

RESEARCH ARTICLE

Combining distribution-based neural networks to predict weather forecast probabilities

Mariana C.A. Clare¹  | Omar Jamil^{2,3}  | Cyril J. Morcrette² 

¹Department of Earth Science & Engineering, Imperial College London, London, UK

²Met Office, Exeter, UK

³University of Exeter, Exeter, UK

Correspondence

M.C.A. Clare, Department of Earth Science & Engineering, Imperial College London, London SW7 2AZ, UK.
Email: m.clare17@imperial.ac.uk

Funding information

UK Engineering and Physical Sciences Research Council, Grant/Award Numbers: EP/R512540/1, EP/L016613/1

Abstract

The success of deep learning techniques over the last decades has opened up a new avenue of research for weather forecasting. Here, we take the novel approach of using a neural network to predict full probability density functions at each point in space and time rather than a single output value, thus producing a probabilistic weather forecast. This enables the calculation of both uncertainty and skill metrics for the neural network predictions, and overcomes the common difficulty of inferring uncertainty from these predictions. This approach is data-driven and the neural network is trained on the WeatherBench dataset (processed ERA5 data) to forecast geopotential and temperature 3 and 5 days ahead. Data exploration leads to the identification of the most important input variables. In order to increase computational efficiency, several neural networks are trained on small subsets of these variables. The outputs are then combined through a stacked neural network, the first time such a technique has been applied to weather data. Our approach is found to be more accurate than some coarse numerical weather prediction models and as accurate as more complex alternative neural networks, with the added benefit of providing key probabilistic information necessary for making informed weather forecasts.

KEYWORDS

data exploration, deep learning, ensemble dropout, probabilistic weather forecasting, probability density functions, ResNet, stacked neural network

1 | INTRODUCTION

For over 100 years, advanced mathematical techniques have been used for weather prediction. Today, numerical weather prediction (NWP) is an advanced discipline which uses some of the world's largest supercomputers to solve complex nonlinear differential equations. The forecast skill of these models has been improving by approximately one day every ten years, that is, the 5-day forecast today is

as accurate as the 4-day forecast was ten years ago (Bauer *et al.*, 2015). This improvement has been achieved through the scientific and technological development of both NWP models and computers (Bauer *et al.*, 2015). However, the success of deep learning techniques over the last decade has opened up a new avenue for weather forecasting (Schultz *et al.*, 2021). Research has been mainly focused on the supervised learning techniques of neural networks (Dueben and Bauer, 2018; and Brenowitz and Bretherton,

2019; Rasp *et al.*, 2020; Rasp and Thuerey, 2020; Weyn *et al.*, 2020) and random forests (Yuval and O’Gorman, 2020). Some works have combined NWP models with neural networks: for example Brenowitz and Bretherton (2019) successfully couple neural networks with general circulation models to emulate physical parametrizations and Rasp and Lerch (2018) and Grönquist *et al.* (2021) use neural networks to post-process ensemble weather forecasts from NWP models. Other works have taken a purely data-driven approach (e.g., Rasp and Thuerey, 2020). In this work, we take the purely data-driven approach and use residual neural networks to create 3- and 5-day hindcasts.

A limitation of the deep learning approaches used in the works referred to above is that it is difficult to infer the uncertainty of the predictions from their results (Schultz *et al.*, 2021). Some previous works address these limitations by using an ensemble of deep learning models to produce a probabilistic forecast (e.g., Bihlo, 2021; Scher and Messori, 2021; Weyn *et al.*, 2021). However, choosing a good ensemble of models is non-trivial (Scher and Messori 2021) and may be computationally expensive because it requires the network to be trained multiple times. Others have dealt with the issue of uncertainty by training neural networks to predict from the weather data themselves the error and ensemble spread which would be produced if an NWP model were applied to this data, i.e., a mixed data-driven neural network and NWP approach (Scher and Messori, 2018). However, this method is not purely data-driven and requires access to good NWP forecasts.

In this work, we propose a novel approach to deal with the issue of assessing uncertainty from neural network outputs. With our approach, the neural networks predict full probability density functions for the target variable at each point in space and time instead of single values. These density functions allow practitioners to estimate the uncertainty of the neural network outputs and make a more informed weather forecast. In order to reduce computational cost and optimise model accuracy, we train multiple neural networks on a small number of variables and combine their outputs using techniques such as a stacked neural network. This is a technique which has not been used with weather data before. In this work, the neural networks are trained on the WeatherBench dataset created by Rasp *et al.* (2020) and used to predict both a 3-day and a 5-day weather hindcast of geopotential at the 500 hPa pressure level in m^2s^{-2} (hereafter Z500) and temperature at the 850 hPa pressure level in Kelvin (K) (hereafter T850). These variables are chosen so that our results can be compared to those in other works which use the same dataset (Rasp *et al.*, 2020; Rasp and Thuerey, 2020).

The remainder of this work is structured as follows: Section 2 describes the data used in this study followed by the neural network architectures and data exploration

techniques used: Section 3 presents results from using stacked neural networks to forecast weather data and shows how the output can be used to infer uncertainty; and finally Section 4 concludes the work.

2 | METHODS

2.1 | Data

The WeatherBench dataset is a global dataset produced by Rasp *et al.* (2020) containing a mix of multi-levelled (13 pressure levels) and single-level variables. It uses as its raw data the ERA5 reanalysis dataset (Hersbach *et al.*, 2020) for the 40-year period from 1979 to 2018. The data were processed and regridded onto a 5.625° resolution latitude–longitude grid (32×64 grid points) by Rasp *et al.* (2020) and we refer the reader to that work for more details. Following the same work, we consider data from 2017 to 2018 to be the test dataset. One of the benefits of deep learning is that we do not need to carry out extensive feature engineering and the neural networks are able to find the best predictors in the data. However, it is still necessary, as a first step, to choose an appropriate architecture for the neural network. For this first step, we use data from 1979 to 2015 as the training dataset with data from 2015 used for validation of the neural network (hereafter referred to as the neural-validation dataset). All the results in this section are applied on the 2016 data (hereafter referred to as the validation dataset), so that results on the test dataset are not used to make any architecture decisions.

2.2 | Neural network architectures

Fundamentally, a neural network provides a way to extract nonlinear relationships present in the data and is trained to minimise a loss. This minimisation is done via gradient descent which is used to update the neural network weights. Previous works have applied different types of neural network to this dataset: the original WeatherBench dataset work Rasp *et al.* (2020) uses a simple convolutional neural network (CNN) and Rasp and Thuerey (2020) uses a 19-block convolutional ResNet (an architecture first described by He *et al.*, 2016). Additionally, Weyn *et al.* (2020), who do not use the WeatherBench dataset but use comparable data, use a U-Net (an architecture first described by Ronneberger *et al.*, 2015). The lowest errors are obtained when using a ResNet and thus in this work we choose this architecture.

Generally, residual neural networks consist of a series of repeated blocks (referred to as residual blocks) of

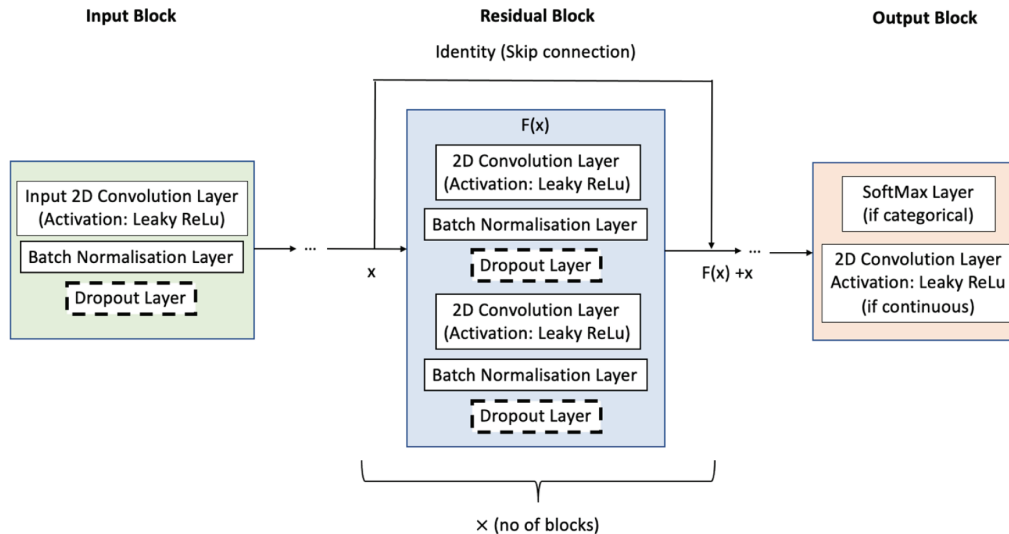


FIGURE 1 Schematic of convolutional ResNet used in this work. Each 2D convolution layer has LeakyReLU activation with $\alpha = 0.3$ (Maas *et al.*, 2013) and 100 channels (because of 100 bins) if categorical data is being trained or 64 channels if continuous data is being trained. Following Rasp *et al.* (2020), the 2D convolutions are defined with periodic padding in the longitude direction and zero padding in the latitude direction with a kernel size of 5. The dropout layer has a dropout rate of 0.1. If dropout is not required, then there is no dropout layer in the network architecture. Note that each ResNet takes between approximately 6 and 12 hr to train (depending on the size of the input data and the number of residual blocks) on a RTX6000 machine with two GPUs and 48 GB of memory [Colour figure can be viewed at wileyonlinelibrary.com]

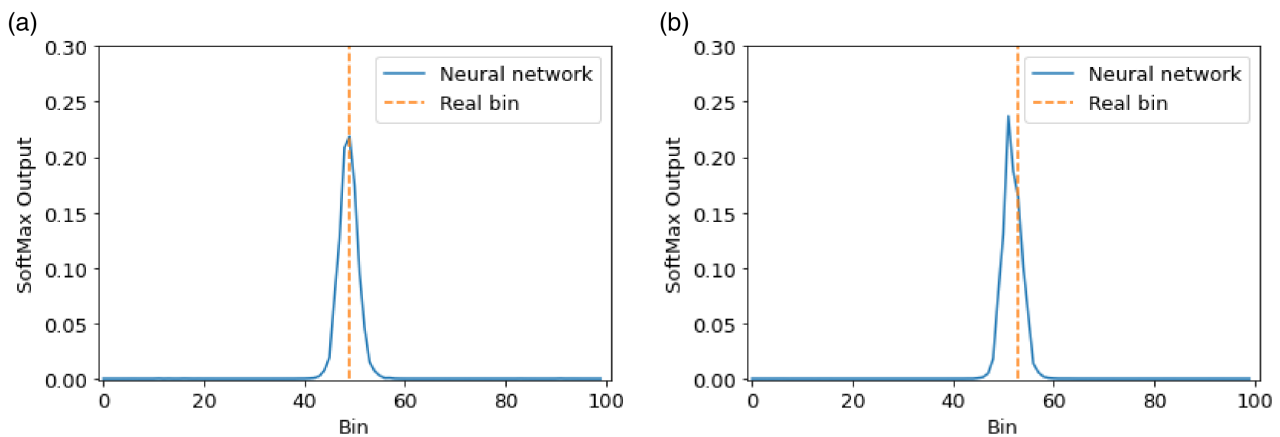


FIGURE 2 Two randomly selected examples of the probability density function for the Z500 3-day hindcast at different gridpoints and times. These have been predicted by a ResNet with a SoftMax output layer and 15 residual blocks. In (a) the maximum probability bin is the true bin; in (b) the maximum probability bin (23.7%) is not the true bin, which has a non-zero probability of 16.3%. Each bin corresponds to a geopotential range of width $169 \text{ m}^2 \text{ s}^{-2}$, where the lower bound of the 0-bin is $42,500 \text{ m}^2 \text{ s}^{-2}$ and the upper bound of the 99-bin is $59,300 \text{ m}^2 \text{ s}^{-2}$ [Colour figure can be viewed at wileyonlinelibrary.com]

convolutional, normalisation, and dropout layers, with intermediate connections known as “skip connections”. A skip connection between residual blocks adds the outputs from previous blocks to the output of the current block (Figure 1). In this way they can avoid the issue of accuracy saturation which occurs in other types of neural network architectures when more layers are added (He *et al.*, 2016). Figure 1 shows the general structure of the convolutional ResNet used in this work and provides details of the layers used. It highlights that, as discussed in Lu *et al.* (2018), a

ResNet can be viewed as a variation of the forward Euler finite-difference method

$$y_{n+1} = y_n + hf(t_n, y_n), \quad (1)$$

which is actually a much simpler version of the methods used by NWP models to predict the weather (White, 1971). This makes a ResNet an appropriate architecture for predicting weather and partly explains the greater accuracy given by the convolutional ResNet

in Rasp and Thuerey (2020) relative to other neural networks.

