arXiv:2008.08626v1 [physics.ao-ph] 19 Aug 2020

Purely data-driven medium-range weather forecasting achieves comparable skill to physical models at similar resolution

Stephan Rasp
Department of Informatics
Technical University of Munich
Munich, Germany
stephan.rasp@tum.de

Nils Thuerey
Department of Informatics
Technical University of Munich
Munich, Germany
nils.thuerey@tum.de

Abstract

Numerical weather prediction has traditionally been based on physical models 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. Here, we train a significantly larger model than in previous studies to predict geopotential, temperature and precipitation up to 5 days ahead and achieve comparable skill to a physical model run at similar horizontal resolution. Crucially, we pretrain our models on historical climate model output before fine-tuning them on the reanalysis data. We also analyze how the neural network creates its predictions and find that, with some exceptions, it is compatible with physical reasoning. Our results indicate that, given enough training data, data-driven models can compete with physical models. At the same time, there is likely not enough data to scale this approach to the resolutions of current operational models.

Introduction

Numerical weather prediction (NWP) is based on physical models of the atmosphere, and ocean for longer forecast times, in which the governing equations are discretized and sub-grid processes are parameterized[13]. 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 four decades[2]. The improvements in the model components and the tuning of free parameters is, in a large majority of cases, guided by scientific expertise rather than using a statistical method[12]. 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; MOS) but in recent years more modern machine learning (ML) techniques, such as random forests and neural networks, have been explored[6,15,18,25].

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[3,21,22,26] showed the general feasibility of such an approach for medium-range weather forecasting. In particular, some researchers were sceptical whether the complex physics described by systems of partial differential equations could be encoded in a neural network. In this paper, we aim to answer this fundamental question and explore the potential for purely data-driven weather forecasting as an alternative to physical modeling.

Specifically, we focus on global, medium-range weather forecasting up to 5 days forecast time. From a societal point of view, this forecast range is particularly important in order to prepare for extreme weather events such as heavy rainfall, tropical cyclones, cold spells or heat waves. Scientifically, producing a good medium-range forecast requires modeling the large-scale dynamics of the atmosphere, especially the initiation and evolution of tropical and extra-tropical cyclones. This is in contrast to very short-term prediction of e.g. precipitation, called "nowcasting", which does not necessarily require knowledge about atmospheric dynamics and can be done by extrapolating observations into the future[24].

We tackle the challenge posed in the WeatherBench benchmark[19], namely predicting 500 hPa geopotential (Z500), 850 hPa temperature (T850), 2-meter temperature (T2M) and 6-hourly accumulated precipitation (PR) up to 5 days ahead. The first two variables represent upper-level atmospheric variables that describe the large-scale flow of the atmosphere, while the latter two represent impact variables. The data used are 40 years of ERA5 reanalyses[9] at 5.625° resolution (32×64 grid points in latitude/longitude, approximately 625 km resolution at the equator). This data is much coarser than current operational weather ($\sim$10 km) or even climate models ($\sim$100 km). However, with a large deep neural network and using significantly more variables and levels than has been done before, using higher-resolution data is technically very challenging. As we will see below, even 5.625° resolution data allow us to draw meaningful conclusions about the potential of data-driven weather forecasting if it was scaled to higher resolutions. In addition to the ERA data, we use 150 years of climate model data from the Climate Model Inter-comparison Project (CMIP)[5] for pretraining (see Methods).

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[24]). 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. Here, we train direct and continuous models using a fully-convolutional Resnet[8] that takes the prognostic variables at seven vertical levels as well as some surface and constant fields at the current time t as well as $t - 6$h and $t - 12$h as input. Z500, T850 and T2M are predicted with a separate set of networks tha PR (see Methods for details on the model). Iterative models, if successfully trained, have some nice properties such long-term stability[27]. However, the computational cost and memory requirements of training such models over several time steps is challenging for a network of the size used here.

Finally, it is important to have meaningful baselines to judge the skill of the data-driven techniques. As a gold standard, we use the operational Integrated Forecasting System (IFS) model from the European Center for Medium-range Weather Forecasting (ECMWF) which currently has a horizontal resolution of around 10 km[29]. Further, we compare our forecasts to IFS forecasts run at lower resolution, T42 (approximately 2.8° or 310 km at the equator) and T63 (approximately 1.9° or 210 km) (see Methods for details). These lower-resolution versions should provide a fair comparison to our data-driven forecasts in terms of resolution and computational expense.

