# WeatherBench 2: A benchmark for the next generation of data-driven weather models

August 31, 2023

Posted by Stephan Rasp, Research Scientist, and Carla Bromberg, Program Lead, Google Research

In 1950, weather forecasting started its digital revolution when researchers used the first programmable, general-purpose computer ENIAC to solve mathematical equations describing how weather evolves. In the more than 70 years since, continuous advancements in computing power and improvements to the model formulations have led to steady gains in weather forecast skill: a 7-day forecast today is about as accurate as a 5-day forecast in 2000 and a 3-day forecast in 1980. While improving forecast accuracy at the pace of approximately one day per decade may not seem like a big deal, every day improved is important in far reaching use cases, such as for logistics planning, disaster management, agriculture and energy

Now we are seeing the start of yet another revolution in weather forecasting, this time fueled by advances in machine learning (ML). Rather than hard-coding approximations of the physical equations, the idea is to have algorithms learn how weather evolves from looking at large volumes of past weather data. Early attempts at doing so go back to 2018 but the pace picked up considerably in the last two years when several large ML models demonstrated weather forecasting skill comparable to the best physics-based models. Google's MetNet [1, 2], for instance, demonstrated state-of-the-art capabilities for forecasting regional weather one day ahead. For global prediction, Google DeepMind created GraphCast, a graph neural network to make 10 day predictions at a horizontal resolution of 25 km, competitive with the best physics-based models in many skill metrics.

Apart from potentially providing more accurate forecasts, one key advantage of such ML methods is that, once trained, they can create forecasts in a matter of minutes on inexpensive hardware. In contrast, traditional weather forecasts require large super-computers that run for hours every day. Clearly, ML represents a tremendous opportunity for the weather forecasting community. This has also been recognized by leading weather forecasting centers, such as the European Centre for Medium-Range Weather Forecasts' (ECMWF) machine learning roadmap or the National Oceanic and Atmospheric Administration's (NOAA) artificial intelligence strategy.

To ensure that ML models are trusted and optimized for the right goal, forecast evaluation is crucial. Evaluating weather forecasts isn't straightforward, however, because weather is an incredibly multi-faceted problem. Different end-users are interested in different properties of forecasts, for example, renewable energy producers care about wind speeds and solar radiation, while crisis response teams are concerned about the track of a potential cyclone or an impending heat wave. In other words, there is no single metric to determine what a "good" weather forecast is, and the evaluation has to reflect the multi-faceted nature of weather and its downstream applications. Furthermore, differences in the exact evaluation setup — e.g., which resolution and ground truth data is used — can make it difficult to compare models. Having a way to compare novel and established methods in a fair and reproducible manner is crucial to measure progress in the field.

To this end, we are announcing WeatherBench 2 (WB2), a benchmark for the next generation of data-driven, global weather models. WB2 is an update to the original benchmark published in 2020, which was based on initial, lower-resolution ML models. The goal of WB2 is to accelerate the progress of data-driven weather models by providing a trusted, reproducible framework for evaluating and comparing different methodologies. The official website contains scores from several state-of-the-art models (at the time of writing, these are Keisler (2022), an early graph neural network, Google DeepMind's GraphCast and Huawei's Pangu-Weather, a

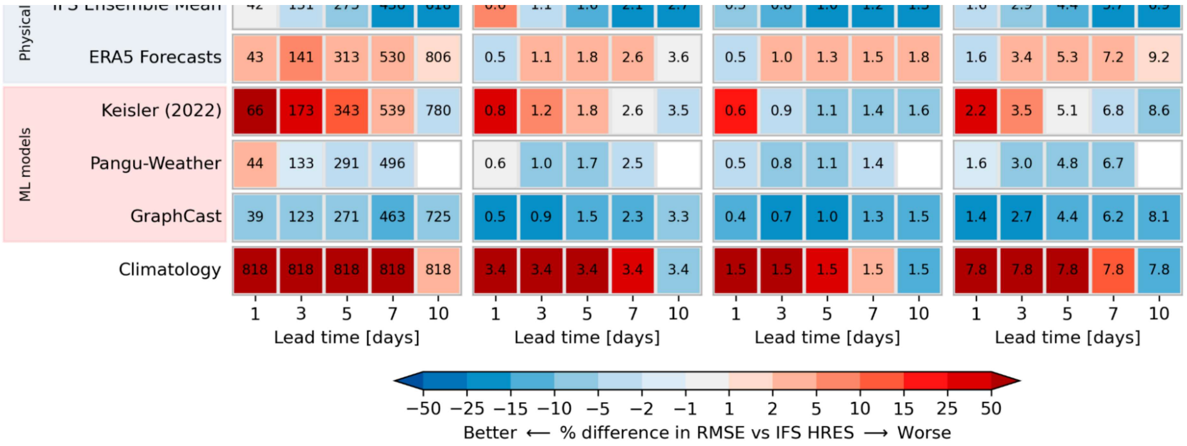
## Making evaluation easier

The key component of WB2 is an open-source evaluation framework that allows users to evaluate their forecasts in the same manner as other baselines. Weather forecast data at high-resolutions can be quite large, making even evaluation a computational challenge. For this reason, we built our evaluation code on Apache Beam, which allows users to split computations into smaller chunks and evaluate them in a distributed fashion, for example using DataFlow on Google Cloud. The code comes with a quick-start guide to help people get up to speed.