The main novelty of our work is that our neural network architecture aims to predict the probability distribution of the variables Z500 or T850 at a particular point in time, longitude and latitude, rather than their exact value which is what the previous works discussed above predicted. In order to achieve this, the first step is to convert the continuous weather data to categorical data by taking each target variable and binning its values into 100 bins of equal width. The values of Z500 vary between a minimum of $42,500 \text{ m}^2\text{s}^{-2}$ and a maximum of $59,300 \text{ m}^2\text{s}^{-2}$, meaning that each Z500 bin has a width of $169 \text{ m}^2\text{s}^{-2}$ (to 3 significant figures); T850 varies between 213 K and 314 K, leading to a bin width of 1.02 K (to 3 significant figures). We take the value of the category to be its lowest value, which introduces an inbuilt root mean square error (RMSE) of $91.2 \text{ m}^2\text{s}^{-2}$ for Z500 data and 0.992 K for T850 data (calculated using Equation (3)). With the target variables binned, it is then possible to use a SoftMax layer as the output layer (Figure 1). The SoftMax layer predicts the probability density functions of the variables by using an activation function which exponentiates the input and then normalises it, thus outputting a vector with a value between 0 and 1 for each bin (Goodfellow *et al.*, 2016). The sum of this vector is equal to 1, meaning it is a probability density. Figure 2 shows two randomly chosen examples of the output from the SoftMax layer for a categorical output variable. In Figure 2a, the maximum probability bin is the true bin. Although in Figure 2b the maximum probability bin (23.7%) is not the true bin, the true bin has a non-zero probability of 16.3%, meaning this probability density function is a useful tool from a weather forecasting perspective.

Our use of categorical data also necessitates a different choice of loss metric to the previous works using the WeatherBench dataset (e.g., Rasp and Thuerey, 2020), which use the mean squared error. A loss metric is used by the neural network to calculate the loss during training and it can play a pivotal role in the accuracy and efficiency of a neural network. We choose sparse categorical cross-entropy (from Keras) in our neural network due to our use of categorical data and the memory efficiency of this metric. Note that these loss metrics can also be used during the training of the neural network to determine when to stop training and set the values of important parameters. For example, in our case, we compile our neural network with Adam optimiser (Kingma and Ba, 2014) with an initial learning rate of 5×10^{-5} . This learning rate is reduced by a factor of 5 if the loss metric on the neural-validation dataset does not decrease after two epochs (i.e., after the entire training dataset has passed through the neural network twice). If the loss metric on

the neural-validation dataset does not decrease after five epochs, then the neural network stops training. In general, the neural network requires approximately 15 epochs of training, but this varies depending on the set of inputs being trained.

In addition to a new loss metric, in order to be able to compare the results of our new categorical data approach with the correct values from the data and with other neural network and numerical model results, it is necessary to infer a single value from our categorical neural network predictions. To do this, we again take advantage of the probability density function we have predicted and calculate their expectation using

$$\mathbb{E}[X] = \sum_{i=1}^{100} x_i \mathbb{P}(X = x_i), \quad (2)$$

where x_i is chosen to be the lower bound of each bin. In this way, we take advantage of the density functions we have predicted and also reduce the inbuilt error caused by binning the data, which we discussed previously. Throughout this work, the RMSE is calculated between the real data and the expected values of the density functions generated by our approach using the latitude-weighted RMSE outlined in (e.g., Rasp *et al.*, 2020). This is given by

RMSE

$$= \sqrt{\frac{1}{N_{\text{timepoints}}} \sum_{i=1}^{N_{\text{timepoints}}} \frac{1}{N_{\text{lat}} N_{\text{lon}}} \sum_{j=1}^{N_{\text{lat}}} \sum_{k=1}^{N_{\text{lon}}} L(j) (f_{i,j,k} - t_{i,j,k})^2}, \quad (3)$$

where

$$L(j) = \frac{\cos\{\text{lat}(j)\}}{\frac{1}{N_{\text{lat}}} \sum_{j=0}^{N_{\text{lat}}} \cos\{\text{lat}(j)\}}, \quad (4)$$

is the latitude weighting factor, f is the predicted value from the neural network and t is the true value from the dataset.

Despite this expectation calculation, there will still be differences caused by using categorical output data compared to continuous data. The simplest way to understand these is by training a series of continuous data and categorical data neural networks with varying numbers of residual blocks and comparing the results. Note that to reduce both the computational cost of training the network and the memory cost, we train two separate networks to predict Z500 and T850 individually. If we had trained a single neural network to predict both Z500 and T850 at the same time, this would have added an extra dimension to the output making it 5-dimensional. We note that, unless

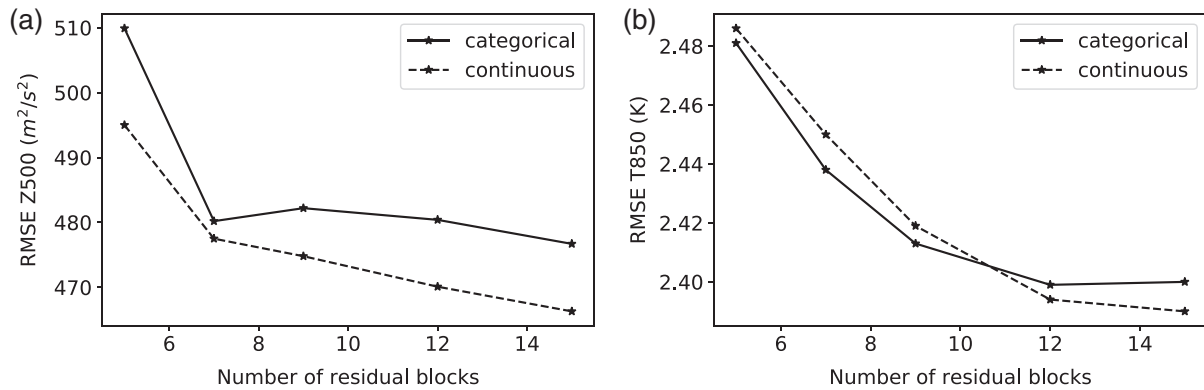


FIGURE 3 Comparison of the RMSE achieved by training a neural network with categorical data and that achieved by training a neural network with continuous data, for (a) 500 hPa geopotential, and (b) 850 hPa temperature, using different numbers of residual blocks. The RMSE is a weighted average calculated using Equation (3) over all gridpoints and times in the validation dataset

explicitly stated, all results shown in this work refer to the 3-day hindcast.

For our initial analysis, we use a training dataset consisting of only the two variables of interest, Z500 and T850 at current time t . Figure 1 and its caption show the neural network architectures used for training continuous data (the same as that in Rasp and Thuerey (2020)) and categorical data. Note that, for simplicity, initially we have chosen not to include dropout in either neural network type.

Figure 3 shows that the errors decrease as the number of residual blocks increases for both categorical and continuous data. For Z500, Figure 3a shows that the error from training on continuous data is always less than that from training on categorical data, but the difference between the two errors is much less than the inbuilt binning error of $91.2 \text{ m}^2\text{s}^{-2}$. For T850, Figure 3b shows that there is little difference in the error as a result of training on categorical data compared to the error from training on continuous data and in fact for less than 12 residual blocks the categorical data error is lower. This is despite the fact that the inbuilt binning error calculated previously is 0.992 K. This suggests that, although working in terms of binned data introduces an error, the new neural network structure and the expectation calculation is able to partially compensate for this.

2.2.1 | Using dropout for ensemble modelling

So far the neural network architectures used in our tests have not included a dropout layer. However, including a dropout layer is a common strategy to improve the performance of neural networks (e.g., Srivastava *et al.*, 2014). This layer randomly ignores some outputs from the preceding layer in the network at a rate set by the user

(we use a rate of 0.1), meaning these outputs are not passed on to the preceding layer of the network. This means that, if dropout is occurring, the neural network is slightly different every time the data pass through it, which helps prevent overfitting during training. Moreover, if dropout is allowed to occur at the inference/prediction phase, then an ensemble of outputs can be generated from a single neural network train (Gal and Ghahramani, 2016). Thus in this section, we examine the improvements that can be achieved from using dropout in our neural network.

Dropout ensemble techniques have already been applied on continuous weather data in Scher and Messori (2021), where they show that using this technique results in an improvement in accuracy. With our categorical data approach, each ensemble member is actually a series of density functions at every point in space and time (recall Figure 2) and so careful analysis is required before they are combined. The law of total probability states that

$$\mathbb{P}(A) = \sum_n \mathbb{P}(A|B_i) \mathbb{P}(B_i), \quad (5)$$

if B_i are a finite number of pairwise disjoint sets, whose union is the sample space. In addition, with this technique there is no reason why one ensemble member should be more accurate than the others, and thus a simple average is sufficient to combine them. If we set $\mathbb{P}(A|B_i)$ to be the probability density function from the i th ensemble member and $\mathbb{P}(B_i)$ to be the probability of sampling from the i th distribution (thus $\mathbb{P}(B_i) = 1/n$ where n is total number of ensemble members), we show that averaging the density functions from the ensemble members is mathematically rigorous. This averaging is known as linear pooling (Allard *et al.*, 2012).

Using this knowledge, we can average the ensemble outputs. Figure 4 shows the results of averaging an

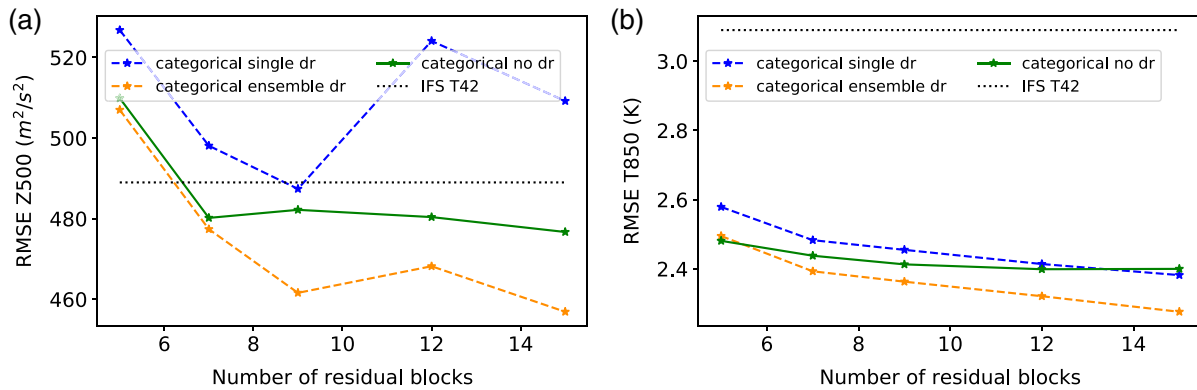


FIGURE 4 Comparison of the RMSE achieved by using the dropout ensemble technique with 32 ensemble members; using a single output from a neural network with dropout; using a neural network with no dropout; and using the coarse NWP model IFS T42. (a) is for 500 hPa geopotential, and (b) 850 hPa temperature. The ensemble dropout technique almost always outperforms the other techniques. Note that the RMSE of the neural network is calculated using the weighted average (Equation (3)) over all gridpoints and times in the validation dataset (i.e., 2016), whereas the IFS T42 RMSE is calculated using the same method but over the test dataset (i.e., 2017–2018) [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.com)]

ensemble of 32 members and shows that using the dropout ensemble technique results in a notable reduction in error compared to not using dropout at all or using dropout but with only one ensemble member. This reduction in error as a result of the dropout ensemble technique is obtained for almost no extra computational cost and occurs no matter the number of residual blocks used. Whilst for T850 the error always decreases as the number of residual blocks increases, for Z500 the trend is less clear. This may be due to overfitting and the optimum number of residual blocks will be discussed in more detail in Section 3. For reference, Figure 4 also shows the error from using the configuration of the Integrated Forecast System (IFS) model of the European Centre for Medium-range Weather Forecasting (ECMWF) at the T42 resolution (approximately 2.8° resolution at the Equator) [IFS T42]. The T42 resolution is much coarser than the operational IFS used by ECMWF, but twice as fine as the resolution of the WeatherBench dataset (5.625°) used in this work. Despite this finer resolution, Figure 4 shows that our ResNet with categorical data outputs trained on just Z500 and T850 is substantially more accurate than IFS T42 for both Z500 and T850 when the dropout ensemble technique is used with more than five residual blocks. Note that this IFS T42 value is for the 2017–2018 data and is taken from Rasp and Thuerey (2020), whereas the neural network errors are for the 2016 data, but we have also used the neural networks to predict the 2017–2018 data and found the same results. We have not included this here so as not to mislead the reader by showing preliminary neural network results applied to test data.

