JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2020MS002203

Key Points:
- Benchmarks with strong baselines are a key ingredient for rapid progress on a problem
- Here, we define a benchmark for data-driven global, medium-range weather prediction
- The data are processed for convenient use in machine learning models, and a quickstart guide is provided

Correspondence to:
S. Rasp,
stephan.rasp@tum.de

Citation:
Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., & Thuerey, N. (2020). WeatherBench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, *12*, e2020MS002203. https://doi.org/10.1029/2020MS002203

Received 12 JUN 2020
Accepted 14 AUG 2020
Accepted article online 19 AUG 2020

WeatherBench: A Benchmark Data Set for Data-Driven Weather Forecasting

Stephan Rasp[1], Peter D. Dueben[2], Sebastian Scher[3], Jonathan A. Weyn[4], Soukayna Mouatadid[5], and Nils Thuerey[1]

[1]Department of Informatics, Technical University of Munich, Munich, Germany, [2]European Centre for Medium-range Weather Forecasts, Reading, UK, [3]Department of Meteorology and Bolin Centre for Climate Research, Stockholm University, Stockholm, Sweden, [4]Department of Atmospheric Sciences, University of Washington, Seattle, WA, USA, [5]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada

Abstract Data-driven approaches, most prominently deep learning, have become powerful tools for prediction in many domains. A natural question to ask is whether data-driven methods could also be used to predict global weather patterns days in advance. First studies show promise but the lack of a common data set and evaluation metrics make intercomparison between studies difficult. Here we present a benchmark data set for data-driven medium-range weather forecasting (specifically 3–5 days), a topic of high scientific interest for atmospheric and computer scientists alike. We provide data derived from the ERA5 archive that has been processed to facilitate the use in machine learning models. We propose simple and clear evaluation metrics which will enable a direct comparison between different methods. Further, we provide baseline scores from simple linear regression techniques, deep learning models, as well as purely physical forecasting models. The data set is publicly available at https://github.com/pangeo-data/WeatherBench and the companion code is reproducible with tutorials for getting started. We hope that this data set will accelerate research in data-driven weather forecasting.

Plain Language Summary WeatherBench provides a new benchmark to test data-driven approaches to weather forecasting. Traditional weather models are based on the discretized equations governing the atmosphere. They perform very well for many tasks but are still found lacking for some others. Data-driven approaches, such as deep learning, directly learn from the best available observations and could potentially produce better forecasts. In this paper, we define a benchmark task—predicting pressure and temperature across the globe 3 and 5 days ahead—which will hopefully lead to progress in data-driven weather prediction and foster collaboration across disciplines.

1. Introduction

Deep learning, a branch of machine learning based on multilayered artificial neural networks, has proven to be a powerful tool for a wide range of tasks, most notably image recognition and natural language processing (LeCun et al., 2015). More recently, deep learning has also been used in many fields of natural science. Much of the success of deep learning is based on the ability of neural networks to recognize patterns in high-dimensional spaces. A natural question to ask then is whether deep learning can also be used to predict future weather patterns.

Currently, weather (and climate) predictions are based on purely physical computer models, in which the governing equations, or our best approximation thereof, of the atmosphere and ocean are solved on a discrete numerical grid (Bauer et al., 2015). Overall, this approach has been very successful. However, today's numerical weather prediction (NWP) models still have shortcoming for many important applications, for example, forecasting mesoscale convective systems over Africa (Vogel et al., 2018). Furthermore, huge amounts of computing power are required, especially for creating probabilistic forecasts which are usually limited to 50 ensemble members or less. For these reasons and the growing popularity of machine learning (ML) there has been a growing interest to improve and speed up NWP with data-driven approaches.

ML can be applied to weather prediction in many different ways. Two long-standing applications of ML are postprocessing—the correction of statistical biases in the output of physical models—and statistical

forecasting—the prediction of variables not directly output by the physical model. Traditionally, this has been done using simple linear techniques but more recently modern machine learning approaches like random forests or neural networks have been explored (Gagne et al., 2014; Lagerquist et al., 2017; McGovern et al., 2017; Rasp & Lerch, 2018; Taillardat et al., 2016). Typically, these approaches target very specific variables or locations whereas the general evolution of the atmosphere is still predicted by a physical model. Another application that has recently been explored using ML is nowcasting, which describes the short range (up to 6 hr) prediction of precipitation by directly extrapolating radar observation without a physical model involved (Agrawal et al., 2019; Grönquist et al., 2020; Shi et al., 2015, 2017).

Yet another direction for ML research is hybrid modeling, in which a physical model is combined with data-driven components, for example, replacing heuristic cloud or radiation parameterizations (Brenowitz & Bretherton, 2018; Chevallier et al., 1998; Krasnopolsky et al., 2005; Rasp et al., 2018; Yuval & O'Gorman, 2020). The key idea behind these approaches is to only replace uncertain (e.g., clouds) or computationally expensive (e.g., line-by-line radiation) model components with machine learning emulators and leave other model components (e.g., large-scale dynamics) untouched. However, such hybrid models also have drawbacks. First, the interaction between physical and machine learning components are poorly understood and can lead to unexpected instabilities and biases (Brenowitz & Bretherton, 2019). Second, they are difficult to implement from a technical perspective because one has to interface the machine learning components with complex climate model code, typically written in Fortran even though recent developments aim to make this step easier (Ott et al., 2020).

Here we focus on purely data-driven prediction of the global atmospheric flow in the medium range. Specifically, we select lead times of 3 and 5 days, for which the atmosphere is still reasonably deterministic but also exhibits complex nonlinear behavior, such as baroclinic instabilities and tropical cyclogenesis. This forecast range is important from a societal point of view because it delivers crucial information for disaster preparation, for example, for flooding, cold and hot spells, or damaging winds (Lazo et al., 2009). Creating a good medium-range forecast requires understanding complex atmospheric dynamics and the interplay between several variables across a range of scales. This sets this challenge apart from postprocessing and statistical forecasting, in which the large-scale dynamics are predicted by a physical model, and nowcasting, in which the considered evolution is univariate and short term. In other words, this benchmark closely emulates the task performed by physical NWP models.

There are several motivations for considering a purely data-driven approach. As mentioned above current NWP is computationally expensive and, nevertheless, has low skill for certain applications. If data-driven models were able to learn a more efficient representation of the underlying dynamical and physical equations, they might enable computationally cheaper forecasts. This can be useful for many applications, for example, creating very large ensembles to better estimate the probability of extreme events. It is also possible that by learning from a diverse set of data sources, data-driven models can outperform physical models in areas where the latter struggle. While in this benchmark challenge the focus is on upper-level fields of pressure and temperature—for which physical models perform very well—the hope is that the insights gained from this task can be leveraged for more impactful application. Further, recent research into interpretable machine learning might provide scientists with new analysis tools (McGovern et al., 2019; Toms et al., 2019). Finally, there is the basic scientific question to what extent purely data-driven models can learn the underlying dynamics of the atmosphere.

Note also that while this benchmark is framed as a data-driven prediction challenge, the proposed framework can also be applied to postprocessing using the same metrics.

In machine learning research, the data-driven prediction of future states is an active area of research with applications from language translation (Sutskever et al., 2014), over audio signals (Oord et al., 2016), to numerical simulations (Morton et al., 2018). In this context, weather forecasts are a particularly challenging task. The behavior is highly complex and nonlinear and also exhibits some recurring patterns, albeit only on local scales, at least for medium-range predictions (Hamill & Whitaker, 2006). As such the the proposed benchmark poses interesting challenges for deep learning algorithms, for example, to evaluate different architectures (He et al., 2015; Huang et al., 2016; Ronneberger et al., 2015), regularization methods (Krogh & Hertz, 1992; Srivastava et al., 2014; D. Xie et al., 2017), or optimizers (Graves, 2013; Kingma & Ba, 2014).

