Foundation Models for Earth Systems

Wessel Bruinsma

Microsoft Research AI for Science

The Lorentz Center Workshop: Advancing Ecosystem Carbon Flux Research
Leiden, 16 Oct 2024

The Aurora Team

Paris Perdikaris

University of Pennsylvania, formerly MSR

Megan Stanley

MSR

Wessel Bruinsma

MSR

Cristian Bodnar

Silurian, formerly MSR

Ana Lučić

University of Amsterdam, formerly MSR

Richard Turner

University of Cambridge, formerly MSR

Elizabeth Heider

Book tour, formerly MSR

Johannes Brandstetter

JKU Linz, NXAI, formerly MSR

Patrick Garvan

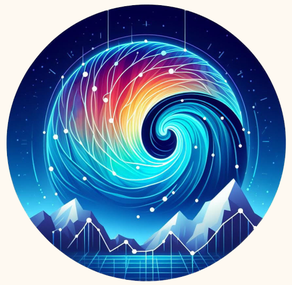
Formerly MSR

Maik Riechert

MSR

Max Welling

University of Amsterdam, CuspAI, formerly MSR

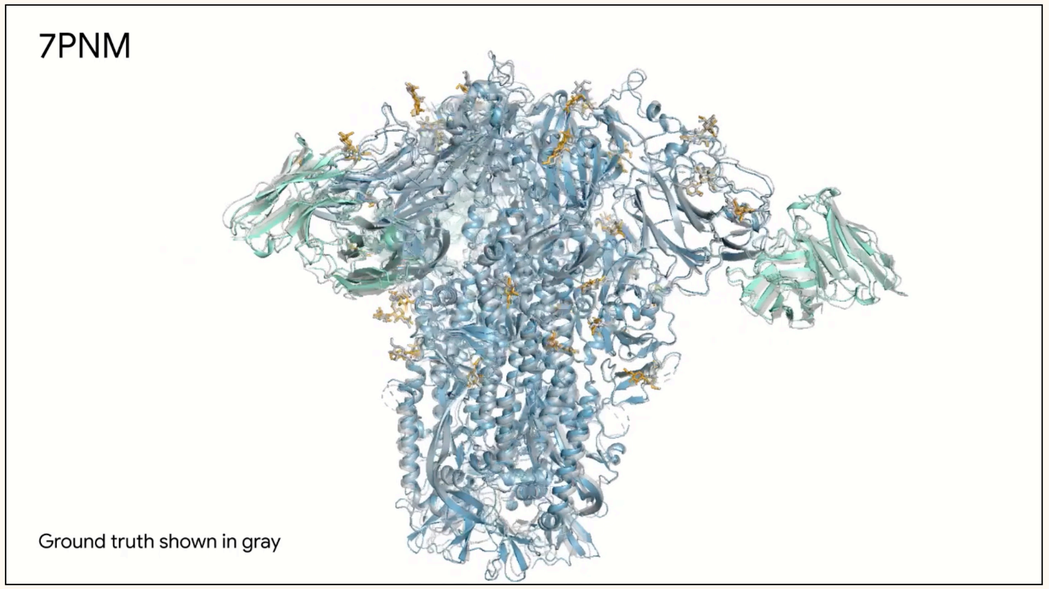


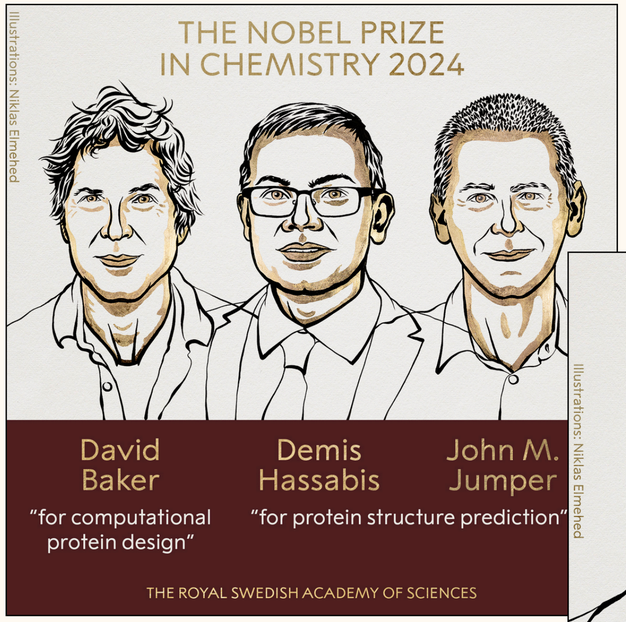
Outline

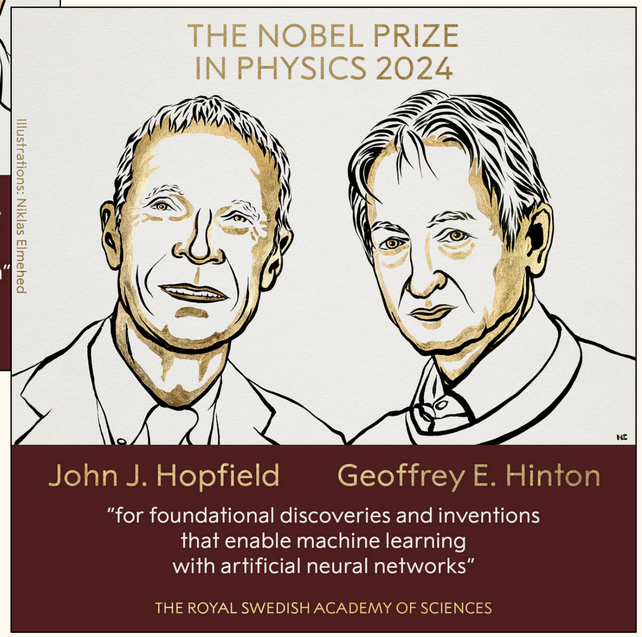
- The AI Revolution in Medium-Term Weather Forecasting
- Aurora: A Foundation Model of the Atmosphere
