

Benchmark Datasets for ML in Weather & Climate



Image: Midjourney

Presenter: **Tom Beucler** (UNIL); Oct 8th, 2025, presentation at **oceanbench2025**

TCBench: M. Gomez, S. Ganesh, F. Tam, S. Darmon, I. Azizi (UNIL), M. McGraw (CIRA), S. Bourdin (Oxford)...

ClimSim: S. Yu, J. Lin, M. Pritchard, L. Peng (UCI), W. Hannah (LLNL), Z. Hu (NVIDIA), Kaggle winners...

HybridESMBench: F. Lan (UNIL), L. Bock, B. Gier, M. Schlund, V. Eyring (DLR), J. Lin (UCI), Z. Hu (NVIDIA)...

Atmospheric Physics + AI

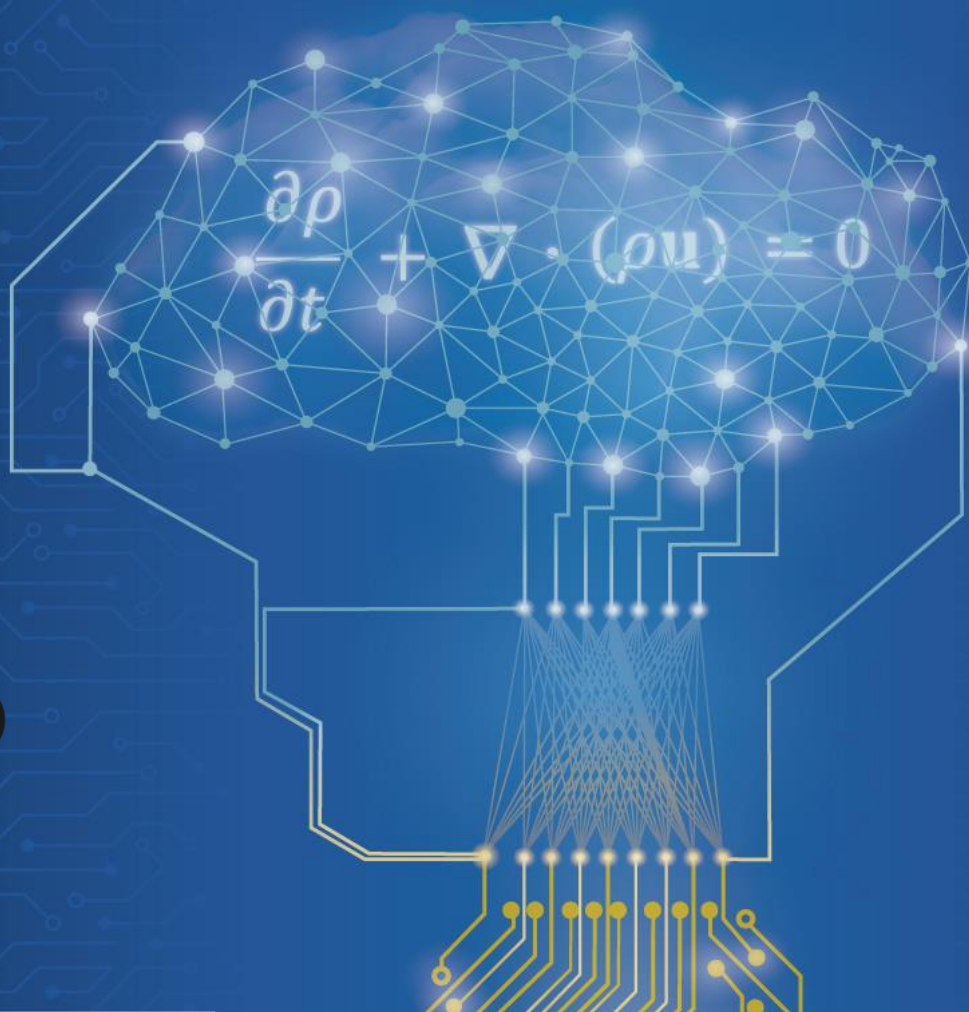
$$\frac{D\vec{v}}{Dt} + 2\vec{\Omega} \times \vec{v} = -\frac{\vec{\nabla} p}{\rho} - \vec{\nabla} \Phi$$

$$\mu \frac{dI}{dT} = I - B$$

$$\frac{de^*}{dT} = \frac{\mathcal{L}_v e^*}{R_v T^2}$$

Physics-Guided ML

Data-Driven Discovery



Extreme Weather Events

Forecasting,
Post-Processing,
Downscaling

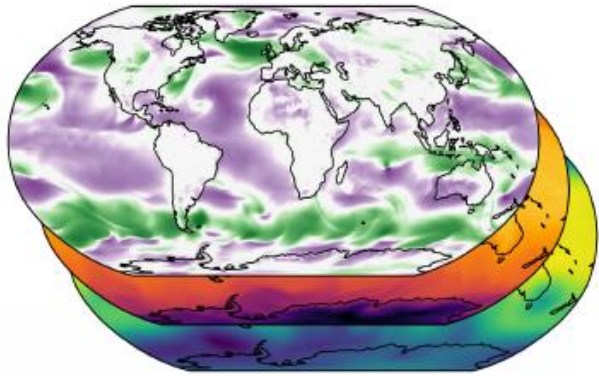
Earth System Modeling
(Parameterization)

Environmental
Data Science



Accelerated progress in data-driven weather forecasting can be traced back to WeatherBench...

a) Input weather state

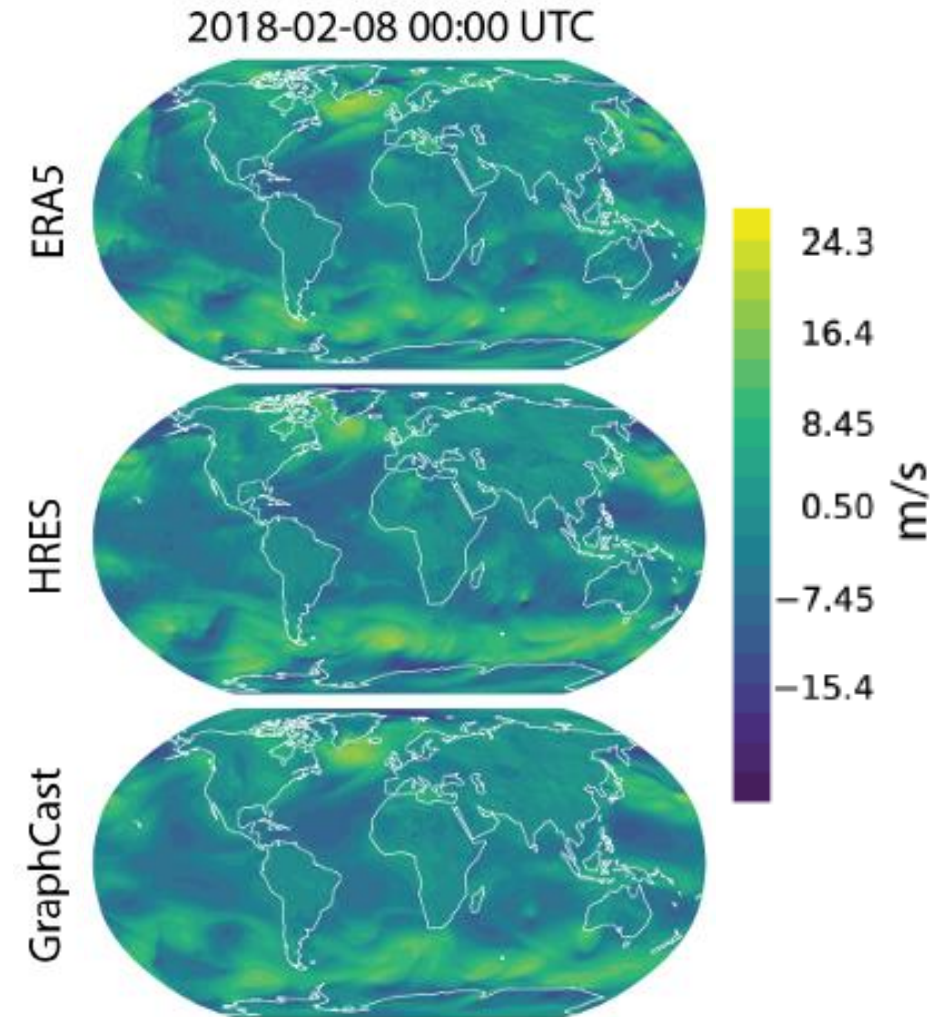


Time = Now

- Temperature
- Humidity
- Winds
- Geopotential h.
- ...

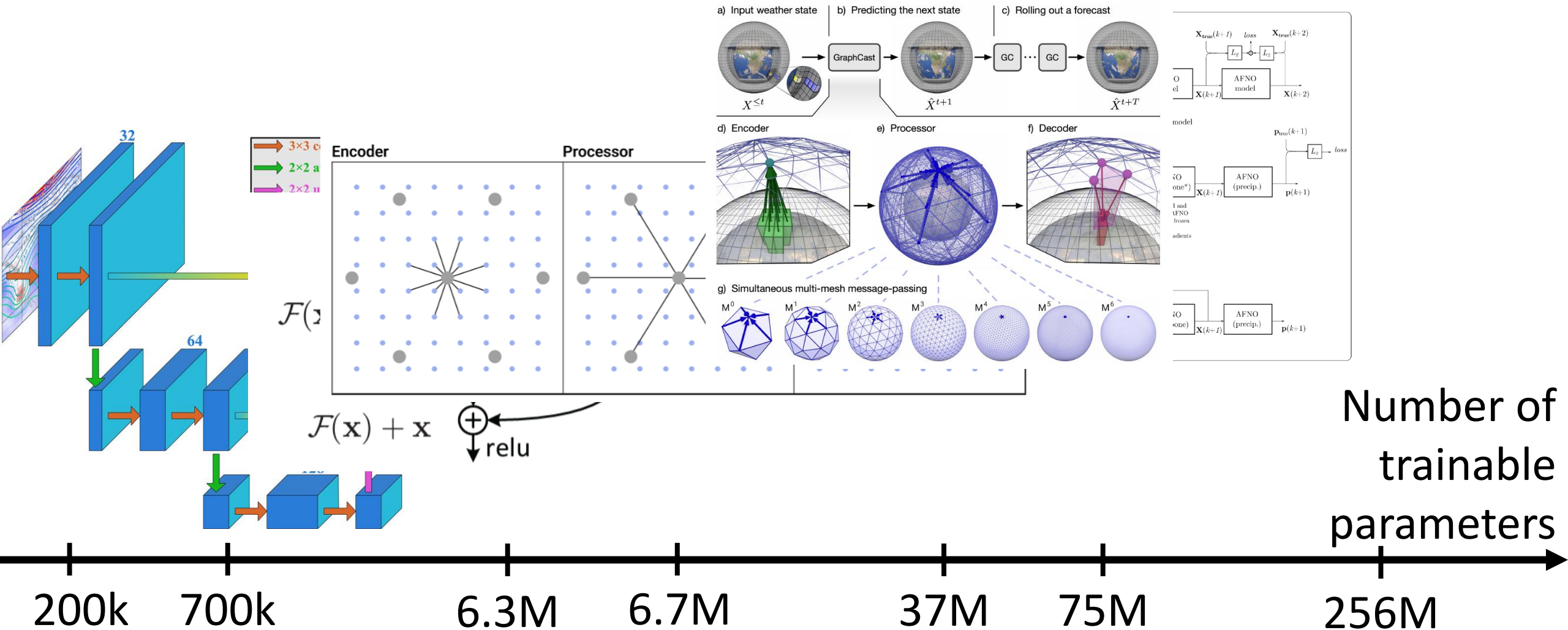
Time = Now+Lead Time

- Temperature
- Humidity
- Winds
- Geopotential h.
- ...

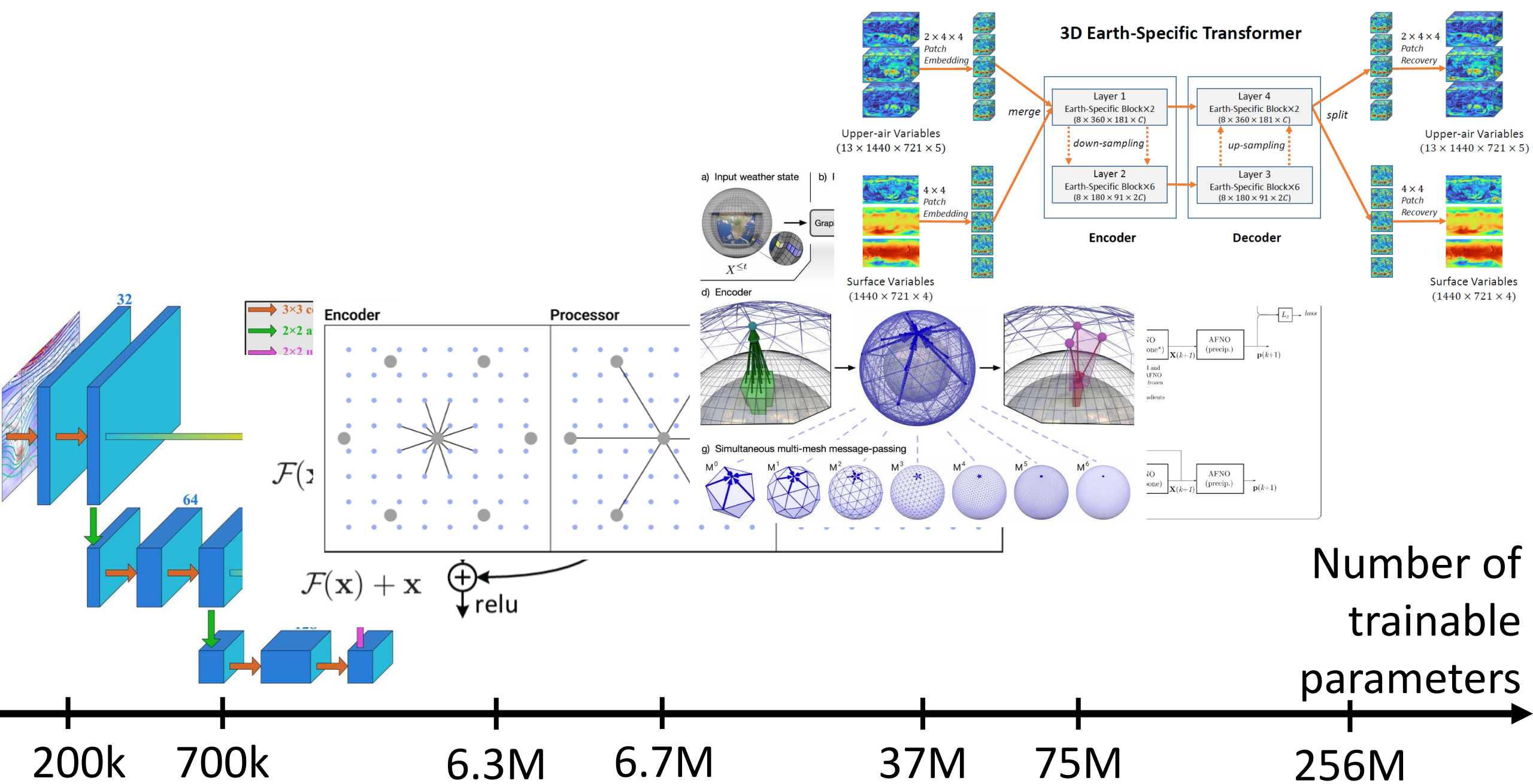


10m wind, 10-day forecast

Accelerated progress in data-driven weather forecasting can be traced back to WeatherBench...

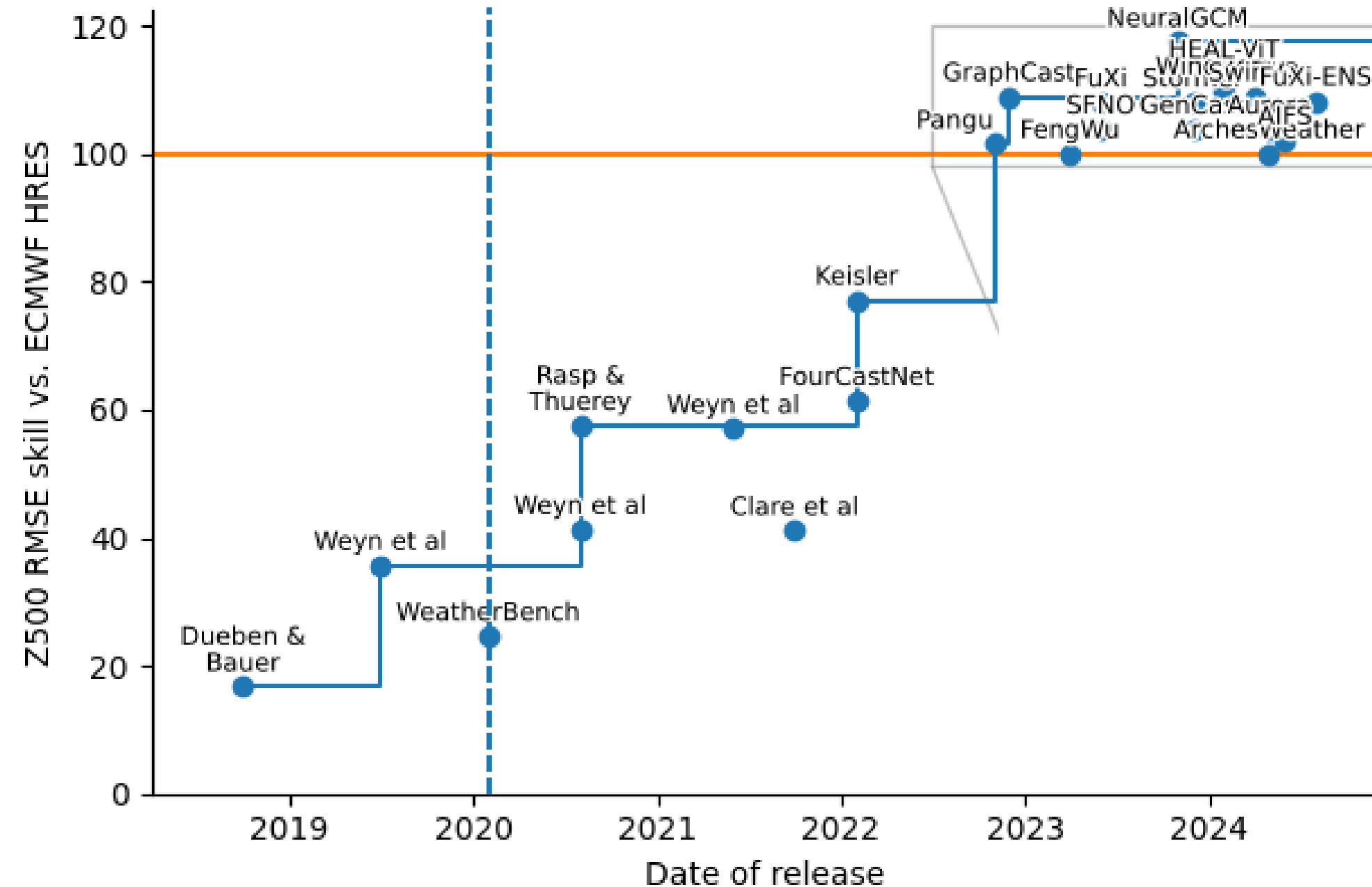


See: Weyn et al. (2019, 2020), Rasp et al. (2021), He et al. (2015), Keisler (2020), Pathak et al. (2022), Lam et al. (2022), Bi et al. (2022)



See: Weyn et al. (2019, 2020), Rasp et al. (2021), He et al. (2015), Keisler (2020), Pathak et al. (2022), Lam et al. (2022), Bi et al. (2022)

State of the art in AI weather prediction



Authors:
shoyer@google.com
srasp@google.com

Accelerated progress in data-driven weather forecasting can be traced back to WeatherBench...

Google Research WeatherBench Scorecards 2020 Deterministic scores Probabilistic scores Spectra FAQ

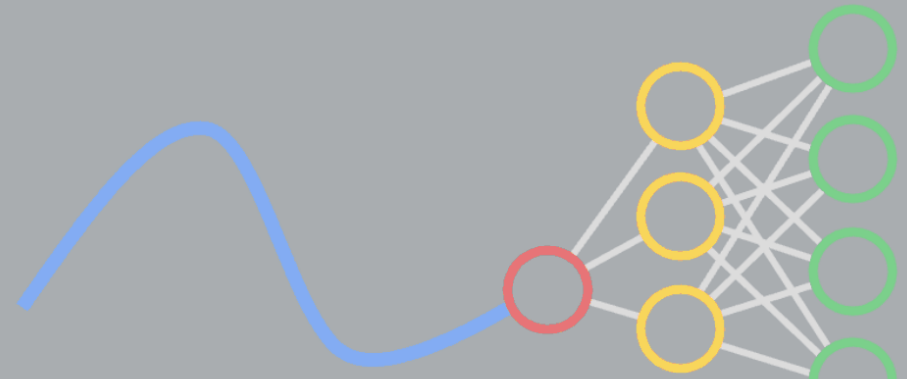
WeatherBench

A benchmark for the next generation of data-driven global weather models

Paper

Code

Blog



Overview

Weather forecasting using machine learning (ML) has seen [rapid progress](#) in recent years. WeatherBench is an open framework for evaluating ML and physics-based weather forecasting models in a like-for-like fashion.

This website contains up-to-date scores of many state-of-the-art global weather models with a focus on medium-range (1-15 day) prediction. In addition, the WeatherBench framework consists of our recently updated [WeatherBench-X evaluation code](#) and publicly available, cloud-optimized ground-truth and baseline [datasets](#), including a comprehensive set of [F2S](#) data for training ML models. For more information, visit our [FAQ](#).

QUICK LINKS

Paper

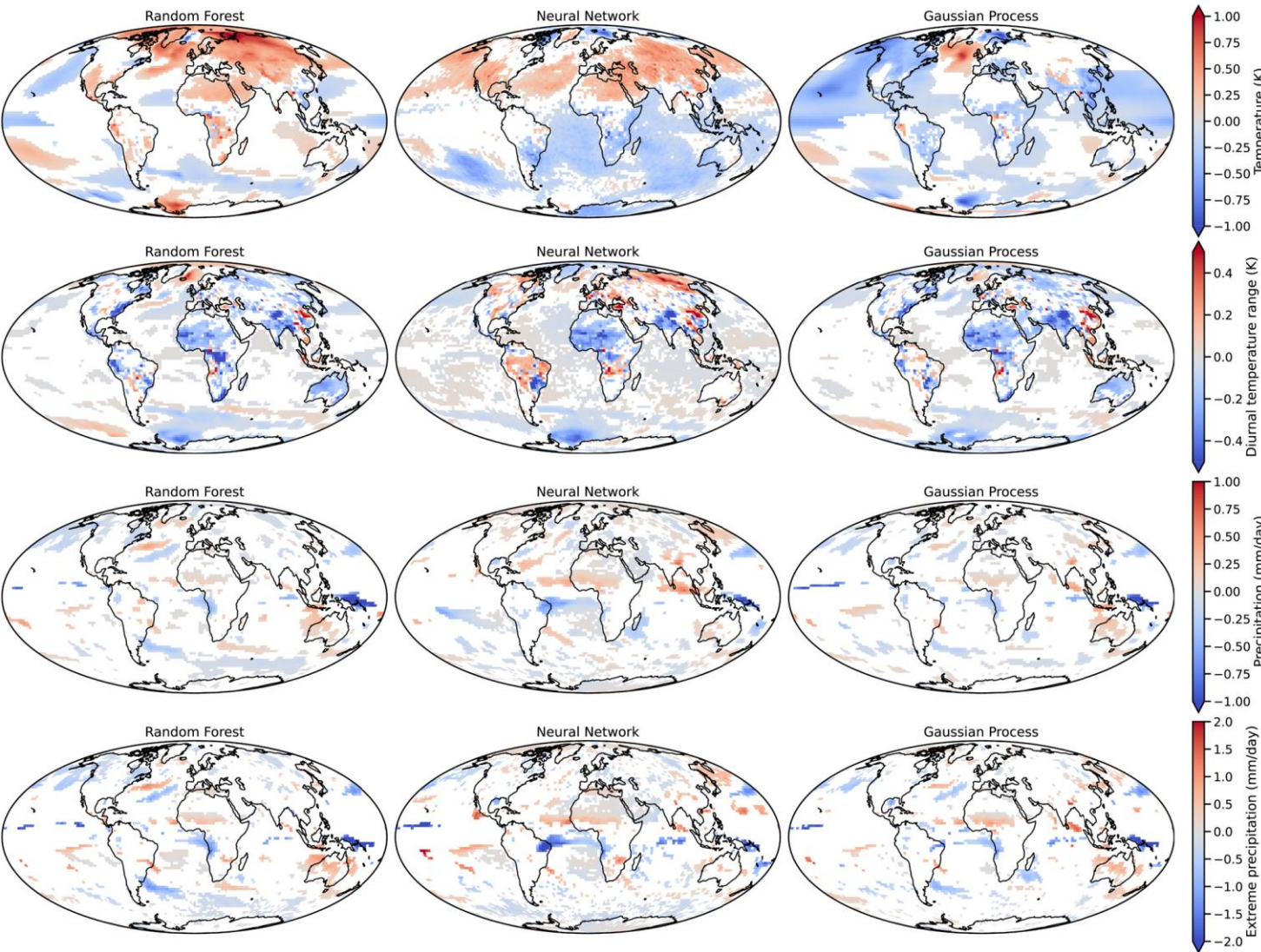
Code

Data

Share

Source:
WeatherBench
(Google)

...but interacting Earth system components warrants multiple benchmarks at the climate timescale



MIT News

ON CAMPUS AND AROUND THE WORLD

SUBSCRIBE

Simpler models can outperform deep learning at climate prediction

New research shows the natural variability in climate data can cause AI models to struggle at predicting local temperature and rainfall.