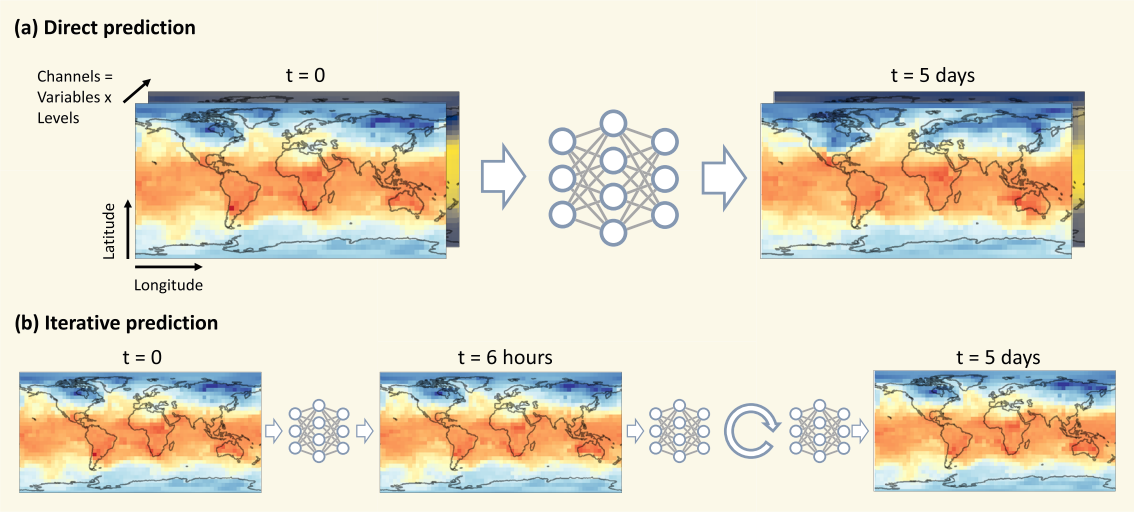

Figure 1. Schematic of data-driven weather forecasting. (a) Example of direct weather prediction for 5 days lead time. The input to the neural network are fields on a latitude-longitude grid. The fields can be several levels of the same variable and/or different variables. The goal is to predict the same fields some time ahead. (b) Iterative forecasts are created from data-driven models trained on a shorter lead time, for example, 6 hr, which are then iteratively called up to the required forecast lead time.

In the last couple of years, several studies (summarized in section 2) have pioneered data-driven, global, medium-range weather prediction. All of them show that there is some potential in this approach but also highlight the need for further research. In particular, we currently lack a common benchmark challenge to accelerate progress. Benchmark data sets can have a huge impact because they make different algorithms quantitatively inter-comparable and foster constructive competition, particularly in a nascent direction of research. Famous examples are the computer vision data sets MNIST (LeCun et al., 1998) and ImageNet (Russakovsky et al., 2015). Further, well-curated benchmark data sets make it easier for people from different fields to work on a problem (Ebert-Uphoff et al., 2017). Some existing examples of benchmark challenges in meteorology are Mudigonda et al. (2020), Rasp et al. (2019), and some Kaggle competitions (https://www.kaggle.com/.

Here, we propose a benchmark problem for data-driven weather forecasting (Figure 1). We provide a ready-to-use data set for download along with specific metrics to compare different approaches. In this paper, we start by reviewing the previous work done on this topic (section 2), describe the data set (section 3) and the evaluation metrics (section 4), and provide several baseline models (section 5). Finally, we will highlight several promising directions for further research (section 6) and conclude with a big picture view (section 7).

2. Overview of Previous Work

In this section, we briefly describe the three existing approaches on predicting the large-scale atmospheric state in the medium range with a focus on the data, methods, and evaluation.

2.1. Dueben and Bauer (2018)

In this study, the authors trained a neural network to predict 500 hPa geopotential (Z500; see section 4 for details on commonly used fields), and in some experiments 2 m temperature, 1 hr ahead. The training data were taken from the ERA5 archive for the time period from 2010 to 2017 and regridded to a 6 degree latitude-longitude grid. Two neural network variants were used, a fully connected neural network and a spatially localized network, similar to a convolutional neural network (CNN). After training they then created iterative forecasts up to 120 hr lead time for 10 month validation period. They compared their data-driven forecasts to an operational NWP model and the same model run at a spatial resolution comparable to the data-driven method. One interesting detail is that their networks predict the difference from one time step to the next, instead of the absolute field. To create these iterative forecasts, they use a third-order Adams-Bashford explicit time-stepping scheme. The CNN predicting only geopotential performed best but was unable to beat the low-resolution physical baseline.

2.2. Scher (2018) and Scher and Messori (2019)

These two studies addressed the issue of data-driven weather forecasting in a simplified reality setting. Long runs of simplified general circulation models (GCMs) were used as "reality." Neural networks were trained to predict the model fields several days ahead. The neural network architecture are CNNs with an encoder-decoder setup. They take as input the instantaneous 3-D model fields at one time step, and output the same model fields at some time later. In Scher (2018), a separate network was trained for each lead time up to 14 days. Scher and Messori (2019b) trained only on 1-day forecasts, and constructed longer forecasts iteratively. Interestingly, networks trained to directly predict a certain forecast time, for example, 5 days, outperformed iterative networks. The forecasts were evaluated using the root-mean-square error and the anomaly correlation coefficient of Z500 and 800 hPa temperature. Scher (2018) used a highly simplified GCM without hydrological cycle and achieved very high predictive skill. Additionally, they were able to create stable "climate" runs (long series of consecutive forecasts) with the network. Scher and Messori (2019b) used several more realistic and complex GCMs. The data-driven model achieved relatively good short-term forecast skill, but was unable to generate stable and realistic "climate" runs. In terms of neural network architectures they showed that architectures tuned on simplified GCMs also work on more complex GCMs and that the same architecture also has some prediction skill on single-level reanalysis data.

2.3. Weyn et al. (2019)

In this study, reanalysis-derived Z500 and 700–300 hPa thickness at 6-hourly time steps are predicted with deep CNNs. The data are from the Climate Forecast System (CFS) Reanalysis from 1979–2010 with 2.5° horizontal resolution and cropped to the Northern Hemisphere. The authors used similar encoder-decoder convolutional networks as those used by Scher (2018) and Scher and Messori (2019b) but also experimented with adding a convolution long short-term memory (Hochreiter & Schmidhuber, 1997) hidden layer. As in Scher and Messori (2019b), forecasts are generated iteratively by feeding the model's outputs back in as inputs. The authors found that using two input time steps, 6 hr apart, and predicting two output time steps, performed better than using a single step. Their best CNN forecast outperforms a climatology benchmark at up to 120 hr lead time, and appears to correctly asymptote toward persistence forecasts at longer lead times up to 14 days.

These three approaches outline promising first steps toward data-driven forecasting. The differences of the proposed methods already highlight the importance of a common benchmark case to compare prediction skill.

3. Data Set

For the proposed benchmark, we use the ERA5 reanalysis data set (Hersbach et al., 2020) for training and testing. Reanalysis data sets provide the best guess of the atmospheric state at any point in time by combining a forecast model with the available observations. The raw data are available hourly for 40 years from 1979 to 2018 on a 0.25° latitude-longitude grid (721 × 1,440 grid points) with 37 vertical levels.

Since this raw data set is very large (a single vertical level for the entire time period amounts to almost 700 GB of data), we regrid the data to lower resolutions. This is also a more realistic use case, since very high resolutions are still hard to handle for 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) resolution for our data. The regridding was done with the xesmf Python package (Zhuang, 2019) using a bilinear interpolation. Powers of two for the grid are used since this is common for many deep learning architectures where image sizes are halved in the algorithm. Further, for 3-D fields we selected 13 vertical levels: 50, 100, 150, 200, 250, 300, 400, 500, 600, 700, 850, 925, and 1,000 hPa. Note that it is common to use pressure in hecto-Pascals as a vertical coordinate instead of physical height. This has practical advantages such as reducing the number of required state variables and simplifying mass conservation (Holton, 2004). The pressure at sea level is ~1,000 hPa and decreases roughly exponentially with height. The 850 hPa is at around 1.5 km height. The 500 hPa is at around 5.5 km height. If the surface pressure is smaller than a given pressure level, for example, at high altitudes, the pressure-level values are interpolated. The selected pressure levels contain the seven pressure levels that are commonly used for 3-D output by the climate models in the Coupled Model Intercomparison Project Phase 6 (CMIP6) (Eyring et al., 2016) which could be useful for pretraining. One regridded historical climate run is also available from the data repository with a template workflow for downloading further CMIP data on the Github repository.

Table 1
List of Variables Contained in the Benchmark Data Set

Long name	Short name	Description	Unit	Levels
Geopotential	z	Proportional to the height of a pressure level	($m^2 s^{-2}$)	13
Temperature	t	Temperature	(K)	13
Specific_humidity	q	Mixing ratio of water vapor	($kg kg^{-1}$)	13
Relative_humidity	r	Humidity relative to saturation	(%)	13
u_component_of_wind	u	Wind in x/longitude-direction	($m s^{-1}$)	13
v_component_of_wind	v	Wind in y/latitude direction	($m s^{-1}$)	13
Vorticity	vo	Relative horizontal vorticity	($1 s^{-1}$)	13
Potential_vorticity	pv	Potential vorticity	($K m^2 kg^{-1} s^{-1}$)	13
2m_temperature	t2 m	Temperature at 2 m height above surface	(K)	1
10m_u_component_of_wind	u10	Wind in x/longitude-direction at 10 m height	($m s^{-1}$)	1
10m_v_component_of_wind	v10	Wind in y/latitude-direction at 10 m height	($m s^{-1}$)	1
total_cloud_cover	tcc	Fractional cloud cover	(0–1)	1
total_precipitation	tp	Hourly precipitation	(m)	1
toa_incident_solar_radiation	tisr	Accumulated hourly incident solar radiation	($J m^{-2}$)	1
Constants		*File containing time-invariant fields*		
land_binary_mask	lsm	Land-sea binary mask	(0/1)	1
soil_type	slt	Soil-type categories	see text	1
orography	orography	Height of surface	(m)	1
latitude	lat2d	2-D field with latitude at every grid point	(°)	1
longitude	lon2d	2-D field with longitude at every grid point	(°)	1

Note. All fields have dimensions *lat* × *lon* × *level*. Latitude and longitude dimensions are dependent on the data resolution. The number of vertical levels is given in the table.

