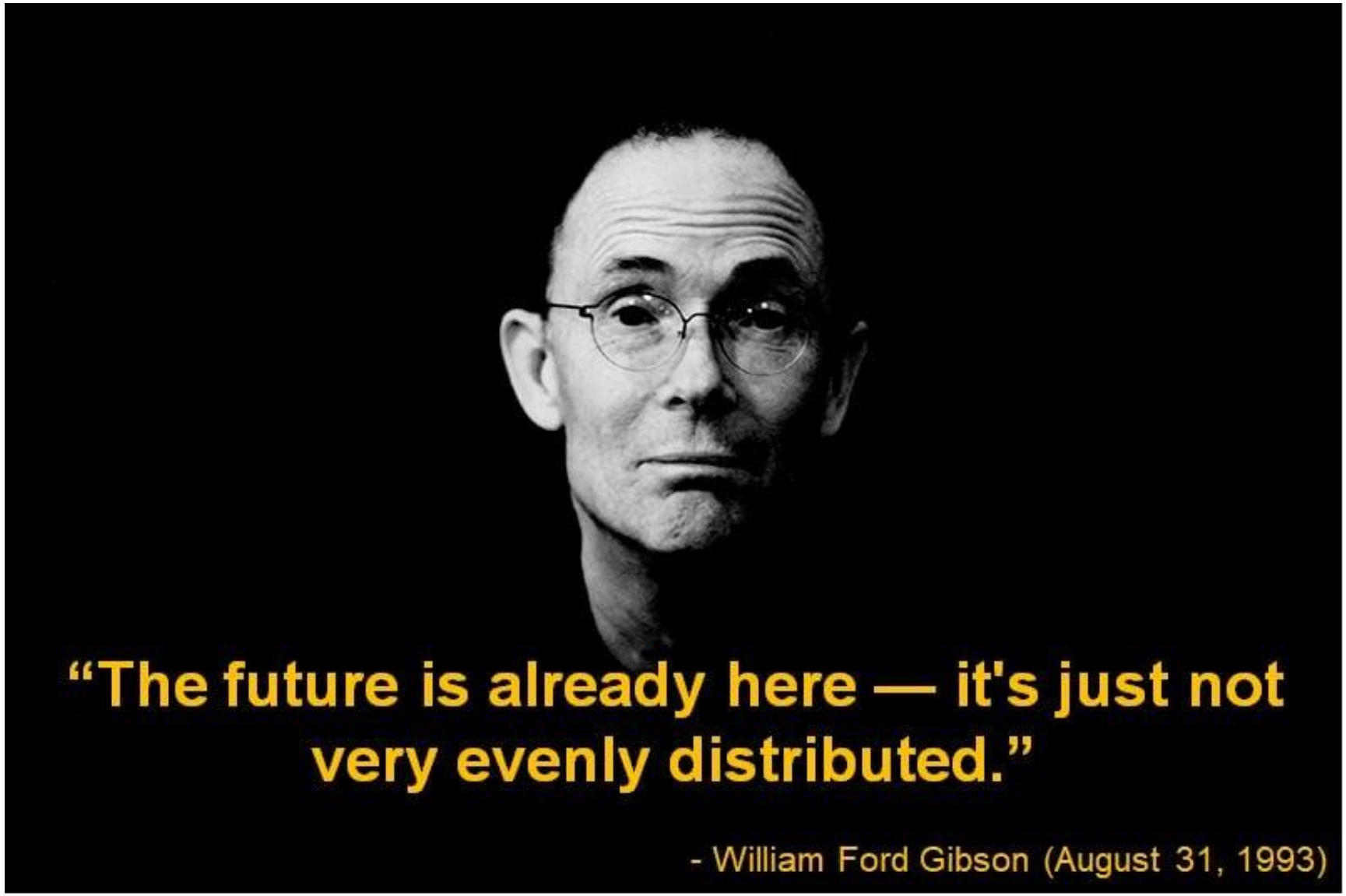
Data-driven Models for Predicting Precipitation

Chi-Jui Chen[1], **R. Saravanan**[1], Courtney Schumacher[1], Istvan Szunyogh[1], Raymond Wong[2]

