Learn More!

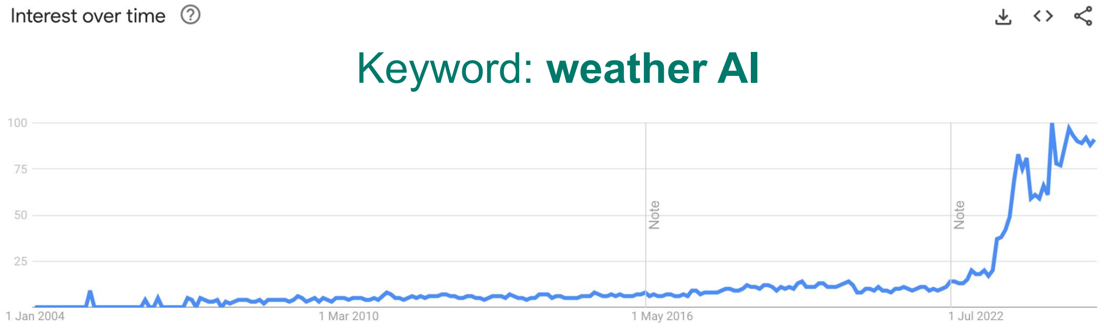
ChaosBench:

A Multi-Channel, Physics-Based Benchmark for Subseasonal-to-Seasonal Climate Prediction

Juan Nathaniel (jn2808@columbia.edu), Yongquan Qu, Tung Nguyen, Sungduk Yu, Julius Busecke, Aditya Grover, Pierre Gentine

The rise of weather ML

Weather (/ ˈwɛð ər /): short-medium term **(up to 2 weeks)** states



Keyword: **weather AI**

Graphcast

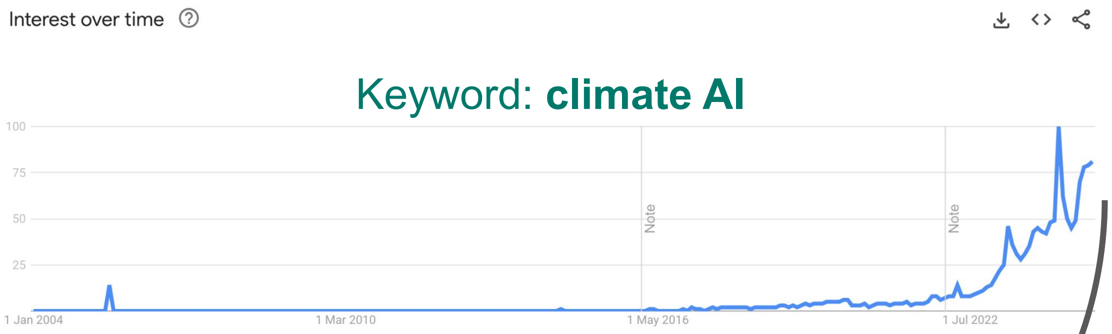
AIFS (operational)

FourCastNet

...

The rise of climate ML

Climate (/ˈklaɪmət/): long-term **(annual/decadal/centuries)** states



DYffusion **ClimateBench**

ACE2 **...**

Prediction across timescale

Ultimate Goal: Climate prediction across timescale…

WeatherBench:
Initial Condition (IC) Sensitivity

ClimateBench:
Boundary Condition (BC) Sensitivity

Weather	?	Climate

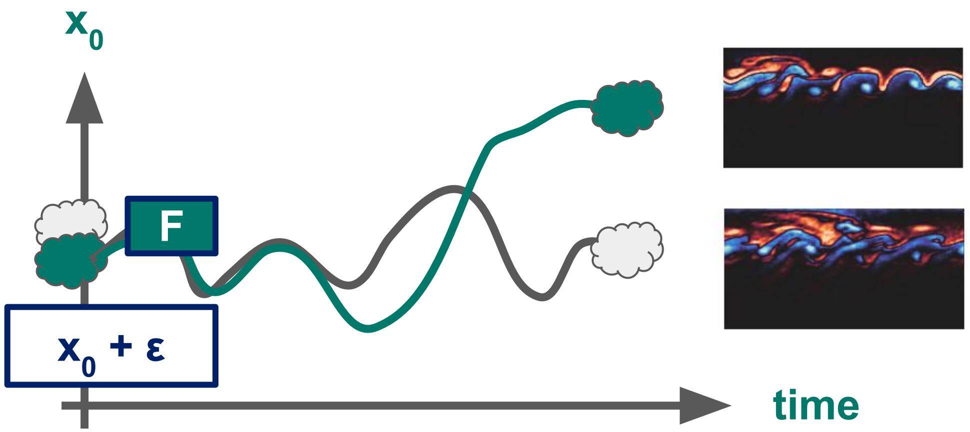
ChaosBench:
IC + BC: Challenging, but can give us clues for across time-scale prediction?

timescales

Skillful forecasting at this scale
i.e., **subseasonal-to-seasonal (S2S)** is **hard**, yet **critical**

Initial condition sensitivity

Weather | ? | Climate



different state!

