JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2020MS002405

Key Points:

- A large convolutional neural network is trained for the WeatherBench challenge
- Pretraining on climate model data improves skill and prevents overfitting
- The model sets a new state-of-the-art for data-driven medium-range forecasting

Correspondence to:
S. Rasp,
raspstephan@gmail.com

Citation:
Rasp, S. & Thuerey, N. (2021). Data-driven medium-range weather prediction with a Resnet pretrained on climate simulations: A new model for WeatherBench. *Journal of Advances in Modeling Earth Systems*, *13*, e2020MS002405. https://doi.org/10.1029/2020MS002405

Received 10 NOV 2020
Accepted 22 JAN 2021

Data-Driven Medium-Range Weather Prediction With a Resnet Pretrained on Climate Simulations: A New Model for WeatherBench

Stephan Rasp[1,2] **and Nils Thuerey**[1]

[1]Department of Informatics, Technical University of Munich, Munich, Germany, [2]Now at ClimateAi, San Francisco, USA

Abstract Numerical weather prediction has traditionally been based on the models that discretize the dynamical and physical equations of the atmosphere. Recently, however, the rise of deep learning has created increased interest in purely data-driven medium-range weather forecasting with first studies exploring the feasibility of such an approach. To accelerate progress in this area, the WeatherBench benchmark challenge was defined. Here, we train a deep residual convolutional neural network (Resnet) to predict geopotential, temperature and precipitation at 5.625° resolution up to 5 days ahead. To avoid overfitting and improve forecast skill, we pretrain the model using historical climate model output before fine-tuning on reanalysis data. The resulting forecasts outperform previous submissions to WeatherBench and are comparable in skill to a physical baseline at similar resolution. We also analyze how the neural network creates its predictions and find that, for the case studies analyzed, the model has learned physically reasonable correlations. Finally, we perform scaling experiments to estimate the potential skill of data-driven approaches at higher resolutions.

Plain Language Summary Weather forecasts are created by running hugely complex computer simulations that encapsulate our knowledge of how the atmosphere works. This approach has served us well but is there a different way? The paradigm of machine learning proposes learning an algorithm from data rather than building it from physical principles. For several areas like computer vision and natural language processing this has worked exceedingly well, so it just makes sense to try it as well for weather forecasting. This paper presents the latest attempt at training a machine learning weather forecasting model. It is shown that the learned model produces reasonable forecasts, approximately on par with traditional models run on much lower resolution. However, there is still a large gap to current state–of–the–art high–resolution weather models that is unlikely to be closed with a purely data–driven approach because not enough training data exists.

1. Introduction

Current numerical weather prediction (NWP) is based on physical models of the atmosphere, and the ocean, in which the governing equations are discretized and sub-grid processes are parameterized (Kalnay, 2003). Continued refinement of these models along with increasing computing power and better observations to create initial conditions has led to steady increases in forecast skill over the last 4 decades (Bauer et al., 2015). The improvements in the model components and the tuning of free parameters is, in a large majority of cases, guided by scientific expertize rather than using a statistical method (Hourdin et al., 2017). In the current operational weather forecasting chain, the only component that includes a learning algorithm is post-processing, the correction of statistical errors from NWP output. Most commonly, post-processing is done using simple linear techniques (model output statistics) but in recent years more modern machine learning techniques, such as random forests and neural networks, have been explored (Grönquist et al., 2020; McGovern et al., 2017; Rasp & Lerch, 2018; Taillardat et al., 2016).

With the apparent successes of deep learning in modeling high-dimensional data in other domains such as computer vision and natural language processing, a natural question to ask is whether numerical weather models can also be learned purely from data. This question sparked some debate after initial studies (Dueben & Bauer, 2018; Scher, 2018; Scher & Messori, 2019; Weyn et al., 2019) showed the general feasibility of such an approach for medium-range weather forecasting. In particular, some researchers were skeptical

whether the complex physics described by systems of partial differential equations could be encoded in a neural network.

To answer this question Rasp et al. (2020b) defined a benchmark challenge for data-driven medium-range weather forecasting called WeatherBench. Specifically the challenge is to predict 500 hPa geopotential (Z500), 850 hPa temperature (T850), 2-m temperature (T2M) and 6-hourly accumulated precipitation (PR) up to 5 days ahead. Here, we train a large neural network for this task. Section 2 describes the data set and the neural network setup. In Section 3, the results for the WeatherBench benchmark are presented and discussed. Section 4 includes several sensitivity experiments followed by an attempt to interpret the neural network's predictions in Section 5. Finally, we discuss the results in Section 6.

2. Materials and Methods

2.1. Data, Evaluation, and Baselines

The data are provided by the WeatherBench challenge. A full description can be found in Rasp et al. (2020b) and the latest version of the data is available at https://github.com/pangeo-data/WeatherBench. WeatherBench contains regridded ERA5 (Hersbach et al., 2020) data from 1979 to 2018 at hourly resolution, of which 2017 and 2018 are set aside for evaluation. In addition, we use climate model simulations to pretrain our simulations as described below. For this, we downloaded a historical simulation from the CMIP6 archive (Eyring et al., 2016). Specifically, we picked the MPI-ESM-HR model since it was one of the only models for which the data was saved at vertical resolution to match the ERA5 data. The temporal resolution of the CMIP data is 6 hours. The regridded climate model data are also available on the WeatherBench data repository. In addition to the data, WeatherBench defines the evaluation metrics. The area-weighted RMSE and ACC are used for evaluating 500 hPa geopotential (Z500), 850 hPa temperature (T850), 2-m temperature (T2M), and 6-hourly accumulated precipitation (PR) at 3 and 5 days lead time. The area-weighted RMSE is defined as

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

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:

$$L(j) = \frac{\cos(\mathrm{lat}(j))}{\frac{1}{N_{\mathrm{lat}}} \sum_{j}^{N_{\mathrm{lat}}} \cos(\mathrm{lat}(j))} \tag{2}$$

The definition of the ACC can be found in the appendix.

