SEP 22-26

AIM for Scale
AI Weather Training Program

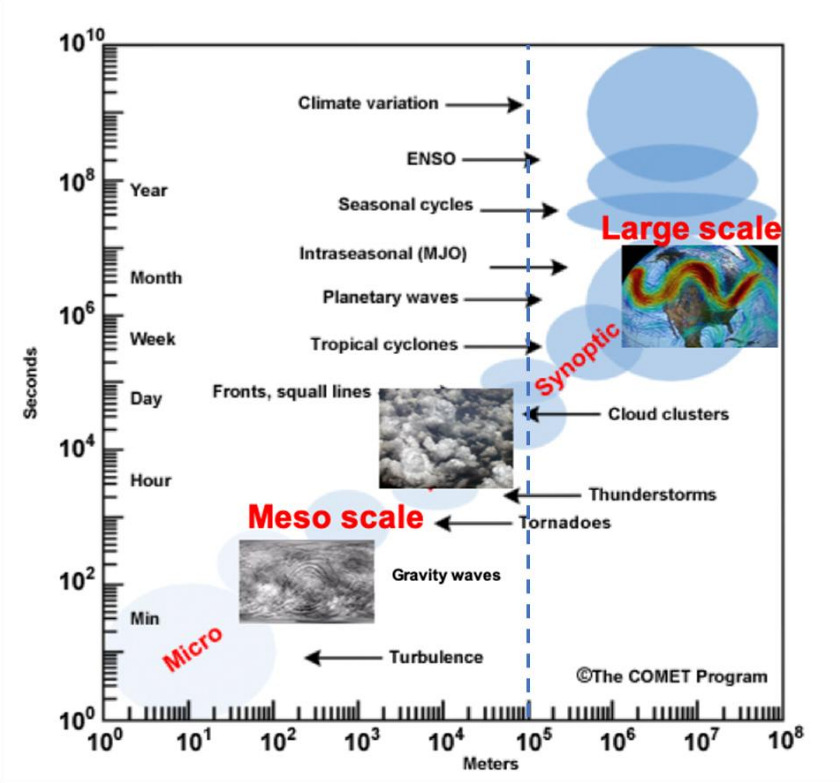
AIM FOR SCALE | WEATHER PACKAGE

AI vs. NWP

Pedram Hassanzadeh

Simulating & Analyzing the Atmosphere is Challenging

Multi-scale, multi-physics, nonlinear, high-dimensional & chaotic/noisy…



$$\boxed{\frac{d(\boldsymbol{U}+\boldsymbol{u})}{dt} = \mathbf{N}(\boldsymbol{U}+\boldsymbol{u})}$$

$\boldsymbol{U}$: large/slow-scale variables

The main variables of interest

$\boldsymbol{u}$: small/fast-scale variables

Influence the spatio-temporal variability of $\boldsymbol{U}$

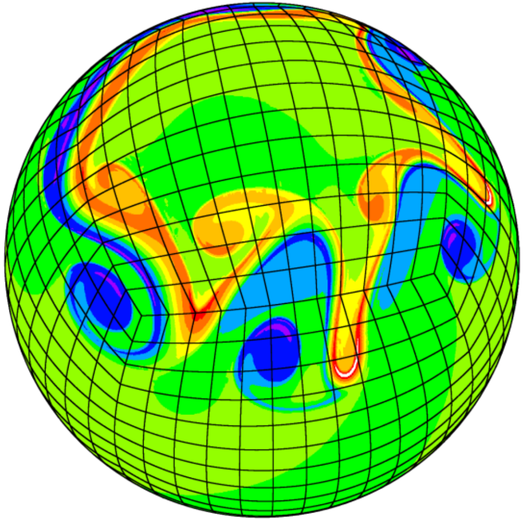
Current Approach:

Low-resolution numerical solver + physics-based subgrid-scale (SGS) models

General circulation models (GCMs): Large-scale processes

$$\frac{d\boldsymbol{U}}{dt} = \mathbf{F}(\boldsymbol{U}, \mathbf{P}(\boldsymbol{U}))$$

solved numerically at $\mathcal{O}(10)$-$\mathcal{O}(100)$ km resolutions



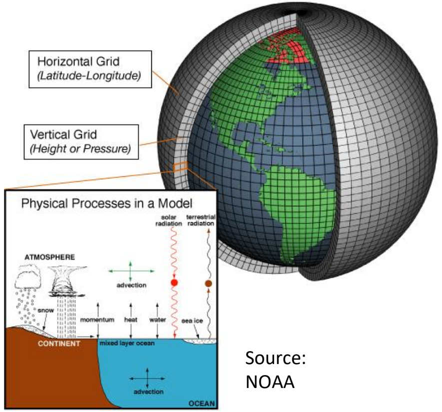
https://admg.engin.umich.edu/

Parameterizations (closures) for SGS processes

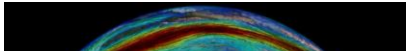
$$\boldsymbol{u} = \mathbf{P}(\boldsymbol{U})$$

1st Revolution in Weather and Climate Prediction: 1950-2000

Numerical solution of PDEs governing atmosphere, ocean etc.



Source:
NOAA



Atmospheric Predictability as Revealed by Naturally Occurring Analogues

Edward N. Lorenz

Dept. of Meteorology, Massachusetts Institute of Technology, Cambridge, Mass.[1]

(Manuscript received 2 April 1969)

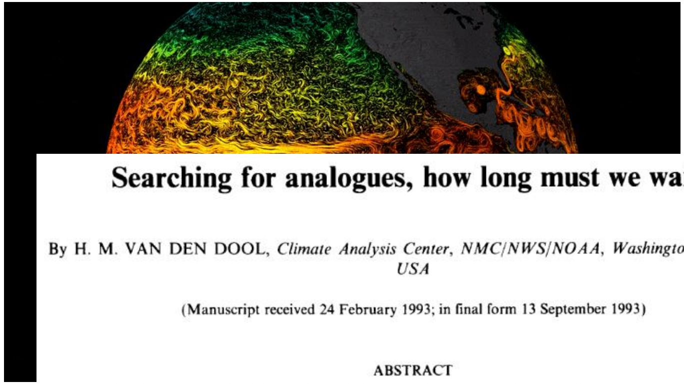
ABSTRACT

Two states of the atmosphere which are observed to resemble one another are termed *analogues*. Either state of a pair of analogues may be regarded as equal to the other state plus a small superposed "error." From the behavior of the atmosphere following each state, the growth rate of the error may be determined.

Five years of twice-daily height values of the 200-, 500-, and 850-mb surfaces at a grid of 1003 points over the Northern Hemisphere are procured. A weighted root-mean-square height difference is used as a measure of the difference between two states, or the error. For each pair of states occurring within one month of the same time of year, but in different years, the error is computed.

There are numerous mediocre analogues but no truly good ones. The smallest errors have an average doubling time of about 8 days. Larger errors grow less rapidly. Extrapolation with the aid of a quadratic hypothesis indicates that truly small errors would double in about 2.5 days. These rates may be compared with a 5-day doubling time previously deduced from dynamical considerations.

The possibility that the computed growth rate is spurious, and results only from having superposed the smaller errors on those particular states where errors grow most rapidly, is considered and rejected. The likelihood of encountering any truly good analogues by processing all existing upper-level data appears to be small.



Searching for analogues, how long must we wait?

By H. M. VAN DEN DOOL, *Climate Analysis Center, NMC/NWS/NOAA, Washington DC 20233, USA*

(Manuscript received 24 February 1993; in final form 13 September 1993)

ABSTRACT

A three-way relationship is derived between the size of a library (M years) of historical atmospheric data, the distance between an arbitrarily picked state of the atmosphere and its nearest neighbor (or analogue), and the size of the spatial domain, as measured by the number of spatial degrees of freedom (N). It is found that it would take a library of order 10^{30} years to find 2 observed flows that match to within current observational error over a large area such as the Northern Hemisphere. Obviously, with only 10–100 years of data, the probability of finding natural analogous is very small, unless one is satisfied with analogy over small areas or in just 2 of 3 degrees of freedom as represented, for instance, by 2 or 3 leading empirical orthogonal modes. We further propose the notion that analogues can be constructed by combining a number of observed flow patterns. We have found at least one application where linearly constructed analogues are conclusively better at specifying US surface weather from concurrent 700 mb geopotential height than natural analogues are.

1910: first numerical weather forecast failed (L. Richardson)

data science:
pattern matching
I. Krick (Caltech)

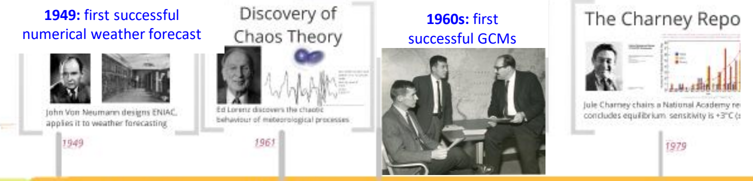
vs.

