PHILOSOPHICAL TRANSACTIONS A

royalsocietypublishing.org/journal/rsta

Research

Check for updates

Cite this article: 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. https://doi.org/10.1098/rsta.2020.0099

Accepted: 12 June 2020

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

Subject Areas:
meteorology, statistics

Keywords:
data integration, uncertainty quantification, quantile regression, model stacking, decision theory, artificial intelligence

Author for correspondence:
Charlie Kirkwood
e-mail: c.kirkwood@exeter.ac.uk

THE ROYAL SOCIETY
PUBLISHING

A framework for probabilistic weather forecast post-processing across models and lead times using machine learning

Charlie Kirkwood[1], Theo Economou[1], Henry Odbert[2] and Nicolas Pugeault[3]

[1]College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK
[2]Met Office, Exeter, UK
[3]School of Computing Science, University of Glasgow, Glasgow, UK

CK, 0000-0003-3218-4097; TE, 0000-0001-8697-1518

Forecasting the weather is an increasingly data-intensive exercise. Numerical weather prediction (NWP) models are becoming more complex, with higher resolutions, and there are increasing numbers of different models in operation. While the forecasting skill of NWP models continues to improve, the number and complexity of these models poses a new challenge for the operational meteorologist: how should the information from all available models, each with their own unique biases and limitations, be combined in order to provide stakeholders with well-calibrated probabilistic forecasts to use in decision making? In this paper, we use a road surface temperature example to demonstrate a three-stage framework that uses machine learning to bridge the gap between sets of separate forecasts from NWP models and the 'ideal' forecast for decision support: probabilities of future weather outcomes. First, we use quantile regression forests to learn the error profile of each numerical model, and use these to apply empirically derived probability distributions to forecasts. Second, we combine these probabilistic forecasts using quantile averaging. Third, we interpolate between the aggregate quantiles in order to generate a full predictive distribution, which we demonstrate has properties suitable for decision support. Our results suggest that this approach provides an effective and operationally viable framework for the cohesive post-processing of weather forecasts across multiple models and lead times to produce a well-calibrated probabilistic output.

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

1. Introduction

The 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.

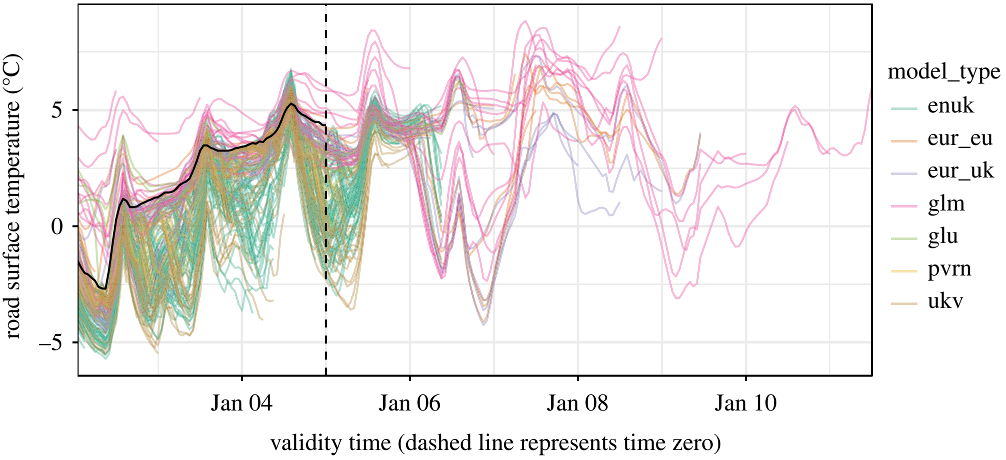

Figure 1. A visualization of the information provided by numerical weather prediction (NWP) forecasts. Each coloured line represents an ensemble member from a different model type. Observations (solid black line) go as far as time zero (vertical dashed line: the 'current time', which is 00:00 on 5 January 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.)

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.

While 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.

2. Post-processing framework

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.

(a) From deterministic to probabilistic forecasts

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

$$\epsilon_{t,m} = y - x_{t,m}, \tag{2.1}$$

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

$$\hat{y}_{t,m} = x_{t,m} + \epsilon_{t,m}. \tag{2.2}$$

Modelling 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.