Furthermore, WeatherBench contains several baselines from physical models: the operational Integrated Forecasting System (IFS) of the European Center for Medium-range Weather Forecasting (ECMWF), the current state-of-the-art in NWP, which currently runs at 9 km horizontal resolution with 137 vertical levels; and the same model run at two lower resolutions, T42 (~2.8° or 310 km at the equator) with 62 vertical levels and T63 (~1.9° or 210 km at the equator) with 137 vertical levels. For an exact definition of the evaluation metrics and the initialization of the baseline models, refer to Rasp et al. (2020b). Furthermore, a climatology and persistence baseline is computed as well as a climatology computed for each calendar week. As an additional baseline, here we include the work by Weyn et al. (2020) who trained an neural network to predict Z500 and T850. Their model is iterative, that is, it consists of a sequence of 6 h forecasts. During training they also trained their neural network over two-time steps (12 h) to ensure stability for longer integrations. Further they mapped the latitude-longitude data to a cube-sphere grid with roughly 1.9° resolution to minimize the distortion during the convolution operations. Their model was trained on 40 years of ERA data.

2.2. Data-Driven Forecasts Using a Pretrained Resnet

There are three fundamental techniques for creating data-driven forecasts: direct, continuous and iterative. For direct forecasts, a separate model is trained directly for each desired forecast time. In continuous models, time is an additional input and a single model is trained to predict all forecast lead times (as in MetNet; Sønderby et al. (2020)). Finally, iterative forecasts are created by training a direct model for a short forecast time (e.g., 6 h) and then running the model several times using its own output from the previous iteration. As mentioned above, this is the approach taken by Weyn et al. (2020).

Here, we train direct and continuous models. Advantages and disadvantages of each technique will be discussed later. All models in this study use the same architecture (except in the network size scaling experiments). The basic structure is a fully convolutional Resnet (He et al., 2015) with 19 residual blocks. Each residual block consists of two convolutional blocks, defined as (two-dimensional [2D] convolution → LeakyReLU → Batch normalization → Dropout), after which the inputs to the residual layer are added to the current signal. The 2D convolutions inside the residual blocks have 128 channels with a kernel size of 3. All convolutions are periodic in longitude with zero padding in the latitude direction. For the first layer a simple convolutional block with 128 channels is used with a kernel size of 7 to increase the field of view. LeakyReLU is used with $\alpha = 0.3$. Weight decay of 1×10^{-5} is used for all layers. Dropout is set to 0.1.

The inputs are geopotential, temperature, zonal and meridional wind and specific humidity at seven vertical levels (50, 250, 500, 600, 700, 850, and 925 hPa), 2-m temperature, 6-h accumulated precipitation, the top-of-atmosphere incoming solar radiation, all at the current time step t, $t - 6$h and $t - 12$h, and, finally three constant fields: the land-sea mask, orography and the latitude at each grid point. All fields were normalized by subtracting the mean and dividing by the standard deviation, with the exception of precipitation for which the mean was not subtracted to keep the lower bound at zero. Additionally, we log-transform of the precipitation to make the distribution less skewed ($\tilde{PR} = \ln(\epsilon + PR) - \ln(\epsilon)$) with $\epsilon = 0.001$. Subtracting the log of ϵ ensures that zero values remain zero. This transformation turns out to be crucial to prevent the network from simply predicting zeros. All variables, levels and time-steps were stacked to create an input signal with 114 channels. For the continuous forecast, in addition, we add 32×64 fields which contains the forecast time in hours divided by 100. During training, a random forecast time from 6 to 120 h is drawn for each sample. Two separate sets of networks were trained, one to predict Z500, T850, and T2M and another one to predict TP. The reason for treating TP separately is that its distribution is significantly more skewed even after the log-transform compared to the other three variables. Predicting all four variables with a single network led to bad predictions for all variables. A loss scaling factor for TP be one potential solution but here, we chose to simply treat it separately.

For our best models, we first train our model using the 150 years of CMIP data described above. We then take the pretrained model and fine-tune it using the ERA data. We will also show results for models trained only with CMIP or ERA data. The loss function is the latitude-weighted mean squared error. The latitudes are weighted proportionally to the area of the grid boxes $\propto \cos\phi$. The Adam optimizer (Kingma & Ba, 2014) is used with a batch size of 32 and an initial learning rate of 5×10^{-5} for the ERA and CMIP only experiments. The learning rate was decreased twice by a factor of five when the validation loss has not decreased for two epochs. Early stopping on the validation loss was used to terminate training with a patience of five epochs. The training period for ERA was from 1979 to 2015, validation was done with a single year (2016). For fine-tuning the CMIP networks on ERA data, a lower initial learning rate of 5×10^{-7} was chosen. Also note that for the pretrained model, no dropout was used as this led to better validation scores. For the direct approach we trained models for 6 h, 1, 3, and 5 days forecast time. We used Tensorflow 2. Training a single model takes around 1 day on a GTX 2080 GPU.

3. WeatherBench Results

Figure 1 and Table 1 show the results of the two networks on the WeatherBench metrics (ACC results can be found in Appendix Table A1. Notably pretraining with CMIP data improves skill significantly for Z500, T850, and T2M over just using ERA data, with increasing impact for longer lead times. This is because overfitting, as measured by the difference between training and testing scores, tends to be worse for longer lead