Results

Forecast skill

For the upper-level fields Z500 and T850, the direct and continuous models achieve comparable skill to the T63 forecast across metrics with better relative scores for short forecast times (Figs. 1 and S1 and Tables 1 and S1). Pretraining with climate model data helps in achieving good skill with increasing impact for longer forecast times as the gap between the ERA only and pretrained networks show. This is because overfitting, as measured by the difference between training and testing scores, tends to be worse for longer lead times (Fig. 3a). As the atmosphere becomes more chaotic 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 is 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

Table 1: RMSE for 3 and 5 days forecast time. All forecasts evaluated at 5.625° resolution. Best physical and data-driven methods are highlighted.

	Latitude-weighted RMSE (3 days / 5 days)			
Model	**Z500 [m^2 s^{-2}]**	**T850 [K]**	**T2M [K]**	**PR [mm]**
Persistence	936 / 1033	4.23 / 4.56	3.00 / 3.27	3.23 / 3.24
Climatology	1075	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

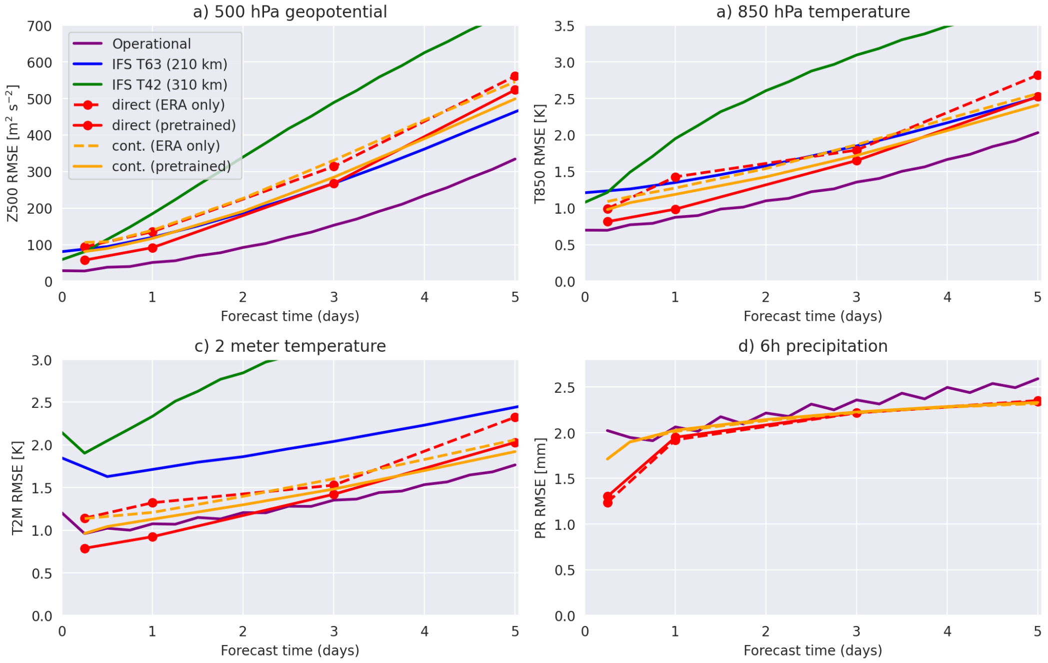

Figure 1: Root mean squared error (RMSE) for a) Z500, b) T850, c) T2M and d) PR evaluated against ERA5 data.

closer together. Pretraining with climate model data helps to prevent overfitting (Fig. 3a) and leads to better testing scores (Fig. 1). 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.

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 have a lower generalization error (Fig. 3a). 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

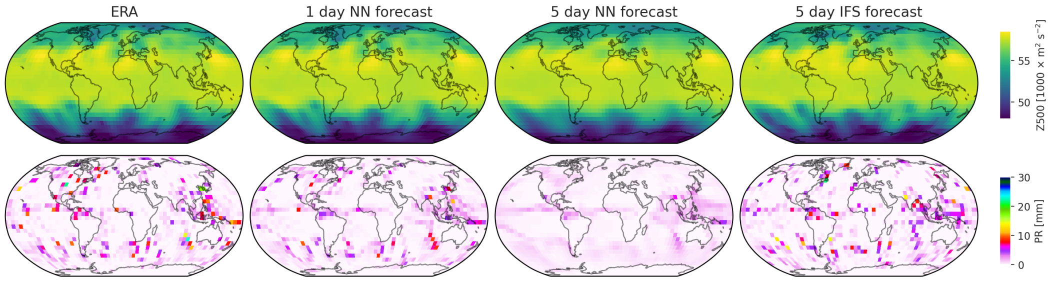

Figure 2: Sample forecasts valid at 1 July 2018 00UTC for 500 hPa geopotential (top row) and 6 h accumulated precipitation (bottom row). 1 and 5 day pretrained, direct neural network forecasts are compared to the 5 day operational IFS forecast and the ERA5 ground truth.