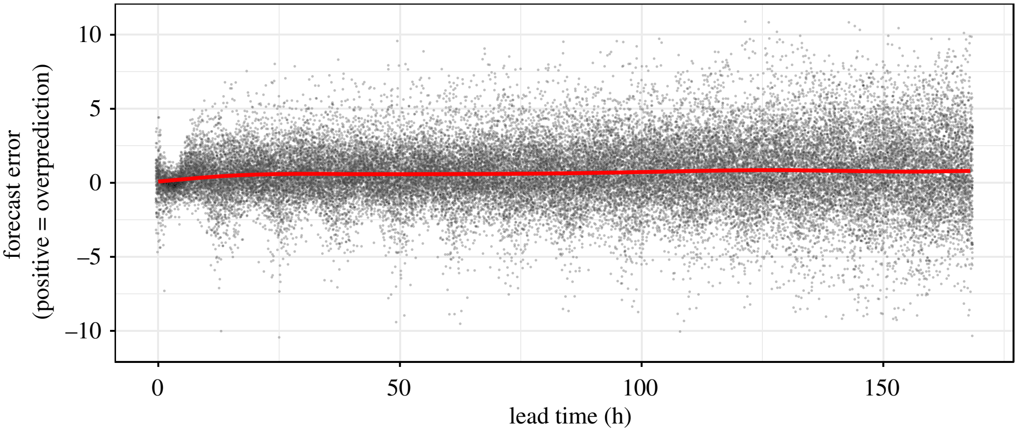

Figure 2. Plot of $\epsilon_{t,m}$ for $m = \text{glm}$ against lead hour (1, 2, . . . , 168) for a random sample of our dataset (spanning multiple months of absolute time). Each point is $\epsilon_{t,m}$ at a single hourly time step. The red line is a smooth estimate of the mean. (Online version in colour.)

Figure 2 shows $\epsilon_{t,m}$ for $m = \text{glm}$ (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

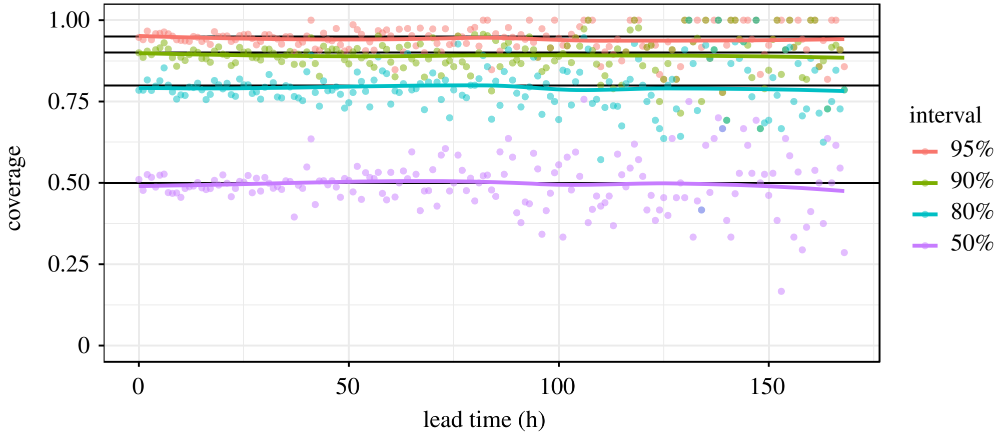

Figure 3. Coverage of the 50%, 80%, 90% and 95% QRF prediction intervals on out-of-bag data from one training scenario (though the picture is indicative of other scenarios). The coverage is the proportion of observations that fall within each prediction interval, and should match the interval (i.e. with 95% of observations falling within the 95% prediction interval) in a well-calibrated set-up. (Online version in colour.)

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.

Once 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.

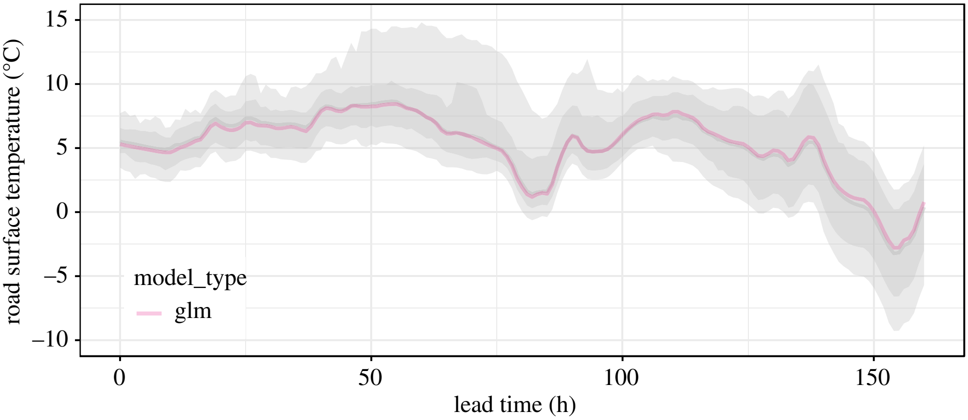

Figure 4. A deterministic NWP forecast for $m =$ glm that has been converted to a probabilistic forecast using equation (2.2). The 80% and 95% prediction intervals are shown as overlain grey ribbons, while the solid grey line is the median (which differs little from the NWP forecast here). (Online version in colour.)