Weather Forecasting: Get the **<u>IC</u>** right!

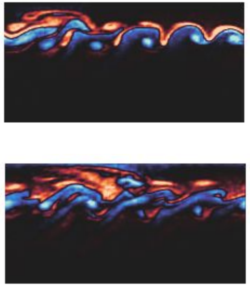
Boundary condition sensitivity

Weather

?

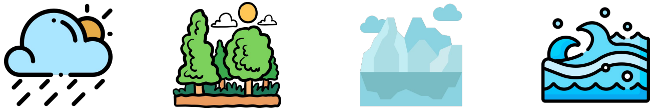
Climate

different state!



Statistics:
mean warming of
1.5^0 by 2050

Boundary conditions
air, land, ice, sea

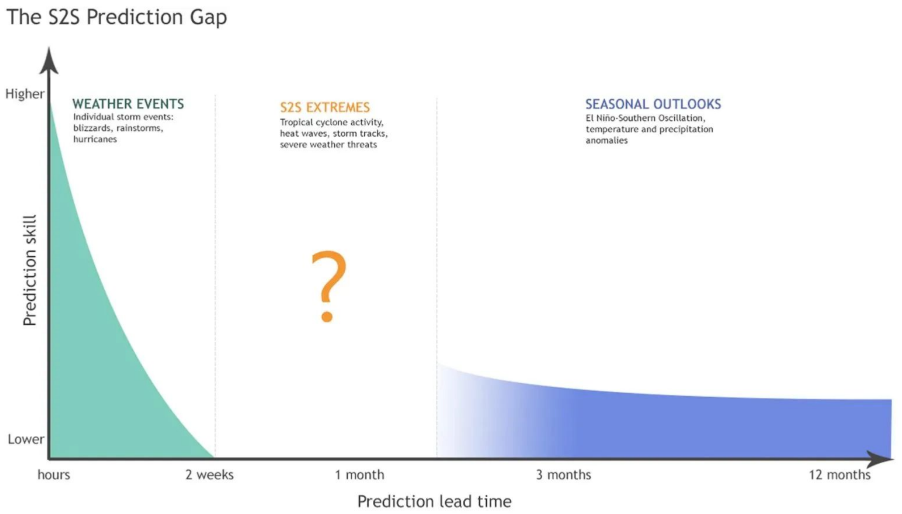


Climate Forecasting: Get the **BC** right!

LEAP

Predictability gap

Skill gap in between weather and climate timescales



Adapted from: iri.columbia.edu/news/qa-subseasonal-prediction-project

S2S: High Impact Events

ChaosBench features

Large benchmarks
Get the **IC** right!

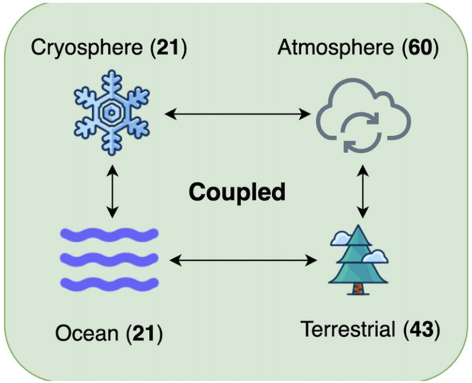
Physics models across weather centers globally

Deep models across SoTAs (Graphcast, …)

Comprehensive metrics (probabilistic, physics)

Coupled dataset
Get the **BC** right!

Beyond ERA5



Looking ahead
Get **both** right!