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 overfit 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 data-driven forecasts for T2M have higher skill than the T63 forecast and even match the raw operational IFS predictions for the first few forecast days. However, as discussed in more detail below the evaluation shown here favors our data-driven forecasts. For PR this is even more pronounced as the data-driven forecasts match the operational IFS skill even at 5 days forecast time. However, looking at a sample prediction (Fig. 2) one can see that with increasing lead time the 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. For precipitation, a much more chaotic variable, the smoothing out is much more pronounced across the globe. At 5 days forecast time the data-driven forecast has no extreme events left.

Sensitivity to resolution and network size

Next, we conducted sensitivity tests to probe the scaling of forecast skill with data resolution and network size. To assess the impact of resolution we trained 3-day direct networks using 11.25° and 22.5° data but an otherwise identical training procedure (Fig. 3b). The skill drops with coarser resolution. This trend is present regardless whether the evaluation was done at 5.625° or 22.5°

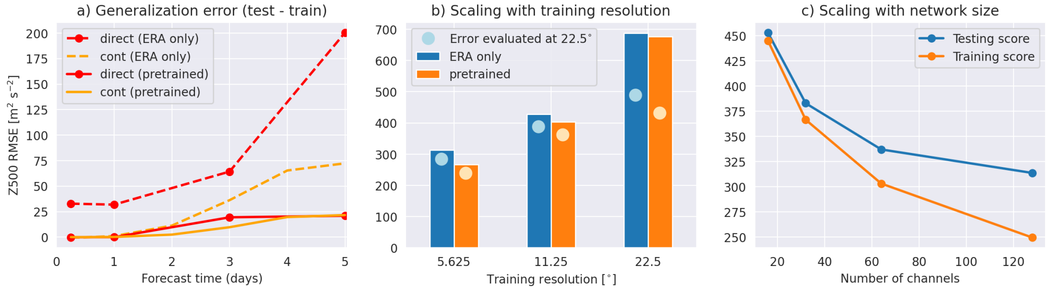

Figure 3: a) Generalization error (testing minus training RMSE for Z500) b) RMSE of Z500 for networks trained with different resolution data. Bars show the RMSE computed at 5.625° resolution. For this the predictions from the lower resolution networks were upscaled. Dots show the RMSE evaluated at 22.5° 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.

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 (Fig. 3c). 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[16] 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 dataset.

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[23]. That is for each sample, we chose a point in space and a specific variable p, e.g. T850 over London. We then compute the gradient G of this scalar p with respect to the entire input array $X \in \mathbb{R}^{samples \times lat \times lon \times variables}$: $G = \partial p / \partial X$ 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[4]. For the purposes of this paper, 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[1].

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 (Fig. 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 towards 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. Further we can look at the mean gradient $\bar{G} = 1/N_{\text{samples}} \sum_i G_i$ of a specific input variable, in this case Z500, for 3 days forecast time (Fig. 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 very different and more circular which is in accordance with calmer meteorological conditions in the subtropics without a persistent preferred wind direction. Note that $\overline{|G|}$ is a mean over all seasons so that seasonally prevailing regimes do not show up.

We can also take the horizontal mean of $|\bar{G}|$ to obtain the mean influence of each normalized input variable (Supp. Fig. S2). 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[11]. Further, 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. The confirms our empirical findings that adding these previous time steps only improved the scores marginally (not shown).

