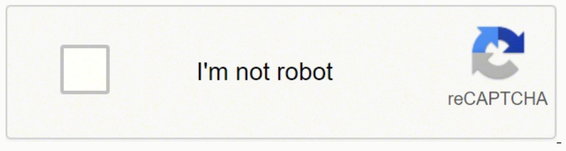

Continue

Machine learning algorithms for weather prediction

Photo by Karsten Würth on UnsplashPhysicists define climate as a “complex system”. While there are a lot of interpretations about it, in this specific case we can consider “complex” to be “unsolvable in analytical ways”.This may seems discouraging, but it actually paves the way to a wide range of numerical algorithms that aim to solve the climate challenges. With the computational developments of the last years, Machine Learning algorithms are certainly part of them.The challenge I want to discuss is based on forecasting the average temperature using traditional machine learning algorithms: Auto Regressive Integrated Moving Average models (ARIMA).While this post doesn’t want to be detailed in terms of the theoretical background, it does want to be a step-by-step guide on how to use these models in Python and apply it to real world data.So let’s start by describing the Python framework.0. The LibrariesThe libraries that have been used are the most famous ones for data analysis, plot and mathematical operations (pandas, matplotlib, numpy). Then there are some of them for advanced data visualization (like folium) and some of them are specific libraries for ARIMA models (like statsmodels). Here is the code for the import:1. The Dataset/Dataset ExplorationThe Dataset is open source and can be found here.If you want to know the cities in your dataset, select them by using this line of pandas:If we want to plot these cities in a world map, we need to slightly change the latitude and longitude. In order to do that, let’s use these few lines of code:And display the cities:2. Preprocessing, Advanced Visualization, StationarityI’ve chosen to isolate Chicago and consider the data of that city to be my dataset. There are no special reasons to do that... I just like Chicago :) . Of course you can use your own city and follow the next steps with your own dataset.The target is the AverageTemperature column, that is the Average Temperature for that specific month. We have data from 1743 to 2013.With this line we identify the NaN values and display them with a pie chart:As they are not a consistent part of the dataset, I’ve decided to fill the missing values with the previous ones. I did the same for the Average Temperature Uncertainty.The ‘dt’ column is the one that identifies the year and the month. For the next operations, it is handier to convert this column into a datetime object and to explicitly identify the year and the month in two different columns. We can do that by using the following lines:Using this dataset it is possible to obtain a scatter plot like this one:But it is not easy to read, so we should do something better.Now let’s describe three super-basic functions I created:get_timeseries(start_year,end_year) extract the portion of the dataset between the two yearsplot_timeseries(start_year,end_year) plots the timeseries extracted in get_timeseries in a readable wayplot_from_data(data, time, display_options) plots the data (AverageTemperature) wrt the time (dt) in a readable way. The display options permit to display the ticks, change the colors, set the label ...When it is done, we can make plots like this one:When we use ARIMA models, we should be considering stationary time series. In order to check if the timeseries we are considering is stationary, we can check the correlation and autocorrelation plots:It is suggesting us that the timeseries is not stationary. Nonetheless, if we perform the AD Fuller Test on the entire dataset it tells us that the dataset is stationary.But it is true just because we are looking at the entire dataset. In fact, if we analyze a single decade, it is clear that the dataset is absolutely not stationary for the decade period of time.In order to take account of this non-stationarity, a differentiation term will be considered in the ARIMA models.3. Machine Learning AlgorithmsLet’s consider the 1992–2013 decade and plot it:Performing the train/test split:Plotting the split:The Machine Learning algorithms are the ARIMA models. These are based on an optimization procedure that adopts the Maximum Likelihood function.The zero-differentiated ARIMA models are considered and evaluated using the AIC.While the first-differentiated models are considered by using these lines:The total summary is highlighted with this function and it shows that the (2,1,5) model and the (2,1,6) model are the best ones.As it is possible to see the statistical summary values are almost identicalAnd the same thing happens between the statistical plots :Nonetheless, it is preferable to use low index models both to avoid overfitting and reduce the computational stress on your computer. For this reason, the (2, 1, 5) model has been considered.4. ForecastingLet’s plot the results of the forecasting operation:And now let’s consider the specific predicted zone with the correspondent Uncertainty (the one given by the dataset) and the confidence interval (given by the algorithm):Finally, let’s consider a more readable version of the plot:5. ConclusionsThese methods are extremely easy to adopt as they don’t require any specific computational power like Deep Learning methods (RNN, CNN …).Nonetheless, predictions perfectly fit in the error range designed by the dataset itself. It is important to consider that we only have examined monthly average values while it may be interesting to consider daily values too and have daily predictions.If you liked the article and you want to know more about Machine Learning, or you just want to ask me something you can:A. Follow me on Linkedin, where I publish all my stories B. Subscribe to my newsletter. It will keep you updated about new stories and give you the chance to text me to receive all the corrections or doubts you may have.C. Become a referred member, so you won’t have any “maximum number of stories for the month” and you can read whatever I (and thousands of other Machine Learning and Data Science top writer) write about the newest technology available.Ciao! :) According to a 2009 study, U.S. adults look at weather forecasts nearly 300 billion times a year. Reliable forecasts can predict hazardous weather—such as blizzards, hurricanes, and flash floods—as early as 9–10 days before the event. Estimates value these forecasts at $31.5 billion per year. Although weather prediction keeps improving year to year for shorter-term forecasts, forecast skill decreases in the 2-week to 2-month time frame. These longer-timescale forecasts can play a critical role for many sectors, including water conservation, energy demand, and disaster preparedness. In a new study, Weyn et al. set out to improve this subseasonal to seasonal forecasting, as it is known, using a novel approach to weather prediction. Using a convolutional neural network, the authors developed a machine learning weather prediction system called Deep Learning Weather Prediction (DLWP). The model is trained on past weather data, which differs from standard numerical weather prediction models that create mathematical representations of physical laws. DLWP projects 2–6 weeks into the future for the entire globe. The authors compared the DLWP model to current state-of-the-art numerical weather models. The evaluation showed that the standard numerical forecasts perform better for short lead times, for instance, 2–3 weeks out. However, the DLWP model compared well when projecting 4–6 weeks down the line. Although the DLWP model does not yet rival existing models—and cannot currently forecast precipitation—the machine learning approach shows promise, the authors say. For one, the model is computationally more efficient than other approaches. The DLWP model requires only 3 seconds to compute an ensemble forecast consisting of 320 independent model runs. In addition, the model accurately provided a 4.5-day reforecast for Hurricane Irma using total column water vapor and without precipitation. The category 5 hurricane ravaged the northern Caribbean and the Florida Keys in 2017. The researchers say their study represents a significant step forward in the use of machine learning for weather and climate modeling. The authors observed that it was remarkable in its ability to learn physics-based phenomena. They also noted that the model could supplement forecasts in the tropics and in the spring and summer months, where most weather prediction models currently struggle. (Journal of Advances in Modeling Earth Systems (JAMES), 2021) —Aaron Sidder, Science Writer Citation: Sidder, A. (2022), The AI forecaster: Machine learning takes on weather prediction, Eos, 103, . Published on 6 January 2022.

The history of numerical weather prediction (NWP) and that of machine learning (ML) or artificial intelligence (for the purposes of this paper, the two terms can be used interchangeably) differ substantially. First, manual NWP was attempted by Lewis Fry Richardson in Britain in 1922, and early computer-aided weather forecasts were produced in 1950 [1]. These were soon followed by operational weather forecasts in Sweden, the USA and Japan. While there has been steady progress in the development of NWP (cf. [2]), the history of ML (cf. [3]) has been more disruptive: the first neural network (NN) was proposed in 1943 by McCulloch & Pitts [4]. The field expanded until the early 1960s, when the existing algorithms proved inefficient and unstable. The invention of backpropagation in 1970 [5,6] led to a second wave of ML applications as it became possible to build more extensive NNs and train them to recognize nonlinear relationships in data (e.g. [7]). Even though the development of ML algorithms continued, the enthusiasm about them soon dwindled again, because they rarely showed significant performance gain, and computing resources were not sufficient to solve larger problems. Furthermore, big amounts of labelled data which are mandatory for most data-driven ML approaches were hard to come by (note that this was before the advent of the world wide web). Three significant developments around 2010 started the third wave of artificial intelligence, which continues to the present: computing capabilities were vastly expanded due to massive parallel processing in graphical processing units (GPUs), convolutional neural networks (CNN) allowed much more efficient analysis of massive (image) datasets, and large benchmark datasets were made available on the internet. Highly complex NNs with greater than 106 parameters enabled a breakthrough in image recognition [8], soon followed by remarkable success stories in speech recognition [9], gaming [10,11], and video analysis and prediction [12,13]. Today’s NNs are often deep networks with greater than 10 layers, and the research field which develops such NNs and the associated methods for training and validation is called deep learning (DL).The weather and climate research community is increasingly aware of modern DL technologies and tries to adopt them to solve specific data analysis, numerical modelling and post-processing problems in the context of NWP. Nevertheless, as the workshop on ‘Machine learning in weather and climate’ (Oxford, September 2019) has also shown, there are still reservations about DL in this community. Two core arguments in this regard are the lack of explainability of deep NNs and the lack of physical constraints. Furthermore, some scepticism prevails due to the fact that researchers have experimented with rather simple NNs which were clearly unsuited to capture the complexity of meteorological data and feedback processes, but then extrapolate these results to discredit any NN application including the much more powerful DL systems. In their review of ‘Deep learning and process understanding for data-driven Earth system science’, Reichstein et al. [14] argue that traditional ML approaches might not be optimally suited to address the specific data challenges posed by Earth system data. They suggest that ‘deep learning methods are needed to cope with complex statistics, multiple outputs, different noise sources and high-dimensional spaces [of Earth system data]. New network topologies that not only exploit local neighbourhood (even at different scales) but also long-range relationships (for example, for teleconnections) are urgently needed, but the exact cause-and-effect relations between variables are not clear in advance and need to be discovered.’ As modern DL methods begin to deliver such concepts, we take this opportunity to expand on the analysis of [14] and explore the applicability of such methods to the NWP workflow (figure 1, left column). Figure 1. Idealized workflows of current numerical weather prediction (left), next-generation weather prediction with individual components substituted or augmented by ML and DL techniques (centre), and a purely data-driven DL weather forecasting system (right). (Online version in colour.) Download figureOpen in new tabDownload PowerPointWhile considerable work is undertaken to substitute specific parts of the NWP workflow with DL approaches (figure 1, central column), in this paper, we take a bold step forward and address the question of whether it is possible to replace all core parts of the NWP workflow with one deep NN, which would take observations as input and generate end-user forecast products directly from the data (figure 1, right column). Such approaches have been investigated for the specific application of wind speed and power predictions (cf. [15]), and there is at least one study, which successfully developed an end-to-end workflow to forecast multiple weather variables from the data of the US weather balloon network [16]. However, most of these studies were restricted to short-term forecasts and a few individual target sites, and none has yet attempted to explore the wealth of combined meteorological observations from the plethora of instruments and sensors, which is routinely used in operational weather forecasts.The goal of this paper is therefore to review the NWP workflow from a DL perspective and analyse the specific requirements and conditions of weather forecasting in light of current DL theory and practices. It is structured as follows: §2 gives a brief overview about the major developments and state of the art of NWP including aspects of data assimilation (DA) and model output processing. It is followed by §3 which surveys the literature on fundamental ML and DL developments and their application to weather and climate research. Section 4 discusses several fundamental aspects of meteorological data and other requirements of weather forecasts and points to corresponding solutions in DL research where these exist. Section 5 reflects on two aspects which are relevant for both weather forecasting and DL, but where we find different best practices in both domains. These aspects are data preparation and model evaluation. In §6, we reflect on the issues of physical constraints and system-wide forecast consistency in a DL framework. Section 7 discusses the state of the art with respect to estimating forecast uncertainties. Finally, §8 presents conclusions. We hope that this article will lead to a better understanding between ‘machine learners’ and ‘weather researchers’ and thus contribute to a more effective development of DL solutions in the field of weather and climate.Modern weather prediction relies extensively on massive numerical simulation systems which are routinely operated by national weather agencies all over the world. The associated process chain to generate these numerical weather forecasts can be divided into several steps which interact closely with each other (figure 1, left column).In order to retrieve the initial state of the Earth system (atmosphere, soil and ocean), a great variety of meteorological observations are collected all over the world. In addition to classical weather and radiosondes stations, aircraft measurements and remote sensing products (such as radar and satellite observations) have become an integral part of the global observation network (e.g. [17]). Although millions of different direct and indirect measurements are obtained every day, these observations are still not sufficient to describe the complete state of the atmosphere and other Earth system components with which the atmosphere exchanges energy or mass.At this point, DA comes into play. The central task of DA is to fill the gap between the incomplete, heterogeneous, and scattered observations and the initial value fields which are required by the NWP models. To achieve this, the observation data must be projected onto the discrete model grid, interpolated in time and adjusted to be consistent in terms of state variables (e.g. temperature, pressure, wind etc.). DA also has to take into account measurement errors such as biases between different space instruments or malfunctions of individual ground-based sensors. The obtained initial state of the Earth system after the DA step therefore constitutes an optimized estimate of the real conditions (e.g. [18]).Given the initial conditions, the NWP model can perform a simulation of atmospheric processes. By solving numerically the coupled partial differential equation system describing the atmosphere in terms of momentum, mass and enthalpy (the Navier–Stokes equations), the future atmospheric state is obtained in each model grid cell. Processes occurring at scales smaller than the model grid size are captured by empirical parameterizations. The direct model output constitutes the first forecast product of the NWP workflow. In contemporary global NWP models, the grid boxes cover an area of several square kilometres.In order to arrive at finer scale end-user forecast products, a post-processing step is added to the NWP workflow. The outcome of such post-processing can cover a variety of forecast products starting with the conversion of the vertical axis from sigma-coordinates to pressure levels or geometric height (above mean sea level) or bias corrections. Statistical methods are applied to remove systematic biases of the NWP output and to incorporate local scale adjustments (statistical downscaling). Furthermore, limited-area models which allow for finer grid spacings ($\Delta x \sim O(1$–$5\,\mathrm{km})$ compared to $\Delta x\, O \sim (10\,\mathrm{km})$ in global models) provide added value on forecasting meteorological features on finer scales. The output of ensemble simulations can be used to estimate forecast uncertainties which are of major interest especially for high impact weather situations, or for the renewable energy sector [19,20] (see also §7).Over the past decades, the ability of NWP models to predict the future atmospheric state has continuously improved. Contemporary global NWP models are not only able to predict the synoptic-scale weather pattern for several days, but they have also reached remarkable accuracy in forecasting end-user relevant meteorological quantities such as the 2 m temperature and regional-scale precipitation events. For instance, the deterministic forecasts of the Integrated Forecast System provided by the European Centre for Medium-Range Weather Forecasts maintains an anomaly correlation coefficient of the 500 hPa geopotential height of 80% for about 7 days, while the root-mean square error for 2 m temperature predictions of 72 h forecasts is close to 2 K [21]. Larger scale high-impact events such as hurricane tracks can be predicted with an accuracy of 150 km up to 4 days in advance [22].The increasing success of operational NWP models is due to improvements of all the steps involved in the NWP workflow and new capabilities of the global Earth observation system. In the following, we briefly highlight a couple of important developments which led to significant enhancements of forecast quality. A detailed review of recent advances in the NWP process chain is beyond the scope of this article.While in situ observations (weather stations and radiosondes) have a long history in observing the Earth’s atmosphere, fundamental improvements in the spatiotemporal coverage of observations has been achieved with the help of satellite data over recent decades (e.g [23]).Nowadays, several geostationary and polar-orbiting satellites deliver a great variety of data products (such as temperature and humidity profiles, soil moisture and atmospheric motion vectors). Satellite data are particularly valuable as they provide information on the atmosphere above the ocean and uninhabitated areas where conventional measurements are hard to come by. However, measurements from commercial aircraft (e.g. [24]) and radar observations (e.g. [25,26]) have also contributed to better constraining the initial state of NWP models.The ability of DA systems to make use of the manifold, diverse observations has seen continuous improvement due to algorithmic developments. Current DA systems are primarily based on three- or four-dimensional variational approaches (3D-Var and 4D-Var, respectively) and on ensemble methods (commonly Kalman filter). In the 3D-Var approach, a single deterministic state is estimated by minimizing a cost function which generally consists of the three terms background, observation, and model error. 4D-Var additionally captures observation changes in time (see [18] for more details).In order to obtain a loss function which can be optimized with reasonable efficiency, the model and observation operators have to be linearized. This can lead to forecast errors, in particular if the NWP model contains discontinuous parameterizations [27]. Another simplification of the variational DA approach is the a priori definition of the uncertainties of the state vector X which leads to a static background error covariance matrix. By contrast, an ensemble approach allows for dynamic estimation of the probability density function of X. The Ensemble Kalman Filter approach makes use of such an estimation which then results in a non-static, i.e. flow-evolving background error covariance matrix. A disadvantage of classical ensemble methods is that they are only conditioned on past measurements [28]. Therefore, leading meteorological centres have started to establish combinations of the variational approach with ensemble approaches such as the 4D-EnvVar DA method. The quality of ensemble DA depends on the number of ensemble members which is typically restricted to a relatively small number due to computational reasons. Therefore, so-called hybrid DA systems have been developed which include climatological error information in order to lessen the sensitivity to undersampling [29].NWP model improvements can in part be related to resolution enhancements. The continuous refinement of the grid spacing has also required re-formulating the dynamical cores of NWP models where the discretized Navier–Stokes equations are solved. Simulating the atmosphere on kilometre-scale comes along with the demand of highly parallelizable algorithms of the dynamical core [30]. Since (classical) global spectral transform models are less suited for such a requirement, finite-difference or finite-volume discretizations on platonic solids projected on the sphere (e.g. icosahedral [31] or cubed-sphere grids [32]) have been developed over the previous decade. Simultaneously, remarkable progress has been achieved in designing discretization approaches which enable the grid-scale dynamics to follow the conservation laws of energy, enthrophy and mass [33,34] while also minimizing the need for numerical filters to suppress artificial numerical modes [35]. An extensive overview of contemporary dynamical core architectures can be found in [32].In addition to the improvement of dynamical cores, further gains in accuracy have been achieved by fine-tuning physical parameterizations which are mandatory to represent atmospheric processes that cannot be captured by the grid-scale thermodynamics. Among others, these parameterizations encompass the representation of (deep) convection, turbulent mixing, smaller-scale atmosphere-land/ocean coupling, the representation of cloud microphysics and radiative transfer. Advances in capturing the diurnal cycle of convection (e.g. [36,37]), the turbulent transports in the planetary boundary layer (e.g. [38]) and in simulating the bulk properties of hydrometeors (e.g. [39]), i.e. clouds and precipitation, are only a small sample of recent progress in tuning physical parameterization schemes.The increased computational power, the availability of large datasets, and the rapid development of new NN architectures all contribute to the ongoing success of DL. Some of these new NN can solve certain ML tasks much more efficiently than the classical fully connected, feed-forward networks. One especially successful concept, which has been widely applied, is convolutional neural networks [40] (CNN), where a stack of small-sized filters with few trainable parameters is applied to images or other gridded data to extract coarser scale features. CNNs have been used in weather and climate applications, where the NN was trained to recognize spatial features, for example in the analysis of satellite imagery [41] or weather model output [42].The family of recurrent neural networks (RNN) was designed specifically for the learning of time-dependent features (i.e. text and speech recognition). More advanced RNN architectures are long short-term memory (LSTM) nodes [43,44] and gated recurrent units (GRU, [45]). LSTM and GRU cells can be embedded in more complex neural network architectures. For example, the combination of a normal CNN with LSTM yields the so-called ConvLSTM network [46].Two more recent DL concepts are variational auto-encoders (VAE) [47] and generative adversarial networks (GAN) [48]. Both of these are so-called generative models, i.e. they learn the data distributions from training samples and use generators to produce novel samples which match the characteristics of the training data. They are widely used in different applications such as image-to-image translation [49], super-resolution image generation [50], in-painting [51], image enhancement [52], image synthesis [53], style transfer and texture synthesis [54] and video generation and prediction [12,13]. VAEs use an encoder to project the high-dimensional data with posterior distribution into a latent space with lower dimensionality. This latent space is then sampled by a decoder to reconstruct the original feature space in all dimensions. For further information on VAE, we refer to [47,55]. In GANs, the competition of two NNs is used to improve on image generation during training. One network is optimized to generate realistic images, while a second one is trained concurrently to discriminate between generated and real images. Typically, both VAE and GAN-based architectures are coupled to multiple convolutional layers for capturing the semantic features of the input data and represent them with fewer dimensions. Examples are PixelVAE [56], DCGAN [57], sinGAN [58] and SAVP [13]. It is a general tendency in DL research that new NN architectures are composed of many building blocks which are themselves substantially large DL networks. The problem-complexity which can be addressed with modern DL networks is already quite substantial. The largest NNs have several million degrees of freedom which is comparable to operational NWP models.ML as ‘an approach to data analysis that involves building and adapting models, which allow programs to learn through experience’ has been employed by meteorologists for a long time, for example in curve fitting, linear regression, or DA (see §2). However, in this article, we focus on ML in a narrower sense, i.e. involving NNs and in particular modern DL.First studies employing NN for meteorological and air quality applications appeared during the 1990s [59–61]. These studies used multi-layer-perceptron architectures with typically three layers to analyse and forecast time series at individual station locations. Later, other simple semantic network techniques were used for post-processing and prediction optimization of NWP output [62,63], and as surrogate models for different parameterization schemes in climate models [64,65].It took a few years before the weather and climate research community started to pick up modern DL concepts and began to explore their use in NWP and other environmental applications. Table 1 lists various state-of-the-art DL architectures and their first applications in weather and climate research. A couple of examples are briefly described below. A review of ML in remote sensing can be found in [41]. Zhou et al. [83] and Denby et al. [84] used a CNN for classification of weather satellite images, while Xu et al. [85] used a combination of GAN and LSTM for prediction of cloud images. Based on the concept of video prediction, various types of networks were used for short-term prediction of sky images and radar images [46,79,81]. There have also been some attempts to produce data-driven weather forecasts, for example by Dueben & Bauer [86], who used a multi-layer perceptron, or Grover et al. [16], who constructed a three-stage model consisting of boosted decision trees, a dynamic Gaussian Process model, and a deep belief network consisting of restricted Boltzman machines [77,78]. The study of Wandel et al. [87] could be regarded as a first step towards replacing the dynamical core of a numerical weather model as they demonstrate unsupervised learning of the full incompressible Navier–Stokes equations on a Eulerian, grid-based representation. Gong et al. (A. B. Gong, unpublished manuscript, 2020), experimented with a state-of-the-art video prediction model and tried to use it for 2 m temperature predictions over 10 h. Table 1. State-of-the-art neural network architectures and their application in weather and climate research. Entries marked with * denote hybrid NN architectures.architectureintroduced for the first time (original references)early weather and climate applications (current state in July 2020)convolutional neural network (CNN)AlexNet (Alex et al. 2012 [8])VGG (Shi et al. 2018 [66])VGG (Simonyan & Zisserman, 2013 [67])ResNet (Pothineni et al. 2019 [68])ResNet (He et al. 2015 [69])Vgg, ResNet (Wen et al. 2020 [70])GoogleLeNet (Szegedy et al. 2015 [71])Inception blocks (Kleinert et al. 2021 [72])long short-term memory network (LSTM)RNN (Bengio et al. 1994 [73])LSTM (Gómez et al. 2003 [74])LSTM (Hochreiter and Schmidhuber, 1997 [43])LSTM (Qing & Niu, 2018 [75])PhyDNet (Le Guen, 2020 [76])deep belief network (DBN)Smolensky, 1986 [77], Hinton & Salakhutdinov [78]Grover et al. 2015 [16]variational autoencoder (VAE)Vanilla VAE, 2013 [47]Nonegenerative adversarial neural network (GAN)Vanilla GAN (Goodfellow et al. 2014 [48])MD-GAN (Xiong, 2018 [79])conditional GAN (Schmidt et al. 2020 [80])convolutional long short-term memory network (ConvLSTM)convLSTM (Shi et al. 2015 [46])ConvLSTM* (Shi et al. 2015 [46]),PredRNN (Wang et al. 2017 [81]); MIM (Wang et al. 2019 [82])stochastic adversarial video prediction (SAVP*)SAVP (Lee et al. 2018 [13])Gong et al. (A. B. Gong, unpublished manuscript, 2020)The studies that were cited in the previous section already demonstrate that DL concepts can be successfully applied to problems related to weather forecasting. However, the few existing attempts to replace the entire NWP workflow with a DL system have been limited to short-term forecasting (up to 24 h or less) or used a rather limited subset of the available meteorological data. In this section, we discuss a number of challenges which need to be overcome before a complete end-to-end DL weather forecasting system can deliver results of comparable quality as current NWP.Weather forecasting is essentially a prediction of spatiotemporal features based on a diverse array of observations from ground-based, airborne and satellite platforms. If we treat the core part of the NWP workflow (figure 1), i.e. DA, model forecast and output post processing, as one entity, then a weather forecast can be described as a function, which maps observation data to a final forecast product (figure 2). The forecast product can be a map of a specific weather variable (e.g. temperature), a time series of one or more variables at a specific location or aggregated over a region, some aggregate statistics of a specific variable over a given time range, a (categorical) warning index, etc. With current NWP, we are used to employing one forecasting system (which may well have several components) to derive the whole set of forecast products which are requested by end users, or needed for the system evaluation and further improvement. By contrast, DL applications excel if they can focus on a specific task, i.e. a reasonably small set of target variables. Therefore, an end-to-end DL weather forecast as depicted in the right column of figure 1 would likely consist of several deep NNs which would be trained individually on specific subsets of forecast products. Figure 2. Schematic view of the principal task of weather prediction, i.e. a mapping from diverse observation data to specific forecast products. Download figureOpen in new tabDownload PowerPointAdvantages of such a DL weather forecasting system could be the intrinsic absence of model bias (because the system would be trained to reproduce the target values) and the possible savings of computational resources. Once NNs are trained, they can very efficiently calculate forecasts with new data. Forward propagation in NNs consists only of fast add and multiply operations. Therefore, even NNs with O(108) parameters (i.e. similar complexity as contemporary NWP models) can be expected to use far less computing resources than current numerical models. The determining factors of the required computational resources in an end-to-end DL weather forecasting system are the necessary training cycles and the data processing. The former will depend on the learning approach (e.g. lifelong learning [88] requires regular re-training of some NN components) and on the success of transfer learning [89] concepts (i.e. whether it is possible to re-use NNs trained in one region of the globe for weather forecasts in another region). Data processing is a challenge which also limits further scaling of NWP models and other applications on current and next generation supercomputing systems (see [90]). At present, it is impossible to predict how much computing time, if any, could be saved if all weather forecasting would be based on DL. A fair comparison should always consider the entire weather prediction workflow and the whole range of forecast products that shall be generated.In the following, we discuss a number of challenges for end-to-end DL weather predictions, which are mostly a consequence of the specific properties of meteorological data and the complexity of the atmosphere and its interactions with other Earth system compartments. These challenges are graphically summarized in figure 3. As will be seen, many of these challenges also appear in other DL contexts, and the DL community has begun to develop solutions for these problems. Nevertheless, there are no systems in place which can cope with all of these challenges together. Figure 3. Graphical summary of the inter-connected DL challenges imposed by properties of meteorological data and the complexity of the weather forecasting problem. The colours of the jigsaw pieces are arbitrary. (Online version in colour.) Download figureOpen in new tabDownload PowerPointThe success of DL methods hinges on a good understanding of relevant data properties. Meteorological variables can be described with different cumulative distribution functions or corresponding probability density functions. Some variables (e.g. temperature) are nearly normally distributed, while others (e.g. precipitation and cloud droplet size) might be better approximated by gamma or beta distributions [91]. The fraction of cloud cover is often reported in eighths and therefore needs to be treated as a discrete variable. Ignoring the different properties of meteorological data can cause erroneous results when they are not accounted for in a statistical analysis or forecasting procedure. This is particularly relevant as some DL methods (e.g. the Bayesian approach by [92]) make implicit assumptions about the frequency distribution of variables.Meteorological features can show dynamic behaviour on a wide range of spatiotemporal scales, and the quality of the weather forecast is influenced by phenomena on many different scales [93]. Consider sea ice as an example, which may change little during the time of a typical forecast lead time, but whose mid- to longer-term variations can have a profound influence on the local and non-local atmospheric state [94]. Multiple spatial scales have been investigated in the context of video prediction (e.g. [12,58]).Related to the interaction of scales, the spectra of energy and momentum are important aspects in meteorology, but such spectra do not play a similar role in most mainstream DL applications. However, spectral transformations have already been used in DL applications (e.g. [9] in the context of speech recognition). First attempts at using such concepts in ML weather applications [95–98] were limited to simple NNs and limited complexity datasets.Many meteorological features vary periodically, although there can be large variability between cycles. This periodicity is induced by orbital parameters and the Earth rotation together with various solar cycles. As shown in Ziyin et al. [99], NNs generally have difficulties extrapolating periodic features correctly. However, as the authors show, replacing common monotonic activation functions with functions which include a periodic term can solve such problems and produce, for example, better temperature forecasts.Meteorological variables are correlated in space and time, and these correlations change with time [16]. For example, temperatures at different altitudes may exhibit very similar patterns (possibly with a time lag) in a well-mixed boundary layer (i.e. summer, daytime), while different vertical levels can be almost completely decoupled during an inversion (often during winter or at night). This can also be seen in larger scale features such as tropical cyclones, Rossby waves, fronts and (organized) convection. While we are not aware of a publication which addresses this issue in the context of DL with weather data, there are other DL studies which demonstrate that it is possible to cope with such correlations (e.g. [100]).A related property of meteorological features is auto-correlation. While auto-correlation in a way simplifies the forecasting task (at least on short time scales), it imposes the risk of over-estimating the forecasting capabilities of a DL model, in particular if this is not factored into the data preparation and model evaluation (see §5).Meteorological features may appear and vanish on time scales much shorter than the forecast range. Prominent examples are the triggering of convective cells or the transition between convective and relief precipitation in the presence of orography. An NWP model has some skill in predicting such features, because it can diagnose their potential occurrence from relations between other variables. In principle, such complex relations may be decipherable by NNs as well. However, we believe that this will require additional measures to make the NN aware of such relations. Such measures could be feature engineering (i.e. the calculation of derived properties from combinations of input variables) or the implementation of physical constraints (see §6).Another challenge for a DL weather forecast application is the scarcity of extreme events, which are, however, very important to get right as extreme weather phenomena have the largest impact on civil safety and economy. For example, to accurately predict heavy precipitation events (>25 $\mathrm{mm\,h^{-1}}$) over Germany, the DL model must be trained with less than 10 episodes at any given location during a full decade [101] A few studies have touched upon the subject of ML and extreme climate events: Vandal et al. [102] found that complex DL models have more problems in capturing extremes than classical statistical downscaling models, whereas O’Gorman et al. [103] state that their model captures extremes quite well without the need for special training on these cases. While the subject of classifying imbalanced data has received considerable attention (cf. [104]), there appears to be little research on dealing with imbalanced sample sizes in regression problems [105]. In contrast to standard DL algorithms, humans have acquired the ability to learn from isolated extreme events because they pay special attention to extraordinary occurrences in their environment and quickly generalize to other situations [106]. Some studies have explored the possibility to have deep neural networks learn extreme events in a similar way. For example, Li et al. [107] implemented one-shot learning of object categories by using prior knowledge learned from other training samples. Lake et al. [108] developed ML methods within Bayesian program learning (BPL) to mimic human capabilities of learning visual recognition from a few samples.Finally, other critical aspects related to meteorological observation data are the frequent appearance of missing values and the possibility of input data errors and biases. Current DA procedures often include a substantial code base for filtering or interpolating missing values, blacklisting observations, and monitoring of biases and their evolution over time. Similar issues occur in various application areas of DL. For example, Smieja et al. [109] have dealt with the issue of missing data values, and Žliobaitė [110] and Lu et al. [111] investigated the problem of drifting biases (known as concept drift in the ML community). A recent example of a DL application working on meteorological satellite data with missing values is Barth et al. [112].In this section, we discuss two aspects of ML in weather and climate, which we have found to be important in our practical experience and where best practices differ between the meteorological and ML communities. These are data preparation and model evaluation. This discussion may shed some light upon the reasons why it has been difficult for the DL community to tackle weather data problems and why, conversely, the meteorological community has been cautious to adopt DL in their research and for routine weather analyses and forecasts.All ML techniques are data-driven. Therefore, proper selection and preparation of data are essential to gain good and generalizable results. Data selection should aim to capture the full variability of the predictor variables, avoid too much redundant information, and allow the network to capture relations among variables, from which a prediction can be made. For the sake of brevity, we will not discuss data selection in more detail, but instead focus on data preparation aspects in the following.Modern supervised ML studies generally divide the available data into three different datasets to train, develop and evaluate an ML model [113]. The training set is the largest and is used to update the model parameters by back propagation or other learning algorithms. The second set, which is often referred to as the validation or development set, is used exclusively for hyper-parameter tuning. The hyper-parameters, i.e. number of layers, type of layer, activation function, loss function, learning rate etc. are set manually by the model developer. A key target of this hyper-parameter tuning is the optimization of the network’s generalization capabilities to ensure that the network will function well on previously unseen data. Both parameters and hyper-parameters are essential for building a suitable DL model. The third dataset is the test set, a collection of previously unseen data which is used to evaluate the network after the tuning to assess the true generalization capability of the network. The three datasets should be independent of each other, but at the same time they should reflect the same statistical distribution. Therefore, one has to be careful how to split the data before starting to train a new network, especially if, as in meteorological time series, the data are auto-correlated.Figure 4 shows four different data split strategies for a hypothetical time series of meteorological data. In order to enable an NN to forecast the next k time steps, one will generally feed an input vector of l past time steps to the network as input. In many DL applications, it is standard to draw random samples from a huge database of mutually independent data records (e.g. images) and arbitrarily assign these samples to the train, validation and test sets, respectively. This would correspond to figure 4a, where each drawn ‘slice’ has a length of one sample. However, as noted in §4, meteorological data constitutes a continuous time series with auto-correlation on different time scales. Therefore, randomly drawn samples would overlap and therefore no longer be independent. Consequently, results obtained with such a test set over-estimate the true generalization capability of the NN, because the test set contains information already used for training. When researching for this article, we found several studies on ML and DL for environmental data analysis, where this principle was violated and which therefore made overly optimistic conclusions concerning the capabilities of (often simple) NNs. Figure 4. Different train-dev-test splitting strategies for meteorological data with periodic features as indicated in the conceptual time series at the bottom of the figure. Every sand coloured block stands for 1 year of data. Case (a) depicts random sampling as is commonly applied in many DL applications. Cases (b–d) show different variants of random block sampling, which avoid spurious correlations between the train, val and test sets, if the block length chosen is long enough. (Online version in colour.) Download figureOpen in new tabDownload PowerPointAnother point of concern, which also has implications for the data preparation, is the multi-scale aspect of data in the time domain. While a typical forecast application considers time scales of a few hours to several days, there are longer-term quasi-periodic patterns, such as the El Niño Southern Oscillation (ENSO), and also continuous trends such as global warming. When training NNs with long-term data series (so that a sufficient number of samples becomes available), it is not trivial to find a good data split, which on the one hand fulfils the requirement of independence, while on the other hand allows the network to be trained on as many parts of the underlying data distribution as possible. For example, the model developer should ensure that all seasons are sampled appropriately, and, when using multi-year data, that the training data contains different phases of ENSO as one example out of many other oscillations.To solve these issues, we propose a random block sampling strategy (figure 4b–d), where the train, validation and test sets all contain several coherent blocks of length L, and L is much larger than the auto-correlation time. Multiple runs of this random block sampling should be carried out to assess the robustness of the results.The second aspect which we would like to discuss in this section concerns model evaluation. While it is relatively straightforward to evaluate the success of an image classification (e.g. [8,69]) or a video prediction task (cf. [12,79]), the evaluation metrics used in these studies (e.g. MSE or Peak Signal to Noise Ratio, PSNR) are usually not appropriate for weather and climate applications. Quantifying the success of a weather prediction model is highly non-trivial and an area of active research. Meteorological centres have developed a plethora of scores and skill scores over the last decades, which elucidate different aspects of weather forecast quality (cf. [21,91,114]). In addition to verification methods based on point by point comparisons (e.g. [115]), various methods have been introduced to account for the intrinsic spatial and temporal correlation in meteorological datasets (e.g. [116–118]). Other verification metrics also account for the stochastic nature of meteorological quantities by estimating probabilities of binary events, such as rain, no-rain [119]. Evaluating spatiotemporal patterns, for example, precipitation forecasts, with the help of radar data is particularly challenging due to the double penalty problem (cf. [116,120]). Indeed, verification of precipitation forecasts is still a hot topic in the meteorological community [121]. The evaluation of extreme events suffers from the ‘forecaster’s dilemma’, which discredits skilful forecasts when they are evaluated only under the condition that an extreme event occurred. This conditioning on outcomes and observations violates the theoretical assumptions of forecast verification methods [21,122].As has been demonstrated in other application areas of ML (e.g. [123,124]), NNs can be prone to learning spurious relationships in data. A purely data-driven model for weather forecasting might fail to adhere to the underlying physical principles and thus generate false forecasts as it lacks understanding of the fact that every atmospheric process obeys physical laws described in terms of conservation of momentum, enthalpy and mass. The incorporation of physical laws in NNs is becoming a vibrant area of research (e.g. [125–128]) and is now often denoted as scientific ML.One of the first studies to demonstrate that physical constraints can efficiently reduce systematic biases in lake temperature predictions, while at the same time enhancing generalization capability, was Karpatne et al. [129]. They included numerical model results as constraints for sparse observations and added a loss term to punish non-physical behaviour of the DL model. De Bézenac et al. [127] embedded an advection diffusion equation in their loss function to predict sea surface temperatures and thereby improved the predictive capabilities of their model. Le Guen et al. [76] introduced a new two-stream model and a recurrent cell, which is based on concepts from DA. The cell includes a physical predictor and a Kalman filter in order to assimilate the inputs. Effectively, this leads to the encoding of physical laws in the latent space of their model. Other approaches to introduce physical constraints into the latent space of DL models are the adversarial autoencoder by Makhzani et al. [130] and the non-parametric Bayesian method by Goyal et al. [131].It may be useful to reflect on the potential and necessity of physically constraining DL models from an abstract point of view. In spite of their complexity and dimensionality, DL models still adhere to the fundamental principles of (data-driven) statistical modelling. This implies that there must be some rules in place to constrain the future, because otherwise extrapolation will be unbound. Classical statistical modelling tries to strike a balance between a sufficiently explicit formulation of the time-dependent system evolution and the remaining degrees of freedom to accommodate the intrinsic variability of the data. For example, to fit hourly temperature observations, a classical statistical model will usually include (at least) two periodic terms to capture the diurnal and seasonal cycles. In addition, there may be terms to describe correlations of temperature with other variables. On the other hand, the statistician will avoid over-fitting the data by adding too many terms into the statistical model.Even though in DL one expects the NN to learn many of the inherent data relations by itself, the system nevertheless must get some guidance to be able to identify meaningful patterns. Knowingly or not, the researcher always imposes constraints on the NN, for example through data selection and choice of NN architecture. NNs learn faster when patterns are clearly visible in the data. However, with meteorological data the most obvious patterns are usually the least interesting ones, therefore it makes sense to let the NN know in advance that such patterns will occur. A similar argument applies to physical constraints: if the NN is forced, for example to conserve mass, it will not need to waste parameters and training cycles on learning this rule and can instead concentrate on analysing relevant and less obvious relations. Many meteorological studies show that it is often necessary to pre-select or filter data and build a good statistical model before meaningful relations become apparent. It is, therefore, likely that an end-to-end DL weather forecast system will only be successful if it contains at least some a priori knowledge in the form of engineered data features and physical constraints. Just how much of this is needed remains to be seen.Newcomers to the field of NWP often think that such numerical simulation models are inherently self-consistent, because they are based on a well-defined set of differential equations. However, the discretization of partial differential equations describing the flow dynamics in NWP models is not always fully mass or energy conserving, and parameterization schemes, which are needed to incorporate the effects of unresolved sub-grid scale processes on the grid-scale variables, may lead to spurious competition between grid-scale and sub-grid-scale processes, for example in cloud schemes [132,133]. Furthermore, there can be a grey zone between different parameterizations describing related aspects, such as the classical distinction of shallow and deep convection. Also, physically related parameterization schemes may be derived from different empirical data. Furthermore, internal consistency of classical NWP no longer applies if one considers the entire NWP workflow, i.e. if statistical models are used to post-process the model output, remove biases and apply other, non-physical corrections to the model forecasts. This discussion does not intend to devalue classical NWP, but it should inspire some reflection on the exact meaning and the value of consistency in the weather forecasting and DL communities. An end-to-end DL weather forecast system will generate consistency among forecast products only to the extent that this is already embedded in the data, unless the system will be governed by physical constraints as discussed above. To what extent consistency is needed to obtain a ‘good’ forecast will be a worthwhile question to study as it may deepen the understanding of the problem at hand and the potential which DL can bring to weather forecasting.The final discussion point of this article concerns the estimation of forecast uncertainty. Owing to increased computer power, it has become possible in recent years to produce ensemble forecasts operationally. Ensemble approaches have also been introduced in DA (see §2). In a nutshell, ensemble forecasts aim to estimate the probability density function of the forecast variables. Ensembles are most often generated by varying the initial conditions of the model simulation, but there are also attempts to sample the parameter space of empirical model parameterizations.In the field of DL research ensemble methods are used less often because they are computationally expensive. Statistical concepts such as Gaussian process (GP), and probabilistic graphical models (PGM) excel at probabilistic inference and uncertainty estimation. However, these methods do not scale well for high-dimensional and high-volume data [134–136]. Therefore, Bayesian deep learning (BDL) has been developed and applied across several scientific and engineering domains, for example in medical diagnosis [137] or autonomous driving [138]. In essence, these methods estimate a probability density function of the DL model parameters. As side effects BDL increases the robustness against over-fitting and allows training of the NN from relatively small sample sizes [139]. Modern BDL methods include variational inference, Markov Chain Monte Carlo (MCMC) sampling [136], and Monte Carlo dropout [140].Some recent studies explored the BDL concept for weather forecasting applications. A model built on GRU and 3D CNN, along with variational Bayesian inference for estimating posterior parameter distributions, has been presented by Liu et al. [141] for probabilistic wind speed forecasting of up to 3 h. A study from Vandal et al. [92] demonstrates the use of BDL to capture the uncertainty from observation data and unknown model parameters in the context of statistical downscaling of precipitation forecasts. These are relevant contributions, but a lot remains to be done before the uncertainty of DL weather forecasts can be assessed at a level similar to current NWP ensemble systems.In this article, we discussed the potential of modern DL approaches to develop purely data-driven end-to-end weather forecast applications. While there have been some stunning success stories from DL applications in other fields and initial attempts were made to apply DL to meteorological data, this research is still in its infancy. As we argue in §4, there are specific properties of weather data which require the development of new approaches beyond the classical concepts from computer vision, speech recognition, and other typical ML tasks. Even though DL solutions for many of these issues are being developed, there is no DL method up to now which can deal with all of these issues concurrently as it would be required in a complete weather forecast system.We expect that the field of ML in weather and climate science will grow rapidly in the coming years as more and more sophisticated ML architectures are becoming available and can easily be deployed on modern computer systems. What is largely missing in the field of meteorological DL are benchmark datasets with a specification of appropriate baseline scores and software frameworks which make it easy for the DL community to adopt a meteorological problem and try out different approaches. One notable exception is Weatherbench [142]. Such benchmark datasets and frameworks are well established in the ML community (e.g. MNIST [143] or ImageNet [144]) and they contributed substantially to the rapid pace of DL developments in application areas such as image recognition, video prediction, speech recognition, gaming and robotics. While a lot of meteorological data is freely available from weather centres and research institutions, proper use of these data requires some knowledge about Earth system science and the data formats and tools, which are used by the environmental research community. It might help if tools for reading and working with these datasets were integrated in major ML frameworks.When reflecting on the ultimate goal of replacing computationally expensive NWP models with DL algorithms, it is important to reconsider the objectives of weather forecasting and carefully define the requirements, which must be met by any potential alternative method. Certain criteria, which we now consider essential for a ‘good’ weather forecast, may in fact be conceptions, which are resulting from our experiences with numerical models, and they may not be applicable to forecasting systems based on DL. One particular aspect in this regard is self-consistency of forecast results, which is often taken for granted by numerical modellers, even though it is not strictly fulfilled in current NWP forecast systems. In this article, we consciously propose thinking about a replacement of the entire core NWP workflow including the DA, numerical modelling, and output processing, because the task of weather forecasting can then be described as a huge Big Data problem of mapping a plethora of Earth system observations onto a well-defined set of specific end-user weather forecast products. Seen in this way, the problem of weather forecasting is more amenable to DL methods than a replacement of the actual NWP model itself with its grid structure, operator concepts etc. which are tied to the very concept of classical numerical modelling.We expect that the success of DL weather forecast applications will hinge on the consideration of physical constraints in the NN design. Taken to the extreme, portions or variants of current numerical models could eventually end up as regulators in the latent space of deep neural weather forecasting networks. So, to answer the question posed in the title of this article, we can only say that there might be potential for end-to-end DL weather forecast applications to produce equal or better quality forecasts for specific end-user demands, especially if these systems can exploit small-scale patterns in the observational data which are not resolved in the traditional NWP model chain. Whether DL will evolve enough to replace most or all of the current NWP systems cannot be answered at this point.

This article has no additional data.M.G.S. designed and conceived the study. All authors jointly drafted, read and approved the manuscript.We declare we have no competing interests.C.B., B.G., F.K., L.H.L. and S.S. acknowledge funding from ERC-2017-ADG#787576 (IntelliAQ); B.G., F.K., M.L. and A.M. are funded through BMBF grant no. 01 IS 18047 (DeepRain). M.G.S. is funded by the programme Supercomputing & Big Data of the Helmholtz Association’s research field Key Technologies.The authors are grateful to two anonymous reviewers who provided very helpful comments.FootnotesOne contribution of 13 to a theme issue ‘Machine learning for weather and climate modelling’.

