# WEATHERBENCH: A BENCHMARK DATASET FOR DATA-DRIVEN WEATHER FORECASTING

**Stephan Rasp**
Technical University
of Munich

**Soukayna Mouatadid**
University of Toronto
soukayna@cs.toronto.edu

**Peter D. Dueben**
ECMWF

**Sebastian Scher**
Stockholm University

**Joanathan A. Weyn**
University of Washington

**Nils Thuerey**
Technical University
of Munich

## ABSTRACT

Accurate weather forecasts are a crucial prerequisite for climate change adaptation. Can these be provided by deep learning? First studies show promise, but the lack of a common dataset and evaluation metrics make inter-comparison between the proposed models difficult. In fact, despite the recent research surge in data-driven weather forecasting, there is currently no standard approach for evaluating the proposed models. Here we introduce WeatherBench, a benchmark dataset for data-driven medium-range weather forecasting. We provide data derived from an archive of assimilated earth observations for the last 40 years that has been processed to facilitate the use in machine learning models. We propose a simple and clear evaluation metric which will enable a direct comparison between different proposed methods. Further, we provide baseline scores from simple linear regression techniques, purely physical forecasting models as well as existing deep learning weather forecasting models. All data and code are made publicly available along with tutorials for getting started. We believe WeatherBench will provide a useful and reproducible way of evaluating data-driven weather forecasting models and we hope that it will accelerate research in this direction.

## 1 INTRODUCTION

Extreme weather events are the most immediate and threatening consequence of climate change to human life, property, agricultural production and natural capital (Bouwer, 2019). As the frequency of extreme weather events increases, it becomes clear that accurate weather forecasts will play a crucial role in local and national crisis management planning and climate adaptation decision-making. Currently, these forecasts are generated using physics-based computer simulations, in which the governing equations (or our best approximation thereof) of the atmosphere and ocean are solved on a discrete numerical grid. These models are generally effective, but computationally expensive. They also perform poorly when it comes to the prediction of extreme events. If data-driven models were able to learn a more efficient representation of the underlying dynamical and physical equations, they would enable computationally cheaper forecasts. This could help improve the probability estimation of extreme events through large ensemble (Monte-Carlo) simulations. It is also possible that by using a diverse set of data sources, data driven models can outperform physical models in areas where the latter struggle, for example predicting rainfall over Africa (Vogel et al., 2018).

In the last couple of years, several studies have pioneered data-driven forecasting methods (Dueben & Bauer, 2018; Scher, 2018; Scher & Messori, 2019; Weyn et al., 2019). These studies considered different settings of general circulation models at different resolutions to be ground truth, and used these simulations to train different neural network architectures evaluated using different metrics. The differences of the proposed methods already highlight the importance of a common benchmark case to compare prediction skill. In particular, benchmark datasets can have a huge impact because they make different algorithms inter-comparable and foster constructive competition, particularly in a nascent direction of research.

Here, we introduce WeatherBench, the first benchmark problem for data-driven weather forecasting. Our initial release provides a ready-to-use dataset for download along with specific metrics to compare different approaches. In this paper, we describe the dataset and evaluation metrics. We show how one can use WeatherBench to perform two types of forecasting experiments: a direct forecast and an iterative forecast. We also provide several baseline models and highlight directions for further research.

## 2 Benchmark overview

The main goal of WeatherBench is to evaluate deep learning models for global medium-range (i.e. several days to two weeks) weather forecasting. By specifying a target variable and a target horizon, WeatherBench allows experimentation and evaluation of new model architectures. For this initial release, we use the ERA5 reanalysis dataset (C3S, 2017) for training, evaluating (1979-2016) and testing (2017-2018). Reanalysis datasets provide the best guess of the atmospheric state at any point in time by combining a forecast model with the available observations. We regrid the data from its original resolution, which comprises several Terabytes, to lower resolutions using bilinear interpolation. This is a more realistic use case, since very high resolutions are hard to handle for most deep learning models because of GPU memory constraints and I/O speed. In particular, we chose 5.625° (32×64 grid points), 2.8125° (64×128 grid points) and 1.40525° (128×256 grid points) degrees latitude by longitude. Further, for 3D fields we selected 10 vertical levels, varying from 1 to 1000 hPa.