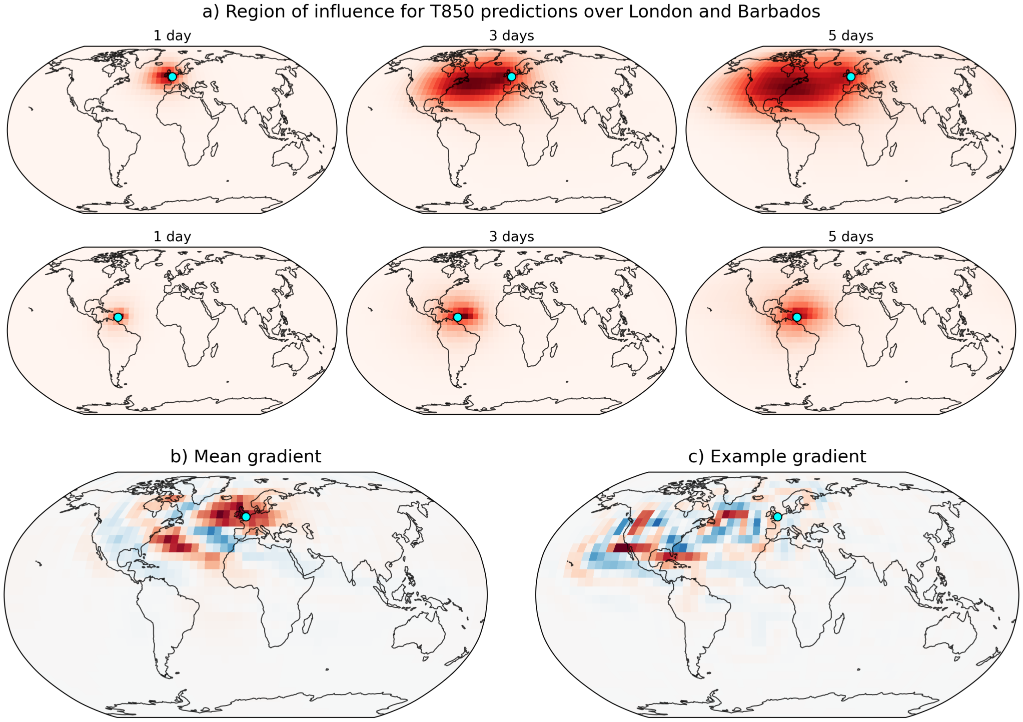

Figure 4: Saliency plots: a) The region of influence $\overline{|G|}$ (see text for explanation) of T850 over London and Barbados with respect to all input variables (averaged). b) Mean gradient over time $\bar{G}$ of T850 with respect to Z500 over London. c) Sample gradient G of Z500 with respect to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

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. Fig. 4c shows the gradient of T850 over London with respect the 250 hPa geopotential for a 3 day forecast. Significant gradients stretch across the Atlantic and North America all the way to Hawaii. This extent of information propagation within 3 days is rather unphysical. Studies using physical models typically estimate that it takes perturbations 5–6 days to cross the Atlantic[20]. 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 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 layer - 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.

Discussion

To accurately assess how the data-driven models stack-up against physical models it is important to highlight some aspects of the comparison conducted in this study. First, the physical models (operational IFS and T63) are initialized from slightly different initial conditions, leading to a non-zero 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 two days[30]. 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 like T2M and PR[10]. In data-driven forecasts the post-processing is implicitly performed. While not a perfect one-to-one comparison we believe that it is fair to say that our data-driven models achieve comparable skill compared to physical models at similar horizontal resolutions.

More generally, our findings suggest that, given enough data, there is no fundamental reason why purely data-driven forecasts could not be as good as state-of-the-art physical models. 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[7] 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 ref.[17] 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. The findings regarding the relative potential skill of data-driven forecasting versus physical modeling are specific to the problem of forecasting global weather in the medium range, however. The applicability of data-driven forecasts has to be assessed for every application separately based on the availability of data and the potential to improve upon physical approaches.

Methods

Data

The data handling follows the WeatherBench paper[19]. The data are freely available. Instruction for downloading can be found at `https://github.com/pangeo-data/WeatherBench`. The ERA5 data[9] was regridded bilinearly to 5.625° resolution using the xesmf Python package[31]. Data is available from 1979 to 2018, with the last two years reserved for testing/evaluation.

For the climate model pretraining, we downloaded a historical simulation from the CMIP6 archive[5]. 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 is six hours. The regridded climate model data are also available on the WeatherBench data repository.

Verification

In this study we use two skill metrics, the latitude-weighted root mean squared error (RMSE), 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:

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

and the anomaly correlation coefficient (ACC; see Section 7.6.4 of ref.[28]). The ACC compares the correlation of anomalies with respect to the respective climate, thereby performing an implicit bias correction. For all scores, the grid points are weighted by their area. We generally follow the WeatherBench methodology[19].

Baselines

WeatherBench contains three physically-based baselines: 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 10 km horizontal resolution with 137 vertical levels; and the same model run at two lower resolutions, T42 (approximately 2.8° or 310 km at the equator) with 62 vertical levels and T63 (approximately 1.9° or 210 km at the equator) with 137 vertical levels. Computationally, a single ten day forecast with the operational IFS models takes roughly one hour of real time on a cluster with 11,664 cores. The T42 and T63 models take 270 s and 503 s and a single XC40 node with 36 cores. For details on the initialization of each model, refer to the WeatherBench paper[19].