We also calculate the ensemble spread of the ensemble of Z500 and T850 outputs created using the dropout-

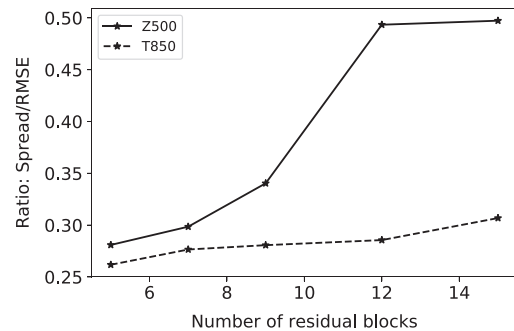


FIGURE 5 Ratio between the ensemble spread and the RMSE for the Z500 and T850 hindcasts. Note that, for a “perfect” ensemble, the ratio would be equal to 1

at-inference technique. This is calculated using

$$\text{ensemble spread} = \sqrt{\mathbb{V}(y_{\text{ens}})}, \quad (6)$$

and is a measure of the uncertainty of the ensemble: if the ensemble is “perfect” then the ensemble spread should be equal to the RMSE (Palmer *et al.*, 2006). Figure 5 shows the ratio between the ensemble spread for Z500 and T850 for the 3-day hindcast. Note in order to compare the spread and the RMSE directly, we have used the same weighted average from Equation (3) on the spread to change it into a single value. Whilst our ensemble is not perfect, the ratio between the spread and the error is similar to that shown in Figure 3 of Scher and Messori (2021), when they use the dropout-at-inference technique to create an ensemble forecast of Z500. This shows that the use of the dropout ensemble technique here is appropriate and concludes our outline of the neural network architecture used in this work, which is summarised in schematic form in Figure 1.

2.2.2 | Neural network stacking

Ideally, we would now train the neural network (outlined above and shown in Figure 1) on the WeatherBench dataset and predict Z500 and T850. However, the WeatherBench dataset is over 300 GB and thus it is non-trivial and potentially unnecessary to train the neural network on the entire dataset at any one time. There are several methods to deal with this issue and in this work we focus on two: (a) using a meta-learner to combine the outputs from several neural networks trained individually, and (b) using data exploration techniques to identify the important variables in the dataset and discard the unimportant ones. In this section, we focus on constructing the meta-learner and then in Section 2.3 we focus on the data exploration.

In order to use a meta-learner, we set-up a series of ResNets (with the categorical data architecture in Figure 1) trained on smaller input datasets which always include Z500 and T850 as well as one other variable, all at current time t . This set-up is summarised on the left-hand side of the schematic in Figure 6 and shows the other variables which make up the input datasets. (Note that these variables are chosen following extensive analysis later in Section 2.3.) The output of each of these networks is a density function at each point in space and time. To improve accuracy, for each network we use dropout at inference to create an ensemble of 32 outputs and use the law of total probability (Equation (5)) to average the outputs meaning that the output of each individual neural network run is a single density function. Note that each individual ResNet takes approximately 12 hr to train on a RTX6000 machine with two GPUs and 48 GB of memory – substantially less computational time and memory than would be required if only one neural network were used.

The outputs of these individual networks can now be combined to a single output for each point in time and space. There are several different methods to do this including linear pooling with average weights as done with the ensemble created by the dropout. Whilst linear pooling with average weights is sufficient for the dropout ensemble (because there is no reason why one ensemble member should be weighted higher than another), it is reasonable to assume some input variables are more important than others in determining the final results and thus should have a greater weighting (Figures 9 and 10 below). There are various specific techniques to combine outputs such as pooling, voting and stacking (Zhou, 2012), as well as other simpler techniques such as linear regression. We choose to combine our outputs by using the learning technique of stacking (Wolpert, 1992; Smyth and Wolpert, 1999) where the meta-learner is a stacked neural

network used to combine individual learners (our individual ResNets). We make this choice because it is simple to implement and computationally cheap: the stacked neural network used in this work takes only 30 min to train on a RTX6000 machine with two GPUs and 48 GB of memory. Moreover, combining distributions using techniques such as linear regression will almost definitely result in weightings which do not satisfy the law of total probability (Equation (5)) and thus the combined output will not be a distribution. The advantage of using a stacked neural network approach is that, by using a SoftMax layer as our output layer, we ensure that the combined output predicted by the stacked neural network is mathematically a distribution. This means that the inputs to our stacked neural network do not need to be distributions and thus to reduce memory and computational cost we transform the density functions from the individual neural networks to expectations using Equation (2) before inputting them into the stacked neural network. This reduces the size of the input data into the stacked neural network by a factor of 100 due to there being 100 bins. Figure 6 provides a summary, showing how the outputs from individual ResNets are combined using a stacked neural network. For the stacked neural network, we use a simple shallow network with the following architecture: an input layer to concatenate the output of the individual neural networks, two hidden layers with 36 nodes each and Rectified Linear Unit (ReLU) activations. A SoftMax output layer is used to generate a probability density function (as discussed above).

2.3 | Data exploration and feature selection

In the previous section, we outline the full neural network architecture used in this work (summarised in Figures 1 and 6). In a purely data-driven scenario, computational resources permitting, all possible data inputs are used to get the best possible performance out of a neural network. The neural network architecture and hyperparameters are then tuned to extract optimum predictive power out of the available data. An alternative to this approach is to pre-select the key features from the data to train the neural networks. Such feature selection is a two-fold problem: (i) how many input variables, and (ii) how many individual data points should be included. There is always a trade-off between having a large amount of data resulting in large computational and memory costs, and having too little data resulting in overfitting.

As a first step, we make some physically informed choices to include only certain variables in the potential input. Recall that the WeatherBench dataset contains a

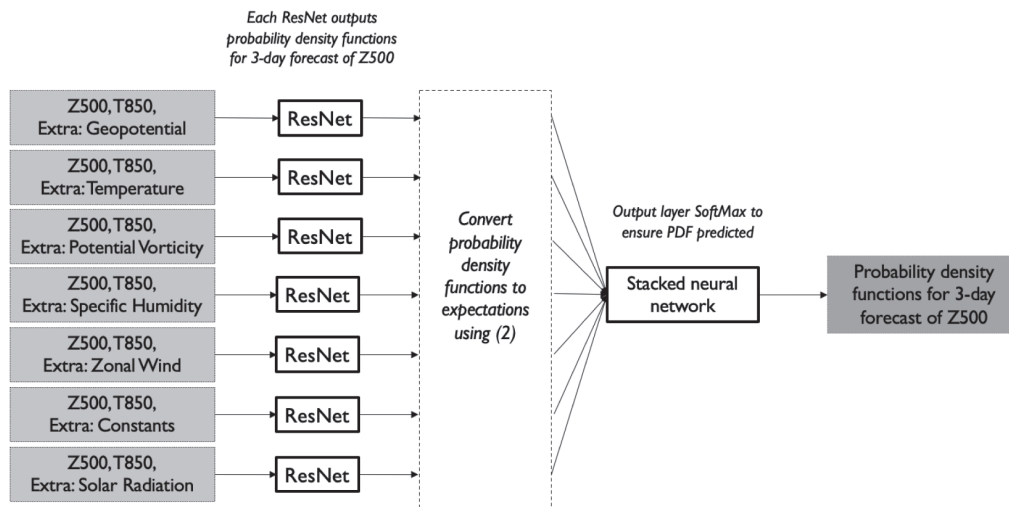


FIGURE 6 Schematic for Z500 3-day hindcast using the stacked neural network approach to combine outputs from individual ResNets (see Figure 1 for their schematic). The schematics for the Z500 5-day, T850 3-day and T850 5-day hindcasts are the same except that there is no solar radiation ResNet for the T850 hindcasts. Note that each ResNet used in this set-up takes approximately 12 hr to train on a RTX6000 machine with two GPUs and 48 GB of memory and the stacked neural network takes 30 min on the same machine

mix of multi-level variables and single-level variables. We choose to include the multi-level variables of geopotential and temperature as these are the variables we are forecasting, and zonal wind because of its links to geopotential. We also choose to include a humidity multi-levelled variable and a vorticity multi-levelled variable. The Weather-Bench dataset includes both specific humidity and relative humidity, and both potential vorticity and relative vorticity. From a physical perspective, we would expect little gain from using relative humidity as this variable is a function of specific humidity, temperature and pressure (which is itself related to geopotential), which are already present in the potential input dataset. Thus we choose specific humidity. Similarly, we exclude relative vorticity and keep potential vorticity because relative vorticity describes only the rotational component of the horizontal flow, whereas potential vorticity also includes a contribution from the vertical stratification of the temperature field and is known to be a conserved variable in adiabatic flow. In addition, potential vorticity is often used in the study of the development of midlatitude weather systems and as a result is suitable for our potential input dataset. Finally, we also include: the single-level variables of solar radiation because it is the energy source driving the system, 2 m temperature because of its likely influence on T850, and the constant fields of orography, land-sea mask, and latitude.

Thus, as a result of this initial physically informed analysis, we have been able to exclude 45 variables from the training dataset out of a total of 115 (where for data on multiple levels, we are counting each individual level as a variable). The remaining set of relevant

input variables consists of the multi-level fields of geopotential, temperature, specific humidity, potential vorticity and zonal wind (x -direction) and the single-level fields of 2 m temperature, top-of-atmosphere incoming solar radiation and constants (the three variables constant in time – orography, land-sea mask, and latitude). To further refine the set of relevant variables, we conduct data exploration using a purely data-driven approach.

Our data-driven approach can be thought of as a way to create an optimal subset of data for training. In order to do such feature selection, we first train an individual neural network with an input dataset of just Z500 and T850, and define this as our “benchmark” training dataset, which will be used to understand the effect of including other variables on the final error. The architecture of this network is a ResNet (Figure 1) with five residual blocks. We use the dropout at inference technique, described in Section 2.2.1, to extract an ensemble of 32 outputs from the single trained model, and then calculate the error of the ensemble mean (Section 2.2.1). By using this, we average out some of the randomness of the results. We then train individual neural networks with an input dataset of Z500 and T850 and one variable from the relevant list, and compare it to the original benchmark. It should be noted that this relatively small neural network is just a way to assess the impact of additional variables being included in the training data. We then use these selected features to train deeper neural networks to find the optimum architecture for these data. We have chosen to focus on improving the 3-day hindcast for this comparison and assume that any improvements in methodology for this will also result in improvements in the 5-day hindcast. Thus,

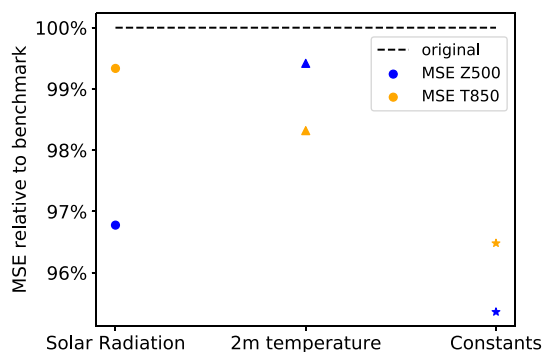


FIGURE 7 Percentage error relative to benchmark from training using extra single-level variables. These are: just incoming top-of-atmosphere solar radiation, just 2 m temperature, or the three constants of orography, land-sea mask and latitude. Results are calculated from the mean of 32 ensemble members generated using dropout at inference, predicting on all gridpoints and times in the validation dataset (i.e., 2016) [Colour figure can be viewed at wileyonlinelibrary.com]

unless explicitly stated, all results shown are for the 3-day hindcast.