The available variables were chosen based on meteorological considerations. Geopotential, temperature, humidity and wind are prognostic state variables in most physical numerical weather prediction (NWP) and climate models, and were hence included along with horizontal relative vorticity and potential vorticity. In addition to the 3D fields, we also include 2D fields: 2 meter-temperature, 10 meter wind, total cloud cover, precipitation and top-of-atmosphere incoming solar radiation. Further, there are several potentially important time-invariant fields: the land-sea mask, the soil type and the topography.

Evaluation is done for the years 2017 and 2018, for a 5.625° resolution. This means that predictions at higher resolutions have to be downscaled to the evaluation resolution. We also evaluated some baselines at higher resolutions and found that the scores were almost identical with differences smaller than 1%. Therefore we are reassured that little information is lost by evaluating at a coarser resolution.

We chose 500 hPa geopotential and 850 hPa temperature as primary verification fields. Geopotential at 500 hPa pressure, typically denoted as Z500 or $\Phi$ with units of $\mathrm{m}^2\mathrm{s}^{-2}$, defined as

$$\Phi = \int_0^{z \text{ at 500hPa}} g\, dz' \tag{1}$$

where $z$ describes height in meters and $g = 9.81$ m s$^{-2}$ is the gravitational acceleration, is a commonly used variable that encodes the synoptic-scale pressure distribution. It is the standard verification variable for most medium-range NWP models. We picked 850 hPa temperature as our secondary verification field because temperature is a more impact-related variable. 850 hPa is usually above the planetary boundary layer and therefore not affected by diurnal variations but provides information about broader temperature trends, including cold spells and heat waves.

We chose the root mean squared error (RMSE) as a metric because it is easy to compute and mirrors the loss used for many ML applications. We define the RMSE as the mean latitude-weighted RMSE over all forecasts:

$$\text{RMSE} = \frac{1}{N_{\text{forecasts}}} \sum_i^{N_{\text{forecasts}}} \sqrt{\frac{1}{N_{\text{lat}} N_{\text{lon}}} \sum_j^{N_{\text{lat}}} \sum_k^{N_{\text{lon}}} L(j)(f_{i,j,k} - t_{i,j,k})^2} \tag{2}$$

where $f$ is the model forecast and $t$ is the ERA5 truth. $L(j)$ is the latitude weighting factor for the latitude at the $j$th latitude index:

$$L(j) = \frac{\cos(\text{lat}(j))}{\frac{1}{N_{\text{lat}}} \sum_j^{N_{\text{lat}}} \cos(\text{lat}(j))} \tag{3}$$

Table 1: Baseline scores for 3 and 5 days (3/5 days) forecast time at 5.625° resolution. Best machine learning model and operational baseline are highlighted.

| **Baseline** | **Z500 RMSE [m$^2$ s$^{-2}$]** | **T850 RMSE [K]** |
|---|---|---|
| Persistence | 936 / 1033 | 4.23 / 4.56 |
| Climatology | 1075 | 5.51 |
| Weekly climatology | 816 | 3.50 |
| Linear regression (direct) | 714 / 814 | 3.19 / 3.52 |
| Linear regression (iterative) | 719 / 812 | 3.17 / 3.48 |
| CNN (direct) | **626 / 757** | **2.87 / 3.37** |
| CNN (iterative) | 1114 / 1559 | 4.48 / 9.69 |
| IFS T42 | 489 / 743 | 3.09 / 3.83 |
| IFS T63 | 268 / 463 | 1.85 / 2.52 |
| Operational IFS | **153 / 334** | **1.36 / 2.03** |

Instructions for downloading the data, as well as all code are available in the Github repository[1]. The repository also contains all scripts for downloading and processing of the data. This enables users to download additional variables or regrid the data to a different resolution. We intend for WeatherBench to be open and evolving, as opposed to a fixed snapshot in time. We encourage contributions of new data pre-processing pipelines, alternative baseline architectures such as recurrent layers and/or attention layers, and evaluation metrics. We hope the benchmark will incite a community discussion around the existing data-driven weather forecasting models and inspire new architectures and tasks.