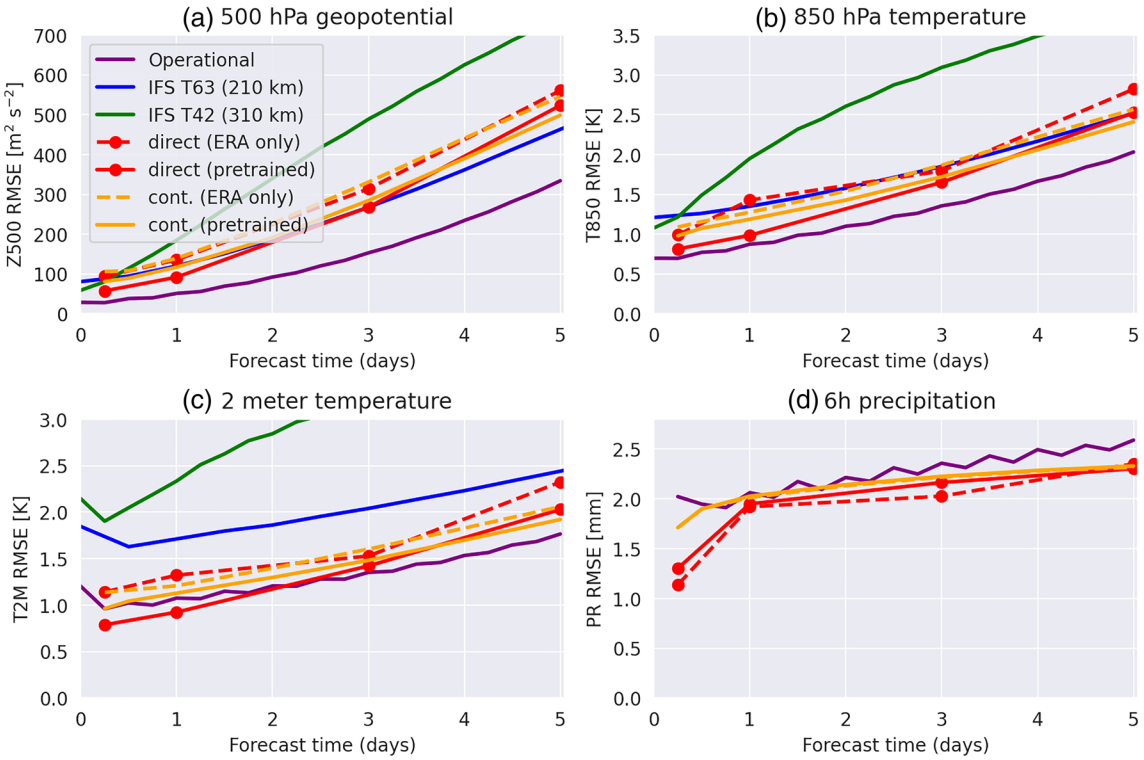

Figure 1. Root mean squared error (RMSE) for (a) Z500, (b) T850, (c) T2M, and (d) PR evaluated against ERA5 data.

Table 1
RMSE for 3–5 days Forecast Time

Model	Latitude-weighted RMSE (3 days/5 days)			
	Z500 [m^2 s^{-2}]	T850 [K]	T2M [K]	PR [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**
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
Weyn et al. (2020)	373/611	1.98/2.87	–	–
Direct (ERA only)	314/561	1.79/2.82	1.53/2.32	**2.03**/2.35
Direct (CMIP only)	323/561	2.09/2.82	1.90/2.32	2.30/2.39
Direct (pretrained)	**268**/523	**1.65**/2.52	**1.42**/2.03	2.16/**2.30**
Continuous (ERA only)	331/545	1.87/2.57	1.60/2.06	2.22/2.32
Continuous (CMIP only)	330/548	2.12/2.75	2.24/2.59	2.29/2.38
Continuous (pretrained)	284/**499**	1.72/**2.41**	1.48/**1.92**	2.23/2.33

Note. All forecasts evaluated at 5.625° resolution. Best physical and data-driven methods are highlighted.

Abbreviations: IFS, integrated forecasting system; RMSE, root mean squared error.

times (Figure 2a). As there is a longer time for errors to grow nonlinearly for longer forecast horizons, similar initial conditions can lead to a wider range of outcomes. In the face of such uncertainty, a model that is trained to minimize the mean squared error, will tend to predict the mean of the distribution of possible outcomes. Our hypothesis is that for a wider distribution (longer forecast time) more training data are required to estimate the mean. In other words, if, because of the intrinsic unpredictability of the atmosphere, a broader range of outcome is physically plausible, then overfitting to individual outcomes encountered in the training data will lead to more overfitting than it would for shorter forecast times, where the plausible forecasts are closer together. Pretraining with climate model data helps to prevent overfitting and leads to better testing scores. Strikingly, even without fine-tuning on ERA data ("CMIP only" in Table 1) the testing scores computed on reanalysis data are not much or not at all worse than the "ERA only" networks. This shows that climate models, even though they do not exactly represent the real atmosphere, provide a good proxy for the general circulation of the atmosphere. For precipitation, pretraining does not improve skill. This is most likely because precipitation skill is low anyway and climate models might not represent precipitation as realistically as the large-scale circulation. Finally, it is important to note that RMSE and ACC are sub-optimal metrics for precipitation.

Comparing the direct and continuous models, direct models tend to be better up to around 3 days forecast time, while the continuous models have more skill for longer forecast horizons. This difference also seems to be caused by overfitting. The continuous models without pretraining

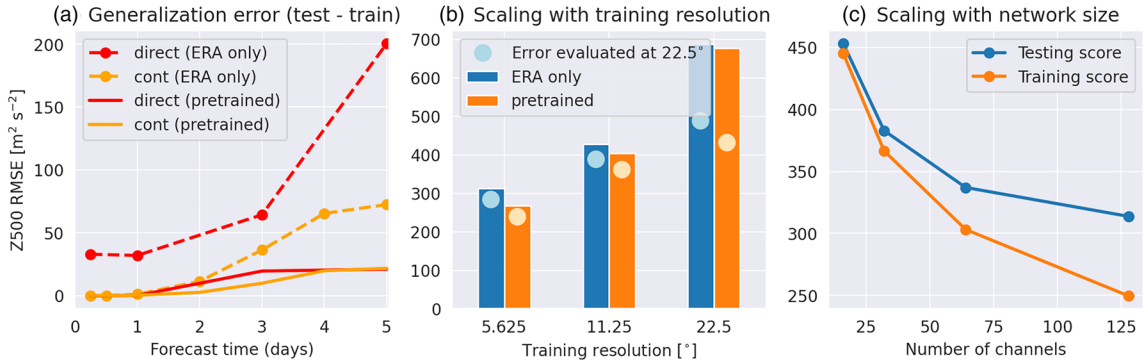

Figure 2. (a) Generalization error (testing minus training RMSE for Z500). (b) RMSE of Z500 for networks trained with different resolution data. Bars show the RMSE computed at 5.625° resolution. For this the predictions from the lower resolution networks were upscaled. Dots show the RMSE evaluated at 22.5° for which all predictions were downscaled. (c) RMSE of Z500 for different network architectures. *y*-axis has the same units (Z500 RMSE in $m^2\ s^{-2}$) for all three panels. RMSE, root mean squared error.

have a lower generalization error (Figure 2a). One hypothesis as to why this is, could be that the fact that, in the continuous approach, a single model has to learn to make predictions for all forecast times acts as data-augmentation. Another plausible hypothesis is that the continuous networks also learn a time-evolution of the flow which helps regularize the network. With pretraining, the difference in the generalization error does not appear as large but this is potentially an artifact of using early stopping for fine-tuning rather than a sign that the direct models do not over-fit more. The two approaches, direct and continuous, therefore, represent a trade-off between specificity and generalization. One advantage of the continuous method is that arbitrary forecast times within the training range can be chosen. However, using the continuous network to predict beyond its training range quickly leads to large errors (not shown).