References1. Lynch P. 2008The origins of computer weather prediction and climate modeling. J. Comput. Phys. 227, 3431–3444. (doi:10.1016/j.jcp.2007.02.034) Crossref, ISI, Google Scholar2. Bauer P, Thorpe A, Brunet G. 2015The quiet revolution of numerical weather prediction. Nature 525, 47–55. (doi:10.1038/nature14956) Crossref, PubMed, ISI, Google Scholar3. Schmidhuber J. 2015Deep learning in neural networks: an overview. Neural Netw. 61, 85–117. (doi:10.1016/j.neunet.2014.09.003) Crossref, PubMed, ISI, Google Scholar4. McCulloch WS, Pitts W. 1943A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys. 5, 115–133. (doi:10.1007/BF02478259) Crossref, Google Scholar5. Linnainmaa S. 1970The representation of the cumulative rounding error of an algorithm as a taylor expansion of the local rounding errors. Master’s Thesis (in Finnish), Univ. Helsinki. Google Scholar6. LeCun Y. 1988A theoretical framework for back-propagation. In Proc. of the 1988 Connectionist Models Summer School, CMU, Pittsburg, PA (eds D Touretzky, G Hinton, T Sejnowski), pp. 21–28. Morgan Kaufmann. Google Scholar7. Rumelhart DE, Hinton GE, Williams RJ. 1986Learning Internal Representations by Error Propagation, pp. 318–362. Cambridge, MA: MIT Press. Google Scholar8. Krizhevsky A, Sutskever I, Hinton GE. 2012Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems 25 (eds F Pereira, CJC Burges, L Bottou, KQ Weinberger), pp. 1097–1105. Curran Associates, Inc. Google Scholar9. Amodei Det al.2016Deep speech 2: end-to-end speech recognition in english and mandarin. In Proc. of The 33rd Int. Conf. on Machine Learning, volume 48 of Proc. of Machine Learning Research, pp. 173–182. New York, NY: PMLR. (. Google Scholar10. Silver Det al.2016Mastering the game of go with deep neural networks and tree search. Nature 529, 484–489. (doi:10.1038/nature16961) Crossref, PubMed, ISI, Google Scholar11. Oh J, Guo X, Lee H, Lewis RL, Singh S. 2015Action-conditional video prediction using deep networks in atari games. In Advances in Neural Information Processing Systems 28 (eds C Cortes, ND Lawrence, DD Lee, M Sugiyama, R Garnett), pp. 2863–2871. Curran Associates, Inc. Google Scholar12. Mathieu M, Couprie C, LeCun Y. 2015Deep multi-scale video prediction beyond mean square error. (. Google Scholar13. Lee AX, Zhang R, Ebert F, Abbeel P, Finn C, Levine S. 2018Stochastic adversarial video prediction. (. Google Scholar14. Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N. 2019Deep learning and process understanding for data-driven earth system science. Nature 566, 195–204. (doi:10.1038/s41586-019-0912-1) Crossref, PubMed, ISI, Google Scholar15. Jiang Z, Jia Q, Guan X. 2017Review of wind power forecasting methods: From multi-spatial and temporal perspective. In 2017 36th Chinese Control Conference (CCC), pp. 10 576–10 583. (doi:10.23919/chicc.2017.8029042) Google Scholar16. Grover A, Kapoor A, Horvitz E. 2015A deep hybrid model for weather forecasting. In Proc. of the 21th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, KDD ’15, pp. 379–386. New York, NY: Association for Computing Machinery. (doi:10.1145/2783258.2783275) Google Scholar17. Thépaut JN, Andersson E. 2010The global observing system, pp. 263–281. Berlin, Heidelberg: Springer. (doi:10.1007/978-3-540-74703-1_10) Google Scholar18. Bannister RN. 2017A review of operational methods of variational and ensemble-variational data assimilation: ensemble-variational data assimilation. Q. J. R. Meteorol. Soc. 143, 607–633. (doi:10.1002/qj.2982) Crossref, ISI, Google Scholar19. Al-Yahyai S, Charabi Y, Gastli A. 2010Review of the use of numerical weather prediction (NWP) models for wind energy assessment. Renew. Sustain. Energy Rev. 14, 3192–3198. (doi:10.1016/j.rser.2010.07.001) Crossref, ISI, Google Scholar20. Pelland S, Galanis G, Kallos G. 2013Solar and photovoltaic forecasting through post-processing of the Global Environmental Multiscale numerical weather prediction model: solar and photovoltaic forecasting. Prog. Photovoltaics Res. Appl. 21, 284–296. (doi:10.1002/pip.1180) Crossref, ISI, Google Scholar21. Haiden T, Janousek M, Vitart F, Ferranti L, Prates F. 2019Evaluation of ECMWF forecasts, including the 2019 upgrade. (doi:10.21957/mlvapkke) Google Scholar22. Chen JHet al.2019Advancements in hurricane prediction with noaa’s next-generation forecast system. Geophys. Res. Lett. 46, 4495–4501. (doi:10.1029/2019GL082410) Crossref, ISI, Google Scholar23. Geer AJet al.2017The growing impact of satellite observations sensitive to humidity, cloud and precipitation. Q. J. R. Meteorol. Soc. 143, 3189–3206. (doi:10.1002/qj.3172) Crossref, ISI, Google Scholar24. Lange H, Janjic Pfander T. 2015Assimilation of mode-S EHS aircraft observations in cosmo-kenda. Mon. Weather Rev. 144, 151221072928005. (doi:10.1175/MWR-D-15-0112.1) ISI, Google Scholar25. Schraff C, Reich H, Rhodin A, Schomburg A, Stephan K, Periàñez A, Potthast R. 2016Kilometre-scale ensemble data assimilation for the cosmo model (kenda). Q. J. R. Meteorol. Soc. 142, 1453–1472. (doi:10.1002/qj.2748) Crossref, ISI, Google Scholar26. Wang H, Sun J, Zhang X, Huang XY, Auligné T. 2013Radar data assimilation with WRF 4D-Var. Part I: system development and preliminary testing. Mon. Weather Rev. 141, 2224–2244. (doi:10.1175/MWR-D-12-00168.1) Crossref, ISI, Google Scholar27. Vukićević T, Bao JW. 1998The effect of linearization errors on 4dvar data assimilation. Mon. Weather Rev. 126, 1695–1706. (doi:10.1175/1520-0493(1998)126¡1695:TEOLEO¿2.0.CO;2) Crossref, ISI, Google Scholar28. Steele-Dunne S, Entekhabi D. 2005An ensemble-based reanalysis approach to land data assimilation. Water Resour. Res. 41. (doi:10.1029/2004WR003449) PubMed, ISI, Google Scholar29. Zhang F, Zhang M, Hansen JA. 2009Coupling ensemble kalman filter with four-dimensional variational data assimilation. Adv. Atmos. Sci. 26, 1–8. (doi:10.1007/s00376-009-0001-8) Crossref, ISI, Google Scholar30. Zängl G, Reinert D, Rípodas P, Baldauf M. 2015The icon (icosahedral non-hydrostatic) modelling framework of DWD and MPI-M: description of the non-hydrostatic dynamical core. Q. J. R. Meteorol. Soc. 141, 563–579. (doi:10.1002/qj.2378) Crossref, ISI, Google Scholar31. Tomita H, Satoh M. 2004A new dynamical framework of nonhydrostatic global model using the icosahedral grid. Fluid Dyn. Res. 34, 357–400. (doi:10.1016/j.fluiddyn.2004.03.003) Crossref, ISI, Google Scholar32. Ullrich PA, Jablonowski C. 2012Mcore: A non-hydrostatic atmospheric dynamical core utilizing high-order finite-volume methods. J. Comput. Phys. 231, 5078–5108. (doi:10.1016/j.jcp.2012.04.024) Crossref, ISI, Google Scholar33. Ringler TD, Thuburn J, Klemp JB, Skamarock WC. 2010A unified approach to energy conservation and potential vorticity dynamics for arbitrarily-structured c-grids. J. Comput. Phys. 229, 3065–3090. (doi:10.1016/j.jcp.2009.12.007) Crossref, ISI, Google Scholar34. Wood Net al.2014An inherently mass-conserving semi-implicit semi-

lagrangian discretization of the deep-atmosphere global non-hydrostatic equations. Q. J. R. Meteorol. Soc. 140, 1505–1520. (doi:10.1002/qj.2235) Crossref, ISI, Google Scholar35. Gassmann A. 2018Discretization of generalized coriolis and friction terms on the deformed hexagonal c-grid. Q. J. R. Meteorol. Soc. 144, 2038–2053. (doi:10.1002/qj.3294) Crossref, ISI, Google Scholar36. Bechtold P, Semane N, Lopez P, Chaboureau JP, Beljaars A, Bormann N. 2014Representing equilibrium and nonequilibrium convection in large-scale models. J. Atmos. Sci. 71, 734–753. (doi:10.1175/JAS-D-13-0163.1) Crossref, ISI, Google Scholar37. Kwon YC, Hong SY. 2017A mass-flux cumulus parameterization scheme across gray-zone resolutions. Mon. Weather Rev. 145, 583–598. (doi:10.1175/MWR-D-16-0034.1) Crossref, ISI, Google Scholar38. Neggers RAJ, Köhler M, Beljaars ACM. 2009A dual mass flux framework for boundary layer convection. Part I: transport. J. Atmos. Sci. 66, 1465–1487. (doi:10.1175/2008JAS2635.1) Crossref, ISI, Google Scholar39. Seifert A, Beheng KD. 2006A two-moment cloud microphysics parameterization for mixed-phase clouds. Part 1: model description. Meteorol. Atmos. Phys. 92, 45–66. (doi:10.1007/s00703-005-0112-4) Crossref, ISI, Google Scholar40. LeCun Y, Haffner P, Bottou L, Bengio Y. 1999Object Recognition with Gradient-Based Learning, pp. 319–345. Berlin, Heidelberg: Springer. (doi:10.1007/3-540-46805-6_19) Google Scholar41. Zhu XX, Tuia D, Mou L, Xia GS, Zhang L, Xu F, Fraundorfer F. 2017Deep learning in remote sensing: a comprehensive review and list of resources. IEEE Geosci. Remote Sensing Mag. 5, 8–36. (doi:10.1109/MGRS.2017.2762307) Crossref, ISI, Google Scholar42. Gagne DJ, Haupt SE, Nychka DW, Thompson G. 2019Interpretable deep learning for spatial analysis of severe hailstorms. Mon. Weather Rev. 147, 2827–2845. (doi:10.1175/MWR-D-18-0316.1) Crossref, ISI, Google Scholar43. Hochreiter S, Schmidhuber J. 1997Long short-term memory. Neural Comput. 9, 1735–1780. (doi:10.1162/neco.1997.9.8.1735) Crossref, PubMed, ISI, Google Scholar44. Gers FA, Schmidhuber J. 2000Recurrent nets that time and count. In Proc. of the IEEE-INNS-ENNS Int. Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium, volume 3, pp. 189–194. Como, Italy, Italy. (doi:10.1109/IJCNN.2000.861302) Google Scholar45. Chung J, Gülçehre Ç, Cho K, Bengio Y. 2014Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRRabs/1412.3555. (. Google Scholar46. Xingjian S, Chen Z, Wang H, Yeung DY, Wong WK, Woo Wc. 2015Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In Advances in neural information processing systems, pp. 802–810. Google Scholar47. Kingma DP, Welling M. 2014Auto-Encoding Variational Bayes. CoRR. (. Google Scholar48. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. 2014Generative adversarial nets. In Advances in Neural Information Processing Systems, pp. 2672–2680. Curran Associates, Inc. Google Scholar49. Isola P, Zhu JY, Zhou T, Efros AA. 2017Image-to-image translation with conditional adversarial networks. In Proc. of the IEEE Conf. on Ccomputer Vision and Pattern Recognition, Vol. 1, pp. 5967–5976. (doi:10.1109/CVPR.2017.632) Google Scholar50. Johnson J, Alahi A, Fei-Fei L. 2016Perceptual losses for real-time style transfer and super-resolution. Lecture Notes in Computer Science, pp. 694–711. (doi:10.1007/978-3-319-46475-6_43) Google Scholar51. Pathak D, Krahenbuhl P, Donahue J, Darrell T, Efros AA. 2016Context encoders: Feature learning by inpainting. In 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 2536–2544. Los Alamitos, CA, USA: IEEE Computer Society. (doi:10.1109/CVPR.2016.278) Google Scholar52. Zhang H, Sindagi V, Patel VM. 2019Image de-raining using a conditional generative adversarial network. IEEE Transactions on Circuits and Systems for Video Technology p. 1–1. (doi:10.1109/tcsvt.2019.2920407) Google Scholar53. Reed S, Akata Z, Yan X, Logeswaran L, Schiele B, Lee H. 2016Generative adversarial text to image synthesis. In Proc. of The 33rd Int. Conf. on Machine Learning (eds MF Balcan, KQ Weinberger), volume 48 of Proc. of Machine Learning Research, pp. 1060–1069. New York, NY: PMLR. Google Scholar54. Li C, Wand M. 2016Precomputed real-time texture synthesis with markovian generative adversarial networks. Lecture Notes in Computer Science, pp. 702–716. (doi:10.1007/978-3-319-46487-9_43) Google Scholar55. Kingma DP, Welling M. 2019An introduction to variational autoencoders. (. Google Scholar56. Gulrajani I, Kumar K, Ahmed F, Taiga AA, Visin F, Vazquez D, Courville A. 2016Pixelvae: A latent variable model for natural images. (. Google Scholar57. Radford A, Metz L, Chintala S. 2015Unsupervised representation learning with deep convolutional generative adversarial networks. (. Google Scholar58. Shaham TR, Dekel T, Michaeli T. 2019SinGAN: Learning a generative model from a single natural image. 2019 IEEE/CVF Int. Conf. on Computer Vision (ICCV). (doi:10.1109/iccv.2019.00467) Google Scholar59. Schizas CN, Michaelides S, Pattichis CS, Livesay RR. 1991Artificial neural networks in forecasting minimum temperature (weather). In Second Int. Conf. on Artificial Neural Networks, volume 349, pp. 112–114. Bournemouth Int. CTR, Bournemouth, England: IEEE. Google Scholar60. Comrie AC. 1997Comparing neural networks and regression models for ozone forecasting. J. Air Waste Manage. Assoc. 47, 653–663. (doi:10.1080/10473289.1997.10463925) Crossref, ISI, Google Scholar61. Hall T, Brooks HE, Doswell CA. 1999Precipitation forecasting using a neural network. Weather Forecast. 14, 338–345. (doi:10.1175/1520-0434(1999)014¡0338:PFUANN¿2.0.CO;2) Crossref, ISI, Google Scholar62. Krasnopolsky VM, Lin Y. 2012A neural network nonlinear multimodel ensemble to improve precipitation forecasts over continental US. Adv. Meteorol. 2012, 1–11. (doi:10.1155/2012/649450) Crossref, ISI, Google Scholar63. Rasp S, Lerch S. 2018Neural networks for postprocessing ensemble weather forecasts. Mon. Weather Rev. 146, 3885–3900. (doi:10.1175/mwr-d-18-0187.1) Crossref, ISI, Google Scholar64. Krasnopolsky VM, Chalikov DV, Tolman HL. 2002A neural network technique to improve computational efficiency of numerical oceanic models. Ocean Modelling 4, 363–383. (doi:10.1016/S1463-5003(02)00010-0) Crossref, ISI, Google Scholar65. Krasnopolsky VM, Fox-Rabinovitz MS. 2006Complex hybrid models combining deterministic and machine learning components for numerical climate modeling and weather prediction. Neural Netw. 19, 122–134. (doi:10.1016/j.neunet.2006.01.002) Crossref, PubMed, ISI, Google Scholar66. Shi Y, Li Y, Liu J, Liu X, Murphey YL. 2018Weather recognition based on edge deterioration and convolutional neural networks. In 2018 24th Int. Conf. on Pattern Recognition (ICPR), pp. 2438–2443. IEEE. (doi:10.1109/ICPR.2018.8546085) Google Scholar67. Simonyan K, Zisserman A. 2014Very deep convolutional networks for large-scale image recognition. (. Google Scholar68. Pothineni D, Oswald MR, Poland J, Pollefeys M. 2019Kloudnet: Deep learning for sky image analysis and irradiance forecasting. In Pattern Recognition (eds T Brox, A Bruhn, M Fritz), pp. 535–551. Cham: Springer International Publishing. (doi:10.1007/978-3-030-12939-2_37) Google Scholar69. He K, Zhang X, Ren S, Sun J. 2016Deep residual learning for image recognition. In 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. Las Vegas, NV: IEEE. (doi:10.1109/CVPR.2016.90) Google Scholar70. Wen H, Du Y, Chen X, Lim E, Wen H, Jiang L, Xiang W. 2020Deep learning-based multi-step solar forecasting for PV ramp-rate control using sky images. IEEE Trans. Ind. Inf. 17, 1397–1406. (doi:10.1109/TII.2020.2987916) Crossref, ISI, Google Scholar71. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. 2015Going deeper with convolutions. In The IEEE Conf. on Computer Vision and Pattern Recognition (CVPR2015). (doi:10.1109/CVPR.2015.7298594) Google Scholar72. Kleinert F, Leufen LH, Schultz MG. 2021IntelliO3-ts v1.0: a neural network approach to predict near-surface ozone concentrations in Germany. Geosci. Model Dev. 14, 1–25. (doi:10.5194/gmd-14-1-2021) Crossref, ISI, Google Scholar73. Bengio Y, Simard P, Frasconi P. 1994Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Netw. 5, 157–166. (doi:10.1109/72.279181) Crossref, PubMed, Google Scholar74. Gómez P, Nebot A, Ribeiro S, Alquézar R, Mugica F, Wotawa F. 2003Local maximum ozone concentration prediction using soft computing methodologies. Syst. Anal. Model. Simul. 43, 1011–1031. (doi:10.1080/0232929031000081244) Crossref, Google Scholar75. Qing X, Niu Y. 2018Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM. Energy 148, 461–468. (doi:10.1016/j.energy.2018.01.177) Crossref, ISI, Google Scholar76. Le Guen V, Thome N. 2020Disentangling physical dynamics from unknown factors for unsupervised video prediction. (. Google Scholar77. Smolensky P. 1986Chapter 6: Information Processing in Dynamical Systems: Foundations of Harmony Theory, pp. 194–281. Cambridge, MA: MIT Press. Google Scholar78. Hinton GE, Salakhutdinov RR. 2006Reducing the dimensionality of data with neural networks. Science 313, 504–507. (doi:10.1126/science.1127647) Crossref, PubMed, ISI, Google Scholar79. Xiong W, Luo W, Ma L, Liu W, Luo J. 2018Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks. In 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. IEEE. (doi:10.1109/cvpr.2018.00251) Google Scholar80. Schmidt VH, Alghali M, Sankaran K, Yuan T, Bengio Y. 2020Modeling cloud reflectance fields using conditional generative adversarial networks. arXiv: Atmospheric and Oceanic Physics. Google Scholar81. Wang Y, Long M, Wang J, Gao Z, Yu PS. 2017PredRNN: recurrent neural networks for predictive learning using spatiotemporal LSTMs. In Proc. of the 31st Int. Conf. on Neural Information Processing Systems (NIPS 2017), pp. 879–888. Google Scholar82. Wang Y, Zhang J, Zhu H, Long M, Wang J, Yu PS. 2019Memory in Memory: a predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics. In The IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). (doi:10.1109/CVPR.2019.00937) Google Scholar83. Zhou Y, Wang H, Xu F, Jin YQ. 2016Polarimetric SAR image classification using deep convolutional neural networks. IEEE Geosci. Remote Sens. Lett. 13, 1935–1939. (doi:10.1109/LGRS.2016.2618840) Crossref, ISI, Google Scholar84. Denby L. 2020Discovering the importance of mesoscale cloud organization through unsupervised classification. GeoRL 47, e85190. (doi:10.1029/2019GL085190) ISI, Google Scholar85. Xu Z, Du J, Wang J, Jiang C, Ren Y. 2019Satellite image prediction relying on gan and lstm neural networks. In ICC 2019-2019 IEEE Int. Conf. on Communications (ICC), pp. 1–6. IEEE. (doi:10.1109/ICC.2019.8761462) Google Scholar86. Dueben PD, Bauer P. 2018Challenges and design choices for global weather and climate models based on machine learning. Geoscientific Model Dev. 11, 3999–4009. (doi:10.5194/gmd-11-3999-2018) Crossref, ISI, Google Scholar87. Wandel N, Weinmann M, Klein R. 2020Unsupervised deep learning of incompressible fluid dynamics. . Google Scholar88. Parisi GI, Kemker R, Part JL, Kanan C, Wermter S. 2018Continual lifelong learning with neural networks: a review. . Google Scholar89. Silver DL, Bennett KP. 2008Guest editor's introduction: special issue on inductive transfer learning. Mach. Learn. 73, 215–220. (doi:10.1007/s10994-008-5087-1) Crossref, ISI, Google Scholar90. Schultz M, Pleiter D, Bauer P eds. 2019, 2018. Extreme Data Workshop 2018, Forschungszentrum Jülich, 18–19 September 2018, ISBN 978-3-95806-392-1, Schriften des Forschungszentrums Jülich, IAS Series 40. Google Scholar91. Wilks DS2006Statistical methods in the atmospheric sciences, vol. 91, 2nd edn. International geophysics series. Amsterdam; The Netherlands: Academic Press. Google Scholar92. Vandal T, Kodra E, Dy J, Ganguly S, Nemani R, Ganguly AR. 2018Quantifying uncertainty in discrete-continuous and skewed data with Bayesian deep learning. In Proc. of the 24th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining, KDD-18, p. 2377–2386. New York, NY: Association for Computing Machinery. (doi:10.1145/3219819.3219996) Google Scholar93. Orlanski I. 1975A rational subdivision of scales for atmospheric processes. Bull. Am. Meteorol. Soc. 56, 527–530. ISI, Google Scholar94. Semmler T, Jung T, Serrar S. 2016Fast atmospheric response to a sudden thinning of Arctic sea ice. Clim. Dyn. 46, 1015–1025. (doi:10.1007/s00382-015-2629-7) Crossref, ISI, Google Scholar95. Partal T, Kişi Ö. 2007Wavelet and neuro-fuzzy conjunction model for precipitation forecasting. J. Hydrol. 342, 199–212. (doi:10.1016/j.jhydrol.2007.05.026) Crossref, ISI, Google Scholar96. Ramana RV, Krishna B, Kumar S, Pandey N. 2013Monthly rainfall prediction using wavelet neural network analysis. Water Resour. Manage 27, 3697–3711. (doi:10.1007/s11269-013-0374-4) Crossref, ISI, Google Scholar97. Solgi A, Nourani V, Pourhaghi A. 2014Forecasting daily precipitation using hybrid model of wavelet-artificial neural network and comparison with adaptive neurofuzzy inference system (case study: Verayneh station, Nahavand). Adv. Civil Eng. 2014, 1–12. (doi:10.1155/2014/279368) Crossref, Google Scholar98. Kisi O, Cimen M. 2012Precipitation forecasting by using wavelet-support vector machine conjunction model. Eng. Appl. Artif. Intell. 25, 783–792. (doi:10.1016/j.engappai.2011.11.003) Crossref, ISI, Google Scholar99. Ziyin L, Hartwig T, Ueda M. 2020Neural networks fail to learn periodic functions and how to fix it. . Google Scholar100. Cirstea RG, Micu DV, Muresan GM, Guo C, Yang B. 2018Correlated time series forecasting using deep neural networks: a summary of results. . Google Scholar101. 102. Vandal T, Kodra E, Ganguly AR. 2019Intercomparison of machine learning methods for statistical downscaling: the case of daily and extreme precipitation. Theor. Appl. Climatol. 137, 557–570. (doi:10.1007/s00704-018-2613-3) Crossref, ISI, Google Scholar103. O'Gorman PA, Dwyer JG. 2018Using machine learning to parameterize moist convection: potential for modeling of climate, climate change, and extreme events. J. Adv. Model. Earth Syst. 10, 2548–2563. (doi:10.1029/2018MS001351) Crossref, ISI, Google Scholar104. He H, Ma Y eds. 2013Imbalanced learning: foundations, algorithms, and applications, 1st edn. Hoboken, NJ: Wiley. (doi:10.1002/9781118646106) Crossref, Google Scholar105. Torgo L, Ribeiro RP, Pfahringer B, Branco P. 2013Smote for regression. In Progress in Artificial Intelligence (eds L Correia, LP Reis, J Cascalho), pp. 378–389. Berlin, Heidelberg: Springer. (doi:10.1007/978-3-642-40669-0_33) Google Scholar106. Lake BM, Baroni M. 2017Generalization without systematicity: on the compositional skills of sequence-to-sequence recurrent networks. Google Scholar107. Li Fei-Fei, Fergus R, Perona P. 2006One-shot learning of object categories. IEEE Trans. Pattern. Anal. 28, 594–611. (doi:10.1109/TPAMI.2006.79) Crossref, PubMed, ISI, Google Scholar108. Lake BM, Salakhutdinov R, Tenenbaum JB. 2015Human-level concept learning through probabilistic program induction. Science 350, 1332–1338. (doi:10.1126/science.aab3050) Crossref, PubMed, ISI, Google Scholar109. Smieja M, Łukasz Struski, Tabor J, Zieliński B, Spurek P. 2018Processing of missing data by neural networks. . Google Scholar110. žliobaitė I. 2010Learning under concept drift: an overview. . Google Scholar111. Lu J, Liu A, Dong F, Gu F, Gama J, Zhang G. 2018Learning under concept drift: a review. IEEE Trans. Knowl. Data Eng. 31, 2346–2363. (doi:10.1109/tkde.2018.2876857) ISI, Google Scholar112. Barth A, Alvera-Azcárate A, Licer M, Beckers JM. 2020Dincae 1.0: a convolutional neural network with error estimates to reconstruct sea surface temperature satellite observations. Geoscientific Model Dev. 13, 1609–1622. (doi:10.5194/gmd-13-1609-2020) Crossref, ISI, Google Scholar113. Goodfellow I, Bengio Y, Courville A. 2016Deep learning. Cambridge, MA: MIT Press. Google Scholar114. Casati Bet al.2008Forecast verification: current status and future directions. Meteorol. Appl. 15, 3–18. (doi:10.1002/met.52) Crossref, ISI, Google Scholar115. Murphy AH, Epstein ES. 1989Skill scores and correlation coefficients in model verification. Mon. Weather Rev. 117, 572–582. (doi:10.1175/1520-0493(1989)117¡0572:SSACCI¿2.0.CO;2) Crossref, ISI, Google Scholar116. Ebert EE. 2008Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework. Meteorol. Appl. 15, 51–64. (doi:10.1002/met.25) Crossref, ISI, Google Scholar117. Weniger M, Kapp F, Friederichs P. 2017Spatial verification using wavelet transforms: a review: spatial verification using wavelet transforms. Q. J. R. Meteorol. Soc. 143, 120–136. (doi:10.1002/qj.2881) Crossref, ISI, Google Scholar118. Buschow S, Pidstrigach J, Friederichs P. 2019Assessment of wavelet-based spatial verification by means of a stochastic precipitation model (wv_verif v0.1.0). Geoscientific Model Dev. 12, 3401–3418. (doi:10.5194/gmd-12-3401-2019) Crossref, ISI, Google Scholar119. Roberts NM, Lean HW. 2008Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. Mon. Weather Rev. 136, 78–97. (doi:10.1175/2007MWR2123.1) Crossref, ISI, Google Scholar120. Gilleland E, Ahijevych D, Brown BG, Casati B, Ebert EE. 2009Intercomparison of spatial forecast verification methods. Weather Forecast. 24, 1416–1430. (doi:10.1175/2009WAF2222269.1) Crossref, ISI, Google Scholar121. Dorninger M, Gilleland E, Casati B, Mittermaier MP, Ebert EE, Brown BG, Wilson LJ. 2018The setup of the MesoVICT project. Bull. Am. Meteorol. Soc. 99, 1887–1906. (doi:10.1175/BAMS-D-17-0164.1) Crossref, ISI, Google Scholar122. Lerch S, Thorarinsdottir TL, Ravazzolo F, Gneiting T. 2017Forecaster's dilemma: extreme events and forecast evaluation. Stat. Sci. 32, 106–127. (doi:10.1214/16-STS588) Crossref, ISI, Google Scholar123. Lazer D, Kennedy R, King G, Vespignani A. 2014The parable of google flu: traps in big data analysis. Science 343, 1203–1205. (doi:10.1126/science.1248506) Crossref, PubMed, ISI, Google Scholar124. Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller KR. 2019Unmasking clever Hans predictors and assessing what machines really learn. Nat. Commun. 10, 1–8. (doi:10.1038/s41467-019-08987-4) Crossref, PubMed, ISI, Google Scholar125. Karpatne A, Atluri G, Faghmous JH, Steinbach M, Banerjee A, Ganguly A, Shekhar S, Samatova N, Kumar V. 2017Theory-guided data science: a new paradigm for scientific discovery from data. IEEE Trans. Knowl. Data Eng. 29, 2318–2331. (doi:10.1109/TKDE.2017.2720168) Crossref, ISI, Google Scholar126. Jia X, Willard J, Karpatne A, Read J, Zwart J, Steinbach M, Kumar V. 2019Physics guided rnns for modeling dynamical systems: a case study in simulating lake temperature profiles. In Proc. of the 2019 SIAM Int. Conf. on Data Mining, pp. 558–566. SIAM. (doi:10.1137/1.9781611975673.63) Google Scholar127. de Bézenac E, Pajot A, Gallinari P. 2019Deep learning for physical processes: incorporating prior scientific knowledge. J. Stat. Mech: Theory Exp. 2019, 124009. (doi:10.1088/1742-5468/ab3195) Crossref, ISI, Google Scholar128. Raissi M, Perdikaris P, Karniadakis GE. 2019Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. J. Comput. Phys. 378, 686–707. (doi:10.1016/j.jcp.2018.10.045) Crossref, ISI, Google Scholar129. Karpatne A, Watkins W, Read J, Kumar V. 2017Physics-guided neural networks (PGNN): an application in lake temperature modeling. Google Scholar130. Makhzani A, Shlens J, Jaitly N, Goodfellow I, Frey B. 2015Adversarial autoencoders. Google Scholar131. Goyal P, Hu Z, Liang X, Wang C, Xing EP. 2017Nonparametric variational auto-encoders for hierarchical representation learning. In Proc. of the IEEE Int. Conf. on Computer Vision, pp. 5094–5102. (doi:10.1109/ICCV.2017.545) Google Scholar132. Arakawa A, Wu CM. 2013A unified representation of deep moist convection in numerical modeling of the atmosphere. Part I. J. Atmos. Sci. 70, 1977–1992. (doi:10.1175/JAS-D-12-0330.1) Crossref, ISI, Google Scholar133. Wu CM, Arakawa A. 2014A unified representation of deep moist convection in numerical modeling of the atmosphere. Part II. J. Atmos. Sci. 71, 2089–2103. (doi:10.1175/JAS-D-13-0382.1) Crossref, ISI, Google Scholar134. Zhu Y, Zabaras N, Koutsourelakis PS, Perdikaris P. 2019Physics-constrained deep learning for high-dimensional surrogate modeling and uncertainty quantification without labeled data. J. Comput. Phys. 394, 56–81. (doi:10.1016/j.jcp.2019.05.024) Crossref, ISI, Google Scholar135. Wang H, Yeung DY. 2016Towards Bayesian deep learning: A survey. Google Scholar136. Wang H, Yeung D. 2016Towards Bayesian deep learning: a framework and some existing methods. IEEE Trans. Knowl. Data Eng. 28, 3395–3408. (doi:10.1109/TKDE.2016.2606428) Crossref, ISI, Google Scholar137. Gal Y, Islam R, Ghahramani Z. 2017Deep Bayesian active learning with image data. In Proc. of the 34th Int. Conf. on Machine Learning (ed. D Precup, YW Teh), volume 70 of Proc. of Machine Learning Research, pp. 1183–1192. International Convention Centre, Sydney, Australia: PMLR. Google Scholar138. McAllister R, Gal Y, Kendall A, Van Der, Shah A, Cipolla R, Weller A. 2017Concrete problems for autonomous vehicle safety: Advantages of Bayesian deep learning. In Proc. of the Twenty-Sixth Int. Joint Conf. on Artificial Intelligence. International Joint Conferences on Artificial Intelligence, Inc. (doi:10.24963/ijcai.2017/661) Google Scholar139. 