Adam Zewe | MIT News

August 26, 2025



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

NeurIPS 2024 Oral

1) TCBench

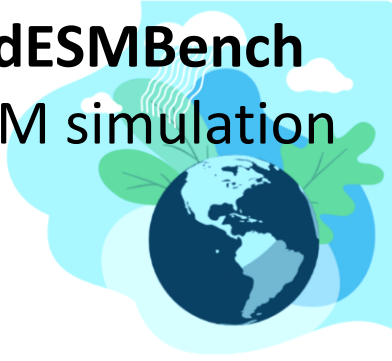
Global TC track and
intensity forecasting

2) ClimSim

Hybrid physics-ML
climate emulation

3) HybridESMBench

Hybrid ESM simulation



1) TCBench

Global TC track and
intensity forecasting

2) ClimSim

Hybrid physics-ML
climate emulation

3) HybridESMBench

Hybrid ESM simulation

Order 1 requirements

R1: Data available online without access restrictions

R2: Clear problem statement for meaningful task in atmospheric science

R3: Data input into high level open data science language provided

R4: Evaluation metrics defined analytically and in code

Order 2 requirements

R5: Simple example machine learning solution provided in code

R6: Visualisation and diagnostics provided in code

R7: Tests for physical consistency and explainability provided

R8: Benchmarks for the computational performance provided

1) TCBench

Global TC track and
intensity forecasting

2) ClimSim

Hybrid physics-ML
climate emulation

3) HybridESMBench

Hybrid ESM simulation

Order 1 requirements

R1: Data available online without access restrictions

R2: Clear problem statement for meaningful task in atmospheric science

R3: Data input into high level open data science language provided

R4: Evaluation metrics defined analytically and in code

Order 2 requirements

R5: Simple example machine learning solution provided in code

R6: Visualisation and diagnostics provided in code

R7: Tests for physical consistency and explainability provided

R8: Benchmarks for the computational performance provided

| Webcams MX

MEXICO

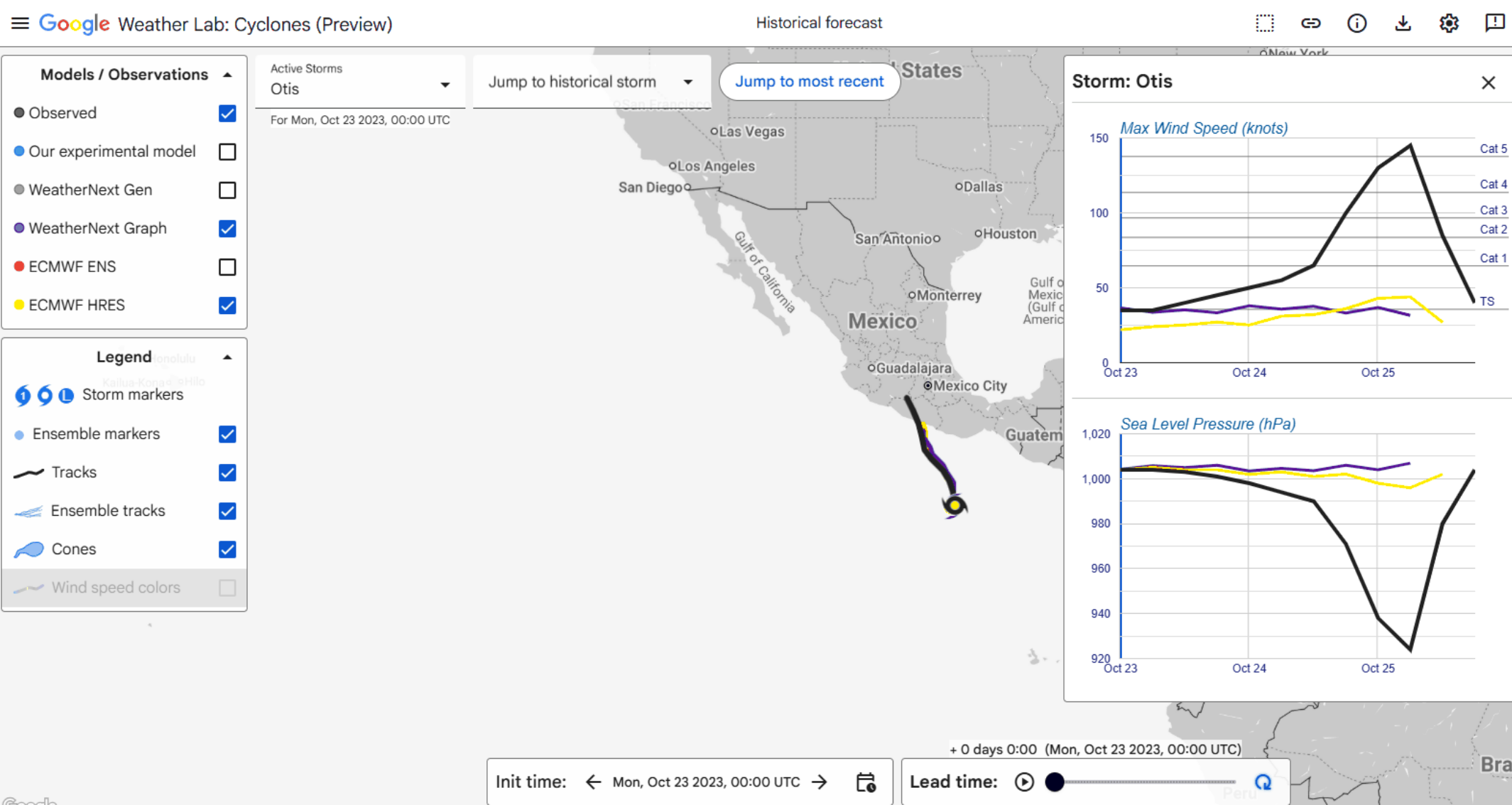
> BREAKING NEWS

OTIS HITS MEXICO AS 165 MPH CAT. 5 HURRICANE

 **MSNBC**

LIVE > 11:38 ^{AM} ET

Rapid intensification of hurricane Otis (Oct 23) exposed limits of both traditional & AI weather models



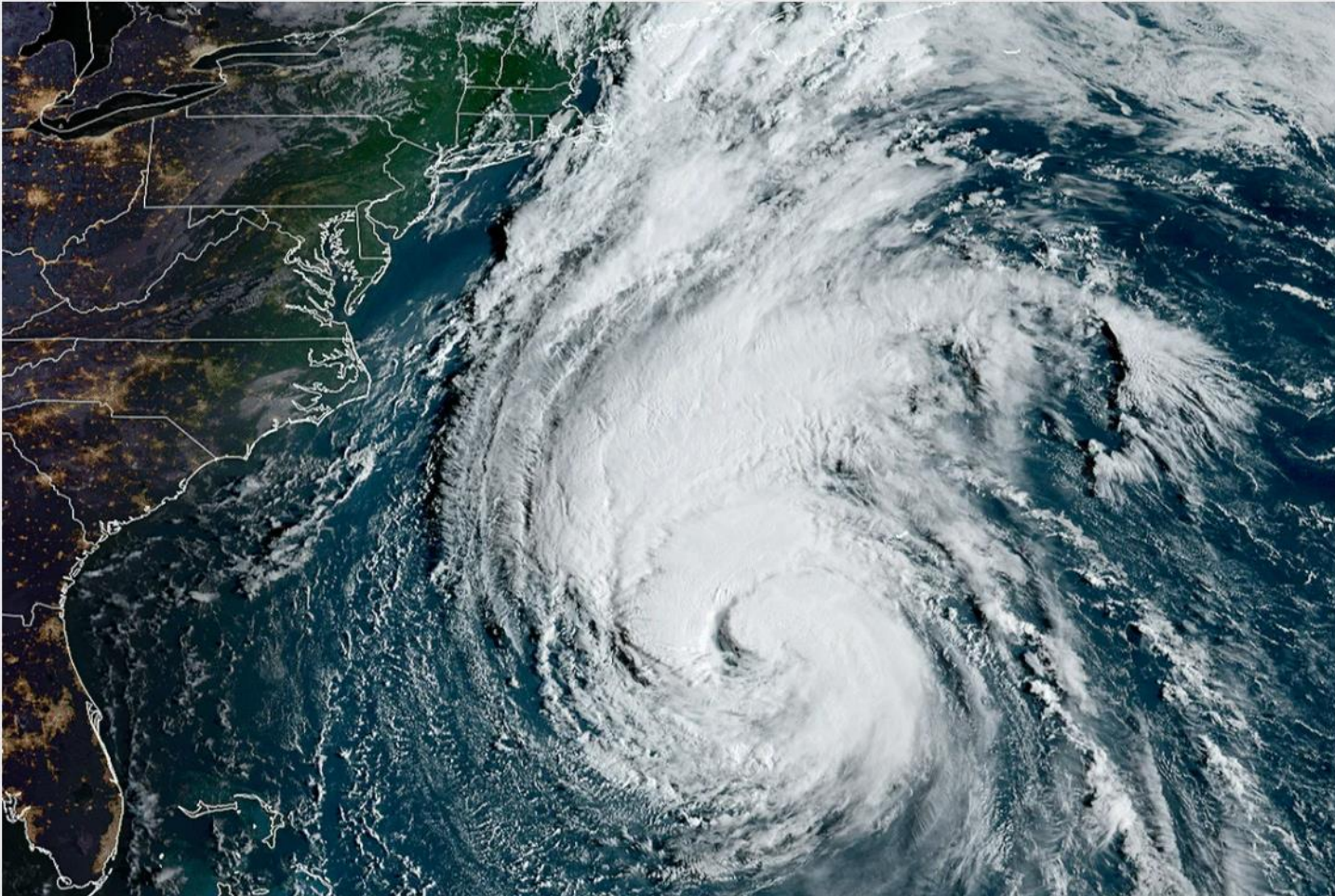
Source: Google
Weather Lab
(Oct 2025)

JANUARY 9, 2024 | 6 MIN READ

AI Weather Forecasting Can't Replace Humans—Yet

GraphCast and other artificial intelligence-based forecasting tools offer a whole new way to predict the weather, but they have limits

BY [LAUREN LEFFER](#) EDITED BY [ANDREA THOMPSON](#)



The sun rises on Hurricane Lee as it casts quite the shadow from New York City to Maine. [CSU/CIRA & NOAA](#)

*Source: Scientific American
(Jan 2024)*

Emergence of probabilistic AI forecasters holds promise for extremes...

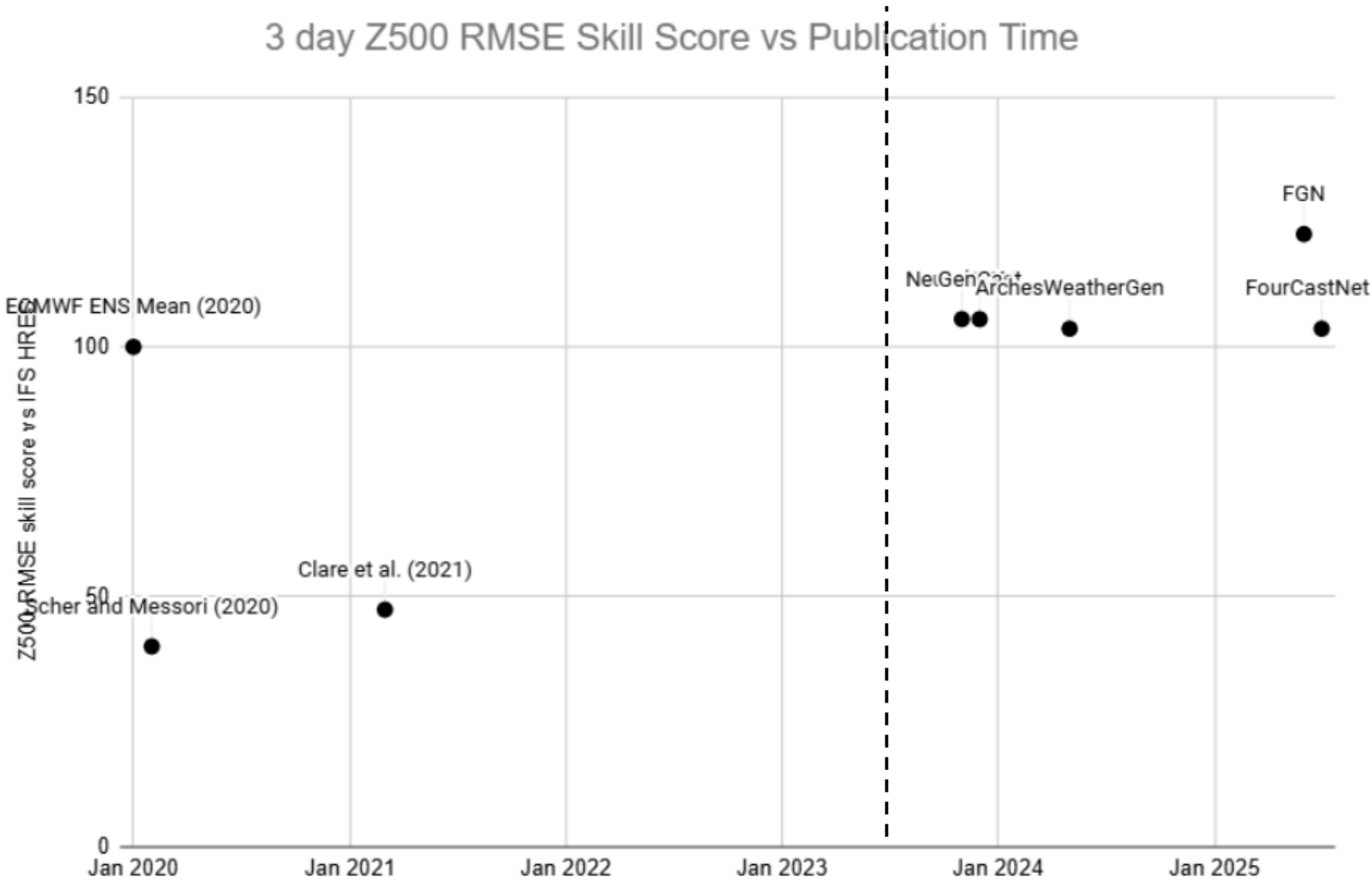
Yet

GraphCast and other artificial intelligence offer a whole new way to predict the weather,

BY LAUREN LEFFER EDITED BY ANDREA THOMPSON



The sun rises on Hurricane Lee as it casts quite the shadow from New York City to Maine. [CSU/CIRA & NOAA](#)



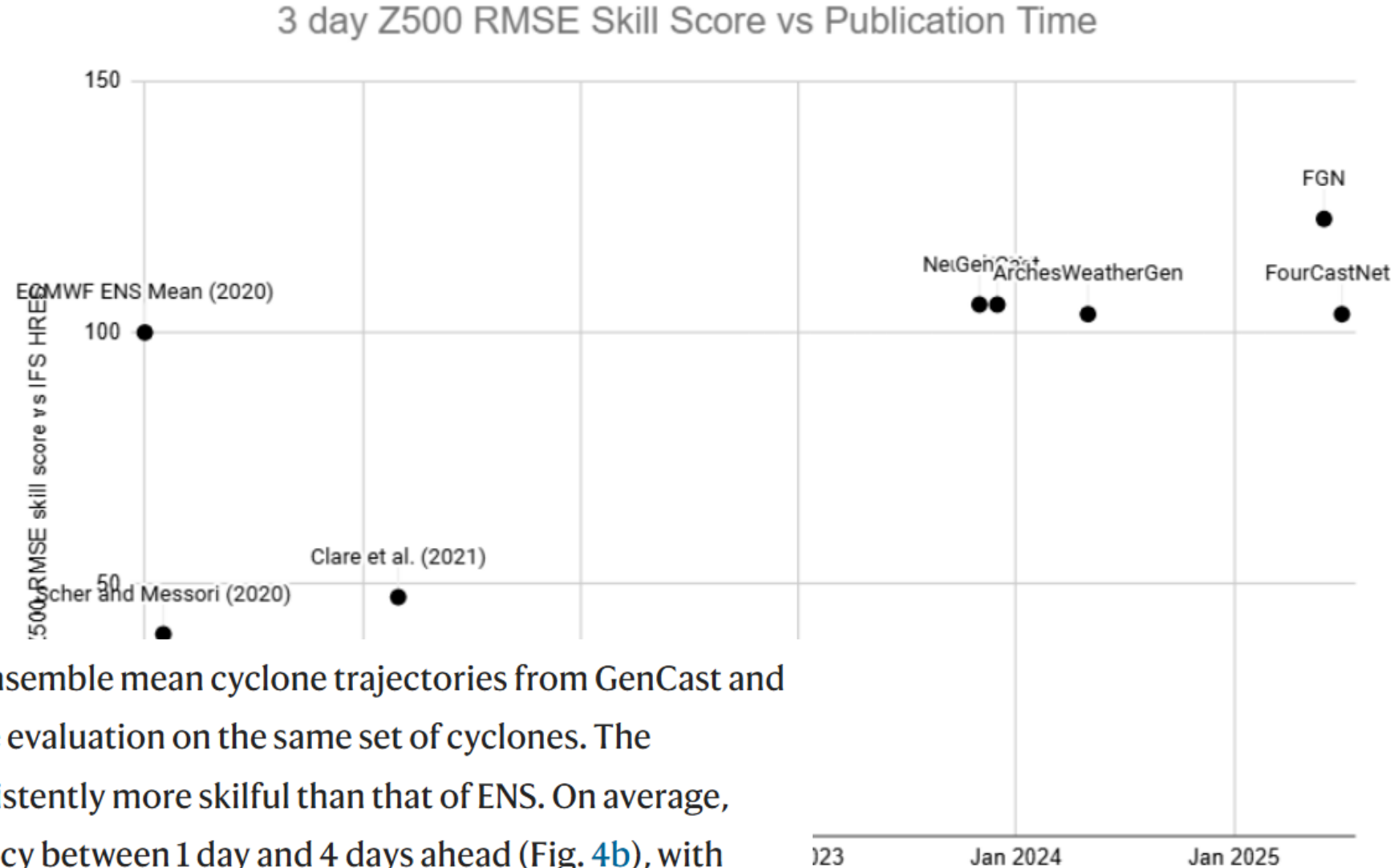
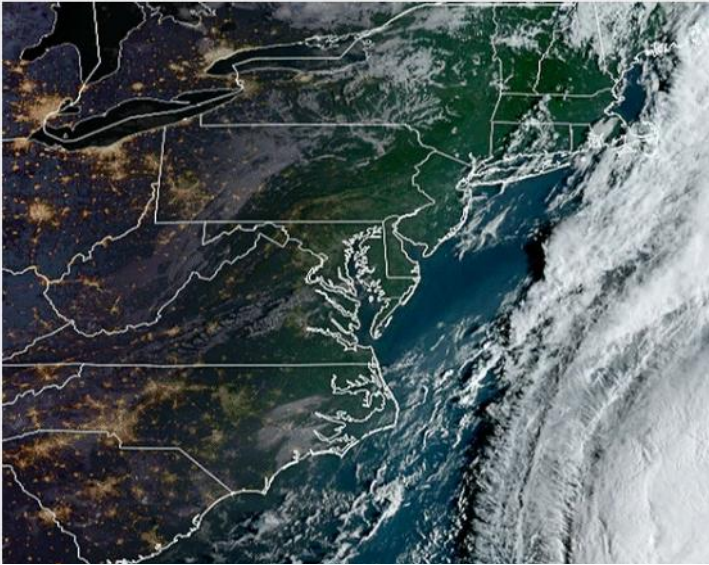
Sources: Scientific American (Jan 2024),
S. Hoyer & S. Rasp

...but how to objectively find best solutions with hand-picked tests?

Yet

GraphCast and other artificial intelligence
whole new way to predict the weather,

BY LAUREN LEFFER EDITED BY ANDREA THOMPSON



First, we evaluate the position error of ensemble mean cyclone trajectories from GenCast and ENS, using a pairing procedure to ensure evaluation on the same set of cyclones. The ensemble mean track of GenCast is consistently more skilful than that of ENS. On average, GenCast gives a 12-h advantage in accuracy between 1 day and 4 days ahead (Fig. [4b](#)), with significantly ($P < 0.05$) lower error between 12 h and 3.5 day lead times (inclusive, Supplementary Fig. [B9](#)).

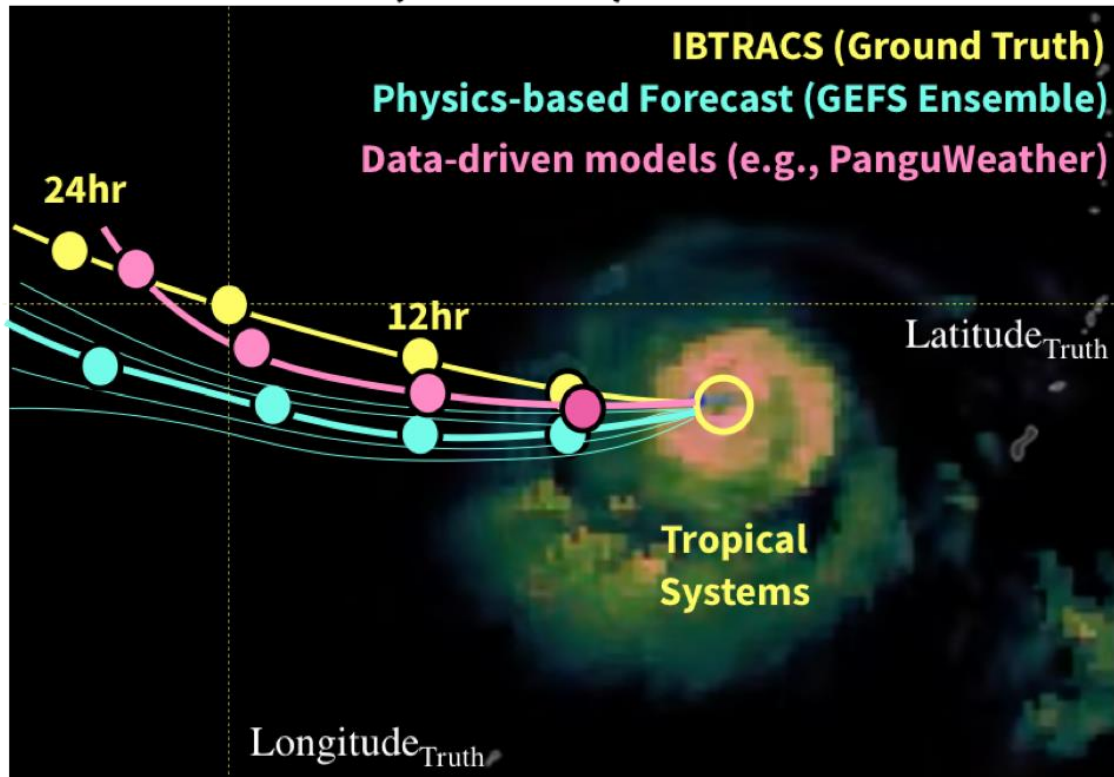
Sources: Scientific American (Jan 2024),
S. Hoyer & S. Rasp, Price et al. (2025)