(b) Combining probabilistic forecasts

The 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.

Our 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

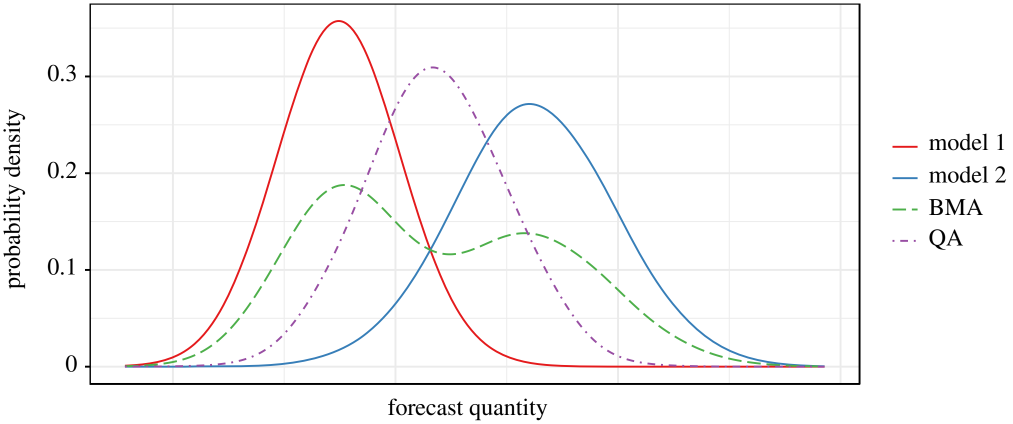

Figure 5. Synthetic example of combining two probabilistic forecasts using Bayesian model averaging (BMA) and quantile averaging (QA), after [29].

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).

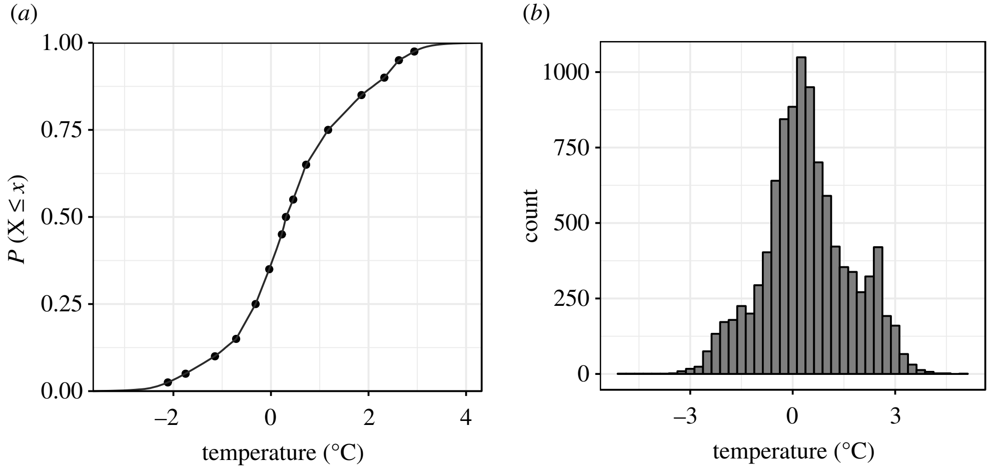

Figure 6. Interpolated CDF of the combined predictive distribution (*a*), and corresponding road surface temperature simulation (*b*) for a particular 50 h ahead forecast.

(c) Simulation from the full predictive distribution

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 $P(X \le x) = 0$ and up to $P(X \le x) = 1$ assuming tails that decay exponentially with a rate that ensures the corresponding PDF sums to one (figure 6*a*; 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 6*b*). 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.

3. Results

To 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.

