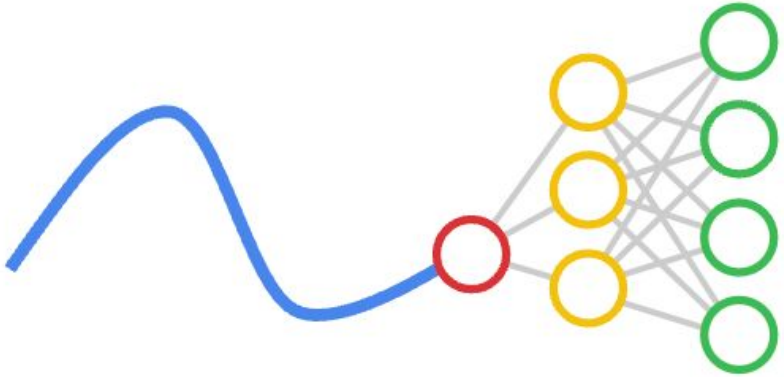


WeatherBench 2.0

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

Stephan Rasp, Google Research
with collaborators from
Google, DeepMind and ECMWF

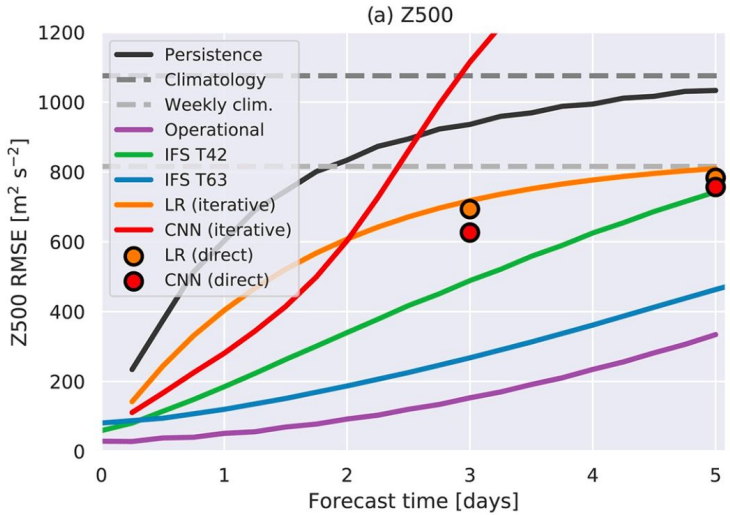
Google Research

WeatherBench v1 and what happened since

Google Research

WeatherBench v1

- Benchmark dataset for global medium-range weather forecasting
- Published in 2020[1] on the back of first studies
- Evaluation
  - Variables: Z500, T850, T2M & Precipitation
  - Metrics: RMSE & ACC
- Ground truth: ERA5 at 5.625°
- Statistical and physical baselines
- Regridded data available on public server
- github.com/pangeo-data/WeatherBench