Well-spread ensembles

Hybrid physics model

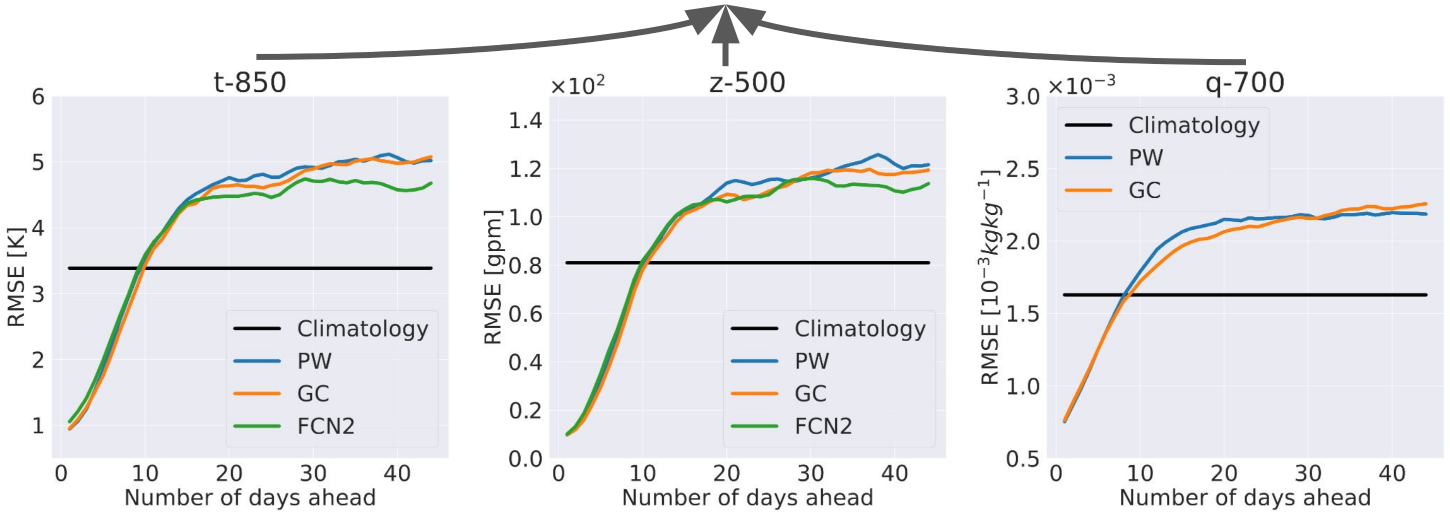
Multiscale approach

https://leap-stc.github.io/ChaosBench/

The collapse of deterministic weather ML

Variables at different vertical atmosphere level

Metrics



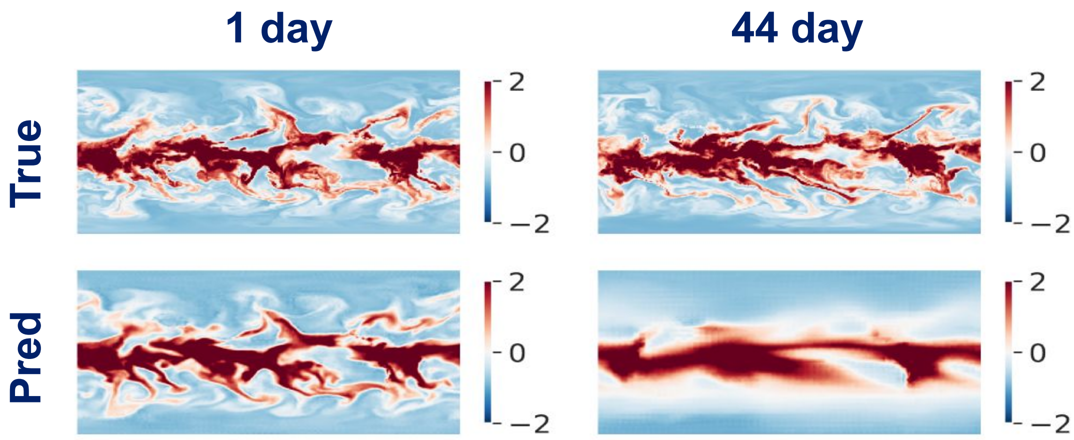
Climatology: long-term average (unskilled baseline)