1928: Courant-Friedrich-Lewy (CFL) condition

first principles, PDEs
CG. Rossby (U. Chicago)

1949: first successful numerical weather forecast

1960s: first successful GCMs



modified
From

www.easterbrook.ca

AI-based approach: Low-resolution numerical solver + data-driven subgrid-scale (SGS) models

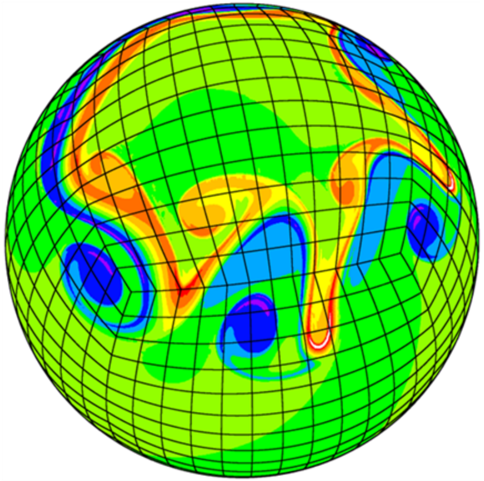
General circulation models (GCMs): Large-scale processes

$$\frac{d\boldsymbol{U}}{dt} = \mathbf{F}(\boldsymbol{U}, \mathbf{D}(\boldsymbol{U}))$$

solved numerically at $\mathcal{O}(10)$-$\mathcal{O}(100)$ km resolutions

Data-driven parameterizations for SGS processes

$$\boldsymbol{u} = \mathbf{D}(\boldsymbol{U})$$



https://admg.engin.umich.edu/

AIM FOR SCALE | WEATHER PACKAGE

2nd Revolution in 1-15 Day Weather Forecasting (2018-now)

10^5x faster and more accurate than the best physics-based models

autoregressive

$$\boldsymbol{u}(t+\Delta t) = \mathbf{NN}(\boldsymbol{u}(t), \boldsymbol{\theta})$$

$$\mathcal{L} = \|\boldsymbol{u}(t+\Delta t) - \mathbf{NN}(\boldsymbol{u}(t), \boldsymbol{\theta})\|_2$$

initial-value solver

2nd Revolution in 1-15 Day Weather Forecasting (2018-now)

10^5x faster and more accurate than the best physics-based models

autoregressive

$$\boldsymbol{u}(t+\Delta t) = \mathbf{NN}(\boldsymbol{u}(t), \boldsymbol{\theta})$$

~75M learnable parameters (transformer + FNO)

FOURCASTNET: A GLOBAL DATA-DRIVEN HIGH-RESOLUTION WEATHER MODEL USING ADAPTIVE FOURIER NEURAL OPERATORS

2022: https://arxiv.org/abs/2202.11214

Jaideep Pathak
NVIDIA Corporation
Santa Clara, CA 95051

Shashank Subramanian
Lawrence Berkeley National Laboratory
Berkeley, CA 94720

Peter Harrington
Lawrence Berkeley National Laboratory
Berkeley, CA 94720

Sanjeev Raja
University of Michigan
Ann Arbor, MI 48109

Ashesh Chattopadhyay
Rice University
Houston, TX 77005

Morteza Mardani
NVIDIA Corporation
Santa Clara, CA 95051

Thorsten Kurth
NVIDIA Corporation
Santa Clara, CA 95051

David Hall
NVIDIA Corporation
Santa Clara, CA 95051

Zongyi Li
California Institute of Technology
Pasadena, CA 91125
NVIDIA Corporation
Santa Clara, CA 95051

Kamyar Azizzadenesheli
Purdue University
West Lafayette, IN 47907

Pedram Hassanzadeh
Rice University
Houston, TX 77005

Karthik Kashinath
NVIDIA Corporation
Santa Clara, CA 95051

Animashree Anandkumar
California Institute of Technology
Pasadena, CA 91125
NVIDIA Corporation
Santa Clara, CA 95051

Trained only on 33 variables of 1979-2017 25km ERA5

2nd Revolution in 1-15 Day Weather Forecasting (2018-now)

2019 Oxford workshop: getting close to IFS is 10 years awa

2022: FourCastNet (done) → 2023: IFS is beat!

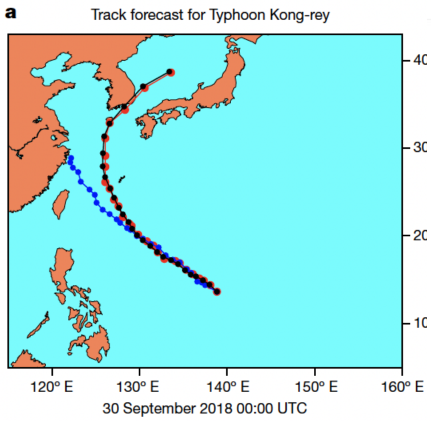
NeuralGCM
Aardvark
Salient
GraphCast
FourCastNet
WeatherMesh
Aurora
ClimaX
FuXi
Pangu
Prophet
Prithvi
From Neil Hausmann

Pangu

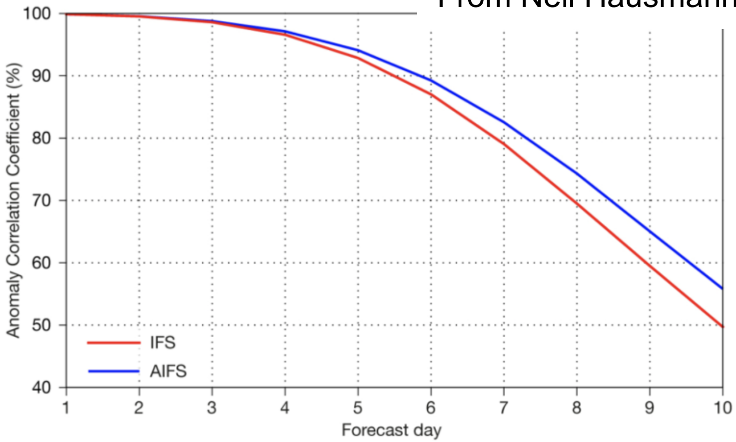
Bi et al. 2023 (Nature): *200M parameters (transformer)*

GraphCast

Lam et al. 2023 (Science): *300M parameters (graph neural nets)*



Pangu-Weather forecast — ECMWF-HRES forecast — Ground truth

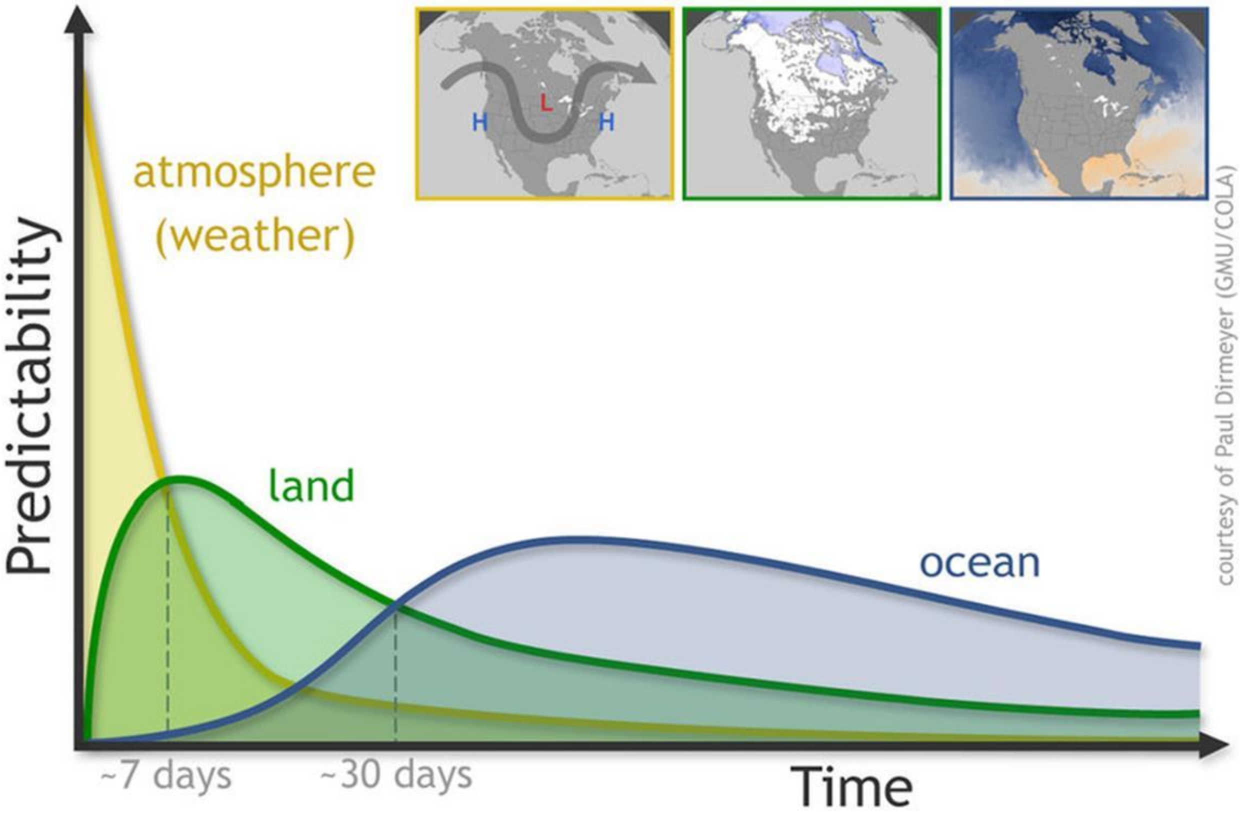


Lam et al. (2023)

www.ecmwf.int

1st Revolution in Subseasonal-to-Seasonal (S2S) Forecasting?

