Predict4 Resilience: Sample Data and Exploratory Analysis

Discovery Phase: University of Glasgow Deliverable Report

Prepared by: Dr Jethro Browell
Date: 18/04/2022

Executive Summary

The Predict4Resilience project aims to build a new fault prediction capability to provide actionable forecasts of network faults to DNO control rooms, enabling more confident decision-making and reductions in customer minutes lost due to faults. This report surveys fault prediction methodologies reported in literature from academia and industry-led innovation projects. It also presents an exploratory analysis of weather and fault data for two DNO regions operated by SPEN, and a proof-of-concept forecasting methodology.

There is an extensive literature on the resilience of electricity networks to adverse weather, and naturally this has been extended to predicting future weather impacts on both operational and planning timescales. The literature on short-term (days-ahead) fault prediction, however, is small. Existing capabilities in this space are limited to forecasting on day-ahead timescales, and do not quantify forecast uncertainty. However, they provide strong evidence that weather-driven fault prediction is possible, and methodologies that are likely extendable to both longer-range and probabilistic forecasting. The following proof-of-concept has been informed by these methods although not all aspects could be explored in the limited time available in the Discovery phase of Predict4Resilience, such as incorporation of high-resolution weather forecasts or vegetation data.

Fault and weather data have been gathered, analysed, and employed in a proof-of-concept fault forecasting model for wind-related faults. The relationship between wind speed, direction and faults is apparent, but significant variation is present representing a challenge for predictive modelling. Nevertheless, a statistical model for describing the possible number of faults in a day for a given district has been developed and evaluated. It is found that faults can be predicted skilfully from zero to five-days-ahead, the key horizon for operational decision-making, beyond which performance declines. There is still some skill in forecasts up to 14-days-ahead, and while this may provide some situational awareness is unlikely to inform any operational decisions. These results are encouraging and provide insight and direction for future developments to improve the quality and usefulness of weather-related fault predictions in future phases of Predict4Resilience.

Contents

Change Log

Version/purpose	Contributor	Changes	Date
V1/First draft	Jethro Browell		13/04/2022
V2/Revision for delivery	Jethro Browell	Corrections and edits following review by JO. Addition of appendix detailing supplementary data.	28/04/2022

Author contact details
Dr Jethro Browell
School of Mathematics and Statistics
University of Glasgow
132 University Place
Glasgow, G12 8QQ
jethro.browell@glasgow.ac.uk

1. Introduction

This deliverable report is the product of the Discovery Phase of the *Predict4Resilience* project by the University of Glasgow. It should be read in conjunction with associated reports produced by the Met Office and ARUP and will be combined with these into a single cohesive document. This document provides a literature review of fault prediction capabilities to supplement the literature review produced by the Met Office, which focuses on weather forecasting technology, plus a description and exploratory analysis of sample data gathered during this discovery phase.

The primary aim of this work is to provide a "proof of concept" for the fault prediction tool that is envisioned by Predict4Resilience. Relevant literature has been examined to determine if similar capabilities have been developed previously, either in academic or commercial settings. The sample dataset comprising weather and fault data specific to UK distribution networks has been compiled and analysed to establish if standard prediction techniques are effective in this setting. Additionally, this literature review and exploratory analysis will inform discussion of priority areas for further development when aiming to produce the most accurate fault predictions possible in later phases of the project. The sample data and proof-of-concept predictions are supplied along with this report.

Exploratory analysis focusing on wind-related faults on the high-voltage and medium voltage distribution networks operated by SPEN has been carried out. Wind-rated faults were chosen because they are the most prevalent weather-related cause of faults and were identified as a priority by the SPEN Control Room. Other types of fault will be considered in future phases of the project. Furthermore, faults are predicted at the *District* level (as defined by SPEN) as this is the level on which associated decisions are based, such as allocation of engineering teams and other resources. The proceeding analysis focuses on establishing whether it is reasonable to predict wind-related faults in this way. It does not seek to develop the most accurate method for doing so, and the reported predictive performance should be considered a minimum that is possible that will almost certainly be possible to improve upon in future phases of the project. The report concludes with suggestions for future developments to that end.

2. Supplementary literature review

This literature review is supplemental to that included in the associated Predict4Resilience report produced by the Met Office, which reviews relevant literature with a focus on ensemble NWP. Here, literature on weather-related fault prediction and forecasting is surveyed. Overall, the literature on this topic is not extensive suggesting that weather-related fault prediction for electricity networks is either not a mature capability, or not seen as a priority for research; some combination is probably true. Several examples of desk-based academic studies have been reviewed plus one example of a well-developed fault prediction capability from the USA. The literature suggests that fault prediction is challenging but possible. It is apparent from broader discussions with project partners that conventional weather forecasts combined with expert judgement have been sufficient in practice to date, but that there is an appetite to develop more sophisticated capabilities.

The most comprehensively reported fault prediction capability is the University of Connecticut's "Outage Prediction Model". This model is used by industry and predicts faults in north-eastern USA on a 4km grid during storms. It is driven by data on infrastructure, soils, topography, land cover, vegetation, and weather characteristics (Watson et al. 2020; Cerrai et al. 2019). They employ a custom Numerical Weather Prediction (NWP) model based on the open-source Weather Research and Forecasting model (WRF), and two decision tree-based machine learning models: random forests, and Bayesian additive regression trees. The model predicts on outages, defined as any incident requiring a repair crew, during extra-tropical storms only. It is found that the errors in the weather forecast are only a small contribution to errors in fault prediction. The major source of error is weather-to-fault modelling. An extension considering thunderstorms has also been developed (Alpay et al. 2020).

Faults are only modelled and predicted during 48h windows containing a storm, and input features, such as number of hours the wind speed exceeds some threshold, vegetation type, density of development, are engineered as is typical for use in tree-based machine learning methods. A schematic overview of the methodology is shown in Figure 1. This approach produces point predictions of the number of faults with an error rate of 59% (mean absolute percentage error). It is unknown how the model would perform during weather conditions severe enough to produce weather-related faults but not classified as storms.

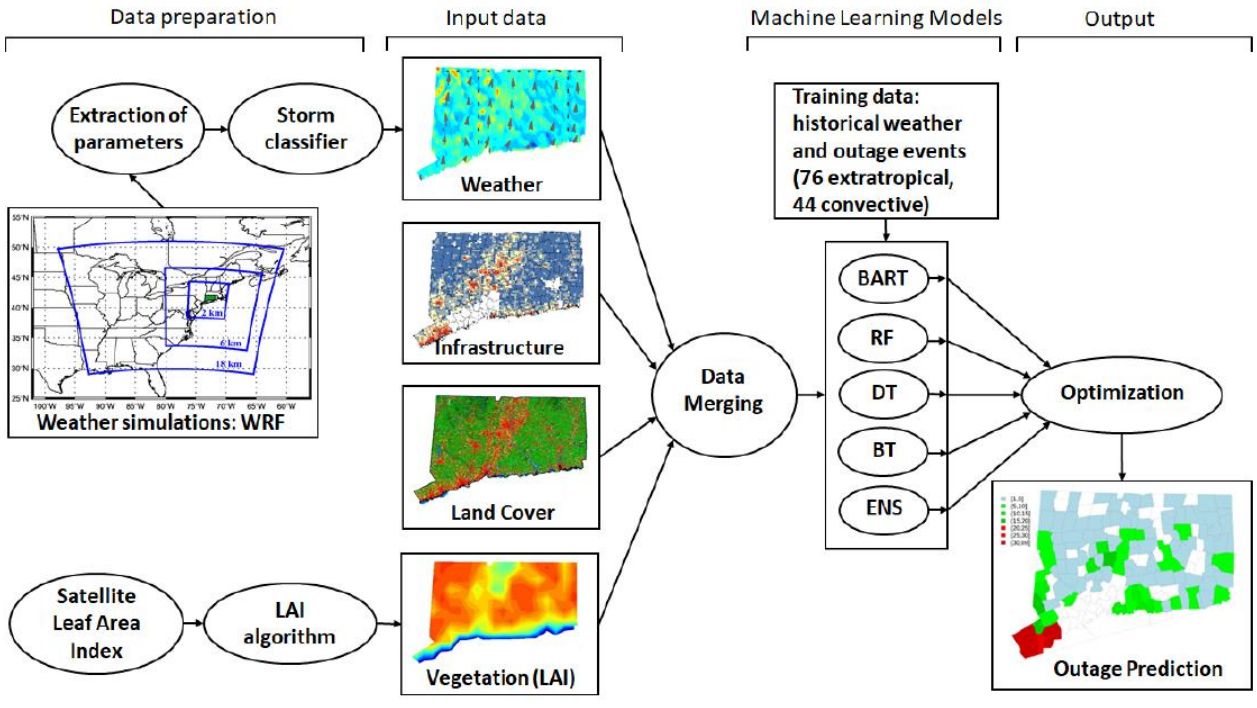

Figure 1: Schematic of the "Outage Prediction Model" from (Cerrai et al. 2019).

