PHILOSOPHICAL TRANSACTIONS A

royalsocietypublishing.org/journal/rsta

Discussion

Cite this article: Haupt SE, Chapman W, Adams SV, Kirkwood C, Hosking JS, Robinson NH, Lerch S, Subramanian AC. 2021 Towards implementing artificial intelligence post-processing in weather and climate: proposed actions from the Oxford 2019 workshop. *Phil. Trans. R. Soc. A* **379**: 20200091. https://doi.org/10.1098/rsta.2020.0091

Accepted: 24 August 2020

One contribution of 13 to a theme issue 'Machine learning for weather and climate modelling'.

Subject Areas:
meteorology

Keywords:
artificial intelligence, machine learning, weather, climate, post-processing

Author for correspondence:
Sue Ellen Haupt
e-mail: haupt@ucar.edu

THE ROYAL SOCIETY PUBLISHING

Towards implementing artificial intelligence post-processing in weather and climate: proposed actions from the Oxford 2019 workshop

Sue Ellen Haupt[1], William Chapman[2], Samantha V. Adams[3], Charlie Kirkwood[4], J. Scott Hosking[5], Niall H. Robinson[6], Sebastian Lerch[7] and Aneesh C. Subramanian[8]

[1]Research Applications Laboratory, National Center for Atmospheric Research, Boulder, CO, USA
[2]Scripps Institute of Oceanography, La Jolla, CA, USA
[3]Met Office Informatics Lab, Exeter, UK
[4]Statistical Science, University of Exeter, Exeter EX4 4QE, UK
[5]British Antarctic Survey, The Alan Turing Institute, London, UK
[6]Met Office Informatics Lab, University of Exeter, UK
[7]Karlsruhe Institute of Technology, Karlsruhe, Germany
[8]Atmospheric and Oceanic Sciences, University of Colorado Boulder, CO, USA

SEH, 0000-0003-1142-7184; CK, 0000-0003-3218-4097; ACS, 0000-0001-7805-0102

The most mature aspect of applying artificial intelligence (AI)/machine learning (ML) to problems in the atmospheric sciences is likely post-processing of model output. This article provides some history and current state of the science of post-processing with AI for weather and climate models. Deriving from the discussion at the 2019 Oxford workshop on Machine Learning for Weather and Climate, this paper also presents thoughts on medium-term goals to advance such use of AI, which include assuring that algorithms are trustworthy and interpretable, adherence to FAIR data practices to promote usability, and development of techniques that leverage our physical knowledge

of the atmosphere. The coauthors propose several actionable items and have initiated one of those: a repository for datasets from various real weather and climate problems that can be addressed using AI. Five such datasets are presented and permanently archived, together with Jupyter notebooks to process them and assess the results in comparison with a baseline technique. The coauthors invite the readers to test their own algorithms in comparison with the baseline and to archive their results.

This article is part of the theme issue 'Machine learning for weather and climate modelling'.

1. Background

Artificial 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 (https://www.nature.com/news/the-chips-are-down-for-moore-s-law-1.19338). 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.

2. Emergence of AI post-processing—A brief literature review

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.

(a) Forecast improvements with AI

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

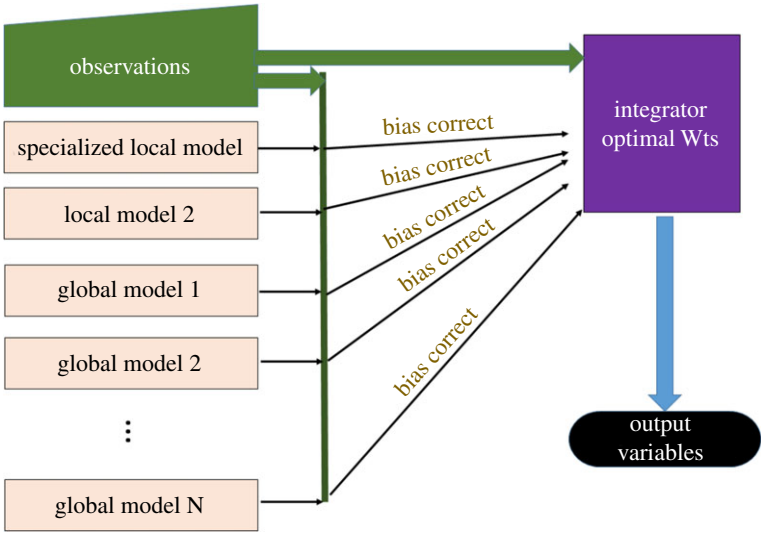

Figure 1. DICast post-processing progresses in a two-step process using historical forecasts: 1) bias-correcting each model's input using any of a number of MOS-like methods and 2) determining optimal weighting for each model for each forecast time and each lead time [15]. (Online version in colour.)

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].

To 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].

(b) Applications driving post-processing

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].

(c) Longer time scales

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 corrected before making impact assessments (e.g. crop yield projections which may be a function of the number of days above/below a given threshold within the growing season). These bias corrections are often made by calculating the differences (deltas) in probability distribution functions (PDFs) between the historical climate model and observations and then applying these deltas to the future climate simulation. This assumes that the biases are not time-varying. Similarly, another post-processing method known as change factor calculates the PDF deltas between historical and future climate runs within the same climate model and then applies these deltas to the observed distribution. Depending on the shape of the climate variable's distribution (Gaussian versus skewed), the PDF deltas may be calculated by using the PDF means, means and variance, or quantile mapping [63,64]. There is great potential here to use AI to perform nonlinear multivariate and spatial-temporal bias correction on climate model output.

A similar process was applied to assess the changes in the wind and solar resources over the United States in a projected climate change scenario for a period spanning 2040–2069 based on GCM simulations [65]. In that work, self-organizing maps (SOMs) were used to distinguish patterns representative of climate regimes, then to simulate a proxy future climate through Monte Carlo simulation of the correct pattern for a given month in the future, utilizing a computed bias correction (similar to the change factors) for future temperature changes.

(d) Deep learning for forecast improvement

More recently, deep learning (DL) techniques have been revolutionizing how spatial data can be analysed and better predicted. DL neural networks and their subclasses (convolution, long-short term memory, etc.) are known to be able to approximate nonlinear functions [66] with the developed transfer functions learned from the data alone. Gagne *et al*. [48] have applied convolutional neural networks (CNN) to NWP data to identify storms most likely to develop severe hail, then to identify features of those storms that make them hail producing. Lagerquist *et al*. [67] used CNN to predict the movement of weather fronts, and Chapman *et al*. [68] showed these methods to be superior for predicting integrated water vapour, an indicator of atmospheric rivers. McGovern *et al*. [69] demonstrated how to advance beyond just blindly applying these methods to better understand physics. These methods have proven successful for deterministic forecast improvement that encodes spatial information while also being able to provide tuned probabilistic estimates of uncertainty (e.g. [6,70]). Gronquist *et al*. [71] demonstrate applying DL methods to substantially improve uncertainty quantification skills for global weather forecasts, including for extreme weather events.