Requires capturing the interactions of atmosphere-ocean-land and generating calibrated ensembles



Source: NOAA

Why generate a forecast with an AI model? Speed is a huge factor

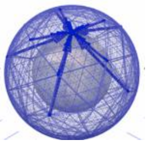
- Several hours to run an NWP forecast but several minutes to run an AI forecast - instead of redoing all of the physics for each simulation, you can emulate patterns much faster
- Thousands of CPUs vs 1 GPU to make a forecast - energy savings
  - Large upfront cost in training, but once these models are trained, much faster to inference
  - Generally only need one GPU for inference (although several GPUs needed for training)
- GPUs - great for parallelism -

AI Weather Model Architecture is on a Spectrum of Data-driven Behavior

Purely data driven ⟵⟶ **Hybrid** ⟵⟶ **Dynamical**

earth-specific transformer mechanisms

FourCastNet
PanguWeather
FuXi
FuXi S2S
GraphCast
AIFS



Gencast

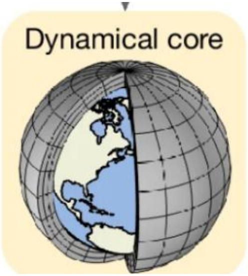
conditional diffusion

[ACE2]

incorporates physical center of mass rules

NeuralGCM

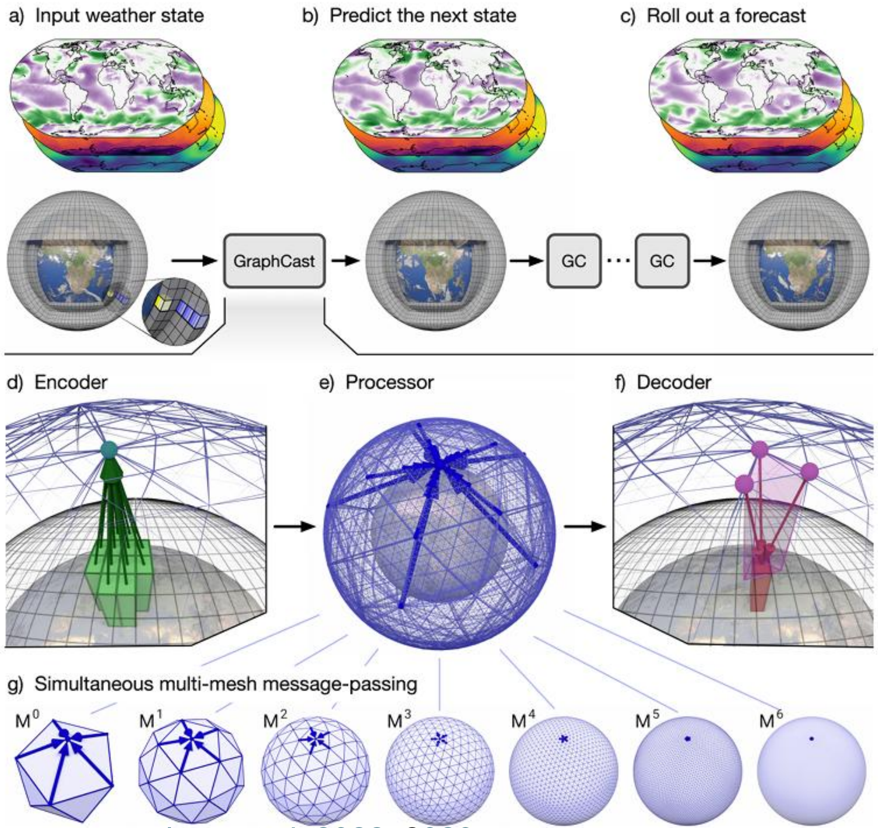
combines a dynamical core with machine learning processes



IFS
GFS

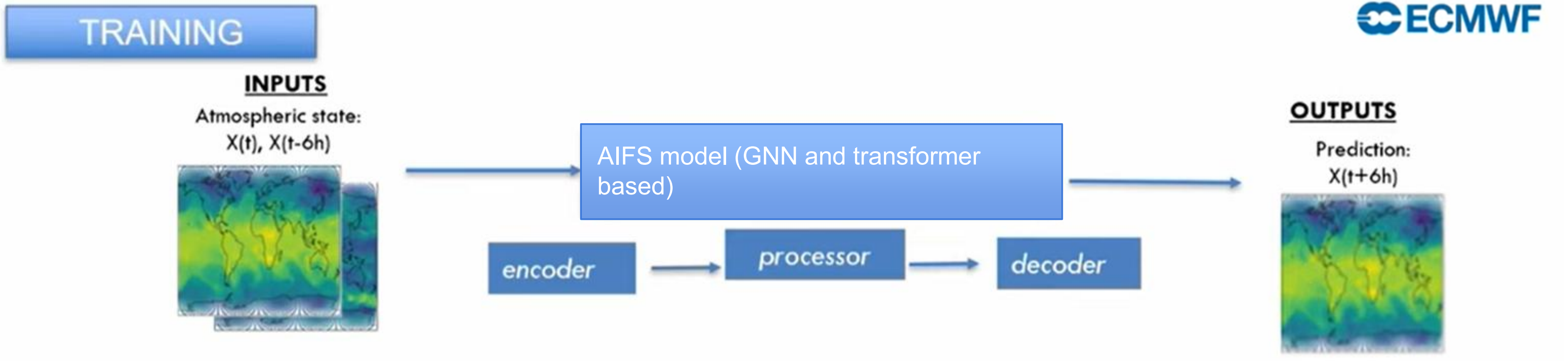
purely dynamical, simulating earth system with physical laws

Graphcast looks at earth as nodes in a mesh, examining how nodes interact



source: Lam et al. 2022, 2023

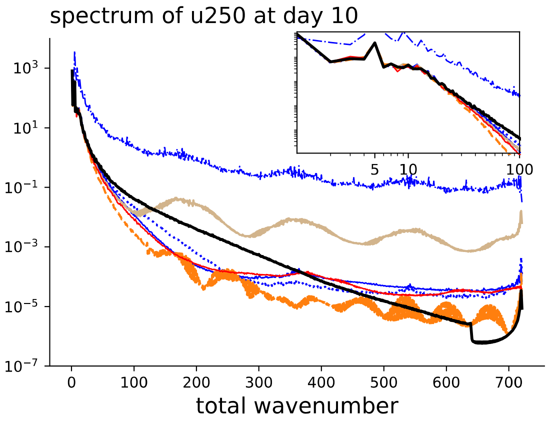
- Initial weather states defined on a 25 km grid - 5 surface variables, 6 atmospheric variables, repeated at 37 pressure levels here
- Multi-mesh grid that enables earth-specific training across nodes (locations) and atmospheric levels

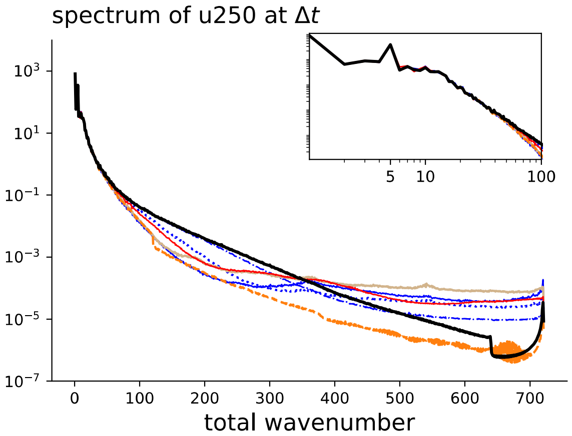


AIFS (deterministic and ensemble) works as a transformer with a sliding attention window

source: ECMWF, see recent seminar on AIFS ENS for more information on recent advances

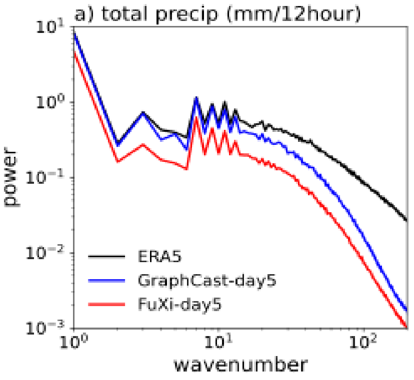
Double penalty problem and how it relates to spectral bias

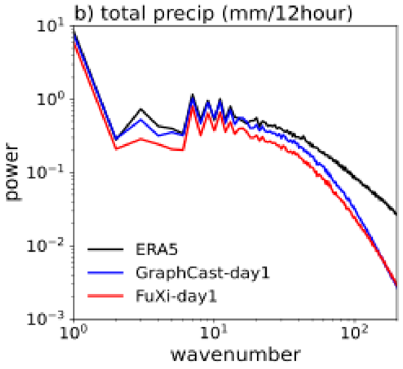




$$\boldsymbol{u}(t+\Delta t) = \mathbf{NN}(\boldsymbol{u}(t), \boldsymbol{\theta})$$

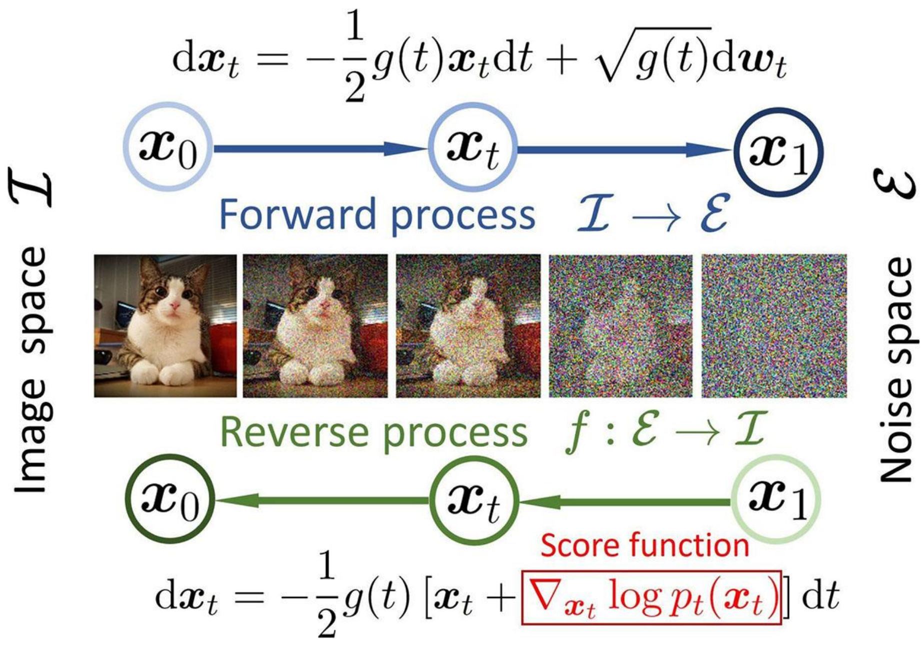
$$\mathcal{L} = \|\boldsymbol{u}(t+\Delta t) - \mathbf{NN}(\boldsymbol{u}(t), \boldsymbol{\theta})\|_2$$