In addition to the Outage Prediction Model described above, IBM offer an outage prediction capability up to 72 hours ahead, although technical details are not available in the public domain "IBM Environmental Intelligence Suite: Outage Prediction" ("IBM Environmental Intelligence Suite - Outage Prediction | IBM" n.d.).

Others have developed similar albeit less sophisticated models for other regions, e.g. Finland (Brester et al. 2020), and in the UK in (Tsioumpri, Stephen, and McArthur 2021; Wilkinson et al. 2022) and as part of a 2016 NIA project (UKPN 2016). These are all desk-based studies, not operational tools. The latter stablished a correlation between weather forecasts of precipitation, wind speed, and lightning and number of faults, and proposed simple statistical models for predicting numbers of faults for UKPN's network in the Southeast of the UK. However, the accuracy of lightning forecasts was found to be poor compared to other weather variables. Uncertainty quantification was not considered. Ice accretion was found to not be a significant factor for UKPN's network.

The impact and "consequence" of faults caused by windstorms are considered in Wilkinson et al. 2022 where both number of faults and customer interruptions are predicted 24h ahead with uncertainty quantification. The authors note that 24h ahead predictions and 0h ahead (proxy observations) are of similar accuracy, suggesting that the weather-to-fault model, fragility curves in this work, is a greater source of uncertainty than that of the short-term weather forecast.

Forecasting weather impact on greater horizons is considered in (Brayshaw et al. 2020) for the case of telecoms infrastructure. It was found that sub-seasonal-to-seasonal weather forecasts could be leveraged to produce skilful forecasts of the weekly fault rate in winter three to four weeks ahead. Such methods are likely transferable to energy networks, though the specificity in time (weekly average faults, several weeks ahead) may not be relevant to current fault-management practices but could yield other benefits to DNOs.

In summary, prediction of weather-related faults is not a new concept but has received some attention in the academic literature, and by energy network operators. Published results and the existence of commercial offerings verifies that fault prediction is possible, but extent capabilities tend to be specific to certain locations/networks, and short forecast horizons of only a few days. The extent to which existing methods will generalise to new locations/network and longer forecast horizons is unknown. Furthermore, quantification of forecast uncertainty has also largely been overlooked, limiting the usefulness of these methods for risk management.

3. Sample dataset

Three types of data have been combined to explore the predictability of wind related faults:

1. Number of wind-related faults in an hour (approx. 2010-2021, prepared by ARUP)
2. Actual weather at the time and approximate location of faults
3. Ensemble numerical weather predictions (NWP) for the time and approximate location of faults

The preparation of fault data is detailed in the associated report by ARUP and is not repeated here. Two sources of actual weather data have been considered: meteorological mast measurements from the MIDAS dataset and the ERA5 reanalysis data. Ensemble NWP produced by the European Centre for Medium Range Weather Forecasting (ECMWF) have been extracted from the TIGGE archive and are used here. Details of these are provided below.

While data relevant to snow/ice and precipitation/flooding have been extracted, they are not investigated in the Discovery phase which had prioritised wind-related faults. These other types of fault and associated data, including additional data relevant to wind-related faults such as elevation and vegetation, will be considered in future phases of the project.

Actual Weather

The MIDAS[1] dataset includes multiple weather stations measuring relevant variables, and most faults (with associated location data) occurred within 20km of a weather station. However, not all weather stations have full coverage of the relevant time period (2010-2021), and the irregular spatial distribution of weather stations and proximity to network assets complicates analysis and modelling of the relationship between weather and faults.

The ERA5[2] reanalysis on the other hand provides consistent estimates of actual weather on a regular grid covering the full spatial and temporal extend of the fault dataset. Therefore, the proceeding analysis and sample data is based on ERA5 only. The weather parameters extracted are listed in Table 1. Data are stored in netCDF format with one file per month from 2010-2021, totalling approximately 3.5GB. Parameters relevant to the top three weather-related causes of faults (wind, precipitation, snow/ice) have been downloaded, although analysis in the Discovery phase focuses only on the more prevalent cause, wind.

Table 1: ERA5 weather parameters contained in the sample dataset. Additional details available from the Climate Data Store[2]

Parameter	Description
100m u component of wind	Zonal component on wind vector (from west) 100m above surface [ms^{-1}]
100m v component of wind	Meridional component on wind vector (from north) 100m above surface [ms^{-1}]
10m v component of wind	Zonal component on wind vector (from west) 10m above surface [ms^{-1}]
10m u component of wind	Meridional component on wind vector (from north) 10m above surface [ms^{-1}]
Instantaneous 10m wind gust	Maximum 3 second gust 10m above surface [ms^{-1}]
2m temperature	Air temperature 2m above surface [K]
2m dewpoint temperature	The temperature to which the air, 2 metres above surface, would have to be cooled for saturation to occur. It is a measure of the humidity. [K]
Total precipitation	Hourly average precipitation rate [kg m^{-2} s^{-1}]
Snowfall	[m of water equivalent]
Snow depth	[m of water equivalent]
Volumetric soil water layer 1	[m^3m^{-3}] Layer 1 is depths from 0 to 7cm, the surface is at 0cm
Volumetric soil water layer 2	[m^3m^{-3}] Layer 2 is depths from 7cm to 28cm, the surface is at 0cm

[1] https://catalogue.ceda.ac.uk/uuid/dbd451271eb04662beade68da43546e1

[2] https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels

Ensemble numerical weather prediction

Historic ensemble forecasts issued by ECMWF have been downloaded from the TIGGE database, which stores ensemble forecast from multiple weather centres for research purposes. This archived data only contains a subset of parameters and is stored at lower resolution than the original but was still produced by the higher resolution operational model in use at the time of the forecast origin. All parameters of the original full resolution data is available but is much slower to retrieve from the MARS archive[3], hence opting to use TIGGE here. The four-year dataset compiled here from TIGGE took approximately two months to retrieve from the MARS archive. Forecasts issued at midnight each day from 2018-2021, from 0h to 360h (15 days) ahead have been collected. The data are on a 0.5^0 x 0.5^0 x 6h grid. The parameters retrieved are listed in Table 2. Some potentially useful parameters are not available in TIGGE, such as 100m winds and wind gust. Approximately 30 days' forecasts are missing due to data storage tapes being damaged. Data are stored in netCDT format with one file per issue time, totalling approximately 26GB.

Table 2: ECMWF ensemble NWP parameters. Additional details available from ECMWF[4].

Parameter	Description
10m v component of wind	Zonal component on wind vector (from west) 10m above surface [ms^{-1}]
10m u component of wind	Meridional component on wind vector (from north) 10m above surface [ms^{-1}]
2m temperature	Air temperature 2m above surface [K]
2m dewpoint temperature	The temperature to which the air, 2 metres above surface, would have to be cooled for saturation to occur. It is a measure of the humidity. [K]
Total precipitation	Hourly average precipitation rate [kg m^{-2} s^{-1}]
Snowfall	[m of water equivalent]
Snow depth	[m of water equivalent]
Soil Moisture	[kg m^{-3}]

[3] https://www.ecmwf.int/assets/elearning/mars/mars1/story_html5.htmlhttps://www.ecmwf.int/assets/elearning/mars/mars1/story_html5.html

[4] https://apps.ecmwf.int/codes/grib/param-db/

4. Exploratory analysis

The sample dataset has been analysed and a proof-of-concept fault forecasting system implemented and evaluated. The objective of this analysis is to:

1. evaluate the skill in ensemble NWP over relevant regions,
2. determine if the relationship between weather and wind-related faults can be learned from historic data, and
3. combine a model for weather-related faults with weather forecasts to produce a fault forecast.

The modelling work presented here is preliminary and it is likely that all aspects could be improved. In all stages of modelling, as simple an approach as possible has been taken to prove concepts. Producing the most accurate predictions possible will be an objective of future phases of the Predict4Resilience project.

4.1. Weather-fault model

In the Discovery phase of Predict4Resilience we focus on wind-related faults, as these are by fare the most prevalent. Control room staff have developed an understanding of this relationship over years of experience, including key combinations of wind speed and direction in certain locations that they expect to result in high numbers of faults. Combining and visualising the weather (EAR5 reanalysis) and fault data introduced above, this relationship is clear.

Faults are most prevalent when wind speeds are high (daily average greater than 40mph) and from the Southwest. The extent to which the directional trend is due to this being the prevailing wind direction over the UK or an affect related to the orientation of power lines has not been explored. This is illustrated below by plotting number of daily faults against wind speed and direction in Figure 2. While there is a clear correlation between wind speed and number of faults, there are also many occasions when wind-related faults occur with relatively modest wind speeds, posing a potential challenge for their prediction. Nevertheless, here we will identify a suitable modelling framework for wind-related faults and explore how well they may be predicted based on a statistical model of the weather-fault relationship and weather forecasts.