The models presented in this study outperform the simple machine learning baselines from the original WeatherBench study and also the approach of Weyn et al. (2020) (Table 1). However, as mentioned previously, their model is an iterative model. Technically this is a more difficult approach because training a neural network for short-term predictions (in their case 6 h) and then calling it iteratively can lead to self-amplifying errors. In fact, this is what we observed when trying this with our model architecture. Furthermore, this requires the output vector to match the input vector, greatly increasing the number of variables to predict which can lead to a loss in specificity. Weyn et al. (2020) trained the model over two time steps. This, however, quickly becomes very computationally expensive for large model such as the ones in this study. Therefore, if the goal is simply to predict a certain field at a predefined time ahead, the direct and continuous approaches will likely lead to better results. On the other hand, iterative models can be used to make arbitrarily long predictions which opens up a range of potential use cases.

Finally, it is interesting to discuss the performance of our models compared to the physical baselines. For Z500 and T850, the skill is comparable to the T63 model, for T2M a little better than that. However, there are big caveats to consider in this comparison. First, the physical models (operational IFS and T63) are initialized from slightly different initial conditions, leading to a nonzero error at $t = 0$. In addition, the coarse resolution models T42 and T63 suffer from errors due to the conversion to spherical coordinates at coarse resolutions. Since error growth is initially exponential, this initial condition difference primarily affects short forecast times up to 2 days (Zhang et al., 2007). A likely more important consideration is that the T42 and T63 models were not tuned for this resolution. This is in contrast to the operational IFS model which is carefully tuned over many years. This means that tuning the lower resolution IFS models would almost certainly lead to increased skill, however it is hard to estimate how much. On the other hand, our models are trained at significantly coarser resolutions and further hyper-parameter/architecture tuning would likely result in better scores. Another limitation is that statistical errors of the physical model were not removed by post-processing and that the evaluation was done at a very coarse grid. This is likely not so important for the upper-level variables Z500 and T850 but very important for surface variables (Hewson & Pillosu, 2020). In data-driven forecasts the post-processing is implicitly performed. Lastly, it is important to consider that our models do not necessarily predict realistic fields. Figure 3 one can see that with increasing lead time the

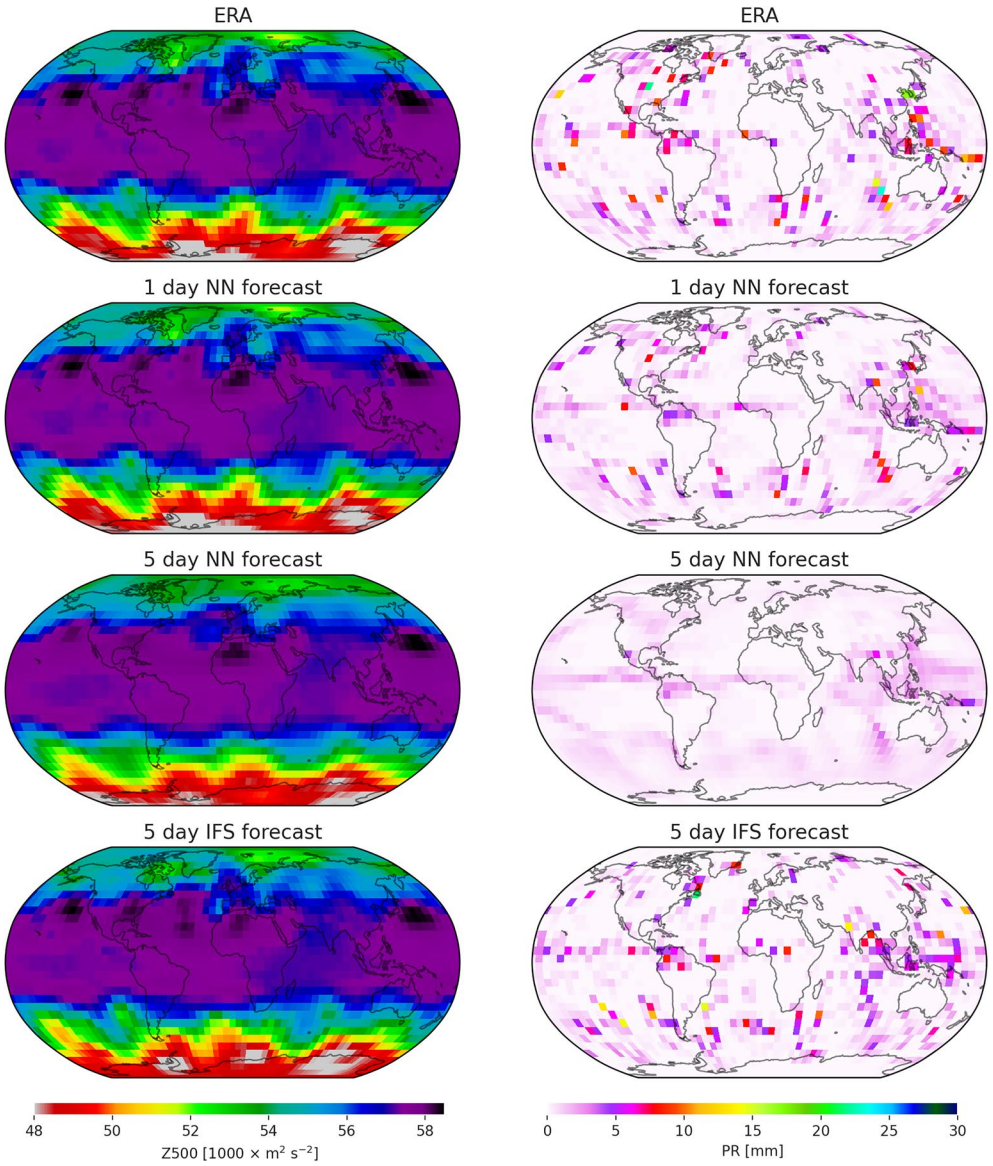

Figure 3. Sample forecasts valid at July 1, 2018 00UTC for 500 hPa geopotential (top row) and 6 h accumulated precipitation (bottom row). 1 and 5 days pretrained, direct neural network forecasts are compared to the 5 days operational IFS forecast and the ERA5 ground truth. IFS, integrated forecasting system.