TC forecasting framed as regression with known initial conditions

R2: Clear problem statement for meaningful task in atmospheric science

1. Forecasting task definition

Track (Location) Prediction

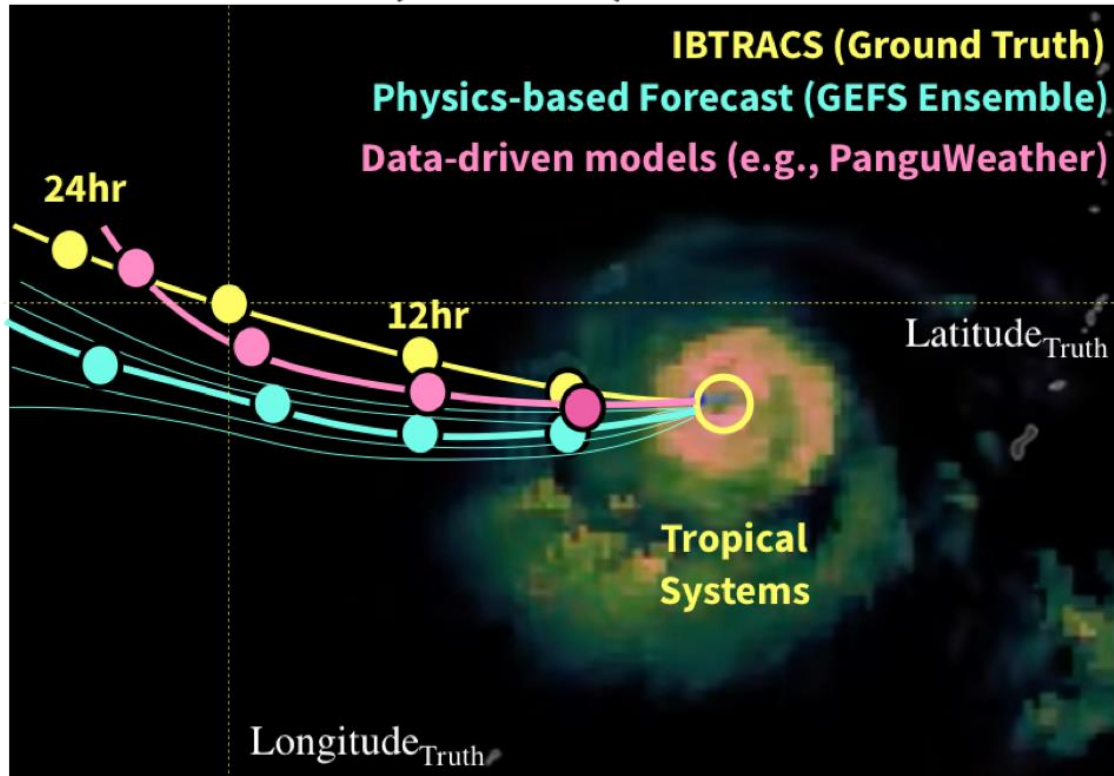


TC forecasting framed as regression with known initial conditions

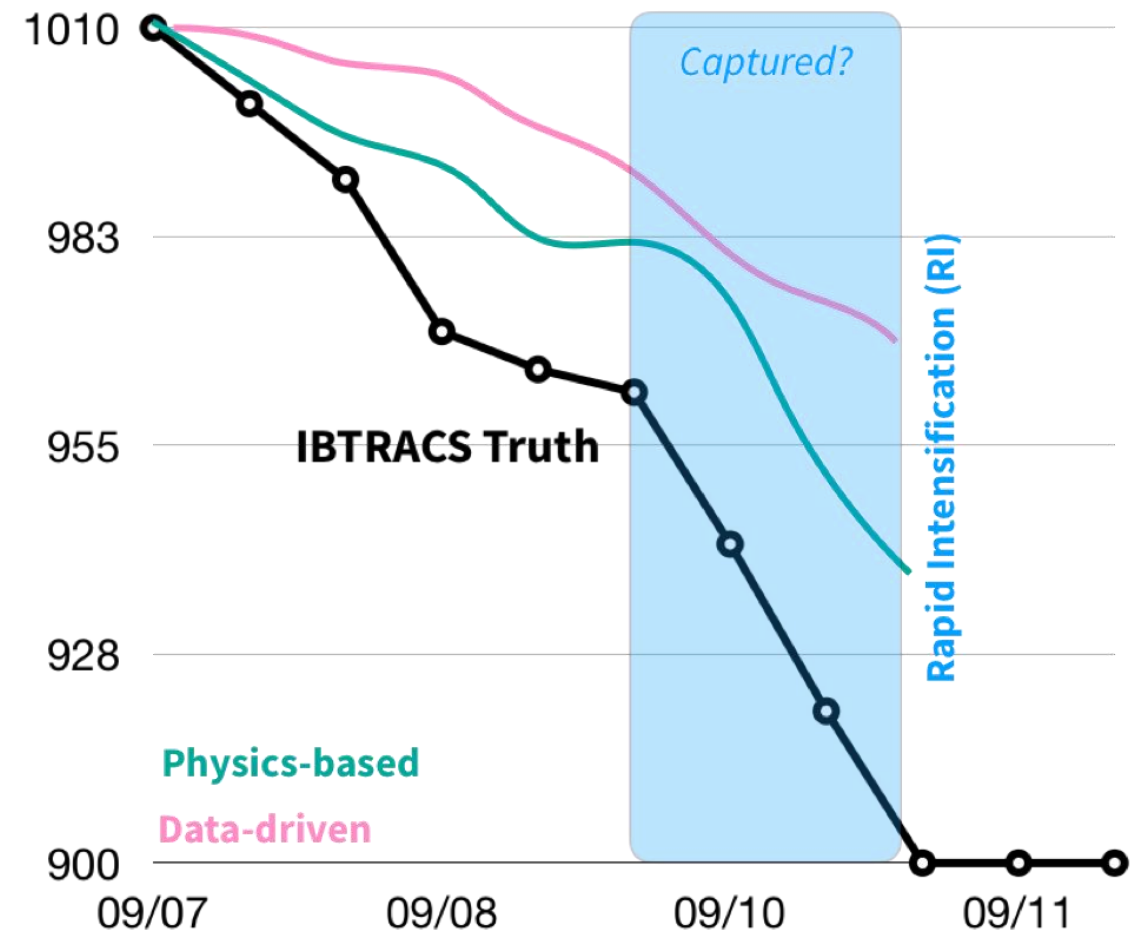
R2: Clear problem statement for meaningful task in atmospheric science

1. Forecasting task definition

Track (Location) Prediction



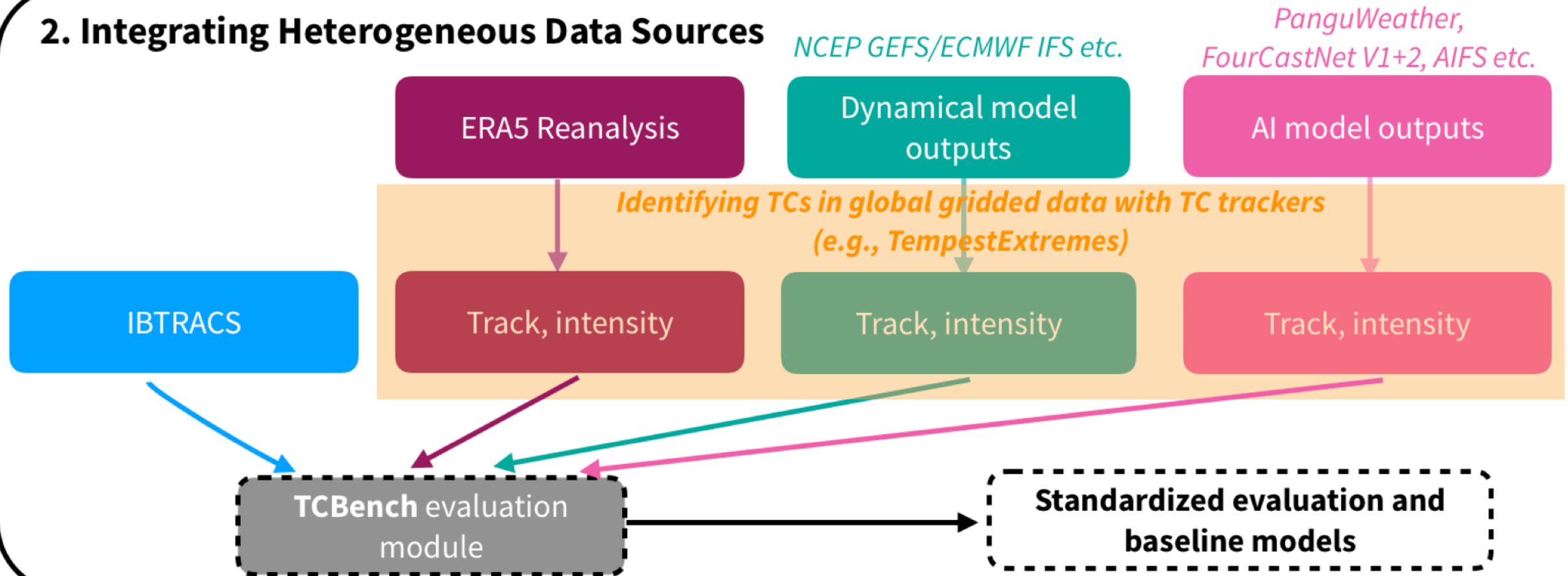
Intensity and RI Prediction



R1: Data available online without access restrictions

R3: Data input into high level open data science language provided

2. Integrating Heterogeneous Data Sources



Test set: 2023 Global Tropical Cyclone Season

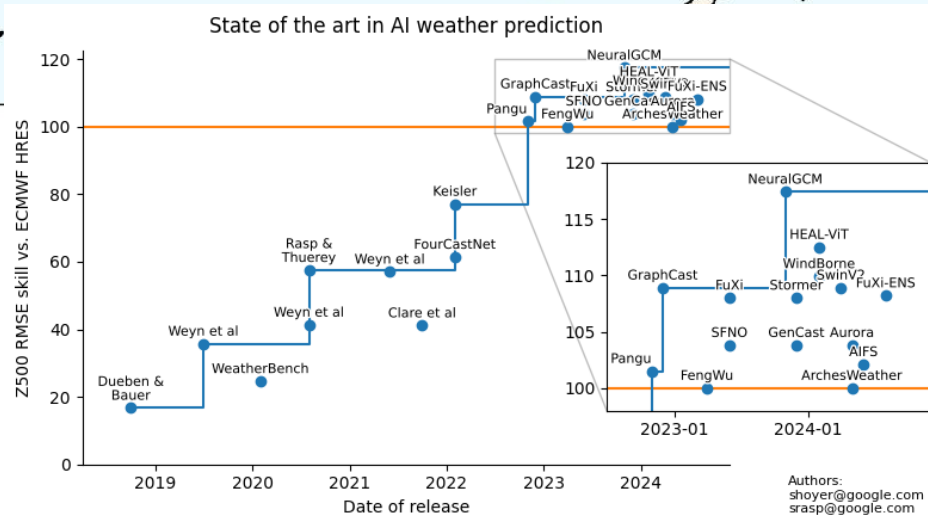
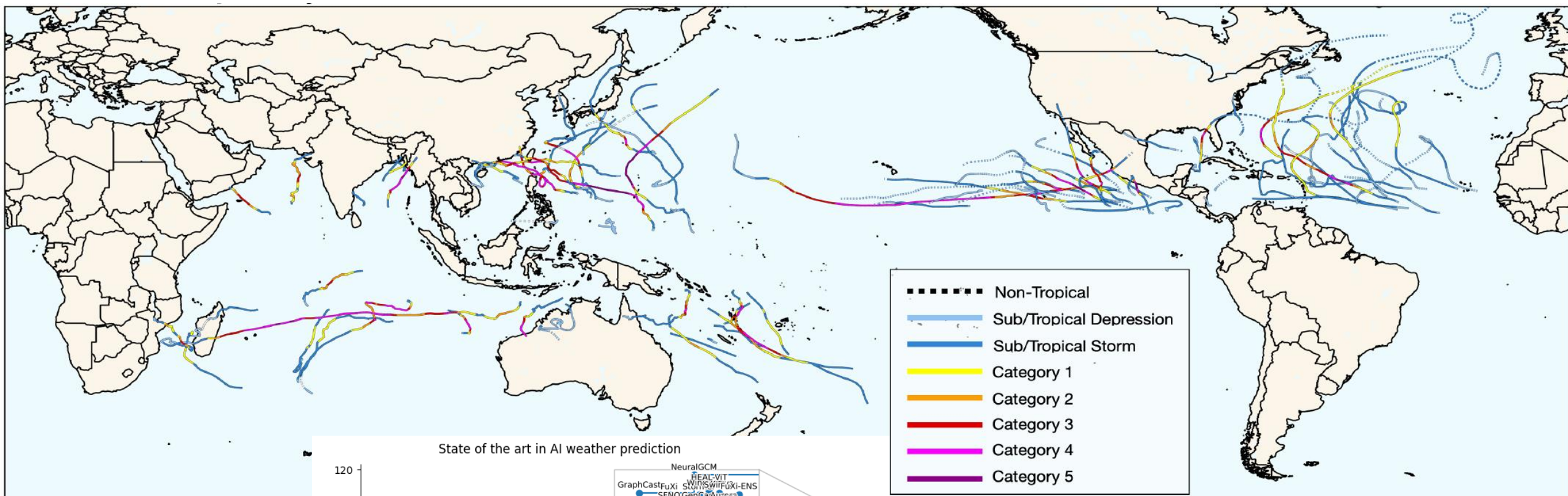


Image source: Marie McGraw (CIRA, now Jupiter Intelligence)

Plot: S. Hoyer & S. Rasp (Google)

R4: Evaluation metrics defined analytically and in code

Track Error Metrics: To evaluate the quality of predicted storm tracks, a set of deterministic error metrics is computed following the methodology described by Heming (Heming, 2017). These include:

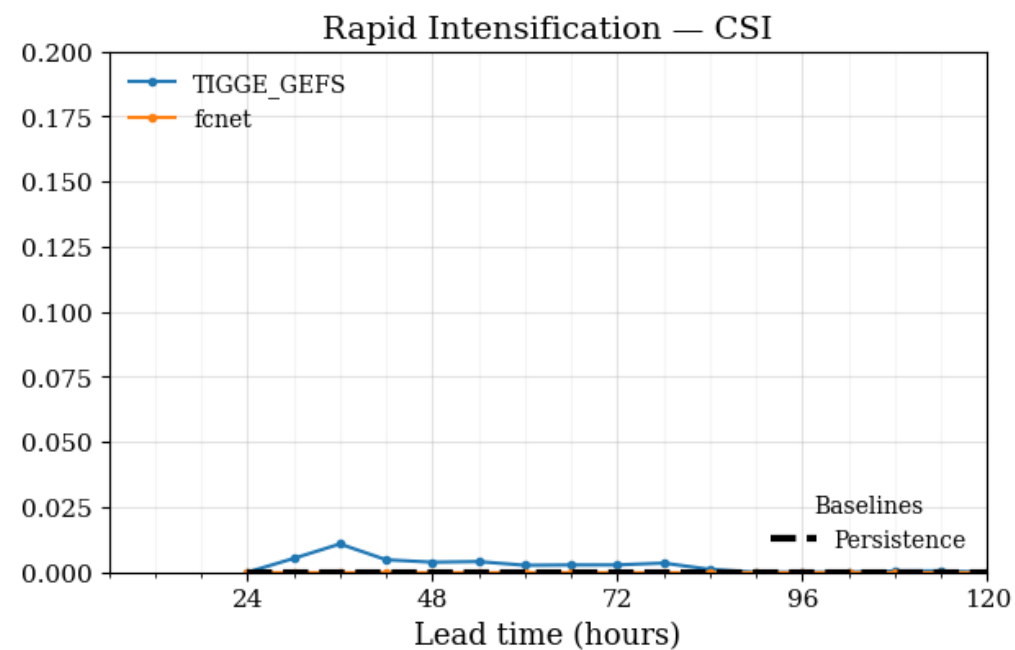
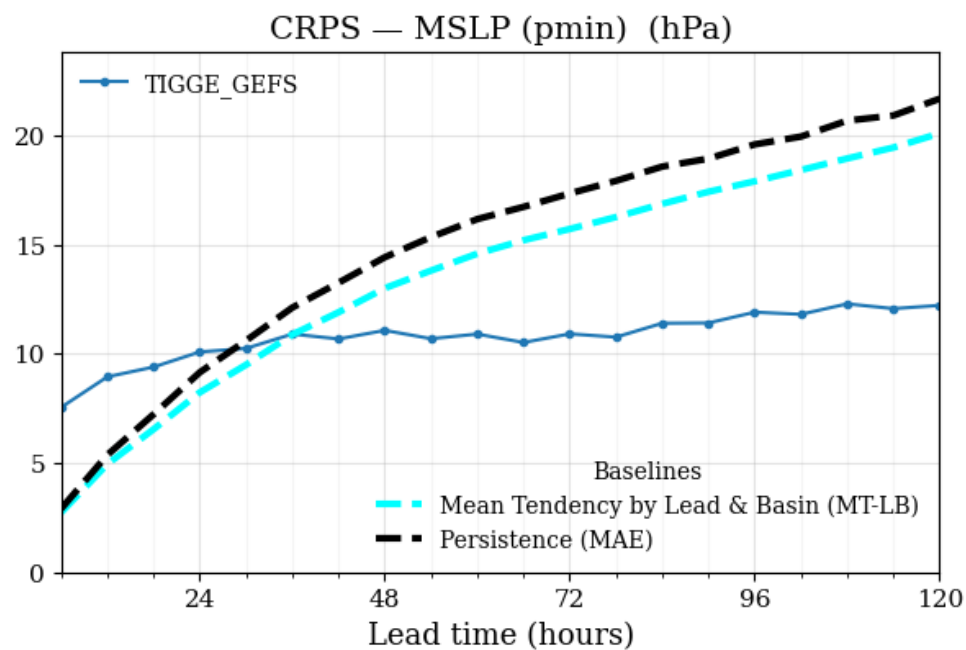
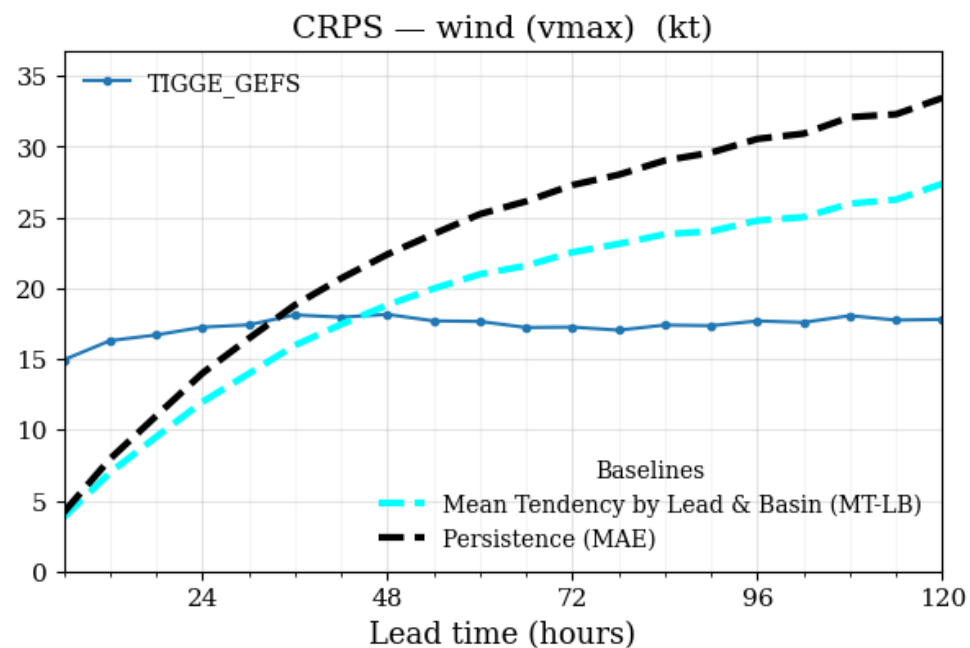
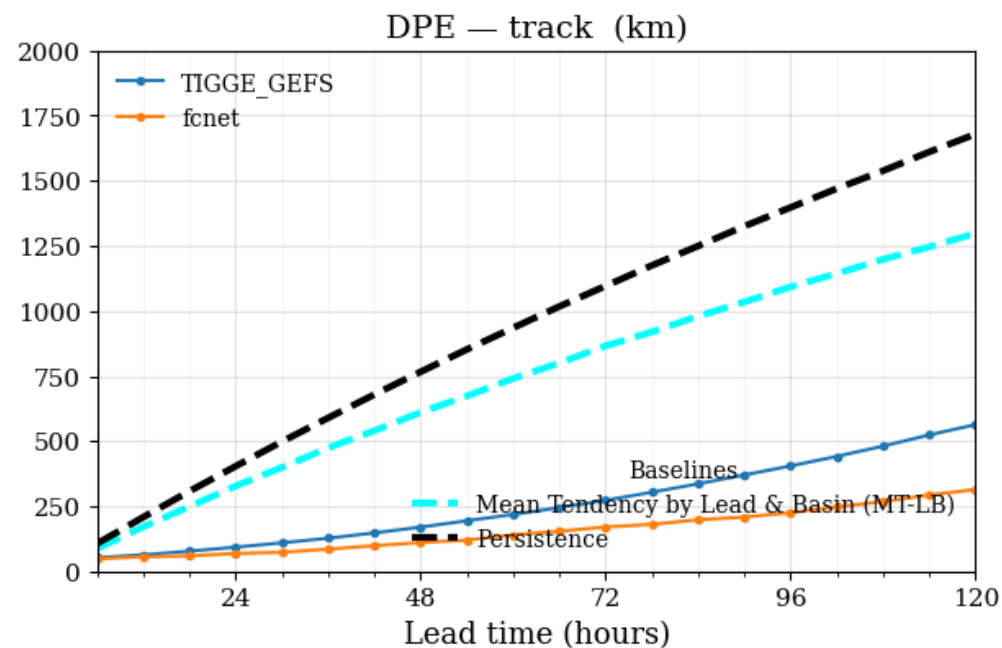
Direct Positional Error (DPE): The straight-line distance between the forecast and observed storm positions at the same verification time.

$$\text{CRPS}(\{x_i\}, y) = \frac{1}{N} \sum_{i=1}^N |x_i - y| - \frac{1}{2N(N-1)} \sum_{i=1}^N \sum_{j=1}^N |x_i - x_j|.$$

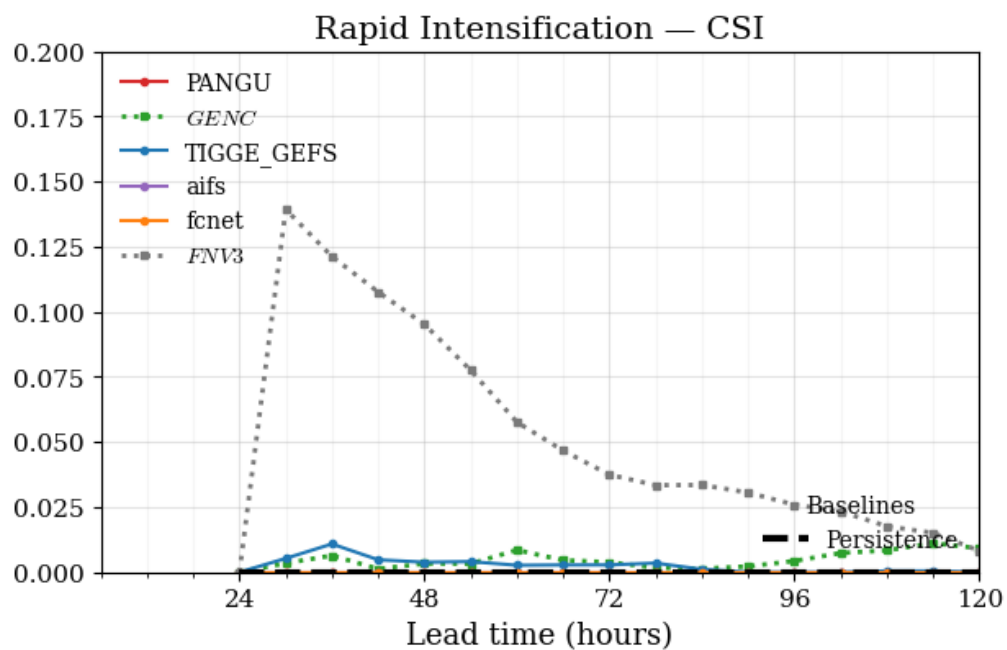
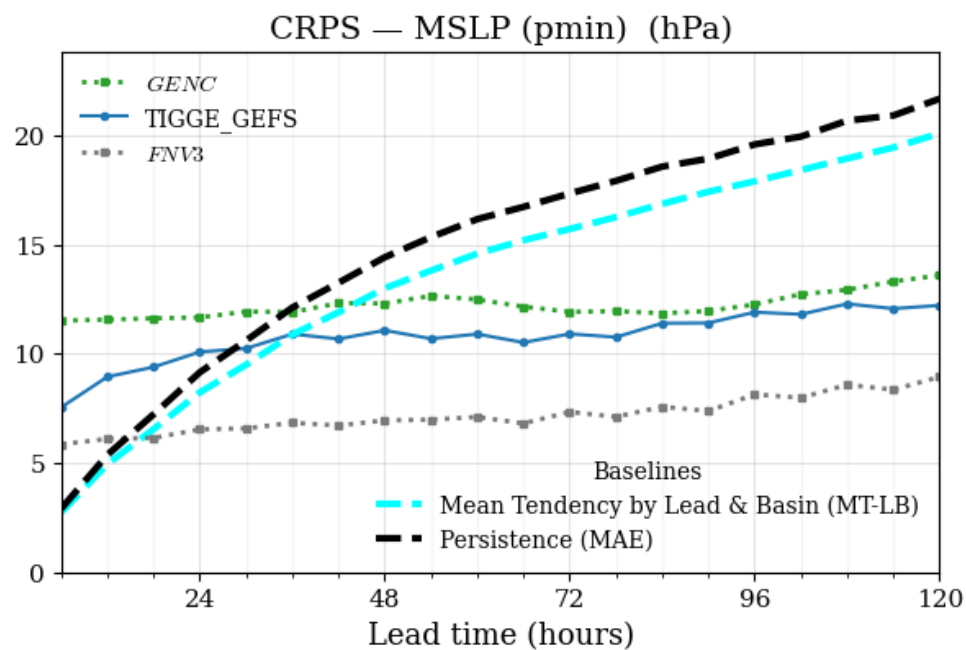
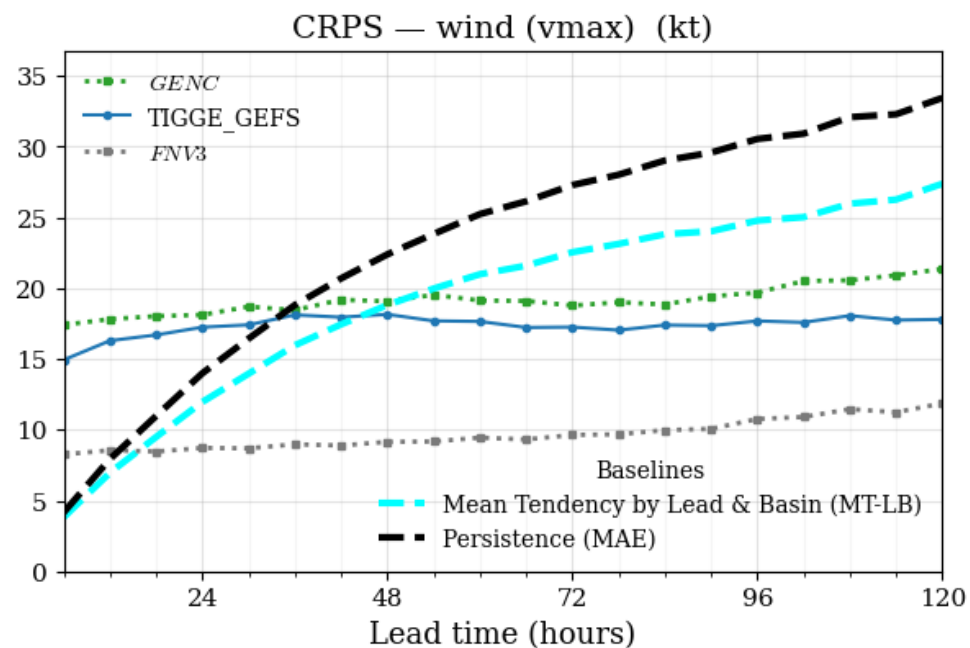
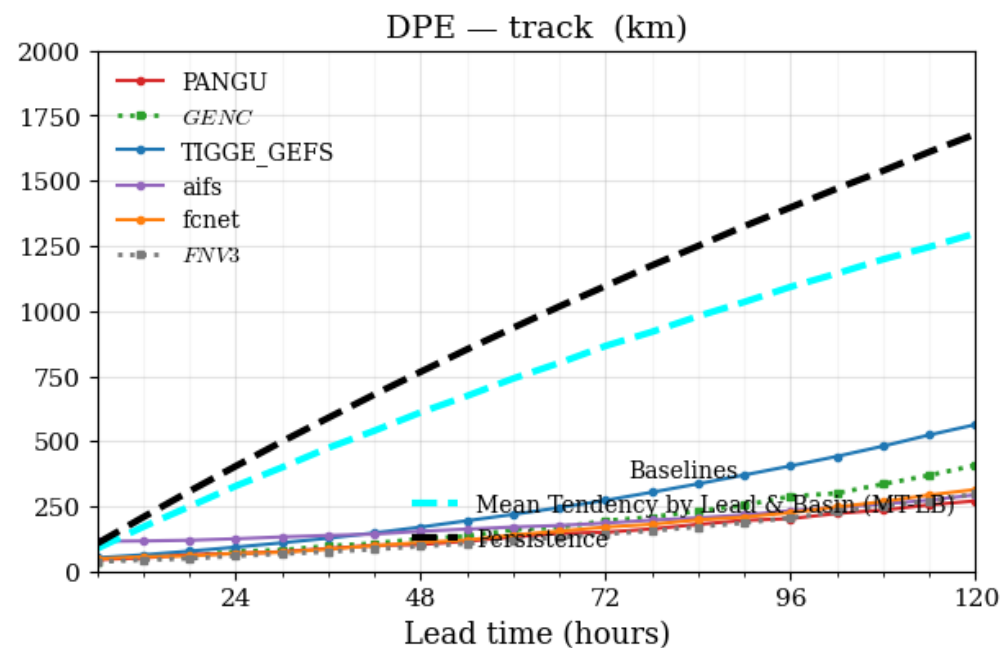
Critical Success Index (CSI): Also known as the Threat Score, measures the ratio of correctly predicted positive observations to the sum of all predicted positives, actual positives, and minus true positives. CSI can be used both for probabilistic track evaluation and RI.