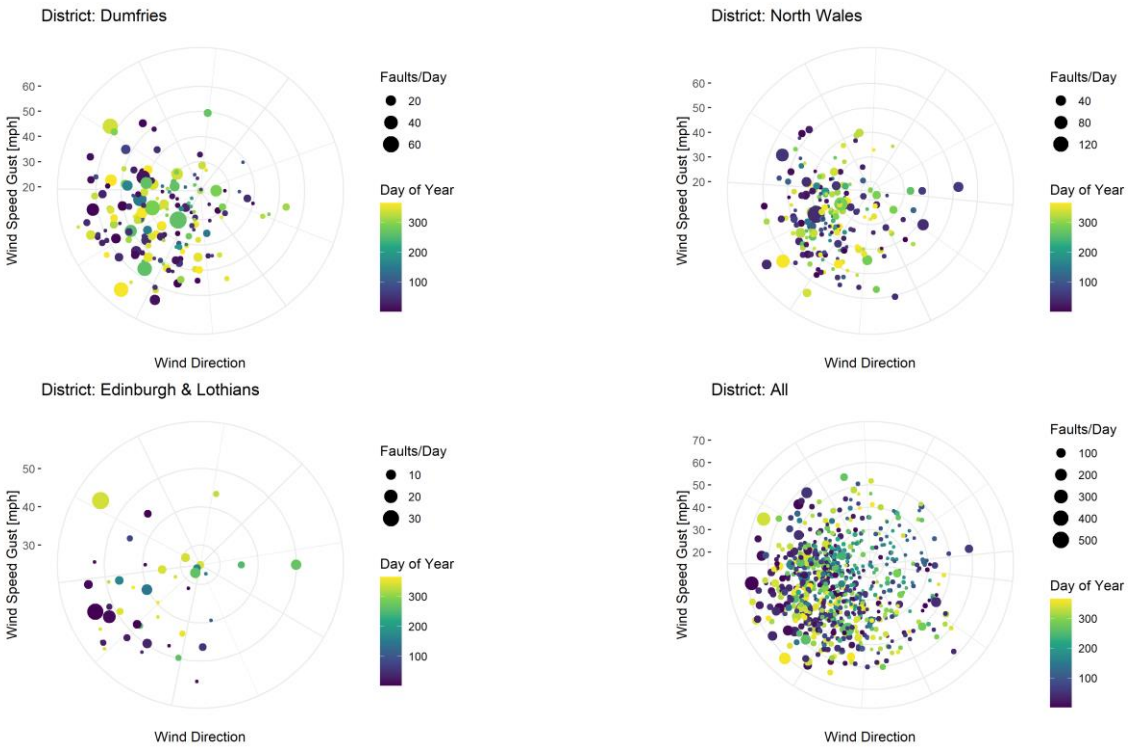

Figure 2: Wind-related faults per day by wind speed, direction, and time of year. Only days with three or more faults are included.

Of the 13 districts considered, Dumfries and North Wales are the most prone to wind-related faults with substantially more than others. These areas are large geographically making them challenging to manage from a maintenance perspective and are therefore of greatest priority for fault prediction. Smaller districts containing major cities exhibit fewer wind-related faults and are therefore of lower priority here.

Uncertainty quantification is a requirement of the fault forecasting tool being developed in this project and is therefore built-in to every stage in the modelling process. The model of weather-related faults developed here produces prediction in the form of a probability distribution for the number of faults given weather conditions, as opposed to predicting a single number. The range of possible outcomes and associated probabilities quantifies the uncertainty associated with this process.

This task requires the parameters of a probability distribution to be predicted and is called distributional regression. The class of Generalised Additive Models for Location Scale and Shape, and associated implementation in software, provide a suitable framework for this work (Rigby and Stasinopoulos 2005). In this setting, the parameters $\theta_k, k = 1, \dots, 4$, of a probability distribution are modelled as additive models of covariates.

The Zero-adjusted Zipf distribution has been identified as a good candidate for modelling the occurrence of wind-related faults, which is heavy-tailed count data, and poorly modelled by more familiar count distributions such as the (zero-adjusted) Poisson and Negative Binomial. Zero-adjustment refers to the necessity to handle occurrences of zero faults, and the resulting distribution is a mixture of σ *'the probability of zero faults*' and $\mathrm{Zipf}(\mu)$ *'the distribution of number of faults given there are greater than zero faults'*, controlled by parameter μ. Both σ and μ may be modelled as additive functions of covariates describing the weather. The model for number of faults y_t at time t may be written

$$P(y_t|\mu_t, \sigma_t) = \sigma_t \delta(y_t) + \big(1 - \delta(y_t)\big)(1 - \sigma_t)\frac{y_t^{-(\mu_t+1)}}{\zeta(\mu_t + 1)} \tag{1}$$

where ζ is the Reimann Zeta function and $\delta(\cdot)$ is the Dirac delta function, $\delta(y_t = 0) = 1$ and $\delta(y_t \neq 0) = 0$, and,

$$\mathrm{logit}(\sigma_t) = \beta_0 + \boldsymbol{x}_{t,0}\boldsymbol{\beta} + \sum_{j=1}^{J} h_j(\boldsymbol{x}_{t,j}) \tag{2}$$

$$\log(\mu_t) = \mu_0 \tag{3}$$

where β_0, $\boldsymbol{\beta}$, and μ_0 are coefficients to be estimated, $\boldsymbol{x}_{t,j}$ are known vectors of explanatory variables, and $h_j(\cdot)$ are flexible basis functions, such as P-splines, also to be estimated.

An important aspect of model specification is the choice of explanatory variables $\boldsymbol{x}_{t,j}$ and basis function. In this exploratory analysis, simple weather-based features have been created for each district to serve as explanatory variables, but this is a particular area for further development. The features used here are:

1. mean wind direction over the district and day,
2. daily average of the maximum hourly 10m wind speed over the district, and
3. the previous two days' values of #2.

P-splines of #2 (with three degrees of freedom) are used, as are sin and cosine terms of #1 to handle the circular nature of wind direction.

The models are fit using weather features derived from ERA5 to predict corresponding wind-related faults, as identified by ARUP. Fault data are aggregated to daily totals, and the daily average of weather features are used for the purpose of this proof-of-concept. This was necessitated by computational burden and the limitation of the available ensemble NWP being 6h resolution. Refining this feature engineering step using higher resolution NWP in combination with asset metadata (elevation, proximity to vegetation etc) could be an easy wind for future model improvement in the next phase of Predict4Resilience. Model parameters are

estimated by maximum-likelihood using numerical methods implemented in the *gamlss* R package (Rigby and Stasinopoulos 2005).

The goodness-of-fit for the probability model described in Equations (1)-(3) is examined via diagnostic plots called "worm plots" (van Buuren and Fredriks 2001). These illustrate the difference between two distributions, conditional on an explanatory variable. Here, we compare the actual and predicted distribution of faults conditional on wind speed. The worm plot for the Dumfries model is shown in Figure 3, which it typical of all Districts. For all wind speed ranges, with predicted distribution is a reasonable fit for the central and left tail of the distribution, but the right tail is "light", meaning that the model tends to underestimate the probability of large numbers of faults occurring. While the fit is not perfect, it is adequate to prove the concept of probabilistic fault prediction and is an area for improvement in future phases of the project.

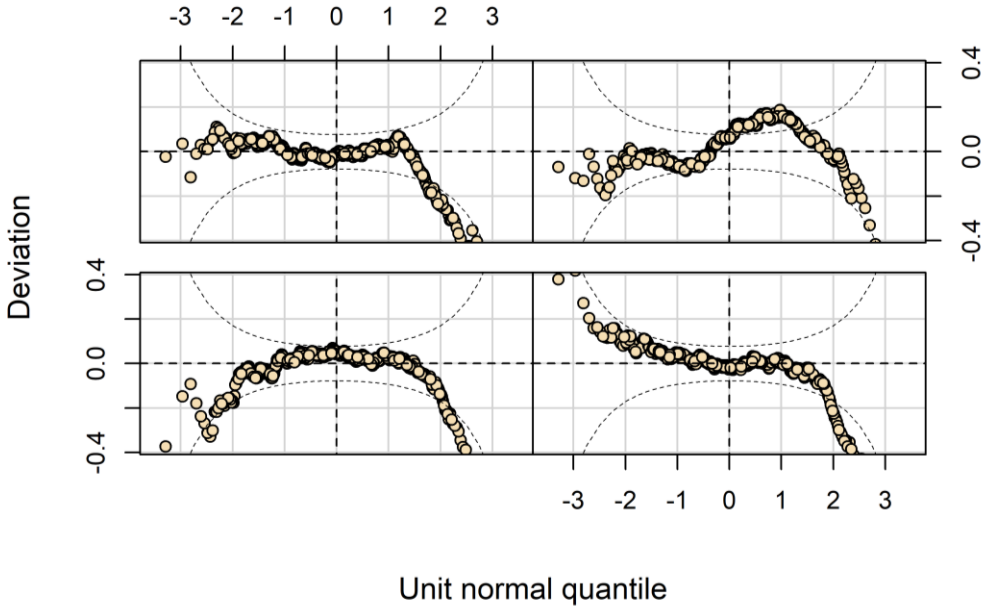

Figure 3: Worm plot for the Dumfries weather-fault model conditional on wind speed quantile (ordered low to high, top left to bottom right). A worm plot is a de-trended QQ-plot and highlights deviation from nominal calibration of probabilistic predictions. Dashed "U" shaped lines indicate 95% consistency intervals. Circles are standardised residuals. Residuals falling outside the consistency intervals indicate inadequate fit in regions of the distribution and value of the conditioning variable. Here, the right tail of the predictive distribution is too "light"; the model will therefore tend to underestimate the probability of large numbers of faults occurring.