The processed data (see Table 1) are available at https://mediatum.ub.tum.de/1524895 (Rasp et al., 2020). The data are split into yearly NetCDF files for each variable and resolution, packed in a zip file. The entire data set at 5.625° resolution has a size of 191 GB. Individual variables amount to around 25 GB three-dimensional and 2 GB for two-dimensional fields. File sizes for 2.8125° and 1.40525° resolutions are a factor 4 and 16 times larger. Data processing was organized using Snakemake (Koster & Rahmann, 2012). For further instructions on data downloading, visit the Github page (https://github.com/pangeo-data/WeatherBench). The available variables were chosen based on meteorological consideration. For an introduction to atmospheric variable, we suggest Wallace and Hobbs (2006). Geopotential, temperature, humidity, and wind are prognostic state variables in most physical NWP and climate models. Geopotential at a certain pressure level p, typically denoted as Φ with units of $m^2 s^{-2}$, is defined as

$$\Phi = \int_0^{z(p)} g \, dz' \tag{1}$$

where z describes height in meters and $g = 9.81 m s^{-2}$ is the gravitational acceleration. Note that *geopotential height*, also commonly used, is defined as Φ/g with units of meters. Horizontal relative vorticity, defined as $\partial v/\partial x - \partial u/\partial y$, describes the rotation of air at a given point in space. Potential vorticity (Holton, 2004; Hoskins et al., 1985) is a commonly used quantity in synoptic meteorology which combines the rotation (vorticity) and vertical temperature gradient of the atmosphere. It is defined as $PV = \rho^{-1}\zeta_a \cdot \nabla\theta$, where ρ is the density, ζ_a is the absolute vorticity (relative plus the Earth's rotation) and θ is the potential temperature. PV is fundamental dynamical unit that is conserved for adiabitic flow (i.e., in the absence of external heating). In addition to the three-dimensional fields, we also include several two-dimensional fields: 2 m temperature is often used as an impact variable because of its relevance for human activities and is directly affected by the diurnal solar cycle; 10 m wind is also an important impact-related forecast variable, for example, for wind energy; similarly, total cloud cover is an essential variable for solar energy forecasting. We also included precipitation but urge caution since precipitation in reanalysis data sets often shows large deviation from observations (Betts et al., 2019; Xu et al., 2019). Finally, we added the top-of-atmosphere incoming solar radiation as it could be a useful input variable to encode the diurnal cycle. Further, there are several

potentially important time-invariant fields, which are contained in the constants file. The first three variables enclose information about the surface: the land-sea mask is a binary field with ones for land points; the soil type consists of seven different soil categories (Coarse = 1, Medium = 2, Medium fine = 3, Fine = 4, Very fine = 5, Organic = 6, Tropical organic = 7, see https://apps.ecmwf.int/codes/grib/param-db?id=43); orography is simply the surface height. In addition, we included two-dimensional fields with the latitude and longitude values at each point. Particularly, the latitude values could become important for the network to learn latitude-specific information such as the grid structure or the Coriolis effect (see section 6). The Github code repository includes all scripts for downloading and processing of the data. This enables users to download additional variables or regrid the data to a different resolution.

4. Evaluation

Evaluation is done for the years 2017 and 2018. To make sure no overlap exists between the training and test data set, the first test date is 1 January 2017 00 UTC plus forecast time (i.e., for a 3-day forecast the first test date would be 4 January 2017 00 UTC) while the last training target is 31 December 2016 23 UTC. Further, the evaluation presented here is done on 5.625° resolution (The evaluation of all baselines in this paper are done in this Jupyter notebook: https://github.com/pangeo-data/WeatherBench/blob/master/notebooks/4-evaluation.ipynb). 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.

A note on validation and testing: In machine learning it is good practice to split the data into three parts: the training, validation, and test sets. The training data set is used to actually fit the model. The validation data set is used during experimentation to check the model performance on data not seen during training. However, there is the danger that through continued tuning of hyperparameters one unwillingly overfits to the validation data set. Therefore it is advisable to keep a third testing data set for final evaluations of model performance. For this benchmark this final evaluation is done for the years 2017 and 2018. Therefore, we strongly encourage users of this data set to pick a period from 1979 to 2016 for validation of their models for hyperparameter tuning. Because meteorological fields are highly correlated in time, it is advisable to choose a longer contiguous chunk of data for validation instead of a completely random split. Here we chose the year 2016 for validation.

We chose 500 hPa geopotential and 850 hPa temperature as primary verification fields. Geopotential at 500 hPa pressure, often abbreviated as Z500, 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, except at high altitudes and regions with deep boundary layers, and therefore not affected by diurnal variations but provides information about broader temperature trends, including cold spells and heat waves. In addition we also provide some baseline scores for total 6-hourly accumulated precipitation (TP; but noting the dubious quality mentioned above) and 2-m temperature (T2M). However, they will not be further discussed here.

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

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

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

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

In addition, we also evaluate the baselines using the latitude-weighted anomaly correlation coefficient (ACC; see Section 7.6.4 of Wilks, 2006) and the mean absolute error (MAE). For smooth fields like Z500 and T850

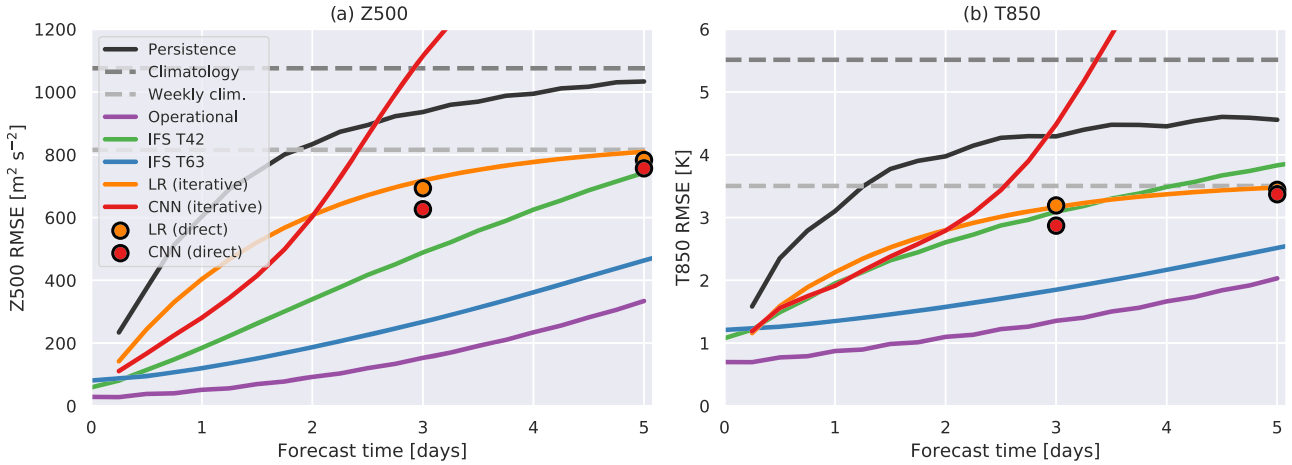

Figure 2. RMSE of (a) 500 hPa geopotential and (b) 850 hPa temperature for different baselines at 5.625° resolution. Solid lines for linear regression and CNN indicate iterative forecasts, while dots represent direct forecasts for 3 and 5 days lead time.

the qualitative differences between the metrics are small. For intermittent fields like precipitation the choice of metric matters a lot more.

5. Baselines

To evaluate the skill of a forecasting model it is important to have baselines to compare to. In this section, we compute scores for several baselines. The results are summarized in Figure 2 and Table 2. The tables and figures for ACC and MAE can be found in Appendix A.

5.1. Persistence and Climatology

The two simplest possible forecasts are (a) a persistence forecast in which the fields at initialization time are used as forecasts ("today's weather is tomorrow's forecast"), and (b) a climatological forecast. For the climatological forecast, two different climatologies were computed from the training data set (1979–2016): first, a single mean over all times in the training data set and, second, a mean computed for each of the 52 calendar weeks. The weekly climatology is significantly better, approximately matching the persistence forecast between 1 and 2 days, since it takes into account the seasonal cycle. This means that to be useful, a forecast system needs to beat the weekly climatology and the persistence forecast.

