Predict4Resilience: Discovery Phase – Review of existing ensemble numerical weather prediction products

Author: Megan Pearce (Scientific Consultant)

Reviewed by: Ian Pearman (Senior Applied Scientist) and Dr Edward Steele (Science Manager)

Approved by: Dr Joe Osborne (Industry Managing Consultant)

Revision Date: April 2022

© Crown Copyright 2022, Met Office

If printing double-sided you will need this blank page. If printing single sided, please delete this page

© Crown copyright 2022, Met Office

Revision History

VERSION	DESCRIPTION	AUTHOR	REVIEWER	APPROVED
0.1	Draft report	Megan Pearce March 2022	Ian Pearman Dr Edward Steele March 2022	
1	Final report	Megan Pearce March 2022	Dr Joe Osborne March 2022	Dr Joe Osborne March 2022

© Crown copyright 2022, Met Office

Contents

© Crown copyright 2022, Met Office

1 Overview of Ensemble Prediction Systems (EPS)

The purpose of a forecast is to estimate the future state of the atmosphere, by evaluating the current state and calculating how this will evolve in time. The current state of the atmosphere is estimated using observations, which are taken from a number of sources such as radiosondes, surface weather stations, aircrafts, radars and satellites[1]. Data assimilation is a procedure that uses the observational information, along with previous forecasts, in a continuous, sequential process to estimate the initial state of the model and nudge the forecast trajectory[2]. A forecast is very sensitive to the accuracy of this initial 'observed' atmosphere input, and a small error in the initial state can lead to large forecast errors as the result of the atmosphere's chaotic system[3].

1.1 What is an ensemble forecast?

As highlighted above the evolution of a forecast is very sensitive to the starting conditions, which creates a source of uncertainty in forecast modelling. Another source of uncertainty is the forecast model itself, which uses equations and approximations to represent complex atmospheric dynamics[4]. A deterministic forecast is generated by running a forecast model once, using one set of initial conditions, to provide a single forecast output. An ensemble forecast is generated by running the forecast model a number of times from slightly different starting conditions and can be used to help quantify sources of uncertainty[3,5]. This is depicted in Figure 1, with each blue line representing an individual forecast (an ensemble member) which can be combined as a set of forecasts to create an ensemble. Each blue line has slightly different initial conditions and results in a different forecast output as the model moves through time, allowing a quantification of both initial condition uncertainty and forecast uncertainty. The red line shows a single deterministic forecast.

[1] https://www.metoffice.gov.uk/research/weather/data-assimilation/use-of-observations
[2] https://www.ecmwf.int/en/research/data-assimilation
[3] https://www.metoffice.gov.uk/research/weather/ensemble-forecasting/what-is-an-ensemble-forecast
[4] https://www.metoffice.gov.uk/research/weather/ensemble-forecasting/mogreps
[5] https://www.ecmwf.int/en/about/media-centre/focus/2017/fact-sheet-ensemble-weather-forecasting

© Crown copyright 2022, Met Office

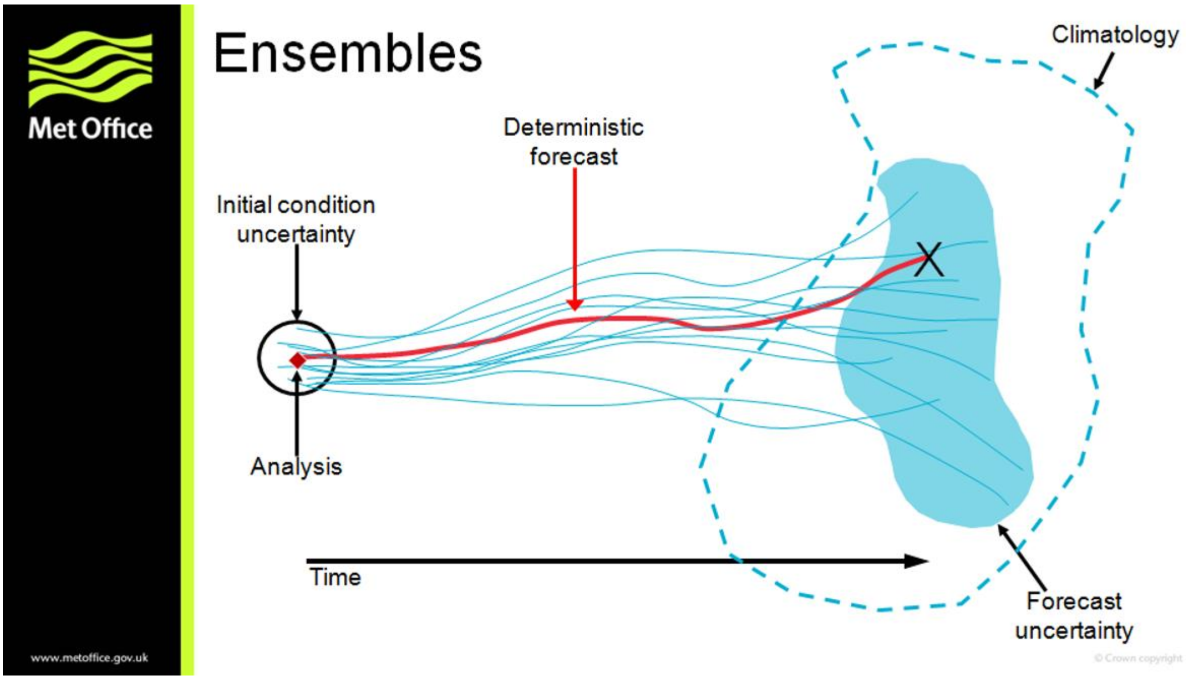

Figure 1: A schematic illustration of an ensemble forecast[3]

1.2 Availability of ensemble prediction systems

Models used for weather forecasting are called numerical weather prediction (NWP) models. At the Met Office, the Unified Model is run in multiple configurations as the operational NWP model used for weather forecasting, however it can also be used for seasonal forecasting and climate modelling[6].

The principal, large-scale weather configuration is the **Global**, which provides deterministic, medium-range (six days) weather forecasts at a 10km horizontal spatial resolution[6]. The global ensemble is **MOGREPS-G**, which provides an ensemble forecast up to seven days ahead with a 20km horizontal resolution over the globe. The development of the Met Office Global and Regional Ensemble Prediction System (MOGREPS) was primarily to aid the forecasting of rapid storm development, wind, rain, snow and fog, to improve Met Office capabilities in forecasting severe weather[4], by sampling a broader set of plausible outcomes.

