Data-driven Models for Predicting Precipitation



<u>Chi-Jui Chen¹</u>, **R. Saravanan**¹, Courtney Schumacher¹, Istvan Szunyogh¹, Raymond Wong²

> ¹Department of Atmospheric Sciences ²Department of Statistics



The rationale for using ML to make predictions



"Future": The thing we want to predict

Can ML/AI "learn" physical equations from training data?



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

WeatherBench 2

Geopotenti 500hPa geopotential RMS			ntial MSE [kg²,	(m ²] 850hPa temperature RMSE [K]				Humidity 700hPa specific humidity RMSE [g/kg]					Wind Vector 850hPa wind vector RMSE [m/s]								
sical models	IFS HRES	42	135	304	521	801	0.62	1.16	1.82	2.63	3.63	0.55	0.96	1.27	1.53	1.81	1.69	3.30	5.21	7.13	9.16
	IFS ENS (mean)	42	132	277	439	621	0.65	1.11	1.62	2.17	2.80	0.51	0.84	1.06	1.22	1.38	1.63	2.99	4.45	5.75	6.95
Phys	ERA5-Forecasts	43	142	316	534	811	0.59	1.19	1.87	2.68	3.66	0.53	1.01	1.33	1.59	1.86	1.63	3.41	5.38	7.27	9.25
	Pangu-Weather (oper.)	45	136	300	510	785	0.65	1.09	1.74	2.54	3.55	0.53	0.86	1.17	1.45	1.76	1.71	3.03	4.86	6.77	8.84
	GraphCast (oper.)	40	124	276	477	754	0.53	0.93	1.56	2.36	3.40	0.48	0.76	1.03	1.29	1.60	1.48	2.74	4.52	6.42	8.57
	GenCast (oper.) (mean)	41	129	274	440	623	0.55	0.96	1.51	2.11	2.76	0.49	0.78	1.00	1.18	1.35	1.54	2.80	4.30	5.67	6.90
	Keisler (2022)	66	174	345	544	787	0.81	1.22	1.87	2.63	3.55	0.65	0.94	1.19	1.41	1.65	2.27	3.51	5.18	6.87	8.64
	Pangu-Weather	44	133	294	501	778	0.62	1.05	1.71	2.51	3.54	0.53	0.88	1.19	1.47	1.79	1.66	3.01	4.83	6.73	8.81
S	GraphCast	39	124	274	467	731	0.51	0.94	1.55	2.33	3.36	0.47	0.79	1.06	1.30	1.59	1.42	2.76	4.45	6.23	8.20
d mode	FuXi	40	125	277	433	631	0.54	0.97	1.59	2.14	2.91						1.47	2.80	4.51	5.66	7.04
L / hybri	NeuralGCM 0.7	37	115	267	469	751	0.54	0.97	1.58	2.38	3.42	0.48	0.83	1.12	1.40	1.71	1.49	2.81	4.59	6.51	8.66
Σ	NeuralGCM ENS (mean)	43	126	266	424	606	0.65	1.02	1.53	2.10	2.75	0.54	0.81	1.02	1.19	1.37	1.76	2.89	4.29	5.60	6.84
	GenCast (mean)	39	123	262	420	606	0.54	0.95	1.48	2.07	2.73	0.49	0.80	1.02	1.19	1.37	1.49	2.78	4.23	5.55	6.83
		1	3 Lead	5 time	7 [days]	10	1	3 Lead	5 time	7 [days]	10	1	3 Lead	5 time [7 [days]	10	1	3 Lead	5 time	7 [days]	10
					<	- 50 -	-20 –	10	-5	-2	-1	1	2	5	1	0 20	0 5	0			

Better \leftarrow % difference in RMSE vs IFS HRES \rightarrow Worse

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

20

10

50

Post-GATE view of tropical cloud population



Houze et al. (1980)

How do physics-based GCMs predict rain?

1-D Column subgrid parameterization



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)



(e) CMIP5 Mean



Physics-based GCMs underestimate extreme precipitation (drizzle problem)



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



Annual RainFall in 2003

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

(a) : STR Observations in 2004

PRED (GLM)

(b) : STR Predictions in 2004

Stratiform



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,

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
		Stratiform	0.154	0.211	0.155	0.094
West Pacific	WP	Deep convective	0.294	0.421	0.296	0.100
		Shallow convective	0.031	0.036	0.030	0.036
	EP	Stratiform	0.174	0.241	0.150	0.103
East Pacific		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)

IMERG: Integrated Multi-satellitE Retrievals for GPM



Overparameterized Neural Network

Network Architecture:

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



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

T, U, V, RH... at 13 layers

Precipitation

1-D ML Prediction of Precipitation

ERA5

ML-PRED ERA5 Precipitation Sum for 20190608 [mm/day] Predicted Precipitation Sum for 20190608 [mm/day] 100°E 110°E 120°E 130°E 100°E 110°E 120°E 130°E 140°E 140°E 10°N 10°N 10°N 10°N 5°N 5°N 5°N 5°N 0° 0° 0° 0° 5°S 5°S 5°S 5°S 10°S 10°S 10°S 10°S 100°E 110°E 120°E 130°E 140°E 100°E 110°E 120°E 130°E 140°E 1.9 3.7 5.6 9.3 11.2 0.0 7.5 13.0 Precipitation [mm/day]

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	
MedAbsErr	or [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

160°E

160°E

160°E

160°E

2019-06-01 06Z

170°E

170°E

170°E

170°E

5°N

0°

5°S

10°S

10°N

5°N

0°

5°S

10°5

2019-06-01 00Z

The model 10°N 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



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?