- Towards a Foundation Model of the Earth

The AI Revolution in Science

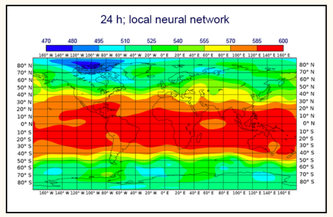
AlphaFold
Protein folding







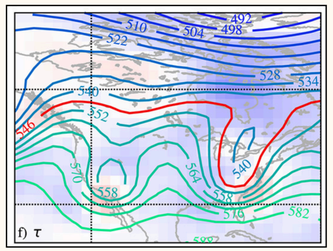
The AI Revolution in Weather Forecasting



2018

First serious efforts to compare AI models to physics baselines

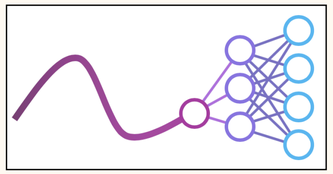
Dueben and Bauer (2018)



2019

AI models skillful to multiple days

Weyn et al. (2019)

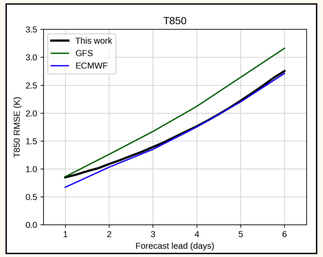


2020

WeatherBench starts to drive ML development

Rasp et al. (2018)

The AI Revolution in Weather Forecasting



HUAWEI

2022

GNN outperforms GFS at 1°

Keisler (2022)

2023

Pangu-Weather outperforms HRES at 0.25°

Bi et al. (2023)

The AI Revolution in Weather Forecasting

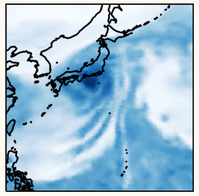
NVIDIA

Microsoft

Google DeepMind

2022–2023

Tech companies start to work in this space



2023

GenCast outperforms IFS ensemble

Price et al. (2023)