The Global deterministic and MOGREPS-G ensemble support higher resolution, nested models by providing boundary data. The nested high resolution models are convection-permitting, which means they are capable of simulating small-scale behaviour seen in the atmosphere (as opposed to parameterising it) as well as the finer influences of mountains, coastlines and urban areas (Kendon et al., 2021). This is important for capturing the detailed behaviour of weather events that vary or change quickly over short spatial or temporal

[6] https://www.metoffice.gov.uk/research/approach/modelling-systems/unified-model/weather-forecasting

© Crown copyright 2022, Met Office

timescales, such as convective rainfall and wind, and capturing the influence of topographical features. The deterministic UK model is **UKV**, a variable resolution model that is nested over the UK providing a deterministic forecast out to five days[6]. The UKV model has a high resolution (1.5km horizontal) grid over the UK domain and a coarser grid (4km horizontal resolution) over an outer domain, with the two grids separated by a variable resolution transition zone that reduces unwanted effects occurring between the two boundaries[6]. **MOGREPS-UK** is the regional ensemble providing a five-day ensemble forecast over the UK[4]. Similar to UKV, **MOGREPS-UK** is a nested, convection permitting model which has a 2.2km horizontal resolution over the inner domain (over the UK) and a 4km horizontal resolution over the outer domain.

In addition to the Met Office, ensemble prediction systems are run at most operational global weather centres. The European Centre for Medium Range Weather Forecasts (ECMWF) and the National Centre for Environmental Prediction (NCEP) are also among the most widely used (Steele et al., 2017).

The ECMWF produced the first ensemble forecasts in 1992 and has since been developing techniques to increase forecast sensitivity and quantify uncertainty in observations[5], routinely upgrading their Integrated Forecasting System (IFS)[7]. The IFS uses two models to produce forecasts in the early-medium time horizon. The **ENS** ensemble is made up fifty-one members with an 18km horizontal spatial resolution, and has a forecast horizon out to fifteen days. ENS consists of a control member (utilising the most accurate estimate of the current conditions and model physics), from which the perturbations for another fifty ensemble members are applied to explore the model and observational uncertainty[8]. Additionally, there is the **HRES**, a singular deterministic forecast, named for its higher spatial resolution (9km) compared to ENS, that is often combined with ENS in the early time horizon (0-10 days) (section 1.3)[8]. The ECMWF also carry out extended (10-46 days)[9] and long range (1-3 months)[10] ensemble forecasting using the IFS.

Another example is the Global Ensemble Forecast System (**GEFS**), which is a twenty-one member ensemble forecast system with global coverage created by the National Centers for Environmental Prediction (NCEP), formed of one control and twenty perturbed members[11]. However, given the spatial resolution and ensemble size, the GEFS is not considered as

[7] https://www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model
[8] https://www.ecmwf.int/en/forecasts/documentation-and-support/medium-range-forecasts
[9] https://www.ecmwf.int/en/forecasts/documentation-and-support/extended-range-forecasts
[10] https://www.ecmwf.int/en/forecasts/documentation-and-support/long-range
[11] https://www.ncei.noaa.gov/products/weather-climate-models/global-ensemble-forecast

© Crown copyright 2022, Met Office

valuable over the region of interest in comparison to MOGREPS or ENS and therefore is not evaluated further in this report.

The deterministic and ensemble weather forecasting models that have been described are summarised in Table 1 in relation to their spatial domain, grid resolution, and forecast horizons. This is not an exhaustive list of available EPS but reflect those reviewed as most suitable for consideration for this project application.

© Crown copyright 2022, Met Office

Table 1: A summary of relevant ensemble prediction systems, noting the resolution and forecast period.

Model *Centre*	Grid length in mid-latitudes	Vertical levels	Forecast length	Run times (UTC)	Ref
Global Deterministic Global *Met Office*	10 km (2560 x 1920 grid points)	70 (lid ~80 km)	6 days (for 00z and 12z)	00, 06, 12, 18	6
MOGREPS-G Ensemble Global *Met Office*	20 km (1280 x 960 grid points)	70 (lid ~80 km)	Control member + 17 perturbed members to 7 days. 36 member ensemble generated by time-lagging over 12 hours (latest 2 cycles).	00, 06, 12, 18	4, 6
UKV Deterministic UK *Met Office*	1.5 km inner domain (622 x 810 grid points); 4 km outer domain (950 x 1025 grid points)	70 (lid ~40 km)	120 hours 54 hours 12 hours	03,15 00, 06, 09, 12, 18, 21 All other hours	6
MOGREPS-UK Ensemble UK *Met Office*	2.2 km inner domain; 4km outer domain (740 x 752 grid points)	70 (lid ~40 km)	Control member + 2 perturbed members to 5 days 3 perturbed members to 5 days 18 member ensemble generated by time-lagging over 6 hours (latest 6 cycles)	05, 11, 17, 23 All other hours Every hour	4, 6
HRES Deterministic *ECMWF*	9km	137 (lid ~80 km)	10 days	00, 12	15, 12
ENS Ensemble *ECMWF*	18km	137 (lid ~80 km)	51 members (control + 50 perturbed members) to 15 days	00, 12	12, 13

Whilst a deterministic forecast may provide a single 'best estimate' outcome (Hagelin et al., 2017), an ensemble forecast provides a range of possible outcomes that can be used by Operational Meteorologists and service providers to gauge a better understanding of what

[12] https://confluence.ecmwf.int/display/FUG/Grid+point+Resolution
[13] https://confluence.ecmwf.int/display/FUG/ENS+-+Ensemble+Forecasts

© Crown copyright 2022, Met Office

weather events may occur and how likely a particular event will be[3]. If there is large spread in the ensemble realisations (e.g. a large shaded blue area in Figure 1) then there is greater uncertainty in the forecast. If each ensemble member (i.e. each blue line) results in a similar forecast output, there is greater confidence in predicting a particular weather event[3]. As a result, ensemble models can be used to generate probabilistic forecasts by utilising post processing techniques to calculate the probability of a pre-defined threshold being exceeded (or not) for a forecast variable. The probability is estimated based on the proportion of the ensemble that forecasts values exceeding that threshold (Hagelin et al., 2017). It should be noted that ensembles are often under-dispersive (the observed outcome falls outside the spread of the ensemble forecast), and this can result in an over-confident probabilistic forecast (Hagelin et al., 2017).

1.3 Introduction to multi-model ensembles

Multi-model ensembles combine two or more ensemble prediction systems, with the aim to reduce forecast uncertainty arising from each of the individual systems in the ensemble. The approach enhances the ensemble spread by utilising different physics parametrizations used in different models, as well the differences in, or perturbations of, initial conditions within a single model – depending on whether the models are deterministic or ensembles (Beck et al., 2016). For a probabilistic multi-model ensemble there are two approaches; either pooling the ensemble members from each model and generating a singular probabilistic forecast, or to separately compute a probabilistic forecast for each model and then combine them[14].