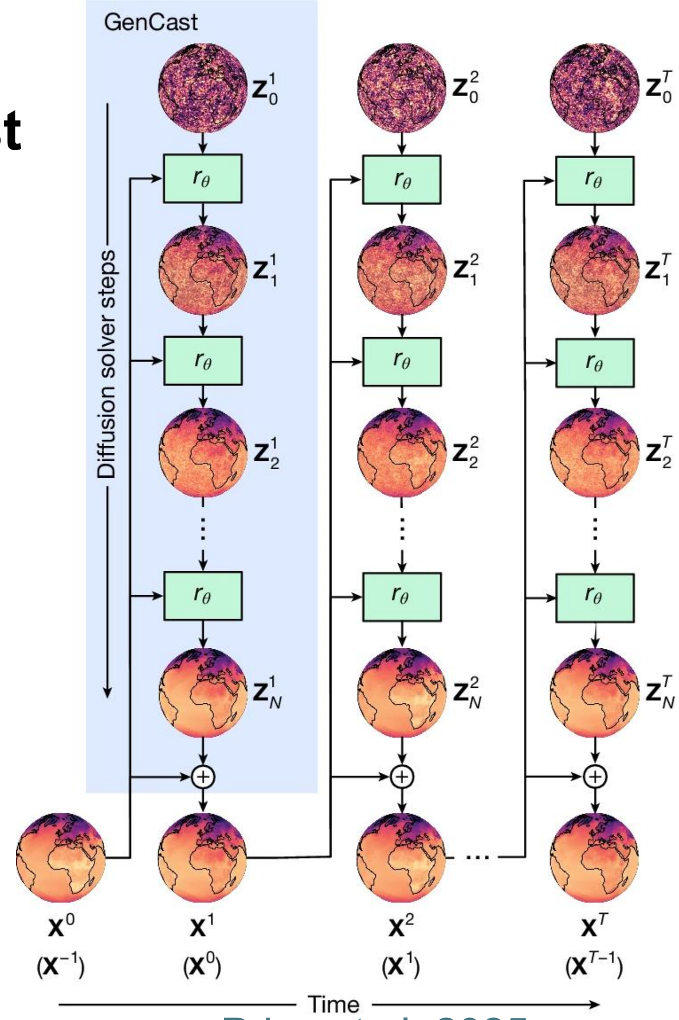


Chattopadhyay, Sun & Hassanzadeh (2023 http

Conditional diffusion models, one way to address spectral bias isse - example **GenCast**

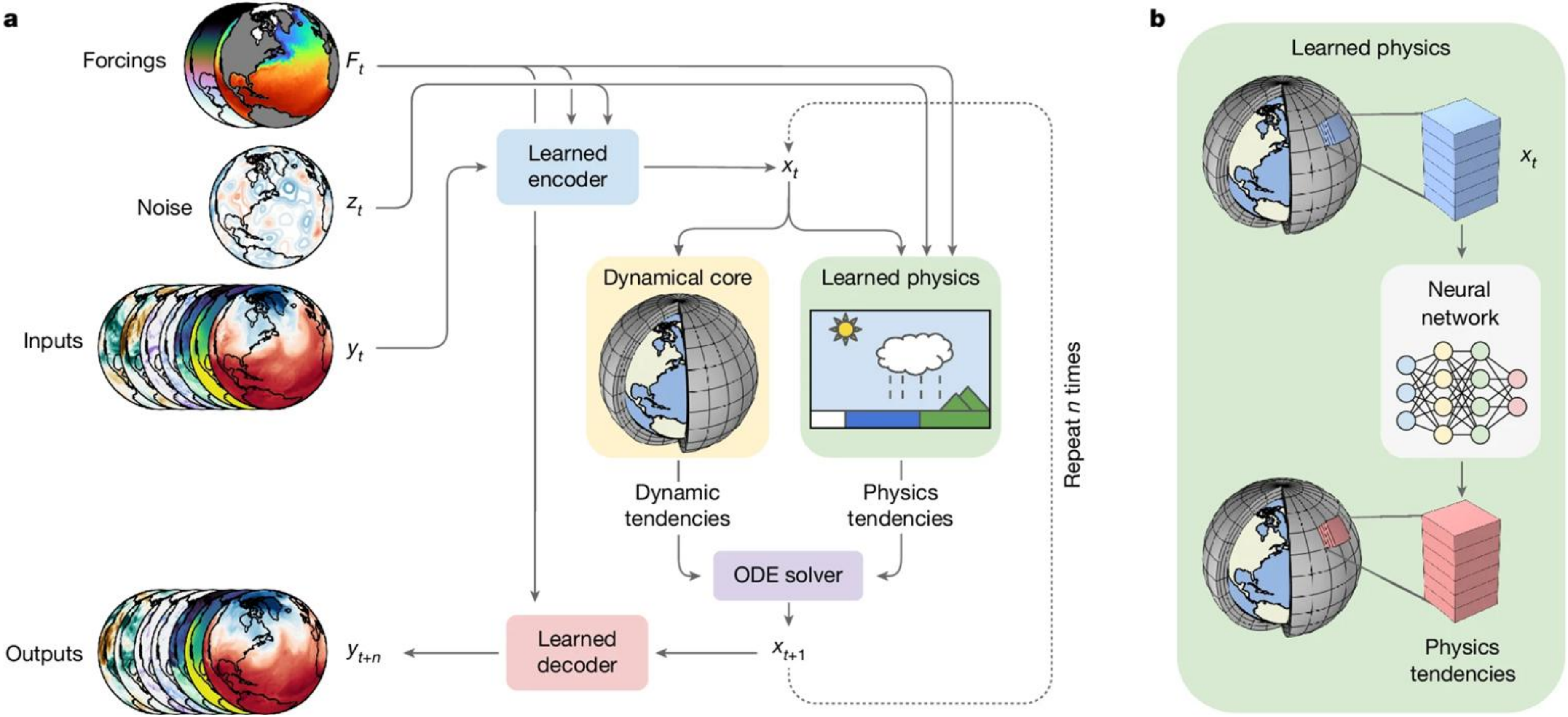


Son et al, 2021



source: Price et al. 2025

NeuralGCM setup - *includes a dynamical core* - aims to tackle spectral bias challenge with learned physics - maintain physical consistency in the subgrid scales



Source: [Kochkov et al. (2024) Nature]

AIFS ensemble tackles spectral bias with improved training metrics - training probabilistically instead of optimizing for deterministic total error

What is **continuous ranked probability score (CRPS)**?

Scoring rule that compares **full predicted probability distribution** to observed outcome, rewarding predictions that capture both large-scale trends and small -scale variations

Training include variance requirements - reflecting how confident a model is about the small- vs large- scale patterns

AIFS-ENS:

Probabilistic training of AIFS:

$$
\begin{aligned}
\mathrm{afCRPS}_\alpha &:= \alpha\,\mathrm{fCRPS} + (1-\alpha)\mathrm{CRPS} \\
&= \frac{1}{M}\sum_{j=1}^{M}|x_j - y| - \frac{M-1+\alpha}{2M^2(M-1)}\sum_{j=1}^{M}\sum_{k=1}^{M}|x_j - x_k| \\
&= \frac{1}{M}\sum_{j=1}^{M}|x_j - y| - \frac{1-\epsilon}{2M(M-1)}\sum_{j=1}^{M}\sum_{k=1}^{M}|x_j - x_k|
\end{aligned}
$$

AIFS ENS is optimized with a modified version of CRPS to make the scoring fair, accounting for ensemble size - see intro to AIFS ENS for more details

Some differences between AIFS single vs. latest ensemble version
How models are trained matters - improvements with CRPS loss

AIFS ensemble training approach

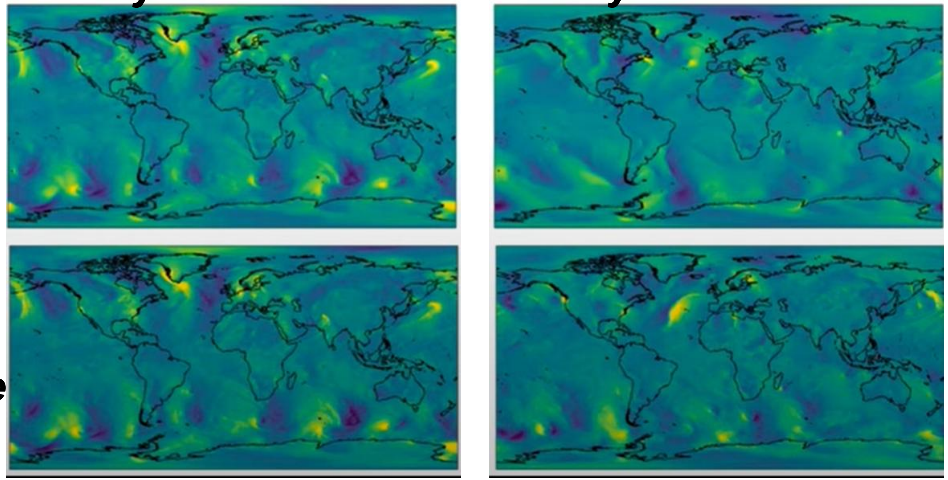
- Trains an ensemble of forecasts gets an injection of noise
- Trained on fair continuous ranked probability score (CRPS) based loss – fair CRPS accounts for number of ensemble members used – **can train AIFS with 2 members only**

Result: CRPS loss not possible for the model – so the model has not lost the small scale of variability to fulfill the training target by Day 10.