The predictions produced by the fault models given actual weather have been visualised to give an impression of how probabilistic fault forecasts may be presented, and a more initiative indication of model performance. The predictions from all models spanning 2010 to 2021 are provided as interactive HTML plots along with this report. A static example of the year centred on winter 2014 for Dumfries is provided in Figure 4. Two elements of the forecast are visualised, the probability of there being zero faults, and the possible number of faults.

The probability of zero faults is communicated via a traffic-light system. There is a clear correlation between days where the probability of faults is "amber" or "red" and the occurrence of faults. Furthermore, the chance of high numbers of faults is illustrated by spikes in the 99th and 95th quantiles of the fault forecast. These too correlate well with actual numbers of faults, although there are several false positives.

In practice, users may set thresholds based on probability and number of faults to define warnings, alarms or similar, and should do so with properties such as false positive rate in mind. It is anticipated that improvements beyond this proof-of-concept will improve the specificity of forecasts – both in terms of timing and number of faults.

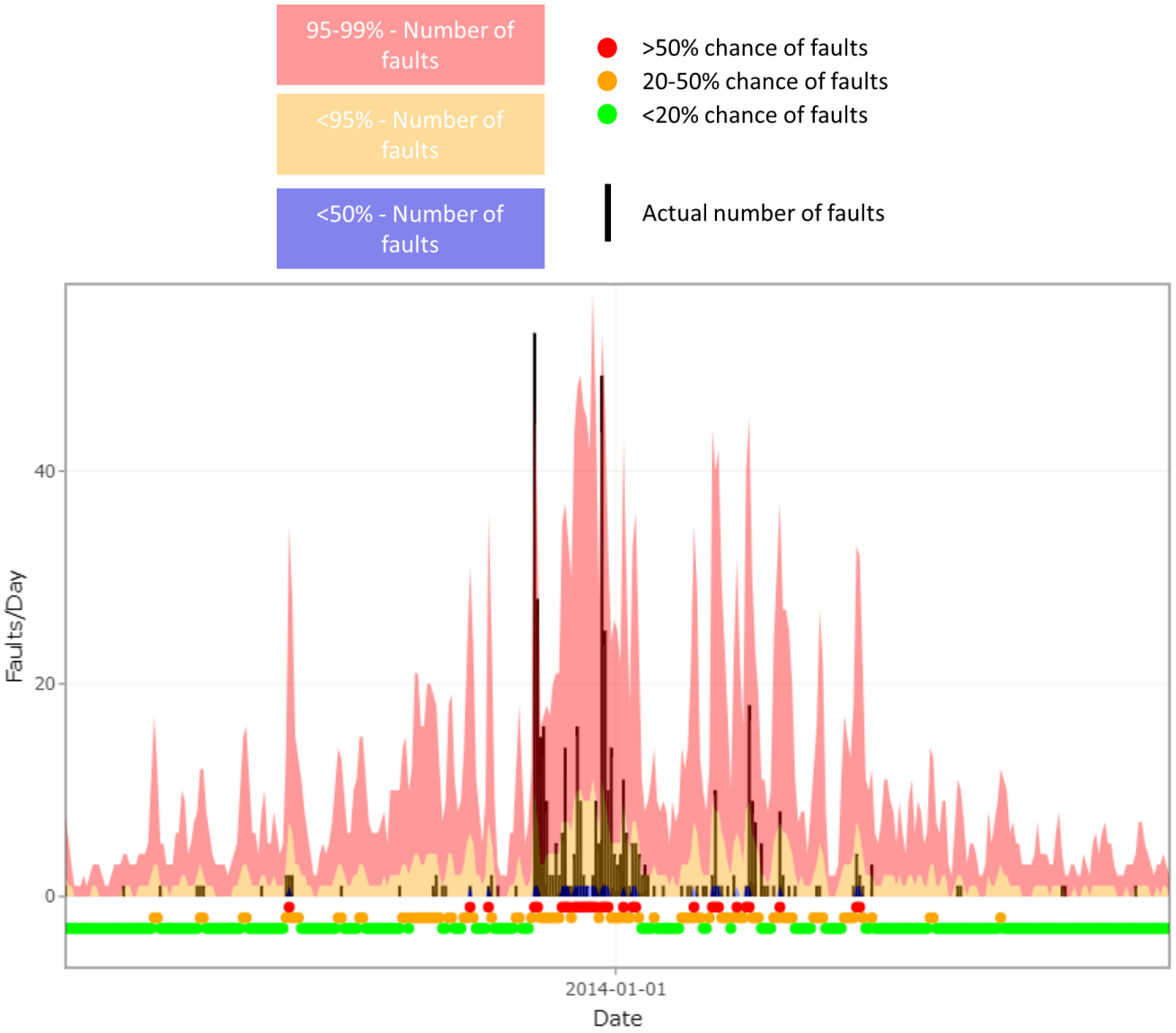

Figure 4: Daily fault predictions (given actual weather) for Dumfries cantered on winter 2014. A key is provided above the main figure.

4.2. Ensemble NWP Calibration

The ensemble NWP considered here comprises 50 members, and two properties of the ensemble are evaluated, and where necessary calibrated. First the forecast bias, and second the ensemble dispersion. Bias is evaluated by comparing the ensemble mean to observed weather. Max wind speed by district derived from ERA5 is the target variable or "truth" here as this is the key input to the fault prediction models. An example of an ensemble weather forecast is provided in Figure 5.

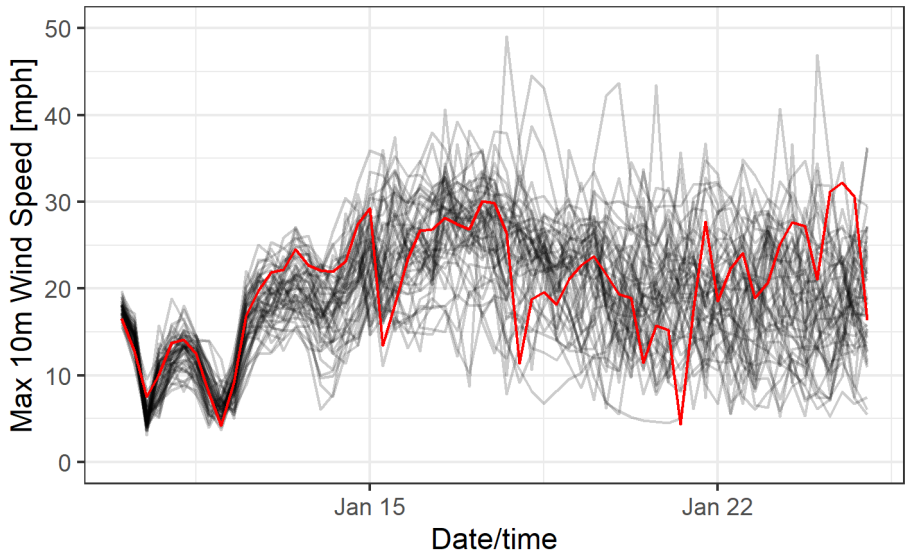

Figure 5: Example ensemble NWP forecast of max wind speed for the Borders issued on 10/01/2018 (50 members, black) and the ERA5 "truth" (red). The spread of the ensemble members quantifies forecast uncertainty, which grows the further into the future we look, eventually converging on the long-run average behaviour or "climatology".

The ensemble is said to be unbiased if the average difference between the ensemble mean and observation is zero, and this should be the case across the whole forecast horizon. If necessary, biases may be corrected by modelling and subtracting the bias from each ensemble member. Here, linear regression models are used to correct biases in wind speed features.

Secondly, the ensemble dispersion/spread should match the ensemble skill. Here we compare the Root Mean Squared Error of the ensemble mean to the standard deviation of the ensemble members. The two should match, and if not, a correction can be made. Here, it was found that after bias correction the ensemble skill-spread relationship was close to nominal, so no dispersion correction has been applied.