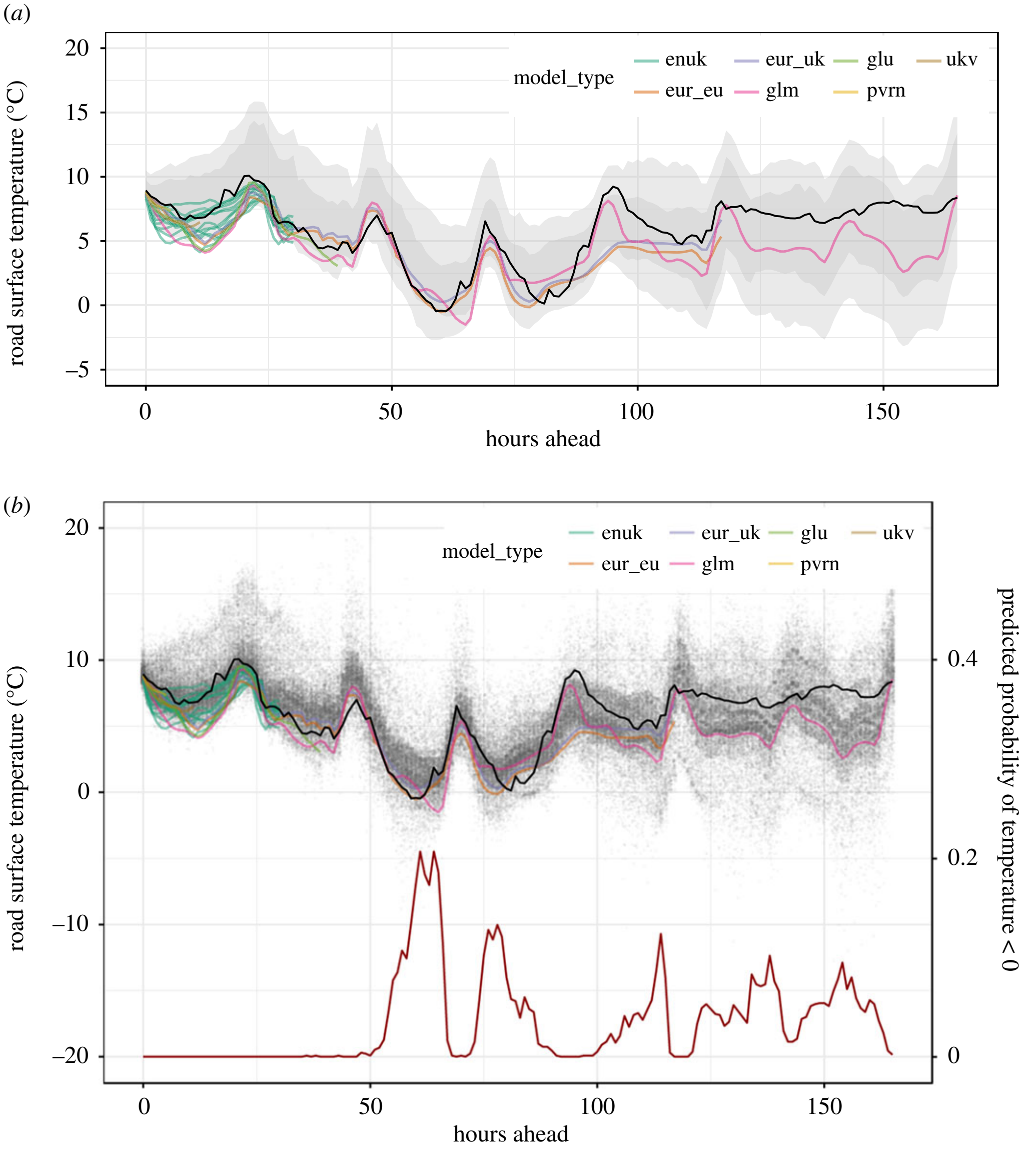

Figure 7. An example of the output of our post-processing framework. (*a*) The probabilistic forecast is visualized by the 80% and 95% prediction intervals. (*b*) Simulations from the full predictive distribution as grey dots, while the red line (right-hand *y*-axis) shows the probability of temperature being <0°C. NWP model forecasts are shown by coloured lines, and the true observed temperature (not known at time of forecasting) is shown by a solid black line. (Online version in colour.)

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.

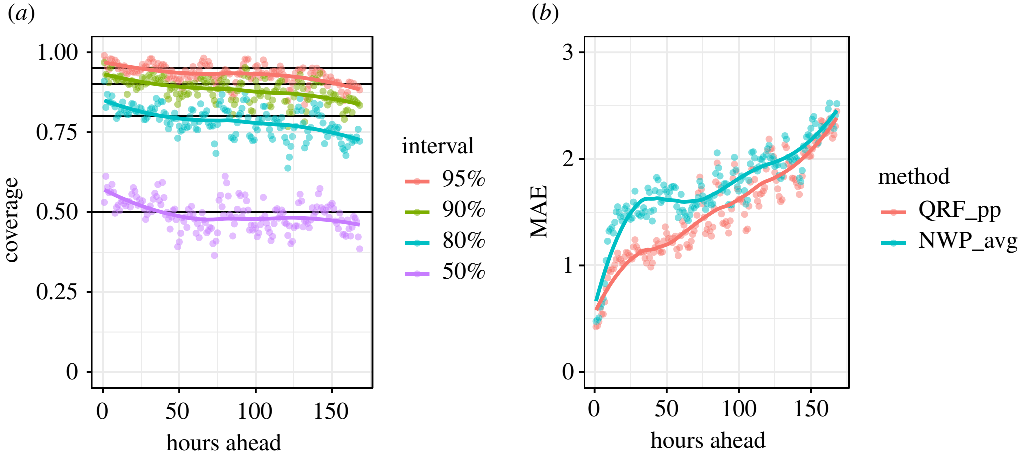

Figure 8. Evaluation metrics from 200 forecast scenarios. (*a*) Coverage of the prediction intervals of the combined probabilistic forecasts. (*b*) The MAE achieved by the median of the combined probabilistic forecast (QRF_pp) compared to taking the median of the available NWP forecasts (NWP_avg). (Online version in colour.)