Table 2
Baseline RMSE for 3 and 5 Days Forecast Time at 5.625° Resolution

	RMSE (3 days/5 days)			
Baseline	Z500 (m^2 s^{-2})	T850 (K)	T2 M (K)	TP (mm)
Persistence	936/1,033	4.23/4.56	3.00/3.27	3.23/3.24
Climatology	1,075	5.51	6.07	2.36
Weekly climatology	816	3.50	3.19	**2.32**
Linear regression (direct)	693/783	3.19/3.44	2.39/2.60	2.37/2.37
Linear regression (iterative)	718/810	3.17/3.48		
CNN (direct)	**626/757**	**2.87/3.37**		
CNN (iterative)	1,114/1,559	4.48/9.69		
IFS T42	489/743	3.09/3.83	3.21/3.69	
IFS T63	268/463	1.85/2.52	2.04/2.44	
Operational IFS	**154/334**	**1.36/2.03**	**1.35/1.77**	2.36/2.59

Note. Best machine learning baseline and physical model are highlighted. TP is 6 hourly accumulated precipitation.

5.2. Operational NWP Model

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 downloaded the forecasts for 2017 and 2018 from the THORPEX Interactive Grand Global Ensemble (TIGGE) (Bougeault et al., 2010) archive, which contains the operational forecasts, initialized at 00 and 12 UTC regridded to a 0.5° by 0.5° grid, which we further regridded to 5.625°. Note that the forecast error starts above zero because the operational IFS is initialized from a different analysis. Operational forecasting is computationally very expensive. The current IFS deterministic forecast is computed on a cluster with 11,664 cores. One 10-day forecast at 10-km resolution takes around 1 hr of real time to compute. It is worth mentioning that the operational IFS model is the product of decades of tightly organized efforts by many hundreds of scientists across disciplines.

5.3. Physical NWP Model Run at Coarser Resolution

To provide physical baselines more in line with the computational resources of a data-driven model, we ran the IFS model at two coarser horizontal resolutions, T42 (~2.8° or 310-km resolution at the equator (NCAR, 2020)) with 62 vertical levels and T63 (~1.9° or 210 km) with 137 vertical levels. The T42 run was initialized from ERA5 whereas the T63 run was initialized from the operational analysis. The gap in skill at $t = 0$ is caused by the conversion to spherical coordinates at coarse resolutions. For Z500 the skill for these two runs lies in-between the operational IFS and the machine learning baselines. For T850, the T42 run is significantly worse. The likely reason for this is that temperature close to the ground is much more affected by the resolution and representation of topography within the model. Further, the model was not specifically tuned for these resolutions. Computationally, a single forecast takes 270 s for the T42 model and 503 s for the T64 model on a single XC40 node with 36 cores. Since the computational costs and resolutions of these runs are much closer to those of a data-driven method, beating those baselines should be a realistic target. note, however, that the model was not tuned to run at such coarse resolutions.

5.4. Linear Regression

As a first purely data-driven baseline we fit a simple linear regression model. For the direct predictions a separate model was trained for each of the four variables, with only this variable in the input/output. For this purpose the 2-D fields were flattened from $32 \times 64 \rightarrow 2{,}048$. This was done for 3- and 5-day forecast time. In addition an iterative model for Z500 and T850 was trained (see Figure 1). Here we use a single linear regression to predict 6 hr ahead where the two fields are concatenated ($2 \times 32 \times 64 \rightarrow 4{,}096$). 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. For iterative forecasts the model takes its previous output as input for the next step. To create a 5-day iterative forecast the model trained to predict 6-hr forecasts is called 20 times. For this model, the 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.

5.5. Simple Convolutional Neural Network

As our deep learning baseline we chose a simple fully convolutional neural network. CNNs are the natural choice for spatial data since they exploit translational invariances in images/fields. Here we train a CNN with five layers. 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 square error loss function. The total number of trainable parameters is 313,858. We implemented periodic convolutions in the longitude direction but not the latitude direction. The implementation can be found in the Github repository. The direct CNN forecasts beat the linear regression forecasts for 3- and 5-day forecast time. However, at 5 days these forecasts are only marginally better than the weekly climatology (see Table 2). This baseline CNN architecture, however, is only a very basic starting point for future research. Adding more variables or making the CNN larger likely leads to much better scores. The iterative CNN forecast, which equivalently to the linear regression iterative forecast was created by chaining together 6-hourly predictions, performs well up to around 1.5 days, but then the network's errors grow quickly and diverge. This confirms the findings of Scher and Messori (2019a) whose experiments showed that training with longer lead time yields better results than chaining together short-term forecasts. However, the poor skill of the iterative forecast could easily be a result of using an overly simplistic network architecture. The iterative forecasts of Weyn et al. (2019), who employ a more complex network structure, show stable long-term performance up to 2 weeks with realistic statistics.

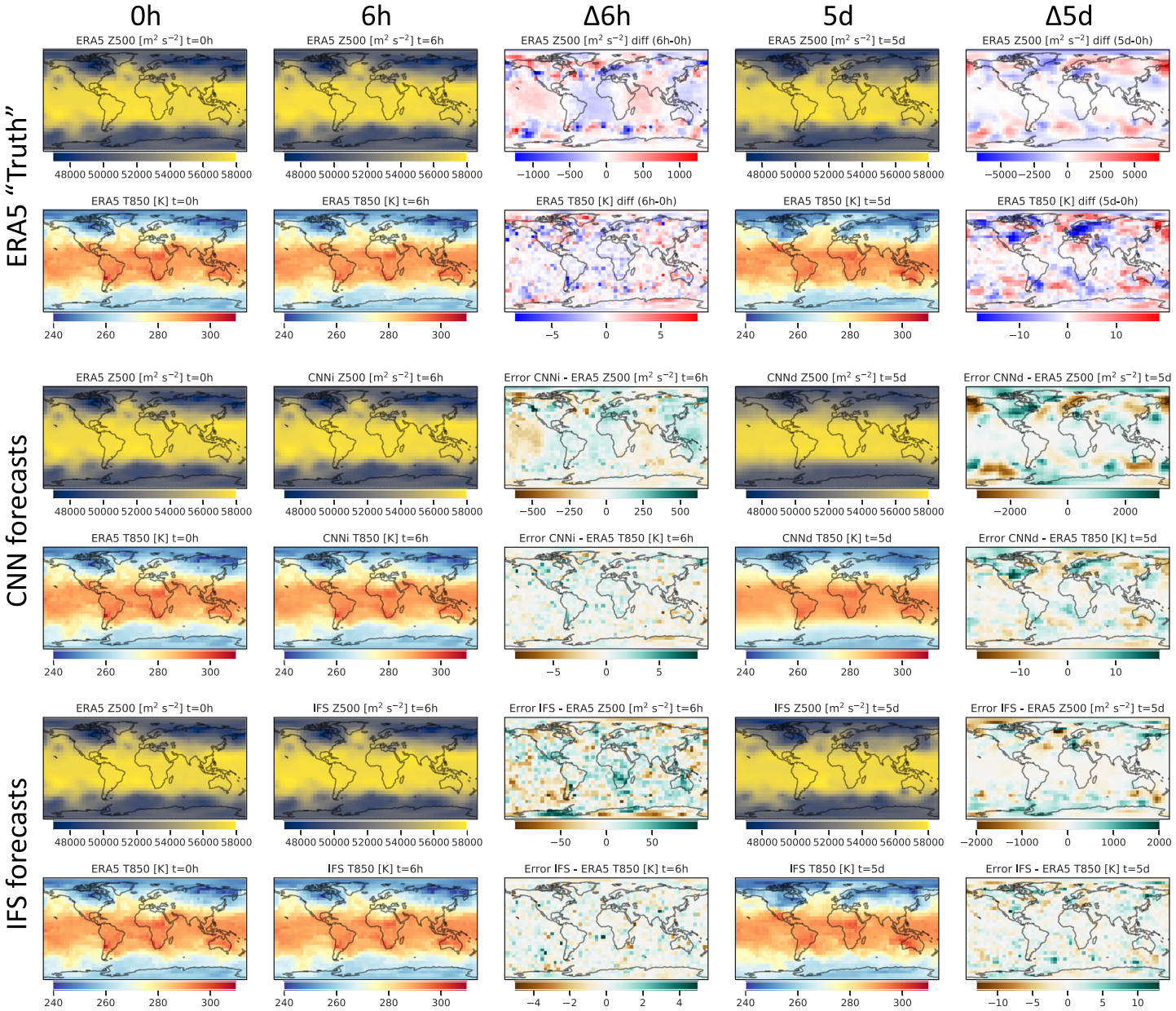

Figure 3. Example fields for 1 January 2017 00 UTC initialization time. The top two rows show the ERA5 "truth" fields for geopotential (Z500) and temperature (T850) at initialization time (t = 0h) and for 6h and 5d forecast time. In addition, the difference between the forecast times and the initialization time is shown. The third and fourth rows show the forecasts from the CNN model. Rows five and six show the IFS operational model. For the CNN forecasts the first column is identical to the ERA5 truth. We selected the 6h iterative CNN model for the 6h forecast but the 5d direct CNN model for the 5 day forecast. For the IFS the initial states (t = 0h) differ slightly albeit not visibly. In addition to the forecast fields the error relative to the ERA5 "truth" is shown in the third and fifth columns. Please note that the colorbars for the difference fields change.