[1]*Department of Atmospheric Sciences*
[2]*Department of Statistics*

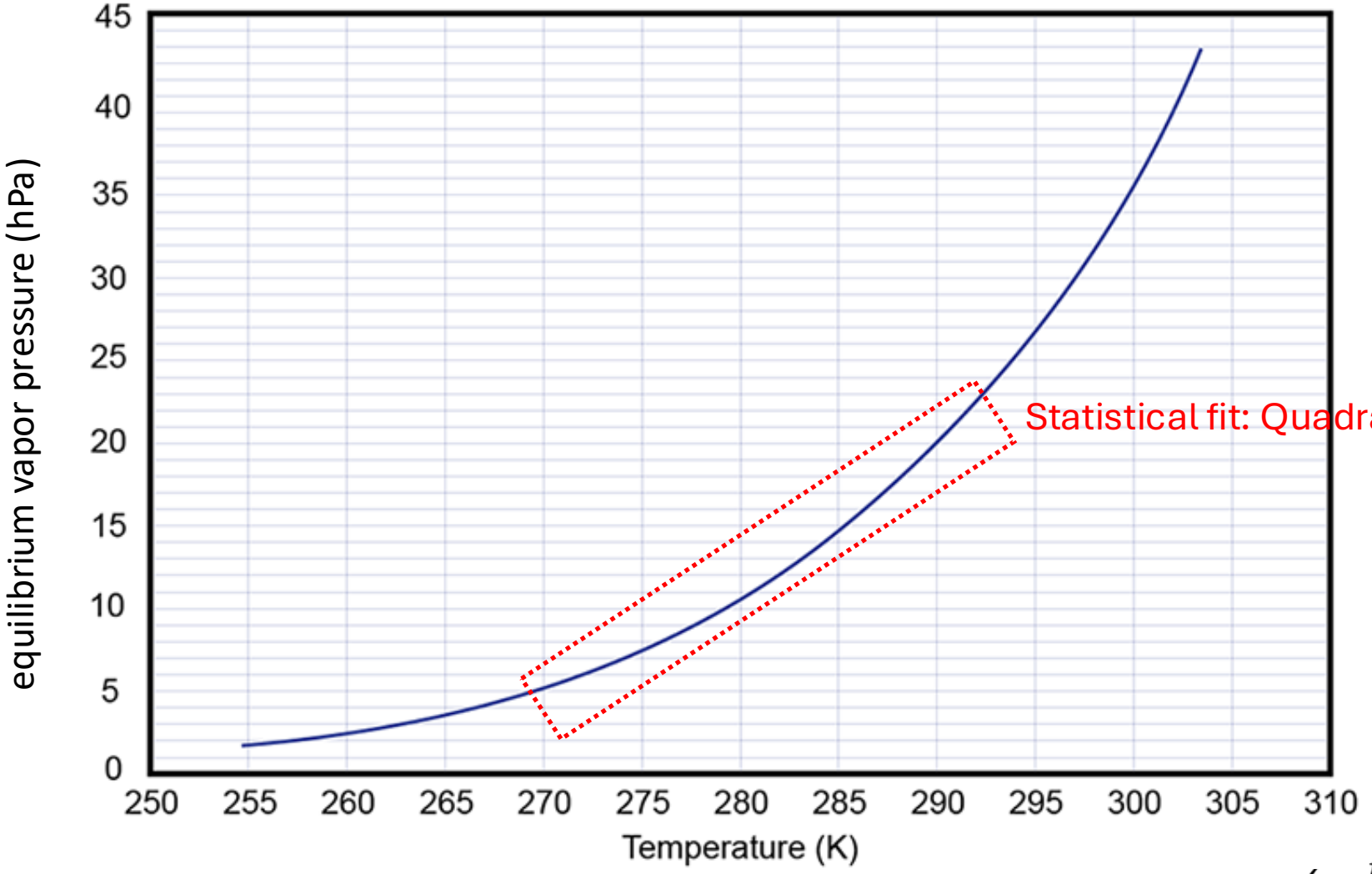
TEXAS A&M UNIVERSITY

The rationale for using ML to make predictions



“Future”: The thing we want to predict

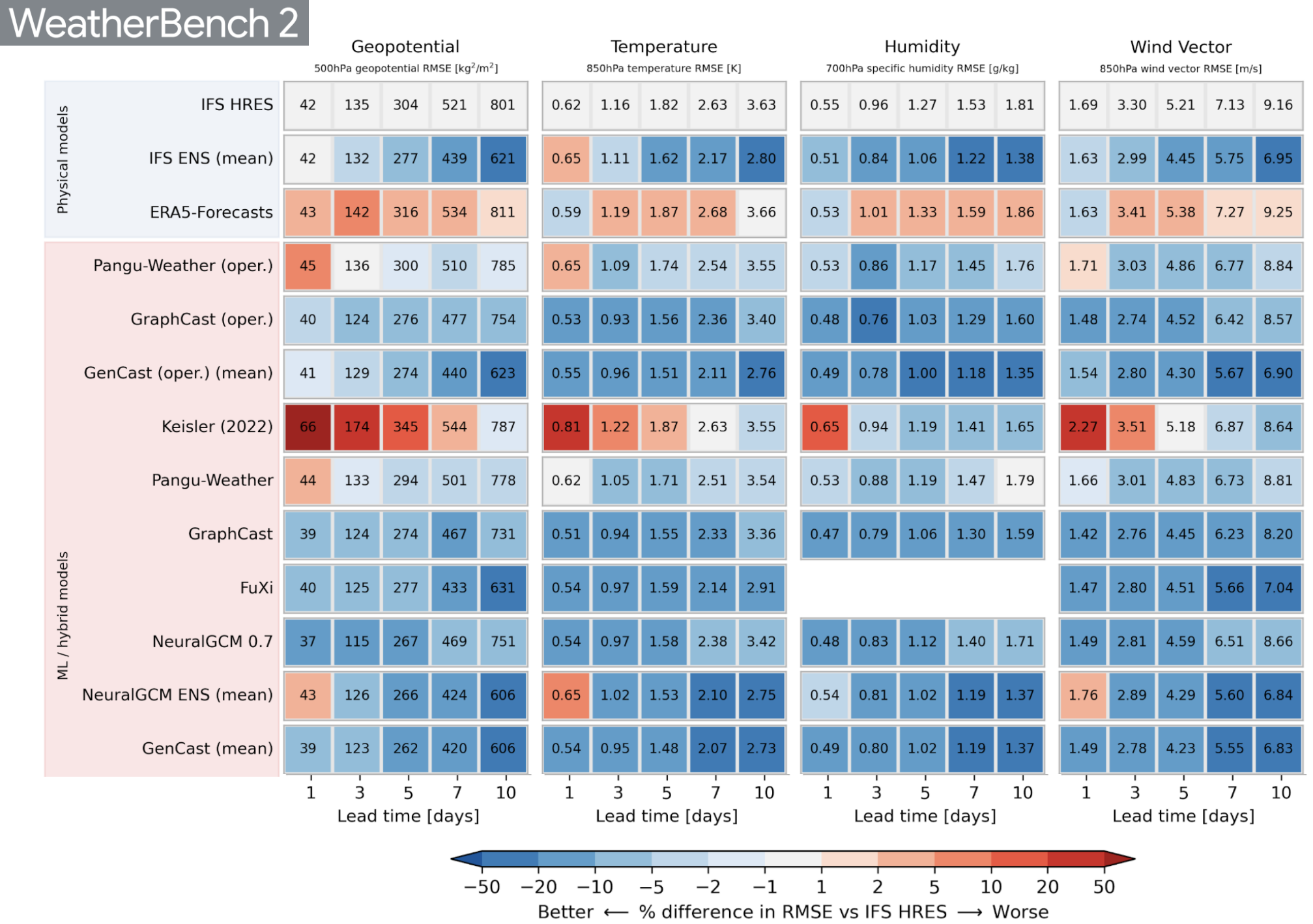
Can ML/AI “learn” physical equations from training data?



Clausius-Clapeyron Equation $e_s = e_{so} \exp\left(\frac{l_v}{R_v T_o}\right) \exp\left(\frac{-l_v}{R_v T}\right)$

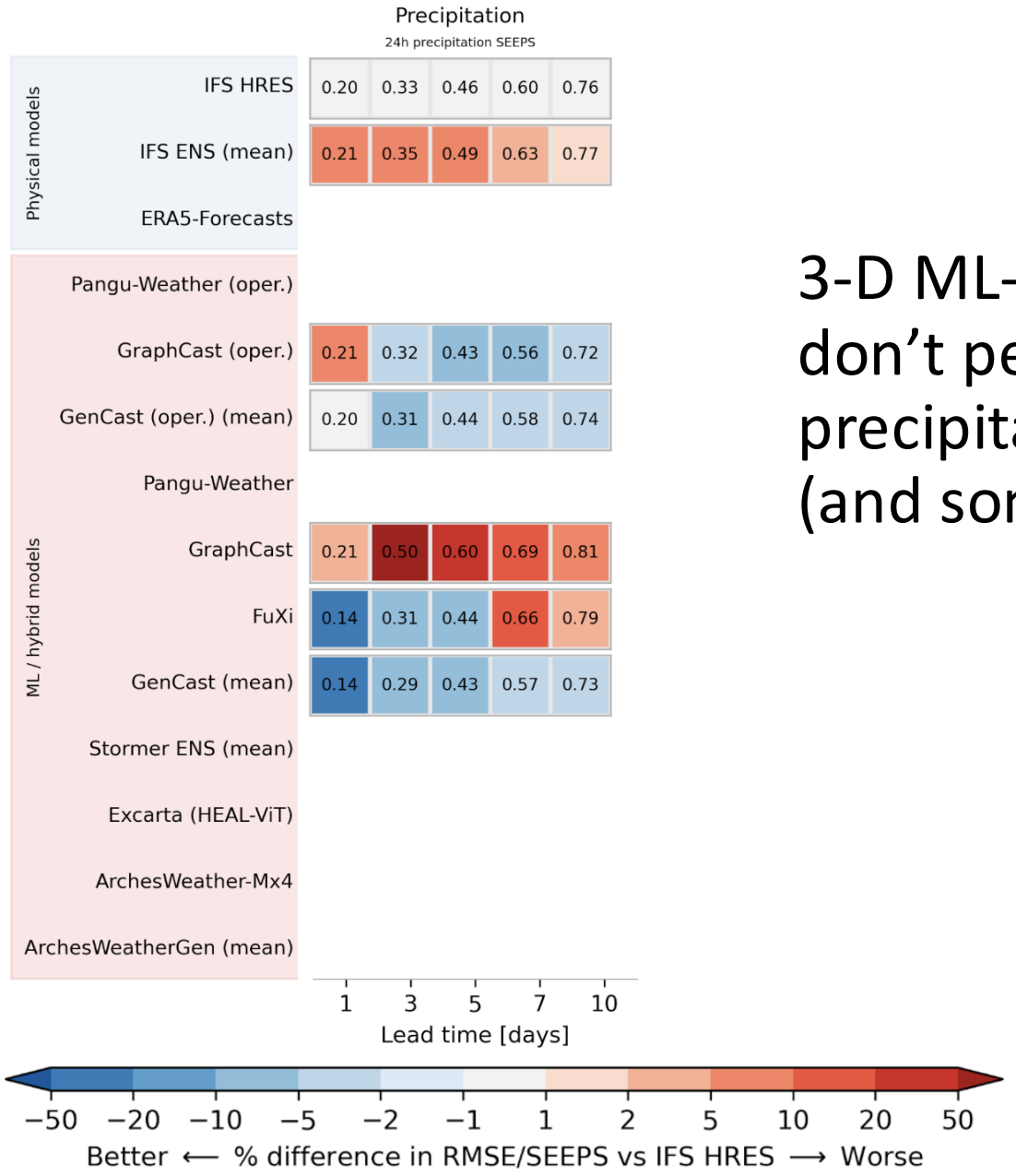
www.e-education.psu.edu/meteo300/node/584

T_o = 273 K, e_{so} = 6.11 hPa



3-D ML-based Weather Models predict some aspects of weather rather well!

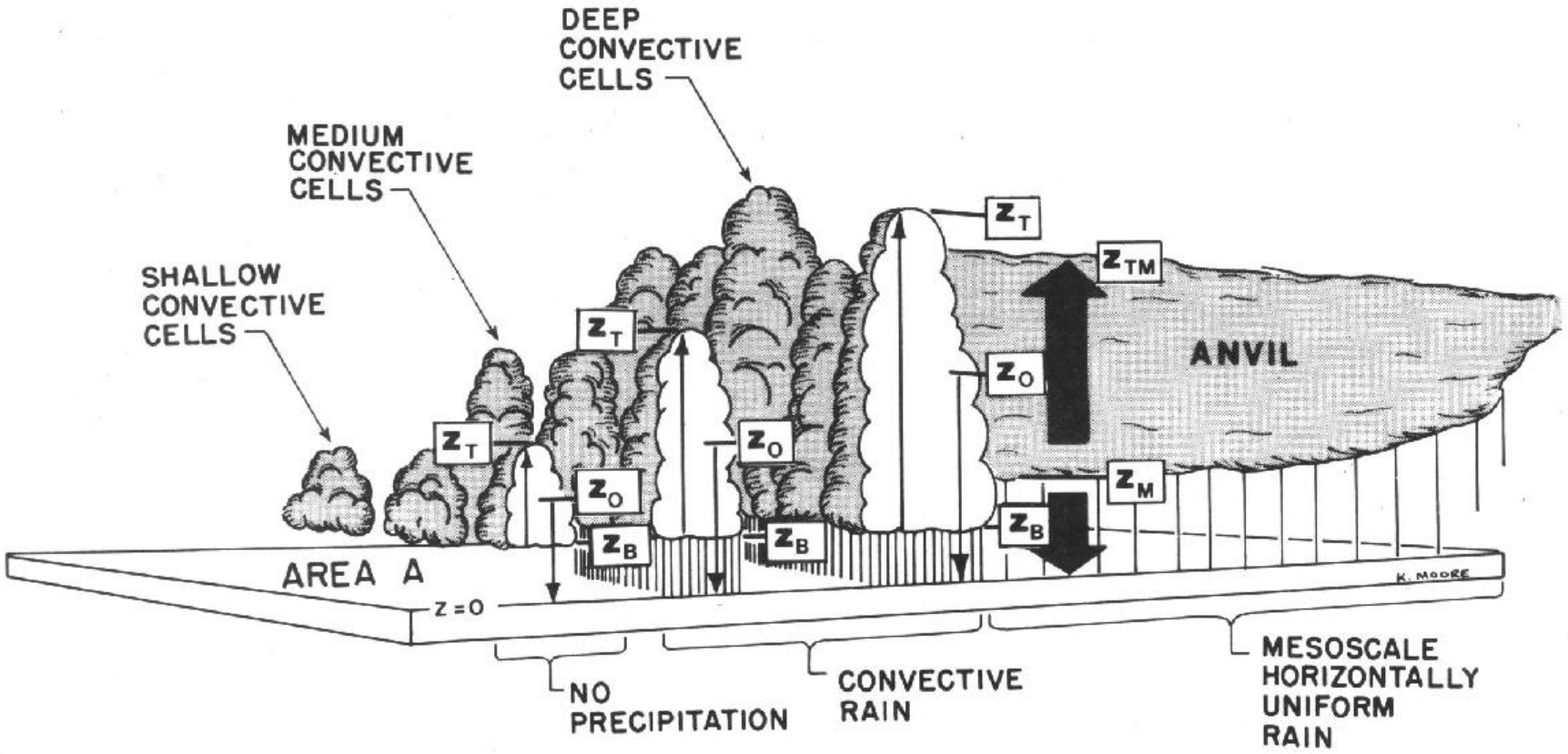
WeatherBench 2



3-D ML-based Weather Models don't perform that well in precipitation forecasting (and some don't even forecast it)*