Various metrics could be used to evaluate the skill of our probabilistic forecasts over multiple scenario runs. From the perspective of decision support, the ideal metric to evaluate would be the change in loss resulting from using our forecasts to make real-world decisions, such as about when to grit roads in our case. However, in the interest of a more general analysis, we use a range of standard metrics. These are: prediction interval coverage (figure 8*a*), the mean-absolute-error (MAE) of the median (figure 8*b*, because sometimes a single 'best' deterministic forecast is still desired), as well as the continuous ranked probability score (CRPS) and logarithmic score of our probabilistic forecast (both in figure 9).

Figure 8 indicates that coverage is good overall, with 94.7% of observations falling within the 95% prediction interval, although there is some over-dispersion of our forecast at the shortest ranges and under-dispersion at the longest ranges. This is an indication that, despite producing near perfect results on out-of-bag training data (figure 3), the QRF performance diminishes slightly when applied to new data. The range dependent over- and under-dispersion may be due to the partitioning process on which the forest is grown—by necessity the partitions that represent the extremes of forecast range must extend some distance towards the middle of the range, and in doing so end up capturing an empirical error distribution that is slightly biased towards the average empirical error distribution, rather than perfectly representing the distribution at the extremes of covariates. It may be the case that other data modelling approaches could do better in this respect.

Although deterministic performance was not our focus, the QRF median prediction does outperform the median of the available NWP models across the entire forecast range in terms of MAE. While only a conceptual benchmark, this can be taken as some indication that we have not 'thrown away' deterministic performance in pursuit of probabilistic calibration. Figure 8 also indicates that our method results in a monotonically increasing error with forecast range, unlike the median of the original NWP forecasts. Similarly, we see a monotonic increase in both the CRPS and the logarithmic score with increasing forecast range (figure 9*a*,*b*), and, when compared with the performance of the raw NWP ensemble on the same metrics, find our QRF post-processing approach to perform better. In the case of CRPS, our QRF post-processing approach reduces the rate at which forecasting skill decreases with forecast range. Also, by looking at the spread of performance across individual forecasting scenarios (represented by individual points in figure 9, rather than the lines, which trace the mean) we can see that our QRF post-processing approach reduces the variance in forecasting skill across different forecasting scenarios, making it a more consistent forecast than raw NWP. In the case of logarithmic score (figure 9*b*), we see again that