Google Research

[1] Rasp et al. 2020. JAMES

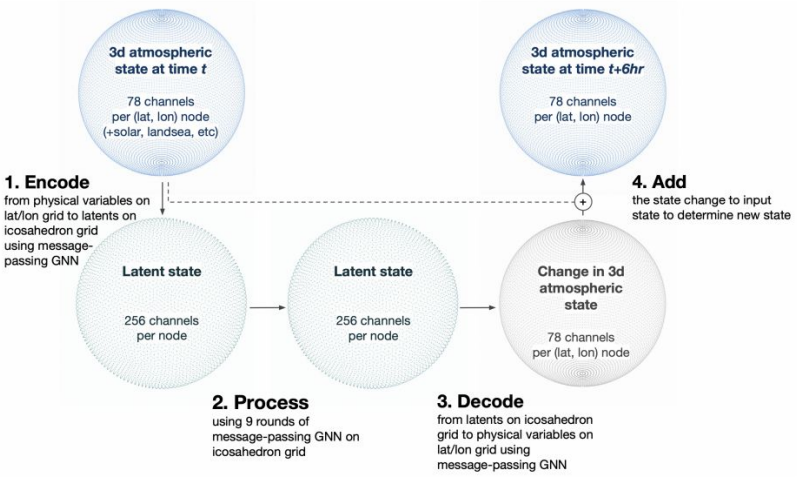
Leaderboard

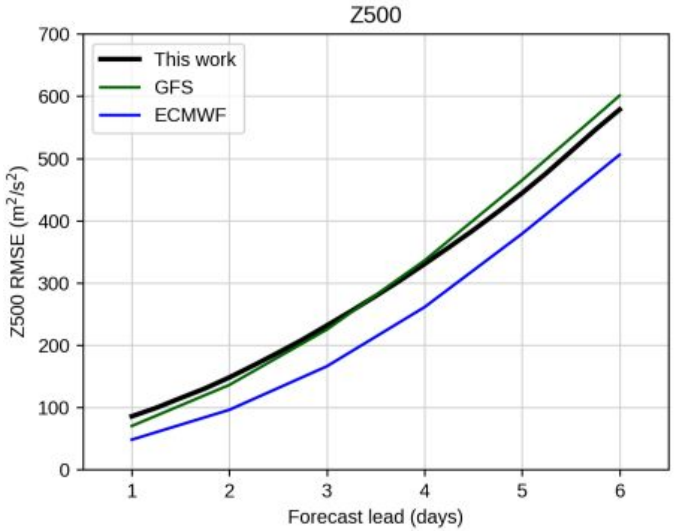
Model	Z500 RMSE (3 / 5 days) [m^2/s^2]	T850 RMSE (3 / 5 days) [K]	Notes	Reference
Operational IFS	154 / 334	1.36 / 2.03	ECWMF physical model (10 km)	Rasp et al. 2020
Rasp and Thuerey 2020 (direct/continuous)	**268 / 499**	**1.65 / 2.41**	Resnet with CMIP pretraining (5.625 deg)	Rasp and Thuerey 2020
IFS T63	268 / 463	1.85 / 2.52	Lower resolution physical model (approx. 1.9 deg)	Rasp et al. 2020
Weyn et al. 2020 (iterative)	**373 / 611**	**1.98 / 2.87**	UNet with cube-sphere mapping (2 deg)	Weyn et al. 2020
Clare et al. 2021 (direct)	375 / 627	2.11 / 2.91	Stacked ResNets with probabilistic output (5.625 deg)	Clare et al. 2021
IFS T42	489 / 743	3.09 / 3.83	Lower resolution physical model (approx. 2.8 deg)	Rasp et al. 2020
Weekly climatology	816	3.50	Climatology for each calendar week	Rasp et al. 2020
Persistence	936 / 1033	4.23 / 4.56		Rasp et al. 2020
Climatology	1075	5.51		Rasp et al. 2020

Google Research

Keisler 2022[2]: Graph Neural Networks

- Atmospheric state represented as a graph
- Upper level mean skill comparable to GFS at 1° resolution



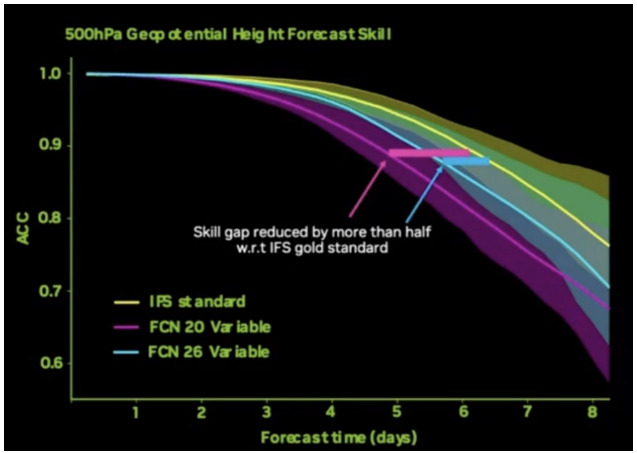


Google Research

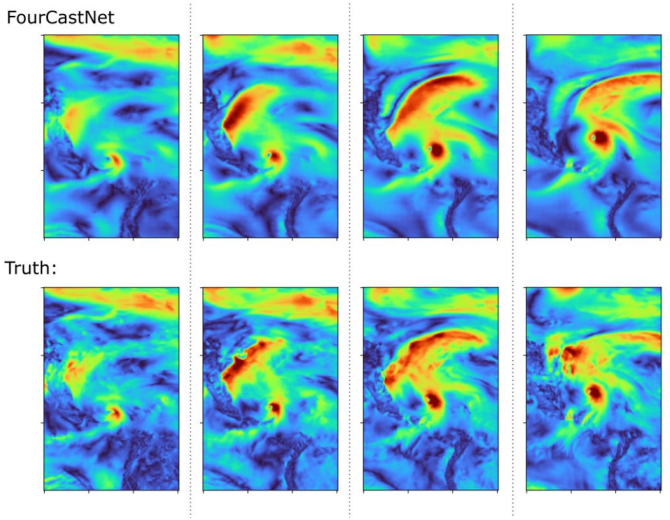
[2] Keisler 2022. Forecasting Global Weather with Graph Neural Networks. arXiv