An example of a multi-model ensemble is described in Beck et al. (2016), where three different ensemble prediction systems were combined over Western Europe to create two different multi-model ensembles. The three systems used were MOGREPS-UK (in the previous twelve-member configuration used operationally at the Met Office at the time of writing of Beck et al. (2016)), COSMO-DE-EPS (twenty-member ensemble with a 2.8km horizontal resolution operationally used by Deutscher Wetterdienst, the German national meteorological service), and AROME-EPS (twelve-member ensemble with 2.5km horizontal resolution that at the time of writing of Beck et al. (2016) was pre-operational for Météo-France, the French national meteorological service). The creation of two multi-model ensembles took the overlapping domain and forecast lead times for each system pair (MOGREPS-UK/AROME-EPS with twenty-four members and COSMO-DE-EPS/AROME-EPS with thirty-two members), using a uniform, fixed resolution after conversion to a common 2.2km grid and comparing a twenty-one hour window between 0300 and 0000 UTC (Beck et al. 2016). The two multi-model

[14] https://www.wmolc.org/seasonPmmeInfo/information

© Crown copyright 2022, Met Office

ensembles were evaluated for wind speed, temperature, relative humidity and six hourly precipitation accumulation during the summer of 2012 and spring of 2014, with each variable bias corrected for each individual ensemble prediction system prior to multi-model analysis. Beck et al. (2016) found that in most cases the multi-model ensembles performed better than the individual ensembles, and, in a precipitation case study, showed that combining the characteristics of two models into a multi-model forecast resulted in a forecast more closely matching observed precipitation distribution and intensity.

An example of a multi-model ensemble is the combination of the ECMWF deterministic (HRES) and ensemble (ENS) models. On average, the HRES provides the most accurate single run realisation of weather patterns, but any individual forecast may not be the most skilful when compared to other ENS members[15]. As it is deterministic, it also cannot provide an estimate of forecast uncertainty or confidence. The confidence placed on HRES and ENS members when combined in an ensemble forecast varies with forecast horizon and meteorological parameter, with greater weighting placed on HRES over short lead-times (day 1) compared to long lead-times (day 10) for surface variables, but a lower weighting overall towards the HRES for upper-air forecast variables where the higher resolution (and subsequent superior surface detail) of HRES has a less significant role[15]. The relative weighting placed between the HRES and ENS can also vary with location and annual cycle, and there a number of methods that can be used to explore the most suitable relative weights for a defined region of interest[16].

15 https://confluence.ecmwf.int/display/FUG/HRES+-+High-Resolution+Forecast
16 https://confluence.ecmwf.int/display/FUG/Relative+weights+for+HRES+and+ENS

© Crown copyright 2022, Met Office

2 Performance and verification

An important step in the forecasting process is to assess the accuracy of the forecasts, as well as the performance of an individual model (or ensemble) compared to other models, using a combination of observational datasets and verification techniques.

2.1 Forecast accuracy

Sources of "truth"

There are a number of sources of 'truth' that can be used for forecast verification, including raw observations at synoptic weather stations, gridded observational datasets and modelled truth such as reanalysis and hindcast datasets.

Data from the United Kingdom's network of surface stations is held in the Met Office Integrated Data Archive System (MIDAS)[17]. A subset of this data (Midas-open)[18] is available via the CEDA archive on an Open Government data license, containing the data of UK mainland land surface observations owned or operated by the Met Office. Midas-open represents approximately 95% of available daily temperature and weather observations, 83% of hourly weather data (wind speed and direction, cloud type and amount, visibility, and temperature), and 13% of daily rainfall within the full MIDAS collection[18]. The latter is significantly lower as a large proportion of rain gauges are operated by other agencies. However, this does not mean it isn't available from other sources. For example, the Scottish Environment Protection Agency (SEPA) provide the capability to download rainfall data from 275 rain gauges across Scotland under an Open Government License[19].

The data from land surface observation stations has also been interpolated onto a uniform 1km grid, in a dataset called HadUK-Grid[20], to provide consistent coverage over the UK. Variables available include air temperature, precipitation, sunshine, mean sea level pressure, wind speed, relative humidity and days of lying snow. The temporal resolution of these variables ranges from daily, monthly, seasonal and annual, as well as long term climatologies. For example, wind speed is only available at the monthly temporal resolution and therefore would not be of much value for this project.

Reanalysis datasets combine models with observations to give a gridded, numerical description of the historical climate. The fifth generation of ECMWF atmospheric reanalyses of the global climate is called ERA5[21]. The ERA5 dataset has global extent and provides global

[17] https://catalogue.ceda.ac.uk/uuid/245df050d57a500c183b88df509f5f5a?jump=related-anchor
[18] https://catalogue.ceda.ac.uk/uuid/dbd451271eb04662beade68da43546e1
[19] https://www2.sepa.org.uk/rainfall/DataDownload
[20] https://www.metoffice.gov.uk/research/climate/maps-and-data/data/haduk-grid/haduk-grid
[21] https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5

© Crown copyright 2022, Met Office

coverage of many atmospheric, land and oceanic climate variables on a ~30km grid (0.25° x 0.25°) for the period 1979 to within 5 days of real time. It offers complete spatial coverage (unlike MIDAS) and wind as a variable (unlike HadUK-Grid). Whereas there would be little merit in using ERA5 data as a source of 'truth' for rainfall and temperature, it might have some value in verifying wind forecasts where the spatial distribution of observing station data from MIDAS is not as good.

Verification

Murphy (1993) identifies three features for distinguishing what makes a forecast "good", which are consistency, quality and value. However, a high-quality forecast (predicting observed conditions with high skill) may not always have good value (support users in decision-making), and vice versa[22]. Verification is used to evaluate these features to monitor and assess forecast/model performance. The choice of verification method will vary depending on the forecast type (deterministic vs probabilistic) and specificity (for example, yes/no, categorical or continuous). A common verification method for dichotomous (yes/no) forecasts is the use of contingency tables (Wilks, 2011a). This is useful in impact modelling when considering whether an event (such as a fault) did or did not happen as forecast. An example structure of a contingency table is shown in Figure 2. There are four combinations of forecast vs observation outcomes, which are known as the joint distribution[22]. These are:

- **hit** - event forecast to occur, and did occur
- **miss** - event forecast not to occur, but did occur
- **false alarm** - event forecast to occur, but did not occur
- **correct rejection** - event forecast not to occur, and did not occur

		OBSERVED		
		YES	**NO**	*TOTAL*
FORECAST	**YES**	Hit	False alarm	*Forecast yes*
	NO	Miss	Correct rejection	*Forecast no*
	TOTAL	*Observed yes*	*Observed no*	

Figure 2: Example structure of a contingency table.