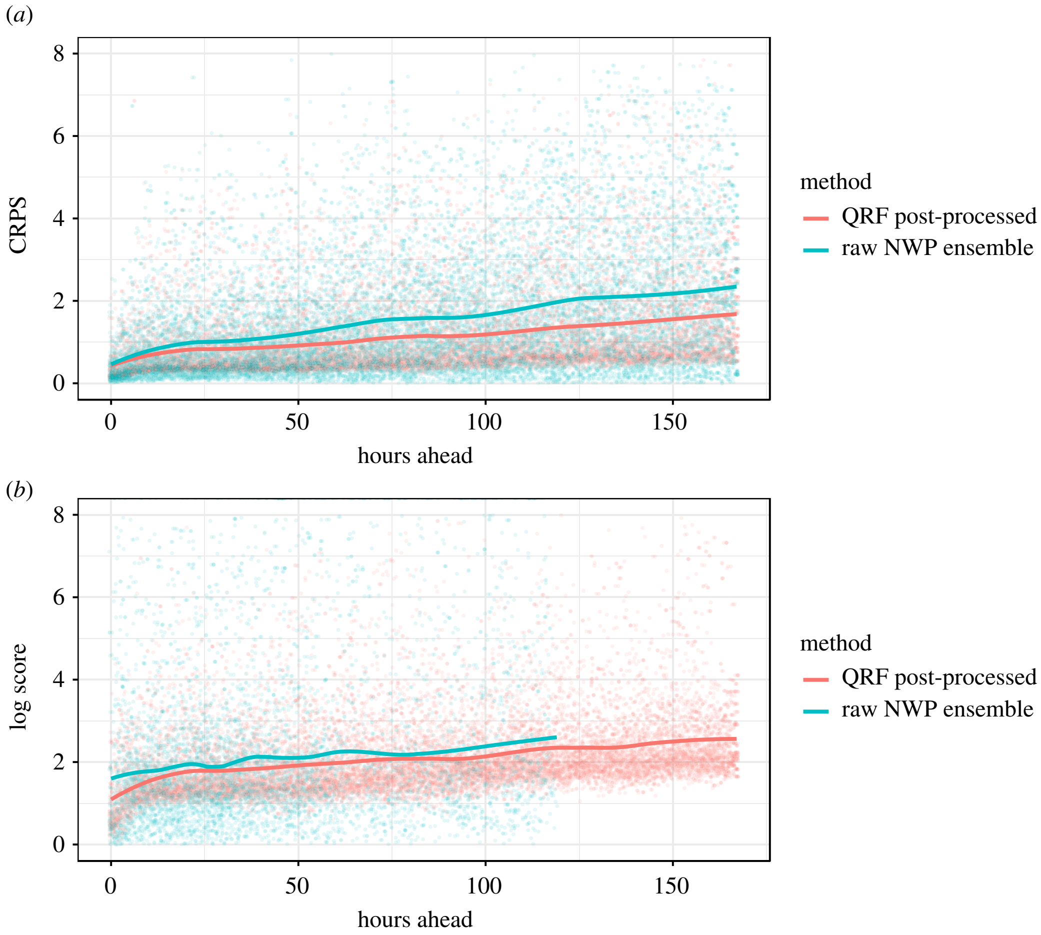

Figure 9. Evaluation metrics of our post-processing framework across all lead times on 200 random forecast scenarios. We compare our QRF post-processed output to the raw NWP ensemble in terms of continuous ranked probability score (CRPS, *a*) and logarithmic score (*b*). (Online version in colour.)

the forecasting skill provided by the QRF post-processed output is more consistent than that of the raw NWP ensemble, although the difference in the mean performance is less pronounced. The logarithmic score of the NWP ensemble cannot be obtained at longer ranges as only a single deterministic forecast is available. The authors recognize that comprehensive comparisons of our approach to other probabilistic post-processing approaches (in addition to raw NWP output) will be important to consider when choosing the best approach for any operational set-up. While we do not offer such comparisons in this study, we have made our dataset openly accessible as one of several benchmark datasets compiled by Haupt *et al.* [39] at https://doi.org/10.6075/J08S4NDM in the hope that it will facilitate comparison of different post-processing approaches on common benchmarks in the future.

In terms of speed, training the QRF for each forecast scenario takes between just 3 and 4 s on an i7-8550U laptop, and so the implementation of this framework can be expected to add very little overhead to a typical operational NWP forecasting set-up.

4. Discussion and conclusion

The conversion of disparate forecasts into a cohesive probabilistic output is important. A key function of weather forecasts is to support decision making, but current numerical methods do not provide the well-calibrated probabilistic output required to do this rigorously. By applying our framework, we compensate for this shortcoming, effectively supplementing forecasts with information from their historic performance in order to combine all available deterministic inputs, for all lead times, into a single well-calibrated probabilistic forecast. While our approach is by no means the first to provide probabilistic post-processing of weather forecasts, we believe the flexibility and speed provided by our use of machine learning, along with our framework's relative simplicity and ability to simultaneously deal with all available models and lead times, makes it a strong option for consideration in operational forecasting settings.

In this study, we have only applied our framework to site-specific forecasting, but there are no fundamental reasons why the same principles cannot be applied to spatial forecasting by providing the QRF with additional spatial covariates against which to learn its error profiles, or by adopting the standardized anomaly model output statistics (SAMOS) approach as proposed by Dabernig *et al.* [24]. The error modelling approach that we use seems a very effective way of minimizing the amount of training data required compared to predicting absolute values. Taillardat *et al.* [14], who also make use of QRF in their post-processing, initially used 4 years of training data for their absolute value forecasting system in 2016, but have since adopted an error modelling approach themselves [23].

There are still several aspects of our framework that are open to further investigation. One significant aspect that we explored in preliminary experiments but have not included in our methodology here, is the opportunity to use weighted quantile averaging for combining forecasts. In our set-up, where all of the inputs are recent NWP forecasts (and therefore similarly skilful), we saw negligible difference in using a weighted averaging approach, but in situations where more diverse forecast types are in use, it may prove beneficial to assign weightings according to forecast skill. A dynamic weighting approach also enables individual models to be updated without jeopardizing the overall post-processed output, as the contribution of the new or updated model will be minimal until it is error profile is well understood. The QRF algorithm provides a convenient means by which skill can be estimated ahead of time, in the form of out-of-bag metrics. For example, we showed earlier the out-of-bag coverage of our trained QRF (figure 3). Metrics such as the CRPS, logarithmic score, and Kullback–Leibler divergence would provide good comparisons of forecast skill on which to base quantile averaging weight, although their calculation would add some additional processing time. Yao *et al.* [40] provide more detail about using such metrics for weighted model stacking, and in fact these weights can be optimized as an additional supervised learning problem [33].

