WeatherBench >= 2

What's next for AI-weather models and evaluation?

Stephan Rasp, Google Research

With help from many others:
Stephan Hoyer, Peter Battaglia, Alex Merose, Ian Langmore, Tyler Russell, Alvaro Sanchez, Rob Carver, Vivian Yang, Shreya Agrawal, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha (Google)

Matthew Chantry, Zied Ben Bouallegue and Peter Dueben (ECMWF)

WeatherBench 2 - Status Quo

WeatherBench 2 is a benchmark for **global, medium-range weather prediction**.

It consists of:

1. **Data**: Relevant data freely available as Zarr on GCS (ERA5, IFS HRES and ENS, ML forecasts).
2. **Code**: Parallelizable and reproducible evaluation code on GitHub.
3. **Website**: Up-to-date platform showing state-of-the-art of AI-weather modeling.

For background information, check out our **paper** (Arxiv, soon to be published in JAMES).

For technical information, visit the **GitHub page** and the **documentation**.

Lessons from the leaderboard

	Pressure: 500hPa geopotential RMSE [kg^2/m^2]					Temperature: 850hPa temperature RMSE [K]					Humidity: 700hPa specific humidity RMSE [g/kg]					Wind Vector: 850hPa wind vector RMSE [m/s]				
Lead time [days]	1	3	5	7	10	1	3	5	7	10	1	3	5	7	10	1	3	5	7	10
Physical models																				
IFS HRES	42	135	304	521	801	0.62	1.16	1.82	2.63	3.63	0.55	0.96	1.27	1.53	1.81	1.69	3.29	5.20	7.11	9.14
IFS ENS Mean	42	132	277	439	621	0.65	1.11	1.62	2.17	2.80	0.51	0.84	1.06	1.22	1.38	1.63	2.98	4.44	5.74	6.94
ERA5 Forecasts	43	142	316	534	811	0.59	1.19	1.87	2.68	3.66	0.53	1.01	1.33	1.59	1.86	1.63	3.40	5.37	7.26	9.23
ML / hybrid models																				
Pangu-Weather (oper.)	45	136	300	510	785	0.65	1.09	1.74	2.54	3.55	0.53	0.86	1.17	1.45	1.76	1.71	3.03	4.85	6.75	8.82
GraphCast (oper.)	40	124	277	477	751	0.53	0.93	1.56	2.36	3.40	0.48	0.76	1.03	1.29	1.59	1.48	2.74	4.52	6.41	8.53
Keisler (2022)	66	174	345	544	787	0.81	1.22	1.87	2.63	3.55	0.65	0.94	1.19	1.41	1.65	2.26	3.51	5.17	6.85	8.62
Pangu-Weather	44	133	294	501	778	0.62	1.05	1.71	2.51	3.54	0.53	0.88	1.19	1.47	1.79	1.66	3.00	4.82	6.71	8.79
GraphCast	39	124	274	468	731	0.51	0.94	1.56	2.33	3.36	0.47	0.79	1.06	1.30	1.59	1.42	2.76	4.44	6.22	8.17
FuXi	40	125	276	433	631	0.54	0.97	1.59	2.14	2.91						1.47	2.80	4.49	5.64	7.02
SphericalCNN	54	161	338	546	815	0.73	1.18	1.86	2.64	3.62	0.59	0.89	1.17	1.43	1.72	2.05	3.38	5.17	7.01	8.98
NeuralGCM 0.7°	37	115	267	469	751	0.54	0.97	1.58	2.38	3.42	0.48	0.83	1.12	1.40	1.71	1.49	2.81	4.57	6.49	8.64
NeuralGCM ENS Mean	43	126	266	424	606	0.65	1.02	1.53	2.10	2.75	0.54	0.81	1.02	1.19	1.37	1.76	2.88	4.28	5.59	6.83
Climatology	820	820	820	820	820	3.44	3.44	3.44	3.44	3.44	1.59	1.59	1.59	1.59	1.59	7.89	7.89	7.89	7.89	7.89

−50 −20 −10 −5 −2 −1 1 2 5 10 20 50

Better ⟵ % difference in RMSE vs IFS HRES ⟶ Worse

- ML models are roughly on par with physics-based models.
- Many deterministic ML models blur. Spectra are somewhere in-between HRES and ENS.
- Overall, most ML models have similar performance. ERA5 seems to be the main limiting factor.

