Machine Learning for Land Modeling: Lessons from Hydrologic Benchmarking

From Hydro-Climatic Robustness to Observation-Based Soil Moisture Benchmarks

Sungmin O

Kangwon National University, Samcheok, Republic of Korea

ECMWF ML4LM

March 2026

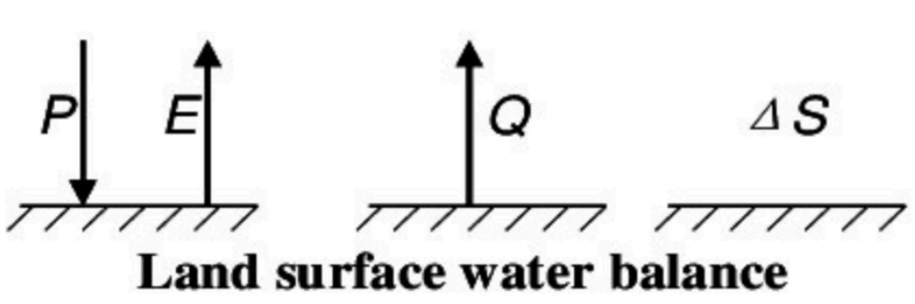
Performance of ML/DL under changing conditions

Process-based vs. Data-driven

Process-based

$$\Delta S = P - E - Q$$

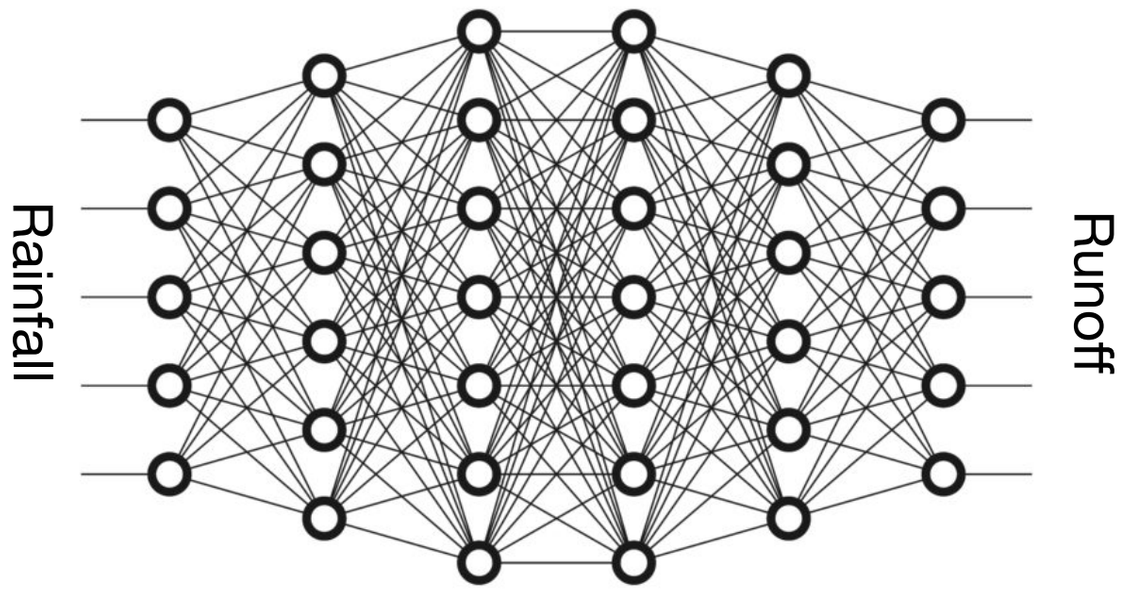
Change in storage = Precipitation − Actual evapotranspiration − Discharge



Based on process knowledge

Mathematical ***representation of processes***; e.g. physically-based or *conceptual models.*

Data-driven



Based on input-output data

No physical knowledge is required (empirical); e.g. *machine-learning, neural network, nonlinear regression.*

Hydrology as a testbed for data-driven modeling

Proceedings of 1993 International Joint Conference on Neural Networks

Application of Artificial Neural Networks in Hydrological Modeling : A Case Study of Runoff Simulation of a Himalayan Glacier Basin.

A. M. Buch [+], Member, IEEE; H. S. Mazumdar [*], P. C. Pandey [+]

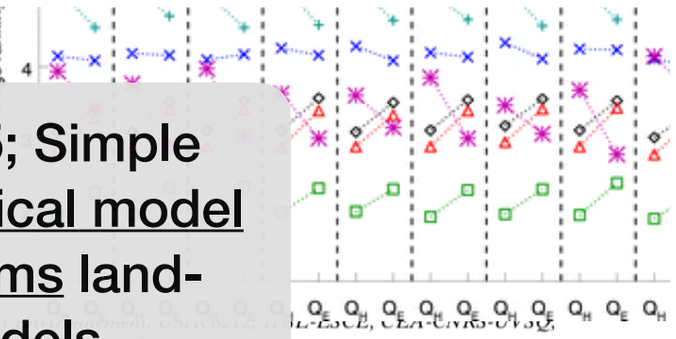
[+] Space Applications Center, Ahmedabad 380 - 053, INDIA.
[*] Physical Research Laboratory, Ahmedabad 380 - 009, INDIA.

Buch et al, 1993; Artificial neural network (ANN) model was found to be superior to the energy balance model

The Plumbing of Land Surface Models: Benchmarking Model Performance

M. J. Best,[a] G. Abramowitz,[b] H. R. Johnson,[a] A. J. Pitman,[b] G. Balsamo,[c] A. Boone,[d] M. Cuntz,[e] B. Decharme,[d] P. A. Dirmeyer,[f] J. Dong,[g] M. Ek,[g] Z. Guo,[f] V. Haverd,[h] B. J. J. van den Hurk,[i] G. S. Nearing,[j] B. Pak,[k] C. Peters-Lidard,[j] J. A. Santanello Jr.,[j] L. Stevens,[k] and N. Vuichard[l]

[a] Met
[b] ARC Centre of Excellence for Cli



Best et al, 2015; Simple regression empirical model often outperforms land-surface models

Artificial neural network modeling of the rainfall-runoff process

Kuo-lin Hsu, Hoshin Vijai Gupta, and Soroosh Sorooshian
Department of Hydrology and Water Resources, University of Arizona, Tucson

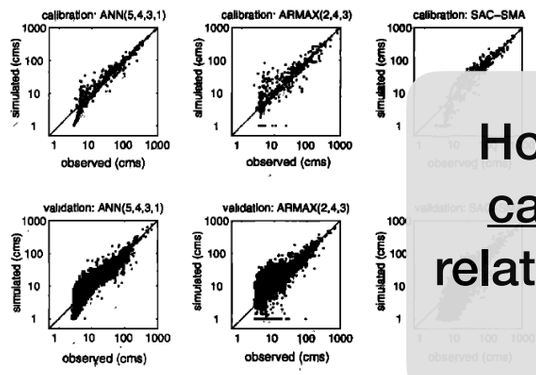


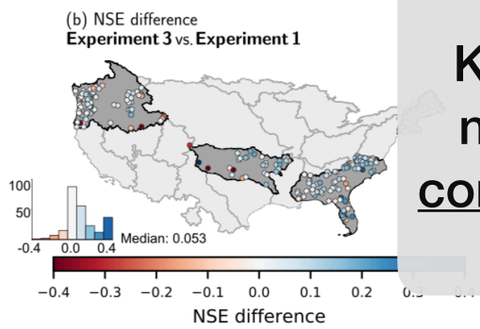
Figure 7. Scatterplots comparing simulated and observed flows for calibration data and validation data.