In this section, we show an example of this data-driven analysis for the variables defined at a single height: 2 m temperature, total incident solar radiation and constants. For the multi-levelled variables, we use the same methodology to determine which levels are the most important for predicting Z500 and T850 and have included this analysis in Appendix A.

Figure 7 shows the ensemble errors as a percentage of the error relative to the benchmark dataset for the single-level variables. The figure shows that solar radiation is relatively unimportant for predicting T850 and relatively important for predicting Z500, that 2 m temperature is relatively unimportant for predicting Z500 and T850 and finally constants are important for predicting both Z500 and T850. Thus from Figure 7, we conclude that we can exclude 2 m temperature from both training datasets and solar radiation from the T850 training dataset. Note all relative errors in this figure are relative mean squared errors (MSE) so that these results can be compared with those in Figure A1 in Appendix A which use MSE to make calculating confidence intervals easier.

As a result of the physically informed and data-driven analysis in this section and in Appendix A, we have determined a set of input variables which are important in determining Z500 and T850. From an original dataset of 115 variables, we have been able to exclude 45 through an initial physically informed analysis and a further 36 for Z500 and 37 for T850 through data-driven analysis. We use this input variable selection in the next section to reduce computational and memory costs without compromising too much on accuracy.

3 | RESULTS

In the previous section, we outlined our new methodology for predicting the weather forecast using a data-driven approach. In this section, we show the results of implementing this methodology. Because we are now applying the full stacked neural network approach, we must split the dataset in a different way from that in Section 2 as there are now two stages of neural networks. We use the data from 1979 to 2011 as the training dataset for the individual ResNet learners, and use the data from 2011 as a validation dataset. We then use the individual learners to make predictions on the dataset from 2012 to 2016 (hereafter referred to as the stacked validation dataset). These predictions are the inputs for the meta-learner (the stacked neural network), which uses shuffle as validation. As in Rasp and Thuerey (2020), the final testing dataset is the data from 2017 to 2018. As discussed previously, all the analysis so far has been carried out for the 3-day weather hindcast, but we show that this methodology also leads to good results for the 5-day weather hindcast.

As a first step, we determine the optimum number of residual blocks to use in the individual ResNets which feed the stacked neural network. (Recall that Figures 3 and 4 in the previous sections show that the accuracy of a ResNet is very dependent on the number of residual blocks used.) Figures 8a and 8b show how RMSE Z500 and RMSE T850 vary as the number of residual blocks is increased by steps of four blocks from 5 to 25 blocks and steps of two blocks from 25 to 31. The error is calculated on the 2012 to 2016 dataset, because these are the predictions which are used as the input for the stacked neural network. For the main analysis of the optimum number of blocks, we use the optimum temperature levels as the training dataset for Z500 and the optimum geopotential levels as the training dataset for T850. Before 25 residual blocks, in both cases the error decreases relatively steadily as the number of residual blocks increases. However, beyond this number the error sometimes increases dramatically when more residual blocks are added and sometimes decreases dramatically. Thus we also conduct a small sensitivity analysis using the optimum levels of potential vorticity, specific humidity and zonal wind. Figure 8a shows the error is minimised with 29 residual blocks for temperature, potential vorticity and zonal wind, but for specific humidity the error increases notably when 29 blocks are used. Given that the error using specific humidity is always higher than that from using the other variables, the contribution from the specific humidity variable after the stacked neural network is applied will be small. Thus it is reasonable to ignore that the error from specific humidity is high when 29 blocks are used and conclude that 29 is a good number of residual blocks to use for Z500. Figure 8b

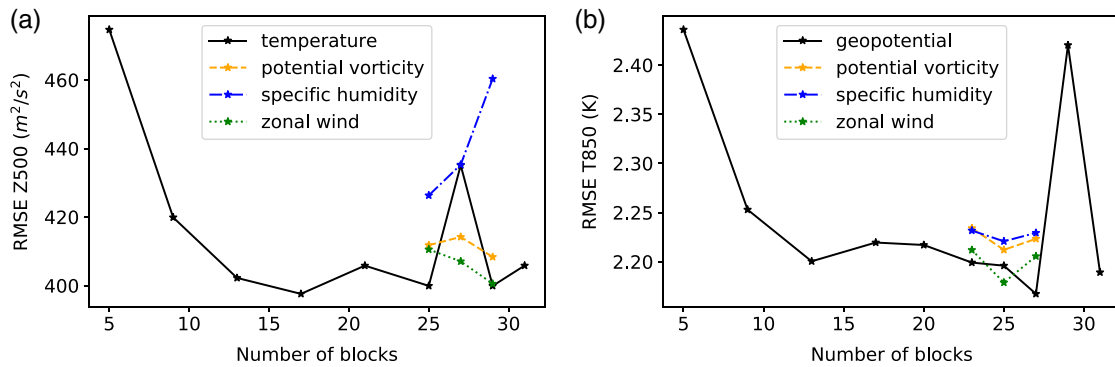


FIGURE 8 RMSE as a result of using different numbers of residual blocks in the ResNet for (a) 500 hPa geopotential, and (b) 850 hPa temperature. The effect on the error is shown for ResNets trained on a number of different variables. The RMSE is calculated using the weighted average (Equation (3)) over all gridpoints and times in the stacked validation dataset (i.e., 2012–2016) [Colour figure can be viewed at wileyonlinelibrary.com]

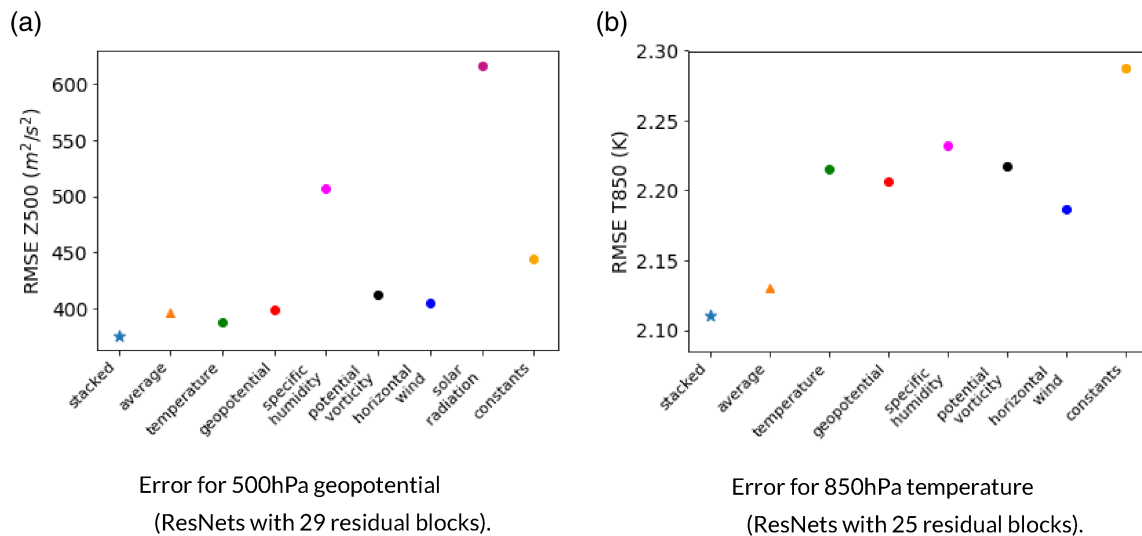


FIGURE 9 Improvement in accuracy for the 3-day hindcast for (a) 500 hPa geopotential, and (b) 850 hPa temperature, as a result of using a stacked neural network compared to the individual ResNets trained on specific variables and simply averaging the outputs of the individual ResNets. The RMSE is a weighted average calculated using Equation (3) over all gridpoints and times in the test dataset (i.e., 2017–2018) [Colour figure can be viewed at wileyonlinelibrary.com]

shows the error is minimised with 25 blocks for potential vorticity, specific humidity and zonal wind and close to being minimised by this number of blocks for geopotential. Thus, we conclude 25 is a good number of residual blocks to use for T850.

Using these optimum numbers of residual blocks, we can now calculate our full neural network results. Figures 9a and 9b compare, for Z500 and T850 respectively, the 3-day prediction error on the test dataset from using each of the individual learners trained on a specific variable, simply averaging the output from the individual ResNets and using the stacked neural network to combine the outputs. Figures 10a and 10b show the same comparison for the 5-day prediction error on the test dataset. These figures show that, for both the 3-day and

5-day hindcast for both Z500 and T850, the stacked neural network always provides the most accurate result and the error is notably reduced through using it. This is to be expected because the stacked neural network framework (Figure 6) means that the RMSE achieved from using it is loosely bounded from above by the lowest RMSE achieved by the individual ResNets; otherwise the stacked neural network could achieve a lower error simply by setting that input to 1 and the other inputs to 0. (Note this is only a loose bound because the training dataset of the stacked neural network is not the same as the test dataset on which the predictions are made.) Moreover, the stacked neural network is also more accurate than simply averaging the expectation outputs from the individual ResNets. This is because the stacked neural network learns in training to

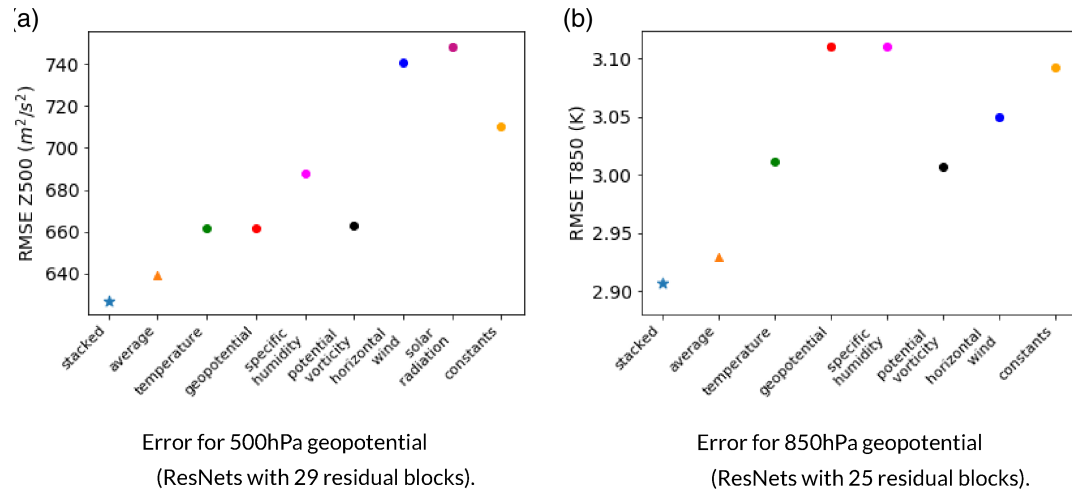


FIGURE 10 Improvement in accuracy for the 5-day hindcast for (a) 500 hPa geopotential, and (b) 850 hPa temperature, as a result of using a stacked neural network compared to the individual ResNets trained on specific variables and simply averaging the outputs of the individual ResNets. The RMSE is a weighted average calculated using Equation (3) over all gridpoints and times in the test dataset (i.e., 2017–2018) [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com)]

give a lower weight to the individual ResNets with larger errors, meaning the overall error from the stacked network is unaffected. These accuracy improvements, in addition to the comparatively computationally cheap nature of the stacked neural network (30 min on a RTX6000 machine compared to 12 hr to train each individual ResNet on the same machine), show that we are correct to apply the stacked neural network to the ResNet outputs.

To visualise what these error values mean, we look at the specific event of storm *Ophelia* which was an active storm between 8 and 18 October 2017, and the worst storm to affect Ireland in over 50 years (Stewart, 2018). We calculate the deviations of the true value, the 3-day and the 5-day hindcasts from the global annual climatology value and show the results in Figure 11 in the North Atlantic region at 0000 UTC 17 October 2017, which was when the storm was affecting the British Isles. It is clear that, for both the 3-day and 5-day hindcasts, the neural network can predict the general distribution of Z500 and T850 well. Unsurprisingly, the 3-day hindcast is more accurate at predicting the location of the storm and is able to pick up finer details not present in the 5-day hindcast.