Google Research

Lessons from the leaderboard

Pressure — 500hPa geopotential CRPS [kg^2/m^2]

Lead time [days]	1	3	5	10	15
IFS ENS	22	58	114	262	323
NeuralGCM ENS	22	54	108	253	318
Climatology	369	369	369	369	369

Temperature — 850hPa temperature CRPS [K]

Lead time [days]	1	3	5	10	15
IFS ENS	0.32	0.53	0.76	1.32	1.54
NeuralGCM ENS	0.33	0.49	0.71	1.29	1.54
Climatology	1.69	1.69	1.69	1.69	1.69

Humidity — 700hPa specific humidity CRPS [g/kg]

Lead time [days]	1	3	5	10	15
IFS ENS	0.22	0.39	0.50	0.69	0.75
NeuralGCM ENS	0.26	0.37	0.48	0.67	0.75
Climatology	0.81	0.81	0.81	0.81	0.81

Wind Speed — 850hPa wind speed CRPS [m/s]

Lead time [days]	1	3	5	10	15
IFS ENS	0.56	0.97	1.35	1.94	2.10
NeuralGCM ENS	0.64	0.95	1.30	1.91	2.10
Climatology	2.19	2.19	2.19	2.19	2.19

−50 −20 −10 −5 −2 −1 1 2 5 10 20 50

Better ⟵ % difference in CRPS vs IFS ENS ⟶ Worse

- Less progress on ensemble methods.
- Existing models roughly on par with IFS ENS.

\+ GenCast (and Pangu ensemble)

Google Research

From research to operations

ERA5 is not available for initialization in real time.

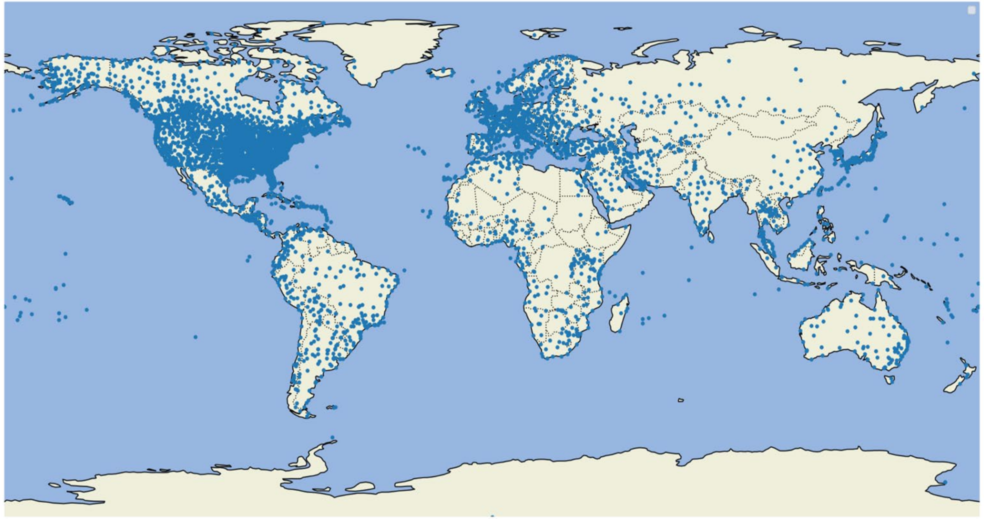
ERA5 is often not the best ground truth for impactful weather (see ECMWF's "lower half of the scorecard").

Many real-world applications require post-processing to higher-quality datasets.

Forecast latency matters.