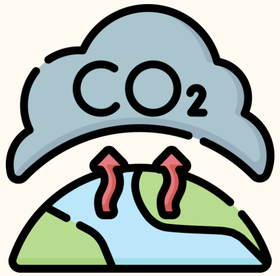
ECMWF

2024

ECMWF launches AIFS

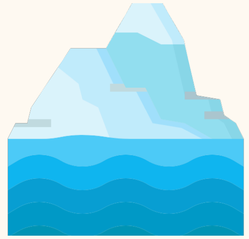
What About Other Forecasting Tasks?

Air pollution



Atmospheric composition

Waves

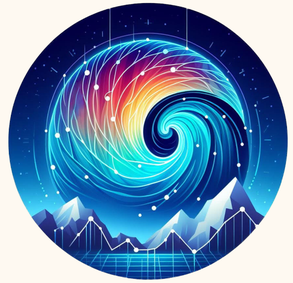


Sea ice

Ocean

- Current models are impressive, but **limited to one setting**.
- Unified approach?

Aurora



pretraining

- Train a single neural network a *large* body of atmospheric data
- Learn universal representation of atmospheric dynamics
- Slow and data hungry

fine-tuning

- Leverage learned representation to **efficiently adapt to new domains**!
- Fast and data efficient

Aurora: a **foundation model** of the atmosphere

Aurora

Pretraining

- Predict global state of *any variables* at any resolution 6 h ahead

Variable	Units	Description
SURFACE-LEVEL METEOROLOGICAL VARIABLES		
2T	K	Temperature at 2 m above surface of land or sea
U10	$m\,s^{-1}$	Eastward component of wind at 10 m
V10	$m\,s^{-1}$	Southward component of wind at 10 m
WS	$m\,s^{-1}$	Wind speed at 10 m; equal to $(U10^2 + V10^2)^{1/2}$
MSL	Pa	Air pressure at mean sea level
ATMOSPHERIC METEOROLOGICAL VARIABLES		
U	$m\,s^{-1}$	Eastward component of wind
V	$m\,s^{-1}$	Southward component of wind
T	K	Temperature
Q	$kg\,kg^{-1}$	Specific humidity
Z	$m^2\,s^{-2}$	Geopotential

Cost:

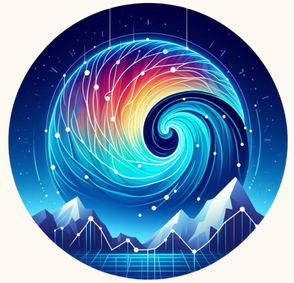
- 150 000 steps
- 32 A100s
- 3 weeks

Pretraining Datasets

Name	Resolution	Timeframe	Surface Variables	Atmospheric Variables	Num levels	Size (TB)	Num frames
ERA5	0.25° × 0.25°	1979-2020	2T, U10, V10, MSL	U, V, T, Q, Z	13	105.43	367,920
HRES-0.25	0.25° × 0.25°	2016-2020	2T, U10, V10, MSL	U, V, T, Q, Z	13	42.88	149,650
IFS-ENS-0.25	0.25° × 0.25°	2018-2020	2T, U10, V10, MSL	U, V, T, Q, Z	3	518.41	6,570,000
GFS Forecast	0.25° × 0.25°	2015-2020	2T, U10, V10, MSL	U, V, T, Q, Z	13	130.39	560,640
GFS Analysis	0.25° × 0.25°	2015-2020	2T, U10, V10, MSL	U, V, T, Q, Z	13	2.04	8,760
GEFS Reforecast	0.25° × 0.25°	2000-2019	2T, MSL	U, V, T, Q, Z	3	194.02	2,920,000
CMCC-CM2-VHR4	0.25° × 0.25°	1950-2014	2T, U10, V10, MSL	U, V, T, Q	7	12.6	94,900
ECMWF-IFS-HR	0.45° × 0.45°	1950-2014	2T, U10, V10, MSL	U, V, T, Q	7	3.89	94,900
MERRA-2	0.625° × 0.5°	1980-2020	2T, U10, V10, MSL	U, V, T, Q	13	5.85	125,560
IFS-ENS-Mean	0.25° × 0.25°	2018-2020	2T, U10, V10, MSL	U, V, T, Q, Z	3	10.37	131,400
					Total	1,219.91	11,023,730