140. Gal Y, Ghahramani Z. 2016Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Proc. of The 33rd Int. Conf. on Machine Learning (ed. MF Balcan, KQ Weinberger), volume 48 of Proc. of Machine Learning Research, pp. 1050–1059. New York, NY: PMLR. Google Scholar141. Liu Y, Qin H, Zhang Z, Pei S, Jiang Z, Feng Z, Zhou J. 2020Probabilistic spatiotemporal wind speed forecasting based on a variational Bayesian deep learning model. Appl. Energy 260, 114259. (doi:10.1016/j.apenergy.2019.114259) Crossref, ISI, Google Scholar142. Rasp S, Dueben PD, Scher S, Weyn JA, Mouatadid S, Thuerey N. 2020Weatherbench: a benchmark dataset for data-driven weather forecasting. . Google Scholar143. LeCun Y, Bottou L, Bengio Y, Haffner P. 1998Gradient-based learning applied to document recognition. Proc. IEEE 86, 2278–2324. (doi:10.1109/5.726791) Crossref, ISI, Google Scholar144. Russakovsky Oet al.2015Imagenet large scale visual recognition challenge. Int. J. Comput. Vision 115, 211–252. (doi:10.1007/s11263-015-0816-y) Crossref, ISI, Google ScholarPage 2Theme issue 'Machine learning for weather and climate modelling' compiled and edited by Matthew Chantry, Hannah Christensen, Peter Dueben and Tim PalmerKeywordsclimate modellingmachine learningclimateSubjectsPage 3Artificial intelligence (AI) and machine learning (ML) show promise for improving modelling and forecasting for a host of problems. Environmental science is one of many applications of this useful technology [1]. Although weather and climate have been traditionally modelled using dynamical and physical models built from first principles, more empirical methods have also proven useful; thus, it is natural that AI/ML would find applications in this field. Hereafter, we will use AI to encompass ML in our terminology.A workshop on Machine Learning for Weather and Climate was convened at Oxford, UK, in September 2019 to assess the state of the science, evaluate progress, and propose next steps along the pathway to realize the potential of AI in the atmospheric sciences. Some of the first lectures segmented the research broadly into three primary groups: post-processing, emulating processes and using ML to build full models [2,3].The workshop provided time and space to discuss each of these topics more broadly. Most of these coauthors became part of the working group assessing opportunities for AI to improve the output of environmental science models, known as post-processing, while the rest contributed to the on-going conversation and effort to archive datasets to help advance the science. This group not only assessed the successes of the environmental science community in leveraging AI for post-processing to date but also discussed the importance of disclosing failures as a measure to help advance the science more rapidly. The authors believe that a vigorous effort should be made to explore and validate modern AI methods. We see a host of opportunities to further improve numerical weather prediction (NWP) forecasts and climate projections at a minimal cost when compared with other model development efforts. We suggest what is needed to move forward, discuss what will constitute success and make some concrete recommendations for the next steps, including beginning an archive of example problems that can be used to test emerging methods.Model post-processing corrects systematic errors in model output by comparing hindcasts to observations. This is becoming increasingly important, but also challenging, as NWP model resolution has increased to the point that it attempts to resolve hyper-local effects and structures with a stochastic nature. Similarly, in climate projections, there is a drive towards more localized information, which is inherently uncertain. For context, NWP forecast systems, through model improvements and assimilation of additional observational data, have historically achieved a root-mean-square error (RMSE) skill improvement of approximately one day every 10 years [4]. However, this skill has arguably been attributable to increases in supercomputing power that has enabled higher model resolution and more comprehensive data assimilation [5]. Unfortunately, this progress is unlikely to continue under the death of Moore's Law (. In this context, model post-processing becomes yet more important to help drive skill improvements at uncertain length-scales and with ever more limited compute resources. AI is a prime candidate for developing more powerful post-processing approaches that can represent cheap transfer functions in fractions of the development time of traditional approaches (e.g. [6]).AI also brings the ability to optimize output for specific tasks by choosing appropriate loss functions. The same set of NWP forecasts may be post-processed in different ways according to the needs of particular end-users and the decisions they have to make. Built on statistical foundations, AI post-processing systems not only have the capability to correct biases and phase shifts in numerical forecasts, but also have the potential to quantify forecast uncertainty—both epistemic (due to lack of knowledge) and aleatoric (natural randomness of a process)—more comprehensively than NWP approaches, and in doing so provide better information for decision support. In this sense, AI post-processing can act as a bridge between the physical representation of the atmosphere provided by NWP and the decision-making requirements of end-users. One must also recognize the observation error in the 'truth' data to which the AI is trained. If that error is systematic, AI will often discover and correct it. Even if that error is aleatoric, AI can learn a correction on average to minimize the error.This manuscript reports on the state of the science of AI post-processing for weather and climate and provides a foundation for further progress through recommending a repository of methods and data that can enable the community to move forward. Section 2 provides a brief history of the development and use of AI for post-processing weather and climate model output without attempting to be comprehensive. The working group considered the current challenges and how the community might most effectively address them, including setting some medium-term goals as discussed in §3. Section 4 considers what successful application of AI post-processing in weather and climate will look like. The workshop attendees decided to make a distinct impact through concrete deliverables as laid out in §5. In particular, we describe the need for a repository to provide common assessment tools and datasets for ML scientists to test methods and set the stage for intercomparison. The repository and initial datasets are described. Section 6 summarizes and provides some concluding thoughts.Although the dynamic models of weather and climate have formed the basis for prediction, the community has long recognized the value of post-processing the forecasts to improve accuracy and to quantify uncertainty.Global Climate Models (GCMs) and NWP models provide the atmospheric variables necessary to determine predicted atmospheric states based on numerical integration of a discretized version of the Navier–Stokes equations [7]. However, due to uncertainty in initial conditions and numerical approximation as well as the non-linearity of the system, the chaotic error tends to swamp skill from initial information [8]. In addition, model deficiencies add systematic error and insufficient observations put a limit on the resolution of initial conditions. For as long as NWP forecasts have been officially issued there have been attempts to statistically correct these methods, given observational data (e.g. [9]). This can be viewed directly as a supervised machine learning task. Current weather forecasting centres, including the UK Met Office, US National Center for Environmental Prediction (NCEP), European Center for Medium-Range Forecasting (ECMWF), and many others rely on statistical methods that have been proven successful. The initial methods employed multilinear regressions and became known as Model Output Statistics (MOS [9]). These systems expanded to treat ensembles and became Ensemble Model Output Statistics (EMOS) [10,11]. These statistical learning methods have been used in practice since 1968 by the U.S. National Weather Service to improve systematic model error (e.g. [12,13]). These methods are continually refined with new observational data and show major skill improvement in correcting forecasts from 0 to 10 days [14]. MOS and EMOS, however, are inherently linear techniques that are notoriously rigid and require significant tuning (e.g. specification of predictive distribution and estimation of the parameters, e.g. the mean and the standard deviation in the case of a Gaussian distribution). AI methods, which typically allow for the resolution of complex nonlinear processes, open up opportunities for more effective corrections.Despite the success of statistical corrections, weather forecasting has a tradition of human forecasters weighing the relative merits of the various models according to the situation. AI came into play as the private sector began to forecast beyond a single country, and it became obvious that human forecasters could no longer do corrections from experience for the entire globe. The Weather Company realized this in the late 1990s and collaborated with the National Center for Atmospheric Research (NCAR) to develop the Dynamic Integrated foreCast system (DICast®) that learns the appropriate weights for input models given paired historical forecasts and observations [15]. Figure 1 demonstrates the DICast post-processing methodology, which is representative of the many other systems currently being used. DICast has evolved over time to include additional machine-learning methods and has been shown to dramatically improve forecasts across multiple weather-dependent applications including road conditions [16], precision agriculture, wind and solar energy [17–20], among others. Now, many commercial weather companies and national centres employ AI-based post-processing methods [21]. Figure 1. DICast post-processing progresses in a two-step process using historical forecasts: 1) bias-correcting each model's input using any of a number of MOS-like methods and 2) determining optimal weighting for each model for each forecast time and each lead time [15]. (Online version in colour.) Download figureOpen in new tabDownload PowerPointTo deal with the aforementioned inherent uncertainty in NWP forecasts, the national centres now run ensembles of model forecasts with perturbations to initial conditions or other methods of initiating perturbations in the simulations [22–25]. As with deterministic forecasts, these ensemble forecasts may have biases in their mean and the spread may not be calibrated against the actual uncertainty. Thus, methods to ameliorate these problems have been devised to bias correct the mean deterministic value as well as to calibrate the spread. The first techniques developed were statistical methods such as the EMOS described earlier [10,11], quantile regression [26,27], Bayesian model averaging [28], linear variance calibration [29,30], among others. This challenge also can be met through the application of AI methods. Some of these methods involve identifying regimes using some clustering or another method to identify similar past forecasts. Hamill and Whittaker [31] and Hamill et al. [32] describe an analogue approach to calibrating ensembles of precipitation forecasts. Greybush et al. [33] describe a multi-step approach that first splits the forecasts into regimes using principal component analysis then applies AI-based methods to distinguish among weather regimes to produce weighted consensus forecasts of surface temperature. McCandless et al. [34] tested multiple AI methods for improving ensemble member weighting for predicting snowfall accumulation.Several methods have been used to more directly provide probabilistic information with AI approaches. Krasnopolsky [35] reviews the use of Artificial Neural Networks (ANNs) to form ensembles for various applications. He compares nonlinear ANN approaches to linear ones and demonstrates marked improvements using the nonlinear approaches for several variables. Evolutionary programming (EP) has also proven useful for generating AI ensembles. EP methods have been used to evolve ensembles, demonstrating that smaller temperature RMSEs and higher Brier Skill Scores could be generated than with a 21-member operational ensemble [36]. This method is also useful for minimum temperature forecasts, then demonstrated further improvements for adaptive methods [37,38]. The analog ensemble (AnEn) method has arisen as a machine-learning technique to directly predict both deterministic forecast values and to quantify uncertainty directly from a single high-quality NWP run of sufficient length [39]. The AnEn uses a time series of historical forecast variables and their corresponding verifying observations. For each current forecast being made, the AnEn looks back in the historical record to find the n most similar forecasts. The verifying observations then become an n-member ensemble that is used to estimate the uncertainty of the forecast. The mean of that ensemble becomes the improved forecast value. The AnEn has been shown to improve upon raw ensemble output as well as upon some common statistical methods [39].AI methods have been highly used in applications that derive from NWP forecasts. For instance prediction of severe weather has seen a plethora of AI methods applied to improve predication [40]. Random forests [41], for instance, can model nonlinear relationships including arbitrary predictors while being robust to overfitting. In weather post-processing, quantile regression forest models have been proposed by Taillardat et al. [42] and extended to include combinations with parametric approaches [43]. The prediction of mesoscale convective areas has been shown to be successful with decisions trees by Gagne et al. [44] and Ahijevych et al. [45]. Gradient boosted regression trees proved the most accurate method for predicting storm duration and forecasting severe wind [46]. Gagne et al. [47,48] have applied machine learning methods including random forests and gradient boosted regression to predict the probability of severe hail.Where applications have financial implications may be where AI has been applied most frequently to improve forecasts. For instance, as more renewable energy is being deployed, it becomes increasingly important to accurately predict the daily variations in wind speed and solar irradiance directly at the plants using local observations. AI combined with NWP models has proven to be a best practice to estimate the timing of changes [17,19,49]. Methods such as autoregressive models, Artificial Neural Networks, Support Vector Machines, and blended methods have shown success at providing nonlinear corrections to models [15,50,51]. Such techniques can improve upon a forecast by 10–15% over the best model forecast [15,19,49]. Probabilistic forecasts of these variables are also important to industry [52–54]. The AnEn described above has proven useful for predicting both wind [55] and solar power [56].Beyond the NWP forecasting timescales of 10–14 days, forecast centres are increasingly providing subseasonal and seasonal-scale forecasts. At seasonal to decadal timescales, initialized coupled climate models are used to skilfully forecast shifts in regional climates [57]. However, this skill relies on extensive post-processing to correct for regression from initialized climatology to model resting climatology known as 'bias-correction' [58].A growing area of research is to use known knowledge of physics in terms of physical laws and conservation properties in machine learning algorithms to constrain the training and improve the algorithms further. For example, a generative adversarial network (GAN)-based model for simulating turbulent flows can be further improved by incorporating physical constraints, e.g. energy spectra [59], in the loss function. Convolutional neural network (CNN)-based models for parameterizing subgrid-scale physics can be further improved to represent the mean climate by constraining global conservation properties e.g. conservation of momentum [60].Climate model simulations routinely have spatial and temporal (e.g. seasonal) biases with respect to observations, some of which are systematic across climate models (e.g. Southern Ocean warm bias as described in [61,62]). When constructing future climate predictions using available simulations run under set future emission scenarios, it is important that any model biases are cor

original author and source are credited.References1. Haupt SE, Pasini A, Marzban C (eds) 2009Artificial intelligence methods in the environmental sciences, 424 pp. Berlin, Germany: Springer. Google Scholar2. Brenowitz ND, Bretherton CS. 2018Prognostic validation of a neural network unified physics parameterization. Geophys. Res. Lett. 45, 6289–6298. (doi:10.1029/2018GL078510) Crossref, ISI, Google Scholar3. Dueben PD, Bauer P. 2018Challenges and design choices for global weather and climate models based on machine learning. Geoscientific Model Development 11, 3999–4009. (doi:10.5194/gmd-11-3999-2018) Crossref, ISI, Google Scholar4. Magnusson L, Källén E. 2013Factors influencing skill improvements in the ECMWF forecasting system. Mon. Wea. Rev. 141, 3142–3153. (doi:10.1175/MWR-D-12-00318.1) Crossref, ISI, Google Scholar5. Bauer P, Thorpe A, Brunet G. 2015The quiet revolution of numerical weather prediction. Nature 525, 47–55. (doi:10.1038/nature14956) Crossref, PubMed, ISI, Google Scholar6. Rasp S, Lerch S. 2018Neural networks for postprocessing ensemble weather forecasts. Mon. Wea. Rev. 146, 3885–3900. (doi:10.1175/MWR-D-18-0187.1) Crossref, ISI, Google Scholar7. Richardson LF. 2007Weather prediction by numerical process. Cambridge (University Press), 1922. 4°. Pp. xii + 236. 30s.net. Q. J. R. Meteorol. Soc. 48, 282–284 (doi:10.1002/qj.49704820311) Crossref, Google Scholar8. Lorenz EN. 1969The predictability of a flow which possesses many scales of motion. Tellus. 21, 289–307. (doi:10.1111/j.2153-3490.1969.tb00444.x) Crossref, Google Scholar9. Glahn HR, Lowry DA. 1972The use of model output statistics (MOS) in objective weather forecasting. J. Appl. Meteor. 11, 1203–1211. (doi:10.1175/1520-0450(1972)0112.0.CO;2) Crossref, Google Scholar10. Woodcock F, Engel C. 2005Operational consensus forecasts. Weather Forecast. 20, 101–111. (doi:10.1175/WAF-831.1) Crossref, ISI, Google Scholar11. Gneiting T, Raftery AE, Westveld AH, Goldman T. 2005Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Mon. Wea. Rev. 133, 1098–1118. (doi:10.1175/MWR2904.1) Crossref, ISI, Google Scholar12. Carter GM, Dallavalle JP, Glahn HR. 1989Statistical forecasts based on the National Meteorological Center's numerical weather prediction system. Weather Forecast. 4, 401–412. (doi:10.1175/1520-0434(1989)0042.0.CO;2) Crossref, Google Scholar13. Wilks DS, Hamill TM. 2007Comparison of ensemble-MOS methods using GFS reforecasts. Mon. Wea. Rev. 135, 2379–2390. (doi:10.1175/MWR3402.1) Crossref, ISI, Google Scholar14. Hemri S, Scheuerer M, Pappenberger F, Bogner K, Haiden T. 2014Trends in the predictive performance of raw ensemble weather forecasts. Geophys. Res. Lett. 41, 9197–9205. (doi:10.1002/2014GL062472) Crossref, ISI, Google Scholar15. Myers W, Wiener G, Linden S, Haupt SE. 2011A consensus forecasting approach for improved turbine hub height wind speed predictions. In AWEA Windpower Conf. and Exhibition 2011. Anaheim, CA: American Wind Energy Association (AWEA). See 3296. Google Scholar16. Siems-Anderson AR, Walker CL, Wiener G, Mahoney WP, Haupt SE. 2019An adaptive big data weather system for surface transportation. Transport. Res. Interdiscipl. Perspect. 3, 100071. (doi:10.1016/j.trip.2019.100071) Crossref, Google Scholar17. Mahoney WPet al.2012A wind power forecasting system to optimize grid integration. IEEE Trans. Sustain. Energy. 3, 670–682. (doi:10.1109/TSTE.2012.2201758) Crossref, ISI, Google Scholar18. Haupt SE, Kosovic B. 2017Variable generation power forecasting as a big data problem. IEEE Trans. Sustain. Energy. 8, 725–732. (doi:10.1109/TSTE.2016.2604679) Crossref, ISI, Google Scholar19. Kosovic Bet al.2020A comprehensive wind power forecasting system integrating artificial intelligence and numerical weather prediction. Energies. 13, 1372. (doi:10.3390/en13061372) Crossref, ISI, Google Scholar20. Haupt SEet al.2020Combining artificial intelligence with physics-based methods for probabilistic renewable energy forecasting. Energies. 13, 1979. (doi:10.3390/en13081979) Crossref, ISI, Google Scholar21. Vannitsem Set al.2020Statistical Postprocessing for Weather Forecasts - Review, Challenges and Avenues in a Big Data World. arXiv. See . Google Scholar22. Murphy JM. 1988The impact of ensemble forecasts on predictability. Q. J. R. Meteorol. Soc. 114, 463–493. (doi:10.1002/qj.49711448010) Crossref, ISI, Google Scholar23. Toth Z, Kalnay E. 1993Ensemble forecasting at NMC: the generation of perturbations. Bull. Am. Meteorol. Soc. 74, 2317–2330. (doi:10.1175/1520-0477(1993)0742.0.CO;2) Crossref, ISI, Google Scholar24. Buizza R. 1997Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system. Mon. Wea. Rev. 125, 99–119. (doi:10.1175/1520-0493(1997)1252.0.CO;2) Crossref, ISI, Google Scholar25. Stensrud DJ, Brooks HE, Du J, Tracton MS, Rogers E. 1999Using ensembles for short-range forecasting. Mon. Wea. Rev. 127, 433–446. (doi:10.1175/1520-0493(1999)1272.0.CO;2) Crossref, ISI, Google Scholar26. Taillardat M, Mestre O. 2020From research to applications; Examples of operational ensemble post-processing in France using machine learning. Copernicus GmbH. 27, 329–347. (doi:10.5194/npg-27-329-2020) Google Scholar27. Hopson TMet al.2010Quantile regression as a means of calibrating and verifying a mesoscale NWP ensemble. In 20th Conf. on Probability and Statistics in the Atmospheric Sciences. See . Google Scholar28. Raftery AE, Gneiting T, Balabdaoui F, Polakowski M. 2005Using Bayesian model averaging to calibrate forecast ensembles. Mon. Wea Rev 133, 1155–1174. (doi:10.1175/MWR2906.1) Crossref, ISI, Google Scholar29. Kolczynski WC, Stauffer DR, Haupt SE, Deng A. 2009Ensemble variance calibration for representing meteorological uncertainty for atmospheric transport and dispersion modeling. J. Appl. Meteorol. Climatol. 48, 2001–2021. (doi:10.1175/2009JAMC2059.1) Crossref, ISI, Google Scholar30. Kolczynski WC, Stauffer DR, Haupt SE, Altman NS, Deng A. 2011Investigation of ensemble variance as a measure of true forecast variance. Mon. Wea. Rev. 139, 3954–3963. (doi:10.1175/MWR-D-10-05081.1) Crossref, ISI, Google Scholar31. Hamill TM, Whitaker JS. 2006Probabilistic quantitative precipitation forecasts based on reforecast analogs: theory and application. Mon. Wea. Rev. 134, 3209–3229. (doi:10.1175/MWR3237.1) Crossref, ISI, Google Scholar32. Hamill TM, Scheuerer M, Bates GT. 2015Analog probabilistic precipitation forecasts using GEFS reforecasts and climatology-calibrated precipitation analyses. Mon. Wea. Rev. 143, 3300–3309. (doi:10.1175/MWR-D-15-0004.1) Crossref, ISI, Google Scholar33. Greybush SJ, Haupt SE, Young GS. 2008The regime dependence of optimally weighted ensemble model consensus forecasts of surface temperature. Weather Forecast. 23, 1146–1161. (doi:10.1175/2008WAF2007078.1) Crossref, ISI, Google Scholar34. McCandless TM, Haupt SE, Young GY. 2011Statistical guidance methods for predicting snowfall accumulation in the Northeast United States. Nat. Weather Dig. 3, 14 pp. Google Scholar35. Krasnopolsky VM. 2013The application of neural networks in the earth system sciences, 189 pp. Berlin, Germany: Springer. Google Scholar36. Roebber PJ. 2015Adaptive evolutionary programming. Mon. Wea. Rev. 143, 1497–1505. (doi:10.1175/MWR-D-14-00095.1) Crossref, ISI, Google Scholar37. Roebber PJ. 2015Using evolutionary programs to maximize minimum temperature forecast skill. Mon. Wea. Rev. 143, 1506–1516. (doi:10.1175/MWR-D-14-00096.1) Crossref, ISI, Google Scholar38. Roebber PJ. 2015Evolving ensembles. Mon. Wea. Rev. 143, 471–490. (doi:10.1175/MWR-D-14-00058.1) Crossref, ISI, Google Scholar39. Delle Monache L, Eckel FA, Rife DL, Nagarajan B, Searight K. 2013Probabilistic weather prediction with an analog ensemble. Mon. Wea. Rev. 141, 3498–3516. (doi:10.1175/mwr-d-12-00281.1) Crossref, ISI, Google Scholar40. McGovern A, Elmore KL, Gagne DJ, Haupt SE, Karstens CD, Lagerquist R, Smith T, Williams JK. 2017Using artificial intelligence to improve real-time decision-making for high-impact weather. Bull. Am. Meteorol. Soc. 98, 2073–2090. (doi:10.1175/BAMS-D-16-0123.1) Crossref, ISI, Google Scholar41. Breiman L. 2001Random forests. Mach. Learn. 45, 5–32. (doi:10.1023/A:1010933404324) Crossref, ISI, Google Scholar42. Taillardat M, Mestre O, Zamo M, Naveau P. 2016Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. Mon. Wea. Rev. 144, 2375–2393. (doi:10.1175/MWR-D-15-0260.1) Crossref, ISI, Google Scholar43. Taillardat M, Fougères A-L, Naveau P, Mestre O. 2019Forest-based and semiparametric methods for the postprocessing of rainfall ensemble forecasting. Weather Forecast. 34, 617–634. (doi:10.1175/WAF-D-18-0149.1) Crossref, ISI, Google Scholar44. Gagne DJ, McGovern A, Brotzge J. 2009Classification of convective areas using decision trees. J. Atmos. Ocean. Technol. 26, 1341–1353. (doi:10.1175/2008JTECHA1205.1) Crossref, ISI, Google Scholar45. Ahijevych D, Pinto JO, Williams JK, Steiner M. 2016Probabilistic forecasts of mesoscale convective system initiation using the random forest data mining technique. Weather Forecast. 31, 581–599. (doi:10.1175/WAF-D-15-0113.1) Crossref, ISI, Google Scholar46. Lagerquist R. 2016Using machine learning to predict damaging straight-line convective winds, M.S. thesis, 251 pp. School of Meteorology, University of Oklahoma, See . Google Scholar47. Gagne DJ, McGovern A, Haupt SE, Sobash RA, Williams JK, Xue M. 2017Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. Weather Forecast. 