Pathak et al. 2022[3]: FourCastNet

- Modified vision transformer
- Predictions at 0.25° resolution



Nvidia GTC Talk

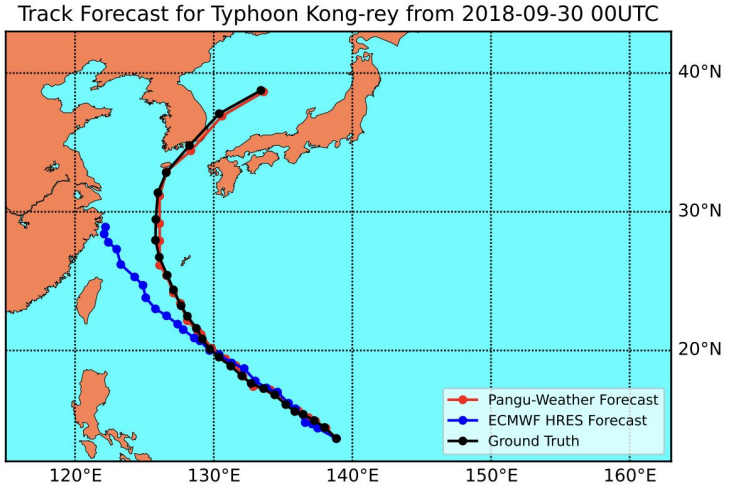


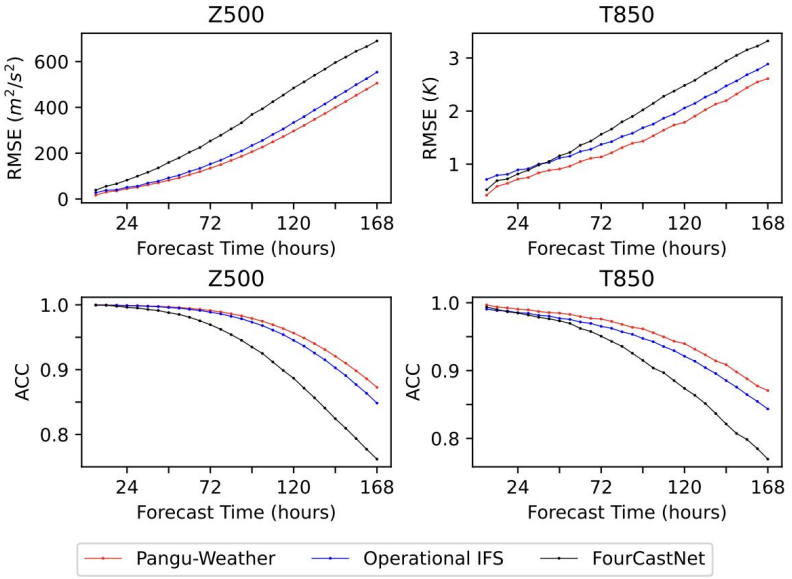
Google Research

[3] Pathak et al. 2022. FourCastNet: [...]. arXiv

Bi et al. 2022[4]: Pangu-Weather

- Another vision transformer at 0.25 degree resolution





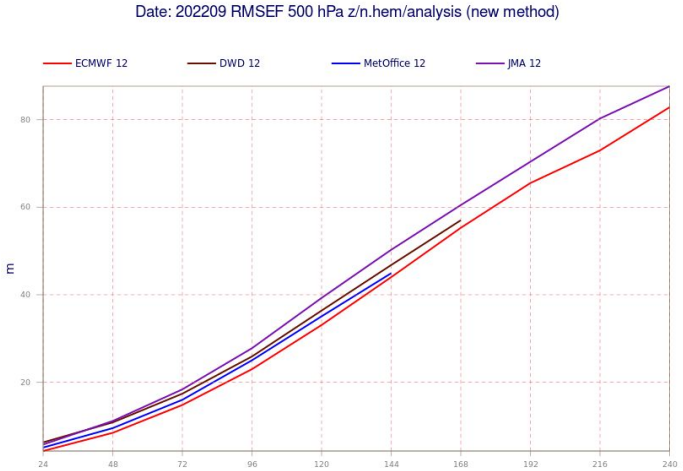
Google Research

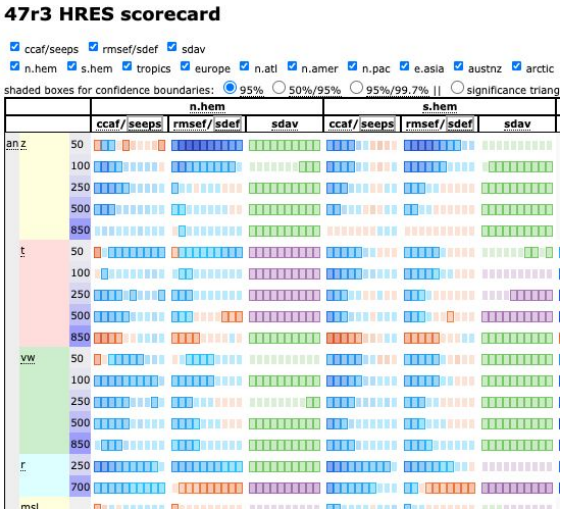
[4] Bi et al. 2022. Pangu-Weather: [...]. arXiv

Design principles for WeatherBench 2.0

Trusted evaluation

- Stay close to operational evaluation at WMO[4] and ECMWF[5]
- ERA5 ground truth
- 1.5° resolution
- Two years of verification
- IFS HRES and ENS baselines