predictions become more smoothed out and lose variability compared to the observations. This is another reflection of predicting the mean of the hypothetical forecast distribution as mentioned above. For geopotential and temperature this is especially grave in the extra-tropics where the largest natural fluctuations occur.

Particular care has to be applied when comparing precipitation between the neural networks and the IFS models. The mean squared error is not an optimal choice for verifying intermittent fields like precipitation. More fitting metrics would have to be applied to get an accurate view of forecast skill but are outside the scope of the WeatherBench challenge. As the snapshots in Figure 3 show, and the scores suggest, for longer lead times the neural network essentially learns to predict a climatological mean rather than anything remotely realistic. The IFS model, on the other hand, forecast a seemingly realistic field, with precipitation in slightly wrong positions. The inherent stochasticity of precipitation is a key reason why probabilistic forecasts are necessary, which are not considered here.

3.1. Sensitivity to Resolution and Network Size

It is interesting to ask how the results might change if the resolution was increased or larger networks were trained. Doing so, however, is technically very challenging and are outside the scope of this study. We can however, assess the scaling to resolution and network size by using lower resolution and smaller networks. For this purpose we trained 3-days direct networks using 11.25° and 22.5° data but an otherwise identical training procedure (Figure 2b). The skill drops with coarser resolution. This trend is present regardless whether the evaluation was done at 5.625° or 22.5° resolution with higher/lower resolution data interpolated to the evaluation resolution. This tendency makes sense since a higher data resolution provides better information to the network. One caveat of this sensitivity test is that we left the model architecture the same for these experiments, which means that the number of parameters relative to the size of the input/output vectors increases with coarser resolutions.

To compare different network sizes we reduced the number of channels in each convolution from 128 to 64, 32 and 16 (Figure 2c). The number of parameters decreases approximately by a factor 4 for each reduction. The testing skill increases with increasing network size but the trend flattens off and overfitting increases. This suggests that, while further improvements are certainly possible, there likely is a ceiling in skill for a given amount of training data. Note that the regularization parameters (weight decay and dropout) are the same across all network sizes. Another way to change the network size would be to change the number of layers. These experiments led to qualitatively similar results. Recent findings in deep learning (Nakkiran et al., 2019) suggest that, further increasing network size can lead to lower testing losses despite increased overfitting. It would be interesting to see whether similar trends hold for this data set.

4. Interpretability

The data-driven weather models predict weather with reasonable skill. One interesting question to ask is whether they do this for the "right reasons." To find out, we test which variables and which geographical region are important for the network to make a prediction. We do this by computing saliency maps (Simonyan et al., 2013). That is for each sample, we chose a point in space and a specific variable p, for example, T850 over London. We then compute the gradient G of this scalar p with respect to the entire input array $X \in \mathcal{R}^{\text{samples}\times\text{lat}\times\text{lon}\times\text{variables}}$: $G = \partial p/\partial X$ with the same shape as X. We do this analysis for two climatologically different locations: London, which is in the mid-latitudes and therefore influenced by eastwards-propagating Rossby waves and Barbados, located in the sub-tropical trade wind zones. This is done for different lead times using the pretrained direct networks.

It is important to highlight that the saliency method does not evaluate which inputs were most important for the prediction but rather which changes in the input would most affect the output. For a discussion on the differences, see (Ebert-Uphoff & Hilburn, 2020). For the purposes of this study, the saliency method is appropriate since it allows us to evaluate effect of small input perturbations which is closely related to the body of work on adjoint sensitivity (Ancell & Hakim, 2007).

First, we investigate the region of influence by computing the mean absolute gradient of T850 over all samples $|\overline{G}| = 1 / N_{\text{samples}} \sum_i |G_i|$ and then taking the mean over all input variables (Figure 4a). Because we compute the gradients for the normalized inputs, the different variables and levels should be comparable in scale and the gradients are dimensionless. It is important to highlight that the saliency analysis is primarily of qualitative nature. The resulting maps show that the networks tends to look at physically reasonable geographical regions. For London the region of influence extends toward the West with increasing forecast time. This is in line with our physical understanding of eastwards traveling Rossby waves being a key factor for weather in the mid-latitudes. Furthermore, we can look at the mean gradient $\overline{G} = 1 / N_{\text{samples}} \sum_i G_i$ of a specific input variable, in this case Z500, for 3 days forecast time (Figure 4b). Here, we see a positive-negative pattern across the Atlantic. Physically, one could interpret this as the signature of Rossby phase shifts influencing the temperature over London several days ahead. Over Barbados the region of influence looks smaller and more circular. This is in accordance with calmer meteorological conditions in the subtropics. We also performed this analysis for specific seasons to check for seasonality but could detect little difference. Meteorologically one would maybe expect such differences. This highlights the difficulties of using