*Rapidly evolving field

Post-GATE view of tropical cloud population

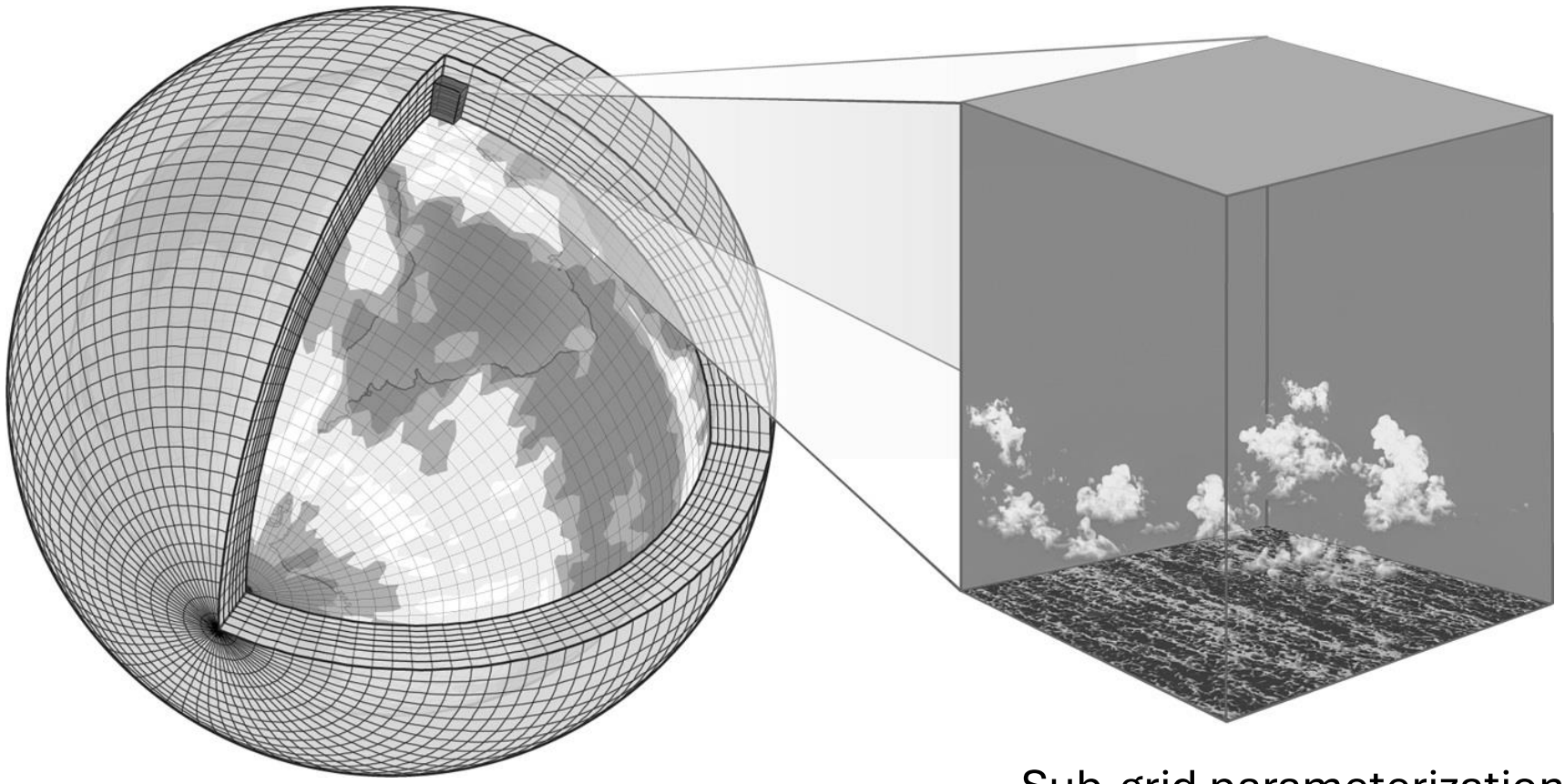


km-scale

Houze et al. (1980)

How do physics-based GCMs predict rain?

1-D Column subgrid parameterization



Model grid

Sub-grid parameterization

Schneider et al., 2017

Geophysical Research Letters / Volume 47, Issue 8 / e2020GL087232

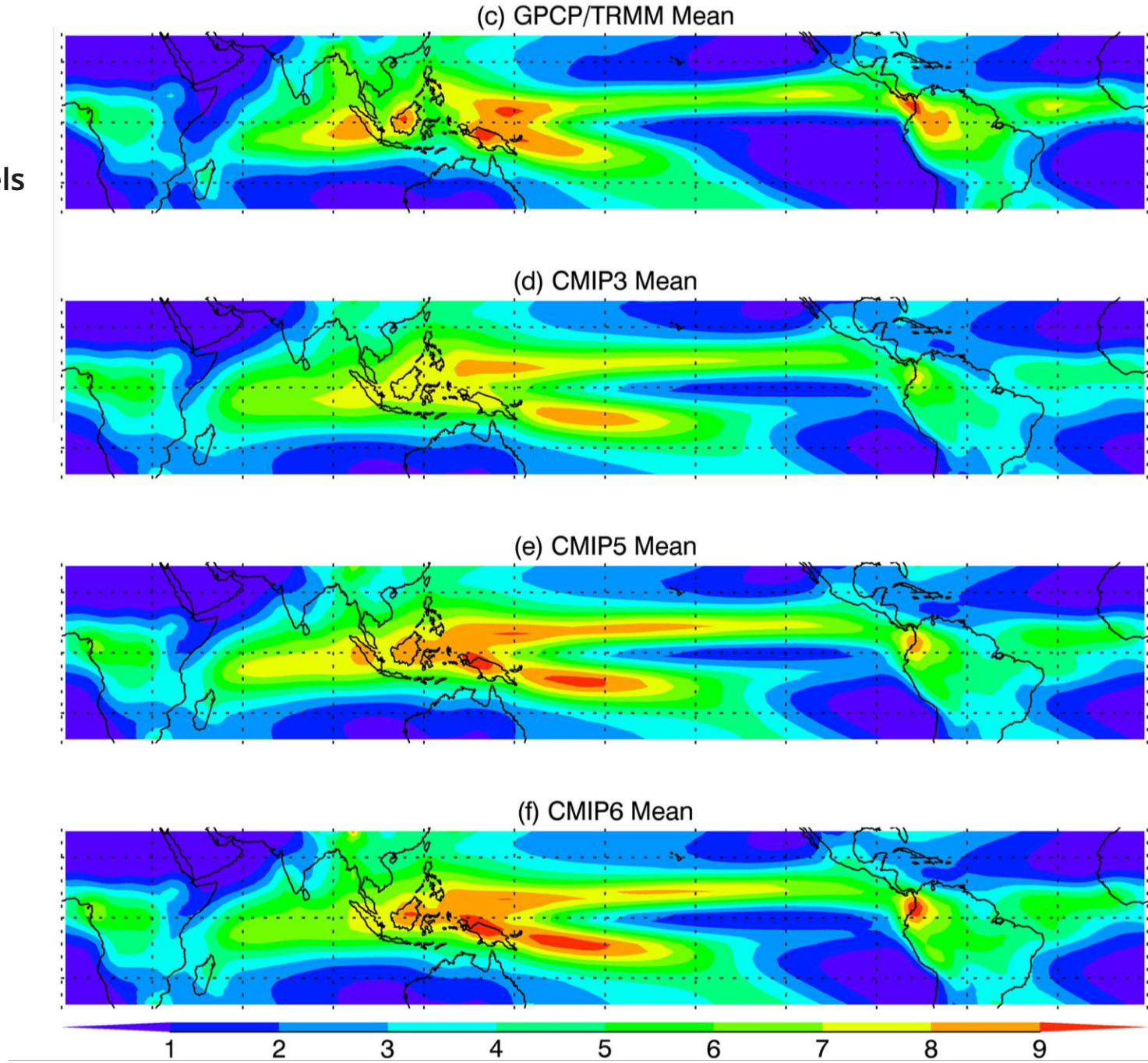
Research Letter | Free Access

The Double-ITCZ Bias in CMIP3, CMIP5, and CMIP6 Models Based on Annual Mean Precipitation

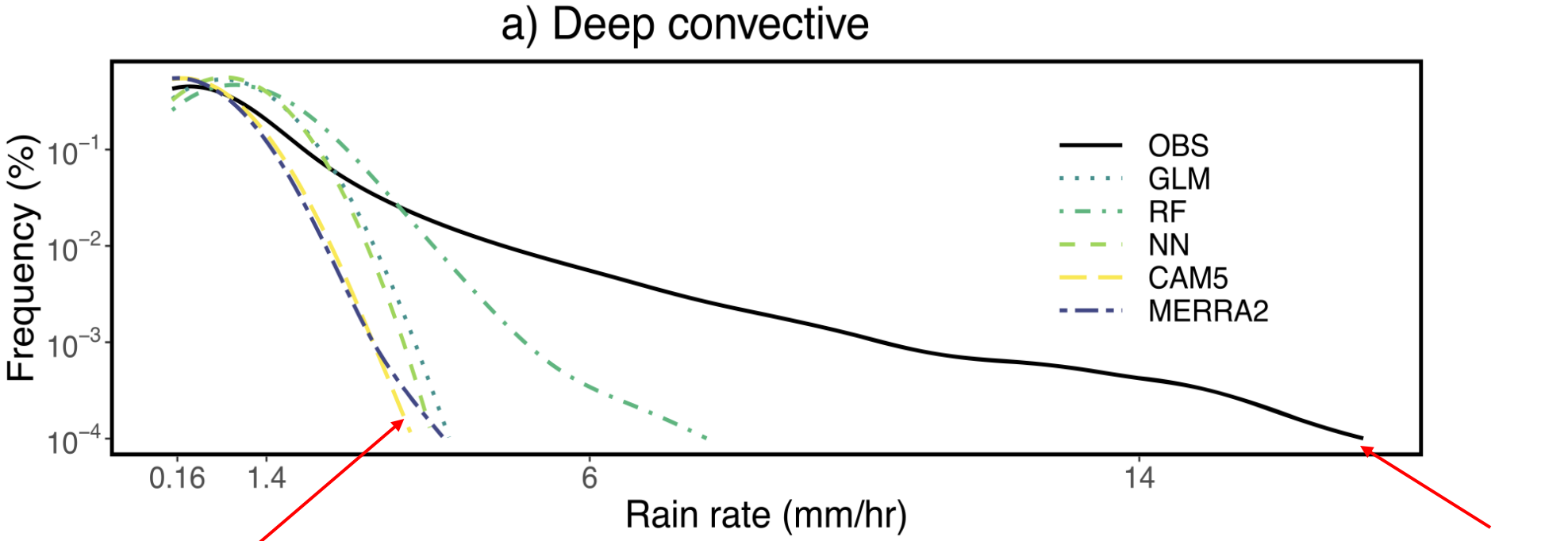
Baijun Tian ✉, Xinyu Dong

First published: 28 March 2020
https://doi.org/10.1029/2020GL087232

The annual double-ITCZ bias persists in CMIP3/5/6 models
(but is slightly reduced from CMIP3/5 to CMIP6)