As an additional reference in Table 1 we include the work by Weyn et al.[27] who trained an neural network to predict Z500 and T850. Their model is iterative, i.e. 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. They trained only on ERA data.

Data-driven models

All models in this study use the same architecture (except in the network size scaling experiments). The basic structure is a fully convolutional Resnet[8] with 19 residual blocks. Each residual block consists of two convolutional blocks, defined as [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. 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-meter temperature, 6-hourly 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 a 32×64 fields which contains the forecast time in hours divided by 100. During training, a random forecast time from 6 to 120 hours is drawn for each sample. For the output layer a simple 2D convolution is used with the number of output channels either being three for the networks that predict Z500, T850 and T2M or one for the PR networks. Zero padding is used in the latitude direction. LeakyReLU is used with $\alpha = 0.3$. Weight decay of 1×10^{-5} is used for all layers. Dropout of 0.1 is only used for the ERA only networks. For the CMIP only and pretrained networks the validation score was better without any dropout.

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[14] 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 5 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 5 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. For the direct approach we trained models for 6h, 1d, 3d and 5d forecast time. We used Tensorflow>=2.0. Training a single model takes around one day on a GTX 2080 GPU.

Data and code availability

All data are available at https://github.com/pangeo-data/WeatherBench. All code is available at https://github.com/raspstephan/WeatherBench.

Acknowledgements

We would like to thank Sebastian Scher and David Greenberg for their valuable comments on the paper as well as George Craig for discussing the saliency analysis. Stephan Rasp acknowledges funding from the German Research Foundation (DFG) under grant no. 426852073.

Author contributions

S.R. conceived the study, trained the models and analyzed results. All authors discussed results and wrote the manuscript.

References

[1] Brian Ancell and Gregory J. Hakim. Comparing Adjoint- and Ensemble-Sensitivity Analysis with Applications to Observation Targeting. *Monthly Weather Review*, 135:4117–4134, 2007. ISSN 0027-0644. doi: 10.1175/2007MWR1904.1.

[2] Peter Bauer, Alan Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55, 9 2015. ISSN 0028-0836. doi: 10.1038/nature14956. URL http://www.nature.com/doifinder/10.1038/nature14956.

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

[4] Imme Ebert-Uphoff and Kyle A. Hilburn. Evaluation, Tuning and Interpretation of Neural Networks for Meteorological Applications. 5 2020. URL http://arxiv.org/abs/2005.03126.

[5] Veronika Eyring, Sandrine Bony, Gerald A. Meehl, Catherine A. Senior, Bjorn Stevens, Ronald J. Stouffer, and Karl E. Taylor. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, 9(5): 1937–1958, 5 2016. ISSN 1991-9603. doi: 10.5194/gmd-9-1937-2016. URL https://www.geosci-model-dev.net/9/1937/2016/.

[6] Peter Grönquist, Chengyuan Yao, Tal Ben-Nun, Nikoli Dryden, Peter Dueben, Shigang Li, and Torsten Hoefler. Deep Learning for Post-Processing Ensemble Weather Forecasts. 5 2020. URL http://arxiv.org/abs/2005.08748.

[7] Reindert J. Haarsma, Malcolm J. Roberts, Pier Luigi Vidale, Catherine A. Senior, Alessio Bellucci, Qing Bao, Ping Chang, Susanna Corti, Neven S. Fučkar, Virginie Guemas, Jost von Hardenberg, Wilco Hazeleger, Chihiro Kodama, Torben Koenigk, L. Ruby Leung, Jian Lu, Jing-Jia Luo, Jiafu Mao, Matthew S. Mizielinski, Ryo Mizuta, Paulo Nobre, Masaki Satoh, Enrico Scoccimarro, Tido Semmler, Justin Small, and Jin-Song von Storch. High Resolution Model Intercomparison Project (HighResMIP v1.0) for CMIP6. *Geoscientific Model Development*, 9(11):4185–4208, 11 2016. ISSN 1991-9603. doi: 10.5194/gmd-9-4185-2016. URL https://gmd.copernicus.org/articles/9/4185/2016/.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. 12 2015. URL http://arxiv.org/abs/1512.03385.