Day 1 Forecast ***Day 10 Forecast***

AIFS Single

AIFS Ensemble



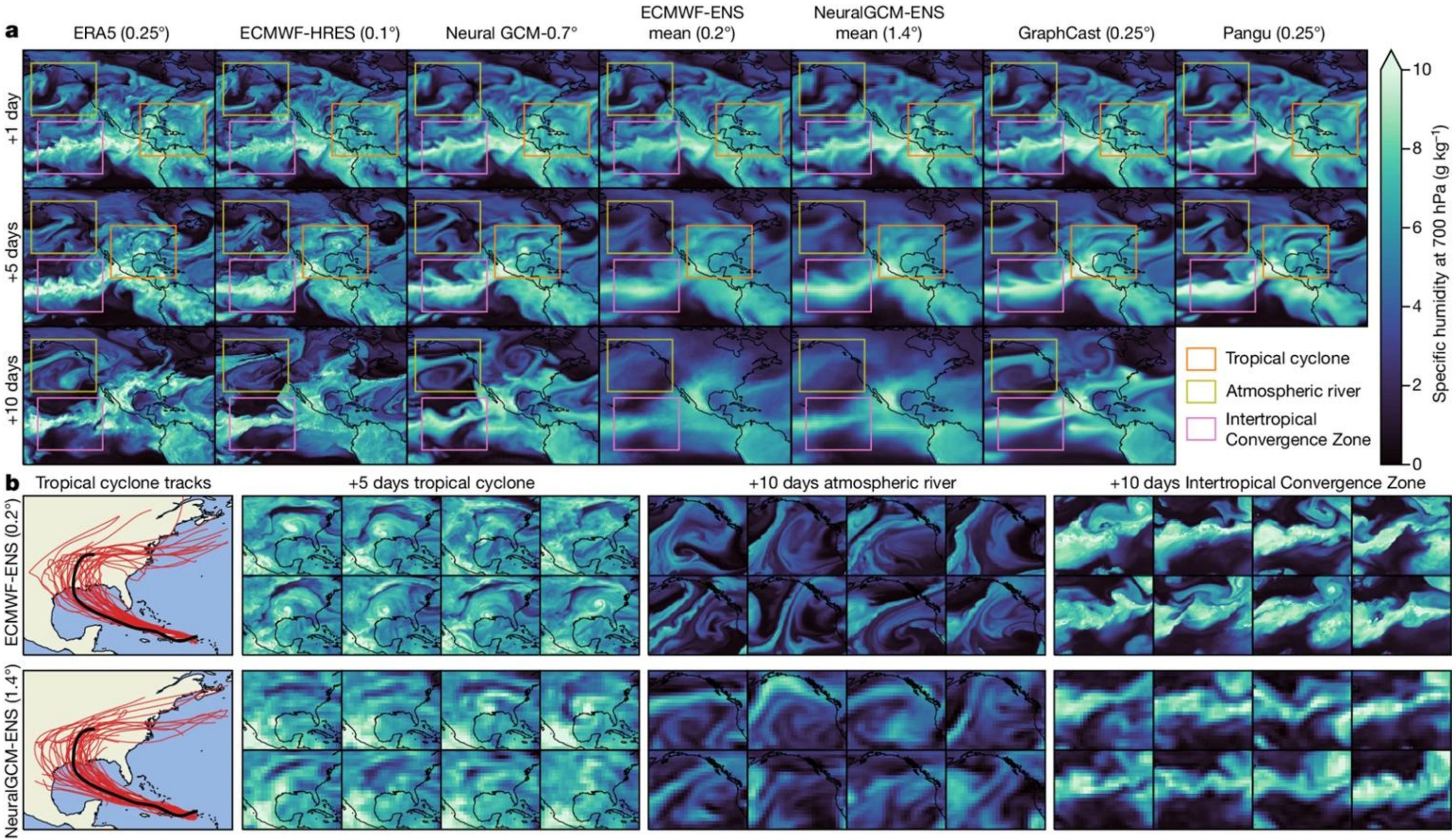
less blurry by day 10, more physical realism

Note, probabilistic version of NeuralGCM also trained to minimize CRPS to reduce spectral bias

Source: ECMWF, see recent seminar on AIFS ENS for more information on recent advances

AIM FOR SCALE | WEATHER PACKAGE

Spectral bias - a thorn in AI models - NGCM big push to cut down on spectral bias (a reason for the ‘blurriness’)



Source: Kochkov et al. (2024) Nature

Practical Considerations for Using AI Models

- Benchmark! Benchmark! Benchmark!
- Understanding the limitations: Data and loss function!

    AI is powerful but not magical: it cannot forecast things it has not seen!*

- Real-time forecasts challenges: see Adam's talk
- Blending (multi AI models, NWP, other data): see the monsoon's talk
- Localization: see DeepMind's talk
- Downscaling: see Forecast-in-Box and G42 talks
- Understanding "uncertainties" of the forecasts

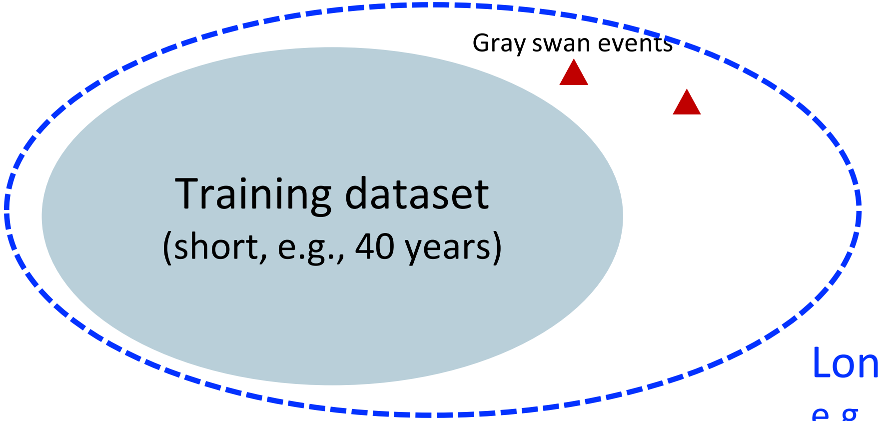
    Area of research across AI: see Souhaib's talk

AIM FOR SCALE | WEATHER PACKAGE

Can AI Predict Events Rarer and Stronger than What is Seen in the Training Set?

Can they extrapolate at the distributions' tails?

Gray swans (AI+climate): Physically possible weather extremes (for a given climate) that have not occurred in the often short training sets



Longer dataset
e.g., 1000 years

Black swan event

Credit: EOS 2022

Can AI Predict Events Rarer and Stronger than What is Seen in the Training Set?

Can they extrapolate at the distributions' tails?

$$\boldsymbol{X}(t+\Delta t) = \mathbf{NN}(\boldsymbol{X}(t), \boldsymbol{\theta})$$

$$\mathcal{L} = \|\boldsymbol{X}(t+\Delta t) - \mathbf{NN}(\boldsymbol{X}(t), \boldsymbol{\theta})\|_2$$

NO!

- Rare events cause "data imbalance" -->
  they do not contribute to the loss function

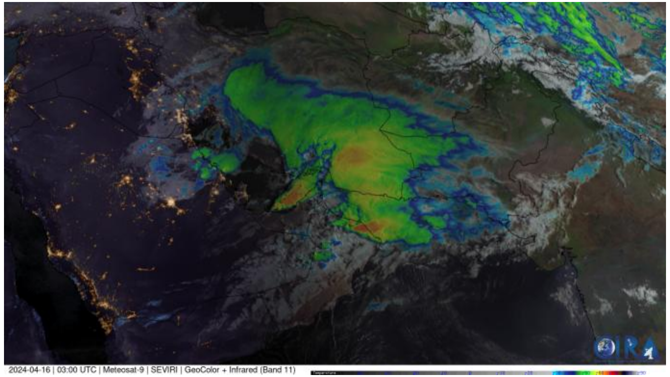
- So rare absent from training set: AI cannot do out-of-distribution generalization (extrapolation)

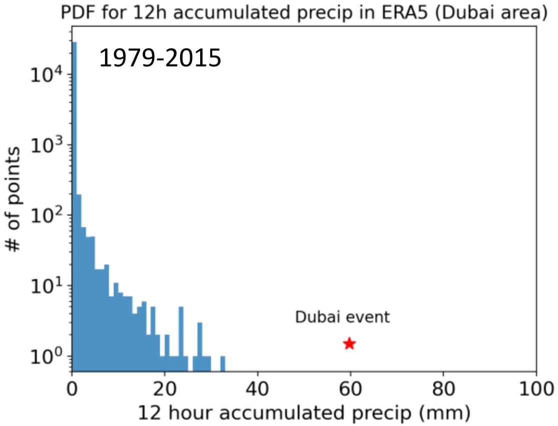
YES!

- AI models learn atmospheric dynamics
  (Hakim & Masanam, 2024 AIES; Rackow et al, 2024 …)

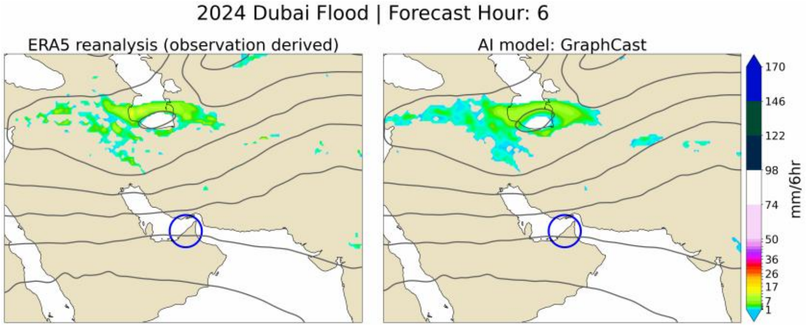
April 2024 Rainfall in UAE: A Gray Swan Event

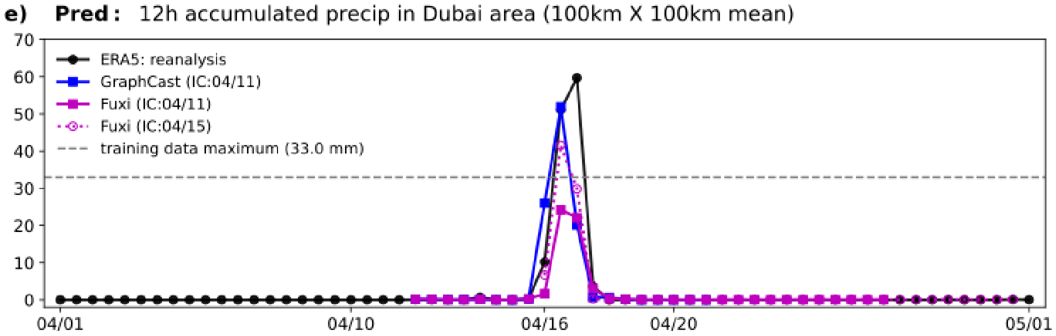
Can AI weather models predict such an unprecedented event?