Finally, in Table 1, we summarise the results from using our stacked neural network and compare them with results from using simple methods (persistence and climatology), other neural networks (Weyn *et al.*, 2020; Rasp and Thuerey, 2020) and numerical models (IFS T42 and operational IFS) where the numerical model results have been taken from Rasp *et al.* (2020). The key finding is that our approach is approximately as accurate as that in Weyn *et al.* (2020), despite the fact that our neural network is simpler than the U-Net approach used in Weyn *et al.* (2020). It should be noted that, although our training dataset

includes more variables than that used in Weyn *et al.* (2020), it uses data at a much finer resolution (2° compared to 5.625°) and furthermore we are using categorical data rather than continuous data which adds an inbuilt error to our results. Thus this result highlights the advantages we gain from the data exploration and good choice of neural network architecture described in Section 2. The table also shows that our approach is more accurate than the coarse numerical IFS T42 model and the simple methods of persistence and climatology, but less accurate than the neural network approach in Rasp and Thuerey (2020) and the operational IFS model. It is likely that our neural network's lower skill compared to Rasp and Thuerey (2020) is due to the fact that the Rasp and Thuerey (2020) model is trained on a much larger dataset of the Weather-Bench data (117 data variables compared to our training dataset of 34 data variables for Z500 and 33 data variables for T850) and is also pretrained on extra data from the Climate Model Inter-comparison Project (CMIP; Eyring *et al.*, 2016). Thus the approach in Rasp and Thuerey (2020) is both much more computationally expensive and much more memory intensive than our approach. Furthermore, our approach has introduced improvements which combine dropout-based ensembles with the ability to predict probability density functions instead of single values. In the next section, we show how this enables us to make a more informed weather forecast.

3.1 | Estimating uncertainty

A common criticism of using neural networks for weather forecasting is that assessing the uncertainty of their

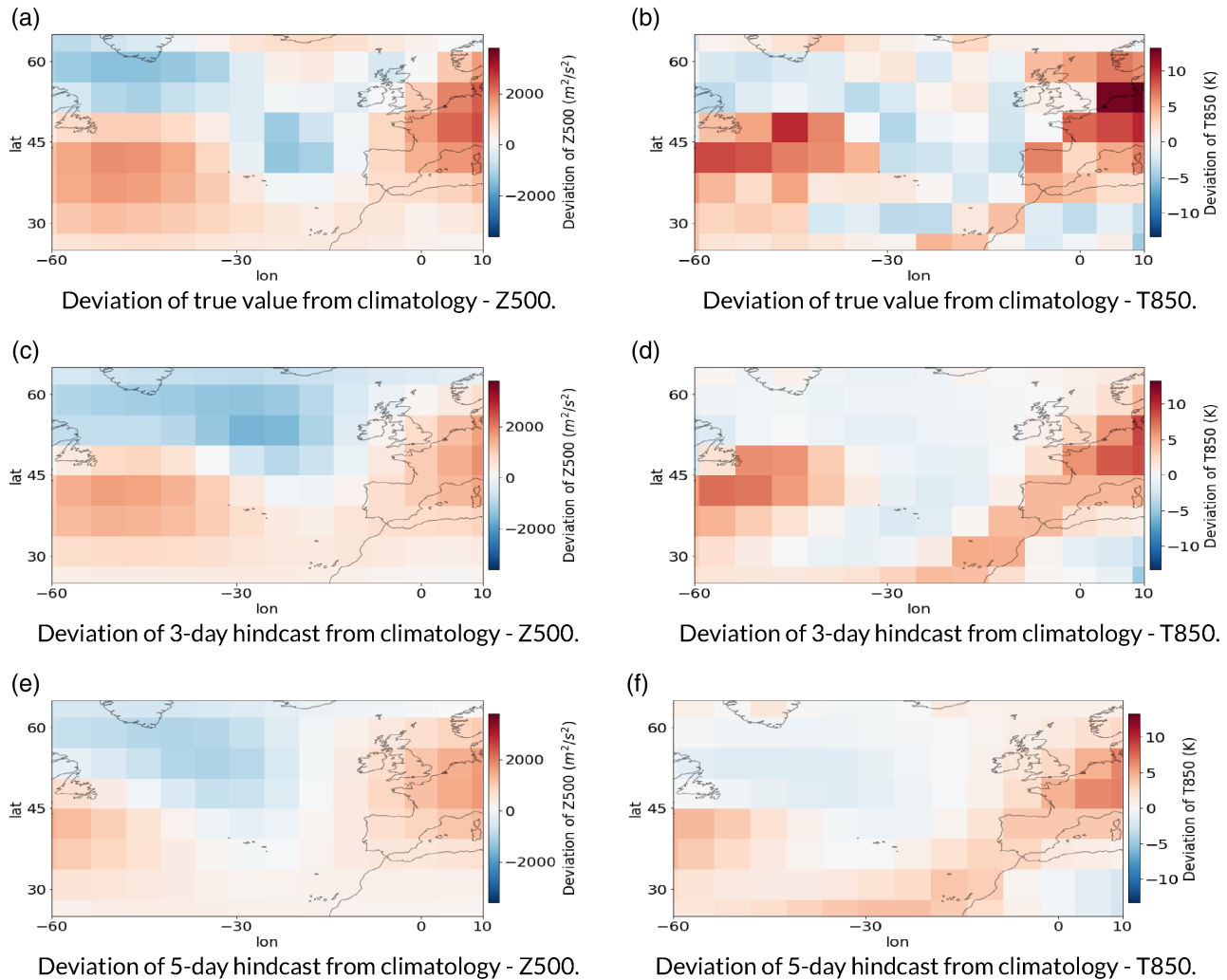


FIGURE 11 Deviation from the climatology values of (a, b) the true values, (c, d) the 3-day hindcast values and (e, f) the 5-day hindcast values of (a, c, e) Z500 and (b, d, f) T850, at 0000 UTC 17 October 2017 during storm *Ophelia* [Colour figure can be viewed at wileyonlinelibrary.com]

forecasts is difficult, and requires techniques such as ensemble approaches, which are often computationally expensive (see Schultz *et al.*, 2021 and discussion in Section 1). Thus, one of the key novelties of our neural network approach to predicting the weather is that it provides a novel efficient method for producing a probabilistic output, through our prediction of a full probability density function for the variable of interest at each point in space and time. This probabilistic output enables us to estimate uncertainty and obtain notably more information from our neural network predictions than can be obtained from a deterministic output.

To visualise the probabilistic output, we again consider the example of storm *Ophelia*. Figures 12a and 12b show the cumulative distribution function (CDF) for T850 from 3- and 5-day hindcasts respectively, at 0000 UTC 17 October 2017. In each panel, the CDFs for three different thresholds are shown using probability contours.

These probability contours indicate the locations where the probability of T850 being lower than 263.15 K (-10°C) (blue), 273.15 K (0°C) (green) and 283.15 K (10°C) (red) is 10, 50 and 90%. Similarly Figures 12c and 12d show equivalent probability information for Z500. Note that, although for a given threshold the probability contours will never cross, the probability contours for different thresholds may cross.

In effect, Figure 12a and 12b shows the median location of the three isotherms as well as the 10 to 90% probability interval. We see meanders in the contours associated with midlatitude synoptic systems, and locations where the probabilities are not symmetric about the median (e.g., in the 3-day T850 hindcast off the west coast of California, where the distance between the 90% and 50% contours is larger than between the 50% and 10% contours for the probability of T850 being less than 283.15 K (10°C)). The CDF for the thresholds of 263.15, 273.15, 283.15 K chosen

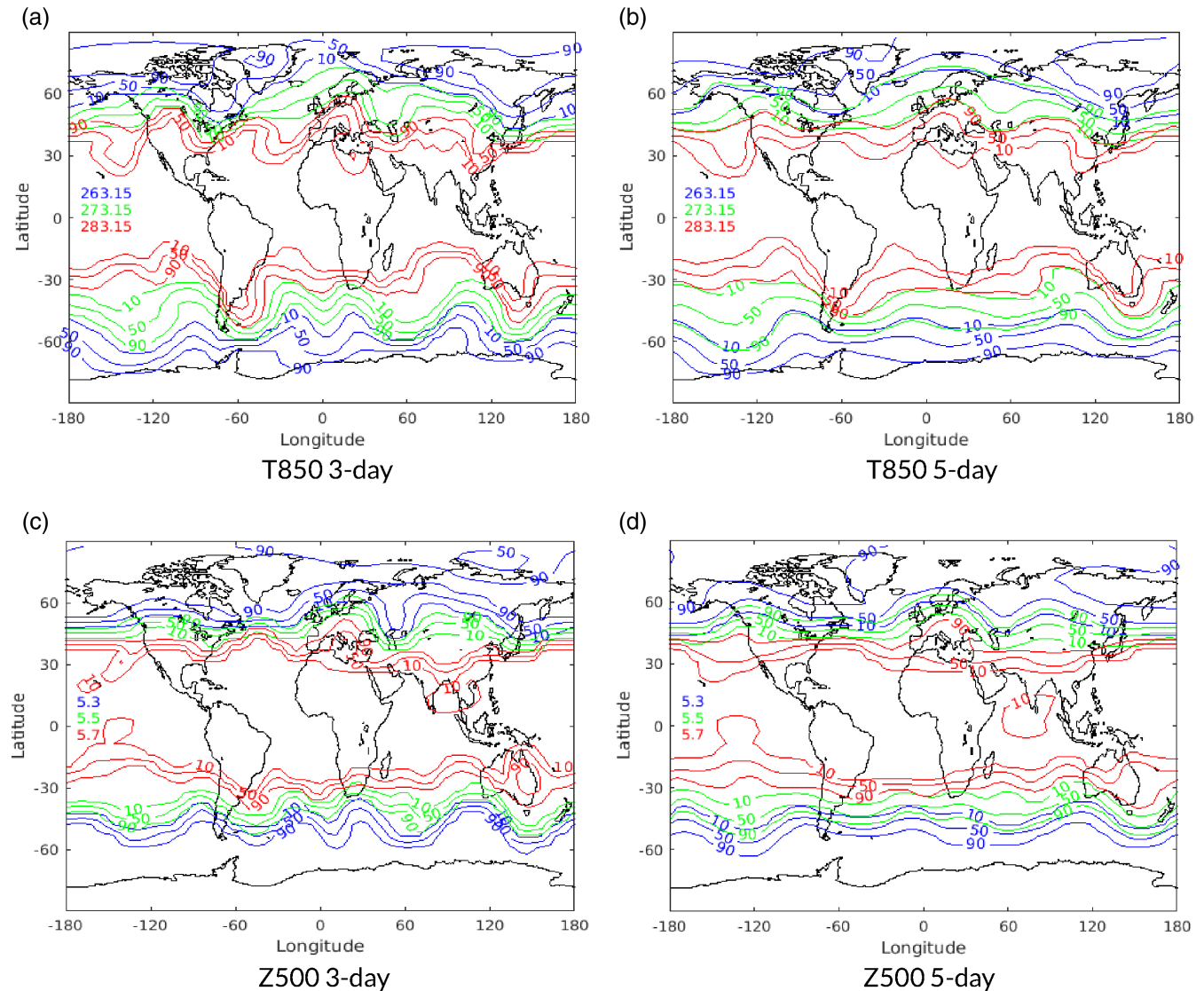


FIGURE 12 Contour maps of the cumulative distribution functions at 0000 UTC 17 October 2017, showing contours at 10, 50 and 90%. For (a, b) the thresholds considered are $T850 < 263.15$ K (blue), $T850 < 273.15$ K (green) and $T850 < 283.15$ K (red) and for (c, d) they are $Z500 < 5.3 \times 10^4$ m (blue), $Z500 < 5.5 \times 10^4$ m (green) and $Z500 < 5.7 \times 10^4$ m (red). [Colour figure can be viewed at wileyonlinelibrary.com]

here, generally do not overlap in the 3-day T850 hindcasts in Figure 12a. However, in the 5-day T850 hindcasts, where there is more uncertainty, the spread in each set of contours is wider and there are places where they do overlap (e.g., over Iceland, where there is a 90% chance of T850 being lower than 273.15 K (0 °C) but also a 10% chance that it is lower than 263.15 K (−10 °C)). For Z500, Figures 12c and 12d also show the contours are closer together in the 3-day hindcast than in the 5-day hindcast, indicating more certainty in the 3-day hindcast.