Hou et al, 1995; ANN better captures the rainfall-runoff relationship than the SCA-SMA conceptual model

Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks

Frederik Kratzert[1,*], Daniel Klotz[1], Claire Brenner[1], Karsten Schulz[1], and Mathew Herrnegger[1]

[1]Institute of Water Management, Hydrology and Hydraulic Engineering, University of Natural Resources and Life Sciences, Vienna, 1190, Austria



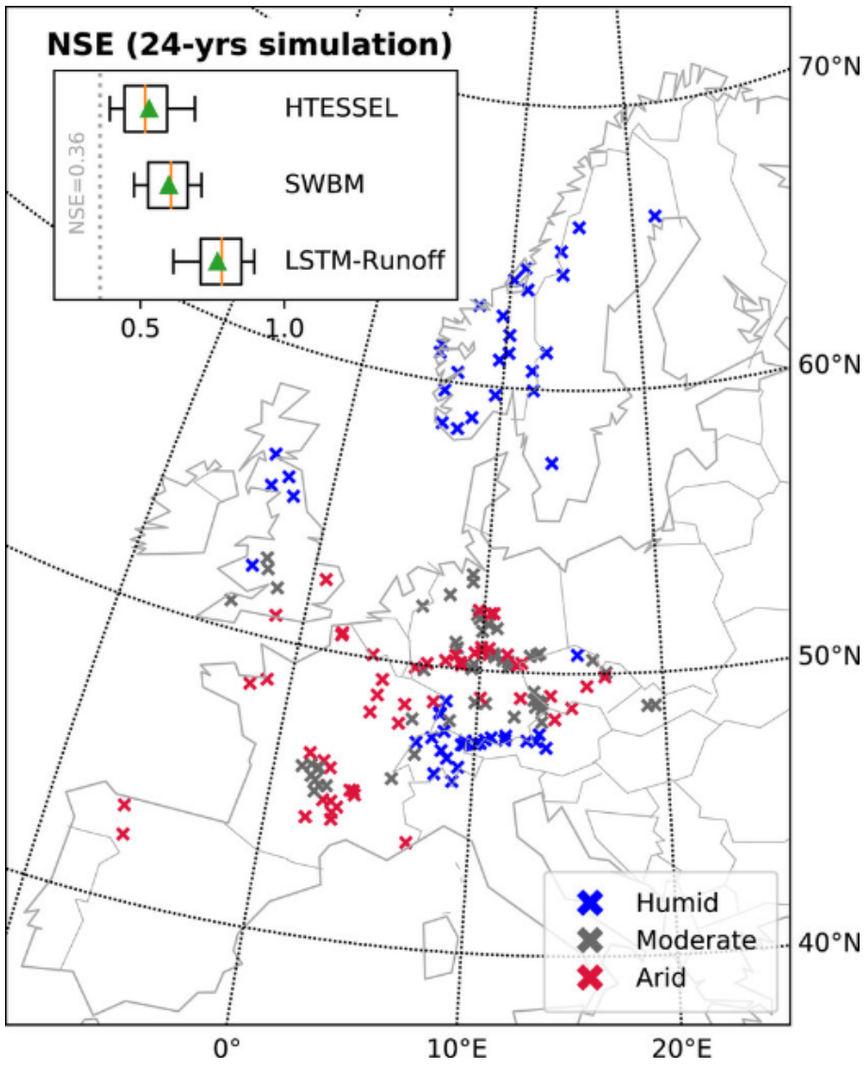
Kratzert et al, 2019; LSTM networks' performance is comparable to the SAC-SMA

Hydrology as a testbed for data-driven modeling

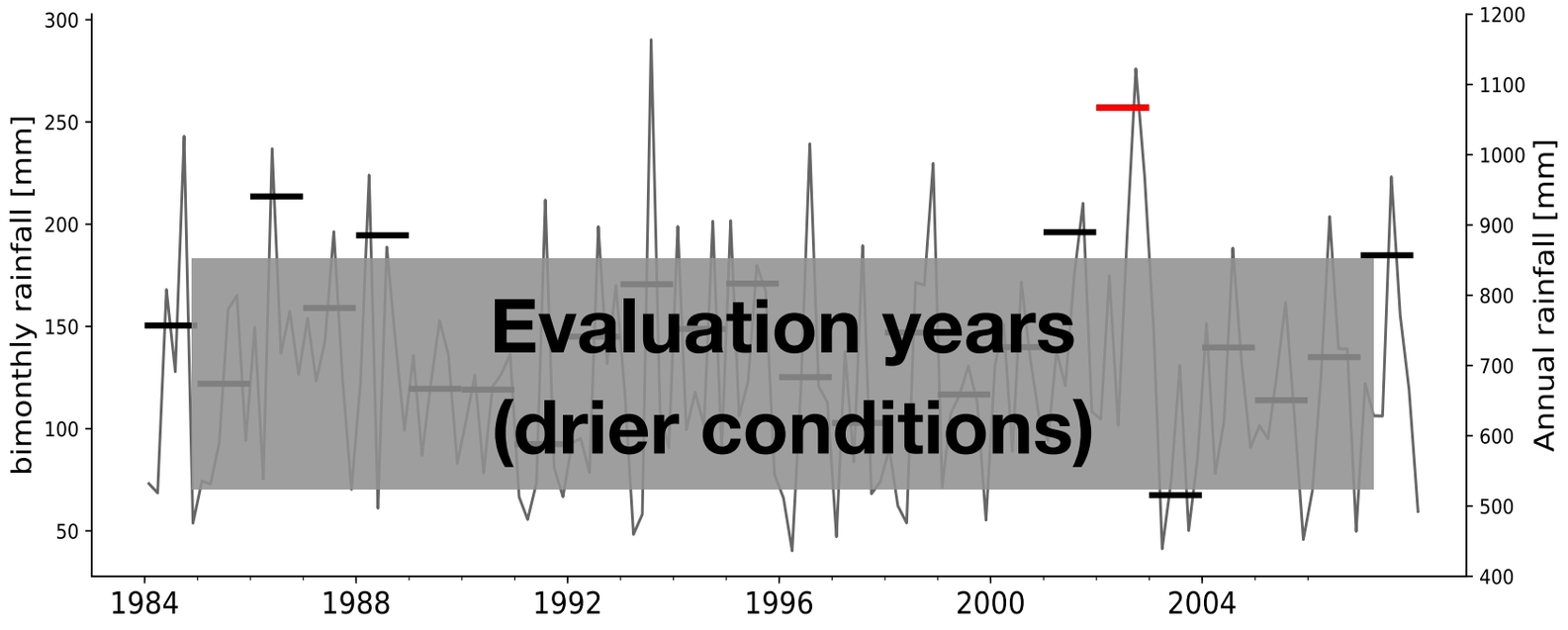
Predictive performance of data-driven models is as good as, or better than, established process-based models

But, do they remain robust under changing hydro-climatic conditions?