The raw and calibrated ensemble forecasts of max wind speed in North Wales are presented in Figure 6. The raw forecasts contain a large negative bias with a diurnal structure. As a result of the bias, the RMSE of the ensemble mean is much greater than the standard deviation of ensemble members. A linear regression model has been used to model and remove the bias. The bias of the calibrated forecast is approximately zero across the whole forecast horizon, and the RMSE matches ensemble standard deviation well, although the forecast is slightly under-dispersed for lead-times of 8 to 15 days.

Equivalent plots for all districts are provided in Figure 7 and Figure 8 which illustrate that ensemble forecasts of max wind speed can be effectively calibrated with simple bias correction, however, some districts may benefit from a more refined approach in future phases of the project. Calibration of wind direction is not considered here but is not expected to have a large impact on fault prediction.

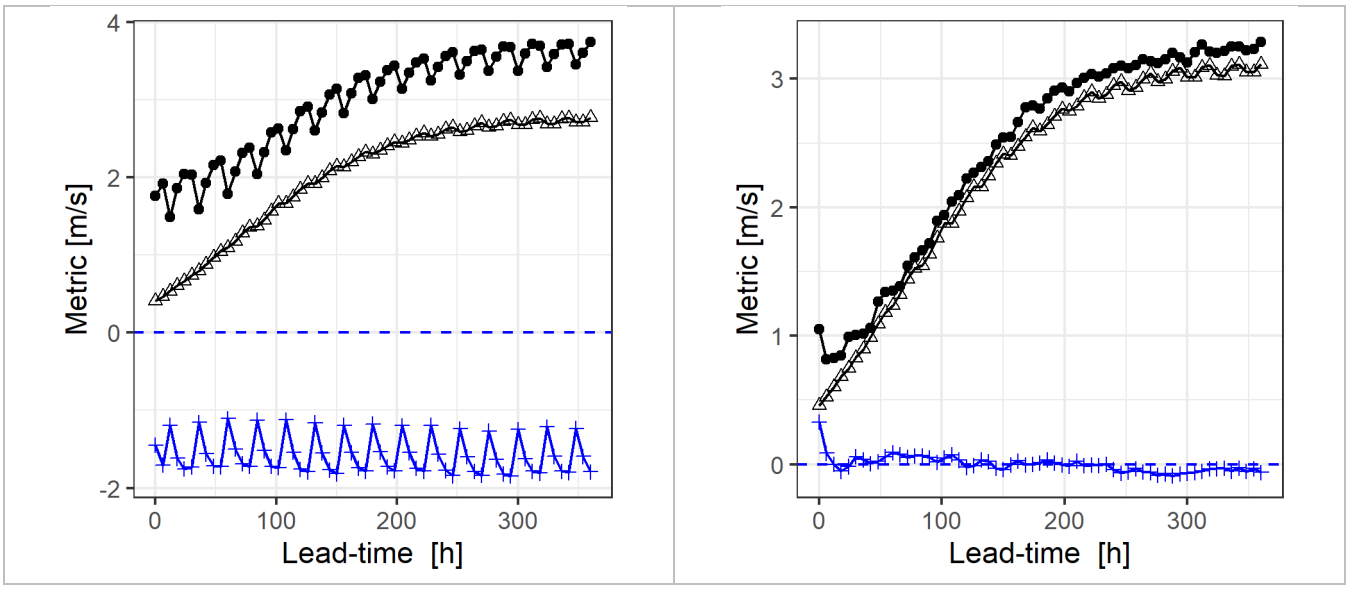

Figure 6: Bias (blue) and skill-spread (Δ=std.dev., •=RMSE) for the ECMWF ensemble predictions of max wind speed for North Wales before (left) and after (right) bias correction.

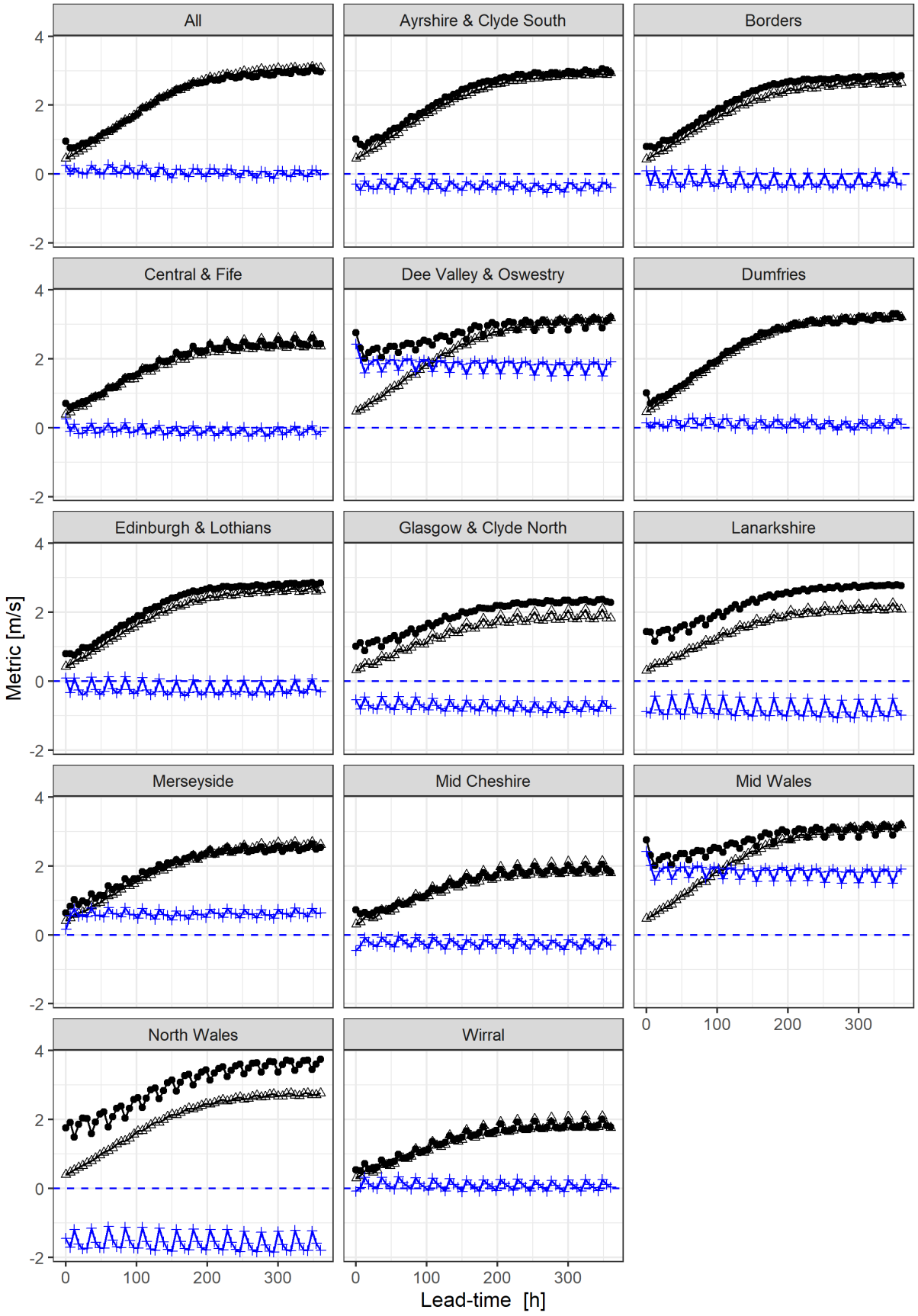

Figure 7: Verification of raw ensemble NWP for max wind speed by district. Bias (blue) and skill-spread (Δ=std.dev., •=RMSE). "All" is a domain containing all SPD and SPM districts.