A 'perfect' forecast system would produce no misses or false alarms. The contingency table results can then be used to calculate a variety of categorical statistics that can be used to describe the forecast performance (Wilks, 2011a), with the choice of statistic depending on the performance question, for example, '*How well did the forecast "yes" events correspond to the observed "yes" events? – Gilbert Score', '*How well did the forecast separate the "yes" events from the "no" events? – Peirces's Score*' or '*What was the accuracy of the forecast*

[22] https://www.cawcr.gov.au/projects/verification/verif_web_page.html

© Crown copyright 2022, Met Office

relative to that of random chance? – Heidke Score'[22]. A contingency table can also be used for multi-category forecasts, applying a distribution approach to examine the relationship between the elements. This can make the nature of forecast errors easier to diagnose, however there are fewer statistics available to summarise the performance. There are also a number of other diagnostic verification methods that can be used for exploration of forecast error. As well as a verification tool for evaluating forecast accuracy, contingency tables can be utilised with ensemble probabilistic forecasts to aid operational decision-making, highlighted in further detail in section 0.

An additional consideration is the spatial choice during verification. The use of point location verification can result in duplicated penalisation arising from both a temporal and spatial mismatch between predicted and observed values (penalty for forecasting feature in wrong place and again for not forecasting in the correct place) (Hagelin et al., 2017).

2.2 Ensemble performance

To overcome limitations associated with duplicated penalisation, so-called neighbourhood methods (Mittermaier, 2014) can also be used to evaluate ensemble members and compare these with deterministic forecasts using a probabilistic approach (Hagelin et al., 2017). Research has been undertaken by Dey et al. (2014, 2016a, 2016b) to develop the neighbourhood technique to better account for day-to-day variability in forecast uncertainty and ensemble spread, by adapting the neighbourhood size in response to the spatial structure of the ensemble spread.

Examples of probabilistic measures that can be used for verification include Ensemble Bias[23], Brier Score[24], Ranked Probability Score (RPS)[25], and Continuous Ranked Probability Score (CRPS)[26] (Hagelin et al., 2017). Other commonly used verification measures include Hinton

[23] Ensemble Bias: the frequency bias of the ensemble and is defined as the ratio of the number of forecasted events multiplied by their probability to the frequency of observed events. The ideal EB value is 1, while a higher or lower value means a poor match.
[24] Brier Score: evaluates the probability of exceeding (or falling below) a pre-defined threshold
[25] Ranked Probability Score: a similar measure to Brier Score but for multiple categories, i.e. it answers the question of how well did a probability forecast predict the category into which the observation fell. Used for variables for which distributions are non-Gaussian and a suitable set of thresholds can be chosen. The lower the score the better.
[26] Continuous Ranked Probability Score: a version of Ranked Probability Score for continuous categories e.g. temperature which is a continuous variable with an approximately Gaussian distribution. Diurnal and seasonal variation limit used of threshold methods. The lower the score the better.

© Crown copyright 2022, Met Office

Diagrams[27], Receiver Operating Characteristic[28] and Reliability Diagrams[29] (Porson et al., 2020). Reliability, or calibration, (i.e. the relationship of the forecast to the distribution of the observations) is particularly important for extracting meaningful outcomes (Wilks, 2011a). For example, a probability forecast is well-calibrated when the probability output from the forecast is equal to the subsequent event frequency (i.e. if the forecast probability is rain for 50% of the time, then there should be should rain for 50% of the time – the forecast means what it says) (Wilks, 2011a).

An example of model evaluation is discussed in Hagelin et al. (2017), where neighbourhood processing is used to compare the performance of MOGREPS-UK to its deterministic counterpart, UKV. Over a 3-month trial between January and March 2016, comparing forecasts initiated at 0300, 0900, 15000 and 2100, MOGREPS-UK was found to perform better than UKV for all the variables considered i.e. the CRPS for temperature and RPS for wind speed and precipitation score better in the MOGREPS-UK ensemble compared to deterministic UKV and the control member of MOGREPS-UK when treated as a singular deterministic. However, it was also noted that the UKV scores better than the MOREGPS-UK control member (when the latter is treated as a singular deterministic forecast), highlighting the benefits of the higher resolution of UKV (Hagelin et al., 2017).

As the computing capacity at the Met Office increases with investment in supercomputing capabilities[30], consideration is given to the future upgrade options of MOGREPS-UK. Hagelin et al. (2017) explored the three main options for model improvement, which are; increasing the domain size (increased coverage), decreasing the horizontal grid-spacing (increased resolution) and increasing the number of ensemble members (reduced uncertainty). The objective verification found that all the options showed a benefit in one month summer and winter trials. The largest improvement was seen when increasing the ensemble size, and this was particularly the case for forecasting precipitation (Hagelin et al. 2017).

Porson et al. (2020) describe the most recent updates to MOGREPS-UK, which became operational in March 2019, and compare this to the previous configuration. The two main changes are the ensemble size and forecast horizon. An eighteen-member ensemble is

---

[27] Hinton diagrams summarise differences in skill (measured by Ranked or Continuous Ranked Probability Scores) between two ensemble forecasts for a range of fields such as temperature, wind speed, cloud amount, cloud base and precipitation for a given neighbourhood size.
[28] Receiver Operating Characteristic: ROC area gives us an indication of the ability of the forecasts to discriminate between events and non-events. The higher its value, the better the discrimination.
[29] Reliability Diagrams: evaluate how close the probability of an event is to the observed frequency of this event, so, the closer the points are to the diagonal the better the reliability.
[30] https://www.metoffice.gov.uk/about-us/press-office/news/corporate/2021/met-office-and-microsoft-announce-supercomputer-project

© Crown copyright 2022, Met Office

created by running three members every hour and time-lagging these over a six-hour window (Porson et al., 2020). Previously, the configuration was set up to run twelve members every six hours. Time lagging allows for the generation of more timely forecasts (reducing time and information constraints for operational meteorologists) from an increased ensemble size (that samples greater small-scale uncertainties in initial conditions) without additional computational costs (Porson et al., 2020).

Additionally, Porson et al. (2020) note that by increasing the time-lagged ensemble size by lagging over further additional cycles also leads to improvements, and these are larger (in most cases) than those obtained through neighbourhood processing. It is also noted that the larger ensemble size is critical in order for a time-lagged configuration to be as good as, or outperform, the previous configuration. The second main configuration change is the extension of forecast lead times to T+ 120 hours (five days), which allows for verification against the deterministic UKV, and the global ensemble, MOGREPS-G, for which MOGREPS-UK is shown to perform better out to T + 120 hours (Porson et al., 2020). Other benefits of the latest configuration are the utilisation of hourly 4D variational data assimilation (Milan et al., 2020) and a reduction in the jumpiness between forecasts times, as the impacts of new global perturbations and data assimilation analysis is slowly introduced over the cycles (Porson et al., 2020).