The overall strategy for combining forecasts is also open to further research. Because it retains the inter-model variance, BMA may be considered to provide a better representation of extreme outcomes at the expense of well-calibrated coverage (at least in set-ups where each input forecast is already well-calibrated, which is likely to become the norm). We also think that the output of BMA would be difficult to make use of in practice when applied across all lead times as in our framework, because of the discrepancy in the number of models available at each time step, and therefore the spurious inconsistency of the inter-model variance across the forecast range. Still, applications where capturing extremes is a priority may wish to investigate further. For general purposes, we are satisfied with our time-consistent and calibration-preserving quantile averaging approach.

It is our belief that, as time goes on, and the number of different forecasting models in use—along with their complexity and resolution—continues to increase, there will be increasing need for algorithmic interfaces such as ours to summarize the otherwise overwhelming sea of forecast information into decision-ready output. This would consist of optimally well-calibrated probabilities of future weather outcomes given all available information. Probabilistic machine learning is a technology that can enable this, and we hope that the work we have demonstrated here will go some way in aiding progression towards this goal.

Data accessibility. Our dataset has been made available with permission from the Met Office and Highways England, for which we are grateful. It is available for download along with several other open weather forecast

post-processing datasets collated by Haupt *et al.* [39] at https://doi.org/10.6075/J08S4NDM. In addition, the code for this study can be accessed at https://github.com/charliekirkwood/mlpostprocessing

Authors' contributions. All authors shared in conceiving and designing the study. C.K. wrote the code and manuscript, and ran the experiments behind the solution we present here, with other authors providing supervision. All authors read and approved the manuscript.

Competing interests. We declare we have no competing interests.

Funding. This work has been conducted as part of CK's PhD research, funded by the UK's Engineering and Physical Sciences Research Council (EPSRC project ref: 2071900) and the Met Office.

Acknowledgements. The lead author is grateful for the insightful discussions and community feedback that came from attending the Machine Learning for Weather and Climate Modelling conference at the University of Oxford in 2019, and the EUMETNET Workshop on Artificial Intelligence for Weather and Climate at the Royal Meteorological Institute of Belgium in 2020. In addition, C.K. would like to thank Thomas Voigt for offering his data engineering expertise. We are grateful to all who have helped to guide this study.

References

1. Rahmstorf S, Coumou D. 2011 Increase of extreme events in a warming world. *Proc. Natl Acad. Sci. USA* **108**, 17 905–17 909. (doi:10.1073/pnas.1101766108)
2. Abbe C. 1901 The physical basis of long-range weather forecasts. *Mon. Weather Rev.* **29**, 551–561. (doi:10.1175/1520-0493(1901)29[551c:TPBOLW]2.0.CO;2)
3. Bjerknes V. 1904 Das Problem der Wettervorhers-age, betrachtet vom Standpunkte der Mechanik und der Physik. *Meteor. Z.* **21**, 1–7.
4. Richardson LF. 1922 *Weather prediction by numerical process*. Cambridge, UK: Cambridge University Press.
5. Bauer P, Thorpe A, Brunet G. 2015 The quiet revolution of numerical weather prediction. *Nature* **525**, 47–55. (doi:10.1038/nature14956)
6. Economou T, Stephenson DB, Rougier JC, Neal RA, Mylne KR. 2016 On the use of Bayesian decision theory for issuing natural hazard warnings. *Proc. R. Soc. A* **472**, 20160295. (doi:10.1098/rspa.2016.0295)
7. Simpson M *et al.* 2016 Decision analysis for management of natural Hazards. *Annu. Rev. Environ. Res.* **41**, 489–516. (doi:10.1146/annurev-environ-110615-090011)
8. Gulshan V *et al.* 2016 Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316**, 2402–2410. (doi:10.1001/jama.2016.17216)
9. Silver D *et al.* 2017 Mastering the game of Go without human knowledge. *Nature* **550**, 354–359. (doi:10.1038/nature24270)
10. Hey T, Butler K, Jackson S, Thiyagalingam J. 2020 Machine learning and big scientific data. *Phil. Trans. R. Soc. A* **378**, 20190054. (doi:10.1098/rsta.2019.0054)
11. Krizhevsky A, Sutskever I, Hinton GE. 2012 Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105.
12. Rasp S, Lerch S. 2018 Neural networks for postprocessing ensemble weather forecasts. *Mon. Weather Rev.* **146**, 3885–3900. (doi:10.1175/MWR-D-18-0187.1)
13. 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)
14. Taillardat M, Mestre O, Zamo M, Naveau P. 2016 Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Mon. Weather Rev.* **144**, 2375–2393. (doi:10.1175/MWR-D-15-0260.1)
15. Meinshausen N. 2006 Quantile regression forests. *J. Mach. Learn. Res.* **7**, 983–999.
16. Stephens E, Cloke H. 2014 Improving flood forecasts for better flood preparedness in the UK (and beyond). *Geogr. J.* **180**, 310–316. (doi:10.1111/geoj.12103)
17. Raftery AE, Gneiting T, Balabdaoui F, Polakowski M. 2005 Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Weather Rev.* **133**, 1155–1174. (doi:10.1175/MWR2906.1)
18. Glahn HR, Lowry DA. 1972 The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteorol.* **11**, 1203–1211. (doi:10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2)