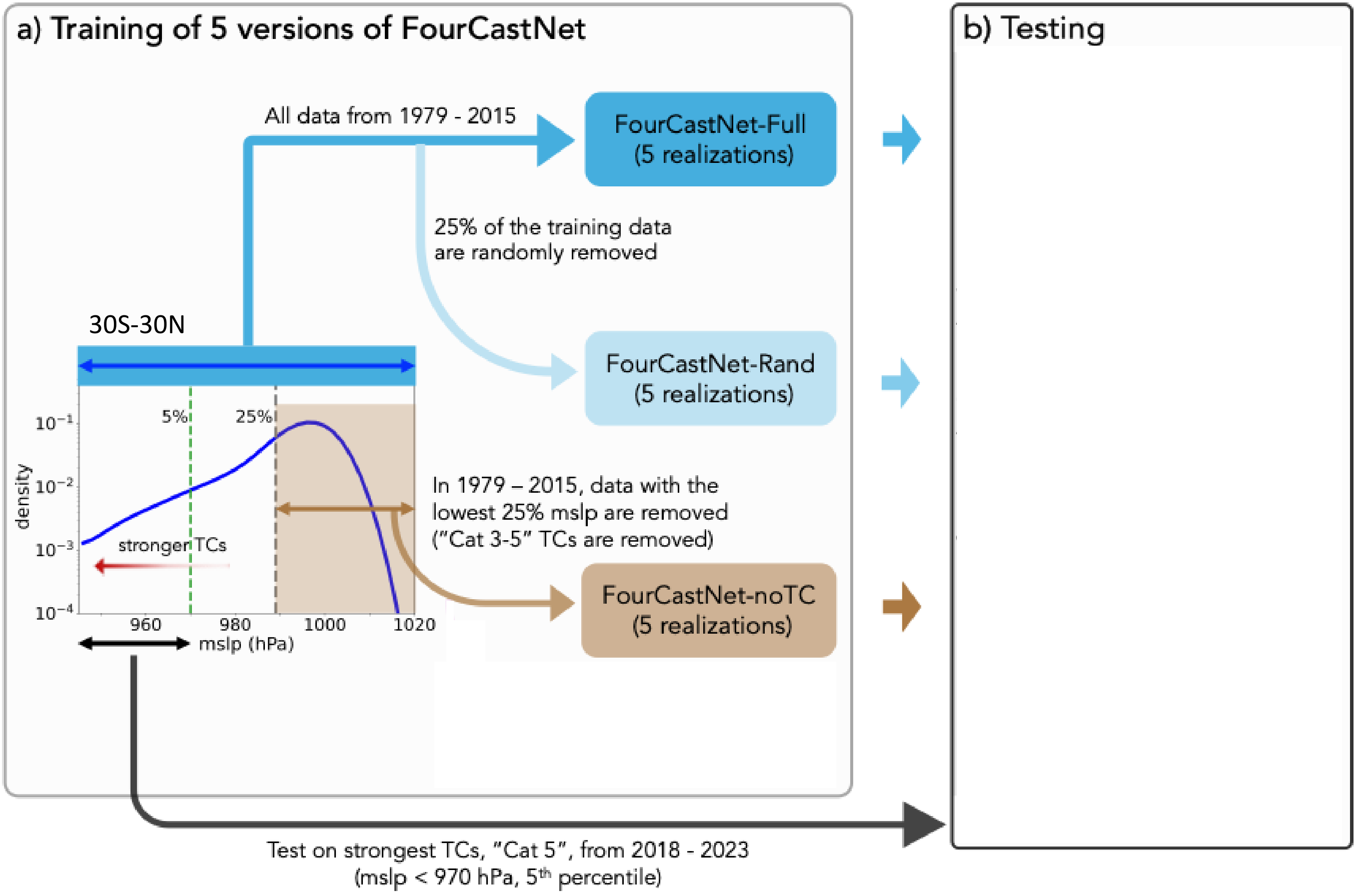


Sun et al. (2025)
http://arxiv.org/abs/2505.10241



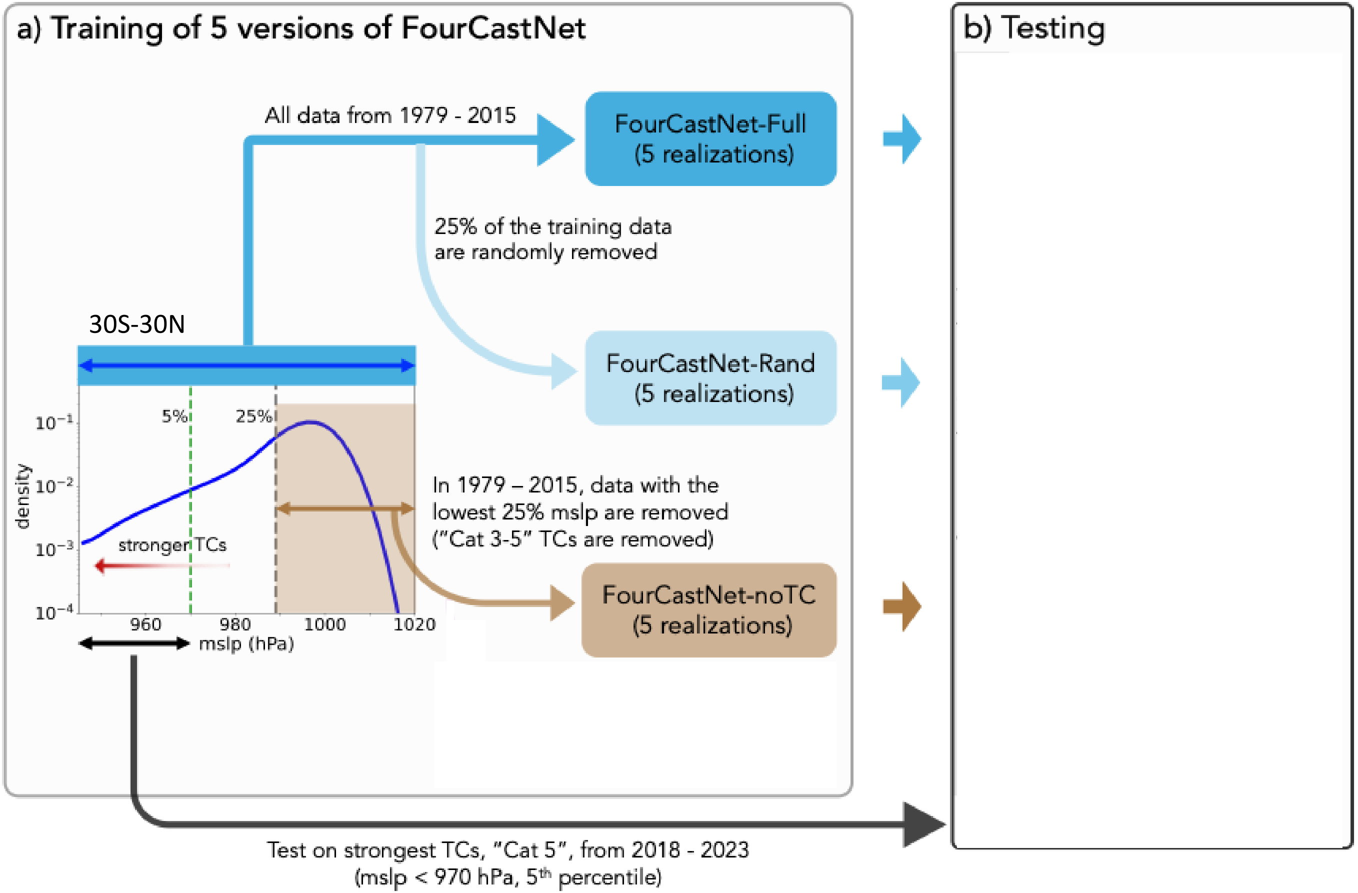


Controlled Experiments with FourCastNet: Rarest Extreme Events are Removed from Training Set



Controlled Experiments with FourCastNet

Can AI weather models predict gray swan tropical cyclones (TCs)?



Why this should not work?