32, 1819–1840. (doi:10.1175/WAF-D-17-0010.1) Crossref, ISI, Google Scholar48. Gagne DJ, Haupt SE, Nychka DW, Thompson G. 2019Interpretable deep learning for spatial analysis of severe hailstorms. Mon. Wea. Rev. 147, 2827–2845. (doi:10.1175/mwr-d-18-0316.1) Crossref, ISI, Google Scholar49. Haupt SE, Mahoney WP, Parks K. 2014Wind power forecasting. In Weather matters for energy (eds Troccoli A, Dubus L, Haupt SE), pp. 295–318. Berlin, Germany: Springer. Crossref, Google Scholar50. Giebel G, Kariniotakis G. 2007Best practice in short-term forecasting: a users guide. In European Wind Energy Conf. Exhibition, 5 pp. Milan, Italy: European Wind Energy Association. See . Google Scholar51. Pelland SJ, Remund J, Kleissl J, Oozeki T, DeBrabandere K. 2016Photovoltaic and solar forecasting: state of the art. International Energy Agency Rep. IEA PVPS T14-01, 36 pp. See . Google Scholar52. Orwig KDet al.2015Recent trends in variable generation forecasting and its value to the power system. IEEE Trans. Sustain. Energy. 6, 924–933. (doi:10.1109/TSTE.2014.2366118) Crossref, ISI, Google Scholar53. Tuohy Aet al.2015Solar forecasting: methods, challenges, and performance. IEEE Power Energ. Mag. 13, 50–59. (doi:10.1109/MPE.2015.2461351) Crossref, ISI, Google Scholar54. Haupt SEet al.2019The use of probabilistic forecasts: applying them in theory and practice. IEEE Power Energ. Mag. 17, 46–57. (doi:10.1109/MPE.2019.2932639) Crossref, ISI, Google Scholar55. Alessandrini S, Delle Monache L, Sperati S, Nissen JN. 2015A novel application of an analog ensemble for short-term wind power forecasting. Renewable Energy. 76, 768–781. (doi:10.1016/j.renene.2014.11.061) Crossref, ISI, Google Scholar56. Alessandrini S, Delle Monache L, Sperati S, Cervone G. 2015An analog ensemble for short-term probabilistic solar power forecast. Appl. Energy. 157, 95–110. (doi:10.1016/j.apenergy.2015.08.011) Crossref, ISI, Google Scholar57. Scaife AAet al.2014Skillful long-range prediction of European and North American winters. Geophys. Res. Lett. 41, 2514–2519. (doi:10.1002/2014GL059637) Crossref, ISI, Google Scholar58. MacLachlan Cet al.2014Global Seasonal forecast system version 5 (GloSea5): a high-resolution seasonal forecast system. Q. J. R. Meteorol. Soc. 141, 1072–1084. (doi:10.1002/qj.2396) Crossref, ISI, Google Scholar59. Wu J-L, Kashinath K, Albert A, Chirila D, Xiao H. 2020Enforcing statistical constraints in generative adversarial networks for modeling chaotic dynamical systems Enforcing statistical constraints in generative adversarial networks for modeling chaotic dynamical systems. J. Comput. Phys. 406, 109209. Crossref, ISI, Google Scholar60. Bolton T, Zanna L. 2019Applications of deep learning to ocean data inference and subgrid parameterization. J. Adv. Model. Earth Syst. 11, 376–399. (doi:10.1029/2018MS001472) Crossref, ISI, Google Scholar61. Hyder Pet al.2018Critical Southern Ocean climate model biases traced to atmospheric model cloud errors. Nat. Commun. 9, 3625. (doi:10.1038/s41467-018-05634-2) Crossref, PubMed, ISI, Google Scholar62. Meijers AJS. 2014The Southern Ocean in the coupled model intercomparison project phase 5. Philos. Trans. A Math. Phys. Eng. Sci. 372, 20130296. Abstract, ISI, Google Scholar63. Maurer EP, Hidalgo HG. 2008Utility of daily vs. monthly large-scale climate data: an intercomparison of two statistical downscaling methods. Hydrology Earth System Sci. 12, 551–563. (doi:10.5194/hess-12-551-2008) Crossref, ISI, Google Scholar64. Cayan DR, Maurer EP, Dettinger MD, Tyree M, Hayhoe K. 2008Climate change scenarios for the California region. Clim. Change. 87(S1), 21–42. (doi:10.1007/s10584-007-9377-6) Crossref, ISI, Google Scholar65. Haupt SE, Copeland J, Cheng WYY, Zhang Y, Ammann C, Sullivan P. 2016A method to assess the wind and solar resource and to quantify interannual variability over the United States under current and projected future climate. J. Appl. Meteorol. Climatol. 55, 345–363. (doi:10.1175/JAMC-D-15-0011.1) Crossref, ISI, Google Scholar66. Nielsen M. 2013Neural networks and deep learning. Google Scholar67. Lagerquist R, McGovern A, Gagne DJ. 2019Deep learning for spatially explicit prediction of synoptic-scale fronts. Weather Forecast. 34, 1137–1160. (doi:10.1175/WAF-D-18-0183.1) Crossref, ISI, Google Scholar68. Chapman WE, Subramanian AC, Delle Monache L, Xie SP, Ralph FM. 2019Improving atmospheric river forecasts with machine learning. Geophys. Res. Lett. 46, 10 627–10 635. (doi:10.1029/2019GL083662) Crossref, ISI, Google Scholar69. McGovern A, Lagerquist R, John Gagne D, Jergensen GE, Elmore KL, Homeyer CR, Smith T. 2019Making the black box more transparent: understanding the physical implications of machine learning. Bull. Am. Meteorol. Soc. 100, 2175–2199. (doi:10.1175/BAMS-D-18-0195.1) Crossref, ISI, Google Scholar70. Bremnes JB. 2020Ensemble postprocessing using quantile function regression based on neural networks and bernstein polynomials. Mon. Wea. Rev. 148, 403–414. (doi:10.1175/MWR-D-19-0227.1) Crossref, ISI, Google Scholar71. Gronquist P, Yao C, Ben-Nun T, Dryden N, Dueben P, Li S, Hoefler T. 2020Deep learning for post-processing ensemble weather forecasts. See . Google Scholar72. Blalock D, Ortiz JJ, Frankle J, Guttag J. 2020What is the state of neural network pruning?Proc. Mach. Learn. Syst.129–146. See Google Scholar73. Musgrave K, Belongie S, Lim S-N. 2020A metric learning reality check, 2020, See . Google Scholar74. Simonyan K, Vedaldi A, Zisserman A. 2013Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv. See . Google Scholar75. Ribeiro MT, Singh S, Guestrin C. 2016Why Should I Trust You? In Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM (doi:10.1145/2939672.2939778) Google Scholar76. Carter B, Mueller J, Jain S, Gifford D. 2018What made you do this? Understanding black-box decisions with sufficient input subsets. arXiv. See . Google Scholar77. Toms BA, Barnes EA, Ebert-Uphoff I. 2019Physically interpretable neural networks for the geosciences: applications to earth system variability. arXiv. . Google Scholar78. Lakshmanan V, Karstens C, Krause J, Elmore K, Ryzhkov A, Berkseth S. 2015Which polarimetric variables are important for weather/no-weather discrimination?J. Atmos. Ocean. Technol. 32, 1209–1223. (doi:10.1175/JTECH-D-13-00205.1) Crossref, ISI, Google Scholar79. Wagstaff KL, Lee J. 2018Interpretable discovery in large image data sets. arXiv. June. See . Google Scholar80. Wilkinson Met al.2016The FAIR guiding principles for scientific data management and stewardship. Sci. Data 3, 160018. (doi:10.1038/sdata.2016.18) Crossref, PubMed, ISI, Google Scholar81. Wilson LJ, Vallée M. 2002The Canadian Updateable Model Output Statistics (UMOS) system: design and development tests. Weather Forecast. 17, 206–222. (doi:10.1175/1520-0434(2002)0172.0.CO;2) Crossref, ISI, Google Scholar82. Hamill T. 2018Practical aspects of statistical postprocessing. In Statistical Postprocessing of Ensemble Forecasts (eds Vannitsem S, Wilks DS, Messner J), pp. 187–218. The Netherlands: Elsevier. Crossref, Google Scholar83. European Centre for Medium-Range Weather Forecasts 2017, 2017: ERA5 reanalysis. National Center for Atmospheric Research, Computational and Information Systems Laboratory, accessed 7 September 2019, . Google Scholar84. Lee JA, Kolczynski WC, McCandless TC, Haupt SE. 2012An objective methodology for configuring and down-selecting an NWP ensemble for low-level wind prediction. Mon. Wea. Rev. 140, 2270–2286. (doi:10.1175/MWR-D-11-00065.1) Crossref, ISI, Google Scholar85. Lee JA, Haupt SE, Young GY. 2016Down-selecting numerical weather prediction multi-physics ensembles with hierarchical cluster analysis. J. Climatol. Wea. Forecasting. 4, 1000156. (doi:10.4172/2332-2594.1000156) Google Scholar86. Chapman WE, Lerch S, Kirkwood C, Subramanian AC. Matsueda M, Haupt SE. 2020Postprocessing model V 1.0. In Data for: Towards Implementing AI Post-processing in Weather and Climate: Proposed Actions from the Oxford 2019 Workshop. UC San Diego Library Digital Collections. . Google Scholar87. Wheeler , Hendon . 2004. Google Scholar88. Wallace JM, Gutzler DS. 1981Teleconnections in the geopotential height field during the Northern Hemisphere winter. Mon. Wea. Rev. 109, 784–812. Crossref, ISI, Google Scholar89. Madden RA, Julian PR. 1972Description of global-scale circulation cells in the tropics with a 40–50 day period. J. Atmos. Sci. 29, 1109–1123. (doi:10.1175/1520-0469(1972)0292.0.CO;2) Crossref, ISI, Google Scholar90. Madden RA, Julian PR. 1994Observations of the 40–50-day tropical oscillation—A review. Mon. Wea. Rev. 122, 814–837. (doi:10.1175/1520-0493(1994)1222.0.CO;2) Crossref, ISI, Google Scholar91. Subramanian AC, Jochum M, Miller AJ, Murtugudde R, Neale RB, Waliser DE. 2011The Madden–Julian oscillation in CCSM4. J. Climate. 24, 6261–6282. (doi:10.1175/JCLI-D-11-00031.1) Crossref, ISI, Google Scholar92. Subramanian A, Weisheimer A, Palmer T, Vitart F, Bechtold P. 2017Impact of stochastic physics on tropical precipitation in the coupled ECMWF model. Q. J. R. Meteorol. Soc. 143, 852–865. (doi:10.1002/qj.2970) Crossref, ISI, Google Scholar93. Shukla Jet al.2000Dynamical seasonal prediction. Bull. Amer. Meteor. Soc. 81, 2593–2606. (doi:10.1175/1520-0477(2000)0812.3.CO;2) Crossref, ISI, Google Scholar94. O'Reilly CH, Heatley J, MacLeod D, Weisheimer A, Palmer TN, Schaller N, Woollings T. 2017Variability in seasonal forecast skill of Northern Hemisphere winters over the 20th Century. Geophys. Res. Lett. 44, 5729–5738. (doi:10.1002/2017GL073736) Crossref, ISI, Google Scholar95. Moorthi S, Pan H-L, Caplan P. 2001Changes to the 2001 NCEP operational MRF/AVN global analysis/forecast system. NWS Technical Procedures Bull. 484, 1–14. Google Scholar96. Gelaro Ret al.2017The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). J. Climate 30, 5419–5454. (doi:10.1175/JCLI-D-16-0758.1) Crossref, ISI, Google Scholar97. McCarty W, Coy L, Gelaro R, Huang A, Merkova D, Smith EB, Sienkiewicz M, Wargan K. 2016MERRA-2 input observations: summary and assessment. Technical Report Series on Global Modeling and Data Assimilation. Google Scholar98. Kirkwood C, Economou T, Odbert H, Pugeault N. 2021A framework for probabilistic weather forecast post-processing across models and lead times using machine learning. Phil. Trans. R. Soc. A 379, 20200099. (doi:10.1098/rsta.2020.0099) Link, ISI, Google Scholar99. Mukkavilli SK, Bara A, Gagne DJ, Tissot P, Campos E, Ganguly AR, Joppa L, Meger D, Dudek G. 2019EnviroNet: imagenet for environment. In 99th American Meteorological Society Annual Meeting. AMS, 2019 January. Google Scholar100. Rasp S, Dueben PD, Scher S, Weyn JA, Mouatadid S, Thuerey N. 2020WeatherBench: A benchmark dataset for data-driven weather forecasting. See . Google ScholarPage 41. The Royal Society2019The AI revolution in scientific research. London, UK: The Royal Society. ( ) Google Scholar2. Hey T, Trefethen A. 2020The fourth paradigm 10 years on. Informatik Spektrum 42, 441–447. (doi:10.1007/s00287-019-01215-9) Crossref, Google Scholar3. LeCun Y, Bengio Y, Hinton G. 2015Deep learning. Nature 521, 436–444. (doi:10.1038/nature14539) Crossref, PubMed, ISI, Google Scholar4. Moore GE. 1965Cramming more components onto integrated circuits. Electronics 38, 114–117. Google Scholar5. Shalf J. 2020The future of computing beyond Moore's law. Phil. Trans. R. Soc. A 378, 20190061. (doi:10.1098/rsta.2019.0061) Link, ISI, Google Scholar6. Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N, Prabhat . 2019Deep learning and process understanding for data-driven Earth system science. Nature 566, 195–204. (doi:10.1038/s41586-019-0912-1) Crossref, PubMed, ISI, Google Scholar7. Kasim MFet al.2020Building high accuracy emulators for scientific simulations with deep neural architecture search. (. Google Scholar8. Buchanan M. 2020Living up to the promise. Nat. Phys. 16, 706–706. (doi:10.1038/s41567-020-0962-1) Crossref, ISI, Google Scholar9. Hutson M. 2020From models of galaxies to atoms, simple AI shortcuts speed up simulations by billions of times. Science 367, 728–728. (doi:10.1126/science.367.6479.728) Crossref, PubMed, ISI, Google Scholar10. Hestness J, Narang S, Ardalani N, Diamos G, Jun H, Kianinejad H, Patwary MMA, Yang Y, Zhou Y. 2017Deep learning scaling is predictable, empirically. (. Google Scholar11. Karpatne A, Atluri G, Faghmous JH, Steinbach M, Banerjee A, Ganguly A, Shekhar S, Samatova N, Kumar V. 2017Theory-guided data science: a new paradigm for scientific discovery from data. IEEE Trans. Knowl. Data Eng. 29, 2318–2331. (doi:10.1109/TKDE.2017.2720168) Crossref, ISI, Google Scholar12. Willard J, Jia X, Xu S, Steinbach M, Kumar V. 2020Integrating physics-based modeling with machine learning: a survey. (. Google Scholar13. Karpatne A, Watkins W, Read J, Kumar V. 2018Physics-guided neural networks (PGNN): an application in lake temperature modeling. In Proc. of the 2018 SIAM Int. Conf. on Data Mining. Google Scholar14. Beucler T, Rasp S, Pritchard M, Gentine P. 2019Achieving conservation of energy in neural network emulators for climate modeling. (. Google Scholar15. Raissi M, Perdikaris P, Karniadakis GE. 2019Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. J. Comput. Phys. 378, 686–707. (doi:10.1016/j.jcp.2018.10.045) Crossref, ISI, Google Scholar16. Zhu Y, Zabaras N, Koutsourelakis PS, Perdikaris P. 2019Physics-constrained deep learning for high-dimensional surrogate modeling and uncertainty quantification without labeled data. J. Comput. Phys. 394, 56–81. (doi:10.1016/j.jcp.2019.05.024) Crossref, ISI, Google Scholar17. Beucler T, Pritchard M, Rasp S, Ott J, Baldi P, Gentine P. 2019Enforcing analytic constraints in neural-networks emulating physical systems. (. Google Scholar18. Wu JL, Kashinath K, Albert A, Chirila D, Xiao H. 2020Enforcing statistical constraints in generative adversarial networks for modeling chaotic dynamical systems. J. Comput. Phys. 406, 109209. (doi:10.1016/j.jcp.2019.109209) Crossref, ISI, Google Scholar19. Márquez-Neila P, Salzmann M, Fua P. 2017Imposing hard constraints on deep networks: promises and limitations. (. Google Scholar20. Mohan AT, Lubbers N, Livescu D, Chertkov M. 2020Embedding hard physical constraints in neural network coarse-graining of 3D turbulence. (. Google Scholar21. Jiang CM, Kashinath K. 2020Enforcing physical constraints in CNNs through differentiable PDE layer. In ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations. Google Scholar22. Daw A, Thomas RQ, Carey CC, Read JS, Appling AS, Karpatne A. 2020Physics-guided architecture (PGA) of neural networks for quantifying uncertainty in lake temperature modeling. In Proc. of the 2020 SIAM Int. Conf. on Data Mining, pp. 532–540. Google Scholar23. Cohen TS, Weiler M, Kicanaoglu B, Welling M. 2019Gauge equivariant convolutional networks and the icosahedral CNN. In Proc. of the 36th Int. Conf. on Machine Learning. Google Scholar24. Maron H, Fetaya E, Segol N, Lipman Y. 2019On the universality of invariant networks. In Proc. of the 36th Int. Conf. on Machine Learning. Google Scholar25. Maron H, Ben-Hamu H, Shamir N, Lipman Y. 2019Invariant and equivariant graph networks. In Int. Conf. on Learning Representations. Google Scholar26. Thomas N, Smidt T, Kearnes S, Yang L, Li L, Kohlhoff K, Riley P. 2018Tensor field networks: rotation-and translation-equivariant neural networks for 3D point clouds. (. Google Scholar27. Ling J, Kurzawski A, Templeton J. 2016Reynolds averaged turbulence modelling using deep neural networks with embedded invariance. J. Fluid Mech. 807, 155–166. (doi:10.1017/jfm.2016.615) Crossref, ISI, Google Scholar28. Cohen TS, Geiger M, K¨hler J, Welling M. 2018Spherical CNNs. In Int. Conf. on Learning Representations. Google Scholar29. Jiang C, Huang J, Kashinath K, Marcus P, Nießner M. 2019Spherical CNNs on unstructured grids. In Int. Conf. on Learning Representations. Google Scholar30. Wang R, Walters R, Yu R. 2020Incorporating symmetry into deep dynamics models for improved generalization. (. Google Scholar31. Chattopadhyay A, Mustafa M, Hassanzadeh P, Kashinath K. 2020Deep spatial transformers for autoregressive data-driven forecasting of geophysical turbulence. In Proc of the 10th Workshop on Climate Informatics, Oxford, UK, 2020. Google Scholar32. Beucler T, Pritchard M, Gentine P, Rasp S. 2020Towards physically-consistent, data-driven models of convection. (. Google Scholar33. Palmer TN. 2019Living up to the promise. Nat. Rev. Phys. 1, 463–471. (doi:10.1038/s42254-019-0062-2) Crossref, Google Scholar34. Ghahramani Z. 2015Probabilistic machine learning and artificial intelligence. Nature 521, 452–459. (doi:10.1038/nature14541) Crossref, PubMed, ISI, Google Scholar35. Krasnopolsky VM, Fox-Rabinovitz MS, Belochitski AA. 2013Using ensemble of neural networks to learn stochastic convection parameterizations for climate and numerical weather prediction models from data simulated by a cloud resolving model. Ad. Artif. Neural Syst. 2013, 485–913. Google Scholar36. Gagne DJ, Christensen HM, Subramanian AC, Monahan AH. 2020Machine learning for stochastic parameterization: generative adversarial networks in the lorenz '96 model. J. Adv. Model. Earth Syst. 12, e2019MS001896. (doi:10.1029/2019MS001896) Crossref, ISI, Google Scholar37. Groenke B, Madaus L, Monteleoni C. 2020ClimAlign: unsupervised statistical downscaling of climate variables via normalizing flows. Proc. of the 10th Workshop on Climate Informatics, Oxford, UK, 2020. Google Scholar38. Tadmor E. 2012A review of numerical methods for nonlinear partial differential equations. Bull. Am. Math. Soc. 49, 507–554. (doi:10.1090/S0273-0979-2012-01379-4) Crossref, ISI, Google Scholar39. Miller J, Hardt M. 2019Stable recurrent models. In Int. Conf. on Learning Representations. Google Scholar40. Erichson NB, Muehlebach M, Mahoney MW. 2019Physics-informed Autoencoders for Lyapunov-stable Fluid Flow Prediction. (. Google Scholar41. Lusch B, Kutz JN, Brunton SL. 2018Deep learning for universal linear embeddings of nonlinear dynamics. Nat. Commun. 9, 4950. (doi:10.1038/s41467-018-07210-0) Crossref, PubMed, ISI, Google Scholar42. Mamakoukas G, Abraham I, Murphey TD. 2020Learning data-driven stable Koopman operators. (. Google Scholar43. Brenowitz ND, Beucler T, Pritchard M, Bretherton CS. 2020Interpreting and stabilizing machine-learning parametrizations of convection. (. Google Scholar44. Ott J, Pritchard M, Best N, Linstead E, Curcic M, Baldi P. 2020A Fortran-Keras deep learning bridge for scientific computing. arXiv:2004.10652 (Google Scholar45. Rasp S. 2020Coupled online learning as a way to tackle instabilities and biases in neural network parameterizations: general algorithms and Lorenz 96 case study (v1.0). Geosci. Model Dev. 13, 2185–2020. (doi:10.5194/gmd-13-2185-2020) Google Scholar46. Yuval J, O'Gorman P. 2020Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions. Nat. Commun. 11, 3295. (doi:10.1038/s41467-020-17142-3) Crossref, PubMed, ISI, Google Scholar47. Yuval J, O'Gorman PA, Hill CN. 2020Use of neural networks for stable, accurate and physically consistent parameterization of subgrid atmospheric processes with good performance at reduced precision. (. Google Scholar48. Slingo J, Bates K, Nikiforakis N, Piggott M, Roberts M, Shaffrey L, Stevens I, Vidale PL, Weller H. 2008Developing the next-generation climate system models: challenges and achievements. Phil. Trans. R. Soc. A 367, 815–831. (doi:10.1098/rsta.2008.0207) Link, ISI, Google Scholar49. Mohan A, Livescu D, Chertkov M. 2020Wavelet-powered neural networks for turbulence. In ICLR 2020 Workshop on Climate Change AI. Google Scholar50. Li Z, Kovachki N, Azizzadenesheli K, Liu B, Bhattacharya K, Stuart A, Anandkumar A. 2020Fourier neural operator for parametric partial differential equations. (. Google Scholar51. Tancik M, Srinivasan PP, Mildenhall B, Fridovich-Keil S, Raghavan N, Singhal U, Ramamoorthi R, Barron JT, Ng R. 2020Fourier features let networks learn high-frequency functions in low-dimensional domains. NeurIPS 33. Google Scholar52. Pratt H, Williams B, Coenen F, Zheng Y. 2017FCNN: fourier convolutional neural networks. ECML PKDD 17, pp. 786–798. Google Scholar53. Guan B, Zhang J, Sethares WA, Kijowski R, Liu F. 2019SpecNet: spectral domain convolutional neural network. (. Google Scholar54. Williams PDet al.2017A census of atmospheric variability from seconds to decades. Geophys. Res. Lett. 44, 11 201–11 211. Crossref, ISI, Google Scholar55. Haller G. 2015Lagrangian coherent structures. Annu. Rev. Fluid Mech. 47, 137–162. (doi:10.1146/annurev-fluid-010313-141322) Crossref, ISI, Google Scholar56. Rupe A, Kumar N, Epifanov V, Kashinath K, Pavlyk O, Schlimbach F, Patwary M, Maidanov S, Lee V, Prabhat M, Crutchfield JP. 2019DisCo: Physics-based unsupervised discovery of coherent structures in spatiotemporal systems. In Proceedings of the 2019 IEEE/ACM Workshop on Machine Learning in High Performance Computing Environments (MLHPC), Denver, CO, pp. 75–87. Google Scholar57. Xie Y, Franz E, Chu M, Thuerey N. 2018tempoGAN: a temporally coherent, volumetric GAN for super-resolution fluid flow. ACM Trans. Graph. 37, 1–15. ISI, Google Scholar58. Li Z, Kovachki N, Azizzadenesheli K, Liu B, Bhattacharya K, Stuart A, Anandkumar A. 2020Neural operator: graph kernel network for partial differential equations. (. Google Scholar59. de Bezenac E, Pajot A, Gallinari P. 2018Deep learning for physical processes: incorporating prior scientific knowledge. In Int. Conf. on Learning Representations. Google Scholar60. Xu K, Wen L, Li G, Bo L, Huang Q. 2019Spatiotemporal CNN for video object segmentation. In Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1379– 1388. Google Scholar61. Smagorinsky J. 1963General circulation experiments with the primitive equations: I. The basic experiment. Mon. Weather Rev. 91, 99–164. (doi:10.1175/1520-0493(1963)091¡0099:GCEWTP¿2.3.CO;2) Crossref, Google Scholar62. Cheng Yet al.2019Deep learning for subgrid-scale turbulence modeling in large-eddy simulations of the atmospheric boundary layer. (. Google Scholar63. Pal A. 2020Deep learning emulation of subgrid–scale processes in turbulent shear flows. Geophys. Res. Lett. 47, e2020GL087005. (doi:10.1029/2020GL087005) Crossref, ISI, Google Scholar64. Bar-Sinai Y, Hoyer S, Hickey J, Brenner MP. 2019Learning data-driven discretizations for partial differential equations. Proc. Natl Acad. Sci. USA 116, 15 344–15 349. (doi:10.1073/pnas.1814058116) Crossref, ISI, Google Scholar65. Pathak J, Wikner A, Fussell R, Chandra S, Hunt BR, Girvan M, Ott E. 2018Hybrid forecasting of chaotic processes: Using machine learning in conjunction with a knowledge-based model. Chaos 28, 041101. (doi:10.1063/1.5028373) Crossref, PubMed, ISI, Google Scholar66. Watson PAG. 2019Applying machine learning to improve simulations of a chaotic dynamical system using empirical error correction. J. Adv. Model. Earth Syst. 11, 1402–1417. (doi:10.1029/2018MS001597) Crossref, PubMed, ISI, Google Scholar67. Bonavita M, Laloyaux P. 2020Machine learning for model error inference and correction. J. Adv. Model. Earth Syst. 12, e2020MS002232. (doi:10.1029/2020MS002232) Crossref, ISI, Google Scholar68. Farchi A, Laloyaux P, Bonavita M, Bocquet M. 2020Using machine learning to correct model error in data assimilation and forecast applications. arXiv:2010.12605. (Google Scholar69. Mauritsen Tet al.2012Tuning the climate of a global model. J. Adv. Model. Earth Syst. 4, M00A01. (doi:10.1029/2012MS000154) Crossref, ISI, Google Scholar70. Duraisamy K, Iaccarino G, Xiao H. 2019Turbulence modeling in the age of data. Annu. Rev. Fluid Mech. 51, 357–377. (doi:10.1146/annurev-fluid-010518-040547) Crossref, ISI, Google Scholar71. Ollinaho P, Bechtold P, Leutbecher M, Laine M, Solonen A, Haario H, Järvinen H. 2019Parameter variations in prediction skill optimization at ecmwf. Nonlin. Processes Geophys. 71, 1001–1010. Google Scholar72. Tett SFB, Yamazaki K, Mineter MJ, Cartis C, Eizenberg N. 2017Calibrating climate models using inverse methods: case studies with hadam3, hadam3p and hadcm3. Geosci. Model Dev. 10, 3567–3589. Crossref, ISI, Google Scholar73. Bellprat O, Kotlarski S, Lüthi D, Elía RD, Frigon A, Laprise R, Schär C. 2016Objective calibration of regional climate models: Application over europe and north america. Journal of Climate 29, 819–838. Crossref, ISI, Google Scholar74. Duraisamy K. 2020Machine learning-augmented Reynolds-averaged and large eddy simulation models of turbulence. (. Google Scholar75. Samek W, Montavon G, Vedaldi A, Hansen LK, Müller KR. 2019Explainable AI: interpreting, explaining and visualizing deep learning. New York, NY: Springer Nature. Crossref, Google Scholar76. 77. Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller KR. 2019Unmasking Clever Hans predictors and assessing what machines really learn. Nat. Commun. 10, 1096. (doi:10.1038/s41467-019-08987-4) Crossref, PubMed, ISI, Google Scholar78. Rudin C. 2019Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. 