$$\text{CSI} = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}},$$

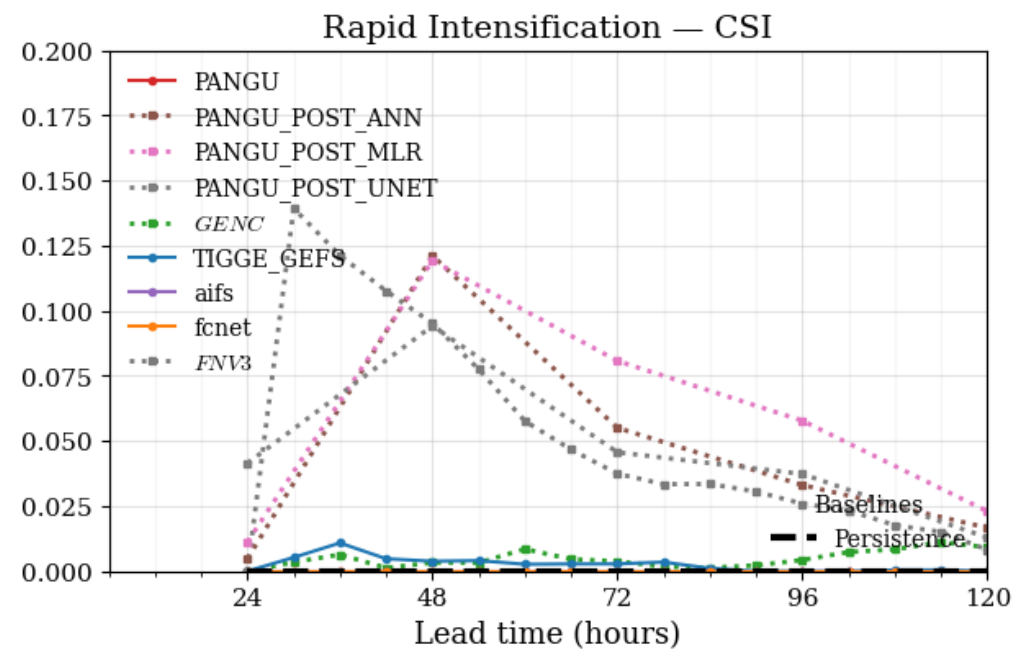
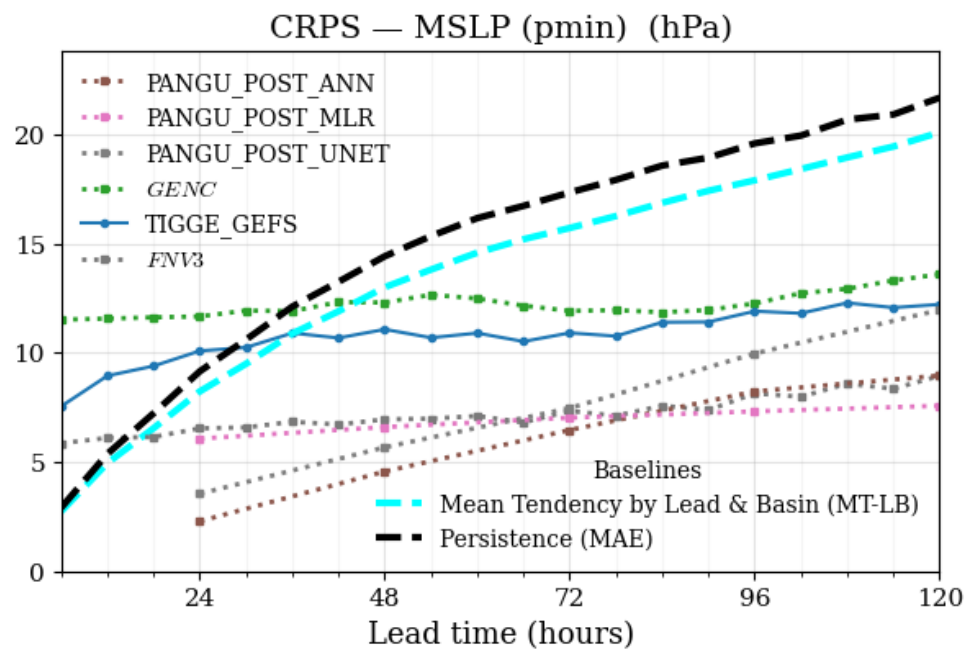
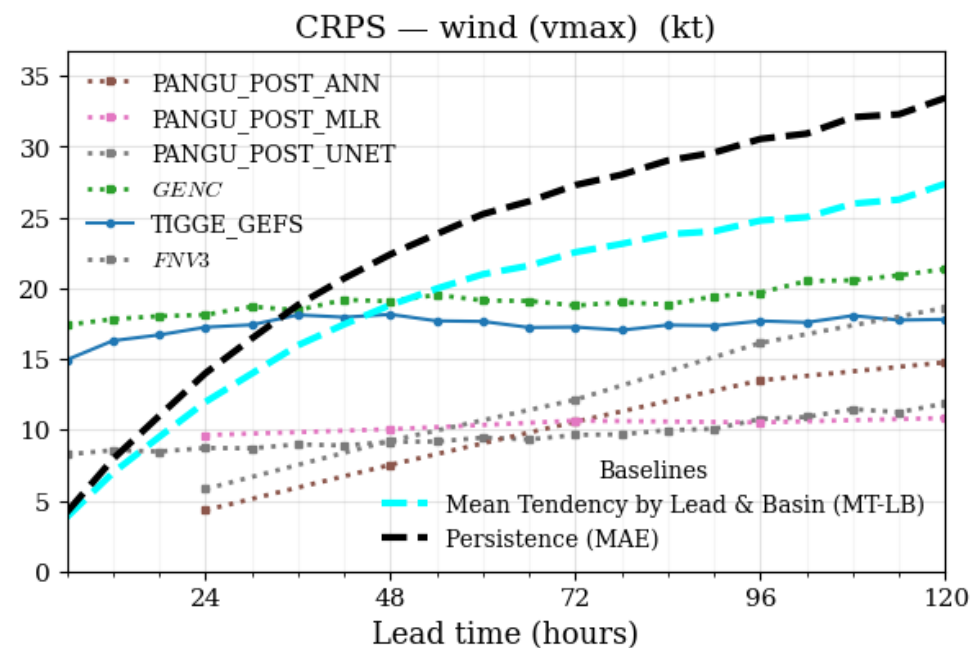
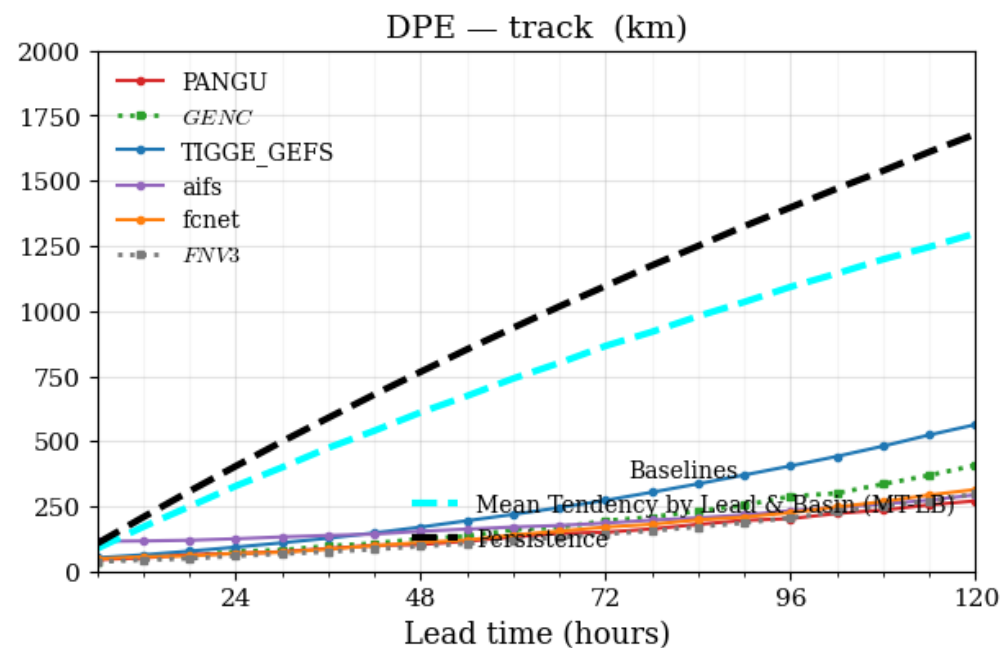
Physical models only

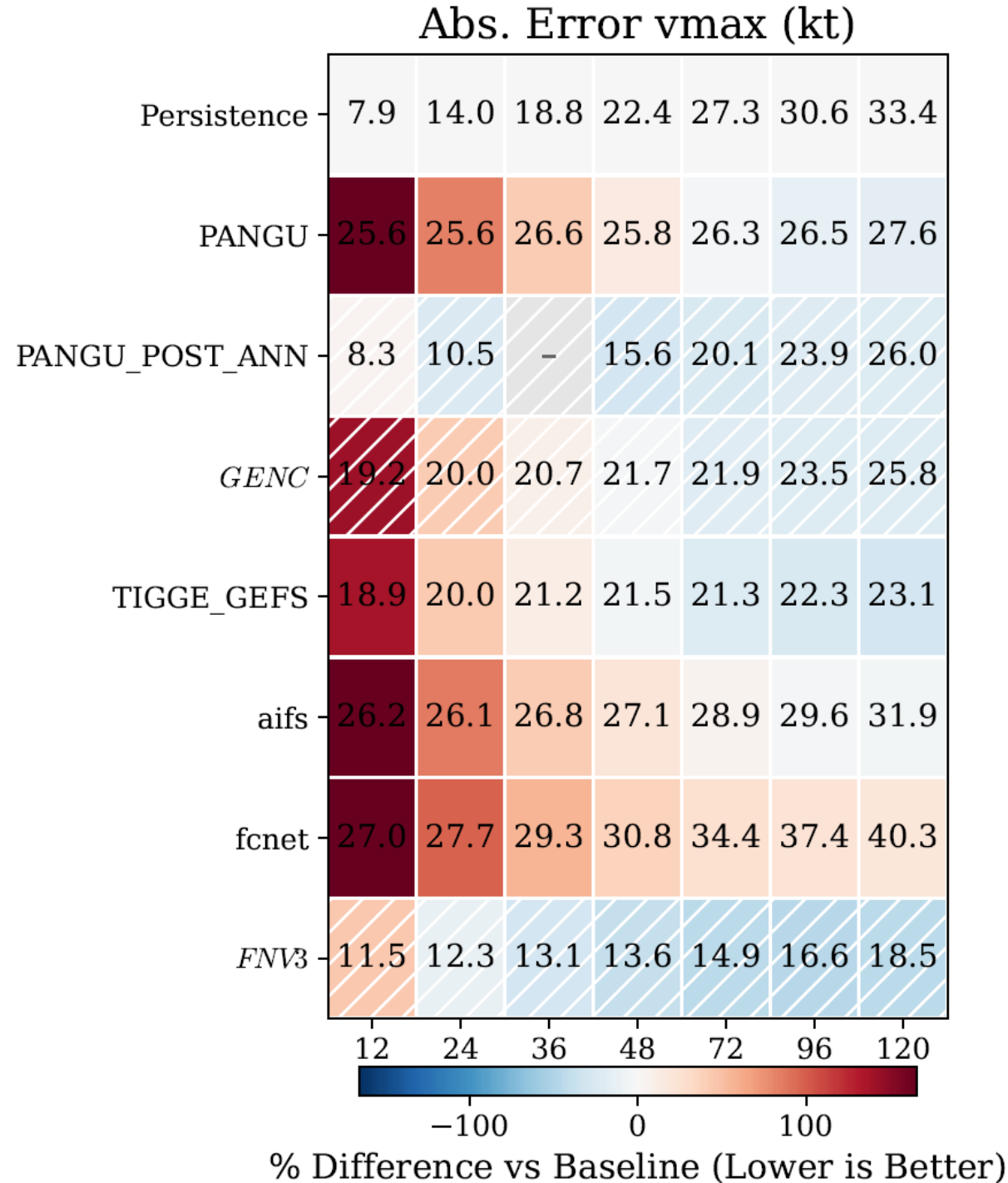


Physical + AI models



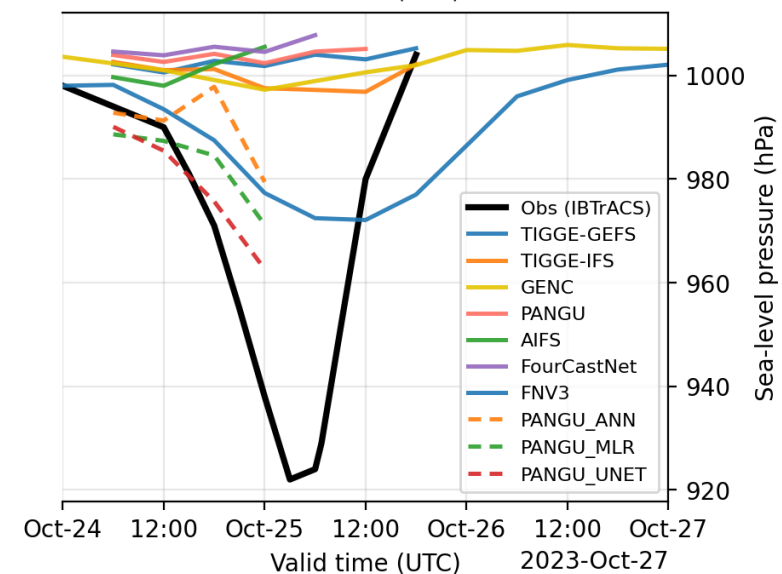
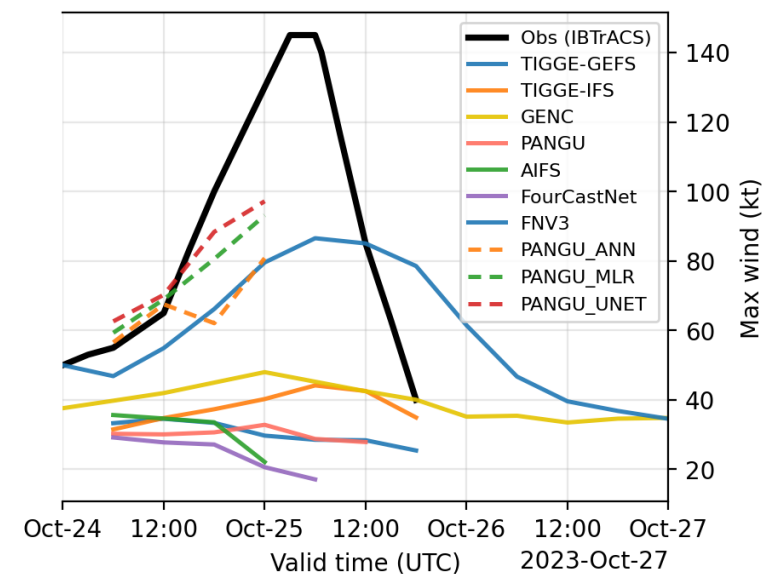
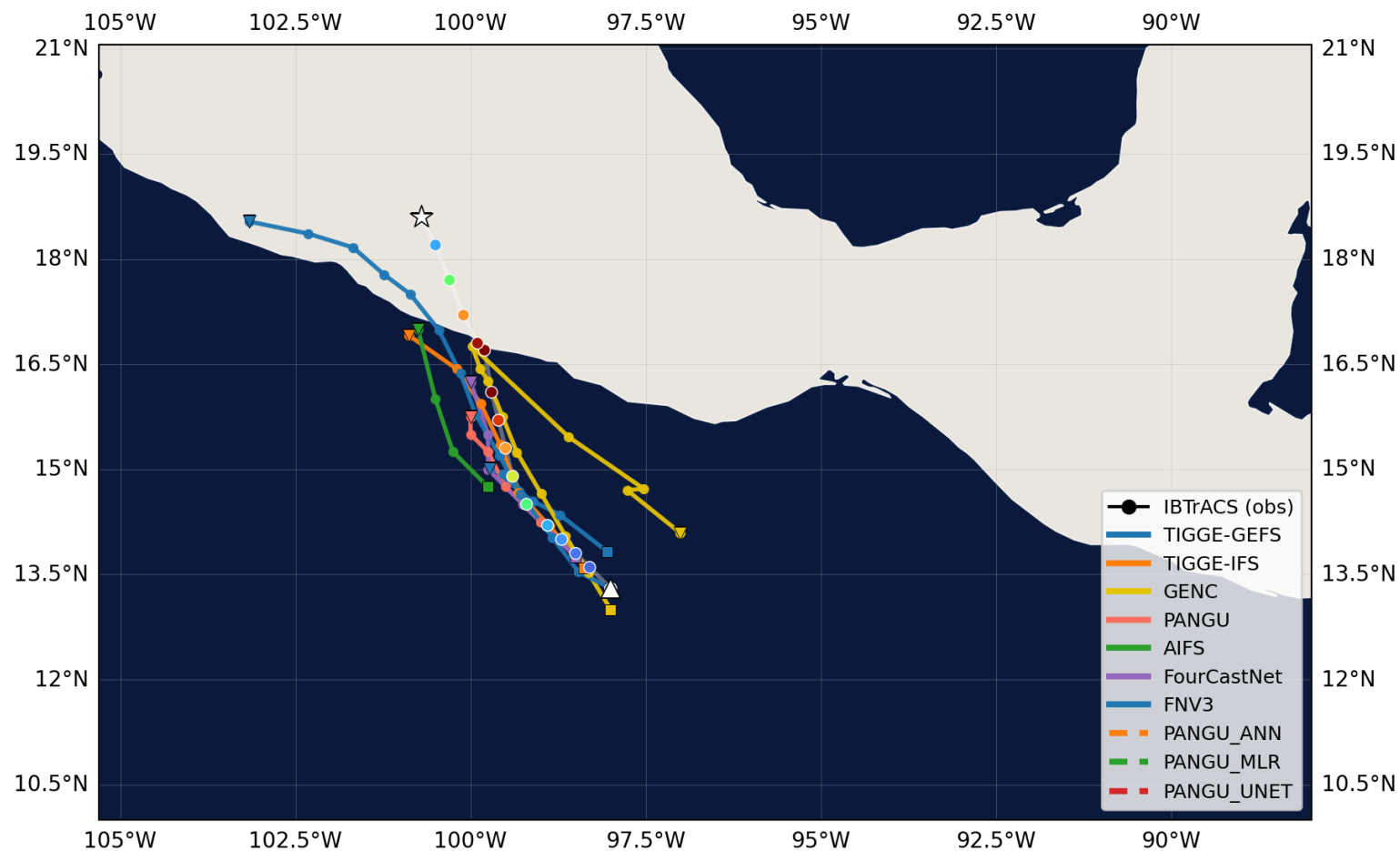
Physical + AI + Post-processing





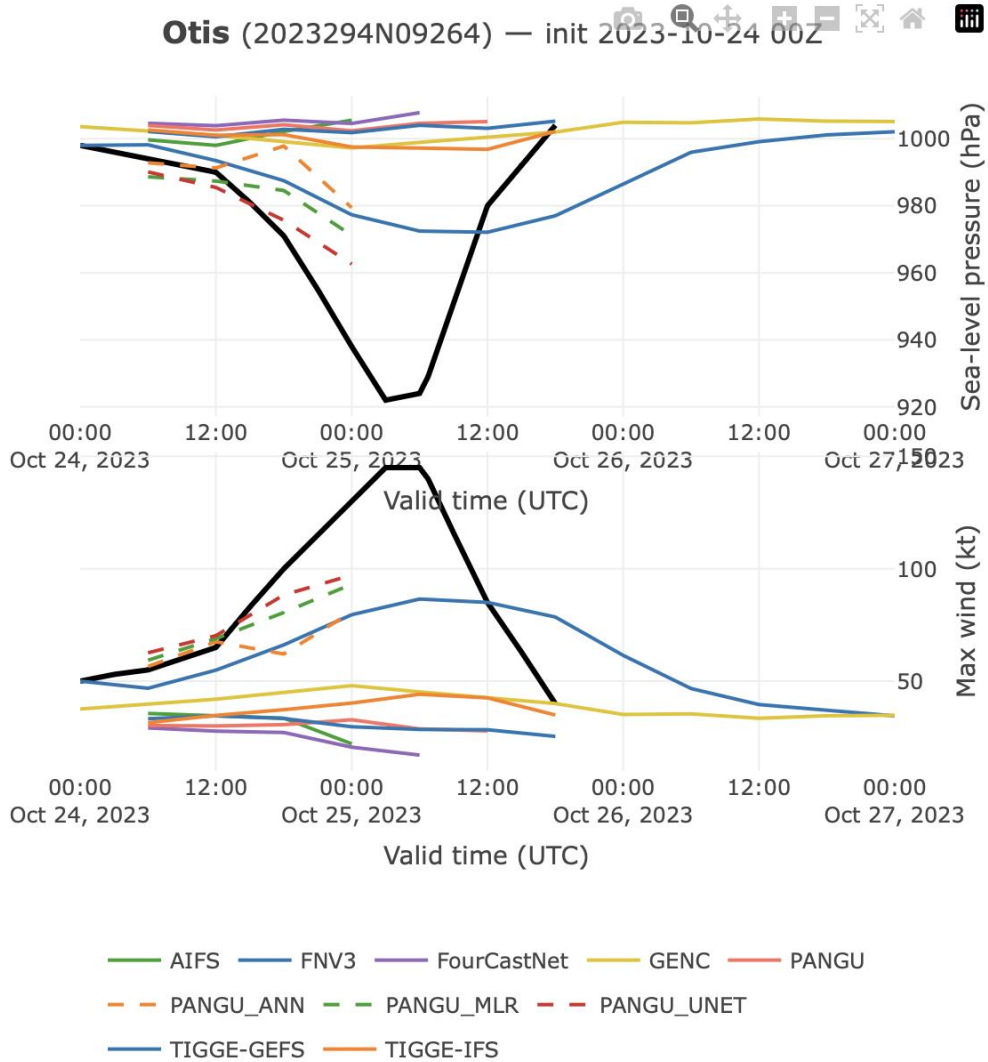
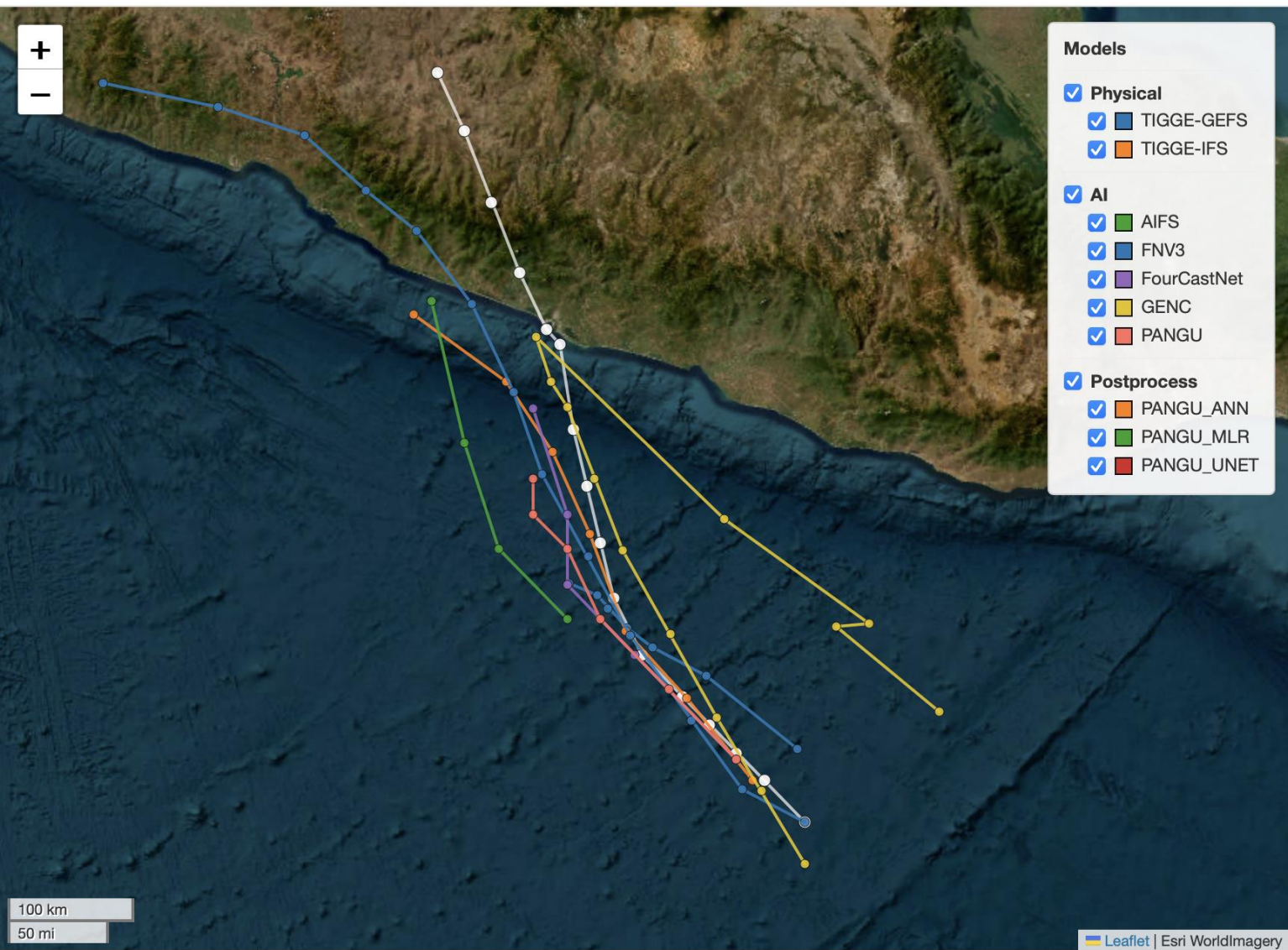
Source: Gomez et al. (2025)