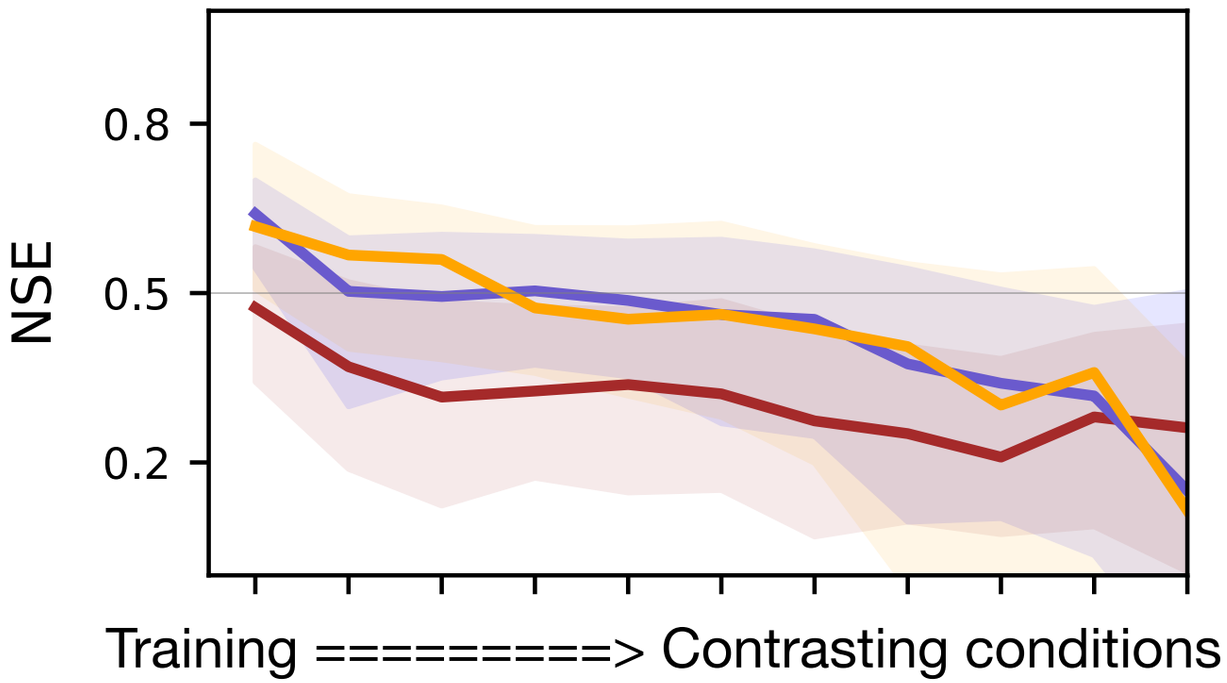
Are data-driven models robust under changing climatic conditions?



Training year (wet conditions)