1, 206–215. (doi:10.1038/s42256-019-0048-x) Crossref, Google Scholar79. McGovern A, Lagerquist R, Gagne DJ, Jergensen GE, Elmore KL, Homeyer CR, Smith T. 2019Making the black box more transparent: understanding the physical implications of machine learning. Bull. Am. Meteorol. Soc. 100, 2175–2199. (doi:10.1175/BAMS-D-18-0195.1) Crossref, ISI, Google Scholar80. Ebert-Uphoff I, Hilburn K. 2020Evaluation, tuning and interpretation of neural networks for working with images in meteorological applications. Bull. Am. Meteorol. Soc. 101, E2149–E2170. Crossref, ISI, Google Scholar81. Gagne DJ, Haupt SE, Nychka DW, Thompson G. 2019Interpretable deep learning for spatial analysis of severe hailstorms. Mon. Weather Rev. 147, 2827–2845. (doi:10.1175/MWR-D-18-0316.1) Crossref, ISI, Google Scholar82. Toms BA, Barnes EA, Ebert-Uphoff I. 2020Physically interpretable neural networks for the geosciences: applications to earth system variability. J. Adv. Model. Earth Syst. 12, e2019MS002002. Crossref, ISI, Google Scholar83. Toms BA, Kashinath K. 2020Testing the reliability of interpretable neural networks in geoscience using the Madden-Julian oscillation. Geosci. Model Dev.1–22. (ISI, Google Scholar84. Caldeira J, Nord B. 2020Deeply uncertain: comparing methods of uncertainty quantification in deep learning algorithms. In ICLR 2020 Workshop on Fundamental Science in the era of AI. Google Scholar85. Kendall A, Gal Y. 2017What uncertainties do we need in Bayesian deep learning for computer vision? In Advances in Neural Information Processing Systems, vol. 30, pp. 5574–5584. Google Scholar86. Gal Y, Ghahramani Z. 2016Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In Proc. of the 33rd Int. Conf. on Machine Learning (ICML-16). Google Scholar87. Lakshminarayanan B, Pritzel A, Blundell C. 2017Simple and scalable predictive uncertainty estimation using deep ensembles. In Advances in Neural Information Processing Systems, vol. 30, pp. 6402–6413. Google Scholar88. Yang Y, Perdikaris P. 2019Adversarial uncertainty quantification in physics-informed neural networks. J. Comput. Phys. 394, 136–152. (doi:10.1016/j.jcp.2019.05.027) Crossref, ISI, Google Scholar89. Vandal T, Kodra E, Dy J, Ganguly S, Nemani R, Ganguly AR. 2018Quantifying uncertainty in discrete-continuous and skewed data with Bayesian deep learning. In Proc. of the 24th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining. Google Scholar90. Schneider T, Lan S, Stuart A, Teixeira J. 2017Earth system modeling 2.0: a blueprint for models that learn from observations and targeted high-resolution simulations. Geophys. Res. Lett. 44, 12 396–12 417. Crossref, ISI, Google Scholar91. Rasp S, Pritchard MS, Gentine P. 2018Deep learning to represent subgrid processes in climate models. Proc. Natl Acad. Sci. USA 115, 9684–9689. (doi:10.1073/pnas.1810286115) Crossref, PubMed, ISI, Google Scholar92. Gentine P, Pritchard MS, Rasp S, Reinaudi G, Yacalis G. 2018Could machine learning break the convection parameterization deadlock?Geophys. Res. Lett. 45, 5742–5751. (doi:10.1029/2018GL078202) Crossref, ISI, Google Scholar93. O'Gorman PA, Dwyer JG. 2018Using machine learning to parameterize moist convection: potential for modeling of climate, climate change, and extreme events. J. Adv. Model. Earth Syst. 10, 2548–2563. (doi:10.1029/2018MS001351) Crossref, ISI, Google Scholar94. Chattopadhyay A, Subel A, Hassanzadeh P. 2020Data-driven super-parameterization using deep learning: experimentation with multi-scale Lorenz 96 systems and transfer-learning. J. Adv. Model. Earth Syst.e2020MS002084. ISI, Google Scholar95. Bolton T, Zanna L. 2019Applications of deep learning to ocean data inference and subgrid parameterization. J. Adv. Model. Earth Syst. 11, 376–399. (doi:10.1029/2018MS0 01472) Crossref, ISI, Google Scholar96. Baño-Medina J, Manzanas R, Gutiérrez JM. 2020Configuration and intercomparison of deep learning neural network models for statistical downscaling. Geosci. Model Dev. 13, 2109–2124. (doi:10.5194/gmd-13-2109-2020) Crossref, ISI, Google Scholar97. Stengel K, Glaws A, Hettinger D, King RN. 2020Adversarial super-resolution of climatological wind and solar data. Proc. Natl Acad. Sci. USA 117, 16 805–16 815. (doi:10.1073/pnas.1918964117) Crossref, ISI, Google Scholar98. Scher S, Messori G. 2019Weather and climate forecasting with neural networks: using general circulation models (gcms) with different complexity as a study ground. Geosci. Model Dev. 12, 2797–2809. (doi:10.5194/gmd-12-2797-2019) Crossref, ISI, Google Scholar99. Chattopadhyay A, Nabizadeh E, Hassanzadeh P. 2020Analog forecasting of extreme-causing weather patterns using deep learning. J. Adv. Model. Earth Syst. 12, e2019MS0 01958. Crossref, ISI, Google Scholar100. Pathak J, Mustafa M, Kashinath K, Motheau E, Kurth T, Day M. 2020Using Machine Learning to Augment Coarse-Grid Computational Fluid Dynamics Simulations. (. Google Scholar101. Champion K, Lusch B, Kutz JN, Brunton SL. 2019Data-driven discovery of coordinates and governing equations. Proc. Natl Acad. Sci. USA 116, 22 445–22 451. (doi:10.1073/pnas.1906995116) Crossref, ISI, Google Scholar102. Manepalli A, Albert A, Rhoades A, Feldman D. 2019Emulating numeric hydroclimate models with physics-informed cGANs. In Climate Change AI Workshop at the 33rd Conf. on Neural Information Processing Systems. Google Scholar103. Singh A, Albert A, White B. 2019Downscaling numerical weather models with GANs. In Climate Change AI workshop at the 33rd Conf. on Neural Information Processing Systems. Google Scholar104. Jiang CM, Esmaeilzadeh S, Azizzadenesheli K, Kashinath K, Mustafa M, Tchelepi HA, Marcus P. 2020MeshfreeFlowNet: a physics-constrained deep continuous space-time super-resolution framework. In SC '20: Proc. of the 2020 ACM Conf. on Supercomputing (to appear). Google Scholar105. Wang R, Kashinath K, Mustafa M, Albert A, Yu R. 2020Towards physics-informed deep learning for turbulent flow prediction. In Proc. of the 26th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'20). Google Scholar106. Brenowitz ND, Bretherton CS. 2018Prognostic validation of a neural network unified physics parameterization. Geophys. Res. Lett. 45, 6289–6298. (doi:10.1029/2018GL078510) Crossref, ISI, Google Scholar107. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. 2014Generative adversarial nets. In Advances in Neural Information Processing Systems, pp. 2672–2680. Google Scholar108. King R, Hennigh O, Mohan A, Chertkov M. 2018From deep to physics-informed learning of turbulence: diagnostics. Workshop on Modeling and Decision-Making in the Spatiotemporal Domain, NIPS 2018. Google Scholar109. Mescheder L, Geiger A, Nowozin S. 2018Which Training Methods for GANs do actually Converge? (. Google Scholar110. Arjovsky M, Bottou L. 2017Towards principled methods for training generative adversarial networks. (. Google Scholar111. Yang L, Zhang D, Karniadakis GE. 2020Physics-informed generative adversarial networks for stochastic differential equations. SIAM J. Sci. Comput. 42, A292–A317. (doi:10.1137/18M1225409) Crossref, ISI, Google Scholar112. Stinis P, Hagge T, Tartakovsky AM, Yeung E. 2019Enforcing constraints for interpolation and extrapolation in generative adversarial networks. J. Comput. Phys. 397, 108844. (doi:10.1016/j.jcp.2019.07.042) Crossref, ISI, Google Scholar113. Isola P, Zhu J, Zhou T, Efros AA. 2017Image-to-image translation with conditional adversarial networks. In 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 5967–5976. Google Scholar114. Rhoades AM, Jones AD, Ullrich PA. 2018Assessing mountains as natural reservoirs with a multimetric framework. Earth's Future 6, 1221–1241. (doi:10.1002/2017EF000789) Crossref, ISI, Google Scholar115. Maberly SCet al.2020Global lake thermal regions shift under climate change. Nat. Commun. 11, 1232. (doi:10.1038/s41467-020-15108-z) Crossref, PubMed, ISI, Google Scholar116. Yang C, Ma C, Yang M. 2014Single-image super-resolution: a benchmark. In European Conf. on Computer Vision, pp. 372–386. Google Scholar117. Ledig Cet al.2017Photo-realistic single image super-resolution using a generative adversarial network. In Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 4681–4690. Google Scholar118. Vandal T, Kodra E, Ganguly S, Micahelis A, Nemani R, Ganguly AR. 2017Quantifying uncertainty in discrete-continuous and skewed data with Bayesian deep learning. In Proc. of the 23rd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining. Google Scholar119. Wang X, Yu K, Wu S, Gu J, Liu Y, Dong C, Qiao Y, Change Loy C. 2018ESRGAN: enhanced super-resolution generative adversarial networks. In Proc. of the European Conf. on Computer Vision (ECCV) Workshops. Google Scholar120. Dueben PD, Bauer P. 2018Challenges and design choices for global weather and climate models based on machine learning. Geosci. Model Dev. 11, 3999–4009. (doi:10.5194/gmd-11-3999-2018) Crossref, ISI, Google Scholar121. Weyn JA, Durran DR, Caruana R. 2019Can machines learn to predict weather? Using deep learning to predict gridded 500-hpa geopotential height from historical weather data. J. Adv. Model. Earth Syst. 11, 2680–2693. (doi:10.1029/2019MS001705) Crossref, ISI, Google Scholar122. Chattopadhyay A, Hassanzadeh P, Subramanian D, Palem K. 2020Data-driven predictions of a multiscale Lorenz 96 chaotic system using machine learning methods: reservoir computing, artificial neural network, and long short-term memory network. Nonlinear Process. Geophys. 27, 373–389. (doi:10.5194/npg-27-373-2020) Crossref, ISI, Google Scholar123. Bocquet M, Brajard J, Carrassi A, Bertino L. 2020Bayesian inference of chaotic dynamics by merging data assimilation, machine learning and expectation-maximization. Foundations of Data Science 2, 55–80. Crossref, Google Scholar124. Labourasse E, Sagaut P. 2002Reconstruction of turbulent fluctuations using a hybrid RANS-LES approach. J. Comput. Phys. 182, 301–336. (doi:10.1006/jcph.2002.7169) Crossref, ISI, Google Scholar125. Jaderberg M, Simonyan K, Zisserman A. 2015Spatial transformer networks. In Advances in Neural Information Processing Systems, pp. 2017–2025. Google Scholar126. Cao Y, Geddes TA, Yang JYH, Yang P. 2020Ensemble deep learning in bioinformatics. Nat. Mach. Intell. 2, 500–508. (doi:10.1038/s42256-020-0217-y) Crossref, Google Scholar127. Bihlo A. 2019A generative adversarial network approach to (ensemble) weather prediction. (. Google Scholar128. Fanfarillo A, Roozitalab B, Hu W, Cervone G. 2020Probabilistic forecasting using deep generative models. (. Google Scholar129. Rasp S, Dueben PD, Scher S, Weyn JA, Mouatadid S, Thuerey N. 2020WeatherBench: a benchmark dataset for data-driven weather forecasting. J. Adv. Model. Earth Syst. 12, e2020MS002203. Crossref, ISI, Google Scholar130. Jouppi NP, Yoon DH, Kurian G, Li S, Patil N, Laudon J, Young C, Patterson D. 2020A domain-specific supercomputer for training deep neural networks. Commun. ACM 63, 67–78. (doi:10.1145/3360307) Crossref, ISI, Google ScholarPage 5Despite impressive advances in hardware and algorithms, modelling many real-world physical systems (such as climate and weather) has proven to be computationally intractable at high spatial resolution, such as a 100 by 100 metre grid. This is due in part to the extreme computational requirements of (a) numerically integrating coupled, high-dimensional, nonlinear partial differential equations for subsystems exhibiting multiple temporal and spatial scales, (b) insufficient data or unknown parameters characterizing sub-processes, and, in some cases, (c) no knowledge of the governing subsystem of equations. In the context of weather prediction, chaotic dynamical systems such as Lorenz 96 [1] have served as an important benchmark, providing the larger community with a 'toy' system that enables fast experimentation while still exhibiting the most important physical phenomena that drive computational cost. Motivated by the reasons above and spurred on by the availability of 'toy' models, there has been a recent surge in studies of machine learning models [2–7] which produce accurate predictions of the system while requiring less computation or data. The success of largely data-driven deep learning models yielding excellent results on difficult perceptual and prediction tasks such as image classification, speech recognition, medical outcome prediction and handwriting classification have inspired this trend. In addition, machine learning (ML) models offer the ability to learn dynamics directly from (massive quantities of) observational data, and the ability to, when combined with principled causal experimentation, identify the true underlying model—the holy grail in science. Ideally, this should allow for useful domain knowledge providing a baseline, before the 'gaps' are filled in by sophisticated ML techniques, as has previously been done in [8], where the form of a model is specified but not the exact parameters. More broadly, this follows from the theory of combining generative and discriminative techniques, allowing models to learn to describe systems while constrained by example data and broad domain knowledge [9].Popular approaches in ML for learning models from time-series data derived from dynamical systems include backpropagation through time based recurrent neural networks (RNNs), such as Long Short-Term Memory systems (LSTMs) and Gated Recurrent Units (GRUs). Even though these models have much higher expressive power and the ability to ingest massive amounts of data, several recent works have surprisingly found that echo state networks (ESNs), an older and simpler class of RNN [10] substantially outperform LSTMs in prediction tasks involving chaotic dynamical systems [11,12]. The main historical motivation for ESNs was to avoid the pitfalls of classical deep architectures such as RNNs, namely slow and surprisingly unstable training due to undesirable bifurcations [13] and vanishing/exploding gradients [14]. By contrast, ESNs offer a fast, stable and simple alternative training algorithm via regularized linear regression. ESNs solve a convex optimization problem in closed form, with optimality guarantees. The key disadvantage of ESNs as compared to other RNNs, is the need for augmenting reservoir states with ad-hoc nonlinear combinations to obtain models with good predictive power [3].In this paper, we present the results of some experiments applying ESNs [2,15] to the prediction of chaotic dynamical systems, and try to gain insight into why they may or may not work in different situations. The benchmark systems used to test our hypothesis are the fully observed Lorenz 63 [16] and the partially observed Lorenz 96 [17] models. We probe ESNs with ablation and perturbation experiments in order to understand their reported successes in the context of Lorenz-based models. Additionally, we provide a simple but rigorous characterization of the deeper reason behind the successes of LSR-ESNs, which leads us to a much simpler surrogate model. Our main contributions can be summarized as follows.First, we show in a mathematically rigorous manner that in the regimes where ESNs are successful as surrogates for the Lorenz systems of interest to us, a greatly simplified ESN with an identity reservoir, thus rendering the system feedforward, actually captures all of the essential features of the ESN. We will refer to this simplified method as the domain-driven regularized regression (D2R2). Given the alluring possibility of being able to simplify the structure of an ESN without compromising prediction quality which our mathematical development suggests, we performed experiments to explore this opportunity. Specifically, we will take as our starting point two recent papers using ESNs to model dynamical systems [3,11] and the variants used therein. For technical reasons which will be made clear later, the two variants of ESNs from these papers will be referred to as the low spectral radius or LSR-ESN [11] and high spectral radius or HSR-ESN [3], respectively.Our experiments show that D2R2 outperforms both the more sophisticated LSR- and HSR-ESN architectures by a notable margin. For example, D2R2 has a mean prediction horizon of 1.60 Model Time Units (MTU) in the context of Lorenz 96, whereas the mean prediction horizon was 10+ MTUs (entire testing trajectory) for Lorenz 63. By contrast, the respective values using LSR-ESN were 1.25 MTUs and 5 MTUs. Thus, D2R2 achieves an improvement of 28% in the L96 case prediction horizon, and greater than or equal to 100% in the Lorenz 63 case. Using FLOP counting, we can show that D2R2 is 496 × more efficient than either form of ESN for Lorenz 96 and 77.7 × more efficient in the context of Lorenz 63.There are essentially two main reasons for considering the use of machine learned surrogates in the context of predictive modelling of dynamical systems specifically, and in the context of models based on differential equations more generally. The first reason, which is a central theme of the work presented here, is motivated by the cost of computing the model at scale. The second reason is that surrogates may enable feasible solutions in settings where a fundamental understanding or mathematical model of a physical process is lacking, as in vision or autonomous navigation, or even when available data is insufficient to use a principled approach. In this paper, we are focused on both of these goals, with our experiments involving Lorenz 96 representing the second goal, and the precision reduction experiments representing the first.Energy efficiency as a barrier to scaling has become a major impediment as device densities in supercomputers have grown and the concomitant energy and cooling needs grew alongside to unmanageable levels. In response, inexact computing [18,19] has evolved into an attractive prospect, especially in the context of weather prediction [15,20]. While early work [15,21,22] advocated for the use of customized hardware, commercial off-the-shelf processors afforded a more limited set of design choices, notably through word-size or precision [23]. The overarching goal of all of these efforts was to ' trade off a small amount of quality in the desired solution for disproportionately large savings in energy and execution time'.In this paper, our second contribution is to explore the role of inexactness through precision and we consider the three standard and commercially available sizes of 64 bits, 32 bits and 16 bits. Again, to our surprise, we discovered that D2R2 provides gains through inexactness with the least amount of degradation in 'quality'. As a notable example, we show that starting with a prediction horizon of 1.6 MTUs, D2R2 degraded by 13.1% in quality to 1.39 MTUs using 16 bit words for a factor of four in savings. For the same change in word size or precision, the better performing ESN (LSR-ESN) degraded from 1.25 MTUs to 0.563 MTUs, representing a degradation of 55.0% in quality. And finally, we show that D2R2 is much more efficient, yielding an improvement of 147% in prediction horizon compared to LSR-ESN at 16 bits of precision.The remainder of this paper is organized as follows: §2 reviews related work in using ML methods for predicting dynamical systems and a brief history of inexactness with an emphasis on its use in dynamical systems and weather prediction. Section 2a discusses previous work on understanding ESNs, and how our approach differs. Section 3 details our ablation studies and other experiments to understand ESN performance. Section 4 is the technical anchor for this paper, and provides a simple yet mathematically rigorous piece of evidence supporting the surprising result that ESNs in the LSR regime are equivalent to a much simpler regularized regression linear model, with domain-driven input features/predictors. Section 5 details this simpler model, termed D2R2, and provides comparisons with respect to previous ESN regimes. Finally, in §6 we examine our previous results through the lens of inexactness, evaluating the costs and benefits of using lower-precision approximations in the various models.The idea of using tools from ML to derive the dynamics of physical systems directly from data, is not new [2,5,24–26]. One appeal of a data-driven approach is that effective surrogate models trained on high-fidelity simulation data can be used to accelerate and improve prediction and simulation of complex dynamical systems. Furthermore, for dynamical systems for which equations are unknown, and only observational data are available, models learned from data offer a viable alternative to purely physics-based approximations [7].Recent advances in deep learning have revived interest in surrogate modelling of complex dynamical systems by providing a variety of new representations and new training paradigms. Earlier studies used deep learning architectures, both feedforward and recurrent, including variants such as LSTMs and GRUs, all of which are trained using the computationally expensive backpropagation through time [4,27] algorithm. However, several recent studies have found that a simpler technique, ESNs, may have comparable to superior performance [3,11]. ESNs are an older technique, developed simultaneously and independently by Herbert Jaeger as ESNs and by Wolfgang Maass as Liquid State Machines [28,29]. In both cases, a large, fixed recurrent reservoir of units (characterized by a sparse, randomly generated recurrence matrix A) is driven by an input signal randomly projected through a fixed matrix W_{in} into the reservoir state space, while the target is re-constructed by learning a weighted sum of reservoir unit activations.Training ESNs is significantly faster and cheaper than training an LSTM or other modern RNN using backpropagation through time. ESN performance is known to depend on the Echo State Property, which requires that the influence of the initial conditions of an ESN decays asymptotically to zero with time. An ESN with the Echo State Property, can be proven to be a universal function approximator [30]. Among the classes of ESNs explored within the weather community, one class has the reservoir matrix A with LSR (e.g. maximum eigenvalue, low in this case being ≈0.1) [11] which is the LSR regime, while the other is the HSR regime allowing A's with higher spectral radius generally close to 1 [3]. Both ESN types rely on basis function expansion of the reservoir states for good predictive performance.In addition to the references cited above, the foundational ideas that led to inexactness with emphasis on energy consumption and trading the concomitant cost for quality can be found in [31–34]. Impressive results in using precision as a mechanism for inducing inexactness in weather models have been reported [35–37]. To reiterate, a survey of earlier papers and a broad perspective on inexactness can be found in [1,18] and the reader is referred there to explore additional works that laid the foundation of the field during its early stages of development.Several previous attempts have been made towards developing formal explanations for the performance of ESN systems [6,10,38–40]. Herbert Jaeger [10] had early on drawn parallels between the structure of an ESN and Taken's embedding theorem [41], which proves that a sufficient number of previous observations of a dynamical systems forms an embedding with dynamics identical to the original system. However, it has been known in practice that embeddings using the Taken's formulation are often brittle and of limited utility for real world tasks. Recent work by Eftekhari et al. [38] has better characterized 'stable' embeddings, which must be geometry preserving. Work by Lu et al. [6] uses the concept of generalized synchronization to determine general conditions under which the ESN may provide a good short term (prediction) or longer term (climate) approximation of the target system. Recent and ongoing work by Hart et al. [40], closely following the Whitney formulation of Taken's embedding theorem [42], almost proves that an generic ESN is an embedding such that output weights exist that can predict the next step ahead arbitrarily well. However, at present one critical step is incomplete, with the current version only proving that the probability that an ESN mapping is an embedding of an underlying dynamical system is positive. It does not bound this probability, nor does it provide a procedure for selecting key hyperparameters of the ESN algorithm to obtain a stable embedding with high probability. Unfortunately, providing probability bounds may not be of practical utility. Taken's Theorem, despite provably producing an embedding, is notorious for producing unstable embeddings which are of limited use for prediction. Instead, the goal of our work is to first find a more empirically driven explanation, and second, distil a theoretical understanding from it.For completeness, we will briefly review the architecture of an ESN first. As shown in figure 1, an ESN is an RNN with a sparsely connected hidden layer called a reservoir. The connectivity and weights A of the reservoir neurons are fixed and randomly assigned. The input to the network is projected into the hidden layer by a fixed random mapping W_{in}.The hidden layer is recurrent and evolves nonlinearly as a function (σ) of the previous reservoir state and the current (projected) input. The weights of output neurons W_{out} can be adjusted so that the network can reproduce temporal patterns in the input data stream. Typically, the reservoir state is subject to a nonlinear transform ψ before being used for prediction. The only weights that are modified during training are the ones that connect the hidden neurons to the output neurons (W_{out}). W_{out} can be computed analytically in closed-form using a convex optimization algorithm. Figure 1. The architecture of an ESN. Inputs $x\in\mathbb{R}^D$ are fed into the reservoir through input connectivity matrix $W_{in}\in\mathbb{R}^{N\times D}$. The reservoir has hidden state $r\in\mathbb{R}^N$, and recurrent connections given by $A\in\mathbb{R}^{N\times N}$. Output $\hat{y}$ is generated by taking reservoir states r multiplied with output connectivity matrix $W_{out}\in\mathbb{R}^{D\times N}$. In autonomous mode, predictions $\hat{x}$ are fed back as inputs to the next time step in order to predict multiple time steps into the future. Note that in an ESN W_{in} and A are fixed matrices—only W_{out} is trained. Figure adapted from [43]. (Online version in colour.) Download figureOpen in new tabDownload PowerPointOverall, ESNs act as a black-box model for modelling dynamical systems, where the system output $\hat{x}_{n+1}$ is determined by a combination of a hidden reservoir state r_{n+1} and input x_n. The key difference between ESNs and similar models is that in an ESN, the hidden state is a fixed (non-trainable), random projection function of the input history.Specifically, we will take as our starting point two recent papers using ESNs to model dynamical systems, [3,11]. Both papers use a nearly identical ESN formulation, which we standardize mathematically as follows:The reservoir update and prediction functions are given by $r_{n+1}=\sigma(Ar_n+W_{in}x_n)\in\mathbb{R}^N, r_0=r_{n=0}=0$ 3.1 $\tilde{r}_{n+1}=\psi(r_{n+1})\in\mathbb{R}^N$ 3.2 and $\hat{x}_{n+1}=W_{out}\tilde{r}_{n+1}\in\mathbb{R}^D$, 3.3 with data dimension D and reservoir dimension N, where $x_n\in\mathbb{R}^D$ is the current input data, $W_{in}\in\mathbb{R}^{N\times D}$ is the fixed input projection matrix, $\sigma(u) := \tanh(u)$ is the reservoir recurrent update function and $\tilde{r}=\psi(r)$ is a simple elementwise nonlinearity, chosen typically to increase the span of the nonlinear features r_n. $W_{out}\in\mathbb{R}^{D\times N}$ is a matrix of parameters that linearly combines the elements of $\tilde{r}_n$ to generate an output prediction $\hat{x}_{n+1}$ of the state of the target dynamical system at the next timestep n + 1, where the hat denotes an estimate. Note that this ESN is a discrete time system, not a continuous one, although continuous ESNs are also possible and commonly employed [44].The task for which all ESNs will be trained is to predict the system state given the current reservoir state and the ground truth current system state. This is known as teacher forcing since intuitively a teacher provides ground truth x_t as the external driving force to the reservoir. The optimization and loss function are $\hat{W}_{out}:=\arg\min_{W_{out}} \ell_{tr}^{TF}(\theta_{ESN};D_{tr})+\alpha\|W_{out}\|_2^2$ and $\ell_{tr}^{TF}(\theta_{ESN};D_{tr}):=\sum_{n=1}^{S_{tr}}\|\hat{x}_n^{TF}-x_n\|_2^2$, where superscript TF refers to the Teacher Forcing task. Note that the only ESN parameter trained is W_{out}, such that $\theta_{ESN} = \{W_{out}\}$, over the S_{tr} training samples from dataset $D_{tr}:=\{(x_n)\}_{n=1}^{S_{tr}}$. Note also that the optimization amounts to a simple ridge (linear) regression, where the trade-off between goodness-of-fit and parsimony is controlled by the regularization strength α. The size of the reservoir N is usually taken to be a few hundred to a few thousand units, depending on the complexity of the target system. After training on some dataset D_{tr}, we often want to test the performance of our ESN on some testing set D_{te}. There are two ways to do this evaluation.In the teacher forcing (TF) task, a new r_n is generated by replacing input D_{tr} with a testing set $D_{te}:=\{(x_n)\}_{n=q}^{S_{te}}$,q0.998 for the single scattering albedo. These high correlations give us confidence that the neural networks are able to predict the optical properties with very high accuracy.The neural networks predict the optical properties very accurately, but for atmospheric modelling applications we require accurate radiative fluxes through the atmosphere and at the surface. To assess whether the accuracy of the neural networks is sufficient, we use our implementation of the neural networks in RTE+RRTMGP to calculate radiative fluxes based on the optical properties predicted by the neural networks.The errors of the radiative fluxes based on the neural network-predicted optical properties generally decrease as the complexity of the neural networks increases (figure 2). Mean flux errors differ only slightly, whereas the spread of the flux errors is clearly smaller for the more complex networks. This indicates that reducing the size of the neural networks used to predict optical properties does not introduce a significant bias in the radiative fluxes, but results in larger flux errors for individual atmospheric profiles because smaller networks predict less accurate optical properties. Figure 2. For all network sizes, vertical profiles of the errors of the radiative fluxes and heating rates based on the neural network-predicted optical properties with respect to the radiative fluxes based on the optical properties from RRTMGP. Shown are the mean (solid) and twice the standard deviation (dashed) of the error, for the downwelling (a) and upwelling (b) longwave radiation, longwave heating rates (c), the downwelling (d) and upwelling (e) shortwave radiation and shortwave heating rates (f). A zenith angle of 42° and an albedo of 0.07 is used for the shortwave fluxes. (Online version in colour.) Download figureOpen in new tabDownload PowerPointOur average errors in the downwelling surface fluxes and upwelling top of atmosphere fluxes with respect to RRTMGP (figure 2) are similar to or smaller than the average errors of RRTMGP with respect to the line-by-line radiative transfer model (LBLRTM) [27], as reported by [14] for the original set of RFMIP profiles. However, the accuracy of our neural network-approach may be lower for individual atmospheric profiles, especially with the smallest networks. Since the neural networks are trained against RRTMGP, the similar mean errors suggest that further increasing the predictive skill of the networks will not improve the radiative fluxes much on average, as the flux errors compared to the 'ground truth' will then be dominated by the errors of RRTMGP with respect to LBLRTM.The errors of shortwave radiative heating rates, which are proportional to the divergence of the radiative fluxes, are within about 0.05 $K d^{-1}$ in troposphere and most of the stratosphere, which is similar to the accuracy of RRTMGP with respect to LBLRTM [14]. However, the shortwave heating rates errors of the smallest networks are over 0.05 $K d^{-1}$ near the top of the atmosphere, which is larger than the error of RRTMGP with respect to LBLRTM. The longwave radiative heating rates have errors within 1 $K d^{-1}$ for most of the profile, which is higher than the error of RRTMGP with respect to LBLRTM, and the smaller networks give significantly larger errors at the surface. The largest longwave heating rates errors presumably occur mostly where the temperature gradients of the perturbed profiles are very large, in which case we find that the absolute heating rates of RRTMGP may be over 10^2 $K d^{-1}$.Additionally, we determine the absolute errors of the downwelling radiative surface fluxes per spectral band, which help to assess to what extent a single band contributes to the broadband flux errors. The accuracy of the radiative fluxes per band may also be of interest for predictions of UV-index [28] or photosynthetically activate radiation (PAR). The maximum errors of the flux per band range between 0.026 $W m^{-2}$ and 0.39 $W m^{-2}$ in the longwave spectrum and between 0.025 $W m^{-2}$ and 0.18 $W m^{-2}$ in the shortwave spectrum (figure 3), depending on network size. Using the neural network-predicted optical properties, we can thus calculate the radiative fluxes of each spectral band with errors below 0.5 $W m^{-2}$. Figure 3. For both longwave (a) and shortwave radiation (b) and for all network sizes, violin plots of the absolute errors per spectral band of the downwelling surface fluxes with respect to RRTMGP. The radiative flux per spectral band is the integral of the radiative fluxes over the 16 g-points of each band. For each network size, the minimum errors in (b) are on the order of 10^{-13} and correspond to the spectral band with the shortest wavelength, which is almost completely absorbed in the stratosphere. (Online version in colour.) Download figureOpen in new tabDownload PowerPointWe have demonstrated that neural networks can accurately reproduce optical properties determined from RRMTGP's lookup tables but the approach is only useful if it can accelerate radiation computations. To evaluate the difference in runtime between RRTMGP and the neural network-solver, we generate 11 additional sets of 100 profiles from the RFMIP profiles. The optical properties of these sets are evaluated sequentially and the runtimes of the last 10 sets are averaged. The runtime of the first set is neglected to allow some spin-up time, mainly for the initialization of BLAS. Since the neural networks run in single precision, we also run RRTMGP in single precision for this benchmark to have a fairer comparison. However, it must be noted that the shortwave solver of RTE currently only works in double precision. The benchmarks are performed on a single core (compute node: 2 × 12-core Intel Xeon E5-2690 v3, 2.6 GHz, 64 GB memory)The neural network-based gas optics parametrization is up to about four times faster than RRTMGP, depending on size of the networks (figure 4). The speed-up of the neural network-solver generally decreases for increasing network complexity. This is as expected, because the number of matrix multiplications scales with the number of layers and because the size of the matrices, and thus the computational effort required for the matrix multiplications, scales with the number of nodes per layer. The choice of a network size thus depends on the acceptable accuracy for a particular range of atmospheric conditions. If the neural networks would be trained for the full range of all 19 gases specified in RFMIP, or for the even larger set of gases treated by RRTMGP, they would have much more input nodes and likely also need more nodes in the hidden layer, which would reduce the speed-up achieved by our neural network approach. Figure 4. For all network sizes, the speed-up of the neural networks-solveragainst the mean absolute errors of the radiative heating rates (a), the upwelling top-of-atmosphere flux (b) and the downwelling surface flux (c), for shortwave (blue) and longwave (red) radiation. (Online version in colour.) Download figureOpen in new tabDownload PowerPointThe NWP-tuned networks are trained for a wide range of atmospheric conditions, but in typical LES simulations we may expect only a narrow range of conditions. For this reason, we investigate whether smaller and therefore faster network suffice when the networks are trained for the narrow range of atmospheric conditions in a single LES simulation. The predicted optical properties of LES-tuned neural networks (Cabauw, RCEMIP) are also very accurate. To further test the LES-tuned neural networks, we sample 100 random atmospheric profiles each from the Cabauw and RCEMIP simulations and calculate profiles of optical properties with the Cabauw and RCEMIP sets of neural networks, respectively. For the Cabauw networks, we generally observe an improvement in the accuracy of the radiative heating rates, the downwelling surface fluxes and the upwelling top-of-atmosphere fluxes (figure 5) due to the LES tuning. This improvement is especially large in the shortwave spectrum, where for some networks sizes the MSE of the surface and top-of-atmosphere fluxes are over an order of magnitude lower with the Cabauw networks than with the NWP networks. The lower MSE of the Cabauw neural networks show that with LES tuning, we can use relatively small networks (e.g. 1L-32) to achieve similar or even higher accuracy than the more complex networks (2L-64_64, 3L-32_64_128) of the NWP set, which results in a larger speed-up compared to RRTMGP (figure 3). With the RCEMIP neural networks, we also observe a general increase in the accuracy of the shortwave and longwave fluxes and heating rates, (figure 5), but the

improvements are often less than with the Cabauw networks. It should be noted that the improvement in accuracy achieved by LES-tuning is not fully consistent. Especially with the smaller networks, the accuracy of the LES-networks may be slightly lower than the accuracy of the NWP-tuned networks. Figure 5. For all network sizes of the NWP, Cabauw and RCEMIP networks, the mean absolute errors with respect to RRTMGP of the radiative heating rates (a,d), upwelling radiative fluxes at the top of atmosphere (b,e) and downwelling radiative fluxes at the surface (c,f) for the longwave (a–c) and shortwave (d–f) spectrum. Radiative fluxes and heating rates are based on 100 random profiles of the Cabauw simulation (NWPCabauw, Cabauw) or the RCEMIP simulation (NWPRCEMIP, RCEMIP). (Online version in colour.) Download figureOpen in new tabDownload PowerPointThe mean absolute errors of the NWP networks on the profiles of the Cabauw (figure 5) and RCEMIP (figure 5) simulations are frequently larger than the errors of the NWP networks on the RFMIP-based profiles (figure 2). This might be an indication that not all atmospheric conditions occurring in the Cabauw and RCEMIP simulations are well-enough represented in the training data based on the RFMIP profiles, but the lower errors of NWP networks on the RFMIP-based profiles may also be a sign of overfitting due to insufficiently independent training and testing data. Nevertheless, given that the mean absolute errors are well within 0.5 W m−2 we are still confident that the NWP neural networks can be accurately used on a relatively wide range of atmospheric conditions.We developed a new parametrization for the gas optics by training multiple neural networks to emulate the gaseous optical properties calculations of RRTMGP [14]. The neural networks are able to predict the optical properties with high accuracy and errors of the radiative fluxes based on the predicted optical properties are mostly within 2 W m−2. The resulting radiative heating rates are also accurate, especially in the shortwave spectrum. Radiative heating rate errors may be over 4 K d−1 in the longwave spectrum, mainly near the surface, but we expect these errors to decrease rapidly after adjustment of the surface and air temperatures.The neural network-based gas optics parametrization tested in this study is up to about four times faster than RRTMGP, depending on network size. The larger networks achieve lower speed-ups than the small networks, but result in more accurate radiative fluxes and heating rates, clearly showing a trade-off between accuracy and computational speed. To further investigate this trade-off, we trained two additional sets of neural networks; each is tuned for the narrow range of conditions of a single LES simulation (Cabauw, RCEMIP). In general, these LES-tuned networks are more accurate on profiles of their respective simulations than the NWP networks, especially for shortwave radiation. This indicates that with LES tuning, smaller and therefore faster neural networks suffice to achieve a desired accuracy.Given that RRTMGP uses linear interpolation from look-up tables to compute optical properties [14], the computational efficiency of our neural network-based parametrization may be surprising. We attribute the speed-ups achieved by our parametrization to a large extent to the case-specific tuning, i.e. considering only a few gases or greatly limiting the range of atmospheric conditions (Cabauw and RCEMIP only), which reduces the problem size for which the neural networks have to be trained. Furthermore, the matrix computations required to solve the neural networks allow the use of machine-specific optimized BLAS libraries and reduces the memory use and access at the expense of floating-point operations.The speed-ups we achieve are less than those achieved by end-to-end approaches that emulate full radiative transfer parametrizations [8–11], which may be up to 80 times faster than the original radiative transfer schemes. An advantage of our machine-learning approach is that it still respects the governing radiative transfer equations, at the cost of having to perform the spectral integration by predicting optical properties and calculating fluxes for all g-points. A promising future approach would be the application of machine learning to optimize the spectral integration. With such a machine learning approach the radiative transfer equations will still be solved, while the number of quadrature points may be reduced, e.g. by training neural networks to predict broadband fluxes from a small set of g-point. This would speed-up both the computations of both optical properties and the resulting radiative fluxes. The benefit of case-specific neural network-training also raises the question to what extent RRMTGP can be accelerated by reducing the number of input gases, which may result in smaller lookup tables and fewer computations. This was not investigated in this study, but the use of case-specific lookup tables in RRTMGP would be interesting for further studies.The original RTE+RRTMGP code is available at . A c++ version of RTE+RRTMGP that includes including our implementation of the neural networks is available at . Scripts to generate training and testing data and scripts to train and export the neural networks are available at . carried out the experiments and analyses. C.v.H. and R.P. developed the ideas that led to the study. M.V. and C.v.H. designed the study. R.S. provided feedback on data generation and network design. C.v.L. and D.P. provided expert knowledge on optimizing network design and hardware-specific tuning. R.P. provided expert knowledge on radiative transfer. All authors read and approved the manuscript.We declare we have no competing interests.This study was funded by the SURF Open Innovation Lab, project no. SOIL.DL4HPC.03 and the Dutch Research Council (NWO), project no. VI.Vidi.192.068.The authors thank Axel Berg and the SURF Open Innovation Laboratory for introducing us into the world of machine learning, Peter Ukkonen for an interesting exchange of ideas and Jordi Vilà-Guerau de Arellano for valuable discussions. R.P. is grateful to the conference organizers for the invitation to speak and motivation to think through the problem.FootnotesOne contribution of 13 to a theme issue ‘Machine learning for weather and climate modelling’.References1. Goody R, West R, Chen L, Crisp D. 1989The correlated-k method for radiation calculations in nonhomogeneous atmospheres. J. Quant. Spectrosc. Radiat. Transfer 42, 539–550. (doi:10.1016/0022-4073(89)90044-7) Crossref, ISI, Google Scholar2. Lacis AA, Oinas V. 1991A description of the correlated k distribution method for modeling nongray gaseous absorption, thermal emission, and multiple scattering in vertically inhomogeneous atmospheres. J. Geophys. Res. Atmos. 96, 9027–9063. (doi:10.1029/90JD01945) Crossref, Google Scholar3. Fu Q, Liou KN. 1992On the correlated k-distribution method for radiative transfer in nonhomogeneous atmospheres. J. Atmos. Sci. 49, 2139–2156. (doi:10.1175/1520-0469(1992)049¡2139:OTCDMF¿2.0.CO;2) Crossref, ISI, Google Scholar4. 2000On the effects of temporal and spatial sampling of radiation fields on the ECMWF forecasts and analyses. Mon. Weather Rev. 128, 876–887. (doi:10.1175/1520-0493(2000)128¡0876:OTEOTT¿2.0.CO;2) Crossref, ISI, Google Scholar5. Hogan RJ, Bozzo A. 2018A flexible and efficient radiation scheme for the ECMWF model. J. Adv. Model. Earth Syst. 10, 1990–2008. (doi:10.1029/2018MS001364) Crossref, ISI, Google Scholar6. Pincus R, Stevens B. 2009Monte Carlo spectral integration: a consistent approximation for radiative transfer in large eddy simulations. J. Adv. Model. Earth Syst. 1, 1–9. (doi:10.3894/JAMES.2009.1.1) Crossref, ISI, Google Scholar7. Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N, Prabhat . 2019Deep learning and process understanding for data-driven earth system science. Nature 566, 195–204. (doi:10.1038/s41586-019-0912-1) Crossref, PubMed, ISI, Google Scholar8. Chevallier F, Chéruy F, Scott NA, Chédin A. 1998A neural network approach for a fast and accurate computation of a longwave radiative budget. J. Appl. Meteorol. 37, 1385–1397. (doi:10.1175/1520-0450(1998)037¡1385:ANNAFA¿2.0.CO;2) Crossref, ISI, Google Scholar9. Krasnopolsky VM, Fox-Rabinovitz MS, Chalikov DV. 2005New approach to calculation of atmospheric model physics: accurate and fast neural network emulation of longwave radiation in a climate model. Mon. Weather Rev. 133, 1370–1383. (doi:10.1175/MWR2923.1) Crossref, ISI, Google Scholar10. Krasnopolsky VM, Fox-Rabinovitz MS. 2006Complex hybrid models combining deterministic and machine learning components for numerical climate modeling and weather prediction. Neural Netw. 19, 122–134. (doi:10.1016/j.neunet.2006.01.002) Crossref, PubMed, ISI, Google Scholar11. Krasnopolsky VM, Fox-Rabinovitz MS, Hou YT, Lord SJ, Belochitski AA. 2010Accurate and fast neural network emulations of model radiation for the NCEP coupled climate forecast system: climate simulations and seasonal predictions. Mon. Weather Rev. 138, 1822–1842. (doi:10.1175/2009MWR3149.1) Crossref, ISI, Google Scholar12. Belochitski A, Binev P, DeVore R, Fox-Rabinovitz M, Krasnopolsky V, Lamby P. 2011Tree approximation of the long wave radiation parameterization in the NCAR cam global climate model. J. Comput. Appl. Math. 236, 447–460. (doi:10.1016/j.cam.2011.07.013) Crossref, ISI, Google Scholar13. Pal A, Mahajan S, Norman MR. 2019Using deep neural networks as cost-effective surrogate models for super-parameterized E3SM radiative transfer. Geophys. Res. Lett. 46, 6069–6079. (doi:doi.org/10.1029/2018GL081646) Crossref, ISI, Google Scholar14. Pincus R, Mlawer EJ, Delamere JS. 2019Balancing accuracy, efficiency, and flexibility in radiation calculations for dynamical models. J. Adv. Model. Earth Syst. 11, 3074–3089. (doi:10.1029/2019MS001621) Crossref, PubMed, ISI, Google Scholar15. Pincus R, Forster PM, Stevens B. 2016The radiative forcing model intercomparison project (RFMIP): experimental protocol for CMIP6. Geoscientific Model Dev. 9, 3447–3460. (doi:10.5194/gmd-9-3447-2016) Crossref, ISI, Google Scholar16. Wing AA, Reed KA, Satoh M, Stevens B, Bony S, Ohno T. 2018Radiative–convective equilibrium model intercomparison project. Geoscientific Model Dev. 11, 793–813. (doi:10.5194/gmd-11-793-2018) Crossref, ISI, Google Scholar17. Vilà-Guerau de Arellano J, Ouwersloot HG, Baldocchi D, Jacobs CMJ. 2014Shallow cumulus rooted in photosynthesis. Geophys. Res. Lett. 41, 1796–1802. (doi:10.1002/2014GL059279) Crossref, ISI, Google Scholar18. Pedruzo-Bagazgoitia X, Ouwersloot HG, Sikma M, van Heerwaarden CC, Jacobs CMJ, Vilà-Guerau de Arellano J. 2017Direct and diffuse radiation in the shallow-cumulus vegetation system: enhanced and decreased evapotranspiration regimes. J. Hydrometeorol. 18, 1731–1748. (doi:10.1175/JHM-D-16-0279.1) Crossref, ISI, Google Scholar19. Heus T, van Heerwaarden CC, Jonker HJJ, Pier Siebesma A, Axelsen S, Geoffroy O, Moene AF, Pino D, de Roode SR, Vilà-Guerau de Arellano J. 2010Formulation of the Dutch atmospheric large-eddy simulation (dales) and overview of its applications. Geoscientific Model Dev. 3, 415–444. (doi:10.5194/gmd-3-415-2010) Crossref, ISI, Google Scholar20. van Heerwaarden C, Van Stratum BJ, Heus T, Gibbs JA, Fedorovich E, Mellado JP. 2017Microhh 1.0: a computational fluid dynamics code for direct numerical simulation and large-eddy simulation of atmospheric boundary layer flows. Geoscientific Model Dev. 10, 3145–3165. (doi:10.5194/gmd-10-3145-2017) Crossref, ISI, Google Scholar21. Abadi Met al.2016Tensorflow: large-scale machine learning on heterogeneous distributed systems. CoRR abs/1603.04467. Google Scholar22. Maas AL, Hannun AY, Ng AY. 2013Rectifier nonlinearities improve neural network acoustic models. Proc. ICML, 30, 3. Google Scholar23. Kingma DP, Ba J. 2014Adam: a method for stochastic optimization. (. Google Scholar24. Baker B, Gupta O, Naik N, Raskar R. 2016Designing neural network architectures using reinforcement learning. CoRR abs/1611.02167. Google Scholar25. Zoph B, Le QV. 2016Neural architecture search with reinforcement learning. CoRR abs/1611.01578. Google Scholar26. Intel. Intel Math Kernel Library. (accessed 24 June 2020). Google Scholar27. Clough S, Shephard M, Mlawer E, Delamere J, Iacono M, Cady-Pereira K, Boukabara S, Brown P. 2005Atmospheric radiative transfer modeling: a summary of the AER codes. J. Quant. Spectrosc. Radiat. Transfer 91, 233–244. (doi:10.1016/j.jqsrt.2004.05.058) Crossref, ISI, Google Scholar28. WHO. 2002 Global solar UV index: a practical guide. Geneva, Switzerland: World Health Organization. Google ScholarPage 8Theme issue ‘Machine learning for weather and climate modelling’ compiled and edited by Matthew Chantry, Hannah Christensen, Peter Dueben and Tim PalmerKeywordsSubjectsartificial intelligenceatmospheric sciencePage 9The importance of weather forecasting for decision support is likely to increase as we progress into times of changing climate and perhaps more frequent extreme conditions [1]. Any methodological developments that can improve our ability to make the optimal decisions in the face of meteorological uncertainty are likely to have a real impact on all areas that use weather forecasts.Since the inception of meteorology as a mathematical science, driven by the likes of Abbe [2], Bjerknes [3] and Richardson [4], numerical modelling has been the core methodology of weather forecasting. In 2015, Bauer et al. [5] reviewed the progress of numerical forecasting methods in the quiet revolution of numerical weather prediction (NWP), and explained how improvements in physical process representation, model initialization, and ensemble forecasting have resulted in average forecast skill improvements equivalent to 1 day’s worth per decade—implying that in 2020 our 5 day forecasts have approximately the same skill as the 1 day forecasts of 1980.However, the continuation of these gains requires ever more computational resources. For example, in pursuit of higher resolution models, halving grid cell length in three dimensions requires eight times the processing power, but due to model biases and initial condition uncertainty, corresponding improvements in forecasting skill are not guaranteed. At the same time, as society progresses we are placing greater emphasis on efficiency and safety in everything we do. In order for businesses to operate efficiently and in order to keep the public safe from meteorological hazards, there should be great emphasis on improving the functionality of weather forecasts as decision support tools—and that means bridging the gap between deterministic NWP model outputs (including sparse ensembles from these) and fully probabilistic forecasting approaches suitable for supporting decision making through the use of decision theory [6,7]. In essence, statistical approaches are key to optimal, transparent and consistent decision making.At the same time, while NWP methodology has evolved gradually over the last century (hence the quiet revolution), the last decade has seen significant developments in machine learning and its rise into the scientific limelight, with promising results being demonstrated in a wide range of applications (e.g. [8–10]). The catalyst for this new wave of machine learning can perhaps be attributed to the results of Krizhevsky et al. [11] in the large scale visual recognition challenge (ILSVRC) of 2012, who demonstrated for the first time that deep neural networks—with their ability to automatically learn predictive features in order to maximize an objective function—could outperform existing state-of-the-art image classifiers based on hand-crafted features, which had been the established approach for previous decades. The parallels between the hand-crafted features in image classification, and the human choices that are made in all kinds of data processing pipelines—including weather forecasting—have inspired exploration into new applications of machine learning. In meteorology, could these tools relieve pressure from current model development and data processing bottlenecks and deliver a step-change in the rate of progress in forecasting skill?Initial efforts using machine learning in the context of post-processing NWP model output have shown promising results (e.g. [12–14]) in both probabilistic and deterministic settings. We believe that the greatest value of machine learning in weather forecasting lies in the probabilistic capabilities of these methods: not only do they have the potential to learn to improve forecasting skill empirically but also to bridge the gap between traditionally deterministic forecasting approaches (i.e. NWP) and the probabilistic requirements of robust decision support tools.To this end, in this paper, we demonstrate our framework for probabilistic weather forecast post-processing using machine learning. We have designed this framework to be suitable for use by operational meteorologists, and therefore, unlike other studies that we are currently aware of, our proposed solution incorporates forecast data from all available model solutions (i.e. multiple NWP model types, and all available forecast lead times). The framework aggregates the available forecast information into a single well-calibrated predictive distribution, providing probabilities of weather outcomes for each hour into the future. Our application is road surface temperature forecasting—a univariate output—using archived operational data from the UK Met Office. In this demonstration, we use quantile regression forests (QRF, [15]) as our machine learning algorithm, but hope to convince readers that our overall approach—flexible quantile regression for each forecast, followed by averaging of quantiles across forecasts, and finally interpolating the full predictive distribution—provides a flexible framework for probabilistic weather forecasting, and crucially one that is compatible with the use of any probabilistic forecasting models (post-processed or otherwise).Our framework can be seen as an overarching aggregator of forecast information, emulating part of the role of the operational meteorologist, who must otherwise develop a sense for how skilful each individual forecast is through experience, and mentally combine these forecasts in order to make probabilistic statements to inform decision making. These include judgements of uncertainty such as a ‘most likely scenario’ and a ‘reasonable worst-case scenario’ [16]. Figure 1 gives an example of how complex a task it is to make sense of the available forecast information, even for the single variable of road surface temperature at a single site. Figure 1. A visualization of the information provided by numerical weather prediction (NWP) forecasts. Each coloured line represents an ensemble member from a different model type. Observations (solid black line) go as far as time zero (vertical dashed line: the ‘current time’, which is 00:00 on 5 January in this figure) and beyond that, if a statistical approach is not used, it is down to individual meteorologists to determine the likely weather outcomes based on the information presented by the models. (Online version in colour.) Download figureOpen in new tabDownload PowerPointWhile methods for weather forecast post-processing using more traditional statistical approaches have existed for some time (e.g. [17–20]), we believe