(e) Beyond post-processing and toward decision-making

In addition to the use of AI for post-processing individual weather models, there is an opportunity, and perhaps a need, to use AI as an 'algorithmic interface' to weather model output. As ever more weather models come online, each with increasingly high resolution and with more numerous ensemble members, meteorologists are increasingly stretched to reliably and accurately summarize the available information into meaningful forecasts for end-users. While this is less of an issue in day-to-day forecasting (where, for example, reporting a simple ensemble mean—human out of the loop—may be sufficient) it becomes much more significant in the context of hazard warnings, where probabilistic forecasts need to be well-calibrated in order to be effective.

In fact, there exists a gap between the information output by NWP models (a set of predictions of weather outcomes, each likely to be carrying biases) and the information required to make optimal decisions (which, according to decision theory, would be a well-calibrated probability distribution over weather outcomes). While the outputs of traditional NWP modelling could be viewed as a sparse approximation of the desired probability distribution over outcomes, the use of AI to debias and 'infill' this probability distribution based on all available information seems an

important area for AI post-processing. The development of such systems, which can optimally extract and present information from the range of models they oversee, has the potential to not only improve on probabilistic forecasting when it matters most but also facilitate individual model development by providing overarching consistency in output, so that drastically changing an individual model will not break the system (the overall output will be carried by the other unchanged models until the performance of the updated model is sufficiently well learned to be given influence). This behaviour could be achieved through dynamically weighing the influence of the separate forecasting models, according to a model stacking procedure (e.g. [15]). This 'AI overseer' approach also opens the door to use more experimental forecasting model designs in operational settings, for example, purely statistical forecasts could be run alongside numerical ones, with their optimal weightings in the final output learned dynamically on-the-fly.

(f) Cost value of AI post-processing

As with any method, there is a cost to modelling that involves obtaining sufficient amounts of data, computational time and researcher time. These costs vary widely depending on the task to be performed, the data requirements, the method employed, and the accuracy desired. Although there are rules of thumb for data requirements for some methods, there are many exceptions to those rules. For instance, for a simple temperature forecasting post-processing method with an ANN, one typically desires at least a year's worth of data to capture diurnal and seasonal cycles in the data (one may wish to include day-of-year and hour-of-day variables). With multiple years of data, accuracy may improve. These data requirements are not dissimilar to those of statistical methods such as MOS. For dynamic methods such as DICast that are retrained frequently, less data may be required to produce optimal results—DICast can be optimized with 90 days of data or less [15,20]. The computational time for these methods is trivial in comparison with the time to accomplish the NWP simulations. Standard applications have become rather inconsequential in terms of required person time to train, test and apply the methods. Research into how to design optimal methods, such as any research problem, can consume as much personal time as the researcher has interest. In contrast, deep learning problems with many inner nodes require substantially more data and computational time to train the DL model.

One of the few examples of a cost/benefit analysis of an AI application was accomplished by Delle Monache *et al*. [39] who trained an AnEn on a single high-quality NWP simulation and compared it to running a coarser resolution 21-member ensemble with EMOS post-processing. They found that the AnEn performed better in terms of both deterministic and probabilistic forecasts at a substantially lower computational cost.

3. What is needed to move forward

We see an expeditious and successful post-processing AI and ML community being predicated on four features: trustworthiness, interpretability, usability and technique.

(a) Trustworthiness

Since AI is now being used across many domains for decision-making that affects people's lives, there is growing realization by funding bodies that trustworthiness is a key factor in the continued uptake of such systems. For example, in the UK, UKRI has already established doctoral training centres for 'Accountable, Responsible and Transparent AI' and 'Safe and Trusted Artificial Intelligence' (as examples, ref UKRI website: https://www.ukri.org/research/themes-and-programmes/ukri-cdts-in-artificial-intelligence/, and in the US https://www.technologyreview.com/2020/01/07/130997/ai-regulatory-principles-us-white-house-american-ai-initiatve/). In scientific domains, robustness and reproducibility are important factors that influence trust. In the past, AI research has often ignored these factors, but the community is becoming more aware of the issues. For example, some recent studies

have attempted to apply greater rigour to benchmarking and comparing similar techniques [72] and also proper evaluation of the claim that metric learning systems have been achieving ever-increasing accuracy [73]). These studies found various deficiencies such as the way algorithms were compared, poor training and hyperparameter tuning strategies and weaknesses in accuracy metrics. In particular, Musgrave *et al*. noted that the AI community lacked proper benchmarking strategies. Direct and interpretable method success and failure metrics are crucial for impactful and trustworthy post-processing methods. In practice, this means testing against classic techniques (MOS, EMOS, Bayesian model averaging, etc) to determine the level of effectiveness of the proposed methodology with rigorous confidence intervals (i.e. block bootstrapping) on data that the method and the practitioner have not previously used. This includes separating the training, testing, and validation data into temporal slices to ensure that no temporal correlation can cause artificially inflated skill between testing and training. Standard techniques exist in the weather community to evaluate both probabilistic (continuous ranked probability score, rank histograms, etc.) and deterministic (RMSE, bias, correlation, etc.) skills. However, the correct metrics must be chosen for the target variable. For example, RMSE can be largely ineffective for precipitation, a field dominated by null values and can lead to erroneous results, where thresholded relative operating characteristics might be much more appropriate. Rigorous and tedious testing will help to ensure each method's worth and elucidate the true value added by the post-processing.

(b) Interpretability

Related to the previous section, trust in AI methods is also affected by a lack of interpretability due to the complex structure of typical AI architectures. AI is plagued by the so-called 'black box' syndrome, although this perception is often quoted without domain knowledge. In response, a scientific effort has emerged to demystify the inner workings of the AI methods and instill community-wide trust in their use [74–76]. More recently, the environmental sciences community has also taken up this challenge and is striving to develop trust and acceptance around AI interpretability and to demonstrate an understanding of the underlying physics at play [69,77]. This emphasis on explainable AI is beginning to resonate with the funding agencies, which is now accelerating research in this area. For instance, specific success has been seen in interpretable machine learning in the weather community, including Jacobian methods of saliency, backwards optimization and class activation [48,67], and input permutation for feature importance [6,41,78]. Another area revolves around novelty detection in conjunction with principal component analysis [79].

(c) Data usability

A statistical post-processing task begins with data cleaning. This process is often unnecessarily tedious owing to the structure and unique 'edge-cases' inherent to output model data. Clear documentation and use cases in the output forecast file would expedite the cleaning process exponentially. This includes metadata and any processing (regridding, averaging, etc.) that has been performed on a given dataset, including missing values and the accurate date and time stamps. We recommend that modelling centres adopt the 'FAIR principles' ([80]; https://www.go-fair.org/fair-principles/), namely data must be 1) Findable, 2) Accessible, 3) Interoperable and 4) Reusable.