Otis (2023294N09264) — init 2023-10-24 00Z



Obs 1-min sustained wind (kt)

Credits: Sam Darmon (UNIL/EPFL)



R6: Visualisation and diagnostics provided in code

1) TCBench

Global TC track and
intensity forecasting

2) ClimSim

Hybrid physics-ML
climate emulation

3) HybridESMBench

Hybrid ESM simulation

Order 1 requirements

R1: Data available online without access restrictions

R2: Clear problem statement for meaningful task in atmospheric science

R3: Data input into high level open data science language provided

R4: Evaluation metrics defined analytically and in code

Order 2 requirements

R5: Simple example machine learning solution provided in code

R6: Visualisation and diagnostics provided in code

R7: Tests for physical consistency and explainability provided

R8: Benchmarks for the computational performance provided

1) TCBench

Global TC track and
intensity forecasting

2) ClimSim

Hybrid physics-ML
climate emulation

3) HybridESMBench

Hybrid ESM simulation

Order 1 requirements

R1: Data available online without access restrictions

R2: Clear problem statement for meaningful task in atmospheric science

R3: Data input into high level open data science language provided

R4: Evaluation metrics defined analytically and in code

Order 2 requirements

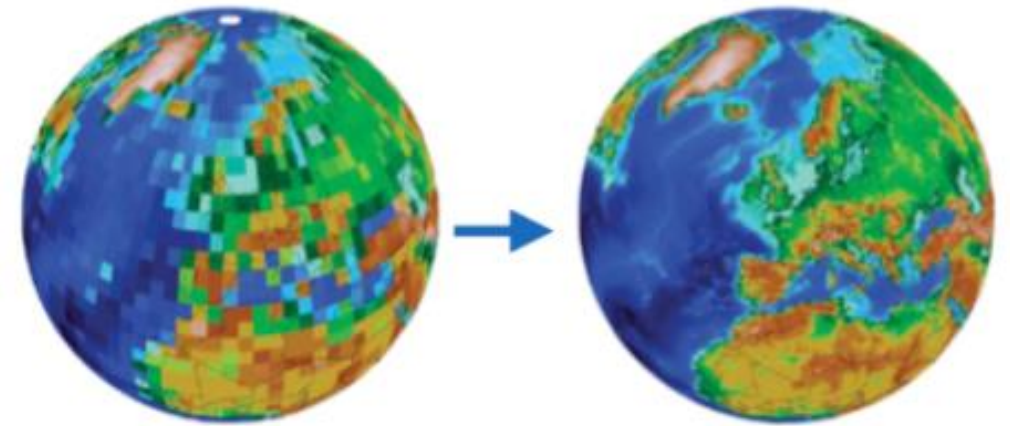
R5: Simple example machine learning solution provided in code

R6: Visualisation and diagnostics provided in code

R7: Tests for physical consistency and explainability provided

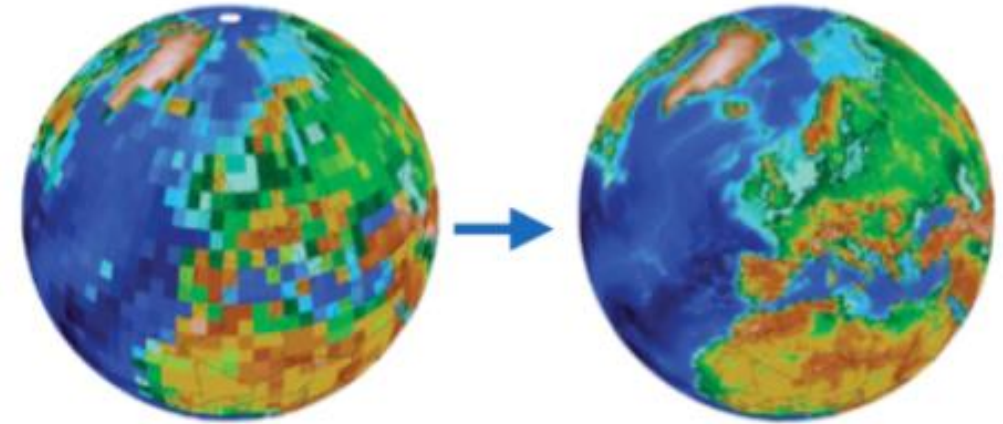
R8: Benchmarks for the computational performance provided

Context: We need a fine-resolution climate models, but they are too expensive

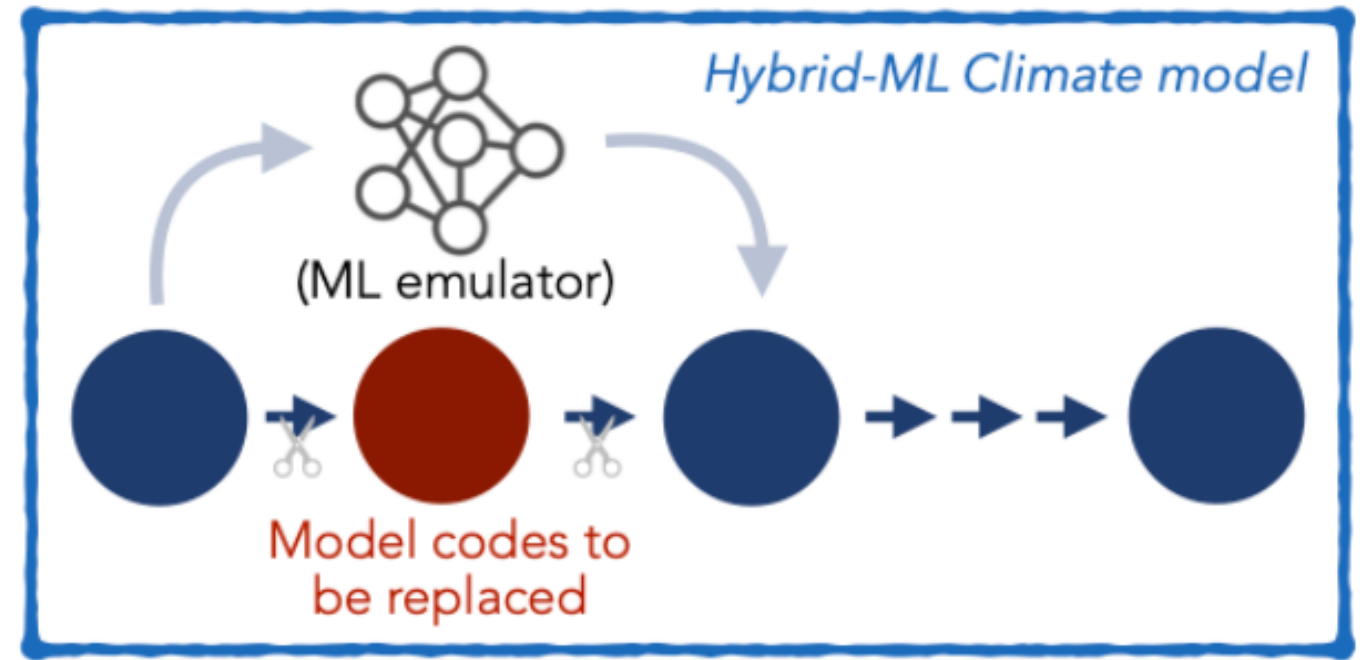


Context: We need a fine-resolution climate models, but they are too expensive

- Current climate models are not accurate enough as an actionable tool for climate mitigation plannings
- We need a fine-resolution climate model at ~ 1 -km scale. HOWEVER, these are computationally too expensive for next 20-30 years

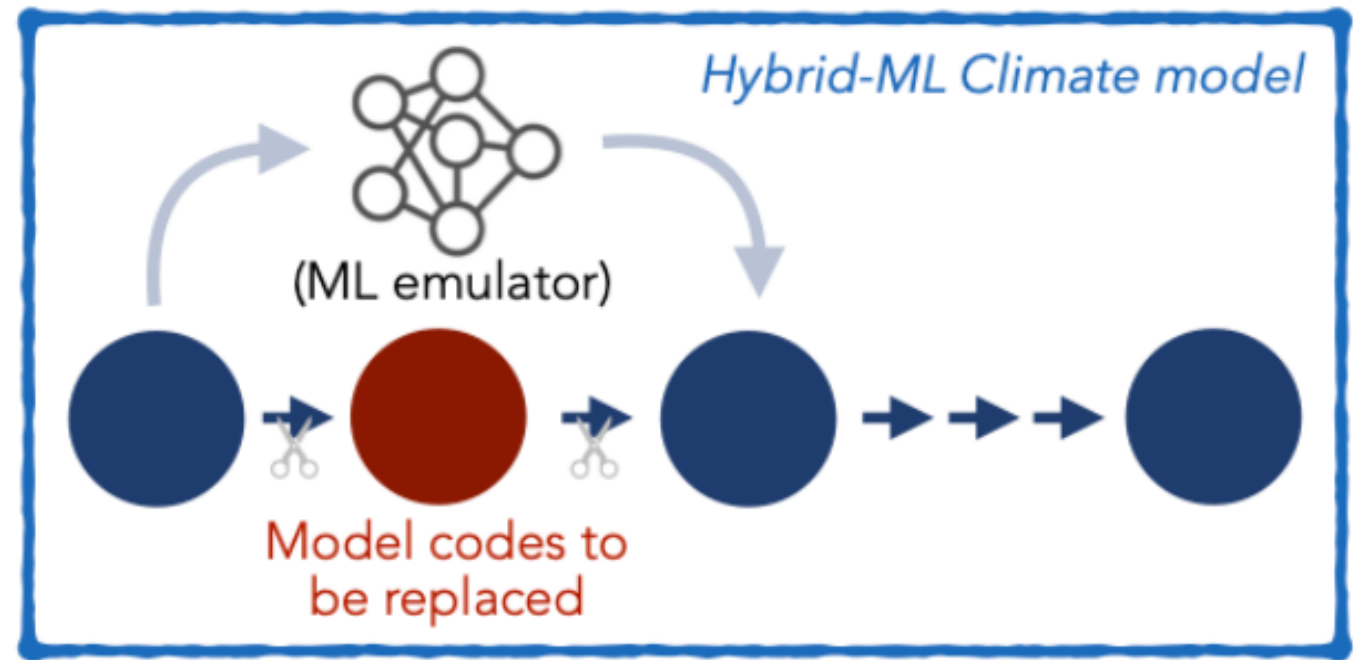


Solution: Hybrid-ML climate models can break accuracy vs. compute tradeoff.



Solution: Hybrid-ML climate models can break accuracy vs. compute tradeoff.

- In a hybrid-ML model, a sub-resolution process known to produce large errors are replaced with an ML emulator trained on a high-fidelity dataset
- The target of ClimSim is cloud formation processes, which are the source of the largest uncertainties in future climate projection



Challenge: Hybrid-ML climate models are not stable and accurate enough yet

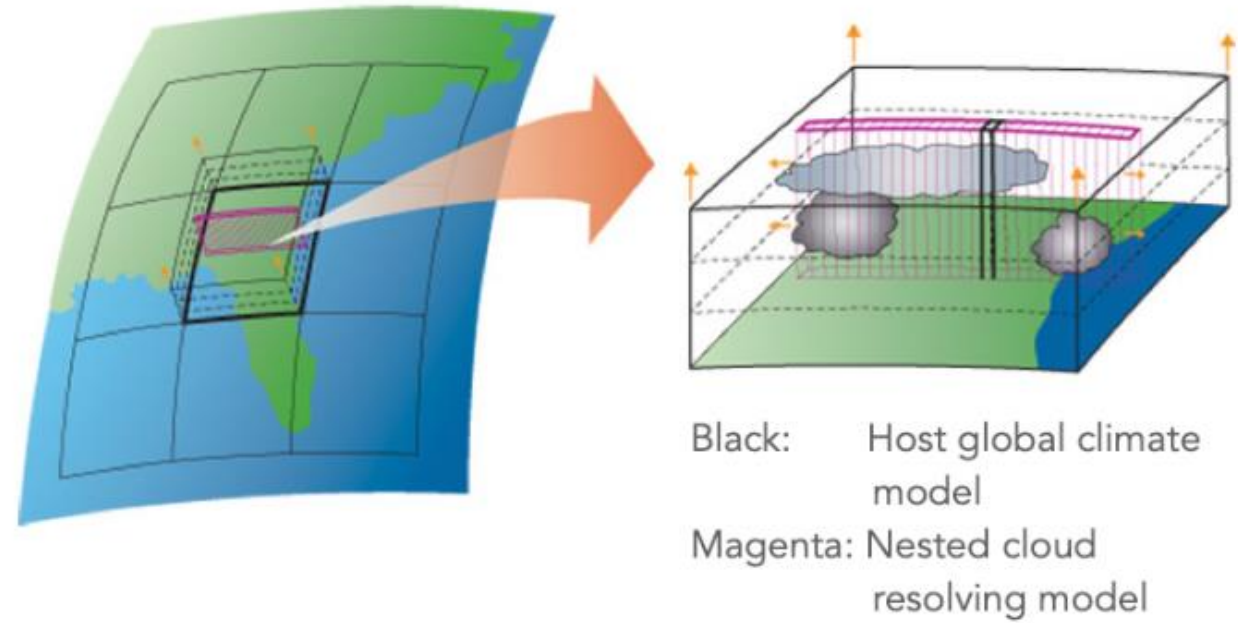


Challenge: Hybrid-ML climate models are not stable and accurate enough yet

- When coupled, small errors can be amplified by interacting with the rest of climate model codes.
- Climate modeling community has been working on this problem for ~7 years now, but no operational success.

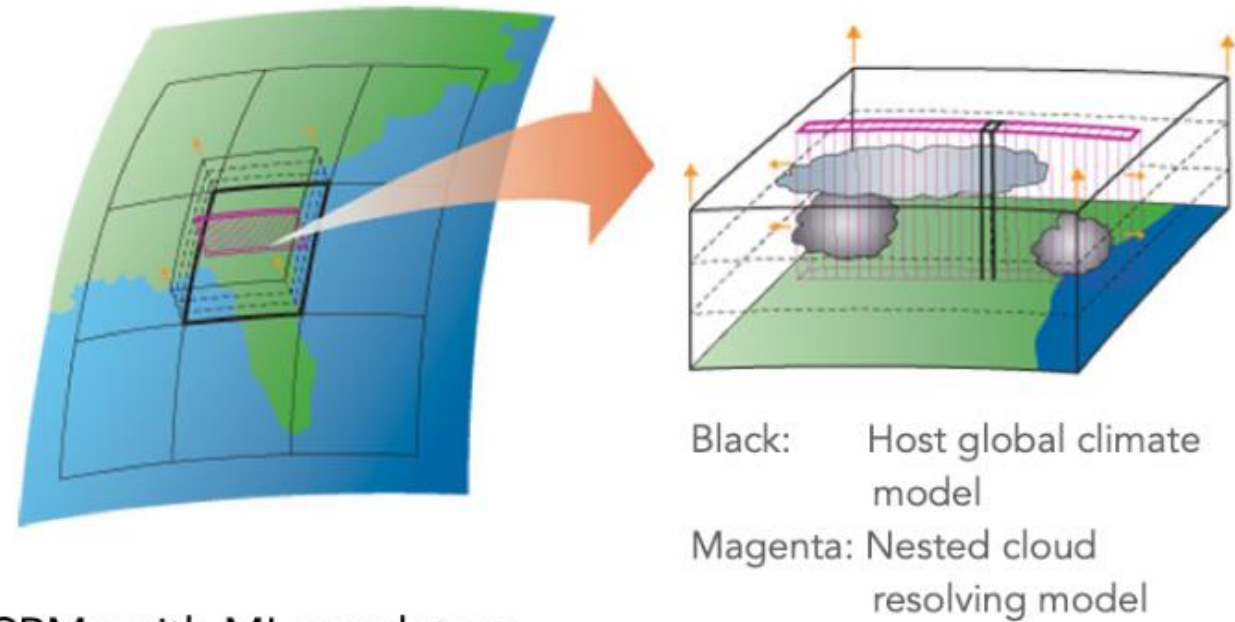


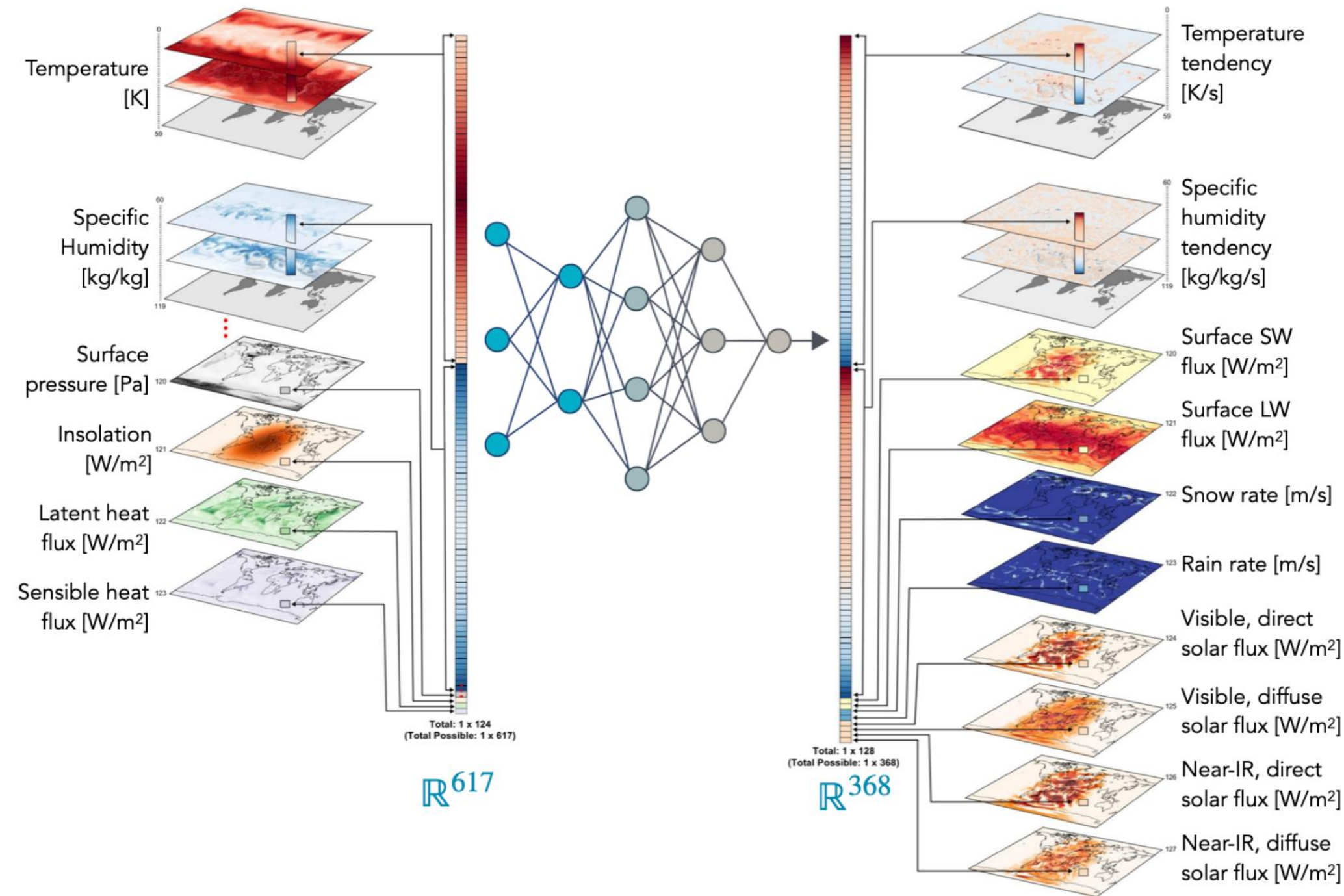
Emulation target: A cloud resolving model nested within each grid cell



Emulation target: A cloud resolving model nested within each grid cell

- We used multiscale-molding framework (MMF) as a backbone of hybrid ML model.
- The MMF approach replaces the conventional approximation with a cloud resolving model (CRM) in each grid cell, explicitly resolving cloud and turbulence.
- The CRM contains complex physical processes to capture clouds, including convection, turbulence, radiation, and phase changes of water
- ClimSim targets to replace *computationally-expensive* CRMs with ML emulators





Source: Yu et al. (2023)

Baseline Evaluation

Baseline experiment

R5: Simple example machine learning solution provided in code

Baseline Evaluation

Baseline experiment

R5: Simple example machine learning solution provided in code

- The low-resolution, real-geography version is used
- A subset of Input/output variables are chosen for its similarity to recent attempts in the literature (Marked by ★ in the “Variables” table)

Deterministic model

- Multilayer perceptron (MLP)
- Convolutional neural network (CNN)
- Encoder decoder (ED)