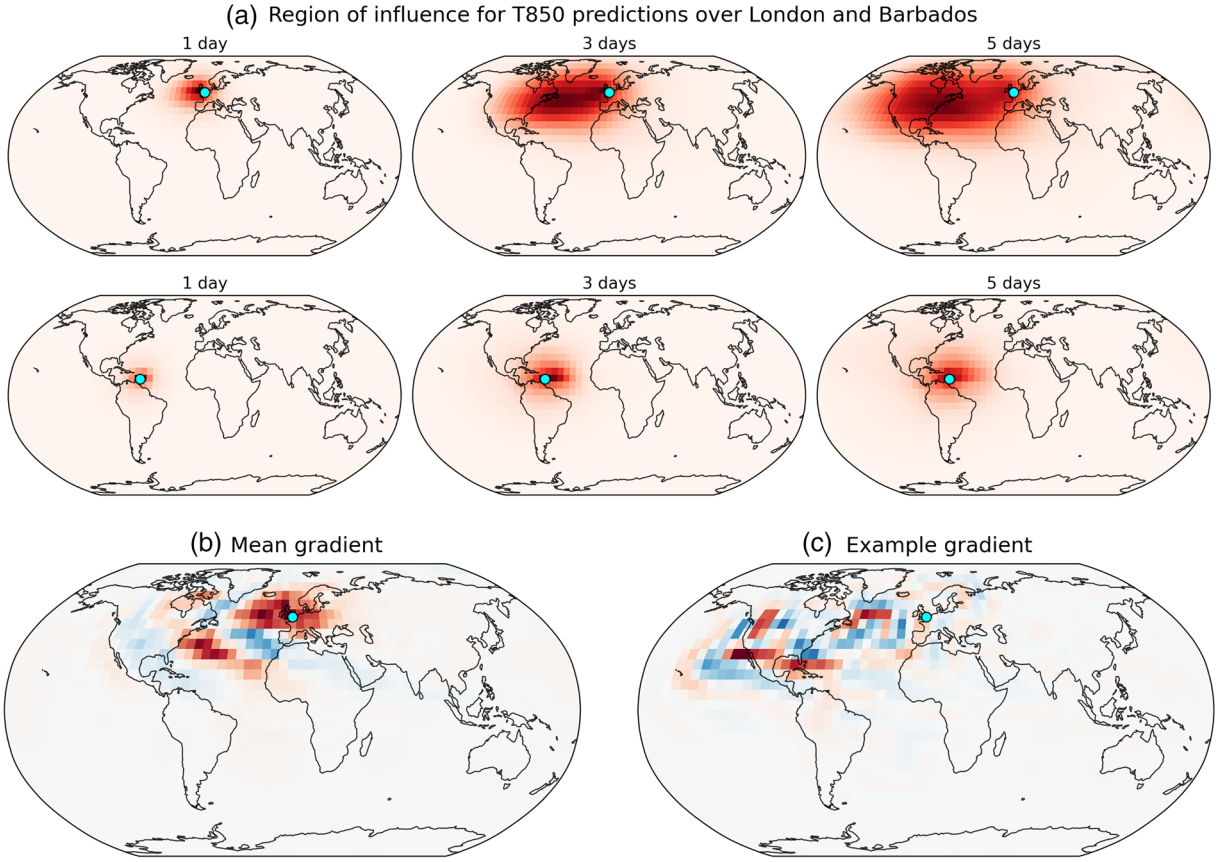

Figure 4. Saliency plots: (a) The region of influence $|\overline{G}|$ (see text for explanation) of T850 over London and Barbados with respect to all input variables (averaged). (b) Mean gradient over time $\overline{G}$ of T850 with respect to Z500 over London. (c) Sample gradient G of Z500 with respect to 250 hPa geopotential for 8 January 2017 12:00UTC.

saliency analyses to discover physical insight. Rather, here we mainly check whether the network learns completely unreasonably correlations.

We can also take the horizontal mean of $|\overline{G}|$ to obtain the mean influence of each normalized input variable (Figure 5). Geopotential and temperature show the largest gradients on average. Specifically changes in the geopotential at 250 hPa appear to have a large effect. This is reasonable since 250 hPa is close to the tropopause and changes in the tropopause height are known to be influential for medium-range weather evolution (Hoskins et al., 1985). Furthermore, the gradient analysis shows that T2M is important for Barbados which reflects the importance of the ocean temperatures. Comparing, the influence of the inputs at the current time step t, $t-6$h and $t-12$h, the current time step is much more important than earlier time steps. This confirms our empirical findings that adding these previous time steps only improved the scores marginally (not shown).

So far, all results are in agreement with physical reasoning and similar results could be expected to come out of adjoint sensitivity studies with physical models. However, looking at G for individual samples, it is evident that this is not always the case. Figure 4c shows the gradient of T850 over London with respect to the 250 hPa geopotential for a 3 days forecast. Significant gradients stretch across the Atlantic and North America all the way to Hawaii. Such long-range correlations might not be completely physical. Studies using physical models typically estimate that it takes perturbations 5–6 days to cross the Atlantic (Rodwell et al., 2013). These results suggest that while the network, on average, learns physically plausible connections from data it appears to make unphysical connections for some samples. This makes sense since, in our setup, the network purely learns correlations between input and output images and there is nothing