2.3 Post-processing

Wilks (2011b) notes that ensemble forecasts require post-processing because they contain errors in both statistical location (ensemble members are away from the actual state of the atmosphere but relatively near each other) and in dispersion (the ensemble can under or over represent forecast uncertainty). As noted in section 1.2, operational ensemble systems are often under-dispersive which, without correction, could lead to overconfident probability estimates. A common statistical post-processing technique applied to ensemble forecasts is Ensemble Model Output Statistics (EMOS). This includes regression and kernel density methods, which use historical error statistics to derive corrected ensemble distributions (Wilks, 2011b).

A widely used adaptive technique within NWP are Kalman filters, which is a statistical technique to improve the prediction of variables with a well-defined cycle, by combining observations and forecasts to minimise bias (Cassola and Burlando, 2012). In the ECMWF User Guide, an example is provided of an adaptive Kalman filter applied to all ensemble members to correct for systematic errors in 2m temperature, noting that a two- or multi-dimensional error equation can be used to correct for mean errors and systematic over- and under-forecasting of variability. Kalman filters are common in meteorological literature, used

© Crown copyright 2022, Met Office

operationally (i.e., for MOGREPS-UK in site specific forecasts) and continue to evolve in their development (and are utilised in other applications under alternative names). For example, Cassola and Burlando (2012) develop a method for applying a Kalman filter to wind speed, which is typically too irregular, and found that this approach, involving tuning the time step and forecast horizon, showed significant improvement in model output for wind speed forecast. It was found to be particularly effective in use for very short-term forecast applications, and has been successful during a two year case study test forecasting wind energy output for a wind farm, maintaining a low percentage error between simulated and measured wind energy (Cassola and Burlando, 2012).

Similarly, there are a number of machine learning techniques that can be used to learn and model systematic biases and then be applied to correct a forecast. These concepts are not new in themselves, however the continued improvement in computing power is facilitating the integration of these techniques into post-processing of numerical weather prediction systems (Loken et al., 2019). Examples of such techniques include Quantile Regression Forests (QRF), a nonparametric approach to estimate desired quantiles, used as an alternative method to traditional EMOS techniques that fit parametric distributions, and neural networks, which identify nonlinear relationships between arbitrary predictor variables and forecast distribution parameters. When applied to a thirty-five member ensemble forecast for surface temperature and wind speed, it was found that QRF performed better than EMOS at lead times up to T+54 hours (Taillardat et al., 2016). Evin et al. (2021) also found QRF as beneficial over a EMOS regression method in forecasting depth of new snow, due to the larger number of predictors.

In a case study application of ensemble forecasting for 2m temperature at surface stations in Germany, the neural network was found to significantly outperform the benchmark EMOS method, while being computationally more affordable (Rasp and Lerch, 2018). However, as machine techniques are data driven, they are reliant on the availability of sufficient historical data for all the variables over the area of interest in order to train a successful model. Brester at al. (2020) explored the use of three neural networks for predicting faults in the electricity network and found that multilayer perceptron[31] (a type of neural network) was most capable of accurately predicting events that would result in a large number of faults. However, the analysis found the optimum data input (the most important information relevant to the prediction) was the observational data of weather conditions six hours prior to the predicted fault, which may be of less value for timescales of response decision-making.

[31] https://www.sciencedirect.com/topics/computer-science/multilayer-perceptron

© Crown copyright 2022, Met Office

Over recent years, the Met Office has been developing IMPROVER (Integrated Model Post-Processing and Verification), a probabilistic post-processing system specifically for use in ensemble weather prediction models, that has integrated verification and open source code[32] for enabling wide use and collaboration. The purpose of IMPROVER is to provide ensemble forecast information in a format suitable for automated weather forecasts, for use by operational meteorologists and the public. It utilises the techniques described in this report so far; using spatial and temporal neighbourhood to increase forecast spread, incorporating topographic variation, statistical post-processing, use of older forecasts in data assimilation (time-lagging), and blending forecasts from multiple models to generate seamless probabilistic output from a consolidation of many potential weather outcomes[33].

[32] https://github.com/metoppv/improver
[33] https://improver.readthedocs.io/en/latest/about.html

© Crown copyright 2022, Met Office

3 Application of Ensemble Prediction Systems to impact forecasting

As explained in section 1.2, an ensemble forecast provides a range of possible outcomes that help users understand what weather events may occur and how likely a particular event will be[6]. The probabilistic output from ensemble forecasts can be used to support a number of tools, such as Public Weather Service (PWS) applications like weather warnings, forecasting weather patterns to support long range forecasting or be incorporated into PWS/customer Hazard Impact Models. This section provides examples of the use of ensembles in impact forecasting, which is non-exhaustive but aims to demonstrate a number of different applications.

The Met Office uses weather patterns, developed by Neal et al. (2016) using mean sea level pressure data from 1850 to 2003 over the UK and a surrounding European domain, in a probabilistic weather pattern forecasting tool called Decider[34] that automatically assigns forecasts from each ensemble member to the closest matching weather pattern, reducing the ensemble forecast to a sequence of circulation type probabilities. The weather patterns can be used in medium-range (out to fifteen days), monthly and seasonal forecasting. By understanding the climatological variability of weather patterns and their occurrence throughout the year, it is possible to identify common trends in pattern seasonality and as a result, unusual patterns at a particular time of year can be an indication of severe weather[35]. This makes it easier to interpret, as the characteristics of each weather patterns (e.g. their associated climatologies and impacts) are understood from evaluating occurrences in the historical period, and is particularly useful for assisting in the production of long-range forecasts. This work can be extended into customer Hazard Impact Models, when an expected impact is associated with a specific weather pattern, and therefore Decider can be used to summarise a broad range of probabilistic information that allows for quick, risk-based advice and decision-making.

Similarly, ensemble prediction systems can be utilised alongside user-defined parameterisation to allow for enhanced operational decision-making. An example of a Hazard Impact Model to support decision-making is the Vehicle OverTurning (VOT) model[36], that highlights areas of the road network at risk from vehicle overturning incidents to support the issuance of National Severe Weather Warning Service (NSWWS) wind warnings. The VOT model works by combining probabilistic wind forecast information with vulnerability and exposure data, which produces a forecast of vehicle overturning risk for the major trunk road

[34] https://www.metoffice.gov.uk/services/business-industry/energy/decider
[35] https://www.metoffice.gov.uk/research/news/2016/new-weather-patterns-for-uk-and-europe
[36] https://www.metoffice.gov.uk/research/news/2019/vehicle-overturning-model

© Crown copyright 2022, Met Office

network across Great Britain (Hemingway and Robbins, 2019). The hazard forecast probability for each vehicle type and forecast lead time is estimated by the proportion of ensemble members that exceed each wind gust threshold over a segment of the road network in MOGREPS-UK[36].