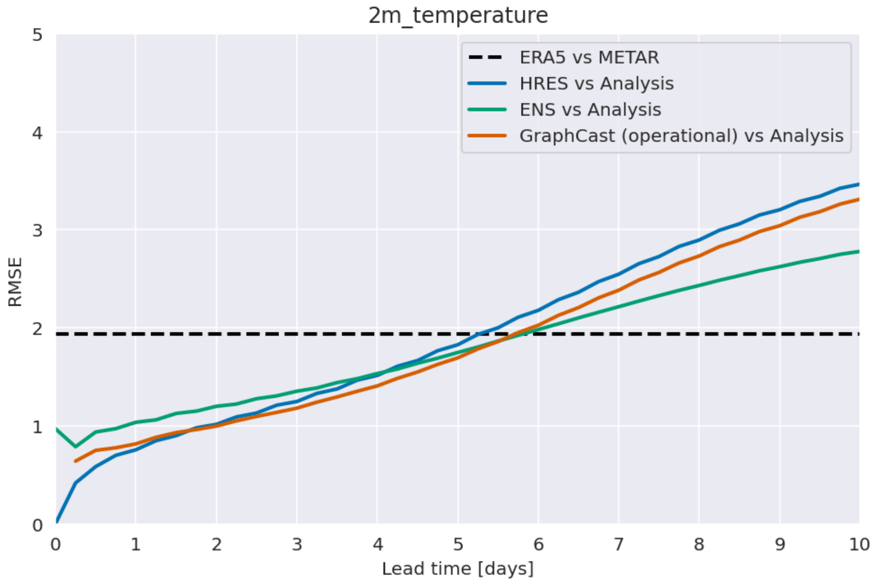
→ The next step in evaluation: operational conditions and "best" ground truth.

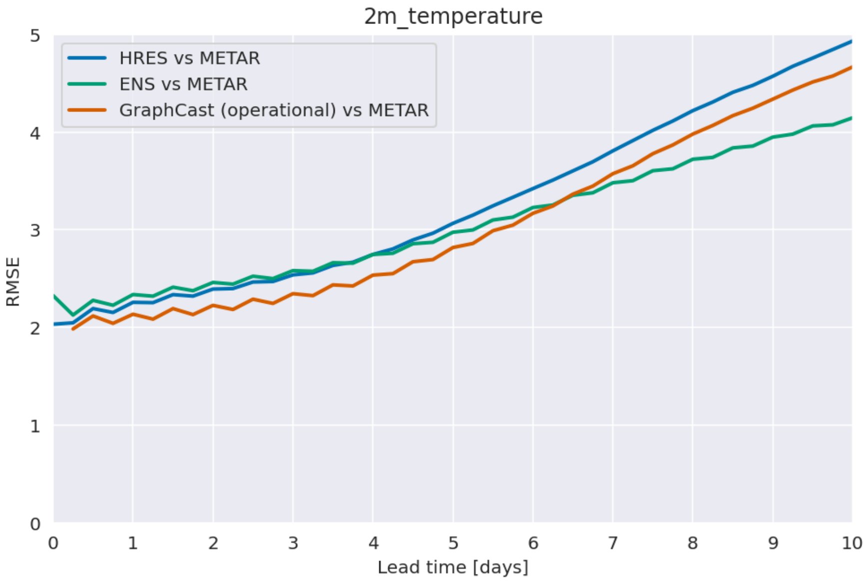
Station evaluation



- Evaluation against ~5000 METAR stations.
- All 00/12 initializations for 2020.
- Gridded fields are bilinearly interpolated to station locations.