Aurora

Air Pollution Forecasting



- **Setup:** model levels of PM_1, $PM_{2.5}$, PM_{10}, CO, NO, NO_2, SO_2, O_3
- **Data:** Copernicus Atmospheric Monitoring Service (CAMS) analysis
- **Baseline:** CAMS forecasts

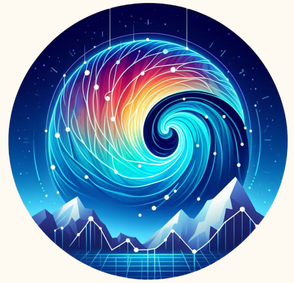
Coupled to IFS, ~10x more expensive:

~16 node-hours per hour lead time!

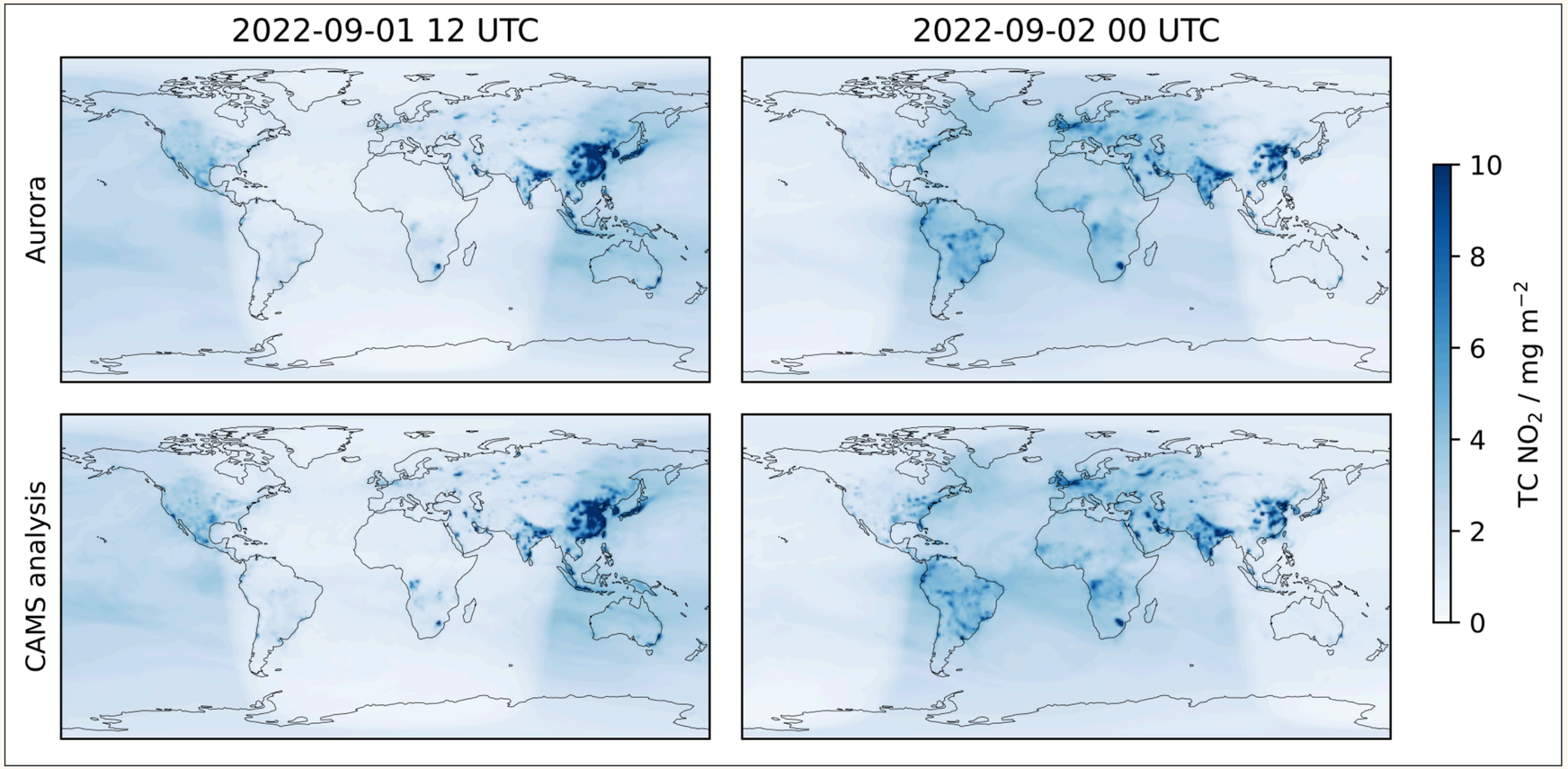
Aurora: **~1.1 s per hour lead time**

Aurora

Air Pollution Forecasting (2)



- Heterogeneous and spikey
- Anthropogenic factors
- Scarce
- Non-stationary



Overall:

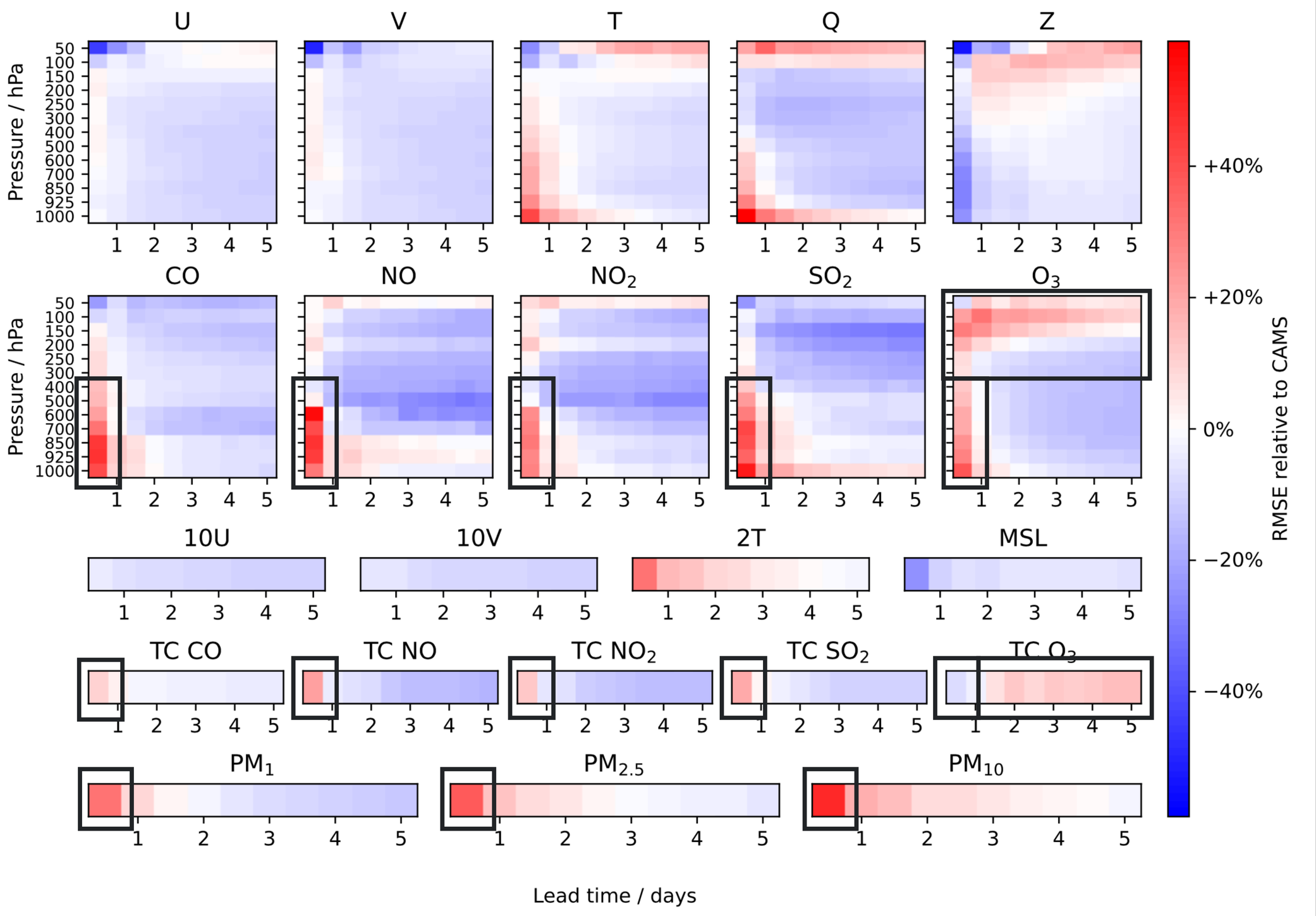
Competitive on 95%
(≤ 20% RMSE)