[9] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna Chiara, Per Dahlgren, Dick Dee, Michail

Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The ERA5 Global Reanalysis. *Quarterly Journal of the Royal Meteorological Society*, page qj.3803, 5 2020. ISSN 0035-9009. doi: 10.1002/qj.3803. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803`.

[10] Tim D. Hewson and Fatima M. Pillosu. A new low-cost technique improves weather forecasts across the world. 3 2020. URL `http://arxiv.org/abs/2003.14397`.

[11] B. J. Hoskins, M. E. McIntyre, and A. W. Robertson. On the use and significance of isentropic potential vorticity maps. *Quarterly Journal of the Royal Meteorological Society*, 111(470): 877–946, 8 1985. ISSN 00359009. doi: 10.1002/qj.49711147002. URL `http://doi.wiley.com/10.1002/qj.49711147002`.

[12] Frédéric Hourdin, Thorsten Mauritsen, Andrew Gettelman, Jean-Christophe Golaz, Venkatramani Balaji, Qingyun Duan, Doris Folini, Duoying Ji, Daniel Klocke, Yun Qian, Florian Rauser, Catherine Rio, Lorenzo Tomassini, Masahiro Watanabe, and Daniel Williamson. The Art and Science of Climate Model Tuning. *Bulletin of the American Meteorological Society*, 98(3):589–602, 3 2017. ISSN 0003-0007. doi: 10.1175/BAMS-D-15-00135.1. URL `http://journals.ametsoc.org/doi/10.1175/BAMS-D-15-00135.1`.

[13] Eugenia Kalnay. *Atmospheric modeling, data assimilation, and predictability*, volume 54. 2003. ISBN 9780521791793. URL `http://books.google.com/books?hl=en&amp;lr=&amp;id=Uqc7zC7NULMC&amp;oi=fnd&amp;pg=PR11&amp;dq=Atmospheric+modeling,+data+assimilation+and+predictability&amp;ots=lI5gpir1RV&amp;sig=FuhXqkYSMxhz2jLI2T8144HX6fs`.

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

[15] Amy McGovern, Kimberly L. Elmore, David John Gagne, Sue Ellen Haupt, Christopher D. Karstens, Ryan Lagerquist, Travis Smith, John K. Williams, Amy McGovern, Kimberly L. Elmore, David John Gagne II, Sue Ellen Haupt, Christopher D. Karstens, Ryan Lagerquist, Travis Smith, and John K. Williams. Using Artificial Intelligence to Improve Real-Time Decision-Making for High-Impact Weather. *Bulletin of the American Meteorological Society*, 98(10):2073–2090, 10 2017. ISSN 0003-0007. doi: 10.1175/BAMS-D-16-0123.1. URL `http://journals.ametsoc.org/doi/10.1175/BAMS-D-16-0123.1`.

[16] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep Double Descent: Where Bigger Models and More Data Hurt. 12 2019. URL `http://arxiv.org/abs/1912.02292`.

[17] Tim Palmer. A Vision for Numerical Weather Prediction in 2030. 7 2020. URL `http://arxiv.org/abs/2007.04830`.

[18] Stephan Rasp and Sebastian Lerch. Neural Networks for Postprocessing Ensemble Weather Forecasts. *Monthly Weather Review*, 146(11):3885–3900, 11 2018. doi: 10.1175/MWR-D-18-0187.1. URL `http://journals.ametsoc.org/doi/10.1175/MWR-D-18-0187.1`.

[19] Stephan Rasp, Peter D. Dueben, Sebastian Scher, Jonathan A. Weyn, Soukayna Mouatadid, and Nils Thuerey. WeatherBench: A benchmark dataset for data-driven weather forecasting. 2 2020. URL `http://arxiv.org/abs/2002.00469`.

[20] Mark J. Rodwell, Linus Magnusson, Peter Bauer, Peter Bechtold, Massimo Bonavita, Carla Cardinali, Michail Diamantakis, Paul Earnshaw, Antonio Garcia-Mendez, Lars Isaksen, Erland Källén, Daniel Klocke, Philippe Lopez, Tony McNally, Anders Persson, Fernando Prates, and Nils Wedi. Characteristics of Occasional Poor Medium-Range Weather Forecasts for Europe. *Bulletin of the American Meteorological Society*, 94(9):1393–1405, 9 2013. ISSN 0003-0007. doi: 10.1175/BAMS-D-12-00099.1. URL `http://journals.ametsoc.org/doi/abs/10.1175/BAMS-D-12-00099.1`.