Another example is presented in Steele et al. (2021), where information about the past performance of the ECMWF ensemble prediction system in accurately predicting wave conditions up to 15 days ahead (based on user-orientated categorical verification statistics underpinned by a contingency table approach as explained in section 2.1) was combined with a bespoke parameterisation of the impact of adverse weather on viability of marine operations. Decision-making during the installation of a subsea pipeline in the North Sea was based on whether or not the relevant go/no-go ensemble probability threshold (i.e. the number of individual/constituent forecast members that must predict favourable/unfavourable conditions) elicited from this analysis was exceeded. This achieved a considerably greater value (benefit) than equivalent deterministic (single) forecasts or traditional climate-based options at all lead times, promising a more robust basis for effective planning than typically considered by the offshore industry, and demonstrating a practical approach for how such data are handled; similarly relevant for maximising the exploitation of any ensemble forecast.

Ensemble forecasts of rainfall are utilised in the Centre of Ecology and Hydrology's Grid-to-Grid model, which takes the probabilistic forecast and converts this into surface water runoff, to allow for a Hazard Impact Model for surface water flooding[37]. Cole at al. (2016) note that the use of high resolution models is important to capture convection dynamics, but also the requirement for the use of ensemble forecasts – in this case MOGREPS-UK – to capture spatio-temporal uncertainty. MOGREPS-UK is blended with extrapolated radar rainfall data and downscaled to 2km within the Short-Term Ensemble Prediction System (STEPS), producing a twenty-four member multi model ensemble. Hydrometeorologists use this information to assess flood risk and the potential impacts, supporting the issue of Flood Guidance Statements to emergency responders. This is one example of a model developed under the Hazard Impact Framework by the Natural Hazards Partnership (Hemingway and Gunawan, 2018), with the purpose of the Hazard Impact Modelling Framework to outline core concepts in impact modelling and provide standard guidelines and protocols for Hazard Impact Model development[38].

[37] http://www.naturalhazardspartnership.org.uk/science/hims/surface-water-flooding/
[38] http://www.naturalhazardspartnership.org.uk/hazard-impact-framework/

© Crown copyright 2022, Met Office

A recent publication by Wilkinson et al. (2022) utilises deterministic forecasting (Met Office UKV) to predict faults prior to an extreme weather event, in particular the threats to the electricity distribution network from approaching windstorms, with the purpose of allowing effective resource and response decision-making. The framework represents a method for translating the outputs of numerical weather prediction into damage and consequences associated with an extreme weather event, known as consequence forecasting, which provides a probabilistic estimate of the number and locations of failed assets as well as customers without power (Wilkinson et al., 2022). A number of quality control measures were implemented to eliminate errors in the historical UKV data, including the removal of systematic bias from a change in post-processing algorithms, and removal of outliers (temporally and spatially anomalous wind speeds). In this example, the probabilistic relationship is established using fragility curves (an analysis to determine the probability of reaching or exceeding a specified damage state i.e. wind speeds > 28 m/s that cause significant damage to assets). Differences between predicted and observed faults are due to uncertainty in the derived fragility curves rather than the forecast. This is a different approach to the use of an ensemble prediction system and categorical forecasts for probability thresholds outlined in Steele et al. (2021).

The Advanced Weather Forecasts for Dynamic Line Rating[39] delivered by the Met Office in collaboration with National Grid Electricity Transmission (NGET) looked to explore the use of NWP models as input to an overhead line (OHL) rating forecasting model, in order to generate ensemble forecasts of overhead line ratings that could be implemented in an operational dynamic line rating (DLR) system. An important aspect of the method used in this project was the retention of physical coherence between multiple weather parameters in order to end up with a physically justifiable distribution. For this model, the aim was to represent the plausible least favourable weather conditions for safe operation (i.e. the line would not be operated at a higher rating than this). Therefore, the most important part of the distribution is the lower tail, where it was found that the uncorrected ensemble approach produced a forecast no worse than the current operational deterministic approach, which signified an opportunity to get an overall improved forecast by applying post processing to ensemble forecasting to handle model biases. This work also looked to explore how the forecast distribution could be used in place of fixed error margins, reducing conservatism when confidence is high and add additional conservatism when confidence is low – highlighting an additional benefit of ensemble forecasting over deterministic. This was underpinned by a cost benefit analysis to

[39] https://www.nationalgrid.com/national-grid-partner-met-office-boost-capacity-electricity-transmission-network

© Crown copyright 2022, Met Office

explore how a DLR system with enhanced weather forecasts can deliver value for both customers and NGET.

The project explored a number of potential post processing techniques for the use in forecast system validation and optimisation. For example, a dependency on topography was shown during the forecast validation, and therefore the opportunity to use LIDAR techniques to derive terrain height and improve wind forecast in sheltered areas was identified. Machine learning methods such as quantile regression forest and neural networks were considered as a theoretically plausible replacement of the physical model for calculating the rating of a section of overhead line, identifying the potential in the future to generate rating forecasts directly from NWP models, although this is not currently operationally implemented. The additional use of machine learning techniques such as generalised linear models to learn and correct for model biases, could also improve the NWP forecast input feeding the machine learning rating forecast model.

A key aspect in all these applications is the importance of early and ongoing user engagement where there should be a prior consideration of how the user must or will respond when presented with a particular probability value. This will ensure maximum use of the ensemble to generate useful and relevant impact forecasts.

© Crown copyright 2022, Met Office

4 Synthesis and recommendations

Ensemble prediction systems generate forecasts indicating a range of possible outcomes, as opposed to a single 'best estimate' outcome, providing a number of advantages and further opportunities for analysis beyond that of a traditional deterministic forecast. The use of an ensemble forecast allows for the quantification of uncertainty. Ensembles can also benefit from post-processing and machine learning techniques to identify systematic bias and further improve forecast accuracy and ensemble performance. Additionally, it is demonstrated in the literature that the use of ensembles to generate probabilistic forecasts for impact modelling can add value to operational decision making across different sectors, including the energy sector.