Additionally, we provide most of the ground-truth and baseline data on Google Cloud Storage in cloud-optimized Zarr format at different resolutions, for example, a comprehensive copy of the ERA5 dataset used to train most ML models. This is part of a larger Google effort to provide analysis-ready, cloud-optimized weather and climate datasets to the research community and beyond. Since downloading these data from the respective archives and converting them can be time-consuming and compute-intensive, we hope that this should considerably lower the entry barrier for the community.

## Assessing forecast skill

Together with our collaborators from ECMWF, we defined a set of headline scores that best capture the quality of global weather forecasts. As the figure below shows, several of the ML-based forecasts have lower errors than the state-of-the-art physical models on deterministic metrics. This holds for a range of variables and regions, and underlines the competitiveness and promise of ML-based approaches.



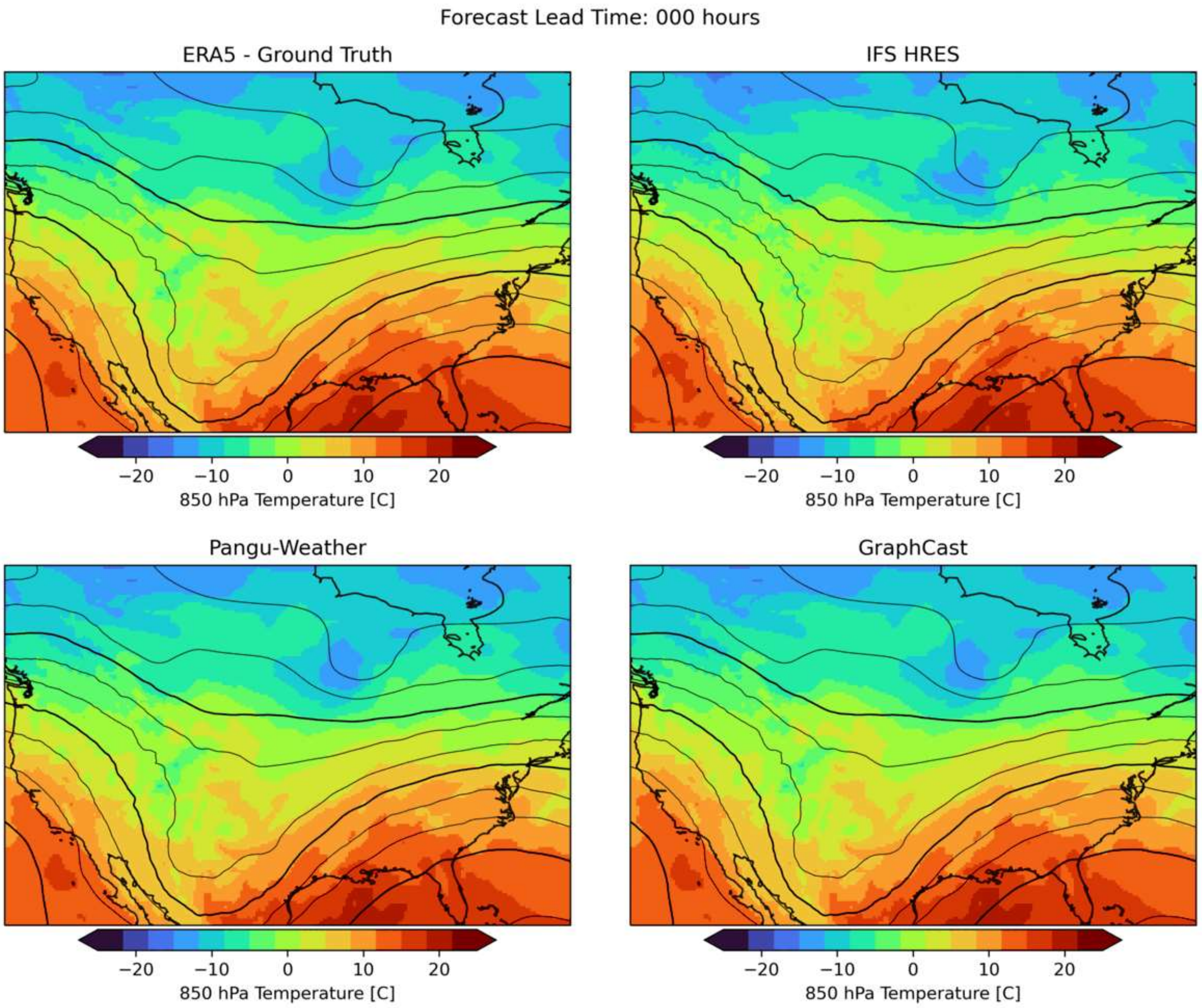
*This scorecard shows the skill of different models compared to ECMWF's Integrated Forecasting System (IFS), one of the best physics-based weather forecasts, for several variables. IFS forecasts are evaluated against IFS analysis. All other models are evaluated against ERA5. The order of ML models reflects publication date.*