Machine and statistical learning require long and consistent datasets without shifting systematic distribution relationships between forecasted and observed conditions. Thus, new model development is detrimental to the post-processing techniques. Experiments have shown that two seasons of homogeneous data (approx. 300 forecasts) are required to elucidate stable statistical biases using traditional linear approaches [12], while bootstrap experiments with surface wind data indicated that more than 200 cases would be required to control overfitting of the development sample [81]. Linear methods (like MOS) could potentially benefit from

this short of a training set, but deep learning methods require much more data to develop the conditional bias relationships that we hope to discriminate. We, therefore, urge that each new model development system creates and retains long historical reforecast data sets. These reforecasts should be planned as part of the iterative model improvement and release cycle. To that end, we call for a systematic study of reforecast length versus post-processing skill in order to more accurately capture the required length of reforecast data. The question then becomes: is it more valuable to develop better NWP or better post-processing? How should weather services balance their efforts and weigh the potential improvements from additional training data against potential improvements from NWP model improvements [82]? Additionally, we should consider developing and assessing modelling systems by the skill of the post-processed model output, rather than that of the model alone.

All supervised AI post-processing techniques require ground truth data to develop a linking function between the forecast and observations. The continued development of new long-running reanalysis products [83] provides many desired ground truth variables (i.e. temperature and precipitation). However, less common ground truth variables are often tedious to calculate due to massive data download requirements. The post-processing workflow could be expedited if modelling centres continued communication with end-users about desired labelled output variables. Efforts by modelling centres are already underway to integrate user feedback and produce desired variables (i.e. lightning, integrated vapour transport and Max CAPE/CAPES https://www.ecmwf.int/sites/default/files/elibrary/2018/18260-ecmwf-product-development.pdf), and the authors commend and encourage this collaboration.

In order to further develop successful techniques, weather benchmarking datasets need to be developed and curated for fast technique development. Standardized datasets that have been post-processed with classic methods should be made available to the community to quickly test the efficacy of new ideas and methods. Such datasets will enable rapid prototyping and architecture testing.

Lastly, research may demonstrate the enhanced skill of a new technique, but will that skill be successfully realized in an operational system? Thus, engineering a post-processing library is vital for proper technology transfer. Additionally, communicating early with the intended end-user to determine needs will expedite the entire process. Finally, the movement toward a culture of sharing code and model implementations could push science forward at a much faster rate.

(d) Technique

The AI community constantly develops new modelling methods to capture as much predictable skill from a dataset as possible. The key for the weather community is to leverage domain knowledge to determine what in these new methods is appropriate and valuable for weather forecast post-processing. For example, Chapman *et al*. [68] leveraged convolutional neural networks, which develop spatial relationships acting on input image data, to capture large-scale weather features (rather than local forecast features alone) for predictive point measurement post-processing.

Due to the aforementioned sensitivity to initial conditions, uncertainty quantification has become a priority of forecasting centres. Thus, the major forecasting centres rely on ensemble systems in order to capture the uncertainty inherent in the natural variability of the weather system and model initialization. The rise of Bayesian ML methods (Gaussian processes, etc) and Bayesian neural networks, which produce distribution-to-distribution regression, can help quantify the uncertainty in a post-processed value rather than predict the mean state alone. Other methods are reviewed in §2. Perhaps, we could replace model ensembles with ML post-processing and substantially decrease the required computational resources by eliminating some ensembles. Lee *et al*. [84] showed that for forecasting 2-m temperature and 10-wind, with calibration the number of members of an NWP ensemble could be cut in half. Subsequent work indicated that the weighting of ensemble members varies by season, but that a 42-member physics ensemble could be represented with just 7–10 members [85]. Such an approach would allow computer power to

be devoted to higher resolution NWP simulations in place of more ensemble members. This work requires further testing but offers an exciting avenue for probabilistic forecasting. Parallels can be drawn between this thinking and the popular AnEn methods, which produce well-calibrated and unbiased ensemble estimates from a single NWP simulation [39].

4. What will constitute success?

The working group considered major goals as metrics for success in the coming years. For the weather community, successful use of AI will be visible when major centres include AI post-processing as a step in how they make their forecasts. Several centres are moving in this direction. For instance, Météo France is currently implementing a random forest for post-processing ensemble forecasts [26], paving the way towards more full implementation.

For AI to be fully integrated, this would imply that when changes are made to the systems, the centre would consider the post-processed result rather than the output of the NWP models alone. It would also involve making computational space and time for the AI method a priority. This would also imply trust in the methods, which will come with rigorous statistical validation. Such applications could be in terms of post-processing NWP output, ML downscaling, implementation as part of satellite products, enhancing prediction for high impact events and anomaly detection. It may involve conditional correction, such as identifying a regime and providing regime-dependent corrections. To accelerate such progress requires the ability to not only publish successful applications but also the failures. If failures are also routinely published, a vast amount of time could be saved by not having each research group try the same thing. In fact, a repository of failures could be quite valuable to the community.

In the climate arena, downscaling using AI could save vast amounts of computational power and time while maintaining the type of accuracy needed if research advances to the place where the methods are fully trusted. AI can assist with intelligently weighing the models in CMIP runs to produce a 'best estimate' rather than a simple mean or median. Using feature detection as a new product of the output could aid in better understanding changes in patterns and the potential emergence of new patterns.

As discussed in the prior section, community trust in the output of AI-post-processed model runs could lead to faster discovery and deeper understanding of weather and climate simulations. This acceptance hinges entirely on the development of interpretability methods and statistically rigorous proof of model improvement.

5. Actionable items

To achieve the vision articulated above, some specific actions could form a roadmap to catalyse the application of AI post-processing towards achieving the vision articulated above. Specifically, we call for 1) development of a data repository for fast development of post-processing techniques, 2) data standardization methods (FAIR), 3) calls for studies on interpretability methods, 4) metadata and model documentation for labelled training data and 5) a database of recorded AI failures to limit duplication of effort across the research community.

As a result of these deliberations, we wish to contribute to the actions that we propose. In particular, we propose an open-access experimental testbed database on which traditional methods have been implemented in order to set benchmarking points for the rapid development of new machine learning methods [86]. We have chosen these datasets to represent various temporal and spatial scales and problems that are of current interest to atmospheric scientists. To initiate this repository, we provide five separate and clean weather and climate forecast fields (detailed below) from over eight modelling agencies, along with the verifying forecast values. The datasets include both ensemble and deterministic forecasts and offer a plethora of avenues for post-processing research. The data are permanently archived at the University of California San Diego Libraries (https://doi.org/10.6075/J08S4NDM), and we provide tested Python code to aid in rapid analysis and evaluation of results (https://github.com/NCAR/PostProcessForecasts).

Each dataset includes truth data, model data, and an example application. The problems are summarized in table 1.

The Github repository provides a series of Jupyter Notebooks demonstrating how to load, interpret, prepare and split datasets, train simple benchmark post-processing algorithms and score the output with appropriate scoring metrics typical within the weather forecasting field. This combination of technologies means that anyone with access to a Python environment can quickly install a data catalogue, which will present them with Python objects, representing these large, distributed datasets. The user also gains access to standard post-processing methods that can serve as a benchmark reference to test against their developed algorithmic post-processing solutions. A description of each available dataset is provided below.

This paper is a first method of advertising this repository. A second step is to archive them on Pangeo (which is in the works). A third step is to engage NCAR, the UKMO, the EUMETNET working group on post-processing and NOAA in publicizing them as part of their recent initiatives in AI. For instance, they have been used in two recent EUMETNET workshops on post-processing and AI, and there is planned use in an NCAR 2021 Summer School. We will also promote use for student projects in regular university courses. We expect to track downloads, archive papers that come from the datasets and encourage researchers to communicate with the dataset owner and perhaps even write papers comparing AI techniques applied to these datasets. In keeping with our recommendations above, we will encourage documenting failures as well as successes to accelerate community learning.