Physics-based GCMs underestimate extreme precipitation (drizzle problem)



CAM5
(GCM)

Observation

Wang et al., 2021: Statistical and machine learning methods applied to the prediction of different tropical rainfall types. Environ. Res. Commun., 3, 111001

1-D data-driven models of precip

- Why focus on precip?
  - Physics-based models still have trouble simulating mean & extreme precip
  - Lots of data available to train statistical/ML models
    (with good global hi-resolution spatial and temporal coverage)

- Why use 1-D data-driven approach?
  - 3-D ML weather models also don't predict precip that well*
  - Affordable (especially in an academic setting)
  - Simple and physically motivated; analogous to GCM parameterization
    - no "action at a distance"
    - translation invariance
  - Generalizable for climate prediction applications?

*Rapidly evolving field

Predicting rainfall using a Statistical Column Parameterization

Predictive Statistical Representations of Observed and Simulated Rainfall Using Generalized Linear Models
J. Yang, M. Jun, C. Schumacher, and R. Saravanan, *Journal of Climate*, 2019

Predict TRMM/GPM rainfall using reanalysis flow fields and a Generalized Linear Model:

- Logistic Regression for occurrence of rainfall
- Gamma Regression for rainfall intensity

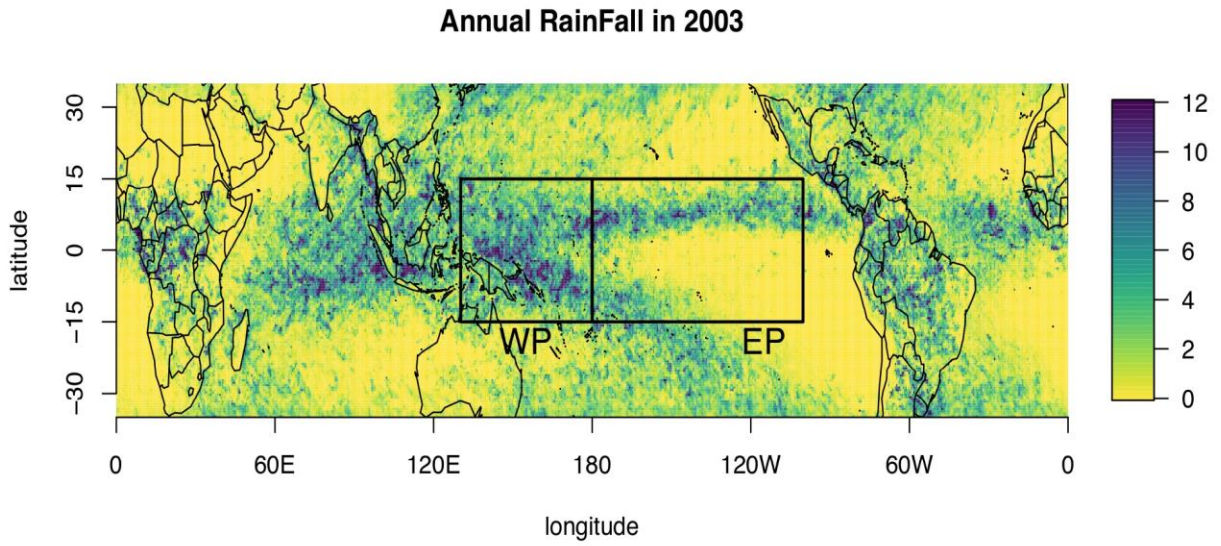


FIG. 1: *Mean annual total precipitation in 2003 from TRMM PR data. The unit is mm/day.*

Observed and predicted rain rates (E. Pacfic)

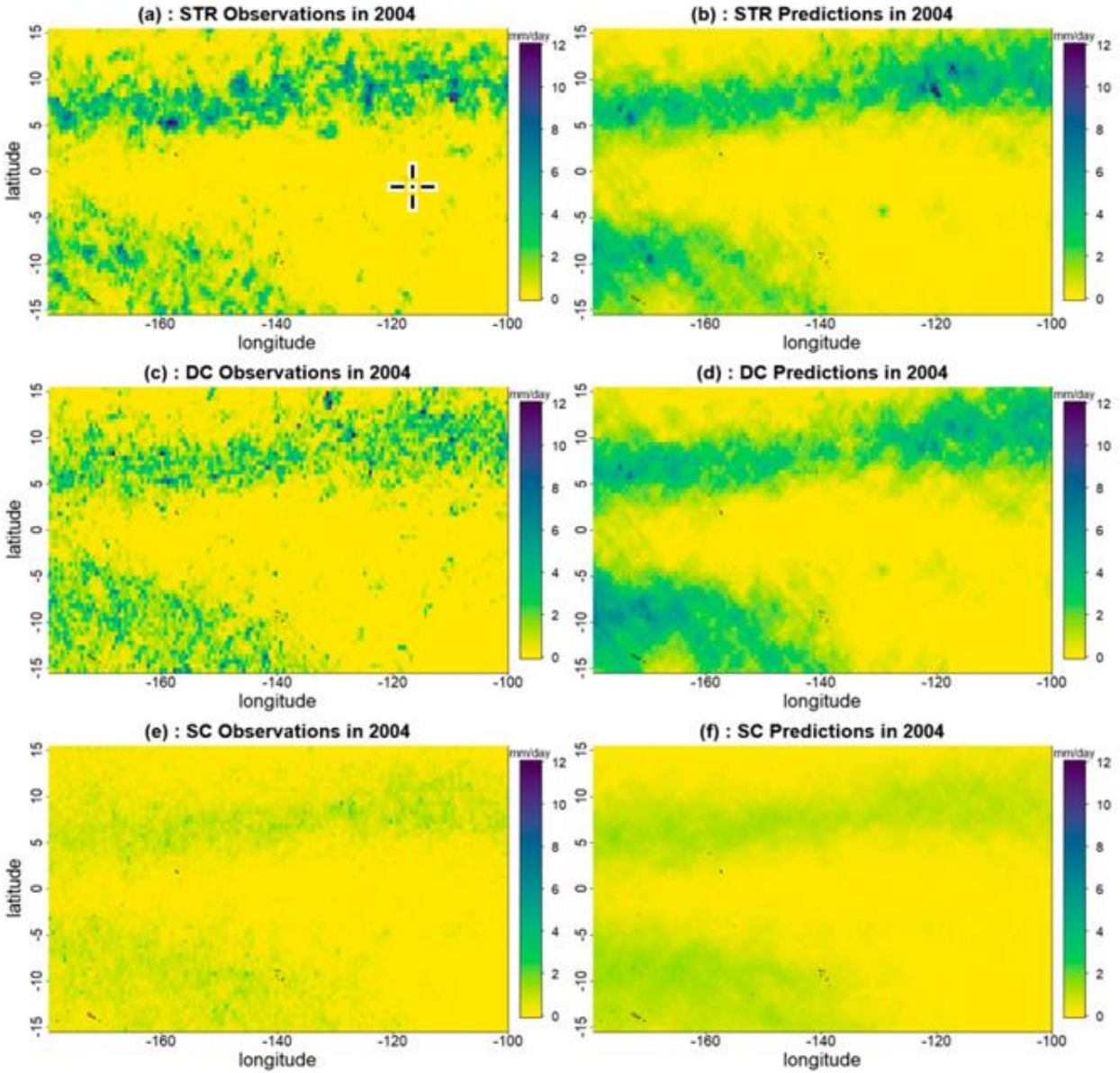
OBS

PRED (GLM)