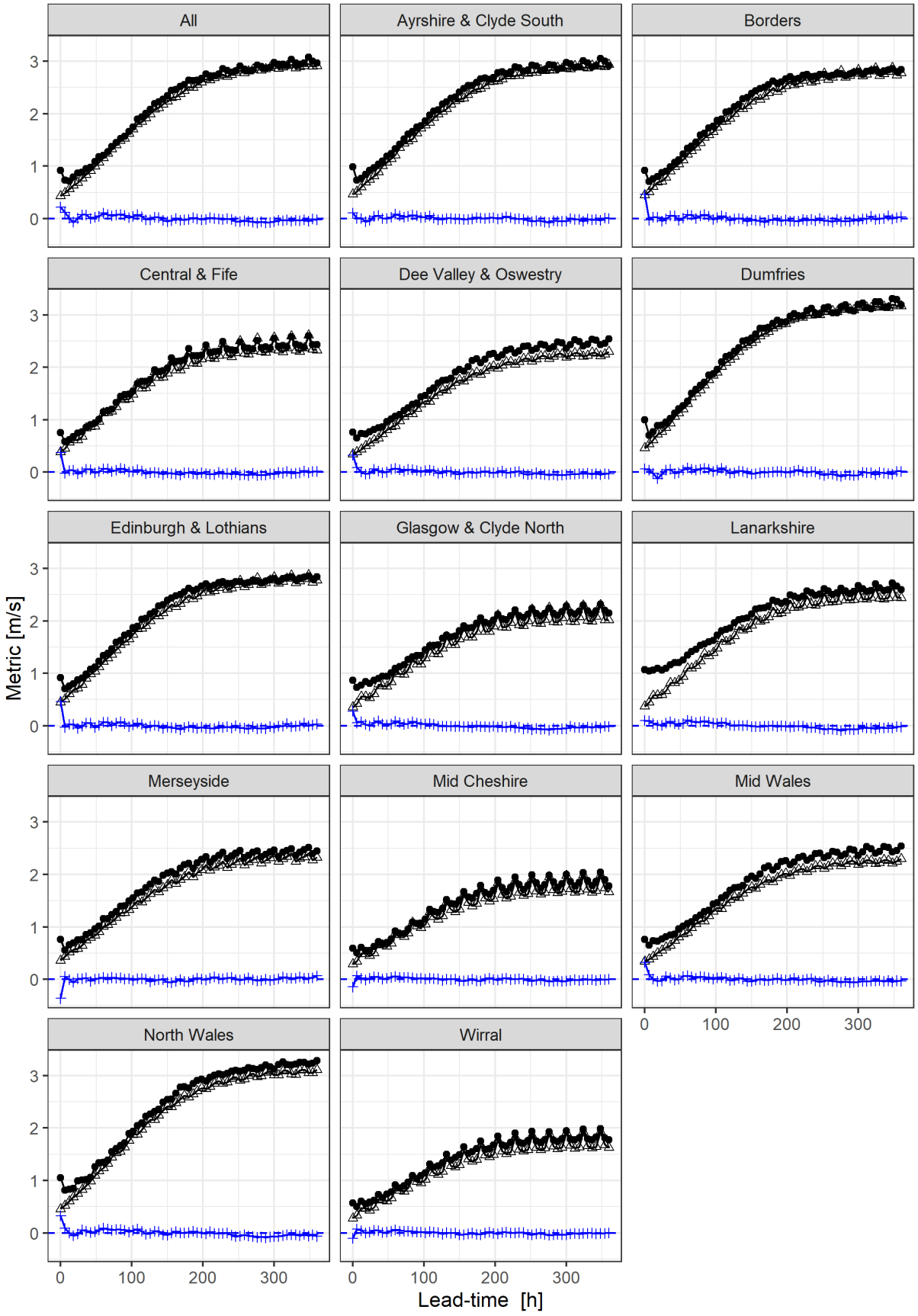

Figure 8: Verification of calibrated ensemble NWP for max wind speed by district. Bias (blue) and skill-spread (Δ=std.dev., •=RMSE). "All" is a domain containing all SPD and SPM districts.

4.3. Fault forecasting

Fault forecasts are produced by using the predicted (and calibrated) weather features as inputs to the weather-fault models described above. Forecasts are produced for each district from 0 to 14 days-ahead. ECMWF weather forecasts are published by 06:40 UTC on day 0, and the time required to convert these into fault forecasts is negligible.

Each prediction is the combination of outputs from the fault model for each of the 50 members of the ECMWF ensemble forecast. For the purpose of this proof-of-concept, we take a simple mixture model approach (similar to Bayesian Model Averaging, see associated Met Office report). However, this may mask some potentially valuable information, particularly regarding the possibility of extreme events, and should be re-visited in future phases of the project.

The weather and fault forecasts are provided in an R data file along with this report along with the R code used to produce it and the following illustrations. An example of a 0-14 day-ahead fault forecast is given in Figure 9 for North Wales in winter 2018. It is presented alongside the ensemble wind speed forecast to provide context for the resulting fault predictions.

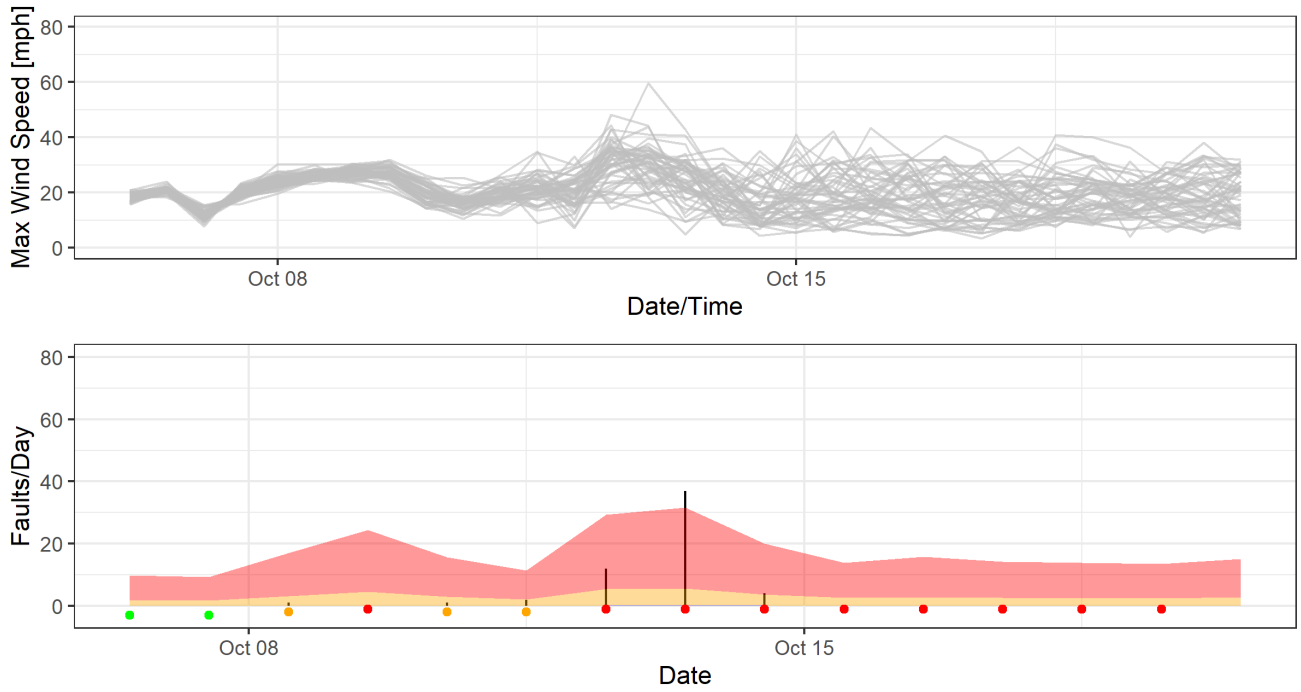

Figure 9: Forecast issued for North Wales on 2018-10-06. The 50-member ensemble forecast of max wind speed (top) and corresponding fault forecast (bottom) from 0 to 14 days-ahead. The wind speed ensemble converges towards climatology by approximately 10 days-ahead, but in this example still provides sufficient signal to provide warning of fault events 6, 7 and 8 days ahead. The traffic-light indicator predicts a >50% of faults in the latter half of the forecast horizon, which reflects climatology for the time of year.

To visually inspect forecast performance, predictions of daily faults are visualised in the same manner as Figure 4 for each district and lead-time. These are similarly provided as interactive HTML plots, extracts form which are shown below in Figure 10 for North Wales. Day-ahead forecasts are comparable to the predictions based on actual weather in Figure 4, and are specific in terms of timing of large fault events and confidently predicting no faults. The five-day-ahead forecasts are comparable, but the specificity of forecasts decreases a little as weather uncertainty has increased. This suggests that the weather-to-fault relationship is a greater source of uncertainty than that of the short-term weather forecast.

By 14-days-ahead, the forecast is dominated by average conditions for the time of year, but there are still some signals that the risk of large fault events may be higher or lower than average in the ten- and 14-day-ahead forecasts.