# Toward reliable probabilistic forecasts

However, a single forecast often isn't enough. Weather is inherently chaotic because of the butterfly effect. For this reason, operational weather centers now run ~50 slightly perturbed realizations of their model, called an ensemble, to estimate the forecast probability distribution across various scenarios. This is important, for example, if one wants to know the likelihood of extreme weather.

Creating reliable probabilistic forecasts will be one of the next key challenges for global ML models. Regional ML models, such as Google's MetNet already estimate probabilities. To anticipate this next generation of global models, WB2 already provides probabilistic metrics and baselines, among them ECMWF's IFS ensemble, to accelerate research in this direction.

As mentioned above, weather forecasting has many aspects, and while the headline metrics try to capture the most important aspects of forecast skill, they are by no means sufficient. One example is forecast realism. Currently, many ML forecast models tend to "hedge their bets" in the face of the intrinsic uncertainty of the atmosphere. In other words, they tend to predict smoothed out fields that give lower average error but do not represent a realistic, physically consistent state of the atmosphere. An example of this can be seen in the animation below. The two data-driven models, Pangu-Weather and GraphCast (bottom), predict the large-scale evolution of the atmosphere remarkably well. However, they also have less small-scale structure compared to the ground truth or the physical forecasting model IFS



*Forecasts of a front passing through the continental United States initialized on January 3, 2020. Maps show temperature at a pressure level of 850 hPa (roughly equivalent to an altitude of 1.5km) and geopotential at a pressure level of 500 hPa (roughly 5.5 km) in contours. ERA5 is the corresponding ground-truth analysis, IFS HRES is ECMWF's physics-based forecasting model.*

# Conclusion

WeatherBench 2 will continue to evolve alongside ML model development. The official website will be updated with the latest state-of-the-art models. (To submit a model, please follow these instructions). We also invite the community to provide feedback and suggestions for improvements through issues and pull requests on the WB2 GitHub page.

Designing evaluation well and targeting the right metrics is crucial in order to make sure ML weather models benefit society as quickly as possible. WeatherBench 2 as it is now is just the starting point. We plan to extend it in the future to address key issues for the future of ML-based weather forecasting. Specifically, we would like to add station observations and better precipitation datasets. Furthermore, we will

We hope that WeatherBench 2 can aid researchers and end-users as weather forecasting continues to evolve.

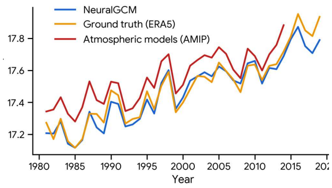
## Acknowledgements

*WeatherBench 2 is the result of collaboration across many different teams at Google and external collaborators at ECMWF. From ECMWF, we would like to thank Matthew Chantry, Zied Ben Bouallegue and Peter Dueben. From Google, we would like to thank the core contributors to the project: Stephan Rasp, Stephan Hoyer, Peter Battaglia, Alex Merose, Ian Langmore, Tyler Russell, Alvaro Sanchez, Antonio Lobato, Laurence Chiu, Rob Carver, Vivian Yang, Shreya Agrawal, Thomas Turnbull, Jason Hickey, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha. We also would like to thank Kunal Shah, Rahul Mahrsee, Aniket Rawat, and Satish Kumar. Thanks to John Anderson for sponsoring WeatherBench 2. Furthermore, we would like to thank Kaifeng Bi from the Pangu-Weather team and Ryan Keisler for their help in adding their models to WeatherBench 2.*

Labels:

[Open Source Models & Datasets]

## Other posts of interest



JULY 22, 2024

Fast, accurate climate modeling with NeuralGCM

JULY 18, 2024

Harnessing hidden genetic information in

JUNE 25, 2024

Efficient data generation for source-grounded information-

*Machine Intelligence ·*
*Open Source Models & Datasets*

*Generative AI ·*
*Health & Bioscience ·*
*Machine Intelligence ·*
*Open Source Models & Datasets*

meeting transcripts

*Machine Intelligence ·*
*Natural Language Processing ·*
*Open Source Models & Datasets*