Data-driven performance drops rapidly under changing conditions



- **Process-based models**

  Physics model

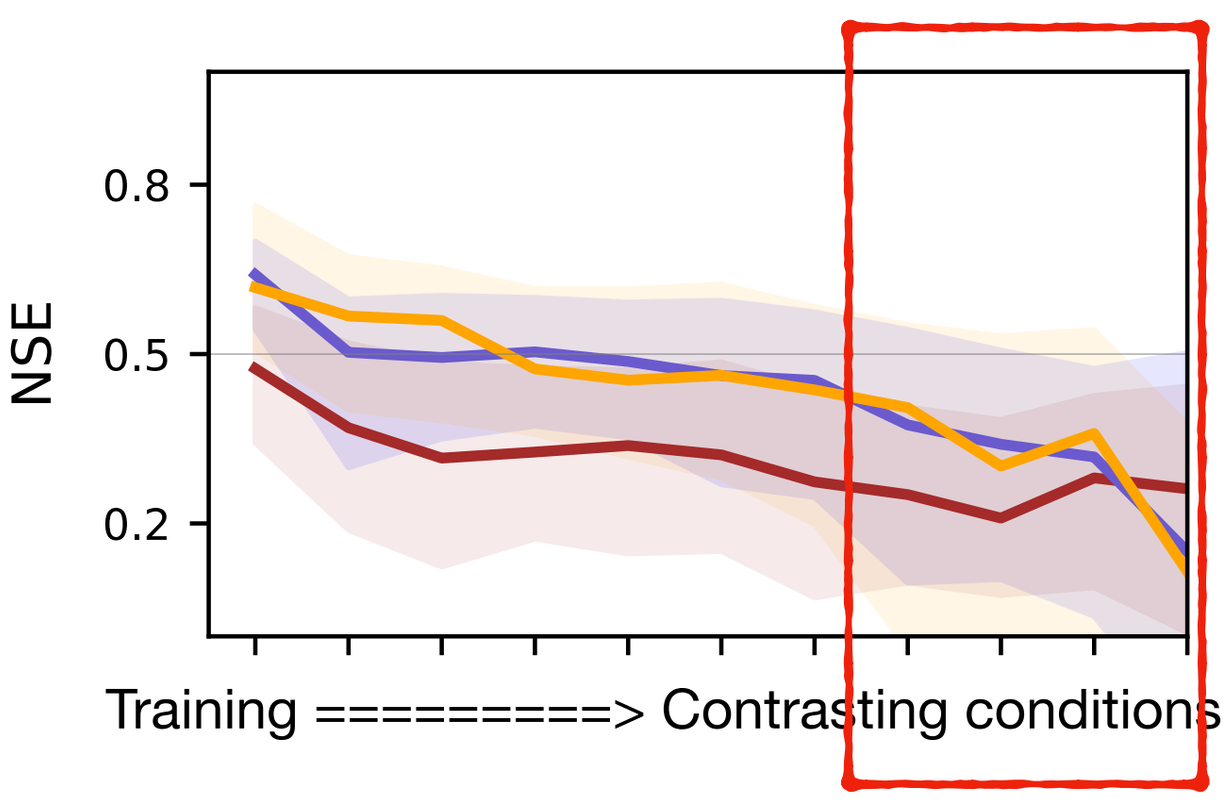
  Conceptual model

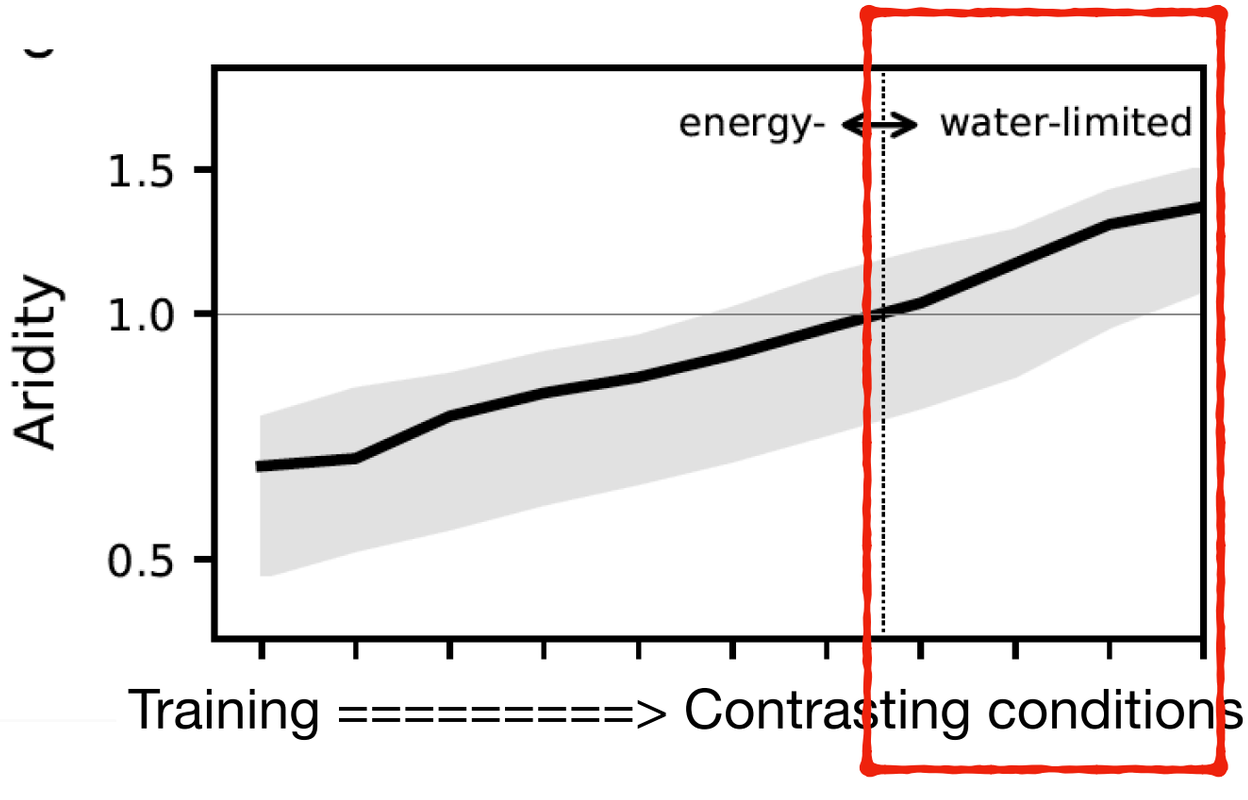
- **Data-driven models**

  DL model (LSTM)

O et al. (J Hydrometeorol, 2020)

Data-driven performance drops rapidly under changing conditions

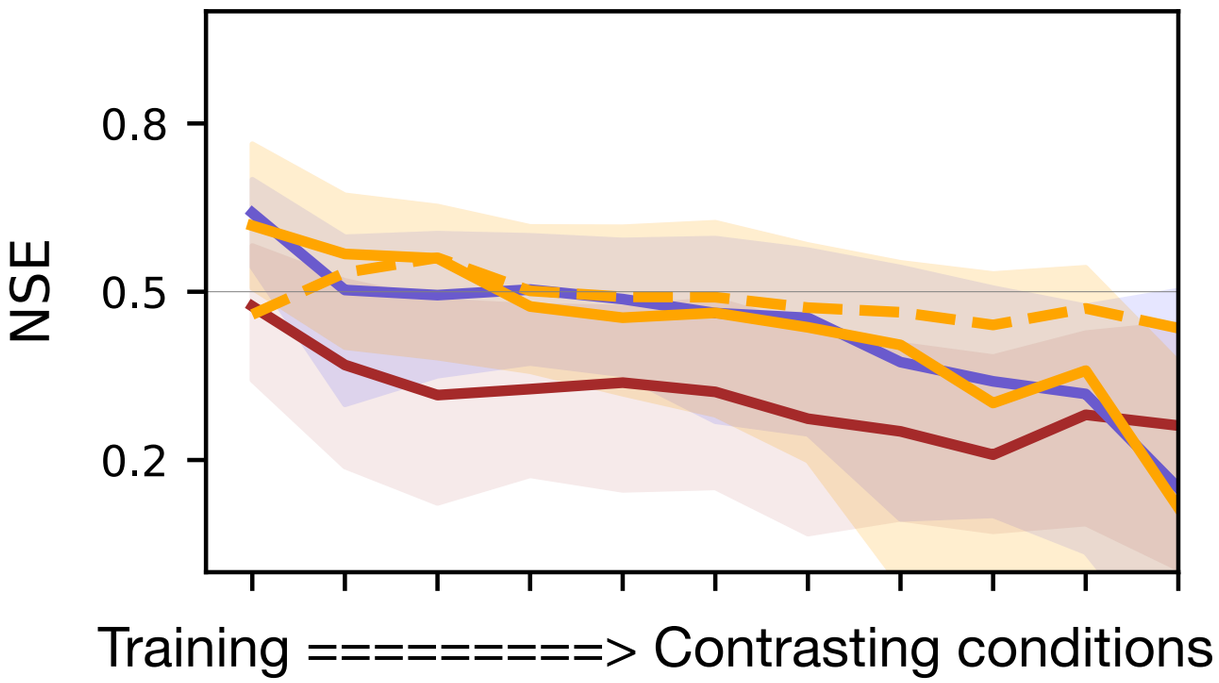




Shifts in hydro-climatic regime from energy-limited to water-limited

O et al. (J Hydrometeorol, 2020)

Diverse training conditions improve generalization



- **Process-based models**

  Physics model

  Conceptual model

- **Data-driven models**

  DL model (LSTM)

  DL model (LSTM*)

**trained on diverse climatic conditions*

O et al. (J Hydrometeorol, 2020)

Main points

[1] Data-driven skill can degrade rapidly under hydro-climatic shifts.

[2] Process-based structure remains valuable for robustness, motivating hybrid approaches.

[3] Diverse training conditions improve generalization, motivating benchmark-ready datasets.