Better on 75%

Three days:

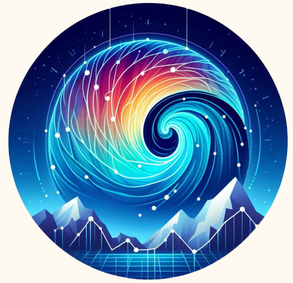
Competitive on 100%
(≤ 20% RMSE)

Better on 86%



Aurora

Wave Forecasting (Fresh off the Press!)

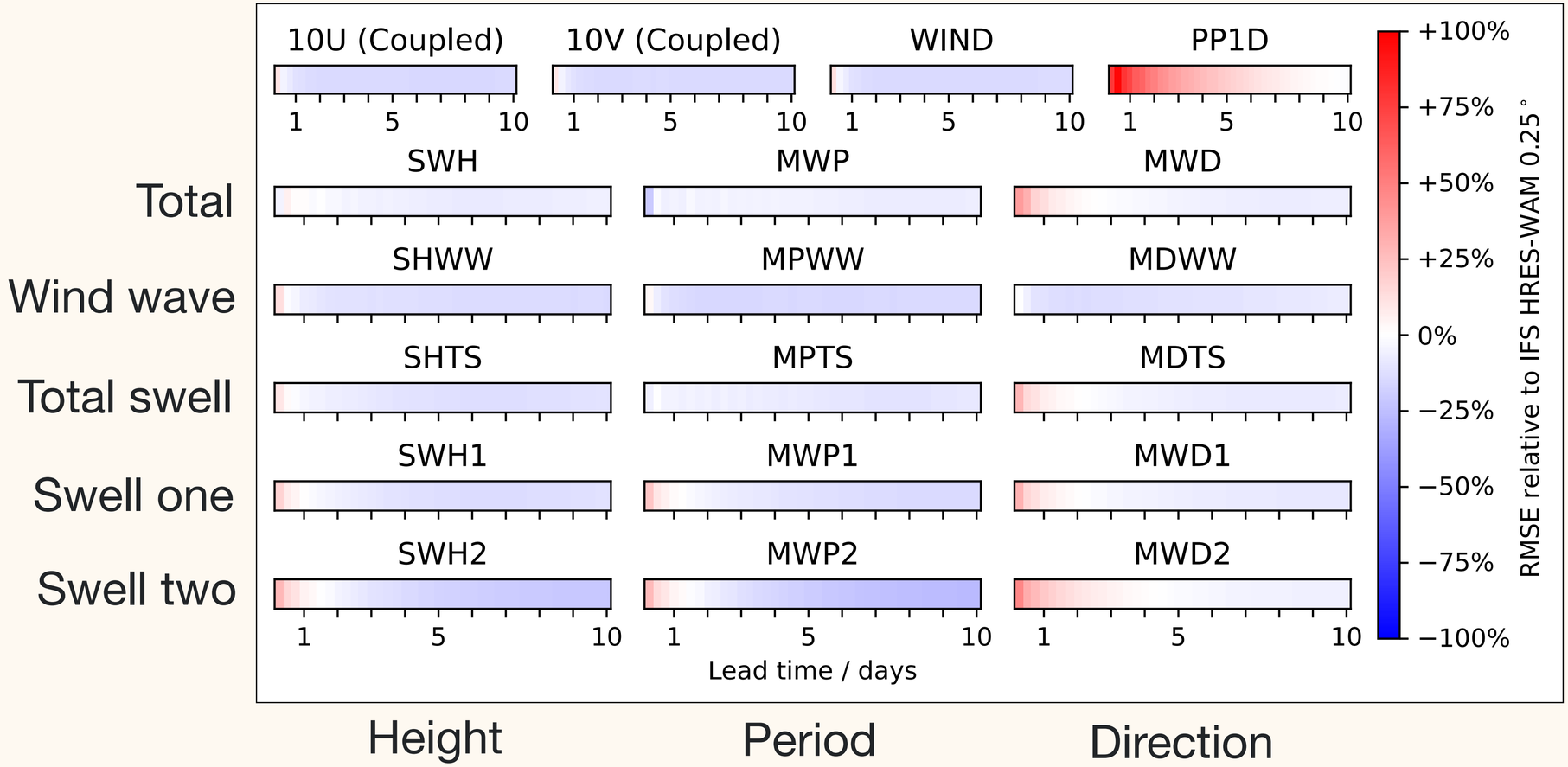


- **Setup:** model height, period, and direction of all wave components
- **Data:** IFS HRES-WAM analysis
- **Baseline:** IFS HRES-WAM forecasts

- Where data is defined is variable (e.g. absence of swell, sea ice)
- How to model angles?

Aurora

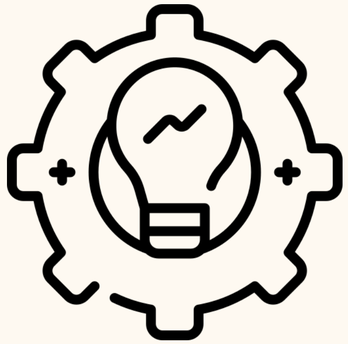
Wave Forecasting (Fresh off the Press!) (2)