Station evaluation



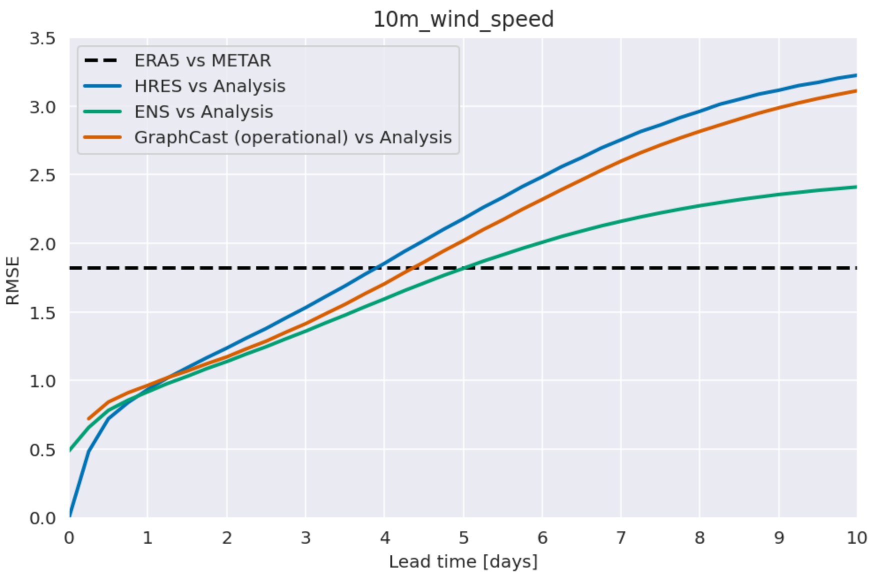


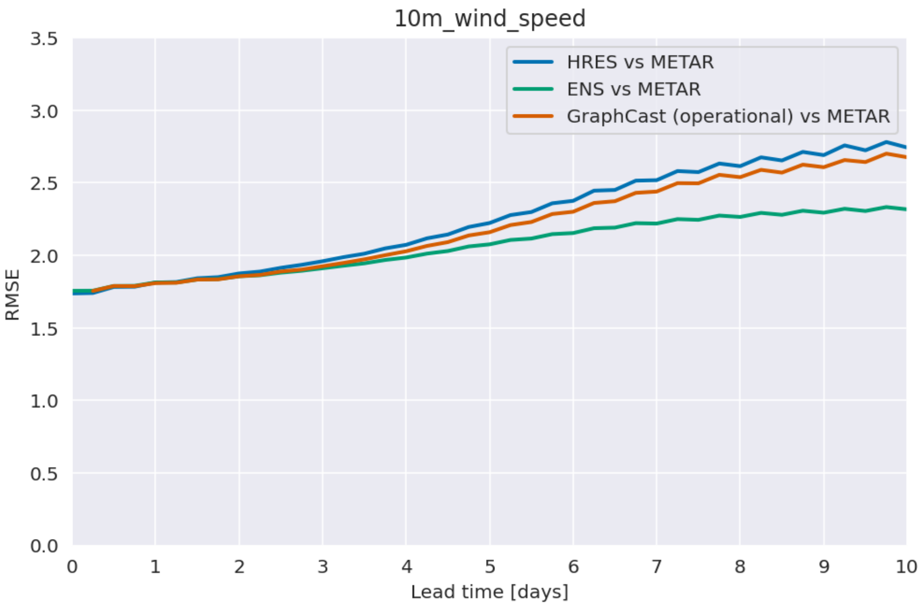
→ ERA5 error ≈ 5 day forecast error.

→ Relative score of models largely unchanged.