From robustness to benchmark-ready soil moisture data

Benchmark data enables ML/DL development and comparison

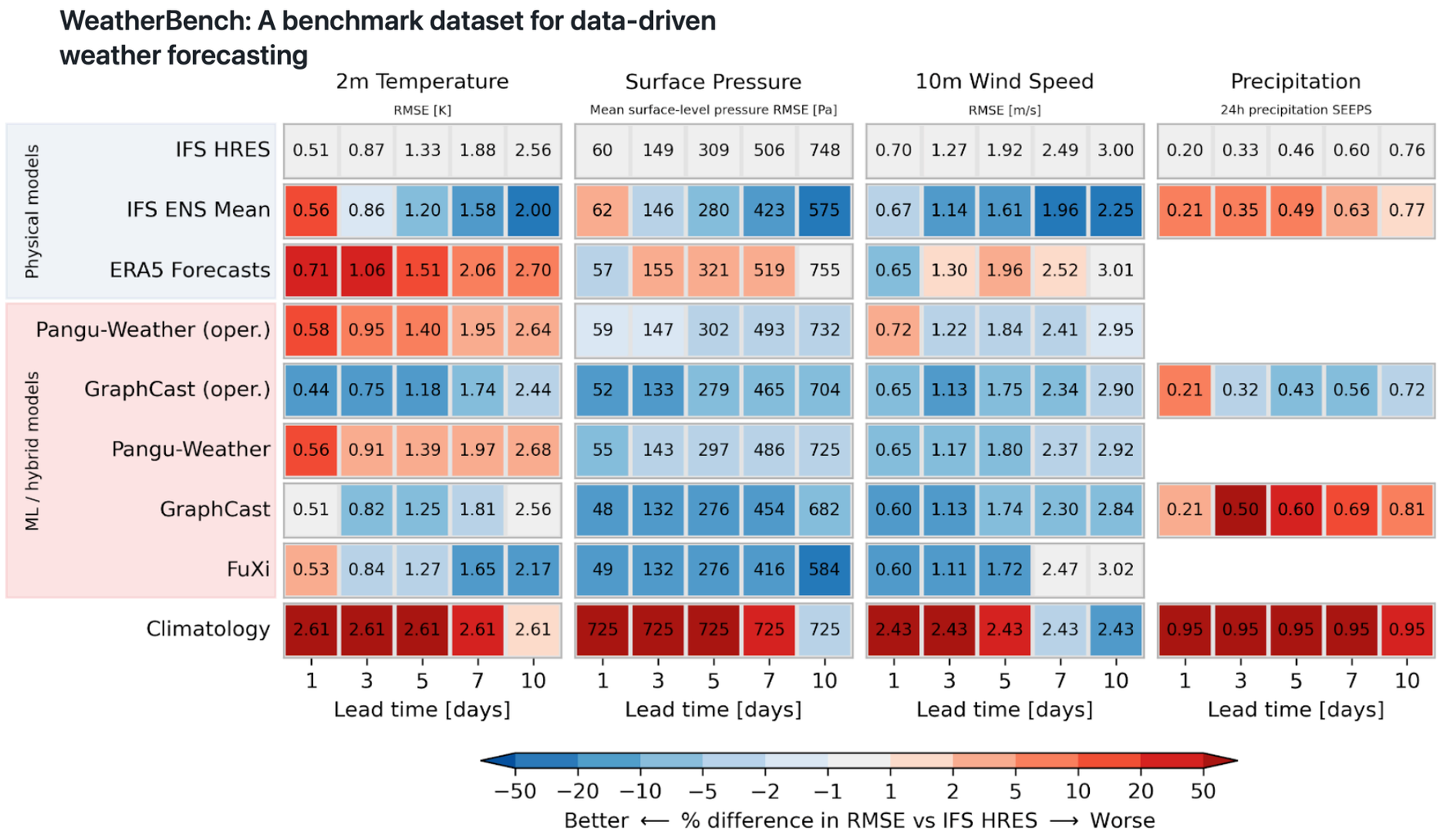
MNIST

ImageNet

https://git-disl.github.io/GTDLBench/datasets/mnist_datasets/
https://www.image-net.org/

Earth science is moving toward benchmark-ready datasets

WeatherBench



https://github.com/pangeo-data/WeatherBench

Earth science is moving toward benchmark-ready datasets

- WeatherBench (Rasp et al., 2020): ERA5
- ClimateBench (Watson-Parris, 2022): climate model emulation
- RainBench (Witt et al., 2021): ERA5+IMERG
- WaterBench (Demir et al., 2022): streamflow and precipitation measurements
- LandBench (Li et al., 2024): ERA5 land variables
- AQ-Bench (Betancourt et al., 2021): TOAR Ozone observation
- MAELSTROM (https://www.maelstrom-eurohpc.eu/): weather and climate
- Caravan (Kratzert et al., 2023): streamflow measurements