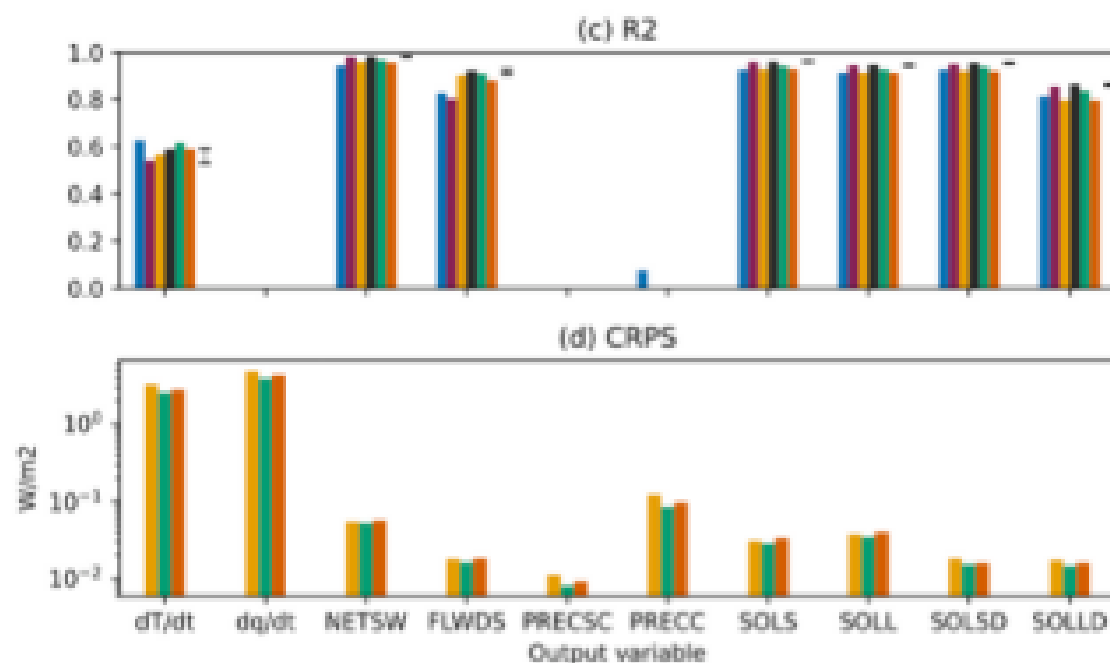
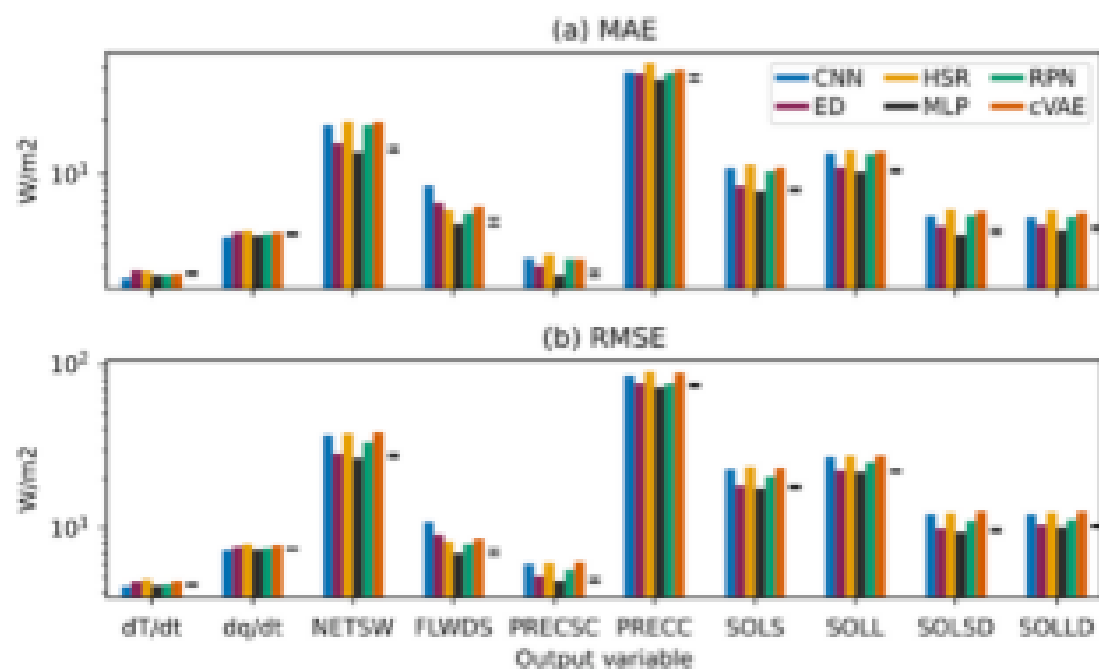
Stochastic model

- Heteroskedastic regression (HSR)
- Conditional VAE (cVAE)
- Randomized prior network (RPN)

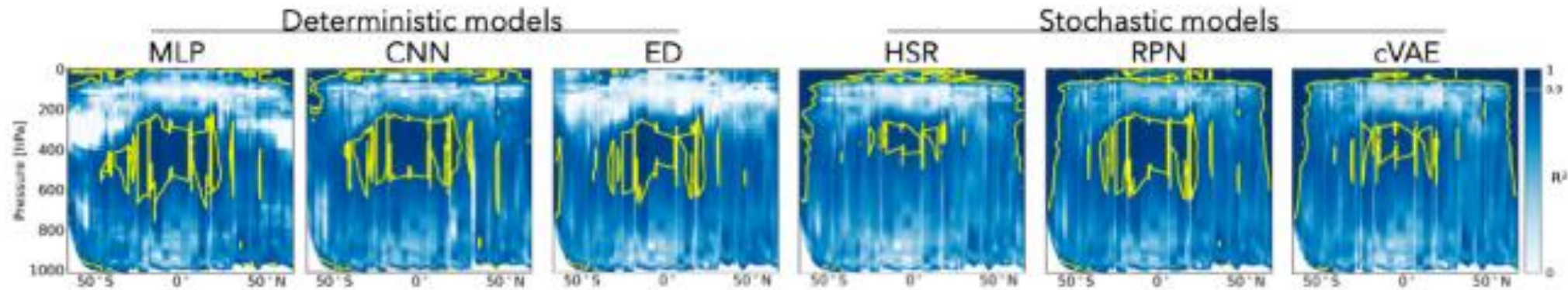
- Six architectures are tested
- To ensure that our evaluation is physically-consistent, we convert all predicted variables to energy flux units W/m^2 (power per unit area), e.g.,
 - Heating tendency $[\text{K/s}] \quad \times \quad c_p \cdot \Delta p/g \quad \rightarrow \quad [\text{W/m}^2]$
 - Moistening tendency $[\text{kg/kg/s}] \quad \times \quad L_v \cdot \Delta p/g \quad \rightarrow \quad [\text{W/m}^2]$

Results

- Summary statistics: first, a metric is calculated at each grid cell; then, averaged globally and vertically with area and mass weighting

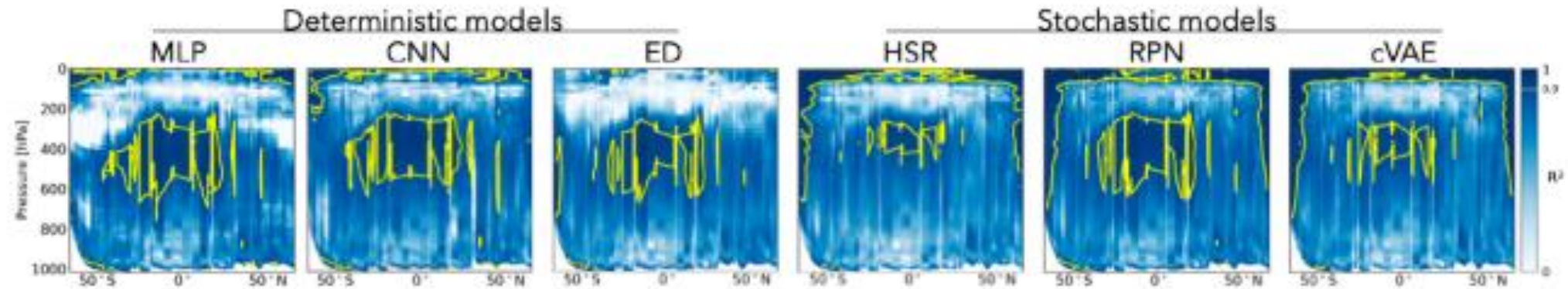


- Stochastic models show a clear advantage in the regions (e.g., the stratosphere and poles) where deterministic models are inaccurate



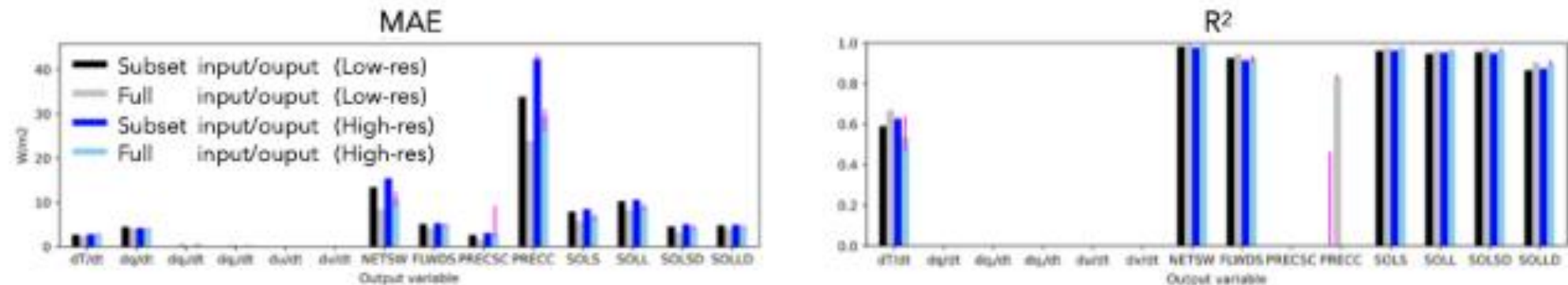
(For MLP, we conducted an ablation study testing all variables and high-res)

- Stochastic models show a clear advantage in the regions (e.g., the stratosphere and poles) where deterministic models are inaccurate



(For MLP, we conducted an ablation study testing all variables and high-res)

- Including all input variables generally improves the model accuracy in both the low-resolution and the high-resolution datasets



Crowdsourcing to accelerate progress in hybrid modeling?

LEAP - Atmospheric Physics using AI (ClimSim)

Simulate higher resolution atmospheric processes within E3SM-MMF, a climate model supported by the U.S. Department of Energy



- Overview
- Data
- Code
- Models
- Discussion
- Leaderboard
- Rules

Overview

In this competition, you'll develop machine learning models that accurately emulate subgrid-scale atmospheric physics in an operational climate model—an important step in improving climate projections and reducing uncertainty surrounding future climate trends.



Competition Host

LEAP

Prizes & Awards

\$50,000

Awards Points & Medals

Participation

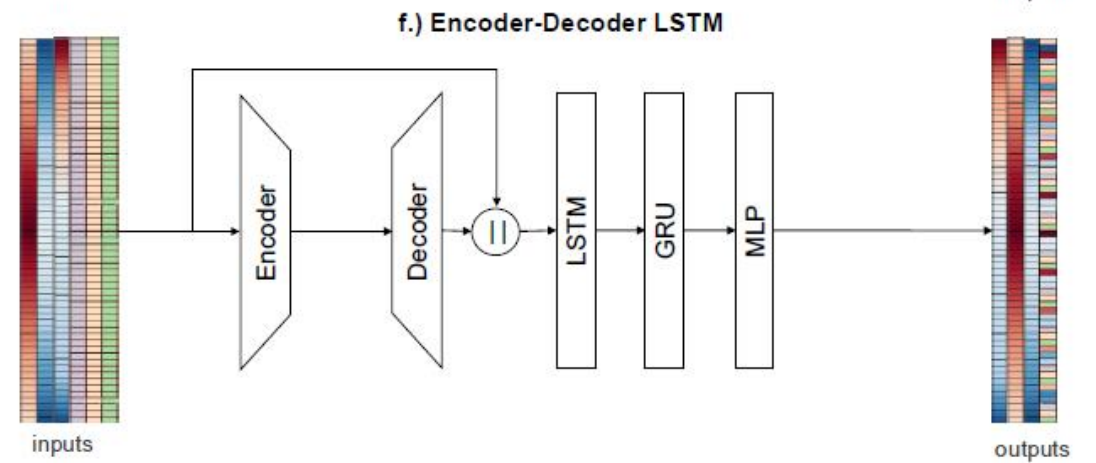
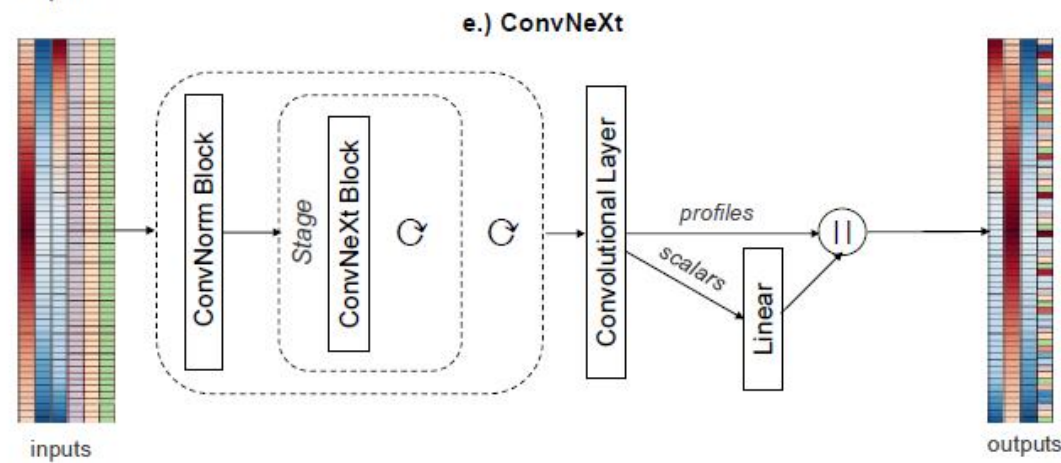
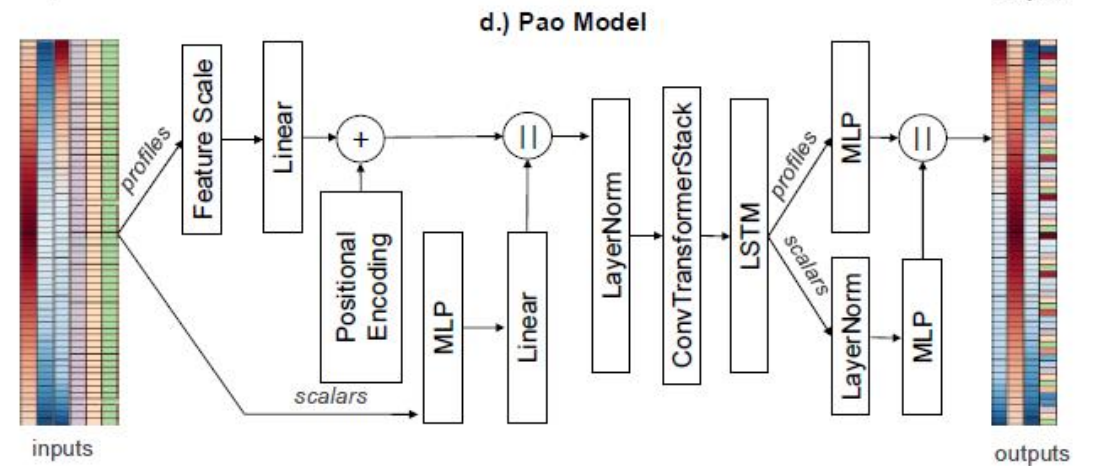
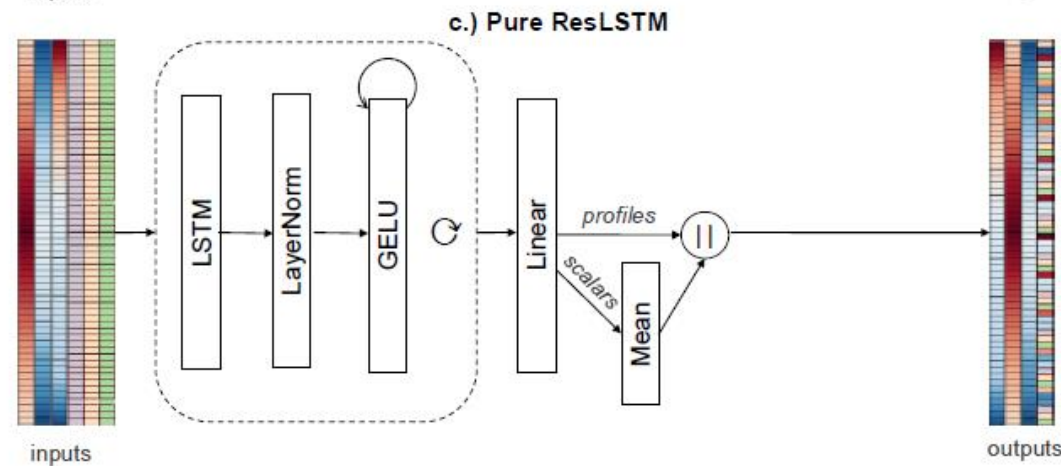
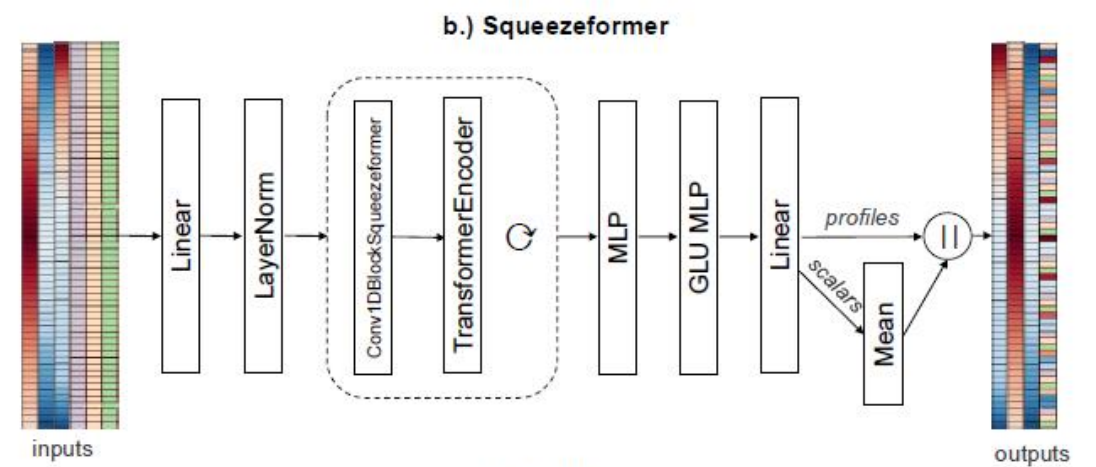
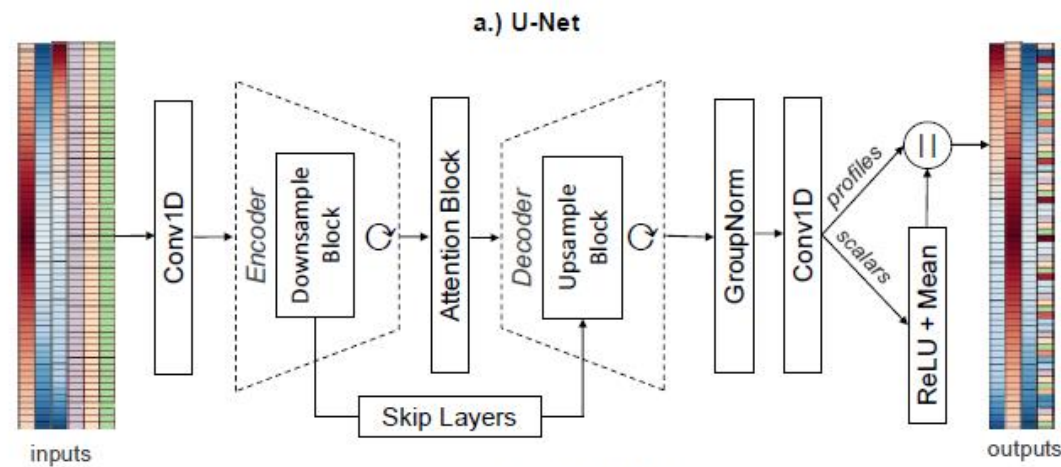
5,177 Entrants

877 Participants

693 Teams

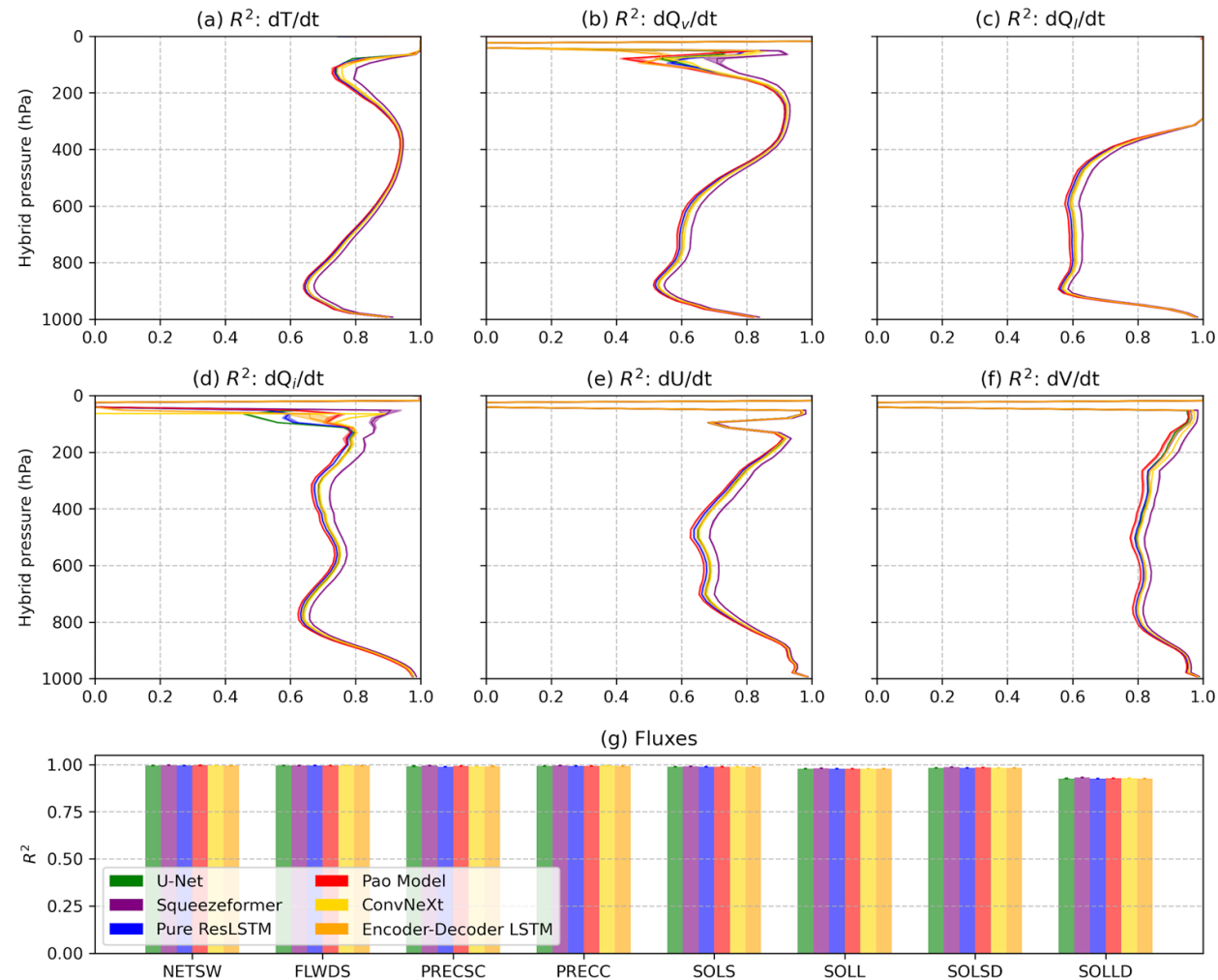
11,059 Submissions

Slide credits: Jerry Lin (UCI, BU)



Offline, the Kaggle architectures have strong R2 skill

Offline R^2 for each model (Expanded Variable List Configuration)



ClimSim Online: Containerized pipeline to integrate ML models into operational climate simulators for hybrid testing

[Home Page](#)[Papers](#)[Submissions](#)[News](#)[Editorial Board](#)[Special Issues](#)[Open Source
Software](#)[Proceedings
\(PMLR\)](#)[Data \(DMLR\)](#)

ClimSim-Online: A Large Multi-Scale Dataset and Framework for Hybrid Physics-ML Climate Emulation

Sungduk Yu, Zeyuan Hu, Akshay Subramaniam, Walter Hannah, Liran Peng, Jerry Lin, Mohamed Aziz Bhowri, Ritwik Gupta, Björn Lütjens, Justus C. Will, Gunnar Behrens, Julius J. M. Busecke, Nora Loose, Charles I Stern, Tom Beucler, Bryce Harrop, Helge Heuer, Benjamin R Hillman, Andrea Jenney, Nana Liu, Alistair White, Tian Zheng, Zhiming Kuang, Fiaz Ahmed, Elizabeth Barnes, Noah D. Brenowitz, Christopher Bretherton, Veronika Eyring, Savannah Ferretti, Nicholas Lutsko, Pierre Gentine, Stephan Mandt, J. David Neelin, Rose Yu, Laure Zanna, Nathan M. Urban, Janni Yuval, Ryan Abernathey, Pierre Baldi, Wayne Chuang, Yu Huang, Fernando Iglesias-Suarez, Sanket Jantre, Po-Lun Ma, Sara Shamekh, Guang Zhang, Michael Pritchard; 26(142):1–85, 2025.