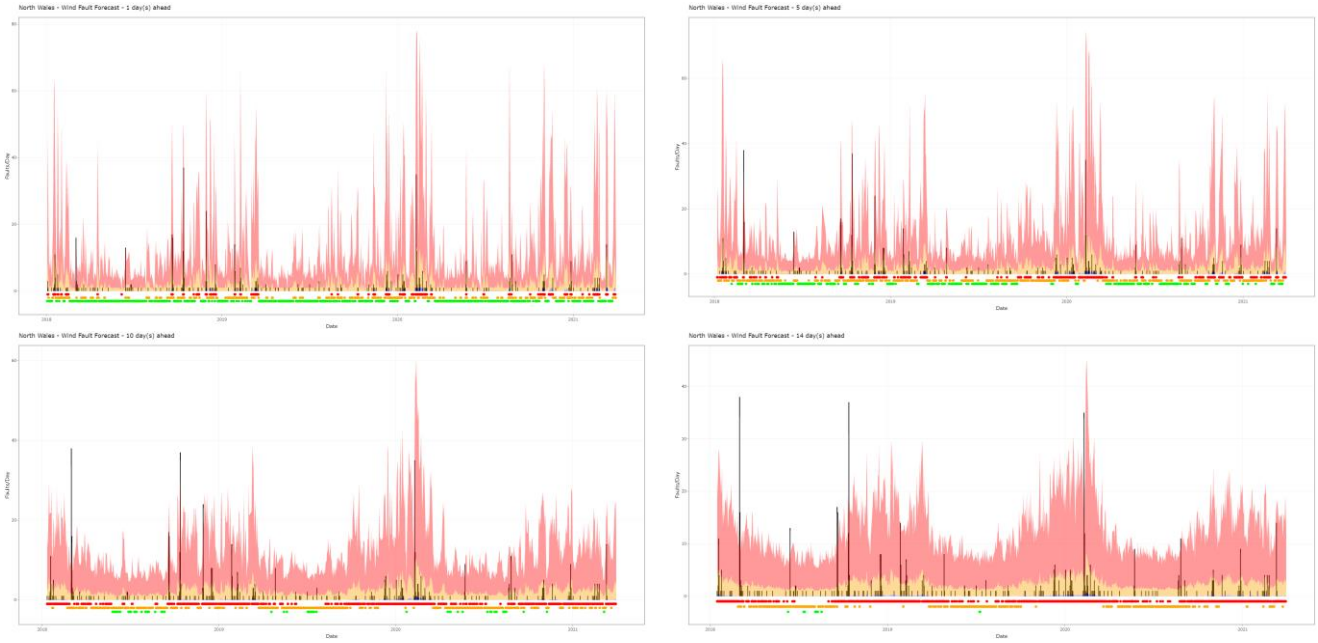

Figure 10: Fault forecasts for North Wales for 2018 to mid-2021. Day-ahead (top-left), five-days-ahead (top-right), ten-days-ahead (bottom-left) and 14—days-ahead (bottom-right) are illustrated for the whole period. The key is the same as that for Figure 4. Day-ahead forecasts are the most specific in terms of timing of large fault events (red spikes) and confidently predicting no faults (green). By 14-days-ahead, the forecast contains little specific information beyond average conditions for the time of year.

Finally, we evaluate forecast performance quantitatively. First, considering the categorical aspect of the forecast – the forecast probability that there will be zero faults given by the parameter σ_t in Equation (2). The Receiver Operating Characteristic (ROC) provides a framework for this analysis, specifically we'll look at associated metric Area Under the [ROC] Curve (AUC). This provides a measure of forecast performance, where 1 is a prefect forecast which predicts 100% chance of occurrence when the event does occur, and 0% when it doesn't. Random guesses with no skill produce an AUC of 0.5, and an AUC of 0 indicates perfectly incorrect forecasts, i.e., predicting 100% chance of an event when the event does not happen and vice versa. The AUC has been calculated by District and lead-time and visualised in Figure 11.

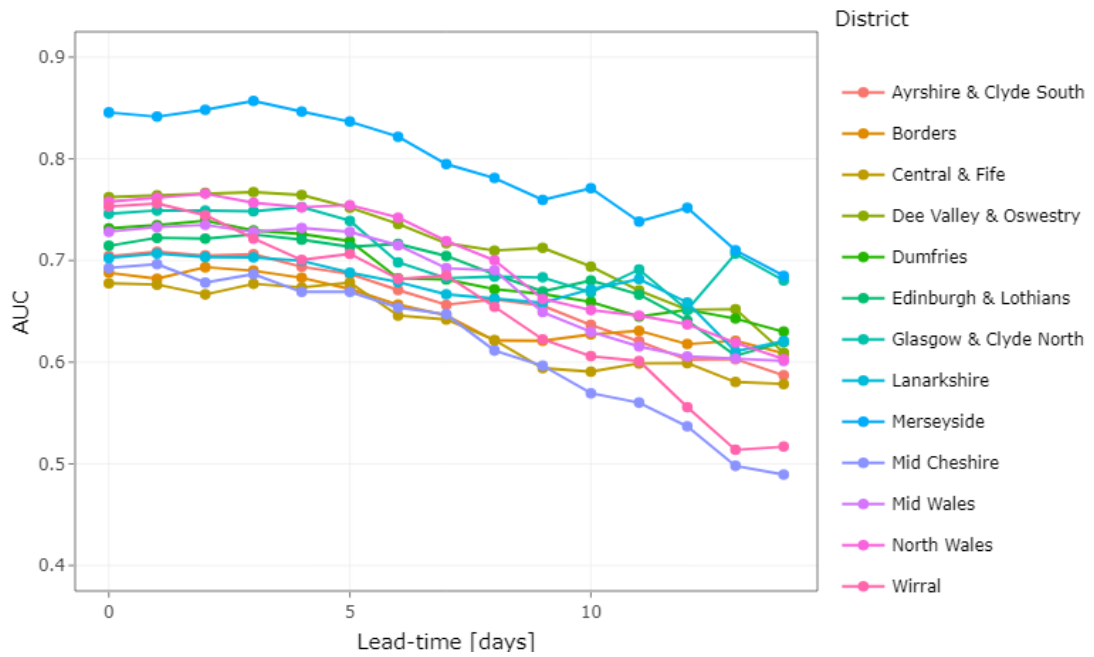

Figure 11: Area Under the [ROC] Curve by District and lead-time. AUC=1 is perfect performance, AUC=0.5 is no skill. By this metric, performance is affected by the number of faults there are to predict, so comparison between Districts is difficult to interpret. However, all districts show consistent forecast performance from zero to five-days-ahead, beyond which performance declines.

By this metric, performance is affected by the number of faults there are to predict, so comparison between Districts is difficult to interpret. However, all districts show consistent forecast performance from zero to five-days-ahead, beyond which performance declines. All districts apart from Mid Cheshire and Wirral retain some skill up to and including 14-days-ahead.

We are also interested in predictions of large numbers of faults. While there is always significant uncertainty on precise number of faults, potential users have indicated that their primary concern is the risk associated with some threshold being exceeded. Therefore, we will consider a threshold of five faults in a day being exceeded with a probability of at least 5% in Dumfries and North Wales. These Districts are chosen as faults are most prevalent in these districts, as are days with large numbers of faults.

First, consider the contingency tables for these forecasts one day-ahead below. Most events are correctly predicted, although results are consistent with forecasts underestimating the probability of large numbers of faults occurring, as previously identified in Section 4.1. The high number of false positives is to be expected given that the threshold is set at the 5% level. Increasing this threshold would decrease the number of false positive, but increase the number of false negatives, or "misses".

Table 3: Contingency tables for predicting the occurrence of five or more faults in a given day one day ahead with probability greater than 5%.

North Wales		OBSERVED		
		YES	**NO**	*TOTAL*
FORECAST	**YES**	10	92	*102*
	NO	2	1028	*1030*
	TOTAL	*12*	*1120*	

Dumfries		OBSERVED		
		YES	**NO**	*TOTAL*
FORECAST	**YES**	20	129	*149*
	NO	2	981	*983*
	TOTAL	*22*	*1110*	

We can examine how the True Positive Rate (TPR, proportion of high-fault days correctly predicted) and False Positive Rate (FPR, proportion of high-fault predictions that were not high fault days) evolve over the forecast horizon. The TPR remains high for the lead-times of zero to four or five days ahead, and then decreases rapidly. The FPR is 8-12% from zero to five-days-ahead and then also decreases. The FPR is high given the number days there are fewer than five faults.

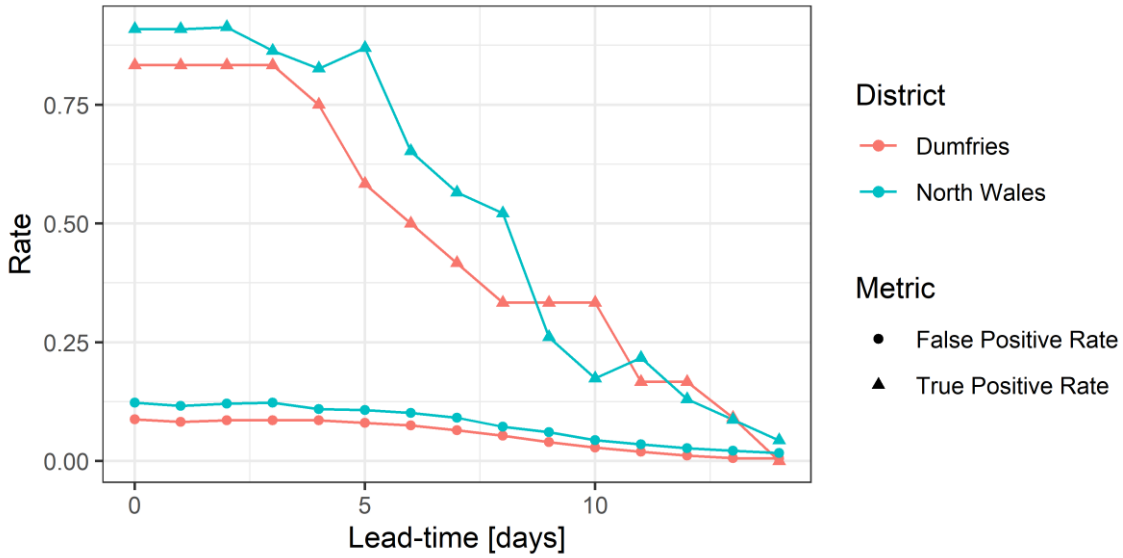

Figure 12: Performance of categorical forecasts of high-fault days (5 or more faults) for Dumfries and North Wales by lead-time based on period from 2018 to mid-2021.