5.6. Example Forecasts

To further illustrate the prediction task, Figure 3 shows example geopotential and temperature fields. The ERA5 temporal differences show several interesting features. First, the geopotential fields and differences are much smoother compared to the temperature fields. The differences in both fields are also much smaller in the tropics compared to the extratropics where propagating fronts can cause rapid temperature changes. An interesting feature is detectable in the 6 hr Z500 difference field in the tropics. These alternating patterns are signatures of atmospheric tides.

The CNN forecasts for 6h lead time are not able to capture these wave-like patterns which hints at a failure to capture the basic physics of the atmosphere. For 5 days forecast time the CNN model predicts unrealistically smooth fields. This is likely a result of two factors: first, the two input fields used in this baseline CNN contain insufficient information to create a skillful 5 day forecast; and second, at 5 days the atmosphere already shows some chaotic behavior which causes a model trained with a simple RMSE loss to predict smooth fields (see section 6). The IFS operational forecast has much smaller errors than the CNN forecast. It is able to capture the propagation of tropical waves. Its main errors appear at 5 days in the midlatitudes where extratropical cycles are in slightly wrong positions.

6. Discussion

6.1. Weather-Specific Challenges

From a ML perspective, state-to-state weather prediction is similar to image-to-image translation. For this sort of problem many deep learning techniques have been developed in recent years (Kaji & Kida, 2019). However, forecasting weather differs in some important ways from typical image-to-image applications and raises several open questions.

First, the atmosphere is three-dimensional. So far, this aspect has not been taken into account. In the networks of Scher and Messori (2019a), for example, the different levels have been treated as separate channels of the CNN. However, simply using a three-dimensional CNN might not work either because atmospheric dynamics and grid spacings change in the vertical, thereby violating the assumption of translation invariance which underlies the effectiveness of CNNs. This directly leads to the next challenge: On a regular latitude-longitude grid, the dynamics also change with latitude because toward the poles the grid cells become increasingly stretched. This is in addition to the Coriolis effect, the deflection of wind caused by the rotation of Earth, which also depends on latitude. A possible solution in the horizontal could be to use spherical convolutions (Cohen et al., 2018; Jiang et al., 2019; Perraudin et al., 2019) or to feed in latitude information to the network.

Another potential issue is the limited amount of training data available. Forty years of hourly data amounts to around 350,000 samples. However, the samples are correlated in time. If one assumes that a new weather situation occurs every day, then the number of samples is reduced to around 15,000. Without empirical evidence it is hard to estimate whether this number is sufficient to train complex networks without overfitting. Should overfitting be a problem, one could try transfer learning. In transfer learning, the network is pretrained on a similar task or data set, for example, climate model simulations, and then finetuned on the actual data. This is common practice in computer vision and has been successfully applied to seasonal ENSO forecasting (Ham et al., 2019). Another common method to prevent overfitting is data augmentation, which in traditional computer vision is done by, for example, randomly rotating or flipping the image. However, many of the traditional data augmentation techniques are questionable for physical fields. Random rotations, for example, will likely not work for this data set since the x and y directions are physically distinct. Thus, finding good data augmentation techniques for physical fields is an outstanding problem. Using ensemble analyses and forecasts could provide more diversity in the training data set.

Finally, there are technical challenges. Data for a single variable with 10 levels at 5.625° resolution take up around 30 GB of data. For a network with several variables or even at higher resolution, the data might not fit into CPU RAM any more and data loading could become a bottleneck. For image files, efficient data loaders have been created. For netCDF files, however, so far no efficient solution exists to our knowledge. Further, one can assume that to create a competitive data-driven NWP model, high resolutions have to be used, for which GPU RAM quickly becomes a limitation. This suggests that multi-GPU training might be necessary to scale up this approach (potentially similar to the technical achievement of Kurth et al., 2018).

6.2. Probabilistic Forecasts and Extremes

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. In physical NWP this is done by running several forecasts, called an ensemble, from slightly different initial conditions and potentially with different or stochastic model physics (Palmer, 2019). From

this Monte Carlo forecast one can then estimate a probability distribution. A different approach, which is often taken in statistical postprocessing of NWP forecasts, is to directly estimate a parametric distribution (Gneiting et al., 2005; Rasp & Lerch, 2018). For a probabilistic forecast to be reliable the forecast uncertainty has to be an accurate indicator of the error. A good first-order approximation for this is the spread (ensemble standard deviation) to error (RMSE) ratio which should be one (Leutbecher & Palmer, 2008). A more stringent test is to use a proper probabilistic scoring rule, for example the continuous ranked probability score (CRPS) (Gneiting & Katzfuss, 2014; Gneiting & Raftery, 2007). For deterministic forecast the CRPS reduces to the mean absolute error. 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. We encourage users of this benchmark to explore this dimension.

A related issue is the question of extreme weather situations, for example heat waves. 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 for now but users are encouraged to chose their own verification of extremes.

6.3. Climate Simulations

Another aspect that is untouched by the benchmark challenge proposed here is climate prediction. Even though weather and climate deal with the same underlying physical system, they pose different forecasting challenges. In weather forecasting the goal is to predict the state of the atmosphere at a specific time into the future. This is only possible up to the prediction horizon of the atmosphere, which is thought to be at roughly two weeks. Climate models, on the other hand, are evaluated by comparing long-term statistics to observations, for example, the mean surface temperature (Stocker et al., 2013). Scher and Messori (2019a) created iterative climate time scale runs with their data-driven models and compared first and second-order statistics. They found that the model sometimes produced a stable climate but with significant biases and a poor seasonal cycle. This indicates that, so far, iterative data-driven models have been unable to produce physically reasonable long-term predictions. This remains a key challenge for future research. While not specifically included in this benchmark, a good test for climate simulations is to look at long-term mean statistics and the seasonal cycle as done in Figures 6 and 7 of Scher and Messori (2019a).

Climate change simulations represent another step up in complexity. To start with, external greenhouse gas forcing would have to be included. Further, future climates will produce atmospheric states that lie outside of the historical manifold of states. Plain neural networks are very bad at extrapolating to climates beyond what they have seen in the training data set (Rasp et al., 2018). For this reason, climate change simulations with current data-driven methods are likely not a good idea. However, research into physical machine learning is ongoing and might offer new opportunities in the near future (Bar-Sinai et al., 2019; Beucler et al., 2019).

6.4. Promising Research Directions

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. This data set provides, so far, an 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 data set offers a 4 times higher horizontal resolution than all previous studies. The hope is that these data will enable researcher to train more complex models than have previously been used.

With regard to model architecture, there is a huge variety of network architectures that can be explored. U-Nets (Ronneberger et al., 2015) have been used extensively for image segmentation tasks that require computations across several spatial scales. Resnets (He et al., 2015) are currently the state of the art for image classification and their residual nature could be a good fit for state-to-state forecasting tasks (see Han et al., 2020 for a recent application in meteorology). For synthesis tasks, generative adverserial networks (GANs) (Goodfellow et al., 2014) were shown to be particularly powerful for creating realistic natural images and fluid flows (Y. Xie et al., 2018). This might be attractive since minimizing a mean loss, such as the MSE, for random or stochastic data leads to unrealistically smooth predictions as seen in Figure 3. Conditional GANs

(Isola et al., 2016; Mirza & Osindero, 2014) could potentially alleviate this issue but it is still unclear to what extent GAN predictions are able to recover the multivariate distribution of the training samples.

7. Conclusions

In this paper a benchmark data set for data-driven weather forecasting is presented. It focuses on global medium-range (roughly 2 days to 2 weeks) prediction. With the rise of deep learning in physics, weather prediction is a challenging and interesting target because of the large overlap with traditional deep learning tasks (Reichstein et al., 2019). While first attempts have been made in this direction, as discussed in section 2, the field currently lacks a common data set which enables the intercomparison of different methods. We hope that this benchmark can provide a foundation for accelerated research in this area. Loosely following Ebert-Uphoff et al. (2017), the key features of this benchmark are as follows:

- *Scientific impact.* Numerical weather forecasting impacts many aspects of society. Currently, NWP model run on massive supercomputers at very high computational cost. Building a capable data-driven model would be beneficial in many ways (see section 1). In addition, there is the open, and highly debated, question whether fully data-driven methods are able to learn a good representation of atmospheric physics.
- *Challenge for data science.* While global weather prediction is conceptually similar an image-to-image task, and therefore allows for the application of many state-of-the-art deep learning techniques, there are some unique challenges to this problem: the three-dimensional, anisotropic nature of the atmosphere; nonuniform-grids; potentially limited amounts of training data and the technical challenge of handling large data volumes.
- *Clear metric for success.* We defined a single metric (RMSE) for two fields (500 hPa geopotential and 850 hPa temperature). These scores provide a simple measure of success for data-driven, medium-range forecast model.
- *Quick start.* The code repository contains a quick-start Jupyter notebook for reading the data, training a neural network and evaluating the predictions against the target data. In addition, the repository contains many functions which are likely to be used frequently, for example, an implementation of periodic convolutions in Keras.
- *Reproducibility and citability.* All baselines and results form this paper are fully reproducible from the code repository. Further, the baseline predictions are all saved in the data repository. The data have been assigned a permanent DOI.
- *Communication platform.* We will use the Github code repository as an evolving hub for this project. We encourage users of this data set to start by forking the repository and eventually merge code that might be useful for others back into the main branch. The main platform for communication, for example, asking questions, about this project will be Github issues.