Abstract

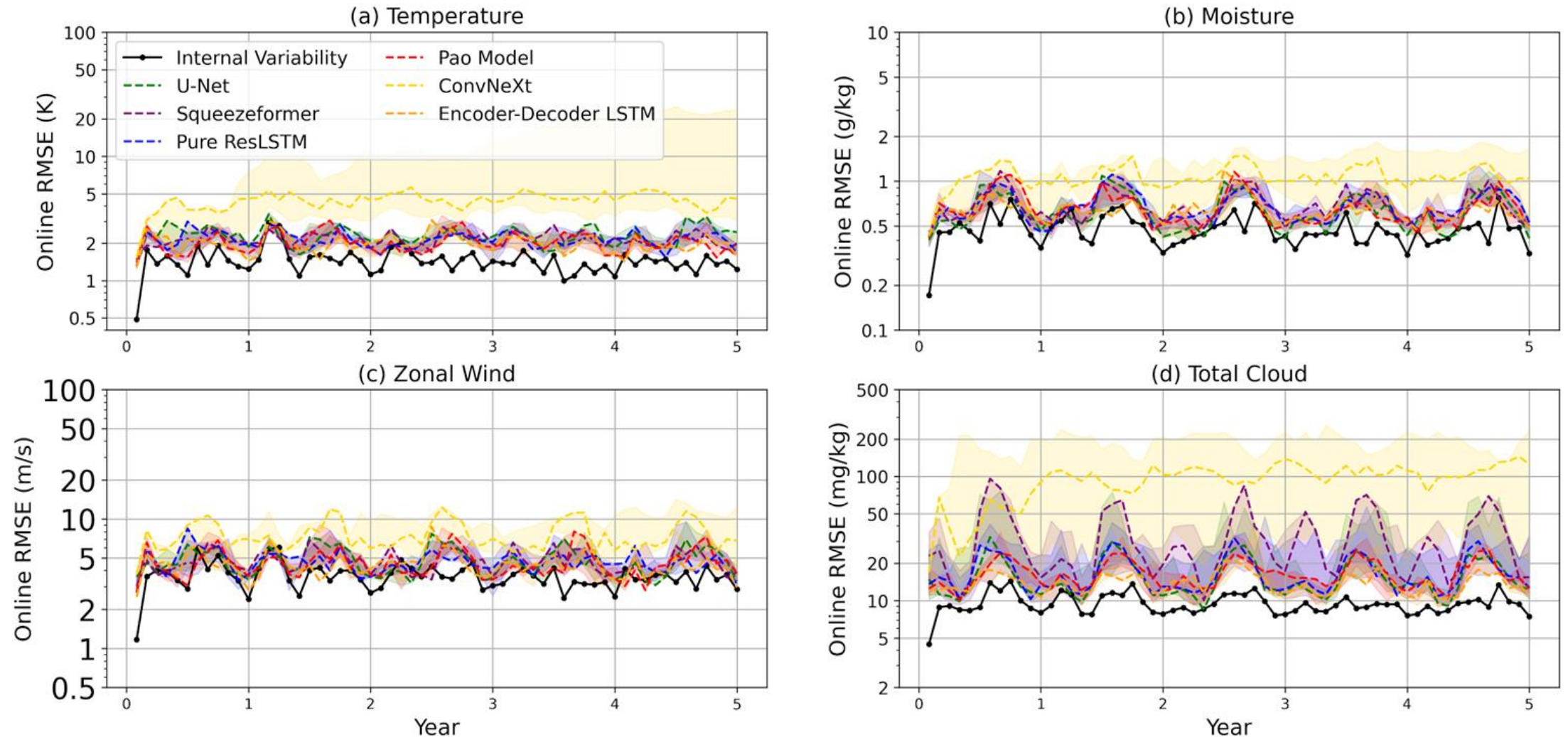
Modern climate projections lack adequate spatial and temporal resolution due to computational constraints, leading to inaccuracies in representing critical processes like thunderstorms that occur on the sub-resolution scale. Hybrid methods combining physics with machine learning (ML) offer faster, higher fidelity climate simulations by outsourcing compute-hungry, high-resolution simulations to ML emulators. However, these hybrid physics-ML simulations require domain-specific data and workflows that have been inaccessible to many ML experts. This paper is an extended version of our NeurIPS award-winning ClimSim dataset paper. The ClimSim dataset includes 5.7 billion pairs of multivariate input/output vectors spanning ten years at high temporal resolution, capturing the influence of high-resolution, high-fidelity physics on a host climate simulator's macro-scale state. In this extended version, we introduce a significant new contribution in Section 5, which provides a cross-platform, containerized pipeline to integrate ML models into operational climate simulators for hybrid testing. We also implement various baselines of ML models and hybrid simulators to highlight the ML challenges of building stable, skillful emulators. The data (https://huggingface.co/datasets/LEAP/ClimSim_high-res, also in a low-resolution version at https://huggingface.co/datasets/LEAP/ClimSim_low-res and an aquaplanet version at https://huggingface.co/datasets/LEAP/ClimSim_low-res_aqua-planet) and code (<https://leap-stc.github.io/ClimSim> and <https://github.com/leap-stc/climsim-online>) are publicly released to support the development of hybrid physics-ML and high-fidelity climate simulations.

[\[abs\]](#)[\[pdf\]](#)[\[bib\]](#) [\[code\]](#)

© 2025 M. Pritchard et al. All rights reserved.

Online, stability is now the norm not the exception

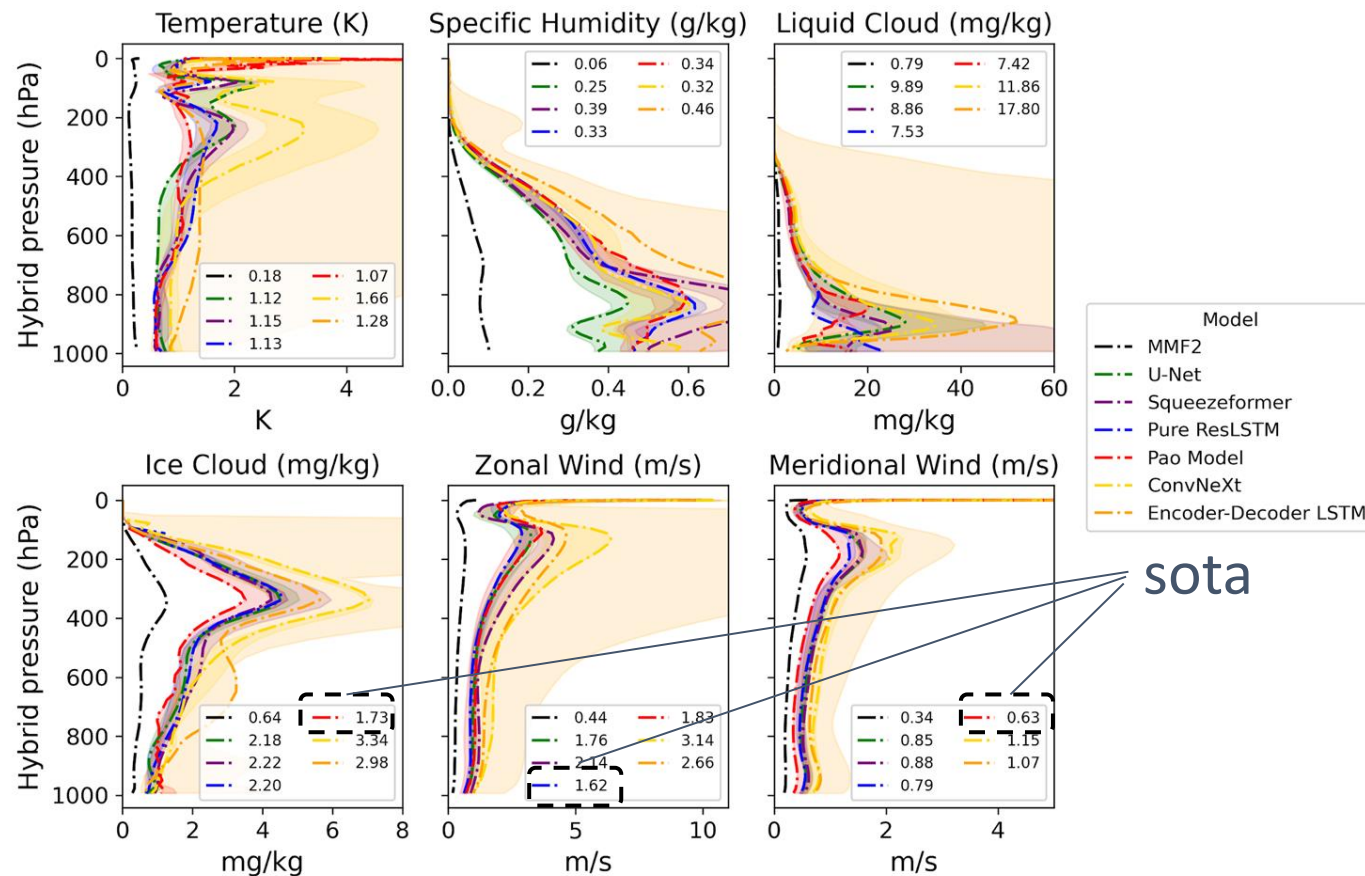
5 Year Online Root Mean Squared Error (Standard Configuration)



Interestingly, no pareto-improvement with a single model

Confidence Loss Configuration

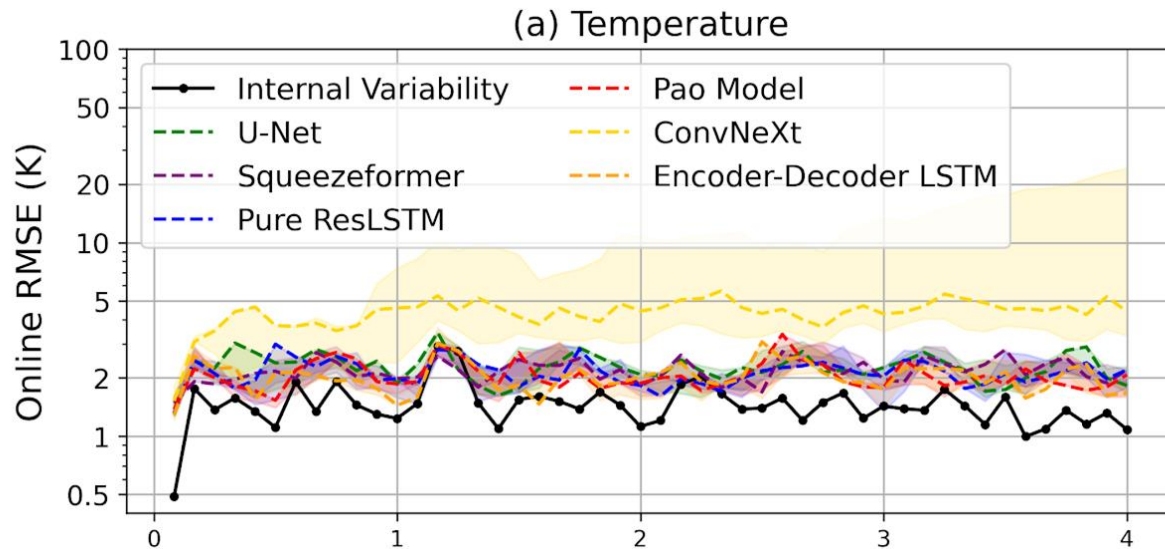
5 Year Global Mean Root Mean Squared Error (Confidence Loss configuration)



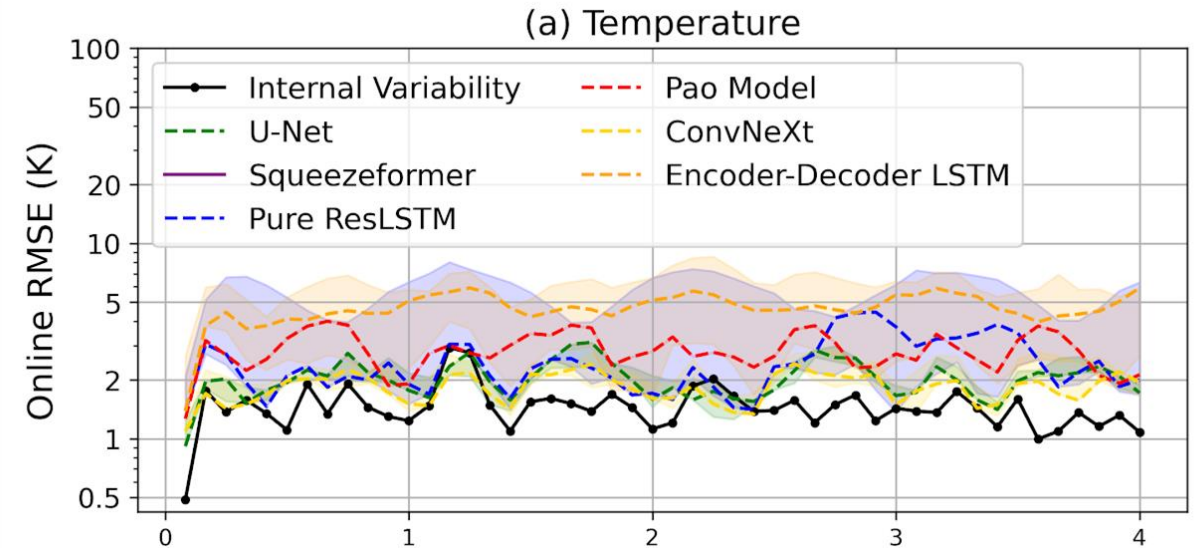
sota

Different architectures respond differently to the same architecture agnostic design decisions

standard configuration

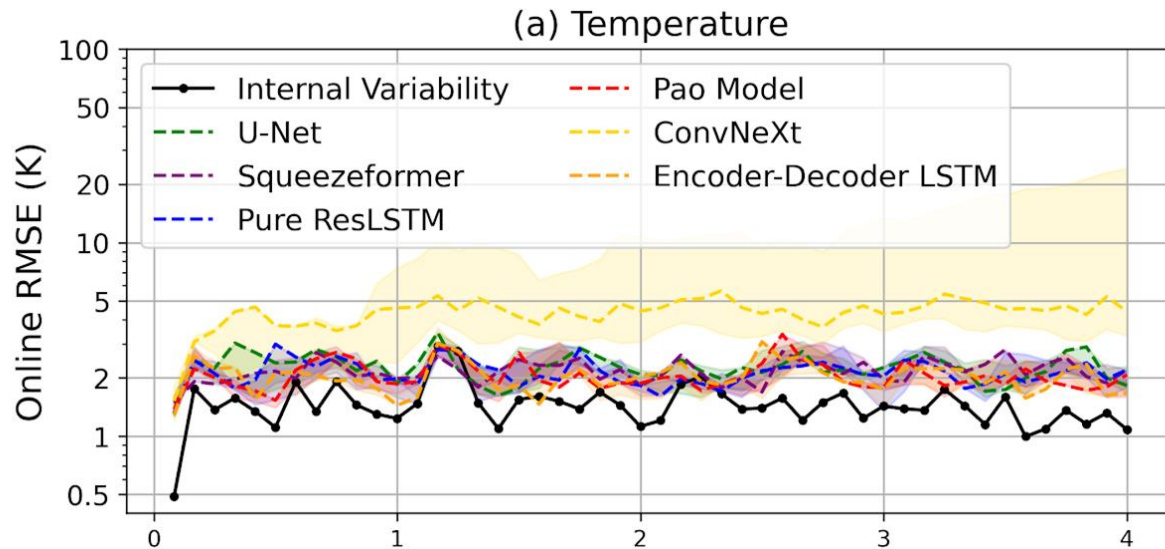


expanded variable list

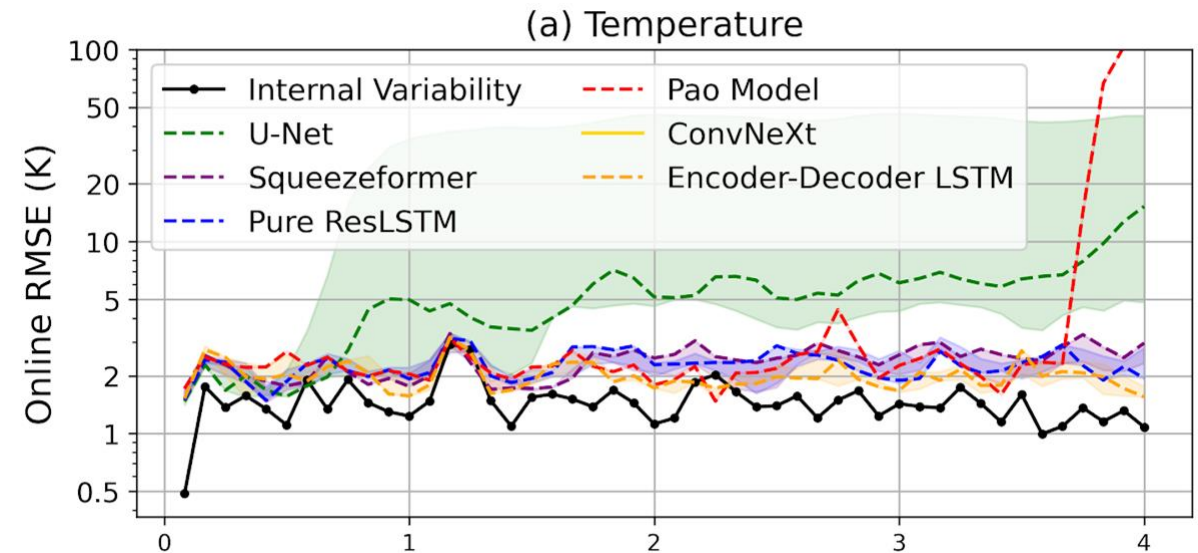


Different architectures respond differently to the same architecture agnostic design decisions

standard configuration

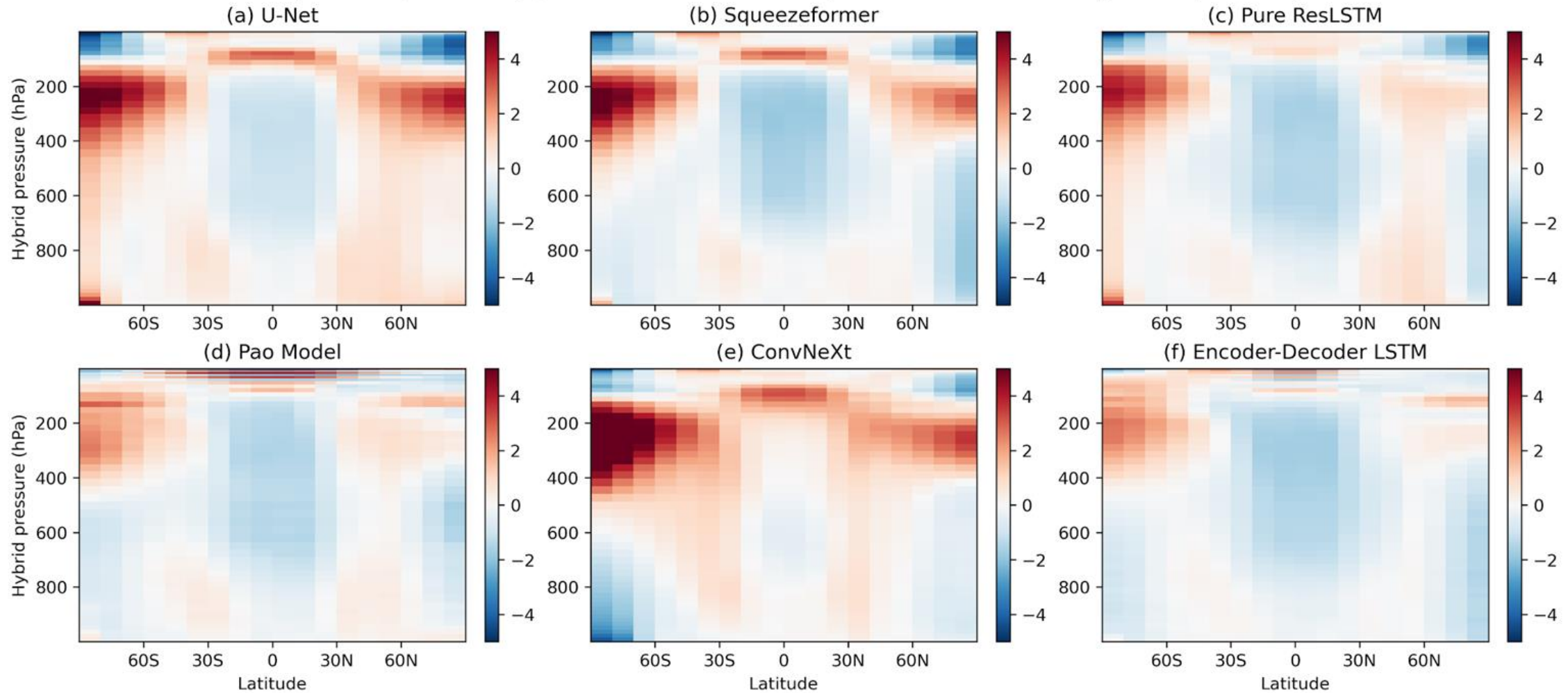


multirepresentation



Online biases are remarkably similar across architectures, underlining the limits of relying on a single ESM

5 year Temperature (K) zonal mean difference (Confidence Loss Configuration, Seed 7)



1) TCBench

Global TC track and
intensity forecasting

2) ClimSim

Hybrid physics-ML
climate emulation

3) HybridESMBench

Hybrid ESM simulation

Order 1 requirements

R1: Data available online without access restrictions

R2: Clear problem statement for meaningful task in atmospheric science

R3: Data input into high level open data science language provided

R4: Evaluation metrics defined analytically and in code

Order 2 requirements

R5: Simple example machine learning solution provided in code

R6: Visualisation and diagnostics provided in code

R7: Tests for physical consistency and explainability provided

R8: Benchmarks for the computational performance provided

1) TCBench

Global TC track and
intensity forecasting

2) ClimSim

Hybrid physics-ML
climate emulation

3) HybridESMBench

Hybrid ESM simulation

Order 1 requirements

R1: Data available online without access restrictions

R2: Clear problem statement for meaningful task in atmospheric science

R3: Data input into high level open data science language provided

R4: Evaluation metrics defined analytically and in code

Order 2 requirements

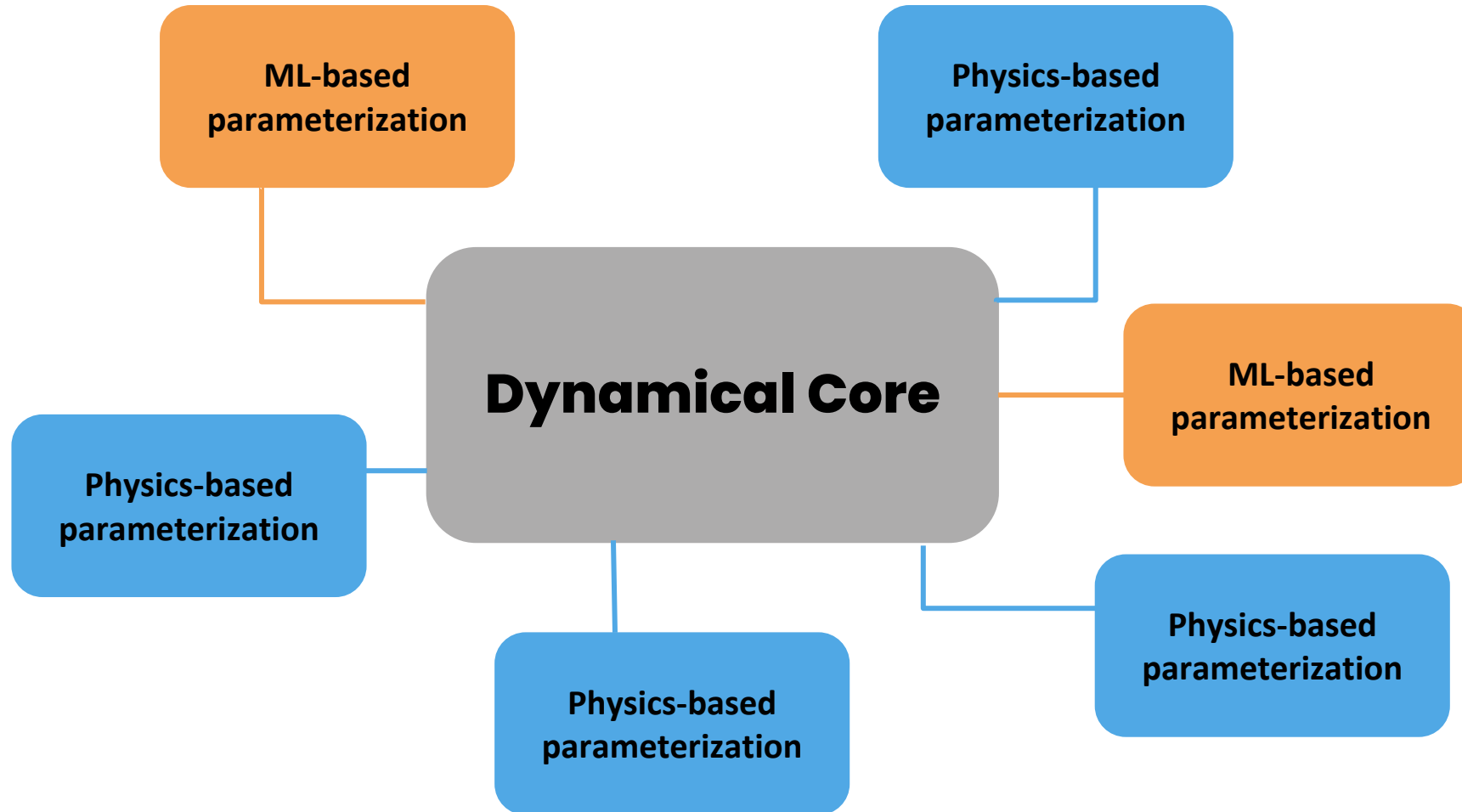
R5: Simple example machine learning solution provided in code

R6: Visualisation and diagnostics provided in code

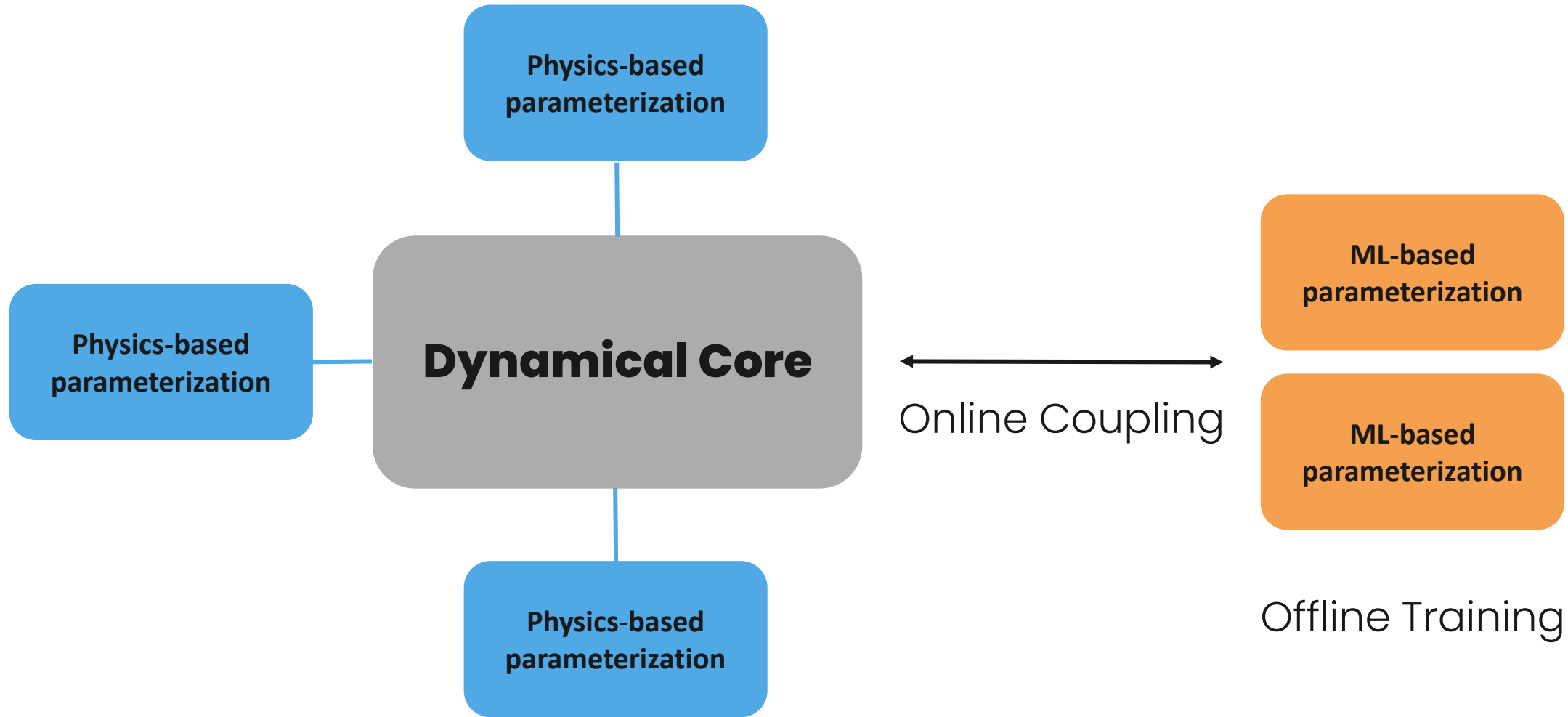
R7: Tests for physical consistency and explainability provided

R8: Benchmarks for the computational performance provided

Hybridizing Established ESMs



Challenges in Hybridizing ESMs



HybridESM-Bench

Goal: promote objective comparisons of hybrid ESMs.