Quantitative evaluation of fault forecasts has verified that it is possible to skilfully predict wind-related faults, with performance greatest on lead-times from zero to five days-ahead and declining thereafter. It was also observed that ensemble NWP resulted in improved performance compared to deterministic NWP (single member) for all lead-times (not shown). The extent to which this skill may be further improved is difficult to estimate, but improvement is most likely possible given the use of limited explanatory data and time available for statistical model development. Also unknown is how forecast skill will translate into business value. In addition to improving performance according to statistical metrics, the aspects of performance that are most valuable to users should be identified and be the target of future refinements and forecast evaluation.

5. Conclusions and suggestions for future work

This report has reviewed existing literature on weather-related fault prediction in electricity networks and performed exploratory analysis on data from SPEN's electricity distribution networks. Both activities indicate that fault prediction is possible, and that uncertainties originating from both the weather-to-fault relationship and weather forecasting play a role in the performance of fault forecasts. Based on preliminary analysis, the uncertainty associated with the weather-to-fault relationship dominates on lead-times of zero to five days-ahead. It is anticipated that fault-prediction will be improved by incorporation of additional explanatory data and refined statistical modelling. There is still skill in fault forecasts up to 14 days-ahead, although performance decreases rapidly beyond one week-ahead with that of the underlying weather forecast.

Based on the exploratory analysis, the aim of which has been proof-of-concept, the following have been identified as potential areas for improvement:

Weather Fault Model:

- Incorporation of additional explanatory data:
    - Extracting weather data from higher-resolution NWP (reanalysis, forecasts, observations), and in a more targeted fashion, such as at locations of assets and more precise timing of faults
    - A physics-inspired approach to feature engineering and modelling, e.g. engineering features based on asset fragility and/or vulnerability based on vegetation, elevation/topography, ground conditions (for precipitation-related faults) or directional (relative to wind direction) characteristics
- Refinements to the statistical modelling framework
    - Exploration of other, potentially more flexible parametric distributions, such as the Discrete Generalised Pareto Distribution (Hitz, Davis, and Samorodnitsky 2017)
    - Alternative estimation frameworks (beyond gamlss) for extremes should be considered, such as Bayesian estimation
    - Other model variations should be explored, including a time varying threshold in the non-stationary peaks-over-threshold setting
    - Treatment of short-term correlation should be considered in statistical analysis and parameter estimation
    - If many candidate features/explanatory variables are generated, regularisation may be employed to automatically select those to retain in the model

Ensemble NWP Calibration:

- Here, basic calibration appears sufficient, but more detailed verification should be carried out.
- Other ensemble NWP products, such as MOGREPS-UK may require different levels of calibration.
- Methods for multi-variate calibration should be considered if multiple weather parameters are used in the weather fault model.

Fault Forecasting:

- Above, the final predictive distribution for each forecast is a simple combination of the predictive distribution implied by individual ensemble members, similar to kernel dressing or Bayesian Model Averaging methods. Refinements and alternative methods should be considered and evaluated.
- There are many ways these forecasts can be evaluated. Some, such as goodness-of-fit are necessary to verify statistical methods are suitable for the task, while others should be chosen to reflect how forecasts will be used. An evaluation framework should be developed with this in mind to ensure future developments are targeted to maximise value from fault forecasts.

Appendix: Supplementary Data

The sample dataset and predictions produced by the proof-of-concept methodology are supplied along with this report in a zip file called "UofG Data.zip". This file contains the following folders, details of which are provided here.

Weather-Fault Plots contains plots of number of faults by wind speed and direction for all distracts, as in Figure 2 of this report.

Weather-Fault Model contains interactive plots (html files) of the weather-fault model (based on ERA5 reanalysis), as in Figure 4, for all districts from 2010-2021.

Fault Forecasts contains interactive plots (html files) of weather-fault forecasts (based on ECMWF ensemble weather predictions), as in Figure 10, for all districts from 2018-2021 for lead-times 1, 3, 5, 7, 10 and 14 days-ahead. The lead-time is indicated in the file name.

The raw EAR5 and ECMWF ensemble NWP are in the "Weather Data" folder on the project SharePoint "Predict4Resilience WIP" owned by Michael Eves (SPEN). These data are in netCDT format. Meta-data on parameters, units, spatial grid, and so on, is included in the files and may be accessed with any netCDF reader. NB: downloading them from SharePoint is a hassle as download of more than 4GB results in corrupt zip files. It is necessary to use a syncing tool, download in batches, of contact Jethro Browell to discuss alternatives. Licence: ERA5 is governed by the Licence to Use Copernicus Products[5], historic ECMWF forecasts are governed by the CC-BY 4.0 licence[6].

[5] https://cds.climate.copernicus.eu/api/v2/terms/static/licence-to-use-copernicus-products.pdf

[6] https://www.ecmwf.int/en/forecasts/accessing-forecasts/licences-available

References

Alpay, Berk A., David Wanik, Peter Watson, Diego Cerrai, Guannan Liang, and Emmanouil Anagnostou. 2020. "Dynamic Modeling of Power Outages Caused by Thunderstorms." *Forecasting* 2 (2): 151–62. https://doi.org/10.3390/forecast2020008.

Brayshaw, David J., Alan Halford, Stefan Smith, and Kjeld Jensen. 2020. "Quantifying the Potential for Improved Management of Weather Risk Using Sub-seasonal Forecasting: The Case of UK Telecommunications Infrastructure." *Meteorological Applications* 27 (1). https://doi.org/10.1002/met.1849.

Brester, Christina, Harri Niska, Robert Ciszek, and Mikko Kolehmainen. 2020. "Weather-Based Fault Prediction in Electricity Networks with Artificial Neural Networks." In *2020 IEEE Congress on Evolutionary Computation (CEC)*, 1–8. IEEE. https://doi.org/10.1109/CEC48606.2020.9185555.

Buuren, Stef van, and Miranda Fredriks. 2001. "Worm Plot: A Simple Diagnostic Device for Modelling Growth Reference Curves." *Statistics in Medicine* 20 (8): 1259–77.

Cerrai, Diego, David W. Wanik, Md Abul Ehsan Bhuiyan, Xinxuan Zhang, Jaemo Yang, Maria E. B. Frediani, and Emmanouil N. Anagnostou. 2019. "Predicting Storm Outages Through New Representations of Weather and Vegetation." *IEEE Access* 7: 29639–54. https://doi.org/10.1109/ACCESS.2019.2902558.

Hitz, Adrien, Richard Davis, and Gennady Samorodnitsky. 2017. "Discrete Extremes," July.

"IBM Environmental Intelligence Suite - Outage Prediction | IBM." n.d. Accessed April 13, 2022. https://www.ibm.com/uk-en/products/environmental-intelligence-suite/outage-prediction.

Rigby, R A, and D M Stasinopoulos. 2005. "Generalized Additive Models for Location, Scale and Shape." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 54 (3): 507–54.

Tsioumpri, Eleni, Bruce Stephen, and Stephen D. J. McArthur. 2021. "Weather Related Fault Prediction in Minimally Monitored Distribution Networks." *Energies* 14 (8): 2053. https://doi.org/10.3390/en14082053.

UKPN. 2016. "Prediction of Weather-Related Faults (NIA Closedown Report NIA_UKPN0006)."

Watson, Peter L., Diego Cerrai, Marika Koukoula, David W. Wanik, and Emmanouil Anagnostou. 2020. "Weather-related Power Outage Model with a Growing Domain: Structure, Performance, and Generalisability." *The Journal of Engineering* 2020 (10): 817–26. https://doi.org/10.1049/joe.2019.1274.

Wilkinson, Sean, Sarah Dunn, Russell Adams, Nicolas Kirchner-Bossi, Hayley J. Fowler, Samuel González Otálora, David Pritchard, Joana Mendes, Erika J. Palin, and Steven C. Chan. 2022. "Consequence Forecasting: A Rational Framework for Predicting the Consequences of Approaching Storms." *Climate Risk Management* 35: 100412. https://doi.org/10.1016/j.crm.2022.100412.