Stratiform

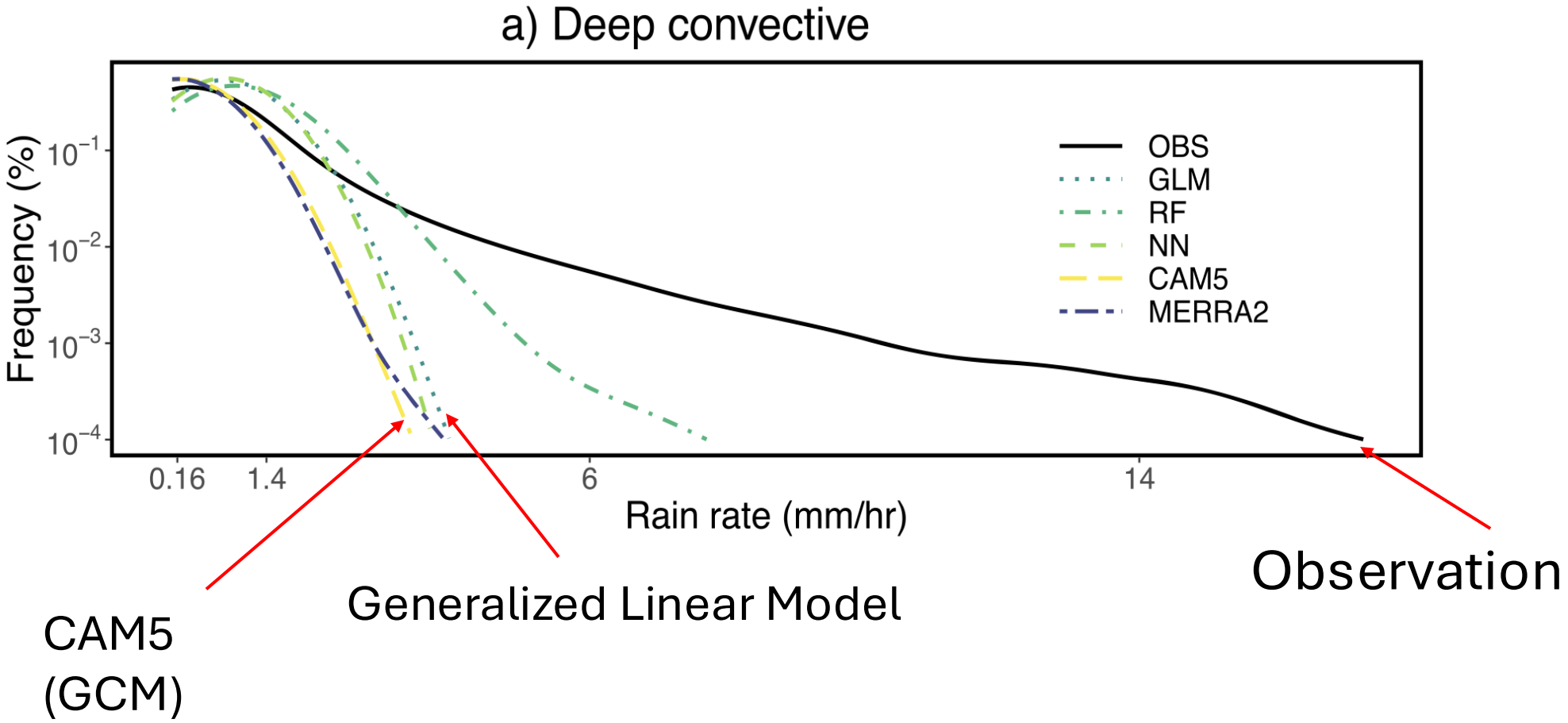
Deep convective

Shallow convective



Yang et al. (2019)

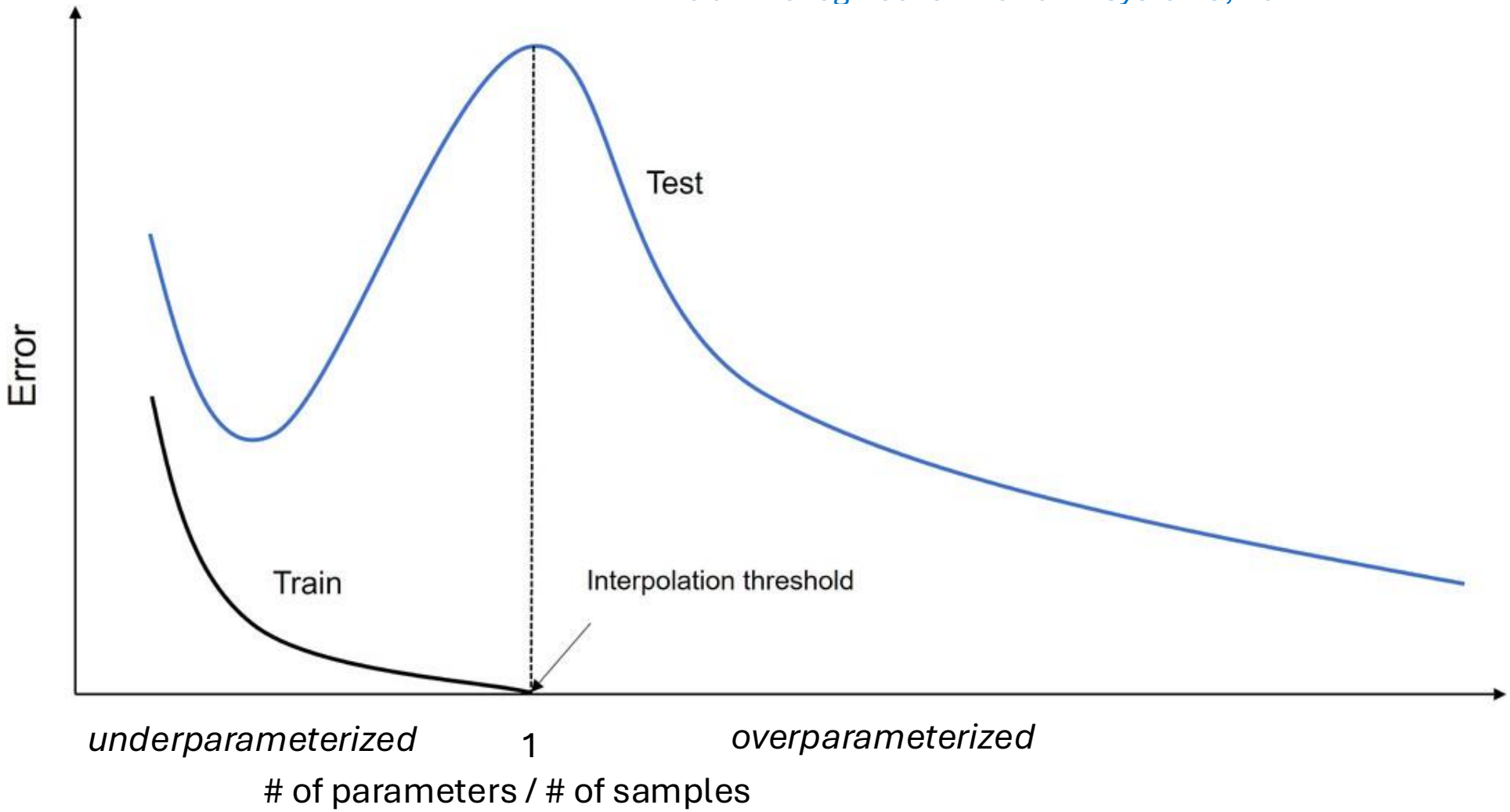
Generalized Linear Model still underestimates extreme precipitation



Wang et al., 2021: Statistical and machine learning methods applied to the prediction of different tropical rainfall types. Environ. Res. Commun., 3, 111001

Prediction of Tropical Pacific Rain Rates with Overparameterized Neural Networks

H. You, J. Wang, R.K.W. Wong, C. Schumacher, R. Saravanan and M. Jun, *Artifiicial Intellegince for the Earth Systems*, 2024



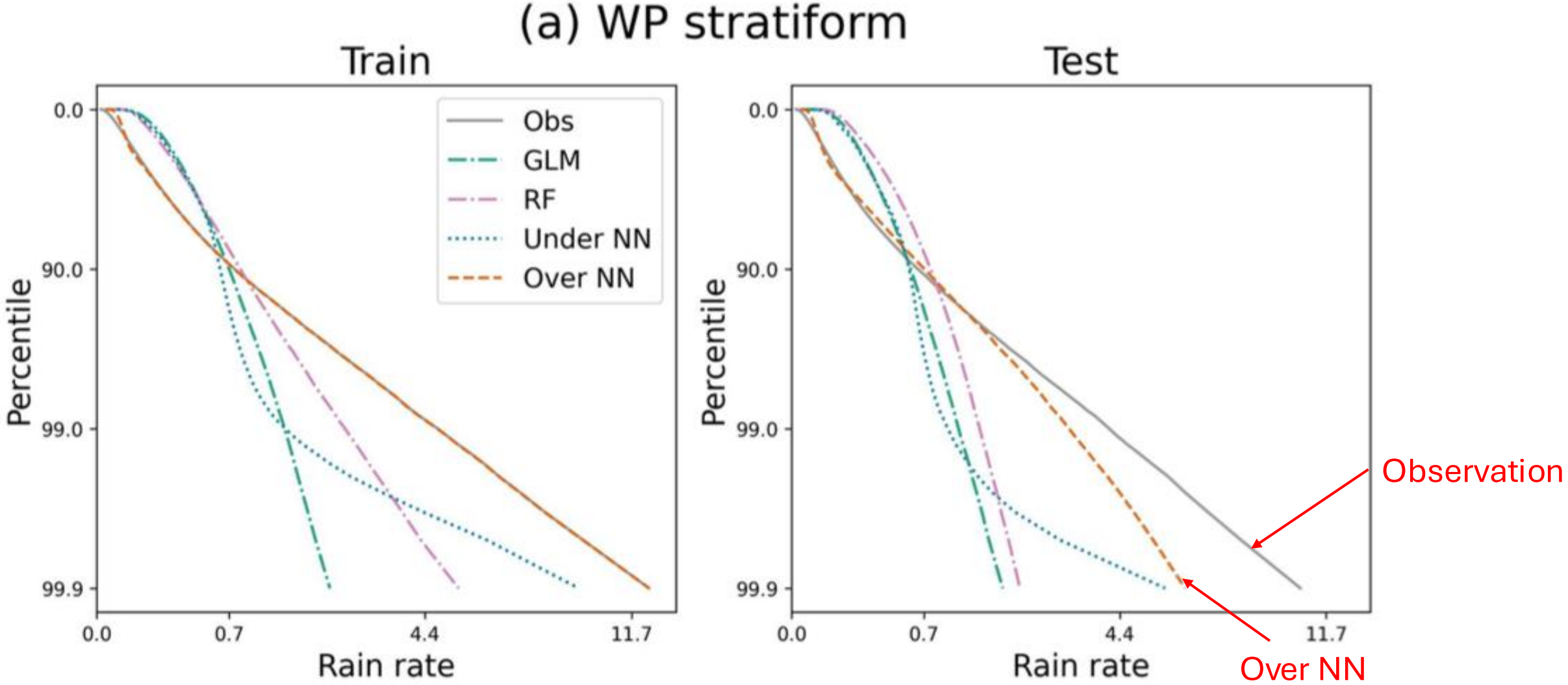
Overparameterized Neural Networks can improve precipitation forecasts (over GLM)