- Support for more variables and levels



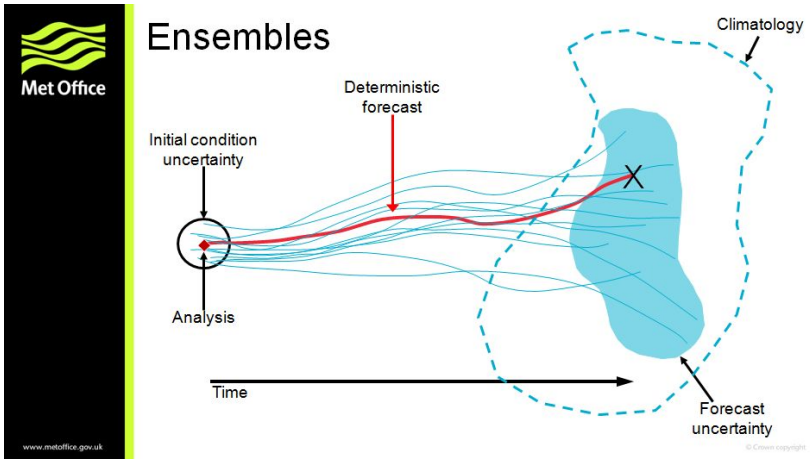


[4] https://apps.ecmwf.int/wmolcdnv/
[5] https://sites.ecmwf.int/ifs/scorecards/scorecards-47r3HRES.html#

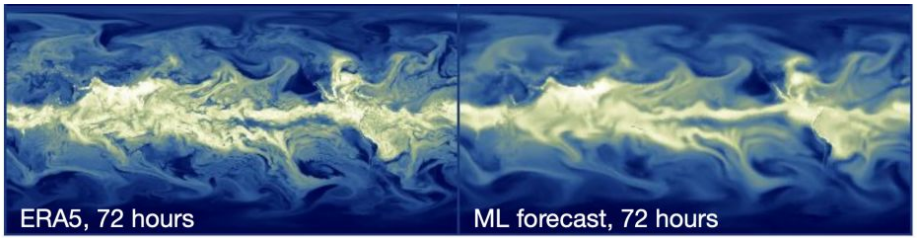
Google Research

Focus on probabilistic forecasting

- Medium-range forecasting is fundamentally probabilistic
- Include CRPS and spread-skill ratio
- Include IFS ENS baseline
- Ensemble mean as a deterministic baseline baseline
- Spectra to evaluate realism



https://www.metoffice.gov.uk/research/weather/ensemble-forecasting/what-is-an-ensemble-forecast



Keisler 2022

Google Research

Headline scores

- 4 variables to capture upper-level, large-scale dynamics
- 4 surface variables to capture weather impact