PW: Panguweather
GC: Graphcast
FCN2: FourCastNetV2

~10 days predictability range for SoTA Weather ML

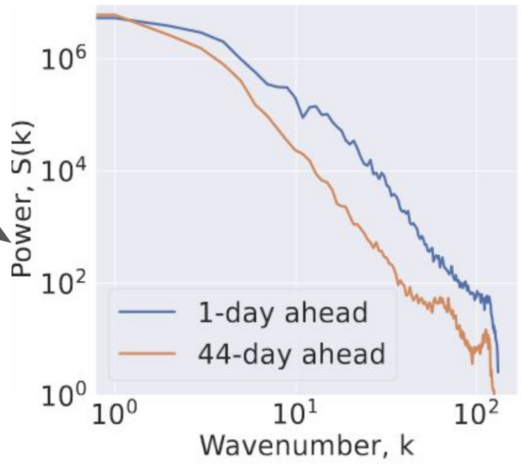
LEAP

The collapse of deterministic weather ML



Blurrier as **forecast time** increases

Failure to capture
high-frequency details



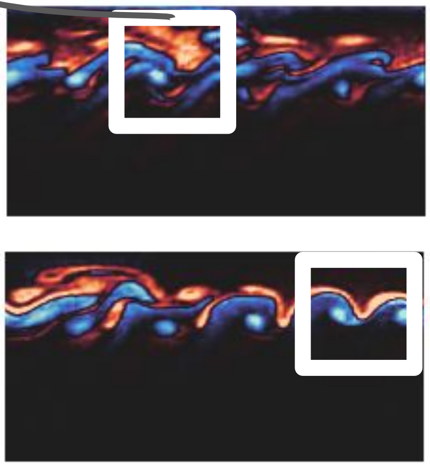
*Analysis done with ViT/ClimaX: specific humidity (q) at pressure level 700 hPa;
** Insights are consistent across the board (models & variables)

The collapse of deterministic weather ML

Extreme heat

x_a

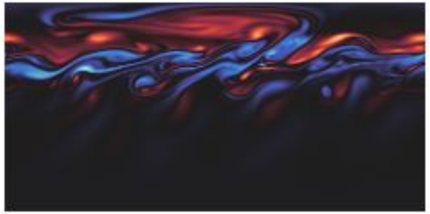
x_b



Extreme cold

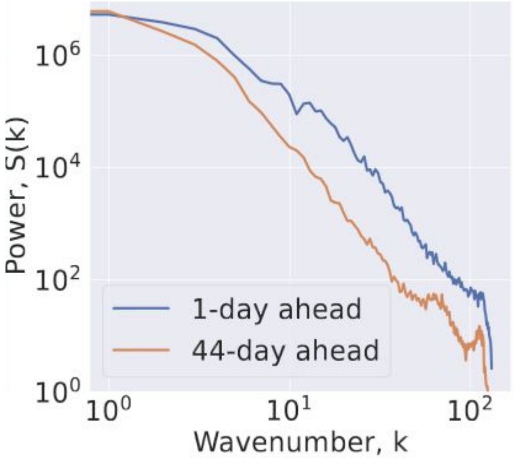
double penalty

distance loss (L1, L2)



Smoothed forecasts (no extremes, etc)

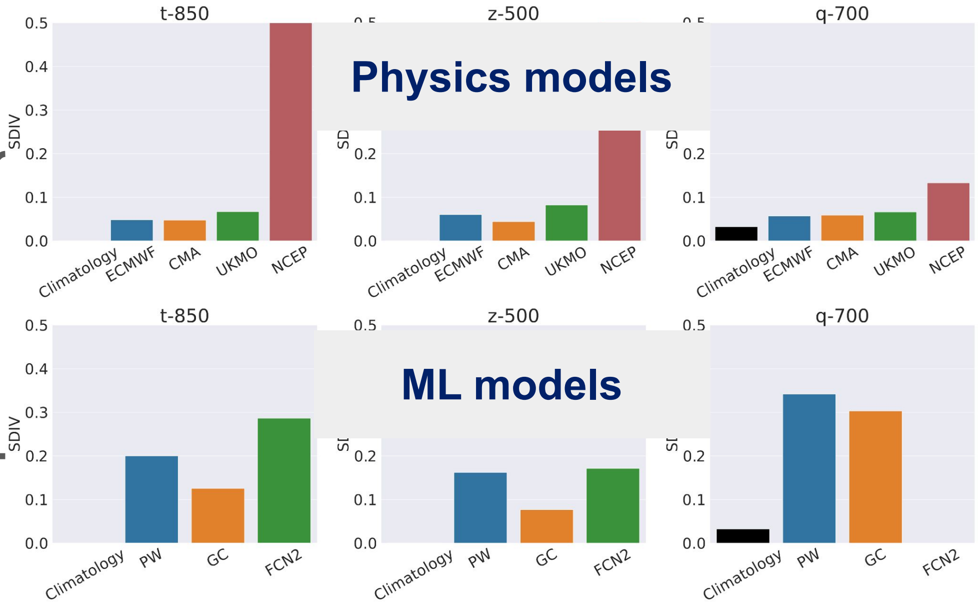
Closing the gap: physics-based metrics



Compute Spectra Discrepancy:

1. Kullback-Leibler Divergence
2. L1 Loss

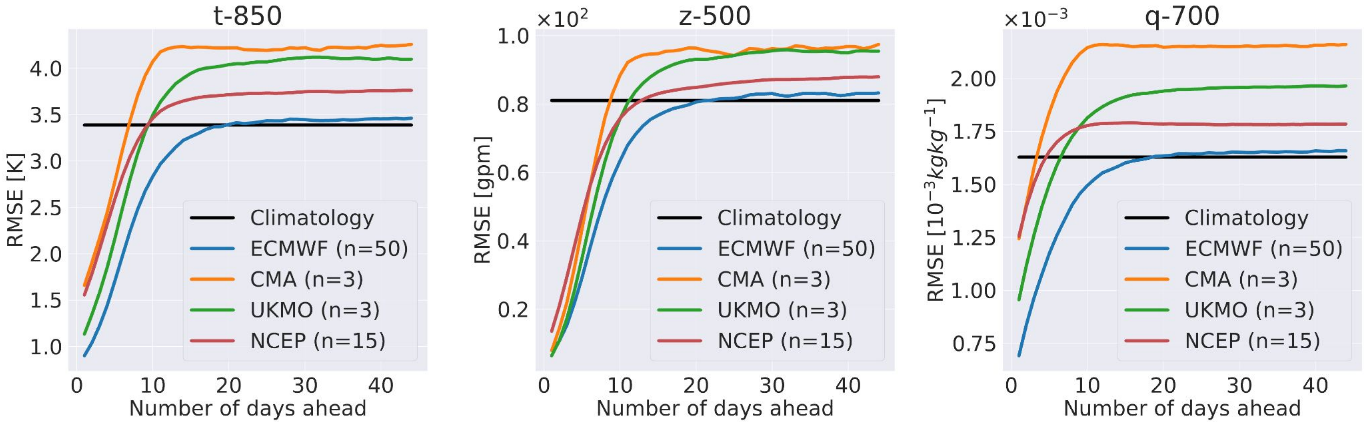
close the gap!



ML models have **higher spectra discrepancy** than **physics models**

What’s next: consulting domain science

Physics models appear to be **better**, but still **plateauing early…**

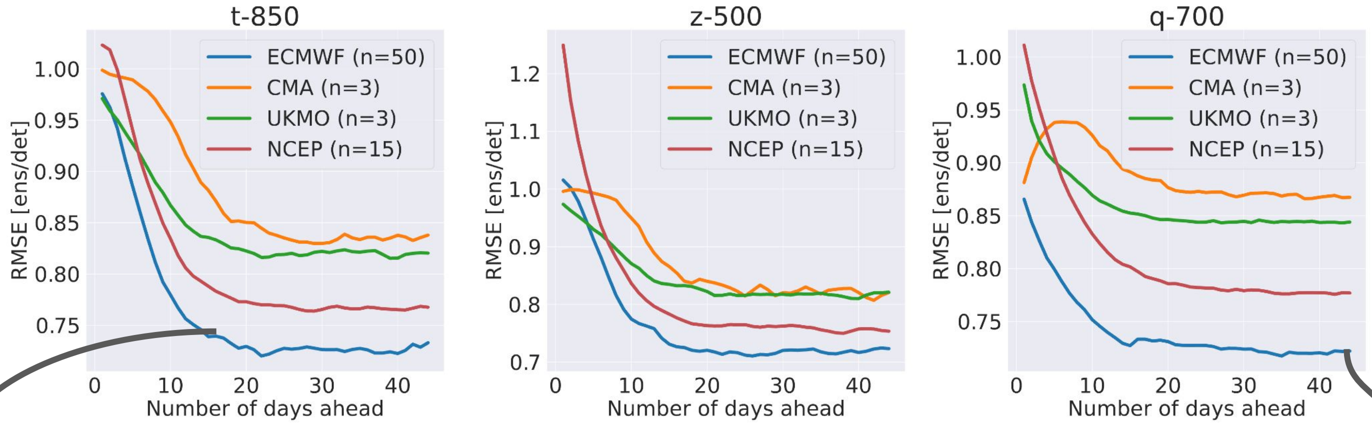


ECMWF: European Centre for Medium-Range Weather Forecasts
CMA: China Meteorological Administration
UKMO: UK Meteorology Office
NCEP: National Centers for Environmental Prediction