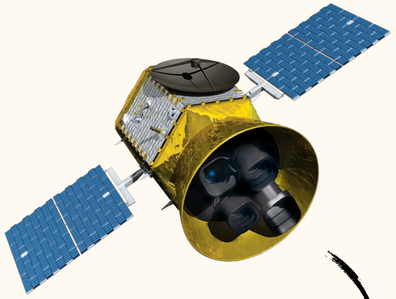
Vision: A Foundation Model of the Earth

New domains

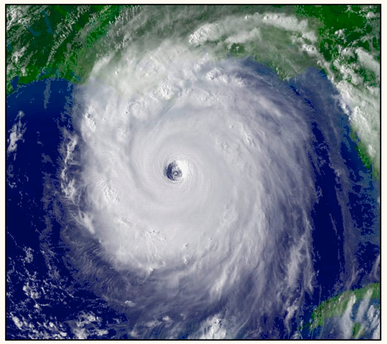
Methodological development



Observational data



Principled evaluation



Aardvark-Weather

End-to-end weather forecasting

Summary

- Medium-term weather forecasting has seen incredible progress
- **Pretraining–fine-tuning paradigm** to extend these advancements to other domains
- Can we build towards a foundation model of the Earth?

wessel.ai/pdf/aurora
wessel.ai/pdf/aardvark
hi@wessel.ai