We hope that this benchmark will foster collaboration between atmospheric and data scientists in the ways we imagined and beyond.

Appendix A: Additional Metrics

The anomaly correlation coefficient (ACC) is defined as

$$\mathrm{ACC} = \frac{\sum_{i,j,k} L(j) f'_{i,j,k} t'_{i,j,k}}{\sqrt{\sum_{i,j,k} L(j) f'^2_{i,j,k} \sum_{i,j,k} L(j) t'^2_{i,j,k}}} \tag{A1}$$

where the prime $'$ denotes the difference to the climatology. Here the climatology is defined as $\mathrm{climatology}_{j,k} = \frac{1}{N_{time}} \sum t_{j,k}$. The mean absolute error is defined just like the MSE (Equation 2) but with the absolute instead of the squared difference. Figure A1 shows the RMSE for T2M and TP. Figure A2 and Table A1 show ACC, Figure A3 and Table A2 show MAE.

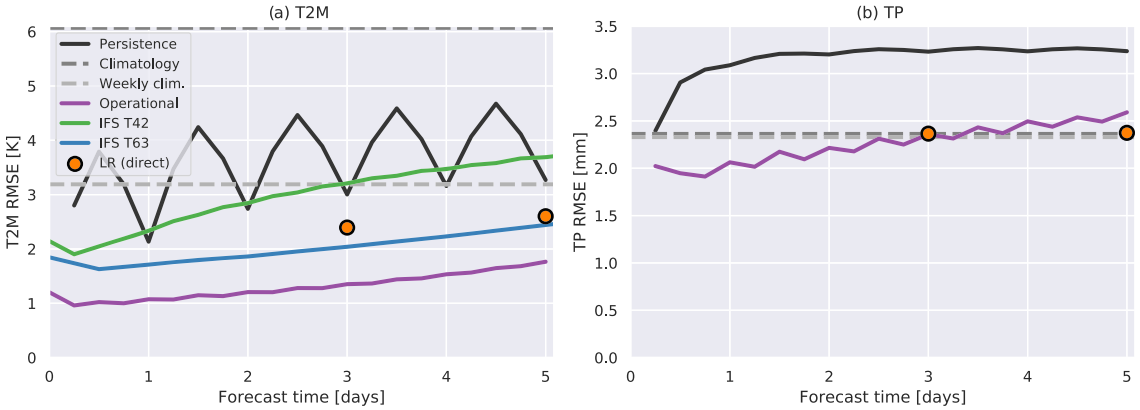

Figure A1. RMSE of (a) 2-meter temperature and (b) 6-hourly accumulated precipitation for different baselines at 5.625° resolution.

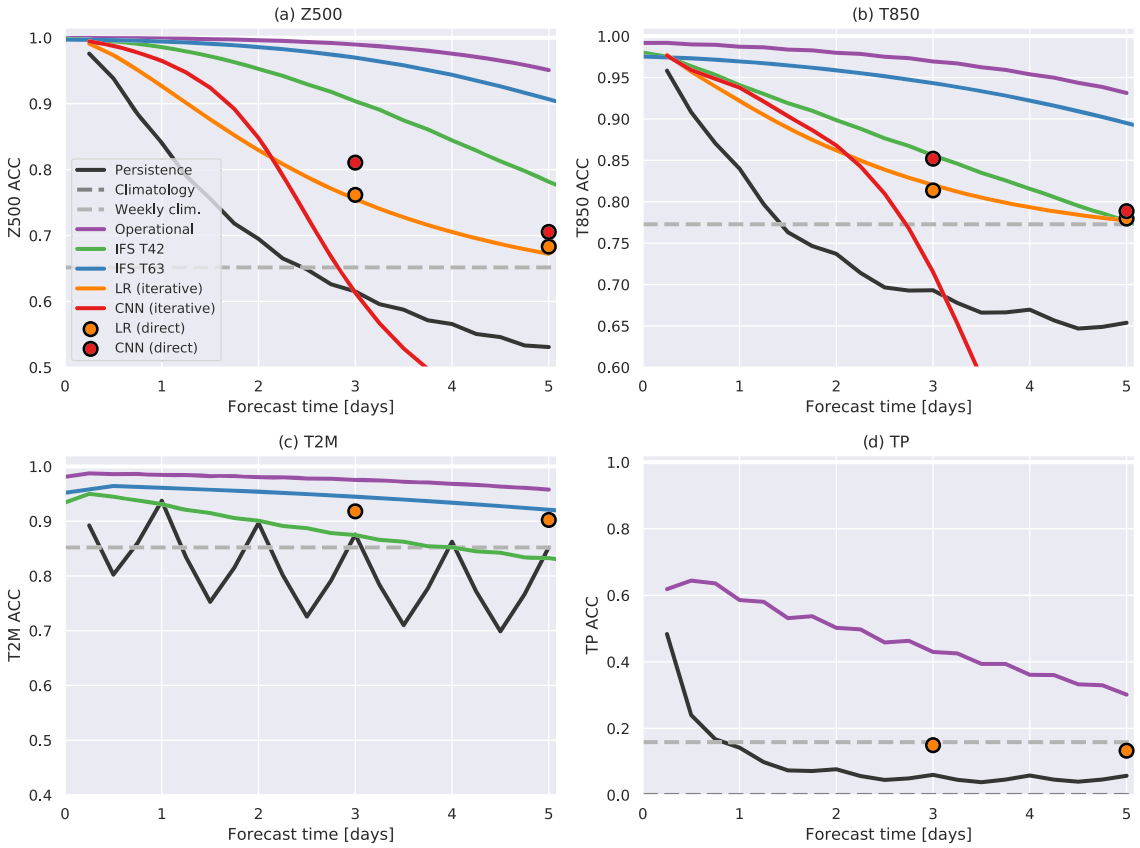

Figure A2. ACC of (a) 500 hPa geopotential, (b) 850 hPa temperature, (c) 2-m temperature, and (d) 6-hourly accumulated precipitation for different baselines at 5.625° resolution.

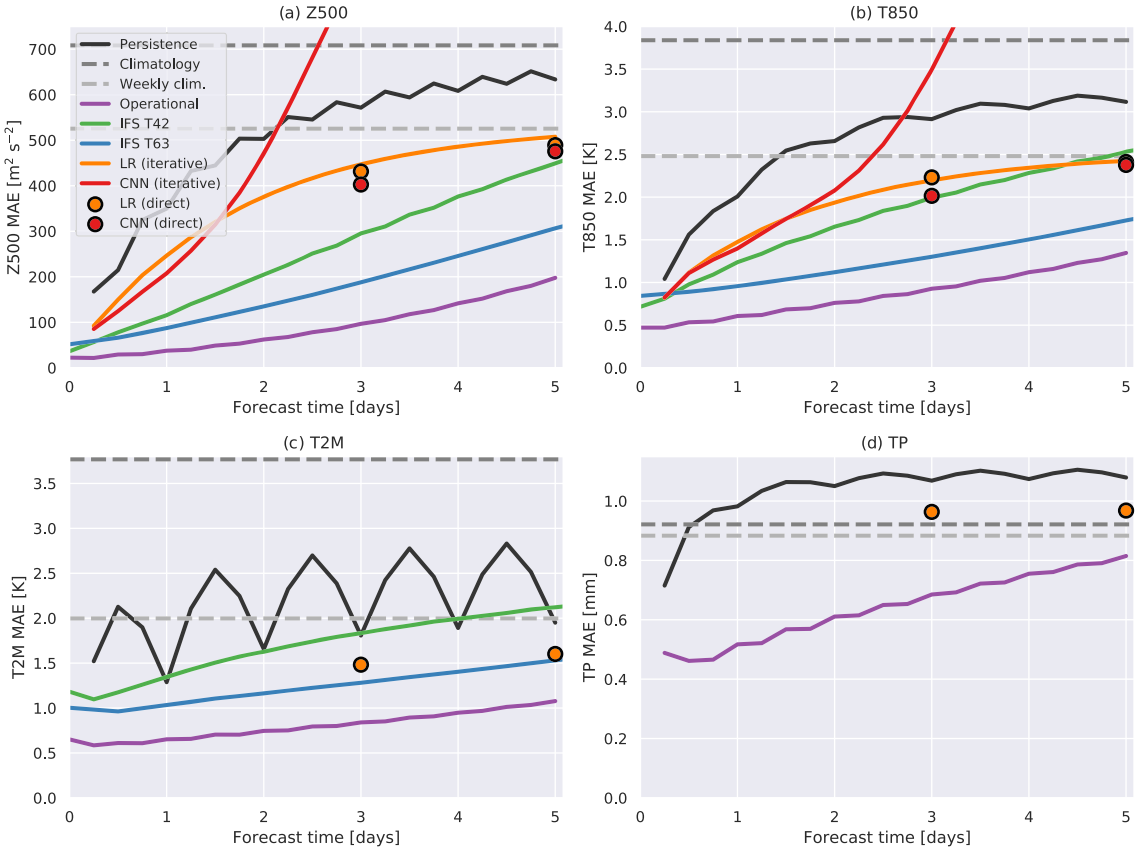

Figure A3. MAE of (a) 500 hPa geopotential, (b) 850 hPa temperature, (c) 2-m temperature, and (d) 6-hourly accumulated precipitation for different baselines at 5.625° resolution.