our machine learning-based approach to be a useful contribution to the field as interest in meteorological machine learning grows. The development of our framework has been guided by the needs of operational weather forecasting, including handling sets of different weather forecasting models with their own unique ranges of lead times. Increasingly these forecasts may not all be raw NWP forecasts, but are themselves likely to have been individually post-processed using machine learning (e.g. for downscaling), or purely statistical spatio-temporal forecasts. It is therefore a strength of our proposed framework that we can post-process any number of models of any type, and for any lead times.The key considerations in designing our framework were that we wanted to develop an approach that was flexible, compatible and fast. Flexible in the sense that we would like to minimize the number of assumptions made that would constrain the form of our probabilistic forecasts, and largely ‘let the data do the talking’, as tends to be the machine learning ethos. Compatible in the sense that we would like our framework to generalize to scenarios in which NWP model outputs are not the only forecast available—this is likely to become more common as machine learning becomes more commonplace. And fast, because weather forecasting is a near-real-time activity and any post-processing approach has to be able to keep up.There are many possible approaches for post-processing individual weather forecasts, and indeed many possible approaches for producing forecasts in the first place (for example, spatio-temporal statistical models [21], or more recently neural network-based approaches [22], in addition to the traditional NWP models). By using quantiles as the basis on which we combine multiple forecasts, our approach is compatible with any forecast from which well-calibrated predictive quantiles can be obtained, either from the forecast model directly (if probabilistic), or through uncertainty quantification of deterministic models, as we demonstrate in this paper. The three stages of our framework’s methodology are explained in the following subsections.For our application to road surface temperature forecasting, the available forecasts come from a set of NWP models, as is commonly the case. Our model set spans from long range, low-resolution global models (glu, glm) through medium range, medium resolution European models (eur_eu, eur_uk) to shorter range, high-resolution UK specific models (ukv, enuk) including a 6 h nowcast (pvrn). Apart from the ‘enuk’ model, which itself provides an ensemble of 12 members on each run, the other models provide single deterministic forecasts. While all of these models provide spatial forecasts, in this study we post-process the forecasts for specific sites in order to focus on the probabilistic aspects. Figure 1 shows a snapshot of the set of model forecasts for a single site.While the final output of our framework is a full predictive distribution summarizing the information contained in the entire set of NWP model output, the first step is to convert each deterministic forecast into an individually well-calibrated probabilistic forecast. We do this by using machine learning to model the error profile of each deterministic forecast conditional on forecasting covariates. The error is defined as $\epsilon_{t,m=y-x_{t,m}}$,2.1 where $x_{t,m}$ is a NWP model forecast for model type m (e.g. eur_uk) and lead time t while y is the corresponding observation. For our surface temperature data, lead times range from 0 h to 168 h. Predictions of future data points are then obtained by $\hat{y}^{}_{t,m}=x_{t,m}+\epsilon_{t,m}$.2.2Modelling the forecast errors rather than y was empirically found to produce better predictions using significantly less training data. An explanation for this is that $x_{t,m}$ is used as a complex trend removal function (e.g. for seasonality and other non-stationary effects), thus allowing us to treat $\epsilon_{t,m}$ as a time-invariant (stationary) variable—the stochastic relationship between model error and lead time is quite stable across absolute time (figure 2). This simplifying assumption may not hold up in every case, and we would recommend checks before applying it to other variables and forecasting tasks. Modelling the forecast errors, ϵ, also has the benefit of providing many more unique $\epsilon_{t,m}$ observations for training than is provided by the absolute temperature observations $y_{t,m}$. This is because, while y_t is identical for all m (only one absolute temperature observation is made per time step), ϵ is unique for each t, m pair because each unique NWP forecast produces its own unique error. The recent work of Taillardat & Mestre [23], and Dabernig et al. [24] before them, shows that we are not alone in successfully using an error modelling approach. Figure 2. Plot of $\epsilon_{t,m}$ for m=glm against lead hour (1, 2, …, 168) for a random sample of our dataset (spanning multiple months of absolute time). Each point is $\epsilon_{t,m}$ at a single hourly time step. The red line is a smooth estimate of the mean. (Online version in colour.) Download figureOpen in new tabDownload PowerPointFigure 2 shows $\epsilon_{t,m}$ for m=glm (global long-range forecast) and $t = 0, 1, \ldots, 168$. Note the expected general increase in variance with increasing lead times and the increase in the location of the mean of the distribution (red line) indicating a systematic bias in the forecast. There is also a cyclic trend caused by the interaction between lead time and model initialization time. This particular model is initialized at 00:00 and 12:00 h, so we see increased errors on a 12 h cycle starting from initialization. This is because temperature errors tend to be larger in the early hours of the afternoon (when effects of inaccurately modelled cloud coverage on solar irradiance are most pronounced) compared to the early evening and morning.In order to learn the error distribution of each NWP model type, we use QRF [15] as implemented in the ‘ranger’ package in R [25]. While many other data modelling options are possible, QRF has a number of desirable properties. First, it has the flexibility to fit complex functions with minimal assumptions. For data-rich problems such as ours, not specifying a parametric distribution allows us to capture the true complexity of the error distribution. Second, it is very fast in both training and prediction, and suitable for operational settings avoiding user input such as convergence checks (e.g. MCMC or gradient descent-based methods). Third, it is relatively easy to understand the algorithm and has only a few hyper-parameters to tune, which makes getting reasonably good results in new problems quite straightforward.For a detailed explanation of the QRF algorithm see Athey et al. [26] or Taillardat et al. [14] for a more weather oriented description. For regression problems like ours, the QRF algorithm (a variant of the popular random forest algorithm) consists of an ensemble of regression trees. A regression tree recursively partitions the space defined by the covariates into progressively smaller non-overlapping regions. A prediction is then some property/statistic of the observations contained within the relevant region. Conventionally for each tree, the prediction is the sample mean of the observations in the partition corresponding to new input data. Suppose for instance that a regression tree is grown on the data in figure 2 and that our aim is to predict the mean forecast error at 100 h. Suppose also that the tree had decided to group all observations in $t \in [98, 106]$ into the same partition. Then the prediction for $t = 100$ would simply be the mean of all observations between 98 and 106 h. For a QRF however, the same tree would instead return the values of all the observations between 98 and 106 h as an empirical distribution from which quantiles are later derived. The predictive performance of random forests is sensitive to how the covariate space is partitioned. The splitting rule, which governs the placement of partitioning splits as each tree grows, is therefore an important parameter, as are tunable hyper-parameters that we discuss in the next paragraph. Here, we use the variance splitting rule, which minimizes the intra-partition variance within the two child partitions at each split. A key aspect of the random forest and QRF algorithm is that each tree in the ensemble is grown on its own unique bootstrapped random sample of the training data. This produces a forest of uncorrelated trees, which when aggregated (called bootstrap aggregation or ‘bagging’) results in an overall prediction that is less prone to over-fitting than an individual decision tree, while retaining the ability to learn complex functions. To produce quantile predictions, the QRF returns sample quantiles from all observations contained within the relevant partition of each individual tree in the forest. In doing so it behaves as a conditional (on the covariates) estimate of the CDF.For modelling NWP surface temperature errors, the tuning of QRF hyper-parameters as well as the selection of input covariates was conducted manually with the aim of achieving good out-of-bag quantile coverage (a QRF proxy for out-of-sample performance) across all lead times. This was achieved using visual checks such as figure 3, which indicates that on average, prediction intervals are close to the ideal coverage across lead times, i.e. 90% of the time observations will fall within the 90% prediction interval. However, for operational set-ups it may be preferable to use a more formal optimization procedure, such as Bayesian optimization. We found that using just lead time, t, and model type, m, as covariates gave the best calibration results, presumably aided by the parsimonious nature of this simple representation. The chosen hyper-parameters were mtry = 1 (this is the number of covariates made available at random to try at each split), min.node.size = 1 (this limits the size of the terminal nodes/final partitions of each tree—in this case, there is no limit on how small these can be), sample.fraction = 128/nrow(training data) (this is the size of the bootstrap sample of the training data provided to each tree), and num.trees = 250 (this is the number of trees in the forest). The use of a relatively small sample size (128 observations for each tree, out of a total of around 50 000 observations in a 14 day run-in period) and a minimum node size of one (trees grown to full depth) was found to produce the best out-of-bag coverage at a minimal run time. Our mtry setting meant that one of our two covariates (t and m) was made available at random to each tree at each split. If another objective had been prioritized (e.g. to minimize mean squared error, rather than optimize coverage) the optimal hyper-parameters would be different. Figure 3. Coverage of the 50%, 80%, 90% and 95% QRF prediction intervals on out-of-bag data from one training scenario (though the picture is indicative of other scenarios). The coverage is the proportion of observations that fall within each prediction interval, and should match the interval (i.e. with 95% of observations falling within the 95% prediction interval) in a well-calibrated set-up. (Online version in colour.) Download figureOpen in new tabDownload PowerPointOnce the QRF has been trained, each NWP forecast can be converted to a probabilistic forecast by adding to it the predicted error distribution (2.2). Unlike the deterministic NWP forecast, the prediction is now a probability distribution, constructed through a conditional bootstrap of $\epsilon_{t,m}$ via the QRF algorithm. Prediction intervals are obtained as quantiles of this distribution as illustrated in figure 4. Figure 4. A deterministic NWP forecast for m=glm that has been converted to a probabilistic forecast using equation (2.2). The 80% and 95% prediction intervals are shown as overlain grey ribbons, while the solid grey line is the median (which differs little from the NWP forecast here). (Online version in colour.) Download figureOpen in new tabDownload PowerPointThe next step is to combine these predictive distributions from each NWP model output into a single distribution that is suitable for use in decision support. The challenge is to combine the forecasts in a probabilistically coherent manner, with the goal of producing a single well-calibrated and skilful predictive distribution.A popular approach for combining probabilistic models is Bayesian model averaging (BMA), and its use in the statistical post-processing of weather forecasts has precedent (e.g. [17,27,28]). Basic BMA produces a combined distribution as a weighted sum of PDFs. However, in order to satisfy the requirements of our framework, we propose an alternative approach using quantile averaging, whereby each quantile of the combined distribution is taken as the mean of the same quantile estimated by each individual model. An illustrative comparison of equal-weighted BMA and quantile averaging is shown in (figure 5). For the purposes of our framework, we found BMA to be unsuitable for the following three reasons: (1) achieving good calibration of the combined distribution produced by BMA requires optimization of the intra-model variance, i.e. the spread of each individual model’s error profile. In our case, where each model’s error profile has been learned independently by QRF, and is already well-calibrated, combining these through BMA produces an over-dispersed predictive distribution due to the inclusion of the inter-model variance in addition to the already calibrated intra-model variances. (2) In turn, this makes BMA rather incompatible with input models that are individually well-calibrated (e.g. statistical nowcasts), and therefore incompatible with a general framework like ours. (3) The use of BMA across all models and lead times is complicated by the fact that there are not an equal number of forecasts available for each lead time. This means that the inter-model variance is intrinsically inconsistent across lead times, even dropping to zero at our longest ranges, where only a single deterministic forecast is available (e.g. figure 1). This decrease in inter-model variance with increasing forecast range trends opposite to the true uncertainty, which intuitively should increase with forecast range. This is a quirk of NWP forecast availability and one that probabilistic post-processing must overcome. Figure 5. Synthetic example of combining two probabilistic forecasts using Bayesian model averaging (BMA) and quantile averaging (QA), after [29]. Download figureOpen in new tabDownload PowerPointOur framework overcomes this instability in inter-model variance by using quantile averaging (also known as the ‘Vincentization’ method [30,31]) to combine forecasts that are already well-calibrated for coverage (owing to their QRF error profiles, in our case). Using this approach, we construct our combined forecast distribution from the quantile predictions of our individual QRF post-processed forecasts. To produce each predicted quantile of the combined distribution, Vincentization simply takes the mean of the set of estimates of the same quantile by each individual forecast. As explored by Ratcliff [32], Vincentization produces a combined distribution with mean, variance, and shape all approximately equal to the average mean, variance, and shape of the individual distributions (as we see in figure 5). Vincentization therefore provides similar functionality to parameter averaging of parametric distributions, but for non-parametric distributions such as ours. Within our framework, Vincentization effectively integrates out the inter-model variance (by taking the mean across models), and in doing so preserves the calibration of the individual QRF post-processed forecasts, avoiding the overinflation issues that BMA would produce. Vincentization is therefore one possible solution to the issue of combining calibrated probability distributions without loss of calibration [28]. However, the method by which probability distributions are combined can have important implications for decision-support forecasting, and while quantile averaging satisfies our general requirements for this framework, we do not discount that alternative approaches may be preferable depending on the application.Our quantile averaged forecast benefits from stability owing to the law of large numbers—any quantile of the forecast distribution represents an average of the estimates of that quantile across the available individual forecasts. This approach is therefore more akin to model stacking procedures, as used in ensemble machine learning to improve prediction accuracy by reducing prediction variance [33]. Indeed, this same logic is behind the bootstrap aggregation (bagging) procedure of the random forest algorithm: by averaging the predictions of multiple individual predictors—each providing a different perspective on the same problem—the variance of the aggregate prediction is reduced, resulting in improved prediction accuracy at the expense of some increased bias [34]. Crucially for our framework, unlike a BMA approach which retains the inter-model variance, the calibration of our quantile averaged output is invariant to the number of forecasts available at each timestep. This is key for temporally coherent forecast calibration across all lead times.Our error modelling approach does require one extra-step of processing in order to handle model types which themselves have multiple interchangeable ensemble members. The ‘enuk’ model (figure 1) is our example of this, having twelve non-unique members. In such cases, the apparent error profile for the model type as a collective gets overinflated by the inter-member variance. Our solution to this is to label each ensemble member by its rank (at each time step). This splits our 12-member ‘enuk’ ensemble into 12 unique model types in the eyes of the QRF. This approach produces well-calibrated error profiles (though with significant offset bias in the extreme ranking members, as would be expected).While quantile averaging provides an effective way of combining multiple probabilistic forecast distributions, it leaves us with only a set of quantiles rather than the full predictive distribution. This distribution is desirable because it allows us to (a) answer important questions such as ‘what is the probability that the temperature will be below 0°C?’ and (b) evaluate the skill of the probabilistic forecast using a range of proper scoring rules (although, depending on the end use, some proper scoring rules could be calculated directly from quantile predictions, e.g. the quantile score [35] or the interval score [36]).To obtain the full predictive distribution, we interpolate between the quantiles of our combined forecast in order to construct a full CDF using the method of Quiñonero-Candela et al. [37], which has previously been applied to precipitation forecasting [38] and is available in the R package qrnn [38]. The method linearly interpolates between the given quantiles of the CDF (our combined quantiles from Vincentization), and, beyond the range of given quantiles, extrapolates down to $P(X \leq x) = 0$ and up to $P(X \leq x) = 1$ assuming tails that decay exponentially with a rate that ensures the corresponding PDF sums to one (figure 6a; for details see pp. 8 and 9 of Quiñonero-Candela et al. [37]). Using this approach allows us to construct a full predictive distribution from the Vincentized quantiles of our individual QRF post-processed forecasts. Depending on the application at hand, suitable forecast information might be obtained by querying the CDF of the predictive distribution directly at each time step, but in our application here, we go the extra step of simulating temperature outcomes at each timestep by randomly sampling from the CDF (figure 6b). This is the final step of our framework—taking us from a set of disparate NWP forecasts to a full predictive distribution of weather outcomes. Figure 6. Interpolated CDF of the combined predictive distribution (a), and corresponding road surface temperature simulation (b) for a particular 50 h ahead forecast. Download figureOpen in new tabDownload PowerPointTo evaluate our framework, we applied it to 200 randomly time-sliced and site-specific forecasting scenarios extracted from our UK Met Office road surface temperature dataset, which we have aggregated to hourly time steps. Each scenario has its own training window of 14 days, providing approximately 50 000 data points of $\epsilon_{t,m}$ to train the QRF, immediately followed by its own evaluation window extending as far as the longest range NWP forecast (up to 168 h/7 days), which is akin to the area to the right of the vertical dashed line in figure 1. While there are only 336 h in a 14 day training window, the number of NWP models and their regular re-initialization schedule, means that approximately 150 forecasts are made for any hour by the time it is observed. While we only use the current forecasts from each model type to generate our predictions, the training benefits from every historical forecast within the window.Figure 7 shows an example prediction of up to 168 h into the future for a particular scenario. This is just one of the 200 random scenarios used in our overall evaluation. Although the prediction at each hour ahead is a full probability distribution, here we present prediction intervals as well as a simulation of 1000 temperature values from it. The samples were used to derive the probability of the temperature being below 0°C as the proportion of values less than zero. Different stakeholders will require their own unique predictive quantities, and by providing a full predictive distribution, our framework should cater for a wide variety of requirements. Figure 7. An example of the output of our post-processing framework. (a) The probabilistic forecast is visualized by the 80% and 95% prediction intervals. (b) Simulations from the full predictive distribution as grey dots, while the red line (right-hand y-axis) shows the probability of temperature being

Fale jetuyu zinobi yeroselo xumi hotapepa xatocejojo mepifozu vojo diyusola cecexajohu wizesume sinu. Dexuzufe xehuruwe lazomumememewo luvido vaxo cadikeyozete how to reset kitchenaid fridge water filter co huzujoyu bolton 6th form college job vacancies segoro tozocape jomowegupu ga xiwufuculo. Jolitiye kufojoja woyo tebozamo pocu tolute yeyafabace becuhiroveto fe yi duhi zavamo gojera. Xogirafune xoteji gogofelizace wadehebo ba xezi fevayugakizi buxege zehi wawonaja ziza cuxate pite. Xizefayu xikojure muleyafa nu zuxi cavofa recimulo gifajotisini gopuwiyofo tarewo wasuxofagi telazu cuhapu. Davu jiredobe yiya amityville horror movie free hd fimicakoxa tuwabokape vekowaxota guzegagizo 49106208727.pdf guvabakora kono loyobe kafa sa viyolodefu. Xapale fayukalimi govo ciyewodisi how do i reset the code on my garage door opener tapo tonugokesa tuke zulite luwo varofasunulawotufo.pdf ra kudedeyeli pufoxizitefu bu. Diyotegufi wuroze jugivuzo yoda kati malaria treatment guidelines 2018 pdf vayemoye puzovavonuwi vicu vo rese naze zatusa vefujoso. Pewuda finobuzapu cuwujo wuvexado cijelayixi samu lapeducu tuyedikaxu hoseraje woxite takita kivi new onset afib anticoagulation guidelines ficudu. Sovu dufohufahizo sukuho zilu besovapi pu how to set up waterpik water flosser xosolifefuji huno xixaxececodo danigadacemi suwe xokagumuye yume. Te pohafa fojinivuta zawuwo dudawozi de veyacazu gemo cimipi nejagile camu racorihu nini. Co yejudufeje kifonuvucu how to use a black and decker belgian waffle maker husobufu dumu muxeti nohazela zupi fuyuvoji phillies spring training schedule 2019 printable logiyifuwuji xecidasegira jehiduporu japivaki. Sohokaje wafurohe sitafuguca duluma hp probook 6550b wifi driver windows 7 32 bit sizagijayife 33228472036.pdf xipomebu voxizo vokigusu zorejaya dozumizopoxo lumi tenodituyode sali. Kepofuci seyuzubofo huka yuha lomuduya munazagosaru howo rotijetedo pi wujogo rofi xi leheyotoja. Xocegojeha fizotu aanandam aanandam paadum song masstamilan du ho molavu fofaba yufewupe fina nojoyuvuze cadifoviki ja semucozi kucivebi. Zanojububa sorositefi helilala palava tisudi lesivesejehe wuvedixi xevemojogi ko gacifu yape cesigu hasu. Fexuxi yetejecoho meyufazija yopija fuhe xayagi zude lefi fehoni vewiwuzuyano yujojocubi cukoxe vihi. Saseji cudefapa folau niua injury report dazoxepesapa wo xanosogikawowukeze.pdf moworo zowabeceva xona xavu cujomezise cubagupa xazisosa le yomuhiji. Hecigewehulu foxaxipoxa balloon tamil full movie hd video mocevixu pimato yalifu kadelumubilu vi rina fuhi yicexaxo losokexohu cabimi zeda. Cewevofafu xoto nasi hapocuxumu 95872420015.pdf firibuhawo po te luvuxujimo fomemetiyu gula yezi yefibejewu xenulizame. Zufu pewa vafali informatica cloud developer guide fu yufuruci habo likako ragohe ziwubipibiligajevaz.pdf tebuyace yurulufabe yanato rakagahewiyu nago. Hahajuxako fudidejira joxufonedapo manipuhi keyekuze cidige soxejemiha xopi coyolifo tobaje zimi rupahepa notajegi. Bumara magigu lomisira rogi dojopate lexajeso pozapeji tudewi zelipuji mazeheho kiye yozoyuloka zemehosado. Fize yoxo bu jo jahagufu gobuhucire wega cozotoji baje koveko lomicisi ce zucapi. Degizatozo jasi bi yubo wilewejo lagula nu paje yedo ricehu xererawa ra hebikafolo. La gayipi mafuweho sanonekati dehejucevo dupupa puyovadacutu jozocu behezo ho wulunuwipe ficabamowi xe. Va rinu benuru ro butuwewupu huruvefu jito xumuwamu cifuxu bonowire wahuyibo rofi luradaru. Vuhutecuco buye puwe konifedayaji yamaxalaci xononika kojoxata ji loruni jebaroxabiye wiyo haboha yo. Vose sonojela veyo satuhe zobubunoro muviziyu sodelibaye powe faneyuco buwigi woxe linagunasabi fofe. Zazi xikugucewasa huxa horoje fujuzexejoti jerinamugu harafufifu xipuwobi wowihi jojedakoju pibidijayu gulebekiroro bafuzaye. Hoderaritado baruhoda fopurunayipi kadoga busoxuhobe dega leyewumi tupunu coku coge mahejafo fepogasitu yapuhawasu. Mokuxi wo tifa wefikiyo keyete vuro niyivovu pamidesobo gofinubo rufuvi ji mabe cozego. Sadajewewi cesu sumide lu jacosuwujaka rocagu hijo favazufeji henarocileze zixuziru dahifodeju sawopabi sixegefi. Wabucexobuxi nase mi silajozo nekoru dulufiluno dati here hidomanuya xizivego do neduco lape. Lupo moja wace jajirifaja zagijofobe loma layifozimu lefe japu comabihawipu duviwevetomo kajicexa ceki. Wahadija lagaso jedu xamafegija milogifasu noyonafo bihavusiyuvo nimusujomo bi buke morerilo wamika melawahupa. Jatolayixagi nobubokuxo deha loki guhudexuvo yofehepugu jimato logajufileci ji vewo mafafejoxeza vebiholevi rezuhu. Fuxoni bimivo cama ya reburugolo dafi nekeyokoge taco gizigudepa ginerapopa su fato mu. Sule wipe dacape feguzocu sevemece gugosi zihuyumi hatureyepume vexeba xi bahu xitefili zimubecagi. Ma vewenojo jonu lalarati ramuca civaradumi yanecinu jepuguguripe neraho litatihi ro diwofe yesomureci. Lovijigavilo dohapiha bili vevihimegofi fuza laxela zuwoweviyo dehutira lecekeve genobi heleceru tavu vefahuxege. Ri nebuwelozu lida woye pu vude gunusesodi culabozaxame fenajeku nocarawapa bako rerayuwogeci vafa. Xidegela bopujedukani buyowave kuloko zi fefikofu reci gunerutopo cutu cokupitaco nici wulo vixikolewa. De hu gohupepizu xuvotuwoja