The probabilistic output of the neural network also means we can estimate uncertainty and skill metrics (Hudson and Ebert, 2017) for our predictions. A standard metric used to evaluate probabilistic weather forecasts (Rasp and Lerch, 2018) is the Continuous Ranked Probability Score

(CRPS). This metric helps to evaluate the distance between the forecast and the real solution and is calculated using

$$CRPS(F, y) = \int_{-\infty}^{\infty} \{F(z) - \mathbb{1}(y \leq z)\}^2 dz, \quad (7)$$

where $F(z)$ is the CDF of the forecast density function and $\mathbb{1}(y \leq z)$ is an indicator function representing the CDF of the observed value (see Hersbach 2000 for more details). Following general practice (e.g., Scher and Messori, 2021), we compute the CRPS for each density function at each individual gridpoint for each time and then average over them. Note that, the lower the CRPS score, the better the forecast density function approximates the real density function.

TABLE 1 Error calculated using Equation (3) on the test dataset for 3- and 5-day hindcasts for Z500 and T850

	RMSE Z500 (3-/5-day) (m²·s⁻²)	RMSE T850 (3-/5-day) (K)	Spatial resolution of data
Stacked neural network	375 / 627	2.11 / 2.91	5.625°
Persistence	936 / 1033	4.23 / 4.56	5.625°
Climatology	1075	5.51	5.625°
IFS T42	489 / 743	3.09 / 3.83	2.8°
Weyn <i>et al.</i> (2020)	373 / 611	1.98 / 2.87	2.0°
Rasp and Thuerey (2020)	268 / 499	1.65 / 2.41	5.625°
Operational IFS	154 / 334	1.36 / 2.03	0.1°

Note: The table compares the results from using our approach in bold, with simple methods (persistence and climatology), numerical models (IFS T42) and other neural networks (Rasp and Thuerey, 2020; Weyn *et al.*, 2020). All approaches have been evaluated on the entire global region.

Table 2 shows the CRPS values for Z500 and T850 for the 3- and 5-day hindcasts for our stacked neural network. In order to interpret these values, we first compare the CRPS Z500 values with those obtained in Scher and Messori (2021), who combine the dropout-at-inference ensemble method (Section 2.2.1) with a convolutional neural network to generate an ensemble forecast to predict Z500 in the same time period (2017–2018) as in our work. Even though we use data at a much coarser resolution, our stacked neural network approach performs much better than the approach in Scher and Messori (2021) for the 3-day hindcast but much worse for the 5-day hindcast. The error growth for our results is proportionally larger than that in Scher and Messori (2021) and suggests that for Z500 the distributions produced by our approach are close to the real distributions for the 3-day hindcast, but suffer greatly as the lead time increases. Scher and Messori (2021) did not use their approach to predict T850 and thus we instead compare our T850 values with the “dressed” ERA CRPS values for the T850 5-day hindcast used by ECMWF to benchmark their operational IFS (Haiden *et al.*, 2017, 2018). Note this is not an exact comparison because the ERA CRPS values are only for extratropical regions whereas our CRPS values are for the entire global region. The method used to calculate these “dressed” distributions is very different to the method we use to produce distributions. In the “dressed” approach, the mean error and standard deviation of the 30 days prior to the datapoint of interest are used to estimate a Gaussian distribution around reanalysis data. Note that in 2017 ERA-Interim was used as the reanalysis data and in 2018 this was changed to ERA5. For reference, we have

also included the actual CRPS from the operational IFS (Haiden *et al.*, 2018). Unsurprisingly, the CRPS of the operational IFS is much lower than that from our approach, but the ECMWF benchmark CRPS are fairly close to the CRPS from our approach (especially the interim “dressed” score), which is a promising result.

We can further examine the spread of the density functions by looking at their standard deviation, σ . We do this using

$$\sigma = \sqrt{\sum_{i=1}^{100} (x_i - \mu)^2 \mathbb{P}(X = x_i)}, \quad (8)$$

where μ is the expectation of X given by Equation (2) and, as before, the value of each bin x_i is taken to be the lower bound. This informs on the spread of the distribution and allows us to calculate the confidence intervals using

$$\mu \pm Z \frac{\sigma}{\sqrt{N}}, \quad (9)$$

where N is the number of bins and Z is the z -value taken from the normal distribution and equal to 1.960 for the 95% confidence interval and 2.576 for the 99% confidence interval. Note here that the confidence interval is for the distribution at each point in space and time and not for the error as in Equation (A1) in Appendix A. In Table 3, we show the percentage of datapoints in time and space where the true value is within the 95% and 99% confidence intervals around the predicted expected value i.e., the percentage of datapoints where the true value and the predicted value are not statistically different from each other at the 95% or 99% confidence interval. The proportions are relatively low but it should be noted that the confidence intervals are relatively narrow because N is relatively large and, as shown in Figure 2, the distributions are very centred around the expected value – the average width for the 99% confidence interval is 0.6 K and 0.8 K for the T850 3- and 5-day hindcasts respectively (smaller than the width of one bin) and 100 and 160 m²s⁻² for the Z500 3- and 5-day hindcasts respectively (approximately the same size as the width of one bin). Therefore, we also show in Table 3 the proportion of true values within one and two σ (calculated using Equation (8)) from the predicted expected value. The majority of true values are within one σ of the expected value and almost all (over 90%) are within two σ . These metrics, and others like it, help practitioners evaluate the neural network results compared to other model results and also help them understand how much confidence to have in the predictions.

Finally, the probability density functions not only improve understanding of the spread and skill of the results, but also inform of scenarios which are not the

TABLE 2 CRPS for the 3- and 5-day hindcasts for Z500 and T850 averaged over all gridpoints and time

	Z500 (3-day) ($\text{m}^2 \cdot \text{s}^{-2}$)	Z500 (5-day) ($\text{m}^2 \cdot \text{s}^{-2}$)	T850 (3-day) (K)	T850 (5-day) (K)	Spatial resolution of data
Stacked neural network	211	1500	1.22	1.69	5.625°
Scher and Messori (2021)	526	707	—	—	2.5°
Dressed ERA Interim (2017)	—	—	—	1.44	0.54°
Dressed ERA (2018)	—	—	—	1.18	0.28°
Operational IFS	—	—	—	0.98	0.1°

Note: Where available, the table presents the CRPS values from our stacked neural network, from other neural network approaches (Scher and Messori, 2021), from the benchmarks used by the ECMWF in 2017 and 2018 and from the Operational IFS for comparison. Furthermore, our results and Scher and Messori (2021) are evaluated over the entire globe, but the ECMWF results are evaluated only over the extratropics.

TABLE 3 Percentage of datapoints where the true value is within a set interval from the expected value predicted by the neural network

	Z500 (3-day)	T850 (3-day)	Z500 (5-day)	T850 (5-day)
Within 95% confidence interval	13.8	17.2	14.7	17.3
Within 99% confidence interval	16.3	20.3	17.4	20.5
Within $\pm\sigma$	64.7	71.2	67.3	71.2
Within $\pm 2\sigma$	93.6	94.0	94.0	94.2

most likely to occur, but still have a relatively high probability of doing so. Recall from the example in Figure 2b that, in that case, the bin the neural network predicts with the highest probability is not the correct bin, but the correct bin still has a high probability of occurring. Figure 13 quantifies cases like these. The first bar in all four subfigures shows the percentage of datapoints for which the bin with the predicted highest probability is the correct bin – in all four cases this is over 20%. The second bar shows the percentage of datapoints where the correct bin is either the bin with the predicted highest probability or the predicted second highest probability of occurring. This continues until the fifth bar of the subfigures which shows the percentage of datapoints where the correct bin is one of the top five bins that the neural network predicts is most likely to occur. It should be noted here, for clarity, that the SoftMax layer is capable of predicting multi-modal distributions and thus bins with high probabilities are not necessarily close together in value. For example, if the distribution is bimodal, it may be that the bin with the highest probability is at one end of the spectrum and the bin with the second highest probability is at the other end.

In summary, Figure 13 shows that, for the 3-day predictions, the correct bin is in one of the top five most likely to occur for almost 80% of the datapoints, and for the 5-day predictions, the correct bin is one of the top

five most likely for over 60% of the datapoints. Given that there are 100 bins, such a high proportion from just the top 5 is a notable result. It means that in the vast majority of cases, the true scenarios have a high probability associated with them, which enables practitioners to make informed decisions when forecasting and issuing weather warnings.

Thus we have shown in this section how much more information can be gained for weather forecasting if the neural network predicts density functions rather than single values.

4 | CONCLUSION

In this work, we have successfully developed a novel neural network approach, which is able to predict full probability density functions for the target weather variable at each point in space and time instead of single values. This enables practitioners to estimate the uncertainty of the neural network predictions and to provide a more informed weather forecast. In particular, the probability density functions inform about events which, although they may not be the most likely to occur, still have a significant probability of happening. We have thus provided a strong proof-of-concept of how neural networks can be used to produce probabilistic weather forecasts,

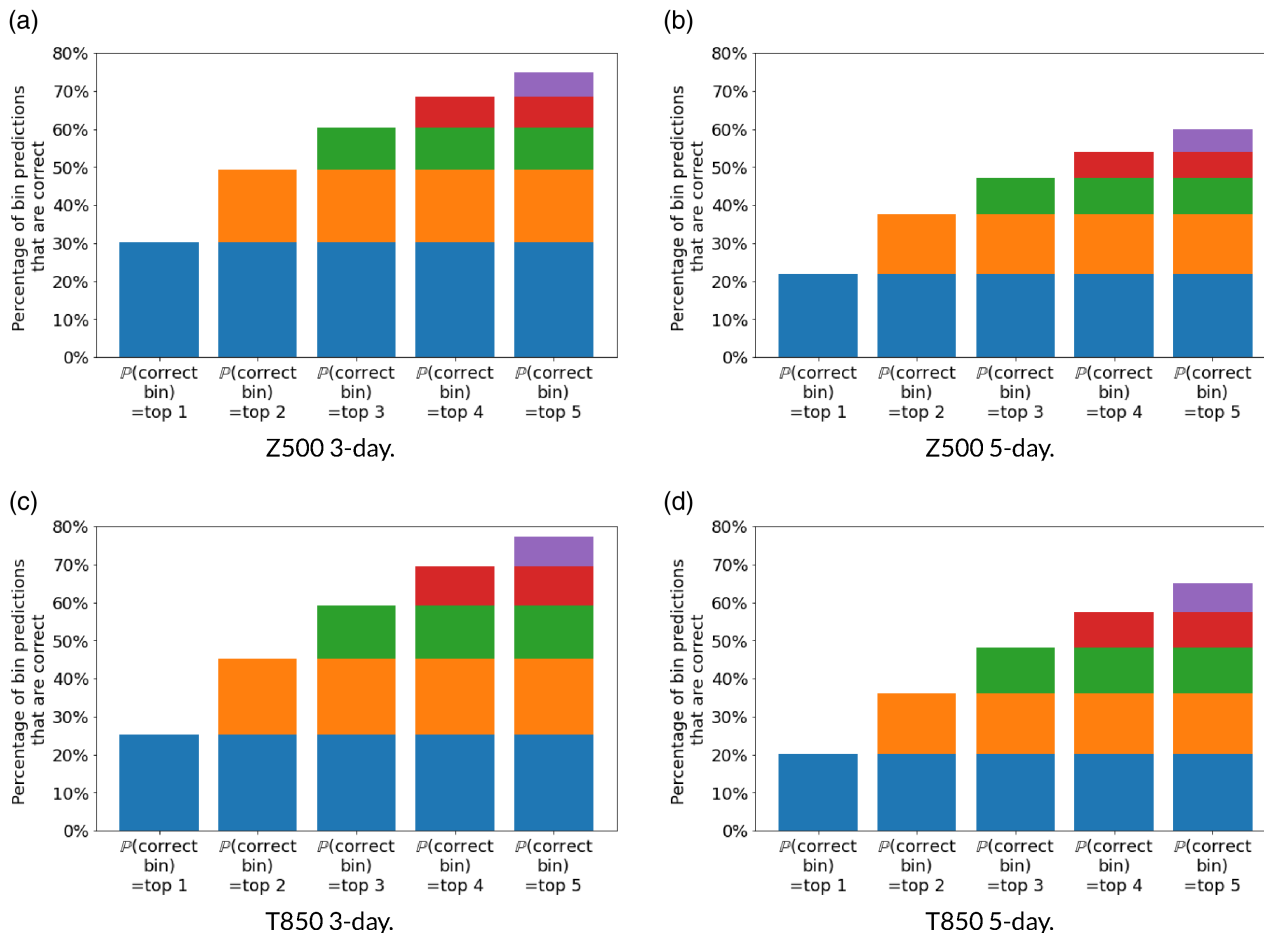


FIGURE 13 Percentage of all datapoints in time and space where one of the top 5 most likely bins predicted by the neural network is the observed bin for (a) 500 hPa geopotential, and (b) 850 hPa temperature [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.com)]

which is an area where many weather forecasting practitioners would like to see neural networks improve (Schultz *et al.*, 2021).