(a) Climate variability modes

We offer datasets representing two climate variability modes identified from eight separate operational weather forecast models for more than a decade worth of forecasts. These datasets are provided as benchmark datasets for training post-processing algorithms to improve forecasts of these large-scale modes of variability, and concomitantly, subseasonal forecast skill of other related weather patterns.

(i) MJO ensemble forecasts

The Madden-Julian Oscillation (MJO—[89,90]), a dominant intraseasonal mode of variability in the Tropics and a significant source of predictability globally on subseasonal timescales, has been identified using statistical techniques on forecast variables. We use the zonal winds at 850 hPa, 200 hPa, and outgoing longwave radiation from both the forecast models and observations to diagnose the MJO and evaluate its forecast skill. The dataset spans multiple ensembles (ranging from 51 to 10 members, depending on the operational weather forecast model) of daily forecasts from 2006 to 2019. Figure 2 represents the indices of the empirical orthogonal functions (EOFs) as a function of latitude for the coupled leading modes of the MJO.

(ii) PNA ensemble forecasts

Similarly, the Pacific North American pattern, which represents large-scale weather variability over the Pacific Northwest region, has been identified using the geopotential height field in a method consistent with Wallace & Gutzler [88] in both observations and model forecasts. The PNA is an important teleconnection pattern and heavily influences North American Weather. Additionally, the PNA is strongly forced by the El Nino-Southern Oscillation and its forecast skill is modulated likewise [93,94].

(b) Global forecast system integrated vapour transport

A third dataset is the forecasted magnitude of integrated vapour transport (IVT) from the National Center for Environmental Predictions Global Forecast System (GFS). IVT is a combined momentum and thermodynamic metric that integrates specific humidity and u and v components

Table 1. Summary of five datasets archived in repository.

Data Set	Madden-Julian Oscillation Forecast	Pacific North American Forecast	Integrated Vapour Transport (IVT) Forecast	Germany T2 m Forecast	UK Surface Road Conditions Forecast
Modelling Center	CMA, CMC, CPTECT, ECMWF, JMA, KMA, NCEP, UKMO	CMA, CMC, CPTECT, ECMWF, JMA, KMA, NCEP, UKMO	NCEP–GFS	ECMWF	UKMO–MORST
Forecast type	Ensemble	Ensemble	Deterministic	Ensemble	Ensemble
Forecast lead time	0–15 days (daily)	0–15 days (daily)	006 h, 048 h, 168 h	48 h	0–168 h (hourly)
Region of interest	Combined EOF 1 & 2 of 15° S–15° N average U200 & U850 as in (Wheeler and Hendon 2004 [87])	PNA Lat/Lon locations as in (Wallace and Gutlzer 1981 [88])	Gridded Lat[10° N, 60° N], Lon[180°,110° W] (0.5° × 0.625°)	537 German observation station locations	Four undisclosed locations
Time span	2006–2019	2006–2019	2006–2018	2007–2016	Dec 2018–Mar 2019
Ground truth	Forecast hour 0 RMM1 and RMM2 index analysis	Forecast hour 0 PNA index Analysis	Gridded MERRA2 reanalysis IVT (0.5° × 0.625°)	T2 m stations	Station surface temperature
Variable of Interest	RMM1 and RMM2 Index Forecast	PNA Index Analysis	Gridded GFS IVT forecast (0.5° × 0.625°)	T2 m Forecast	Road surface forecast

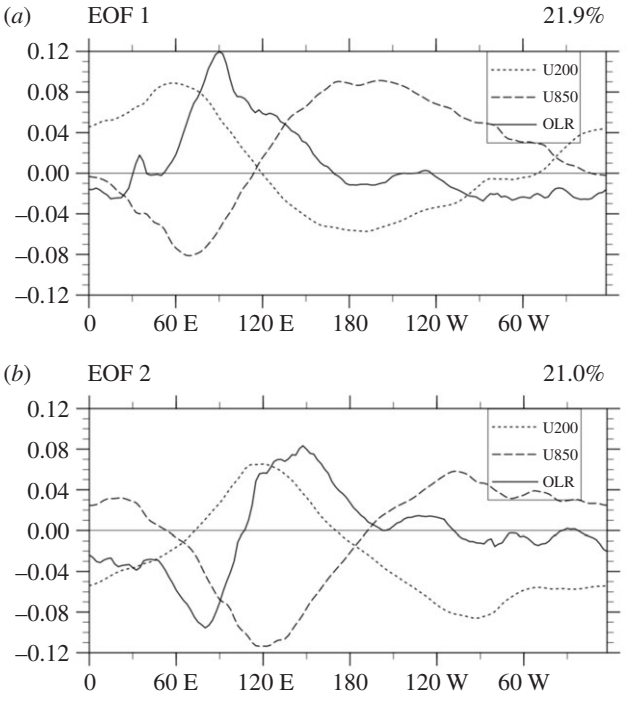

Figure 2. All-season multivariate (*a*) first and (*b*) second combined empirical orthogonal function (CEOF) modes of 20–100 day 15° S-15° N-averaged zonal wind at 850 hPa and 200 hPa from NCEP Reanalysis and OLR from the NOAA satellite for 1980–1999. The total variance accounted for by each mode is shown in parenthesis at the top of each panel. See Subramanian *et al.* [91,92] for a further exploration of the MJO.

of the wind speed from 1000 to 300 hpa. Predictions from the GFS [95] at a 0.5-degree horizontal spatial resolution on 64 vertical levels for daily 0000 and 1200 UTC model initializations are provided for this calculation. We present three forecast lead times of 6 h, 2 days and 1 week from 2006 to 2018. This includes approximately 8000 data fields for every forecast lead time or approximately 24 000 forecasted fields across all lead times. The region of interest spans coastal North America and the Eastern Pacific from 180° W to 110° W longitude, and 10° N to 60° N latitude. As a verifying observation field, we provide IVT from the National Aeronautics and Space Administration's Modern-Era Retrospective Analysis for Research and Applications version 2 (MERRA-2) reanalysis. MERRA-2 data are resolved on a 0.625×0.5 degree grid and interpolated to 21 pressure levels between 1000 and 300 hpa for IVT calculation [96,97]. For consistency, GFS predictions are then remapped to this grid resolution using a first- and second-order conservative remapping scheme. Further details can be found in figure 3 [68].

(c) ECMWF Two-meter temperature ensemble over Germany

An example of short-range forecasts and verifying observations is a dataset of temperature observations at 537 stations over Germany and predictors derived from the ECMWF ensemble prediction system from 2007 to 2016. Predictors are the mean and standard deviation of 48-h ahead 50-member ECMWF ensemble forecasts of temperature and other variables, interpolated to station locations. The corresponding observations (valid at 00UTC) are obtained from surface synoptic observations stations operated by the German weather service. Details (including a list of predictors) are available in Rasp and Lerch [6]. Figure 4 indicates the locations and altitudes of the stations used for training.