**(n = number of ensemble members)

Ensemble scaling

Increasing the number of well-spread ensemble member (n) improves skillfulness



Improvement gain as <u>number of members</u> increases
Beneficial especially for <u>long-range forecasts</u>

ECMWF: European Centre for Medium-Range Weather Forecasts
CMA: China Meteorological Administration
UKMO: UK Meteorology Office
NCEP: National Centers for Environmental Prediction

**(n = number of ensemble members)

Hybrid modeling

Hybrid physics-informed ML model* shows promise for long-range modeling**

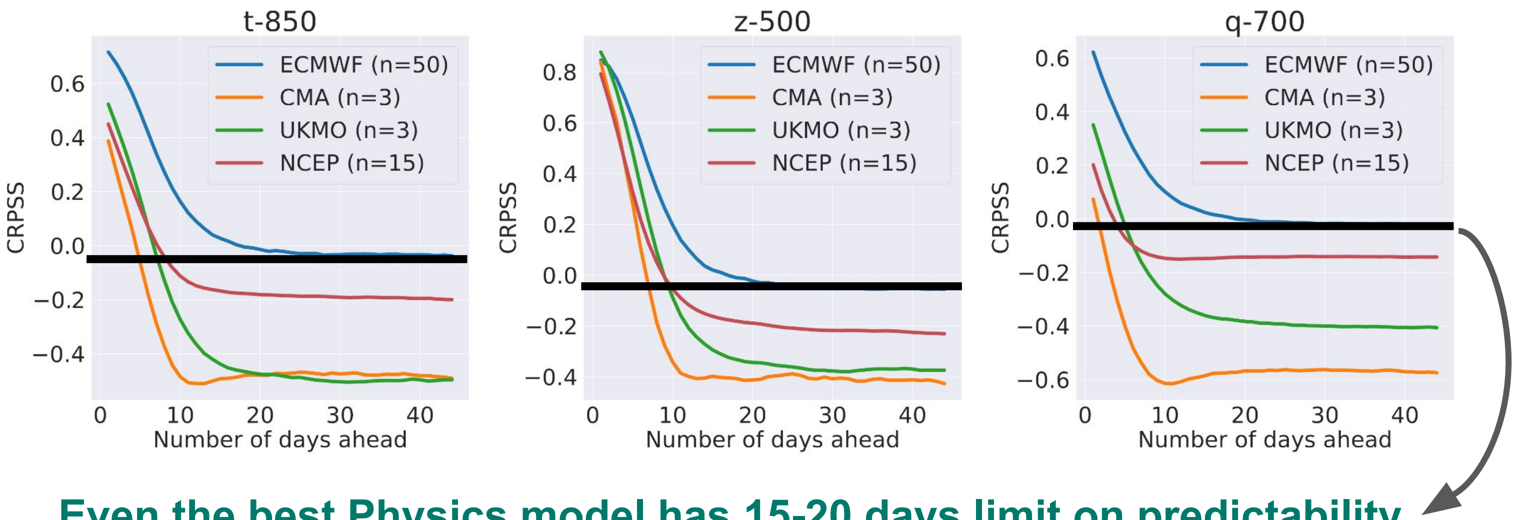
Models	RMSE ↓			MS-SSIM ↑			SpecDiv ↓		
	T850 (K)	Z500 (gpm)	Q700 ($\times 10^{-3}$)	T850	Z500	Q700	T850	Z500	Q700
Lagged AE	5.55	122.4	2.03	0.74	0.71	0.47	0.18	2.44	0.21
ResNet	5.67	125.3	2.07	0.73	0.70	0.47	0.21	0.37	0.26
UNet	5.47	121.5	2.13	0.73	0.71	0.45	0.30	1.16	2.20
FNO	**5.06**	**112.5**	**1.95**	**0.75**	**0.73**	**0.51**	**0.18**	**0.11**	**0.10**

FNO preserves some spectral physical information

*all experiments are performed with identical number of trainable parameters, hyperparameters
**results at final timestep T = 44 days ahead

On the *limit* of current predictability

Measuring Skillfulness: CRPSS* → 0 (Unskilled)



Even the best Physics model has 15-20 days limit on predictability

Challenge: Can we extend the predictability range with ML?

*CRPSS: Continuous Ranked Probability Skill Score

The path forward: ML + Physics synthesis

Predictability can be extended (some strategies analyzed in the paper):

- Well-spread ensemble → w/ Probabilistic ML
- Physics-based ML
- Robust control of error propagation

3-easy step Quickstart!

https://leap-stc.github.io/ChaosBench/

LEAP