Benchmark-ready means more than open data

SoMoBench: an observation-based soil moisture benchmark

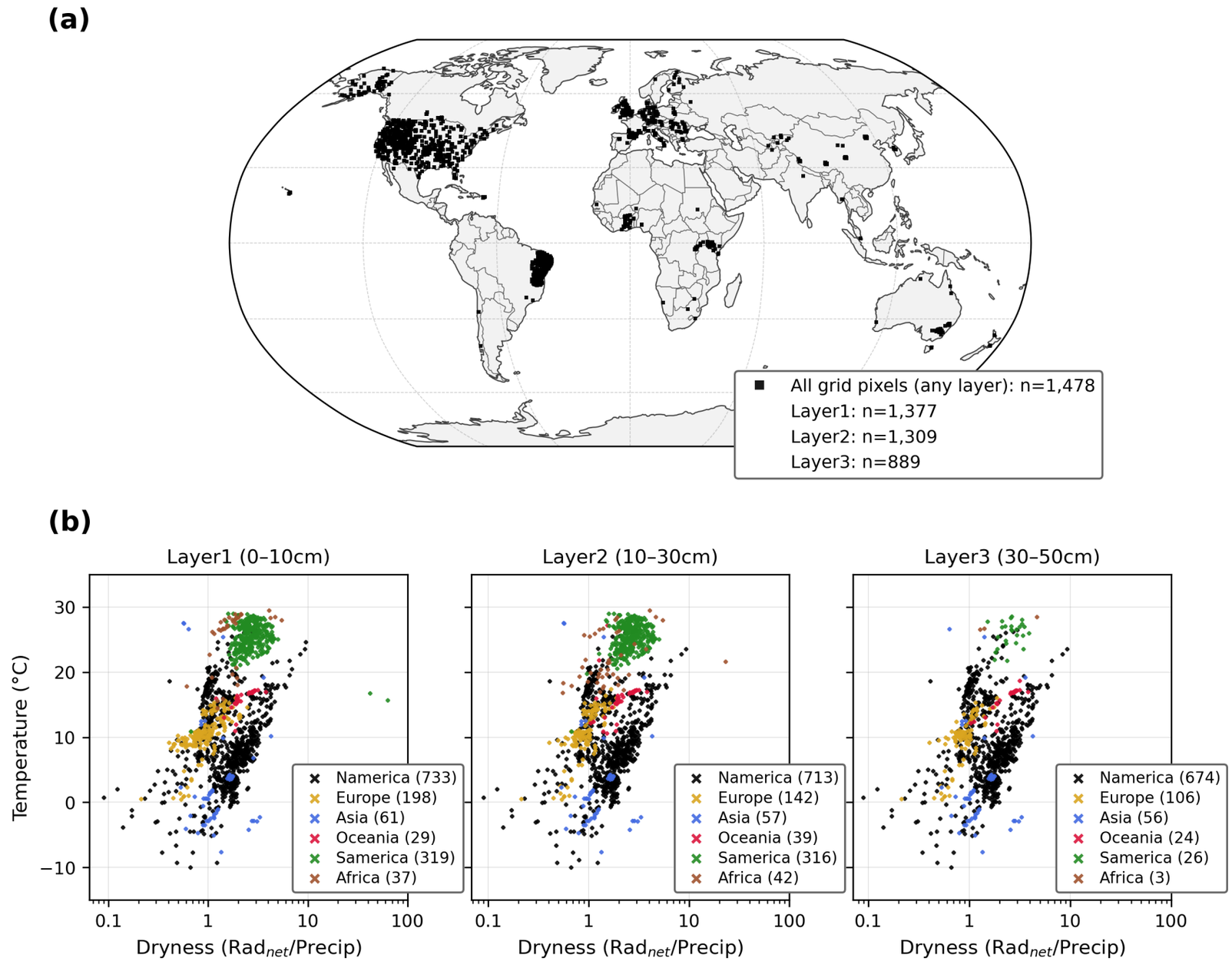
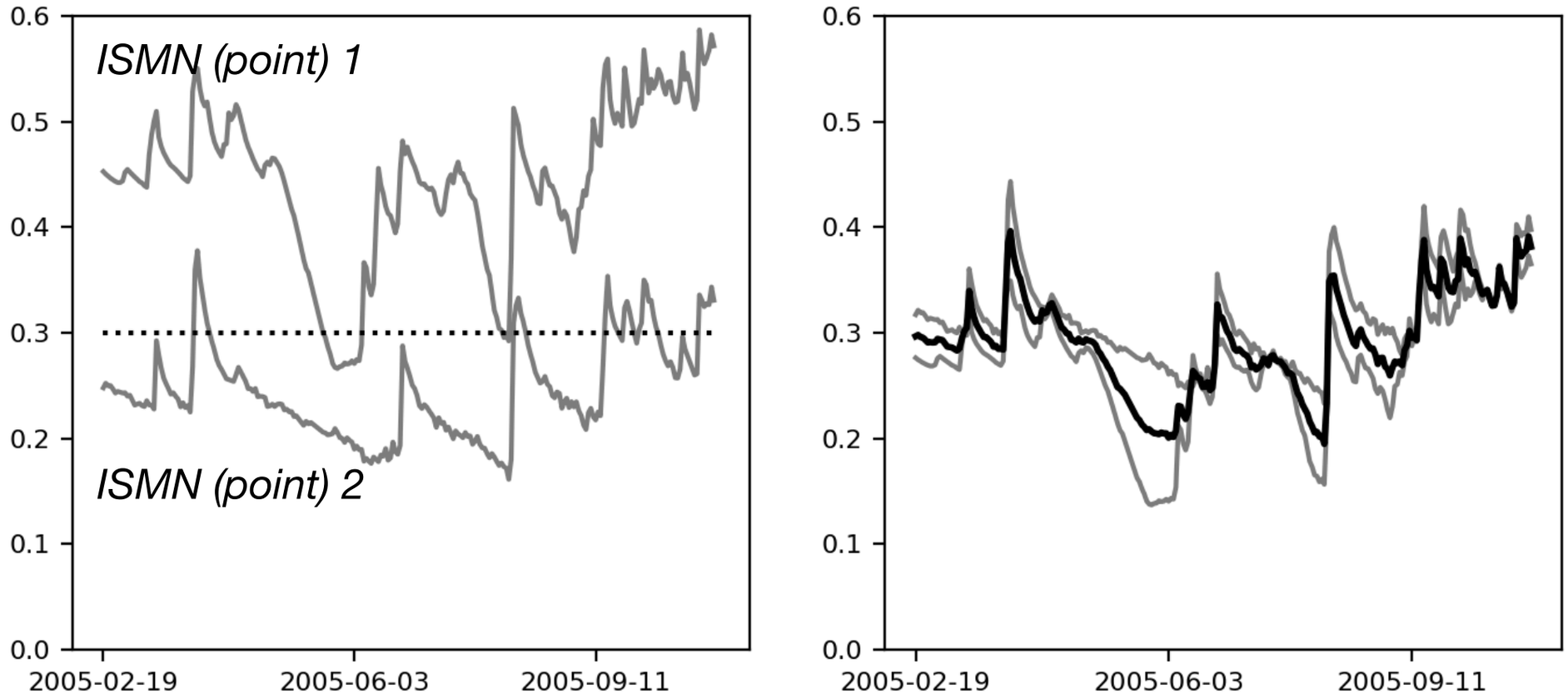


Fig. 2 Data availability

O et al. (*in preparation*, 2026)