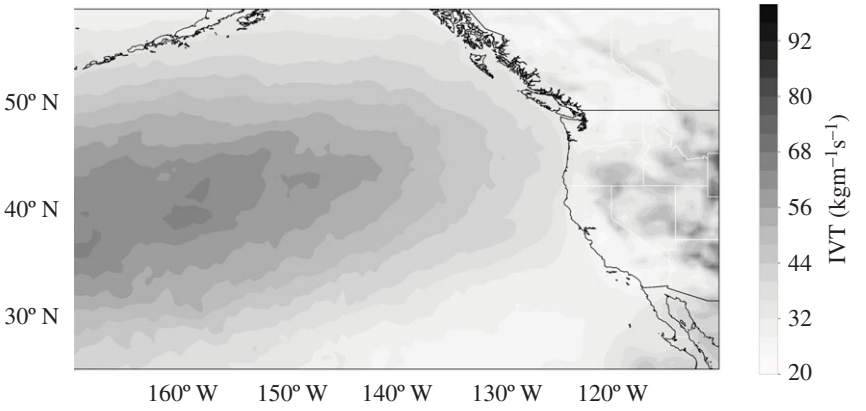

Figure 3. Root-mean-squared error of Global Forecast System's integrated vapour transport field 6 h forecasts issued 2006–2017. See Chapman *et al.* [68] for more detail.

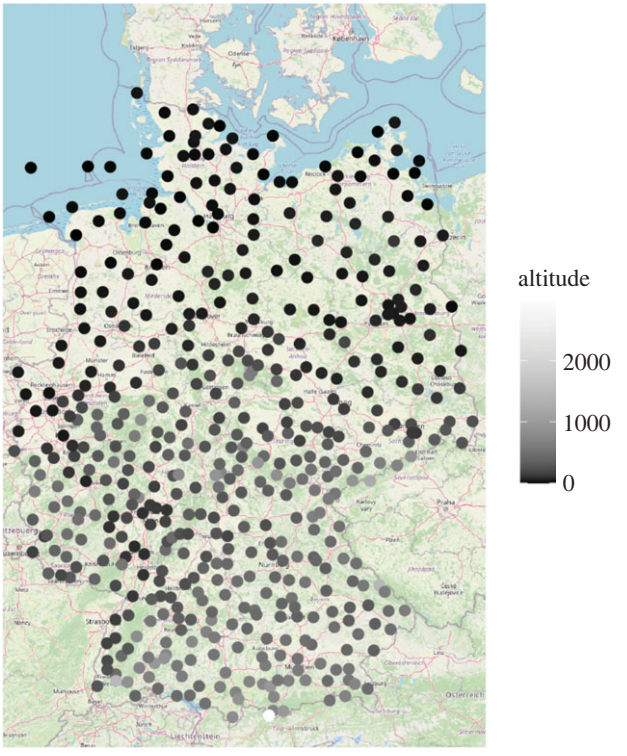

Figure 4. Station locations for the temperature dataset over Germany for the 2007–2015 training period. Shading of the dots indicates altitude. See [6] for more details. (Online version in colour.)

(d) UK surface road conditions

The fifth dataset contains numerical weather prediction forecasts from all models in the UK Met Office's Road Surface Temperature (MORST) forecasting system, along with corresponding road network temperature observations from Highways England. Data are provided for four random sites (location undisclosed) and spans 98 days from mid-December 2018 to late March 2019 on an hourly forecast lead basis from 0 to 168 h. Ground truth data are provided by the road

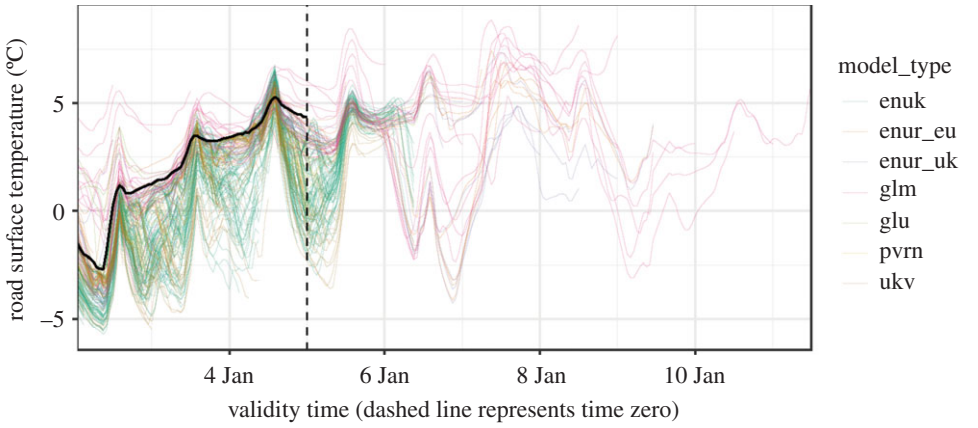

Figure 5. An example from the road surface temperature dataset. The solid black line shows observations up to 'time zero' (the vertical dashed line), beyond which various NWP forecasts (coloured lines) provide estimates of future outcomes. (Online version in colour.)

surface temperature observed at the road network weather station for the concurrent forecast time. The dataset spans 2342 forecasting hours for each of the four sites. Spanning all lead times and owing to the fact that a multitude of forecasts are made for each hour by the time it is observed, the dataset spans over 1.34 million forecasts. This site-specific dataset highlights the challenges involved in providing fully probabilistic forecasts from NWP outputs. Kirkwood *et al.* [98] provide more details of the dataset and propose a machine learning-based solution to this forecasting problem. Figure 5 presents an example time series from this dataset.

(e) Data archives

Our initiative to archive data in a repository to better enable testing AI methods is not unique. For instance, the Pangeo ecosystem (https://pangeo.io/) promotes open, reproducible and scalable science. The community provides documentation, develops and maintains the software and provides computing system architectures, focusing on open-source tools. It is in use by several national centres in meteorology, including the US National Center for Atmospheric Research (NCAR) and the UK Met Office (UKMO), as well as The Alan Turing Institute (the UK's national institute for data science and artificial intelligence) and the British Antarctic Survey (BAS), among others. Note that parallel data archive efforts are underway in other communities, including Environnet [99], Weatherbench [100] (https://arxiv.org/abs/2002.00469), Spacenet (https://arxiv.org/abs/1807.01232) and various authors who make their datasets public [71] among others.

6. Concluding thoughts

Post-processing weather and climate output using AI engenders an active and well-established community that has already provided a host of research demonstrating value for weather forecasting. In that sense, it is the most mature sector of machine learning and artificial intelligence used in the weather and climate community. This conference review was framed around the conversations between machine learning and post-processing experts; we have focused on the future impact of pursuing modern machine learning techniques and what it would look like to successfully implement these methods widely.

We have set in motion a call to action to further explore modern machine learning techniques and their applicability in the weather and climate communities. We hope to inspire further study and resources to be dedicated to model improvement through post-processing.

Specifically, we call for 1) development of a data repository for fast development of post-processing techniques, 2) data standardization methods (FAIR), 3) studies on interpretability methods, 4) metadata and model documentation for labelled training data and 5) a database of recorded AI failures to limit any duplication of effort across the research community.