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

[22] Sebastian Scher and Gabriele Messori. Generalization properties of neural networks trained on Lorenzsystems. *Nonlinear Processes in Geophysics Discussions*, pages 1–19, 6 2019. ISSN 2198-5634. doi: 10.5194/npg-2019-23. URL `https://www.nonlin-processes-geophys-discuss.net/npg-2019-23/`.

[23] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *2nd International Conference on Learning Representations, ICLR 2014 - Workshop Track Proceedings*, 12 2013. URL `http://arxiv.org/abs/1312.6034`.

[24] Casper Kaae Sønderby, Lasse Espeholt, Jonathan Heek, Mostafa Dehghani, Avital Oliver, Tim Salimans, Shreya Agrawal, Jason Hickey, and Nal Kalchbrenner. MetNet: A Neural Weather Model for Precipitation Forecasting. 3 2020. URL `http://arxiv.org/abs/2003.12140`.

[25] Maxime Taillardat, Olivier Mestre, Michaël Zamo, and Philippe Naveau. Calibrated Ensemble Forecasts using Quantile Regression Forests and Ensemble Model Output Statistics. *Monthly Weather Review*, page 160301131220006, 3 2016. ISSN 0027-0644. doi: 10.1175/MWR-D-15-0260.1. URL `http://journals.ametsoc.org/doi/abs/10.1175/MWR-D-15-0260.1?af=R`.

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

[27] Jonathan A Weyn, Dale Richard Durran, and Rich Caruana. Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. 2020. doi: 10.1002/ESSOAR.10502543.1.

[28] Daniel S Wilks. *Statistical Methods in the Atmospheric Sciences*. Elsevier, 2006. ISBN 0127519661. URL `http://cds.cern.ch/record/992087`.

[29] WMO. WMO Lead Centre for Deterministic Forecast Verification, 2020. URL `https://apps.ecmwf.int/wmolcdnv/`.

[30] Fuqing Zhang, Naifang Bei, Richard Rotunno, Chris Snyder, and Craig C Epifanio. Mesoscale predictability of moist baroclinic waves: Convection-permitting experiments and multistage error growth dynamics. *Journal of the Atmospheric Sciences*, 64(10):3579–3594, 2007.

[31] Jiawei Zhuang. xESMF: v0.2.1, 10 2019. URL `https://xesmf.readthedocs.io/`.

Supplement

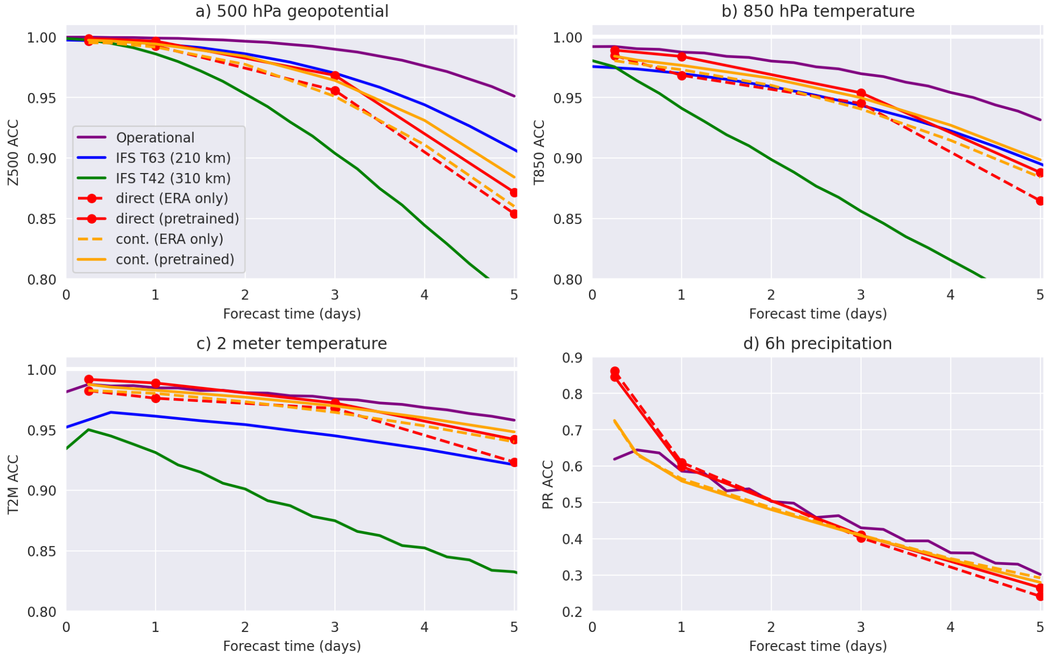

Figure S1: Anomaly correlation coefficient (ACC) for a) Z500, b) T850, c) T2M and d) PR evaluated against ERA5 data.