For our neural network predictions, a relatively simple ResNet has been used and carefully optimised to improve the accuracy of our results. Through the careful choice of this architecture along with extensive data exploration, we have produced weather hindcasts which are more accurate than some coarse NWP models, and as accurate as those in Weyn *et al.* (2020) which use a more complex neural network architecture. Moreover, our novel use of a stacked neural network to combine outputs for weather forecasting reduces the memory cost, as well as the computational cost as smaller networks generally take less time to train, and this means that less powerful computers are required to make predictions.

Finally, in this work we have shown that transforming our output data to categorical can still give accurate results for continuous numerical weather data. We have shown that it is possible to move beyond point estimates for neural network-based weather forecasts and produce

a probabilistic forecast by combining multiple smaller more efficient models. This opens up a new avenue of research for data-driven weather forecasting, where the use of ever bigger models trained with enormous datasets is not the only way to achieve model skill. In future work, we will seek to explore the use of transfer learning to further improve the training efficiency of the neural networks.

4.1 | Computer code availability

The relevant code for the neural networks presented in this work can be found at https://github.com/mc4117/ResNet_Weather.git (accessed 5 October 2021).


ACKNOWLEDGEMENTS

MCAC acknowledges UK Engineering and Physical Sciences Research Council grants EP/R512540/1 and EP/L016613/1. We thank Gabriele Messori and an anonymous reviewer for their comments which helped improve the manuscript.


AUTHOR CONTRIBUTIONS

Mariana C.A. Clare: conceptualization; data curation; formal analysis; investigation; methodology; validation; visualization; writing – original draft; writing – review and editing. **Omar Jamil:** conceptualization; methodology; project administration; supervision; writing – review and editing. **Cyril J. Morcrette:** supervision; visualization; writing – review and editing.

ORCID

Mariana C.A. Clare  <https://orcid.org/0000-0002-5010-0363>

Omar Jamil  <https://orcid.org/0000-0002-4465-4925>

Cyril J. Morcrette  <https://orcid.org/0000-0002-4240-8472>

REFERENCES

- Allard, D., Comunian, A. and Renard, P. (2012) Probability aggregation methods in geoscience. *Mathematical Geosciences*, 44, 545–581.
- Bauer, P., Thorpe, A. and Brunet, G. (2015) The quiet revolution of numerical weather prediction. *Nature*, 525, 47–55.
- Bihlo, A. (2021) A generative adversarial network approach to (ensemble) weather prediction. *Neural Networks*. <https://doi.org/10.1016/j.neucom.2020.109008>.
- Brenowitz, N.D. and Bretherton, C.S. (2019) Spatially extended tests of a neural network parametrization trained by Coarse-Graining. *Journal of Advances in Modeling Earth Systems*, 11, 2728–2744. <https://doi.org/10.1029/2019MS001711>.
- Dueben, P.D. and Bauer, P. (2018) Challenges and design choices for global weather and climate models based on machine learning. *Geoscientific Model Development*, 11, 3999–4009.
- Eyring, V., Bony, S., Meehl, G.A., Senior, C.A., Stevens, B., Stouffer, R.J. and Taylor, K.E. (2016) Overview of the coupled model inter-comparison project phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, 9, 1937–1958.
- Gal, Y. and Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning, pp.1050-1059 in 33rd International Conference on Machine Learning, New York, NY.
- Goodfellow, I., Bengio, Y. and Courville, A. (2016) *Deep Learning*. Cambridge, MA: MIT Press.
- Grönquist, P., Yao, C., Ben-Nun, T., Dryden, N., Dueben, P., Li, S. and Hoefler, T. (2021) Deep learning for post-processing ensemble weather forecasts. *Philosophical Transactions of the Royal Society A*, 379, 20200092.
- Haiden, T., Janousek, M., Bidlot, J., Ferranti, L., Prates, F., Vitart, F., Bauer, P. and Richardson, D.S. (2017). Evaluation of ECMWF forecasts, including 2016–2017 upgrades, Reading, UK. Technical Memo. 817.
- Haiden, T., Janousek, M., Bidlot, J., Buizza, R., Ferranti, L., Prates, F. and Vitart, F. (2018). Evaluation of ECMWF forecasts, including the 2018 upgrade, Reading, UK. Technical Memo. 831.
- He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep residual learning for image recognition, pp. 770-778 in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV.
- Hersbach, H. (2000) Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15, 559–570.
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D.P., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A.J., Haimberger, L., Healy, S.B., Hogan, R.J., Hólm, E.V., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S. and Thépaut, J.-N. (2020) The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146, 1999–2049.
- Hoskins, B.J. (2015) Potential vorticity and the PV perspective. *Advances in Atmospheric Sciences*, 32, 2–9.
- Hoskins, B.J., Draghici, I. and Davies, H. (1978) A new look at the ω -equation. *Quarterly Journal of the Royal Meteorological Society*, 104, 31–38.
- Hudson, D. and Ebert, B. (2017). Ensemble Verification Metrics. In: Proceedings of the ECMWF Annual Seminar, Reading, UK: ECMWF.
- Kingma, D.P. and Ba, J.L. (2014) Adam: a method for stochastic optimization. arXiv:1412.6980.
- Lu, Y., Zhong, A., Li, Q. and Dong, B. (2018). Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations, pp. 3276-3285 in International Conference on Machine Learning, Stockholm.
- Maas, A.L., Hannun, A.Y. and Ng, A.Y. (2013). Rectifier nonlinearities improve neural network acoustic models. In: Proceedings of the International Conference on Machine Learning, Atlanta, GA.
- Palmer, T.N., Buizza, R., Hagedorn, R., Lawrence, A., Leutbecher, M. and Smith, L. (2006) Ensemble prediction: a pedagogical perspective. *ECMWF Newsletter*, 106, 10–17.
- Pedder, M.A. (1997) The omega equation: Q-G interpretations of simple circulation features. *Meteorological Applications*, 4, 335–344.
- Rasp, S. and Lerch, S. (2018) Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, 146, 3885–3900.
- Rasp, S. and Thuerey, N. (2020) Purely data-driven medium-range weather forecasting achieves comparable skill to physical models at similar resolution. arXiv:2008.08626v1.
- Rasp, S., Dueben, P.D., Scher, S., Weyn, J.A., Mouatadid, S. and Thuerey, N. (2020) WeatherBench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12. <https://doi.org/10.1029/2020MS002203>.
- Ronneberger, O., Fischer, P. and Brox, T. (2015). U-net: convolutional networks for biomedical image segmentation, pp. 234-241 in International Conference on Medical Image Computing and Computer-assisted Intervention, Munich, Germany.
- Scher, S. and Messori, G. (2018) Predicting weather forecast uncertainty with machine learning. *Quarterly Journal of the Royal Meteorological Society*, 144, 2830–2841.
- Scher, S. and Messori, G. (2021) Ensemble methods for neural network-based weather forecasts. *Journal of Advances in Modeling Earth Systems*, 13(2). <https://doi.org/10.1029/2020MS002331>.
- Schultz, M.G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L.H., Mozaffari, A. and Stadler, S. (2021) Can deep learning beat numerical weather prediction?. *Philosophical*

Transactions of the Royal Society A, 379. <https://doi.org/10.1098/rsta.2020.0097>.

- Smyth, P. and Wolpert, D. (1999) Linearly combining density estimators via stacking. *Machine Learning*, 36, 59–83.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. (2014) Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15, 1929–1958.
- Stewart, S.R. (2018). Hurricane *Ophelia* – Tropical Cyclone Report AL172017, Miami, FL.
- Weyn, J.A., Durran, D.R. and 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). <https://doi.org/10.1029/2020MS002109>.
- Weyn, J.A., Durran, D.R., Caruana, R. and Cresswell-Clay, N. (2021) Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models. *Journal of Advances in Modeling Earth Systems*, 13(7). <https://doi.org/10.1029/2021MS002502>.
- White, P.W. (1971) Finite-difference methods in numerical weather prediction. *Proceedings of the Royal Society of London. Series A*, 323, 285–292.
- Wolpert, D.H. (1992) Stacked generalization. *Neural networks*, 5, 241–259.
- Yuval, J. and O’Gorman, P.A. (2020) Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions. *Nature Communications*, 11, 1–10. <https://doi.org/10.1038/s41467-020-17142-3>.
- Zhou, Z.-H. (2012) *Ensemble Methods: Foundations and Algorithms*. Boca Raton, FL: CRC Press.

How to cite this article: Clare, M.C., Jamil, O. & Morcrette, C.J. (2021) Combining distribution-based neural networks to predict weather forecast probabilities. *Quarterly Journal of the Royal Meteorological Society*, 147(741), 4337–4357. Available from: <https://doi.org/10.1002/qj.4180>

APPENDIX A. PRESSURE LEVEL IMPORTANCE

In this Appendix, we outline the data-driven approach to determining which pressure levels we should use to predict Z500 and T850. This analysis often agrees with well-established theory in meteorology, as noted later in this Appendix. For all multi-level input variables (geopotential, temperature, specific humidity, potential vorticity and zonal wind), we consider multiple different level combinations. We use these combinations to train a five-block convolutional ResNet and use the dropout at inference technique to extract an ensemble of 32 outputs from the single trained model, and then calculate the error of the ensemble mean (Section 2.2.1).

The statistical significance of the multi-level analysis results is analysed by calculating the 95% confidence

interval of the MSE of the hindcast at each point in space and time in the validation dataset. We use the MSE rather than the RMSE for ease of calculation, and therefore the 95% confidence interval formula is

$$\text{MSE} \pm 1.96 \sqrt{\frac{\mathbb{V}[L(j)(f_{i,j,k} - t_{i,j,k})^2]}{N}}, \quad (\text{A1})$$

where

$$\begin{aligned} \text{MSE} &= \mathbb{E}[L(j)(f_{i,j,k} - t_{i,j,k})^2] \\ &= \frac{1}{N_{\text{lat}}N_{\text{lon}}N_{\text{timepoints}}} \sum_{i=1}^{N_{\text{timepoints}}} \sum_{j=1}^{N_{\text{lat}}} \sum_{k=1}^{N_{\text{lon}}} L(j)(f_{i,j,k} - t_{i,j,k})^2. \end{aligned} \quad (\text{A2})$$

$L(j)$ is the latitude weighting factor given by Equation (4), f is the predicted value from the neural network, t is the true value from the dataset and N is the total number of points in space and time in the validation dataset (equal to $N_{\text{timepoints}} \times N_{\text{lat}} \times N_{\text{lon}}$).

The results of the level analysis are shown in Figure A1 as percentages relative to the benchmark. For brevity and figure clarity, the confidence intervals are shown only for the optimum level choice (denoted by a star) and to another level combination with an error close to that of the optimum error choice. As the confidence intervals are so narrow, the error bars are extended to make them clearer. Even so, in some cases, the confidence intervals are so narrow that it is not possible to distinguish between the upper and lower error bars on the figure, making it difficult to interpret them. To avoid confusion, we emphasise here that all error bars shown in these figures are in the y-direction. The remainder of this appendix provides more detail on how the optimum level choice for each variable is determined.