19. Wilks DS, Hamill TM. 2007 Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Weather Rev.* **135**, 2379–2390. (doi:10.1175/MWR3402.1)
20. Gneiting T, Raftery AE, Westveld III AH, Goldman T. 2005 Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Weather Rev.* **133**, 1098–1118. (doi:10.1175/MWR2904.1)
21. Hengl T, Heuvelink GBM, Tadić MP, Pebesma EJ. 2012 Spatio-temporal prediction of daily temperatures using time-series of MODIS LST images. *Theor. Appl. Climatol.* **107**, 265–277. (doi:10.1007/s00704-011-0464-2)
22. Akbari Asanjan A, Yang T, Hsu K, Sorooshian S, Lin J, Peng Q. 2018 Short-term precipitation forecast based on the PERSIANN system and LSTM recurrent neural networks. *J. Geophys. Res.: Atmos.* **123**, 12 543–12 563. (doi:10.1029/2018JD028375)
23. Taillardat M, Mestre O. 2020 From research to applications - examples of operational ensemble post-processing in France using machine learning. *Nonlinear Process. Geophys. Discuss.* **27**, 329–347. (doi:10.5194/npg-2019-65)
24. Dabernig M, Mayr GJ, Messner JW, Zeileis A. 2017 Spatial ensemble post-processing with standardized anomalies. *Q. J. R. Meteorol. Soc.* **143**, 909–916. (doi:10.1002/qj.2975)
25. Wright MN, Ziegler A. 2017 ranger: a fast implementation of random forests for high dimensional data in C++ and R. *J. Stat. Softw.* **77**, 1–17. (doi:10.18637/jss.v077.i01)
26. Athey S, Tibshirani J, Wager S. 2019 Generalized random forests. *Ann. Stat.* **47**, 1148–1178. (doi:10.1214/18-AOS1709)
27. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. 1999 Bayesian model averaging: a tutorial (with comments by M Clyde, D Draper and EI George, and a rejoinder by the authors. *Stat. Sci.* **14**, 382–417. (doi:10.1214/ss/1009212519)
28. Gneiting T, Ranjan R. 2013 Combining predictive distributions. *Electron. J. Stat.* **7**, 1747–1782. (doi:10.1214/13-EJS823)
29. Schepen A, Wang QJ. 2015 Model averaging methods to merge operational statistical and dynamic seasonal streamflow forecasts in Australia. *Water Resour. Res.* **51**, 1797–1812. (doi:10.1002/2014WR016163)
30. Genest C. 1992 Vincentization revisited. *Ann. Stat.* **20**, 1137–1142. (doi:10.1214/aos/1176348676)
31. Vincent SB. 1912 The function of the vibrissae in the behavior of the white rat. *Animal Behavior Monographs*, 1, 5, 84–84.
32. Ratcliff R. 1979 Group reaction time distributions and an analysis of distribution statistics. *Psychol. Bull.* **86**, 446–461. (doi:10.1037/0033-2909.86.3.446)
33. Ren Y, Zhang L, Suganthan PN. 2016 Ensemble classification and regression-recent developments, applications and future directions [Review Article]. *IEEE Comput. Intell. Mag.* **11**, 41–53. (doi:10.1109/MCI.2015.2471235)
34. Belkin M, Hsu D, Ma S, Mandal S. 2019 Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proc. Natl Acad. Sci. USA* **116**, 15 849–15 854. (doi:10.1073/pnas.1903070116)
35. Bentzien S, Friederichs P. 2014 Decomposition and graphical portrayal of the quantile score. *Q. J. R. Meteorol. Soc.* **140**, 1924–1934. (doi:10.1002/qj.2284)
36. Gneiting T, Raftery AE. 2007 Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **102**, 359–378. (doi:10.1198/016214506000001437)
37. Quiñonero-Candela J, Rasmussen CE, Sinz F, Bousquet O, Schölkopf B. 2006 Evaluating predictive uncertainty challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, Lecture Notes in Computer Science, pp. 1–27. Berlin, Heidelberg: Springer.
38. Cannon AJ. 2011 Quantile regression neural networks: implementation in R and application to precipitation downscaling. *Comput. Geosci.* **37**, 1277–1284. (doi:10.1016/j.cageo.2010.07.005)
39. 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. (doi:10.1098/rsta.2020.0091)
40. Yao Y, Vehtari A, Simpson D, Gelman A. 2018 Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Anal.* **13**, 917–1007. (doi:10.1214/17-BA1091)