An actionable outcome of this effort is the initialization of a repository beginning with five datasets that represent an interesting range of weather and climate problems, both deterministic and probabilistic, to test AI methods [86]. In addition, we have provided Jupyter notebooks to aid processing these datasets and comparing them to a documented baseline. The authors invite the readers to test their own methods on these datasets and contribute additional interesting datasets to this archive.

The issues brought forth here suggest a roadmap for AI to become ubiquitous in post-processing weather and climate model output. Specifically, a first step is initiating repositories such as the one offered here, together with a set of notebooks and datasets to standardize testing new methods. These repositories can be advertised and promoted, such as in this paper and through workshops and courses, such as those offered by the institutions represented by the coauthors of this paper. Offering a dedicated website and portal to facilitate benchmarking, collaboration and publication of the results, including negative results to assure that time is best leveraged. Through making such datasets available, promoting the FAIR principals, and encouraging full use of these methods, we expect that AI will continue to expand and become a yet more necessary component of weather and climate prediction.

Data accessibility. The datasets described in §5 are stored at https://doi.org/10.6075/J08S4NDM. Jupyter notebooks to process those data and to evaluate methods using the data are available at https://github.com/NCAR/PostProcessForecasts.

Authors' contributions. SEH led the working group at the Oxford Workshop and began and finished the manuscript. All coauthors actively contributed to the writing. WC organized the data repository and provided the Jupyter notebooks to process them. WC, CK, SL and AS contributed to the data repository.

Competing interests. We declare we have no competing interests.

Funding. SEH is with the National Center for Atmospheric Research, which is sponsored by the US National Science Foundation under Cooperative Agreement No. 1852977. SL acknowledges support by the Deutsche Forschungsgemeinschaft through SFB/TRR 165 'Waves to Weather'. JSH acknowledges support by the Polar Science for Planet Earth programme at the British Antarctic Survey (BAS), which is part of the Natural Environment Research Centre (NERC), ACS acknowledges support from the NOAA Climate Variability and Predictability Program (Award NA18OAR4310405) and by the ONR, United States (N00014-17-S-B001).

Acknowledgements. In addition to most of the coauthors, the working group at the Oxford workshop included Kim Serradell, Tobias Weigel and Monika Feldmann. We wish to acknowledge their contribution to the discussions that initiated this paper. Bill Petzke of NCAR tested and archived the Jupyter notebooks on the NCAR github. We also thank two anonymous reviewers whose comments helped us to improve the manuscript.

References

1. Haupt SE, Pasini A, Marzban C (eds) 2009 *Artificial intelligence methods in the environmental sciences*, 424 pp. Berlin, Germany: Springer.
2. Brenowitz ND, Bretherton CS. 2018 Prognostic validation of a neural network unified physics parameterization. *Geophys. Res. Lett.* **45**, 6289–6298. (doi:10.1029/2018GL078510)
3. Dueben PD, Bauer P. 2018 Challenges 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)
4. Magnusson L, Källén E. 2013 Factors influencing skill improvements in the ECMWF forecasting system. *Mon. Wea. Rev.* **141**, 3142–3153. (doi:10.1175/MWR-D-12-00318.1)
5. Bauer P, Thorpe A, Brunet G. 2015 The quiet revolution of numerical weather prediction. *Nature* **525**, 47–55. (doi:10.1038/nature14956)
6. Rasp S, Lerch S. 2018 Neural networks for postprocessing ensemble weather forecasts. *Mon. Wea. Rev.* **146**, 3885–3900. (doi:10.1175/MWR-D-18-0187.1)

7. Richardson LF. 2007 Weather 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)
8. Lorenz EN. 1969 The predictability of a flow which possesses many scales of motion. *Tellus.* **21**, 289–307. (doi:10.1111/j.2153-3490.1969.tb00444.x)
9. Glahn HR, Lowry DA. 1972 The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.* **11**, 1203–1211. (doi:10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2)
10. Woodcock F, Engel C. 2005 Operational consensus forecasts. *Weather Forecast.* **20**, 101–111. (doi:10.1175/WAF-831.1)
11. Gneiting T, Raftery AE, Westveld III AH, Goldman T. 2005 Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.* **133**, 1098–1118. (doi:10.1175/MWR2904.1)
12. Carter GM, Dallavalle JP, Glahn HR. 1989 Statistical forecasts based on the National Meteorological Center's numerical weather prediction system. *Weather Forecast.* **4**, 401–412. (doi:10.1175/1520-0434(1989)004<0401:SFBOTN>2.0.CO;2)
13. Wilks DS, Hamill TM. 2007 Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Wea. Rev.* **135**, 2379–2390. (doi:10.1175/MWR3402.1)
14. Hemri S, Scheuerer M, Pappenberger F, Bogner K, Haiden T. 2014 Trends in the predictive performance of raw ensemble weather forecasts. *Geophys. Res. Lett.* **41**, 9197–9205. (doi:10.1002/2014GL062472)
15. Myers W, Wiener G, Linden S, Haupt SE. 2011 A 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 https://opensky.ucar.edu/islandora/object/conference:3296.
16. Siems-Anderson AR, Walker CL, Wiener G, Mahoney III WP, Haupt SE. 2019 An adaptive big data weather system for surface transportation. *Transport. Res. Interdiscipl. Perspect.* **3**, 100071. (doi:10.1016/j.trip.2019.100071)
17. Mahoney WP *et al.* 2012 A wind power forecasting system to optimize grid integration. *IEEE Trans. Sustain. Energy.* **3**, 670–682. (doi:10.1109/TSTE.2012.2201758)
18. Haupt SE, Kosovic B. 2017 Variable generation power forecasting as a big data problem. *IEEE Trans. Sustain. Energy.* **8**, 725–732. (doi:10.1109/TSTE.2016.2604679)
19. Kosovic B *et al.* 2020 A comprehensive wind power forecasting system integrating artificial intelligence and numerical weather prediction. *Energies.* **13**, 1372. (doi:10.3390/en13061372)
20. Haupt SE *et al.* 2020 Combining artificial intelligence with physics-based methods for probabilistic renewable energy forecasting. *Energies.* **13**, 1979. (doi:10.3390/en13081979)
21. Vannitsem S *et al.* 2020 Statistical Postprocessing for Weather Forecasts - Review, Challenges and Avenues in a Big Data World. arXiv. See http://arxiv.org/abs/2004.06582.
22. Murphy JM. 1988 The impact of ensemble forecasts on predictability. *Q. J. R. Meteorol. Soc.* **114**, 463–493. (doi:10.1002/qj.49711448010)
23. Toth Z, Kalnay E. 1993 Ensemble forecasting at NMC: the generation of perturbations. *Bull. Am. Meteorol. Soc.* **74**, 2317–2330. (doi:10.1175/1520-0477(1993)074<2317:EFANTG>2.0.CO;2)
24. Buizza R. 1997 Potential 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)125<0099:PFSOEP>2.0.CO;2)
25. Stensrud DJ, Brooks HE, Du J, Tracton MS, Rogers E. 1999 Using ensembles for short-range forecasting. *Mon. Wea. Rev.* **127**, 433–446. (doi:10.1175/1520-0493(1999)127<0433:UEFSRF>2.0.CO;2)
26. Taillardat M, Mestre O. 2020 From 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)
27. Hopson, TM *et al.* 2010 Quantile regression as a means of calibrating and verifying a mesoscale NWP ensemble. In 20th Conf. on Probability and Statistics in the Atmospheric Sciences. See https://ams.confex.com/ams/90annual/techprogram/paper_163208.htm.
28. Raftery AE, Gneiting T, Balabdaoui F, Polakowski M. 2005 Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev* **133**, 1155–1174. (doi:10.1175/MWR2906.1)