SoMoBench links sparse observations to gridded soil moisture targets

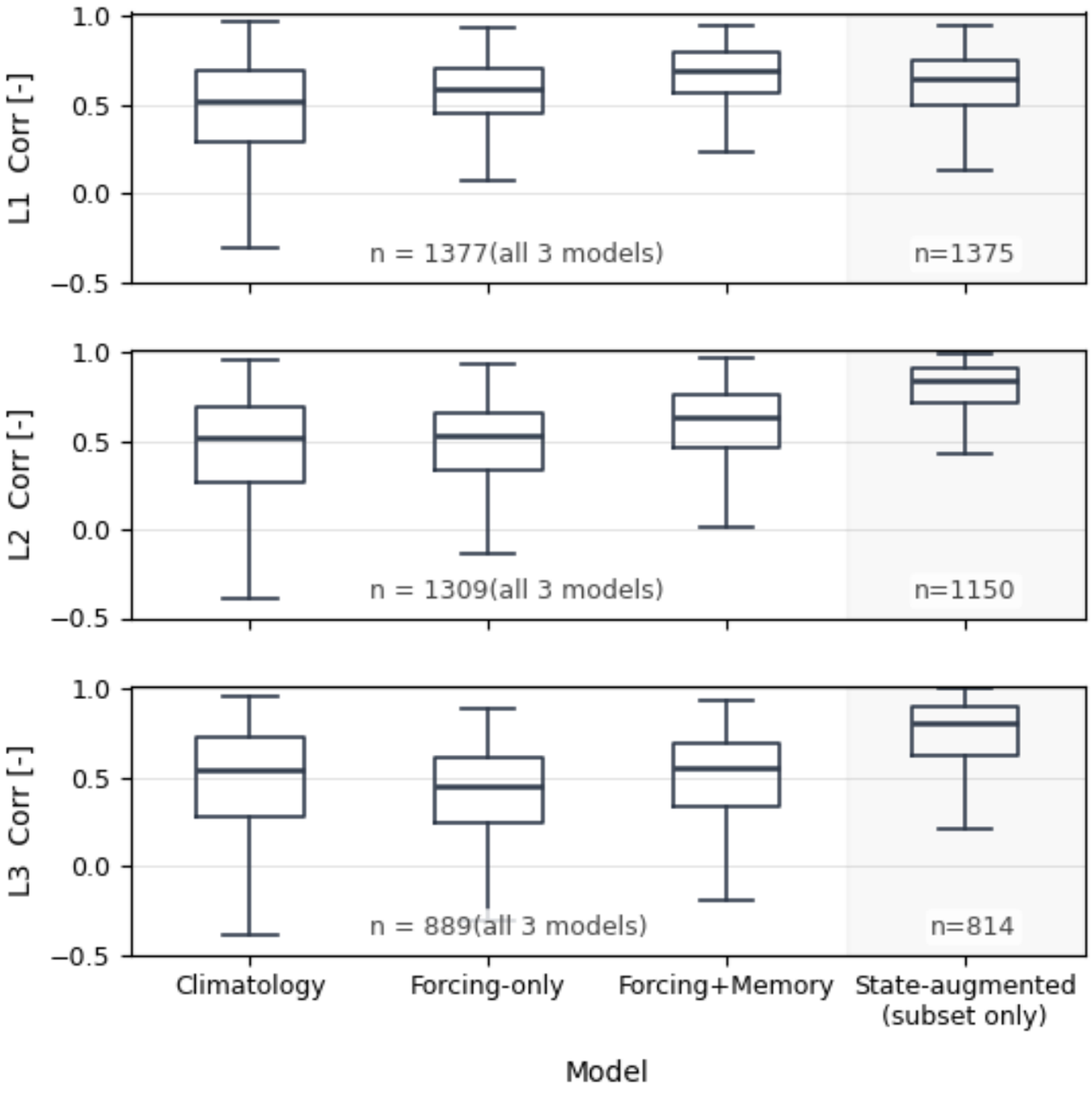


Long-term mean of ERA5 gridded soil moisture

Daily soil moisture data from ISMN point data

Daily SoMoBench gridded soil moisture

Benchmark value depends on both targets and forcing features



1) **Climatology**
2) **Forcing-only**
3) **Forcing + Memory**
   - forcing history (d-7 day)
4) **State-augmented**
   - skin temperature
   - upper-layer soil moisture

Fig. 6 consistency and usefulness of the forcing features.

O et al. (*in preparation*, 2026)

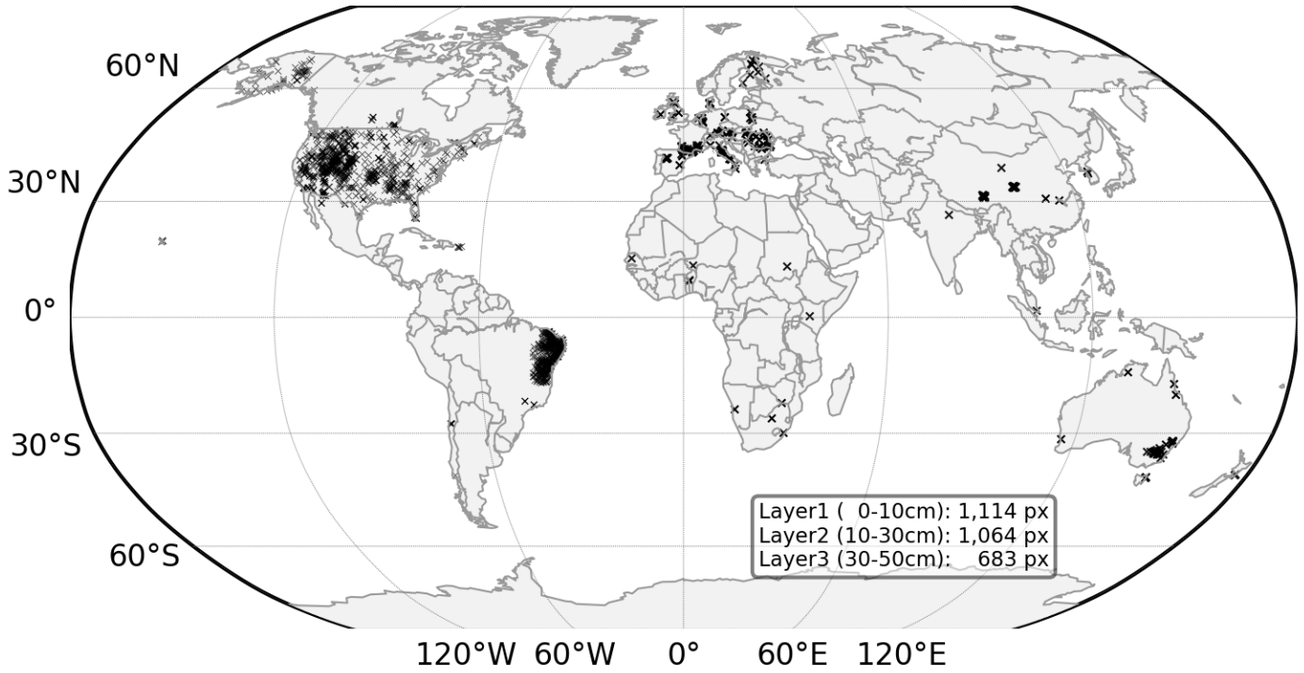
Main points

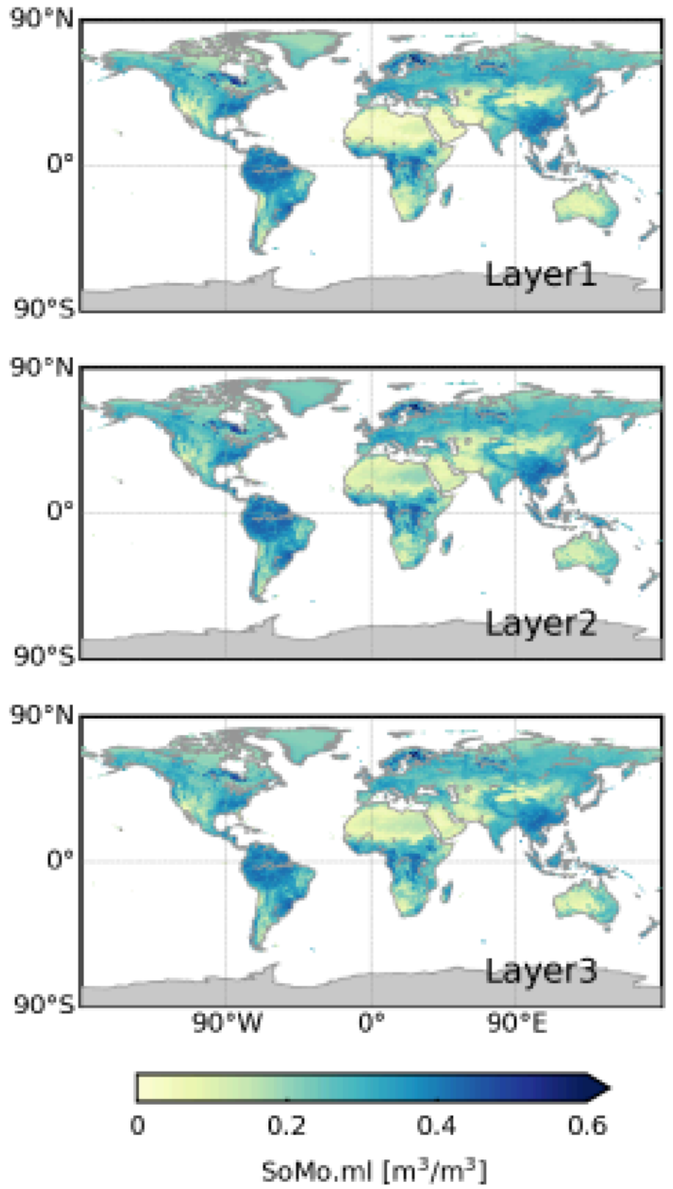
[1] SoMoBench provides observation-based, benchmark-ready soil moisture targets.

[2] It enables consistent comparison across models, inputs, and evaluation strategies.

[3] It supports testing generalization across hydro-climatic regimes, not only average skill.