## 3 Baselines

To evaluate the skill of a forecasting model it is important to have baselines to compare to. The two simplest possible forecasts one can consider are a) a persistence forecast in which the fields at initialization time are used as forecasts (”tomorrow’s weather is today’s weather”), and b) two climatological forecasts: a single mean is computed over all times in the training set (1979 – 2016) and a mean is computed for each of the 52 calendar weeks in the training set. In addition, considering that the gold standard of medium-range NWP is the operational IFS (Integrated Forecast System) model of the European Center for Medium-range Weather Forecasting (ECMWF), we present scores for the operational IFS 2016–2018 forecasts, regridded to 5.625. We also ran the IFS model at a lower resolution (T42 $\approx$ 2.8° and T43 $\approx$ 1.9°) to provide an intermediate target for data-driven methods.

Finally, we consider two purely data-driven baselines. We start by fitting a simple linear regression model to directly predict the Z500 and T850 fields at different lead times, 3 days and 5 days. Then, we train a five-layers fully convolutional neural network. Each hidden layer has 64 channels with a convolutional kernel of size 5 and ELU activations (Clevert et al., 2015). The input and output layers have two channels, representing Z500 and T850. The model was trained using the Adam optimizer (Kingma & Ba, 2014) and a mean squared error loss function. We implemented periodic convolutions in the longitude direction but not the latitude direction.

We trained the machine learning baselines a) to directly predict the fields 3 and 5 days ahead and b) to make a 6 hour prediction, which we use to create an iterative forecast. For these iterative forecasts the model takes its previous output as input for the next forecast. To create a 5 day iterative forecast the model trained to predict 6 hour forecasts is called 20 times. The advantage of iterative forecasts is that a single model is able to make predictions for any forecast time rather than having to train several models.

Table 1 shows the RMSE of the different baseline models. The weekly climatology is significantly better than the standard climatological forecast, approximately matching the persistence forecast between 1 and 2 days, since it takes into account the seasonal cycle. For the linear model, the

[1] `https://anonymous.4open.science/r/6bb0b0c0-e929-4fa0-a7b7-a42361bcf7dd/` hosts the anonymized repositery. Upon acceptance, the full repository and data will be released.

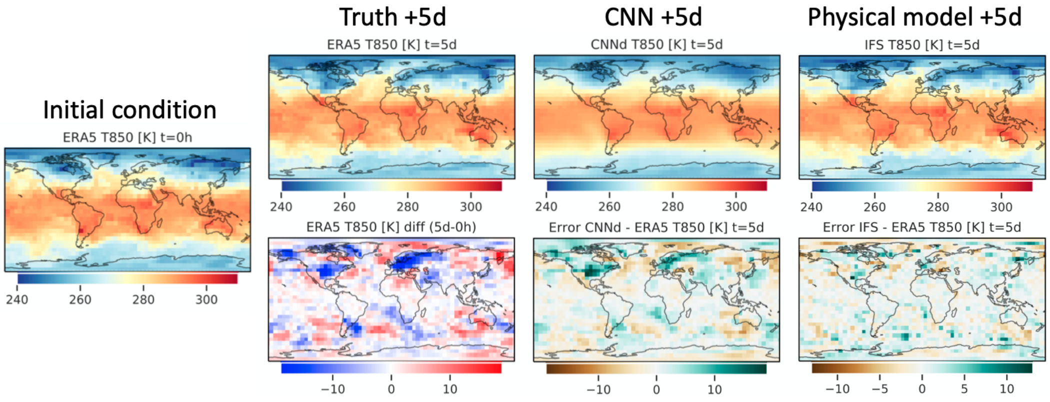

Figure 1: Example fields for 5 day forecasts of 850 hPa temperature initialized at 2017-01-01 00UTC for 850 hPa temperature. (Left) Truth, bottom row shows difference to initial condition. CNN (middle) and operational IFS (right) forecasts. Bottom row shows errors relative to truth.

iterative forecast performs just as well as the direct forecast due to its linear nature. At 5 days, the linear regression forecast is about as good as the weekly climatology. The direct CNN forecasts beat the linear regression forecasts for 3 and 5 days forecast time. However, at 5 days these forecasts are only marginally better than the weekly climatology. The iterative CNN forecast performs well up to around 1.5 days but then the network's errors grow non-linearly. Note, however, that the simple CNN trained for this baseline is to be seen simply as a starting point for more complex models.