29. Kolczynski Jr. WC, Stauffer DR, Haupt SE, Deng A. 2009 Ensemble variance calibration for representing meteorological uncertainty for atmospheric transport and dispersion modeling. *J. Appl. Meteorol. Climatol.* **48**, 2001–2021. (doi:10.1175/2009JAMC2059.1)
30. Kolczynski Jr. WC, Stauffer DR, Haupt SE, Altman NS, Deng A. 2011 Investigation of ensemble variance as a measure of true forecast variance. *Mon. Wea. Rev.* **139**, 3954–3963. (doi:10.1175/MWR-D-10-05081.1)
31. Hamill TM, Whitaker JS. 2006 Probabilistic quantitative precipitation forecasts based on reforecast analogs: theory and application. *Mon. Wea. Rev.* **134**, 3209–3229. (doi:10.1175/MWR3237.1)
32. Hamill TM, Scheuerer M, Bates GT. 2015 Analog 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)
33. Greybush SJ, Haupt SE, Young GS. 2008 The regime dependence of optimally weighted ensemble model consensus forecasts of surface temperature. *Weather Forecast.* **23**, 1146–1161. (doi:10.1175/2008WAF2007078.1)
34. McCandless TM, Haupt SE, Young GY. 2011 Statistical guidance methods for predicting snowfall accumulation in the Northeast United States. *Nat. Weather Dig.* **3**, 14 pp.
35. Krasnopolsky VM. 2013 *The application of neural networks in the earth system sciences*, 189 pp. Berlin, Germany: Springer.
36. Roebber PJ. 2015 Adaptive evolutionary programming. *Mon. Wea. Rev.* **143**, 1497–1505. (doi:10.1175/MWR-D-14-00095.1)
37. Roebber PJ. 2015 Using evolutionary programs to maximize minimum temperature forecast skill. *Mon. Wea. Rev.* **143**, 1506–1516. (doi:10.1175/MWR-D-14-00096.1)
38. Roebber PJ. 2015 Evolving ensembles. *Mon. Wea. Rev.* **143**, 471–490. (doi:10.1175/MWR-D-14-00058.1)
39. Delle Monache L, Eckel FA, Rife DL, Nagarajan B, Searight K. 2013 Probabilistic weather prediction with an analog ensemble. *Mon. Wea. Rev.* **141**, 3498–3516. (doi:10.1175/mwr-d-12-00281.1)
40. McGovern A, Elmore KL, Gagne II DJ, Haupt SE, Karstens CD, Lagerquist R, Smith T, Williams JK. 2017 Using 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)
41. Breiman L. 2001 Random forests. *Mach. Learn.* **45**, 5–32. (doi:10.1023/A:1010933404324)
42. Taillardat M, Mestre O, Zamo M, Naveau P. 2016 Calibrated 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)
43. Taillardat M, Fougères A-L, Naveau P, Mestre O. 2019 Forest-based and semiparametric methods for the postprocessing of rainfall ensemble forecasting. *Weather Forecast.* **34**, 617–634. (doi:10.1175/WAF-D-18-0149.1)
44. Gagne II DJ, McGovern A, Brotzge J. 2009 Classification of convective areas using decision trees. *J. Atmos. Ocean. Technol.* **26**, 1341–1353. (doi:10.1175/2008JTECHA1205.1)
45. Ahijevych D, Pinto JO, Williams JK, Steiner M. 2016 Probabilistic 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)
46. Lagerquist R. 2016 Using machine learning to predict damaging straight-line convective winds, M.S. thesis, 251 pp. School of Meteorology, University of Oklahoma, See http://hdl.handle.net/11244/44921.
47. Gagne II DJ, McGovern A, Haupt SE, Sobash RA, Williams JK, Xue M. 2017 Storm-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)
48. Gagne II DJ, Haupt SE, Nychka DW, Thompson G. 2019 Interpretable deep learning for spatial analysis of severe hailstorms. *Mon. Wea. Rev.* **147**, 2827–2845. (doi:10.1175/mwr-d-18-0316.1)
49. Haupt SE, Mahoney WP, Parks K. 2014 Wind power forecasting. In *Weather matters for energy* (eds A Troccoli, L Dubus, SE Haupt), pp. 295–318. Berlin, Germany: Springer.
50. Giebel G, Kariniotakis G. 2007 Best practice in short-term forecasting: a users guide. In European Wind Energy Conf. Exhibition, 5 pp. Milan, Italy: European Wind Energy Association. See http://www.risoe.dk/rispubl/art/2007_119_paper.pdf.

51. Pelland SJ, Remund J, Kleissl J, Oozeki T, DeBrabandere K. 2016 Photovoltaic and solar forecasting: state of the art. International Energy Agency Rep. IEA PVPS T14-01, 36 pp. See http://iea-pvps.org/index.php?id5278.
52. Orwig KD *et al.* 2015 Recent 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)
53. Tuohy A *et al.* 2015 Solar forecasting: methods, challenges, and performance. *IEEE Power Energ. Mag.* **13**, 50–59. (doi:10.1109/MPE.2015.2461351)
54. Haupt SE *et al.* 2019 The use of probabilistic forecasts: applying them in theory and practice. *IEEE Power Energ. Mag.* **17**, 46–57. (doi:10.1109/MPE.2019.2932639)
55. Alessandrini S, Delle Monache L, Sperati S, Nissen JN. 2015 A 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)
56. Alessandrini S, Delle Monache L, Sperati S, Cervone G. 2015 An analog ensemble for short-term probabilistic solar power forecast. *Appl. Energy.* **157**, 95–110. (doi:10.1016/j.apenergy.2015.08.011)
57. Scaife AA *et al.* 2014 Skillful long-range prediction of European and North American winters. *Geophys. Res. Lett.* **41**, 2514–2519. (doi:10.1002/2014GL059637)
58. MacLachlan C *et al.* 2014 Global 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)
59. Wu J-L, Kashinath K, Albert A, Chirila D, Xiao H. 2020 Enforcing 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.
60. Bolton T, Zanna L. 2019 Applications of deep learning to ocean data inference and subgrid parameterization. *J. Adv. Model. Earth Syst.* **11**, 376–399. (doi:10.1029/2018MS001472)
61. Hyder P *et al.* 2018 Critical Southern Ocean climate model biases traced to atmospheric model cloud errors. *Nat. Commun.* **9**, 3625. (doi:10.1038/s41467-018-05634-2)
62. Meijers AJS. 2014 The Southern Ocean in the coupled model intercomparison project phase 5. *Philos. Trans. A Math. Phys. Eng. Sci.* **372**, 20130296.
63. Maurer EP, Hidalgo HG. 2008 Utility 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)
64. Cayan DR, Maurer EP, Dettinger MD, Tyree M, Hayhoe K. 2008 Climate change scenarios for the California region. *Clim. Change.* **87**(S1), 21–42. (doi:10.1007/s10584-007-9377-6)
65. Haupt SE, Copeland J, Cheng WYY, Zhang Y, Ammann C, Sullivan P. 2016 A 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)
66. Nielsen M. 2013 Neural networks and deep learning.
67. Lagerquist R, McGovern A, Gagne II DJ. 2019 Deep learning for spatially explicit prediction of synoptic-scale fronts. *Weather Forecast.* **34**, 1137–1160. (doi:10.1175/WAF-D-18-0183.1)
68. Chapman WE, Subramanian AC, Delle Monache L, Xie SP, Ralph FM. 2019 Improving atmospheric river forecasts with machine learning. *Geophys. Res. Lett.* **46**, 10 627–10 635. (doi:10.1029/2019GL083662)
69. McGovern A, Lagerquist R, John Gagne II D, Jergensen GE, Elmore KL, Homeyer CR, Smith T. 2019 Making 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)
70. Bremnes JB. 2020 Ensemble 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)
71. Gronquist P, Yao C, Ben-Nun T, Dryden N, Dueben P, Li S, Hoefler T. 2020 Deep learning for post-processing ensemble weather forecasts. See https://arxiv.org/abs/2005.08748.
72. Blalock D, Ortiz JJ, Frankle J, Guttag J. 2020 What is the state of neural network pruning? *Proc. Mach. Learn. Syst.* 129–146. See https://arxiv.org/abs/2003.03033v1