Through the combination of ensemble forecasts (to sample a broader range of plausible outcomes), statistical post-processing (to improve upon forecast skill, calibration to truth and ensemble performance) and an understanding of impact thresholds for the power system (to create relevant and useful data), we have introduced concepts providing a basis for the prediction of weather-related faults that can support decision-making and reduce customer disruption. It is noted from conversations with the operational team at Scottish Power Energy Networks that decisions relating to staff and resource planning for fault response typically occur at shorter time leads running up to a forecasted impactful scenario (0-3 days), whilst awareness and preparation for a response is considered over the short-medium time horizons (5-15 days). As noted in section 1, different EPS provide forecasts over different time horizons and spatial resolutions, and therefore it may be appropriate to consider the use of multiple EPS, either individually or as a multi-model configuration, to achieve useful forecasts over the full time horizon of interest that utilise the highest resolution models (which is important for capturing the impact of topography and the behaviour of small scale processes as skilfully as possible). Further, it should be noted that the energy sector is too reliant on deterministic forecasts and adopting the use of probabilistic forecasts should lead to improved decision making.

Based on the available EPS over the UK, a number of recommendations are presented for exploration in the Alpha phase of the Predict4Resilience project. These recommendations take into consideration the requirements of the project, data availability and the conclusions of this literature review, with the aim to put forward a practical framework that will support desired operational outcomes, but also allow opportunities for innovative exploration of methods.

The recommendations for exploration in Alpha phase are as follows:

© Crown copyright 2022, Met Office

1. Two separate ensemble prediction systems, run in parallel, with differing time horizon focus. In this set up, it is suggested that an ensemble over the short lead time (up to 5 days) would be utilised in a risk threshold approach for operational decision making e.g. Steele et al. 2021 (ensemble 1.a), whilst a second ensemble would be utilised within a weather pattern approach for awareness of the developing broader scale conditions on a medium range time horizon (up to 15 days) e.g. Decider[34] (ensemble 1.b).
    a. **MOGREPS-UK / UKV** (Met Office) – The use of a high resolution 18-member ensemble generating hourly forecasts using time-lagging techniques. In literature, the current configuration of MOGREPS-UK was shown to perform better out to 5 days when compared to UKV, but pairing MOGREPS-UK with UKV is also recommended.

       Cons: An additional ensemble is required for forecasting beyond 5 days. Currently, there is no access to historical MOGREPS-UK / UKV forecasts, however access to this data for this type of project is in progress (reference data sheet).
    b. **ENS / HRES** (ECMWF) – ECMWF forecast user guidance notes that the use of the HRES deterministic model as part of the ensemble (with appropriate weighting), is considered the recommended approach when looking at the early-medium time horizon (however, ENS can be used in isolation and potentially allows for a more seamless approach if there is an interest in forecast horizons beyond ten days). The aim of this approach would be to create a seamless forecast output in which the HRES provides information on finer scale detail and holds more weighting in the 0-2 day horizons, whilst the ENS provides an indication of forecast condition out to medium horizons (up to 15 days) and supports quantification of uncertainty. ECMWF are focussed on forecasting in the medium-range and are the most sensible choice for ensemble forecasting over the UK on those time horizons.

       Cons: The spatial resolution of the HRES and ENS are 9km and 18km, respectively, compared to the higher resolution of MOGREPS-UK. Additionally, time availability in the Alpha phase may limit the ability to retrieve all of the required data for this approach. There is the potential that the spatial resolution of the ENS may be upgraded to a 9km configuration over the project

© Crown copyright 2022, Met Office

timescales, which is likely to influence the performance of the ENS (potentially improving the performance).

2. Blend models from two centres to create one seamless, probabilistic forecast over full time horizon. This forecast could then be utilised for either risk-based threshold decision-making or weather pattern analysis.
    a. **MOGREPS-UK / ENS** (Met Office/ECMWF) – Utilise techniques in literature, such as Beck et al. (2016) and Met Office IMPROVER, to blend models from different forecasting centres. The aim would be to produce one consistent, seamless forecast from 0-15 days, with relative weighting to models as appropriate.

       <u>Cons:</u> An additional layer of complexity to create an multi-model with improved skill over the individual EPS (introducing spread can increase error as well as increased post-processing requirements to seamlessly blend forecasts from different physical models with varying spatial resolution).

It is anticipated that the verification of forecast accuracy (forecast of meteorological variables rather than the faults themselves) could be undertaken using the Midas-open (temperature and wind), HadUK-Grid (precipitation) datasets and ERA5 (wind, taking advantage of the complete spatial coverage but recognising the limitations that exists with reanalysis wind fields). The importance of observations and the error associated with them should not be overlooked and testing against more than one dataset for historical forecast verification is recommended.

The Alpha phase should look to implement standard measures for verification as noted from the literature (such as Wilks, 2011a; Porson et al., 2020), both in terms of forecast accuracy and ensemble performance, with a particular consideration towards the use of neighbourhood approaches over point verification (section 2.2). Additionally, the literature has highlighted a number of post-processing techniques that can be utilised to improve both forecast skill and ensemble performance such as batch-processing using EMOS, Kalman filtering and machine learning. Given the active research in post-processing of ensemble forecasts, there is interesting scope to explore the implementation of adaptive and real-time processing methods that use the ensemble forecast spread to further improve upon the forecast skill and ensemble performance – a distinct advantage of ensemble over deterministic forecasting. However, the extent to which more involved techniques, such as machine learning, could be explored in Alpha will depend on the availability of suitable training data over the region of interest and may be beyond the timescales and scope of this project phase.

© Crown copyright 2022, Met Office

More generally, the above highlights the recommended approach that would be taken in an Alpha phase. It does not lay out a full methodology, or consider the timescales needed for the recommended approach (and indeed whether the ideal approach would fit into the timescale of the Alpha phase). However, it serves as a basis to develop a more detailed scope, and some of the key areas that require consideration in the development of this scope.

© Crown copyright 2022, Met Office

5 References

Beck, J., Bouttier, F., Wiegand, L., Gebhardt, C., Eagle, C. and Roberts, N., 2016. Development and verification of two convection-allowing multi-model ensembles over Western Europe. Quarterly Journal of the Royal Meteorological Society 142, pp2808-2826. https://doi.org/10.1002/qj.2870

Brester, C., Niska, H., Ciszek, R. and Kolehmainen, M., 2020. Weather-based Fault Prediction in Electricity Networks with Artificial Neural Networks. IEEE Congress on Evolutionary Computation (CEC), pp. 1-8. https://doi.org/10.1109/CEC48606.2020.9185555

Cassola, F. and Burlando, M., 2012. Wind speed and wind energy forecast through Kalman filtering of Numerical Weather Prediction model output. Applied Energy 99, pp154-166. http://dx.doi.org/10.1016/j.apenergy.2012.03.054

Cole, S., Moore, R., Wells, S. and Mattingley, P., 2016. Real-time forecasts of flood hazard and impact: some UK experiences. E3S Web of Conferences 7, p18015. https://doi.org/10.1051/e3sconf/20160718015

Dey, S., Leoncini, G., Roberts, N., Plants, R., Migliorini, S., 2014. A spatial view of ensemble spread in convection-permitting ensembles. Monthly Weather Review 142, pp4091–4107. https://doi.org/10.1175/MWR-D-14-00172.1

Dey, S., Plant, R., Roberts, N., Migliorini, S., 2016a. Assessing spatial precipitation uncertainties in a convective-scale ensemble. Quarterly Journal of the Royal Meteorological Society 142, pp2935–2948. https://doi.org/10.1002/qj.2893.