Table S1: ACC for 3 and 5 days forecast time. All forecasts evaluated at 5.625° resolution.

	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	
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

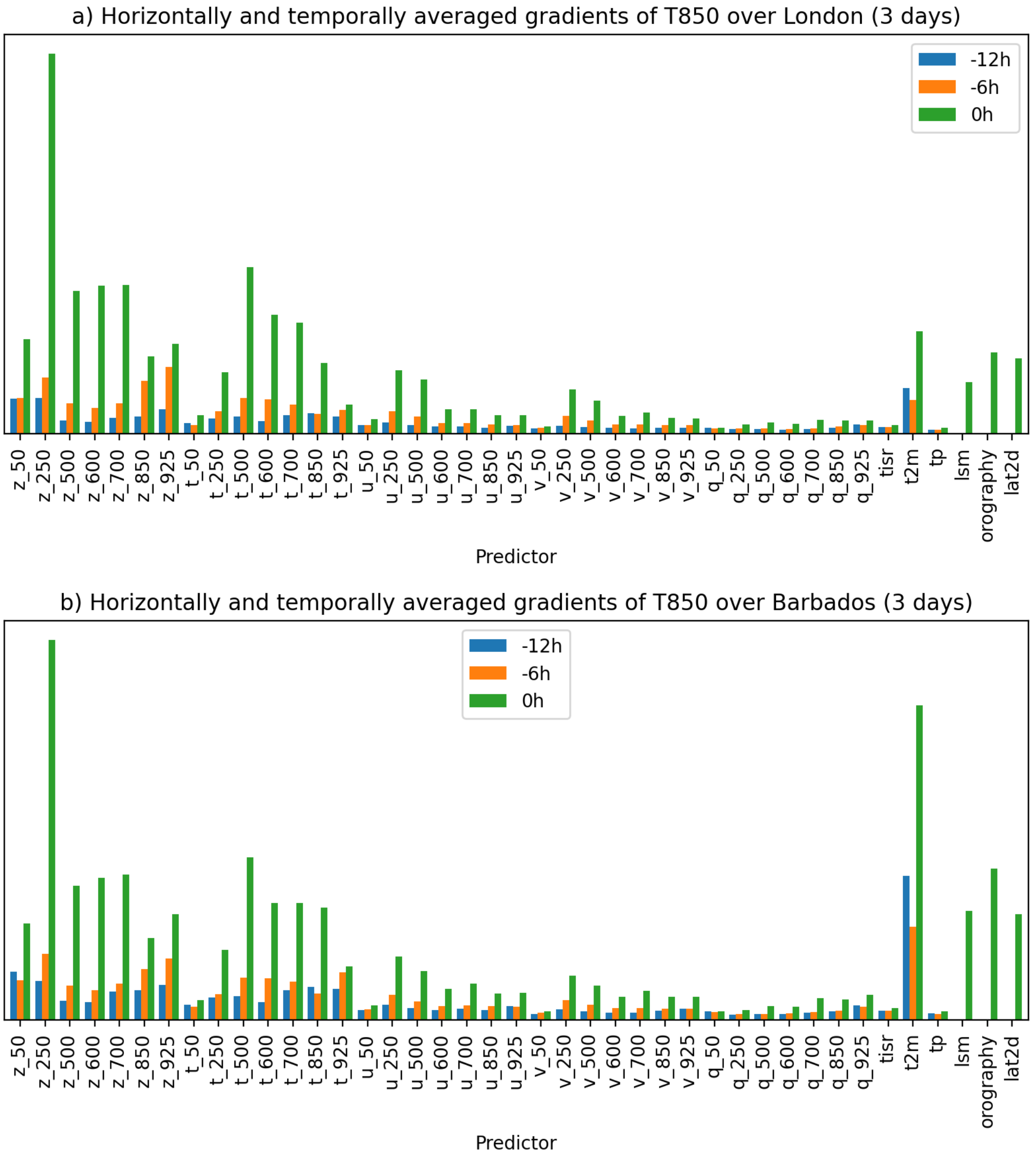

Figure S2: Horizontally averaged saliency $\overline{|G|}$ of T850 over a) London b) Barbados for 3 day direct forecasts. tisr is the top-of-atmosphere incoming solar radiation, tp is precipitation, lsm is the land sea mask.