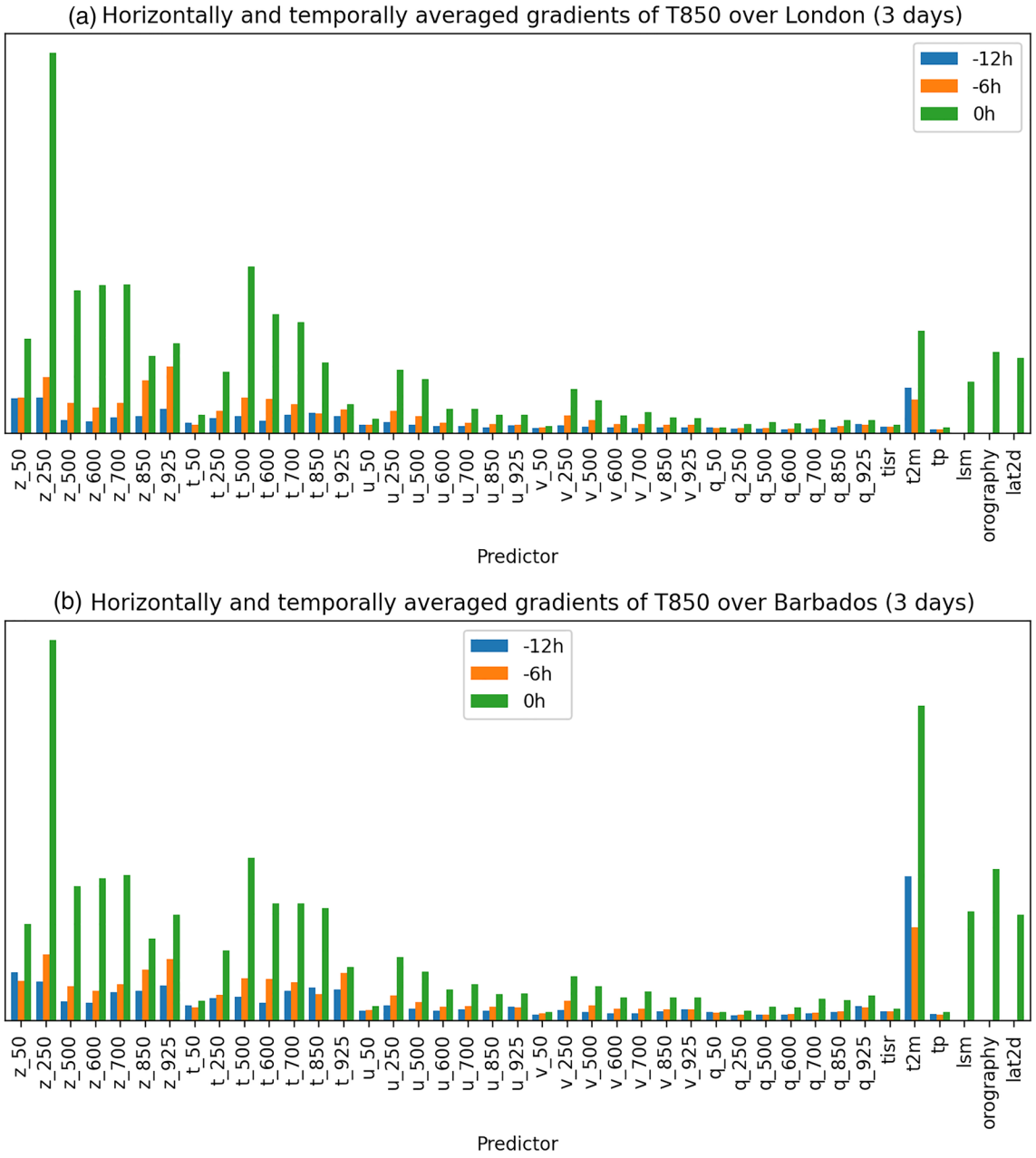

Figure 5. Horizontally averaged saliency $|\overline{G}|$ of T850 over (a) London (b) Barbados for 3 days direct forecasts. tisr is the top-of-atmosphere incoming solar radiation, tp is precipitation, lsm is the land sea mask.

stopping it from learning "unphysical" correlations. If, for example, a certain pattern over eastern North America—which likely has an influence on European weather 3 days later—also concurs in the training data with some pattern over the eastern Pacific, the network will pick up that connection between Pacific and European weather even if it might not be a causal relationship. In a way, such "unphysical" relations are also a sign of overfitting.

5. Discussion and Conclusion

In this study, we presented a data-driven method for medium-range weather forecasting using a Resnet neural network architecture. Specifically, we trained models to predict 500 hPa geopotential, 850 hPa temperature, 2 m temperature and precipitation up to 5 days ahead following the WeatherBench challenge. To avoid overfitting, we pretrained the networks using climate model data from the CMIP archive. Our models set a new data-driven state-of-the-art for WeatherBench. Most previous approaches on similar problems used a U-Net (Ronneberger et al., 2015) architecture, which in our experiments did not work as well as a simple Resnet without any changes in dimensionality. Compared to physical models, the Resnet achieves comparable scores to a physical model at comparable resolution. However, it is important not to over-interpret these results for several reasons discussed in the study. More detailed evaluation would also be needed to accurately compare the two. The focus in this study is primarily on the challenge set by WeatherBench and the methodological development of the neural network models.

It is more interesting to discuss the relevance of the findings presented here for data-driven weather forecasting in the future. It appears that with sufficient training data for pretraining purely data-driven forecasting can achieve reasonable skill. Our scaling analysis indicates that going to higher resolutions and larger networks leads to better scores. It is an interesting question whether the resolution scaling continues for higher resolutions than those considered here. However, the increased overfitting for larger networks already suggests that large amounts of data are required to train competitive data-driven models. One can also assume that larger models are needed for higher-resolutions to maintain a reasonable receptive field. Here, we used climate model simulations to combat overfitting. Current CMIP models, however, are run at around 100 km resolution, and therefore cannot be used for forecasts at higher-resolutions. There are several atmosphere-only climate simulations (Haarsma et al., 2016) run at resolutions comparable to the ERA5 resolution of 25 km. It can be assumed that using all this available data at the highest possible resolution for training would greatly increase the forecast skill of data-driven methods. However, for the resolutions of current operational NWP models (10 km) is it unlikely that there is sufficient data to challenge these models (see Palmer, 2020, for a theoretical argument). As an aside, even if data-driven models matched physical models at forecasting, creating an initial condition currently requires data-assimilation systems that are currently based on physical models.

However, the findings regarding relative skill of data-driven versus physical forecasting are specific to the particular problem at hand. Data-driven methods could still play a large role in the broad field of weather forecasting. Two crucial questions to ask are how much training data is available for a particular problem and how much potential there is to improve upon physical approaches. For medium-range forecasting, physical modeling has achieved impressive skill in recent decades which makes it hard to do better with the observations at hand. Other task in numerical weather prediction could offer a much bigger potential for data-driven methods.

Appendix A: ACC Results

Table A1 shows the ACC skill for all experiments. ACC is defined as

Table A1
ACC for 3–5 Days Forecast Time

	Latitude-weighted ACC (3 days/5 days)			
Model	Z500 [$m^2\ s^{-2}$]	T850 [K]	T2M [K]	PR [mm]
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
IFS T42	0.90/0.78	0.86/0.78	0.87/0.83	–

Table A1
Continued

Model	Latitude-weighted ACC (3 days/5 days)			
	Z500 [$m^2 s^{-2}$]	T850 [K]	T2M [K]	PR [mm]
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**
Direct (ERA only)	0.96/0.85	0.94/0.86	0.97/0.92	**0.55**/0.24
Direct (CMIP only)	0.95/0.85	0.93/0.86	0.95/0.92	0.32/0.20
Direct (pretrained)	**0.97**/0.87	**0.95**/0.89	**0.97**/0.94	0.45/**0.29**
Continuous (ERA only)	0.95/0.86	0.94/0.88	0.96/0.94	0.41/0.29
Continuous (CMIP only)	0.95/0.86	0.93/0.87	0.93/0.91	0.41/0.29
Continuous (pretrained)	0.96/**0.88**	**0.95**/**0.90**	0.97/**0.95**	**0.41**/0.28

Note. All forecasts evaluated at 5.625° resolution.
Abbreviation: IFS, integrated forecasting system.