Dey, S., Roberts, N., Plant, R., Migliorini, S., 2016b. A new method for the characterization and verification of local spatial predictability for convective-scale ensembles. Quarterly Journal of the Royal Meteorological Society 142, pp1932–1996. https://doi.org/10.1002/qj.2792

Evin, G., Lafaysse, M., Taillardat, M., and Zamo, M., 2021. Calibrated ensemble forecasts of the height of new snow using quantile regression forests and ensemble model output statistics, Nonlinear Processes in Geophysics 28, pp467–480. https://doi.org/10.5194/npg-28-467-2021

Hagelin, S., Son, J., Swinbank, R., McCabe, A., Roberts, N. and Tennant, W., 2017. The Met Office convective-scale ensemble, MOGREPS-UK. Quarterly Journal of the Royal Meteorological Society 143, pp2846-2861. https://doi.org/10.1002/qj.3135

Hemingway, R. and Gunawan, O., 2018. The Natural Hazards Partnership: A public-sector collaboration across the UK for natural hazard disaster risk reduction. International Journal of Disaster Risk Reduction 27. https://doi.org/10.1016/j.ijdrr.2017.11.014

Hemingway, R. and Robbins, J., 2019. Developing a hazard-impact model to support impact-based forecasts and warnings: The Vehicle OverTurning (VOT) Model. Meteorological Applications 27. https://doi.org/10.1002/met.1819

Kendon, E., Short, C., Pope, J., Chan, S., Wilkinson, J., Tucker, S., Bett, P., Harris, G., 2021. Update to UKCP Local (2.2km) projections. Technical report. https://www.metoffice.gov.uk/pub/data/weather/uk/ukcp18/science-reports/ukcp18_local_update_report_2021.pdf

Loken, E., Clark, A., McGovern, A., Flora, M. and Knopfmeier, K., 2019. Postprocessing Next-Day Ensemble Probabilistic Precipitation Forecasts Using Random Forests. Weather and Forecast 34, pp2017-2044. https://doi.org/10.1175/WAF-D-19-0109.1

© Crown copyright 2022, Met Office

Milan, M., Macpherson, B., Tubbs, R., Dow, G., Inverarity, G., Mittermaier, M., Halloran, G., Kelly, G., Li, D., Maycock, A., Payne, T., Piccolo, C., Stewart, L. and Wlasak, M., 2019. Hourly 4D-Var in the Met Office UKV operational forecast model. Quarterly Journal of the Royal Meteorological Society 146, pp1281-1301. https://doi.org/10.1002/qj.3737

Mittermaier, M., 2014. A Strategy for Verifying Near-Convection-Resolving Model Forecasts at Observing Sites. Weather and Forecasting 29, pp185-204. https://doi.org/10.1175/WAF-D-12-00075.1

Murphy, A., 1993. What is a good forecast? An essay on the nature of goodness in weather forecasting. Weather Forecasting 8, pp281-293. https://doi.org/10.1175/1520-0434(1993)008%3C0281:WIAGFA%3E2.0.CO;2

Neal, R., Fereday, D., Crocker, R. and Comer, R., 2016. A flexible approach to defining weather patterns and their application in weather forecasting over Europe. Meteorological Applications 23, pp389-400. https://doi.org/10.1002/met.1563

Pelosi, A., Medina, H., Van den Bergh, J., Vannitsem, S. and Chirico, G., 2017. Adaptive Kalman Filtering for Postprocessing Ensemble Numerical Weather Predictions. Monthly Weather Review 145, pp4837-4854. https://doi.org/10.1175/MWR-D-17-0084.1

Porson, A., Carr, J., Hagelin, S., Darvell, R., North, R., Walters, D., Mylne, K., Mittermaier, M., Willington, S. and Macpherson, B., 2020. Recent upgrades to the Met Office convective-scale ensemble: An hourly time-lagged 5-day ensemble. Quarterly Journal of the Royal Meteorological Society 146, pp3245-3265. https://doi.org/10.1002/qj.3844

Rasp, S. and Lerch, S., 2018. Neural networks for postprocessing ensemble weather forecasts. Monthly Weather Review 146, pp3885-3900. https://doi.org/10.1175/MWR-D-18-0187.1

Steele, E., Neal, R., Bunney, C., Evans, B., Fournier, N., Gill, P., Mylne, K. and Saulter, A., 2017. Making the Most of Probabilistic Marine Forecasts on Timescales of Days, Weeks and Months Ahead. Offshore Technology Conference. https://doi.org/10.4043/27708-MS

Steele, E., Brown, H., Bunney, C., Gill, P., Mylne, K., Saulter, A., Standen, J., Blair, L., Cruickshank, S. and Gulbrandsen, M., 2021. Using Metocean Forecast Verification Information to Effectively Enhance Operational Decision-Making. Offshore Technology Conference. http://dx.doi.org/10.4043/31253-MS

Taillardat, M., Mestre, O., Zamo, M. and Naveau, P., 2016. Calibrated Ensemble Forecasts Using Quantile Regression Forests and Ensemble Model Output Statistics. Monthly Weather Review 144. http://dx.doi.org/10.1175/MWR-D-15-0260.1

Wilkinson, S., Dunn, S., Adams, R., Kirchner-Bossi, N., Fowler, H., Otalora, S., Pritchard, D., Mendes, J., Palin, E., Chan, S., 2022. Consequence forecasting: A rational framework for predicting the consequences of approaching storms. Climate Risk Management 35. https://doi.org/10.1016/j.crm.2022.100412

Wilks, D., 2011a. 'Chapter 8: Forecast Verification' in Statistical Methods in the Atmospheric Sciences. DOI: 10.1016/B978-0-12-385022-5.00008-7

Wilks, D., 2011b. 'Chapter 7: Statistical Forecasting in Statistical Methods in the Atmospheric Sciences. DOI: 10.1016/B978-0-12-385022-5.00008-7

© Crown copyright 2022, Met Office