- Rare events cause "data imbalance"
- AI cannot do out-of-distribution generalization (extrapolation)

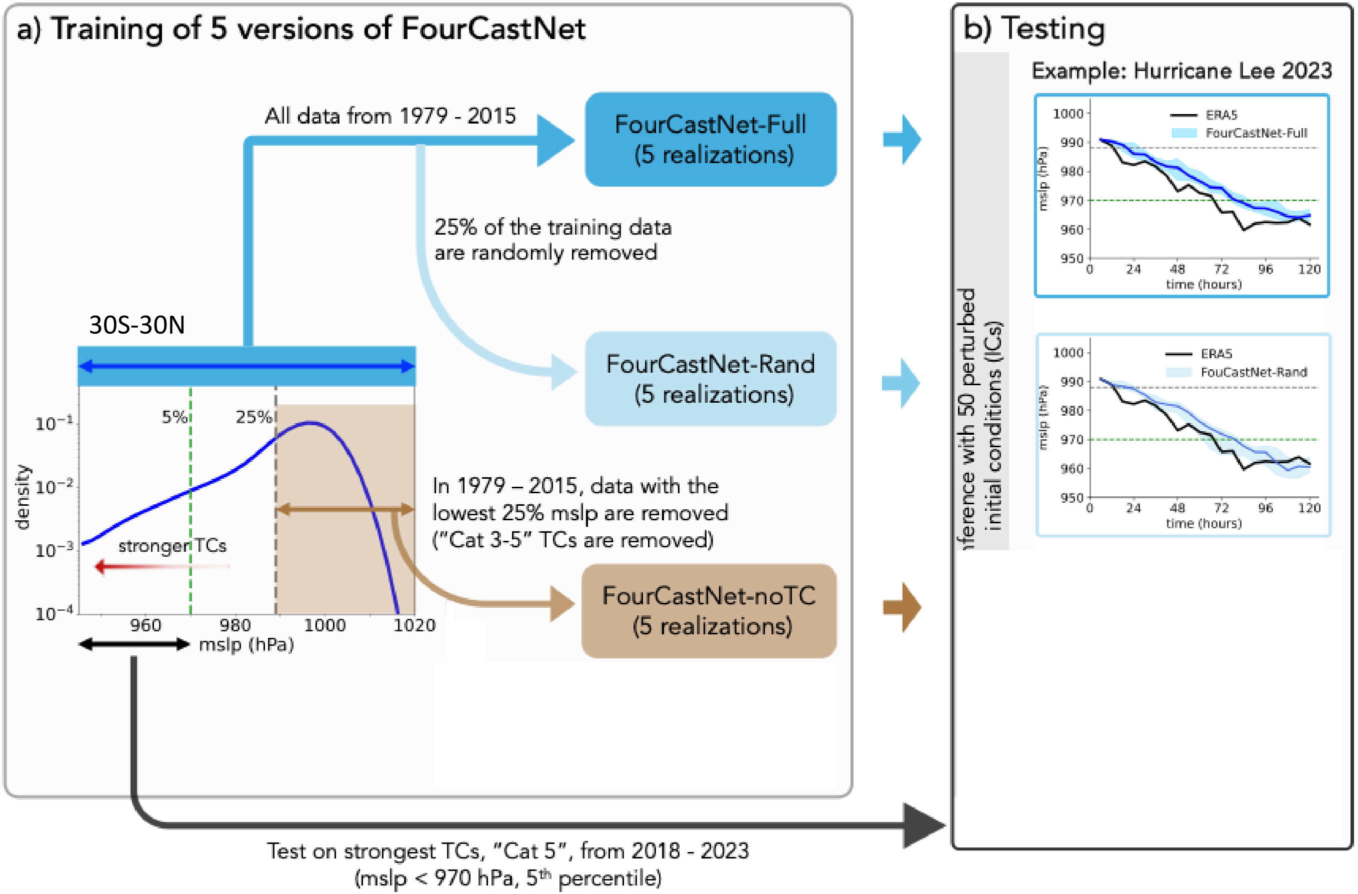
Why this might work?

- Learning dynamics vs memorization

- FourCastNet might learn unseen Cat-5 TCs (gray swans) from weaker (Cat 1-2) TCs in the training set

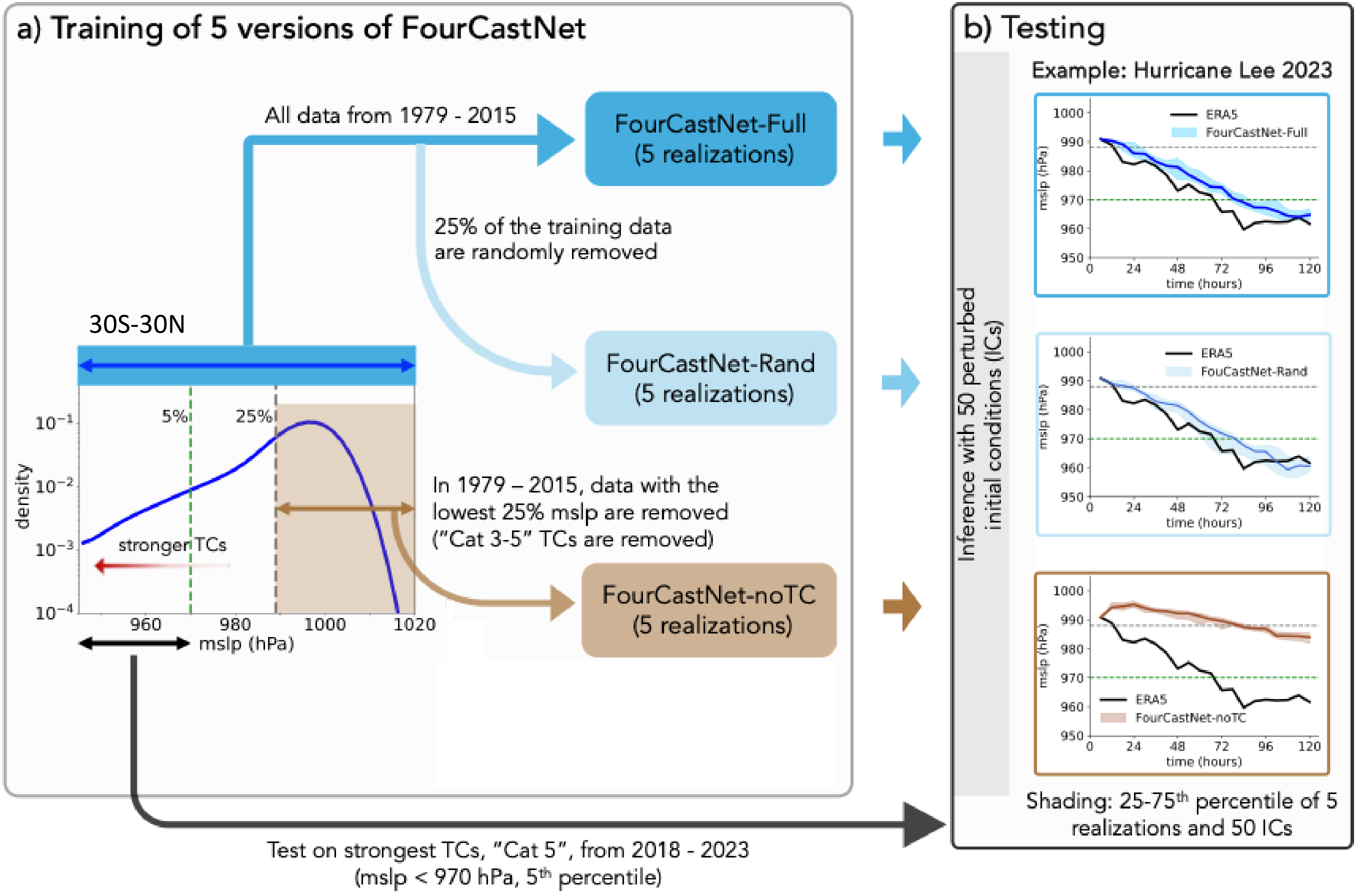
Controlled Experiments with FourCastNet

Can AI weather models predict gray swan tropical cyclones (TCs)?



FourCastNet Cannot Predict Gray Swans

No out-of-distribution generalization/extrapolation at the tails

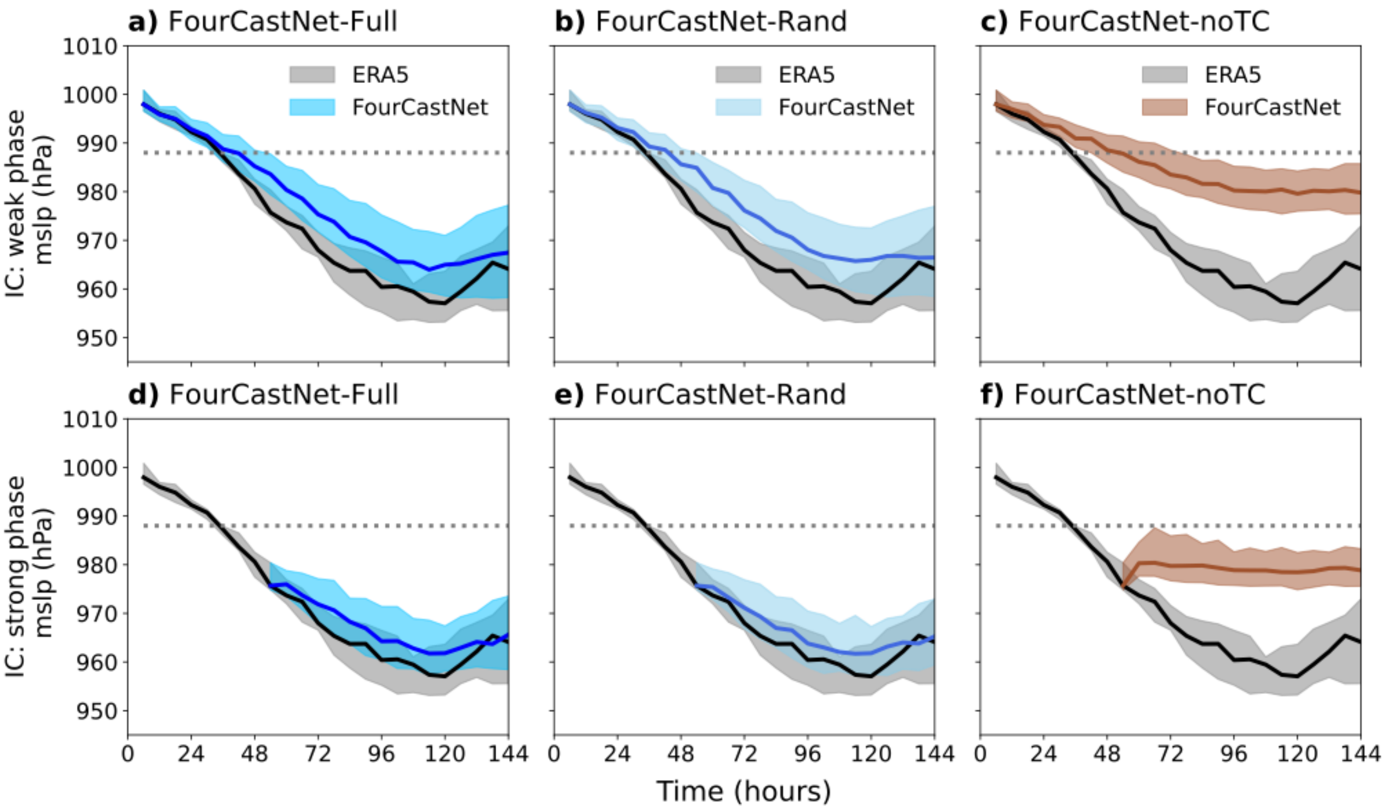


It did not work!