Rain rate forecast: Median Absolute Error (mm/hr)

	Region	Rain type	GLM	RF	Under NN	Over NN
West Pacific	WP	Stratiform	0.154	0.211	0.155	**0.094**
		Deep convective	0.294	0.421	0.296	**0.100**
		Shallow convective	0.031	0.036	**0.030**	0.036
East Pacific	EP	Stratiform	0.174	0.241	0.150	**0.103**
		Deep Convective	0.301	0.416	0.279	**0.117**
		Shallow Convective	0.047	0.058	**0.044**	0.055

You et al. (2024)

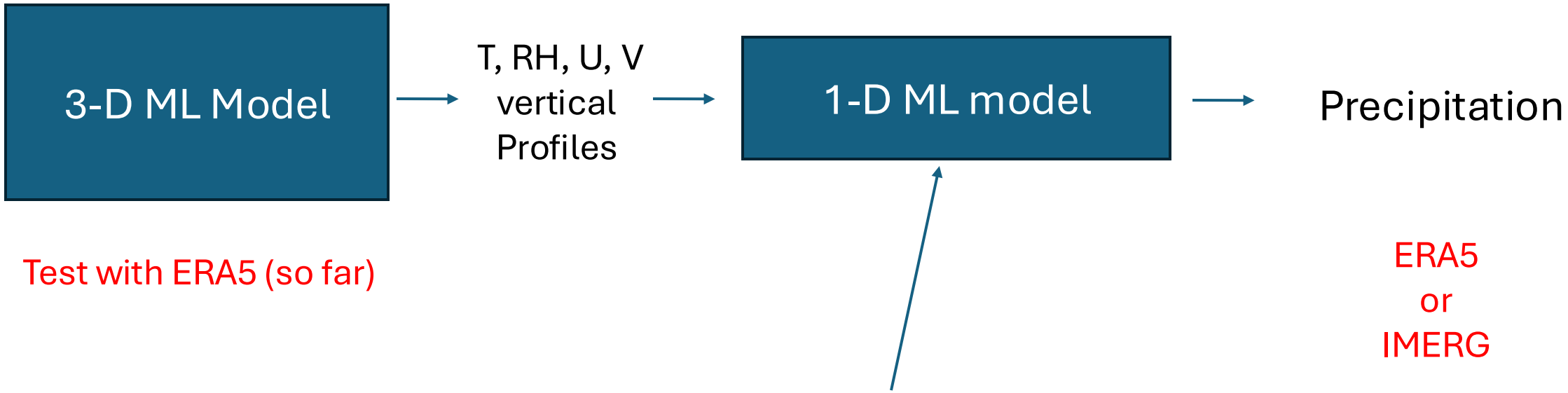
Overparameterized Neural Networks also do better at simulating extreme precipitation



You et al., 2024

TEXAS A&M

2-step approach to precip prediction



Train a 1-D column model using an overparameterized neural network

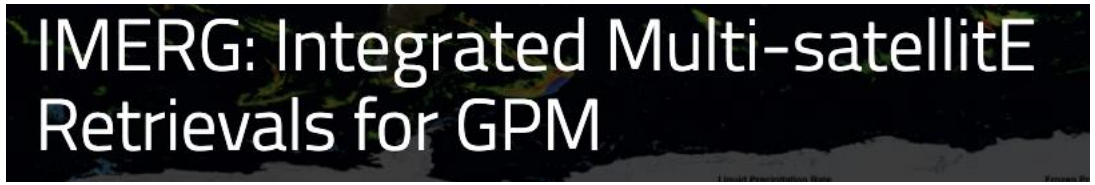
Data-driven 1-D Column Parameterization

- Data: 0.25° ERA5 2015-2018 as training set and 2019-2022 as test set
  - Jun-Jul-Aug only
- Region: West Pacific (130E – 180E, 15S – 15N)
- **Input variables**
  1. 13 levels: Temperature, Wind, Relative Humidity, Divergence, Vorticity
  2. Elevation
  3. Optional: Lat, Lon, Local Hour
- **Output variable**: 1-hourly accumulated Precipitation

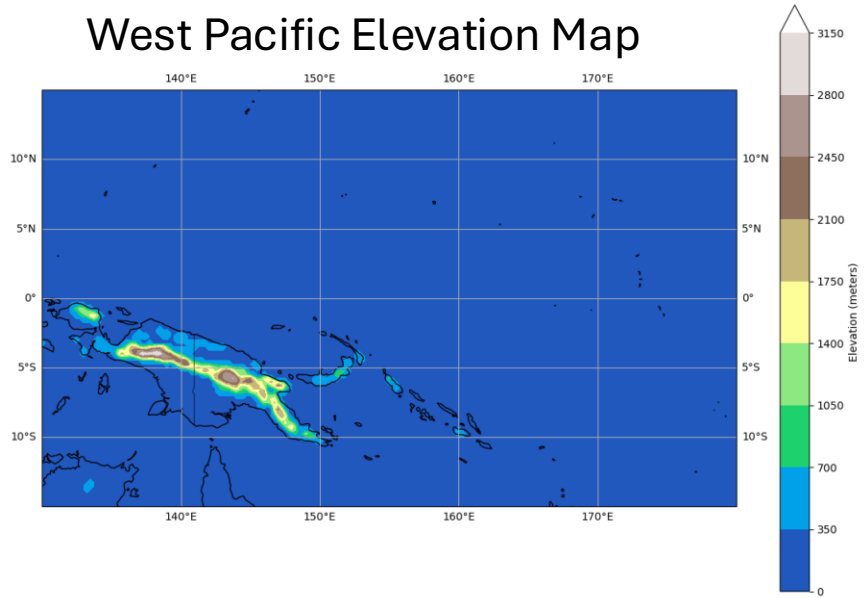
  ERA5: 25km

  or

  IMERG: (10km, 30 min) regridded to (25km, 1 hr)



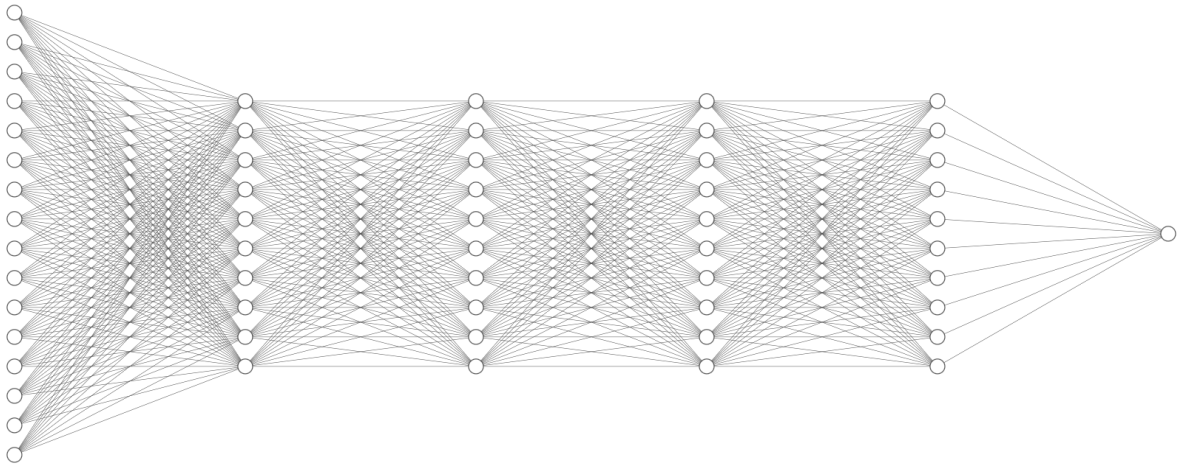
West Pacific Elevation Map



Overparameterized Neural Network

Network Architecture:

- 9 layers with 2,000 neurons per layer
- Total parameters: ~ 24 million
- Training Data: ~ 3 million samples



T, U, V, RH…
at 13 layers

Precipitation

Input Variables:

- Atmospheric variables
- Elevation

Sensitivity Study:

1. No Lat/Lon/Hour: 79 variables
2. Lat/Lon: 81 variables
3. Hour: 81 variables
4. Lat/Lon/Hour: 83 variables

1-D ML Prediction of Precipitation

ERA5

ML-PRED

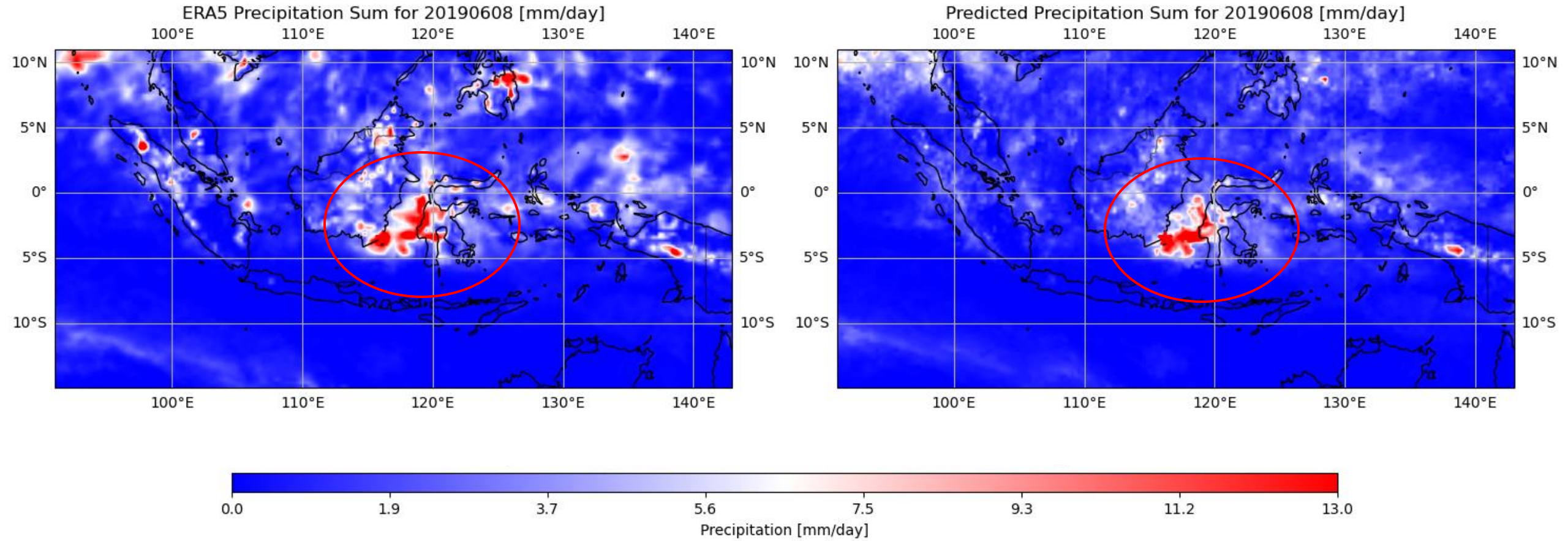


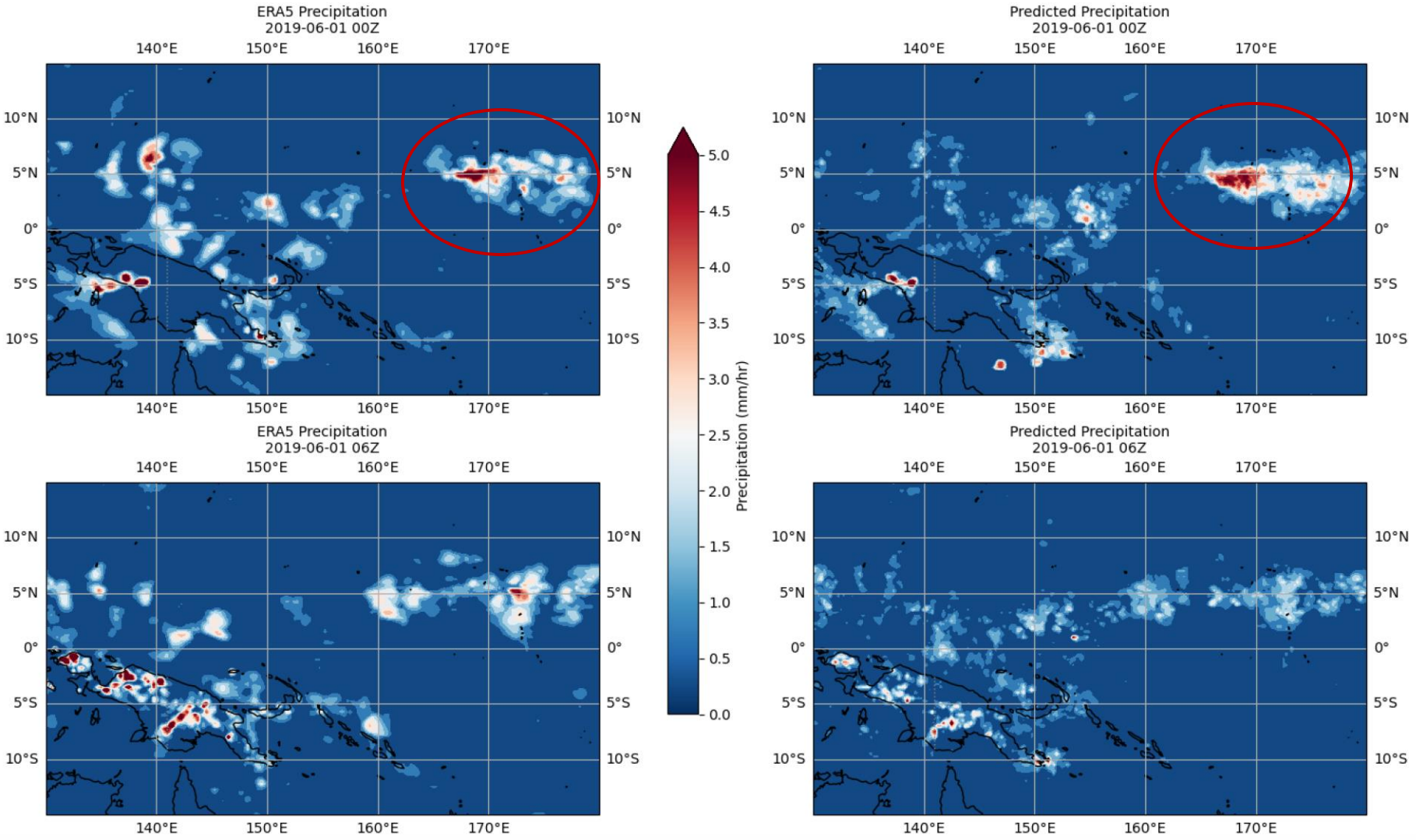
Figure courtesy of Destry Intan Syafitri J.

Model Comparison: with and without coordinate variables as predictors

	Only TUVRH	Lat/Lon	Hour	Lat/Lon/Hour
RMSE [mm/hr]	0.781	0.783	0.734	0.732
MedAbsError [mm/hr]	0.144	0.150	0.134	0.137

The model can predict precipitation based solely on environmental variables, without explicit latitude, longitude, or time information (“translation invariance”).

ERA5



1-D ML Model

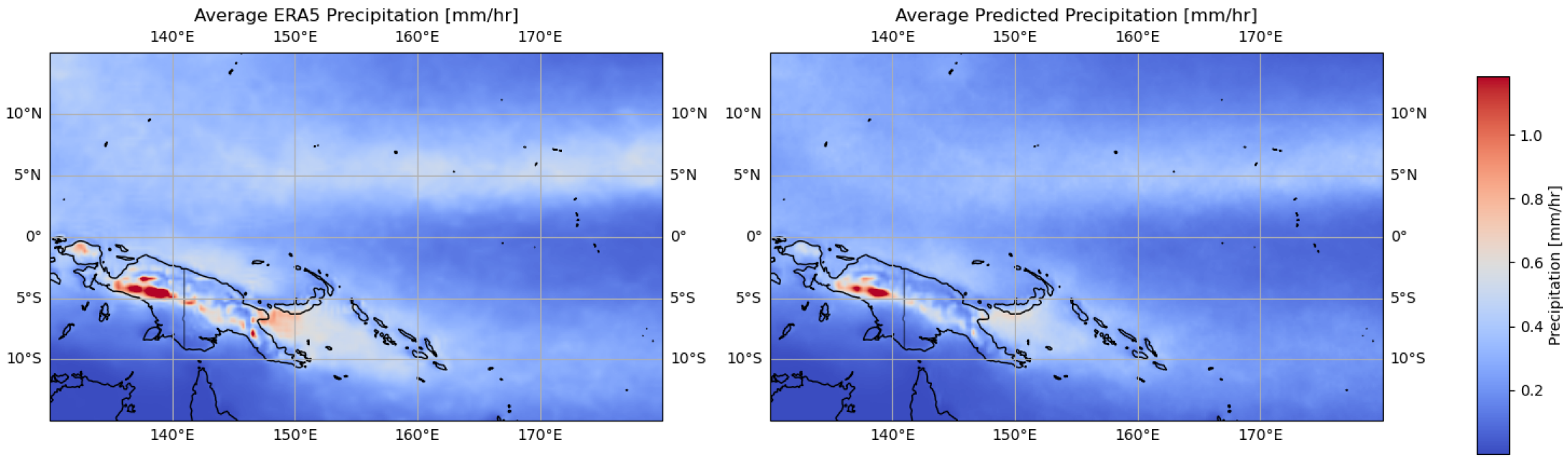
The model predicts well for synoptic-scale weather systems.

The model underestimates precip for smaller scale systems.