Application 1
Global soil moisture products





Inputs
0.25°/daily

Long Short-Term Memory

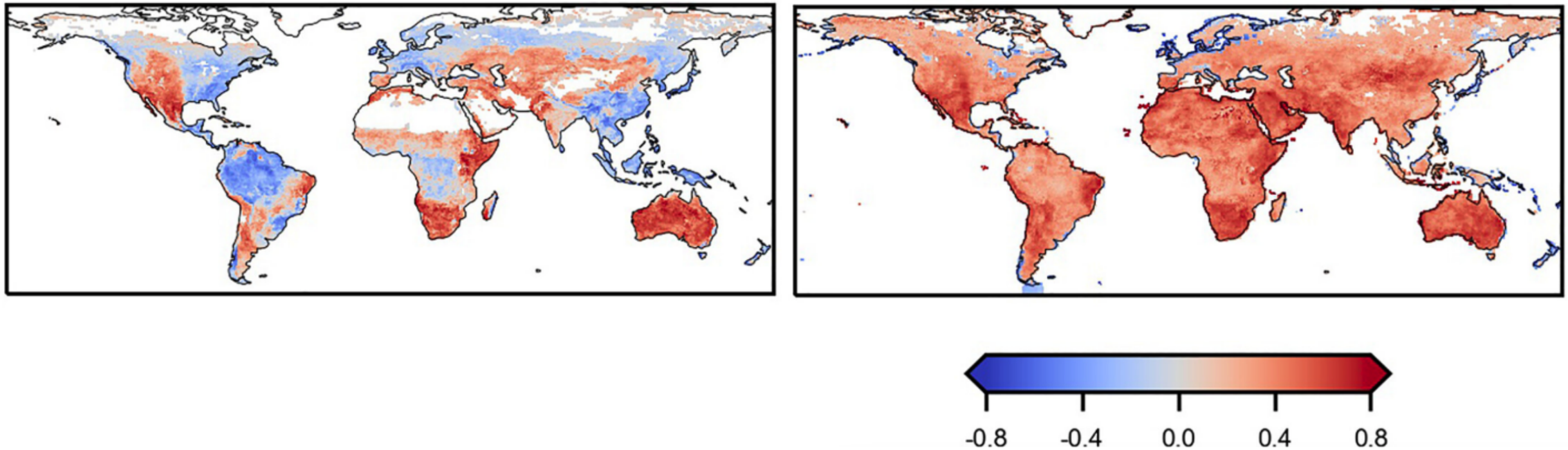
Target
0.25°/daily

O and Orth (Scientific Data, 2021), O et al. (Scientific Data, 2022)

Application 2
Drought–ET: Observations vs Physics

Soil Moisture-ET Corr.
(ML+Observation)

Soil Moisture-ET Corr.
(Physics model)



Li et al. (Earth's Future, 2023)

Main points

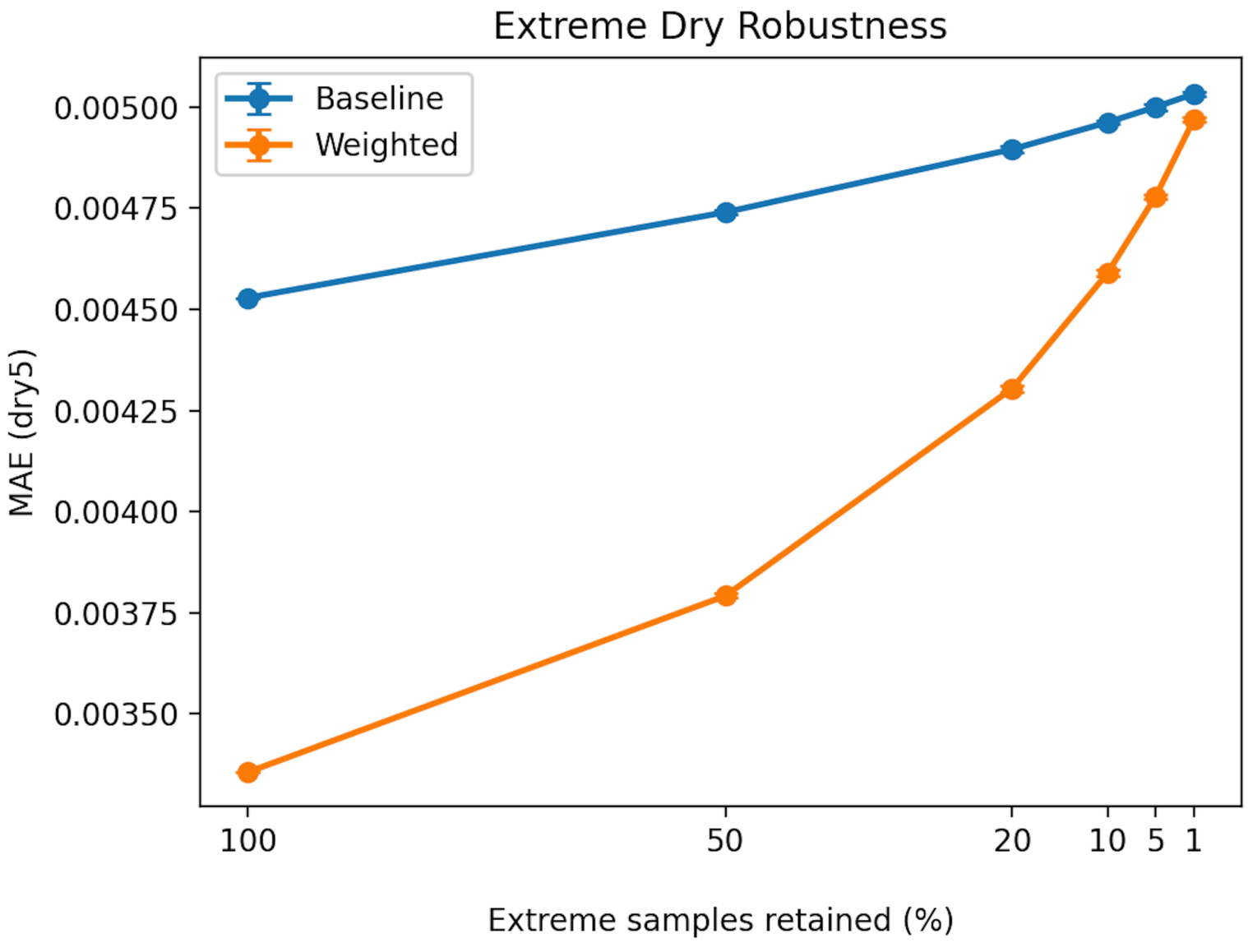
[1] Global benchmark datasets allow ML/DL models to be trained under more diverse conditions, which can improve model generalization.

[2] Observation-based datasets can serve as independent references to evaluate model behavior.

[3] Benchmark datasets therefore matter not only for predictive skill, but also for scientific insight.

Toward robustness benchmarks for ML4LM

Extreme-weighted training improves skill under dry extremes



Reflections on Benchmarking for ML4LM

[1] Benchmark datasets enable systematic evaluation of ML approaches.

[2] Observation-based benchmarks can provide independent constraints for model evaluation.

[3] Robustness under hydro-climatic shifts and extremes remains an important challenge.