Table A1
Baseline ACC for 3 and 5 Days Forecast Time at 5.625° Resolution

	ACC (3 days/5 days)			
Baseline	Z500	T850	T2M	TP
Persistence	0.62/0.53	0.69/0.65	0.88/0.85	0.06/0.06
Climatology	0	0	0	0
Weekly climatology	0.65	0.77	0.85	0.16
Linear regression (direct)	0.76/0.68	0.81/0.78	0.92/0.90	0.15/0.13
Linear regression (iterative)	0.76/0.67	0.82/0.78		
CNN (direct)	0.81/0.71	0.85/0.79		
CNN (iterative)	0.61/0.41	0.72/0.31		
IFS T42	0.90/0.78	0.86/0.78	0.87/0.83	
IFS T63	0.97/0.91	0.94/0.90	0.94/0.92	
Operational IFS	0.99/0.95	0.97/0.93	0.98/0.96	0.43/0.30

Note. TP is 6-hourly accumulated precipitation.

Table A2
Baseline MAE for 3 and 5 Days Forecast Time at 5.625° Resolution

	MAE (3 days/5 days)			
Baseline	Z500 ($m^2 s^{-2}$)	T850 (K)	T2 M (K)	TP (mm)
Persistence	572/634	2.91/3.12	1.81/1.95	1.07/1.08
Climatology	708	3.84	3.77	0.92
Weekly climatology	525	2.48	2.00	0.88
Linear regression (direct)	431/489	2.23/2.41	1.48/1.60	0.96/0.97
Linear regression (iterative)	447/508	2.20/2.42		
CNN (direct)	403/476	2.02/2.38		
CNN (iterative)	892/1,263	3.49/7.49		
IFS T42	295/449	1.99/2.53	1.83/2.12	
IFS T63	188/307	1.30/1.73	1.28/1.53	
Operational IFS	97/198	0.93/1.35	0.84/1.08	0.69/0.81

Note. TP is 6-hourly accumulated precipitation.

Data Availability Statement

The WeatherBench data set is available for download at https://mediatum.ub.tum.de/1524895 (Rasp et al., 2020)

Acknowledgments
Stephan Rasp acknowledges funding from the German Research Foundation (DFG). We thank the Copernicus Climate Change Service (C3S) for allowing us to redistribute the data. Peter D. Dueben gratefully acknowledges funding from the Royal Society for his University Research Fellowship and the ESIWACE2 project. The ESIWACE2 project have received funding from the European Unions Horizon 2020 research and innovation programme under Grant Agreement 823988.

References

Agrawal, S., Barrington, L., Bromberg, C., Burge, J., Gazen, C., & Hickey, J. (2019). Machine learning for precipitation nowcasting from radar images. Retrieved from https://arxiv.org/abs/1912.12132

Bar-Sinai, Y., Hoyer, S., Hickey, J., & Brenner, M. P. (2019). Learning data-driven discretizations for partial differential equations. *Proceedings of the National Academy of Sciences*, *116*(31), 15,344–15,349. https://doi.org/10.1073/PNAS.1814058116

Bauer, P., Thorpe, A., & Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, *525*(7567), 47–55. https://doi.org/10.1038/nature14956

Betts, A. K., Chan, D. Z., & Desjardins, R. L. (2019). Near-surface biases in ERA5 over the Canadian prairies. *Frontiers in Environmental Science*, *7*, 129. https://doi.org/10.3389/fenvs.2019.00129

Beucler, T., Rasp, S., Pritchard, M., & Gentine, P. (2019). Achieving conservation of energy in neural network emulators for climate modeling. Retrieved from http://arxiv.org/abs/1906.06622

Bougeault, P., Toth, Z., Bishop, C., Brown, B., Burridge, D., Chen, D. H., et al. (2010). The THORPEX interactive grand global ensemble. *Bulletin of the American Meteorological Society*, *91*(8), 1059–1072. https://doi.org/10.1175/2010BAMS2853.1

Brenowitz, N. D., & Bretherton, C. S. (2018). Prognostic validation of a neural network unified physics parameterization. https://doi.org/10.17605/OSF.IO/EU3AX

Brenowitz, N. D., & Bretherton, C. S. (2019). Spatially extended tests of a neural network parametrization trained by coarse-graining. *Journal of Advances in Modelling Earth Systems*, *11*, 2728–2744. https://doi.org/10.1029/2019MS001711

Chevallier, F., Chéruy, F., Scott, N. A., & Chédin, A. (1998). A neural network approach for a fast and accurate computation of a longwave radiative budget. *Journal of Applied Meteorology*, *37*(11), 1385–1397. https://doi.org/10.1175/1520-0450(1998)037<1385:ANNAFA>2.0.CO;2

Clevert, D.-A., Unterthiner, T., & Hochreiter, S. (2015). Fast and accurate deep network learning by exponential linear units (ELUs). Retrieved from http://arxiv.org/abs/1511.07289

Cohen, T. S., Geiger, M., Köhler, J., & Welling, M. (2018). Spherical CNNs. In *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR.

Ebert-Uphoff, I., Thompson, D. R., Demir, I., Gel, Y. R., Hill, M. C., Karpatne, A., et al. (2017). A vision for the development of benchmarks to bridge geoscience and data science. In *7th International Workshop on Climate Informatics*. Boulder. Retrieved from https://par.nsf.gov/servlets/purl/10057023

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Gagne, D. J., McGovern, A., Xue, M., Gagne, D. J. II, McGovern, A., & Xue, M. (2014). Machine learning enhancement of storm-scale ensemble probabilistic quantitative precipitation forecasts. *Weather and Forecasting*, *29*(4), 1024–1043. https://doi.org/10.1175/WAF-D-13-00108.1

Gneiting, T., & Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, *1*(1), 125–151. https://doi.org/10.1146/annurev-statistics-062713-085831

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378 en. https://doi.org/10.1198/016214506000001437

Gneiting, T., Raftery, A. E., Westveld, A. H., & Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, *133*(5), 1098–1118. https://doi.org/10.1175/MWR2904.1