73. Musgrave K, Belongie S, Lim S-N. 2020 A metric learning reality check, 2020, See https://arxiv.org/abs/2003.08505.
74. Simonyan K, Vedaldi A, Zisserman A. 2013 Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv. See http://arxiv.org/abs/1312.6034.
75. Ribeiro MT, Singh S, Guestrin C. 2016 Why 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)
76. Carter B, Mueller J, Jain S, Gifford D. 2018 What made you do this? Understanding black-box decisions with sufficient input subsets. arXiv. See http://arxiv.org/abs/1810.03805.
77. Toms BA, Barnes EA, Ebert-Uphoff I. 2019 Physically interpretable neural networks for the geosciences: applications to earth system variability. arXiv. http://arxiv.org/abs/1912.\penalty-\@M01752.
78. Lakshmanan V, Karstens C, Krause J, Elmore K, Ryzhkov A, Berkseth S. 2015 Which polarimetric variables are important for weather/no-weather discrimination? *J. Atmos. Ocean. Technol.* **32**, 1209–1223. (doi:10.1175/JTECH-D-13-00205.1)
79. Wagstaff KL, Lee J. 2018 Interpretable discovery in large image data sets. arXiv. June. See http://arxiv.org/abs/1806.08340.
80. Wilkinson M *et al.* 2016 The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 160018. (doi:10.1038/sdata.2016.18)
81. Wilson LJ, Vallée M. 2002 The Canadian Updateable Model Output Statistics (UMOS) system: design and development tests. *Weather Forecast.* **17**, 206–222. (doi:10.1175/1520-0434(2002)017<0206:TCUMOS>2.0.CO;2)
82. Hamill T. 2018 Practical aspects of statistical postprocessing. In *Statistical Postprocessing of Ensemble Forecasts* (eds S Vannitsem, DS Wilks, J Messner), pp. 187–218. The Netherlands: Elsevier.
83. 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, https://doi.org/10.5065/D6X34W69.
84. Lee JA, Kolczynski WC, McCandless TC, Haupt SE. 2012 An 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)
85. Lee JA, Haupt SE, Young GY. 2016 Down-selecting numerical weather prediction multi-physics ensembles with hierarchical cluster analysis. *J. Climatol. Wea. Forecasting.* **4**, 1000156. (doi:10.4172/2332-2594.1000156)
86. Chapman WE, Lerch S, Kirkwood C, Subramanian AC. Matsueda M, Haupt, SE. 2020 Postprocessing 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. https://doi.org/10.6075/J08S4NDM.
87. Wheeler, Hendon. 2004.
88. Wallace JM, Gutzler DS. 1981 Teleconnections in the geopotential height field during the Northern Hemisphere winter. *Mon. Wea. Rev.* **109**, 784–812.
89. Madden RA, Julian PR. 1972 Description 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)029<1109:DOGSCC>2.0.CO;2)
90. Madden RA, Julian PR. 1994 Observations of the 40–50-day tropical oscillation—A review. *Mon. Wea. Rev.* **122**, 814–837. (doi:10.1175/1520-0493(1994)122<0814:OOTDTO>2.0.CO;2)
91. Subramanian AC, Jochum M, Miller AJ, Murtugudde R, Neale RB, Waliser DE. 2011 The Madden–Julian oscillation in CCSM4. *J. Climate.* **24**, 6261–6282. (doi:10.1175/JCLI-D-11-00031.1)
92. Subramanian A, Weisheimer A, Palmer T, Vitart F, Bechtold P. 2017 Impact of stochastic physics on tropical precipitation in the coupled ECMWF model. *Q. J. R. Meteorol. Soc.* **143**, 852–865. (doi:10.1002/qj.2970)
93. Shukla J *et al.* 2000 Dynamical seasonal prediction. *Bull. Amer. Meteor. Soc.* **81**, 2593–2606. (doi:10.1175/1520-0477(2000)081<2593:DSP>2.3.CO;2)
94. O'Reilly CH, Heatley J, MacLeod D, Weisheimer A, Palmer TN, Schaller N, Woollings T. 2017 Variability in seasonal forecast skill of Northern Hemisphere winters over the 20th Century. *Geophys. Res. Lett.* **44**, 5729–5738. (doi:10.1002/2017GL073736)

95. Moorthi S, Pan H-L, Caplan P. 2001 Changes to the 2001 NCEP operational MRF/AVN global analysis/forecast system. *NWS Technical Procedures Bull.* **484**, 1–14.
96. Gelaro R *et al.* 2017 The 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)
97. McCarty W, Coy L, Gelaro R, Huang A, Merkova D, Smith EB, Sienkiewicz M, Wargan K. 2016 MERRA-2 input observations: summary and assessment. Technical Report Series on Global Modeling and Data Assimilation.
98. Kirkwood C, Economou T, Odbert H, Pugeault N. 2021 A 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)
99. Mukkavilli SK, Bara A, Gagne DJ, Tissot P, Campos E, Ganguly AR, Joppa L, Meger D, Dudek G. 2019 EnviroNet: imagenet for environment. In 99th American Meteorological Society Annual Meeting. AMS, 2019 January.
100. Rasp S, Dueben PD, Scher S, Weyn JA, Mouatadid S, Thuerey N. 2020 WeatherBench: A benchmark dataset for data-driven weather forecasting. See https://arxiv.org/abs/2002.00469.