Google Research

Station evaluation

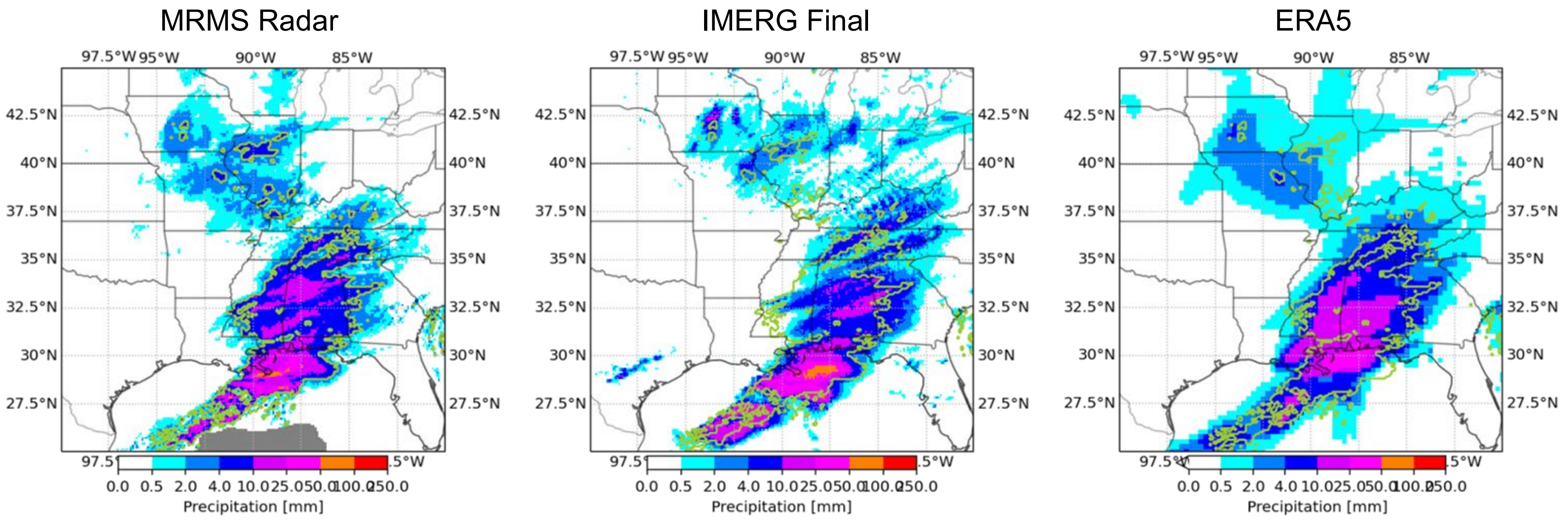




→ Same applies to wind speed.

