(11), 8558-8593, doi: 10.1029/2018WR022643 (2018)
14. **Shen, CP.**, Deep learning: A next-generation big-data approach for hydrology, ***Eos***, 99, https://doi.org/10.1029/2018EO095649 (2018)
15. **Shen, CP.**, XY. Chen and E. Laloy, Editorial: Broadening the Use of Machine Learning in Hydrology, ***Frontiers in Water –Water and Hydrocomplexity***, doi: 10.3389/frwa.2021.681023 (2021)

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