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets, *Advances in neural information processing systems* (pp. 2672–2680).
Grönquist, P., Yao, C., Ben-Nun, T., Dryden, N., Dueben, P., Li, S., & Hoefler, T. (2020). Deep learning for post-processing ensemble weather forecasts. Retrieved from http://arxiv.org/abs/2005.08748
Graves, A. (2013). Generating sequences with recurrent neural networks. Retrieved from http://arxiv.org/abs/1308.0850
Ham, Y. G., Kim, J. H., & Luo, J. J. (2019). Deep learning for multi-year ENSO forecasts. (Vol. 573) (No. 7775). Nature Publishing Group. https://doi.org/10.1038/s41586-019-1559-7
Hamill, T. M., & Whitaker, J. S. (2006). Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Monthly Weather Review*, *134*(11), 3209–3229. https://doi.org/10.1175/MWR3237.1
Han, Y., Zhang, G. J., Huang, X., & Wang, Y. (2020). A moist physics parameterization based on deep learning. *Journal of Advances in Modeling Earth Systems*, *12*, e2020MS002076. https://doi.org/10.1029/2020MS002076
He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. Retrieved from http://arxiv.org/abs/1512.03385
Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*, 1999–2049. https://doi.org/10.1002/qj.3803
Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–80. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9377276
Holton, J. R. (2004). *An introduction to dynamic meteorology* (Vol. 88). https://doi.org/10.1119/1.1987371
Hoskins, B. J., McIntyre, M. E., & Robertson, A. W. (1985). On the use and significance of isentropic potential vorticity maps. *Quarterly Journal of the Royal Meteorological Society*, *111*(470), 877–946. https://doi.org/10.1002/qj.49711147002
Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2016). Densely connected convolutional networks. http://arxiv.org/abs/1608.06993
Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2016). Image-to-image translation with conditional adversarial networks. Retrieved from https://arxiv.org/pdf/1611.07004v1.pdf
Jiang, C. M., Huang, J., Kashinath, K., Prabhat, Marcus, P., & Niessner, M. (2019). Spherical CNNs on unstructured grids. Retrieved from http://arxiv.org/abs/1901.02039
Kaji, S., & Kida, S. (2019). Overview of image-to-image translation by use of deep neural networks: Denoising, super-resolution, modality conversion, and reconstruction in medical imaging. Retrieved from http://arxiv.org/abs/1905.08603
Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv, 1412.6980. Retrieved from http://arxiv.org/abs/1412.6980
Koster, J., & Rahmann, S. (2012). Snakemake—A scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520–2522. https://doi.org/10.1093/bioinformatics/bts480
Krasnopolsky, V. M., Fox-Rabinovitz, M. S., & Chalikov, D. V. (2005). New approach to calculation of atmospheric model physics: Accurate and fast neural network emulation of longwave radiation in a climate model. *Monthly Weather Review*, *133*(5), 1370–1383. https://doi.org/10.1175/MWR2923.1
Krogh, A., & Hertz, J. A. (1992). A simple weight decay can improve generalization. In *Advances in neural information processing systems* (pp. 950–957). Retrieved from http://papers.nips.cc/paper/563-a-simple-weight-decay-can-improve-generalization.pdf
Kurth, T., Treichler, S., Romero, J., Mudigonda, M., Luehr, N., Phillips, E., et al. (2018). Exascale deep learning for climate analytics. Retrieved from http://arxiv.org/abs/1810.01993
Lagerquist, R., McGovern, A., & Smith, T. (2017). Machine learning for real-time prediction of damaging straight-line convective wind. *Weather and Forecasting*, *32*(6), 2175–2193. https://doi.org/10.1175/WAF-D-17-0038.1
Lazo, J. K., Morss, R. E., & Demuth, J. L. (2009). 300 billion served: sources, perceptions, uses, and values of weather forecasts. *Bulletin of the American Meteorological Society*, *90*(6), 785–798. https://doi.org/10.1175/2008BAMS2604.1
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539
LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2323. https://doi.org/10.1109/5.726791
Leutbecher, M., & Palmer, T. N. (2008). Ensemble forecasting. *Journal of Computational Physics*, *227*(7), 3515–3539. https://doi.org/10.1016/j.jcp.2007.02.014
McGovern, A., Elmore, K. L., Gagne, D. J., Haupt, S. E., Karstens, C. D., Lagerquist, R., et al. (2017). Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bulletin of the American Meteorological Society*, *98*(10), 2073–2090. https://doi.org/10.1175/BAMS-D-16-0123.1
McGovern, A., Lagerquist, R., Gagne, D. J., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., & Smith, T. (2019). Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, *100*, 2175–2199. https://doi.org/10.1175/bams-d-18-0195.1
Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. arXiv, 1411.1784. Retrieved from http://arxiv.org/abs/1411.1784
Morton, J., Witherden, F. D., Jameson, A., & Kochenderfer, M. J. (2018). Deep dynamical modeling and control of unsteady fluid flows. Retrieved from http://arxiv.org/abs/1805.07472
Mudigonda, M., Kashinath, K., Kapp-Schwoerer, L., Graubner, A., Karaismailoglu, E., von Kleist, L., et al. (2020). Climatenet: Bringing the power of deep learning to weather and climate sciences via open datasets and architectures. ICLR 2020 workshop.
NCAR (2020). The climate data guide: Common spectral model grid resolutions. Retrieved from https://climatedataguide.ucar.edu/climate-model-evaluation/common-spectral-model-grid-resolutions
Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., et al. (2016). WaveNet: A generative model for raw audio. Retrieved from http://arxiv.org/abs/1609.03499
Ott, J., Pritchard, M., Best, N., Linstead, E., Curcic, M., & Baldi, P. (2020). A Fortran-Keras deep learning bridge for scientific computing. Retrieved from http://arxiv.org/abs/2004.10652
Palmer, T. (2019). The ECMWF ensemble prediction system: Looking back (more than) 25 years and projecting forward 25 years. *Quarterly Journal of the Royal Meteorological Society*, *145*(S1), 12–24. https://doi.org/10.1002/qj.3383
Perraudin, N., Defferrard, M., Kacprzak, T., & Sgier, R. (2019). DeepSphere: Efficient spherical convolutional neural network with HEALPix sampling for cosmological applications. *Astronomy and Computing*, *27*, 130–146. https://doi.org/10.1016/j.ascom.2019.03.004
Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., & Thuerey, N. (2020). *WeatherBench: A benchmark dataset for data-driven weather forecasting*: Technical University of Munich. https://mediatum.ub.tum.de/1524895
Rasp, S., & Lerch, S. (2018). Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, *146*(11), 3885–3900. https://doi.org/10.1175/MWR-D-18-0187.1

Rasp, S., Pritchard, M. S., & Gentine, P. (2018). *Deep learning to represent subgrid processes in climate models*. In *Proceedings of the National Academy of Sciences of the United States of America, 201810286*. https://doi.org/10.1073/pnas.1810286115

Rasp, S., Schulz, H., Bony, S., & Stevens, B. (2019). Combining crowd-sourcing and deep learning to explore the meso-scale organization of shallow convection. Bulletin of the American Meteorological Society. https://doi.org/10.1175/BAMS-D-19-0324.1

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. Retrieved from http://arxiv.org/abs/1505.04597

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Scher, S. (2018). Toward data-driven weather and climate forecasting: Approximating a simple general circulation model with deep learning. *Geophysical Research Letters*, *45*, 616–12. https://doi.org/10.1029/2018GL080704

Scher, S., & Messori, G. (2019a). Generalization properties of neural networks trained on Lorenz systems. *Nonlinear Processes in Geophysics Discussions*, *26*, 381–399. https://doi.org/10.5194/npg-2019-23

Scher, S., & Messori, G. (2019b). 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. https://doi.org/10.5194/gmd-12-2797-2019

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W., & Woo, W. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. Retrieved from http://arxiv.org/abs/1506.04214

Shi, X., Gao, Z., Lausen, L., Wang, H., Yeung, D.-Y., Wong, W.-K., & Woo, W.-C. (2017). Deep learning for precipitation nowcasting: A benchmark and a new model. Retrieved from http://arxiv.org/abs/1706.03458

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958. Retrieved from http://www.jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf?utm_content=buffer79b43&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer

Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M. M. B., Allen, S. K., Boschung, J., et al. (2013). Climate change 2013. The physical science basis. Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Retrieved from https://inis.iaea.org/Search/search.aspx?origq=RN:45042273

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. http://arxiv.org/abs/1409.3215

Taillardat, M., Mestre, O., Zamo, M., & Naveau, P. (2016). Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Monthly Weather Review*, *144*, 2375–2393. https://doi.org/10.1175/MWR-D-15-0260.1

Toms, B. A., Barnes, E. A., & Ebert-Uphoff, I. (2019). Physically interpretable neural networks for the geosciences: applications to earth system variability. Retrieved from http://arxiv.org/abs/1912.01752

Vogel, P., Knippertz, P., Fink, A. H., Schlueter, A., & Gneiting, T. (2018). Skill of global raw and postprocessed ensemble predictions of rainfall over Northern Tropical Africa. *Weather and Forecasting*, *33*(2), 369–388. https://doi.org/10.1175/WAF-D-17-0127.1

Wallace, J. M., & Hobbs, P. V. (2006). *Atmospheric science: An introductory survey*. Cambridge, MA: Academic Press. https://doi.org/10.1021/jp112019s

Weyn, J. A., Durran, D. R., & Caruana, R. (2019). Can machines learn to predict weather? Using deep learning to predict gridded 500hPa geopotential height from historical weather data. *Journal of Advances in Modeling Earth Systems*, *11*, 2680–2693. https://doi.org/10.1029/2019MS001705

Wilks, D. S. (2006). *Statistical methods in the atmospheric sciences*. Cambridge, MA: Academic Press. Retrieved from http://cds.cern.ch/record/992087

Xie, Y., Franz, E., & Chu, M. (2018). tempoGAN: A temporally coherent, volumetric GAN for super-resolution fluid flow method using a variety of complex inputs and applications in two and three dimensions. *ACM Transactions on Graphics*, *37*, 15. https://doi.org/10.1145/3197517.3201304

Xie, D., Xiong, J., & Pu, S. (2017). All you need is beyond a good init: Exploring better solution for training extremely deep convolutional neural networks with orthonormality and modulation. Retrieved from http://arxiv.org/abs/1703.01827

Xu, Z., Bi, S., Sunkavalli, K., Hadap, S., Su, H., & Ramamoorthi, R. (2019). Deep view synthesis from sparse photometric images. *ACM Transactions on Graphics*, *38*(4), 1–13. https://doi.org/10.1145/3306346.3323007

Yuval, J., & O'Gorman, P. A. (2020). Use of machine learning to improve simulations of climate. Retrieved from http://arxiv.org/abs/2001.03151

Zhuang, J. (2019). xESMF: V0.2.1. Retrieved from https://xesmf.readthedocs.io/ doi: http://10.5281/ZENODO.3475638