Google Research

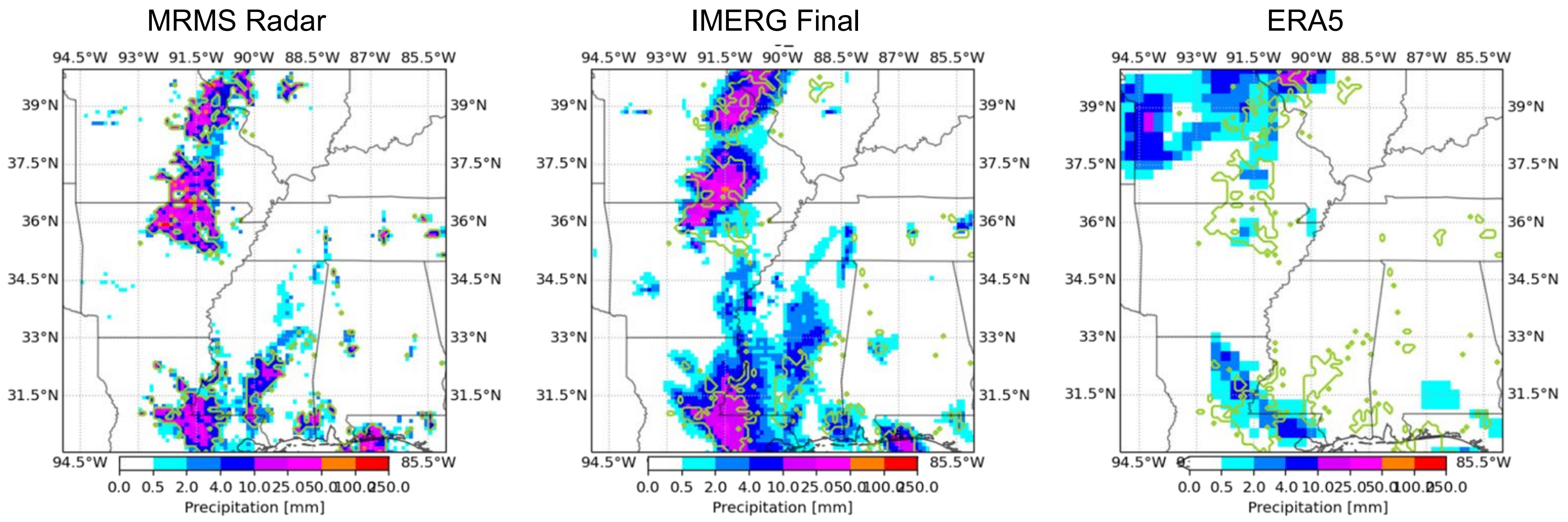
Precipitation



- No single "best" precipitation ground truth.
- Rain gauges are sparse and noisy.
- Radar derived products (e.g. MRMS in the US) are only regional.

Google Research

Precipitation



- IMERG (and other satellite derived products) are global but not perfectly accurate (CSI 4mm/6hr ≈ 0.4 for IMERG vs 0.35 for ERA5).
- No shortcut to evaluating against a range of “ground truths”.

Google Research

An observation benchmark?

Next-gen ML models will likely be trained directly against observations.

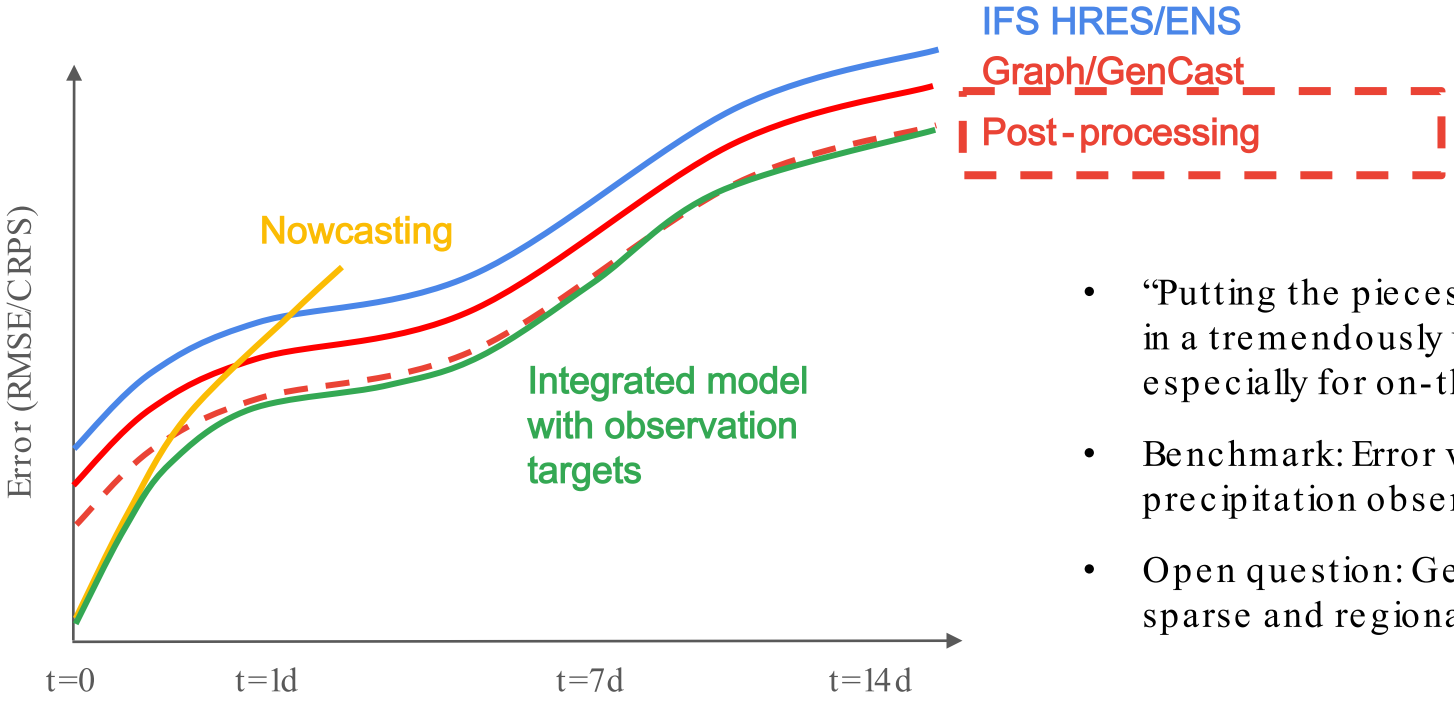
Therefore, WeatherBench 3(?) should be an observation benchmark but …