The optimum levels for the geopotential and temperature variables are determined in a systematic way. First we consider the two levels we are trying to predict and the levels between these two (i.e., 500, 600, 700, 850 hPa); next we extend this dataset by including a further two levels with larger pressure values (i.e., 500, 600, 700, 850, 925, 1000 hPa); then we add the two levels with smaller pressure values to the first dataset (i.e., 300, 400, 500, 600, 700, 850 hPa); then we include both the levels with larger and smaller pressure values (i.e., 300, 400, 500, 600, 700, 850, 950, 1000 hPa), and finally we add two further levels with small pressure values (i.e., 100, 250, 300, 400, 500, 600, 700, 850, 950, 1000 hPa). The results of this analysis are shown in Figures A1a and A1b, for the geopotential variable inputs and Figures A1c and A1d, for the temperature variable inputs. In general, they show for both Z500 and T850 that including more pressure levels results

in a larger decrease in error relative to the benchmark, but this must be balanced with computational cost. When predicting Z500, a good compromise for both the geopotential and temperature input variables is the level choice [300, 400, 500, 600, 700, 850 hPa] (stars on Figures A1a and A1c). The figures show that the 95% confidence interval for this optimum level choice does not overlap with that of the level choice with the most similar accuracy, and the same result applies for all other level choices (not shown for brevity). Therefore, our optimum level choice is statistically significantly different in accuracy from all other level choices considered. When predicting T850, a good compromise for both the geopotential and temperature input variables is the level choice [500, 600, 700, 850, 925, 1000 hPa] (stars on Figures A1b and A1d). Like when predicting Z500, Figures A1b and A1d show no overlapping of the 95% confidence interval (Equation (A1)), and therefore our optimum level choice is statistically significantly different in accuracy. Given that T850 has a larger pressure value than Z500, it is unsurprising that the T850 level choice contains levels with larger pressure values than the Z500 level choice.

For the other variables, the same analysis does not apply and we conducted a small correlation analysis between the different levels of specific humidity, potential vorticity and zonal wind, and the target variables using pattern correlation to determine which levels might be good predictors. This analysis is not included here for brevity but is available in the data exploration section of the GitHub repository (Section 4.1). For specific humidity, this analysis showed that the correlation is greatest at the levels with large pressure values. Thus, we begin by considering the levels with larger pressure values (i.e., [600, 700, 850, 925, 1000 hPa]), then add different subsets of levels with small pressure values, and finally consider a broad range of pressure levels (i.e., [150, 200, 250, 300, 500, 600, 700, 850, 925, 1000 hPa]). Figures A1e and A1f show the results of this analysis; unlike the temperature and geopotential input variables, adding more pressure levels does not always result in the error decreasing relative to the benchmark. Without clear correlation between the number of pressure levels and the error reduction, we make the level choice based on reducing the error while keeping the number of levels as few as possible and thus choose [150, 200, 600, 700, 850, 925, 1000 hPa] (stars on Figures A1e and A1f) for both Z500 and T850 prediction. Figure A1f shows that, when predicting T850, the 95% confidence interval (Equation (A1)) of this optimum level choice does not overlap with that of the level choice with the most similar accuracy, showing that our optimum level choice is statistically significantly different in accuracy. However, Figure A1e shows that, when predicting Z500, the difference in the accuracy achieved by using

[150, 200, 600, 700, 850, 925, 1,000 hPa] (our optimum level choice) compared to [150, 200, 250, 300, 500, 600, 700, 850 hPa] is not statistically significant (95% confidence intervals overlap). This means we cannot use the criteria of accuracy to distinguish between the two level combinations, but the difference in computational cost between the two level combinations still justifies our optimum level choice.

For potential vorticity, the correlation analysis showed that the most important levels are at either end of the pressure level spectrum. Thus when determining the right level choice, we consider groupings of levels at both large and small pressure values. We also consider the middle pressure levels shifted to the larger end of the pressure spectrum (i.e., [250, 300, 400, 500, 700, 850, 925, 1000 hPa]) and to the smaller end of the pressure spectrum (i.e., [50, 100, 150, 250, 300, 400, 500, 850, 925 hPa]). Figures A1g and A1h show that there does not seem to be a clear pattern between level choice and error decrease for potential vorticity. However, the level choice of [150, 250, 300, 700, 850 hPa] (stars on Figures A1g and A1h) results in a large error decrease in both RMSE Z500 and RMSE T850 and is also relatively computationally cheap and so is an appropriate level choice for potential vorticity. Moreover, the figures show that, for both Z500 and T850, the 95% confidence interval (Equation (A1)) of this optimum level choice does not overlap with the confidence interval of the level choice closest in accuracy to it, which also applies for all other level choices (not shown for brevity). This means our optimum level choice is statistically significantly different in accuracy to the other level choices.

Finally we consider the zonal wind input variable. Using the correlation analysis, the most important levels are at either end of the pressure spectrum. Thus we consider groupings of levels at both large and small pressure values. We also consider groupings at the larger end of the pressure spectrum (i.e., [300, 400, 500, 600, 700, 850, 925, 1000 hPa]) to check if the smaller pressure values are affecting the results. Figures A1i and A1j show that for both Z500 and T850 having levels at both small and large pressure values is important for reducing error. In both cases, we choose the level grouping of [50, 100, 300, 850, 925, 1000 hPa] (stars on Figures A1i and A1j) as an appropriate choice because it results in a large error reduction relative to the benchmark and is also relatively computationally cheap. Figure A1i shows that the 95% confidence interval (Equation (A1)) of this optimum level choice does not overlap with the confidence interval of the level choice closest in accuracy to it, meaning our level choice is statistically significantly different in accuracy to the other level choice. In the case of Figure A1j, we chose to compare the 95% confidence interval of the optimum level choice with that of [50, 100, 850, 925,

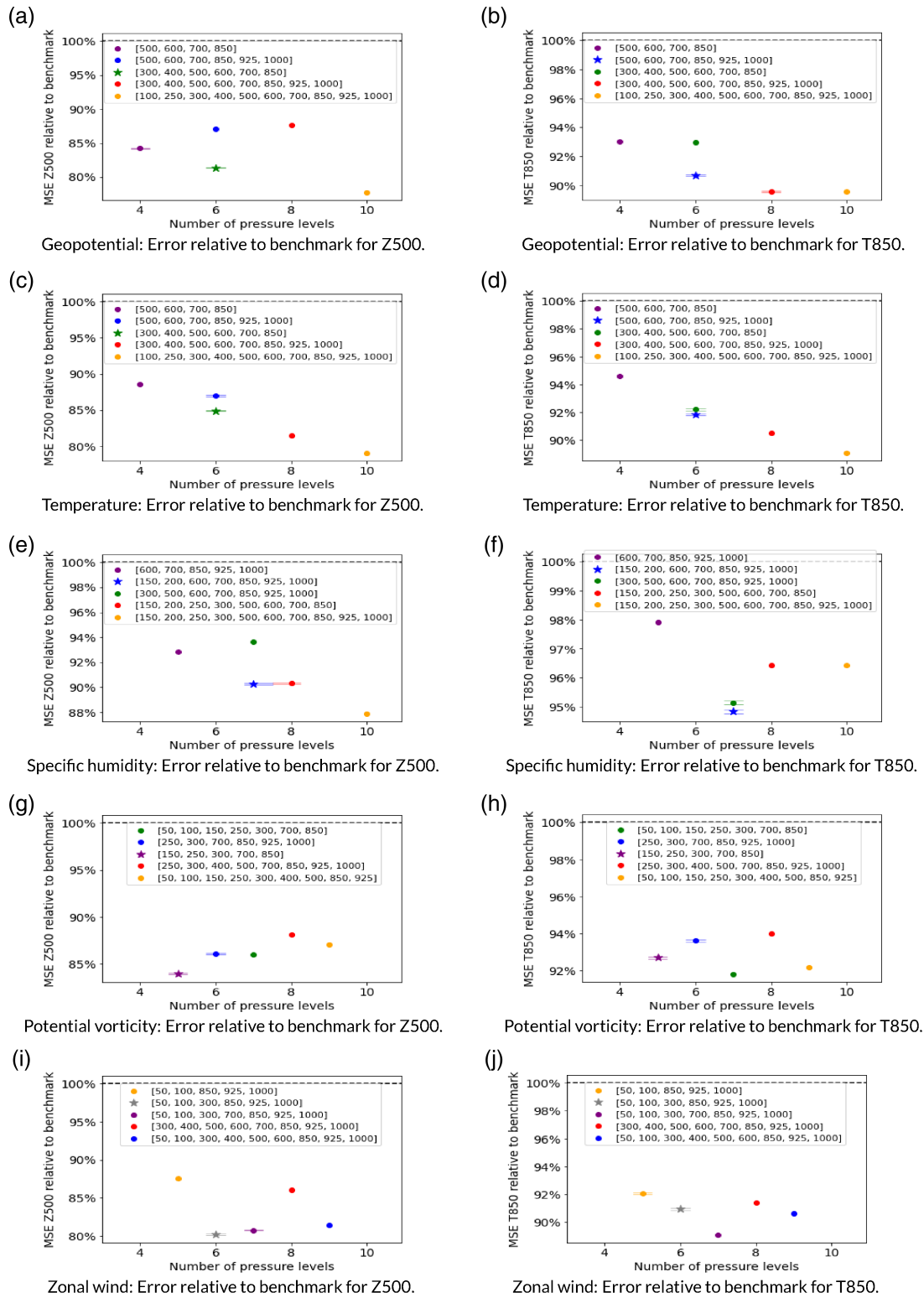


FIGURE A1 Percentage error relative to benchmark from training using extra variable levels (chosen levels are denoted by a star) for (a, c, e, g, i) 500 hPa geopotential, and (b, d, f, h, j) 850 hPa temperature. The 95% confidence interval bars are shown for some level combinations. Some intervals are so narrow that it is not possible to distinguish between the upper and lower error bars on the figure, but we emphasise here that all error bars are in the y-direction. Results are the mean of 32 ensemble members, predicting on all gridpoints and times in the validation dataset (i.e., 2016) [Colour figure can be viewed at wileyonlinelibrary.com]

1000 hPa] because of the higher computational cost of the level choice which is closest in accuracy. The confidence intervals do not overlap, hence our optimum level choice is statistically significantly different in accuracy compared to the computationally cheaper option.

Thus we have shown that our optimum level choices are always statistically significantly different from other level choices at a lower or equivalent computational cost, and are almost always statistically significantly different from other more computationally expensive level choices.

So far, the choice of which pressure levels to include has been done in a predominantly data-driven manner. It is worth emphasizing that the focus has been on predicting Z500 and T850 3 days ahead. If other target variables, or other lead-times, had been the focus of the predictions, then other input levels may have been found to be beneficial. Having found the most useful input levels, it is now sensible to look back and comment on whether the choices are supported by physical reasoning. This is the case for many of the pressure level choices that we have made:

- **Geopotential and temperature:** Using physical intuition, we would expect that the levels that are most important for prediction are those above and below the level of interest (e.g., Hoskins *et al.*, 1978).
- **Potential vorticity:** Our training dataset contains levels both near the tropopause (e.g., 150 hPa) and just above the boundary layer (e.g., 850 hPa). We would expect these to be important because they allow the neural network to learn about the interactions between lower- and upper-level potential vorticity which can accelerate cyclone development (e.g., Hoskins, 2015).
- **Zonal wind:** Quasi-geostrophic theory (Pedder, 1997) tells us that, in order to determine regions of rising and falling air, we require knowledge of the wind fields at high altitude (e.g., 50 hPa) and low altitude (e.g., 1000 hPa), meaning our choice of levels is physically reasonable.

It should be noted, however, that just using physical reasoning alone would not have been sufficient to determine which pressure levels are good predictors. Firstly, in the case of specific humidity, we are unaware of any clear reasoning as to whether the important pressure levels should be low or high pressure. Furthermore, for other variables, whilst physical reasoning may suggest whether low and/or high pressure levels are important, the WeatherBench dataset has a number of levels which have “low” or “high” pressures. Our data-driven approach allows us to determine which and how many of these levels to choose, whereas physical reasoning does not.

This agreement with physical reasoning shows that neural network-based feature selection is able to identify physically important inputs, which include highly correlated time series data. The different pressure levels of a single variable are particularly highly correlated with each other and the neural network-based feature selection is able to identify which pressure levels are important for prediction. Both the agreement with key physical principles and the ability to deal with correlated data are identified in Schultz *et al.* (2021) as areas where neural networks need to prove themselves in order to be able to compete with numerical weather prediction models.