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

where the prime $'$ denotes the difference to the climatology. Here, the climatology is defined as $\mathrm{climatology}_{j,k} = \frac{1}{N_{time}} \sum t_{j,k}$

Data Availability Statement

The data set is available at https://mediatum.ub.tum.de/1524895 (Rasp et al., 2020a). The code for the WeatherBench challenge is at https://github.com/pangeo-data/WeatherBench. The code for this study specifically is at https://github.com/raspstephan/WeatherBench.

Acknowledgments
The authors would like to thank Sebastian Scher and David Greenberg for their valuable comments on the study as well as George Craig for discussing the saliency analysis. S. Rasp acknowledges funding from the German Research Foundation (DFG) under grant no. 426852073.
Open access funding enabled and organized by Projekt DEAL.

References

Ancell, B., & Hakim, G. J. (2007). Comparing adjoint- and ensemble-sensitivity analysis with applications to observation targeting. *Monthly Weather Review*, *135*, 4117–4134. https://doi.org/10.1175/2007MWR1904.1

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

Dueben, P. D., & Bauer, P. (2018). Challenges and design choices for global weather and climate models based on machine learning. *Geoscientific Model Development*, *11*, 3999–4009. https://doi.org/10.5194/gmd-2018-148

Ebert-Uphoff, I., & Hilburn, K. A. (2020). Evaluation, tuning and interpretation of neural networks for meteorological applications. *Bulletin of the American Meteorological Society*, *101*, E2149–E2170.

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

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

Haarsma, R. J., Roberts, M. J., Vidale, P. L., Senior, C. A., Bellucci, A., Bao, Q., & von Storch, J.-S. (2016). High resolution model intercomparison project (HighResMIPv1.0) for CMIP6. *Geoscientific Model Development*, *9*(11), 4185–4208. https://doi.org/10.5194/gmd-9-4185-2016

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., MuñozSabater, J., & Thépaut, J. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*, 1999–2049. https://doi.org/10.1002/qj.3803

Hewson, T. D., & Pillosu, F. M. (2020). *A new low-cost technique improves weather forecasts across the world*. Retrieved from http://arxiv.org/abs/2003.14397

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep residual learning for image recognition*. Las Vegas, NV:IEEE. Retrieved from http://arxiv.org/abs/1512.03385

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*, 877–946. https://doi.org/10.1002/qj.49711147002

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J.-C., Balaji, V., Duan, Q., et al. (2017). The art and science of climate model tuning. *Bulletin of the American Meteorological Society*, *98*(3), 589–602. https://doi.org/10.1175/BAMS-D-15-00135.1

Kalnay, E. (2003). *Atmospheric modeling, data assimilation, and predictability* (Vol. 54). Cambridge University Press. Retrieved from http://books.google.com/books?hl=en&lr=&id=Uqc7zC7NULMC&oi=fnd&pg=PR11&dq=Atmospheric+modeling,+data+assimilation+and+predictability&ots=lI5gpir1RV&sig=FuhXqkYSMxhz2jLI2T8144HX6fs

Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization*. Retrieved from http://arxiv.org/abs/1412.6980

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

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., & Sutskever, I. (2019). *Deep double descent: Where bigger models and more data hurt*. Retrieved from http://arxiv.org/abs/1912.02292

Palmer, T. (2020, 7). *A vision for numerical weather prediction in 2030*. Retrieved from http://arxiv.org/abs/2007.04830

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., & Thuerey, N. (2020a). *WeatherBench: A benchmark dataset for data-driven weather forecasting*. Technical University of Munich. Retrieved from https://mediatum.ub.tum.de/1524895

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., & Thuerey, N. (2020b). *WeatherBench: A benchmark dataset for data-driven weather forecasting*. Retrieved from http://arxiv.org/abs/2002.00469

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

Rodwell, M. J., Magnusson, L., Bauer, P., Bechtold, P., Bonavita, M., Cardinali, C., et al. (2013). Characteristics of occasional poor medium-range weather forecasts for Europe. *Bulletin of the American Meteorological Society*, *94*(9), 1393–1405. https://doi.org/10.1175/BAMS-D-12-00099.1

Ronneberger, O., Fischer, P., & Brox, T. (2015). *U-Net: Convolutional networks for biomedical image segmentation*. Paper presented at international Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Cham: Springer Retrieved from http://arxiv.org/abs/1505.04597

Scher, S. (2018). Toward data-driven weather and climate forecasting: approximating a simple general circulation model with deep learning. *Geophysical Research Letters*, *45*(22), 616–712. https://doi.org/10.1029/2018GL080704

Scher, S., & Messori, G. (2019). Generalization properties of neural networks trained on Lorenzsystems. *Nonlinear Processes in Geophysics Discussions*, *26*, 1–19. https://doi.org/10.5194/npg-2019-23

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*. Paper presented at 2nd International Conference on Learning Representations ICLR 2014 - Workshop Track Proceedings. Retrieved from http://arxiv.org/abs/1312.6034

Sønderby, C. K., Espeholt, L., Heek, J., Dehghani, M., Oliver, A., Salimans, T., et al. (2020). *MetNet: A neural weather model for precipitation forecasting*. Retrieved from http://arxiv.org/abs/2003.12140

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

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

Weyn, J. A., Durran, D. R., & Caruana, R. (2020). Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2020MS002109. https://doi.org/10.1029/2020MS002109

Zhang, F., Bei, N., Rotunno, R., Snyder, C., & Epifanio, C. C. (2007). Mesoscale predictability of moist baroclinic waves: Convection-permitting experiments and multistage error growth dynamics. *Journal of the Atmospheric Sciences*, *64*(10), 3579–3594.