Fig. 1 shows an example forecasts of temperature. The ERA5 temporal differences show smaller variability in the tropics compared to the extratropics where propagating fronts can cause rapid changes in temperature. The CNN forecast is noticeably smoother than the truth or the physical forecast. This likely has two reasons: first, the baseline model likely is not powerful enough to make a good prediction including small scale details; second, at 5 days forecast time the atmosphere is slightly chaotic. In face of such uncertainty, a MSE loss will push the network to predict the mean of the distribution rather than a realistic realization. These two issues should be addressed in future work on this dataset.

## 4 Discussion and Conclusion

One important aspect that is not currently addressed by this benchmark is probabilistic forecasting. Because of the chaotic evolution of the atmosphere, it is very important to also have an estimate of the uncertainty of a forecast. Extending this benchmark to probabilistic forecasting simply requires computing a probabilistic score. How to produce probabilistic data-driven forecasts is a very interesting research question in its own right. For the forecast times targeted in the benchmark it still makes sense to look at deterministic forecasts.

A related issue is the question of extreme weather situations. These events are, by definition, rare, which means that they will contribute little to regular verification metrics like the RMSE. However, for society these events are highly important. For this reason, it would make sense to evaluate extreme situations separately. But defining extremes is ambiguous which is why there is no standard metric for evaluating extremes. The goal of this benchmark is to provide a simple, clear problem. Therefore, we decided to omit extremes in this initial release but users are encouraged to choose and contribute their own verification of extremes.

There is a wide variety of promising research directions for data-driven weather forecasting. The most obvious direction is to increase the amount of data used for training and the complexity of the network architecture. WeatherBench provides a, so far, unexploited volume and diversity of data for training. It is up to future research to find out exactly which combination of variables will turn out to be useful. Further, this dataset offers a four times higher horizontal resolution than all previous studies. We intend to keep WeatherBench up-to-date and enrich it with new learning tasks. The hope is that this data will enable researchers to train more complex models than have previously been used and that it will foster collaboration between atmospheric and data scientists in the ways we imagined and beyond.

## References

Laurens M Bouwer. Observed and projected impacts from extreme weather events: implications for loss and damage. In *Loss and damage from climate change*, pp. 63–82. Springer, 2019.

C3S. Copernicus Climate Change Service (C3S) (2017): ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate . Copernicus Climate Change Service Climate Data Store (CDS), 2017. URL `https://cds.climate.copernicus.eu/cdsapp#!/home`.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). 11 2015. URL `http://arxiv.org/abs/1511.07289`.

Peter D Dueben and Peter Bauer. Challenges and design choices for global weather and climate models based on machine learning. *Geosci. Model Dev.*, 2018. doi: 10.5194/gmd-2018-148. URL `https://www.geosci-model-dev-discuss.net/gmd-2018-148/gmd-2018-148.pdf`.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv*, 1412.6980, 12 2014. URL `http://arxiv.org/abs/1412.6980`.

S. Scher. Toward Data-Driven Weather and Climate Forecasting: Approximating a Simple General Circulation Model With Deep Learning. *Geophysical Research Letters*, 45(22):616–12, 11 2018. ISSN 0094-8276. doi: 10.1029/2018GL080704. URL `https://onlinelibrary.wiley.com/doi/abs/10.1029/2018GL080704`.

Sebastian Scher and Gabriele Messori. Weather and climate forecasting with neural networks: using general circulation models (GCMs) with different complexity as a study ground. *Geoscientific Model Development*, 12(7):2797–2809, 7 2019. ISSN 1991-9603. doi: 10.5194/gmd-12-2797-2019. URL `https://www.geosci-model-dev.net/12/2797/2019/`.

Peter Vogel, Peter Knippertz, Andreas H. Fink, Andreas Schlueter, and Tilmann Gneiting. Skill of global raw and postprocessed ensemble predictions of rainfall over Northern Tropical Africa. *Weather and Forecasting*, 33(2):369–388, 2018. ISSN 15200434. doi: 10.1175/WAF-D-17-0127.1.

Jonathan A. Weyn, Dale R. Durran, and Rich Caruana. Can machines learn to predict weather? Using deep learning to predict gridded 500-hPa geopotential height from historical weather data. *Journal of Advances in Modeling Earth Systems*, pp. 2019MS001705, 7 2019. ISSN 1942-2466. doi: 10.1029/2019MS001705. URL `https://onlinelibrary.wiley.com/doi/abs/10.1029/2019MS001705`.