Annual Climatology

ERA5

1-D ML Model

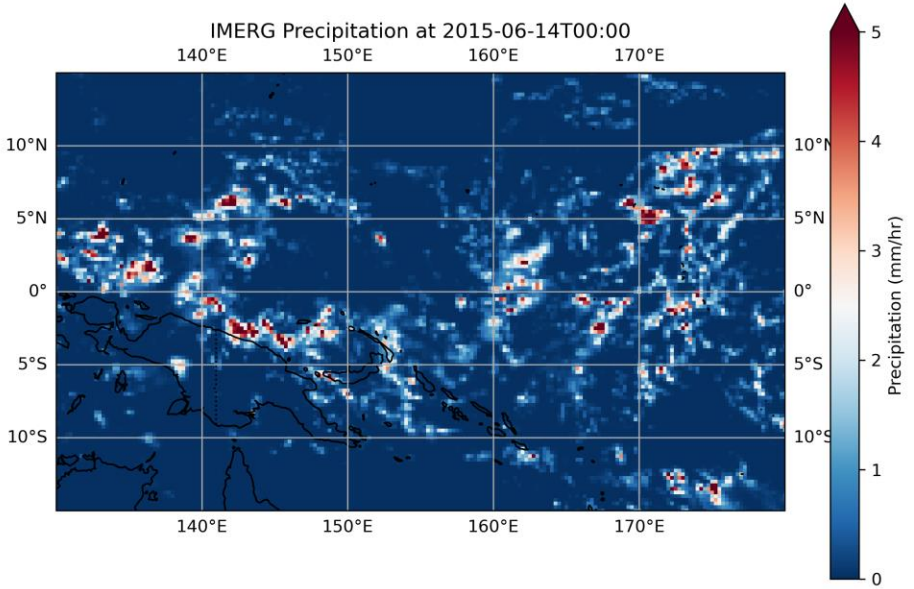


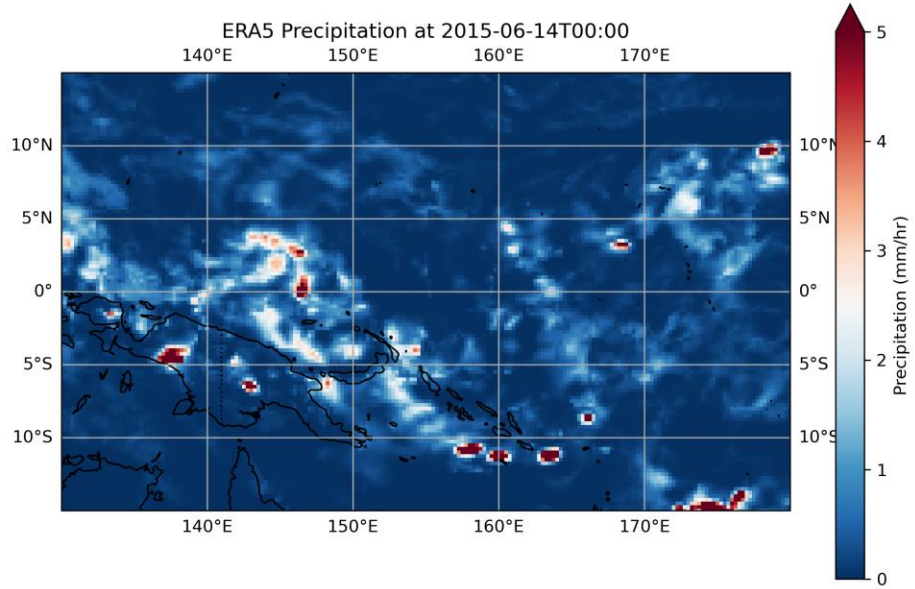
The overall rain structure is realistic, but the heavy precipitation regions (ITCZ and high-elevation areas on Papua New Guinea) are slightly underestimated.

Discrepancy between IMERG and ERA5 precip

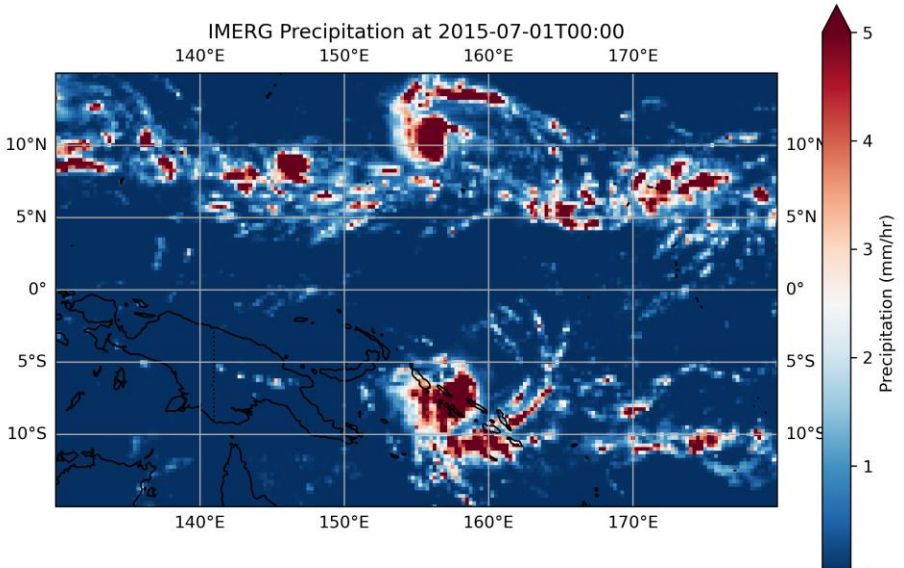
IMERG precip

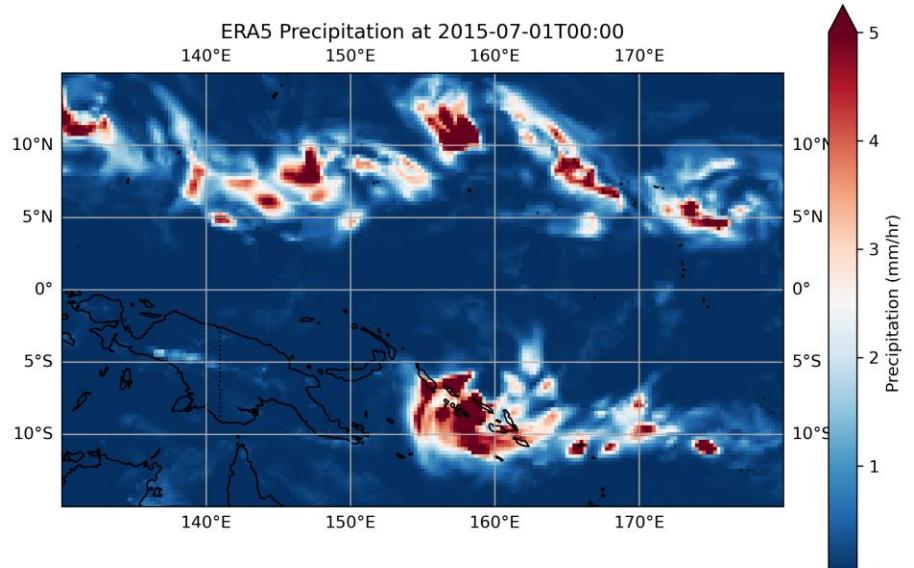
ERA5 precip





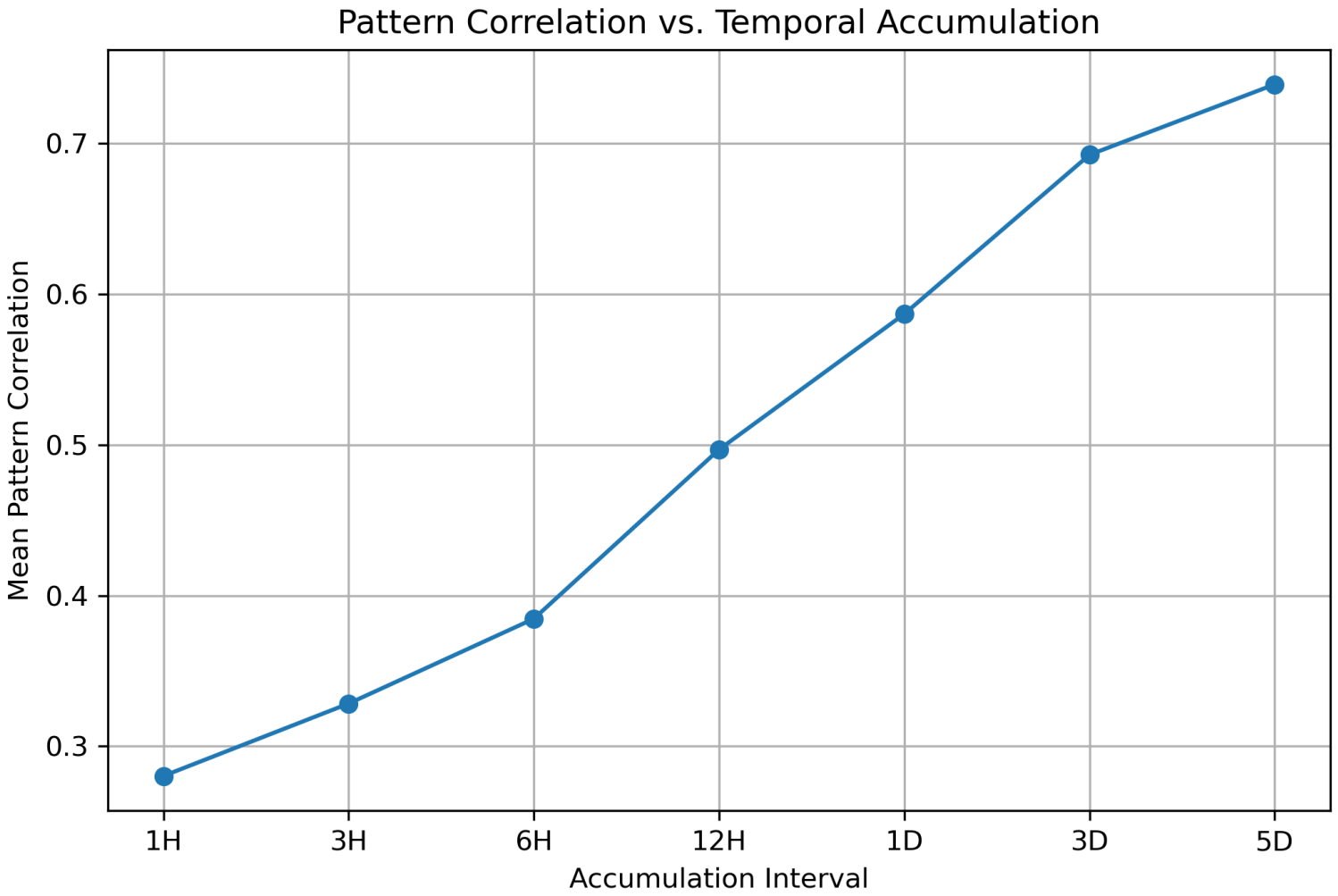
2015-06-14, 0Z





2015-07-01, 0Z

Pattern correlation between IMERG and ERA5 precip for different time-averaging



(Preliminary) Conclusions

- E2E: 1-D ML model trained to **use ERA5 flow** (T, q, U, V) to **predict ERA5 precip** works fairly well in predicting mean *and* extreme precip
  - E2E model is not good at predicting IMERG precip
- E2I: 1-D ML model trained to **use ERA5 flow** to **predict IMERG precip** also doesn't work as well as the E2E model in predicting precip

Hypotheses:

- Small-scale flow features in ERA5 are consistent with ERA5 precip but not with IMERG precip
- Only spatially and temporally smoothed precip may be predictable with ML (lack of consistent hi-res input *and* output training data)
  - Smoothed prediction maybe OK for weather but raises generalizability questions for climate?