Four Hybrid ESMs

ICON-A-MLe (tuned)

- Cloud cover ML parameterization
- Coupled to ICON-A

ARP-GEM (untuned)

- Deep convection ML parameterization
- Coupled to ARP-GEM

ClimSim-Online (untuned)

- Full subgrid-scale physics ML parameterization (atmospheric storms, clouds, turbulence, rainfall, and radiation)
- Coupled to E3SM

HadGEM3-GC5 (untuned)

- Cloud cover ML parameterization
- Coupled to HadGEM

HybridESMEval: Fast and Accessible Evaluation of AI-Powered Climate Models

Compare

Compare Hybrid Model Simulations to pre-processed CMIP6 simulations
Also usable for AI-MIP simulations

Easy to use

```
from hybridesmbench.eval import evaluate

simulation_path = "/path/to/simulation"
model_type = "icon" # or cmip, soon ClimSim
work_dir = "/path/to/output/hybridesmbench"

output = evaluate(icon_output, model_type,
                  work_dir,
                  diagnostics=["timeseries"])
```

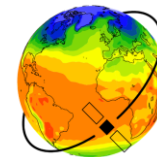
Open source



<https://github.com/HybridESM/HybridESMBench>

Diagnostics based on the Earth System Model Evaluation Tool (ESMValTool):

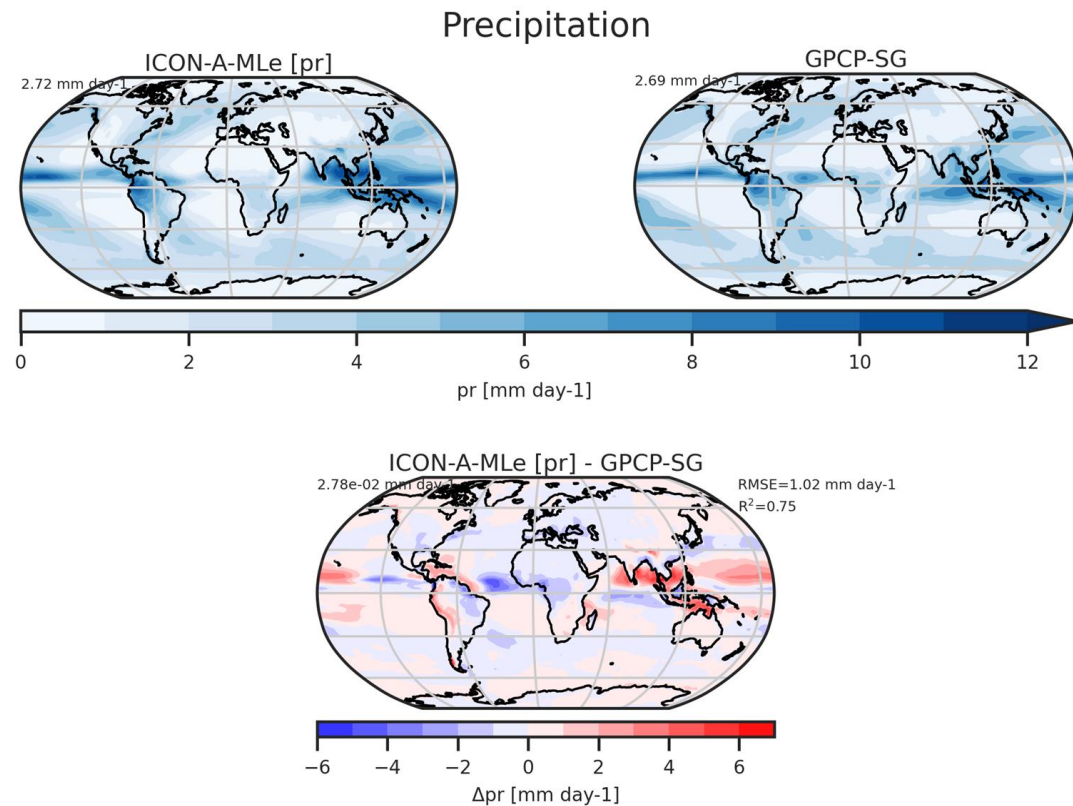
<https://github.com/ESMValGroup/ESMValTool>



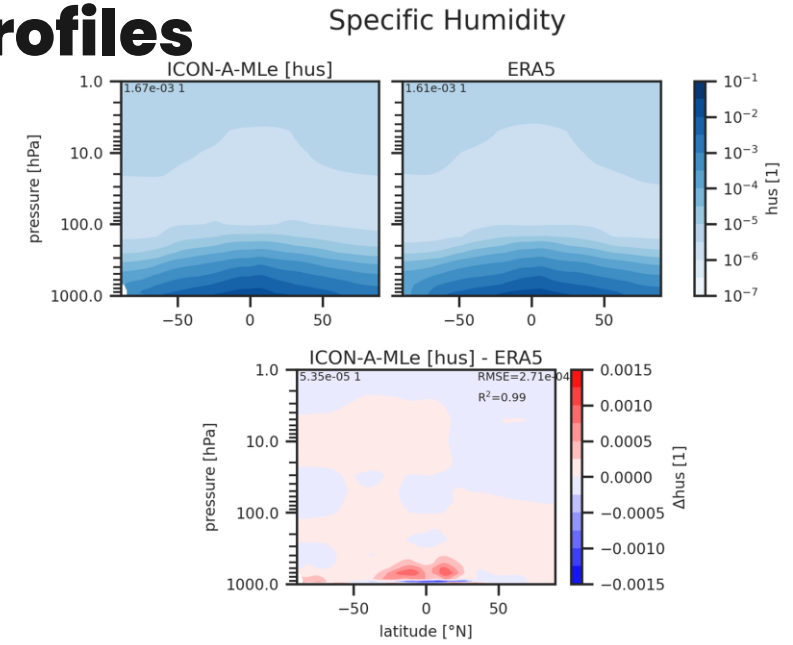
ESMValTool

Earth System Model Evaluation Tool

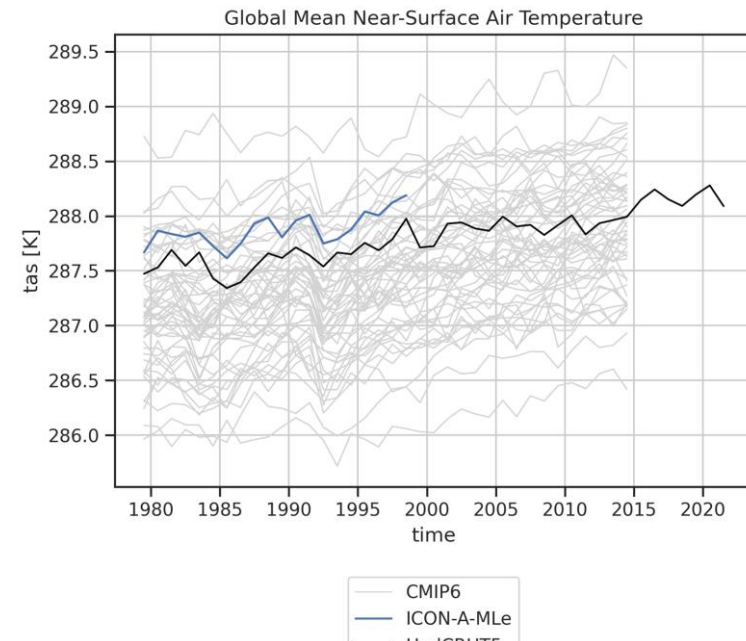
Current Evaluation Plots



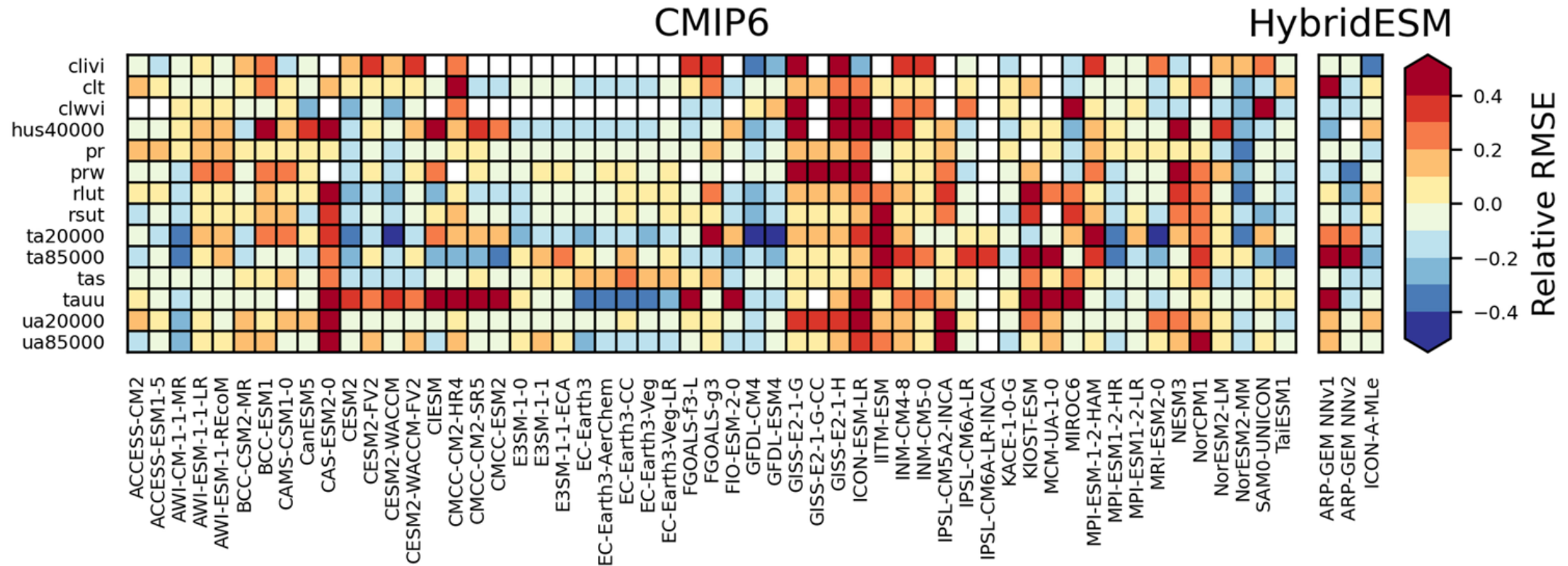
Profiles



Time Series



Portrait Plot

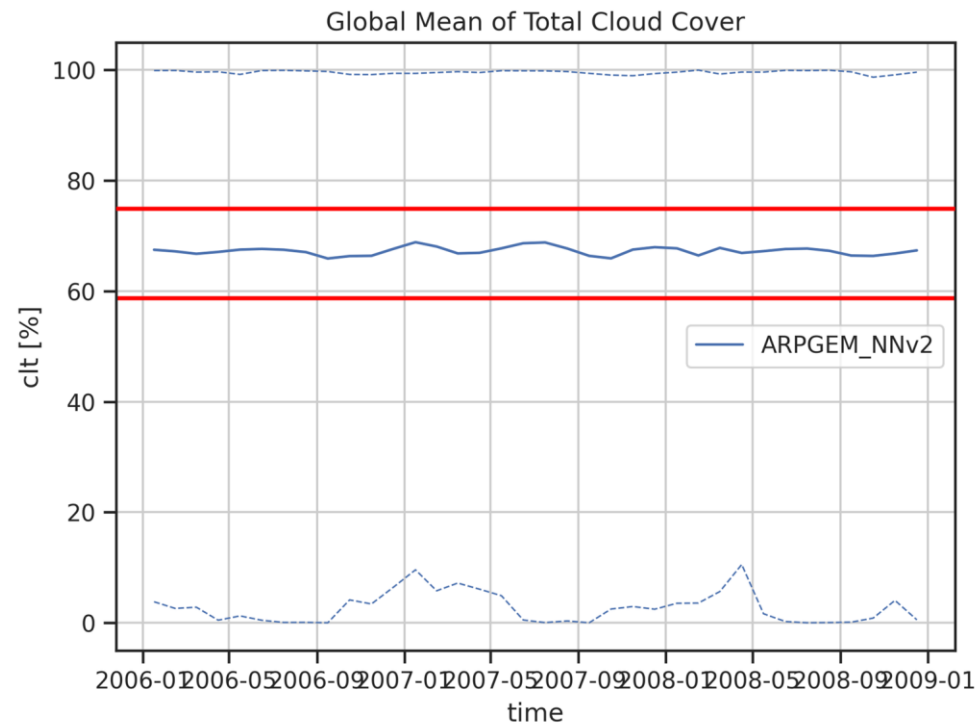
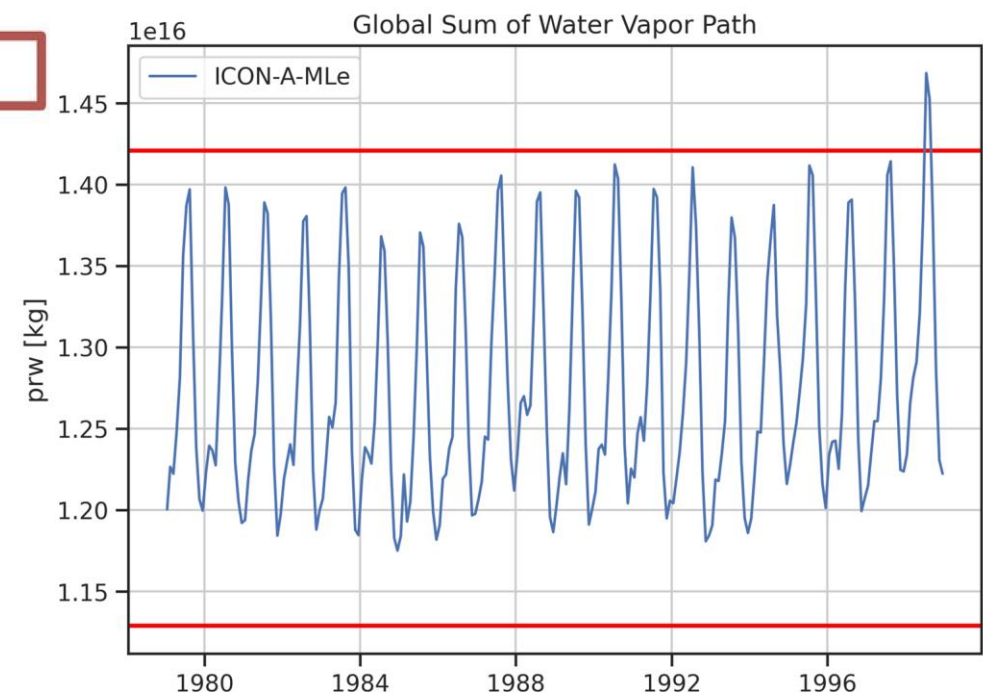


Compare Hybrid Model Simulations to pre-processed CMIP6 simulations

Relative Root Mean-Square Error compared to observations. The bluer the color, the better, the redder, the worse.

New in HybridESMEval: Basic Sanity Checks

- Check for negative mass concentrations occurring at any grid cell at any time (minimum is less than zero) or whether individual grid cells exceed physically reasonable values (e.g. total cloud fraction greater than 100%)
- Compare global monthly means to minimum and maximum values found in reference datasets, with reasonable limits shown as red lines



1) TCBench

Global TC track and
intensity forecasting

2) ClimSim

Hybrid physics-ML
climate emulation

3) HybridESMBench

Hybrid ESM simulation

Order 1 requirements

R1: Data available online without access restrictions

R2: Clear problem statement for meaningful task in atmospheric science

R3: Data input into high level open data science language provided

R4: Evaluation metrics defined analytically and in code

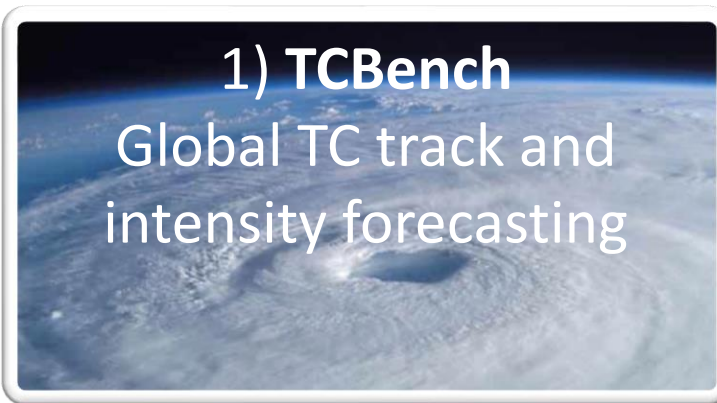
Order 2 requirements

R5: Simple example machine learning solution provided in code

R6: Visualisation and diagnostics provided in code

R7: Tests for physical consistency and explainability provided

R8: Benchmarks for the computational performance provided



1) TCBench

Global TC track and intensity forecasting

TCBENCH: A BENCHMARK FOR TROPICAL CYCLONE TRACK AND INTENSITY FORECASTING AT THE GLOBAL SCALE

Anonymous authors
Paper under double-blind review

Poster

ClimSim: A large multi-scale dataset for hybrid physics-ML climate emulation

Sungduk Yu · Walter Hannah · Liran Peng · Jerry Lin · Mohamed Aziz Bhouiri · Ritwik Gupta · Björn Lütjens · Justus C. Will · Gunnar Behrens · Julius Busecke · Nora Loose · Charles Stern · Tom Beucler · Bryce Harrop · Benjamin Hillman · Andrea Jenney · Savannah L. Ferretti · Nana Liu · Animashree Anandkumar · Noah Brenowitz · Veronika Eyring · Nicholas Geneva · Pierre Gentile · Stephan Mandt · Jaideep Pathak · Akshay Subramaniam · Carl Vondrick · Rose Yu · Laure Zanna · Tian Zheng · Ryan Abernathey · Fiaz Ahmed · David Bader · Pierre Baldi · Elizabeth Barnes · Christopher Bretherton · Peter Caldwell · Wayne Chuang · Yilun Han · YU HUANG · Fernando Iglesias-Suarez · Sanket Jantre · Karthik Kashinath · Marat Khairoutdinov · Thorsten Kurth · Nicholas Lutsko · Po-Lun Ma · Griffin Mooers · J. David Neelin · David Randall · Sara Shamekh · Mark Taylor · Nathan Urban · Janni Yuval · Guang Zhang · Mike Pritchard



Outstanding Paper

[Abstract] [Project Page]

Paper Poster OpenReview

2023 Poster



Home Page

Papers

Submissions

News

Editorial Board

Special Issues

Open Source

Software

Proceedings (PMLR)

Data (DMLR)

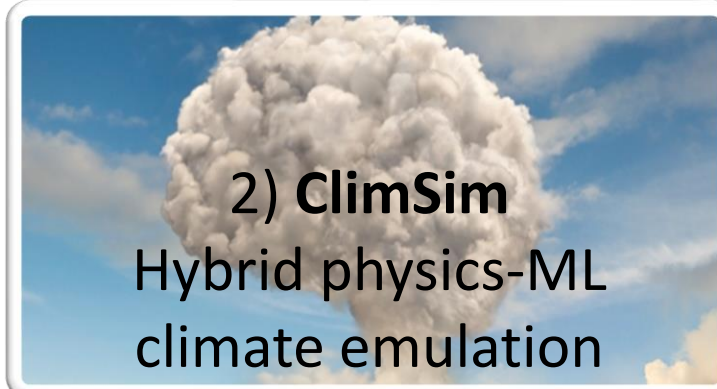
ClimSim-Online: A Large Multi-Scale Dataset and Framework for Hybrid Physics-ML Climate Emulation

Sungduk Yu, Zeyuan Hu, Akshay Subramaniam, Walter Hannah, Liran Peng, Jerry Lin, Mohamed Aziz Bhouiri, Ritwik Gupta, Björn Lütjens, Justus C. Will, Gunnar Behrens, Julius J. M. Busecke, Nora Loose, Charles I. Stern, Tom Beucler, Bryce Harrop, Helge Heuer, Benjamin R. Hillman, Andrea Jenney, Nana Liu, Alistair White, Tian Zheng, Zhiming Kuang, Fiaz Ahmed, Elizabeth Barnes, Noah D. Brenowitz, Christopher Bretherton, Veronika Eyring, Savannah L. Ferretti, Nicholas Lutsko, Pierre Gentile, Stephan Mandt, J. David Neelin, Rose Yu, Laure Zanna, Nathan M. Urban, Janni Yuval, Ryan Abernathey, Pierre Baldi, Wayne Chuang, Yu Huang, Fernando Iglesias-Suarez, Sanket Jantre, Po-Lun Ma, Sara Shamekh, Guang Zhang, Michael Pritchard, 26(142):1–85, 2025.

Abstract

Modern climate projections lack adequate spatial and temporal resolution due to computational constraints, leading to inaccuracies in representing critical processes like thunderstorms that occur on the sub-resolution scale. Hybrid methods combining physics with machine learning (ML) offer faster, higher fidelity climate simulations by outsourcing compute-hungry, high-resolution simulations to ML emulators. However, these hybrid physics-ML simulations require domain-specific data and workflows that have been inaccessible to many ML-experts. This paper is an extended version of our NeurIPS award-winning ClimSim dataset paper. The ClimSim dataset includes 5.7 billion pairs of multivariate input/output vectors spanning ten years at high temporal resolution, capturing the influence of high-resolution, high-fidelity physics on a host climate simulator's macro-scale state. In this extended version, we introduce a significant new contribution in Section 5, which provides a cross-platform, containerized pipeline to integrate ML models into operational climate simulators for hybrid testing. We also implement various baselines of ML models and hybrid simulators to highlight the ML challenges of building stable, skilful emulators. The data (https://huggingface.co/datasets/LEAP/ClimSim_high-res), also in a low-resolution version at https://huggingface.co/datasets/LEAP/ClimSim_low-res and an aquaplanet version at https://huggingface.co/datasets/LEAP/ClimSim_low-res_aqua-planet) and code (<https://leap-stc.github.io/ClimSim> and <https://github.com/leap-stc/climsim-online>) are publicly released to support the development of hybrid physics-ML and high-fidelity climate simulations.

[arXiv] [pdf] [bib] [code]



2) ClimSim

Hybrid physics-ML climate emulation

Crowdsourcing Ideas for Hybrid Physics-ML Climate Simulation from a \$50,000 Kaggle Competition



3) HybridESMBench

Hybrid ESM simulation

HybridESM-Bench: Towards an End-to-End Benchmark for Hybrid Earth System Models

Website: <https://wp.unil.ch/dawn/publications/>