FourCastNet cannot learn unseen Cat-5 TCs (gray swans) from weaker (Cat 1-2) TCs in the training set

FourCastNet Cannot Predict Gray Swans: Gives False Negative!

Expected to be the case with all current AI models



Shading: 25-75 percentiles

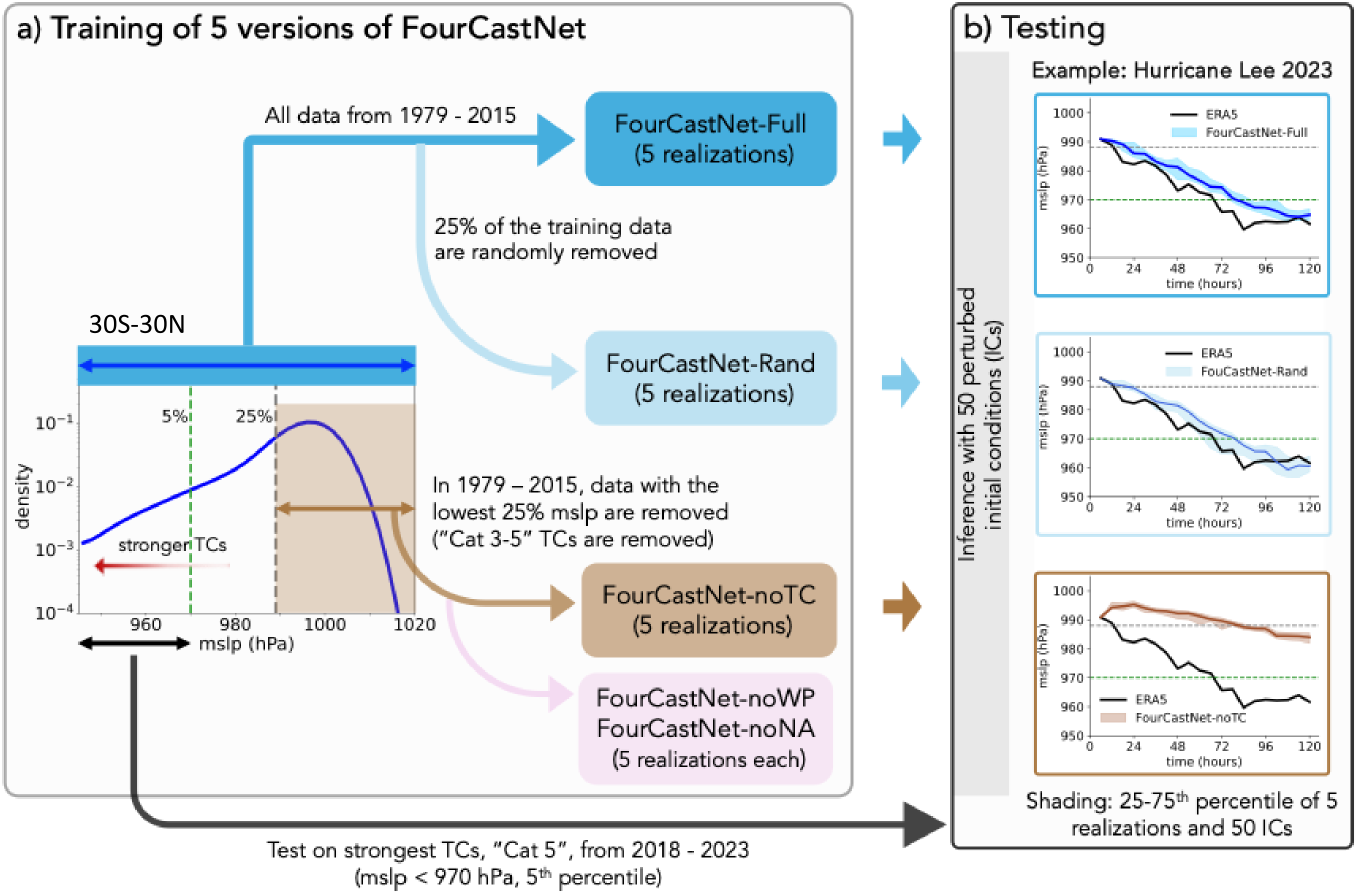
20 Cat 5 TCs in testing set (2018-2024)

50 perturbed ICs

5 realizations of each model

Controlled Experiments with FourCastNet

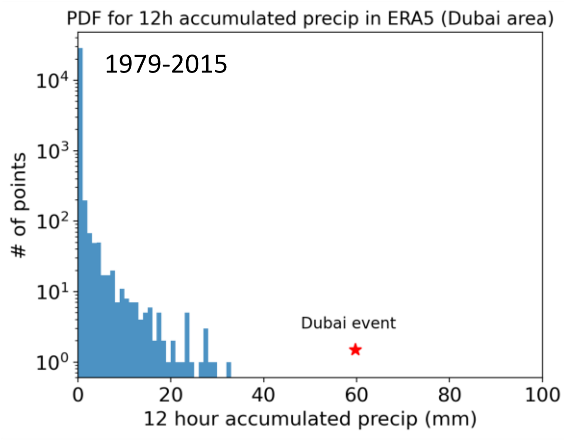
Why did the Dubai forecasts work?!

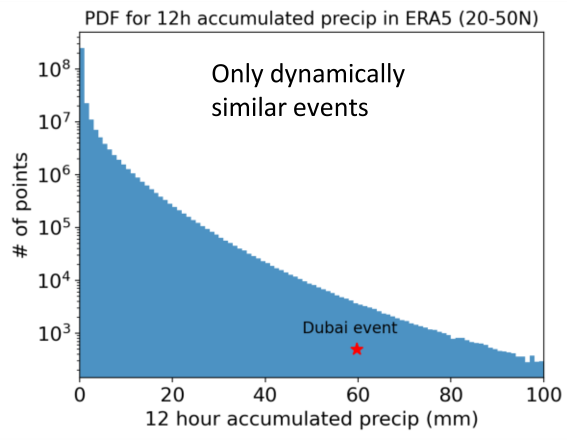


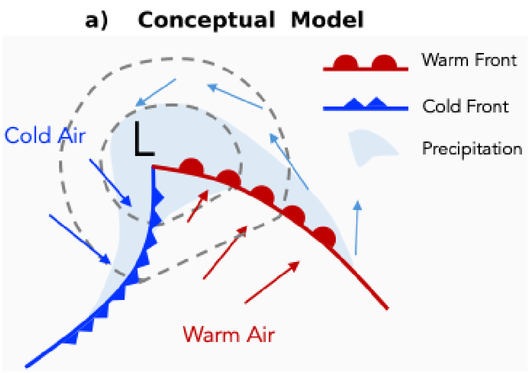
Can the AI model translate learning in one region to another for "dynamically similar" events? YES!

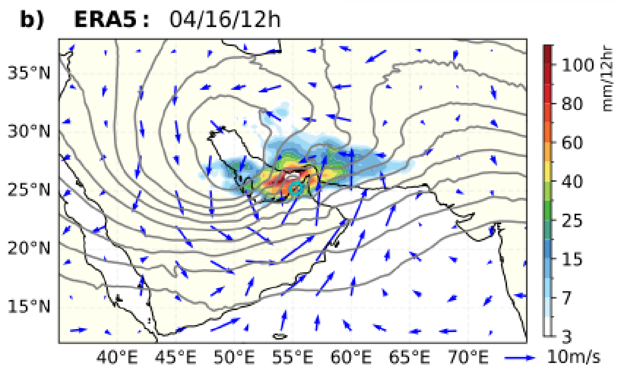
April 2024 Rainfall in UAE: A Regional Gray Swan Event

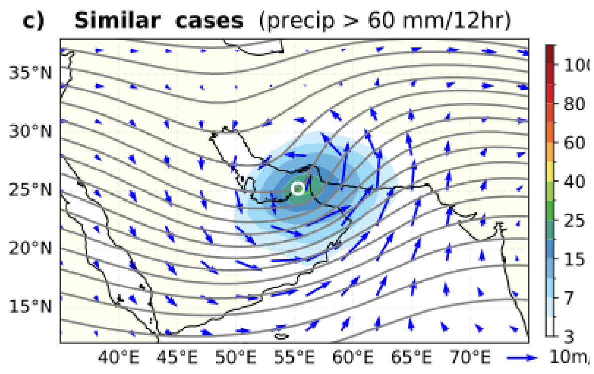
There are many stronger “dynamically similar“ events in the training set in other regions











SEP 22-26

AIM for Scale
AI Weather Training Program

AIM FOR SCALE | WEATHER PACKAGE

Setting up your AI Weather Lab

Tech Team


SEP 22-26
AIM for Scale
AI Weather Training Program
AIM FOR SCALE
WEATHER PACKAGE
JOINT
13:30 - 17:00