- Is there agreement on the "best" ground truth? Especially, for precipitation?
- Sparse observations (e.g. weather stations) require generalization. Therefore, we need a hold-out set of stations (and agree what that would be).
- What should the test period be? Many high-quality observations only available for recent years.
- How can we compare against the current "state-of-the-art", i.e. commercial forecast providers?

Please let me know: What should an observation benchmark look like?

Google Research

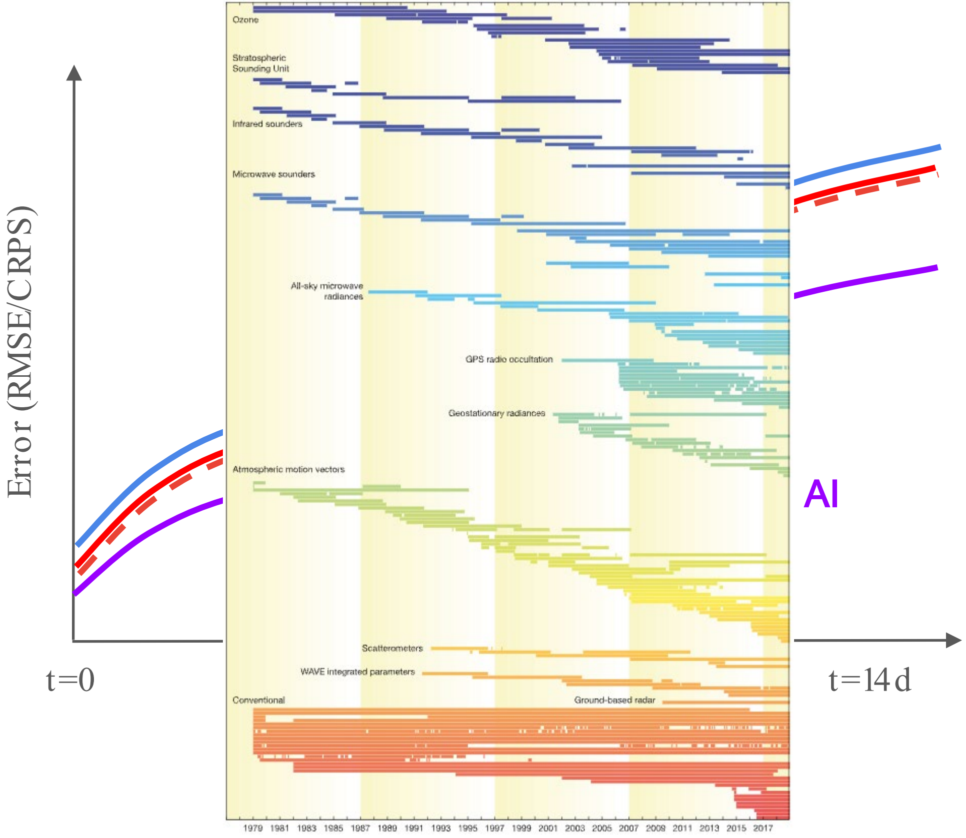
The grand challenge for AI models



- "Putting the pieces together" can result in a tremendously useful model, especially for on-the-ground weather.
- Benchmark: Error vs surface and precipitation observations
- Open question: Generalization of sparse and regional observations

Google Research

The grand challenge for AI models



- Post-processing/nowcasting has little impact on large scale.
- >50% of potential improvements in initial conditions.
- Challenge: Exploit existing observations to improve ICs and large-scale forecasts.
- Benchmark: Z500, TC track and intensity, etc.
- Requires significant investment in data infrastructure.

Google Research