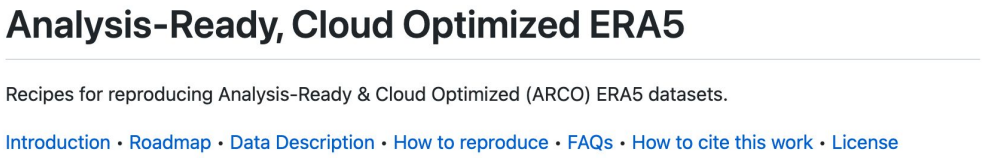
Variable	Short name	Deterministic metric	Probabilistic metric
Upper-level variables			
500hPa Geopotential	Z500	RMSE	CRPS
850hPa Temperature	T850	RMSE	CRPS
700hPa Relative humidity	R700	RMSE	CRPS
850hPa Wind vector	FF850	RMSE	CRPS
Surface variables			
2m Temperature	T2M	RMSE	CRPS
10m Wind speed	WS10	RMSE	CRPS
Mean sea level pressure	MSLP	RMSE	CRPS
6h precipitation	PR	SEEPS(TBD)	CRPS

Google Research

An expandable, open framework

- All data will be available on GCS
- All code will be available on GitHub
- Space for new analyses in the future
- ETA for WB2.0: Q1 2023
- Note: Not only forecasting, also post-processing

Google's effort to make weather data available for free: ARCO-ERA5
https://github.com/google-research/arco-era5



Google Research

Frontiers of data-driven weather forecasting

Google Research

WeatherBench 2.X

1) Resolution and ground truth
    - WB2.0 will probably also offer 0.25 degree evaluation but...
    - ERA5 is not the truth (especially for precipitation)
    - At small scales, station observations are the ground truth
    - Regional observation datasets (e.g. radar networks)
2) Focus on extreme events
    - Anecdotal: Looking at individual extreme events
    - Statistical: Metrics designed to capture extremes (heat, heavy precipitation, etc.)
3) Evaluating realism
    - How can we measure whether forecasts are realistic?

And more based on feedback from you!

Google Research

How to compare traditional and ML models?

1) Physics-based NWP serves a huge number of use cases beyond WB2 output

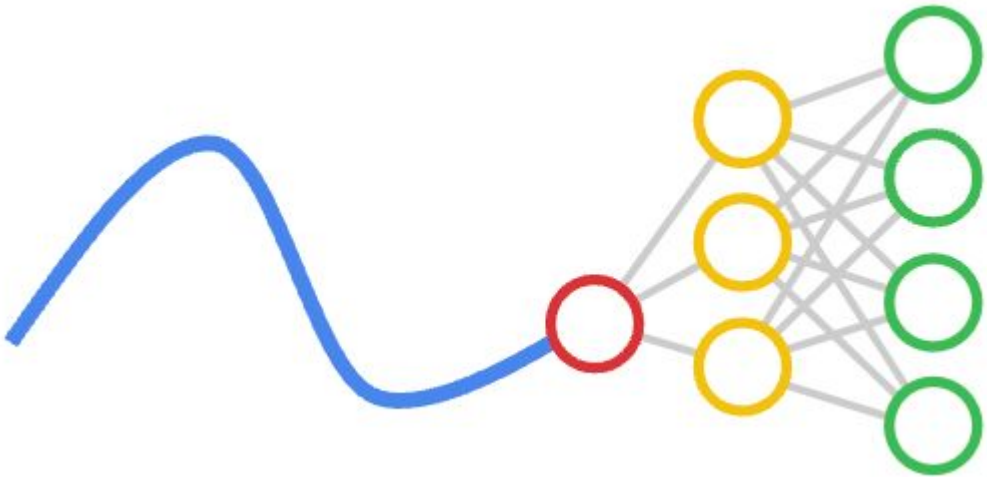
2) Data assimilation is equally as important as the forecast model

3) Usefulness of forecasting system defined by its reliability of predicting extremes; needs to "gain" trust

→ Warning: Good WeatherBench2 scores necessary but not sufficient

→ But also: Good WeatherBench2 scores provide evidence that ML can be seriously competitive

Google Research



Stephan Rasp (srasp@google.com)

With lots of help from
Stephan Hoyer, Alex Merose, Fei Sha and many others @ Google Research
Peter Battaglia @ DeepMind
Zied Ben Bouallegue and Matthew Chantry @ ECMWF
And conversations with many others, e.g. @ Nvidia

Google Research